# CHAPTER 1

# INTRODUCTORY CONCEPTS AND CALCULUS REVIEW

COPYRIGHTED MATERIAL

## 1.1 BASIC TOOLS OF CALCULUS

## Exercises:

1. Show that the third-order Taylor polynomial for $f(x) = (x+1)^{-1}$, about $x_0 = 0$, is

$$p_3(x) = 1 - x + x^2 - x^3.$$

**Solution:** We have $f(0) = 1$ and

$$f'(x) = -\frac{1}{(x+1)^2}, \quad f''(x) = \frac{2}{(x+1)^3}, \quad f'''(x) = -\frac{6}{(x+1)^4},$$

so that $f'(0) = -1$, $f''(0) = 2$, $f''' = -6$. Therefore,

$$\begin{aligned} p_3(x) &= f(0) + xf'(0) + \frac{1}{2}x^2 f''(0) + \frac{1}{6}x^3 f'''(x) \\ &= 1 + x(-1) + \frac{1}{2}x^2(2) + \frac{1}{6}x^3(-6) \\ &= 1 - x + x^2 - x^3. \end{aligned}$$

2. What is the third-order Taylor polynomial for $f(x) = \sqrt{x+1}$, about $x_0 = 0$?

*Solutions Manual to Accompany An Introduction to Numerical Methods and Analysis*, Third Edition.
James F. Epperson.
© 2021 John Wiley & Sons, Inc. Published 2021 by John Wiley & Sons, Inc.

**Solution:** We have $f(x_0) = 1$ and

$$f'(x) = \frac{1}{2(x+1)^{1/2}}, \quad f''(x) = -\frac{1}{4(x+1)^{3/2}}, \quad f'''(x) = \frac{3}{8(x+1)^{5/2}},$$

so that $f'(0) = 1/2$, $f''(0) = -1/4$, $f''' = 3/8$. Therefore

$$\begin{aligned} p_3(x) &= f(0) + xf'(0) + \frac{1}{2}x^2 f''(0) + \frac{1}{6}x^3 f'''(x) \\ &= 1 + x(1/2) + \frac{1}{2}x^2(-1/4) + \frac{1}{6}x^3(3/8) \\ &= 1 - (1/2)x - (1/8)x^2 + (1/16)x^3. \end{aligned}$$

3. What is the sixth-order Taylor polynomial for $f(x) = \sqrt{1+x^2}$, using $x_0 = 0$? Hint: Consider the previous problem.

4. Given that

$$R(x) = \frac{|x|^6}{6!}e^{\xi}$$

for $x \in [-1, 1]$, where $\xi$ is between $x$ and 0, find an upper bound for $|R|$, valid for all $x \in [-1, 1]$, that is independent of $x$ and $\xi$.

5. Repeat the above, but this time require that the upper bound be valid only for all $x \in [-\frac{1}{2}, \frac{1}{2}]$.

**Solution:** The only significant difference is the introduction of a factor of $2^6$ in the denominator:

$$|R(x)| \le \frac{\sqrt{e}}{2^6 \times 720} = 3.6 \times 10^{-5}.$$

6. Given that

$$R(x) = \frac{|x|^4}{4!}\left(\frac{-1}{1+\xi}\right)$$

for $x \in [-\frac{1}{2}, \frac{1}{2}]$, where $\xi$ is between $x$ and 0, find an upper bound for $|R|$, valid for all $x \in [-\frac{1}{2}, \frac{1}{2}]$, that is independent of $x$ and $\xi$.

7. Use a Taylor polynomial to find an approximate value for $\sqrt{e}$ that is accurate to within $10^{-3}$.

**Solution:** There are two ways to do this. We can approximate $f(x) = e^x$ and use $x = 1/2$, or we can approximate $g(x) = \sqrt{x}$ and use $x = e$. In addition, we can be conventional and take $x_0 = 0$, or we can take $x_0 \neq 0$ in order to speed convergence.

The most straightforward approach (in my opinion) is to use a Taylor polynomial for $e^x$ about $x_0 = 0$. The remainder after $k$ terms is

$$R_k(x) = \frac{x^{k+1}}{(k+1)!}e^{\xi}.$$

We quickly have that

$$|R_k(x)| \le \frac{e^{1/2}}{2^{k+1}(k+1)!}$$

and a little playing with a calculator shows that

$$|R_3(x)| \leq \frac{e^{1/2}}{16 \times 24} = 0.0043$$

but

$$|R_4(x)| \leq \frac{e^{1/2}}{32 \times 120} = 4.3 \times 10^{-4}.$$

So we would use

$$e^{1/2} \approx 1 + \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2}\right)^2 + \frac{1}{6}\left(\frac{1}{2}\right)^3 + \frac{1}{24}\left(\frac{1}{2}\right)^4 = 1.6484375.$$

To fourteen digits, $\sqrt{e} = 1.64872127070013$, and the error is $2.84 \times 10^{-4}$, much smaller than required.

