CSE 4345 / CSE 5315 - *Computational Methods*

Homework 1- Sample Solution - Fall 2011

Due Date: Sept. 20 2011

Problems marked with * are required only for students of CSE 5315 but will be graded for extra credit for students of CSE 4345.

Sources of Errors

1. Floating Point Precision and Rounding Errors

a) Compute the machine precision, ϵ_{mach}, overflow number, OFL, and underflow number, UFL, for floating point systems with rounding and the following number of bits for mantissa and exponent:

You can assume that one bit of the mantissa is used for the sign of the number and that the exponent is split evenly between positive and negative exponents.

mantissa/[bits]	base	exponent/[bits]	ϵ_{mach}	OFL	UFL
5	2	3	2^{-4}	$2^4(1-2^{-4})$	2^{-3}
11	2	5	2^{-10}	$2^{16}(1-2^{-10})$	2^{-15}
32	2	16	2^{-31}	$2^{32768}(1-2^{-31})$	2^{-32767}

b) The use of floating point operations can lead to loss of significance (or cancelation) when two large numbers are subtracted to yield a significantly smaller result. For the following equations, compute the error that the use of single precision floating point numbers (float) would lead to for very small numbers, x (e.g. 10^{-6}). For each one, construct an equivalent formula that would result in substantially less error (and compute the corresponding error).

- $\frac{1-(1+(x^3-x^5))}{x^3}$

Modified equation to eliminate loss of significance: $x^2 - 1$

x	original float result	original double result	modified float result	modified double result	original float absolute error	original float relative error	modified float absolute error	modified float relative error
1.000000e-01	-9.900331e-01	-9.900000e-01	-9.900000e-01	-9.900000e-01	-3.309011e-05	3.342436e-05	-9.536743e-09	9.633074e-09
1.000000e-02	-9.536743e-01	-9.999000e-01	-9.999000e-01	-9.999000e-01	4.622568e-02	-4.623031e-02	1.659393e-08	-1.659559e-08
1.000000e-03	0.000000e+00	-9.999990e-01	-9.999990e-01	-9.999990e-01	9.999990e-01	-1.000000e+00	1.327896e-08	-1.327897e-08
1.000000e-04	0.000000e+00	-1.000089e+00	-1.000000e+00	-1.000000e+00	1.000000e+00	-1.000000e+00	-1.000000e-08	1.000000e-08
1.000000e-05	0.000000e+00	-1.110223e+00	-1.000000e+00	-1.000000e+00	1.000000e+00	-1.000000e+00	-1.000000e-10	1.000000e-10
1.000000e-06	0.000000e+00	0.000000e+00	-1.000000e+00	-1.000000e+00	1.000000e+00	-1.000000e+00	-9.999779e-13	9.999779e-13

Note that at 10^{-4} the original double precision result starts to diverge markedly from the correct solution and at 10^{-6} it becomes 0 due to loss of sigificance.

- $((x-2)+(2+x))x^2$

  Modified equation to eliminate loss of significance: $2*x^3$

x	original float result	original double result	modified float result	modified double result	original float absolute error	original float relative error	modified float absolute error	modified float relative error
1.000000e-01	1.999999e-03	2.000000e-03	2.000000e-03	2.000000e-03	-6.034970e-10	-3.017485e-07	9.499490e-11	4.749745e-08
1.000000e-02	1.999998e-06	2.000000e-06	2.000000e-06	2.000000e-06	-2.051413e-12	-1.025706e-06	-5.049515e-15	-2.524757e-09
1.000000e-03	1.999974e-09	2.000000e-09	2.000000e-09	2.000000e-09	-2.603578e-14	-1.301789e-05	-2.786085e-16	-1.393042e-07
1.000000e-04	1.999139e-12	2.000000e-12	1.999999e-12	2.000000e-12	-8.606477e-16	-4.303238e-04	-6.585129e-19	-3.292565e-07
1.000000e-05	2.002716e-15	2.000000e-15	1.999999e-15	2.000000e-15	2.715595e-18	1.357797e-03	-8.397780e-22	-4.198890e-07
1.000000e-06	1.907348e-18	2.000000e-18	1.999999e-18	2.000000e-18	-9.265179e-20	-4.632589e-02	-7.355532e-25	-3.677766e-07

- $\sqrt{9-x^3}-3$

  Modified equation to eliminate loss of significance: $\frac{-x^3}{\sqrt{9-x^3}+3}$

x	original float result	original double result	modified float result	modified double result	original float absolute error	original float relative error	modified float absolute error	modified float relative error
1.000000e-01	-1.666546e-04	-1.666713e-04	-1.666620e-04	-1.666620e-04	7.450502e-09	-4.470426e-05	-7.836825e-14	4.702225e-10
1.000000e-02	-2.384186e-07	-1.666667e-07	-1.666667e-07	-1.666667e-07	-7.175192e-08	4.305115e-01	-4.208836e-15	2.525302e-08
1.000000e-03	0.000000e+00	-1.666667e-10	-1.666666e-10	-1.666667e-10	1.666667e-10	-1.000000e+00	2.783867e-17	-1.670320e-07
1.000000e-04	0.000000e+00	-1.665335e-13	-1.666666e-13	-1.666667e-13	1.666667e-13	-1.000000e+00	5.939358e-20	-3.563615e-07
1.000000e-05	0.000000e+00	-4.440892e-16	-1.666666e-16	-1.666667e-16	1.666667e-16	-1.000000e+00	6.998150e-23	-4.198890e-07
1.000000e-06	0.000000e+00	0.000000e+00	-1.666666e-19	-1.666667e-19	1.666667e-19	-1.000000e+00	6.129610e-26	-3.677766e-07

