Notes of a Numerical Analyst

Floating point numbers and physics

NICK TREFETHEN FRS

The laws of classical continuum mechanics describe the motion and deformation of fluids and solids. They involve quantities such as density, pressure, and temperature, and they are written as partial differential equations. Of course, these laws are approximations, for the world is not continuous but is made of discrete atoms and molecules. Density is an average, pressure is an average, temperature is an average. But this is the right thing to do for most applications in science and engineering: to ignore the molecules and regard the physical world as continuous.

Physicists understand very well the "implementation details" by which the continuum is built up from discrete particles. For example:

• *If you halve the volume of a box, keeping temperature constant, the pressure of a gas inside doubles. Reason: twice as many impacts of molecules per unit cross-section per unit time.*

• *If you double the temperature of a box, keeping volume constant, the pressure doubles. Reason: the momentum of each particle increases by a factor of $\sqrt{2}$, since energy and temperature scale with velocity-squared; the number of impacts per unit cross-section per unit time also goes up by $\sqrt{2}$.*

How fine is the physical continuum? The famous Avogadro's number, about 6×10^{23}, is the number of molecules in a mole. There are about 50 moles of gas in a cubic meter at ordinary conditions, so this comes to about 3×10^{25} molecules per cubic meter in a gas (Loschmidt's constant). The cube root of 3×10^{25} is about 3×10^{8}. Thus there are about 3×10^{8} molecules per linear meter in an ordinary gas. For a solid, the figure is about ten times higher: 3×10^{9}. Thus, roughly speaking,

A gas or solid has around 10^9 particles per meter.

This is how fine the discretisation is in our physical world. It's interesting to compare it with the floating-point arithmetic on our computers. In the IEEE double-precision standard that has prevailed since the 1980s, the real line is discretized by $2^{52} \approx 10^{16}$ numbers between 1 and 2, the same between 2 and 4, and so on. Thus we find:

Computer arithmetic is a million times finer than physics.

If I gave floating point arithmetic coordinates to the desk I'm sitting at, for example, I would find there were around a million coordinate points between each adjacent pair of molecules. In fact, 10^{16} is more or less the number of molecules I'd encounter in a line going all the way through the earth from here to New Zealand.

Figure 1. Turing Award winner William Kahan of UC Berkeley, the man behind the IEEE floating-point arithmetic standard.

Another angle on the extraordinary resolution of floating point numbers is the fact that in the physical world, essentially nothing is known to 16 digits of accuracy. Quantities like the electron mass or the gravitational constant are known to between 5 and 12 digits. Well, the speed of light is *exactly* 299,792,458 meters per second! — but only because the meter is defined thereby.

FURTHER READING

[1] J.-M. Muller et al., *Handbook of Floating-Point Arithmetic*, 2nd ed., Springer, 2018.
[2] *Constants, Units, and Uncertainty*, `physics.nist.gov/cuu/Constants/`.
[3] M. L. Overton, *Numerical Computing with IEEE Floating Point Arithmetic*, SIAM, 2001.

Nick Trefethen
Trefethen is Professor of Numerical Analysis and head of the Numerical Analysis Group at the University of Oxford.