Midterm Exam

Eshan Uniyal eshanuniyal@g.ucla.edu
205172354

1. Problem 1

Consider computing the quantity:

$$y = \frac{1}{x} - \frac{1}{x+1}, \qquad x > 0$$

(a) For what values of x do you expect cancellation of significant digits? Explain.

We expect catastrophic cancellation of significant digits for values of x such that:

$$\frac{1}{x} \approx \frac{1}{x+1} \quad \Longrightarrow \quad \frac{x}{x+1} \approx 1$$

i.e. catastrophic cancellation may occur for large values of x.

(b) Rewrite the expression for computing y so that it avoids cancellation for those values of x identified in part (a).

$$y = \frac{1}{x} - \frac{1}{x+1} = \frac{x+1}{x\cdot(x+1)} - \frac{x}{x\cdot(x+1)} = \frac{x+1-x}{x\cdot(x+1)} = \frac{1}{x\cdot(x+1)}$$

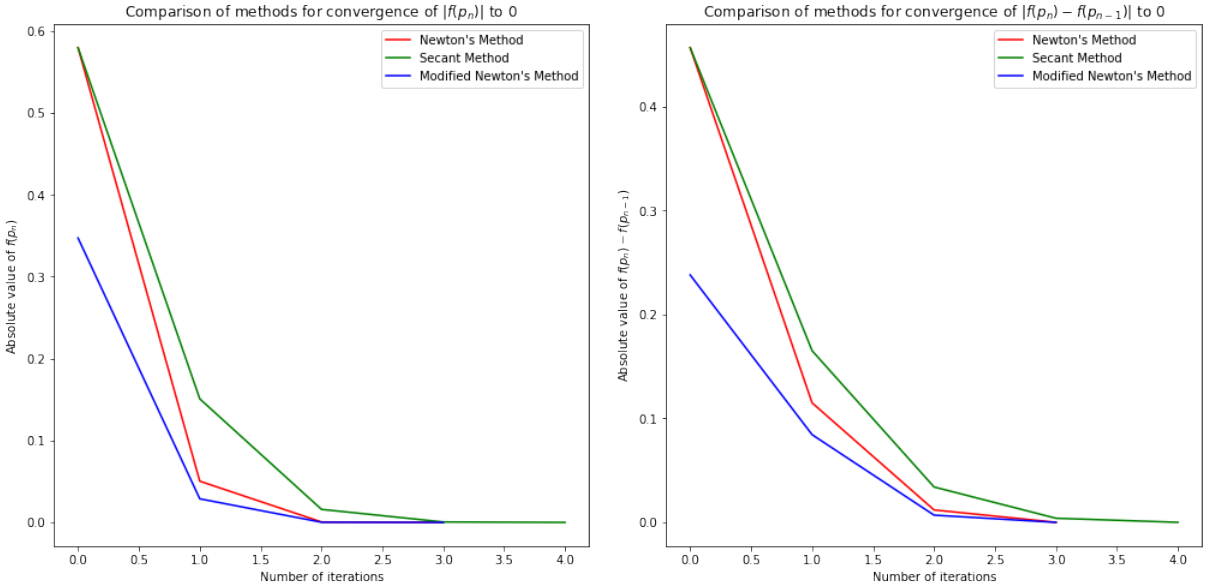
$$\Longrightarrow \quad y = \frac{1}{x\cdot(x+1)}$$

Since the above expression does not involve any subtraction, we can avoid cancellation of significant digits by using it to compute y. However, note that for extremely large values of x, we may have underflow errors involving rounding off to zero.

2. Problem 2

Consider the function $f(x) = e^x + 2^{-x} + 2cos(x) - 6$ for $1 \leq x \leq 2$. Use Newton's method, secant method, and the modified Newton's method to find a root of f to accuracy within 10^{-5} for the following two criteria: $|f(p_n)|$ and $|p_n - p_{n-1}|$. Plot your solutions and explain your findings.

