ECE 204 *Numerical methods*

FINAL EXAMINATION

Douglas Wilhelm Harder dwharder@uwaterloo.ca EIT 4018 x37023

1. **The exam will be graded out of 50.**
2. No notes and no calculators.
3. Turn off all electronic media and store them under your desk.
4. If there is insufficient room, use the back of the last page.
5. You may ask only one question during the examination: "May I go to the washroom?"
6. Asking **any** other question **will** result in a deduction of 5 marks from the exam grade.
7. If you think a question is ambiguous, write down your assumptions and continue.
8. **Do not leave during first hour or after there are only 15 minutes left.**
9. Do not stand up until all exams have been picked up.
10. If a question only asks for an answer, you do not have to show your work to get full marks; however, if your answer is wrong and no rough work is presented to show your steps, no part marks will be awarded.
11. The questions are approximately in the order of the course material.

1 [2] Add the two numbers stored as double-precision floating-point numbers

```
1 01101001101 10100000…0
1 01101001111 01000000…0
```

showing the addition in binary and write the resulting representation in binary.

2 [3] Solve the following system of linear equations using Gaussian elimination with partial pivoting. Zero marks will be awarded if partial pivoting is not used.

$$\left(\begin{array}{rrr|r} -3.5 & 0.6 & 3.3 & 17.5 \\ 1.5 & 1.8 & -0.3 & -2.1 \\ 5 & 2 & 1 & -5 \end{array}\right)$$

3 [3] Find the quadratic polynomial $at^2 + bt + c$ in the variable t that passes through the three points $(-1, y_{-1})$, $(0, y_0)$ and $(1, y_1)$. You will find a, b and c in terms of the three unknowns y_{-1}, y_0, and y_1. You will need to use a Vandermonde matrix to find your solution, but you need not use partial pivoting if you prefer not to.

4 [2] Write down the 2nd-order Taylor series approximation of e^{0+h} and use this to approximate the value of $e^{0.1}$. The correct answer, to 10 decimal digits, is 1.105170918. What is the ratio that must be calculated to find the relative error of using the 2nd-order Taylor series approximation of $e^{0.1}$?

5 [3] Approximate the integral of $y(t) = t^2 + 1$ from 0 to 1 by first using two steps of the trapezoidal rule, and then one step of Simpson's rule. Recall that the error of the composite trapezoidal rule is proportional to the 2nd derivative while the error of Simpson's rule is proportional to the 4th derivative, so what can you say about the second approximation?

6 [2] You have a reasonably exact reading at the previous time step, $y(t-h)$, and you have another reasonably exact reading at the next time step, $y(t-2h)$, but an error in transmission resulted in the reading $y(t)$ to be lost. You will estimate the reading using the average $\frac{y(t-2h)+y(t-h)}{2}$. What is the error of this approximation? You suspect the error is $O(h)$, so you will use a zeroth-order Taylor series, so, for example, $y(t+h) = y(t) + y^{(1)}(\tau)h$.

7 [3] Continuing from Question 6, you instead estimate the reading using the weighted average $2y(t-h) - y(t-2h)$. What is the error of this approximation? You suspect the error is $O(h^2)$, so you will use a first-order Taylor series, so, for example, $y(t + h) = y(t) + y^{(1)}(t)h + ½\, y^{(2)}(\tau)h^2$. Which formula should you use, that in Question 6 or 7?

8 [4] Use two steps of Newton's method to approximate the root of $f(x) = x^3 - x^2 + x - 2$ starting with the approximation $x_0 = 1$. Suppose you factor out the root you find, and applying Newton's method one more time, this time again starting with $x_0 = 1$, and you find it does not converge. What can you say about the two remaining roots, and what should you use as an initial point to find those two remaining roots?

9 [2] Suppose you wanted to find a simultaneous root to the system of non-linear equations $x^2 + 4y^2 - x + y = 1$ and $3x^2 + 2y^2 + x - y = 2$ and you intend to use Newton's method in two dimensions, and your first approximation of that simultaneous root is $x = y = 0.5$. What is the system of linear equations you must solve (do not solve it).

