Approximation and Errors

Rules for Determining Significant Digits

The following rules can be used to determine the number of significant digits in a number x's representation. We assume $x \geq 0$.

1. If $x = d_1 \cdots d_m.e_1 \cdots e_n$, where $d_1 \neq 0$, then x has $m + n$ significant digits.
2. If $x = 0.0 \cdots 0 d_1 d_2 \cdots d_n$, where $d_1 \neq 0$, then x has n significant digits.
3. If $x = a \times 10^n$, where a is a nonnegative real number, then the number of significant digits of x equals the number of significant digits of a.

Example 1. Provide the number of significant digits for each representation.

1. 2340000
2. 0.02965
3. 1.011
4. 2.23×10^3
5. 9.569×10^2
6. $2,314 \times 10^5$
7. 200.000
8. 30.001

Sources of Numerical Error

Round-off Error the error incurred when representing a number with fewer digits than are required to completely capture its exact numerical value.

Truncation Error the error incurred when limiting the accuracy and/or running-time of a procedure that requires an impractical amount of time (possibly infinite) in order to obtain the exact true answer.

Suppose true value x is approximated as y. Then we have the following definitions with respect to x and y.

True Error $E_t = x - y$

Absolute True Error $|E_t| = |x - y|$

Relative True Error $\epsilon_t = (x - y)/x = 1 - y/x$

Relative Absolute True Error $|\epsilon_t| = |(x - y)/x| = |1 - y/x|$

Example 2. Determine E_t, and ϵ_t when approximating the derivative of $f(x) = -x^2 + 5x$ at $x = 2$, using the approximation formula

$$f'(x) \approx \frac{f(x+h) - f(x)}{h},$$

with $h = 0.1$.

Example 2 Solution. $f'(x) = -2x + 5$, and so $f'(2) = 1$. On the other hand,

$$\frac{f(x+h) - f(x)}{h} = \frac{(-(2.1)^2 + 5(2.1)) - (-(2)^2 + 5(2))}{0.1} = 0.9.$$

Therefore, $E_t = 1 - 0.9 = 0.1$, and $\epsilon_t = 0.1/1 = 0.1$. □

Note: sometimes we prefer to discuss the **percentage relative error** which is defined as $100 \times \epsilon_t$.

Example 3. Determine E_t, and ϵ_t when approximating the integral $\int_0^3 2e^{0.1x}dx$ using 10 rectangles under the graph of $f(x) = 2e^{0.1x}$.

Example 3 Solution. Each rectangle has a width of $(3-0)/10 = 0.3$. Moreover, the rectangle heights are

$$f(0), f(0.3), f(0.6), \ldots, f(2.7) = 2, 2.06091, 2.12367, \ldots, 2.61993.$$

Summing these and multiplying by 0.3 yields 6.89274.

Moreover $\int_0^3 2e^{0.1x}dx = 20e^{0.1x}\,|_0^3 = 20(e^{0.3}-1) = 6.99718$. Therefore, $E_t = 0.10443$, while $\epsilon_t = 0.014925$ for an approximately 1.5% relative error. □

Example 4. Provide a general formula for the true and relative true error when approximating the derivative of $f(x) = -x^2 + 5x$ at $x = 2$ using an arbitrary value for h. Determine the largest value of h that can achieve an absolute relative error of no more than 0.5%.

Approximation Error

True error and relative true error do not seem very practical, since the true value is usually not known. Instead, we may compute an approximate error by comparing one approximation with a previous one.

Suppose a numerical value v is first approximated as x, and then is subsequently approximated by y. Then the **approximate error**, denoted E_a, in approximating v as y is defined as $E_a = x - y$. Similarly, the **relative approximate error**, denoted ϵ_a, is defined as $\epsilon_a = (x - y)/x = 1 - y/x$.

Example 5. Recall the function $f(x)$ from Example 2. If a second approximation of $f'(2)$, uses $h = 0.05$, then the approximate error is $E_a = 0.9 - 0.95 = -0.05$, while the relative approximate error is $\epsilon_a = -0.05/0.9 = -0.0555556$.

Taylor Series

Given function $f(x)$, and a real value a for which the n th derivative $f^{[n]}(a)$ is defined for all $n \geq 0$ (note: $f^0(a) \equiv f(a)$), the **Talylor series** of f about a is defined as

$$\sum_{n=0}^{\infty} f^{[n]}(a) \frac{(x-a)^n}{n!}.$$

Moreover, the **radius of convergence** of the Taylor series is defined as the largest $r \geq 0$ such that, for all $x \in (a-r, a+r)$, the series converges. Note that a Taylor series is called a **Maclaurin series** in case $a = 0$.

Example 6. Determine the Taylor series for each of the following functions and values of a: i) $f(x) = e^x$, $a = 0$, ii) $f(x) = \sin x$, $a = 0$, iii) $f(x) = \cos x$, $a = 0$, and iv) $f(x) = \ln x$, $a = 1$.

The n th partial sum, $n \geq 0$, of a Taylor series, denoted $S_{f,n}(x)$ is defined as

$$S_{f,n}(x) = f(a) + f'(a)(x-a) + \cdots + f^{[n]}(a)(x-a)^n/n!.$$

Remainder Theorem. Suppose the Taylor series of $f(x)$ about a converges for some x. Then

$$|E_t| = |R_{f,n}(x)| = |f(x) - S_{f,n}(x)| = |\frac{f^{[n+1]}(\xi)}{(n+1)!}(x-a)^{n+1}|,$$

where ξ is a number between a and x. Here, $R_{f,n}(x)$ is referred to as the **nth remainder**.

