COLUMBIA UNIVERSITY
Intro to Numerical Methods
APAM E4300 (1)

MIDTERM EXAM SOLUTIONS – MARCH 11, 2013

INSTRUCTOR: SANDRO FUSCO

FAMILY NAME: ______________________________________________________________
GIVEN NAME:   ______________________________________________________________
UNI:     ______________________________________________________________

Problem 1: (10 Points)
   a) [3 points] In finding a root with Newton's method, an initial guess of $x_0 = 4$ with $f(x_0) = 1$ leads to $x_1 = 3$. What is the derivative of $f$ at $x_0$?
   b) [3 points] In using the secant method to find a root, $x_0 = 2$, $x_1 = -1$ and $x_2 = -2$ with $f(x_1) = 4$ and $f(x_2) = 3$. What is $f(x_0)$?
   c) [4 points] Can the bisection method be used to find the roots of the function $f(x) = \sin(x) + 1$? Why or why not? Can Newton's method be used to find the roots (or a root) of this function? If so, what will be its order of convergence and why?

Solution:
   a) Since $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$, we have $3 = 4 - \frac{1}{f'(x_0)}$. Hence $f'(x_0) = 1$.
   b) Since $x_2 = x_1 - f(x_1) \frac{x_1-x_0}{f(x_1)-f(x_0)}$, we have $-2 = -1 - 4 \cdot \frac{-3}{4-f(x_0)}$. Hence $f(x_0) = 16$.
       [Note that the value of $f(x_2)$ was not needed for this problem.]
   c) Bisection cannot be used because $f(x)$ is always nonnegative. Newton's method can be used for this problem but its convergence will be only linear since $f'(x) = \cos(x)$ and $\cos(x) = 0$ at the roots of $f$ since at these points $\sin(x) = -1$. 
Problem 2: (15 Points)

a) [5 points] Use Taylor series expansion

\[ f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2!}f''(x_0)(x - x_0)^2 + \frac{1}{3!}f'''(x_0)(x - x_0)^3 + \cdots \]

with \( n = 0 \) to \( 6 \) to approximate \( f(x) = \cos(x) \) at \( x = \frac{\pi}{3} \) on the basis of the value of \( f(x) \) and its derivatives at \( x_0 = \frac{\pi}{4} \).

b) [5 points] After each new term is added, compute the true percent relative error \( \varepsilon_t \).

c) [5 points] What value of \( n \) is required for the absolute value of the true percent error \( |\varepsilon_t| \) to fall below a pre-specified error criterion \( \varepsilon_c \) conforming to six (6) significant figures?

Solution:

a) For the function \( f(x) = \cos(x) \) at the point \( x_0 = \frac{\pi}{4} \), we have:

\[ f(\pi/4) = \cos(\pi/4) = \sqrt{2}/2 = 0.707106781 \]
\[ f'(\pi/4) = -\sin(\pi/4) = -\sqrt{2}/2 \]
\[ f''(\pi/4) = -\cos(\pi/4) = -\sqrt{2}/2 \]
\[ f'''(\pi/4) = \sin(\pi/4) = \sqrt{2}/2 \]
\[ f^{(4)}(\pi/4) = \cos(\pi/4) = \sqrt{2}/2 \]

We have \(-x_0 = \pi/3 - \pi/4 = \pi/12 \). Hence the Taylor series expansion is:

\[ f(\pi/3) = \cos(\pi/4) \cdot \sin(\pi/4) \cdot \frac{\pi}{12} - \frac{\cos(\pi/4)}{2} \cdot \left( \frac{\pi}{12} \right)^2 + \frac{\sin(\pi/4)}{3!} \cdot \left( \frac{\pi}{12} \right)^3 + \cdots \]

b) We know the true value of the function \( f(\pi/3) = \cos(\pi/3) = 0.5 \). The zero-order approximation of \( f(\pi/3) \approx \cos(\pi/4) = \sqrt{2}/2 = 0.707106781 \), which represents a percent relative error of \( \varepsilon_t = \frac{0.5 - 0.707106781}{0.5} \times 100\% = 41.4\% \). For the first-order approximation, we have:

\[ f(\pi/3) \approx \cos(\pi/4) \cdot \sin(\pi/4) \cdot \frac{\pi}{12} = 0.521986659 \],

which has \( \varepsilon_t = \frac{0.5 - 0.521986659}{0.5} \times 100\% = 4.40\% \), and so on.