10 [2] Using Heun's method, approximate y(1) with h = 1 for the initial-value problem defined by

$$y^{(1)}(t) = -y(t) - 2t + 1$$
$$y(0) = 2$$

11 [2] The mid-point method is a variation on Huen's method, where you are given a system of initial-value problems

$$\mathbf{w}^{(1)}(t) = \mathbf{f}(t, \mathbf{w}(t))$$
$$\mathbf{w}(t_0) = \mathbf{w}_0$$

and we approximate $\mathbf{w}_{k+1}$ with the formula:

$$\mathbf{s}_0 = \mathbf{f}(t_k, \mathbf{w}_k)$$
$$\mathbf{s}_1 = f\left(t_k + \frac{h}{2}, \mathbf{w}_k + \frac{h}{2}\mathbf{s}_0\right)$$
$$\mathbf{w}_{k+1} = \mathbf{w}_k + h\mathbf{s}_1$$

Use this technique to approximate $y(1)$ and $z(1)$ with h = 1 given the system of two initial-value problems

$$y^{(1)}(t) = 2y(t) - z(t) + 1$$
$$z^{(1)}(t) = -y(t) + z(t) + 1$$
$$y(0) = 1$$
$$z(0) = 2$$

12 [3] Write down the system of linear equations that must be solved to approximate a solution to the boundary-value problem

$$u^{(2)}(x) + u^{(1)}(x) + u(x) + 1 = 0$$
$$u(0) = 2,$$
$$u(1) = 1$$

using h = ¼. You should not solve this system of linear equations.

13 [3] Show how we convert the heat-conduction equation in one dimension into a finite-difference equation starting with $\frac{\partial}{\partial t}u(t,x) = \alpha \frac{\partial^2}{\partial x^2}u(t,x)$. You must explain each step.

14 [2] Provide two justifications for using adaptive techniques (for example, the Dormand-Prince method) as being more appropriate to getting an acceptable approximation to an initial-value problem as compared to fixed step size methods (for example, the 4th-order Runge Kutta method).

15 [1] In the wave equation $\frac{\partial^2}{\partial t^2}u(t,\mathbf{x}) = c^2\nabla^2 u(t,\mathbf{x})$, suppose we are applying this to model incoming radio signals sent from the space craft Voyager 1. What is c?

16 [2] You have a function that you are attempting to maximize, and you have evaluated the function at three points: (0, 1), (1, 1.25) and (2, 0.5). Apply one step of successive parabolic interpolation to find a better approximation of the maximum of this function.

17 [1] What is the benefit of the golden-ratio search over simply dividing the interval $[a, b]$ into three equal sub-intervals (so a, $a + (b - a)/3$, $a + 2(b - a)/3$ and b)?

18 [3] Suppose you are attempting to use gradient descent to find a minimum of the function $f(x, y) = x^2 - xy - 3x + 4y + 2y^2 + 1$. What is the function of one variable that you would subsequently have to find the minimum of if you started with the initial point $x_0 = 1$ and $y_0 = -1$. Suppose you used another function like successive parabolic interpolations to find an approximation of that minimum in one dimension. How would you use this to find the next approximation x_1 and y_1? Do not find the actual minimum of your function of one variable, just assume that you have found an appropriate approximation and explain how you would use that to get the next step.

19 [2] Suppose you are trying to find a global minimum of a signal $y(t)$ on the interval [0, 10] and you know the signal is sinusoidal with a minimum period of 2. Explain why you cannot just use successive parabolic interpolation to find that minimum starting with the points $t = 0$, $t = 5$ and $t = 10$. You can give a graphical example where such an approach may fail.

20 [1] Suppose that a polynomial is of degree n and its coefficients are being stored in an array of that capacity. What is the run-time of the algorithm that divides out the root r resulting in a polynomial of degree $n - 1$?

21 [3] For each of the seven tools we learned, give one application of that tool in this course.

Floating-point representations

±EEMNNN ±M.NNN × 10^{EE-49}

seeeeeeeeeeebbbbbb…b $(-1)^s$ 1.bbbbbb…b × $10^{eeeeeeeeeee - 01111111111}$

where $01111111111_2 = 1023$.

