Queens College, CUNY, Department of Computer Science

Numerical Methods

CSCI 361 / 761

Spring 2018

Instructor: Dr. Sateesh Mane

Midterm 2 Spring 2018

due Wednesday April 25, 2018, 11:59 pm

- **NOTE: It is the policy of the Computer Science Department to issue a failing grade to any student who either gives or receives help on any test.**
- This is an **open-book** test.
- **Any problem to which you give two or more (different) answers receives the grade of zero automatically.**
- This is a **take home exam.**
  Please submit your solution via email, as a file attachment, to `Sateesh.Mane@qc.cuny.edu`.
  The file name should have either of the formats:

  `StudentId_first_last_CS361_midterm2_Apr2018`

  `StudentId_first_last_CS761_midterm2_Apr2018`

  Acceptable file types are txt, doc/docx, pdf (also cpp, with text in comment blocks).
- **In all questions where you are asked to submit programming code, programs which display any of the following behaviors will receive an automatic F:**
  1. Programs which do not compile successfully (compiler warnings which are not fatal are excluded, e.g. use of deprecated features).
  2. Array out of bounds.
  3. Dereferencing of uninitialized variables (including null pointers).
  4. Operations which yield NAN or infinity, e.g. divide by zero, square root of negative number, etc. *Infinite loops.*
  5. Programs which do NOT implement the public interface stated in the question.
- **In addition, note the following:**
  1. Programs which compile and run successfully but have memory leaks will receive a poor grade (but not F).
  2. All debugging and/or output statements (e.g. `cout` or `printf`) will be commented out.
  3. Program performance will be tested solely on function return values and the values of output variable(s) in the function arguments.
  4. In other words, program performance will be tested solely via the public interface presented to the calling application. (I will write the calling application.)

General information

- **You are permitted to copy and use the code in the online lecture notes.**
- **Value of π to machine precision on any computer.**
  1. Some compilers support the constant `M_PI` for π, in which case you can write

     ```
     const double pi = M_PI;
     ```
  2. If your compiler does not support `M_PI`, the value of π can be computed via

     ```
     const double pi = 4.0*atan2(1.0,1.0);
     ```
- **64–bit computers**
  1. The questions in this exam do not involve problems of overflow.
  2. Solutions involving the writing of algorithms will not be judged if they work on a 64–bit instead of a 32–bit computer.

- **If you submit code, put all your code in ONE cpp file.
  Else include all the code in your main docx or pdf or txt file.
  DO NOT SUBMIT MULTIPLE CPP FILES.**

1 Question 1 no code

- **Finite differences with unequal steps.**
- Suppose the forward and backward steps are not equal.
- Suppose the forward step is h_1 and the backward step is h_2 and $h_2 \neq h_1$.
- **Write the Taylor series for $f(x+h_1)$ and $f(x-h_2)$ up to $O(f''''(x))$.**
- **Derive a numerical expression for the first derivative as follows:**

$$\frac{f(x+h_1) - f(x-h_2)}{h_1 + h_2} = f'(x) + \text{first two error terms}\,. \tag{1.1}$$

- **Show that if $h_2 \neq h_1$ the leading error term in eq. (1.1) is $O((h_1 - h_2)f''(x))$.**
- **Using only $f(x)$, $f(x+h_1)$ and $f(x-h_2)$, derive a finite difference approximation for the first derivative $f'(x)$, where the leading error term is $O(f'''(x))$.**
  1. **Write your answer up to and including the term is $O(f''''(x))$.**
  2. **Simplify your expression in the special case $h_1 = h$ and $h_2 = 2h$.**
- **Using only $f(x)$, $f(x+h_1)$ and $f(x-h_2)$, derive a finite difference approximation for the second derivative $f''(x)$. The leading error term is $O(f'''(x))$ if $h1 \neq h_2$.**
  1. **Write your answer up to and including the term is $O(f''''(x))$.**
  2. **Simplify your expression in the special case $h_1 = h$ and $h_2 = 2h$.**
- **Your expressions in the special case $h_1 = h$ and $h_2 = 2h$ will be different from the expressions derived in the lectures.**

- **This was intended as a gift question, which all students should have aced.**
- **It was verbatim from HW #5, except for the final step to set** $h_1 = h$ **and** $h_2 = 2h$**.**
- **Nevertheless there were several poor quality solutions, including wrong answers.**
- **The relevant Taylor series are:**

$$\begin{aligned}
f(x+h_1) &= f(x) + h_1 f'(x) + \frac{h_1^2}{2!}\,f''(x) + \frac{h_1^3}{3!}\,f'''(x) + \frac{h_1^4}{4!}\,f''''(x) + \cdots \\
f(x-h_2) &= f(x) - h_2 f'(x) + \frac{h_2^2}{2!}\,f''(x) - \frac{h_2^3}{3!}\,f'''(x) + \frac{h_2^4}{4!}\,f''''(x) + \cdots
\end{aligned}$$

- **Subtract to obtain:**

$$f(x+h_1) - f(x-h_2) = (h_1+h_2)f'(x) + \frac{h_1^2 - h_2^2}{2!}\,f''(x) + \frac{h_1^3 + h_2^3}{3!}\,f'''(x) + \frac{h_1^4 - h_2^4}{4!}\,f''''(x) + \cdots$$

- **Divide by** $h_1 + h_2$ **and rearrange terms (and simplify the coefficients) to obtain the finite difference:**

$$\begin{aligned}
f'(x) = {} & \frac{f(x+h_1) - f(x-h_2)}{h_1+h_2} - \frac{h_1 - h_2}{2!}\,f''(x) \\
& - \frac{h_1^2 - h_1h_2 + h_2^2}{3!}\,f'''(x) - \frac{(h_1-h_2)(h_1^2+h_2^2)}{4!}\,f''''(x) + \cdots
\end{aligned}$$

- **If** $h_1 \neq h_2$**, the leading error term is** $\frac{1}{2}(h_1 - h_2)f''(x)$**, which is** $O((h_1 - h_2)f''(x))$**.**
- **Improved expression for** $f'(x)$**.**

1. **Let** a **and** b **be arbitrary coefficients. Form the weighted difference**

$$\begin{aligned}
af(x+h_1) - bf(x-h_2) = {} & (a-b)f(x) + (ah_1 + bh_2)f'(x) + \frac{ah_1^2 - bh_2^2}{2!}\,f''(x) \\
& + \frac{ah_1^3 + bh_2^3}{3!}\,f'''(x) + \frac{ah_1^4 - bh_2^4}{4!}\,f''''(x) + \cdots
\end{aligned}$$

2. **Set** $ah_1^2 - bh_2^2 = 0$ **to make the term in** $f''(x)$ **vanish.**
3. **Choose** $a = h_2^2$ **and** $b = h_1^2$ **(simplest choice) to obtain**

$$\begin{aligned}
h_2^2 f(x+h_1) - h_1^2 f(x-h_2) = {} & (h_2^2 - h_1^2)f(x) + (h_1h_2^2 + h_1^2h_2)f'(x) \\
& + \frac{h_1^3h_2^2 + h_1^2h_2^3}{3!}\,f'''(x) + \frac{h_1^4h_2^2 - h_1^2h_2^4}{4!}\,f''''(x) + \cdots
\end{aligned}$$

4. **Divide by** $h_1h_2(h_1+h_2)$**, rearrange terms and simplify the coefficients to obtain the finite difference:**

$$f'(x) = \frac{h_2^2 f(x+h_1) - h_1^2 f(x-h_2) - (h_2^2 - h_1^2)f(x)}{h_1h_2(h_1+h_2)} - \frac{h_1h_2}{3!}\,f'''(x) - \frac{h_1h_2(h_1-h_2)}{4!}\,f''''(x) + \cdots$$

5. **Special case** $h_1 = h$ **and** $h_2 = 2h$**:**

$$f'(x) = \frac{4f(x+h_1) - f(x-h_2) - 3f(x)}{6h} - \frac{h}{3}\,f'''(x) + \frac{h^3}{12}\,f''''(x) + \cdots$$

- **Expression for $f''(x)$.**

  1. **Let c and d be arbitrary coefficients. Form the weighted difference**

  $$\begin{aligned} cf(x+h_1) - df(x-h_2) &= (c-d)f(x) + (ch_1 + dh_2)f'(x) + \frac{ch_1^2 - dh_2^2}{2!} f''(x) \\ &+ \frac{ch_1^3 + dh_2^3}{3!} f'''(x) + \frac{ch_1^4 - dh_2^4}{4!} f''''(x) + \cdots \end{aligned}$$

  2. **Set $ch_1 + dh_2 = 0$ to make the term in $f''(x)$ vanish.**
  3. **Choose $c = h_2$ and $d = -h_1$ (simplest choice) to obtain**

  $$\begin{aligned} h_2 f(x+h_1) + h_1 f(x-h_2) &= (h_2 + h_1)f(x) + \frac{h_1^2 h_2 + h_1 h_2^2}{2!} f''(x) \\ &+ \frac{h_1^3 h_2 - h_1 h_2^3}{3!} f'''(x) + \frac{h_1^4 h_2 + h_1 h_2^4}{4!} f''''(x) + \cdots \end{aligned}$$

  4. **Divide by $\frac{1}{2}h_1h_2(h_1+h_2)$, rearrange terms and simplify the coefficients to obtain the finite difference:**

  $$\begin{aligned} f''(x) = 2\,&\frac{h_2 f(x+h_1) + h_1 f(x-h_2) - (h_1+h_2)f(x)}{h_1h_2(h_1+h_2)} \\ &- \frac{h_1 - h_2}{3} f'''(x) - \frac{h_1^2 - h_1h_2 + h_2^2}{12} f''''(x) + \cdots \end{aligned}$$

  5. **Special case $h_1 = h$ and $h_2 = 2h$:**

  $$f''(x) = \frac{2f(x+h_1) + f(x-h_2) - 3f(x)}{3h^2} + \frac{h}{3} f'''(x) - \frac{h^2}{4} f''''(x) + \cdots$$

2 Question 2 show code, possible partial credit

- Instead of a unit circle described by the equation $x^2 + y^2 = 1$, it is also possible to have curves described by the equation
$$|x|^\alpha + |y|^\beta = 1\,. \tag{2.1}$$
Here $\alpha, \beta > 0$ and $\alpha \neq \beta$ in general. The curve with $\alpha = 1.7$ and $\beta = 3.2$ is plotted in Fig. 1.