8. What is the fourth-order Taylor polynomial for $f(x) = 1/(x+1)$, about $x_0 = 0$?

   **Solution:** We have $f(0) = 1$ and

$$f'(x) = -\frac{1}{(x+1)^2}, \quad f''(x) = \frac{2}{(x+1)^3}, \quad f'''(x) = -\frac{6}{(x+1)^4}, \quad f''''(x) = \frac{24}{(x+1)^5},$$

   so that $f'(0) = -1$, $f''(0) = 2$, $f''' = -6$, $f''''(0) = 24$. Thus,

$$p_4(x) = 1 + x(-1) + \frac{1}{2}x^2(2) + \frac{1}{6}x^3(-6) + \frac{1}{24}x^4(24) = 1 - x + x^2 - x^3 + x^4.$$

9. What is the fourth-order Taylor polynomial for $f(x) = 1/x$, about $x_0 = 1$?

10. Find the Taylor polynomial of third-order for $\sin x$, using:

    (a) $x_0 = \pi/6$.

    **Solution:** We have

$$f(x_0) = \frac{1}{2}, \quad f'(x_0) = \frac{\sqrt{3}}{2}, \quad f''(x_0) = -\frac{1}{2}, \quad f'''(x_0) = -\frac{\sqrt{3}}{2},$$

    so

$$p_3(x) = \frac{1}{2} + \frac{\sqrt{3}}{2}\left(x - \frac{\pi}{6}\right) - \frac{1}{4}\left(x - \frac{\pi}{6}\right)^2 - \frac{\sqrt{3}}{12}\left(x - \frac{\pi}{6}\right)^3;$$

    (b) $x_0 = \pi/4$;

    (c) $x_0 = \pi/2$.

11. For each function below construct the third-order Taylor polynomial approximation, using $x_0 = 0$, and then estimate the error by computing an upper bound on the remainder, over the given interval.

    (a) $f(x) = e^{-x}$, $x \in [0, 1]$;

    (b) $f(x) = \ln(1 + x)$, $x \in [-1, 1]$;

    (c) $f(x) = \sin x$, $x \in [0, \pi]$;

    (d) $f(x) = \ln(1 + x)$, $x \in [-1/2, 1/2]$;

(e) $f(x) = 1/(x+1)$, $x \in [-1/2, 1/2]$.

**Solution:**

(a) The polynomial is

$$p_3(x) = 1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3,$$

with remainder

$$R_3(x) = \frac{1}{24}x^4 e^{-\xi}.$$

This can be bounded above, for all $x \in [0, 1]$, by

$$|R_3(x)| \leq \frac{1}{24}e$$

(b) The polynomial is

$$p_3(x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3,$$

with remainder

$$R_3(x) = \frac{1}{4}x^4 \frac{1}{(1+\xi)^4}.$$

We *can't* bound this for all $x \in [-1, 1]$, because of the potential division by zero.

(c) The polynomial is

$$p_3(x) = x - \frac{1}{6}x^3,$$

with remainder

$$R_3(x) = \frac{1}{120}x^5 \cos\xi.$$

This can be bounded above, for all $x \in [0, \pi]$, by

$$|R_3(x)| \leq \frac{\pi^5}{120}.$$

(d) The polynomial is the same as in (b), of course,

$$p_3(x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3,$$

with remainder

$$R_3(x) = \frac{1}{4}x^4 \frac{1}{(1+\xi)^4}.$$

For all $x \in [-1/2, 1/2]$ this can be bounded by

$$R_3(x) \leq \frac{1}{4}(1/2^4)\frac{1}{(1-(1/2))^4} = \frac{1}{4}.$$

(e) The polynomial is

$$p_3(x) = 1 - x + x^2 - x^3,$$

with remainder

$$R_3(x) = x^4 \frac{1}{(1+\xi)^5}.$$

This can be bounded above, for all $x \in [-1/2, 1/2]$, by

$$|R_3(x)| \leq (1/2)^4 \frac{1}{(1-1/2)^5} = 2.$$

Obviously, this is not an especially good approximation.

12. Construct a Taylor polynomial approximation that is accurate to within $10^{-3}$, over the indicated interval, for each of the following functions, using $x_0 = 0$.

    (a) $f(x) = \sin x$, $x \in [0, \pi]$;
    (b) $f(x) = e^{-x}$, $x \in [0, 1]$;
    (c) $f(x) = \ln(1 + x)$, $x \in [-1/2, 1/2]$;
    (d) $f(x) = 1/(x + 1)$, $x \in [-1/2, 1/2]$;
    (e) $f(x) = \ln(1 + x)$, $x \in [-1, 1]$.