Fixed-point theorem: Solving $x = f(x)$, choose x_0 and let $x_{k+1} \leftarrow f(x_k)$.

$$f(x+h) = \left(\sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(x) h^k\right) + \frac{1}{(n+1)!} f^{(n+1)}(\xi) h^{n+1} \quad \text{where } x < \xi < x+h.$$

$$f(x) = \left(\sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(x_0)(x-x_0)^k\right) + \frac{1}{(n+1)!} f^{(n+1)}(\xi) h^{n+1} \quad \text{where } x_0 < \xi < x.$$

Averaging noisy values with zero bias mitigates the effect, while differentiating noisy values magnifies the effect.

```
double horner( double a[], unsigned int degree; double x ) {
    double result{a[0]};
    for ( std::size_t k{1}; k <= degree; ++k ) {
        result += result*x + a[k];
    }
    return 0;
}
```

Formula of interest:

$$f^{(1)}(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{1}{6} f^{(3)}(\xi) h^2 \qquad y^{(1)}(t) = \frac{y(t) - y(t-h)}{h} + \frac{1}{2} y^{(2)}(\tau_-) h$$

$$y^{(1)}(t) = \frac{3y(t) - 4y(t-h) + y(t-2h)}{2h} + \frac{1}{3} y^{(3)}(\tau) h^2$$

$$f^{(2)}(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{1}{12} f^{(4)}(\xi) h^2 \qquad y^{(2)}(t) = \frac{y(t) - 2y(t-h) + y(t-2h)}{h^2} + y^{(3)}(\tau) h$$

$$\int_a^b f(x)\,\mathrm{d}x = \left(\frac{1}{2} f(a) + \frac{1}{2} f(b)\right)(b-a) - \frac{1}{12} f^{(2)}(\xi)(b-a)^3$$

$$\int_a^b f(x)\,\mathrm{d}x = \frac{1}{6}(f_0 + 4f_1 + f_2)(b-a) - \frac{1}{2880} f^{(4)}(\xi)(b-a)^5$$

$$\int_a^b f(x)\,\mathrm{d}x = \frac{1}{8}(f_0 + 3f_1 + 3f_2 + f_3)(b-a) - \frac{1}{6480} f^{(4)}(\xi)(b-a)^5$$

$$\int_a^b f(x)\,\mathrm{d}x = \frac{1}{2}\left(f(a) + 2\left(\sum_{k=1}^{n-1} f(a+kh)\right) + f(b)\right)h - f^{(2)}(\xi)\frac{b-a}{12}h^2$$

$$\int_a^b f(x)\,\mathrm{d}x = \frac{1}{3}\left(f_0 + 4\sum_{k=1}^{\frac{n}{2}} f_{2k-1} + 2\sum_{k=1}^{\frac{n}{2}-1} f_{2k} + f_n\right)h - f^{(4)}(\xi)\frac{b-a}{180}h^4$$

$$\int_a^b f(x)\,\mathrm{d}x = \left(-\frac{1}{24} f(a-h) + \frac{1}{2} f(a) + \frac{25}{24} f(a+h) + \left(\sum_{k=2}^{n-2} f(a+kh)\right) + \frac{25}{24} f(b-h) + \frac{1}{2} f(b) - \frac{1}{24} f(b+h)\right)h + f^{(4)}(\xi)\frac{11}{80}(b-a)h^4$$

```
>> A = vander( -n:0, 2 ); # n + 1 points
>> detAtA = round( det( A'*A ) );   % This should be an integer
>> round( detAtA*inv( A'*A )*A' )   % This should be an integer matrix
>> ans/detAtA
```

For five points, $a_1 \leftarrow \frac{-2y_{n-4} - y_{n-3} + y_{n-1} + 2y_n}{10}$ and $a_0 \leftarrow \frac{-2y_{n-4} + 2y_{n-2} + 4y_{n-1} + 6y_n}{10}$.