- You are required to numerically calculate the area enclosed by the curve $|x|^{1.7} + |y|^{3.2} = 1$.
  1. *Let's not panic and be stupid and think this is a two-dimensional integral.*
  2. Observe that the curve is symmetric around both the x and y axes.
  3. Hence we calculate the area in the first quadrant $x, y \geq 0$ and multiply the result by 4.
  4. **The Cartesian plane is divided into four quadrants.**
  5. See Fig. 1.
     (a) First quadrant: top right $x \geq 0$, $y \geq 0$.
     (b) Second quadrant: top left $x \leq 0$, $y \geq 0$.
     (c) Third quadrant: bottom left $x \leq 0$, $y \leq 0$.
     (d) Fourth quadrant: bottom right $x \geq 0$, $y \leq 0$.

- The area A_1 under the curve in the first quadrant is given by the following integral:
$$A_1 = \int_0^1 y(x)\,dx\,. \tag{2.2}$$

- **Write down the equation for $y(x)$ in the first quadrant.**

- **Prove that the integral in eq. (2.2) is a proper integral.**

- **Compute the value of A_1 using**
(a) midpoint rule, (b) trapezoid rule, (c) Simpson's rule.
  1. Use $n = 1000$ subintervals for all three calculations.
  2. **You are permitted to use the code displayed in the online lecture notes.**

- The total area A is given by $A = 4A_1$.

- **State your computed values for the total area A to four decimal places.**

- **Denote your values by A_{M}, A_{T}, A_{S}.**

- *Do not worry if the three results you obtain are not equal to four decimal places.*

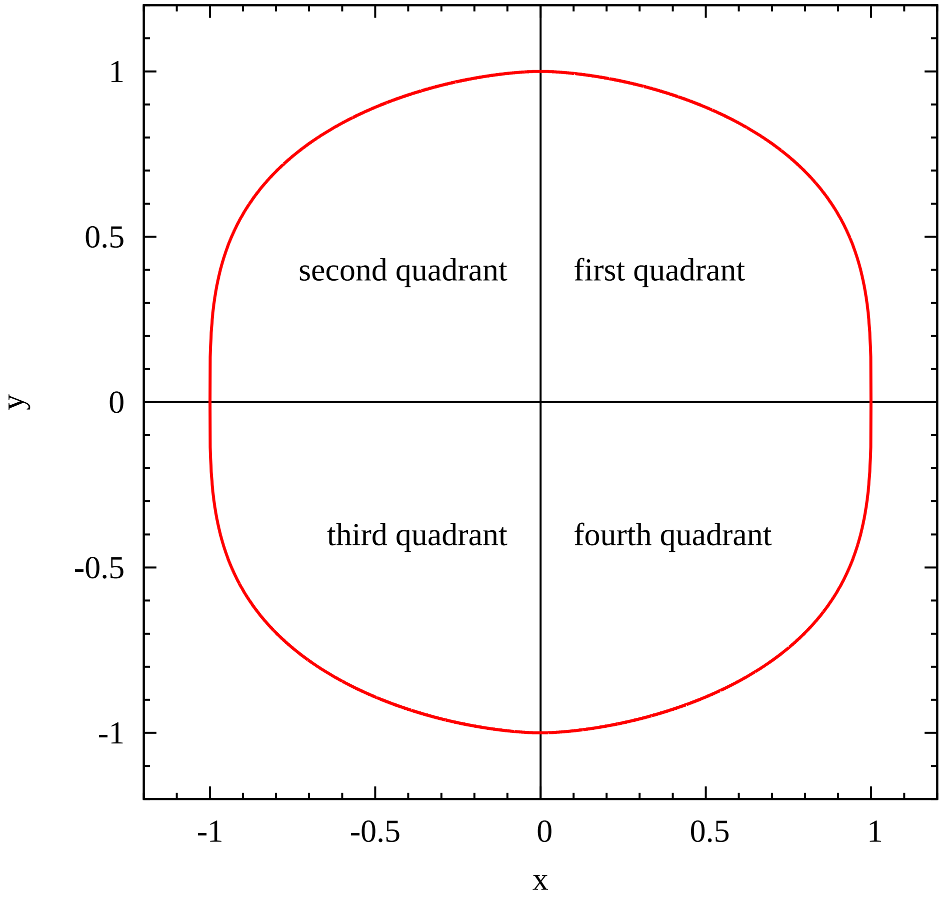

Figure 1: Graph of the curve $|x|^{1.7} + |y|^{3.2} = 1$ in Question 2.

- **In the first quadrant, $|x| = x$ and $|y| = y$ by definition.**
- **Hence the equation of the curve is (with $0 \le x \le 1$ and $0 \le y \le 1$)**

$$x^{1.7} + y^{3.2} = 1\,.$$

- **Express y as a function of x:**

$$y = \left(1 - x^{1.7}\right)^{1/3.2}\,.$$

- **Then**

$$A_1 = \int_0^1 y(x)\,dx = \int_0^1 \left(1 - x^{1.7}\right)^{1/3.2}\,dx\,.$$

- **This is a proper integral because both limits of integration are finite and the integrand is bounded and continuous and has no $0/0$ problems throughout the domain of integration.**
- **To avoid problems at $x = 1$ (in case the numerical calculations produce a value of x slightly greater than 1), it is safer to compute $y = |1 - x^{1.7}|^{1/3.2}$:**

$$\texttt{y} = \texttt{pow}(\texttt{std}::\texttt{abs}(1.0 - \texttt{pow}(\texttt{x},\texttt{alpha})), 1.0/\texttt{beta})\,.$$

- **The algorithms for the midpoint, trapezoid and Simpson's rules are given in the lectures.**
- **The question asked for the total area $A = 4A_1$, but many students also printed the value of A_1.**

	Midpoint	Trapezoid	Simpson
A_1	**0.8307**	**0.8307**	**0.8307**
A	**3.3228**	**3.3226**	**3.3227**

- **The rounded values of A_1 are equal to 4 decimal places, but small differences appear in the values of A.**
- **Student values might differ slightly. Experience has shown that the results vary with the computer and compiler.**

3 Question 3 no code

3.1 Matrix 1

- Let α, β and θ be real numbers and R be the following matrix:

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} . \tag{3.1}$$

- **Calculate the trace and determinant of R.**

- **Calculate the inverse matrix R^{-1}.**

- **Solve the following equation for x and y.**
  Express your answer as a function of α, β and θ.

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} . \tag{3.2}$$

- **Prove the following identity:**

$$x^2 + y^2 = \alpha^2 + \beta^2 . \tag{3.3}$$

- Relevant definitions and identities:

$$\begin{aligned} \cos^2\theta + \sin^2\theta = 1 . \\ \cos\theta &= \frac{e^{i\theta} + e^{-i\theta}}{2} . \\ \sin\theta &= \frac{e^{i\theta} - e^{-i\theta}}{2i} . \end{aligned}$$

3.2 Matrix 2

- Let u, v and ω be real numbers and B be the following matrix:

$$B = \begin{pmatrix} \cosh\omega & \sinh\omega \\ \sinh\omega & \cosh\omega \end{pmatrix} . \tag{3.4}$$

- **Calculate the trace and determinant of B.**

- **Calculate the inverse matrix B^{-1}.**

- **Solve the following equation for x and t.**
  Express your answer as a function of u, v and ω.

$$\begin{pmatrix} \cosh\omega & \sinh\omega \\ \sinh\omega & \cosh\omega \end{pmatrix} \begin{pmatrix} t \\ x \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix} . \tag{3.5}$$

- **Prove the following identity:**

$$t^2 - x^2 = u^2 - v^2 . \tag{3.6}$$

- Relevant definitions and identities:

$$\cosh^2\omega - \sinh^2\omega = 1 .$$

$$\cosh\omega = \frac{e^{\omega} + e^{-\omega}}{2} .$$

$$\sinh\omega = \frac{e^{\omega} - e^{-\omega}}{2} .$$

3.3 For your information (no calculations required):

- The matrix R is a **rotation matrix**.
- The values of x and y (also α and β) are coordinates in a plane.
- The value of $x^2 + y^2$ is the squared length of the vector with coordinates (x, y).
- The value of $\alpha^2 + \beta^2$ is the squared length of the vector with coordinates (α, β).
- A rotation does not change the length of a vector.

