Appendix to "Algorithms for triple-word arithmetic"

N. Fabiano J.M. Muller J. Picot

1 Analysis of Algorithm 13 and Proof of Theorem 9 (reciprocal of a TW)

1.1 Bound on the error

Momentarily, the computation of $\bar{i}$ is seen as a black box (see Section **??**). We have

$$\begin{aligned}\left|a-\tfrac{1}{x_0}\right| &\leq \left|a-\tfrac{1+2u}{x_0}\right|+\left|\tfrac{1+2u}{x_0}-\tfrac{1}{x_0}\right| \\ &\leq u\left|\tfrac{1+2u}{x_0}\right|+\left|\tfrac{2u}{x_0}\right| \\ &\leq (3u+2u^2)\cdot\left|\tfrac{1}{x_0}\right|,\end{aligned}$$

which implies $\left|\frac{1-3.1u}{x_0}\right| \leq |a| \leq \left|\frac{1+3.1u}{x_0}\right|$, so that $|h_1| \leq (1+u)(u\cdot|ax_0|+|a|\cdot 2u|x_0|) \leq 3u+13u^2$. Now,

$$\begin{aligned}\left|\tfrac{1}{x_0}-\tfrac{1}{x_0+x_1}\right| &= \left|\tfrac{1}{x_0+x_1}\right|\left|1-\tfrac{x_0+x_1}{x_0}\right| \\ &\leq (2u-2u^2)\left|\tfrac{1}{x_0+x_1}\right| \\ \left|a-\tfrac{1}{x_0+x_1}\right| &\leq (5u+6u^2)\left|\tfrac{1}{x_0+x_1}\right| \\ \left|a(2-a(x_0+x_1))-\tfrac{1}{x_0+x_1}\right| &= |x_0+x_1|\left(a-\tfrac{1}{x_0+x_1}\right)^2 \\ &\leq (25u^2+61u^3)\left|\tfrac{1}{x_0+x_1}\right|\end{aligned}$$

Hence, $|a| \leq \left|\frac{1+6u}{x_0+x_1}\right|$. Now, we have,

$$\begin{aligned}\left|\bar{h}-(2-a(x_0+x_1))\right| &= |h_1-(-h_{1,1}-ax_1)| \\ &\leq u\cdot|h_1| \\ &\leq 3u^2+13u^3 \\ \left|\bar{b}-a\bar{h}\right| &= |b_{1,2}-(b_{1,1}+ah_1)| \\ &\leq u|b_{1,1}+ah_1| \\ &\leq u\left(u|a|(1-2u)+|ah_1|\right) \\ &\leq (4u^2+11u^3)|a|\end{aligned}$$

$$\begin{aligned}
\left|\bar{b} - \tfrac{1}{x_0+x_1}\right| &\leq (7u^2+24u^3)|a| + (25u^2+61u^3)\left|\tfrac{1}{x_0+x_1}\right| \\
&\leq (32u^2+121u^3)\left|\tfrac{1}{x_0+x_1}\right| \\
\\
\left|\tfrac{1}{x_0+x_1} - \tfrac{1}{\bar{x}}\right| &= \left|\tfrac{1}{\bar{x}}\right|\left|1 - \tfrac{\bar{x}}{x_0+x_1}\right| \\
&\leq \tfrac{2u^2}{1-2u}\left|\tfrac{1}{\bar{x}}\right| \\
&\leq (2u^2+5u^3)\left|\tfrac{1}{\bar{x}}\right| \\
\left|\bar{b} - \tfrac{1}{\bar{x}}\right| &\leq (34u^2+126u^3)\left|\tfrac{1}{\bar{x}}\right| \\
\left|\bar{b}(2-\bar{b}\bar{x}) - \tfrac{1}{\bar{x}}\right| &= |\bar{x}|(\bar{b} - \tfrac{1}{\bar{x}})^2 \\
&\leq 1165u^4\left|\tfrac{1}{\bar{x}}\right|
\end{aligned}$$

$$\rightarrow \left|\tfrac{1-35u^2}{\bar{x}}\right| \leq |\bar{b}| \leq \left|\tfrac{1+35u^2}{\bar{x}}\right|$$

Remark 1. *This term is ultimately negligible if p is large, because the accuracy is doubled at each step so it jumps from roughly 2 words to roughly 4 instead of just 3. Thus it is not a problem that computations were not performed accurately, for instance starting with $1+2u$ instead of 1.*

We denote δ_1 the relative error committed when computing $\bar{i}$ (taken relatively to $\bar{x}\bar{b}$) and δ_2 the one for $\bar{y}$. They are supposed less than $20u^3$ (see later).