**Solution:**

(a) The remainder here is

$$R_n(x) = \frac{(-1)^{n+1}}{(2n+1)!} x^{2n+1} \cos c,$$

for $c \in [0, \pi]$. Therefore, we have

$$|R_n(x)| \leq \frac{1}{(2n+1)!} |\pi|^{2n+1} \leq \frac{\pi^{2n+1}}{(2n+1)!}.$$

Simple manipulations with a calculator then show that

$$\max_{x \in [0,\pi]} |R_6(x)| \leq 0.4663028067 \times 10^{-3}$$

but

$$\max_{x \in [0,\pi]} |R_5(x)| \leq 0.7370430958 \times 10^{-2}.$$

Therefore the desired Taylor polynomial is

$$p_{11}(x) = 1 - x + \frac{1}{6}x^3 - \frac{1}{120}x^5 - \frac{1}{7!}x^7 + \frac{1}{9!}x^9 + \frac{1}{11!}x^{11}.$$

(b) The remainder here is

$$R_n(x) = \frac{(-1)^{n+1}}{(n+1)!} x^{n+1} e^{-c},$$

for $c \in [0, 1]$. Therefore, we have

$$|R_n(x)| \leq \frac{1}{(n+1)!} |x|^{n+1} \leq \frac{1}{(n+1)!}.$$

Simple manipulations with a calculator then show that

$$\max_{x \in [0,1]} |R_6(x)| \leq 0.0001984126984$$

but

$$\max_{x\in[0,1]} |R_5(x)| \le 0.1388888889 \times 10^{-2}$$

Therefore the desired Taylor polynomial is

$$p_6(x) = 1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \frac{1}{24}x^4 - \frac{1}{120}x^5 + \frac{1}{720}x^6.$$

(c) $f(x) = \ln(1+x)$, $x \in [0, 3/4]$.

(d) **Solution:** The remainder is now

$$|R_n(x)| \le \frac{(1/2)^{n+1}}{(n+1)},$$

and $n = 8$ makes the error small enough.

(e) $f(x) = \ln(1+x)$, $x \in [0, 1/2]$.

13. Repeat the above, this time with a desired accuracy of $10^{-6}$.

14. Since

$$\frac{\pi}{4} = \arctan 1,$$

we can estimate $\pi$ by estimating $\arctan 1$. How many terms are needed in the Gregory series for the arctangent to approximate $\pi$ to 100 decimal places? 1,000? Hint: Use the error term in the Gregory series to predict when the error gets sufficiently small.

**Solution:** The remainder in the Gregory series approximation is

$$R_n(x) = (-1)^{n+1} \int_0^x \frac{t^{2n+2}}{1+t^2} dt,$$

so to get 100 decimal places of accuracy for $x = 1$, we require

$$|R_n(1)| = \left| \int_0^1 \frac{t^{2n+2}}{1+t^2} dt \right| \le \int_0^1 t^{2n+2} dt = \frac{1}{2n+3} \le 10^{-100},$$

thus, we have to take $n \ge (10^{100} - 3)/2$ terms. For 1,000 places of accuracy we therefore need $n \ge (10^{1000} - 3)/2$ terms.

Obviously, this is not the best procedure for computing many digits of $\pi$!

15. Elementary trigonometry can be used to show that

$$\arctan(1/239) = 4\arctan(1/5) - \arctan(1).$$

This formula was developed in 1706 by the English astronomer John Machin. Use this to develop a more efficient algorithm for computing $\pi$. How many terms are needed to get 100 digits of accuracy with this form? How many terms are needed to get 1,000 digits? Historical note: Until 1961, this was the basis for the most commonly used method for computing $\pi$ to high accuracy.

**Solution:** We now have two Gregory series, thus complicating the problem a bit. We have

$$\pi = 4\arctan(1) = 16\arctan(1/5) - 4\arctan(1/239).$$

Define $p_{m,n} \approx \pi$ as the approximation generated by using an $m$ term Gregory series to approximate $\arctan(1/5)$ and an $n$ term Gregory series for $\arctan(1/239)$. Then we have

$$p_{m,n} - \pi = 16R_m(1/5) - 4R_n(1/239),$$

where $R_k$ is the remainder in the Gregory series. Therefore,

$$\begin{aligned} |p_{m,n} - \pi| &\leq \left| 16(-1)^{m+1} \int_0^{1/5} \frac{t^{2m+2}}{1+t^2} dt - 4(-1)^{n+1} \int_0^{1/239} \frac{t^{2n+2}}{1+t^2} dt \right| \\ &\leq \frac{16}{(2m+3)5^{2m+3}} + \frac{4}{(2n+3)239^{2n+3}}. \end{aligned}$$

To finish the problem we have to apportion the error between the two series, which introduces some arbitrariness into the problem. If we require that they be equally accurate, then we have that

$$\frac{16}{(2m+3)5^{2m+3}} \leq \epsilon$$

and

$$\frac{4}{(2n+3)239^{2n+3}} \leq \epsilon.$$

Using properties of logarithms, these become

$$\log(2m+3) + (2m+3)\log 5 \geq \log 16 - \log \epsilon$$

and

$$\log(2n+3) + (2n+3)\log 239 \geq \log 4 - \log \epsilon.$$

For $\epsilon = (1/2) \times 10^{-100}$, these are satisfied for $m = 70$, $n = 20$. For $\epsilon = (1/2) \times 10^{-1000}$, we get $m = 712$, $n = 209$. Changing the apportionment of the error doesn't change the results by much at all.