- The matrix B is a **boost matrix**, from Einstein's Special Theory of Relativity.
- The values of x and t (also u and v) are coordinates in space-time.
- The value of $t^2 - x^2$ is the squared **invariant separation** of the **space-time event** with space-time coordinates (t, x).
- The value of $u^2 - v^2$ is the squared invariant separation of the space-time event with space-time coordinates (u, v).
- A boost does not change the invariant separation.
- Note that the value of $t^2 - x^2$ can be **negative** even though it is called a "squared" invariant separation.

  1. If $t^2 - x^2 > 0$ it is called **timelike separation.**
  2. If $t^2 - x^2 < 0$ it is called **spacelike separation.**
  3. If $t^2 - x^2 = 0$ it is called **lightlike (or null) separation.**

- **Solution for rotation.**
- **Trace**$(R) = 2\cos\theta$**.**
- **Determinant**$(R) = \cos^2\theta + \sin^2\theta = 1$ **for all values of** θ**.**
- **Matrix inverse**

$$R^{-1} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} .$$

- **Solution**

$$\begin{pmatrix} x \\ y \end{pmatrix} = R^{-1} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha\cos\theta + \beta\sin\theta \\ -\alpha\sin\theta + \beta\cos\theta \end{pmatrix} .$$

- **Proof of identity:**

$$\begin{aligned} x^2 + y^2 &= (\alpha\cos\theta + \beta\sin\theta)^2 + (-\alpha\sin\theta + \beta\cos\theta)^2 \\ &= \alpha^2\cos^2\theta + 2\alpha\beta\cos\theta\sin\theta + \beta^2\sin^2\theta \\ &\quad + \alpha^2\sin^2\theta - 2\alpha\beta\cos\theta\sin\theta + \beta^2\cos^2\theta \\ &= \alpha^2(\cos^2\theta + \sin^2\theta) + \beta^2(\cos^2\theta + \sin^2\theta) \\ &= \alpha^2 + \beta^2 . \end{aligned}$$

- **Solution for boost.**
- **Trace**$(B) = 2\cosh\omega$**.**
- **Determinant**$(B) = \cosh^2\omega - \sinh^2\omega = 1$ **for all values of** ω**.**
- **Matrix inverse**

$$B^{-1} = \begin{pmatrix} \cosh\omega & -\sinh\omega \\ -\sinh\omega & \cosh\omega \end{pmatrix} .$$

- **Solution**

$$\begin{pmatrix} t \\ x \end{pmatrix} = B^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cosh\omega & -\sinh\omega \\ -\sinh\omega & \cosh\omega \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u\cosh\omega - v\sinh\omega \\ -u\sinh\omega + v\cosh\omega \end{pmatrix} .$$

- **Proof of identity:**

$$\begin{aligned} t^2 - x^2 &= (u\cosh\omega - v\sinh\omega)^2 - (-u\sinh\omega + v\cosh\omega)^2 \\ &= u^2\cosh^2\omega - 2uv\cosh\omega\sinh\omega + v^2\sinh^2\omega \\ &\quad - u\sinh^2\omega + 2uv\cosh\omega\sinh\omega - v^2\cosh^2\omega \\ &= u^2(\cosh^2\omega - \sinh^2\omega) - v^2(\cosh^2\omega - \sinh^2\omega) \\ &= u^2 - v^2 . \end{aligned}$$

- **Some of you need to learn how to write mathematical proofs properly.**
- **I accepted your answers, but there is a concept of correct exposition, or good writing.**

4 Question 4 do by hand, code is optional for last step

- **Solve the following three sets of equations for x_1, x_2 and x_3:**

$$\begin{aligned} -x_1 + 2x_2 + x_3 &= 3\,, \\ 2x_1 + 5x_2 + 7x_3 &= -6\,, \\ -3x_1 + 2x_2 + 4x_3 &= -1\,. \end{aligned} \tag{4.1}$$

$$\begin{aligned} -x_1 + 2x_2 + x_3 &= 3\,, \\ 2x_1 + 5x_2 + 7x_3 &= 3\,, \\ -3x_1 + 2x_2 + 4x_3 &= -1\,. \end{aligned} \tag{4.2}$$

$$\begin{aligned} -x_1 + 2x_2 + x_3 &= 3\,, \\ 2x_1 + 5x_2 + 7x_3 &= 12\,, \\ -3x_1 + 2x_2 + 4x_3 &= -1\,. \end{aligned} \tag{4.3}$$

- *Let's not panic. It is only necessary to perform the LU decomposition once.*
- **Write down the matrix A associated with eqs. (4.1), (4.2) and (4.3).**
- **Write out the steps in the LU decomposition of A.**
- **Display the final matrix in LU form.**
- **Also write down the final value of the array of the swap indices S.**
- **Also write down the total number of swaps performed.**
- **Calculate the determinant of the matrix A.**
- **Solve eqs. (4.1), (4.2) and (4.3) for x_1, x_2 and x_3.**

1. **To answer this part of the question, you are permitted to use the functions displayed in the online lectures, for LU decomposition and backsubstitution.**
2. **You do not need to display all the backsubstitution steps.**
3. **Just state the answer.**

- **The matrix A associated with the above equations is**

$$A = \begin{pmatrix} -1 & 2 & 1 \\ 2 & 5 & 7 \\ -3 & 2 & 4 \end{pmatrix} .$$

- **Scaled pivots: $\hat{a}_1 = 2$, $\hat{a}_2 = 7$, $\hat{a}_3 = 4$, hence**

$$\frac{|a_{11}|}{\hat{a}_1} = \frac{1}{2} = 0.5 \,, \qquad \frac{|a_{21}|}{\hat{a}_2} = \frac{2}{7} \simeq 0.2857 \,, \qquad \frac{|a_{31}|}{\hat{a}_3} = \frac{3}{4} = 0.75 \,.$$

- **Hence swap rows 1 and 3. The swap index is $(3, 2, 1)$.**

$$A_1 = \begin{pmatrix} -3 & 2 & 4 \\ 2 & 5 & 7 \\ -1 & 2 & 1 \end{pmatrix} .$$

- **Subtract $-2/3$ times row 1 from row 2 and $1/3$ times row 1 from row 3:**

$$A_1 = \begin{pmatrix} -3 & 2 & 4 \\ 0 & \frac{19}{3} & \frac{29}{3} \\ 0 & \frac{4}{3} & -\frac{1}{3} \end{pmatrix} .$$

- **Fill in the multipliers. It is acceptable if you combine the two steps.**

$$A_2 = \begin{pmatrix} -3 & 2 & 4 \\ -\frac{2}{3} & \frac{19}{3} & \frac{29}{3} \\ \frac{1}{3} & \frac{4}{3} & -\frac{1}{3} \end{pmatrix} .$$

- **Calculate new scaled pivots: $\hat{a}'_2 = \frac{29}{3}$, $\hat{a}'_3 = \frac{4}{3}$, hence**

$$\frac{|a'_{22}|}{\hat{a}'_2} = \frac{19}{29} \simeq 0.655 \,, \qquad \frac{|a'_{32}|}{\hat{a}'_3} = 1 \,.$$

- **Hence swap rows 2 and 3. The swap index is $(3, 1, 2)$.**

$$A_3 = \begin{pmatrix} -3 & 2 & 4 \\ \frac{1}{3} & \frac{4}{3} & -\frac{1}{3} \\ -\frac{2}{3} & \frac{19}{3} & \frac{29}{3} \end{pmatrix} .$$

- **Subtract $19/4$ times row 2 from row 3:**

$$A_3 = \begin{pmatrix} -3 & 2 & 4 \\ \frac{1}{3} & \frac{4}{3} & -\frac{1}{3} \\ -\frac{2}{3} & 0 & \frac{45}{4} \end{pmatrix} .$$

- **Fill in the multiplier. It is acceptable if you combine the two steps.**

$$A_3 = \begin{pmatrix} -3 & 2 & 4 \\ \frac{1}{3} & \frac{4}{3} & -\frac{1}{3} \\ -\frac{2}{3} & \frac{19}{4} & \frac{45}{4} \end{pmatrix} .$$

- **This is the final matrix in LU form. It is acceptable to write decimals.**
- **The final swap index is $(3, 1, 2)$. Total two swaps.**
- **The determinant is**

$$\det(A) = (-1)^2 \det(U) = (-3)(\frac{4}{3})(\frac{45}{4}) = -45 .$$

- **To solve $A\boldsymbol{x} = \boldsymbol{b}$ using LU decomposition we requirw two backsubstitution steps. First we permute the entries of $\boldsymbol{b}$ using the swap index to obtain $LU\boldsymbol{x} = \boldsymbol{b}_{\text{swap}}$. We introduce a temporary column vector $\boldsymbol{y}$ and perform backsubstitution to solve $L\boldsymbol{y} = \boldsymbol{b}_{\text{swap}}$. Then we perform a second backsubstitution step to solve $U\boldsymbol{x} = \boldsymbol{y}$.**
- **I bypassed that and solved the equations using the code supplied in the lectures.**
- **First set of equations:**

$$LU\,\boldsymbol{x} = \boldsymbol{b}_{\text{swap}} = \begin{pmatrix} -1 \\ 3 \\ -6 \end{pmatrix} , \qquad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \\ -2 \end{pmatrix} .$$

- **Second set of equations:**

$$LU\,\boldsymbol{x} = \boldsymbol{b}_{\text{swap}} = \begin{pmatrix} -1 \\ 3 \\ 3 \end{pmatrix} , \qquad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.2 \\ 2.2 \\ -1.2 \end{pmatrix} .$$

