Solutions to Worksheet on Decimal and Binary Representations

The names like "System B16" are mine. Note B16 is not exactly used on any real machine. However, with slight modifications, B16 is the IEEE-standard **binary16** ("half precision") system. Modern computers support two systems, **binary32** ("single precision") and **binary64** ("double precision"), as described in section 1.3. MATLAB uses **binary64** by default.

System D-. **a)** The largest is $x = +9.99 \times 10^{+9} \approx 10^{10}$ while the smallest is $-x$.

b) The smallest positive representable number is $+0.01 \times 10^{-9} = 10^{-11}$.

c) Do the same question for System D, below, first. For System D- the counting question is difficult because of numbers of the form $\pm 0.XY \times 10^{\pm Z}$ and $\pm 0.0X \times 10^{\pm Z}$. They are called "subnormal" in the literature. They can, *for most but not all values of* Z, be represented by $\pm X.Y0 \times 10^{\pm(Z\pm1)}$ and $\pm X.00 \times 10^{\pm(Z\pm2)}$, respectively. *Extra credit* for getting the precise count.

System D. **a)** Same as for D-.

b) $+1.00 \times 10^{-9} = 10^{-9}$.

c) Each different choice of symbols in the six locations corresponds to a distinct number. There are 2 choices for $\Diamond$, 9 choices for $\hat{\square}$, and 10 choices for $\square$. The number of distinct numbers is

$$2 \cdot 9 \cdot 10 \cdot 10 \cdot 2 \cdot 10 = 36000.$$

d) Unlike System D-, you cannot represent zero in System D. You can add it as an exception, like "$+0.00 \times 10^{+0}$."

System B16. **a)** The largest is $+1.111111111_2 \times 2^{+1111_2} = (2 - 2^{-10}) \times 2^{15} = 65504$. Note this is approximately $2^{16} = 65536$.

b) $+1.000000000_2 \times 2^{-1111_2} = 2^{-15}$.

c) Again, each different choice gives a distinct number:

$$2 \cdot 1 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \times 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^{15} = 32768.$$

d) As with System D, you cannot represent zero in System B16. You can, and **binary16** does, add it as an exception. (*I believe that any number with all zeros in the exponent is an exception, and that if all 16 bits are zero then the number is zero. But I am no expert on IEEE 754-2008. Look it up.*)

e) $\Diamond = (-1)^{\square}$.

f) The smallest representable number larger than one is

$$x = +1.000000001_2 \times 2^{+0000_2} = 1 + 2^{-9}.$$

Thus the gap is $\epsilon = x - 1 = 2^{-9}$. This number says, essentially, that System B16 gives 9 binary digit accuracy, or about 3 decimal digits.

g) The inefficiency is that there is no reason to store a 1 in the $\hat{\square}$ location because it is always a 1. So don't store it, and add a little more accuracy by extending the "mantissa" to 10 bits:

$$\Diamond 1.\square\square\square\square\square\square\square\square\square\square_2 \times 2^{\Diamond\square\square\square\square_2}.$$