16. In 1896, a variation on Machin's formula was found:

$$\arctan(1/239) = \arctan(1) - 6\arctan(1/8) - 2\arctan(1/57),$$

and this began to be used in 1961 to compute $\pi$ to high accuracy. How many terms are needed when using this expansion to get 100 digits of $\pi$? 1,000 digits?

**Solution:** We now have three series to work with, which complicates matters only slightly more compared to the previous problem. If we define $p_{k,m,n} \approx \pi$ based on

$$\pi = 4\arctan(1) = 24\arctan(1/8) + 8\arctan(1/57) + 4\arctan(1/239),$$

taking $k$ terms in the series for $\arctan(1/8)$, $m$ terms in the series for $\arctan(1/57)$, and $n$ terms in the series for $\arctan(1/239)$, then we are led to the inequalities

$$\log(2k+3) + (2k+3)\log 8 \geq \log 24 - \log \epsilon,$$

$$\log(2m+3) + (2m+3)\log 57 \geq \log 8 - \log \epsilon,$$

and

$$\log(2n+3) + (2n+3)\log 239 \geq \log 4 - \log \epsilon.$$

For $\epsilon = (1/3) \times 10^{-100}$, we get $k = 54$, $m = 27$, and $n = 19$; for $\epsilon = (1/3) \times 10^{-1000}$ we get $k = 552$, $m = 283$, and $n = 209$.

Note: In both of these problems a slightly more involved treatment of the error might lead to fewer terms being required.

17. What is the Taylor polynomial of order 3 for $f(x) = x^4 + 1$, using $x_0 = 0$?

    **Solution:** This is very direct:

$$f'(x) = 4x^3, \quad f''(x) = 12x^2, \quad f'''(x) = 24x,$$

    so that

$$p_3(x) = 1 + x(0) + \frac{1}{2}x^2(0) + \frac{1}{6}x^3(0) = 1.$$

18. What is the Taylor polynomial of order 4 for $f(x) = x^4 + 1$, using $x_0 = 0$? Simplify as much as possible.

19. What is the Taylor polynomial of order 2 for $f(x) = x^3 + x$, using $x_0 = 1$?

20. What is the Taylor polynomial of order 3 for $f(x) = x^3 + x$, using $x_0 = 1$? Simplify as much as possible.

    **Solution:** We note that $f'''(1) = 6$, so we have (using the solution from the previous problem)

$$p_4(x) = 3x^2 - 2x + 1 + \frac{1}{6}(x-1)^3(6) = x^3 + x.$$

    The polynomial is its own Taylor polynomial.

21. Let $p(x)$ be an arbitrary polynomial of degree less than or equal to $n$. What is its Taylor polynomial of degree $n$, about an arbitrary $x_0$?

22. The Fresnel integrals are defined as

$$C(x) = \int_0^x \cos(\pi t^2/2)dt,$$

    and

$$S(x) = \int_0^x \sin(\pi t^2/2)dt.$$

    Use Taylor expansions to find approximations to $C(x)$ and $S(x)$ that are $10^{-4}$ accurate for all $x$ with $|x| \leq \frac{1}{2}$. Hint: Substitute $x = \pi t^2/2$ into the Taylor expansions for the cosine and sine.

    **Solution:** We will show the work for the case of $S(x)$, only. We have

$$S(x) = \int_0^x \sin(\pi t^2/2)dt = \int_0^x p_n(t^2)dt + \int_0^x R_n(t^2)dt.$$

    Looking more carefully at the remainder term, we see that it is given by

$$r_n(x) = \pm \int_0^x \frac{t^{2(2n+3)}}{(2n+3)!} \cos \xi dt.$$

Therefore,

$$|r_n(x)| \le \int_0^{1/2} \frac{t^{2(2n+3)}}{(2n+3)!} dt = \frac{(1/2)^{4n+7}}{(4n+7)(2n+3)!}.$$

A little effort with a calculator shows that this is less than $10^{-4}$ for $n \ge 1$; therefore the polynomial is

$$p(x) = \int_0^x (t^2 - (1/6)t^6) dt = -\frac{x^7}{42} + \frac{x^3}{3}.$$

23. Use the Integral Mean Value Theorem to show that the "pointwise" form (1.3) of the Taylor remainder (usually called the *Lagrange* form) follows from the "integral" form (1.2) (usually called the *Cauchy* form).

24. For each function in Problem 11, use the Mean Value Theorem to find a value $M$ such that

$$|f(x_1) - f(x_2)| \le M|x_1 - x_2|$$

is valid for all $x_1$, $x_2$ in the interval used in Problem 11.

**Solution:** This amounts to finding an upper bound on $|f'|$ over the interval given. The answers are as given below.

(a) $f(x) = e^{-x}$, $x \in [0, 1]$; $M \le 1$.

(b) $f(x) = \ln(1 + x)$, $x \in [-1, 1]$; $M$ is unbounded, since $f'(x) = 1/(1 + x)$ and $x = -1$ is possible.