- **Second set of equations:**

$$LU\,\boldsymbol{x} = \boldsymbol{b}_{\text{swap}} = \begin{pmatrix} -1 \\ 3 \\ 12 \end{pmatrix} , \qquad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1.4 \\ 2.4 \\ -0.4 \end{pmatrix} .$$

- **Some students solved without swapping rows.**
- **Some students employed a different implementation of LU decomposition.**
- **I accepted all such answers.**

5 Question 5 no code

- You are given the following equations to solve for x_1, x_2 and x_3:

$$\begin{aligned} -x_1 + 2x_2 + x_3 &= 3\,, \\ 2x_1 + 5x_2 + 7x_3 &= \boldsymbol{r_2}\,, \\ \boldsymbol{a_{31}}x_1 + 2x_2 + 4x_3 &= -1\,. \end{aligned} \tag{5.1}$$

- **Find the value of a_{31} such that the LU decomposition encounters a zero pivot.**
- Denote that value of a_{31} by α_{31}.
- Set $a_{31} = \alpha_{31}$. **Then find the value of r_2 such that the equations are consistent.**
- **Note: Do NOT attempt to solve the resulting equations.**

- **The matrix for the above equations is**

$$A = \begin{pmatrix} -1 & 2 & 1 \\ 2 & 5 & 7 \\ a_{31} & 2 & 4 \end{pmatrix} .$$

- **We do not know the value of a_{31} so do not swap rows.**

- **Note that $a_{11} = -1$ so that is not a zero pivot, hence no problem there.**

- **Subtract multiples of row 1 from rows 2 and 3 to eliminate the entries in the (1,2) and (1,3) positions.**

- **Actually add 2 times row 1 to row 2 and a_{31} times row 1 to row 3:**

$$A_1 = \begin{pmatrix} -1 & 2 & 1 \\ 0 & 9 & 9 \\ 0 & 2+2a_{31} & 4+a_{31} \end{pmatrix} .$$

- **"LU decomposition encounters a zero pivot" means it is impossible to avoid an entry of zero on the main diagonal (of the upper triangular matrix), no matter how we try to swap, etc. Then the LU decomposition cannot avoid a zero pivot.**

- **We have to find the value of a_{31} so that the LU decomposition cannot avoid a zero pivot. There are now various ways to proceed.**

- **Rows not linearly independent.**

  1. **One method is to note that we set the value of a_{31} so that the second and third row are proportional.**
  2. **Then the rows will not be linearly independent.**
  3. **The condition for this is that $(A_1)_{23} = (A_1)_{33}$, hence $2 + 2a_{31} = 4 + a_{31}$.**
  4. **The solution is $\boldsymbol{a_{31} = 2}$.**

- **Gaussian elimination.**

  1. **Subtract $(2 + 2a_{31})/9$ times row 2 from row 3 to obtain**

$$A_2 = \begin{pmatrix} -1 & 2 & 1 \\ 0 & 9 & 9 \\ 0 & 0 & 4+a_{31}-(2+2a_{31}) \end{pmatrix} .$$

  2. **Now demand that the final entry must be zero $(A_2)_{33} = 0$: this is a zero pivot.**
  3. **The condition for this is that $4 + a_{31} - (2 + 2a_{31}) = 0$.**
  4. **The solution is $\boldsymbol{a_{31} = 2}$.**

- **Determinant $= 0$.**

  1. **We do not need to perform any subtraction of rows at all.**
  2. **LU decomposition will encounter a zero pivot if the determinant of A is zero.**

  $$\det(A) = -1(20-14) - 2(8-7a_{31}) + 1(4-5a_{31}) = -18 + 9a_{31}\,.$$

  3. **The condition for a zero pivot is $\det(A) = 0$.**
  4. **The solution is $a_{31} = 2$.**

- **Hence the solution is $a_{31} = 2$.**

- **Then the equations, after eliminating x_1 from the second and third equations, are**

  $$\begin{aligned} -x_1 + 2x_2 + x_3 &= 3\,, \\ 9x_2 + 9x_3 &= r_2 + 6\,, \\ 6x_2 + 6x_3 &= -1 + 6 = 5\,. \end{aligned}$$

  We subtract $6/9$ times the second equation from the third equation.

- **For the equations to be consistent, the right hand side must equal zero when we do this:**

  $$\frac{2}{3}(r_2 + 6) - 5 = 0\,.$$

  The solution is

  $$r_2 = \frac{3}{2}\,.$$

6 Question 6 no code

- You are given the following set of equations for four unknowns x_1, x_2, x_3, x_4:

$$\begin{aligned} 4x_1 - x_2 &= 2\,, \\ 2x_1 + 4x_2 - x_3 &= 6\,, \\ 2x_2 + 4x_3 - x_4 &= 12\,, \\ 2x_3 + 4x_4 &= 40\,. \end{aligned} \tag{6.1}$$

- **Write the set of equations in eq. (6.1) in tridiagonal matrix form as follows:**

$$\begin{pmatrix} a_1 & c_1 & 0 & 0 \\ b_2 & a_2 & c_2 & 0 \\ 0 & b_3 & a_3 & c_3 \\ 0 & 0 & b_4 & a_4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 6 \\ 12 \\ 40 \end{pmatrix} . \tag{6.2}$$

- If you have done your work correctly, the values of a_i, b_i and c_i will not depend on i.

- **Prove that the tridiagonal matrix in eq. (6.2) is strongly diagonally dominant.**

- **Solve eq. (6.2) for the unknowns x_1, x_2, x_3, x_4.**

  1. **Display the steps in your calculation.**
  2. If you do your work correctly, there should be about eight steps, four forward and four backsubstitution steps backward.

- **(Not a question.) For your information:**

- A matrix where all the elements are equal down the diagonals is called a **Toeplitz matrix.**

- Toeplitz matrices can be stored using less storage (only one number per diagonal).

- Technically, a Toeplitz matrix need not be square.

- However, the definitions are simpler for square matrices.

- There are efficient numerical algorithms to process Toeplitz matrices.

- **The equations in tridiagonal matrix form are**

$$\begin{pmatrix} 4 & -1 & 0 & 0 \\ 2 & 4 & -1 & 0 \\ 0 & 2 & 4 & -1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 6 \\ 12 \\ 40 \end{pmatrix} .$$

- **To prove strong diagonal dominance we must show $|a_i| > |c_i|$ in row 1, $|a_i| > |b_i| + |c_i|$ in rows 2 and 3 and $|a_i| > |b_i|$ in row 4.**

  1. **First row $a_i = 4$, $c_i = -1$ and $4 > |-1| = 1$.**
  2. **Second and third rows $a_i = 4$, $b_i = 2$, $c_i = -1$ and $4 > |-1| + 2 = 3$.**
  3. **Fourth row $a_i = 4$, $b_i = 2$ and $4 > 2$.**
  4. **This proves the strong diagonal dominance.**

- **Use the first equation to eliminate x_1:**

$$x_1 = \frac{1}{2} + \frac{x_2}{4} .$$

- **Substitute in second equation and collect terms**

$$x_2 = \frac{10}{9} + \frac{2x_3}{9} .$$

- **Substitute in third equation and collect terms**

$$x_3 = \frac{11}{5} + \frac{9x_4}{40} .$$

- **Substitute in fourth equation and collect terms and solve for x_4:**

$$x_4 = \frac{40}{5} = 8 .$$

- **Backsubstitution solution for x_3:**

$$x_3 = \frac{11}{5} + \frac{9}{5} = 4 .$$

- **Backsubstitution solution for x_2:**

$$x_2 = \frac{10}{9} + \frac{8}{9} = 2 .$$

- **Backsubstitution solution for x_1:**

$$x_1 = \frac{1}{2} + \frac{2}{4} = 1 .$$

- **Any guesses how I chose the right hand side column vector?**

7 Question 7 no code

- Let μ be a real number and let T be the following tridiagonal matrix:

$$T = \begin{pmatrix} 2+\mu^2 & \mu-1 & 0 & 0 & 0 \\ \mu-1 & 2+\mu^2 & \mu-1 & 0 & 0 \\ 0 & \mu-1 & 2+\mu^2 & \mu-1 & 0 \\ 0 & 0 & \mu-1 & 2+\mu^2 & \mu-1 \\ 0 & 0 & 0 & \mu-1 & 2+\mu^2 \end{pmatrix} \tag{7.1}$$

- **You will need to consider the cases $\mu \geq 1$ and $\mu < 1$ separately.**
  1. If $\mu \geq 1$ then $|\mu - 1| = \mu - 1$.
  2. If $\mu < 1$ then $|\mu - 1| = -(\mu - 1) = 1 - \mu$.
  3. **Remember to pay attention to the special cases in the first and last rows.**

- **Prove that the matrix T in eq. (7.1) is strongly diagonally dominant for all $\mu \geq 1$.**

- **Find all values of μ for which the matrix T in eq. (7.1) is:**
  1. Strongly diagonally dominant.
  2. Weakly **but not strongly** diagonally dominant.
  3. **Not diagonally dominant.**

- For your information, the matrix T in eq. (7.1) is a symmetric tridiagonal Toeplitz matrix.

