SIMON FRASER UNIVERSITY
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS

M. Alamgir Hossain
John M. Stockie

Clicker Question Bank for Numerical Analysis (Version 1.0 – May 14, 2020)

1. Introduction

Q1–1[1]. Select the best definition for "numerical analysis":

(A) the study of round-off errors

(B) the study of algorithms for computing approximate solutions to problems from continuous mathematics

(C) the study of quantitative approximations to the solutions of mathematical problems including consideration of and bounds for the errors involved

(D) the branch of mathematics that deals with the development and use of numerical methods for solving problems

(E) the branch of mathematics dealing with methods for obtaining approximate numerical solutions of mathematical problems

Answer: (B). All 5 definitions are valid in some sense since they reflect some aspect of the field (most are pulled off the internet). But my favourite definition is (B) because it contains three very important keywords underlined below:

the study of algorithms for computing approximate solutions to problems from continuous mathematics
[algorithms $\Longleftrightarrow$ computing, approximate $\Longleftrightarrow$ floating point arithmetic, continuous $\Longleftrightarrow$ solutions are smooth f'ns]

{ Source: JMS }

1a. Floating Point Arithmetic and Error

Q1a–1[2]. How many significant digits does the floating point number 0.03140×10^3 have?

(A) 6

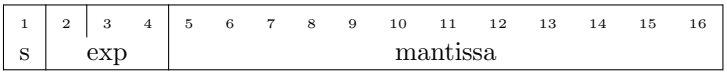
(B) 5

(C) 4

(D) 3

Answer: (C).

Q1a–2[3]. Suppose that a hypothetical binary computer stores floating point numbers in 16-bit words as shown:



Bit 1 is used for the sign of the number, bit 2 for the sign of the exponent, bits 3-4 for the magnitude of the exponent, and the remaining twelve bits for the magnitude of the mantissa. What is machine epsilon for this computer?

(A) 2^{-16}

(B) 2^{-12}

(C) 2^{-8}

(D) 2^{-4}

Answer: (B). Assume that rounding is used and recall that ε_{M} is essentially the same as unit round-off error $u = \frac{1}{2}B^{1-t}$, where $B = 2$ is the base and t is the number of significant digits. The number of digits stored in the mantissa is $t = 12$ and so $\varepsilon_{M} \approx \frac{1}{2}2^{1-12} = 2^{-12}$.

Q1a–3[4]. You are working with a hypothetical binary computer that stores integers as unsigned 4-bit words. What is the largest non-negative integer that can be represented on this computer?

(A) 64

(B) 63

(C) 31

(D) 15

(E) 7

Answer: (D). $(1111)_2 = 1 \times 2^0 + 1 \times 2^1 + 1 \times 2^2 + 1 \times 2^3 = 15$.

Q1a–4[5]. In 1958 the Russians developed a ternary (base-3) computer called *Setun*, after the Setun River that flows near Moscow State University where it was built. In contrast with today's binary computers, this machine used "trits" (ternary bits) whose three possible states can be represented as $\{0, 1, 2\}$. Its floating-point number system was based on 27-trit numbers, with 9 trits reserved for the exponent and 18 for the mantissa. What was the value of machine epsilon ϵ_{M} for the *Setun*?

(A) 3^{-19}

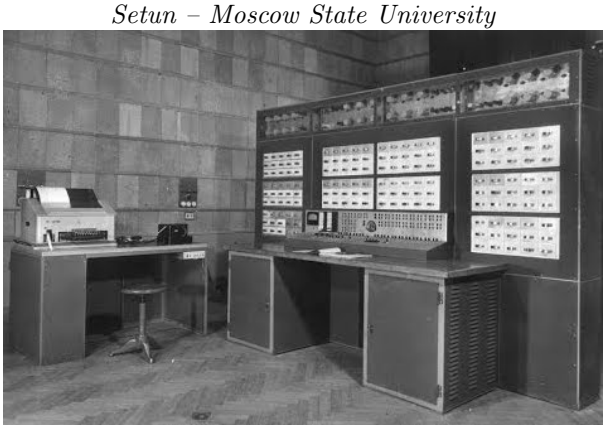
(B) 3^{-18}

(C) 3^{-9}

(D) $\frac{1}{3} \cdot 2^{-18}$

Answer: (B).
Apply the formula $\epsilon_{M} = B^{-t}$ from the notes, where $B = 3$ is the base and $t = 18$ is the number digits in the mantissa. You may have noticed that I didn't mention a "sign trit" for the mantissa. In actual fact, the floating-point representation on Setun *was more complicated than this and the sign of a number came from interpreting one specific trit as $\{-1, 0, +1\}$ instead.*

Setun – Moscow State University



{ Source: JMS, plus info from `http://homepage.divms.uiowa.edu/~jones/ternary/numbers.shtml` }

Q1a–5[6]. In Canada, the total for any store purchase paid in cash is rounded to the nearest 5 cents, whereas no rounding is done if the payment is by credit/debit card. Suppose that when you return home after purchasing your groceries with cash, you notice that your bill was \$10.07. What is the absolute error in your actual cash payment?

(A) 2 cents

(B) 3 cents

(C) 4 cents

(D) 5 cents

Answer: (A).

Q1a–6[7]. Let $\hat{x}$ be some approximation of x. Which of the following error definitions is correct?

(A) absolute error $= |x - \hat{x}|$, relative error $= \dfrac{|x - \hat{x}|}{|x|}$

(B) absolute error $= \dfrac{|x - \hat{x}|}{|x|}$, relative error $= |x - \hat{x}|$

(C) absolute error $= \dfrac{|x - \hat{x}|}{|x|}, x \neq 0$, relative error $= |x - \hat{x}|$

(D) absolute error $= |x - \hat{x}|$, relative error $= \dfrac{|x - \hat{x}|}{|x|}, x \neq 0$

Answer: (D).

Q1a–7[8]. For a base-10 (decimal) floating point number x having t significant digits, the relative error satisfies

$$R_x = \frac{|x - fl(x)|}{|x|} \leqslant u$$

where u denotes unit round-off error. Which of the following is true about u?

(A) $u = \begin{cases} 10^{1-t}, & \text{chopping} \\ \frac{1}{2}10^{1-t}, & \text{rounding} \end{cases}$

(B) $u = \begin{cases} \frac{1}{2}10^{1-t}, & \text{chopping} \\ 10^{1-t}, & \text{rounding} \end{cases}$

(C) $u = \begin{cases} \frac{1}{2}10^{1-t}, & \text{rounding} \\ 10^{1-t}, & \text{chopping} \end{cases}$

(D) $u = \begin{cases} 10^{1-t}, & \text{rounding} \\ \frac{1}{2}10^{1-t}, & \text{chopping} \end{cases}$

Answer: (A).

Q1a–8[9]. *Fill in the blank:* If $f(x)$ is a real-valued function of a real variable, then the _____ error in the difference approximation for the derivative $f'(x) \approx \dfrac{f(x+h) - f(x)}{h}$ goes to zero as $h \to 0$.

(A) absolute

(B) relative

(C) cancellation

(D) truncation

Answer: (D). Strictly, response (A) is also correct since truncation error is an (absolute) difference from the exact derivative.

Q1a–9[10]. The two solutions of the quadratic equation $ax^2 + bx + c = 0$ given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

are computed using floating point arithmetic. Which of the statements below is TRUE?

(A) For some values of the coefficients, this formula can generate cancellation errors.

(B) If the coefficients a, b and c are very small or very large, then b^2 or $4ac$ may overflow or underflow.

(C) The expression $x = \dfrac{2c}{-b \mp \sqrt{b^2 - 4ac}}$ is an alternative formula for x that avoids truncation error.

(D) All of the above.

Answer: (D).

Q1a–10[11]. In floating-point arithmetic, which of the following operations on two positive floating-point numbers can produce an overflow?

(A) addition

(B) subtraction

(C) multiplication

(D) division

Answer: (A). But (C) and (D) are also valid responses. Let x be the largest number that can be represented. Then the operations $x + 1.0$, $x * 2.0$ and $x \div 0.3$ all generate an overflow.

{ Source: Heath [?], Review Question 1.29, p. 40 }

Q1a–11[12]. In floating-point arithmetic, which of the following operations on two positive floating-point numbers can produce an underflow?

(A) addition

(B) subtraction

(C) multiplication

(D) division

Answer: (C). But (D) is also a valid response. Let x be the smallest positive number that can be represented. Then the operations $x * 0.5$ and $x \div 2.3$ both generate an underflow.

{ Source: Heath [?], Review Question 1.30, p. 40 }

Q1a–12[13]. Let $\{x_k\}$ be a decreasing sequence of positive numbers with $x_{k+1} < x_k$ for $k = 1, 2, \ldots$. In what order should the sum $\displaystyle\sum_{k=1}^{N} x_k$ be computed so as to minimize round-off error?

(A) Order the x_k from largest to smallest $(1, 2, \ldots, N)$.

(B) Order the x_k from smallest to largest $(N, \ldots, 2, 1)$.

(C) Sum the terms in random order.

(D) It doesn't matter.

Answer: (B).

{ Source: Heath [?], adapted from Review Question 1.45, p. 41 }

Q1a–13[14]. *True or False:* If two real numbers can be represented exactly as floating-point numbers, then the result of a real arithmetic operation on them can also be represented exactly as a floating-point number.

Answer: FALSE. As a counterexample, let $x_1 = \varepsilon_M$ (machine epsilon) and $x_2 = 2$, which are exact in any other binary floating point system (like the IEEE standard). Then x_1/x_2 has no floating point representation.

{ Source: Heath [?], Review Question 1.7, p. 39 }

Q1a–14[15]. Below are four floating point approximations, each accompanied by its corresponding exact value. Which approximation is the most accurate?

(A) 315700, exact value 315690

(B) 0.0005500, exact value 0.0005510

(C) 8.7362×10^{-5}, exact value 8.7743×10^{-5}

(D) ε_M (machine epsilon), exact value 0

Answer: (A). Accuracy is measured either by counting significant digits or computing relative error. Answer (A) has the most significant digits of accuracy (4 after rounding), whereas choices (B) and (C) have 2 and 3 significant digits. The accuracy of the answer from (D) can't be compared because relative error formula is undefined when the exact answer is zero.

{ Source: JMS }

Q1a–15[16]. You are computing using a floating-point number system that stores a total of t decimal digits in the mantissa. If you perform an arithmetic computation that has a result with relative error R_x, which of the statements below is TRUE?

(A) The number of significant digits in the answer is t.

(B) The number of significant digits cannot be predicted because of cancellation or round-off errors.

(C) The number of significant digits is roughly $-\log_{10}(R_x)$.

(D) None of the above.

Answer: (D).

{ Source: JMS }

Q1a–16[17]. The relative error in an approximate solution is 0.004%. The number of significant digits in the solution that we can trust is:

(A) 2

(B) 3

(C) 4

(D) 5

Answer: (C). This percentage corresponds to a relative error of 4×10^{-5}. With rounding this means that the solution is accurate to within only 4 significant digits.

{ Source: Holistic Numerical Methods [?], quiz_01_02 }

Q1a–17[18]. Let $\hat{x}$ be some approximation of an exact value x. Which of the statements below is FALSE?

(A) The relative error is always smaller than the absolute error.

(B) The relative error can be smaller than the absolute error provided x is large enough.

(C) The relative error gives a better idea of the number of correct significant digits.

(D) None of the above.

Answer: (A). Because absolute error is $E_x = |x - \hat{x}|$ and relative error is $R_x = |x - \hat{x}|/|x| = E_x/|x|$, the only time $R_x < E_x$ is when $|x| > 1$.

{ Source: JMS }

Q1a–18[19]. What is the decimal equivalent of the binary number 110010_2?

(A) 100

(B) 50

(C) 48

(D) 25

Answer: (B). $110010_2 = 2^5 + 2^4 + 2^1 = 32 + 16 + 2 = 50_{10}$.

{ Source: Holistic Numerical Methods [?], quiz_01_04 }

Q1a–19[20]. What is the binary equivalent of the decimal number 25.375_{10}?

(A) 100110.011

(B) 10011.110

(C) 11001.011

(D) 10011.0011

Answer: (C). Rewrite the decimal number as a sum of powers of 2:

$$25.375_{10} = 16 + 8 + 1 + \frac{1}{4} + \frac{1}{8} = 2^4 + 2^3 + 2^0 + 2^{-2} + 2^{-3} = 11001.011_2$$

{ Source: Holistic Numerical Methods [?], quiz_01_04 }

Q1a–20[21]. Suppose that you are working with some small quantity $h \ll 1$. Which of the following expressions is NOT of $O(h^4)$ as $h \to 0$?

(A) $h^5 + 7h^4 - 0.1h^3$

(B) $56h^4$

(C) h^{27}

(D) $27h^{4.3}$

(E) All of the above

Answer: (A). Only response (A) has a term of the form ch^p with $p < 4$ that goes to zero slower than h^4.

{ Source: JMS }

Q1a–21[22]. Suppose that you are working with some large quantity $N \gg 1$. Which of the following expressions is NOT of $O(N^4)$ as $N \to \infty$?

(A) $\dfrac{N^6 + 1}{2N^2 - N + 1}$

(B) $0.0001N^4$

(C) $N^{3.6}$

(D) $N^4 + e^N$

(E) All of the above

Answer: (D). Only response (D) has a term (e^N) that grows faster than N^4.

{ Source: JMS }

1b. Calculus Review

Q1b–1[23]. When estimating $f'(2)$ for $f(x) = x$ using the formula $f'(x) \approx \dfrac{f(x+h) - f(x)}{h}$ and $h = 0.1$, the truncation error is:

(A) 0

(B) 0.1

(C) 1

(D) 2

Answer: (A). This formula can be derived from the linear Taylor approximation, which is exact for linear functions.

Q1b–2[24]. When estimating $f'(2)$ for $f(x) = x^2$ using the formula $f'(x) \approx \dfrac{f(x+h) - f(x)}{h}$ and $h = 0.1$, the truncation error is:

(A) 0

(B) 0.1

(C) 1

(D) 2

Answer: (B).

Q1b–3[25]. Which of the following is NOT a Taylor series for $f(x)$?

(A) $f(x+h) = f(x) + f'(x)h + \dfrac{f''(x)}{2!}h^2 + \dfrac{f'''(x)}{3!}h^3 + \cdots$

(B) $f(x) = f(x_0) + f'(x_0)(x-x_0) + \dfrac{f''(x_0)}{2!}(x-x_0)^2 + \dfrac{f'''(x_0)}{3!}(x-x_0)^3 + \cdots$

(C) $f(x) = f(0) + f'(0)x + \dfrac{f''(0)}{2!}x^2 + \dfrac{f'''(0)}{3!}x^3 + \cdots$

(D) $f(x) = f(x_0) + f'(x_0)x + \dfrac{f''(x_0)}{2!}x^2 + \dfrac{f'''(x_0)}{3!}x^3 + \cdots$

Answer: (D).

Q1b–4[26]. The coefficient of x^3 in the Taylor series for $f(x) = \ln(x)$ centered about $x = 1$ is:

(A) $\dfrac{1}{4}$

(B) $\dfrac{1}{3}$

(C) $\dfrac{2}{3}$

(D) $-\dfrac{2}{3}$

Answer: (B).

Q1b–5[27]. Using the Taylor series expansion of $f(x) = \cos(x)$ near $x = 0$, an approximation of $f(h)$ for small h is:

(A) $\dfrac{\sin(h)}{h}$

(B) $\dfrac{\cos(h)-1}{h}$

(C) $\sin(0)$

(D) $\cos(0)$

(E) $1 - \frac{1}{2}h^2$

Answer: (D). The first term in the Taylor series is $\cos(0) = 1$. *However, response (E) is also correct since it comes from expanding the Taylor series by an additional two terms:*

$$\cos(h) \approx \cos(0) - \sin(0)h - \frac{1}{2}\cos(0)h^2$$

Q1b–6[28]. Using the Taylor series expansion of $f(x) = \cos(x)$ near $x = 0$, an approximation of $f'(0)$ for small h is:

(A) $\dfrac{\sin(h)}{h}$

(B) $\dfrac{\cos(h)-1}{h}$

(C) $\sin(0)$

(D) $\cos(0)$

(E) $1 - \frac{h^2}{2}$

Answer: (B). Expand to get $f(0+h) \approx f(0) + hf'(0)$ and then solve for $f'(0)$.

Q1b–7[29]. Let $f(x)$ be a function that is differentiable on the interval $[-5, 5]$. If $f(-5) = 10$, $f(0) = -10$, and $f(5) = 10$, then which of the statements below must be TRUE?

I. For some $c \in (-5, 5)$, $f(c) = 0$.

II. For some $c \in (-5, 5)$, $f'(c) = 0$.

(A) I only

(B) II only

(C) Both I and II

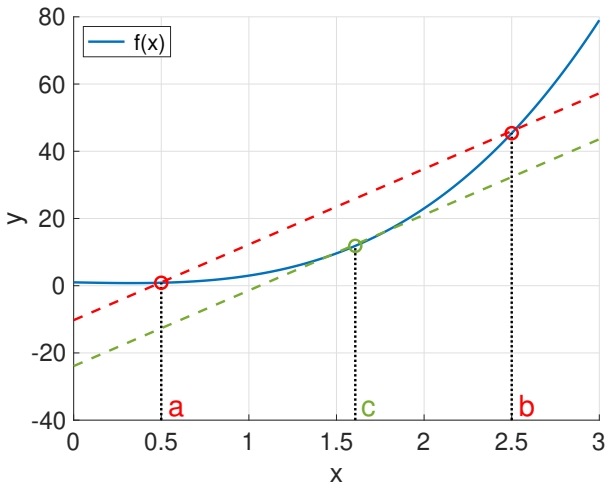
(D) Neither I nor II

Answer: (C). II follows from the Mean Value Theorem.

Regarding I, the Intermediate Value Theorem (IVT) doesn't apply directly to the interval $(-5, 5)$ since $f(-5) = f(5)$. However, f changes sign on both intervals $[-5, 0]$ and $[0, 5]$. So there are at least two roots in $(-5, 5)$ which means that I is also true.

{ Source: JMS }

Q1b–8[30]. Let $f(x)$ be continuous and differentiable for all $x \in [a, b]$ as shown in the plot below. Then there exists some real number $c \in (a, b)$ satisfying:



(A) $f(c) = \dfrac{f(b) - f(a)}{b - a}$

(B) $f'(c) = \dfrac{f(b) - b}{f(a) - a}$

(C) $f'(c) = 0$

(D) $f'(c) = \dfrac{f(b) - f(a)}{b - a}$

Answer: (D). This is an application of the Mean Value Theorem.

Q1b–9[31]. The series $\displaystyle\sum_{n=0}^{\infty} (-4)^n \frac{x^{2n}}{(2n)!}$ is the Taylor series at $x = 0$ for the following function:

(A) $\cos(x)$

(B) $\cos(2x)$

(C) $\sin(x)$

(D) $\sin(2x)$

Answer: (B).

{ Source: Holistic Numerical Methods [?], quiz_01_07 }

Q1b–10[32]. Given that $f(2) = 6$, $f'(2) = -\frac{1}{2}$ and $f''(2) = 10$, what is the most accurate Taylor polynomial approximation of $f(2.2)$ that you can find?

(A) 5.9

(B) 6.1

(C) 6.2

(D) 7

Answer: (B). These three point values determine the second degree Taylor polynomial:

$$f(x) \approx f(2) + f'(2)(x-2) + \frac{1}{2}f''(2)(x-2)^2 = 6 - \frac{1}{2}(x-2) + \frac{10}{2}(x-2)^2$$

and then substituting $x = 2.2$ in the last expression gives

$$f(2.2) \approx 6 - \frac{1}{2}(0.2) + 5(0.2)^2 = 6 - 0.1 + 0.2 = 6.1$$

{ Source: JMS }

Q1b–11[33]. Which is the Taylor series for the function $\ln(x)$ at the point $a = 1$?

(A) $(x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \frac{1}{4}(x-1)^4 + \cdots$

(B) $(x-1) - (x-1)^2 + 2(x-1)^3 - 6(x-1)^4 + \cdots$

(C) $\ln(x) + \frac{1}{x}(x-1) - \frac{1}{x^2}(x-1)^2 + \frac{2}{x^3}(x-1)^3 - \frac{6}{x^4}(x-1)^4 + \cdots$

(D) $\ln(x) + \frac{1}{x}(x-1) - \frac{1}{2x^2}(x-1)^2 + \frac{1}{3x^3}(x-1)^3 - \frac{1}{4x^4}(x-1)^4 + \cdots$

Answer: (A). It's easy to eliminate (C) and (D) since they are not polynomials.

{ Source: MathQuest [?], Taylor series }

Q1b–12[34]. What function is represented by the Taylor series $1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720} + \cdots$ at $a = 0$?

(A) e^x

(B) $\sin x$

(C) $\cos x$

(D) This is not a Taylor series

Answer: (C).

{ Source: MathQuest [?], Taylor series }

Q1b–13[35]. Suppose that you determine the Taylor series for some function $f(x)$ centered at $a = 5$. At which point x would you expect a finite number of terms from this series to give a better approximation?

(A) $x = 0$

(B) $x = 3$

(C) $x = 8$

(D) There is no way to tell

Answer: (B). This is the closest point to $x = 5$ and so the remainder term suggests that the error is smallest there. However, the precise error also depends on the derivative factor $f^{(n)}(c)$ for some value of c between a and x, which can't be determined without knowing $f(x)$. So (D) is strictly correct.

{ Source: MathQuest [?], Taylor series }

2. Nonlinear Equations

Q2–1[36]. How many zeroes does a polynomial of degree n have?

(A) $n+2$

(B) $n+1$

(C) n

(D) $n-1$

Answer: (C). The fundamental theorem of algebra guarantees that there are exactly n complex roots (or zeroes), some of which might be repeated roots. However, if you are interested in real roots, then they could number anywhere between zero and n.

{ Source: Holistic Numerical Methods [?] }

Q2–2[37]. How many real roots does the equation $\sin(x) - x = 0$ have?

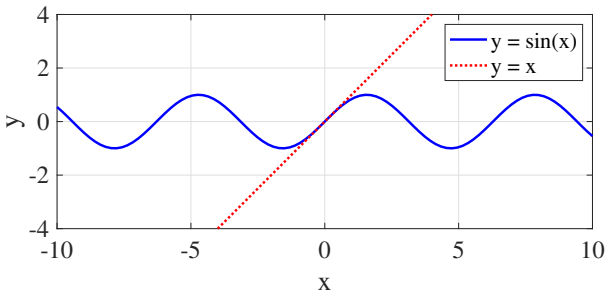
(A) 0

(B) 1

(C) 3

(D) Infinitely many

Answer: (B). Rewriting the equation as $\sin(x) = x$ suggests sketching the two functions $y = x$ and $y = \sin(x)$ on the same axes and looking for intersection points. Then it's easy to confirm that $x = 0$ is the only real root.



{ Source: MAH }

Q2–3[38]. How many real roots does the equation $\sin(x) - ax = 0$ have? (a is some real constant)

(A) 0

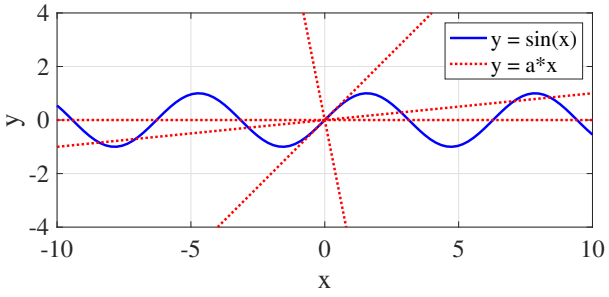
(B) 1

(C) 3

(D) Infinitely many

(E) It depends on a

Answer: (E). By sketching the function $y = \sin(x)$ alongside various straight lines $y = ax$ it is easy to show that there is always at least 1 root ($x = 0$) and that there are infinitely many roots if $a = 0$. With a bit more investigation, as a varies there can be any odd, positive number of roots. So strictly, responses (B)–(E) are all correct, while response (A) is not.



{ Source: JMS }

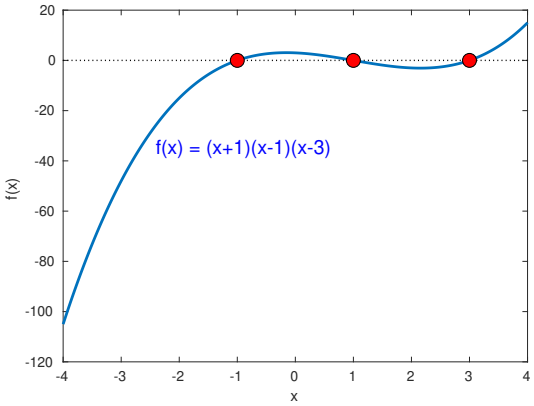
Q2–4[39]. *True or False:* When searching iteratively for a solution x^* of $f(x) = 0$, having a small value of the residual $|f(x_k)|$ guarantees that x_k is an accurate approximation of x^*.

Answer: FALSE.

{ Source: Heath [?], Review Question 5.1, p. 245 }

2a. Bisection Method

Q2a–1[40]. The function $f(x) = (x+1)(x-1)(x-3)$ is pictured in the plot. If the bisection algorithm is applied with initial interval $[-4, 4]$, how many roots of $f(x)$ will you be able to compute?



(A) 3

(B) 2

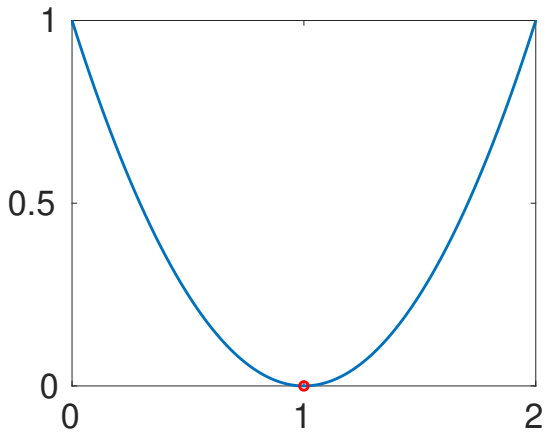
(C) 1

(D) none of the above

Answer: (C). All three roots $(-1,\ 1$ *and* $3)$ *lie within the interval* $[-4, 4]$*, so the bisection algorithm is guaranteed to converge. However, the algorithm will only converge to one of those roots.*

{ Source: MAH }

Q2a–2[41]. *True or False:* Let $f(x) = x^2 - 2x + 1$. The bisection method can be used to approximate the root of the function $f(x)$ pictured.



Answer: FALSE.

{ Source: MAH }

Q2a–3[42]. You are provided with a list of point values for the function $f(x) = x + 10 - e^x$:

x	0	1	2	3	4
$f(x)$	9.000	8.282	4.611	−7.086	−40.598

Use this data to perform two steps of the bisection method for solving $f(x) = 0$, assuming the initial interval $[0, 4]$. What is the approximation of the root?

(A) 1

(B) 2

(C) 3

(D) 4

Answer: (C).

{ Source: MACM 316 final exam question (Fall 2018) }

Q2a–4[43]. You are computing a root of $f(x) = \cos(x) - x$ on the interval $[0, 2]$ using the bisection method. If you require a result that accurate to within five significant digits, which of the following bounds holds for the minimum number of bisection iterations k required?

(A) $\dfrac{1}{2^{k-1}} < 10^{-5}$

(B) $\dfrac{1}{2^{k}} < 10^{-5}$

(C) $\dfrac{2}{2^{k+1}} < 10^{-5}$

(D) none of the above

Answer: (A).

{ Source: MAH }

Q2a–5[44]. Which of the statements below regarding the convergence of the bisection method is TRUE?

I. The iteration is always guaranteed to converge.

II. The order of convergence is linear.

III. The asymptotic rate of convergence is $\frac{1}{2}$.

(A) I and II
(B) II and III
(C) I and III
(D) I, II and III

Answer: (D). All three statements are true, but only assuming that you choose an initial interval $[a, b]$ having $f(a) \cdot f(b) < 0$. If that's not the case, then the method will not converge so none of the statements hold.

