Quotient approximation for schoolbook division

Niels Möller

October 8, 2021

Abstract

A new more efficient way to compute an approximative quotient suitable for school book division of multi-precision integers.

1 Background

This section gives a short overview of schoolbook division history, as it has been applied to the GMP [1] library.

Knuth, 1969

The classic description of schoolbook division is Knuth's, see [4, Sec. 4.3.1, Alg. D]. It works as follows.

To compute the most significant quotient word, start by dividing the two most significant words of the numerator with the most significant word of the divisor (the later is assumed normalized, i.e., most significant bit set). Next, take one more word into account of both numerator and divisor, to check if the approximation is a correct quotient for dividing the three most significant numerator words with the two most significant. If it isn't, it's at most two units too large, and is adjusted accordingly.

After these preliminary adjustments, the quotient word is usually correct, with a small probability of it being one too large. So go ahead and compute the full multi-precision remainder; check for the unlikely underflow, and do a final adjustment of both the quotient word and the multi-precision remainder when underflow happens.

Granlund-Montgomery, 1994

When computing one quotient word at a time, to produce a multi-precision quotient, the numerator is updated incrementally by subtracting multiples of the divisor, but the divisor itself is unchanged; it is a loop invariant. One can therefore speed up the computation of the quotient approximation by precomputing an approximate reciprocal of the most significant divisor word. The initial divison is then replaced by a few multiplications and adjustments, which is a big win since division instructions are usually vastly slower than multiplication instructions[1] The paper [3] describes one clever way to make use of a precomputed

[1]For example, on Intel Skylake processors, a 128-bit by 64-bit division takes 86 cycles, while a 64-bit by 64-bit multiply with 128-bit result takes 3 cycles, and in addition, multiplication is pipelined so that a new multiply can be started every each cycle, see [2].

reciprocal in schoolbook division, even though schoolbook division was not the main topic of that paper.

Möller-Granlund, 2011

The main idea of the next paper, when applied to schoolbook division, is to use a reciprocal based on the two most significant divisor words. The reciprocal is still a single word, but we can simplify the adjustment steps needed for each quotient word, by using a slightly different reciprocal. In effect, moving some of the adjustment work out of the loop and doing it as part of the precomputation of the reciprocal. See [5].

This algorithm divides the three most significant words of the numerator with the two most significant words of the divisor, producing the same candidate quotient as used in Knuth, but with simpler adjustment steps than earlier methods.

Current work, 2018

The algorithms above all produce a correct quotient of three words by two. When this is ready, it is applied to compute the full multi-precision remainder. We need a final adjustment in the unlikely case that the computation of the full remainder underflows.

Further improvement is based on two observations. First, since we do have a final adjustment step, we don't need a three-by-two quotient that is correct in all cases. Second, the influence on the correct quotient from the third most significant word of the numerator is very small.

We therefore aim to compute a candidate quotient based on the two most significant words of both numerator and divisor. When applied to compute the multi-precision remainder, it must be either correct or one too large, and the probability of error should be small. The resulting algorithm features simpler adjustment steps than the earlier methods.

2 Notation and requirements

Let ℓ denote the computer word size, and let $\beta = 2^\ell$ denote the base implied by the word size. **(FIXME: ℓ mustn't be too small. I think these arguments should work wor $\ell \geq 2$, i.e., $\beta \geq 4$.)** Lower-case letters denote single-word numbers, and upper-case letters represent numbers of any size. We use the notation $X = \langle x_{n-1}, \ldots, x_1, x_0 \rangle = x_{n-1}\beta^{n-1} + \cdots + x_1\beta + x_0$, where the n-word integer X is represented by the words x_i, for $0 \leq i < n$.

We consider only one iteration of the schoolbook division algorithm, computing a single quotient word; organizing the outer loop is out of scope for these notes. Let the divisor $D = \langle d_{n-1}, \ldots d_0 \rangle$ consist of $n > 2$ words, and the numerator $U = \langle u_n, \ldots n_0 \rangle$ consist of $n + 1$ words.

We assume that $U < \beta D$, so that the correct quotient $\lfloor U/D \rfloor$ is a single word, and that $d_{n-1} \geq \beta/2$ (normalization).