- **Solve the following matrix equation for the unknowns x_1, x_2, x_3, x_4, x_5, for $\mu = 1$:**

$$T_{(\mu=1)} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} . \tag{7.2}$$

- *Don't laugh.*

- **Case $\mu \geq 1$.**
  1. **Then $|\mu - 1| = \mu - 1$.**
  2. **First row:**
  $$|a_i| - |c_i| = 2 + \mu^2 - (\mu - 1) = \mu^2 - \mu + 3 = (\mu - \tfrac{1}{2})^2 + \frac{11}{4} \geq \frac{11}{4} > 0\,.$$
  3. **Last row:**
  $$|a_i| - |b_i| = 2 + \mu^2 - (\mu - 1) = \mu^2 - \mu + 3 = (\mu - \tfrac{1}{2})^2 + \frac{11}{4} \geq \frac{11}{4} > 0\,.$$
  4. **Second, third and fourth rows:**
  $$|a_i| - |b_i| - |c_i| = 2 + \mu^2 - 2(\mu - 1) = \mu^2 - 2\mu + 4 = (\mu - 1)^2 + 3 \geq 3 > 0\,.$$
  5. **Hence the matrix is strongly diagonally dominant for all $\mu \geq 1$.**
- **Case $\mu < 1$.**
  1. **Then $|\mu - 1| = 1 - \mu$.**
  2. **First row:**
  $$|a_i| - |c_i| = 2 + \mu^2 - (1 - \mu) = \mu^2 + \mu + 1 = (\mu + \tfrac{1}{2})^2 + \frac{3}{4} \geq \frac{3}{4} > 0\,.$$
  3. **Last row:**
  $$|a_i| - |b_i| = 2 + \mu^2 - (1 - \mu) = \mu^2 + \mu + 1 = (\mu + \tfrac{1}{2})^2 + \frac{3}{4} \geq \frac{3}{4} > 0\,.$$
  4. **Second, third and fourth rows:**
  $$|a_i| - |b_i| - |c_i| = 2 + \mu^2 - 2(1 - \mu) = \mu^2 + 2\mu = \mu(\mu + 2)\,.$$
  5. **This is not always positive.**
  6. **It is zero for $\mu = 0$ and $\mu = -2$ (matrix is weakly diagonally dominant).**
  7. **It is negative for $-2 < \mu < 0$ (matrix is not diagonally dominant).**
  8. **It is positive for $\mu < -2$ and $0 < \mu < 1$ (matrix is strongly diagonally dominant).**
- **For $\mu = 1$, the matrix is diagonal, equal to three times the unit matrix.**
- **The solution is obvious:**
$$\begin{aligned} x_1 &= \frac{1}{3}\,, \\ x_2 &= \frac{2}{3}\,, \\ x_3 &= \frac{3}{3} = 1\,, \\ x_4 &= \frac{4}{3}\,, \\ x_5 &= \frac{5}{3}\,. \end{aligned}$$

8 Question 8 submit code, possible partial credit

- See eq. (4.1). Let $(\gamma_1, \gamma_2, \gamma_3)$ denote the solutions for (x_1, x_2, x_3) of the equations below:

$$\begin{aligned} -x_1 + 2x_2 + x_3 &= 3\,, \\ 2x_1 + 5x_2 + 7x_3 &= -6\,, \\ -3x_1 + 2x_2 + 4x_3 &= -1\,. \end{aligned} \tag{8.1}$$

- **For each case $i = 1, 2, 3$, determine if the integral in eq. (8.2) is proper.**

$$I(\gamma_i) = \int_{-1}^{1} x^{2\gamma_i} \cos(x^2)\, dx\,. \tag{8.2}$$

- If the integral in eq. (8.2) is proper, **compute its value numerically using the extended trapezoid rule with $n = 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024$ subintervals.**
- Denote the results by $R(j, 0)$, where $n = 2^j$, so $j = 0, 1, \dots, 10$.
- Then use first order Romberg integration. Define $R(j, 1)$ as follows:

$$R(j,1) = \frac{4R(j,0) - R(j-1,0)}{3}\,, \qquad j = 1, 2, \dots, 10\,. \tag{8.3}$$

- Then use second order Romberg integration. Define $R(j, 2)$ as follows:

$$R(j,2) = \frac{16R(j,1) - R(j-1,1)}{15}\,, \qquad j = 2, 3, \dots, 10\,. \tag{8.4}$$

- **Fill the table below with values to six decimal places.**

n	j	$R(j,0)$	$R(j,1)$	$R(j,2)$
1	0	6 d.p.		
2	1	6 d.p.	6 d.p.	
4	2	6 d.p.	6 d.p.	6 d.p.
8	3	6 d.p.	6 d.p.	6 d.p.
16	4	6 d.p.	6 d.p.	6 d.p.
32	5	6 d.p.	6 d.p.	6 d.p.
64	6	6 d.p.	6 d.p.	6 d.p.
128	7	6 d.p.	6 d.p.	6 d.p.
256	8	6 d.p.	6 d.p.	6 d.p.
512	9	6 d.p.	6 d.p.	6 d.p.
1024	10	6 d.p.	6 d.p.	6 d.p.

- **State the smallest value of j for which $R(j, 2)$ converges to six decimal places.**
- **State the value of the integral $I(\gamma_i)$ to six decimal places.**

- **The solutions are $\gamma_1 = -1$, $\gamma_2 = 2$ and $\gamma_3 = -2$.**

- **The corresponding integrals are**

$$I(\gamma_1) = \int_{-1}^{1} \frac{1}{x^2} \cos(x^2)\,dx\,, \qquad I(\gamma_2) = \int_{-1}^{1} x^4 \cos(x^2)\,dx\,, \qquad I(\gamma_3) = \int_{-1}^{1} \frac{1}{x^4} \cos(x^2)\,dx\,.$$

- **The cases γ_1 and γ_3 yield improper integrals because the integrand diverges at $x = 0$, which is within the domain of integration.**

- **The case γ_2 yields a proper integral because the domain of integration is finite and the integrand is bounded and continuous has no $0/0$ problems throughout the domain of integration.**

- **The table of values using extended trapezoid and Romberg integration is**

n	j	$R(j,0)$	$R(j,1)$	$R(j,2)$
1	0	**1.080600**		
2	1	**0.540302**	**0.360202**	
4	2	**0.330708**	**0.260843**	**0.254220**
8	3	**0.301131**	**0.291272**	**0.293301**
16	4	**0.296447**	**0.294886**	**0.295126**
32	5	**0.295453**	**0.295122**	**0.295138**
64	6	**0.295216**	**0.295137**	**0.295138**
128	7	**0.295158**	**0.295138**	**0.295138**
256	8	**0.295143**	**0.295138**	**0.295138**
512	9	**0.295139**	**0.295138**	**0.295138**
1024	10	**0.295138**	**0.295138**	**0.295138**

- **The smallest value of j for which $R(j,2)$ converges to six decimal places is $j = 5$ (hence $n = 32$).**

- **The value of the integral $I(\gamma_2)$ to six decimal places is 0.295138.**

- **Student values might differ slightly. Experience has shown that the results vary with the computer and compiler.**

9 Question 9 submit code

9.1 Complete elliptic integral of the second kind

- There are three kinds of complete elliptic integrals (and three incomplete elliptic integrals).
- The **complete elliptic integral of the second kind** is given by the following integral:

$$E(x) = \int_0^1 \frac{\sqrt{1-x^2t^2}}{\sqrt{1-t^2}}\,dt \qquad (0 \le x \le 1)\,. \tag{9.1}$$

- For your information, for an ellipse with semi–major axis a and semi–minor axis b and eccentricity $e_{\rm ell} = \sqrt{1-b^2/a^2}$, the circumference of the ellipse, say c, is given by $c = 4aE(e_{\rm ell})$.
- The function $E(x)$ can be expressed as power series as follows:

$$E(x) = \frac{\pi}{2}\left[1 - \left(\frac{1}{2}\right)^2\frac{x^2}{1} - \left(\frac{1\cdot3}{2\cdot4}\right)^2\frac{x^4}{3} - \left(\frac{1\cdot3\cdot5}{2\cdot4\cdot6}\right)^2\frac{x^6}{5} - \cdots\right]. \tag{9.2}$$

- We shall compute the value of $E(x)$ in two ways, by using the integral in eq. (9.1) and by approximating the infinite series in eq. (9.2) by a finite sum.
- The integral in eq. (9.1) is improper because the integrand diverges at $t = 1$ (unless $x = 1$, in which case we obtain $0/0$, which a computer also cannot evaluate).
  1. However, we can compute the integral in eq. (9.1) using the midpoint rule.
  2. **Let $I_n(x)$ denote the value of the integral in eq. (9.1) using the midpoint rule with n subintervals.**
- Let us approximate the infinite series in eq. (9.2) by a finite sum.
  1. **Let $S_m(x)$ denote the value of the series in eq. (9.2), when the series is terminates at the term in x^{2m}.**

$$S_m(x) = \frac{\pi}{2}\left[1 - \left(\frac{1}{2}\right)^2\frac{x^2}{1} - \left(\frac{1\cdot3}{2\cdot4}\right)^2\frac{x^4}{3} - \cdots - \left(\frac{1\cdot3\cdot5\ldots(2m-1)}{2\cdot4\cdot6\ldots(2m)}\right)^2\frac{x^{2m}}{2m-1}\right]. \tag{9.3}$$

  2. **Formulate an efficient algorithm to compute the sum for $S_m(x)$ in eq. (9.3).**
- **(Optional/bonus) Prove using the sum in eq. (9.2) that $E(x)$ is a decreasing function of x, i.e. if $x_2 > x_1$ then $E(x_2) < E(x_1)$.**
- **(Not a question.) For your information:**
  1. Observe from the series in eq. (9.2) that $E(0) = \frac{1}{2}\pi \simeq 1.57$ for $x = 0$.
  2. Observe from the integral in eq. (9.1) that $E(1) = 1$ for $x = 1$.
  3. Hence $E(x)$ decreases monotonically from $\frac{1}{2}\pi$ to 1 as x increases from 0 to 1.