{ Source: JMS, MACM 316 lecture notes }

Q2a–6[45]. You are given an interval $[a, b]$ where $f(a) \cdot f(b) < 0$. Which of the statements below regarding the bisection method is TRUE?

I. There must be a root x^* somewhere in the interval (a, b).

II. The absolute error in the first step of bisection is $|x_1 - x^*| \leqslant \dfrac{b-a}{2}$.

III. The bisection iteration is guaranteed to converge if $|f'(x)| < 1$.

(A) I and II
(B) II and III
(C) I and III
(D) I, II and III

Answer: (A).

{ Source: JMS, MACM 316 lecture notes }

Q2a–7[46]. Which of the following is a suitable initial interval for computing a positive root of the equation

$$g(x) = 2x\cos(\pi x) - e^{x-1} = 0$$

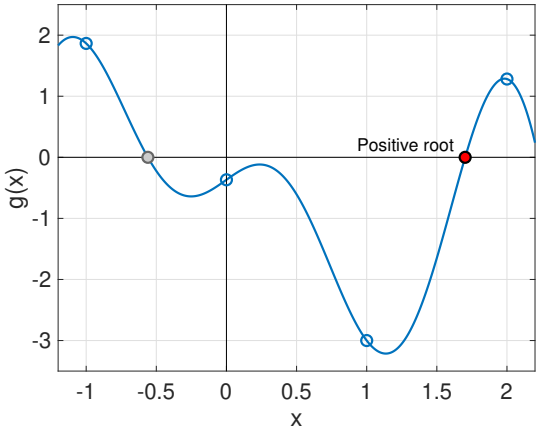
using the bisection method?

(A) $[-1, 0]$
(B) $[-1, 1]$
(C) $[0, 1]$
(D) $[0, 2]$

Answer: (D). From the interval end-point values

$$g(-1) = 2 - e^{-2} > 0, \quad g(0) = -1 < 0, \quad g(1) = -2 - e^{0} = -3 < 0, \quad g(2) = 4 - e > 0,$$

the only interval that is a valid bracket for a positive root is $[0, 2]$, since $g(0) \cdot g(2) < 0$. Response (A) and (B) are also brackets, but $[-1, 0]$ can only contain negative roots, while $[-1, 1]$ may or may not have a positive root. This plot displays the function, the interval end-points, and the desired positive root:



{ Source: JMS }

Q2a–8[47]. This partial code implements the bisection method for finding a root of the nonlinear equation $f(x) = 0$. The Matlab function `f` computes the function values and you can assume that you start with an interval $[a, b]$ satisfying $f(a) \cdot f(b) < 0$. Select the suitable replacement code for the blanks marked (1) and (2) from the list below.

```
tol = 1e-6; % tolerance
fa = f(a); fb = f(b);
done = 0;
while( (b-a)/2 > tol ),
   mid = (a+b)/2;
   fmid = f(mid);
   if fmid == 0, % mid is a solution
      break;
   end
   if sign(fmid)*sign(fa) < 0,
      ...(1)...
   else
      ...(2)...
   end
end
```

(A) (1) `b=mid; fb=fmid;` (2) `a=mid; fa=fmid;`

(B) (1) `a=mid; fa=fmid;` (2) `b=mid; fb=fmid;`

(C) (1) `a=mid; fb=fmid;` (2) `b=mid; fa=fmid;`

(D) (1) `a=mid;` (2) `b=mid;`

Answer: (A).

{ Source: MACM 316 midterm question (Fall 2018) }

2b. Fixed Point Method

Q2b–1[48]. Consider the fixed point iteration $x_{k+1} = g(x_k)$ with $g(x) = \dfrac{x}{3} + \dfrac{4}{3x}$. Which root-finding problem is this equivalent to?

(A) $x^2 - 2 = 0$

(B) $\dfrac{x}{3} + \dfrac{4}{3x} = 0$

(C) $\dfrac{1}{3} - \dfrac{4}{3x^2} = 0$

(D) $x - \dfrac{1}{3} + \dfrac{4}{3x^2} = 0$

Answer: (A).

{ Source: MAH }

Q2b–2[49]. Assuming that the following fixed point iteration converges

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{3}{x_k}\right),$$

to which fixed point will it converge?

(A) $\sqrt{3}$

(B) $\sqrt{2}$

(C) $\sqrt{5}$

(D) none of the above

Answer: (A).

{ Source: MAH }

Q2b–3[50]. For which of the functions below is $x^* = \sqrt{5}$ a fixed point?

(A) $g(x) = \dfrac{x}{\sqrt{5}}$

(B) $g(x) = \sqrt{5}\,x$

(C) $g(x) = x^2 - 4x$

(D) $g(x) = 1 + \dfrac{4}{x+1}$

Answer: (D).

{ Source: MAH }

Q2b–4[51]. Which of the following is a suitable fixed point iteration for solving the equation $\cos(x) - x = 0$?

I. $x_{k+1} = \cos(x_k)$

II. $x_{k+1} = \cos^{-1}(x_k)$

III. $x_{k+1} = x_k + \tan(x_k)$

(A) I

(B) I and II

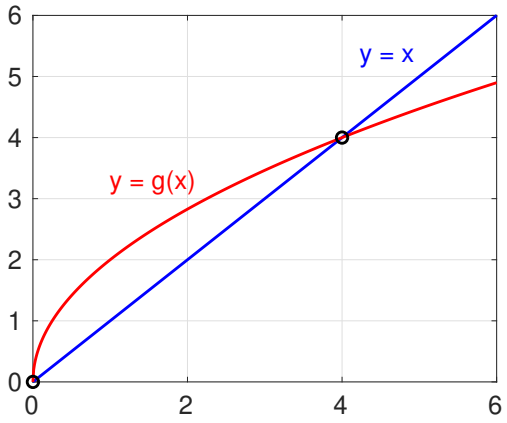
(C) I and III

(D) I, II and III

Answer: (B).

{ Source: MAH }

Q2b–5[52]. The intersection points between the curves $y = x$ and $y = g(x)$ are $x = 0$ and $x = 4$, as shown in the plot. Which of the statements below regarding the fixed point iteration $x_{k+1} = g(x_k)$ is TRUE?



I. If $x_0 = 2$ then x_k converges to 4.

II. If $x_0 = 1$ then x_k converges to 0.

III. If $x_0 = 6$ then x_k converges to 4.

(A) I and II

(B) II and III

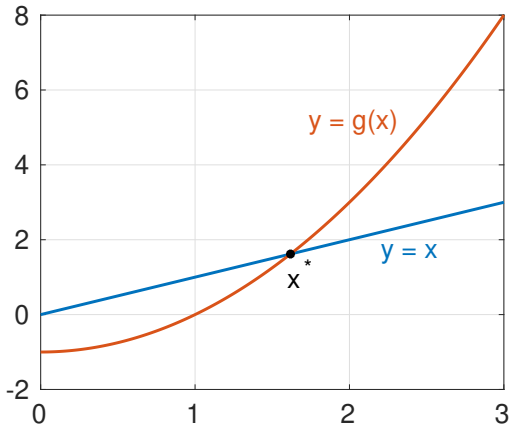
(C) I and III

(D) I, II and III

Answer: (C).

{ Source: MAH }

Q2b–6[53]. The intersection point between the curves $y = x$ and $y = g(x)$ is $x = x^*$, as shown in the plot. Which of the statements below regarding the fixed point iteration $x_{k+1} = g(x_k)$ is TRUE?



I. Starting at $x_0 = 2$, the iteration converges to x^*.

II. Starting at $x_0 = 1$, the iteration converges to x^*.

(A) I

(B) II

(C) I and II

(D) Neither I nor II

Answer: (D).

{ Source: MAH }

Q2b–7[54]. Which of these statements is TRUE?

I. The fixed point and bisection methods have the same order of convergence.

II. If the iteration function $g(x)$ for the fixed point method satisfies $|g'(x)| < 1$, then the fixed point algorithm converges faster than bisection.

III. When it converges, the fixed point method always converges slower than bisection.

(A) I and II

(B) II and III

(C) I and III

(D) I, II and III

Answer: (C). Statement II is false in general, but could be true if $|g'(x)| < \frac{1}{2}$ since then the asymptotic rate constant for the fixed point method is smaller than that of bisection.

{ Source: JMS, MACM 316 lecture notes }

Q2b–8[55]. This partial code implements the fixed point algorithm for finding a root of the equation $x = g(x)$, where the Matlab function `g` returns the value of the fixed point iteration function. Select the best choice of terminating condition for this algorithm to fill in the missing condition in the `if` statement.

```
tol = 1e-6; % tolerance
x = xguess; % initial guess
done = 0;
while( ~done ),
   niter = niter + 1;
   x0 = x;
   x = g(x);
   if ...[fill in blank]...
      done = 1;
   end
end
```

(A) `abs(x) < tol`

(B) `abs(x - x0) < tol`

(C) `x - x0 < tol`

(D) `abs(g(x)) < tol`

Answer: (B).

{ Source: MACM 316 midterm question (Fall 2018) }

2c. Secant and Newton's Methods

Q2c–1[56]. The irrational number $\sqrt{2}$ can be approximated by applying the secant method to the nonlinear equation $x^2 - 2 = 0$. What is the iteration formula?

(A) $x_{k+1} = x_k - \dfrac{x_k^2 - 2}{x_k + x_{k-1}}$

(B) $x_{k+1} = x_k - \dfrac{(x_k^2 - 2)(x_k - x_{k-1})}{x_k + x_{k-1}}$

(C) $x_{k+1} = x_k - \dfrac{(x_k^2 - 2)(x_k - x_{k-1})}{x_k^2 + x_{k-1}^2}$

(D) $x_{k+1} = x_k - \dfrac{(x_k^2 - 2)(x_k - x_{k-1})}{x_k^2 - x_{k-1}^2 - 4}$

Answer: (A). $x_{k+1} = x_k - \dfrac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})} = x_k - \dfrac{(x_k^2 - 2)(x_k - x_{k-1})}{x_k^2 - x_{k-1}^2} = x_k - \dfrac{x_k^2 - 2}{x_k + x_{k-1}}$

{ Source: MAH }

Q2c–2[57]. *True or False:* The convergence of the secant method to simple roots is quadratic.

Answer: FALSE. The convergence of the secant method to simple roots is superlinear, and lies between linear and quadratic convergence (the order of convergence is actually $p \approx 1.618$).

{ Source: MAH }

Q2c–3[58]. *True or False:* Both Newton's method and the secant method are examples of fixed-point iterations.

Answer: FALSE. Newton's method is a fixed point iteration, but secant method is not.

Q2c–4[59]. The irrational number e can be approximated by applying Newton's method to solve the nonlinear equation $f(x) = \ln x - 1 = 0$. What is the Newton iteration formula?

(A) $x_{k+1} = \ln x_k$

(B) $x_{k+1} = x_k - \ln x_k$

(C) $x_{k+1} = x_k - (\ln x_k - 1)$

(D) $x_{k+1} = 2x_k - x_k \ln x_k$

Answer: (D).

{ Source: MAH }

Q2c–5[60]. The irrational number $\sqrt{2}$ can be approximated by applying Newton's method to the nonlinear equation $f(x) = x^2 - 2 = 0$. What is the Newton iteration formula?

(A) $x_{k+1} = x_k + \dfrac{x_k^2 - 2}{2x_k}$

(B) $x_{k+1} = x_k - \dfrac{x_k^2 - 2}{x_k}$

(C) $x_{k+1} = x_k - \dfrac{x_k^2 - 2}{x_k}$

(D) $x_{k+1} = x_k - \dfrac{x_k^2 - 2}{2x_k}$

(E) $x_{k+1} = \dfrac{x_k}{2} + \dfrac{1}{x_k}$

Answer: (D). Response (E) is also correct, since it's just a simplified version of (D).

Indeed, (E) is a remarkably efficient formula that has been used to implement the square root operation on many modern floating-point processors! Generalizing to the case of finding $\sqrt{c}$ as a root of $f(c) = x^2 - c$, the Newton iteration is

$$x_{k+1} = \frac{x_k}{2} + \frac{c}{2x_k} = \frac{1}{2}\left(x_k + c\,\frac{1}{x_k}\right)$$

Counting the flops involved, each iteration requires a reciprocal $(1/x_k)$, one multiplication by c, one addition, and one division by 2 (which is just a bit-shift in binary). This is an incredibly cheap way to approximate $\sqrt{c}$ when the number of Newton iterations is small.

{ Source: MAH }

Q2c–6[61]. Apply Newton's method to the equation $f(x) = x^2 - 2 = 0$ to estimate the root $x^* = \sqrt{2}$. Starting with the initial guess $x_0 = 1$, the first iteration x_1 is:

(A) 0.5

(B) 1.0

(C) 1.5

(D) 2.0

Answer: (C). The Newton iteration is $x_1 = x_0 - \frac{x_0^2-2}{2x_0} = \frac{x_0}{2} + \frac{1}{x_0}$. Setting $x_0 = 1$ gives $x_1 = \frac{3}{2}$.

{ Source: MAH }

Q2c–7[62]. For which values of x does Newton's method exhibit linear and quadratic order of convergence when applied to the function $f(x) = x(x-3)^2$?

(A) linear near $x = 0$ and quadratic near $x = 3$

(B) linear near $x = 3$ and quadratic near $x = 0$

(C) linear convergence everywhere

(D) quadratic convergence everywhere

Answer: (B).

{ Source: MAH }

Q2c–8[63]. Which of these statements is TRUE?

I. Newton's method may converge linearly

II. Newton's method may converge quadratically

III. Newton's method may not converge at all

(A) none is true

(B) I and II only

(C) II and III only

(D) I, II and III

Answer: (D).

{ Source: Heath [?], Review Question 5.16, p. 246 }

Q2c–9[64]. Which of these statements is TRUE?

I. Newton's method is guaranteed to converge

II. Newton's method has a quadratic order of convergence

(A) I
(B) II
(C) I and II
(D) Neither I nor II

Answer: (D).

{ Source: JMS, MACM 316 lecture note }

Q2c–10[65]. The Newton iteration $x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)}$ can be rewritten as a fixed point iteration

$$x_{k+1} = g(x_k) \text{ where } g(x) = x - \frac{f(x)}{f'(x)}.$$

Which of these statements is TRUE?

I. Newton's method fails when $f'(x_k) = 0$.
II. The iteration converges if $|g'(x_k)| < 1$.

(A) I
(B) II
(C) I and II
(D) Neither I nor II

Answer: (C).

{ Source: MAH }

Q2c–11[66]. Which of these statements regarding the secant method is TRUE?

I. It is an example of a fixed-point iteration.
II. It requires two initial guesses.
III. No derivative calculation is needed.

(A) I
(B) I and II only
(C) II and III only
(D) I, II and III

Answer: (C).

{ Source: JMS, MACM 316 lecture notes }

Q2c–12[67]. This partial code implements Newton's method for finding a root to the nonlinear equation $f(x) = 0$, where Matlab functions `f` and `fprime` compute the function and its derivative. Select a suitable terminating condition (to replace the blank `if` statement) that will best improve the robustness of the code.

```
tol = 1e-6; % tolerance
x = 2.0; % initial guess
done = 0;
while( ~done ),
   xlast = x;
   fx = f(x);
   fpx = fprime(x);
   x = xlast - fx / fpx;
   if ...[fill in blank]...
      done = 1;
   end
end
```

(A) `abs(x) < tol`

(B) `abs(x - xlast)/abs(x) < tol`

(C) `abs(fx) < tol`

(D) `abs(x - xlast)/abs(x) < tol & abs(fx) < tol`

Answer: (D).

{ Source: MACM 316 midterm question (Fall 2018) }

Q2c–13[68]. This partial code implements the method of false position for finding a root of the nonlinear equation $f(x) = 0$, where the Matlab function `f` computes the function value. Provide suitable code to replace the blanks marked ① and ②.

```
while( abs(x0-x1) > tol ), % absolute error
tolerance
   x2 = x1 - f(x1)*(x1-x0)/(f(x1)-f(x0));
   if f(x2)*f(x1) > 0,
      ...①...
   else
      ...②...
   end
end
```

(A) ① `x1 = x2;` ② `x0 = x2;`

(B) ① `x2 = x1;` ② `x2 = x0;`

(C) ① `x0 = x2;` ② `x1 = x2;`

(D) ① `x2 = x0;` ② `x2 = x1;`

Answer: (A).

{ Source: MACM 316 midterm question (Fall 2018) }

Q2c–14[69]. After writing a Matlab code that implements Newton's method, you apply it to the function $f(x) = 2x^3 - 3x^2 + 4x - 6$. Using an initial guess of $x_0 = 1.5$, your code returns a result of "`-Inf`" on the first iteration. What is the most likely explanation for this behaviour?

(A) $f(x)$ has a simple root at $x = 1.5$.

(B) $f(x)$ has a multiple root at $x = 1.5$.

(C) $f'(x)$ has a root at $x = 1.5$.

(D) Accumulation of round-off error due to subtractive cancellation.

Answer: (C). This one is a little bit tricky. You need to look at the Newton iteration

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

and see there are two possible invalid floating point operations:

- `±Inf` *or "*$\frac{1}{0}$*": occurs if* $f'(x_k) = 0$ *and* $f(x_k) \neq 0$*. This is case (C).*
- `NaN` *or "*$\frac{0}{0}$*": occurs if both* $f'(x_k) = 0$ *and* $f(x_k) = 0$*. This is response (B) – a root with multiplicity 2.*

{ Source: JMS }

Q2c–15[70]. Which of the following is an advantage of the secant method over Newton's method?

(A) lower cost per iteration

(B) faster convergence rate

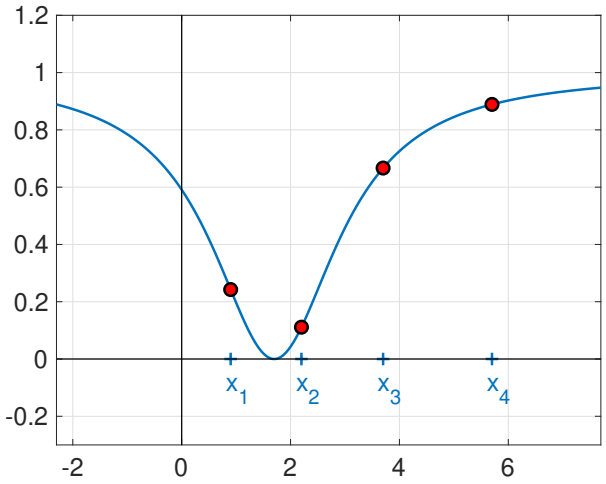
(C) more robust convergence far away from the solution

(D) avoids computing the derivative

Answer: (A). The secant method only requires a single function evaluation in each step (as long as the previous value is saved), whereas the Newton iteration must compute values of both f and f' in every step. This means that (D) is also correct.

{ Source: Heath [?], Review Question 5.38, p. 247 }

Q2c–16[71]. For the function $f(x)$ plotted below, use each of the four points x_1, x_2, x_3 and x_4 as the starting guess for Newton's method. For which point(s) do you expect the iteration to converge to the root shown?



(A) x_1 and x_2 only

(B) x_2 only

(C) x_1, x_2 and x_3 only

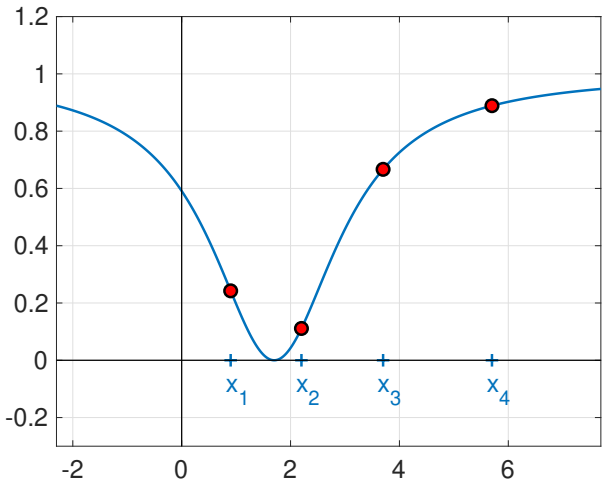
(D) All four points

(E) none of the points

Answer: (C). Check by plotting the Newton iterations on the graph.

{ Source: GoodQuestions [?], Newton's method }

Q2c–17[72]. For the function $f(x)$ plotted below, use each of the intervals $[x_1, x_2]$, $[x_1, x_3]$ and $[x_1, x_4]$ as the starting guess for the secant method. For which interval do you expect the iteration to converge to the root shown?



(A) $[x_1, x_2]$

(B) $[x_1, x_3]$

(C) $[x_1, x_4]$

(D) none of the intervals

Answer: (A). Check by plotting the secant lines joining each pair of points.

{ Source: GoodQuestions [?], Newton's method }

Q2c–18[73]. Let f be a differentiable function that is defined for all x. Starting Newton's method at a point x_0 where $f'(x_0) = 0$ is ...

(A) a good choice, because $x = x_0$ is a critical point of f and Newton's method will converge most rapidly to the root from there

(B) a bad choice, because the Newton iteration fails

(C) usually a bad choice, but it might work if we're lucky

Answer: (B). Strictly, (C) is also correct since it's possible we might have starting with an x_0 that is a root – this is what is meant by "lucky". In every other case, the Newton iteration formula fails because of division by zero.

{ Source: GoodQuestions [?], Newton's method }

Q2c–19[74]. Newton's method is an appealing root-finding technique because ...

(A) it can be used to find quick and accurate approximations to radical numbers like $\sqrt[4]{3}$ or $\sqrt[8]{13}$

(B) it can reliably determine all real roots of an n^{th} degree polynomial

(C) it can find a solution to $7x^8 + x^4 + 1 = 0$

(D) it converges faster than other root-finding methods

Answer: (A). The radical expression $\sqrt[n]{a}$ is a root of $f(x) = x^n - a = 0$, which can always be found with Newton's method (plot $f(x)$ to see why). Response (B) is incorrect because Newton's method requires a good initial guess for each root, which can be difficult/impossible in practice. Response (C) is incorrect because there is no real root (just rewrite as $7x^8 = -1 - x^4$, "positive equals negative"). Response (D) is often true, but not always.

{ Source: GoodQuestions [?], Newton's method }

2d. General Aspects of Convergence

Q2d–1[75]. Suppose you apply an iterative method and obtain the following errors from the first four steps:

$$10^{-2},\ 10^{-4},\ 10^{-6},\ 10^{-8},\ \ldots$$

How would you characterize the order of convergence of this method?

(A) linear

(B) super-linear

(C) quadratic

(D) faster than quadratic

Answer: (B). The error E_k is reduced by a factor of 100 in every iteration and so can be written $E_k = 10^{-2}E_{k-1}$. According to the definition of convergence, this is linear (order 1) convergence with an asymptotic rate constant equal to 10^{-2}.

{ Source: Heath [?], adapted from Review Question 5.9, p. 245 }

Q2d–2[76]. Suppose you apply an iterative method and obtain the following errors from the first four steps:

$$10^{-2},\ 10^{-4},\ 10^{-8},\ 10^{-16},\ \ldots$$

How would you characterize the order of convergence of this method?

(A) linear

(B) super-linear

(C) quadratic

(D) faster than quadratic

Answer: (C). The error E_k is squared in every iteration and so can be written $E_k = E_{k-1}^2$. According to the definition, this is quadratic (order 2) convergence with asymptotic rate constant equal to 1. Note that the response (B) is strictly also correct, although (C) is more precise.

{ Source: Heath [?], adapted from Review Question 5.9, p. 245 }

Q2d–3[77]. Suppose you apply an iterative method and obtain the following errors from the first six steps of an iterative method:

$$0.01369,\ 0.02553,\ 0.01158,\ 0.01044,\ 0.009781,\ 0.008235,\ \ldots$$

How would you characterize the order of convergence of this method?

(A) not converging

(B) linear

(C) quadratic

(D) impossible to determine

Answer: (A). It's actually very hard to tell in this case, and it would be helpful if there were more iterations to based the decision on. But we can make an educated guess. At best, the method is converging very slowly, perhaps linear with a rate constant just slightly less than 1. So (A), (B) and (D) are all reasonable responses. The anomalous "blip" at the start can probably be ignored because convergence isn't always uniform.

{ Source: JMS }

Q2d–4[78]. If you have an iterative method that converges linearly with rate constant $\frac{1}{2}$, how many iterations are required for the initial error to be reduced by a factor of at least 1000?

(A) 5

(B) 10

(C) 20

(D) 100

Answer: (B). The error from step k obeys $E_k \leqslant \frac{1}{2}E_{k-1} \implies E_k \leqslant (\frac{1}{2})^k E_0$. This means that we need $\left(\frac{1}{2}\right)^k \leqslant \frac{1}{1000} \implies 2^k \geqslant 1000 \implies k \geqslant \log_2(1000) \approx 9.97$.

{ Source: JMS }

Q2d–5[79]. If an iterative method approximately squares the error in every <u>two</u> iterations then what is its order of convergence?

(A) 0.5

(B) 1

(C) $\sqrt{2}$

(D) 2

Answer: (C). We're given that the error in step k satisfies $E_k \approx E_{k-2}^2$. But from the definition of convergence for a method of order p, the error in step k also satisfies $E_k \approx E_{k-1}^p \approx E_{k-2}^{p^2}$. Comparing these two expressions for E_k suggests that $p^2 = 2$ or $p = \sqrt{2}$.

{ Source: Heath [?], adapted from Review Question 5.14, p. 246 }

Q2d–6[80]. For some root finding method, you determine that the absolute error in step k satisfies $E_k \leqslant 0.2E_{k-1}^3$. Which of the statements below is TRUE?

(A) The method is superlinearly convergent.

(B) The order of the method is 0.2.

(C) The asymptotic rate of convergence is $\log_{10}(0.2)$.

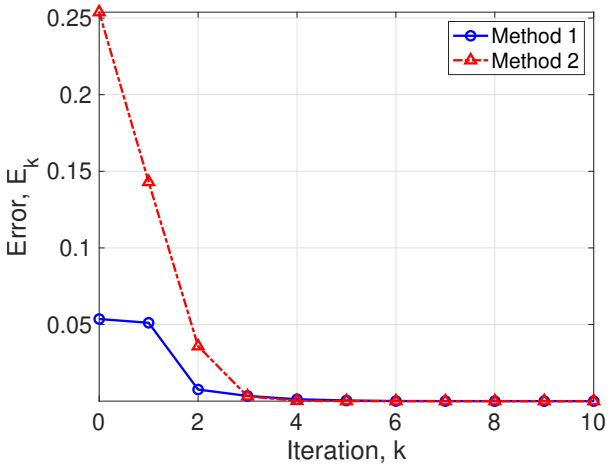
(D) The iteration will always converge.

(E) All of the above.

Answer: (A). Based on our definition of convergence, $\alpha = 0.2$ is the asymptotic rate constant and the order of convergence is $p = 3$ (faster than linear). Response (D) is not always true because the iteration may not converge if the initial guess has an error that satisfies $E_0 > 1$.

{ Source: JMS }

Q2d–7[81]. You use two different iterative methods to compute an approximation, and the errors from each method are plotted below versus the iteration number:



What can you conclude about the iterative convergence of the two methods?

(A) Both methods converge quadratically, but Method 1 has a smaller rate constant.

(B) The order of convergence for Method 1 is linear and Method 2 is quadratic.

(C) Both methods converge faster than linear, but the order cannot be estimated.

(D) No conclusion is possible based on this plot.