- $\frac{5*cos^2(x)-5}{x^2}$

  Modified equation to eliminate loss of significance: $\frac{-5*sin^2(x)}{x^2}$

x	original float result	original double result	modified float result	modified double result	original float absolute error	original float relative error	modified float absolute error	modified float relative error
1.000000e-01	-4.983329e+00	-4.983356e+00	-4.983356e+00	-4.983356e+00	2.624358e-05	-5.266246e-06	-4.593033e-07	9.216748e-08
1.000000e-02	-5.002022e+00	-4.999833e+00	-4.999833e+00	-4.999833e+00	-2.188454e-03	4.377054e-04	2.285609e-07	-4.571371e-08
1.000000e-03	-4.768372e+00	-4.999998e+00	-4.999999e+00	-4.999998e+00	2.316263e-01	-4.632527e-02	-2.361550e-07	4.723101e-08
1.000000e-04	0.000000e+00	-5.000000e+00	-5.000000e+00	-5.000000e+00	5.000000e+00	-1.000000e+00	-1.666667e-08	3.333333e-09
1.000000e-05	0.000000e+00	-5.000000e+00	-5.000000e+00	-5.000000e+00	5.000000e+00	-1.000000e+00	-1.666667e-10	3.333334e-11
1.000000e-06	0.000000e+00	-5.000445e+00	-5.000000e+00	-5.000000e+00	5.000000e+00	-1.000000e+00	-1.667111e-12	3.334222e-13

Note: If you are using a system that does not have single precision floating point numbers, you can reduce the precision of double precision calculations (and numbers) by following each operation by the operation $((2^{\lfloor log_2(x) \rfloor} + (x/2^n)) - 2^{\lfloor log_2(x) \rfloor}) * 2^n$ (or the slightly less precise version $((x + (x/2^n)) - x) * 2^n$) which reduces the precision of x by n bits (i.e. for reducing double precision to single precision $n = 29$). It is important that the operations are performed in the exact order indicated.

2. Data Propagation and Computation Error
   For an experiment, a sphere with diameter d has to be completely filled with helium of a fixed temperature. Fluctuations in the outside temperature of the spherical container cause proportional fluctuations, Δd, in the dimensions of the sphere (but not of the inside temperature) and to measure them a sensor is mounted on the sphere to measure its diameter. To correct the amount of gas inside the container it is necessary to compute by how much the volume has changed due to the change in the diameter of the sphere.

There are multiple ways the change in volume can be calculated (or approximated) by the following 3 formulas:

- $\Delta v = \frac{1}{6}\pi(d + \Delta d)^3 - \frac{1}{6}\pi d^3$
- $\Delta v = \frac{1}{2}\pi d\Delta d^2 + \frac{1}{2}\pi d^2\Delta d + \frac{1}{6}\pi\Delta d^3$
- $\Delta v \approx \frac{1}{2}\pi d^2\Delta d$

a) For each of the three formulas ("algorithms") determine the total error for $d = 1m$ and diameter changes of $10^{-2}m$, $10^{-3}m$, $10^{-4}m$, and $10^{-5}m$ when the formulas are calculated using single precision floating point numbers (float). You can assume that the calculation of the precise formula using double precision floating point numbers (double) provides the accurate result.

Note: If you are using a system that does not have single precision floating point numbers, you can use the precision reduction method from problem 1b).

Δl	correct result	formula 1 error	formula 2 error	formula 3 error
1.000000e-02	1.586557e-02	-2.153091e-09	-2.904454e-10	-1.576024e-04
1.000000e-03	1.572368e-03	1.220917e-07	-1.443440e-10	-1.571402e-06
1.000000e-04	1.570953e-04	8.210711e-08	-2.390320e-11	-1.572542e-08
1.000000e-05	1.570812e-05	2.750587e-08	-2.704152e-12	-1.591372e-10

Note that the approximate formula (formula 3) eventually outperforms the version of the correct formula that is subject to cancellation (formula 1) as x becomes small.

b) Break the total error for each of the formulas (from part a) into data propagation error, truncation error, and rounding error.

