Solving a quadratic equation with precision when using floating point variables

I know how to solve a basic quadratic equation with the formula

$$t_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

but I learned that if $b \approx \sqrt{b^2 - 4ac}$ *floating point* precision may give slightly wrong results and this approach is better. It works, indeed. But why? Is there a simple explanation on why this works?

(analysis) (quadratics) (floating-point)

edited Aug 3 '17 at 10:06 Rodrigo de Azevedo 10.6k 4 13 46

asked Feb 22 '13 at 18:09 John Smith 325 3 11

4 Answers

We are accustomed to solving the equation $ax^2 + bx + c = 0$ by using the *Quadratic Formula*

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Alternately, we can use the *Citardauq Formula*

$$x = \frac{2c}{-b \mp \sqrt{b^2 - 4ac}}.$$

When $|4ac|$ is small in comparison with $|b|$, one of the roots as computed by the Quadratic Formula may suffer serious loss of precision, because in the numerator we are finding the difference of large nearly equal quantities. Precisely that root is then nicely computed with no loss of precision by the Citardauq Formula.

A simpler way of dealing with the problem is to note that the product of the roots of $ax^2 + bx + c = 0$ is c/a. So if we can compute one of the roots r_1 to high precision, the other root r_2 can be computed from $r_1 r_2 = c/a$ with no loss of precision.

edited Jun 20 '17 at 17:42 pyrrhic 258 1 9

answered Feb 22 '13 at 18:34 André Nicolas 440k 35 394 770

Thank you and sorry for making you post this twice (math.stackexchange.com/questions/56942/…) – John Smith Feb 22 '13 at 18:44

4 No problem. For me, *finding* would require much more time than rewriting. – André Nicolas Feb 22 '13 at 18:49

Any time you subtract two nearly equal numbers there is loss of precision. Imagine we work in scientific notation with five decimal digits available. The number 10100 is represented as $1.0100E5$ and you can think of it having a potential error of ± 0.5 as anything from 10099.5 to 10100.5 would get the same representation. If we now subtract it from $10200 = 1.0200E5$ we get $100 = 1.0000E2$ but the error could be ± 1 (in the worst case). We only have three digits of precision now, instead of 5, though it will be represented still with 5 digits.

For the specific case of the quadratic formula, if $4ac \ll b^2$, you have $-b + \sqrt{b^2 - 4ac} = -b + b\sqrt{1 - \frac{4ac}{b^2}} \approx -b + b(1 - \frac{2ac}{b^2}) = -\frac{2ac}{b}$ and the much larger b has canceled, losing precision.

answered Feb 22 '13 at 18:18
Ross Millikan
256k 21 175 322

Thank you, any chance to proof that the two methods are equivalent except empirically by testing it with values? – John Smith Feb 22 '13 at 18:26

@JohnSmith: empirically testing will not prove them equivalent. The idea is that you compute one root by choosing the $\pm$ sign so the two terms are the same sign and add. Then you use the fact that if $r1, r2$ are the two roots of the quadratic, they multiply to $\frac{c}{a}$-the quadratic factors as $a(x-r1)(x-r2) = ax^2 - a(r1+r2)x + ar1r2$ – Ross Millikan Feb 22 '13 at 18:30

If you have two roots r_1 and r_2, then the quadratic can be written in factored form $a(x-r_1)(x-r_2) = 0$. Expanded we get $ax^2 - a(r_1+r_2)x + ar_1r_2 = 0$. This is basically Vieta's formulas consolidated into one equation. As you can see, given any quadratic $ax^2 + bx + c = 0$, the coefficients are $b = -a(r_1 + r_2)$ and $c = ar_1r_2$, or as Vieta put it, $-\frac{b}{a} = r_1 + r_2$ and $\frac{c}{a} = r_1r_2$.

Ignore the first of the two and focus on the second. We can express one root in terms of the other as $\frac{c}{ar_1} = r_2$. This should not be unreasonable to you.

This is not an approximate relationship, its an exact relationship and should work all the time, not just when $b^2 \gg 4ac$. This argument, though, is from a pure-mathematical perspective.

answered Feb 22 '13 at 18:39
CogitoErgoCogitoSum
1,146 1 7 19

As the other comments mention, there is loss of precision when subtracting similar values. So you want to compute one root r_1 of $ax^2 + bx + c = 0$ *without* substraction of possibly similar values and get the other from the relation $r_2 = \frac{c}{ar_1}$. If $4ac$ is small with respect to b^2, $\sqrt{b^2 - 4ac} \approx b$. So the general strategy is:

- If $b < 0$, compute $r_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$
- If $b > 0$, compute $r_1 = \frac{-b-\sqrt{b^2-4ac}}{2a}$
- If $b = 0$, $r_1 = \sqrt{\frac{c}{a}}$, $r_2 = -\sqrt{\frac{c}{a}}$

edited Jun 16 '15 at 9:28
lumbric
275 1 9

answered Feb 22 '13 at 22:42
vonbrand
19.2k 6 29 55