We graph the convergence rate for each of the methods (taking $p_0 = 1.5$) as follows:



From each graph, it is clear that Modified Newton's method converges most quickly, followed by Newton's method and secant method in that order. This is consistent with our theoretical analysis of each method. Modified Newton's method improves on the rate of convergence of Newton's method and is quadratic regardless of the multiplicity of p, and should therefore be faster. Furthermore, since the secant method is only an approximation of Newton's method, it appropriately converges more slowly than Newton's method.

3. Problem 3

(a) By a theorem from class, show that the function $g(x) = 1 + e^{-x}$ has a unique fixed point on $[1, 2]$ (given values: $e^{-1} = 0.3679$, $e^{-2} = 0.1353$).

Let $a := 1$, $b := 2$. We note that $g(x) = e^{-x} + 1$ is continuous for all $x \in \mathbb{R} \implies g(x)$ is continuous on $[a, b]$.

Computing the derivative of $g(x)$, we have $g'(x) = -e^{-x} < 0 \ \forall x \in \mathbb{R}$.

$\implies g'(x) < 0 \ \forall \ x \in [1, 2]$

$\implies g(x)$ is strictly or monotonically decreasing.

Since $g(a) = e^{-1} + 1 = 1.3679$, and $g(b) = e^{-2} + 1 = 1.1353 > g(a) > a$, and $g(x)$ is strictly decreasing on $[a, b]$, we have $g(x) \in [a, b] \ \forall x \in [a, b]$.

$\therefore$ By the Existence of Fixed Point Theorem, there exists a fixed point $p \in [a, b]$ such that $g(p) = p$.

Furthermore, since g is an exponential function, it is continuously differentiable on R. $\therefore g$ is continuously differentiable on $[a, b]$.

To prove uniqueness, we now only need to show there exists $k \in (0, 1)$ s.t. $|g'(x)| \leq k \ \forall x \in (a, b)$. Since $g'(x) = -e^{-x}$, we have

$$|g'(x)| = |-e^{-x}| = e^{-x} \leq e^{-1} \ \forall x \in [1, 2]$$

$$\implies |g'(x)| \leq e^{-1} = 0.3679 \ \forall x \in [a, b]$$

Therefore, $k = e^{-2} < 1$ satisfies the remaining condition for proving uniqueness by the Uniqueness of Fixed Point Theorem. Hence shown there exists a unique fixed point of $g(x)$ on $[1, 2]$. ■

(b) Using $p_0 = 1$, how many iterations does the theory predict it will take to achieve 10^{-5} accuracy, to approximate the fixed point, starting with $p_0 = 1$?

The error bound for p_n in fixed point iteration can be given by:

$$|p_n - p| \ \leq \ k^n \cdot max\{|p_0 - a|, |p_0 - b|\}$$

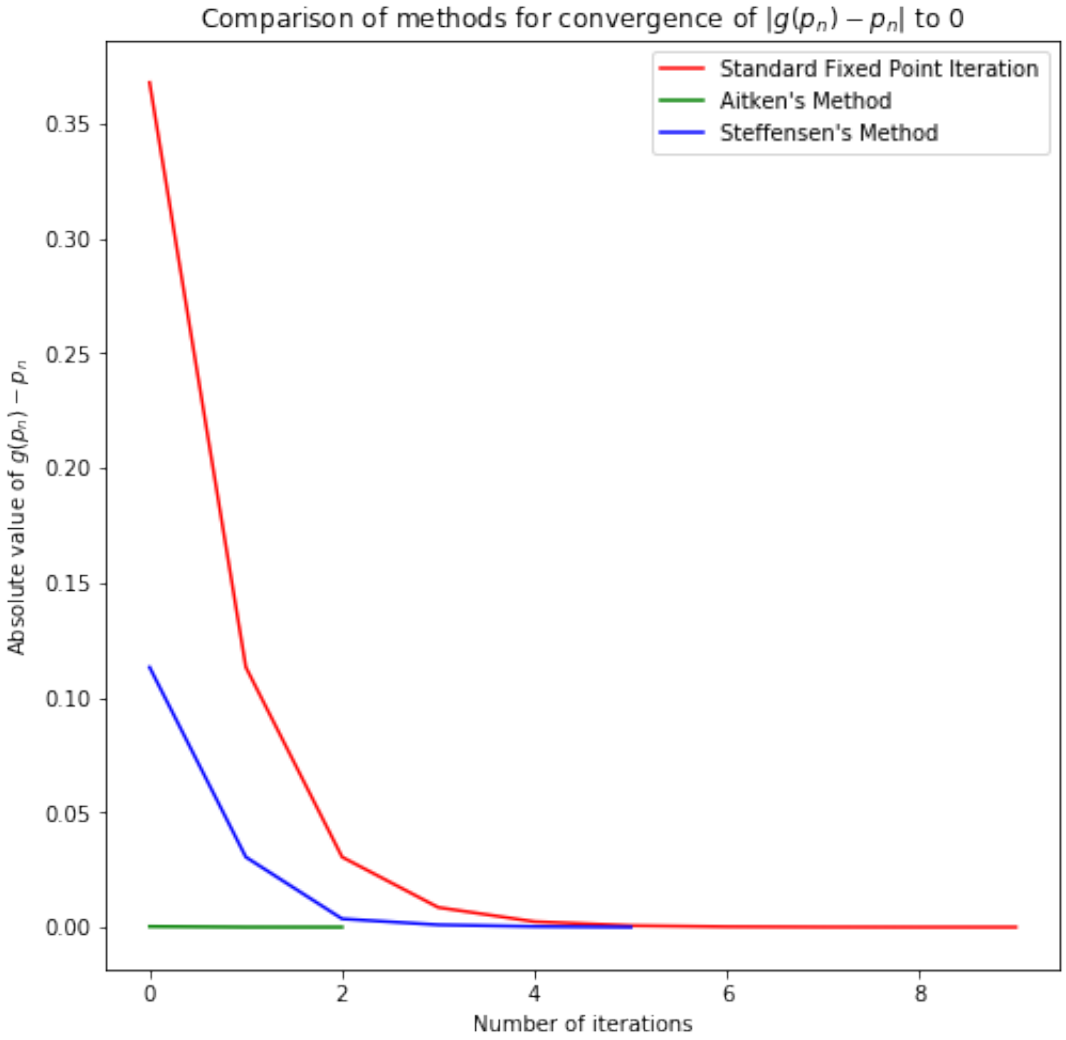
For $[a, b] = [1, 2\pi]$, $p_0 = 1$ gives $max\{|p_0 - a| \, |p_0 - b|\} = max\{|1 - 1|, |1 - 2| = 1$. From (a), we also have $k = 0.3679$.

$$\therefore |p_n - p| \ \leq \ k^n \cdot max\{|p_0 - a|, |p_0 - b|\} \ \leq \ 10^{-5}$$

$$\begin{aligned}
\implies \quad 0.3679^n \cdot 1 &\leq 10^{-5} \\
\implies \quad 0.3679^n &\leq 10^{-5} \\
\implies \quad n \cdot \log_{10}(0.3679) &\leq \log_{10}(10^{-5}) \\
\implies \quad n &\geq \frac{\log_{10}(10^{-5})}{\log_{10}(0.3679)} \quad (\because \log_{10}(0.3.679) < 0) \\
\implies \quad n &\geq 11.52
\end{aligned}$$

We therefore find that for $p_0 = 1$, 11 fixed point iterations give a result with at least 10^{-5} accuracy.

(c) Program a standard fixed point iteration, Aitken's method, and Steffensen's method to find a fixed point with accuracy 10^{-5} using $p_0 = 1$. Use the stopping criteria: $|g(p_n) - p_n|$. Plot your solutions and explain your findings.