The process can be continued and the results are listed in the table below:

<table>
<thead>
<tr>
<th>Term</th>
<th>( f(\pi/3) )</th>
<th>( \varepsilon_t ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.707106781</td>
<td>41.4</td>
</tr>
<tr>
<td>1</td>
<td>0.521986659</td>
<td>4.40</td>
</tr>
<tr>
<td>2</td>
<td>0.497754491</td>
<td>0.449</td>
</tr>
<tr>
<td>3</td>
<td>0.499869147</td>
<td>2.62x10^-2</td>
</tr>
<tr>
<td>4</td>
<td>0.500007551</td>
<td>1.51x10^-2</td>
</tr>
<tr>
<td>5</td>
<td>0.50000304</td>
<td>6.08x10^-3</td>
</tr>
<tr>
<td>6</td>
<td>0.499999988</td>
<td>2.44x10^-6</td>
</tr>
</tbody>
</table>

c) The error criterion that ensures a result that is correct to at least six significant figures is given by the formula \( \varepsilon_x = 0.5 \times 10^{2-6} \% = 0.00005\% \). Thus, we will add terms to the series until \( \varepsilon_t \) falls below this level. Thus, for \( n = 6 \) the percent error falls below \( \varepsilon_x = 0.00005\% \) and the computation is terminated.
Problem 3: (15 Points)

Consider IEEE double precision floating point arithmetic, using round to nearest. Let a, b, and c be normalized double precision floating point numbers, and let ⊕, ⊗, and ∅ denote correctly rounded floating point addition, multiplication, and division.

a) [5 points] Is it necessarily true that \( a \oplus b = b \oplus a \)? Explain why or give an example where this does not hold.

b) [5 points] Is it necessarily true that \((a \oplus b) \odot c = a \odot (b \odot c)\)? Explain why or give an example where this does not hold.

c) [5 points] Determine the maximum possible relative error in the computation \((a \odot b) \odot c\) assuming that \(c \neq 0\). [You may omit terms of order \(O(\varepsilon^2)\) and higher.] Suppose \(c = 0\). What are the possible values that \((a \odot b) \odot c\) could be assigned?

Solution:

a) \( a \oplus b = b \oplus a \), since both must be the correctly rounded value of \(a + b = b + a\).

b) This is not necessarily true. The machine precision of a double precision system is \(2^{-52}\). Hence \((1 \oplus 2^{-53}) \oplus 2^{-52} = 1\) but \(1 \oplus (2^{-53} \oplus 2^{-52}) = 1 + 2^{-52}\).

c) \((a \odot b) = a x b x (1 + \delta_1)\) where \(|\delta_1| < \varepsilon\) (or \(\leq \varepsilon/2\) for round to nearest). \(a x b x (1 + \delta_1) \odot c = (a x b/c) x (1 + \delta_1)(1 + \delta_2)\) where \(|\delta_2| < \varepsilon\) (or \(\leq \varepsilon/2\) for round to nearest).

The relative error is \(|(1 + \delta_1)(1 + \delta_2) - 1| = |\delta_1 + \delta_2 + \delta_1\delta_2|\) which, ignoring terms of order \(\varepsilon^2\), is at most \(2\varepsilon\) (or \(\varepsilon\) for round to nearest).

If \(c = 0\), then if \((a \odot b)\) is positive we get \(+\infty\), if \((a \odot b)\) is negative we get \(-\infty\), and if \((a \odot b)\) is 0 we get NaN.

