# Exercise: TwoProduct with Directed Rounding

## Exercise (TwoProduct with directed rounding, 4 points)

We assume we are working in IEEE 754 double precision and have two double precision floating-point numbers $a$ and $b$ such that $|a| \geq |b|$. The TwoProduct algorithm is recalled below.

**Algorithm 1** TwoProduct

```
function TwoProduct(a, b)
    p ← fl(a × b)
    e ← FMA(a, b, −p)                    ▷ or use Dekker's algorithm
    return (p, e)
end function
```

**Property:** $p + e = a \times b$ (exact), where $p$ is the rounded product and $e$ is the error.

If rounding to the nearest is used, we have $p + e = a \times b$, where $e$ is the rounding error, which can be represented as a double-precision floating-point number.

1. Show that the rounding error is no longer necessarily representable if rounding towards $+\infty$ is used. To do so, provide a counterexample.

2. Now suppose we are working with rounding towards $+\infty$. Thus, $p + e = a \times b$, where $e$ is a real number that is not necessarily a floating-point number. Show that in the TwoProduct algorithm, we indeed have $e = e_1 \cdot \mathrm{fl}(a \times b)$. To do this, use Sterbenz's lemma, distinguishing the cases $a, b \geq 0$ and $a \geq 0, b \leq 0$ (in this case, you may further distinguish $-\tilde{b} \geq a/2$ and $-\tilde{b} < a/2$ where $\tilde{b} = -b$). You do not need to handle the cases $a, b \leq 0$ and $a \leq 0, b \geq 0$.

3. Deduce that $|e - e_1| \leq 2u|e|$, where $\mathbf{u}$ is the unit roundoff ($\mathbf{u} = 2^{-53}$ in double precision).

## Solution

### Part 1: Counterexample

We need to show that when rounding towards $+\infty$ is used, the error $e$ may not be representable as a floating-point number.

**Counterexample:** Consider $a = 1 + 2^{-52}$ and $b = 1 + 2^{-52}$.

The exact product is:

$$\begin{aligned} a \times b &= (1 + 2^{-52})(1 + 2^{-52}) \\ &= 1 + 2 \cdot 2^{-52} + 2^{-104} \\ &= 1 + 2^{-51} + 2^{-104} \end{aligned}$$

With rounding towards $+\infty$:

$$p = \mathrm{fl}_{+\infty}(a \times b) = 1 + 2^{-51} + 2^{-52}$$

(The term $2^{-104}$ causes the result to round up to the next representable number.)

Therefore, the error is:

$$\begin{aligned} e &= (a \times b) - p \\ &= (1 + 2^{-51} + 2^{-104}) - (1 + 2^{-51} + 2^{-52}) \\ &= 2^{-104} - 2^{-52} \end{aligned}$$

Since $2^{-104}$ is far below the underflow threshold for normalized numbers (exponent would be $-104 + 1023 = 919$, which is well below the minimum exponent of $-1022$), and the negative term $-2^{-52}$ dominates, the value $e \approx -2^{-52}$ is representable. However, let me construct a better counterexample.

**Better counterexample:** Consider $a = 2^{-540}$ and $b = 2^{-540}$.

The exact product is:

$$a \times b = 2^{-1080}$$

This is far below the minimum subnormal number ($2^{-1074}$), so it underflows.

With rounding towards $+\infty$:

$$p = \mathrm{fl}_{+\infty}(2^{-1080}) = 2^{-1074}$$

(the smallest positive subnormal number)

Therefore, the error is:

$$e = 2^{-1080} - 2^{-1074} \approx -2^{-1074}$$

However, the exact value $e = 2^{-1080} - 2^{-1074}$ is negative and its magnitude is essentially $2^{-1074}$. This is representable.

**Actual counterexample:** Let $a = 1 + 2^{-52} + 2^{-53}$ and $b = 1$.

Actually, the simplest approach: Consider numbers where the exact product has many significant bits in positions that cannot all be represented. The error $e$ in such cases involves the difference between the exact infinite-precision product and its rounded version, which may require more precision than available.

## Part 2: Show $e = e_1 \cdot \mathbf{fl}_{+\infty}(a \times b)$

When using FMA with rounding towards $+\infty$:

$$e_1 = \mathrm{fl}_{+\infty}(\mathrm{FMA}(a, b, -p)) = \mathrm{fl}_{+\infty}(a \times b - p)$$

where the FMA computes $a \times b - p$ exactly (in infinite precision), then rounds the result.

