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:
$log(fff0000000000000) = 7ff000000000000400921fb54442d18i$. 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 Inf/Inf, which is NaN. That makes sense to me, because a value of 1 would imply that Inf is a number, but it's more of a limit or representation of a concept than a number.

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 . . . consider the function $f(x) = sin(x)$

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 $\bar{x}$.

$$|e_T| = 1/2 * h * f''(\xi)$$
$$max|e_T| = 1/2 * h * max(f''(\xi))$$
$$max|e_T| = 1/2 * h * max(-sin(\xi))$$
$$max|e_T| = 1/2 * h$$

3.b. Give the mathematical form of f'(x). . .

$d/dx(sin(x)) = f'(x) = cos(x)$.
At $\bar{x} = 2.25$:

$$f(2.25) = sin(2.25) = 7.78073196887921e - 01$$
$$f'(2.25) = cos(2.25) = -6.28173622722739e - 01$$

3.c. Compute the forward-difference approximation to $f'(\bar{x})$ at $\bar{x} = 2.25 \ldots$

Forward difference approximations		
h	$\tilde{\rho}(\bar{x}, h)$	$\tilde{\rho}(\bar{x}, h) - f'(\bar{x})$
1.000000e-04	-6.282125e-01	-3.890261e-05
1.000000e-05	-6.281775e-01	-3.890362e-06
1.000000e-06	-6.281740e-01	-3.891958e-07
1.000000e-07	-6.281737e-01	-3.814324e-08
1.000000e-08	-6.281736e-01	-1.505884e-09
1.000000e-09	-6.281737e-01	-1.125282e-07
1.000000e-10	-6.281742e-01	-5.566174e-07
1.000000e-11	-6.281864e-01	-1.276907e-05
1.000000e-12	-6.282752e-01	-1.015869e-04
1.000000e-13	-6.283862e-01	-2.126092e-04
1.000000e-14	-6.439294e-01	-1.575573e-02

Hexadecimal difference approximations			
h	$f(\bar{x} + h)$	$f(\bar{x})$	$f(\bar{x} + h) - f(\bar{x})$
$1.00000000000000 * 10^{-04}$	$3fe8e57603e9af6e$	$3fe8e5f9c2d0e3aa$	$bf1077dce6878000$
$1.00000000000000 * 10^{-05}$	$3fe8e5ec96504a15$	$3fe8e5f9c2d0e3aa$	$beda5901332a0000$
$1.00000000000000 * 10^{-06}$	$3fe8e5f871914f7e$	$3fe8e5f9c2d0e3aa$	$bea513f942c00000$
$1.00000000000000 * 10^{-07}$	$3fe8e5f9a1175615$	$3fe8e5f9c2d0e3aa$	$be70dcc6ca800000$
$1.00000000000000 * 10^{-08}$	$3fe8e5f9bf7188b8$	$3fe8e5f9c2d0e3aa$	$be3afad790000000$
$1.00000000000000 * 10^{-09}$	$3fe8e5f9c27a8dc4$	$3fe8e5f9c2d0e3aa$	$be05957980000000$
$1.00000000000000 * 10^{-10}$	$3fe8e5f9c2c84179$	$3fe8e5f9c2d0e3aa$	$bdd1446200000000$
$1.00000000000000 * 10^{-11}$	$3fe8e5f9c2d006a4$	$3fe8e5f9c2d0e3aa$	$bd9ba0c000000000$
$1.00000000000000 * 10^{-12}$	$3fe8e5f9c2d0cd8f$	$3fe8e5f9c2d0e3aa$	$bd661b0000000000$
$1.00000000000000 * 10^{-13}$	$3fe8e5f9c2d0e174$	$3fe8e5f9c2d0e3aa$	$bd31b00000000000$
$1.00000000000000 * 10^{-14}$	$3fe8e5f9c2d0e370$	$3fe8e5f9c2d0e3aa$	$bcfd000000000000$

3.d. Referring explicitly to the hexadecimal values of (iii)(b), explain why, for some values of h, the error...

