McGill University April 2012
Faculty of Science Final examination

Honours Numerical Analysis

Math 387

Friday, April 20th, 2012

Time: 9am-12pm

Examiner: Prof. A.R. Humphries Associate Examiner: Prof. J.C. Nave

INSTRUCTIONS

1. All questions carry equal weight.
2. **Answer 6 or 7 questions; credit will be given for the best 6 answers.**
3. Answer questions in the exam book provided. Start each answer on a new page.
4. This is a closed book exam.
5. Notes and textbooks are not permitted.
6. Non-programmable calculators are permitted.
7. Translation dictionaries (English-French) are permitted.

This exam comprises the cover page and three pages of questions, numbered 1 to 7.

1. (a) Let $x = a_1.a_2 \dots a_k a_{k+1} \dots \times 10^e$ where e is an integer, the a_i's are decimal digits and $a_1 \geqslant 1$. Describe how an approximation, $fl(x)$, to x is obtained by

   i. k (significant) digit decimal chopping,

   ii. k (significant) digit decimal rounding.

   Show in the case of chopping that

$$\frac{|x - fl(x)|}{|x|} \leqslant 10^{1-k}.$$

   (b) Using the inequality

$$\frac{|xy - \widetilde{x}\widetilde{y}|}{|xy|} \leqslant \frac{|x-\widetilde{x}|}{|x|} + \frac{|y-\widetilde{y}|}{|y|} + \frac{|x-\widetilde{x}|}{|x|}\,\frac{|y-\widetilde{y}|}{|y|},$$

   show that in the case of k digit chopping,

$$\frac{|xy - fl(x)fl(y)|}{|xy|} \leqslant 2 \times 10^{1-k} + 10^{2-2k}$$

   and hence that

$$\frac{|xy - fl\big(fl(x)fl(y)\big)|}{|xy|} \leqslant 3 \times 10^{1-k} + c \times 10^{2-2k}$$

   where c is a constant of order 1 that you do not need to find.

   (c) Using the formula for roots of a quadratic and 3 significant digit decimal chopping compute numerical approximations $\widetilde{x}_1$, $\widetilde{x}_2$ to the roots x_1, x_2 of $x^2 - 100x - 1 = 0$. Organise your calculations so as to minimise the effect of the errors.

2. (a) State the (Banach) fixed point theorem for a fixed point iteration $x_{n+1} = g(x_n)$ (You may omit the error bounds).

   (b) Show that under the conditions of the theorem that in a suitable interval there is at least one point x^* such that $g(x^*) = x^*$, and moreover that for any point x_0 contained in the interval $\lim_{n\to\infty} x_n = x^*$.

   (c) The iteration $g(x) = x - \frac{1}{2}(e^x - 2)$ is proposed to find $\ln 2$. Show that the conditions of the fixed point theorem are satisfied on a suitable interval which you should identify.

   (d) Let x_0 be the midpoint of the interval you identified in (c). Identify the constant L in the error formula

$$|x_n - x^*| \leqslant \frac{L^n}{1-L}|x_1 - x_0|,$$

   and use the formula to find an upper bound on N such that $|x_n - x^*| \leqslant 10^{-8}$ for all $n \geqslant N$. Briefly, do you expect the actual N for which $|x_n - x^*| \leqslant 10^{-8}$ for all $n \geqslant N$ to be close to the number that you just found or not, and why do you expect this?

3. (a) Write Newton's method for finding x^* such that $f(x^*) = 0$ as a fixed point iteration $x_{n+1} = g(x_n)$ and show that $g'(x^*) = 0$ when $f'(x^*) \neq 0$.

   (b) Let f have a zero of multiplicity m at x^*. Define what this means. Show that in this case Newton's method satisfies $\lim_{x \to x^*} g'(x) = 1 - 1/m$.

   (c) Newton's method applied to a certain $f(x)$ gives iterates

$$x_0 = 1, \quad x_1 = 1.25, \quad x_2 = 1.3375, \quad x_3 = 1.3769568$$

$$x_4 = 1.3958372, \quad x_5 = 1.4050859, \quad x_6 = 1.4096645$$

   which appear to converge linearly. Evaluate $\hat{x}_4$ where

$$\hat{x}_n = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}$$

   is defined by Aitken's Δ^2 method. Use x_5, x_6 and $\hat{x}_4$ to estimate the error constant for the Newton iterates above. What is the apparent multiplicity of the zero that Newton's method is converging to?