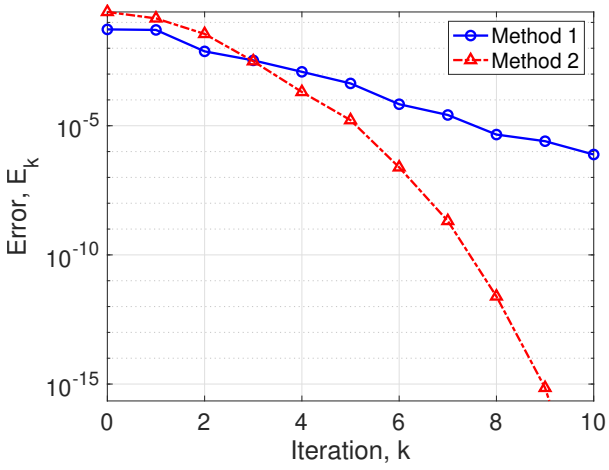
Answer: (D). The order of convergence can't be estimated using this error plot with both axes having linear scales. This is because the definition of convergence reads

$$E_k \leqslant \alpha E_{k-1}^p \quad \Longrightarrow \quad E_k \leqslant \alpha^k E_0^{p^k}$$

(for order p and rate α) which requires taking a logarithm of E_k to reliably estimate either p or α.

{ Source: JMS }

Q2d–8[82]. You use two different iterative methods to compute an approximation, and the errors from each method are plotted below versus the iteration number:



What can you conclude about the iterative convergence of the two methods?

(A) Both methods converge quadratically, but Method 1 has a smaller rate constant.

(B) The order of convergence for Method 1 is linear and Method 2 is quadratic.

(C) Method 1 converges linearly and Method 2 is super-linear.

(D) No conclusion is possible based on this plot.

Answer: (C). The points from Method 1 appear to lie on a straight line. Method 2 curves downward so it's definitely superlinear. It might converge quadratically but that's not possible to estimate from this semi-log plot. Instead, you would have to plot $\log\log E_k$ *(double-log) versus* k*.*

{ Source: JMS }

2e. Nonlinear Systems of Equations

[nothing here yet]

3. Linear Systems of Equations

3a. Review of Linear Algebra

Q3a–1[83]. Which of the following properties of the matrix determinant is TRUE?

(A) Only square matrices have a determinant

(B) For the identity matrix, $\det(I) = 1$

(C) $\det(AB) = \det(A)\det(B)$

(D) $\det(A^T) = \det(A)$

(E) All of the above

Answer: (E).

{ Source: JMS, MACM 316 lecture notes }

Q3a–2[84]. If A is an invertible matrix then which of these statements is TRUE?

(A) $\det(A) \neq 0$

(B) $Ax = b$ has a unique solution for any vector b

(C) $Ax = 0$ has only the trivial solution $x = 0$

(D) A^T is invertible

(E) All of the above

Answer: (E).

{ Source: JMS, MACM 316 lecture notes }

Q3a–3[85]. *True or False:* If a matrix A is nonsingular then the number of solutions to the linear system $Ax = b$ depends on the particular choice of right hand side vector b.

Answer: FALSE. There is a unique solution for any b.

{ Source: Heath [?], Review Question 2.1, p. 92 }

Q3a–4[86]. *True or False:* If a triangular matrix has a zero entry on its main diagonal then the matrix must be singular.

Answer: TRUE. The determinant of any triangular matrix is the product of the diagonal entries and so its determinant of this matrix is zero.

{ Source: Heath [?], Review Question 2.3, p. 92 }

Q3a–5[87]. If A is a 5×5 matrix then what is the value of $\det(3A)$?

(A) $\frac{1}{3^5}\det(A)$

(B) $\frac{1}{5^3}\det(A)$

(C) $3^5\det(A)$

(D) $5^3\det(A)$

Answer: (C).

{ Source: MAH }

Q3a–6[88]. If A is a 2×2 invertible matrix then what is the inverse of $2A$?

(A) $\frac{1}{2}A^{-1}$

(B) $\frac{1}{4}A^{-1}$

(C) $2A^{-1}$

(D) $4A^{-1}$

Answer: (A). Notice that $2A \cdot \frac{1}{2}A^{-1} = AA^{-1} = I$.

{ Source: MathQuest [?], Linear Algebra, Matrix Inverses }

Q3a–7[89]. *True or False:* $\|A\|_1 = \|A^T\|_\infty$.

Answer: TRUE. The 1-norm is the maximum column sum, the ∞*-norm is the maximum row sum, and the transpose "flips" the matrix so the columns turn into rows.*

{ Source: Heath [?], Review Question 2.24, p. 93 }

Q3a–8[90]. Which of the following is a reliable indicator that a matrix is nearly singular?

(A) a small determinant

(B) a small norm

(C) a large norm

(D) a large condition number

Answer: (D). Both (A) and (D) are acceptable answers, at least in exact arithmetic. But floating-point calculations of determinants can lead to very large growth in round-off error, so the condition number tends to be a more reliable test.

{ Source: Heath [?], Review Question 2.62, p. 95 }

Q3a–9[91]. Which of the following matrices is ill-conditioned?

(A) $\begin{bmatrix} 10^{10} & 0 \\ 0 & 10^{-10} \end{bmatrix}$

(B) $\begin{bmatrix} 10^{10} & 0 \\ 0 & 10^{10} \end{bmatrix}$

(C) $\begin{bmatrix} 10^{-10} & 0 \\ 0 & 10^{-10} \end{bmatrix}$

(D) $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

(E) All of the above

Answer: (A). This matrix has condition number (in the 1-norm)

$$\|A\|_1 \cdot \|A^{-1}\|_1 = \left\| \begin{matrix} 10^{10} & 0 \\ 0 & 10^{-10} \end{matrix} \right\|_1 \cdot \left\| \begin{matrix} 10^{-10} & 0 \\ 0 & 10^{10} \end{matrix} \right\|_1 = 10^{10} \cdot 10^{10} = 10^{20} \quad \textit{(HUGE)}$$

But response (D) is also correct since that matrix is singular and so has condition number ∞*. The matrices in (B) and (C) are multiples of the identity matrix, so they are perfectly conditioned.*

{ Source: Heath [?], Review Question 2.61, p. 95 }

Q3a–10[92]. You are given a system of equations containing a real parameter a:

$$\begin{aligned} 2x + 3y &= 5 \\ 4x + ay &= 8 \end{aligned}$$

Which of these statements is TRUE?

(A) For $a = 6$, the system has no solution.

(B) For $a = 6$, the system has infinitely many solutions.

(C) The system has a unique solution for any real value of a.

(D) The two lines intersect when $a = 6$.

Answer: (A).

{ Source: MAH }

Q3a–11[93]. What are the eigenvalues of the matrix

$$\begin{bmatrix} 1 & 4 & 6 \\ 0 & 2 & 5 \\ 0 & 0 & 3 \end{bmatrix} \ ?$$

(A) 1, 2, 3

(B) 1, 4, 6

(C) 1, 0, 0

(D) 4, 5, 6

Answer: (A).

{ Source: MAH }

Q3a–12[94]. Which of the following matrices is strictly diagonally dominant, and hence invertible?

I. $\begin{bmatrix} -6 & 0 & 3 \\ 7 & 8 & -2 \\ 1 & -1 & 2 \end{bmatrix}$ II. $\begin{bmatrix} -6 & 0 & 3 \\ 5 & 8 & -2 \\ 1 & -1 & 2 \end{bmatrix}$ III. $\begin{bmatrix} -6 & 0 & 3 \\ 5 & 8 & -2 \\ 1 & -1 & 3 \end{bmatrix}$ IV. $\begin{bmatrix} -6 & 0 & 3 \\ 1 & 2 & -2 \\ 1 & -1 & 8 \end{bmatrix}$

(A) II

(B) III

(C) II and III

(D) IV

(E) All are strictly diagonally-dominant

Answer: (A).

{ Source: JMS }

Q3a–13[95]. *True or False:* It's a well-known fact that every strictly diagonally-dominant matrix is nonsingular. Is it also true that every nonsingular matrix is strictly diagonally-dominant?

Answer: FALSE. A simple counterexample is the matrix $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ *(the permuted identity matrix). This is clearly invertible* ($A^{-1} = A$) *but not diagonally-dominant (DD). This brings home an important point: checking for DD isa really easy test to determine whether a matrix is nonsingular, but it doesn't work for all nonsingular matrices (that is, some matrices that aren't DD can still be invertible, like A above).*

{ Source: JMS }

Q3a–14[96]. Find $w = \begin{bmatrix} -1 \\ 4 \\ 12 \end{bmatrix}$ as a linear combination of $u = \begin{bmatrix} -2 \\ 2 \\ 3 \end{bmatrix}$ and $v = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$.

(A) $w = 2u + 3v$

(B) $w = 2u - 3v$

(C) $w = u + 3v$

(D) $w = 2u + v$

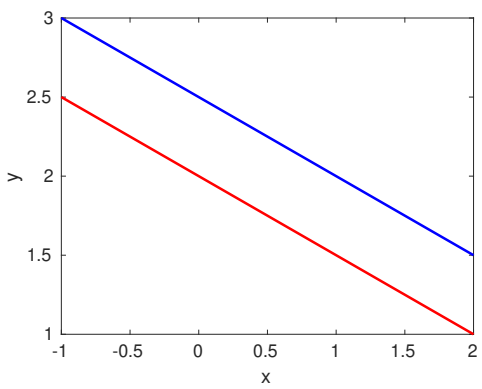
Answer: (A). Check: $\begin{bmatrix} -1 \\ 4 \\ 12 \end{bmatrix} = 2 \begin{bmatrix} -2 \\ 2 \\ 3 \end{bmatrix} + 3 \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$.

{ Source: MAH }

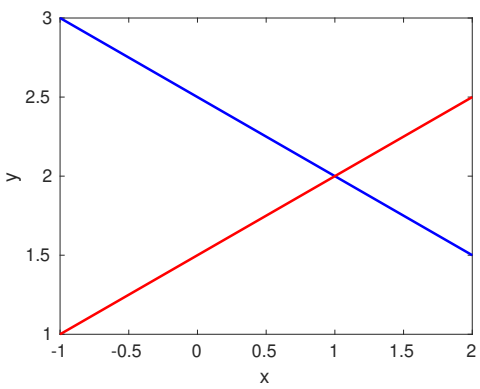
Q3a–15[97]. Which of the graphs below represents the following system of equations?

$$x + 2y = 5$$
$$2x + 4y = 10$$

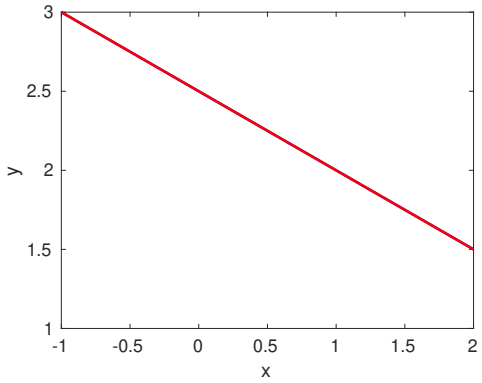
(A)



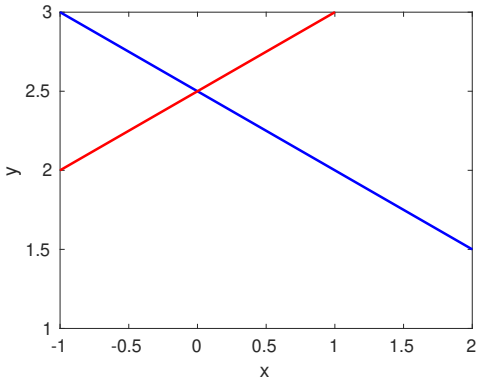
(C)



(B)



(D)



Answer: (B). The second equation is simply the first multiplied by 2, so these are just two equations for the same line (both with negative slope).

{ Source: MAH }

Q3a–16[98]. What is the solution to the following system of equations?

$$-3x + 6y = 3$$
$$2x - 4y = 4$$

(A) $x = -1$ and $y = 0$

(B) $x = 3$ and $y = 2$

(C) $x = 3$ and $y = \frac{1}{2}$

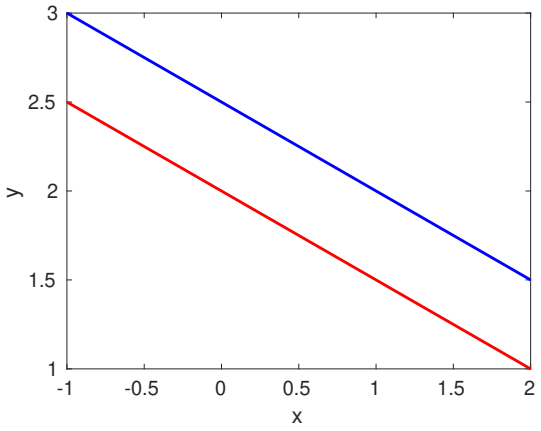
(D) There are no solutions

(E) There are an infinite number of solutions

Answer: (D). The equations can be scaled to get $x - 2y = -1$ and $x - 2y = 2$, which are inconsisent with each other.

{ Source: JMS }

Q3a–17[99]. Which of the following linear systems is depicted in the graph?



(A) $x + 2y = 5, \quad 2x + 4y = 10$

(B) $x + 2y = 5, \quad 2x + 4y = 8$

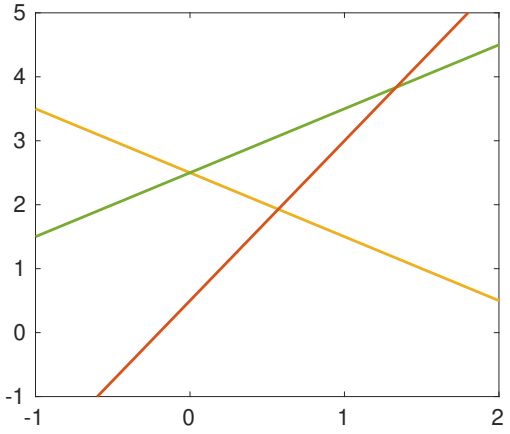
(C) $x + 2y = 5, \quad 2x - 4y = 10$

(D) $x - 2y = 5, \quad 2x + 4y = 8$

Answer: (B). Both lines in the plot have the same slope ($-\frac{1}{2}$), so we need to choose between (A) and (B). Option (A) gives two identical lines and so the correct answer must be (B).

{ Source: MAH }

Q3a–18[100]. A system of three linear equations in two unknowns is pictured in the plot. How many solutions does this system have?



(A) 0

(B) 1

(C) 3

(D) Infinitely many

Answer: (A). There is no common intersection point and so this system has no solution.

{ Source: MathQuest [?], Linear Algebra, Systems of Equations }

Q3a–19[101]. *True or False:* For any $n \times n$ matrix A, the equation $Ax = 0$ has the unique solution $x = 0$.

Answer: FALSE. The equation $Ax = 0$ has $x = 0$ as a unique solution if and only if A is nonsingular, which is equivalent to saying that A^{-1} exists or $\det(A) \neq 0$.

{ Source: MAH }

Q3a–20[102]. Suppose that both A and B are $n \times n$ matrices. Which of the following statement is FALSE?

(A) $\det(A+B) = \det(A) + \det(B)$

(B) $\det(AB) = \det(A)\det(B)$

(C) $\det(kA) = k^n \det(A)$

(D) None of the above

Answer: (A).

{ Source: MAH }

Q3a–21[103]. *True or False:* If A is any $n \times n$ nonsingular matrix, then $\text{cond}(A) = \text{cond}(A^{-1})$.

Answer: TRUE. $\text{cond}(A^{-1}) = ||A^{-1}|| \cdot ||(A^{-1})^{-1}|| = ||A^{-1}|| \cdot ||A|| = \text{cond}(A)$

{ Source: Heath [?], Review Question 2.25, p. 93 }

3b. Gaussian Elimination and Pivoting

Q3b–1[104]. *Fill in the blank:* The goal of the row-reduction phase in the Gaussian elimination algorithm is to convert the coefficient matrix into a _____ matrix.

(A) diagonal

(B) identity

(C) lower triangular

(D) upper triangular

Answer: (D).

{ Source: JMS }

Q3b–2[105]. If A is a singular matrix, then which of the following statements is FALSE?

(A) $\det(A) = 0$

(B) A zero entry arises in the pivot position as Gaussian elimination with partial pivoting is applied to A

(C) $Ax = 0$ has only the trivial solution $x = 0$

(D) $\text{cond}(A) = \infty$

Answer: (B).

{ Source: JMS }

Q3b–3[106]. When solving an upper or lower triangular system of size $n \times n$, the computational cost (measured in terms of multiplication and division operations) is roughly equal to ...

(A) n^2

(B) $\frac{1}{2}\, n^2$

(C) n^3

(D) n

Answer: (A).

{ Source: JMS }

Q3b–4[107]. Which of the following matrices CANNOT be obtained from

$$A = \begin{bmatrix} 2 & 1 & 3 & 1 \\ 0 & 1 & 3 & 4 \\ 1 & 2 & 0 & 4 \end{bmatrix}$$

using elementary row operations?

(A) $\begin{bmatrix} 2 & 4 & 0 & 8 \\ 0 & 1 & 3 & 4 \\ 2 & 1 & 3 & 1 \end{bmatrix}$

(B) $\begin{bmatrix} 2 & 1 & 3 & 1 \\ 0 & 1 & 3 & 4 \\ 1 & 3 & 3 & 8 \end{bmatrix}$

(C) $\begin{bmatrix} 1 & 2 & 3 & 1 \\ 1 & 0 & 3 & 4 \\ 2 & 1 & 0 & 4 \end{bmatrix}$

Answer: (C). Response (A) results from $r_3 \leftarrow 2r_3$ followed by the row swap $r_1 \leftrightarrow r_3$. Response (B) results from $r_3 \leftarrow r_2 + r_3$. Response (C) involves a column swap.

{ Source: MathQuest [?], Gaussian elimination }

Q3b–5[108]. Which of the following operations on an augmented matrix could change the solution of the corresponding linear system?

(A) Interchanging two rows

(B) Multiplying one row by any constant

(C) Adding one row to another

(D) None of the above

Answer: (B). Because every response seems to correspond to a row operation, it's very easy to make the mistake of choosing (D). However, response (B) could be an invalid row operation if the multiple is zero. So response (D) is only correct as long as it's understood that the constant multiple is non-zero.

{ Source: MathQuest [?], Gaussian elimination }

Q3b–6[109]. What is the value of α so that the linear system represented by the following augmented matrix has infinitely many solutions?

$$\left[\begin{array}{cc|c} 2 & 6 & 8 \\ 1 & \alpha & 4 \end{array}\right]$$

(A) $\alpha = 0$

(B) $\alpha = 2$

(C) $\alpha = 3$

(D) $\alpha = 4$

(E) There is always a unique solution

Answer: (C). The two rows are linearly independent as long as $\alpha \neq 3$, in which case row 2 is multiple of row 1 and the system becomes underdetermined.

{ Source: MathQuest [?], Gaussian elimination }

Q3b–7[110]. Let R be the row-reduced echelon form of an $n \times n$ matrix A. Then ...

(A) R is the identity

(B) R has at least one row of zeroes

(C) Neither (A) nor (B)

(D) Both (A) and (B)

(E) The answer depends on the matrix A

Answer: (E).

{ Source: MathQuest [?], Gaussian elimination }

Q3b–8[111]. When solving a linear system of size $n \times n$ using Gaussian elimination with partial pivoting, the computational cost (measured in terms of multiplication and division operations) of the backward substitution step is ...

(A) $O(n)$
(B) $O(n^{3/2})$
(C) $O(n^2)$
(D) $O(n^3)$

Answer: (C).

{ Source: JMS }

Q3b–9[112]. When computing the LU factorization of an $n \times n$ matrix, the computational cost (measured in terms of multiplication and division operations) is ...

(A) $O(n)$
(B) $O(n^{3/2})$
(C) $O(n^2)$
(D) $O(n^3)$

Answer: (D).

{ Source: JMS }

Q3b–10[113]. The matrix B is singular if ...

(A) B is not square
(B) Gaussian elimination with partial pivoting fails on B
(C) B is its own inverse
(D) B has no inverse

Answer: (D).

{ Source: JMS }

Q3b–11[114]. If you have to solve $Ax = b$ many times for different vectors b but the same matrix A, it is best to ...

(A) compute the inverse of A
(B) determine the LU decomposition once, then apply forward/backward substitution for the different b's
(C) use an iterative method
(D) use Gaussian elimination with partial pivoting

Answer: (B).

{ Source: JMS }

Q3b–12[115]. *True or False:* If a linear system is well-conditioned, then pivoting is unnecessary in Gaussian elimination.

Answer: FALSE. A simple counterexample is given by the matrix

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

This is a permutation of the 3×3 *identity matrix so that* $\text{cond}(A) = 1$*, which is perfectly well-conditioned. However, Gaussian elimination will fail without partial pivoting, because the first pivot element is* $a_{11} = 0$*.*

{ Source: Heath [?], Review Question 2.14, p. 92 }

Q3b–13[116]. Consider the matrix

$$A = \begin{bmatrix} 4 & -8 & 1 \\ 6 & 5 & 7 \\ 0 & -10 & -3 \end{bmatrix}$$

whose LU factorization we want to compute using Gaussian elimination. What will the initial pivot element be without pivoting, and with partial pivoting?

(A) 0 (no pivoting), 6 (partial pivoting)

(B) 4 (no pivoting), 0 (partial pivoting)

(C) 4 (no pivoting), 6 (partial pivoting)

Answer: (C).

{ Source: Heath [?], adapted from Review Question 2.39, p. 93 }

Q3b–14[117]. You have a system of three linear equations with three unknowns. If you perform Gaussian elimination and obtain the row-reduced echelon form

$$\left[\begin{array}{ccc|c} 1 & -2 & 4 & 6 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 3 & 0 \end{array}\right]$$

then the system has ...

(A) a unique solution

(B) no solution

(C) infinitely many solutions

(D) more than one solution

Answer: (A). The row-reduced matrix is upper triangular with nonzero diagonal entries.

{ Source: MAH }

Q3b–15[118]. You have a system of three linear equations with three unknowns. If you perform Gaussian elimination and obtain the row-reduced echelon form

$$\left[\begin{array}{ccc|c} 1 & -2 & 4 & 6 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & 3 \end{array}\right]$$

then the system has ...

(A) a unique solution

(B) no solution

(C) infinitely many solutions

(D) more than one solution

Answer: (B). The last equation reads "$0 = 3$" which is a contradiction.

{ Source: MAH }

Q3b–16[119]. You have a system of three linear equations with three unknowns. If you perform Gaussian elimination and obtain the reduced row echelon form

$$\left[\begin{array}{ccc|c} 1 & -2 & 4 & 6 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

then the system has ...

(A) no solution

(B) a unique solution

(C) more than one solution

(D) infinitely many solutions

Answer: (D). The last equation reads "$0 = 0$" so x_3 can be any real number. Strictly (C) is also correct, but (D) is the most accurate answer.

{ Source: MAH }

Q3b–17[120]. Suppose that a square matrix A is perfectly well-conditioned, meaning that $\text{cond}(A) = 1$. Which of the following matrices shares this same property?

(A) cA where c is any nonzero scalar

(B) DA where D is any nonsingular diagonal matrix

(C) PA where P is any permutation matrix

(D) A^{-1}, the inverse of A

(E) A^T, the transpose of A

Answer: (A). In fact, all of the responses are correct.

{ Source: Heath [?], Review Question 2.58, p. 94 }

Q3b–18[121]. *True or False:* Every nonsingular $n \times n$ matrix A can be written as a product $A = LU$.

Answer: FALSE. For example, $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ *is nonsingular because* $\det(A) = -1 \neq 0$. *Any LU factorization of A must have the form*

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ a & 1 \end{bmatrix} \begin{bmatrix} b & c \\ 0 & d \end{bmatrix} = \begin{bmatrix} b & c \\ ab & ac+d \end{bmatrix}.$$

Equating coefficients in the first column yields $b = 0$ and $ab = 1$, which has no solution. So this A has no LU decomposition even though it's nonsingular. A simpler way to recognize this is that very first step of the LU algorithm fails when it encounters a zero in the $1,1$ pivot entry.

{ Source: MAH }

Q3b–19[122]. If $A = PI$ is some permutation P of the identity matrix, then Gaussian elimination yields the following L and U factors:

(A) $L = U = I$

(B) $L = U = PI$

(C) $L = I,\ U = 0$

(D) The factors depend on the permutation matrix P

Answer: (A). If P is a permutation matrix then its inverse is another permutation matrix, $P^{-1} = \widetilde{P}$, and so $\widetilde{P}A = I$. This is already in row-reduced form, and so GE with partial pivoting yields the LU factorization of the same matrix $\widetilde{P}A = LU = I$. The factors are clearly $L = U = I$.

{ Source: JMS }

Q3b–20[123]. *True or False:* The *LU* factorization is unique, so the output from the following Matlab code must mean that there is a bug in the built-in function "`lu`":

```
>> L = [1 0 0; 2 1 0; 3 1 1];
>> U = [2 0 1; 0 2 1; 0 0 2];
>> [L2, U2] = lu(L*U)

L2 = 0.3333  -1.0000   1.0000
     0.6667   1.0000   0
     1.0000   0        0

U2 = 6.0000   2.0000   6.0000
     0        0.6667  -1.0000
     0        0       -2.0000
```

Answer: FALSE. The LU decomposition computed by Matlab involves at least one row swap, so that the L and U factors are different.

{ Source: JMS }

Q3b–21[124]. You perform Gaussian elimination with partial pivoting to solve a linear system on a single-precision floating point computer with roughly 7 decimal digits of accuracy. If the coefficient matrix has a condition number of 10^3 and the the matrix and right hand side are both measured to within full machine precision, about how many digits of accuracy would you expect in the approximate solution?

(A) 2
(B) 3
(C) 4
(D) 7

Answer: (C). The error estimate tells us that $\begin{pmatrix} \textit{relative} \\ \textit{error} \end{pmatrix} \leqslant \begin{pmatrix} \textit{condition} \\ \textit{number} \end{pmatrix} \cdot \begin{pmatrix} \textit{error} \\ \textit{in data} \end{pmatrix} \leqslant 10^{-7} \cdot 10^3 = 10^{-4}$*, which corresponds to 4 decimal digits. This is why we compute in double precision!!*

{ Source: Heath [?], Review Question 2.64, p. 95 }

Q3b–22[125]. Suppose you are solving a linear system $Ax = b$ on a computer with 12 decimal digits of floating-point precision, and that the matrix and right hand side are correct to within full machine precision. Roughly how large can the condition number of the matrix A be before the computed solution x contains NO significant digits?

(A) 10^{10}
(B) 10^{12}
(C) 10^{16}
(D) 10^{22}

Answer: (B). The error estimate tells us that $\frac{\|x-\hat{x}\|}{\|x\|} \leqslant \text{cond}(A)\,\frac{\|r\|}{\|b\|}$*. If the relative solution error is* $O(1) = O(10^0)$ *(no digits of accuracy) and the "data error" (*$\|r\|/\|b\|$*) is* $O(\varepsilon_{_M}) \approx 10^{-12}$*, then the condition number must be* $O(10^{12})$*.*

{ Source: Heath [?], Review Question 2.65, p. 95 }

Q3b–23[126]. You have a system of linear equations $Ax = b$ with $\|A\| = 250$ and $\|A^{-1}\| = 40$ that you solve using Gaussian elimination on a computer with machine epsilon $\varepsilon_{_M} = 1.19 \times 10^{-7}$. What is the largest number of significant digits that you can trust in the solution?

(A) 1

(B) 2

(C) 3

(D) 4

Answer: (D).

{ Source: Holistic Numerical Methods [?] }

Q3b–24[127]. You have a 100×100 matrix A and your computer is able to solve the linear system $Ax = b$ in exactly one minute using the LU factorization. Roughly how much of that minute is spent computing the factorization $A = LU$?

(A) 30 secs

(B) 45 secs

(C) 53 secs

(D) 59 secs

Answer: (D). From the lecture notes, the ratio of time spent on the row-reduction to the total is

$$\frac{\mathcal{R}(n)}{\mathcal{T}(n)} = \frac{2n^3 + 3n^2 - 5n}{2n^3 + 6n^2 - 2n} \approx \frac{2 + \frac{3}{n}}{2 + \frac{6}{n}}$$

When $n = 100$ this ratio is $\frac{2.03}{2.06} \approx 0.99$ which corresponds to roughly 59 seconds.