Goal	Estimation
Estimate $y(t_n + \delta h)$	$a_0 + \delta a_1$
Estimate the rate of change of y over time	a_1/h
Estimate the integral $\int_{t_n-h}^{t_n} y(t)\,\mathrm{d}t$	$(a_0 - a_1/2)h$
Estimate the integral $\int_{t_n}^{t_n+h} y(t)\,\mathrm{d}t$	$(a_0 + a_1/2)h$

For five points, $a_2 \leftarrow \frac{2y_{n-4} - y_{n-3} - 2y_{n-2} - y_{n-1} - 2y_n}{14}$, $a_1 \leftarrow \frac{26y_{n-4} - 27y_{n-3} - 40y_{n-2} - 13y_{n-1} + 54y_n}{70}$ and $a_0 \leftarrow \frac{31y_n + 9y_{n-1} - 3y_{n-2} - 5y_{n-3} + 3y_{n-4}}{35}$.

Goal	Estimation
Estimate $y(t_n + \delta h)$	$a_0 + \delta(a_1 + \delta a_2)$
Estimate the rate of change of y over time at time $t_n + \delta h$	$(a_1 + 2\delta a_2)/h$
Estimate the acceleration of y over time at time t_n	$2a_2/h^2$
Estimate the integral $\int_{t_n - h}^{t_n} y(t)\,\mathrm{d}t$	$(a_0 - a_1/2 + a_2/3)h$
Estimate the integral $\int_{t_n}^{t_n + h} y(t)\,\mathrm{d}t$	$(a_0 + a_1/2 + a_2/3)h$

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \text{ or } \frac{-2c}{b \mp \sqrt{b^2 - 4ac}}$$

Method	Requirements	Iteration step	Rate of convergence	Is convergence guaranteed?
Bisection	An interval $[a, b]$ with $f(a)$ having the opposite sign of $f(b)$	Let $c \leftarrow \frac{a+b}{2}$ and update whichever endpoint has the same sign as $f(c)$.	$O(h)$	Yes
Bracketed secant	An interval $[a, b]$ with $f(a)$ having the opposite sign of $f(b)$	Let $c \leftarrow \frac{af(b) - bf(a)}{f(b) - f(a)}$ and update whichever endpoint has the same sign as $f(c)$.	$O(h)$	Yes
Secant	Two initial approximations x_0 and x_1 with $\lvert f(x_0)\rvert > \lvert f(x_1)\rvert$	Let $x_k \leftarrow \frac{x_{k-2} f(x_{k-1}) - x_{k-1} f(x_{k-2})}{f(x_{k-1}) - f(x_{k-2})}$.	$O(h^{\phi})$	No
Newton's	An initial approximation x_0	Let $x_k \leftarrow x_{k-1} - \frac{f(x_{k-1})}{f^{(1)}(x_{k-1})}$.	$O(h^2)$	No
Muller's	Three initial approximations x_0, x_1 and x_2 with $\lvert f(x_0)\rvert > \lvert f(x_1)\rvert > \lvert f(x_2)\rvert$	The root of the interpolating quadratic polynomial passing through $(x_k, f(x_k))$, $(x_{k-1}, f(x_{k-1}))$ and $(x_{k-2}, f(x_{k-2}))$	$O(h^{1.839})$	No
Inverse quadratic interpolation	Three initial approximations x_0, x_1 and x_2 with $\lvert f(x_0)\rvert > \lvert f(x_1)\rvert > \lvert f(x_2)\rvert$	Evaluate the interpolating quadratic polynomial passing through $(f(x_k), x_k)$, $(f(x_{k-1}), x_{k-1})$ and $(f(x_{k-2}), x_{k-2})$ and evaluate this at 0.	$O(h^{1.839})$	

Given $\mathbf{u}_k$, $\mathbf{u}_{k+1} \leftarrow \mathbf{u}_k$ and then for each entry, $u_{k+1;i} \leftarrow \frac{1}{a_{i,i}}\left(v_i - A_{\text{off};i,\cdots}\mathbf{u}_{k+1}\right)$.

Given $\mathbf{u}_k$, calculate $u_{k+1;i}$ as above, but then set $u_{k+1;i} \leftarrow (1-\omega)u_{k;i} + \omega u_{k+1;i}$.