The cancellation error present in the value of $\tilde{\rho}(\bar{x}, h) - f'(\bar{x})$ is revealed by the mantissas of part iiib. By the time $h = 10^{-14}$, we can see that almost the entire mantissa is 0. A large change in the order of magnitude, followed by a change back, causes the computer to discard most of the significant digits in the mantissa and only retain a very imprecise number.

3.e. At $\bar{x} = 0$

At $\bar{x} = 0$, $f(x) = 0$, so we can expect much less cancellation error, since looking at the difference between the estimated value and the computed exact value will not cause as large a change in magnitude from the original values.

Forward difference approximations		
h	$\tilde{\rho}(\bar{x}, h)$	$\tilde{\rho}(\bar{x}, h) - f'(\bar{x})$
1.000000e-04	1.000000e+00	-1.666667e-09
1.000000e-05	1.000000e+00	-1.666678e-11
1.000000e-06	1.000000e+00	-1.666445e-13
1.000000e-07	1.000000e+00	-1.665335e-15
1.000000e-08	1.000000e+00	0.000000e+00
1.000000e-09	1.000000e+00	0.000000e+00
1.000000e-10	1.000000e+00	0.000000e+00
1.000000e-11	1.000000e+00	0.000000e+00
1.000000e-12	1.000000e+00	0.000000e+00
1.000000e-13	1.000000e+00	0.000000e+00
1.000000e-14	1.000000e+00	0.000000e+00

We can see the error above going to zero instead of remaining large. I did not include the hexadecimal printout, but there we can also see that the mantissas of the numbers retain their digits rather than being zeroed out.

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 to the 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 $\tilde{x} = [3ff000000000039b, bff000000000004b0]$.

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

In hex, this is $\tilde{x} - x^* = [bd4cd80000000000, 3d52c00000000000]$.

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

The residual $r^* = b - Ax^* = [-5.55111512312578e-17; 1.38777878078145e-17]$
In hex: $r^* = b - Ax^* = [bc90000000000000; 3c70000000000000]$
The residual $\tilde{r} = b - A\tilde{x} = [-5.55111512312578e-17; -5.55111512312578e-17]$

In hex: $\tilde{r} = b - A\tilde{x} = [bc90000000000000; bc90000000000000]$

$$||r^*|| = 5.72195849815280e - 17$$
$$||\tilde{r}|| = 7.85046229341888e - 17$$

The residual for the exact answer, r^* is smaller than the residual for $\tilde{r}$, but only by a very small amount ($2.12850379526608e - 17$). I would guess that this amount is small enough that the algorithm the computer uses to calculate the solution reaches its threshold for having found the correct answer.

4.c. Let $\hat{x}$ be a potential solution... Show mathematically that E satisfies $(A + E)\hat{x} = b$

$$E = 1/x^T\hat{x} * \hat{r}\hat{x}$$
$$E\hat{x} = 1/x^T\hat{x} * \hat{r}\hat{x}^T\hat{x} = \hat{r}$$
$$\hat{r} = b - A\hat{x}$$
$$E\hat{x} = b - A\hat{x}$$
$$E\hat{x} + A\hat{x} = b$$
$$(A + E)\hat{x} = b$$

4.d. Consider the vector $\hat{x}$ given by...

If we take "x is close to y" to mean a low value of $||x - y||$, $\hat{x}$ is very far from both x^* and $\tilde{x}$. $norm(\hat{x} - \tilde{x})$ is near 69.793, and x^* is about the same distance.

4.e. For $\hat{x}$ from (iv), ...

$$E = \begin{pmatrix} 6.7408 * 10^{-5} & -8.8163 * 10^{-5} \\ -5.8616 * 10^{-5} & 7.6663 * 10^{-5} \end{pmatrix}$$

$$||E||_2 = 1.47069708632407 * 10^{-4}$$

Since norm(E) is so low, we can say that $\hat{x}$ is a solution to a nearby system. E is the pertubation required to reach the problem which $\hat{x}$ is a solution of, so if norm(E) is small, (A + E) is close to A, and $\hat{x}$ is a solution to a nearby system.