ECE 204 *Numerical methods*
Sections 001/002
FINAL EXAMINATION

Douglas Wilhelm Harder dwharder@uwaterloo.ca EIT 4018 x37023

1. **The exam will be graded out of 51.**
2. No aids (including, 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 and clearly indicate that is where you are answering the question.
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 in the order of the course material.

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

```
1 01111111111 10100000…0
1 10000000001 01000000…0
```

showing the multiplication in binary, write the resulting representation in binary and determine the corresponding number in decimal.

2 [4] Demonstrate that if the absolute error for an approximation x_0 of a root r is $|x_0 - r|$, show that after one iteration of Newton's method, the absolute error is now proportional to $|x_0 - r|^2$ assuming that the second derivative is bounded between the approximation x_0 and the root r.

3 [4] Show that the error of the approximations

$$f^{(1)}(x) \approx \frac{f(x+h) - f(x-h)}{2h} \quad \text{and} \quad f^{(2)}(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

are both $O(h^2)$ by using the appropriate 3rd and 4th-order Taylor series, respectively.

4 [2] You have a sensor that is sending back a reasonably accurate power reading, and you know that there are no significant fluctuations or discontinuities in the power use. At any one time, you would like to know a reasonably accurate measurement of the total energy use up to that point in time. What algorithm would you use? Justify your answer.

5 [1] How would you find the best approximation of the vector $\begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$ as a linear combination of the two vectors $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$?

6 [3] Write down the system of equations that must be solved in order to find a better approximation of the simultaneous root of the following two algebraic equations

$$x^2 + xy + y^2 - 3 = 0$$
$$x^3 - 4xy - 5 = 0$$

assuming your initial guess is $x = 1.7$ and $y = 0.0$.

7 [3] For a given IVP, you have applied Euler's method with $n = 4$ and then again with $n = 8$ steps to get the approximations at the following points:

t_k	0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1
y_k	1		0.98		0.96		0.9		0.82
z_k	1	0.99	0.97	0.94	0.9	0.85	0.79	0.72	0.64

What is a reasonable approximation of the error of z_8, and what is the best estimate of $y(0.75)$ given the information in this table?

8 [3] Apply one step of Heun's method to this system to approximate $y(2)$ and $z(2)$:

$$y^{(1)}(t) = y(t) + tz(t)$$
$$z^{(1)}(t) = z(t) - tz(t)$$
$$y(1) = 4$$
$$z(1) = 5$$

9 [2] Write the following 3rd-order IVP as a system of 1st-order IVPs:

$$y^{(3)}(t) = ty^{(2)}(t) + t^2 y(t) + 1$$
$$y(1) = 4.5$$
$$y^{(1)}(1) = 6.7$$
$$y^{(2)}(1) = 8.9$$

10 [1] List all the restrictions on an IVP (i.e., the ODE and the initial conditions) that may allow us to use methods such as 4^{th}-order Runge Kutta but prevent us from using an adaptive technique like Euler-Heun, Runge-Kutta-Fehlberg or Dormand-Prince?

11 [3] Given the IVP,

$$y^{(1)}(t) = -20y(t)$$
$$y(0) = 4$$

approximate $y(0.1)$ using one step of Euler's method and one step of the backward Euler's method. Given your knowledge of the actual solution to this IVP, which is likely closer to the correct answer?

12 [4] Using the shooting method and two steps of Euler's method, approximate the value of $u(0.5)$ for the boundary-value problem

$$u^{(2)}(x) + u(x) = 1$$
$$u(0) = 2$$
$$u(1) = 3$$

13 [2] Explain, in your own words, why when applying the shooting method for a non-linear BVP $u^{(2)}(t) = f\left(t, u(t), u^{(1)}(t)\right)$ with $u(a) = u_a$ and $u(b) = u_b$, once we have found two slopes s_0 and s_1, why we apply the secant method to the problem

$$\hat{u}_b(s) - u_b$$

to find the appropriate initial slope s.

14 [2] Write down the systems of equations (using matrices and vectors) to find an approximation to the boundary value problem given by

$$u^{(2)}(t) + u(t) = 1$$
$$u(0) = 2$$
$$u(1) = 3$$

with $n = 4$.

