COMPUTER SCIENCE 349A

Handout Number 4

SUBTRACTIVE CANCELLATION (pages 73-76 of the 6th edition; pages 76-79 of the 7th edition)

Subtractive cancellation refers to the loss of significant digits during a floating-point computation due to the subtraction of nearly equal floating-point numbers.

Note that if $\hat{x}$ is an approximation to $x > 0$ and $\hat{y}$ is an approximation to $y > 0$, and if for example $\hat{x}$ agrees with x to 8 significant digits and $\hat{y}$ agrees with y to 8 significant digits, then

$$\hat{x} \times \hat{y} \approx x \times y$$
$$\hat{x} / \hat{y} \approx x / y$$
$$\hat{x} + \hat{y} \approx x + y$$

and these values will agree to about 8 significant digits. However, this may not be true for subtraction: it is possible that none of the significant digits in $\hat{x} - \hat{y}$ and $x - y$ agree.

The following examples illustrate subtractive cancellation, and show how it can be avoided in each of these cases.

Example 1.

The evaluation of

$$f\ell\left(\sqrt{x^2+1} - x\right)$$

will be inaccurate if x is large and positive. For example, using $b = 10$, $k = 4$, idealized <u>rounding</u> floating-point arithmetic and $x = 65.43$, we obtain the following (where $f\ell\left(\sqrt{z}\right)$ is computed using idealized floating-point arithmetic; that is, the exact value of $\sqrt{z}$ is rounded to 4 significant digits).

$$f\ell\left(x^2\right) = f\ell(4281.0849) = 4281 \quad \text{or} \quad 0.4281 \times 10^4$$
$$f\ell\left(x^2+1\right) = f\ell(4281+1) = 4282$$
$$f\ell\left(\sqrt{x^2+1}\right) = f\ell\left(\sqrt{4282}\right) = f\ell(65.43699\cdots) = 65.44$$
$$f\ell\left(\sqrt{x^2+1} - x\right) = f\ell(65.44 - 65.43) = 0.01 \quad \text{or} \quad 0.1000 \times 10^{-1}$$

However, the true (exact) value of $\sqrt{x^2+1} - x$ is $0.0076413\cdots$. The relative error in $f\ell\left(\sqrt{x^2+1} - x\right)$ is about 0.31 or 31%.

To avoid the subtractive cancellation above and to obtain an accurate floating-point result, note that

$$\left(\sqrt{x^2+1}-x\right)\left(\frac{\sqrt{x^2+1}+x}{\sqrt{x^2+1}+x}\right)=\frac{1}{\sqrt{x^2+1}+x}.$$

The latter expression gives an extremely accurate result in floating-point arithmetic when $x = 65.43$ (and indeed for <u>all</u> "large" positive values of x).

$$\begin{aligned}
f\ell\left(x^2\right) &= f\ell(4281.0849) = 4281 \quad \text{or} \quad 0.4281\times 10^4 \\
f\ell\left(x^2+1\right) &= f\ell(4281+1) = 4282 \\
f\ell\left(\sqrt{x^2+1}\right) &= f\ell\left(\sqrt{4282}\right) = f\ell(65.43699\cdots) = 65.44 \\
f\ell\left(\sqrt{x^2+1}+x\right) &= f\ell(65.44+65.43) = f\ell(130.87) = 130.9 \\
f\ell\left(\frac{1}{\sqrt{x^2+1}+x}\right) &= f\ell\left(\frac{1}{130.9}\right) = f\ell(0.00763941\cdots) = 0.007639
\end{aligned}$$

which has a relative error of 0.0003 or 0.03%.

Example 2.

The evaluation of

$$f\ell(x-\sin x)$$

will be inaccurate if x is close to 0. For example, using $b=10$, $k=4$, idealized <u>chopping</u> floating-point arithmetic and $x = 0.01234$, we obtain the following (note that the argument for sine is in radians).

$$\begin{aligned}
f\ell(\sin x) &= f\ell(0.01233968\cdots) = 0.01233 \quad \text{or} \quad 0.1233\times 10^{-1} \\
f\ell(x-\sin x) &= f\ell(0.01234-0.01233) = 0.00001 \quad \text{or} \quad 0.1000\times 10^{-4}
\end{aligned}$$

However, the true (exact) value of $x-\sin x$ is $0.313177\cdots\times 10^{-6}$, giving a relative error in the computed approximation of

$$\left|1-\frac{0.00001}{0.313177\times 10^{-6}}\right| = 30.93 \quad \text{or} \quad 3093\%.$$

To avoid the catastrophic loss of significant digits in this example, use the Taylor series approximation for $f(x)=\sin x$ expanded about $x_0=0$ (see Chapter 4 of the textbook):

$$x-\sin x = x-\left(x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\frac{x^9}{9!}-\cdots\right).$$

Thus, if x is close to 0, a very good approximation to $x-\sin x$ is, for example,

$$x - \sin x \approx \frac{x^3}{6} - \frac{x^5}{120}$$

With $x = 0.01234$ as above, we obtain

$$\begin{aligned}
f\ell(x^3) &= f\ell(0.18790809\cdots\times10^{-5}) = 0.1879\times10^{-5} \\
f\ell(x^3/6) &= f\ell(0.1879\times10^{-5}/6) = f\ell(0.3131666\cdots\times10^{-6}) = 0.3131\times10^{-6} \\
f\ell(x^5) &= f\ell(0.28613817\cdots\times10^{-9}) = 0.2861\times10^{-9} \\
f\ell(x^5/120) &= f\ell(0.2861\times10^{-9}/120) = f\ell(0.2384166\cdots\times10^{-11}) = 0.2384\times10^{-11} \\
f\ell(x^3/6 - x^5/120) &= f\ell(0.3131\times10^{-6} - 0.2384\times10^{-11}) = f\ell(0.3130976\cdots\times10^{-6}) \\
&= 0.3130\times10^{-6}
\end{aligned}$$

which has a very small relative error of

$$\left|1 - \frac{0.3130\times10^{-6}}{0.313177\times10^{-6}}\right| = 0.000565 \quad \text{or} \quad 0.0565\%$$

Example 3.