We need a function `divappr`, that lets us compute a candidate quotient $q \leftarrow \texttt{divappr}(\langle u_n, u_{n-1} \rangle, \langle d_{n-1}, d_{n-2} \rangle)$.

3 The `divappr` function

We define a function $q \leftarrow \texttt{divappr}(\langle u_1, u_0\rangle, \langle d_1, d_0\rangle)$, with the following requirements. Input consists of four single-word numbers. We require that $d_1 \geq \beta/2$ and $\langle u_1, u_0\rangle \leq \langle d_1, d_0\rangle$. The output q is also a single word. Let R denote the corresponding remainder

$$R = \langle u_1, u_0, 0\rangle - q\langle d_1, d_0\rangle$$

In the borderline case $\langle u_1, u_0\rangle = \langle d_1, d_0\rangle$, `divappr` must produce $q = \beta - 1$. This corresponds to $R = \langle d_1, d_0\rangle$, and in the context of schoolbook divison, $q = \beta - 1$ is the correct quotient, thanks to the requirement that $U < \beta D$.

When $\langle u_1, u_0\rangle < \langle d_1, d_0\rangle$, we require R to belong to the range

$$-2\beta < R \leq \begin{cases} \langle d_1, d_0\rangle - 1 & \text{if } q = \beta - 1 \\ \langle d_1 - 1, d_0\rangle & \text{otherwise} \end{cases}$$

When q is applied to the full multi-precision numbers, the corresponding remainder

$$R_{\text{bignum}} = U - qD$$

satisfies $R_{\text{bignum}} < D$ and

$$R_{\text{bignum}} > -3\beta^{n-1} \geq -\frac{6}{\beta}D$$

ensuring that $R < 0$ is unlikely for random inputs.

To compute `divappr`, we will make use of the same approximate reciprocal as for three-by-two division, defined as

$$v = \lfloor(\beta^3 - 1)/\langle d_1, d_0\rangle\rfloor - \beta$$

4 The algorithm

We now present an alrorithm that computes an approximation $q \approx \langle u_1, u_0, 0\rangle/\langle d_1, d_0\rangle$. When we say that q is the *correct* quotient, we mean that $q = \lfloor\langle u_1, u_0, 0\rangle/\langle d_1, d_0\rangle\rfloor$. A correct quotient in this sense doesn't necessarily imply that it's a quotient that the algorithm should return, nor that its application to schoolbook division won't need any final adjustment.

```
q ← DIVAPPR2(⟨u1, u0⟩, ⟨d1, d0⟩)
     In: β/2 ≤ d1 < β, ⟨u1, u1⟩ ≤ ⟨d1, d0⟩,
 1   v = ⌊(β^3 − 1)/⟨d1, d0⟩⌋ − β                    // Should be precomputed
 2   if ⟨u1, u0⟩ ≥ ⟨d1, d0⟩ − d1
 3       return β − 1
 4   ⟨q1, q0⟩ ← vu1 + ⟨u1, u0⟩
 5   q ← q1 + 1
 6   p1 ← ⌊qd0/β⌋
 7   r ← (u0 − qd1 − p1 − 1) mod β
 8   if r ≥ q0
 9       q ← (q − 1) mod β
10       r ← (r + d1 + 1) mod β
11   if r ≥ d1 − 1
12       q ← (q + 1) mod β
13   return q
```

Let us state the desired properties of the return value in the form of a theorem.

Theorem 1 *Assume that $d_1 \geq \beta/2$ and $\langle u_1, u_0\rangle \leq \langle d_1, d_0\rangle$. Let q be the return value of the* ***divappr*** *algorithm, and let R be the corresponding remainder,*

$$R = \langle u_1, u_0, 0\rangle - q\langle d_1, d_0\rangle$$

Then the following holds:

1. *If $\langle u_1, u_0\rangle = \langle d_1, d_0\rangle$, then $q = \beta - 1$ and $R = \langle d_1, d_0\rangle$.*
2. *Otherwise, if $q = \beta - 1$, then $-2\beta < R < \langle d_1, d_0\rangle$*
3. *If $q < \beta - 1$, then $-2\beta < R \leq \langle d_1 - 1, d_0\rangle$.*

To prove this theorem, we need to consider several different cases. Let us start with some preliminaries.

