Cancellation Error

Cancellation error is typically caused by subtracting two numbers that are nearly equal, or adding two numbers that are nearly opposite.

Here is an example in 10-digit arithmetic, where it is easier to visualize the cancellation.

$$x = 123.4567899, \quad y = 123.4555555 \quad \Rightarrow \quad x - y = 0.0012344$$

Both x and y could be accurate to 10 significant digits compared to their respective true values, but since they are nearly equal their leading 5 digits cancelled out, leaving the result of their subtraction accurate to only 5 significant digits.

Some books (and Wikipedia) call this result "catastrophic cancellation", because most or all of the correct significant digits for the entire calculation can be lost in just a single operation.

By using 50-digit precision, Calc-50 can often avoid problems caused by cancellation. If we need a result correct to 10 or 15 significant digits and the formula being evaluated loses 20 digits to cancellation, there is no problem since we will have over 30 digits of precision remaining.

But in some cases the amount of cancellation error can be so great that even 50 digits is not enough for us to finish with enough accuracy.

We need to be able to recognize when cancellation error can happen in a given calculation, when enough digits would be lost that it would keep the calculation from achieving the accuracy we want, and what we can do to work around the problem in that case.

Not every expression that looks like it might have cancellation really does, and sometimes an expression that does have some cancellation in it can recover and get an accurate result.

Example 1. There is cancellation, but the final result is still accurate.

$$x^2 - x + 2$$

It looks like the $x^2 - x$ part of this would have cancellation when x^2 is close to x, or the $x^2 - x + 2$ when $x^2 - x$ is close to -2.

Here the addition cannot cause cancellation because $x^2 - x$ is never less than -0.25, so it will never be close to -2

There can be cancellation in $x^2 - x$ near $x = 1$, but it is not harmful because of what happens next. For x close to 1, $x^2 - x$ is close to $x - 1$, which is small. Next we add 2, leaving the result close to 2. Here is a 10-digit base 10 example showing the intermediate results, along with the same calculation using 20 digits to show the correct values. The same thing can happen with higher precision, but the effects of the cancellation are easier to see with 10 digits.

value	10-digit result	20-digit result
x	0.9999954321	0.99999543210000000000
x^2	0.9999908642	0.99999086422086571041
$x^2 - x$	-0.0000045679	-0.00000456787913428959
$x^2 - x + 2$	1.999995432	1.99999543212086571041

Here in 10-digit arithmetic the $x^2 - x$ has lost half of its significant digits to cancellation, and is accurate to only 5 significant digits. But its missing 6^{th} significant digit would be in the 11^{th} place to the right of the decimal, so after adding 2 those missing digits for $x^2 - x$ don't have any effect on the computed value of $x^2 - x + 2$, and the final result is accurate to 10 significant digits.

Example 2. Cancellation, but it can be avoided with some algebra.

$$\sqrt{x^2 + 5x + 1} - \sqrt{x^2 + 3x + 1}$$

Solving $\sqrt{x^2 + 5x + 1} = \sqrt{x^2 + 3x + 1}$ for x to see where two nearly equal numbers might be subtracted shows that near $x = 0$ there could be cancellation.

Checking $x = 10^{-5}/3$ finds that the difference of the square roots is correct to only 4 significant digits using 10-digit arithmetic.

But there is another less obvious region where cancellation can occur. Suppose we are doing a sum or integral to infinity of a function that involves this expression. When x gets large, there can also be cancellation.

value	10-digit result	20-digit result
x	34567.12345	34567.123450000000000
$\sqrt{x^2 + 5x + 1}$	34569.62339	34569.623374066283028
$\sqrt{x^2 + 3x + 1}$	34568.62343	34568.623431920020591
$\sqrt{x^2 + 5x + 1} - \sqrt{x^2 + 3x + 1}$	0.9999600000	0.99994214626243700000

For this x, 6 significant digits have been lost to cancellation in this evaluation.

Doing some algebra to re-arrange this expression can get rid of the cancellation.

$$\sqrt{x^2 + 5x + 1} - \sqrt{x^2 + 3x + 1} = \frac{\left(\sqrt{x^2 + 5x + 1} - \sqrt{x^2 + 3x + 1}\,\right)\,\left(\sqrt{x^2 + 5x + 1} + \sqrt{x^2 + 3x + 1}\,\right)}{\sqrt{x^2 + 5x + 1} + \sqrt{x^2 + 3x + 1}}$$

$$= \frac{(x^2 + 5x + 1) - (x^2 + 3x + 1)}{\sqrt{x^2 + 5x + 1} + \sqrt{x^2 + 3x + 1}}$$

$$= \frac{2x}{\sqrt{x^2+5x+1}+\sqrt{x^2+3x+1}}$$

Now there are no subtractions in this expression, and all additions involve only positive numbers when $x > 0$, so there will be no cancellation.