(c) $f(x) = \sin x$, $x \in [0, \pi]$; $M \le 1$.

(d) $f(x) = \ln(1 + x)$, $x \in [-1/2, 1/2]$; $M \le 2$.

(e) $f(x) = 1/(x + 1)$, $x \in [-1/2, 1/2]$. $M \le 4$.

25. A function is called *monotone* on an interval if its derivative is strictly positive or strictly negative on the interval. Suppose $f$ is continuous and monotone on the interval $[a, b]$, and $f(a)f(b) < 0$; prove that there is exactly one value $\alpha \in [a, b]$ such that $f(\alpha) = 0$.

**Solution:** Since $f$ is continuous on the interval $[a, b]$ and $f(a)f(b) < 0$, the Intermediate Value Theorem guarantees that there is a point $c$ where $f(c) = 0$, i.e., there is at least one root. Suppose now that there exists a second root, $\gamma$. Then $f(c) = f(\gamma) = 0$. By the Mean Value Theorem, then, there is a point $\xi$ between $c$ and $\gamma$ such that

$$f'(\xi) = \frac{f(\gamma) - f(c)}{\gamma - c} = 0.$$

But this violates the hypothesis that $f$ is monotone, since a monotone function must have a derivative that is strictly positive or strictly negative. Thus we have a contradiction, thus there cannot exist the second root.

A very acceptable argument can be made by appealing to a graph of the function.

26. Finish the proof of the Integral Mean Value Theorem (Theorem 1.5) by writing up the argument in the case that $g$ is negative.

**Solution:** All that is required is to observe that if $g$ is negative, then we have

$$\int_a^b g(t)f(t)dt \leq \int_a^b g(t)f_m dt = f_m \int_a^b g(t)dt,$$

and

$$\int_a^b g(t)f(t)dt \geq \int_a^b g(t)f_M dt = f_M \int_a^b g(t)dt.$$

The proof is completed as in the text.

27. Prove Theorem 1.6, providing all details.

28. Let $c_k > 0$, be given, $1 \leq k \leq n$, and let $x_k \in [a, b]$, $1 \leq k \leq n$. Then, use the Discrete Average Value Theorem to prove that, for any function $f \in C([a, b])$,

$$\frac{\sum_{k=1}^n c_k f(x_k)}{\sum_{k=1}^n c_k} = f(\xi),$$

for some $\xi \in [a, b]$.

**Solution:** We can't apply the Discrete Average Value Theorem to the problem as it is posed originally, so we have to manipulate a bit. Define

$$\gamma_j = \frac{c_j}{\sum_{k=1}^n c_k};$$

then

$$\sum_{j=1}^n \gamma_j = 1$$

and now we can apply the Discrete Average Value Theorem to finish the problem.

29. Discuss, in your own words, whether or not the following statement is true: "The Taylor polynomial of degree $n$ is the best polynomial approximation of degree $n$ to the given function near the point $x_0$."

◁ • • • ▷

## 1.2 ERROR, APPROXIMATE EQUALITY, AND ASYMPTOTIC ORDER NOTATION

# Exercises:

1. Use Taylor's Theorem to show that $e^x = 1 + x + \mathcal{O}(x^2)$ for $x$ sufficiently small.

2. Use Taylor's Theorem to show that $\frac{1-\cos x}{x} = \frac{1}{2}x + \mathcal{O}(x^3)$ for $x$ sufficiently small.

**Solution:** We can expand the cosine in a Taylor series as

$$\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 \cos \xi.$$

If we substitute this into $(1-\cos x)/x$ and simplify, we get

$$\frac{1-\cos x}{x} = \frac{1}{2}x - \frac{1}{24}x^3\cos\xi,$$

so that we have

$$\left|\frac{1-\cos x}{x} - \frac{1}{2}x\right| = \left|\frac{1}{24}x^3\cos\xi\right| \le \frac{1}{24}|x^3| = C|x^3|$$

where $C = 1/24$. Therefore, $\frac{1-\cos x}{x} = \frac{1}{2}x + \mathcal{O}(x^3)$.

3. Use Taylor's Theorem to show that

$$\sqrt{1+x} = 1 + \frac{1}{2}x + \mathcal{O}(x^2)$$

for $x$ sufficiently small.

**Solution:** We have, from Taylor's Theorem, with $x_0 = 0$,

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2(1+\xi)^{-3/2},$$

for some $\xi$ between 0 and $x$. Since

$$\left|\frac{1}{8}x^2(1+\xi)^{-3/2}\right| \le C|x^2|$$

for all $x$ sufficiently small, the result follows. For example, we have

$$\left|\frac{1}{8}x^2(1+\xi)^{-3/2}\right| \le \frac{1}{8} \times 2\sqrt{2}|x^2|$$

for all $x \in [-1/2, 1/2]$.

4. Use Taylor's Theorem to show that

$$(1+x)^{-1} = 1 - x + x^2 + \mathcal{O}(x^3)$$

for $x$ sufficiently small.

