

Floating Point Numbers

Most numbers *cannot* be represented in a computer. We are forced to use approximations which *can* be represented on the computer. Let the floating point approximation of x be called the float of x and write it as $\text{fl}(x)$. In our floating point number system, we can give a bound on the relative error that is made when approximating a number by its float. We assume that our floating point numbers have the form

$$\bar{x} = \pm 0.b_1b_2 \dots b_t \times 2^e, \quad \text{where } e_n \leq e \leq e_p \text{ and } b_k \text{ is 0 or 1, but } b_1 = 1.$$

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 $\hat{x}$ and an adjacent float is no more than 2^{e-t} . Dividing this by $\hat{x}$ 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 t a digit 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

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

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

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

The set of floats is not closed under our arithmetic operations. For example, when we add two floats, the result is not necessarily a float, and it will need to be approximated by another float. Computers today almost always satisfy the following rule:

The Fundamental Axiom of Floating Point Arithmetic.

Let $x \text{ op } y$ be some arithmetic operation. That is, op is one of $+$, $-$, $\times$ or $\div$. Suppose x and y are floats and that $x \text{ op } y$ doesn't underflow or overflow. Then

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

Notice that this is a statement about floats. Real numbers need to be represented by floats before we can do the arithmetic!