value	10-digit result	20-digit result
x	34567.12345	34567.123450000000000
$\sqrt{x^2+5x+1}$	34569.62339	34569.623374066283028
$\sqrt{x^2+3x+1}$	34568.62343	34568.623431920020591
$2x \,/\, \left(\sqrt{x^2+5x+1}+\sqrt{x^2+3x+1}\,\right)$	0.9999421461	0.99994214626243665049

The 10-digit evaluation of the second form of the expression has an error of 2 units in the 10^{th} significant digit, consistent with normal rounding of the expression with no cancellation error.

When we did the algebraic simplification of the numerator above, we cancelled the two x^2 terms and 1 terms analytically instead of numerically. There were no numerical errors made when we did the algebra, so that was how we were able to make the cancellation go away.

Example 3. Hidden cancellation may still be present.

$\ln(x) - \ln(y)$

It is a mistake to think that re-formulating an expression to get rid of the minus sign will automatically get rid of the cancellation. It worked in the example above, but in that case we could see the step in the algebra where the cancellation went away.

When x is close to y, the "obvious" simplification in this case is to re-write the expression as $\ln(x/y)$.

value	10-digit result	20-digit result
x	53.12345678	53.123456780000000000
y	53.12342222	53.123422220000000000
$\ln(x) - \ln(y)$	0.0000006510000000	0.00000065056028610500
$\ln(x/y)$	0.0000006509997881	0.00000065056028610501

The evaluation of $\ln(x) - \ln(y)$ has indeed lost 7 digits to cancellation, and has only 3 correct significant digits. But the $\ln(x/y)$ has not gotten rid of the cancellation — it also has only 3 correct significant digits.

This mistake is subtle, there are even some numerical analysis textbooks that promote the myth that $\ln(x/y)$ will get rid of the cancellation.

What has happened is that we have just moved the cancellation inside the logarithm function. When x and y are close, x/y is close to 1. For the values above, $x/y = 1.000000651$.

For t close to 1, $\ln(t) = (t-1) - (t-1)^2/2 + (t-1)^3/3 - \cdots$. Here $t = x/y$ means $t - 1 = 0.000000651$, so rounding x/y to 10 digits and then subtracting 1 leaves $t - 1$ accurate to only 3 significant digits.

When we do not find any simple way to eliminate the cancellation, the brute force solution is to do the calculation at higher precision as we did above. Using 20 digits in this case gave about 13 digits correct.

(For experts only) Just because we did not find a way to compute $\ln(x) - \ln(y)$ accurately without raising precision does not mean there is no way to do it. Here is a method, due to W. Kahan:

z = (x−y) / y
r = 1 + z
if r=1 then w = 1 else w = ln(r) / (r-1)
w ∗ z

Doing this with 10-digit arithmetic for x and y above gives 0.0000006505602859, which is accurate to 9 digits.

Example 4. Use a trigonometric identity to remove cancellation.

$$1 - \cos(x)$$

This expression can have cancellation when $\cos(x)$ is close to 1. That means x is close to zero, $\pm 2\pi$, $\pm 4\pi$, $\pm 6\pi$, etc. We can try

$$\begin{aligned} 1 - \cos(x) &= \frac{(1 - \cos(x))\ (1 + \cos(x))}{1 + \cos(x)} \\ &= \frac{1 - \cos^2(x)}{1 + \cos(x)} \\ &= \frac{\sin^2(x)}{1 + \cos(x)} \end{aligned}$$

This form should help when x is close to zero, since $\sin(x) = x - x^3/6 + \cdots$ near $x = 0$. There is no further cancellation inside $\sin(x)$ as there was for the logarithm in the previous example.

However, for x near 2π or other multiples of 2π the $\sin(x)$ function will do an argument reduction by subtracting the nearby multiple of 2π from x before using the series. That subtraction puts the cancellation error back inside the sine function, so this re-formulation is used mostly for small x.

value	10-digit result	20-digit result
x	7.123456789e−4	7.1234567890000000000e−4
$1 - \cos(x)$	1.000000000e−10	7.7286978300000000000e−11
$\sin^2(x)/(1 + \cos(x))$	7.728697830e−11	7.7286978295133103092e−11

For this x, the $1 - \cos(x)$ form gets only one significant digit correct, while the other form has full 10-digit accuracy. For x much closer to zero than this, the $1 - \cos(x)$ returns zero, meaning all 10 digits have been lost.

