Raphael Sofaer

Numerical Computing: Homework 3

Due on February 21, 2012

Margaret Wright

1. Exercise 3.1

1.a. Compute $y = x^0$ when:

a. $0^0 = 1$. In hex:
$0000000000000000^{0000000000000000} = 3ff0000000000000$, which makes sense, since $x^0 = 1$.

b. $inf^0 = 1$. In hex:
$7ff0000000000000^{0000000000000000} = 3ff0000000000000$, which makes sense by the same logic.

c. $NaN^0 = 1$. In hex:
$7ff8000000000000^{0000000000000000} = 3ff0000000000000$, which is the same.

Conceptually, this rule makes sense to me if I think of x^y as $1 * x_1 * x_2 * \cdots * x_y$. If $y = 0$, $x^y = 1$.

1.b. Do the same for:

a. $1^{Inf} = 1$. In hex:
$3ff0000000000000^{7ff0000000000000} = 3ff0000000000000$, which makes sense, since no number of 1s multiplied together can give anything but 1.

b. $-1^{Inf} = NaN - NaNi$. In hex:
$bff0000000000000^{7ff0000000000000} = fff8000000000000fff8000000000000i$. This seems arbitrary and a little strange to me, since -1^n is one of 1,-1.

c. $log(0.0) = -Inf$. In hex:
$log(0000000000000000) = fff0000000000000$. This makes sense, since the limit of log(x) as x goes to 0 is -infinity.

d. $log(-Inf) = Inf + 3.132i$. In hex:
$log(fff0000000000000) = 7ff0000000000000400921fb54442d18i$. This makes sense because if $log(x) = y$, $e^y = x$, and since by Euler's identity

$$e^{\pi i} = -1$$

$$log(-1) = \pi i$$

$$log(-Inf) = log(-1) + log(Inf)$$

$$log(Inf) = Inf$$

$$log(Inf) + log(-1) = Inf + \pi i$$

e. $exp(-Inf) = 0$. In hex:
$exp(fff0000000000000) = 0000000000000000$. This makes sense because $e^{-Inf} = 1/e^{Inf} = 1/Inf = 0$.

1.c. A non-standard calculation of my own:

I tried $NaN * -NaN$, which gave me NaN. I first expected expected to get $-NaN$, but I suppose that once in the NaN realm, the sign bit isn't very significant.

2. Hexadecimal IEEE numbers in decimal

2.a. 4059000000000000

The sign bit and exponent of this number is 405 in hex, which is 0100 0000 0101. Without the sign bit, the exponent bitstring is 100 0000 0101, or still 405 in hex, or $4 * 16^2 + 5$, which comes out to 1029. The exponent is 1029-1023, or 6.
The mantissa of this number is 9000000000000, or 1001 followed by 48 zeros in binary. When the implied 1 is included, we get 1.1001 followed by 48 zeros.
In decimal, we have $(1 + 1/2 + 1/16) * 2^6 = 1.5625 * 2^6 = 100$.

2.b. 3f847ae147ae147b

The sign bit is 0, and the exponent bitstring is 3f8, so the exponent is $-1023 + 3 * 16^2 + 15 * 16 + 8 = 1016 - 1023 = -7$.
That leaves a mantissa of 1.47ae147ae147b.
In decimal, this is $(1 + 4/16 + 7/16^2 + ... + 11/16^{13}) * 2^{-7}$. I used a ruby interpreter to calculate this, and it came out to $1.28 * 2^{-7} = 0.01$.

2.c. 3fe921fb54442d18

The sign bit is 0, and the exponent bitstring is 3fe, which gives an exponent of $-1023 + 3 * 16^2 + 15 * 16 + 14 = -1$.
That leaves a mantissa of 1.921fb54442d18.
In decimal, I used the ruby interpreter again to come out with $1.5707963267948966/2 = \pi/4$.

3. Error analysis of the Taylor series expansion of a smooth function f

3.a. Give an upper bound to the truncation error $|e_\tau|$ of (1.2) expressed in terms of the finite-difference interval h, that is valid for all $\bar{x}$.

4. Consider $Ax = b$, where A is a nonsingular n x n matrix with $n > 1$.

4.a. Solve the given linear system

By inspection, the solution given linear system is $x^* = [1, -1]$.
In hex, this is $x^* = [3ff0000000000000, bff0000000000000]$
Matlab gives the answer $\tilde{x} = [1.00000000000020, -1.00000000000027]$.
In hex, this is $\tilde{x} = [3ff000000000039b, bff00000000004b0]$.

$$\tilde{x} - x^* = [-2.04947170345804 * 10^{-13}, 2.66453525910038 * 10^{-13}]$$

In hex, this is $\tilde{x} - x^* = [bd4cd800000000003d52c00000000000]$.

4.b. Compute the residuals $r*$ and $\tilde{r}$