9.2 Computation

- **Set $x = 0$. We know $E(0) = \frac{1}{2}\pi$ is the exact value.**
- Obviously the sum in eq. (9.2) will yield $S_m(0) = \frac{1}{2}\pi$ for all values of m.
- **Compute the value of $I_n(0)$ and fill in the table below.**

n	$I_n(0)$	$\lvert I_n(0) - \frac{1}{2}\pi \rvert$
10^1	6 d.p.	
10^2	6 d.p.	
10^3	6 d.p.	
10^4	6 d.p.	
10^5	6 d.p.	
10^6	6 d.p.	

- **Set $x = 1$. We know $E(1) = 1$ is the exact value.**
- **Compute the values of $I_n(1)$ and $S_m(1)$ and fill in the table below.**

n	$I_n(1)$	$\lvert I_n(1) - 1 \rvert$	m	$S_m(1)$	$\lvert S_m(1) - 1 \rvert$
10^1	6 d.p.		10^1	6 d.p.	
10^2	6 d.p.		10^2	6 d.p.	
10^3	6 d.p.		10^3	6 d.p.	
10^4	6 d.p.		10^4	6 d.p.	
10^5	6 d.p.		10^5	6 d.p.	
10^6	6 d.p.		10^6	6 d.p.	

- Hence observe that the sum $S_m(x)$ converges more rapidly than $I_n(x)$ for small values $x \simeq 0$ and the integral $I_n(x)$ converges more rapidly than $S_m(x)$ for large values $x \simeq 1$.
- **Fill the table below for $I_n(x)$ and $S_m(x)$. Use $m = n = 1000$.**
  Write your answers to 4 decimal places.

x	$I_n(x)$	$S_m(x)$
0	4 d.p.	4 d.p.
0.1	4 d.p.	4 d.p.
0.2	4 d.p.	4 d.p.
0.3	4 d.p.	4 d.p.
0.4	4 d.p.	4 d.p.
0.5	4 d.p.	4 d.p.
0.6	4 d.p.	4 d.p.
0.7	4 d.p.	4 d.p.
0.8	4 d.p.	4 d.p.
0.9	4 d.p.	4 d.p.
1.0	4 d.p.	4 d.p.

- **(Optional) Plot a graph of $S_m(x)$ for $x = 0, 0.1, \ldots, 1.0$.**

9.3 Taylor series: remainder term

- **Let us estimate the remainder term of the Taylor series.**
- If we sum eq. (9.2) to m terms, i.e. use $S_m(x)$, the remainder term $R_m(x)$ is

$$R_m(x) = -\left(\frac{1\cdot 3\cdots(2m+1)}{2\cdot 4\cdots(2m+2)}\right)^2 \frac{x^{2m+2}}{2m+1} - \left(\frac{1\cdot 3\cdots(2m+1)(2m+3)}{2\cdot 4\cdots(2m+2)(2m+4)}\right)^2 \frac{x^{2m+4}}{2m+3} - \cdots \tag{9.4}$$

- All the terms are negative (i.e. same sign) and the coefficients decrease in magnitude.
- Hence an upper bound on the magnitude of the remainder term is $|R_m(x)| \le U_m(x)$, where

$$\begin{aligned} U_m(x) &= \left(\frac{1\cdot 3\cdots(2m+1)}{2\cdot 4\cdots(2m+2)}\right)^2 \frac{x^{2m+2}}{2m+1}\left(1+x^2+x^4+\cdots\right) \\ &= \frac{1}{2m+1}\left(\frac{1\cdot 3\cdots(2m+1)}{2\cdot 4\cdots(2m+2)}\right)^2 \frac{x^{2m+2}}{1-x^2}\,. \end{aligned} \tag{9.5}$$

- Observe that $U_m(x) = 0$ for $x = 0$ and $U_m(x) \to \infty$ for $x \to 1$.
- We saw that the sum $S_m(x)$ is reliable for small $x \simeq 0$ and not so accurate for large $x \simeq 1$.
- **Set $x = 0.5$ and calculate the value of $U_m(0.5)$ and fill the table below.**
  State your results in the 'scientific notation' form $a.bc \times 10^{-d}$.

m	$U_m(0.5)$
2	
3	
4	
5	
6	

- *If you have done your work correctly, then to obtain accuracy for $E(0.5)$ to 4 decimal places it should be sufficient to use $m = 5$ or 6.*

9.4 Root finding

- **Find x such that $E(x) = 1.48$.**
- **Use bisection and use $I_n(x)$ with $n = 1000$.**
  1. **Use numbers from the previous table.**
  2. **State the shortest initial bracket which encloses the root.**
  3. Denote the initial iterates by x_0 for x_{low} and x_1 for x_{high}
  4. Iterate until the value of the root converges to 4 decimal places.

i	x_i	$I_n(x) - 1.48$
0	x_0	5 d.p.
1	x_1	5 d.p.
$\vdots$	$\vdots$	
	converged 4 d.p.	

- **Use the secant method and use $S_m(x)$ with $m = 5$.**
  1. **For the two initial iterates, use the same two values for x_0 and x_1 which were employed above for bisection.**
  2. Iterate until the value of the root converges to 4 decimal places.

i	x_i	$S_m(x) - 1.48$
0	x_0	5 d.p.
1	x_1	5 d.p.
$\vdots$	$\vdots$	
	converged 4 d.p.	

- *If you have done your work correctly, the values for the root, using $I_n(x)$ and $S_m(x)$, will* **not** *be equal.*
- This illustrates the difficulty when trying to compute roots (or function values in general) using numerical algorithms.
- The answers we get depend on external parameters such as m and n.
- **Explain why the answer computed using $S_5(x)$ is reliable to 4 decimal places.**
- For $I_n(x)$, we need to go up to about $n \simeq 5 \times 10^6$ (five million) for the value of the root to agree with the calculation using $S_5(x)$.

See next page.

- From eq. (9.1), for $x = 0.5$, the value of the integral is

$$E(0.5) = \int_0^1 \frac{\sqrt{1-\frac{1}{4}t^2}}{\sqrt{1-t^2}}\,dt\,. \tag{9.6}$$

- The integrand diverges to ∞ for $t \to 1$.
- **Calculate the value of $\sqrt{(1-(t^2/4))/(1-t^2)}$ and fill in the table below.**

n	$h = 1/n$	$t = 1 - h$	$\sqrt{(1-(t^2/4))/(1-t^2)}$	$h \times \sqrt{(1-(t^2/4))/(1-t^2)}$
10^3	10^{-3}	0.999	4 d.p.	4 d.p.
10^4	10^{-4}	0.9999	4 d.p.	4 d.p.
10^5	10^{-5}	0.99999	4 d.p.	4 d.p.
10^6	10^{-6}	0.999999	4 d.p.	4 d.p.

- *If you have done your work correctly, the final number in the last column should be about* 0.0006.
- Because of the divergence of the integrand as $t \to 1$, it requires $n > 10^6$ subintervals to compute the integral in eq. (9.1) to 4 decimal places, for $x = 0.5$.
- There are of course other ways of computing the integral.
- We can transform the integral to obtain

$$E(x) = \int_0^{\pi/2} \sqrt{1 - x^2 \sin^2\theta}\,d\theta\,. \tag{9.7}$$

- In this form the integrand is bounded for all values of θ. It is a proper integral for $|x| \le 1$.

- **Nest the sum in eq. (9.3) as "1−(nested sum of terms)" as follows:**

$$S_m(x) = \frac{\pi}{2}\left[1 - \left(\frac{1}{2}\right)^2 x^2 \left(\frac{1}{1} + \left(\frac{3}{4}\right)^2 x^2 \left(\frac{1}{3} + \cdots + \left(\frac{2m-1}{2m}\right)^2 x^2 \,\frac{1}{2m-1}\right)\cdots\right)\right)\right].$$

- **Proof that $E(x)$ is a decreasing function of x using the sum in eq. (9.2).**

  1. **First we apply the ratio test for the sum in eq. (9.2).**
  2. **The ratio of successive terms is**

$$\begin{aligned}\frac{|a_{m+1}|}{|a_m|} &= \left(\frac{1\cdot3\cdot5\ldots(2m+1)}{2\cdot4\cdot6\ldots(2m+2)}\right)^2 \frac{x^{2(m+1)}}{2m+1}\left[\left(\frac{1\cdot3\cdot5\ldots(2m-1)}{2\cdot4\cdot6\ldots(2m)}\right)^2 \frac{x^{2m}}{2m-1}\right]^{-1} \\ &= \frac{(2m+1)^2}{(2m+2)^2}\,\frac{2m-1}{2m+1}\,x^2\,.\end{aligned}$$