First check what happens when $u_1 = u_0 = 0$. We then get $q_0 = q_1 = p_1 = 0$. We get the initial quotient candidate $q \leftarrow 1$ on line 5. On line 7 we get $r \leftarrow (-d_1 - 1) \bmod \beta$, and trivially $r \geq q_0$. Hence the first adjustment is applied, and the adjusted r on line 10 is zero. The second adjustment isn't applied, and the returned quotient is $q = 0$, corresponding to $R = 0$, which is perfectly right. In the following, let us therefore assume that $\langle u_1, u_0\rangle > 0$.

Since

$$\langle d_1, d_0\rangle(\beta - 1) = \beta^2 d_1 + \beta(d_0 - d_1) - d_0$$

we have $\lfloor\langle u_1, u_0, 0\rangle/\langle d_1, d_0\rangle\rfloor \geq \beta - 1$ if and only if $\langle u_1, u_0\rangle \geq \langle d_1, d_0\rangle - d_1$. This is the condition on line 2, and it follows that we return $q = \beta - 1$ for all inputs where it's the correct quotient, and in the borderline case $\langle u_1, u_0\rangle = \langle d_1, d_0\rangle$. Hence, when the algorithm terminates at line 3, the return value satisfies the theorem, either case 1 or 2.

So let us assume that $\langle u_1, u_0\rangle < \langle d_1, d_0\rangle - d_1$; then the correct quotient is at most $\beta - 2$. This ensures that in the cases that we return a quotient that is one too large (i.e., $R < 0$), that incorrect quotient still fits in one word.

The value q_1 is always upper bounded by the correct quotient (since the reciprocal v is rounded down). Define $\tilde{q} = q_1 + 1$, which also fits in single word.

This is the initial candidate quotient, computed on line 5. Let $\tilde{R}$ denote the corresponding remainder,

$$\tilde{R} = \langle u_1, u_0, 0\rangle - (q_1 + 1)\langle d_1, d_0\rangle$$

We first prove upper and lower bounds for this quantity.

Lemma 2 *Assume that* $0 < \langle u_1, u_0\rangle < \langle d_1, d_0\rangle - d_1$. *Then* $\tilde{R}$ *is bounded as follows:*

$$\tilde{R} > -D \tag{1}$$

$$\tilde{R} > q_0\beta - \beta^2 \tag{2}$$

$$\tilde{R} < \max(\beta^2 - D, q_0\beta) - \beta \tag{3}$$

Proof: We follow the analysis of three-by-two division in [5, Theorem 3] closely, but we can get slightly tighter bounds since (i) the third most significant numerator word is zero, and (ii) we take care of the largest possible values of $\langle u_1, u_0\rangle$ separately.

The definition of v implies that $(\beta + v)\langle d_1, d_0\rangle = \beta^3 - K$, for some K in the range $1 \le K \le \langle d_1, d_0\rangle$. In this proof, also use the notation $D = \langle d_1, d_0\rangle$. Substitution into the expression for $\tilde{R}$ gives

$$\tilde{R} = \frac{u_1 K + u_0(\beta^2 - D) + q_0 D}{\beta} - D$$

Since all terms but the last are non-negative, and at least one of the terms involving u_1 and u_0 is non-zero, Eq. (1) follows immediately. If we keep the term depending on q_0, we get Eq. (2) as

$$\tilde{R} > \frac{q_0 D}{\beta} - D = -D\left(1 - \frac{q_0}{\beta}\right) > -\beta^2\left(1 - \frac{q_0}{\beta}\right) = q_0\beta - \beta^2$$

To derive the upper bound is a bit more involved. Recall that we assume that $\langle u_1, u_0\rangle < D - d_1 \le D - \beta/2$, which implies that $u_1 < (D - u_0 - \beta/2)/\beta$. We then get

$$\begin{aligned}\tilde{R} &< \frac{(D - u_0 - \beta/2)D}{\beta^2} + \frac{u_0(\beta^2 - D)}{\beta} + \frac{q_0 D}{\beta} - D \\ &= \frac{D^2}{\beta^2} + \frac{u_0(\beta^3 - \beta D - D)}{\beta^2} + \frac{q_0 D}{\beta} - D - \frac{D}{2\beta}\end{aligned}$$