Given a function $f(x, y)$ and an approximation to a root (x_k, y_k), we can solve

$$\begin{pmatrix} \frac{\partial}{\partial x} f(x_k, y_k) & \frac{\partial}{\partial y} f(x_k, y_k) \\ \frac{\partial}{\partial x} g(x_k, y_k) & \frac{\partial}{\partial y} g(x_k, y_k) \end{pmatrix} \begin{pmatrix} \Delta x_k \\ \Delta y_k \end{pmatrix} = \begin{pmatrix} -f(x_k, y_k) \\ -g(x_k, y_k) \end{pmatrix}$$

and then let $x_{k+1} \leftarrow x_k + \Delta x_k$, $y_{k+1} \leftarrow y_k + \Delta y_k$.

$y_{k+1} = y_k + hf(t_k, y_k)$	$\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}(t_k, \mathbf{y}_k)$	$O(h)$
$s_0 = f(t_k, y_k)$ $s_1 = f(t_k + \mathrm{h}, y_k + hs_0)$ $y_{k+1} = y_k + h\frac{s_0 + s_1}{2}$	$\mathbf{s}_0 = \mathbf{f}(t_k, \mathbf{y}_k)$ $\mathbf{s}_1 = \mathbf{f}(t_k + \mathrm{h}, \mathbf{y}_k + h\mathbf{s}_0)$ $\mathbf{y}_{k+1} = \mathbf{y}_k + h\frac{\mathbf{s}_0 + \mathbf{s}_1}{2}$	$O(h^2)$
$s_0 = f(t_k, y_k)$ $s_1 = f(t_k + \frac{\mathrm{h}}{2}, y_k + \frac{\mathrm{h}}{2}s_0)$ $s_2 = f(t_k + \frac{\mathrm{h}}{2}, y_k + \frac{\mathrm{h}}{2}s_1)$ $s_3 = f(t_k + h, y_k + hs_2)$ $y_{k+1} = y_k + h\frac{s_0 + 2s_1 + 2s_2 + s_3}{6}$	$\mathbf{s}_0 = \mathbf{f}(t_k, \mathbf{y}_k)$ $\mathbf{s}_1 = \mathbf{f}(t_k + \frac{\mathrm{h}}{2}, \mathbf{y}_k + \frac{\mathrm{h}}{2}\mathbf{s}_0)$ $\mathbf{s}_2 = \mathbf{f}(t_k + \frac{\mathrm{h}}{2}, \mathbf{y}_k + \frac{\mathrm{h}}{2}\mathbf{s}_1)$ $\mathbf{s}_3 = \mathbf{f}(t_k + h, \mathbf{y}_k + h\mathbf{s}_2)$ $\mathbf{y}_{k+1} = \mathbf{y}_k + h\frac{\mathbf{s}_0 + 2\mathbf{s}_1 + 2\mathbf{s}_2 + \mathbf{s}_3}{6}$	$O(h^4)$

With n, calculate $y_1, \ldots, y_n$, with $2n$, calculate $z_1, \ldots, z_{2n}$, and use $|y_n - z_{2n}|$ appropriately to estimate the error of z_{2n}. If the error is small enough, extrapolate to get an even better approximation. The approximation of the error depends on the error of the method used.

Given a target error per unit step in time of ε_{abs}, ensure the error contributed to the total error when approximating y_{k+1} is less than $h\varepsilon_{abs}$. Do this by finding a better approximation z_{k+1}, and overestimating the error of y_{k+1} by $2|y_{k+1} - z_{k+1}|$ and calculating $a = \frac{h\varepsilon_{abs}}{2|y_{k+1} - z_{k+1}|}$. Based on the magnitude of a, either recalculate y_{k+1} or continue to approximate y_{k+2}, in either case using $0.9ah$.