{ Source: JMS }

Q3b–25[128]. Suppose that your computer can solve 100 problems of the form $Ux = c$ in 1 second, where U is a 50×50 upper triangular matrix. Roughly how long will it take to solve a single 500×500 problem of the form $Ax = b$ where A is a full matrix?

(A) 10 seconds

(B) 1 minute

(C) 5 minutes, 30 seconds

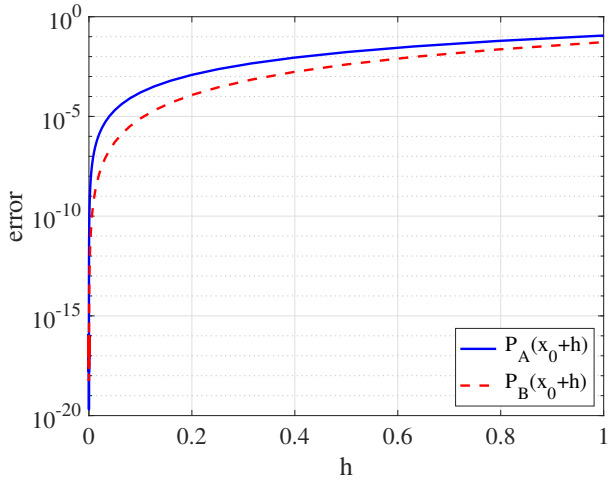
(D) 5 hours

Answer: (C). The cost of solving $Ux = c$ is a backward substitution, which we know has cost with leading order term $\frac{n^2}{2}$ so the total cost is $\frac{100(50^2)}{2} = 50^3$. The full system has cost $\frac{n^3}{3} = \frac{500^3}{3}$ which is a fraction $\frac{1000}{3}$ larger. That translates into about 333 seconds. A rough order of magnitude estimate (missing the constants) yields a ratio of $\frac{500^3}{100 \cdot 50^2} = 500$, and so choice (C) is still the closest answer.

{ Source: JMS }

3c. Plotting Power Laws

Q3c–1[129]. A function is approximated near some given point x_0 using two Taylor polynomials, $P_A(x)$ and $P_B(x)$. On the right is a plot of the absolute error in these two polynomials as a function of h, where both are evaluated at the nearby point $x_0 + h$. What is the degree of the two Taylor polynomials?



(A) P_A is degree 2, P_B is degree 3

(B) P_A is degree 3, P_B is degree 4

(C) P_A and P_B are both degree 3

(D) The degree cannot be determined from this plot

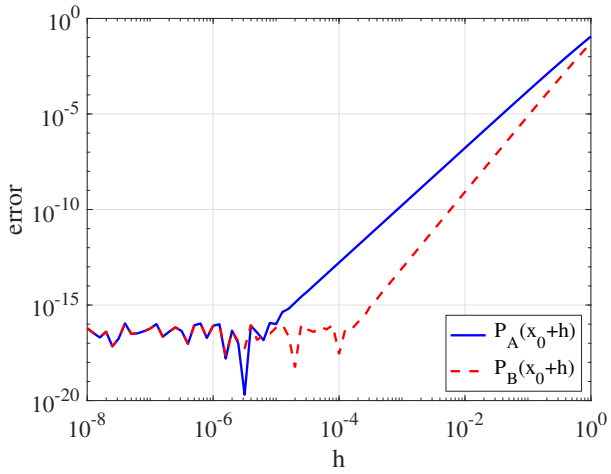
Answer: (D). The remainder/error term for a Taylor polynomial of degree n takes the form

$$R_n = \frac{f^{(n+1)}(c)}{(n+1)!}\,h^{n+1} \quad \Longrightarrow \quad \log R_n \sim (n+1)\log h + \textit{constant}$$

and so the degree can only be determined by plotting the data on a log-log scale. This is a semi-log plot and so the nature of the polynomials can't be determined.

{ Source: JMS }

Q3c–2[130]. A function is approximated near some given point x_0 using two Taylor polynomials, $P_A(x)$ and $P_B(x)$. On the right is a plot of the absolute error in these two polynomials as a function of h, where both are evaluated at the nearby point $x_0 + h$. What is the degree of the two Taylor polynomials?



(A) P_A is degree 2, P_B is degree 3

(B) P_A is degree 3, P_B is degree 4

(C) P_A and P_B are both degree 3

(D) The degree cannot be determined from this plot

Answer: (A). The remainder/error term for a Taylor polynomial of degree n takes the form

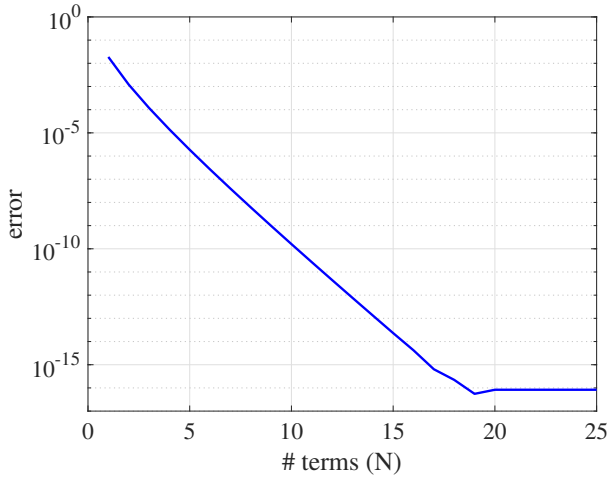
$$R_n = \frac{f^{(n+1)}(c)}{(n+1)!}\,h^{n+1} \quad \Longrightarrow \quad \log R_n \sim (n+1)\log h + \textit{constant}$$

On this log-log plot, the errors behave linearly with slopes 3 and 4, for P_A and P_B respectively. The slope is $n+1$, one more than the degree.

Notice the oscillations appearing on the left half of the plot for small h – these are due to floating-point round-off error that dominates when the truncation error in the Taylor polynomials approaches the value of machine epsilon $\varepsilon_M \approx 2 \times 10^{-16}$.

{ Source: JMS }

Q3c–3[131]. A function is approximated with Taylor polynomials of degrees 1 through 25, and the error in each approximation is plotted against the degree N of the polynomial. What happens when $N \gtrapprox 18$?



(A) The Taylor remainder term blows up when N gets too large.

(B) The accuracy is limited by the floating point machine epsilon.

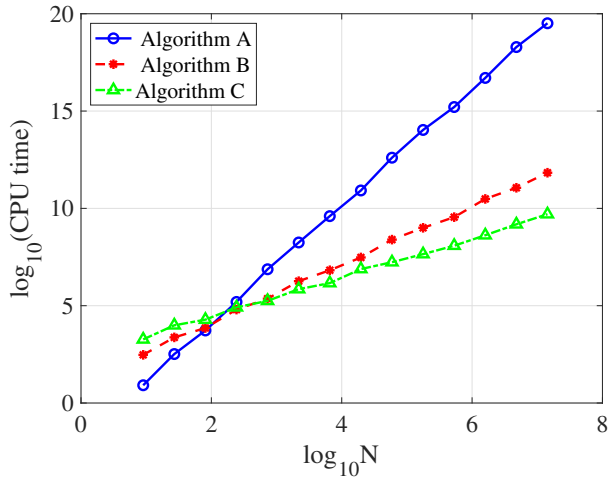
(C) Subtractive cancellation errors dominate when too many terms of differing signs are added to the polynomial.

Answer: (B). As N increases, the error in the Taylor approximation gets smaller until it reaches machine epsilon, $\varepsilon_M \approx 2\times10^{-16}$, after which no more improvement in accuracy is possible. This is not really a question about power laws, but the semi-log plot does clearly illustrate the linear dependence of error on N (for fixed h) that we expect from:

$$R_N(h) = \frac{f^{(N+1)}(c)}{(N+1)!}\,h^{N+1} \quad \Longrightarrow \quad \log R_N \sim (N+1)\log h + \textit{constant}$$

{ Source: JMS }

Q3c–4[132]. Suppose you are comparing three different algorithms and you use all three to solve the same problem for increasing values of problem size N. A plot of the CPU time required for each algorithm is shown on the right. Which algorithm has a cost that scales as $O(N^{3/2})$?



(A) Algorithm A

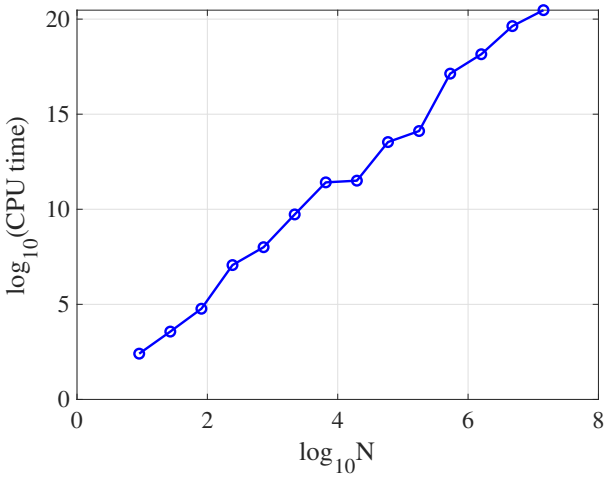
(B) Algorithm B

(C) Algorithm C

Answer: (B). This plot has a log-log scale. If cost scales like $C = \alpha N^{3/2}$, then $\log C = \log\alpha + \frac{3}{2}\log N$. So we are looking for the straight line with slope $\frac{3}{2}$, which is (B).

{ Source: JMS }

Q3c–5[133]. You have a code that you are using to compute several problems of increasing size N. A plot of the CPU time required for each computation is shown on the right. What is the best estimate of the leading-order cost of your algorithm as a function of N?



(A) $O(N)$
(B) $O(N^{3/2})$
(C) $O(N^2)$
(D) $O(N^3)$

Answer: (D). This plot has a log-log scale. If cost scales like $C = \alpha N^p$, then $\log C = \log \alpha + p \log N$ and should appear as a straight line with slope p. The plotted curve is roughly a straight line with slope $\approx \frac{20-2}{7-1} = 3$.

{ Source: JMS }

Q3c–6[134]. You write a code for an iterative algorithm that you believe converges quadratically. To test whether you have implemented the algorithm correctly, you plot the relative error in each step, $R_k = |x_k - x^*|/|x^*|$, as a function of the iteration number k. Using which type of plot is it easiest to recognize the quadratic convergence?

(A) A linear plot showing R_k versus k.
(B) A semi-log plot showing $\log R_k$ versus k.
(C) A semi-log plot showing R_k versus $\log k$.
(D) A log-log plot showing $\log R_k$ versus $\log k$.
(E) A double-log plot showing $\log(\log R_k)$ versus k.

Answer: (E). Because the error in a quadratic method behaves like $R_k \sim R_0^{2^k}$, then $\log R_k \sim (const) \cdot 2^k$ and $\log(\log R_k) \sim (const) \cdot k$. It is easiest to recognize a straight line and so you should plot $\log(\log R_k)$ versus k.

{ Source: JMS }

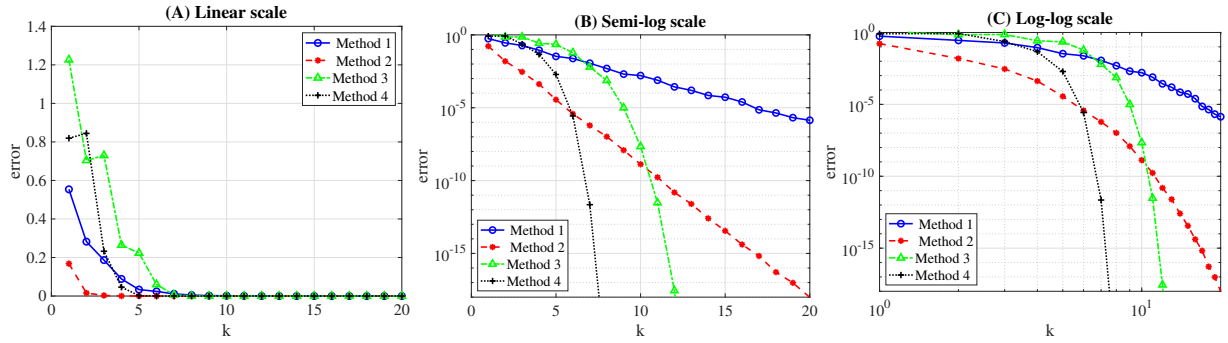
Q3c–7[135]. You write a code for an iterative algorithm that you believe converges superlinearly. To test whether you have implemented the algorithm correctly, you plot the relative error in each step $R_k = |x_k - x^*|/|x^*|$ as a function of the iteration number k. Using which type of plot is it easiest to recognize the superlinear convergence?

(A) A linear plot showing R_k versus k.
(B) A semi-log plot showing $\log R_k$ versus k.
(C) A semi-log plot showing R_k versus $\log k$.
(D) A log-log plot showing $\log R_k$ versus $\log k$.

Answer: (B). Because the error in a linearly convergent method behaves like $R_k \sim \alpha^k R_0$, then $\log R_k \sim k \log \alpha + \log R_0$ which is a linear function of k. So the easiest way to test for superlinear convergence is to plot $\log R_k$ versus k and then look for downward curvature, which indicates a faster decay to zero than linear).

{ Source: JMS }

Q3c–8[136]. You use four different iterative methods to obtain a sequence of approximate solution values x_k, for $k = 1, 2, \ldots, 20$. Below are shown graphs of the absolute solution error, plotted versus k using three different axis scales. Which plot gives you the most useful information about the convergence of the four methods?



(A) Linear scale
(B) Semi-log scale
(C) Log-log scale

Answer: (B). You want to choose an axis scaling where the behaviour of the method is represented as a straight line, since this linear dependence is easy to recognize visually. The semi-log plot (B) is the only one that contains linear features (for Methods 1 & 2). We know that a linearly convergent method with $E_k \sim cE_0^k$ appears as a straight line on a semi-log scale. So Methods 1 & 2 are linearly convergent, while Methods 3 & 4 are superlinear (faster than linear, but we can't say more than that.

{ Source: JMS }

3d. Iterative Methods for Sparse Systems

Q3d–1[137]. Consider the $n \times n$ matrix

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 4 & 8 & \cdots & 2^{n-1} \\ 1 & 3 & 9 & 27 & \cdots & 3^{n-1} \\ 1 & 4 & 16 & 64 & \cdots & 4^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & n & n^2 & n^3 & \cdots & n^{n-1} \end{bmatrix}$$

and suppose that for some given n you know that $\det(A) = 2.49\text{e+}07$, $\|A\|_1 = 1.85\text{e+}05$ and $\text{cond}_1(A) = 4.14\text{e+}07$. Based on this information, which is the best method for solving the linear system $Ax = b$?

(A) Gaussian elimination
(B) Gaussian elimination with partial pivoting
(C) Jacobi's iteration
(D) Gauss-Seidel iteration

Answer: (B). This matrix is not sparse (far from it, since there's not a single zero entry!) and so iterative methods are a bad idea and Gaussian elimination is the only option. The condition number is large, indicating that A is ill-conditioned and partial pivoting is necessary.

{ Source: JMS }

Q3d–2[138]. Consider the $n \times n$ matrix

$$A = \begin{bmatrix} -4 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -4 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -4 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -4 & 1 \\ 0 & \cdots & 0 & 0 & 1 & -4 \end{bmatrix}$$

and suppose that for some given n you know that $\det(A) = -10864$, $\|A\|_1 = 6$ and $\text{cond}_1(A) = 2.97$. Based on this information, identify the best method for solving the linear system $Ax = b$.

(A) Gaussian elimination

(B) Gaussian elimination with partial pivoting

(C) Jacobi's iteration

(D) Gauss-Seidel iteration

Answer: (D). This is a sparse system and so Gauss-Seidel is the "obvious" answer. However, we've seen in lectures that GE+PP can row-reduce this tridiagonal matrix with no fill-in and so it is a very efficient method – so (B) is likely the best answer.

Q3d–3[139]. If the $n \times n$ matrix A is ill-conditioned (that is, A has a very large condition number) then ...

(A) Solving $Ax = b$ accurately is difficult using the LU-decomposition, but iterative methods (Jacobi, Gauss-Seidel) would not have a problem.

(B) Solving $Ax = b$ accurately with iterative methods (Jacobi, Gauss-Seidel) would be difficult, but LU-decomposition with partial pivoting would not have a problem.

(C) Solving $Ax = b$ accurately will be challenging regardless of the method we use.

Answer: (C).

{ Source: JMS }

Q3d–4[140]. *True or False:* You are given the 3×3 linear system in augmented matrix form

$$\left[\begin{array}{ccc|c} 5 & 0 & -1 & 2 \\ 0 & 0 & 2 & 4 \\ 0 & 1 & 0 & 2 \end{array}\right]$$

Because the diagonal of the coefficient matrix contains zero entries, it is not possible to apply the Jacobi iterative method.

Answer: FALSE. Swapping rows 2 and 3 converts the matrix to upper triangular form with non-zeroes on the diagonal, and Jacobi can then be applied.

{ Source: JMS }

Q3d–5[141]. You are given the 3×3 linear system in augmented matrix form

$$\left[\begin{array}{ccc|c} 5 & 0 & -1 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 2 & 4 \end{array}\right]$$

Starting from an initial guess of $x_0 = [1, 1, 1]^T$, what is the result of the first iteration of the Jacobi method?

(A) $x_1 = \left[\frac{3}{5}, 2, 2\right]^T$

(B) $x_1 = \left[\frac{2}{5}, 2, 1\right]^T$

(C) $x_1 = \left[\frac{3}{5}, 2, 5\right]^T$

Answer: (A).
$x_1 = \frac{1}{5}(2 - 0 \cdot 1 + 1 \cdot 1) = \frac{3}{5}$

$x_2 = \frac{1}{1}(2 - 0 \cdot 1 - 0 \cdot 1) = 2$

$x_3 = \frac{1}{2}(4 - 0 \cdot 1 - 0 \cdot 1) = 2$

Interesting observation: Jacobi applied to any upper triangular $n \times n$ matrix converges to the exact solution in at most n steps AND it's equivalent to doing backward substitution ... although it can do a lot of extra work to get there!

{ Source: JMS }

Q3d–6[142]. You are given the 3×3 linear system in augmented matrix form

$$\left[\begin{array}{ccc|c} 5 & 0 & -1 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 2 & 4 \end{array}\right]$$

Starting from an initial guess of $x_0 = [1,1,1]^T$, what is the result of the first iteration of the Gauss-Seidel method?

(A) $x_1 = \left[\frac{3}{5}, 2, 2\right]^T$

(B) $x_1 = \left[\frac{2}{5}, 2, 1\right]^T$

(C) $x_1 = \left[\frac{3}{5}, 2, 5\right]^T$

Answer: (A). This is identical to the Jacobi iteration because the lower triangular part of the matrix is 0!

{ Source: JMS }

Q3d–7[143]. You are given the 3×3 linear system in augmented matrix form

$$\left[\begin{array}{ccc|c} 5 & 0 & -1 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & -3 & 2 & 4 \end{array}\right]$$

Starting from an initial guess of $x_0 = [1,1,1]^T$, what is the result of the first iteration of the Gauss-Seidel method?

(A) $x_1 = \left[\frac{3}{5}, 2, 2\right]^T$

(B) $x_1 = \left[\frac{2}{5}, 2, 1\right]^T$

(C) $x_1 = \left[\frac{3}{5}, 2, 5\right]^T$

Answer: (C).
$x_1 = \frac{1}{5}\left(2 - 0 \cdot 1 + 1 \cdot 1\right) = \frac{3}{5}$
$x_2 = \frac{1}{1}\left(2 - 0 \cdot \frac{3}{5} - 0 \cdot 1\right) = 2$
$x_3 = \frac{1}{2}\left(4 - 0 \cdot \frac{3}{5} + 3 \cdot 2\right) = 5$

{ Source: JMS }

Q3d–8[144]. You apply the Gauss-Seidel method to solve a linear system having an iteration matrix $T = -(D+L)^{-1}U$ with spectral radius $\rho(T) = 0.998$. Using the fact that $\log_{10}(0.998) \approx -0.0009$, roughly how many iterations are needed to reduce the error in the initial guess by a factor of 10^{-6}?

(A) 6

(B) 600

(C) 6000

Answer: (C). The error estimate for a matrix iteration takes the form of an inequality for the error E_k in step k: $E_k \leqslant \rho(T)^k E_0$. If the error needs to reduce by a factor of 10^{-6} then

$$\rho(T)^k \leqslant 10^{-6} \quad \Longrightarrow \quad k \cdot \underbrace{\log_{10}(\rho(T))}_{-0.0009} \leqslant -6 \quad \Longrightarrow \quad k \geqslant \frac{-6}{-0.0009} \approx 6000$$

{ Source: JMS }

Q3d–9[145]. Below is the output from the Jacobi and Gauss-Seidel iterations for solving the linear system $Ax = b$, in each case showing the iteration number k and the norm of the solution difference $\|x_k - x_{k-1}\|$:

Method A:		Method B:	
k= 1:	6.073730e+00	k= 1:	3.501795e+00
k= 2:	4.961063e+00	k= 2:	2.710771e+00
k= 3:	4.408955e+00	k= 3:	2.623906e+00
k= 4:	3.969368e+00	k= 4:	2.353607e+00
k= 5:	3.620131e+00	k= 5:	2.294563e+00
⋮	⋮	⋮	⋮
k= 184:	1.366841e-06	k= 354:	1.177728e-06
k= 185:	1.258350e-06	k= 355:	1.130021e-06
k= 186:	1.158471e-06	k= 356:	1.084248e-06
k= 187:	1.066519e-06	k= 357:	1.040328e-06
k= 188:	9.818663e-07	k= 358:	9.981874e-07

Which method corresponds to the Jacobi iteration?

(A) Method A

(B) Method B

Answer: (B). Both converge, but Method A does so in about half the number of iterations. This is characteristic of the Gauss-Seidel method.

{ Source: JMS }

3e. Eigenvalue Problems

[nothing here yet]

4. Function Approximation

4a. Polynomial Interpolation

Q4a–1[146]. *True or False:* Suppose that you have $n+1$ data points (x_0, y_0), (x_1, y_1), $\ldots$, (x_n, y_n). The interpolating polynomial on the given points is unique.

Answer: FALSE. The n^{th} degree polynomial interpolating these points is unique, provided only that the x_i are all distinct.

{ Source: Heath [?], adapted from Review Question 7.3, p. 333 }

Q4a–2[147]. Suppose that you have n data points (x_1, y_1), $\ldots$, (x_n, y_n), with the x_i all distinct. Which of the following statements about the n^{th} degree interpolating polynomial is TRUE?

(A) The interpolating polynomial is unique.

(B) There are infinitely many such interpolating polynomials.

(C) There is no polynomial of this degree that interpolates all n points.

Answer: (B). An n^{th} degree polynomial has $n+1$ coefficients which are constrained by n interpolating conditions. So there remains one degree of freedom.

{ Source: Heath [?], adapted from Review Question 7.3, p. 333 }

Q4a–3[148]. You determine a function $f(x)$ that passes through the points (x_0, y_0), (x_1, y_1), $\ldots$, (x_n, y_n) with $x_0 < x_1 < x_2 \cdots < x_n$. For some other point $x^* \in (x_0, x_n)$, you then use $f(x^*)$ to approximate the value of the actual smooth function that underlies the data. This procedure is called:

(A) interpolation

(B) extrapolation

(C) curve fitting

(D) regression

Answer: (A).

{ Source: Holistic Numerical Methods [?] }

Q4a–4[149]. *Fill in the blank:* Suppose you have $n+1$ data points (x_0, y_0), (x_1, y_1), $\ldots$, (x_n, y_n). A unique polynomial of degree _____ can be found that passes through these points.

(A) n

(B) n or less

(C) $n+1$

(D) $n+1$ or more

Answer: (B). Assuming the x_i are distinct, then there is a unique polynomial of degree exactly n that passes through the $n+1$ data points. However, response (B) is the correct answer because if any of the x_i are repeated, then some of the higher degree coefficients could be zero.

{ Source: Holistic Numerical Methods [?], MC Question_Solution Ch 05.01 Background of Interpolation.pdf }

Q4a–5[150]. How many different polynomials can be found that pass through the points $(1, 2)$ and $(4, 5)$?

(A) 1

(B) 2

(C) 1 or 2

(D) infinitely many

Answer: (D).

{ Source: Holistic Numerical Methods [?], MC Question_Solution Ch 05.01 Background of Interpolation.pdf }

Q4a–6[151]. You are given the point values $(1,1)$, $(2,4)$, $(3,9)$ for some function $f(x)$. You determine an approximation $P(x)$ that passes through the 3 given points, and you then approximate $f(4)$ using $P(4)$. What is this procedure called?

(A) extrapolation

(B) interpolation

(C) curve fitting

(D) none of the above

Answer: (A). Because $x = 4$ lies outside the interval $[1,3]$ covered by the given x–coordinates.

{ Source: Holistic Numerical Methods [?] }

Q4a–7[152]. The following data describes the velocity of a rocket during lift-off as a function of time:

t (s)	0	14	15	20	30	35
$v(t)$ (m/s)	0	227	363	517	603	902

In order to determine the velocity at $t = 25$ s, you decide to use a quadratic polynomial $v(t) = a+bt+ct^2$ to approximate the velocity profile. The most appropriate system of equations for determining the coefficients is ...

(A) $\begin{bmatrix}1 & 14 & 176\\ 1 & 15 & 225\\ 1 & 20 & 400\end{bmatrix}\begin{bmatrix}a\\ b\\ c\end{bmatrix} = \begin{bmatrix}227\\ 363\\ 517\end{bmatrix}$

(B) $\begin{bmatrix}1 & 15 & 225\\ 1 & 20 & 400\\ 1 & 30 & 900\end{bmatrix}\begin{bmatrix}a\\ b\\ c\end{bmatrix} = \begin{bmatrix}363\\ 517\\ 603\end{bmatrix}$

(C) $\begin{bmatrix}1 & 0 & 0\\ 1 & 15 & 225\\ 1 & 30 & 900\end{bmatrix}\begin{bmatrix}a\\ b\\ c\end{bmatrix} = \begin{bmatrix}0\\ 363\\ 603\end{bmatrix}$

(D) $\begin{bmatrix}1 & 20 & 400\\ 1 & 30 & 900\\ 1 & 35 & 1225\end{bmatrix}\begin{bmatrix}a\\ b\\ c\end{bmatrix} = \begin{bmatrix}517\\ 603\\ 902\end{bmatrix}$

Answer: (B). Response (A) corresponds to extrapolation and (C) doesn't use the three data points closest to $t = 25$, and so both will probably be less accurate. Responses (B) and (D) should give comparable accuracy since they are based on intervals that are closest to the interpolation point.

{ Source: JMS }

Q4a–8[153]. *Fill in the blanks:* For the $n+1$ data points (x_0, y_0), (x_1, y_1), ..., (x_n, y_n), the n^{th} degree Lagrange polynomial is $P_n(x) = \sum_{j=0}^{n} y_j L_j(x)$ where $L_j(x) = \prod_{\substack{i=0 \\ i\neq j}}^{n} \frac{(x - x_i)}{(x_j - x_i)}$.

To compute the computational cost, constructing each L_j requires ___ ①___ multiplications and ___ ②___ divisions, and there are ___ ③___ such L_j to compute. Evaluating the interpolating polynomial then requires a further $n+1$ multiplications and n additions. Therefore, assuming all operations involve the same work, the total cost is ___ ④___ $= O(n^2)$ operations.

(A) ① $n+1$, ② $n+1$, ③ $n+1$, ④ $2(n+1)^2 + (2n+1)$

(B) ① n, ② n, ③ n, ④ $2n(n+1)(2n+1)$

(C) ① n, ② n, ③ n, ④ $2n^2 + (2n+1)$

(D) ① n, ② n, ③ $n+1$, ④ $2n(n+1) + (2n+1)$

Answer: (D).