Problem 4: (15 Points)

Suppose that you are given a polynomial \(P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0\) of degree \(n\).

a) [7 points] Write a short MATLAB function (mine is 4 lines) utilizing the Horner’s method for evaluating polynomials at a given point \(x\). The first line can be written as follows:

```matlab
function  y  = hornersPoly(p,x)
```

where \(p\) is the vector of the polynomial coefficients, \(x\) is the value where the polynomial is to be evaluated, and \(y\) is the output value.

b) [3 points] Change your MATLAB function in part (a) to allow for vectorized arguments. In other words, suppose that \(x\) is now a vector of values where the polynomial is to be evaluated, and \(y\) is a vector of outputs.

c) [5 points] Use part (a) to find \(P(3)\) for the polynomial \(P(x) = x^5 - 6x^4 + 8x^3 + 4x - 40\).

Solution:

a) function \( y = \text{hornersPoly}(p,x)\)

% hornersPoly - evaluates Polynomials using Horner's rule
% \( y = \text{hornersPoly}(p,x)\)
% \( p:\) - vector of polynomial coefficients such that
Problem 5: (15 Points)

a) [7 points] In class we have seen one way to approximate the derivative of a function \( f \):

\[
 f'(x) \approx \frac{f(x + h) - f(x - h)}{2h}
\]

for some small number \( h \) (centered difference formula). Assuming that \( f \in C^2 \), use Taylor's Theorem to determine the accuracy of this approximation.

b) [8 points] Show that, with this formula, we can approximate a derivative to about the \( 2/3 \) power of the machine precision.

Solution:

a) To determine the accuracy of this approximation, we use Taylor's Theorem, assuming that \( f \in C^2 \):
\[
f(x + h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f^{(3)}(\xi), \quad \xi \in [x, x+h]\\
f(x - h) = f(x) - hf'(x) + \frac{h^2}{2} f''(x) - \frac{h^3}{6} f^{(3)}(\eta), \quad \eta \in [x-h, x]
\]

\[
\Rightarrow \frac{f(x + h) - f(x - h)}{2h} = \frac{2hf'(x)}{2h} + \frac{h^3}{12h} (f^{(3)}(\xi) + f^{(3)}(\eta))
\]

\[
\Rightarrow f'(x) = \frac{f(x + h) - f(x - h)}{2h} - \frac{h^2}{12} (f^{(3)}(\xi) + f^{(3)}(\eta)).
\]

This shows that the truncation error is \(O(h^2)\) and the approximation is second-order accurate.

b) The roundoff also plays a role in the evaluation of the centered finite difference. For example, if \(h\) is so small that \(x \pm h\) are rounded to \(x\), then the computed finite difference is zero. More generally, even if the only error made is in rounding the values \(f(x+h)\) and \(f(x-h)\), then the computed difference quotient will be:

\[
\frac{f(x + h)(1 + \delta_1) - f(x - h)(1 + \delta_2)}{2h} = \frac{f(x + h) - f(x - h)}{2h} + \frac{\delta_1 f(x + h) - \delta_2 f(x - h)}{2h}
\]

Since each \(|\delta_i|\) is less than the machine precision \(\varepsilon\), this implies that the rounding error is less than or equal to

\[
\varepsilon \cdot \left( |f(x + h)| + |f(x - h)| \right)
\]

Since the truncation error is proportional to \(h^2\) and the rounding error is proportional to \(1/h\), the best accuracy is achieved when the two quantities are approximately equal. Ignoring the constants, this means that

\[
h^2 \approx \frac{\varepsilon}{h} \Rightarrow h \approx \sqrt[3]{\varepsilon}
\]

Hence the truncation error is \(\varepsilon^{2/3}\). With the centered finite difference, we can achieve greater accuracy to about the 2/3 power of the machine precision.
**Problem 6: (15 Points)**
Consider a forward difference approximation for the second derivative of the form

Use Taylor’s theorem to determine the coefficients A, B, and C that give the maximal order of accuracy and determine what this order is.

