2005 Paper 10 Question 7

Numerical Analysis I

(*a*) The parameters for *IEEE* Single Precision are: $\beta = 2$, $p = 24$, $e_{min} = -126$, $e_{max} = 127$. Explain the terms *significand*, *sign bit*, *exponent*, *normalised number*, *denormal number*, *hidden bit*, *precision* as used in *IEEE* Single Precision. [7 marks]

(*b*) Let ω represent any of the operations $+ \quad - \quad * \quad /$. Let x be a positive finite representable number. List what each of the following evaluates to for each operation:

$$(+\infty)\ \omega\ x$$
$$x\ \omega\ (-\infty)$$

[Show the sign of your answer in each case.] [4 marks]

(*c*) Suppose the principles of *IEEE* arithmetic are applied to a floating-point representation with 6 bytes (48 stored bits). If $\beta = 2$, $e_{max} = 511$ and a hidden bit is used, deduce the values of e_{min} and p. [4 marks]

(*d*) Define *machine epsilon* ε_m. [1 mark]

(*e*) The function

$$f(x) = \frac{(x+1)^2}{x^2+1}$$

is to be evaluated using *IEEE* arithmetic for $x \geq 0$. Re-write the formula so that $f(x)$ can be evaluated in the case where x is representable but x^2 overflows. What does your formula evaluate to in the case that $(1/x) < \varepsilon_m$? [4 marks]