**Solution:** This time, Taylor's Theorem gives us that

$$(1+x)^{-1} = 1 - x + x^2 - x^3/(1+\xi)^4$$

for some $\xi$ between 0 and $x$. Thus, for all $x$ such that $|x| \le m$,

$$\left|(1+x)^{-1} - (1-x+x^2)\right| = \left|x^3/(1+\xi)^4\right| \le |x|^3/(1-m)^4 = C|x|^3,$$

where $C = 1/(1-m)^4$.

5. Show that

$$\sin x = x + \mathcal{O}(x^3).$$

6. Recall the summation formula

$$1 + r + r^2 + r^3 + \cdots + r^n = \sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r}.$$

Use this to prove that

$$\sum_{k=0}^{n} r^k = \frac{1}{1-r} + \mathcal{O}(r^{n+1}).$$

Hint: What is the *definition* of the $\mathcal{O}$ notation?

7. Use the above result to show that 10 terms ($k = 9$) are all that is needed to compute

$$S = \sum_{k=0}^{\infty} e^{-k}$$

to within $10^{-4}$ absolute accuracy.

**Solution:** The remainder in the 9 term partial sum is

$$|R_9| = \left| \frac{e^{-10}}{1-e^{-1}} \right| = 0.000071822 < 10^{-4}.$$

8. Recall the summation formula

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}.$$

Use this to show that

$$\sum_{k=1}^{n} k = \frac{1}{2}n^2 + \mathcal{O}(n).$$

9. State and prove the version of Theorem 1.7 which deals with relationships of the form $x = x_n + \mathcal{O}(\beta(n))$.

**Solution:** The theorem statement might be something like the following:

**Theorem:** *Let $x = x_n + \mathcal{O}(\beta(n))$ and $y = y_n + \mathcal{O}(\gamma(n))$, with $b\beta(n) > \gamma(n)$ for all $n$ sufficiently large. Then*

$$\begin{aligned} x + y &= x_n + y_n + \mathcal{O}(\beta(n) + \gamma(n)), \\ x + y &= x_n + y_n + \mathcal{O}(\beta(n)), \\ Ax &= Ax_n + \mathcal{O}(\beta(n)). \end{aligned}$$

*In the last equation, $A$ is an arbitrary constant, independent of $n$.*

The proof parallels the one in the text almost perfectly, and so is omitted.

10. Use the definition of $\mathcal{O}$ to show that if $y = y_h + \mathcal{O}(h^p)$, then $hy = hy_h + \mathcal{O}(h^{p+1})$.

11. Show that if $a_n = \mathcal{O}(n^p)$ and $b_n = \mathcal{O}(n^q)$, then $a_n b_n = \mathcal{O}(n^{p+q})$.

**Solution:** We have

$$|a_n| \le C_a |n^p|$$

and

$$|b_n| \le C_b |n^q|.$$

These follow from the definition of the $\mathcal{O}$ notation. Therefore,

$$|a_n b_n| \le C_a |n^p| |b_n| \le (C_a |n^p|)(C_b |n^q|) = (C_a C_b)|n^{p+q}|$$

which implies that $a_n b_n = \mathcal{O}(n^{p+q})$.

12. Suppose that $y = y_h + \mathcal{O}(\beta(h))$ and $z = z_h + \mathcal{O}(\beta(h))$, for $h$ sufficiently small. Does it follow that $y - z = y_h - z_h$ (for $h$ sufficiently small)?

13. Show that
$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + \mathcal{O}(h^2)$$
for all $h$ sufficiently small. Hint: Expand $f(x \pm h)$ out to the fourth order terms.

    **Solution:** This is a straight-forward manipulation with the Taylor expansions
$$f(x+h) = f(x) + hf'(x) + \frac{1}{2}h^2 f''(x) + \frac{1}{6}h^3 f'''(x) + \frac{1}{24}h^4 f''''(\xi_1)$$
and
$$f(x-h) = f(x) - hf'(x) + \frac{1}{2}h^2 f''(x) - \frac{1}{6}h^3 f'''(x) + \frac{1}{24}h^4 f''''(\xi_2).$$
Add the two expansions to get
$$f(x+h) + f(x-h) = 2f(x) + h^2 f''(x) + \frac{1}{24}h^4(f''''(\xi_1) + f''''(\xi_2)).$$
Now solve for $f''(x)$.

14. Explain, in your own words, why it is necessary that the constant $C$ in (1.8) be independent of $h$.

◁ • • • ▷

## 1.3 A PRIMER ON COMPUTER ARITHMETIC

# Exercises:

1. In each problem below, $A$ is the exact value, and $A_h$ is an approximation to $A$. Find the absolute error and the relative error.

   (a) $A = \pi$, $A_h = 22/7$;
   (b) $A = e$, $A_h = 2.71828$;
   (c) $A = \frac{1}{6}$, $A_h = 0.1667$;
   (d) $A = \frac{1}{6}$, $A_h = 0.1666$.

