Math 317 Assignment 1

Due in class: September 29, 2016

Instructions: Submit a hard copy of your solution with your name and student number. (**No name = zero grade!**) You must include all relevant program code, electronic output and explanations of your results. Write your own codes and comment them. Late assignment will not be graded and will receive a grade of zero.

1. (10 points) Let a be a nonzero real number and $fl(a)$ be a k-digit rounding approximation to a is base 10. Show that
$$\left|\frac{a - fl(a)}{a}\right| \le \frac{1}{2}10^{1-k}.$$

Solution: The idea is similar to the truncating approximation shown in class. We write

$$a = \pm(0.d_1d_2\dots d_kd_{k+1}d_{k+2}\dots)_{10} \times 10^e \quad \text{and} \quad fl(a) = \pm(0.d_1d_2\dots d_k^*)_{10} \times 10^e,$$

where d_k^* is the rounded digit defined as

$$d_k^* = \begin{cases} d_k & \text{if } d_{k+1} < 5, \\ d_k + 1 & \text{if } d_{k+1} \ge 5. \end{cases}$$

We look at each case seperately. If $d_{k+1} < 5$, then $d_{k+1} \le 4$ and $d_k^* = d_k$. Thus,

$$\begin{aligned}
\left|\frac{a - fl(a)}{a}\right| &= \frac{|(0.d_1d_2\dots d_kd_{k+1}d_{k+2}\dots)_{10} - (0.d_1d_2\dots d_k)_{10}| \times 10^e}{(0.d_1d_2\dots)_{10} \times 10^e} \\
&= \frac{(0.0\dots 0d_{k+1}d_{k+2}\dots)_{10}}{(0.d_1d_2\dots)_{10}} \\
&= \frac{(0.d_{k+1}d_{k+2}\dots)_{10} \times 10^{-k}}{(0.d_1d_2\dots)_{10}} \\
&\le \frac{(0.4d_{k+2}\dots)_{10} \times 10^{-k}}{(0.d_1d_2\dots)_{10}} \\
&\le \frac{(0.499\dots)_{10} \times 10^{-k}}{(0.d_1d_2\dots)_{10}} \\
&\le \frac{(0.500\dots)_{10} \times 10^{-k}}{(0.100\dots)_{10}} \\
&= \frac{1}{2}10^{1-k},
\end{aligned}$$

where we used the fact that $(0.d_1d_2\dots)_{10} \ge (0.100\dots)_{10}$, which follows from $d_1 \ge 1$ since $a \ne 0$.

On the other hand, if $d_{k+1} \ge 5$, then $d_k^* - d_k = 1$ and

$$1 - (0.d_{k+1}d_{k+2}\dots)_{10} \le 1 - (0.50\dots)_{10} = (0.50\dots)_{10}.$$

Hence,

$$\begin{aligned}\left|\frac{fl(a)-a}{a}\right| &= \frac{|(0.d_1d_2\ldots d_k^*)_{10} - (0.d_1d_2\ldots d_kd_{k+1}d_{k+2}\ldots)_{10}| \times 10^e}{(0.d_1d_2\ldots)_{10} \times 10^e} \\ &= \frac{|(d_1d_2\ldots d_{k-1}d_k^*)_{10} - (d_1d_2\ldots d_k.d_{k+1}d_{k+2}\ldots)_{10}| \times 10^{-k}}{(0.d_1d_2\ldots)_{10}} \\ &= \frac{|1-(0.d_{k+1}d_{k+2}\ldots)_{10}| \times 10^{-k}}{(0.d_1d_2\ldots)_{10}} \\ &\leq \frac{(0.500\ldots)_{10} \times 10^{-k}}{(0.d_1d_2\ldots)_{10}} \\ &\leq \frac{(0.500\ldots)_{10} \times 10^{-k}}{(0.100\ldots)_{10}} \\ &= \frac{1}{2}10^{1-k},\end{aligned}$$

where we used again the fact that $(0.d_1d_2\ldots)_{10} \geq (0.100\ldots)_{10}$.