{ Source: JMS, MACM 316 lecture notes }

Q4a–9[154]. For the $n+1$ data points $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, the n^{th} degree Lagrange polynomial is $P_n(x) = \sum_{j=0}^{n} y_j L_j(x)$ where $L_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{(x - x_i)}{(x_j - x_i)}$.

The given partial code calculates a single Lagrange polynomial L_j, where the given x-coordinates are assigned to the variable `xdata`. Provide suitable code to fill the blank spaces labelled (1) and (2).

```
Lj = ...(1)...
for i = 1:n,
   if i ~= j,
      Lj = ...(2)...
   end
end
```

(A) (1) `0.0;` (2) `Lj*(x-xdata(i))/(xdata(j)-xdata(i));`

(B) (1) `1.0;` (2) `Lj*(x-xdata(i))/(xdata(j)-xdata(i));`

(C) (1) `1.0;` (2) `(x-xdata(i))/(xdata(j)-xdata(i));`

(D) (1) `0.0;` (2) `(x-xdata(i))/(xdata(j)-xdata(i));`

Answer: (B).

{ Source: JMS, MACM 316 lecture notes }

Q4a–10[155]. The linear Lagrange polynomial that interpolates the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ is

(A) $P(x) = \dfrac{(x - x_2)}{(x_1 - x_2)} f(x_1) + \dfrac{(x - x_1)}{(x_2 - x_1)} f(x_2)$

(B) $P(x) = \dfrac{(x_1 - x_2)}{(x - x_2)} f(x_1) + \dfrac{(x_2 - x_1)}{(x - x_1)} f(x_2)$

(C) $P(x) = \dfrac{f(x_1) - f(x_2)}{(x_1 - x_2)} x + \dfrac{x_1 f(x_2) - x_2 f(x_1)}{(x_1 - x_2)}$

(D) $P(x) = \dfrac{x}{(x_1 - x_2)} f(x_1) + \dfrac{x}{(x_2 - x_1)} f(x_2)$

Answer: (A).

Q4a–11[156]. This table lists the population of Canada every 5 years from 2000 to 2015:

Year	2000	2005	2010	2015
Population (in millions)	30.69	32.24	34.01	35.83

Using the Lagrange interpolating polynomial to express the population as

$$P(x) = 30.69\, L_0(x) + 32.24\, L_1(x) + 34.01\, L_2(x) + 35.83\, L_3(x),$$

one can approximate the population in any year x between 2000 to 2015. What is $L_0(x)$?

(A) $\dfrac{(x - 2005)(x - 2010)(x - 2015)}{750}$

(B) $\dfrac{(x - 2005)(x - 2010)(x - 2015)}{-750}$

(C) $\dfrac{(x - 2000)(x - 2005)(x - 2010)}{-50}$

(D) $\dfrac{(x - 2000)(x - 2010)(x - 2015)}{-150}$

Answer: (B). This is the only polynomial that satisfies $L_0 = 1$ at $x = 2000$ and $L_0 = 0$ at the other three data points.

{ Source: MAH }

Q4a–12[157]. Fill in the blanks in the divided difference table:

x	y		
1	1		
		ⓐ	
3	9		ⓒ
		ⓑ	
4	16		

(A) ⓐ $= 4$, ⓑ $= 7$, ⓒ $= 1$
(B) ⓐ $= 8$, ⓑ $= 7$, ⓒ $= 3$
(C) ⓐ $= 8$, ⓑ $= 4$, ⓒ $= 3$
(D) ⓐ $= 7$, ⓑ $= 4$, ⓒ $= 3$

Answer: (A). Following the Newton divided difference formula, we have

$$ⓐ = \frac{9-1}{3-1} = 4, \quad ⓑ = \frac{16-9}{4-3} = 7 \quad \textit{and then} \quad ⓒ = \frac{7-4}{4-1} = 1.$$

{ Source: MAH }

Q4a–13[158]. The second degree polynomial

$$P_2(x) = a_1 + a_2(x-3) + a_3(x-3)(x-4)$$

is determined using the given table of Newton divided differences. The correct value of a_2 is ...

x	y			
1	1			
		4		
3	9		1	
		7		0
4	16		1	
		9		
5	25			

(A) $a_2 = 1$
(B) $a_2 = 4$
(C) $a_2 = 7$
(D) $a_2 = 9$

Answer: (C). It's the second diagonal in the table that starts at the point $x = 3$, for which the interpolating polynomial is:

$$P_2(x) = 9 + 7(x-3) + 1(x-3)(x-4).$$

{ Source: Holistic Numerical Methods [?] }

Q4a–14[159]. The second degree polynomial

$$P_2(x) = 4 + 6(x-a) + (x-b)(x-c).$$

is determined using the given table of Newton divided differences. The correct values of the coefficients are ...

x	y			
1	1			
		3		
2	4		1	
		6		0
4	16		1	
		9		
5	25			

(A) $a = 1, \quad b = 2, \quad c = 4$
(B) $a = 2, \quad b = 2, \quad c = 4$
(C) $a = 1, \quad b = 3, \quad c = 1$
(D) $a = 2, \quad b = 4, \quad c = 5$

Answer: (B). These coefficients come from the second diagonal in the table, which starts at the point $x = 2$, so the interpolating polynomial is

$$P_2(x) = 4 + 6(x-2) + (x-2)(x-4).$$

{ Source: MAH }

Q4a–15[160]. Which of the following is a third degree interpolating polynomial for the data given in the Newton divided difference table.

x_i	a_0	a_1	a_2	a_3	a_4
0	0				
		1			
1	1		3		
		7		1	
2	8		6		0
		19		1	
3	27		9		
		37			
4	64				

(A) $P_3(x) = x + 3x(x-1) + x(x-1)(x-2)$
(B) $P_3(x) = 1 + 7(x-1) + 6(x-1)(x-2) + (x-1)(x-2)(x-3)$
(C) $P_3(x) = 64 + 37(x-4) + 9(x-4)(x-3) + (x-4)(x-3)(x-2)$
(D) All of the above

Answer: (D).

{ Source: MAH }

Q4a–16[161]. Consider the interpolating polynomial for $f(x) = x^3$ based on the points $(1,1)$, $(2,8)$, $(3,27)$, and $(4,64)$. Find an upper bound for the interpolation error on $[1,4]$.

(A) $\frac{1}{6}$
(B) $\frac{15}{8}$
(C) $\frac{105}{16}$
(D) 0

Answer: (D). There are four points so the interpolating polynomial is a cubic. Since the original function is a cubic, then $P_3(x) = x^3$, so the error is zero! You could also use the error formula

$$|f(x) - P_3(x)| \leqslant \left| \frac{f^{(4)}(c)}{4!}(x-1)(x-2)(x-3)(x-4) \right| = 0$$

because $f^{(4)}(x) \equiv 0$.

{ Source: MAH }

Q4a–17[162]. Consider the interpolating polynomial for $f(x) = x^3$ based on the points $(0,0)$, $(1,1)$ and $(2,8)$. Find an upper bound for the interpolation error at the point $x = \frac{1}{2}$.

(A) $\frac{1}{6}$
(B) 1
(C) $\frac{15}{8}$
(D) $\frac{3}{8}$

Answer: (D). Since there are three points, the interpolating polynomial is a quadratic for which the error formula gives

$$|f(x) - P_2(x)| \leqslant \left| \frac{f'''(c)}{3!}\, x(x-1)(x-2) \right|$$

$$\implies \left|f(\tfrac{1}{2}) - P_2(\tfrac{1}{2})\right| \leqslant \left|\tfrac{6}{6}\ (\tfrac{1}{2})\ (\tfrac{1}{2}-1)\ (\tfrac{1}{2}-2)\right| = \tfrac{3}{8}.$$

{ Source: MAH }

Q4a–18[163]. How many fourth-degree polynomials pass through the points $(1,1)$, $(2,8)$, $(3,27)$ and $(4,64)$?

Hint: You might be able to use the Newton divided difference table to help you decide.

x	y			
1	1			
		7		
2	8		6	
		19		1
3	27		9	
		37		
4	64			

(A) 1

(B) 2

(C) none

(D) infinitely many

Answer: (D). From the table, it's easy to determine a cubic polynomial that interpolates the four points:

$$P_3(x) = 1 + 7(x-1) + 6(x-1)(x-2) + 1(x-1)(x-2)(x-3)$$

Then, we can generate a fourth-degree polynomial by adding any other point (a,b) which will generate an extra coefficient c in the table. This corresponds to a new polynomial

$$P_4(x) = P_3(x) + c(x-1)(x-2)(x-3)(x-4)$$

that interpolates the original four points, plus (a,b).

{ Source: MAH }

Q4a–19[164]. How many third-degree polynomials pass through the points $(1,1)$, $(2,8)$, $(3,27)$ and $(4,64)$?

Hint: You might be able to use the Newton divided difference table to help you decide.

x	y			
1	1			
		7		
2	8		6	
		19		1
3	27		9	
		37		
4	64			

(A) 1

(B) 2

(C) 3

(D) infinitely many

Answer: (A). The divided difference table gives us a unique third-degree polynomial that can be written in two ways

$$\begin{aligned} P_3(x) &= 1 + 7(x-1) + 6(x-1)(x-2) + 1(x-1)(x-2)(x-3) \\ &= 64 + 37(x-4) + 9(x-4)(x-3) + 1(x-4)(x-3)(x-2) \end{aligned}$$

(expand them both to show they're the same polynomial).

{ Source: MAH }

Q4a–20[165]. Fill in the blanks in the divided difference table:

x	y			
1	1			
		(b)		
2	(a)		-1	
		4		
4	16			

(A) ⓐ $= 8$, ⓑ $= 7$
(B) ⓐ $= 7$, ⓑ $= -\frac{3}{4}$
(C) ⓐ $= 8$, ⓑ $= -3$
(D) ⓐ $= 8$, ⓑ $= -\frac{3}{2}$

Answer: (A).

{ Source: MAH }

4b. Piecewise Polynomial or Spline Interpolation

Q4b–1[166]. Which of the following is NOT a valid reason for choosing piecewise cubic over piecewise linear interpolation?

(A) Piecewise cubics can yield an interpolating function that has continuous derivatives.
(B) Piecewise cubics generally converge to the underlying function faster than linear interpolants when the number of data points is increased.
(C) Piecewise cubics are cheaper to compute.
(D) Piecewise cubics can predict derivative values better than linear interpolants.

Answer: (C). Linear splines are built from polynomials of degree 1, which require many fewer operations to construct and evaluate than cubic splines built with degree 3 polynomials.

{ Source: JMS }

Q4b–2[167]. *True or False:* This piecewise polynomial is a quadratic spline:

$$S(x) = \begin{cases} 0, & \text{if } -1 \leqslant x \leqslant 0 \\ x^2, & \text{if } 0 \leqslant x \leqslant 1 \end{cases}$$

Answer: TRUE. The piecewise functions are both quadratic, and $S(x)$ and $S'(x)$ match at $x = 0$.

{ Source: MAH }

Q4b–3[168]. This piecewise polynomial is NOT a cubic spline:

$$S(x) = \begin{cases} S_0(x) = 1, & \text{if } 0 \leqslant x \leqslant 1 \\ S_1(x) = 1 + (x-1)^2, & \text{if } 1 \leqslant x \leqslant 2 \end{cases}$$

Why not?

(A) $S_0(1) \neq S_1(1)$
(B) $S_0'(1) \neq S_1'(1)$
(C) $S_0''(1) \neq S_1''(1)$
(D) $S_0'''(1) \neq S_1'''(1)$

Answer: (C).

{ Source: MAH }

Q4b–4[169]. *True or False:* This piecewise polynomial cannot be a cubic spline:

$$S(x) = \begin{cases} S_0(x) = 1 + x + x^2 + x^3, & \text{if } 0 \leqslant x \leqslant 1 \\ S_1(x) = 4 + (x-1) + (x-1)^2 + (x-1)^3, & \text{if } 1 \leqslant x \leqslant 2 \end{cases}$$

Answer: TRUE. Although $S_0(1) = S_1(1) = 4$, the derivatives don't match at $x = 1$.

{ Source: MAH }

Q4b–5[170]. In cubic spline interpolation, which spline derivatives must be continuous at interior points?

(A) first derivatives

(B) second derivatives

(C) first and second derivatives

(D) third derivatives

Answer: (C).

{ Source: Holistic Numerical Methods [?] }

Q4b–6[171]. Suppose that

$$S(x) = \begin{cases} S_0(x) = 1 + ax + x^2 - x^3, & \text{if } 0 \leqslant x \leqslant 1 \\ S_1(x) = 1 + b(x-1) - 2(x-1)^2 + (x-1)^3, & \text{if } 1 \leqslant x \leqslant 2 \end{cases}$$

is a clamped cubic spline. What are the clamped end-point slopes?

(A) $S'(0) = 0$ and $S'(2) = -1$

(B) $S'(0) = 1$ and $S'(2) = -2$

(C) $S'(0) = 0$ and $S'(2) = -2$

(D) none of the above

Answer: (C). The matching conditions at $x = 1$ require that

$$\begin{aligned} S_0(1) = S_1(1) \quad &\Longrightarrow \quad 1 + a + 1 - 1 = 1 \quad \Longrightarrow \quad a = 0 \\ S_0'(1) = S_1'(1) \quad &\Longrightarrow \quad a + 2 - 3 = b \quad \Longrightarrow \quad b = -1 \end{aligned}$$

Then $S_0'(0) = a = 0$ and $S_1'(2) = b - 1 = -2$.

{ Source: MAH }

Q4b–7[172]. A cubic spline is defined as

$$S(x) = \begin{cases} S_0(x) = 1 + cx^2 + x^3, & \text{if } 0 \leqslant x \leqslant 1 \\ S_1(x) = 2 + 3(x-1) + 3(x-1)^2 + d(x-1)^3, & \text{if } 1 \leqslant x \leqslant 2 \end{cases}$$

Assuming that natural end-point conditions are used, determine the constants c and d.

(A) $c = 0$ and $d = -1$

(B) $c = 0$ and $d = 1$

(C) $c = 1$ and $d = 1$

(D) $c = 1$ and $d = -1$

Answer: (A). Differentiate to get $S_0'' = 2c + 6x$ and $S_1'' = 6 + 6d(x-1)$. Then the natural end-point conditions $S_0''(0) = 2c = 0$ and $S_1''(2) = 6 + 6d = 0$ are easy to solve.

{ Source: MAH }

Q4b–8[173]. A cubic spline is defined by

$$S(x) = \begin{cases} S_0(x) = 1 + bx + cx^2 - 2x^3, & \text{if } 0 \leqslant x \leqslant 1 \\ S_1(x) = 2 - 3(x-1)^2 + 2(x-1)^3, & \text{if } 1 \leqslant x \leqslant 2 \end{cases}$$

Assuming that periodic end-point conditions are used, determine the constants b and c.

(A) $b = 0$ and $c = 3$

(B) $b = 3$ and $c = 0$

(C) $b = 1$ and $c = 3$

(D) $b = 0$ and $c = 1$

Answer: (A).

{ Source: MAH }

Q4b–9[174]. A cubic spline is defined by

$$S(x) = \begin{cases} S_0(x) = a_0 + b_0 x + c_0 x^2 + d_0 x^3, & \text{if } 0 \leqslant x \leqslant 1 \\ S_1(x) = a_1 + b_1(x-1) + c_1(x-1)^2 + d_1(x-1)^3, & \text{if } 1 \leqslant x \leqslant 2 \\ S_2(x) = a_2 + b_2(x-2) + c_2(x-2)^2 + d_2(x-2)^3, & \text{if } 2 \leqslant x \leqslant 3 \end{cases}$$

Assuming that not-a-knot end-point conditions are used, which of the following statements about the coefficients is TRUE?

(A) $a_0 = a_1 = a_2$

(B) $b_0 = b_1 = b_2$

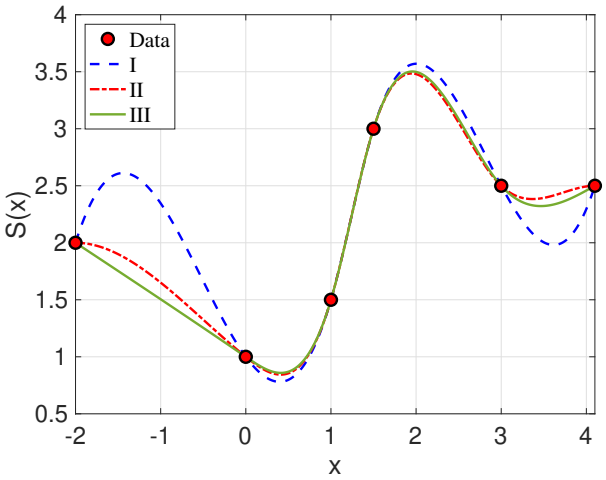
(C) $c_0 = c_1 = c_2$

(D) $d_0 = d_1 = d_2$

Answer: (D). The not-a-knot end-point conditions match $S_0'''(1) = S_1'''(1)$ and $S_1'''(2) = S_2'''(2)$. Using $S_0''' = 6d_0$, $S_1''' = 6d_1$ and $S_2''' = 6d_2$, it's easy to see that $d_0 = d_1 = d_2$.

{ Source: MAH }

Q4b–10[175]. The data shown in the plot is interpolated with three splines using different end-point conditions: natural, not-a-knot and clamped (zero slopes). Identify the spline corresponding to each choice of end-point condition.



(A) I = not-a-knot, II = clamped, III = natural

(B) I = natural, II = not-a-knot, III = clamped

(C) I = natural, II = clamped, III = not-a-knot

Answer: (A). Spline III has the smallest curvature near end-points, so this must be the natural spline. Spline II is the only one with zero slope at both ends, so this is the clamped spline.

{ Source: JMS }

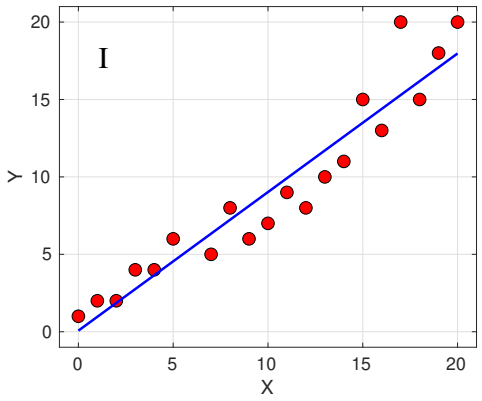
Q4b–11[176]. *True or False:* Suppose you have $n+1$ data points and interpolate them using splines with degree n (that is, $n+1$ piecewise polynomials each having degree n). The resulting splines are all identical and equal to the interpolating polynomial that you would have obtained using a method like Lagrange interpolation.

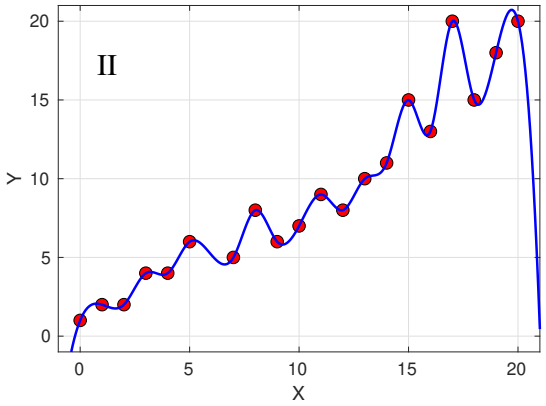
Answer: TRUE.

{ Source: JMS }

4c. Least Squares Fitting or Regression

Q4c–1[177]. For each plot, identify the type of approximation that has been performed:





(A) I = interpolation, II = curve fitting

(B) I = curve fitting, II = interpolation

(C) I = interpolation, II = extrapolation

(D) I = extrapolation, II = interpolation

Answer: (B).

{ Source: MAH }

Q4c–2[178]. You are writing a Matlab code that computes a least-squares fit `yfit` to a set of data values `ydata`, both of which are vectors of length `m`. Which of the following lines of code computes the root mean square (RMS) error for the fit?

(A) `RMS = sqrt(m * sum((yfit-y).^2));`

(B) `RMS = sqrt(sum((yfit-y).^2)) / m;`

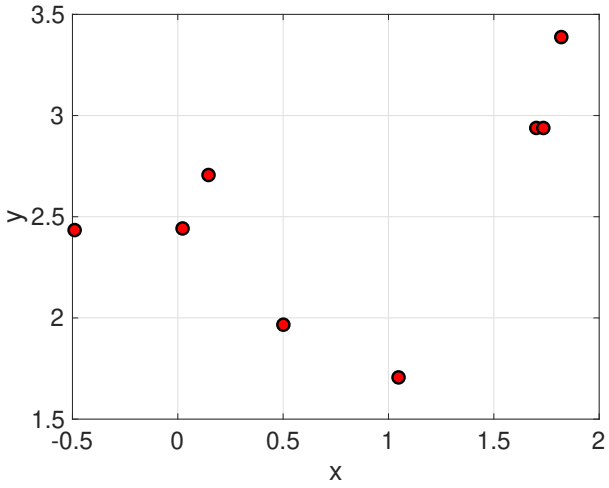
(C) `RMS = sqrt(sum(yfit-y).^2 / m);`

(D) `RMS = sqrt(sum((yfit-y).^2) / m);`

Answer: (D).

{ Source: JMS, MACM 316 lecture notes }

Q4c–3[179]. The plot depicts noisy data points that were measured in an experiment. What is the best choice of method for computing a smooth function that best approximates such noisy data?



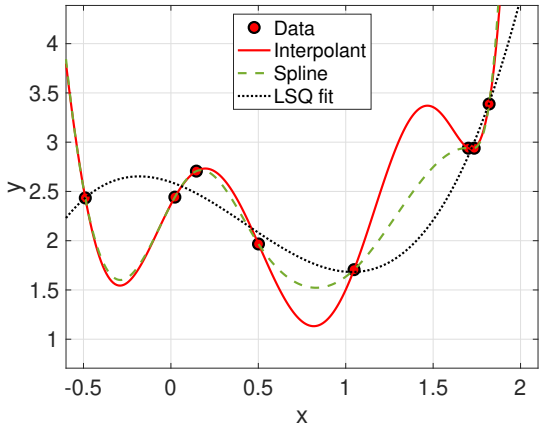
(A) A single polynomial interpolant, because it represents the data exactly.

(B) A piecewise cubic spline, because the resulting curve will be smooth at each data point.

(C) A least-squares fit, because it can capture the overall trends in the data.

(D) None will work because this data is too noisy.

Answer: (C). The data follow a trend that seems close to a cubic, but it's a bit noisy which suggests that neither interpolation method (A,B) will work very well. The plot below demonstrates clearly how the polynomial and spline interpolants can be extremely "wiggly" even in the presence of small amounts of noise. On the other hand, the least-squares cubic fit is smooth and also matches the data pretty closely. In fact, if it weren't for the two points closest to $x = 0$, both interpolants would be much closer to the cubic fit!!



{ Source: JMS }

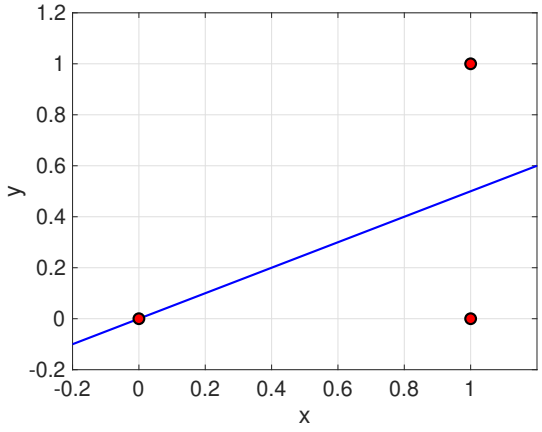
Q4c–4[180]. *True or False:* Given n data points (x_i, y_i) with all of the x_i distinct, you apply the method of least squares to fit a p^{th} degree polynomial to the given data. If $p = n - 1$ then this is equivalent to polynomial interpolation.

Answer: TRUE. When $p = n - 1$, the least squares matrix is size $n \times n$ and has the same structure as the Vandermonde matrix.

{ Source: JMS }

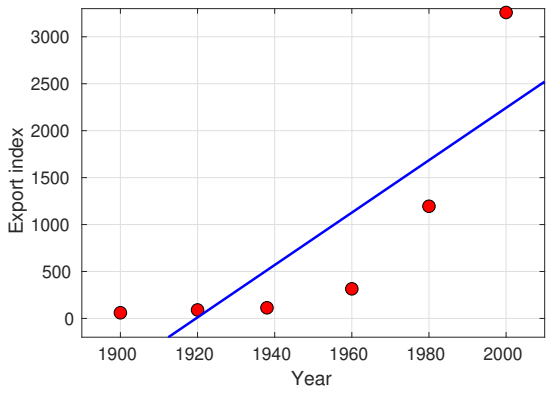
Q4c–5[181]. *True or False:* When fitting a straight line $y = a_0 + a_1 x$ to the three data points $(x_i, y_i) = (0, 0)$, $(1, 0)$ and $(1, 1)$, the least squares solution is unique.

Answer: TRUE. The method of least squares has no difficulty with the fact that the last two points have the same x-coordinate. In this case, the linear fit bisects the repeated points as shown in the plot.



{ Source: Heath [?], Exercise 3.4, p. 148 }

Q4c–6[182]. Over the last century trade has grown remarkably. The graph shows the value of the world export index over the period 1900–2000 along with a least squares fit. What type of fit is it?



(A) linear fit

(B) quadratic fit

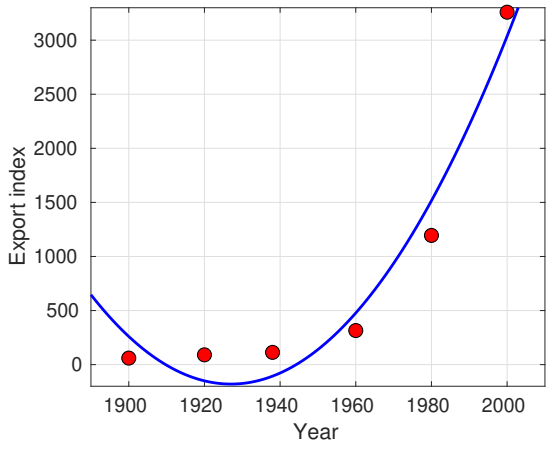
(C) cubic fit

(D) exponential fit

Answer: (A).

{ Source: Data source: https://ourworldindata.org/trade-and-globalization }

Q4c–7[183]. Over the last century trade has grown remarkably. The following graph shows the value of an world export index over the period 1900-2000. We have shown a least square fit. What type of fit is it?



(A) linear fit

(B) quadratic fit

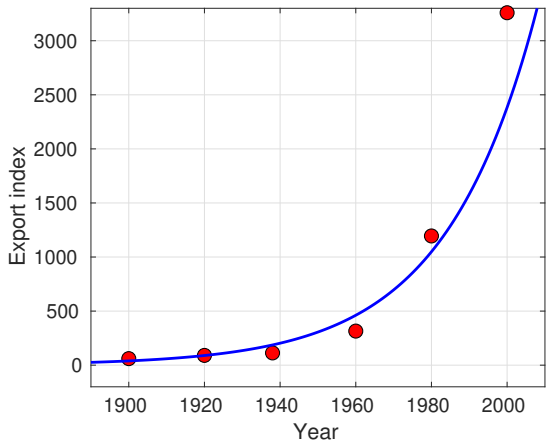
(C) cubic fit

(D) exponential fit

Answer: (B).

{ Source: Data source: https://ourworldindata.org/trade-and-globalization }

Q4c–8[184]. Over the last century trade has grown remarkably. The following graph shows the value of an world export index over the period 1900-2000. We have shown a least square fit. What type of fit is it?