For a moment, assume that $\beta^3 - \beta D - D \ge 0$. It then follows that

$$\begin{aligned}\tilde{R} &< \frac{D^2}{\beta^2} + \frac{(\beta - 1)(\beta^3 - \beta D - D)}{\beta^2} + \frac{q_0 D}{\beta} - D - \frac{D}{2\beta} \\ &= \frac{D^2}{\beta^2} + \beta^2 - 2D - \beta + \frac{D}{\beta^2} + \frac{q_0 D}{\beta} - \frac{D}{2\beta} \\ &= (\beta^2 - D)\left(1 - \frac{D}{\beta^2}\right) + q_0\beta\frac{D}{\beta^2} - \beta - \frac{(\beta - 2)D}{2\beta^2} \\ &< \max(\beta^2 - D, q_0\beta) - \beta\end{aligned}$$

where the final inequality follows from recognising the expression as a convex combination.

Finally, assume that $\beta^3 - \beta D - D < 0$. This implies that $D \geq \beta^2 - \beta + 1$, or $d_1 = \beta - 1$ and $d_0 \geq 1$. It follows that $u_1 \leq \beta - 2$, $v = 0$, $\tilde{q} = u_1 + 1$, and $q_0 = u_0$. We then get

$$\begin{aligned}
\tilde{R} &= \langle u_1, u_0, 0\rangle - \tilde{q}\langle d_1, d_0\rangle \\
&= \beta^2 u_1 + \beta q_0 - (u_1 + 1)(\beta^2 - \beta + d_0) \\
&= \beta q_0 + (u_1 + 1)(\beta - d_0) - \beta^2 \\
&\leq \beta q_0 + (\beta - 1)(\beta - d_0) - \beta^2 \\
&= \beta q_0 - \beta - (\beta - 1)d_0 \\
&\leq \beta q_0 - (2\beta - 1) < \beta q_0 - \beta
\end{aligned}$$

The final expression is smaller than the bound in Eq (3), and this concludes the proof of this lemma. □

This lemma also implies the bounds

$$-\beta^2 + 1 < \tilde{R} < \beta^2 - \beta$$

After these preliminaries, let us complete the proof of Theorem 3.

Proof: Let $\tilde{r}$ denote the value computed on line 7. Let p_0 denote the low half $p_0 = qd_0 \bmod \beta$, then we have

$$\tilde{R} = \beta\left[\langle u_1, u_0\rangle - (q_1 + 1)d_1 - p_1\right] - p_0$$

We see that $\tilde{r}$ is second least significant word of R, except that is is one too small when $p_0 = 0$. This can be expressed as

$$(\tilde{R} - 1) \bmod \beta^2 = \langle \tilde{r}, \beta - 1 - p_0\rangle$$

First, consider the case that $\tilde{R} \leq 0$. Then $\tilde{R} = \beta\tilde{r} + \beta - p_0 - \beta^2$, and the lower bound, Eq. (2), implies

$$\langle \tilde{r}, \beta - 1 - p_0\rangle = \tilde{R} + \beta^2 - 1 \geq q_0\beta$$

Hence, $\tilde{r} \geq q_0$, and so the first adjustment condition applies.

The other lower bound, Eq. (1), implies that

$$\beta\tilde{r} + \beta - p_0 = \tilde{R} + \beta^2 \geq \beta^2 - \langle d_1, d_0\rangle$$

It follows that

$$\tilde{r} + d_1 + 1 \geq \beta - \frac{d_0}{\beta}$$

Since the left hand side is an integer, and $d_0 < \beta$, it follows that $r + d_1 + 1 \geq \beta$.