  3. **We take the limit as $m\to\infty$, for a fixed value of x. Then**

$$\lim_{m\to\infty}\frac{|a_{m+1}|}{|a_m|} = \lim_{m\to\infty}\frac{(2m+1)^2}{(2m+2)^2}\,\frac{2m-1}{2m+1}\,x^2 = x^2\,.$$

  4. **The series converges if the limit is < 1.**
  5. **Hence for $x^2 < 1$, or $-1 < x < 1$, the sum in eq. (9.2) converges and does not run away to ±∞.**
  6. **The first term in the sum in eq. (9.2) is positive and all the rest are negative. The value of x^{2m} increases as x increases from 0 to 1, for all integers $m \geq 1$. Hence for $0 \leq x < 1$, the sum in eq. (9.2) decreases as the value of x increases**

$$E(x_2) < E(x_1) \qquad\qquad (x_2 > x_1)\,.$$

  7. **Exactly at $x = 1$ the series is at its radius of convergence and it is more difficult to justify that the sum in eq. (9.2) converges. That was a weak point in the formulation of the question. I should have restricted the values of x to the interval $0 \leq x < 1$.**

- **Set $x = 0$. Compute the value of $I_n(0)$ and fill in the table below.**

n	$I_n(0)$	$\lvert I_n(0) - \frac{1}{2}\pi\rvert$
10^1	**1.435881**	**0.134916**
10^2	**1.528034**	**0.042762**
10^3	**1.557271**	**0.013526**
10^4	**1.566519**	**0.004278**
10^5	**1.569444**	**0.001353**
10^6	**1.570369**	**0.000428**

- **Notice the rate of convergence is terrible.**

- **The numerical error decreases by a factor of 10 when the value of n increases by a factor of 100, so the rate of convergence is $O(1/\sqrt{n})$.**

- **The fact that the integrand is singular at $t = 1$ reduces the rate of convergence.**

- **Set $x = 1$. Compute the values of $I_n(1)$ and $S_m(1)$ and fill in the table below.**

n	$I_n(1)$	$\lvert I_n(1) - 1\rvert$	m	$S_m(1)$	$\lvert S_m(1) - 1\rvert$
10^1	**1**	**0**	10^1	**1.024086**	$2.408\,10^{-2}$
10^2	**1**	**0**	10^2	**1.002491**	$2.49\,10^{-3}$
10^3	**1**	**0**	10^3	**1.000259**	$2.5\,10^{-4}$
10^4	**1**	**0**	10^4	**1.000025**	$2.5\,10^{-5}$
10^5	**1**	**0**	10^5	**1.000002**	$2.5\,10^{-6}$
10^6	**1**	**0**	10^6	**1.000000**	$2.5\,10^{-7}$

- **The value of $I_n(1)$ is exact, and the value of $S_m(1)$ converges as $O(1/n)$.**

- **Fill the table below for $I_n(x)$ and $S_m(x)$. Use $m = n = 1000$.**
  Write your answers to 4 decimal places.

x	$I_n(x)$	$S_m(x)$
0	**1.5573**	**1.5708**
0.1	**1.5534**	**1.5669**
0.2	**1.5417**	**1.5550**
0.3	**1.5219**	**1.5348**
0.4	**1.4935**	**1.5059**
0.5	**1.4557**	**1.4675**
0.6	**1.4073**	**1.4181**
0.7	**1.3460**	**1.3557**
0.8	**1.2682**	**1.2763**
0.9	**1.1658**	**1.1717**
1.0	**1.0000**	**1.0002**

- **(Optional) Plot a graph of $S_m(x)$ for $x = 0, 0.1, \ldots, 1.0$.**

  1. **A graph of the complete elliptic integral of the second kind $E_2(x)$ is plotted in Fig. 2, for $0 \le x \le 1$.**
  2. **The solid curve is obtained by summing the series $S_m(x)$ with $m = 1000$.**
  3. **The dashed curve is obtained from the numerical integral $I_n(x)$ with $n = 1000$.**

- **Set $x = 0.5$ and calculate the value of $U_m(0.5)$ and fill the table below.**

m	$U_m(0.5)$
2	$4.07\,10^{-4}$
3	$5.56\,10^{-5}$
4	$8.76\,10^{-6}$
5	$1.51\,10^{-6}$
6	$2.75\,10^{-7}$

- Bisection: from the table, the shortest initial bracket which encloses the root is $(0.4, 0, 5)$.
- Students who did not use the initial bracket $(0.4, 0, 5)$ lost points.

i	x_i	$I_n(x) - 1.48$
0	**0.4**	**0.01355**
1	**0.5**	**-0.02425**
2	**0.45**	-0.00411
3	**0.425**	0.00502
4	**0.4375**	0.00053
5	**0.4438**	-0.00177
6	**0.4406**	-0.00062
7	**0.4391**	-0.00004
8	**0.4383**	0.00025
9	**0.4387**	0.00010
10	**0.4389**	0.00003
11	**0.4390**	-0.000005
12	**0.4389**	0.00001

- Student results which differed slightly from the above were accepted. Experience has shown that the results vary with the computer and compiler.
- Use the secant method and use $S_m(x)$ with $m = 5$.

i	x_i	$S_m(x) - 1.48$
0	**0.4**	**0.02594**
1	**0.5**	**-0.01254**
2	**0.4674**	**0.00112**
3	**0.4701**	**0.00004**
4	**0.4702**	$-1.5\,10^{-7}$
5	**0.4702**	$2.0\,10^{-11}$

- Student results which differed slightly from the above were accepted. Experience has shown that the results vary with the computer and compiler.
- The answer computed using $S_5(x)$ is reliable to 4 decimal places because we have a reliable upper bound on the accuracy of the computed value of $S_m(x)$.
  1. We calculated the upper bound $U_m(0.5) = 1.51\,10^{-6}$ for $m = 5$ and $x = 0.5$.
  2. We know from the structure of the sum in eq. (9.2) that the accuracy of $S_m(x)$ is even better for smaller values of x, i.e. $0 \le x < 0.5$.
  3. Since the secant iteration converged to a value of x whose magnitude is less than 0.5, the accuracy of the computed value of $S_m(x)$ is better than $1.51\,10^{-6}$.
  4. Hence value of the root, calculated via iteration using $S_m(x)$, is reliable to 4 decimal places.

- **Calculate the value of $\sqrt{(1-(t^2/4))/(1-t^2)}$ and fill in the table below.**

n	$h = 1/n$	$t = 1 - h$	$\sqrt{(1-(t^2/4))/(1-t^2)}$	$h \times \sqrt{(1-(t^2/4))/(1-t^2)}$
10^3	10^{-3}	0.999	**19.3762**	**0.0194**
10^4	10^{-4}	0.9999	**61.2408**	**0.0061**
10^5	10^{-5}	0.99999	**193.6500**	**0.0019**
10^6	10^{-6}	0.999999	**612.3730**	**0.0006**

- **Essentially, this is a demonstration that for $x \simeq 0.5$, the iteration using the integral $I_n(x)$ is not as reliable as the iteration using the sum $S_m(x)$.**

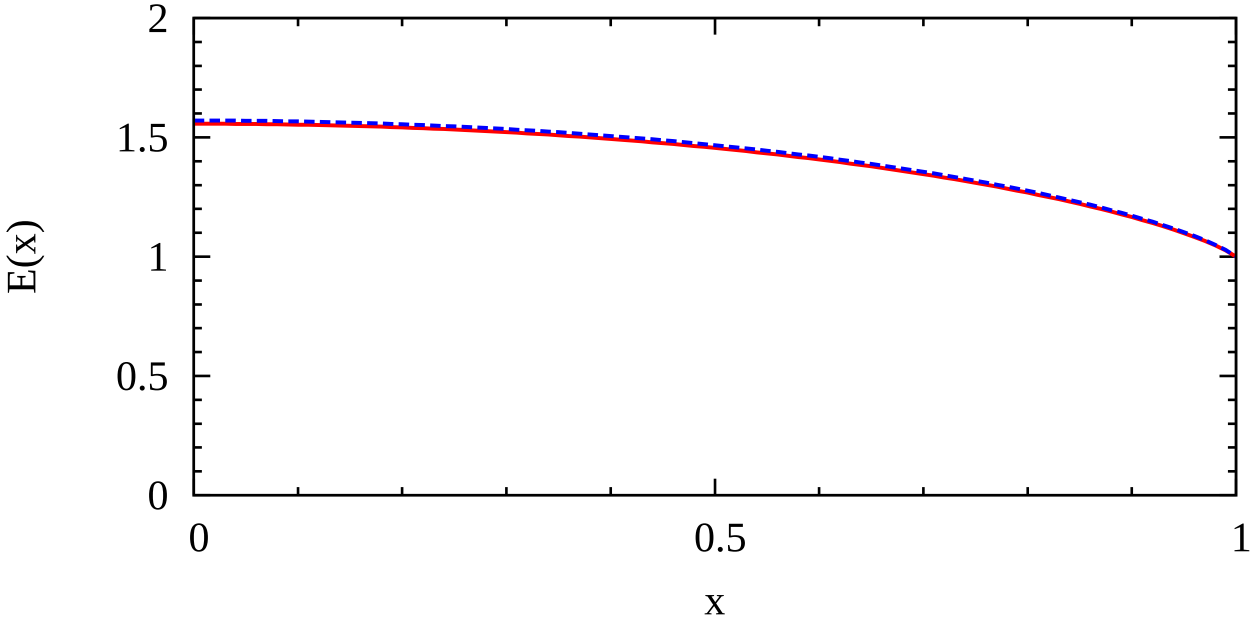

Figure 2: Graph of the complete elliptic integral of the second kind $E_2(x)$ for $0 \leq x \leq 1$. The solid curve is obtained by summing a series with 1000 terms The dashed curve is obtained by numerical integration using 1000 subintervals.