   **Solution:**

   (a) Abs. error $\leq 1.265 \times 10^{-3}$, rel. error $\leq 4.025 \times 10^{-4}$;
   (b) Abs. error $\leq 1.828 \times 10^{-6}$, rel. error $\leq 6.72 \times 10^{-7}$;
   (c) Abs. error $\leq 3.334 \times 10^{-5}$, rel. error $\leq 2.000 \times 10^{-4}$;
   (d) Abs. error $\leq 6.667 \times 10^{-5}$, rel. error $\leq 4 \times 10^{-4}$.

2. Perform the indicated computations in each of three ways: (i) Exactly; (ii) Using three-digit decimal arithmetic, with chopping; (iii) Using three-digit decimal arithmetic, with rounding. For both approximations, compute the absolute error and the relative error.

   (a) $\frac{1}{6} + \frac{1}{10}$;
   (b) $\frac{1}{6} \times \frac{1}{10}$;
   (c) $\frac{1}{9} + \left(\frac{1}{7} + \frac{1}{6}\right)$;
   (d) $\left(\frac{1}{7} + \frac{1}{6}\right) + \frac{1}{9}$.

3. For each function below explain why a naive construction will be susceptible to significant rounding error (for $x$ near certain values), and explain how to avoid this error.

   (a) $f(x) = (\sqrt{x+9} - 3)x^{-1}$;
   (b) $f(x) = x^{-1}(1 - \cos x)$;
   (c) $f(x) = (1-x)^{-1}(\ln x - \sin \pi x)$;
   (d) $f(x) = (\cos(\pi + x) - \cos \pi)x^{-1}$;
   (e) $f(x) = (e^{1+x} - e^{1-x})(2x)^{-1}$.

   **Solution:** In each case, the function is susceptible to subtractive cancellation which will be amplified by division by a small number. The way to avoid the problem is to use a Taylor expansion to make the subtraction and division both explicit operations. For instance, in (a), we would write

   $$f(x) = ((3+(1/6)x-(1/216)x^2+\mathcal{O}(x^3))-3)x^{-1} = (1/6)-(1/216)x+\mathcal{O}(x^2).$$

   To get greater accuracy, take more terms in the Taylor expansion.

4. For $f(x) = (e^x - 1)/x$, how many terms in a Taylor expansion are needed to get single precision accuracy (7 decimal digits) for all $x \in [0, \frac{1}{2}]$? How many terms are needed for double precision accuracy (14 decimal digits) over this same range?

5. Using single precision arithmetic, only, carry out each of the following computations, using first the form on the left side of the equals sign, then using the form on the right side, and compare the two results. Comment on what you get in light of the material in § 1.3.

   (a) $(x+\epsilon)^3 - 1 = x^3 + 3x^2\epsilon + 3x\epsilon^2 + \epsilon^3 - 1$, $x = 1.0$, $\epsilon = 0.000001$.
   (b) $-b + \sqrt{b^2 - 2c} = 2c(-b - \sqrt{b^2 - 2c})^{-1}$, $b = 1,000$, $c = \pi$.

   **Solution:** "Single precision" means 6 or 7 decimal digits, so the point of the problem is to do the computations using 6 or 7 digits.

   (a) Using MATLAB's `single` command on the author's laptop (running MATLAB R2019b), we get

   $$(x+\epsilon)^3 - 1 = 3.000002999797857 \times 10^{-6}$$

   but

   $$x^3 + 3x^2\epsilon + 3x\epsilon^2 + \epsilon^3 - 1 = 3.000003000019902 \times 10^{-6}.$$

(b) Using the same software and hardware, we get

$$-b + \sqrt{b^2 - 2c} = -0.003141597588410$$

but

$$2c\left(-b - \sqrt{b^2 - 2c}\right)^{-1} = -0.003141597588407.$$

What is interesting is how modern hardware and software have dramatically improved the results here. Earlier editions, which relied upon results using FORTRAN or C on a late 1990s Sun workstation, showed much more of a difference.

6. Consider the sum

$$S = \sum_{k=0}^{m} e^{-14(1-e^{-0.05k})}$$

where $m = 2 \times 10^5$. Again using only single precision, compute this two ways: First, by summing in the order indicated in the formula; second, by summing *backwards*, i.e., starting with the $k = 200,000$ term and ending with the $k = 0$ term. Compare your results and comment upon them.

7. (a) Using the computer of your choice, find three values $a$, $b$, and $c$, such that

$$(a + b) + c \neq a + (b + c).$$

(b) Repeat for your favorite calculator app.

(c) Do this for single precision in your preferred computing environment.

