YOUR NAME:

Numerical Analysis, Quiz # 2 with Answers, April 16, 2015.
Give brief explanations of your answers.
Cross out what is not meant to be part of your answers.

Recall that double precision floating point numbers are of the form

$$1.b_1b_2\ldots b_{52}\times 2^E.$$

1. (a) Let a and b be real numbers, not necessarily part of the double precision floating point system. What is the best we can say about the relative accuracy of $c = a^2 + b^2$?

   (b) What is the value of c if $a = b =$the largest of all finite IEEE floating point numbers?

****

(a) The answer will be of excellent quality. Recall that the way to think about the evaluation of arithmetic expressions is that each step is computed exactly and then rounded. Rounding replaces the exact result by one of the two nearest number in the floating point system.

The steps are as follows. a and b are rounded which introduces relative errors bounded by 2^{-54}. The two results are squared; each of the squares have a relative error which is bounded by 3×2^{-54}. Adding two numbers of the same sign is also fine. The result of the addition is of course rounded but in all the relative error is bounded by 7×2^{-54}.

(b) The standard rounding rule will give the value $+\infty$ since in this case c clearly will exceed the largest finite positive number of the floating point system.

2. Consider the integers $I_k := 5^k$, where k is a nonnegative integer. As k increases, there comes a k such that I_k no longer can be represented as a double precision floating point number. Explain how we can determine this value of k and give an estimate of the smallest such k. It might be useful to know that $10^{0.3010}$ is a close approximation of 2.
****

Given that $10 = 5 \times 2$ this is really almost the same as one of the questions in the home work set.

Let us first note that all powers of 5 are odd integers. Also note that we can rewrite any floating point number as

$$1b_1b_2 \ldots b_{52} \times 2^{E-52}.$$

Therefore there will be an error when 5^k results in a value of E that exceeds 52. This is so since the gap between consecutive floating point numbers then will be at least 2 and we can no longer find any odd numbers.

From the fact that $10^{0.3010}$ is a close approximation of 2, we find that $2^{3.222}$ is approximately equal to 10 and that $2^{2.222}$ approximates 5 and that we are OK as long as $k \leq 23$ but that 5^{24} will results in a rounding error.