

Three Views of Cancellation

Let x and y be real numbers such that x , y and $x + y$ do not overflow or underflow. How good is the floating point approximation $\text{fl}(\text{fl}(x) + \text{fl}(y))$ to the true value $x + y$? Write $\bar{x} = \text{fl}(x)$, $\bar{y} = \text{fl}(y)$, and $\bar{z} = \text{fl}(\bar{x} + \bar{y})$. If $z = x + y$, then the relative error in the computed sum is

$$\frac{|z - \bar{z}|}{|z|}.$$

- First, an algorithmic perspective: Suppose $x = 0.d_1d_2 \dots d_s d_{s+1} \dots d_t d_{t+1} \dots \times \beta^e$, and $y = -0.d_1d_2 \dots d_s e_{s+1} \dots e_t e_{t+1} \dots \times \beta^e$, with $\bar{x} = 0.d_1d_2 \dots d_s d_{s+1} \dots d_t \times \beta^e$ and $\bar{y} = -0.d_1d_2 \dots d_s e_{s+1} \dots e_t \times \beta^e$. We have set this up so that x and y are opposite numbers up to s digits. Then (without loss of generality take $e_{s+1} \leq d_{s+1}$)

$$\bar{x} + \bar{y} = \pm 0.00 \dots 0 f_{s+1} f_{s+2} \dots f_t f_{t+1} \times \beta^e,$$

giving

$$\bar{z} = \text{fl}(\bar{x} + \bar{y}) = \pm f_{s+1} f_{s+2} \dots f_t g_1 g_2 \dots g_s \times \beta^{e-s}.$$

Now \bar{z} carries with it the s digits g_1, \dots, g_s which are *completely meaningless!* The first s digits of x and y cancelled out, and in normalization those zeros slid off to the left; they were replaced by garbage on the right. If x and y have the same sign, there is *no* cancellation, but if s is very large the result can be catastrophic.

Notice that s can be large if $x + y \approx 0$.

- Now an error analysis: By the FAFA and the FRT there exist $|\epsilon_x|, |\epsilon_y|, |\epsilon| \leq \mu$ such that

$$\bar{z} = \text{fl}(\bar{x} + \bar{y}) = (x(1 + \epsilon_x) + y(1 + \epsilon_y))(1 + \epsilon),$$

so

$$|z - \bar{z}| = |x(\epsilon_x + \epsilon) + y(\epsilon_y + \epsilon) + O(\mu^2)| \leq 2\mu(|x| + |y|) + O(\mu^2)$$

This gives an upper bound on the relative error:

$$\frac{|z - \bar{z}|}{|z|} \leq 2\mu \frac{|x| + |y|}{|x + y|} + O(\mu^2)$$

Notice that this can be large if $x + y \approx 0$.

- Finally, we do a sensitivity analysis: Consider the problem “evaluate the function $f(z) = x + z$ at $z = y$ ”. Small relative perturbations in z can be magnified in $f(z)$ by the relative condition number

$$\nu = \frac{|y| |f'(y)|}{|f(y)|} = \frac{|y|}{|x + y|}.$$

Notice that this can be large if $x + y \approx 0$.