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 $\mathrm{fl}(\mathrm{fl}(x)+\mathrm{fl}(y))$ to the true value $x+y$? Write $\bar{x}=\mathrm{fl}(x)$, $\bar{y}=\mathrm{fl}(y)$, and $\bar{z}=\mathrm{fl}(\bar{x}+\bar{y})$. If $z=x+y$, then the relative error in the computed sum is

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

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

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

giving

$$\bar{z}=\mathrm{fl}(\bar{x}+\bar{y})=\pm f_{s+1}f_{s+2}\ldots f_tg_1g_2\ldots g_s\times\beta^{e-s}.$$

Now $\bar{z}$ carries with it the s digits $g_1,\ldots,g_s$ *which are completely meaningless!* The first s digits of x and y *cancelled* out, and as 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 $|\epsilon_x|,|\epsilon_y|,|\epsilon|\le\mu$ such that

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

so

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

This gives an upper bound on the relative error:

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

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 $x+y\approx 0$.