Let r' denote the value after adjustment on line 10, it's

$$r' = \tilde{r} + d_1 + 1 \bmod \beta = \tilde{r} + d_1 + 1 - \beta$$

The corresponding two-word remainder is

$$\begin{aligned}
\tilde{R} + \langle d_1, d_0\rangle &= \beta\tilde{r} + \beta - p_0 - \beta^2 + \langle d_1, d_0\rangle \\
&= \beta(\tilde{r} + d_1 + 1 - \beta) - p_0 + d_0 \\
&= \beta r' - p_0 + d_0
\end{aligned}$$

If $r' \le d_1 - 2$, then this is the final remainder R, and it follows that $-\beta < R < \beta(d_1 - 2) + d_0 \le \langle d_1, d_0 \rangle - 2\beta$, which is in the desired range. On the other hand, if $r' \ge d_1 - 1$, then the second adjustment cancels the first leaving $\tilde{R}$ as the final remainder, and we find that

$$\begin{aligned}\tilde{R} &= \beta r' - p_0 + d_0 - \langle d_1, d_0 \rangle \\ &\ge \beta(d_1 - 1 - d_1) - p_0 \\ &> -2\beta\end{aligned}$$

(and still, by assumption, also $\tilde{R} \le 0$). This concludes the proof when $\tilde{R} \le 0$.

Next, consider the case $\tilde{R} > 0$. Then $\tilde{R} = \beta\tilde{r} + \beta - p_0$. First assume that $\tilde{r} < q_0$, so the first adjustment step isn't done. If $r' \le d_1 - 2$, $\tilde{R}$ is the final remainder, it it is bounded as

$$0 < \tilde{R} \le \beta(d_1 - 2) + \beta - p_0 \le \langle d_1, d_0 \rangle - \beta$$

as required. On the other hand, if $\tilde{r} \ge d_1 - 2$, the final remainder is $R = \tilde{R} - \langle d_1, d_0 \rangle$. Note that Eq (3) implies that $\tilde{R} < \beta^2 - \beta$. In addition, $d_1 \ge \beta/2$ implies that $\beta^2 - \langle d_1, d_0 \rangle \le \langle d_1, d_0 \rangle$. It follows that

$$R = \tilde{R} - \langle d_1, d_0 \rangle < \beta^2 - \beta - \langle d_1, d_0 \rangle \le \langle d_1 - 1, d_0 \rangle$$

For the lower bound, we have

$$R = \tilde{R} - \langle d_1, d_0 \rangle \ge \beta(d_1 - 1) + \beta - p_0 - \langle d_1, d_0 \rangle = -\beta + (\beta - p_0) - d_0 > -2\beta$$

Hence, in both cases

$$-2\beta < \tilde{R} \le \langle d_1, d_0 \rangle \le \langle d_1 - 1, d_0 \rangle$$

But what happens if $\tilde{r} \ge q_0$? Then the upper bound, Eq. (3) implies

$$\beta\tilde{r} + \beta - p_0 = \tilde{R} < \beta^2 - \langle d_1, d_0 \rangle - \beta = \beta(\beta - d_1 - 1) - d_0$$

It follows that $\tilde{r} < \beta - d_1 - 1$, Hence, the value after the update is

$$\tilde{r} + d_1 + 1 \bmod \beta = r + d_1 + 1 \ge d_1 - 1$$

so we get two adjustments canceling out, so in this case, $\tilde{R}$ is the final remainder. Furthermore, since we have $\beta^2 - \langle d_1, d_0 \rangle \le \langle d_1, d_0 \rangle$, the upper bound implies

$$\tilde{R} < \beta^2 - \langle d_1, d_0 \rangle - \beta \le \langle d_1 - 1, d_0 \rangle$$

□

References

[1] Torbjörn Granlund. GNU multiple precision arithmetic library. `http://gmplib.org/`.

[2] Torbjörn Granlund. Instruction latencies and throughput for AMD and Intel x86 processors, 2017. `http://gmplib.org/~tege/x86-timing.pdf`.

[3] Torbjörn Granlund and Peter L. Montgomery. Division by invariant integers using multiplication. In *Proceedings of the SIGPLAN PLDI'94 Conference*, June 1994.

[4] Donald E. Knuth. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*. Addison-Wesley, Reading, Massachusetts, third edition, 1998.

[5] Niels Möller and Torbjörn Granlund. Improved division by invariant integers. *IEEE Transactions on Computers*, 60:165–175, 2011.