Numerical Analysis I

(a) For IEEE Double Precision $\beta = 2$, $p = 53$, $e_{\text{min}} = -1022$, $e_{\text{max}} = 1023$. Explain the meaning of these parameters and deduce the number of bits required to store the sign, exponent and significand. How many bytes are required in total? [5 marks]

(b) What is the hidden bit and what is its value for normalised numbers, and for denormal numbers? [2 marks]

(c) Define machine epsilon $\epsilon_m$. What is its value for IEEE Double Precision? [3 marks]

(d) Suppose $f(x) = O(1)$, $f'(x) = O(1)$ and

$$\frac{f(x + h) - f(x)}{h}$$

is to be used with IEEE Double Precision to estimate $f'(x)$ and $f''(x)$. State what value of $h$ you would use in each case, and what absolute accuracy (as a power of 2) you would expect to achieve. [4 marks]

(e) Special purpose floating-point hardware is to be designed with the following specification. Each number is to occupy 6 bytes but otherwise obey the principles of IEEE arithmetic as far as possible. The arithmetic must be sufficiently accurate that second derivatives can be computed to an absolute accuracy of $10^{-3}$ if $f(x) = O(1)$, $f'(x) = O(1)$. Deduce the parameters of this arithmetic. [Hint: $10^{-3} \approx 2^{-10}$ is sufficiently accurate.] [6 marks]