**Solution:**

Expand $f(x+h)$ and $f(x+2h)$ about $x$ as in the previous exercise:

\[
\begin{align*}
  f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + O(h^4), \\
  f(x+2h) &= f(x) + 2hf'(x) + \frac{(2h)^2}{2}f''(x) + \frac{(2h)^3}{6}f'''(x) + O(h^4).
\end{align*}
\]

Combining series, we find

\[
Af(x)+Bf(x+h)+Cf(x+2h) = (A+B+C)f(x)+(B+2C)hf'(x)+(B+4C)\frac{h^2}{2}f''(x)+
\]

\[
(B+8C)\frac{h^3}{6}f'''(x) + (B + 16C)O(h^4).
\]

In order for this to approximate $f''(x)$, we need

\[
\begin{align*}
  A + B + C &= 0, \\
  B + 2C &= 0, \\
  B + 4C &= \frac{h^2}{2}.
\end{align*}
\]

Solving for $A$, $B$, and $C$, we find $A = C = \frac{1}{2}$, $B = -\frac{2}{h^2}$. The coefficient of $f'''(x)$ above is then $(B + 8C)\frac{h^3}{6} = h$, so the maximal order of accuracy is just 1.

**Problem 7: (15 Points)**

Steffensen’s method for solving $f(x) = 0$ is defined by:

\[
x_{k+1} = x_k - \frac{f(x_k)}{g_k},
\]

where

\[
g_k = \frac{f(x_k + f(x_k)) - f(x_k)}{f(x_k)}
\]

Show that this is quadratically convergent, under suitable hypotheses.

[**Hint:** Proceed as we did in the proof of quadratic convergence of Newton’s method.]
Solution:

We will proceed as we did in the proof of quadratic convergence of Newton’s method. If \( x_* \) is a root of \( f \), then from Taylor’s theorem with remainder,

\[
0 = f(x_*) = f(x_k) + (x_* - x_k)f'(x_k) + \frac{(x_* - x_k)^2}{2} f''(\xi_k)
\]

for some \( \xi_k \) between \( x_k \) and \( x_* \). Moving the second term to the left and dividing by \( f'(x_k) \), we find

\[
x_* = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{(x_* - x_k)^2}{2} \frac{f''(\xi_k)}{f'(x_k)}.
\]

Subtracting this from the equation for \( x_{k+1} \) gives

\[
x_{k+1} - x_* = \left( -\frac{f(x_k)}{g_k} + \frac{f(x_k)}{f'(x_k)} \right) + \frac{f''(\xi_k)}{2f'(x_k)} (x_* - x_k)^2.
\]

Now we will use Taylor’s theorem with remainder to estimate the term in parentheses in (2). Let \( y_k = f(x_k) \). Then

\[
f(x_k + y_k) = f(x_k) + y_k f'(x_k) + \frac{y_k^2}{2} f''(\eta_k),
\]

for some \( \eta_k \) between \( x_k \) and \( x_k + y_k \). Using this expression to estimate \( g_k \), we find

\[
g_k = \frac{y_k}{y_k} = \frac{f(x_k + y_k) - f(x_k)}{g_k} = f'(x_k) + \frac{y_k}{2} f''(\eta_k).
\]

Using this expression for \( g_k \) to estimate the term in parentheses in (2), we obtain

\[
\left( -\frac{f(x_k)}{g_k} + \frac{f(x_k)}{f'(x_k)} \right) = \frac{f(x_k)(g_k - f'(x_k))}{f'(x_k)g_k} = \frac{f(x_k)^2 f''(\eta_k)}{2f'(x_k)g_k}.
\]

From (1) it follows that \( f(x_k) = O(x_* - x_k) \); that is,

\[
f(x_k) = -(x_* - x_k)f'(x_k) + O((x_* - x_k)^2),
\]

where \( O((x_* - x_k)^2) \) denotes terms with a factor \( (x_* - x_k)^2 \) multiplied by other factors such as constants and second derivatives of \( f \) that remain bounded as \( x_k \) approaches \( x_* \). Making this substitution in (3), we find

\[
\left( -\frac{f(x_k)}{g_k} + \frac{f(x_k)}{f'(x_k)} \right) = O((x_* - x_k)^2).
\]