2. (a) (10 points) Find $P_9(x)$ about $x_0 = 0$ of

$$f(x) = \frac{\sin(x^3) + e^{x^2} - 1}{x^2}.$$

Solution: First, we recall the Taylor series for both $\sin(x)$ and e^x:

$$\sin(x) = x - \frac{x^3}{6} + \frac{x^5}{120} + \mathcal{O}(x^7) \quad \text{and} \quad e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \mathcal{O}(x^6).$$

Then

$$\sin(x^3) = x^3 - \frac{x^9}{6} + \mathcal{O}(x^{15}) \quad \text{and} \quad e^{x^2} = 1 + x^2 + \frac{x^4}{2} + \frac{x^6}{6} + \frac{x^8}{24} + \frac{x^{10}}{120} + \mathcal{O}(x^{12}))$$

and therefore

$$\begin{aligned}\sin(x^3) + e^{x^2} - 1 &= x^3 - \frac{x^9}{6} + \mathcal{O}(x^{15}) + 1 + x^2 + \frac{x^4}{2} + \frac{x^6}{6} + \frac{x^8}{24} + \frac{x^{10}}{120} + \mathcal{O}(x^{12}) - 1 \\ &= x^2 + x^3 + \frac{x^4}{2} + \frac{x^6}{6} + \frac{x^8}{24} - \frac{x^9}{6} + \frac{x^{10}}{120} + \mathcal{O}(x^{12}).\end{aligned}$$

Finally, we can write

$$f(x) = 1 + x + \frac{x^2}{2} + \frac{x^4}{6} + \frac{x^6}{24} - \frac{x^7}{6} + \frac{x^8}{120} + \mathcal{O}(x^{10})$$

and conclude that

$$P_9(x) = 1 + x + \frac{x^2}{2} + \frac{x^4}{6} + \frac{x^6}{24} - \frac{x^7}{6} + \frac{x^8}{120}.$$

(b) (10 points) Let $p \geq 0$. For $n = 1, 2, \ldots$ find $P_n(x)$ about $x_0 = 0$ of $f(x) = (1+x)^p$ and find an upper bound depending on n, p for the error on $x \in [0, 1]$.

Solution: We start by observing that

$$\begin{aligned} f^{(1)}(x) = p(1+x)^{p-1} &\Longrightarrow f^{(1)}(0) = p \\ f^{(2)}(x) = p(p-1)(1+x)^{p-2} &\Longrightarrow f^{(2)}(0) = p(p-1) \\ &\dots \\ f^{(n)}(x) = p(p-1)\dots(p-n+1)(1+x)^{p-n} &\Longrightarrow f^{(n)}(0) = p(p-1)\dots(p-n+1) \end{aligned}$$

Then by Taylor's theorem with $x_0 = 0$, we have

$$P_n(x) = 1 + px + p(p-1)\frac{x^2}{2!} + \dots + p(p-1)\dots(p-n+1)\frac{x^n}{n!}.$$

In addition, for any $x \in [0,1]$ there exists a number $\xi(x)$ between 0 and x such that

$$f(x) = P_n(x) + R_n(x)$$

where

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}x^{n+1} = \frac{p(p-1)\dots(p-n)(1+\xi(x))^{p-(n+1)}}{n+1)!}x^{n+1}.$$

Therefore finding an upper bound for the error $|f(x) - P_n(x)|$ on $[0,1]$ is the same as finding an upper bound for the remainder term $|R_n(x)|$ on $[0,1]$. Clearly the term $|x|^{n+1} \leq 1$ on $x \in [0,1]$. For any $x \in [0,1]$, we must have $\xi(x) \in [0,1]$ and so

$$\max_{x\in[0,1]} |1+\xi(x)|^{p-(n+1)} \leq \max_{y\in[0,1]} (1+y)^{p-(n+1)}.$$

Now we have two cases to consider depending on the sign of $p - (n+1)$.
If $p - (n+1) \geq 0$, then for any $y \in [0,1]$