We see that Aitken's method seems to converge significantly faster than Steffensen's method, which in turn is much faster than standard fixed point iteration. While the faster convergence rate of Steffensen's method is accurate with our theoretical analysis (since we take Aitken's δ^2 term every third iteration to speed up the fixed point iteration), Aitken's method does not give us faster convergence than Steffensen's method. Convergence for Aitken's method only seems much faster above, since we are only graphing Aitken's δ^2 sequence, generating only the first time of which requires two standard fixed point iterations that are not plotted above. If we were to also plot the initial iterations for Aitken's method, we would see Aitken's and Steffensen's methods converge at nearly the same rate, and both are much faster than the standard fixed point iteration.

4. Problem 4

Suppose that a function f has m continuous derivatives on the interval $[a, b]$ containing p. Show: f has a zero of multiplicity m at p if and only if

$$0 = f(p) = f'(p) = ... = f^{(m-1)}(p) \text{ but } f^{(m)}(p) \neq 0$$

We know that, by definition of multiplicity, p is a zero of f multiplicity m if and only if for $x \neq p$, f can be written as $f(x) = (x-p)^m \cdot q(x)$ for some function $q(x)$ s.t. $\lim_{x\to p} q(x) \neq 0$. We take $f(x) = (x-p)^m \cdot q(x)$ and first show, by induction on m, that $0 = f(p) = f'(p) = ... = f^{(m-1)}(p)$.

By definition, $f(p) = 0$. We only need to show that the j^{th} derivative of f for all $j \in \{1, 2, ..., m-2, m-1\}$ is 0 at p.

Base case: For $j = 1$, we have $f^{(j)}(x) = f'(x) = m \cdot (x-p)^{m-1} \cdot q(x) + (x-p)^m \cdot q'(x)$.
$\implies f'(x) = (x-p)^{m-1} \cdot (m \cdot q(x) + (x-p) \cdot q'(x))$

If we define $q_1(x) := m \cdot q(x) + (x-p) \cdot q'(x)$, it is clear that $\lim_{x\to p} q_1(x) = \lim_{x\to p} m \cdot q(x) \neq 0$ (since by definition, $\lim_{x\to p} q(x) \neq 0$). Therefore, by definition, p is a zero of multiplicity $m-1$ for $f'(x) = (x-p)^{m-1} \cdot q_1(x)$. It follows naturally that that $f'(p) = 0$.

Inductive step: Let $j \in \{1, 2, ..., m-2\}$ be given such that p is a zero of multiplicity $m-j$ for $f^{(j)}(x)$. Therefore, by definition, $f^{(j)}$ can be expressed as $f^{(j)}(x) = (x-p)^{m-j} \cdot q_j(x)$ for some $q_j(x)$ s.t. $\lim_{x\to p} q_j(x) \neq 0$.

$$\begin{aligned}
\therefore f^{(j+1)}(x) &= \frac{d}{dx}(x-p)^{m-j} \cdot q_j(x) \\
&= (m-j) \cdot ((x-p)^{m-(j+1)} \cdot q_j(x) + (x-p)^{m-j} \cdot q_j'(x) \\
&= (x-p)^{m-(j+1)} \cdot ((m-j) \cdot q_j(x) + (x-p) \cdot q_j'(x) \\
&= (x-p)^{m-(j+1)} \cdot q_{j+1}(x) \qquad \text{where } q_{j+1}(x) = (m-j) \cdot q_j(x) + (x-p) \cdot q_j'(x)
\end{aligned}$$

Since $\lim_{x\to p} q_j(x) \neq 0$ (by assumption),

$$\lim_{x\to p} q_{j+1}(x) = \lim_{x\to p} (m-j) \cdot q_j(x) + (x-p) \cdot q_j'(x) = \lim_{x\to p} (m-j) \cdot q_j(x) \neq 0$$

.