$$\begin{aligned}
\left|\bar{i} - (2-\bar{x}\bar{b})\right| &\leq \delta_1|\bar{x}\bar{b}| \\
&\leq \delta_1(1+35u^2) \\
|\bar{i}| &\leq |2-\bar{x}\bar{b}| + \delta_1(1+35u^2) \\
&\leq 1+35u^2 \\
\left|\bar{y} - \bar{b}\bar{i}\right| &\leq \delta_2|\bar{b}\bar{i}| \\
&\leq \delta_2(1+35u^2)|\bar{b}| \\
\\
\left|\bar{y} - \tfrac{1}{\bar{x}}\right| &\leq (\delta_1+\delta_2)(1+35u^2)|\bar{b}| + 1165u^4\left|\tfrac{1}{\bar{x}}\right| \\
&\leq \left((\delta_1+\delta_2)(1+70u^2) + 1165u^4\right)\left|\tfrac{1}{\bar{x}}\right|.
\end{aligned}$$

To conclude, we only have to bound δ_1 and δ_2 as accurately as possible depending on the algorithms used.

1.2 Computation of $\bar{i}$

We have seen that $|\bar{b}\bar{x} - 1| \leq 35u^2$. Inside the computation of $3Prod_{2,3}(\bar{b}, \bar{x})$, we have seen that $|\bar{e} - \bar{b}\bar{x}| \leq 20u^3$. Thus $|\bar{e} - 1| \leq 36u^2$. Given that $\bar{e}$ is F-nonoverlapping, we have $|e_0 - \bar{e}| \leq (1-2^{-4})\mathrm{uls}(e_0)$. If $|e_0| < 1$, then $|\bar{e}| \geq (1-u) + (1-2^{-4})u = 1 - 2^{-4}u < 1 - 36u^2$, which is excluded, and we can exclude similarly $|e_0| > 1$. Thus $e_0 = 1$.

This property is the one that necessitates $p \geq 10$. It probably works for some smaller values of p, but we have no proof of that.

Therefore, in order to compute $2 - 3Prod_{2,3}(\bar{b}, \bar{x})$, we can simply replace $e_1, \ldots$ by their opposites by turning some $+$ into $-$ operations and conversely in order not to waste any operation. Correctness and the error bound are still ensured for the same reasons. For instance, if we choose to use the accurate

version, we obtain Algorithm **??**. One operation (the computation of e_0) is saved compared to the general version.

Algorithm 16 – 2 - 3Prod$^{acc}_{2,3}(b_0, b_1, x_0, x_1, x_2)$. (44 operations & 2 tests)

$z^+_{00}, z^-_{00} \leftarrow \text{2Prod}(b_0, x_0)$
$z^+_{01}, z^-_{01} \leftarrow \text{2Prod}(b_0, x_1)$
$z^+_{10}, z^-_{10} \leftarrow \text{2Prod}(b_1, x_0)$
$b'_0, b'_1, b'_2 \leftarrow \text{VecSum}(z^-_{00}, z^+_{01}, z^+_{10})$
$c \leftarrow \text{RN}(b'_2 + b_1 x_1)$ (FMA)
$z_{3,1} \leftarrow \text{RN}(z^-_{10} + b_0 x_1)$ (FMA)
$z_3 \leftarrow \text{RN}(z_{3,2} + z^-_{01})$
$s_1, e_2, e_3, e_4 \leftarrow \text{VecSum}(-b'_0, -b'_1, -c, -z_3)$
$(i_0 = 1)$
$e_1 \leftarrow \text{Fast2Sum}_2(1)(-z^+_{00}, s_1)$
$i_1, i_2 \leftarrow \text{VSEB}(2)(e_1, e_2, e_3, e_4)$
return $(1, i_1, i_2)$

1.3 Computation of $\bar{y}$

We have very precise information about $\bar{i}$, so we use a modified version of the fast algorithm.

Algorithm 17 – 3Prod$^{fast}_{2,3}(b_0, b_1, (1), i_1, i_2)$. (20 operations)

$z^+_{01}, z^-_{01} \leftarrow \text{2Prod}(b_0, i_1)$
$b'_0, b'_1 \leftarrow \text{Fast2Sum}(b_1, z^+_{01})$
$z_{3,1} \leftarrow \text{RN}(z^-_{01} + b_1 i_1)$ (FMA)
$z_3 \leftarrow \text{RN}(z_{3,1} + b_0 i_2)$ (FMA)
$s_3 \leftarrow \text{RN}(b'_1 + z_3)$
$e_0, e_1, e_2 \leftarrow \text{VecSum}(b_0, b'_0, s_3)$
$y_0 \leftarrow e_0$
$y_1, y_2 \leftarrow \text{Fast2Sum}(e_1, e_2)$
return (y_0, y_1, y_2)

Remark 2. *In VecSum(b_0, b'_0, s_3), the sum of b'_0 and s_3 is performed with a 2Sum in order to ensure correctness, but we can still use a Fast2Sum for the second one.*

A Fast2Sum can be used for the sum of b_1 and z^+_{01} because if the condition for Fast2Sum to be errorless is not satisfied, this means that $|b_1|$ is very small, so that the global error will be small anyway.