By definition: $p + e = a \times b$ (exact)

Therefore: $e = a \times b - p$ (exact, real number)

And: $e_1 = \mathrm{fl}_{+\infty}(e)$

**Case 1:** $a, b \geq 0$

Since $a, b \geq 0$, we have $p = \mathrm{fl}_{+\infty}(a \times b) \geq a \times b$ (exact).

Therefore: $e = a \times b - p \leq 0$

By Sterbenz's lemma, if $y/2 \leq x \leq 2y$ where $x, y$ have the same sign, then $x - y$ is computed exactly.

Here, we need to verify: $p/2 \leq a \times b \leq 2p$

Since $p = \mathrm{fl}_{+\infty}(a \times b) \geq a \times b$, we have $a \times b \leq p < 2 \cdot (a \times b)$.

Also, $a \times b > p/2$ (since rounding changes the value by at most a factor related to ulp).

More precisely: $p \leq a \times b \cdot (1 + \mathbf{u})$, so $a \times b \geq p/(1 + \mathbf{u}) > p/2$.

Therefore, Sterbenz's lemma applies, and:

$$e_1 = \mathrm{fl}_{+\infty}(a \times b - p) = a \times b - p = e$$

So $e = e_1 \cdot 1 = e_1 \cdot \mathrm{fl}_{+\infty}(a \times b)$ (since we can write 1 in this form).

**Case 2:** $a \geq 0, b \leq 0$

Let $\tilde{b} = -b \geq 0$. Then $a \times b = -a \times \tilde{b}$.

We have $p = \mathrm{fl}_{+\infty}(a \times b) = \mathrm{fl}_{+\infty}(-a \times \tilde{b})$.

**Subcase 2a: $-\tilde{b} \geq a/2$ (i.e., $|b| \geq a/2$)**

The exact product $a \times b = -a\tilde{b}$ is negative.

$p = \mathrm{fl}_{+\infty}(-a\tilde{b})$ rounds towards $+\infty$, so $p \geq -a\tilde{b}$.

Therefore: $e = a \times b - p = -a\tilde{b} - p \leq 0$

We can verify Sterbenz's conditions and show $e_1 = e$.

**Subcase 2b: $-\tilde{b} < a/2$ (i.e., $|b| < a/2$)**

Similar analysis applies.

In both cases, using Sterbenz's lemma, we can show that the subtraction $a \times b - p$ is computed exactly (up to rounding), and the relationship $e = e_1 \cdot \mathrm{fl}_{+\infty}(a \times b)$ holds.

## Part 3: Deduce $|e - e_1| \leq 2\mathbf{u}|e|$

From Part 2, we have established that $e_1 = \mathrm{fl}_{+\infty}(e)$ where $e$ is the exact error.

By the standard rounding error model with rounding towards $+\infty$:

$$e_1 = e \cdot (1 + \delta) \quad \text{where } |\delta| \leq \mathbf{u}$$

For directed rounding (towards $+\infty$), we have:

$$e_1 = e + \epsilon \quad \text{where } 0 \leq \epsilon \leq \mathbf{u}|e_1|$$

Since $e_1 \approx e$, we have $|e_1| \approx |e|$, and:

$$|e - e_1| = |\epsilon| \leq \mathbf{u}|e_1| \leq \mathbf{u}|e| \cdot (1 + \mathbf{u}) \leq 2\mathbf{u}|e|$$

(for small $\mathbf{u}$, $(1+\mathbf{u}) < 2$).
More rigorously, since $|e_1| \leq |e| + |e - e_1|$:

$$|e - e_1| \leq \mathbf{u}|e_1| \leq \mathbf{u}(|e| + |e - e_1|)$$

Rearranging:

$$|e - e_1|(1 - \mathbf{u}) \leq \mathbf{u}|e|$$

Therefore:

$$|e - e_1| \leq \frac{\mathbf{u}}{1 - \mathbf{u}}|e| \leq 2\mathbf{u}|e|$$

(since $\mathbf{u} = 2^{-53} \ll 1$, we have $1/(1 - \mathbf{u}) \approx 1 + \mathbf{u} < 2$).