Thus, assuming that \( |f''| \) is bounded by some constant \( M \), that \( f'(x_*) \neq 0 \) and hence \( g_k \neq 0 \) for \( x_k \) sufficiently close to \( x_* \), and assuming that \( x_0 \) is sufficiently close to \( x_* \) to guarantee that future iterates only get closer and that \( g_k \) is nonzero for all \( k \), both terms in (2) are \( O((x_* - x_k)^2) \), so convergence will be quadratic.
Extra Credit Problem: (10 Points)

The conditioning of a problem measures how sensitive the answer is to small changes in the input. Let \( f: \mathbb{R} \rightarrow \mathbb{R} \), and suppose that \( x^* \) is close to \( x \) (e.g., \( x^* \) might be equal to \( \text{round}(x) \)). The conditioning of a problem measures how close \( y = f(x) \) is to \( y^* = f(x^*) \).

If
\[
|y^* - y| \approx C(x) \cdot |x^* - x|
\]
then \( C(x) \) is called the absolute condition number of the function \( f \) at the point \( x \).

If
\[
\left| \frac{y^* - y}{y} \right| \approx \kappa(x) \cdot \left| \frac{x^* - x}{x} \right|
\]
then \( \kappa(x) \) is called the relative condition number of the function \( f \) at the point \( x \).

1) [4 points] Explain why \( C(x) = |f'(x)| \) and \( \kappa(x) = \frac{|x^* - x|}{f(x)} \).

2) [6 points] What are the absolute and relative condition numbers of the following functions? Where are they large?
   a. \((x - 1)^a\)
   b. \(1/(1 + x^{-1})\)
   c. \(\ln(x)\)

Solution:

1) To determine a possible expression for \( C(x) \), note that
\[
y^* - y = f(x^*) - f(x) = \frac{f(x^*) - f(x)}{(x^* - x)} \cdot (x^* - x),
\]
and for \( x^* \) very close to \( x \), \( \frac{f(x^*) - f(x)}{(x^* - x)} \approx f'(x) \). Therefore we can define \( C(x) = |f'(x)| \).

To define the relative condition number \( \kappa(x) \), note that:
\[
\left| \frac{y^* - y}{y} \right| = \frac{f(x^*) - f(x)}{f(x)} = \frac{f(x^*) - f(x)}{(x^* - x)} \cdot \frac{(x^* - x)}{x} \cdot \frac{x}{f(x)}.
\]
Again we use the approximation \( \frac{f(x^*) - f(x)}{(x^* - x)} \approx f'(x) \) to determine \( \kappa(x) = \left| \frac{x f'(x)}{f(x)} \right| \).

2) From the formulae found in point 1), we have:

(a) \((x - 1)^a\)

Assuming \( a \neq 0 \) and \( x - 1 > 0 \) if necessary for \((x - 1)^a\) to be defined (e.g., if \( a = 1/2 \)), \( C(x) = |a(x - 1)^{a-1}| \), \( \kappa(x) = |ax/(x - 1)| \). If \( a > 1 \), then \( C(x) \) is large for \( |x| \) very large, while if \( a < 1 \) then \( C(x) \) is large for \( x \) near 1. If \( a = 1 \), then \( C(x) = 1 \) for all \( x \). \( \kappa(x) \) is large for \( x \) near 1.

(b) \(1/(1 + x^{-1})\)

\( C(x) = 1/(1 + x)^2 \), \( \kappa(x) = 1/|x + 1| \). Both are large when \( x \) is near \(-1\).

(c) \(\ln(x)\)

Assuming \( x > 0 \), \( C(x) = 1/x \), \( \kappa(x) = 1/\ln(x) \). \( C(x) \) is large when \( x \) is near 0, while \( \kappa(x) \) is large for \( x \) near 1.