Δl	Formula 1 (prop, trunc, round)	Formula 2 (prop, trunc, round)	Formula 3 (prop, trunc, round)
1.000000e-02	(-3.581575e-10,0.000000e+00,-1.794933e-09)	(-3.581575e-10,0.000000e+00,6.771203e-11)	(-3.581575e-10,-1.576032e-04,1.167800e-09)
1.000000e-03	(-1.084726e-10,0.000000e+00,1.222002e-07)	(-1.084726e-10,0.000000e+00,-3.587147e-11)	(-1.084726e-10,-1.571320e-06,2.645540e-11)
1.000000e-04	(-1.540029e-11,0.000000e+00,8.212251e-08)	(-1.540029e-11,0.000000e+00,-8.502908e-12)	(-1.540029e-11,-1.570849e-08,-1.532572e-12)
1.000000e-05	(-1.825484e-12,0.000000e+00,2.750770e-08)	(-1.825484e-12,0.000000e+00,-8.786680e-13)	(-1.825484e-12,-1.570802e-10,-2.316004e-13)

c)* Discuss why the different formulas ("algorithms") introduce different amounts of error. Which of the formulas is the best to use ?

The first formula, while accurate (i.e. no truncation error) is prone to cancellation since it subtracts two approximately equal large numbers to obtain a small one.

The second formula avoids the truncation problem by reformulating the equation. It still has no truncation error but much less rounding error than the first formula (due to the absence of cancellation).

The third formula makes a simplification. This introduces truncation error since some terms are omitted but reduces rounding error over the first formula. For small Δd values it has even lower rounding error than the second formula.

For the argument range used here the second formula is always the best one overall and will stay this way until x becomes even smaller. Then the reduced complexity of the third formula might make it more desirable.

3. Sensitivity and Conditioning
   The (relative) condition number represents the relation between the relative forward and the relative backward error for a given problem. For the evaluation of a function $f(x)$ this yields:

$$cond = \frac{|\text{relative forward error}|}{|\text{relative backward error}|} = \frac{\frac{|f(x)-f(\hat{x})|}{|f(x)|}}{\frac{|x-\hat{x}|}{|x|}} = \frac{|f(x)-f(\hat{x})|\ |x|}{|f(x)|\ |x-\hat{x}|} = \frac{|f(x)-(f(x)+(\hat{x}-x)f'(x)+\frac{1}{2}(\hat{x}-x)^2 f''(c)|\ |x|}{|f(x)|\ |x-\hat{x}|} \approx \frac{|(x-\hat{x})f'(x)|\ |x|}{|f(x)|\ |x-\hat{x}|} = \frac{|f'(x)|\ |x|}{|f(x)|}$$

   Consider the following functions and values:

   - $f(x) = x^2 - 5x + \frac{1}{x^2}\ ,\ x = 0.2$
   - $f(x) = (cos^2(x) + sin^2(x))^2\ ,\ x = \pi - 10^{-8}$
   - $f(x) = 1 + tan(x-6) * x\ ,\ x = 6.01$

   a) Compute the condition number for the functions at the indicated values of x and discuss what these numbers tell us about the accuracy with which we can expect to be able to compute the corresponding function values for different machine precisions.

   - $f(x) = x^2 - 5x + \frac{1}{x^2}\ ,\ x = 0.2$
     $cond = \left|\frac{(2*x-5-2*x^{-3})x}{x^2-5x+x^{-2}}\right| = 2.118$
   - $f(x) = (cos^2(x) + sin^2(x)^2\ ,\ x = \pi - 10^{-8}$
     $f(x) = 1^2 = 1 \Rightarrow cond = 0$
   - $f(x) = 1 + tan(x-6) * x\ ,\ x = 6.01$
     $cond = \left|\frac{\left(tan(x-6)+x*\frac{1}{cos^2(x-6)}\right)x}{1+tan(x-6)*x}\right| = 34.1324$

   The condition number measures the "natural" amplification of data error through the problem. The random error that should be expected due to the condition number is $cond * \epsilon_{mach}$.

   b) For each of the functions, compute the error indicated by the condition number for input data with float (single precision floating point numbers) and the actual error and compare them. To compute the actual error you can assume that the result obtained with double precision numbers is accurate.

	Formula 1	Formula 2	Formula 3
Actual error	-9.918213e-07	0.000000e+00	-1.413029e-06
Propagation error from $cond$	2.524853e-07	0.000000e+00	4.068899e-06

   This shows that the third formula is accruing significant additional rounding error during calculation (as opposed to the other two formulas).

   c)* Derive the relative forward and backward errors and the (relative) condition number for function inversion (i.e. the calculation of $f^{-1}(x)$ for a scalar function $f(x)$ with a scalar parameter x.)

   - Relative forward error : $\frac{(\hat{x}-x)}{x}$

- Relative backward error : $\frac{(f(\hat{x})-f(x))}{f(x)}$
- Relative condition number :
$\frac{\left|\frac{\hat{x}-x}{x}\right|}{\left|\frac{f(\hat{x})-f(x)}{f(x)}\right|} = \frac{|(\hat{x}-x)||f(x)|}{|x|\left|f(x)+(\hat{x}-x)f'(x)+\frac{1}{2}(\hat{x}-x)^2f''(c)-f(x)\right|} \approx \frac{|(\hat{x}-x)||f(x)|}{|x||(\hat{x}-x)f'(x)|} = \frac{|f(x)|}{|x||f'(x)|}$