Floating point Algebra

(This is just me working through Robin Leroy's essay from the Principia library. Nothing innovative here from me...)

Base Exponential form

(In this section, we assume that a is positive, or that a can be replaced with its absolute value, $|a|$. If a is negative, then the following analysis is be applied to its absolute value, and the sign is to be tracked separately.)

First define the fractional part of a real number a as

$$\text{frac } a \equiv a - \lfloor a \rfloor \in [0, 1).$$

The fractional part is truly fractional and is always less than one, and it is zero when there is no fractional part.

Next, consider the term $\lfloor \log_2 a \rfloor$, which tells you the power-of-two lower bound of a. For example, $\lfloor \log_2 15.5 \rfloor$ is 3, since it is bounded by 2^3 and 2^4. The remaining bit of a is the part which cannot be expressed as a power of two, and is the "fraction" with respect to this factor.

Using these two operators, we seek to define a by a representative exponent and a fractional correction. Begin with the following decomposition,

$$a = 2^{\log_2 a} = 2^{\lfloor \log_2 a \rfloor} 2^{\text{frac}(\log_2 a)}.$$

Next, split the fractional correction so that

$$a = 2^{\lfloor \log_2 a \rfloor} \left(\lfloor 2^{\text{frac}(\log_2 a)} \rfloor + \text{frac}\left(2^{\text{frac}(\log_2 a)} \right) \right).$$

Since $\text{frac}(\ldots) \in [0, 1)$, the first term is always one, and

$$a = 2^{\lfloor \log_2 a \rfloor} \left(1 + \text{frac}\left(2^{\text{frac}(\log_2 a)} \right) \right).$$

The final term can be rewritten as

$$\begin{aligned} \text{frac}\left(2^{\text{frac}(\log_2 a)} \right) &= \text{frac}\left(2^{\log_2 a - \lfloor \log_2 a \rfloor} \right) \\ &= \text{frac}\left(2^{-\lfloor \log_2 a \rfloor} a \right). \end{aligned}$$

This form makes explicit that the fractional term is the value of a after it has been rescaled by $2^{-\lfloor \log_2 a \rfloor}$, so that it is between $2^{\lfloor \log_2 a \rfloor}$ and $2^{\lfloor \log_2 a \rfloor + 1}$.

We could also simply expand the frac operator and write the fractional part of a as a single algebraic expression,

$$\begin{aligned}\text{frac}\left(2^{-\lfloor \log_2 a \rfloor}\right) &= 2^{-\lfloor \log_2 a \rfloor} a - \lfloor 2^{-\lfloor \log_2 a \rfloor} a \rfloor \\ &= 2^{-\lfloor \log_2 a \rfloor} a - 1.\end{aligned}$$

Bringing it all together, we can write any real number a as

$$\begin{aligned}a &= 2^{\lfloor \log_2 a \rfloor} \left(1 + \text{frac}\ 2^{-\lfloor \log_2 a \rfloor}\right) \\ &= 2^{\lfloor \log_2 a \rfloor} \left(1 + \left(2^{-\lfloor \log_2 a \rfloor} a - 1\right)\right)\end{aligned}$$

which is the standard form of a modern floating point number.

This does not yet represent an actual floating point number. For example, we must also consider the truncation of the fractional part as the exponent increases. But all floating point numbers can be represented by an expression of this form.

Fixed point form

Each floating point number can therefore be represented by three numbers:

- A *base* for the exponent, b, which is almost universally set to two in both computing hardware and theoretical analysis,
- an *exponent*, e, denoting the value for which $2^e \leq |a|$. This is equal to $\lfloor \log_2 |a| \rfloor$.
- The *fraction*, f, which

Not only can the usual floating

First consider a number a and its computational counterpart A.

$$A \equiv b + \lfloor \log_2 a \rfloor + \text{frac}$$

TODO...