Given the estimates on $\bar{i}$, basically all errors are negligible, except the one when computing s_3. We get:

$$\left| \frac{\bar{y} - \bar{b}\bar{i}}{\bar{b}\bar{i}} \right| \leq \frac{u^3 + 256u^4}{(1-2u)(1-35u^2)}$$

Thus we can take $\delta_2 = u^3 + 260u^4$.

1.4 Final error bound and number of operations

We can take $\delta_1 = 10.5u^3 + 39u^4$ if we use the accurate version for the first $3Prod_{2,3}$, and $\delta_1 = 18u^3 + 75u^4$ if we use the fast version. Theorem 9 follows from that.

2 Analysis of Algorithm 14 and Proof of Theorem 10 (division of two TW numbers)

For the computation of $\bar{z}\bar{b}$ in Algorithm 14, $3Prod_{2,3}$ is used (the relative error is denoted by δ_3). The multiplication algorithm, to be used at the last line of Algorithm 14, that takes into account that $i_0 = 1$ is Algorithm **??** below

---

Algorithm 18 – 3Prod$^{fast}_{3,3}(a_0, a_1, a_2, (1), i_1, i_2)$. (21 operations)

$z^+_{01}, z^-_{01} \leftarrow \text{2Prod}(a_0, i_1)$
$b'_0, b'_1 \leftarrow \text{Fast2Sum}(a_1, z^+_{01})$
$z_{3,1} \leftarrow \text{RN}(z^-_{01} + a_1 i_1)$ (FMA)
$z_{3,2} \leftarrow \text{RN}(z_{3,1} + a_0 i_2)$ (FMA)
$z_3 \leftarrow \text{RN}(z_{3,2} + b'_1)$
$s_3 \leftarrow \text{RN}(z_3 + a_2)$
$e_0, e_1, e_2 \leftarrow \text{VecSum}(b_0, b'_0, s_3)$
$y_0 \leftarrow e_0$
$y_1, y_2 \leftarrow \text{Fast2Sum}(e_1, e_2)$
return (y_0, y_1, y_2)

The error analysis is similar to the one of Algorithm **??** with an additional $2u^3$, and we get

$$\left|\frac{\bar{y} - \bar{a}\bar{i}}{\bar{a}\bar{i}}\right| \leq \frac{3u^3 + 256u^4}{(1-2u)(1-35u^2)} \leq 3u^3 + 264u^4.$$

Globally, we have,

$$\left|\bar{y} - \frac{\bar{z}}{\bar{x}}\right| \leq \left((\delta_1 + \delta_2 + \delta_3)(1 + 70u^2) + 1165u^4\right)\left|\frac{\bar{z}}{\bar{x}}\right|$$

If we use the "accurate" versions of the multiplication algorithms, we can use $\delta_1 = 10.5u^3 + 39u^4 = \delta_3$, and if we use the "fast" versions, we can use $\delta_1 = 18u^3 + 75u^4 = \delta_3$. This gives Theorem 10.

3 Analysis of Algorithm 15 and Proof of Theorem 11 (square root of a TW number)

Much of the analysis is very similar to what was done for reciprocal and division, so we only focus on the differences.

The computation of $h_0^{(2)}$ in Algorithm 15 is exact because $h_{0,1}^{(2)} \geq 0.5$ (this is why we started with $1 + 4u$ instead of 1).

The computation of $\overline{i^{(1)}}$ is performed with one of the $3Prod_{2,3}$ algorithms (relative error bounded by δ_1).

The computation of $\overline{i^{(2)}}$ is performed using Algorithm **??** (or its fast version), where the penultimate line is replaced by $e_1 \leftarrow \text{Fast2Sum}(.5)(-z_{00}^+, s_1)$ (relative error bounded by δ_2). This works provided that $p \geq 11$.

The computation of $\bar{y}$ is performed using Algorithm **??** (relative error bounded by $\delta_3 = 3u^3 + 263u^4$).

We obtain

$$\left|\bar{b} - \frac{1}{\sqrt{\bar{x}}}\right| \leq (81u^2 + 622u^3)\left|\frac{1}{\sqrt{\bar{x}}}\right|, \text{and}$$

$$\left|\bar{b}\bar{x}(1.5 - \frac{1}{2}\bar{b}^2\bar{x}) - \sqrt{\bar{x}}\right| \leq 9916u^4\sqrt{\bar{x}}$$

Globally, we have

$$\left|\bar{y} - \sqrt{\bar{x}}\right| \leq \begin{pmatrix} \delta_1(1.5 + 287u^2) \\ +\delta_2(0.5 + 123u^2) \\ +\delta_3(1 + 162u^2) \\ +9916u^4 \end{pmatrix} \sqrt{\bar{x}}$$

which gives Theorem 11.