One other feature of this cancellation "fix" is that if we need to compute this expression for a wide range of x values, we can't just always use the second form. Sometimes $\cos(x)$ is close to -1, making the denominator in the second form have cancellation while the first form is fine.

If we want a single formula that is good near zero and avoids problems when $\cos(x)$ is close to -1, there are lots of trigonometric identities. One is: $1 - \cos(x) = 2 * \sin^2(x/2)$. Using this formula with the x above gives 7.728697828e–11, good to 9 significant digits.

Example 5. Cancellation in the quadratic formula

The two roots of the quadratic equation $ax^2 + bx + c = 0$ are: $\dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

One possible cancellation is when b^2 is close to $4ac$. That makes the square root term small, which means the two roots are close together. Finding closely spaced roots is a problem that is known to be ill-conditioned, so we usually have to resort to higher precision to compute them accurately. Therefore we probably cannot use algebra to get an equivalent formula that is accurate.

The other possible cancellation is that one of the two roots will have opposite signs for the plus or minus outside the square root. The other root will have the same sign for that operation, so that root will be accurate. We can algebraically simplify the product of the two roots and then use the accurate one to get the other. Call the two roots x_1 and x_2.

$$x_1\, x_2 = \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(\frac{-b - \sqrt{b^2 - 4ac}}{2a}\right) = \frac{b^2 - (b^2 - 4ac)}{4a^2} = \frac{4ac}{4a^2} = \frac{c}{a}$$

Case 1. $b > 0$. Compute $x_2 = \dfrac{-b - \sqrt{b^2 - 4ac}}{2a}$ and $x_1 = \dfrac{c}{ax_2}$

Case 2. $b < 0$. Compute $x_1 = \dfrac{-b + \sqrt{b^2 - 4ac}}{2a}$ and $x_2 = \dfrac{c}{ax_1}$

Example 6. Use a series to analyze and treat cancellation in $f(x)$ near $x = 0$.

An article in the October, 2022 issue of Mathematics Magazine shows some functions of interest to physicists that need to be examined near zero. Two of these functions are

$$f(z) = \frac{2\big((1 - z^2)^{-1/2} - 1\big)}{z^2} \quad , \qquad \eta(\omega) = \frac{3}{\omega}\left(\frac{1}{\tanh(\omega)} - \frac{1}{\omega}\right)$$

They try plotting these two functions from 0 to 10^{-7} using 16-digit double precision arithmetic. The plots come out looking messy and not very continuous near zero.

We can see that in both cases there is cancellation near zero. Trying to plot either function this close to zero means the computed function values are just roundoff noise.

Instead of trying to algebraically remove the cancellation, this time we will look at the Taylor series expanded about zero. Either Wolfram Alpha, available on the internet, or the Mathematica computer algebra system will give us these two series using the commands

Series[2*((1 - z^2)^(-1/2) - 1)/z^2, {z, 0, 9}]

gives $1 + \frac{3z^2}{4} + \frac{5z^4}{8} + \frac{35z^6}{64} + \frac{63z^8}{128} + \cdots$

Series[(3/w)*(1/Tanh[w]-1/w), {w, 0, 9}]

gives $1 - \frac{w^2}{15} + \frac{2w^4}{315} - \frac{w^6}{1575} + \frac{2w^8}{31185} + \cdots$

For both of these series, we can get full 16-digit accuracy for any input within 10^{-7} of zero by using just the first two terms. There will be no cancellation in either series when z or w is that close to zero.

value	10-digit result	20-digit result
z or ω	7.123456789e−8	7.1234567890000000000e−8
$f(z)$	0.000000000	1.00000716100133888010
$1 + 3z^2/4$	1.000000000	1.00000000000000380580
$\eta(\omega)$	0.000000000	0.99996395163084353483
$1 - \omega^2/15$	1.000000000	0.99999999999999966171

For this input, both the original forms for the two functions lose all their digits to cancellation when carrying 10 digits. The 20-digit results show that $f(z)$ has lost about 15 digits to cancellation, and $\eta(\omega)$ has lost about 16 digits. Both of the 2-term series approximations give results correct to full precision.

Example 7. Use a series for more serious cancellation of $f(x)$ near $x = 0$.

$$f(x) = \frac{\tan(\sin(x)) - \sin(\tan(x))}{x^7}$$

This is the example mentioned on the main Calc-50 page. It suffers from more loss of accuracy near $x = 0$ than the other expressions above. Looking at the series can show how many digits will be lost for different x's, as well as giving an accurate approximation for small x.