$\therefore$ By definition of multiplicity, p is a zero of multiplicity $m-(j+1) > 0$ (by definition of j) for $f^{(j+1)}(x) = (x-p)^{m-(j+1)} \cdot q_{j+1}(x)$. It follows naturally that $f^{(j+1)}(p) = 0$.

$\therefore p$ is a zero of multiplicity $m-j$ for $f^{(j)}(x) \implies p$ is a zero of multiplicity $m-(j+1)$ for $f^{(j+1)}(x)$ $(j \in 1, 2, \ ... \ , m-2$.

By induction, it follows that p is a zero of multiplicity $m-j > 0$ for all $j \in \{1, 2, \ ... \ , m-1\}$. Furthermore, since p is therefore a zero of multiplicity 1 for $f^{(m-1)}(p)$, by definition of multiplicity, p is not a zero for $f^{(m)}(x)$ i.e. $f^{(m)}(p) \neq 0$.

$$\therefore 0 = f(p) = f'(p) = ... = f^{(m-1)}(p) \text{ but } f^{(m)}(p) \neq 0$$

5. Problem 5

(a) State two equivalent definitions for a zero of multiplicity m for a function $f \in C^m[a,b]$.

Definition 1: p is a zero of f multiplicity m if for $x \neq p$, f can be written as

$$f(x) = (x-p)^m \cdot q(x)$$

for some function $q(x)$ s.t. $\lim_{x\to p} q(x) \neq 0$.

Definition 2: p is a zero of f multiplicity m if

$$0 = f(p) = f'(p) = ... = f^{(m-1)}(p) \text{ but } f^{(m)}(p) \neq 0$$

(b) Suppose p is a zero of multiplicity m of f, where $f^{(m)}$ is continuous on an open interval containing p. Show that the following fixed point method has $g'(p) = 0$:

$$g(x) = x - \frac{mf(x)}{f'(x)}$$

Since p is a zero of multiplicity m of f, by definition of multiplicity, we may express f as $f(x) = (x-p)^m \cdot q(x)$ for some $q(x)$ s.t. $\lim_{x\to p} q(x) \neq 0$.

$$f(x) = (x-p)^m \cdot q(x) \implies f'(x) = (x-p)^{m-1} \cdot (m \cdot q(x) + (x-p) \cdot q'(x))$$

$$\begin{aligned}
\therefore g(x) &= x - \frac{mf(x)}{f'(x)} \\
\implies g(x) &= x - \frac{m(x-p)^m q(x)}{(x-p)^{m-1} \cdot (mq(x) + (x-p)q'(x))} \\
&= x - \frac{m(x-p)q(x)}{mq(x) + (x-p)q'(x)} \\
\implies g'(x) &= 1 - \frac{(mq(x) + (x-p)q'(x)) \cdot (mq(x) + m(x-p)q'(x))}{(mq(x) + (x-p)q'(x))^2} \\
&\quad + \frac{m(x-p)q(x) \cdot (mq'(x) + q'(x) + (x-p)q''(x))}{(mq(x) + (x-p)q'(x))^2} \qquad \text{(quotient rule)} \\
\implies g'(p) &= 1 - \frac{(mq(x) + 0 \cdot q'(x)) \cdot (mq(x) + m \cdot 0 \cdot q'(x))}{(mq(x) + 0 \cdot q'(x))^2} \\
&\quad + \frac{m \cdot 0 \cdot q(x) \cdot (mq'(x) + q'(x) + 0 \cdot q''(x))}{(mq(x) + 0 \cdot q'(x))^2} \\
&= 1 - \frac{mq(x) \cdot mq(x)}{(mq(x))^2} \\
&= 1 - \frac{(mq(x))^2}{(mq(x))^2} = 1 - 1 = 0
\end{aligned}$$

$\therefore$ The given fixed point method has $g'(p) = 0$. ■