Floating Point Numbers

Most numbers cannot be represented in a computer. Those that are not representable are approximated by a relatively small few that are. We will let the floating point approximation of x be called the *float* of x and write it as $\mathrm{fl}(x)$. A floating point number represents the whole interval of reals near it. We can bound the length of this interval, and therefore the relative error that is made when approximating a number by its float. We assume that floating point numbers have the form

$$\bar{x} = \pm 0.b_1 b_2 \ldots b_t \times 2^e, \quad \text{where} \quad e_n \leq e \leq e_p \ \text{ and } \ b_k \text{ is 0 or 1, but } b_1 = 1.$$

Think of it as a (base-2) fraction times 2^e. Numbers too |large| for this representation are said to *overflow*, and numbers too |small| are said to *underflow*. Since we have allotted t bits for the fractional part, the distance between $\bar{x}$ and an adjacent float is no more than 2^{e-t}. Dividing this by $\bar{x}$ gives an upper bound on the relative distance between any two floats: 2^{1-t}. We define the *machine precision*, μ, to be half of this quantity: For a floating point system with a t bit fractional part, the machine precision is $\mu = 2^{-t}$.

The Floating Point Representation Theorem.
Suppose x is a real number which is in the range of the floating point system (doesn't underflow or overflow). Then

$$\mathrm{fl}(x) = x(1+\epsilon), \quad \text{where} \ \ |\epsilon| \leq \mu$$

This is a statement about relative error, and can also be written as

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \leq \mu.$$

The set of floats is not closed under arithmetic operations. For example, when we add two floats, the result is not necessarily a float, but will instead be represented by its float. Computers today follow an industry standard called the IEEE 754, which among many other things guarantees the following:

The Fundamental Axiom of Floating Point Arithmetic.
Let $x \ op \ y$ be some arithmetic operation. That is, *op* is one of $+$, $-$, $\times$ or $\div$. Suppose x and y are floats and that $x \ op \ y$ doesn't underflow or overflow. Then

$$\mathrm{fl}(x \ op \ y) = (x \ op \ y)(1+\epsilon), \quad \text{where} \ \ |\epsilon| \leq \mu$$

The geometry is simple: When doing arithmetic with floats, we always get the float closest to the answer. But be careful: this is a statement about floats; real numbers need to be represented by floats before we can do the arithmetic!