(A) linear fit

(B) quadratic fit

(C) cubic fit

(D) exponential fit

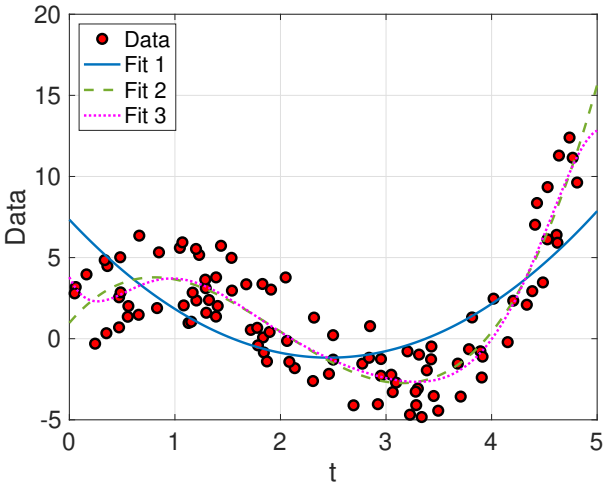
Answer: (D).

{ Source: Data source: https://ourworldindata.org/trade-and-globalization }

Q4c–9[185]. The plot shows a set of data points along with three least-squares fits (numbered 1–3). The root mean square (RMS) errors are listed below:

Fit 1: RMS = 2.8120
Fit 2: RMS = 1.7655
Fit 3: RMS = 1.7241

Which least-squares fit provides the best match with the data?



(A) Fit 1

(B) Fit 2

(C) Fit 3

Answer: (B). It would be easy to choose Fit 3 judging by the RMS error alone, but the difference from Fit 2 is very small. So in such cases, it's always important to evaluate the fit qualitatively or visually ... in the "eyeball norm". Note that Fit 3 exhibits some worrying wiggles near the end-points $t = 0$ and 5 that deviate from the overall trends in the data.

{ Source: JMS }

Q4c–10[186]. *Fill in the blank:* To find the exponential fit $Q(x) = \alpha e^{\beta x}$ using a given data set $\{(x_0, y_0), (x_1, y_1), \ldots, (x_m, y_m)\}$, you can follow these steps:

- Take the natural logarithm of both sides

$$\log(Q(x)) = \log(\alpha) + \beta x.$$

- Apply linear least squares to the points _____ to obtain the coefficients $a_0 = \log(\alpha)$ and $a_1 = \beta$.
- Then $\alpha = e^{a_0}$ and $\beta = a_1$ and $Q(x) = e^{a_0} e^{a_1 x}$.

(A) $\{(x_0, y_0),\ (x_1, y_1),\ \ldots,\ (x_m, y_m)\}$
(B) $\{(\log(x_0),\ \log(y_0)),\ (\log(x_1), \log(y_1)),\ \ldots,\ (\log(x_m), \log(y_m))\}$
(C) $\{(\log(x_0), y_0),\ (\log(x_1), y_1),\ \ldots,\ (\log(x_m), y_m)\}$
(D) $\{(x_0, \log(y_0)),\ (x_1, \log(y_1)),\ \ldots,\ (x_m, \log(y_m))\}$

Answer: (D).

{ Source: JMS, MACM 316 lecture notes }

Q4c–11[187]. *Fill in the blank:* To find the power-law fit $Q(x) = \alpha x^\beta$ using a given data set $\{(x_0, y_0), (x_1, y_1), \ldots, (x_m, y_m)\}$, you can follow these steps:

- Take the natural logarithm of both sides

$$\log(Q(x)) = \log(\alpha) + \beta \log(x).$$

- Apply linear least squares to the points _____ to obtain the coefficients $a_0 = \log(\alpha)$ and $a_1 = \beta$.
- Then $\alpha = e^{a_0}$ and $\beta = a_1$ and $Q(x) = e^{a_0} x^{a_1}$.

(A) $\{(x_0, y_0),\ (x_1, y_1),\ \ldots,\ (x_m, y_m)\}$
(B) $\{(\log(x_0), \log(y_0)),\ (\log(x_1), \log(y_1)),\ \ldots,\ (\log(x_m), \log(y_m))\}$
(C) $\{(\log(x_0), y_0),\ (\log(x_1), y_1),\ \ldots,\ (\log(x_m), y_m)\}$
(D) $\{(x_0, \log(y_0)),\ (x_1, \log(y_1)),\ \ldots,\ (x_m, \log(y_m))\}$

Answer: (B).

{ Source: JMS, MACM 316 lecture notes }

Q4c–12[188]. To find the exponential fit $Q(x) = \alpha e^{\beta x}$ using a given data set $\{(x_0, y_0), (x_1, y_1), \ldots, (x_m, y_m)\}$, the sum of squared residuals to be minimized is

(A) $\sum_{i=1}^{m} \left(y_i - \alpha e^{\beta x_i}\right)^2$
(B) $\sum_{i=1}^{m} \left(\log(y_i) - \alpha - \beta x_i\right)^2$
(C) $\sum_{i=1}^{m} \left(\log(y_i) - \alpha - \beta \log(x_i)\right)^2$
(D) $\sum_{i=1}^{m} \left(\log(y_i) - \log(\alpha) - \beta x_i\right)^2$

Answer: (D). Taking the log *of both sides of the exponential fit*

$$Q(x) = \alpha e^{\beta x}$$

gives

$$\log(Q(x)) = \log(\alpha) + \beta x \implies \log(y_i) = log(\alpha) + \beta x_i \implies E_i = \log(y_i) - \log(\alpha) - \beta x_i$$

The sum of the square of the residuals

$$\sum_{i=1}^{m} E_i^2 = \sum_{i=1}^{m} \left(\log(y_i) - \log(\alpha) - \beta x_i\right)^2.$$

{ Source: Holistic Numerical Methods [?] }

Q4c–13[189]. To find the power law fit $Q(x) = \alpha x^\beta$ using a given data set $\{(x_0, y_0), (x_1, y_1), \ldots, (x_m, y_m)\}$, the sum of squared residuals to be minimized is

(A) $\sum_{i=1}^{m} \left(y_i - \alpha x_i^\beta\right)^2$

(B) $\sum_{i=1}^{m} \left(\log(y_i) - \alpha - \beta x_i\right)^2$

(C) $\sum_{i=1}^{m} \left(\log(y_i) - \alpha - \beta \log(x_i)\right)^2$

(D) $\sum_{i=1}^{m} \left(\log(y_i) - \log(\alpha) - \beta x_i\right)^2$

Answer: (C). Taking the log *of both sides of the exponential fit* $Q(x) = \alpha x^\beta$ *gives*

$$\log(Q(x)) = \log(\alpha) + \beta \log(x) \implies \log(y_i) = \log(\alpha) + \beta \log(x_i) \implies E_i = \log(y_i) - \log(\alpha) - \beta \log(x_i)$$

The sum of the square of the residuals

$$\sum_{i=1}^{m} E_i^2 = \sum_{i=1}^{m} \left(\log(y_i) - \log(\alpha) - \beta \log(x_i)\right)^2 .$$

{ Source: Holistic Numerical Methods [?] }

Q4c–14[190]. Suppose you apply the method of least squares to obtain a linear fit $y = ax + b$ to the four points (x_1, y_1), (x_2, y_2), (x_3, y_3) and (x_4, y_4). You then repeat the process by fitting the same data points to a function of the form $x = cy + d$. Which of the following statements is FALSE?

(A) The two fitted lines are identical

(B) The two linear fits may be different

(C) The procedure fails when $ac = 0$

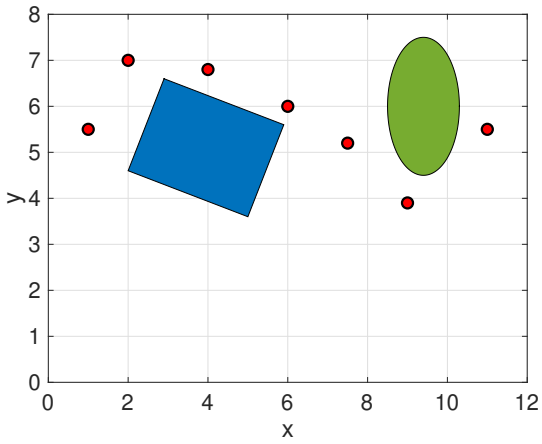
(D) $a = \dfrac{1}{c}$

Answer: (B). The first fit minimizes the (vertical) difference in the y-values, while the second fit minimizes the (horizontal) difference in the x-values.

{ Source: JMS }

4d. Applications

Q4d–1[191]. A robot has to follow a path that passes through seven planning points in order to avoid two obstacles, as shown in the plot. To find the shortest path that is also smooth, which of the following solution strategies would you recommend?



(A) Pass a sixth degree polynomial through the data

(B) Pass linear splines through the data

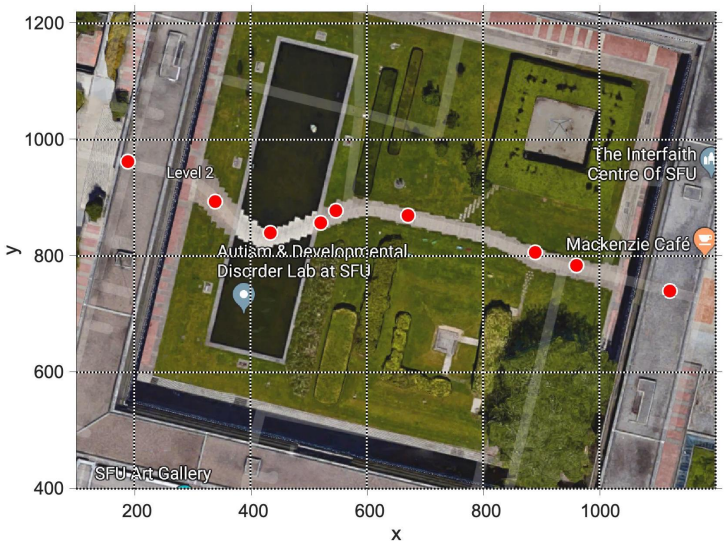
(C) Pass cubic splines through the data

(D) Use least squares to fit a second degree polynomial

Answer: (C). Linear splines would certainly give the shortest possible path, but then the path isn't smooth (maybe not the best for a robot). A least squares fit doesn't interpolate the points and so it's of no use. Among the remaining two choices, (A) is likely to be more "wiggly" and so will either generate a longer path or run into one of the objects.

{ Source: JMS }

Q4d–2[192]. If you have ever watched an SFU convocation ceremony, then you have no doubt observed the students, faculty and SFU Pipe Band process through the Academic Quadrangle and over the reflecting pool. What you may not realize is the fear this event strikes in the hearts of its participants, where one tiny misstep on the treacherous pond walk can turn a beautiful day into disaster.



In order to help graduands in selecting a safe procession path, you have taken GPS measurements of nine strategically-placed points along the procession route. To determine the safest path over the pool, which of the following strategies would you recommend?

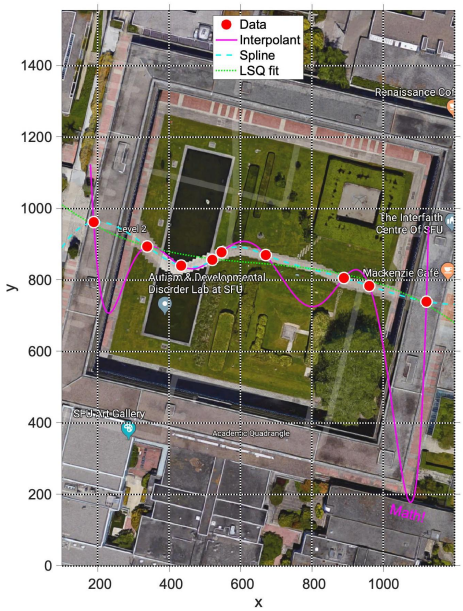
(A) Pass an eighth degree interpolating polynomial through all data points

(B) Use a linear spline to interpolate the data

(C) Use a cubic spline to interpolate the data

(D) Use least squares to fit a third degree polynomial

Answer: (C).



{ Source: MAH and JMS }

5. Differentiation and Integration

5a. Numerical Differentiation

Q5a–1[193]. Which of the following limit statements regarding the first derivative of a smooth function $f(x)$ is TRUE?

(A) $f'(x) = \lim_{h \to 0} \dfrac{f(x+h) - f(x)}{h}$

(B) $f'(x) = \lim_{h \to 0} \dfrac{f(x-h) - f(x)}{-h}$

(C) $f'(x) = \lim_{h \to 0} \dfrac{f(x+h) - f(x-h)}{2h}$

(D) All of the above

Answer: (D).

{ Source: MAH }

Q5a–2[194]. Which of the following difference formulas is a valid approximation of $f'(x_i)$? Here, the x_i represent equally-spaced points with $x_i - x_{i-1} = h$.

(A) $\dfrac{f(x_{i+1}) - f(x_i)}{h}$

(B) $\dfrac{f(x_i) - f(x_{i-1})}{h}$

(C) $\dfrac{f(x_{i+1}) - f(x_{i-1})}{2h}$

(D) All of the above

Answer: (D).

{ Source: MAH }

Q5a–3[195]. What form does the truncation error take for the difference formula $f'(x) \approx \dfrac{f(x+h) - f(x-h)}{2h}$?

(A) $O(1)$

(B) $O(h)$

(C) $O(2h)$

(D) $O(h^2)$

Answer: (D).

{ Source: MAH }

Q5a–4[196]. How would you classify the following difference formula for the derivative

$$f'(x) \approx \frac{-3f(x) + 4f(x+h) - f(x+2h)}{2h} \ ?$$

(A) forward difference

(B) backward difference

(C) centered difference

(D) one-sided difference

Answer: (D). You can also think of this as a forward difference formula, in the sense that the stencil consists of points located "forward" of the approximation point x, so response (A) is equally valid.

{ Source: MAH }

Q5a–5[197]. You want to estimate the first derivative of $f(x)$, given values of the function at discrete points $x = 0, 0.1, 0.2, \ldots, 1$. Which of these formulas is appropriate for estimating $f'(1)$?

(A) $f'(x) \approx \dfrac{-3f(x) + 4f(x+h) - f(x+2h)}{2h}$

(B) $f'(x) \approx \dfrac{f(x+h) - f(x-h)}{2h}$

(C) $f'(x) \approx \dfrac{3f(x) - 4f(x-h) + f(x-2h)}{2h}$

(D) All of the above

Answer: (C). The other two involve points $x = 1.1$ or $x = 1.2$ where f is not known.

{ Source: MAH }

Q5a–6[198]. You want to estimate the first derivative of $f(x)$, given values of the function at discrete points $x_0, x_1, x_2, \ldots, x_n$. Which of these formulas is appropriate for estimating $f'(x_0)$?

(A) $f'(x_i) \approx \dfrac{-3f(x_i) + 4f(x_{i+1}) - f(x_{i+2})}{2h}$

(B) $f'(x_i) \approx \dfrac{f(x_{i+1}) - f(x_{i-1})}{2h}$

(C) $f'(x_i) \approx \dfrac{3f(x_i) - 4f(x_{i-1}) + f(x_{i-2})}{2h}$

(D) All of the above

Answer: (A). The other two involve points to the left of x_0 where f is not known.

{ Source: MAH }

Q5a–7[199]. Which of the statements below is TRUE regarding the difference formula

$$f'(x) = \frac{1}{12h}\Big[f(x-2h) - 8f(x-h) + 8f(x+h) - f(x+2h)\Big] + \frac{h^4}{30}\, f^{(5)}(c) \ ?$$

(A) it is a 5-point centered formula

(B) the truncation error is $O(h^4)$

(C) when h is large, the truncation error dominates

(D) All of the above

Answer: (D). Regarding response (A): this formula spans a five-point stencil centered at the approximation point x, but the coefficient multiplying $f(x)$ is zero.

Q5a–8[200]. The given table lists the absolute errors from three finite difference approximations of the derivative $f'(x)$ (labelled A, B, C). The errors are computed for a sequence of decreasing values of the grid spacing h.

Which of the statements below regarding the order of accuracy for the three formulas is TRUE?

h	Formula A	Formula B	Formula C
0.500000	5.764e-01	6.670e-02	9.329e-03
0.250000	3.096e-01	1.683e-02	3.565e-04
0.125000	1.596e-01	4.218e-03	5.181e-05
0.062500	8.093e-02	1.055e-03	4.116e-06
0.031250	4.074e-02	2.638e-04	2.836e-07
0.015625	2.044e-02	6.595e-05	1.853e-08
0.007812	1.024e-02	1.649e-05	1.183e-09
0.003906	5.122e-03	4.122e-06	7.513e-11

(A) Formulas A and B are order 1, Formula C is order 2.

(B) Formula A is order 1, Formula B is order 2, Formula C is order 3.

(C) Formula A is order 1, Formula B is order 2, Formula C is order 4.

(D) All formulas have the same order of accuracy, but different error constants.

Answer: (C). Notice first that h is reduced by a factor of 2 in each step. Then observe that the error in Formula A goes down roughly by a factor of 2 in each step, Formula B by $4 = 2^2$, and Formula C by $16 = 2^4$.

{ Source: JMS }

Q5a–9[201]. The given table lists approximations for $f'(x)$ using three finite difference formulas (labelled A, B, C). The approximations are computed for a sequence of decreasing values of the grid spacing h.

Consider the convergence of the approximations as $h \to 0$, and notice that the some of the results behave anomalously for the smallest values of h. What is the most likely explanation for this behaviour?

h	Formula A	Formula B	Formula C
1.0e-00	2.474412954	1.263946140	1.974278619
1.0e-01	1.649322255	1.518206757	1.520883362
1.0e-02	1.534001862	1.520879903	1.520906914
1.0e-03	1.522218854	1.520906647	1.520906918
1.0e-04	1.521038136	1.520906915	1.520906918
1.0e-05	1.520920040	1.520906918	1.520906918
1.0e-06	1.520908230	1.520906918	1.520906917
1.0e-07	1.520907045	1.520906916	1.520906919
1.0e-08	1.520906956	1.520906934	1.520906645
1.0e-09	1.520906512	1.520906956	1.520908177
1.0e-10	1.520903403	1.520905624	1.520903403
1.0e-11	1.520916726	1.520916726	1.520794601
1.0e-12	1.520561455	1.520783499	1.520672477

(A) Formula A converges with first order accuracy, while Formulas B and C diverge.

(B) All three formulas suffer from subtractive cancellation errors for small enough h.

(C) Local truncation error dominates at small h, and is largest for Formula C.

(D) Convergence stalls for Formulas B and C around $h \approx 10^{-5}$, but then resumes when h gets small enough.

Answer: (B). Difference formulas are just that – differences – and so they experience subtractive cancellation errors when h is small and function values get very close to each other. These errors are then magnified by a factor of $\frac{1}{h}$. As an example, consider the first-order one-sided difference formula $\frac{f(x)-f(x-h)}{h}$ (this is Formula A).

{ Source: JMS }

Q5a–10[202]. An electric circuit contains a resistor (resistance R), an inductor (inductance L) and a variable voltage source $E(t)$ that obeys

$$E(t) = L\frac{di}{dt} + Ri,$$

where $i(t)$ is the electric current flowing through the circuit. You have measurements of current at several times:

time, t (seconds)	1.00	1.01	1.05	1.10
current, i (amperes)	4.01	4.05	4.10	4.50

If the inductance is $L = 0.82$ henries and the resistance is $R = 0.21$ ohms, the most accurate approximation for $E(1.10)$ is

(A) $0.82\left(\dfrac{4.50 - 4.10}{0.05}\right) + 0.21(4.50)$

(B) $0.21(4.50)$

(C) $0.82\left(\dfrac{4.05 - 4.01}{0.01}\right) + 0.21(4.50)$

(D) $0.82\left(\dfrac{4.50 - 4.01}{0.1}\right) + 0.21(4.50)$

Answer: (A). This uses a backward approximation to $i'(1.1)$ that uses the points closest to $x = 1.1$. Response (D) uses a higher order centered approximation of the derivative, but it's centered at the wrong point – $x = 1.055$!

{ Source: Holistic Numerical Methods [?] and Burden & Faires [?], p. 182 }

Q5a–11[203]. The table below lists the distance $s(t)$ that a car travels along a straight road in time t:

Time, t (seconds)	0	5	10	15
Distance, s (m)	0	100	300	700

Using the forward, backward or centered difference approximation, what is the best estimate you can find for the velocity $v = \frac{ds}{dt}$ of the car at $t = 10$ seconds?

(A) 50

(B) 60

(C) 100

(D) 200

Answer: (B). The centered difference approximation gives $v(10) \approx \left(\dfrac{700-100}{15-5}\right) = 60$.

{ Source: Holistic Numerical Methods [?] and Burden & Faires [?], p. 182 }

Q5a–12[204]. In the following centered difference formula for the second derivative

$$f''(x) = \frac{1}{12h^2}\Big[Af(x-2h) + 16f(x-h) - 30f(x) + 16f(x+h) + Af(x+2h)\Big] + O(h^4),$$

what must the constant A be for the formula to be correct?

(A) -1

(B) 1

(C) -2

(D) 2

Answer: (A). You could answer this by expanding in Taylor series. But it's easier to simply recall that in any difference approximation of a derivative, the stencil coefficients must sum to zero. Then $A+16-30+16+A = 0$.

{ Source: MAH }

Q5a–13[205]. A first-order difference approximation of the first derivative is written in the form $f'(x) = G(h)+a\cdot h+O(h^2)$, where a is some constant. Which of the extrapolation formulas below is an $O(h^2)$ approximation of $f'(x)$?

(A) $\dfrac{4G\left(\frac{h}{2}\right) - G(h)}{3}$

(B) $\dfrac{2G\left(\frac{h}{2}\right) - G(h)}{3}$

(C) $2G\left(\dfrac{h}{2}\right) - G(h)$

(D) none of the above

Answer: (C). This is analogous to Richardson extrapolation, just applied to an $O(h)$ *formula.*

{ Source: MAH }

Q5a–14[206]. *Fill in the blank:* The truncation error terms in a difference approximation $G(h)$ for the first derivative of a function can be written as $f'(x) = G(h) + a_1h + a_2h^2 + a_3h^2 + \ldots$. Then one step of extrapolation can be used to write $f'(x) = 2G\left(\frac{h}{2}\right) - G(h) + O(h^2)$. Applying one more extrapolation step yields the next higher order formula $f'(x) =$ _____ $+ O(h^3)$.

(A) $\dfrac{4G\left(\frac{h}{2}\right) - G(h)}{3}$

(B) $\dfrac{8G\left(\frac{h}{4}\right) - G\left(\frac{h}{2}\right)}{3}$

(C) $\dfrac{8G\left(\frac{h}{4}\right) + G(h)}{3}$

(D) $\dfrac{8G\left(\frac{h}{4}\right) - 6G\left(\frac{h}{2}\right) + G(h)}{3}$

Answer: (D).

{ Source: MAH }

Q5a–15[207]. Write the truncation error terms in the centered difference approximation for the first derivative as

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + a_1 h^2 + a_2 h^4 + a_3 h^6 + \dots$$

where the a_i are constants. After applying $\underline{n \text{ steps}}$ of Richardson extrapolation, what is the order of the resulting approximation for $f'(x)$?

(A) $O(h^2)$
(B) $O(h^n)$
(C) $O(h^{2n})$
(D) $O(h^{2n+2})$

Answer: (D). Each step increases the order by a factor of h^2 so the error is $O(h^2 \cdot (h^2)^n) = O(h^{2n+2})$.

{ Source: MAH }

5b. Numerical Integration or Quadrature

Q5b–1[208]. The Fundamental Theorem of Calculus is

(A) $\displaystyle\int_a^b f(x)\,dx = f(b) - f(a)$
(B) $\displaystyle\int_a^b f(x)\,dx = F(b) - F(a)$ where $f'(x) = F(x)$
(C) $\displaystyle\int_a^b f(x)\,dx = F(b) - F(a)$ where $F'(x) = f(x)$
(D) $\displaystyle\int_a^b f(x)\,dx = \frac{f(b) - f(a)}{b - a}$

Answer: (C). In the theorem statement, F is an antiderivative and so $F' = f$.

Q5b–2[209]. The definite integral $\displaystyle\int_a^b f(x)\,dx$ can be interpreted as ...

(A) the average value of $f(x)$ on the interval $[a, b]$
(B) the area under the curve $y = f(x)$ between $x = a$ and $x = b$
(C) the work done by a force $f(x)$ that acts on an object moving from position $x = a$ to b
(D) all of the above

Answer: (B). Both (B) and (C) are valid interpretations. Response (A) might also be correct, but only if $b - a = 1$ since the $\underline{\text{average}}$ value of a function is defined as $\frac{1}{b-a}\int_a^b f(x)\,dx$.

{ Source: JMS, MACM 316 lecture notes }

Q5b–3[210]. The definite integral $\displaystyle\int_a^b f(x)\,dx$ can be interpreted as

(A) the area below the curve $y = f(x)$ from a to b
(B) the area above the curve $y = f(x)$ from a to b
(C) the difference in end-point values, $f(b) - f(a)$
(D) the arithmetic average, $\dfrac{f(a) + f(b)}{2}$

Answer: (A). Response (A) is correct, but only as long as $f(x)$ is positive on $[a, b]$. Wherever $f(x)$ is negative, then (B) is the correct interpretation.

{ Source: MAH }

Q5b–4[211]. The average value of $f(x)$ on $[a, b]$ is

(A) $\dfrac{f(a) + f(b)}{2}$

(B) $\dfrac{f(a) + f(b)}{a + b}$

(C) $\displaystyle\int_a^b f(x)\,dx$

(D) $\dfrac{1}{b-a}\displaystyle\int_a^b f(x)\,dx$

Answer: (D).

{ Source: Adapted from Holistic Numerical Methods [?], quiz_07_01 }

Q5b–5[212]. *True or False:* The left and right rectangle rules are exact for constant functions, $f(x) = c$.

Answer: TRUE.

{ Source: JMS, MACM 316 lecture notes }

Q5b–6[213]. Which of the statements below is TRUE? Trapezoidal rule is exact for:

I. constant functions, $f(x) = c$

II. linear functions, $f(x) = ax + b$

III. quadratic functions, $f(x) = ax^2 + bx + c$

(A) I

(B) II

(C) I and II

(D) I, II and III

Answer: (C).

{ Source: JMS, MACM 316 lecture notes }

Q5b–7[214]. Which of the following statements is TRUE?

I. Simpson's rule is exact for linear functions, $f(x) = ax + b$.

II. Simpson's rule is exact for second-degree polynomials (quadratics), $f(x) = ax^2 + bx + c$.

III. Simpson's rule is exact for fourth-degree polynomials.

(A) none is true

(B) I

(C) II

(D) I and II

(E) I, II and III

Answer: (D).

{ Source: JMS, MACM 316 lecture notes }

Q5b–8[215]. Numerical integration is called "quadrature" because ...

(A) it is exact for quadratic functions.

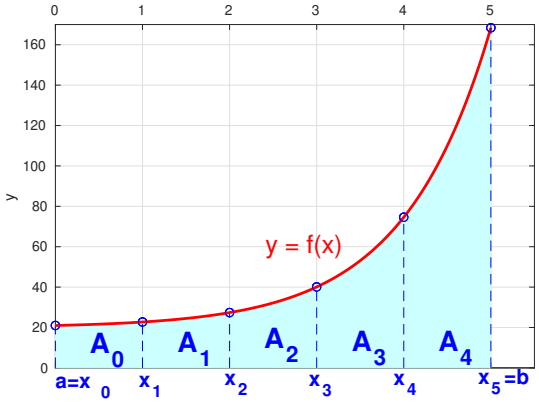
(B) the Greeks conceived of computing the area of a complex shape by replacing it with a square having the same area, and "quad" $\sim$ square.

(C) the simplest formulas involve dividing up the area into quadrangles, which are just rectangles (the Latin word *quadrangulum* means "four-cornered".)

Answer: (B). This is what Wikipedia tells us, but (C) has got to be a reasonable explanation as well!

{ Source: JMS }

Q5b–9[216]. The integral $\int_0^5 f(x)\,dx$ is approximated using the left rectangle rule (A_L), right rectangle rule (A_R) and midpoint rule (A_M) for the function $f(x)$ shown in the plot.