The evaluation of

$$f\ell(1-\sin x)$$

will be inaccurate if $\sin x$ is close to 1; for example, if $x \approx \pi/2 = 1.5707963\cdots$ (radians).

For example, this floating-point computation will be inaccurate if $x = 1.56$. The cancellation of significant digits in this case can be seen since $\sin(1.56) = 0.99994172\cdots$.

To avoid such a loss of significant digits whenever x is close to $\pi/2$, note that

$$(1-\sin x)\left(\frac{1+\sin x}{1+\sin x}\right) = \frac{1-\sin^2 x}{1+\sin x} = \frac{\cos^2 x}{1+\sin x}$$

Evaluation of this expression in floating-point arithmetic when x is close to $\pi/2$ will not result in any large loss of significant digits. For example, with $x = 1.56$ we obtain (using <u>rounding</u>) the following, which is very close to the true value of $0.00005827977\cdots$.

$$\begin{aligned}
f\ell(\cos x) &= f\ell(0.010796\cdots) = 0.01080 \\
f\ell(\cos^2 x) &= f\ell(0.00011664) = 0.0001166 \\
f\ell(\sin x) &= f\ell(0.99994172\cdots) = 0.9999 \\
f\ell(1+\sin x) &= f\ell(1.9999) = 2.000 \\
f\ell(\cos^2 x/(1+\sin x)) &= f\ell(0.0001166/2.000) = 0.00005830
\end{aligned}$$

Example 4.

Provided that $x \neq 1/2$ or $x \neq 2$,

$$\frac{1}{2x-1} - \frac{x+2}{x-2} = \frac{-2x(x+1)}{(2x-1)(x-2)}.$$

However, if evaluated in floating-point arithmetic, these two expressions may give very different results; that is, if we let

$$f(x) = \frac{1}{2x-1} - \frac{x+2}{x-2} \quad \text{and} \quad g(x) = \frac{-2x(x+1)}{(2x-1)(x-2)},$$

then for some values of x, $f\ell(f(x))$ and $f\ell(g(x))$ may differ greatly.

In each of the following cases, assume that our usual floating-point system with $b = 10, k = 4$ and rounding is used.

Case (i).

Suppose that ***x* is an exact valid floating-point number** (in whatever floating-point system you are using) and that ***x* is close to (but not equal to)** -1. Then the evaluation of $f\ell(f(x))$ will be very inaccurate since

$$f\ell\left(\frac{1}{2x-1}\right) \approx -\frac{1}{3} \quad \text{and} \quad f\ell\left(\frac{x+2}{x-2}\right) \approx -\frac{1}{3}$$

so that $f\ell(f(x))$ is computed as the difference of two almost equal numbers, which will result in a loss of significant digits due to subtractive cancellation. However, this does not occur in the evaluation of $f\ell(g(x))$ when x is close to -1.

For example, if $x = -0.9986$, then you can verify the following. The exact value is $f(x) = g(x) = 0.0003111109\cdots$; $f\ell(f(-0.9986)) = 0.0001000$ or 0.1000×10^{-3} is very inaccurate; $f\ell(g(-0.9986)) = 0.0003111$ is very accurate.

Case (ii).

Suppose that ***x* is an exact valid floating-point number** (in whatever floating-point system you are using) and that ***x* is close to (but not equal to) 2**. Then the evaluation of both of $f\ell(f(x))$ and $f\ell(g(x))$ will be very accurate because there is no subtractive cancellation in either expression. Although $f\ell(x-2)$ occurs in the denominator of each expression, if x is an exact valid floating-point number, then there is no round-off error in $f\ell(x-2)$; that is, $f\ell(x-2)$ is exactly equal to the value of $x-2$. (To see this, consider values such as 1.997 or 2.023 .) Thus, $f\ell(1/(x-2))$ will be very accurate (as the round-off error in a single floating-point division is small). Since the value of $f\ell(1/(x-2))$ is also very large relative to all other parts of the expressions for $f(x)$ and $g(x)$, the values of both $f\ell(f(x))$ and $f\ell(g(x))$ will be very accurate.

For example, if $x = 1.997$, then you can verify the following. The exact value is $f(x) = g(x) = 1332.6673\cdots$; $f\ell(f(1.997)) = 1332$ and $f\ell(g(1.997)) = 1333$.

Case (iii).

Suppose that ***x* is NOT a valid floating-point number and that *x* is close to 2.** For example, if $b = 10, k = 4$ suppose that $x = 2.001234$. Then both of $f\ell(f(x))$ and $f\ell(g(x))$ will be very inaccurate because they both are computed using the value of $f\ell(2.001234) = 2.001$ rather than the exact value of x. In such a case, note that the value of

$$f\ell\left(\frac{1}{x-2}\right) \text{ and the exact value of } \frac{1}{x-2}$$

will differ greatly. For example, using $x = 2.001234$,

$$f\ell\left(\frac{1}{x-2}\right) = 1000 \text{ whereas } \frac{1}{x-2} = 810.37277\cdots.$$

Using $x = 2.001234$, the exact value is $f(x) = g(x) = -3242.1580\cdots$. Using $f\ell(2.001234) = 2.001$, you can verify that $f\ell(f(2.001)) = -4001$ and $f\ell(g(2.001)) = -4001$, both of which are very inaccurate.