CS412: Review Notes

Mridul Aanjaneya

April 30, 2015

Arithmetic operations are subject to roundoff error when performed on a finite precision computer. In order to perform an operation x op y on the *real* numbers x and y, we deviate from the analytic result when discretizing those values to machine precision as well as when we store the resulting value.

Let $\bar{x}$ denote the discretized, floating point version of x that is stored on the computer. You may assume that

$$\bar{x} = (1+\varepsilon)x$$

where ε is bounded as $0 \leq |\varepsilon| < \varepsilon_{\mathsf{max}}$ where $\varepsilon_{\mathsf{max}} << 1$ is the machine roundoff precision.

Assume that the result of the arithmetic operation between two floating point numbers $\bar{x}$ and $\bar{y}$ is computed exactly, but when stored on the computer it is once again subject to roundoff error as

$$\overline{\bar{x}\ \mathsf{op}\ \bar{y}} = (1+\varepsilon')(\bar{x}\ \mathsf{op}\ \bar{y})$$

where the roundoff error obeys the same bounds $0 \leq |\varepsilon'| < \varepsilon_{\mathsf{max}}$.

The relative error of a computation is defined as

$$E = \left|\frac{\mathsf{Computed\ Result} - \mathsf{Analytic\ Result}}{\mathsf{Analytic\ Result}}\right|$$

Using this relation, we can provide a bound for various arithmetic operations, or prove that the relative error is unbounded. For the following derivations, we use the lemma: If $0 \leq |\varepsilon_1|, |\varepsilon_2|, \ldots, |\varepsilon_k| < \varepsilon_{\mathsf{max}}$, then there exists an $\varepsilon \in [0, \varepsilon_{\mathsf{max}})$ such that $(1+\varepsilon_1)(1+\varepsilon_2)\ldots(1+\varepsilon_k) = (1+\varepsilon)^k$, which holds by virtue of the intermediate mean value theorem.

1. **Subtraction:** We have $\bar{x} = (1+\varepsilon_1)x$ and $\bar{y} = (1+\varepsilon_2)y$. We will show that there is no bound on the relative error. Consider $\bar{x} - \bar{y} = (1+\varepsilon_1)x - (1+\varepsilon_2)y$. Let $x = a + \theta$ and $y = a$ so $x - y = \theta$. Then

$$
\begin{aligned}
E &= \left|\frac{\overline{\bar{x}-\bar{y}}-(x-y)}{x-y}\right| \\
&= \left|\frac{\theta(1+\varepsilon_3)+a(\varepsilon_1-\varepsilon_2)(1+\varepsilon_3)+\theta\varepsilon_1(1+\varepsilon_3)-\theta}{\theta}\right| \\
&= \left|\varepsilon_3+\varepsilon_1(1+\varepsilon_3)+\frac{a}{\theta}(\varepsilon_1-\varepsilon_2)(1+\varepsilon_3)\right|
\end{aligned}
$$

which becomes unbounded as $\theta \to 0$.

2. **Addition:**

$$
\begin{aligned}
E_+ &= \left|\frac{\overline{\bar{x}+\bar{y}}-(x+y)}{x+y}\right| = \left|\frac{(1+\varepsilon_3)\{(1+\varepsilon_1)x+(1+\varepsilon_2)y\}-(x+y)}{x+y}\right| \\
&= \left|\frac{(1+\varepsilon_3)(1+\varepsilon_1)x+(1+\varepsilon_3)(1+\varepsilon_2)y-(x+y)}{x+y}\right|
\end{aligned}
$$

From the above lemma, $(1+\varepsilon_3)(1+\varepsilon_1)=(1+\varepsilon_4)^2$ and $(1+\varepsilon_3)(1+\varepsilon_2)=(1+\varepsilon_5)^2$. Without loss of generality, assume $\varepsilon_4=\max\{\varepsilon_4,\varepsilon_5\}$. Then,

$$
E_+ \le \left|\frac{(1+\varepsilon_4)^2(x+y)-(x+y)}{x+y}\right| = |2\varepsilon_4+O(\varepsilon_{\mathsf{max}}^2)| = O(|\varepsilon_{\mathsf{max}}|)
$$

3. **Multiplication:**

$$
E_\times = \left|\frac{\overline{\bar{x}\cdot\bar{y}}-xy}{xy}\right| = \left|\frac{xy(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)-xy}{xy}\right|
$$

From the above lemma, we can write $(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)=(1+\varepsilon_4)^3$. This gives,

$$
E_\times = |(1+\varepsilon_4)^3-1| = |3\varepsilon_4+O(\varepsilon_{\mathsf{max}}^2)| = O(|\varepsilon_{\mathsf{max}}|)
$$

4. **Division:**

$$
E_\div = \left|\frac{\overline{\bar{x}/\bar{y}}-x/y}{x/y}\right| = \left|\frac{(x/y)\frac{(1+\varepsilon_1)(1+\varepsilon_3)}{1+\varepsilon_2}-x/y}{x/y}\right| = \left|\frac{(1+\varepsilon_1)(1+\varepsilon_3)}{1+\varepsilon_2}-1\right|
$$

Using the geometric series, $\frac{1}{1+\varepsilon_2}=1-\varepsilon_2+O(\varepsilon_{\mathsf{max}}^2)$ gives

$$
\begin{aligned}
E_\div &= |(1+\varepsilon_1)(1+\varepsilon_3)\{(1-\varepsilon_2)+O(\varepsilon_{\mathsf{max}}^2)\}-1| = |(1+\varepsilon_4)^3-1+O(\varepsilon_{\mathsf{max}}^2)| \\
&= |3\varepsilon_4+O(\varepsilon_{\mathsf{max}}^2)| = O(|\varepsilon_{\mathsf{max}}|)
\end{aligned}
$$