Which of the following inequalities must be TRUE?

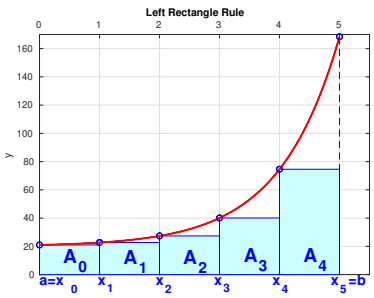
(A) $A_L < A_M < A_R$

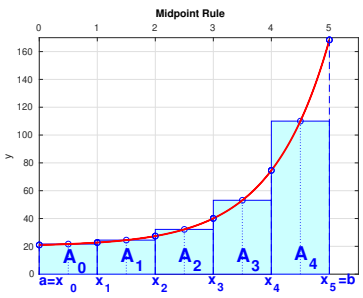
(B) $A_M < A_L < A_R$

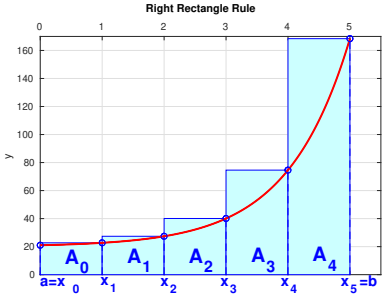
(C) $A_R < A_M < A_L$

(D) $A_M < A_R < A_L$

Answer: (A).

A_L 

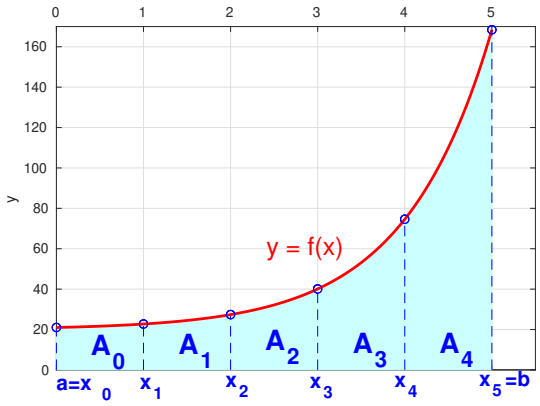
$<$ A_M 

$<$ A_R 

{ Source: MAH }

Q5b–10[217]. Apply the trapezoidal rule to approximate the integral $A = \int_0^5 f(x)\,dx$ for the given function $f(x)$ shown in the plot.

If A_T is the approximation and A is the exact value, which of the following relations is TRUE?



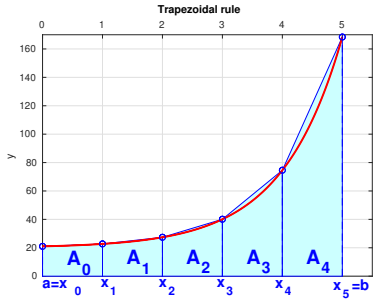
(A) $A_T > A$

(B) $A_T < A$

(C) $A_T = A$

(D) none of the above

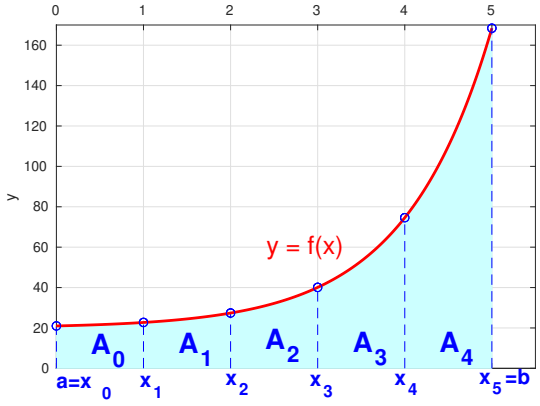
Answer: (A). The function is concave upwards and so the trapezoidal areas (defined by secant lines) always lie slightly above the curve. Then A_T is an overestimate of A.



{ Source: MAH }

Q5b–11[218]. Approximate the integral $A = \int_0^5 f(x)\,dx$ using the left rectangle rule (A_L), right rectangle rule (A_R) and trapezoidal rule (A_T) for the function $f(x)$ shown in the plot.

Which of the inequalities below must be TRUE?



(A) $A_L < A_T < A < A_R$

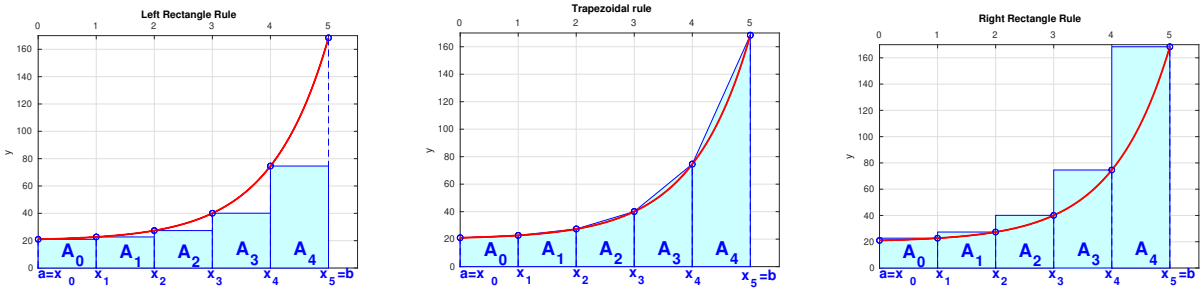
(B) $A_L < A < A_T < A_R$

(C) $A_R < A_T < A < A_L$

(D) $A_T < A_L < A < A_R$

Answer: (B).

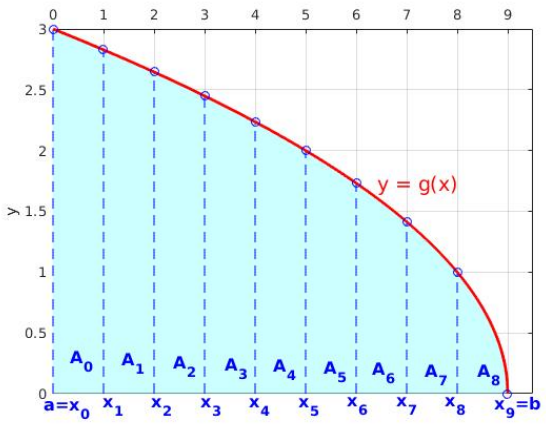
A_L $< A <$ A_T $<$ A_R



{ Source: MAH }

Q5b–12[219]. Approximate the integral $\int_0^9 g(x)\,dx$ using the left rectangle rule (A_L) and right rectangle rule (A_R) for the function $g(x)$ shown in the plot.

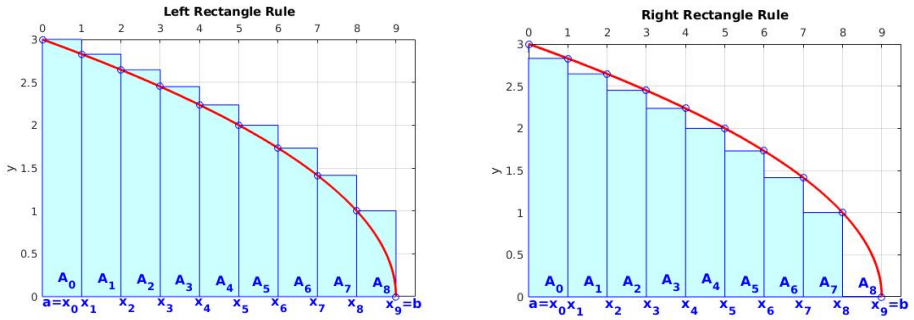
Which of the following relations is TRUE?



(A) $A_L > A_R$
(B) $A_L < A_R$
(C) $A_L = A_R$
(D) none of the above

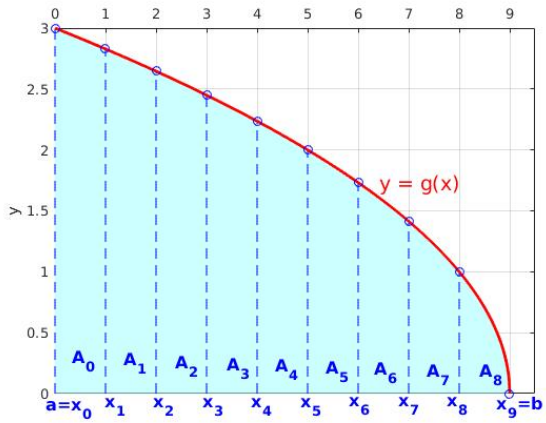
Answer: (A).

A_L $>$ A_R



{ Source: MAH }

Q5b–13[220]. Approximate the integral $A = \int_0^9 g(x)\,dx$ using the left rectangle rule (A_L), right rectangle rule (A_R) and trapezoidal rule (A_T) for the function $g(x)$ shown in the plot.



Which of the inequalities below must be TRUE?

(A) $A_L < A_T < A < A_R$
(B) $A_L < A < A_T < A_R$
(C) $A_R < A_T < A < A_L$
(D) $A_T < A_L < A < A_R$

Answer: (C).

{ Source: MAH }

Q5b–14[221]. Approximate the integral $\int_a^b h(x)\,dx$ using the left rectangle rule (A_L), right rectangle rule (A_R) and trapezoidal rule (A_T) for some arbitrary function $y = h(x)$.

Which of the following relations must be TRUE?

(A) $A_L < A_T < A_R$
(B) $A_R < A_T < A_L$
(C) $A_R < A_L < A_T$
(D) $A_T = \frac{1}{2}(A_L + A_R)$

Answer: (D). The left and right rectangle rules are

$$A_L = h\sum_{i=0}^{n-1} f(x_i) \quad \textit{and} \quad A_R = h\sum_{i=1}^{n} f(x_i),$$

and their sum is just

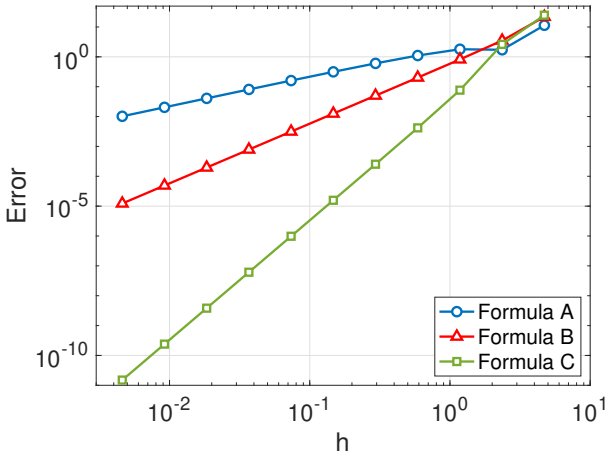
$$A_L + A_R = h\left[f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right] = 2A_T \quad \textit{(implies (D))}$$

Whether inequalities (A), (B) or (C) are true depends on the shape of $h(x)$.

{ Source: MAH }

Q5b–15[222]. The given plot shows the absolute error from three different quadrature approximations of the integral $\int_a^b f(x)\,dx$ (labelled Formulas A, B, C). The errors are computed for a sequence of decreasing values of h, the grid spacing.

Which of these statements regarding the order of accuracy for the three formulas is TRUE?



(A) Formulas A and B are order 1, Formula C is order 2.

(B) Formula A is order 1, Formula B is order 2, Formula C is order 3.

(C) Formula A is order 1, Formula B is order 2, Formula C is order 4.

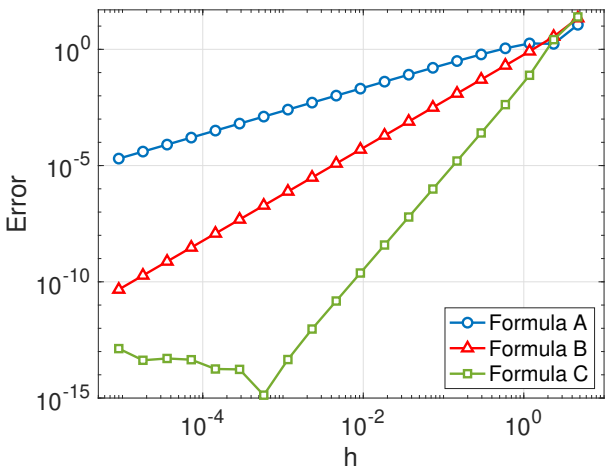
(D) All formulas have the same order of accuracy, but different error constants.

Answer: (C). On a log-log scale, the order of the scheme is given by the slope of each error curve. For example, with Formula B, h decreases by roughly 1000 (3 orders of magnitude) while error decreases by a factor of 10^{-6} (6 orders of magnitude). This means that the error behaves like $E \propto h^{6/3} \propto h^2$, so it's a second-order method.

{ Source: JMS }

Q5b–16[223]. The given plot shows the absolute error in three different quadrature approximations of the integral $\int_a^b f(x)\,dx$ (labelled Formulas A, B, C). The errors are computed for a sequence of decreasing values of h, the grid spacing.

Consider the convergence of the approximations as $h \to 0$, and notice that there is some odd behaviour with Formula C. What is the most likely explanation for this behaviour?



(A) Formulas A and B converge with order 1 and 2 respectively, whereas Formula C diverges.

(B) Formula C suffers from subtractive cancellation errors when h is small enough.

(C) The accuracy of all three formulas is limited by round-off errors that dominate the calculation when h is small.

(D) There is probably a coding error in Formula C.

Answer: (C). The error in Formula C levels off at around 10^{-15} which is pretty close to machine epsilon in double-precision floating point arithmetic. This is an indication that round-off errors are what's limiting the accuracy. Response (B) isn't correct because subtractive cancellation errors tend not to occur in quadrature formulas, since they generally involve summing terms with mostly the same sign.

{ Source: JMS }

Q5b–17[224]. The quadrature formula

$$\int_0^1 f(x)\,dx = c_0 f(0) + c_1 f(1)$$

is exact for all constant and linear functions. Determine the values of c_0 and c_1.

(A) $c_0 = 1, \quad c_1 = 1$
(B) $c_0 = -1, \ c_1 = 1$
(C) $c_0 = -\frac{1}{2}, \ c_1 = \frac{1}{2}$
(D) $c_0 = \frac{1}{2}, \quad c_1 = \frac{1}{2}$

Answer: (D). Take the simplest possible constant function $f(x) = 1$ and linear function $f(x) = x$:

$$f(x) = 1 \implies \int_0^1 dx = x\Big|_0^1 = 1 = c_0 + c_1$$
$$f(x) = x \implies \int_0^1 x\,dx = \frac{1}{2}x^2\Big|_0^1 = \frac{1}{2} = c_1$$

Solving the two linear equations yields $c_0 = \frac{1}{2}$ and $c_1 = \frac{1}{2}$, so the quadrature formula is just the arithmetic average of $f(0)$ and $f(1)$.

{ Source: MAH }

Q5b–18[225]. The quadrature formula

$$\int_{-1}^1 f(x)\,dx = f(w_1) + f(w_2)$$

is exact for all constant and linear functions. Determine all possible values of the parameters w_1 and w_2.

(A) $w_1 = -\frac{1}{2}, \ w_2 = \frac{1}{2}$
(B) $w_1 = -1, \ w_2 = 1$
(C) $w_1 = w_2$
(D) $w_1 = -w_2$

Answer: (D). Take the simplest possible constant function $f(x) = 1$ and linear function $f(x) = x$,

$$f(x) = 1 \implies \int_{-1}^1 dx = x\Big|_{-1}^1 = 2 = 1 + 1 \qquad \textit{(always satisfied)}$$
$$f(x) = x \implies \int_{-1}^1 x\,dx = \frac{1}{2}x^2\Big|_{-1}^1 = 0 = w_1 + w_2$$

Solving the two linear equations gives $w_1 = -w_2$. This means that the quadrature formula can use <u>any two points</u> as long as they are symmetric about the origin. Note that for nonlinear functions f, the formula won't be very accurate if w_1 and w_2 are not within (or at least close to) the interval $[-1, 1]$.

{ Source: MAH }

Q5b–19[226]. The velocity of an object as a function of time is given by

Time (s)	2	5	7	9	12
Velocity (m/s)	12	16	24	15	33

You use the trapezoidal rule to calculate the distance (in metres) travelled by the object between $t = 2$ and $t = 12$ seconds, and the result is ...

(A) 155.0
(B) 161.0

(C) 193.0

(D) 232.5

Answer: (C). Trapezoidal rule for this unequally-spaced set of points is

$$\frac{1}{2}\Big[3(12+16)+2(16+24)+2(24+15)+3(15+33)\Big] = \frac{1}{2}\Big[84+80+78+144\Big] = 193$$

{ Source: JMS }

Q5b–20[227]. Let I_n be the trapezoidal rule approximation of the integral $\int_a^b f(x)\,dx$ on n subintervals, then a more accurate estimate is obtained using Richardson's extrapolation as ...

(A) $I_{2n} + \dfrac{I_{2n} - I_n}{15}$

(B) $I_{2n} + \dfrac{I_{2n} - I_n}{3}$

(C) I_{2n}

(D) $I_{2n} + \dfrac{I_{2n} - I_n}{I_{2n}}$

Answer: (B). The leading order error term from trapezoidal rule is $O(h^2)$ and so the Richardson's extrapolation formula is $\dfrac{4I_{2n} - I_n}{3}$.

{ Source: JMS }

6. Initial Value Problems for Ordinary Differential Equations (ODEs)

6a. Background on ODEs

Q6a–1[228]. What is the most accurate classification of the equation:

$$\frac{dy}{dt} = cy(1-y) \ ?$$

(A) first-order linear ordinary differential equation

(B) first-order nonlinear ordinary differential equation

(C) first-order linear partial differential equation

(D) first-order nonlinear partial differential equation

Answer: (B).

{ Source: MAH }

Q6a–2[229]. A differential equation is classified as "ordinary" if it has:

(A) one dependent variable

(B) more than one dependent variable

(C) one independent variable

(D) more than one independent variable

Answer: (C).

{ Source: Holistic Numerical Methods [?], quiz_08_01 }

Q6a–3[230]. The deflection $w(x)$ of the free end of a flexible beam in response to a load force $q(x)$ is described by the differential equation

$$\frac{d^4w}{dx^4} + a^2\frac{d^2w}{dx^2} = q(x)$$

What is the order of this ODE?

(A) 1

(B) 2

(C) 4

(D) 6

Answer: (C).

{ Source: MAH }

Q6a–4[231]. Which of the following is a well-posed ODE initial value problem?

(A) $\dfrac{dy}{dx} = y^2 - e^y, \quad y(0) = 0$

(B) $y = 4y' - \log(t), \quad y(0) = 4$

(C) $t^2\dfrac{df}{dt} - t|f| + \sin(f+t) - \dfrac{1}{3}t^3 = 0, \quad f(\pi) = -1$

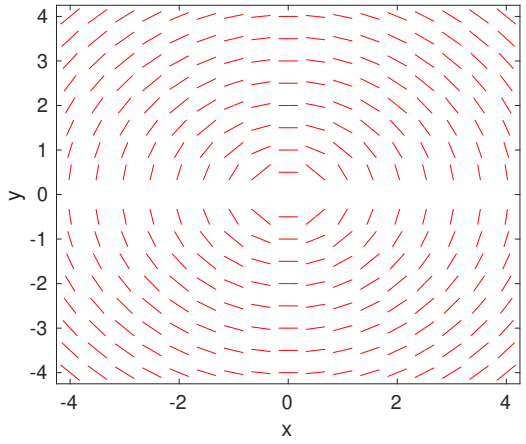
(D) $\dfrac{z'+t}{z} = 1, \quad z(1) = 0$

(E) $xx' - 3xy + 4x' = y^2$

Answer: (A). Responses (C) and (D) are also correct. But (B) is not well-posed because $\log(t)$ *isn't continuous at the initial point* $t = 0$*. And neither is (E) since no initial condition is given.*

{ Source: JMS }

Q6a–5[232]. Which of the ODEs listed below could have generated this slope field plot?



(A) $\dfrac{dy}{dx} = f(y)$

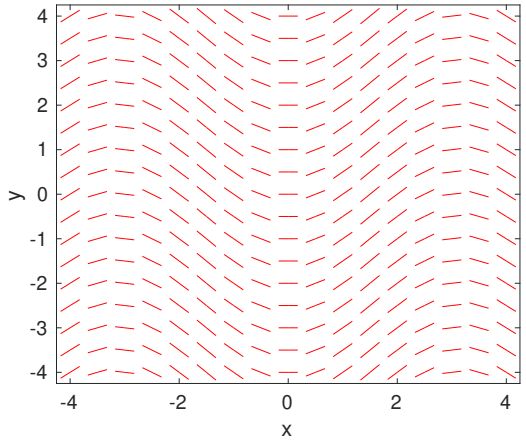
(B) $\dfrac{dy}{dx} = f(x)$

(C) $\dfrac{dy}{dx} = f(x, y)$

Answer: (C). Consider any horizontal line (y = constant) and notice that the slope varies with x. Similarly, the slope depends on y along vertical lines. So the ODE RHS must depend on both x and y. This is actually a slope field of $\frac{dy}{dx} = -\frac{x}{y}$.

{ Source: MAH }

Q6a–6[233]. Which of the ODEs listed below could have generated this slope field plot?



(A) $\dfrac{dy}{dx} = f(x)$

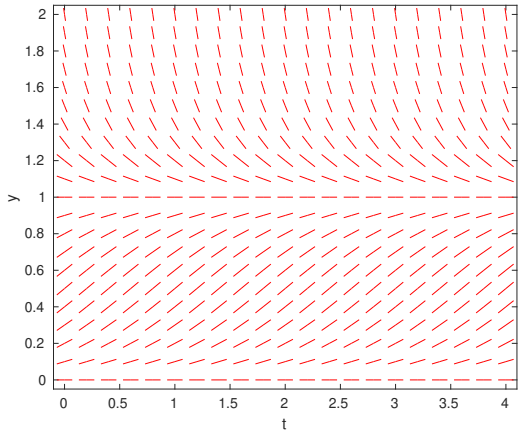
(B) $\dfrac{dy}{dx} = f(y)$

(C) $\dfrac{dy}{dx} = f(x, y)$

Answer: (A). The slope does not change with y along any vertical line (x = constant) and so the ODE RHS only depends on x. This is actually the slope field of $\frac{dy}{dx} = \sin(x)$.

{ Source: MAH }

Q6a–7[234]. Which of the ODEs listed below could have generated this slope field plot?



(A) $\dfrac{dy}{dt} = f(t)$

(B) $\dfrac{dy}{dt} = f(y)$

(C) $\dfrac{dy}{dt} = f(t, y)$

Answer: (B). The slope does not change with t along any horizontal line (y = constant) and so the ODE RHS only depends on y. This is actually the slope field of $\frac{dy}{dt} = 2y(1-y)$.

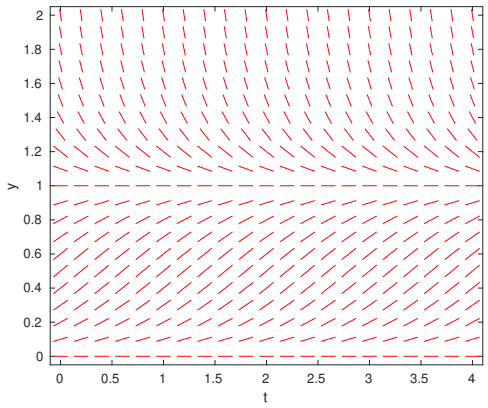
{ Source: MAH }

Q6a–8[235]. The slope field for an ODE is shown. Which of these statements is TRUE?

I. For $y(0) > 1$ all solutions are decreasing.

II. For $y(0) = 1$ the solution remains the same.

III. For $y(0) < 1$ all solutions are increasing.



(A) I

(B) II

(C) I and II

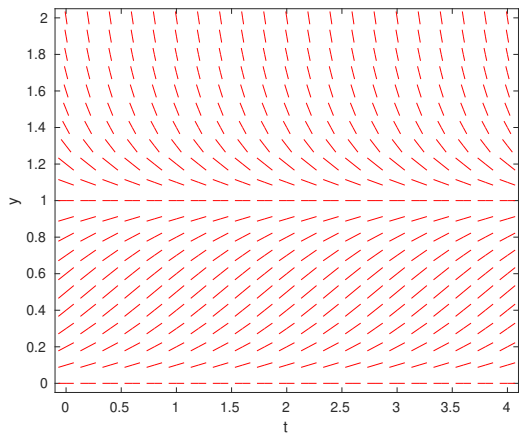
(D) I and III

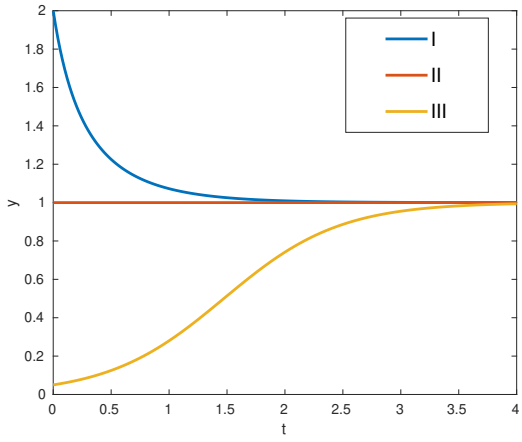
(E) I, II and III

Answer: (C).

{ Source: MAH }

Q6a–9[236]. The slope field for an ODE is shown below on the left. Which of the solution curves (I–III) in the right hand plot could be solutions?





(A) I

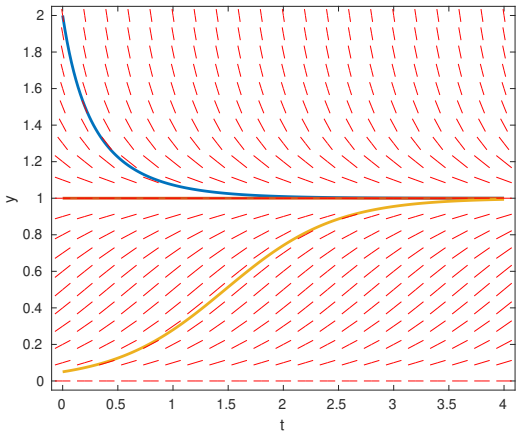
(B) II

(C) I and II

(D) I and III

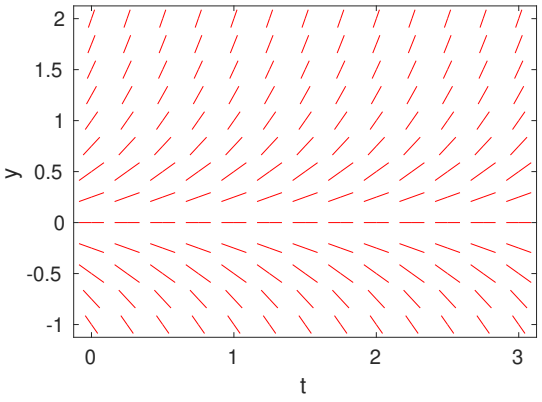
(E) I, II and III

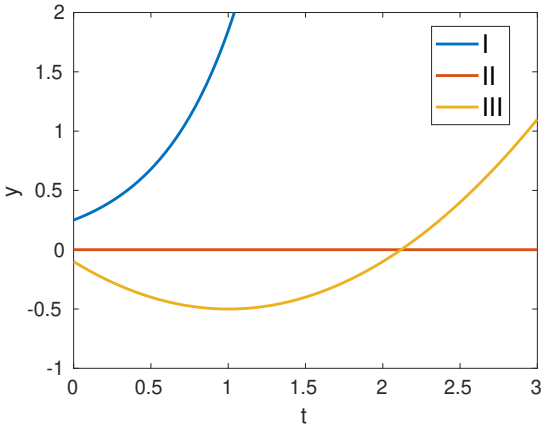
Answer: (E).