$$(1+y)^{p-(n+1)} \leq 2^{p-(n+1)}.$$

Otherwise, if $p - (n+1) < 0$, then for any $y \in [0,1]$

$$(1+y)^{p-(n+1)} = \frac{1}{(1+y)^{(n+1)-p}} \leq 1.$$

Thus,

$$\max_{y\in[0,1]} (1+y)^{p-(n+1)} \leq \max\left\{1, 2^{p-(n+1)}\right\}.$$

Applying the above estimates to $|R_n(x)|$ we get the following upper bound

$$\max_{x\in[0,1]} |R_n(x)| \leq \frac{|p(p-1)\dots(p-n)|}{(n+1)!} \max\left\{1, 2^{p-(n+1)}\right\}.$$

3. Consider the following function $f(x) = x^3 - 2x - 5$.

   (a) (5 points) Show that f has one unique root x^* with $x^* \in [2,3]$.

Solution: We have $f(2) = 8 - 4 - 5 = -1$ and $f(3) = 27 - 6 - 5 = 16$. Then by the Intermediate Value Theorem f has a root in $[2,3]$. We are left to prove that the root is unique.

We have $f'(x) = 3x^2 - 2$ and so the zeros of f' are $x_1 = -\sqrt{\frac{2}{3}}$ and $x_2 = \sqrt{\frac{2}{3}}$. In addition

	$-\infty$	x_1		x_2	∞
$f'(x)$	$+$	0	$-$	0	$+$
$f(x)$	$\nearrow$	≈ -3.91	$\searrow$	≈ -6.09	$\nearrow$

Table 1: Behavior of f.

In addition,

$$\lim_{x\to-\infty} f(x) = -\infty \quad \text{and} \quad \lim_{x\to\infty} f(x) = \infty.$$

Therefore f has only one root.

(b) (10 points) Estimate the number of iterations for the bisection method to reach an accuracy of 10^{-10} and compare with the actual number of iterations.

Solution: We apply the bisection starting with the interval $[2, 3]$. Then, using the estimate derived in class, we have that

$$k > \frac{\log\left(\frac{3-2}{10^{-10}}\right)}{\log(2)} - 1 = 32.22...$$

Hence we need 33 iterations to guarantee an accuracy of 10^{-10}.

In MATLAB running the script "bisection.m" gives $k = 31$, where the stopping criteria was $|x_k - x^*| \leq 10^{-10}$.

```
 >> f = @(x)x^3-2*x-5;
 >> bisection(2,3,f,1e-10);

x = 2.5000000000000000 for k=0
x = 2.2500000000000000 for k=1
x = 2.1250000000000000 for k=2
x = 2.0625000000000000 for k=3
x = 2.0937500000000000 for k=4
x = 2.1093750000000000 for k=5
x = 2.1015625000000000 for k=6
x = 2.0976562500000000 for k=7
x = 2.0957031250000000 for k=8
x = 2.0947265625000000 for k=9
x = 2.0942382812500000 for k=10
x = 2.0944824218750000 for k=11
x = 2.0946044921875000 for k=12
x = 2.0945434570312500 for k=13
x = 2.0945739746093750 for k=14
x = 2.0945587158203125 for k=15
x = 2.0945510864257812 for k=16
x = 2.0945549011230469 for k=17
x = 2.0945529937744141 for k=18
x = 2.0945520401000977 for k=19
```

```
x = 2.0945515632629395 for k=20
x = 2.0945513248443604 for k=21
x = 2.0945514440536499 for k=22
x = 2.0945515036582947 for k=23
x = 2.0945514738559723 for k=24
x = 2.0945514887571335 for k=25
x = 2.0945514813065529 for k=26
x = 2.0945514850318432 for k=27
x = 2.0945514831691980 for k=28
x = 2.0945514822378755 for k=29
x = 2.0945514817722142 for k=30
x = 2.0945514815393835 for k=31
```

(c) (5 points) Consider the fixed point iteration $x_{n+1} = g(x_n)$ with $x_0 \in [2,3]$, where $g(x) = (5+2x)^{\frac{1}{3}}$. Prove that x_n converges to x^*.