15 [2] In class, it was claimed that the finite-difference method used to approximate a solution to a linear BVP is $O(h^2)$ where h is the width between the points at which the approximation occurs. Provide a heuristic justification for this result based on how the finite-difference equations were obtained.

16 [2] Given the state of a system that evolves over time according to the heat-conduction diffusing equation on the interval [0, 4] with initial and boundary conditions

$$u_0(x) = \begin{cases} 1-x & 0 \le x \le 0 \\ 0 & 1 < x \le 4 \end{cases}, \; u_a(t) = 1 \text{ and } u_b(t) = 0,$$

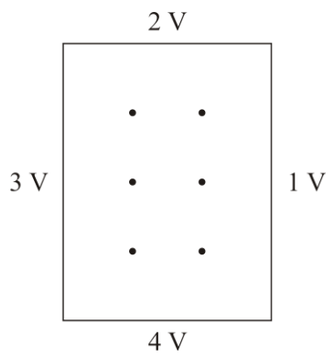
what is the state when $t = 0.1$ and $t = 0.2$ when using $h = 1$?

17 [2] Given the state of a system that evolves over time according to the wave equation on the interval [0, 4] with initial and boundary conditions

$$u_0(x) = \begin{cases} 1-x & 0 \le x \le 0 \\ 0 & 1 < x \le 4 \end{cases}, , \; u_a(t) = 1 \text{ and } u_b(t) = 0,$$

what is the state when $t = 0.1$ and $t = 0.2$ when using $h = 1$?

18 [2] You are given a rectangular wafer that has four constant voltages applied to each of the sides, as specified here:



Write down the system of linear equations, using matrices and vectors, that must be solved to estimate the potential at the six interior points.

19 [2] For the golden-ratio search, given that c_2 for one step must be the same as c_1 for the next step if the maximum is determined to be on the interval $[c_1, b]$, find the value of γ.

20 [3] Suppose we are finding the root of the interpolating line that connects two points (1000, 5) and (1001, 4) and then the root closest to 1001 of the three points (999, 5), (1000, 4) and (1001, 1). Explain why, in the first case, it is safe to find the root directly by simply solving the equation while in the second, it is better to shift the problem to (–2, 5), (–1, 4) and (0, 1) and then find the root as an offset to 1001? The quadratic formula has two roots: which root would you choose in this case?

21 [2] Suppose that you have a device where you can control an input voltage and make a reading, but that reading is noisy. You'd like to estimate the input voltage that maximizes the reading in question. Explain and justify the approach you would use, but you do not have to give explicit formulas.

Floating-point representations

±EEMNNN	$\pm \text{M.NNN} \times 10^{\text{EE}-49}$
seeeeeeeeeeebbbbbb…b	$(-1)^{\text{s}}\,1.\text{bbbbbb…b} \times 10^{\text{eeeeeeeeeee}-01111111111}$

where $01111111111_2 = 1023$.

Fixed-point theorem: Solving $x = f(x)$, choose x_0 and let $x_{k+1} \leftarrow f(x_k)$ for $k = 0, \ldots$.

Gaussian elimination with partial pivoting is the Gaussian elimination algorithm but always swapping appropriate rows so that the largest entry is in the row that will be used to eliminate that term in all subsequent rows.

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

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

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:

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

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

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

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

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

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

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

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

$$\int_a^b f(x)\,\mathrm{d}x = \frac{3}{8}\left(f(a) + 3\left(\sum_{k=1}^{\frac{n}{3}} f(a+(3k-2)h) \right) + 3\left(\sum_{k=1}^{\frac{n}{3}} f(a+(3k-1)h) \right) + 2\left(\sum_{k=1}^{\frac{n}{3}-1} f(a+3kh) \right) + f(b) \right) h - f^{(4)}(\xi) \frac{b-a}{80} h^4$$

Goal	Estimation
Estimate $y(t_n)$	$0.6\,y_n + 0.4\,y_{n-1} + 0.2\,y_{n-2} - 0.2\,y_{n-4}$
Estimate $y(t_n + h)$	$0.8\,y_n + 0.5\,y_{n-1} + 0.2\,y_{n-2} - 0.1\,y_{n-3} - 0.4\,y_{n-4}$
Estimate the rate of change of y over time	$\frac{0.2y_n + 0.1y_{n-1} - 0.1y_{n-3} - 0.2y_{n-4}}{h}$
Estimate the integral $\int_{t_n-4h}^{t_n} y(t)\,\mathrm{d}t$	$(4h)(0.2y_n + 0.2y_{n-1} + 0.2y_{n-2} + 0.2y_{n-3} + 0.2y_{n-4})$
Estimate the integral $\int_{t_n-h}^{t_n} y(t)\,\mathrm{d}t$	$h(0.5y_n + 0.35y_{n-1} + 0.2y_{n-2} + 0.05y_{n-3} - 0.1y_{n-4})$