Given $y^{(n)}(t) = f\left(t, y(t), y^{(1)}(t), \ldots, y^{(n-1)}(t)\right)$ with $y(t) = y_0, y^{(1)}(t) = y_0^{(1)}, \ldots, y(t)^{(n-1)} = y_0^{(n-1)}$, define

$$\mathbf{w}(t) = \begin{pmatrix} w_0(t) \\ w_1(t) \\ \vdots \\ w_{n-1}(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ y^{(1)}(t) \\ \vdots \\ y^{(n-1)}(t) \end{pmatrix}, \ \mathbf{w}_0 = \begin{pmatrix} y_0 \\ y_0^{(1)} \\ \vdots \\ y_0^{(n-1)} \end{pmatrix} \text{ and } \mathbf{w}^{(1)}(t) = \mathbf{f}(t, \mathbf{w}(t)) = \begin{pmatrix} w_0^{(1)}(t) \\ w_1^{(1)}(t) \\ \vdots \\ w_{n-2}^{(1)}(t) \\ w_{n-1}^{(1)}(t) \end{pmatrix} = \begin{pmatrix} w_1(t) \\ w_2(t) \\ \vdots \\ w_{n-1}(t) \\ f(t, \mathbf{w}(t)) \end{pmatrix}.$$

If $u^{(2)}(x) = f\left(x, u(x), u^{(1)}(x)\right)$, $u(a) = u_a$ and $u(b) = u_b$, solve two IVPs with

1. $u(a) = u_a$ and $u^{(1)}(a) = s_0 \leftarrow \frac{u_b - u_a}{b - a}$ with solution $u_0(x)$, and
2. $u(a) = u_a$ and $u^{(1)}(a) = s_1 \leftarrow \frac{2u_b - u_0(b) - u_a}{b - a}$ with solution $u_1(x)$;

and then continue with the initial slope

$$s_{k+1} \leftarrow s_k - \frac{\left(u_b - u_k(b)\right)\left(s_k - s_{k-1}\right)}{u_{k-1}(b) - u_k(b)}$$

with solution $u_{k+1}(x)$ with an approximation of u_b of $u_{k+1}(b)$.

Given $u^{(2)}(x) + \alpha_1(x)u^{(1)}(x) + \alpha_0(x)u(x) = g(x)$ and $x_k = a + kh$ and u_k approximates $u(x_k)$, we have

$$\left(2 - \alpha_1(x_k)h\right)u_{k-1} + \left(-4 + 2\alpha_0(x_k)h^2\right)u_k + \left(2 + \alpha_1(x_k)h\right)u_{k+1} = 2h^2 g(x_k).$$

If the ode has constant coefficients, the super-diagonal, diagonal and sub-diagonal entries are all

$$d_+ = 2 + \alpha_1 h, d_0 = -4 + 2\alpha_0 h^2, d_- = 2 - \alpha_1 h.$$

Apply this twice to get an approximation of the error of the better approximation.

$$u_{k,\ell+1} = u_{k,\ell} + \frac{\alpha \Delta t}{h^2}\left(u_{k-1,\ell} - 2u_{k,\ell} + u_{k+1,\ell}\right)$$

$$u_{k,\ell+1} = 2u_{k,\ell} - u_{k,\ell-1} + \left(\frac{c\Delta t}{h}\right)^2 \left(u_{k-1,\ell} - 2u_{k,\ell} + u_{k+1,\ell}\right)$$

$$u_{k,1} = u_{k,0} + \Delta t \dot{u}(x_k) + \frac{1}{2}\left(\frac{c\Delta t}{h}\right)^2 \left(u_{k-1,0} - 2u_{k,0} + u_{k+1,0}\right)$$

For 1, 2 and 3 dimensions, each point is the average of the 2, 4 or 6 points immediately surrounding it.

For an appropriate value of $\frac{1}{2} < \gamma < 1$ (generally the reciprocal of the golden ratio), calculate $c_1 = b - \gamma(b - a)$ and $c_2 = a + \gamma(b - a)$ and choose the appropriate sub-interval to continue the algorithm.

Given three approximations to a local minimum, we find that

$$\Delta x_k \leftarrow \frac{1}{2}\frac{\left(f(x_k) - f(x_{k-1})\right)(x_{k-1} - x_{k-2})(x_{k-2} - x_k)}{\left(f(x_k) - f(x_{k-1})\right)(x_{k-1} - x_{k-2}) + \left(f(x_{k-2}) - f(x_{k-1})\right)(x_k - x_{k-1})}$$

and $x_{k+1} = x_{k+1} \leftarrow \frac{x_k + x_{k-1}}{2} + \Delta x_k$.