{ Source: MAH }

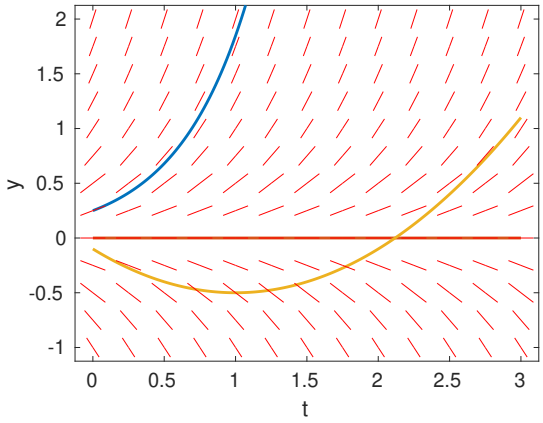
Q6a–10[237]. The slope field for an ODE is shown below on the left. Which of the solution curves (I–III) in the right hand plot could be solutions?





(A) I
(B) II
(C) I and II
(D) I and III
(E) I, II and III

Answer: (C).



{ Source: MAH }

6b. Euler's Method

Q6b–1[238]. Consider the population growth model $\frac{dN}{dt} = aN\left(1 - \frac{N}{K}\right)$, where $N(t)$ is population as a function of time, a is the growth rate, and K is the maximum sustainable population. If you were to approximate the solution using Euler's method with time step h, what is the difference equation you would use?

(A) $N_{j+1} = N_j + haN_j\left(1 - \frac{N_j}{K}\right)$
(B) $N_{j+1} = hN_j + aN_j\left(1 - \frac{N_j}{K}\right)$
(C) $N_{j+1} = aN_j + hN_j\left(1 - \frac{N_j}{K}\right)$

(D) $N_{j+1} = aN_j - \dfrac{h}{K} N_j^2$

Answer: (A).

{ Source: MAH }

Q6b–2[239]. Using Euler's method to approximate an ODE, you obtain the difference equation $y_{j+1} = y_j + hay_j$ where a is a constant and h is the time step. What is the ODE?

(A) $\dfrac{dy}{dt} = ay$

(B) $\dfrac{dy}{dt} = a$

(C) $\dfrac{dy}{dt} = hay$

(D) $\dfrac{dy}{dt} = ha$

Answer: (A).

{ Source: MAH }

Q6b–3[240]. Recall the well-posedness (existence-uniqueness) theorem from class:

Theorem: *For the initial value problem*

$$y' = f(t, y) \quad \textit{for} \quad t \in [a, b] \quad \textit{with } y(a) = y_0,$$

suppose that:
- $f(t, y)$ *is continuous for all* y *and* $t \in [a, b]$, *and*
- $f(t, y)$ *satisfies the "Lipschitz condition"*

$$|f(t, y_1) - f(t, y_2)| \leqslant L|y_1 - y_2|.$$

Then the IVP has a unique solution $y(t)$ *for all* $t \in [a, b]$.

Consider the IVP $y' = -y + t + 1$ with $y(0) = 1$, for $t \in [0, 5]$. What is the corresponding Lipschitz constant that ensures this problem is well-posed?

(A) 0

(B) 1

(C) 5

(D) 6

Answer: (B). *Substituting the RHS function into the Lipschitz condition:*

$$|f(t, y_1) - f(t, y_2)| = \big| - y_1 + t + 1 - (-y_2 + t + 1)\big| = |y_2 - y_1| = |y_1 - y_2|.$$

{ Source: MAH }

Q6b–4[241]. The local truncation error for Euler's method is

(A) $\mathcal{O}(1)$

(B) $\mathcal{O}(h)$

(C) $\mathcal{O}(h^2)$

(D) $\mathcal{O}(h^3)$

Answer: (C). *This is the error in a single step, which comes from the Taylor series expansion.*

{ Source: MAH }

Q6b–5[242]. The global truncation error for Euler's method is

(A) $\mathcal{O}(1)$
(B) $\mathcal{O}(h)$
(C) $\mathcal{O}(h^2)$
(D) $\mathcal{O}(h^3)$

Answer: (B). Each step has LTE$= O(h^2)$, which is multiplied by $N = O\left(\frac{1}{h}\right)$ steps.

{ Source: MAH }

Q6b–6[243]. *True or False:* When you solve an ODE using a time-stepping algorithm like Euler's method, the local truncation errors in every step add up and so the error will always grow with time.

Answer: FALSE. The sign of the error in each step can be positive or negative. If errors always have the same sign, then they can grow with time as they accumulate. But if the signs alternate, then errors can cancel.

{ Source: JMS }

Q6b–7[244]. Consider the IVP

$$\frac{dy}{dx} = \frac{y-2}{x+2}, \quad y(2) = -1.$$

Using a single step of Euler's method, what is the approximate solution at $x = 2.5$?

(A) 0.75
(B) 0.25
(C) -0.625
(D) -1
(E) -1.375

Answer: (E). The RHS function is $f(x,y) = \frac{y-2}{x+2}$ and we start with $y_0 = -1$. Taking one step of Euler's method with $h = 0.5$ gives

$$y_1 = y_0 + hf(x_0, y_0) = -1 + 0.5\left(\frac{-1-2}{2+2}\right) = -1.375$$

{ Source: JMS }

Q6b–8[245]. Consider the IVP

$$2\,\frac{dy}{dx} + y\cos y = 0, \quad y(\pi) = \pi.$$

Using a single step of Euler's method, what is the approximate solution at $x = \dfrac{3\pi}{2}$?

(A) $\pi + \frac{1}{4}\,\pi^2$
(B) $\pi - \frac{1}{2}\,\pi^2$
(C) $\frac{3}{4}\,\pi$
(D) $\frac{1}{2}\,\pi^2$
(E) 0

Answer: (A). The RHS function is $f(y) = -\frac{1}{2}\,y\cos y$ and we start with $y_0 = \pi$. Taking one step of Euler's method with $h = \frac{\pi}{2}$ gives

$$y_1 = y_0 + hf(y_0) = \pi + \frac{\pi}{2}\left(-\frac{\pi}{2}\cos\pi\right) = \pi + \frac{\pi^2}{4}$$

{ Source: JMS }

Q6b–9[246]. Consider the IVP

$$\frac{dy}{dt} - y^2 - t = 0, \quad y(0) = 1.$$

Using Euler's method with time step $h = 1$, what is the approximate value for $y(2)$?

(A) 1
(B) 2
(C) 7
(D) 9

Answer: (C). The RHS function is $f(t, y) = t + y^2$ and we start with $y_0 = 1$. Taking two steps of Euler's method gives

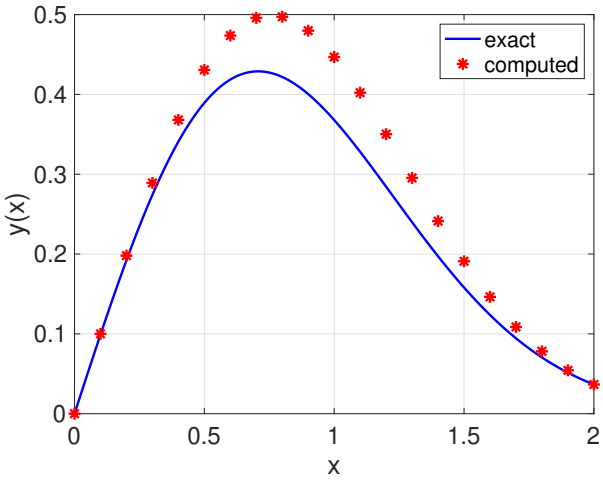
$$y_1 = y_0 + hf(t_0, y_0) = 1 + 1(0 + 1^2) = 2$$
$$y_2 = y_1 + hf(t_1, y_1) = 2 + 1(1 + 2^2) = 7$$

{ Source: JMS }

Q6b–10[247]. You apply Euler's method to solve the initial value problem

$$y' = \begin{cases} y\left(-2x + \frac{1}{x}\right), & \text{if } x \neq 0 \\ 1, & \text{if } x = 0 \end{cases}$$

on the interval $x \in [0, 2]$ with initial condition $y(0) = 0$. The numerical solution for $N = 20$ points is plotted alongside the exact solution $y(x) = xe^{-x^2}$. Which of the following statements is wrong?

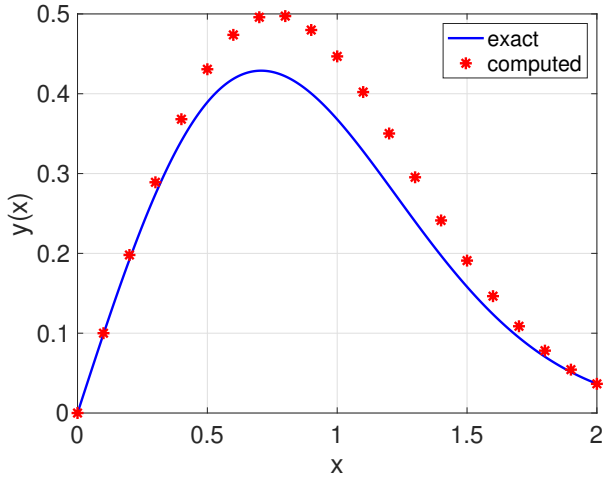


(A) Euler's method does not converge for this problem.
(B) The local truncation error is sometimes large, and sometimes small.
(C) The global truncation error is small near $x = 2$.
(D) The accuracy of the solution could likely be improved by increasing N.

Answer: (A). Responses (B)–(D) are correct. And although (A) might be true, there is no way to determine convergence with only results from a single N.

{ Source: Based on Fausett [?], p. 450 }

Q6b–11[248]. You apply Euler's method to compute the results displayed in the plot, but you decide that the numerical solution is much too inaccurate for your purposes. What could you do to increase the accuracy?



(A) Reduce the time step

(B) Use a higher order method

(C) Use adaptive time-stepping

(D) Switch to an implicit method like backward Euler

(E) All of the above

Answer: (A). Responses (A)–(C) are all reasonable answers. Switching to backward Euler won't help with accuracy, but could improve stability.

{ Source: JMS }

6c. Higher Order Methods

Q6c–1[249]. This partial code implements the modified Euler method for solving the IVP $y' = f(t, y)$, $y(a) = y_0$, for $a \leqslant t \leqslant b$. The Matlab function `f` returns the value of $f(t, y)$. Select suitable replacement code for the `[blank]` from the list below.

```
function [t, y] = meuler(f,a,b,y0,h)
t = a :  h :  b;
y(1) = y0;
for k = 1 :  length(t),
   ystar = y(k) + h * f(t(k), y(k));
   ...[fill in blank]...
end
```

(A) `y(k+1) = y(k) + h/2 * (f(t, y(k)) + f(t+h, ystar));`

(B) `y(k+1) = y(k) + h/2 * (f(t(k), y(k)) + f(t(k)+h, ystar));`

(C) `y(k+1) = y(k) + h * f(t(k), y(k)) + f(t(k)+h, y(k)));`

(D) `y(k+1) = y(k) + h * f(t(k)+h, ystar);`

Answer: (B).

{ Source: JMS, MACM 316 lecture notes }

Q6c–2[250]. You perform the following two calculations on the same initial value problem:

- two steps of Euler's method with step size 0.25
- one step of modified Euler's method with step size 0.5

Which method do you expect to give the more accurate solution?

(A) Euler's method

(B) modified Euler's method

(C) both give the same accuracy

Answer: (C). Euler's method has LTE $= O(h^2)$ so that two steps of size 0.25 give LTE ≈ 0.125. Modified Euler's method has LTE $= O(h^3)$ so that one step of size 0.5 give the same LTE ≈ 0.125.

{ Source: JMS }

Q6c–3[251]. You perform the following two calculations on the same initial value problem:

- two steps of Euler's method with step size 0.1
- one step of modified Euler's method with step size 0.2

Which method do you expect to give the more accurate solution?

(A) Euler's method

(B) modified Euler's method

(C) both give the same accuracy

Answer: (B). Euler's method has LTE $= O(h^2)$ so that two steps of size 0.1 give LTE ≈ 0.02. Modified Euler's method has LTE $= O(h^3)$ so that one step of size 0.2 gives LTE ≈ 0.008, which is much smaller!

{ Source: JMS }

Q6c–4[252]. When comparing Euler's method with standard fourth-order Runge-Kutta (RK4), which of the following is <u>not</u> an advantage of RK4?

(A) The cost per time step is lower.

(B) The local truncation error is much smaller.

(C) It is more stable and a much larger time step can usually be used.

Answer: (A). The cost of an RK4 step is roughly 4 times that of Euler's method. The cost savings come in because many fewer time steps are usually needed.

{ Source: JMS }

Q6c–5[253]. This partial code implements the midpoint method for solving the differential equation $y' = f(t,y)$, $a \leqslant t \leqslant b$, $y(a) = y_0$, where the Matlab function `f` returns the value of $f(t,y)$. Select suitable replacement code for the `[blank]`.

```
function [t, y] = midpointmethod(f,a,b,y0,h)
t = a :  h :  b;
y(1) = y0;
for k = 1 :  length(t),
   ystar = y(k) + h/2 * f(t(k), y(k));
   ...[fill in blank]...
end
```

(A) `y(k+1) = y(k) + h/2 * (f(t(k), y(k)) + f(t(k)+h, ystar));`

(B) `y(k+1) = y(k) + h * f(t(k)+h/2, y(k));`

(C) `y(k+1) = y(k) + h * f(t(k)+h/2, ystar);`

(D) `y(k+1) = y(k) + h * f(t(k)+h, ystar);`

Answer: (C).

{ Source: JMS, MACM 316 lecture notes }

Q6c–6[254]. You have a code for solving ODEs that predicts the solution at time t_{j+1} using

$$y^* = y_j + hf(t_j, y_j),$$

and then computes the actual solution approximation as

$$y_{j+1} = y_j + \frac{h}{2}\Big(f(t_j, y_j) + f(t_{j+1}, y^*)\Big).$$

This approach is known as

(A) Euler's method

(B) modified Euler's method

(C) Runge-Kutta method

(D) midpoint method

Answer: (B). This is a Runge-Kutta method of order 2, and so (C) is also correct.

{ Source: MAH }

Q6c–7[255].

$$y^* = y_j + \frac{h}{2} f(t_j, y_j)$$

$$y_{j+1} = y_j + hf\left(t_j + \frac{h}{2}, y^*\right)$$

This iteration is called

(A) Euler's method

(B) modified Euler method

(C) Runge-Kutta method

(D) midpoint method

Answer: (D).

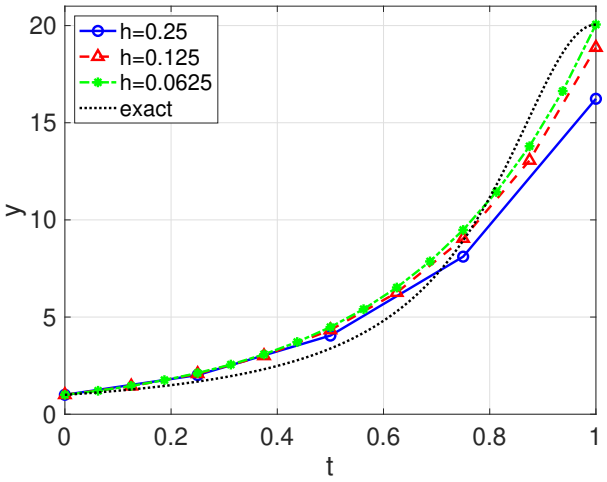
{ Source: MAH }

Q6c–8[256]. *True or False:*

Your friend writes a Matlab code that integrates the ODE

$$\frac{dy}{dt} = \frac{2(1-t)}{(0.05 + (t-1)^2)^2}, \quad y(0) = 1$$

from $t = 0$ to 1 using the modified Euler method. The plot shows the computed solution for a decreasing sequence of step sizes $h = 0.25, 0.125$ and 0.0625, alongside the exact solution. Based on these results, your friend claims that their code is correct.



Answer: FALSE. Well, it's most likely false. The right hand end-point seems to approach the exact solution $y(1) \approx 20$ as h gets smaller. But the middle values (near $t \approx 0.5$) are not getting very close to the exact solution. Furthermore, the numerical approximations all have a slope at $t = 1$ that is nonzero and increasing with h, which seems to diverge from the exact slope of $\frac{dy}{dt}(1) = 0$.

{ Source: Based on Recktenwald [?], p. 728 }

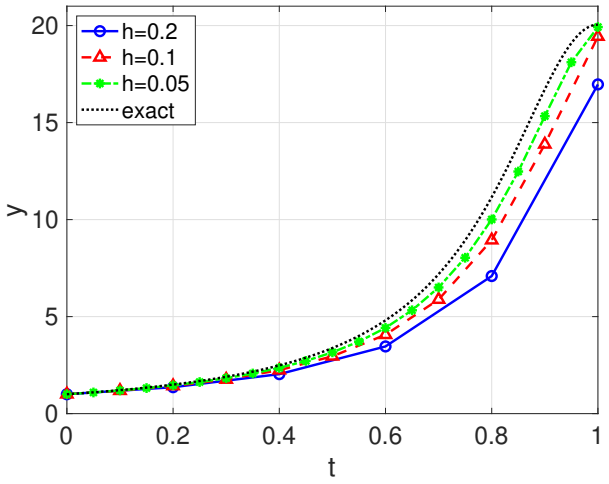
Q6c–9[257]. *True or False:*

Your friend writes a Matlab code that integrates

$$\frac{dy}{dt} = \frac{2(1-t)}{(0.05+(t-1)^2)^2}, \quad y(0) = 1$$

from $t = 0$ to 1 using the modified Euler method. The plot shows the computed solution for a decreasing sequence of step sizes alongside the exact solution. He also calculates the max-norm error in the three solutions:

h	$\max_j \lvert y_j - y(t_j)\rvert$
0.2	4.0680
0.1	2.8348
0.05	1.3826



Based on these results, your friend claims that their code is correct.

Answer: FALSE. The approximations do seem to be converging to the exact solution as $h \to 0$. However, the error is only going down roughly by a factor of 2 in each step, which suggests a first-order method. Modified Euler should be second-order, and so there must be a bug in the code.

{ Source: Based on Recktenwald [?], p. 728 }

Q6c–10[258]. You have coded up an ODE solver and to test it you execute your code twice, first with time step h and second with step $h/2$. Suppose these two solutions have absolute errors E_h and $E_{h/2}$ respectively. Which expression is an estimate for the order of accuracy?

(A) $\dfrac{\log_2 E_h}{\log_2 E_{h/2}}$

(B) $\dfrac{\log_2 E_{h/2}}{\log_2 E_h}$

(C) $\log_2\left(\dfrac{E_h}{E_{h/2}}\right)$

(D) $\log_2\left(\dfrac{E_{h/2}}{E_h}\right)$

Answer: (C). If the error in the method is order p, then $E_h = ch^p$ and $E_{h/2} = c(h/2)^p$. Taking the ratio $E_h/E_{h/2} = 2^p$, it follows that $p = \log_2(E_h/E_{h/2})$.

{ Source: JMS }

Q6c–11[259]. You have coded up an ODE solver and to test it, you execute your code twice, once with time step h and a second time with step $h/10$. Suppose these two solutions have absolute errors E_h and $E_{h/10}$ respectively. Which expression is an estimate for the order of accuracy?

(A) $\log_{10}\left(\dfrac{E_h}{E_{h/10}}\right)$

(B) $\log_2\left(\dfrac{E_h}{E_{h/10}}\right)$

(C) $\dfrac{1}{\log_2 10}\log_2\left(\dfrac{E_h}{E_{h/10}}\right)$

(D) $\dfrac{\log_{10} E_h}{\log_{10} E_{h/10}}$

Answer: (A). Response (C) is also correct. If the error in the method is order p, then $E_h = ch^p$ and $E_{h/10} = c(h/10)^p$. Taking the ratio $E_h/E_{h/10} = 10^p$, it follows that $p = \log_{10}(E_h/E_{h/10})$. Applying the logarithmic identity $\log_{10} x = \frac{\log_2 x}{\log_2 10}$ gives the formula in (C).

{ Source: JMS }

Q6c–12[260]. The integral $f(x) = \int_0^x e^{-t^2}\,dt$ can also be written as an ODE, $\frac{df}{dx} = e^{-x^2}$ (by the Fundamental Theorem of Calculus). With this in mind, consider two methods for estimating the value $f(1)$:

- Apply Simpson's rule to approximate the integral $\int_0^1 e^{-x^2}\,dx$.
- Use the RK4 method to solve $\frac{df}{dx} = e^{-x^2}$ on the interval $[0, 1]$ using initial condition $f(0) = 0$.

Discretize both problems at the same points $x_i = ih$ for $i = 0, 1, \ldots, N$, with grid spacing $h = \frac{1}{N}$. Which of the statements below regarding the accuracy and efficiency of the methods is TRUE?

(A) Both methods have the same accuracy, but Simpson's rule is less expensive.

(B) Simpson's rule is more accurate and less expensive.

(C) RK4 is more accurate and less expensive.

(D) Both methods have the same accuracy and cost.

Answer: (A). Both RK4 and Simpson's rule have error that is $O(h^4)$. Each step of RK4 requires 4 function evaluations, for a total cost of $4N$. In comparison, Simpson's rule involves only $N + 1$ function evaluations (one per point). So the cost of both is $O(N)$, but the quadrature approach will be roughly four times cheaper/faster.

{ Source: Based on Cheney and Kincaid [?], p. 390. }

6d. Systems of First-Order ODEs

Q6d–1[261]. Which of the following is a well-posed ODE initial value problem?

(A) $\dfrac{d^2y}{dx^2} = y^2 - e^y, \quad y(0) = 0$

(B) $z'' - 4xz' + 3z - xe^x = 4, \quad z(1) = z'(1) = 0$

(C) $\dfrac{y''' - y}{y'' + y'} = e^t, \quad y(0) = 1,\ y'(0) = 2,\ y''(0) = -3$

(D) $\dfrac{d^2z}{dt^2} + t\,\dfrac{dz}{dt} - \dfrac{1}{1+z} = 0, \quad z(1) = -1,\ z'(1) = \pi$

(E) $ff'' = 1, \quad f(0) = f'(0) = 3$

Answer: (B). Responses (C) and (E) are also correct. But (A) is not well-posed because it is missing an initial condition on $\frac{dy}{dx}$. And (D) isn't well-posed because the RHS function is discontinuous at the initial value of $z = -1$.

{ Source: JMS }

Q6d–2[262]. A mass–spring problem is described by the ODE $y'' + 2y = \cos(2t)$. Which of the following is the equivalent first-order system?

(A) $y' = x, \quad x' + 2y = \cos(2t)$

(B) $x' = y, \quad y' = -2y + \cos(2t)$

(C) This second-order ODE can't be converted to a first-order system.

Answer: (A).

{ Source: MAH }

Q6d–3[263]. Which second-order ODE is equivalent to this first-order system?

$$x' = y$$
$$y' = -2x + y$$

(A) $y'' + 2y' - y = 0$

(B) $y'' - 2y' - y = 0$

(C) $y'' + 2y' - y = 0$

(D) $y'' - y' + 2y = 0$

Answer: (D).

{ Source: MAH }

Q6d–4[264]. Which second-order ODE is equivalent to this first-order system?

$$x' = x - 2y$$
$$y' = -2x + y$$

(A) $x'' - 2x' - 3x = 0$

(B) $y'' - 2y' - 3y = 0$

(C) $y'' + 2y' + 3y = 0$

(D) This first-order system can't be converted to a second-order ODE.

Answer: (A). We're going to eliminate y and so differentiate the first equation to get $x'' = x' - 2y'$. Then substitute for y' from the second equation to get $x'' = x' - 2(-2x + y) = x' + 4x - 2y$. Finally, replace $-2y = x' - x$ from the first equation, which yields $x'' = 2x' + 3x$.

{ Source: MAH }

Q6d–5[265]. A mass m is attached to a fixed wall with a spring–damper device having spring constant k and damping parameter c. Our aim is to find the horizontal displacement $x = x(t)$ as a function of time $t \geqslant 0$. Applying Newton's second law leads to the second-order ODE

$$mx'' + cx' + kx = f(t), \quad x(0) = x_0, \quad x'(0) = v_0,$$

where x_0 denotes the initial displacement, v_0 the initial velocity, and $f = f(t)$ describes external forces acting on the body. Re-write this IVP as a first-order system by using the new variable $v(t) = x'(t)$.

(A) $x' = v,\ mv' + cv + kx = f,\ x(0) = x_0,\ v(0) = v_0$

(B) $v' = x,\ mv' + cv + kx = f,\ x(0) = x_0,\ v(0) = v_0$

(C) $x' = v,\ mx' + cx + kv' = f,\ x(0) = x_0,\ v(0) = v_0$

Answer: (A).

{ Source: MAH }

Q6d–6[266]. You are given the system of ODEs $x' = 3x - 2y$ and $y' = 4y^2 - 7x$, along with initial values $x(0) = 2$ and $y(0) = 1$. Using one step of Euler's method, what is the approximate solution at $t = 0.1$?

(A) $x(0.1) = 4, \quad y(0.1) = -10$

(B) $x(0.1) = 6, \quad y(0.1) = -9$

(C) $x(0.1) = 2.4, \quad y(0.1) = 0$

(D) $x(0.1) = 2.4, \quad y(0.1) = -1$

(E) none of the above

Answer: (C). Euler's method with $h = 0.1$ gives:

$$x_1 = x_0 + h(3x_0 - 2y_0) = 2 + 0.1(6 - 2) = 2.4$$
$$y_1 = y_0 + h(4y_0^2 - 7x_0) = 1 + 0.1(4 - 14) = 0$$

{ Source: MathQuest [?], Euler's method and systems of equations }

Q6d–7[267]. You are given the system of ODEs $x' = x(-x - 2y + 5)$ and $y' = y(-x - y + 10)$, along with initial values $x(4.5) = 3$ and $y(4.5) = 2$. What are approximate values of x and y at $t = 4$?

(A) $x(4) = 0, \quad y(4) = -3$

(B) $x(4) = 6, \quad y(4) = 10$

(C) $x(4) = 6, \quad y(4) = 7$

(D) none of the above

Answer: (D). There's nothing wrong with taking a negative step size, $h = -0.5$! Euler's method gives:

$$x_1 = x_0 + hx_0(-x_0 - 2y_0 + 5) = 3 - 0.5(3)(-3 - 4 + 5) = 6$$
$$y_1 = y_0 + hy_0(-x_0 - y_0 + 10) = 2 - 0.5(2)(-3 - 2 + 10) = -3$$

{ Source: MathQuest [?], Euler's method and systems of equations }

6e. Stability and Implicit Methods

[nothing here yet]

SUMMARY STATISTICS:

Total questions: 267 (25 TF, 242 MC)
MC responses: A=65, B=55, C=59, D=56, E=7

Printed: June 3, 2020

References

[1] Richard L. Burden, J. Douglas Faires, and Annette M. Burden. *Numerical Analysis*. Cengage Learning, 10th edition, 2015.

[2] Ward Cheney and David Kincaid. *Numerical Mathematics and Computing*. Brooks/Cole, Pacific Grove, CA, 4th edition, 1999.

[3] Laurene V. Fausett. *Applied Numerical Analysis Using MATLAB*. Pearson Education, 2nd edition, 2008.

[4] Michael T. Heath. *Scientific Computing: An Introductory Survey*, revised second edition, volume 80 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 2018.

[5] Autar Kaw. *Holistic numerical methods: Multiple choice questions*. `http://mathforcollege.com/nm/assessment_text.html`. Accessed 18 March 2019.

[6] Gerald W. Recktenwald. *Introduction to Numerical Methods and MATLAB: Implementations and Applications*. Pearson, 2000.

[7] Maria Terrell and Robert Conelly. *GoodQuestions project*. Department of Mathematics, Cornell University, Ithaca, NY. `http://pi.math.cornell.edu/~GoodQuestions`. Accessed 5 April 2019.

[8] Holly Zullo and Kelly Cline. *MathQUEST/MathVote: Resources for clickers and classroom voting in collegiate mathematics*. Mathematics Program, Carroll College, Helena, MT. `http://mathquest.carroll.edu`. Accessed 5 April 2019.