10 Question 10 no code

- **This is a question about a question in Midterm 1.**
- **You do not have to compute numbers or submit code for this question.**
- The Bessel function $J_0(x)$ can be computed by evaluating the following integral:

$$J_0(x) = \frac{1}{\pi}\int_0^{\pi} \cos(x\sin\theta)\,d\theta\,. \tag{10.1}$$

- **Write an expression (sum of terms) to compute the integral in eq. (10.1) using the trapezoid rule with n subintervals.**
- In midterm 1, I approximated the integral in eq. (10.1) using the following sum (here I use n instead of N as I did in Midterm 1):

$$S(x) = \frac{1}{n}\sum_{j=0}^{n-1} \cos\Bigl(x\,\sin\frac{j\pi}{n}\Bigr)\,. \tag{10.2}$$

- **State the difference between the trapezoid rule formula and my sum in eq. (10.2).**
- **Explain why the difference does not matter and the sum in eq. (10.2) yields the same result as the trapezoid rule.**

- **Using $a = 0$ and $b = \pi$ and n subintervals, the stepsize is $h = (b - a)/n = \pi/n$.**
- **The points at which the function is evaluated are $\theta_j = jh = j\pi/n$.**
- **The trapezoid rule with n subintervals yields (say T_n)**

$$\begin{aligned}
T_n &= \frac{1}{\pi}\frac{\pi}{n}\left[\frac{\cos(0) + \cos(x\sin\pi)}{2} + \sum_{j=1}^{n-1}\cos\Bigl(x\,\sin\frac{j\pi}{n}\Bigr)\right] \\
&= \frac{1}{n}\left[\frac{\cos(0) + \cos(0)}{2} + \sum_{j=1}^{n-1}\cos\Bigl(x\,\sin\frac{j\pi}{n}\Bigr)\right] \\
&= \frac{1}{n}\left[1 + \sum_{j=1}^{n-1}\cos\Bigl(x\,\sin\frac{j\pi}{n}\Bigr)\right].
\end{aligned}$$

- **The difference between the trapezoid rule formula and my sum in eq. (10.2) is that the sum in eq. (10.2) extends from $j = 0$ to $n - 1$, whereas the sum in the trapezoid rule extends from $j = 0$ to $n - 1$, and there is a separate term to sum the function values at the end points.**
- **The difference does not matter because the $j = 0$ term in eq. (10.2) has the value $\cos(0) = 1$ and this is the same as the average of the endpoint values in the trapezoid rule (see expression for T_n above).**
- **Hence both expressions for the numerical integral are equal.**

11 Question 11 (bonus question) no code

- Let I_n be the unit matrix of size $n \times n$.
- **Calculate the trace of I_n.**
- Let M be an arbitrary square matrix of size $n \times n$.
- Define a matrix D as follows:
$$D = \frac{\mathrm{trace}(M)}{n}\, I_n \,. \tag{11.1}$$
- **Prove that D is diagonal. Also prove the following:**
$$\mathrm{trace}(D) = \mathrm{trace}(M) \,. \tag{11.2}$$
- Define a matrix A as follows:
$$A = \tfrac{1}{2}(M - M^T) \,. \tag{11.3}$$
- **Prove that A is antisymmetric.**
- Define a matrix T as follows:
$$T = \tfrac{1}{2}(M + M^T) - \frac{\mathrm{trace}(M)}{n}\, I_n \,. \tag{11.4}$$
- **Prove that T is symmetric and traceless.**
- **Prove the following:**
$$M = D + A + T \,. \tag{11.5}$$
- For your information, the matrices D, A and T are the **irreducible components** of the matrix M.
- Every square matrix can be decomposed this way.

- **The matrix D is diagonal by construction since it is proportional to the identity matrix.**
- **The trace of D is**

$$\mathrm{trace}(D) = \frac{\mathrm{trace}(M)}{n}\,\mathrm{trace}(I_n) = \frac{\mathrm{trace}(M)}{n}\,n = \mathrm{trace}(M)\,.$$

- **The matrix A is antisymmetric because its transpose A^T is the negative of A:**

$$A^T = \tfrac{1}{2}(M - M^T)^T = \tfrac{1}{2}(M^T - M) = -A\,.$$

- **The matrix T is symmetric because its transpose T^T is equal to T itself:**

$$T^T = \tfrac{1}{2}(M + M^T)^T - \frac{\textbf{trace}(M)}{n}\,I_n^T = \tfrac{1}{2}(M^T + M) - \frac{\textbf{trace}(M)}{n}\,I_n = T\,.$$

- **The matrix T is traceless because**

$$\begin{aligned}\mathrm{trace}(T) &= \tfrac{1}{2}\mathrm{trace}(M + M^T) - \frac{\textbf{trace}(M)}{n}\,\mathrm{trace}(I_n)\\ &= \tfrac{1}{2}(\mathrm{trace}(M) + \mathrm{trace}(M)) - \frac{\textbf{trace}(M)}{n}\,n\\ &= \mathrm{trace}(M) - \textbf{trace}(M)\\ &= 0\,.\end{aligned}$$

- **Proof of decomposition:**

$$\begin{aligned}D + A + T &= \frac{\textbf{trace}(M)}{n}\,(I_n) + \tfrac{1}{2}(M - M^T) + \tfrac{1}{2}(M + M^T) - \frac{\textbf{trace}(M)}{n}\,(I_n)\\ &= \tfrac{1}{2}(M - M^T) + \tfrac{1}{2}(M + M^T)\\ &= M\,.\end{aligned}$$

Statistics

- **There are 33 registered students, and the breakdown of grades is as follows.**

A+	**7**
A	**9**
A-	**2**
B+	**4**
B	**7**
B-	**3**
C	**1**

- **Two students obtained perfect scores (A+ for every question).**
- **A histogram of the grades is plotted in Fig. 3**
- **The contents are informative.**
- **Approximately half the students scored A or A+ (16 out of 33).**
- **It is slightly higher than in Fall 2017 (4 A+ and 4 A, total 8 out of 20 students).**
- **They form a clear spike of A and A+ grades.**
- **The grade distribution of the rest of the students forms a bell-shaped curve.**
- **I do not “grade on a curve” and the evidence indicates I do not need to.**
- **The students form the curve by themselves.**

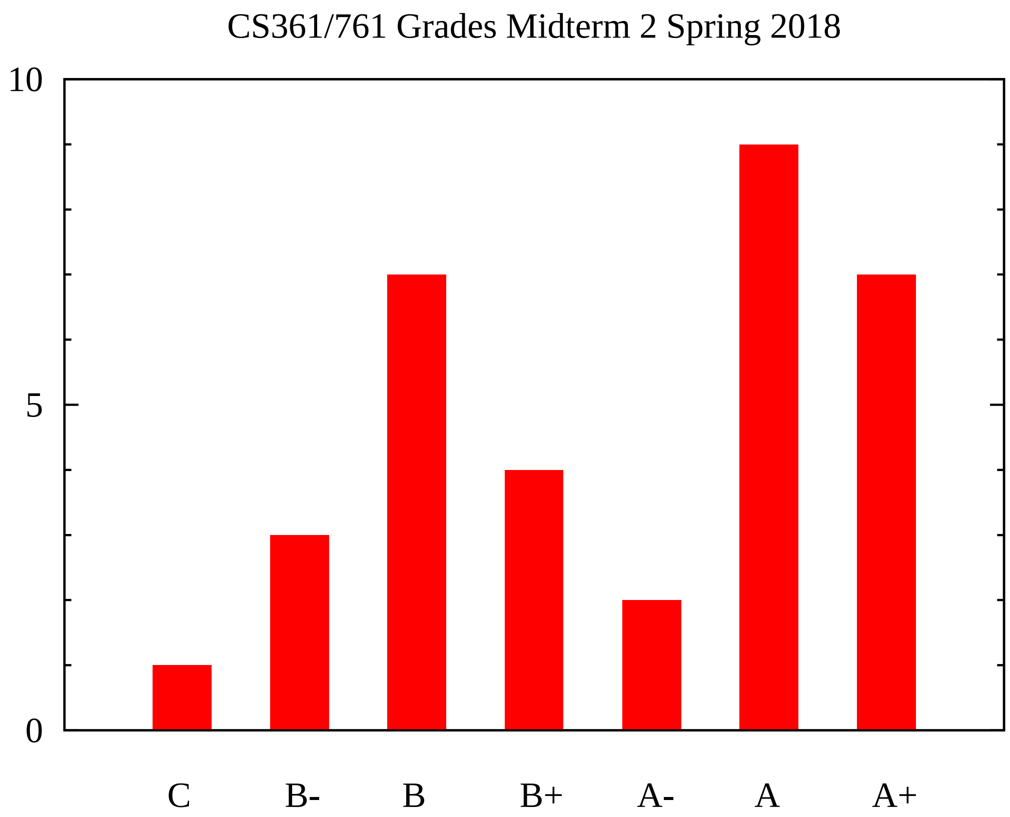

Figure 3: Histogram of grades for CS361/761 midterm 2 Spring 2018.