Solution: It's easy to see that x^* is a fixed point of g since

$$f(x^*) = 0 \iff (x^*)^3 - 2x^* - 5 = 0 \iff (x^*)^3 = 2x^* + 5 \iff x^* = g(x^*).$$

To prove convergence, we need to show $2 \leq g(x) \leq 3$ for all $x \in [2,3]$ and that there is $L > 0$ such that $|g'(x)| \leq L < 1$ for all $x \in [2,3]$. We start by computing the derivative of g:

$$g'(x) = \frac{2}{3(5+2x)^{3/2}}$$

In the interval $[2,3]$, $g' > 0$ and so g is strictly increasing there. Hence for all $x \in [2,3]$,

$$2 < \sqrt[3]{9} = g(2) \leq g(x) \leq g(3) = \sqrt[3]{11} < 3$$

and so $2 \leq g(x) \leq 3$ for all $x \in [2,3]$. In addition,

$$|g'(x)| = \left|\frac{2}{3(5+2x)^{2/3}}\right| \leq \left|\frac{2}{3(5+4)^{2/3}}\right| = \frac{2}{9 \times 3^{1/3}} := L < 1.$$

Note that $L \approx 0.15048$. Hence we can conclude that x_k converges to x^*.

(d) (15 points) For $x_0 = 2.5$, compute the number of iterations needed to reach an accuracy of 10^{-10} for

i. the fixed point method in part (c),

ii. Newton's method,

iii. secant method (use $x_1 = 3$).

Solution: In MATLAB, we run the scripts "fixedPoint.m", "newton.m" and "secant.m". The results follow.

```
 >> g = @(x)nthroot(5+2*x,3);
 >> fixedPoint(2.5,g,1e-10);

x = 2.1544346900318838 for k=1
x = 2.1036120286023885 for k=2
```

```
x = 2.0959274099028793 for k=3
x = 2.0947605454854048 for k=4
x = 2.0945832502254325 for k=5
x = 2.0945563090711339 for k=6
x = 2.0945522151289175 for k=7
x = 2.0945515930174659 for k=8
x = 2.0945514984819864 for k=9
x = 2.0945514841164616 for k=10
x = 2.0945514819334896 for k=11
x = 2.0945514816017674 for k=12

 >> f = @(x)x^3-2*x-5;
 >> fprime = @(x)3*x^2-2;
 >> newton(2.5,f,fprime,1e-10);

x = 2.1641791044776117 for k=1
x = 2.0971353558105545 for k=2
x = 2.0945552323904479 for k=3
x = 2.0945514815502468 for k=4

 >> f = @(x)x^3-2*x-5;
 >> fprime = @(x)3*x^2-2;
 >> secant(2.5,4,f,1e-10);

x = 2.3140495867768598 for k=1
x = 2.2174702908134281 for k=2
x = 2.1078895308294938 for k=3
x = 2.0954260818947001 for k=4
x = 2.0945580114467495 for k=5
x = 2.0945514847563875 for k=6
x = 2.0945514815423385 for k=7
```

The actual number of iterations is 12, 4 and 7, respectively.

(e) (5 points) Rank all four methods by how fast they converge.

Solution: Newton's method was the fastest, followed by the secant method, the fixed point method and finally the bisection method, which was expected due to their order of convergence and asymptotic error constants. Newton's method has at least quadratic convergence (since $f'(x^*) \neq 0$) and the secant method has order of convergence $\frac{1+\sqrt{5}}{2} \approx 1.61$. The fixed point method method has linear convergence with A.E.C $|g'(x^*)| \approx 0.15$ and we expect the bisection to have a linear convergence with A.E.C $\frac{1}{2}$. This explains why the fixed point method converged faster than the bisection method.

