Numerical Analysis I

p397
MRO'D

Solution Notes – Question A 2002

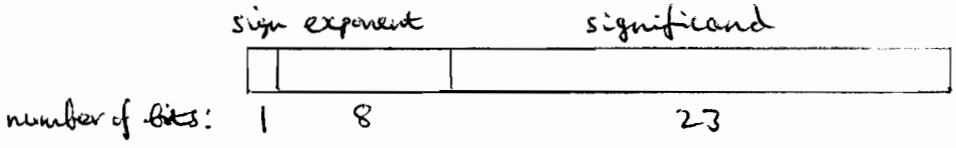
Question concerns: Floating-point arithmetic, IEEE arithmetic.

(a) A number is represented as

$$\pm\, d_0 . d_1 d_2 \ldots d_{p-1} \times \beta^e$$

where β is the base of the representation, p is the number of digits (of base β) of precision and e_{max}, e_{min} are the maximum and minimum values of the exponent e. [3 marks]

(b) IEEE single precision has the following layout of bits:



$\beta = 2$, $p = 24$ (including the hidden bit),

$e_{max} = +127$, $e_{min} = -126$. [5 marks]

(c) IBM System/370 arithmetic uses $\beta = 16$, so there can be no hidden bit. Each significant hexadecimal digit requires 4 bits, so

$$p = \frac{32 - 7 - 1}{4} = 6.$$

The total exponent range is 16^7, so

$$e_{max} = 16^6,$$
$$e_{min} = -16^6 + 1.$$

[5 marks]

(d) 6.789 rounds to 6.79
6.785 " " 6.78
6.755 " " 6.76 [3 marks]

(e) 01101011 0 rounds to 01101011
10111010 1 " " 10111010
11010010 0 " " 11010010
01111111 1 " " 10000000 [4 marks]