Goal	Estimation
Estimate $y(t_n)$	$\frac{1}{35}\left(31y_n + 9y_{n-1} - 3y_{n-2} - 5y_{n-3} + 3y_{n-4}\right)$
Estimate $y(t_n + h)$	$1.8y_n - 0.8y_{n-2} - 0.6y_{n-3} + 0.6y_{n-4}$
Estimate the rate of change of y over time at time t_n	$\frac{54y_n - 13y_{n-1} - 40y_{n-2} - 27y_{n-3} + 26y_{n-4}}{70h}$
Estimate the acceleration of y over time at time t_n	$\frac{2y_n - y_{n-1} - 2y_{n-2} - y_{n-3} + 2y_{n-4}}{7h^2}$
Estimate the integral $\int_{t_n-4h}^{t_n} y(t)\,\mathrm{d}t$	$(4h)\frac{11y_n + 26y_{n-1} + 31y_{n-2} + 26y_{n-3} + 11y_{n-4}}{105}$
Estimate the integral $\int_{t_n-h}^{t_n} y(t)\,\mathrm{d}t$	$h\frac{230y_n + 137y_{n-1} + 64y_{n-2} + 11y_{n-3} - 22y_{n-4}}{420}$

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^ϕ)	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

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 $\lvert y_n - z_{2n}\rvert$ 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 ε_{abs}, ensure the error contributed to the total error when approximating y_{k+1} is less than $\frac{h}{t_f - t_0}\varepsilon_{abs}$. 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 $a = \dfrac{h\varepsilon_{abs}}{2|y_{k+1} - z_{k+1}|(t_f - t_0)}$. Based on the magnitude of a, either recalculate y_{k+1} or continue to approximate y_{k+2}, in either case using $0.9ah$.

For *stiff* ODEs, $y_{k+1} = y_k + hf(t_{k+1}, y_{k+1})$ or $\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}(t_{k+1}, \mathbf{y}_{k+1})$.

Given $y^{(n)}(t) = f\left(t, y(t), y^{(1)}(t), ..., y^{(n-1)}(t)\right)$ with $y(t) = y_0, y^{(1)}(t) = y_0^{(1)}, ..., y(t)^{(n-1)} = y_0^{(n-1)}$, 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 $u^{(2)}(x) + \alpha_1(x)u^{(1)}(x) + \alpha_0(x)u(x) = g(x)$, $u(a) = u_a$ and $u(b) = u_b$, solve two IVPs:

1. $u^{(2)}(x) + \alpha_1(x)u^{(1)}(x) + \alpha_0(x)u(x) = g(x)$ with $u(a) = u_a$ and $u^{(1)}(a) = 0$ with solution $u_g(x)$ and
2. $u^{(2)}(x) + \alpha_1(x)u^{(1)}(x) + \alpha_0(x)u(x) = 0$ with $u(a) = 0$ and $u^{(1)}(a) = 1$ with solution $u_0(x)$.

Add a scalar multiple c of the solution to IVP 2 to the solution of IVP 1 so that $u_g(b) + cu_0(b) = u_b$.

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 $u^{(1)}(a) = s_0$ with solution $u_0(x)$ and
2. $u(a) = u_a$ and $u^{(1)}(a) = s_1$ with solution $u_1(x)$

for appropriate values of the slopes. Define $\hat{u}_b(s)$ appropriately and then

$$s_{k+1} = \frac{s_{k-1}\left(\hat{u}_b(s_k) - u_b\right) - s_k\left(\hat{u}_b(s_{k-1}) - u_b\right)}{\hat{u}_b(s_k) - \hat{u}_b(s_{k-1})}$$

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 $\frac{1}{2} < \gamma < 1$, 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 maximum, we find that

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

and $x_{k+1} = x_k + \Delta x_{k+1}$.

This page is intentionally left blank.