Series[(Tan[Sin[x]] - Sin[Tan[x]])/x^7, {x, 0, 9}]

gives $\frac{1}{30} + \frac{29x^2}{756} + \frac{1913x^4}{75600} + \frac{95x^6}{7392} + \frac{311148869x^8}{54486432000} + \cdots$

Suppose we want at least 10 significant digit accuracy near zero and we carry 16 digits to evaluate $f(x)$. As x gets closer to zero, at what point do we need to switch from the original formula to the series? Looking at the three series for $\tan(\sin(x))$, $\sin(\tan(x))$, and $\tan(\sin(x)) - \sin(\tan(x))$ can answer that question.

Series[(Tan[Sin[x]]), {x, 0, 9}]

gives $x + \frac{x^3}{6} - \frac{x^5}{40} - \frac{107x^7}{5040} - \frac{73x^9}{24192} + \cdots$

Series[Sin[Tan[x]]), {x, 0, 9}]

gives $x + \frac{x^3}{6} - \frac{x^5}{40} - \frac{55x^7}{1008} - \frac{143x^9}{3456} + \cdots$

Series[Tan[Sin[x]] - Sin[Tan[x]], {x, 0, 9}]

gives $\frac{x^7}{30} + \frac{29x^9}{756} + \cdots$

For small x, both terms in the subtraction are about the size of x, but their difference is about $x^7/30$. If $x = 10^{-2}$, then $x^7/30$ is about $10^{-15}/3$. That means the formula would have lost about $-2 - (-15) = 13$ significant digits of accuracy.

Generalizing this to get a formula for the number of digits lost to cancellation for this function, we could say

$$c(x) = \log\left(\frac{x}{(x^7/30)}\right) = \log\left(\frac{30}{x^6}\right)$$

The log is the base 10 logarithm, since we want the number of base 10 digits lost. Checking with $x = 10^{-2}$ gives $c(0.01) = 13.5$ digits lost, which agrees with our previous estimate.

To answer the question about where to switch to the series when we want 10 accurate digits while doing the arithmetic with 16, we can tolerate losing up to 6 digits in the formula, so solve the equation

$$\log\left(\frac{30}{x^6}\right) = 6$$

The solution is about $x = 0.176$. For smaller x we would use the series.

If we were using Calc-50 instead of double precision on a computer, we might want 16 digits accurate from the formula while carrying 50 digits. That means we could afford to lose up to 34 digits to cancellation.

Solving $c(x) = 34$ gives $x = 3.8\text{e}{-6}$ as the point to switch to the series.

Example 8. Cancellation while summing a series.

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Spotting cancellation in a sum can be trickier, since it can occur gradually and not just in a single addition or subtraction.

Here there is no cancellation for $x \geq 0$, since the terms are all positive. When $x < 0$ half the terms are negative, so cancellation is possible.

Theoretically, the e^x series converges for all x. In practice, when $|x|$ is large there can be problems.

Consider $x = -20$. We think of convergent series as having terms that decrease in size as we use more terms. But for $x = -20$ the terms increase in size until the 20^{th} term, and only after that do they decrease.

The largest term in this sum is $(-20)^{20}/20! = 4.31\text{e+}7$. We know that $0 < e^x < 1$ for negative x, so we can see there must be at least 7 digits lost to cancellation in this sum.

It is actually worse than that, since the final value of the sum is about 2.06e−9. That means we expect around 16 digits lost to cancellation.

For e^x we can use an identity to avoid cancellation: $e^x = 1/e^{-x}$. This sums the series accurately for $x = 20$, then takes the reciprocal.

value	10-digit result	20-digit result
x	−20.00000000	−20.000000000000000000
$1 + x + \frac{x^2}{2!} + \cdots$	−1.211525237e−2	2.0622936270509912793e−9
$1/(1 + (-x) + \frac{(-x)^2}{2!} + \cdots)$	2.061153623e−9	2.0611536224385578280e−9

Using the series with $x = -20$ with 10-digit arithmetic loses all the digits, even giving a negative answer when we know e^x is always positive. The reciprocal series where $-x$ is positive gets 9 digits correct.

Comparing the two 20-digit results shows that the original series loses 17 digits to cancellation.

For more complicated series where there is no known identity or algebraic trick to get rid of the cancellation, doing the sum with two different precisions as we did here can reveal how many digits are being lost for a given x.

With Calc-50 the user can't change precision, but the sum key returns an estimated relative error that also gives an estimate for the lost accuracy. See example 5 on the "Infinite sums" page.

The "Experimental math" and "Asymptotics" pages have more examples where cancellation error needs to be recognized and dealt with.