Numerical Computations on the Computer

Limits of Integers

Typically in C++ we use type int (or sometimes type long) for our integers. These integer types have a maximum possible value. This is dependent on both the type of computer you use and on the compiler you use. With Visual Studio on the PCs in the CIS lab, the largest int value we can get is:

$$2,147,483,647$$

One amazing fact is that if we add 1 to the largest positive integer we get the smallest negative integer! Similarly, if we subtract 1 from the smallest negative integer we get the largest positive one. It's as if the integers wrap around in one big circle.

By default most compilers do not check for integer overflow. Your programs will happily add 2 large positive numbers and produce a negative answer because of overflow. Not only does this give a completely wrong answer, but in some cases it causes a security flaw in your software that could allow an attacker to break into the computer on which your software is running -- especially if it is web application software handling users from the Internet.

It is possible to have your code generate an "exception" if integer overflow occurs and to have exception handler code that says what to do in such a case. That is beyond the scope of this course, however.

Limits of Floats

In C++ we use the float or double type to handle what in math we would call real numbers. A float is a single-precision floating point number and a double is a double-precision floating point number. Doubles use more bits and so can represent any given real number more accurately than a float.

As with integers, there is a largest float value. On our systems it is around

10^{38} (which is a 1 followed by 38 zeros)

That is a very large number, but if your calculations might exceed that, then you should switch to doubles. Of course, you might even exceed the size of the largest double!

Another feature of floats or doubles is that there are gaps between them. They do not form a continuous line like reals in mathematics. Furthermore, the size of the gaps vary, with the gap getting larger the further you get from 0. It is at best a poor imitation of the real number line!

This means that there are many decimal numbers that cannot be represented exactly as a float (which is stored in binary, of course).

Example: $0.4 = \frac{4}{10} = \frac{2}{5}$ is a nice easy number to represent in decimal.

However in binary it is: $A = 0.011001100110$ with the 0110 repeating forever

It is somewhat hard to derive this from 0.4, but we can easily show that A is correct:

Since $A = 0.011001100110$

$16A = 2^4 \cdot A = 0110.011001100110$ since multiplying by 2 shifts the number left

Now subtract: $A = 0.011001100110$

$15A = 0110. = 6$

Thus $A = \frac{6}{15} = \frac{2}{5} = \frac{4}{10} = 0.4$

Since it takes an infinite sequence of bits to represent the number 0.4 exactly, we cannot do this on a computer. We have to cut off the bits somewhere and make do with an approximate value.

Ill-conditioned Problems

Suppose you needed to calculate (by calculator or computer) some y values for this function:

$$y = \frac{1}{x - 2.95708}$$

Thus for a given x value, you use this formula to get the desired y value.

However, there can be some error in the input value for x. Part of this error could be round-off error like that described above for the number 0.4, but it can also include some error due to the fact that there is some inaccuracy in any measurement we make in the real world. For example, if we measure the length of an object, we cannot pin that down exactly. Similarly, if we need to measure a voltage we will get an approximate value due to the limits of our measuring instruments.

For the sake of this example, suppose we have an item whose x value happens to be exactly 2.9571. We can't measure it that accurately, so we do not really know this number. Our measured value is x = 2.957, let's say.

The correct y value would be $y = \frac{1}{2.9571 - 2.95708} = \frac{1}{.00002} = 50,000$

However, using our meaured x value we get a vastly different y value:

$$y = \frac{1}{2.957 - 2.95708} = \frac{-1}{.00008} = -12,500$$

Just a slight change in the input value x produces a huge change in the output value y.

The error in our x value is .0001 and the relative error is $\frac{.0001}{2.9571} = .0034\%$

The error in y is $50,000 - (-12,500) = 62,500$

and the relative error in y is $\frac{62,500}{50,000} = 1.25 = 125\%$

This huge relative error in y is due to the subtraction in the denominator for y of items that are almost equal. This gives "catastrophic cancellation" in which we lose most or all of our significant digits. That is why we can get such wildly different outputs for inputs which are very nearly equal.

In your computer programs watch out for calculations that might be similarly ill-conditioned!

Solving Quadratic Equations

Background:

These are equations of the form $ax^2 + bx + c = 0$ where a is not 0.

A typical solution method is to use the quadratic formula, which gives the following solutions.

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

If the discriminant $b^2 - 4ac$ is zero, then there is really only one solution, namely

$$x_1 = \frac{-b}{2a}$$

If the discriminant is positive, there are 2 solutions, namely the x1 and x2 noted above.

If the discriminant is negative, then there are no real solutions. (Any solutions would be non-real complex numbers.)

Example 1:

$3x^2 + 5x - 2 = 0$

Thus a = 3, b = 5, and c = -2. The discriminant is $5^2 - 4 \times 3 \times -2 = 49$, which is positive, so there are 2 solutions.

The quadratic formula gives these 2 solutions as follows:

$$x_1 = \frac{-5 + \sqrt{49}}{2 \times 3} \qquad x_2 = \frac{-5 - \sqrt{49}}{2 \times 3}$$

$$= \frac{-5 + \sqrt{49}}{6} \qquad = \frac{-5 - \sqrt{49}}{6}$$

$$= \frac{2}{6} = \frac{1}{3} \qquad = \frac{-12}{6} = -2$$

We can also easily solve this quadratic equation by factoring:

$3x^2 + 5x - 2 = 0$

$(3x - 1)(x + 2) = 0$

$3x - 1 = 0$ or $x + 2 = 0$

$x = \frac{1}{3}$ $x = -2$

Example 2:

$x^2 + 6x + 9 = 0$

Thus a = 1, b = 6, and c = 9.

Since the discriminant is $6^2 - 4 \times 1 \times 9 = 36 - 36 = 0$ there is only 1 answer:

$x_1 = \frac{-b}{2a} = \frac{-6}{2} = -3$

You can also see that -3 is the only solution by factoring the quadratic:

$x^2 + 6x + 9 = 0$

$(x + 3)(x + 3) = 0$

$x + 3 = 0$ or $x + 3 = 0$

$x = -3$ $x = -3$

Example 3:

$2x^2 + 3x + 4 = 0$

Thus a = 2, b = 3, and c = 4.

Since the discriminant is $3^2 - 4 \times 2 \times 4 = 9 - 32 < 0$ there are no real solutions.

Problems with using the quadratic formula

In certain cases, the quadratic formula can suffer from "catastropic cancellation" in finding one of the two solutions. This means that that solution might be very inaccurate. However, there is a way to calculate that solution so that it is more accurate.

See the following web page for detailed information:

http://cis.stvincent.edu/carlsond/cs270/quad/quadratics.html

Also see the example program QuadEqSmart.cpp on your course network drive.