Hint: to compute the actual number of iterations for each method, you may use the fact that $x^* = 2.0945514815423\ldots$.

4. Consider the iteration $x_{k+1} = g(x_k)$ where

$$g(x) = \frac{\lambda x - \log(x) - 2(x+1)}{\lambda - 2}$$

with $\lambda \neq 2$.

(a) (5 points) Compute the fixed point x^* of g.

Solution: The fixed point x^* satisfies $g(x) = x$. Hence

$$\begin{aligned} g(x) = x \iff & \frac{\lambda x - \log(x) - 2(x+1)}{\lambda - 2} = x \\ \iff & \frac{\lambda x - \log(x) - 2(x+1) - \lambda x + 2x}{\lambda - 2} = 0 \\ \iff & \frac{-\log(x) - 2}{\lambda - 2} = 0 \\ \iff & -2 = \log(x) \\ \iff & x = e^{-2}. \end{aligned}$$

We then have $x^* = e^{-2}$.

(b) (5 points) Determine the values of λ for which $|g'(x^*)| < 1$. Why are these values important?

Solution: We first compute $g'(x)$:

$$g'(x) = \frac{\lambda - x^{-1} - 2}{\lambda - 2}.$$

Then we want

$$\left| \frac{\lambda - e^2 - 2}{\lambda - 2} \right| < 1.$$

It's easy to see by drawing the graphs of $|\lambda - e^2 - 2|$ and $|\lambda - 2|$ that

$$\left| \frac{\lambda - e^2 - 2}{\lambda - 2} \right| < 1 \Leftrightarrow \lambda > \lambda^*$$

where λ^* satisfies $-\lambda + e^2 + 2 = \lambda - 2$. Thus $\lambda^* = \frac{e^2+4}{2}$ and therefore the values of λ for which $|g'(x^*)| < 1$ are $\left(\frac{e^2+4}{2}, \infty\right)$. This is important since it guarantees that the fixed point iteration converges locally, i.e., if x_0 is close enough to x^*.

(c) (5 points) Determine the value of λ such that x_k converges as fast as possible to x^*.

Solution: For x_k to converge as fast as possible to x^* we need λ such that $g'(x^*) = 0$. Hence $\lambda = e^2 + 2$.

5. (15 points) Assume that $g \in C^p[a, b]$ has a fixed point x^* and that the fixed point iteration of g converges. Show that if $g^{(i)}(x^*) = 0$ for all $i = 1, \ldots, p-1$ and $g^{(p)}(x^*) \neq 0$, then the fixed point iteration converges at order p. State the asymptotic error constant in this case. *Hint: Expand g using Taylor's polynomial around $x_0 = x^*$.)*

Solution: Since $g \in C^p[a, b]$, we can apply Taylor's theorem. Expanding g around $x_0 = x^*$ and evaluating at x_k, we get

$$g(x_k) = g(x^*) + g'(x^*)(x_k - x^*) + \cdots + g^{(p-1)}(x^*)\frac{(x_k - x^*)^{p-1}}{(p-1)!} + g^{(p)}(\xi_k)\frac{(x_k - x^*)^p}{p!},$$

where ξ_k is between x^* and x_k. Since the fixed point iteration is defined as $x_{k+1} = g(x_k)$, the above equation implies

$$\frac{x_{k+1} - x^*}{(x_k - x^*)^p} = \frac{g^{(p)}(\xi_k)}{p!}.$$

Hence by the continuity of $g^{(p)}$ on $[a, b]$,

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = \lim_{k\to\infty} \frac{\left|g^{(p)}(\xi_k)\right|}{p!} = \frac{\left|g^{(p)}\left(\lim\limits_{k\to\infty} \xi_k\right)\right|}{p!} = \frac{\left|g^{(p)}(x^*)\right|}{p!}.$$

We conclude then that the fixed point iteration converges at order p with A.E.C of $\frac{\left|g^{(p)}(x^*)\right|}{p!}$.