**Solution:** (a) The key issue is to get an approximation to the machine epsilon, then take $a = 1$, $b = c = (2/3)\mathbf{u}$ or something similar. This will guarantee that $(a + b) + c = a$ but $a + (b + c) > a$. There is an additional issue, in that MATLAB always rounds unformatted output, so to see that you got a different result you have to use (ugh!) `fprintf` to print out enough digits. On my laptop, I was able to use

$$\begin{aligned} a &= 1 \\ b &= 1.101642356786233 \times 10^{-16} \\ c &= 1.101642356786233 \times 10^{-16} \end{aligned}$$

and then `fprintf` told me that

$$\begin{aligned} D &= (a + b) + c = 1.00000000000000000, \\ E &= a + (b + c) = 1.00000000000000022. \end{aligned}$$

It is an interesting aspect of the history of this book, that when this exercise was first written ("a long time ago, in a computational environment far, far, away"), actual physical calculators were still commonplace (as opposed to smartphone/tablet apps). On an elderly Sharp calculator, circa 1997, the author found that $a = 1$, $b = 4 \times 10^{-10}$, and $c = 4 \times 10^{-10}$ worked. Using a scientific calculator app on his phone, the author found that $a = 1$, $b = 4 \times 10^{-16}$, and $c = 4 \times 10^{-16}$ worked.

(However, he was not able to get this to work on the Windows 10 calculator app. It would make an interesting quasi-research question to explain why.)

(b) Using MATLAB's `single` command (carefully), I used

$$\begin{aligned} a &= 1 \\ b &= 3.9572964 \times 10^{-8} \\ c &= b. \end{aligned}$$

Then,

$$\begin{aligned} D &= (a+b)+c = 1 \\ E &= a+(b+c) = 1.0000001. \end{aligned}$$

8. Assume we are using 3-digit decimal arithmetic. For $\epsilon = 0.0001$, $a_1 = 5$, compute

$$a_2 = a_0 + \left(\frac{1}{\epsilon}\right) a_1$$

for $a_0$ equal to each of 1, 2, and 3. Comment.

9. Let $\epsilon \leq \mathbf{u}$. Explain, in your own words, why the computation

$$a_2 = a_0 + \left(\frac{1}{\epsilon}\right) a_1$$

is potentially rife with rounding error. (Assume that $a_0$ and $a_1$ are of comparable size.) Hint: See previous problem.

**Solution:** This is just a generalization of the previous problem. If $\epsilon$ is small enough, then $a_2$ will be independent of $a_0$.

10. Using the computer and language of your choice, write a program to estimate the machine epsilon.

**Solution:** There are lots of ways to do this. The basic idea is to add a small number to 1, and then check to see if the result is different from one, otherwise continue on. One possible solution is the following:

**Algorithm 1.1**

***Computation of the machine epsilon.***

```
x = 1.e-10;
for k=1:6000
   y = 1 + x;
   if y <= 1
  disp('macheps = ')
   disp(x)
  break
 end
 x = x*.99;
end
x
```

This produces (on the author's laptop) $\mathbf{u} = 1.101642356786233 \times 10^{-16}$. If we change the initial `x` to 0.5, and decrement by a factor of 2 each step, we get $\mathbf{u} = 1.110223024625157 \times 10^{-16}$, which, being larger, is a better estimate. (Why?)

11. We can compute $e^{-x}$ using Taylor polynomials in two ways, either using

$$e^{-x} \approx 1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \dots$$

or using

$$e^{-x} \approx \frac{1}{1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \dots}.$$

Discuss, in your own words, which approach is more accurate. In particular, which one is more (or less) susceptible to rounding error?

**Solution:** Because of the alternating signs in the first approach, there is some concern about subtractive cancellation when it is used.

12. What is the machine epsilon for a computer that uses binary arithmetic, 24 bits for the fraction, and rounds? What if it chops?

**Solution:** Recall that the machine epsilon is the *largest* number $x$ such that the computer returns $1 + x = x$. We therefore need to find the largest number $x$ that can be represented with 24 binary digits such that $1 + x$, when rounded to 24 bits, is still equal to 1. This is perhaps best done by explicitly writing out the addition in binary notation. We have

$$\begin{aligned} 1 + x = {} & 1.000\ 0000\ 0000\ 0000\ 0000\ 0000_2 \\ & + 0.000\ 0000\ 0000\ 0000\ 0000\ 0000\ dddd\ dddd\ dddd\ dddd\ dddd\ dddd_2. \end{aligned}$$

If the machine chops, then we can set all of the $d$ values to 1 and the computer will still return $1 + x = 1$; if the machine rounds, then we need to make the first digit a zero. Thus, the desired values are

$$\mathbf{u}_{\text{round}} = \sum_{k=1}^{23} 2^{-k-24} = 0.596 \times 10^{-7},$$

and

$$\mathbf{u}_{\text{chop}} = \sum_{k=1}^{24} 2^{-k-23} = 0.119 \times 10^{-6}.$$

13. What is the machine epsilon for a computer that uses *octal* (base 8) arithmetic, assuming it retains 8 octol digits in the fraction?

◁ • • • ▷

## 1.4 A WORD ON COMPUTER LANGUAGES AND SOFTWARE

(No exercises in this section.)

## 1.5 A BRIEF HISTORY OF SCIENTIFIC COMPUTING

(No exercises in this section.)