Error Propagation in Arithmetic

Let $\circ$ represent an arithmetic operation, $+,-,\ast,$ or $\slash$. Let $x \circ y$ be evaluated on a machine. Assuming that $\tilde{x}, \tilde{y}$ are the machine representations of $x, y$ and $\tilde{x} \circ \tilde{y}$ is the output, write the error as a sum of two terms,

$$x \circ y - \tilde{x} \circ \tilde{y} = [x \circ y - \tilde{x} \circ \tilde{y}] + [\tilde{x} \circ \tilde{y} - \tilde{x} \circ \tilde{y}].$$

The second term on the right has to do with how $\circ$ is implemented and how the result is rounded. These are specifics of a computing device and its rounding regime.

The first term on the right has to do with how sensitive $\circ$ is to perturbations. This does not depend on the device and can be fully analyzed.

**MULTIPLICATION** The easy-to-check identity

$$xy - \tilde{x}\tilde{y} = (x - \tilde{x})y + (y - \tilde{y})x - (x - \tilde{x})(y - \tilde{y})$$

may be written in the form

$$\text{rel}(\tilde{x}\tilde{y}) = \text{rel}(\tilde{x}) + \text{rel}(\tilde{y}) - \text{rel}(\tilde{x})\text{rel}(\tilde{y}).$$

If $\text{rel}(\tilde{x})$ and $\text{rel}(\tilde{y})$ are sufficiently small, then their product is negligible and we have the approximate equality

$$\text{rel}(\tilde{x}\tilde{y}) \approx \text{rel}(\tilde{x}) + \text{rel}(\tilde{y}).$$

The product is therefore not sensitive to small perturbations of individual factors.

**DIVISION** From the expression for the relative error of the reciprocal,

$$\text{rel}(1/\tilde{y}) = -\frac{y - \tilde{y}}{\tilde{y}} = -\frac{y - \tilde{y}}{1 - \frac{y - \tilde{y}}{y}} = -\frac{\text{rel}(\tilde{y})}{1 - \text{rel}(\tilde{y})},$$

it follows that $\text{rel}(1/\tilde{y}) \approx -\text{rel}(\tilde{y})$, as long as $\text{rel}(\tilde{y})$ is sufficiently small ($|\text{rel}(\tilde{y})| << 1$).

So, using the approximate equality for multiplication, we obtain

$$\text{rel}(\tilde{x}/\tilde{y}) \approx \text{rel}(\tilde{x}) - \text{rel}(\tilde{y}).$$

The quotient is not sensitive to small perturbations if the divisor has a small relative error.

Observe that the behavior of the relative error is similar to that of a logarithm.

**ADDITION/SUBTRACTION** The error of the sum is the sum of errors,

$$(x + y) - (\tilde{x} + \tilde{y}) = (x - \tilde{x}) + (y - \tilde{y}),$$

and a similar equality holds for differences. The same formula may also be written in terms of the relative errors,

$$\text{rel}(\tilde{x} + \tilde{y}) = \frac{x}{x + y} \text{rel}(\tilde{x}) + \frac{y}{x + y} \text{rel}(\tilde{y}).$$

Perhaps surprisingly, this expression reveals a subtlety: when $x + y$ is close enough to zero, $\text{rel}(\tilde{x} + \tilde{y})$ may be large compared to $\text{rel}(\tilde{x}), \text{rel}(\tilde{y})$, which may lead to a loss of significant digits.
Here are several examples where the loss-of-significance error occurs.

**EXAMPLE**

\[
\begin{array}{cccc}
2 & 3 & 5 & 2 \\
2 & 3 & 5 & 1 \\
1 & \text{unreliable digits} & \times 10^{-2} & \times 10^{-2} & \times 10^{-5}
\end{array}
\]

**EXAMPLE** Let \( x = 20 \) and \( y = \sqrt{399} \). Note that \( x \) and \( y \) are close to each other. The values of \( x, y, x + y, \) and \( x - y \), rounded to four decimal places, are \( \tilde{x} = 20.0000, \tilde{y} = 19.9750, \tilde{x} + \tilde{y} = 39.9750 \) and \( \tilde{x} - \tilde{y} = 0.0250 \). The absolute errors of \( \tilde{x} + \tilde{y} \) and \( \tilde{x} - \tilde{y} \) are equal and bounded by \( 5 \times 10^{-5} \). The relative error of \( \tilde{x} + \tilde{y} \) is approximately \(-4 \times 10^{-7}\) (6 significant digits). But the relative error of \( \tilde{x} - \tilde{y} \) is only about \( 6 \times 10^{-4}\): significant digits are lost.

**EXAMPLE** As \( x \to 0^{+} \),

\[
\frac{\sqrt{1 + x} - 1}{x} = \frac{1}{1 + \sqrt{1 + x}}
\]

increases to 0.5 monotonically. The following table was obtained using MATLAB. It exhibits noise in evaluation, loss of significant digits, and underflow. The ‘error’ column gives the differences between 0.5 and the entries of the second column.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( \frac{\sqrt{1 + x} - 1}{x} )</th>
<th>error</th>
<th># signif. digits</th>
<th>( \frac{1}{1 + \sqrt{1 + x}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 10^{-5} )</td>
<td>0.4999998750</td>
<td>0.000001250</td>
<td>5</td>
<td>0.49999875000625</td>
</tr>
<tr>
<td>( 10^{-6} )</td>
<td>0.499999875</td>
<td>0.000000125</td>
<td>6</td>
<td>0.49999987500006</td>
</tr>
<tr>
<td>( 10^{-7} )</td>
<td>0.499999988</td>
<td>0.000000125</td>
<td>7</td>
<td>0.49999998750000</td>
</tr>
<tr>
<td>( 10^{-8} )</td>
<td>0.499999997</td>
<td>0.000000125</td>
<td>8</td>
<td>0.49999999875000</td>
</tr>
<tr>
<td>( 10^{-9} )</td>
<td>0.500000041</td>
<td>-0.000000041</td>
<td>7</td>
<td>0.49999999998750</td>
</tr>
<tr>
<td>( 10^{-10} )</td>
<td>0.500000041</td>
<td>-0.000000041</td>
<td>7</td>
<td>0.49999999999875</td>
</tr>
<tr>
<td>( 10^{-11} )</td>
<td>0.500000041</td>
<td>-0.000000041</td>
<td>7</td>
<td>0.49999999999875</td>
</tr>
<tr>
<td>( 10^{-12} )</td>
<td>0.500000041</td>
<td>-0.000000041</td>
<td>7</td>
<td>0.49999999999987</td>
</tr>
<tr>
<td>( 10^{-13} )</td>
<td>0.499600361</td>
<td>0.000399639</td>
<td>3</td>
<td>0.49999999999999</td>
</tr>
<tr>
<td>( 10^{-14} )</td>
<td>0.488498131</td>
<td>0.011501869</td>
<td>1</td>
<td>0.50000000000000</td>
</tr>
<tr>
<td>( 10^{-15} )</td>
<td>0.444089210</td>
<td>0.055910790</td>
<td>0</td>
<td>0.50000000000000</td>
</tr>
<tr>
<td>( 10^{-16} )</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.50000000000000</td>
</tr>
</tbody>
</table>

The last column gives an accurate alternative. Note that the calculation of \( \frac{1}{1 + \sqrt{1 + x}} \) does not involve subtraction of nearby numbers.