   (d) State a variant of Newton's method which will converge quadratically for the problem in (c) (you do not need to show that the convergence is quadratic).

4. (a) Let $f(x)$ be $n+1$ times continuously differentiable on $[a, b]$ and $x_0, x_1, \ldots, x_n$ be distinct interpolation points in $[a, b]$. Define the fundamental Lagrange polynomials $l_0(x), l_1(x), \ldots, l_n(x)$ for the interpolation points and show that

$$p_n(x) = \sum_{j=0}^{n} f(x_j) l_j(x)$$

   interpolates f at $x_0, x_1, \ldots, x_n$.

   (b) Show that $p_n(x)$, defined in (a), is the unique interpolating polynomial of degree less than or equal to n.

   (c) Suppose that $n = 3$

$$x_0 = 0, \quad x_1 = 1, \quad x_2 = 2, \quad x_3 = 3,$$

$$f(x_0) = 0, \quad f(x_1) = 0, \quad f(x_2) = 4, \quad f(x_3) = 6.$$

   Find $p_3(x)$ and evaluate $p_3(2.5)$. Find a bound for the error in this approximation of $f(2.5)$, when $\max_{x \in [0,3]} |f^{(4)}(x)| \leqslant 10$, using the error formula

$$f(x) = p_n(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \prod_{i=0}^{n} (x - x_i).$$

5. (a) Suppose that $f(x)$ is four times continuously differentiable, and show that

$$f''(x_0) = \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} - \frac{h^2}{12} f^{(4)}(\xi), \qquad \xi \in [x_0-h, x_0+h].$$

   (b) Suppose $|f^{(4)}(x)| \leqslant M$ for all $x \in [x_0-h, x_0+h]$, and that $h > 0$. Let $f_h(x)$ be a finite precision evaluation of $f(x)$, and assume that there exists $\delta > 0$ such that $|f_h(x) - f(x)| \leqslant \delta$ for all $x \in [x_0-h, x_0+h]$. Derive a formula for an upper bound on the error

$$\left| f''(x_0) - \frac{f_h(x_0+h) - 2f_h(x_0) + f_h(x_0-h)}{h^2} \right|.$$

   Determine the value of h which minimises this bound when $M = 100$ and $\delta = 10^{-16}$, and state the bound in this case.

   (c) Is the error formula given in (a) appropriate for the direct application of Richardson extrapolation? Whether it is or not, apply one step of Richardson extrapolation to the finite difference approximation in (a) to obtain an approximation to $f''(x_0)$ with $\mathcal{O}(h^4)$ error.

6. (a) Define the degree of accuracy (also known as the degree of precision) of a quadrature formula $I_h(f)$ for approximating the integral

$$I(f) = \int_a^b f(x)dx.$$

   (b) Find constants α, β and γ such that the degree of accuracy of the quadrature formula

$$I_h(f) = h[\alpha f(a) + \beta f(a + \gamma h)]$$

   is as large as possible, where $h = (b-a)$.

   (c) What is the degree of accuracy p of the method in (b)? Given that $I(f) = I_h(f) + kh^{p+2} f^{(p+1)}(\xi)$, find k.

   (d) Use this method and the fact that

$$\ln(2) = \int_1^2 \frac{1}{x} dx$$

   to obtain an approximation to $\ln(2)$. Use the error formula from above to derive an upper bound for the error in this approximation.

7. Consider the initial value problem

$$y' = f(y), \qquad 0 \leqslant t \leqslant T, \quad y(0) = \alpha.$$

Suppose you approximate the solution $y(t)$ using the Runge-Kutta method

$$y_0 = \alpha,$$
$$y_{n+1} = y_n + hf\left(y_n + \frac{1}{2} h f(y_n)\right), \quad n = 0, ..., N$$

with time-step h.

   (a) Define the local truncation error $\tau_{n+1}(h)$ of the method, and determine the order of the method.

   (b) Consider the case where

$$f(y) = \lambda y, \quad \lambda < 0,$$

   and

   i. show that $y_{n+1} = \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right) y_n$.

   ii. Under what conditions on h does $\lim_{n\to\infty} y_n = 0$?