Example 7. Determine the true error when approximating $f(x) = \sin(0.5)$ using $S_{f,3}(0.5)$. Also determine an upper bound for the absolute true error as determined by $|R_{f,3}(0.5)|$. Finally, determine the absolute relative approximate error in approximating $\sin(0.5)$ with $S_{f,5}(0.5)$ instead of $S_{f,3}(0.5)$.

Example 7 Solution. $\sin(0.5) = 0.479426$, while

$$S_{f,3}(0.5) = 0.5 - (0.5)^3/6 = 0.479167.$$

Thus, $E_t = 2.58872 \times 10^{-4}$. Also,

$$|R_{f,3}(0.5)| \leq f^4(\xi)(0.5)^4/24 = \sin(\xi)(0.5)^4/24 \leq 0.479426(0.5)^4/24 = 0.00124850 = 1.24850 \times 10^{-3}.$$

Finally,

$$S_{f,5}(0.5) = 0.5 - (0.5)^3/6 + (0.5)^5/120 = 0.479427.$$

Therefore, $|E_a| = (0.5)^5/120 = 0.000260417$ and $|\epsilon_a| = 0.0125000$ for an error of 1.25%. □

Error Propagation

Suppose x is an approximation of some value v, in which the absolute true error is bounded by $\delta \geq 0$. This error can propagate once a function is applied to x. For example, if x is multiplied by 5, then the absolute true error with respect to $5v$ and $5x$ is now 5δ.

Example 8. Suppose values $x = 3.11034$, $y = 7.76436$, and $z = 1.45981$ have respective absolute true errors of $\delta_1 = 0.01$, $\delta_2 = 0.05$, and $\delta_3 = 0.025$. Then provide a bound on the absolute tree error iherent in the expression $(xy)/z$.

Example 8 Solution. The upper bound on the true value is

$$(x + \delta_1)(y + \delta_2)/(z - \delta_3) = 16.99421.$$

while the lower bound is

$$(x - \delta_1)(y - \delta_2)/(z + \delta_3) = 16.10788.$$

Finally, $(xy/z) = 16.54311$. Thus, the absolute true error in this expression is bounded by

$$\max(|16.54311 - 16.99421|, |16.54311 - 16.10788|) = 0.4511$$

□

At times it may seem difficult to compute an exact bound of the true error that is induced by evaluating an expression, with approximation inputs When this happens we may obtain a **first-order** approximation of the bound as follows.

Exercises

Note: for this and all subsequent assignments round all answers to 6 significant digits. Note: for numbers such as 6.01000, you may simply write 6.01.

1. Use the formula
$$\frac{f(x+h)-f(x)}{h}$$
to approximate the derivative of $f(x) = 3x^2$ at $x = 1$ using $h = 0.1$. Compute both the absolute true error $|E_t|$, and absolute relative true error $|\epsilon_t|$.

2. Provide a general formula for determining both the absolute true error and absolute relative true error when approximating the derivative of $f(x) = x^2$ at $x = a$ using a value h in the expression
$$\frac{(a+h)^2-a^2}{h}.$$

3. Using ϵ_t from the previous exercise, what is the greatest value of h that can be used to approximate the derivative of $f(x) = x^2$ at $x = 4$ with an error of no more than 1%?

4. Determine the absolute true error $|E_t|$, and absolute relative true error $|\epsilon_t|$ that occurs when approximating the integral $\int_0^3 2x\mathrm{d}x$ using three rectangles, each with width $\Delta x = 1$, and for which the heights are $f(0)$, $f(1)$, and $f(2)$, where $f(x) = 2x$.

5. Repeat the previous exercise, but now assume an approximation that uses n rectangles, each having width $\Delta x = 3/n$ and where the height of the i th rectangle is $f(i\Delta x)$, $i = 0, \ldots, n-1$.

6. Using ϵ_t from the previous exercise, what is the least value of n that can be used to approximate $\int_0^3 2x\mathrm{d}x$ with an error of no more than 1%?

7. Provide a formula for the n th term of the Taylor series for $f(x) = \ln x$ about the point $a = 1$.

8. Provide a formula for the n th term of the Taylor series for $f(x) = \sqrt{x}$ about the point $a = 1$.

9. Determine the relative approximate error when approximating e^{-1} using a fifth-degree Taylor polynomial with respect to e^x, compared with using a sixth-degree Taylor polynomial.

10. Make a table with the following columns: i) The value of $n = 0, 1 \ldots,$, ii) the approximation of cos(1 rad) using $P_n(1)$, the n th degree Taylor polynomial with respect to $\cos(x)$, iii) the approximate error in approximating $P_n(1)$ with $P_{n-1}(1)$, and iv) the relative approximate error. Continue the table until the relative approximate error drops below 1%. Hint: you may skip over the odd values of n since only even-degree terms are nonzero.

11. Compute an upper bound for the Taylor series remaininder $R_7(x+h) = (h^8 \cos^{[8]}(c))/8!$, where $x = 0$, $h = 1$, and c is some number in the closed interval $[0, 1]$.

12. How many significant digits does each of the following numbers have?

    a. 185000

b. 0.0185

c. 1.0185

d. 1.85×10^3

e. 1.850×10^2

f. 1.8500×10^{-2}

g. 100.00

h. 100.001

13. Let $v > 0$ be a true value, and $a > 0$ be an approximation of v. Prove that $|\epsilon_t|$ is invariant with respect to scalar multiplication. In other words, for scalar $c > 0$, the absolute relative true error for ca approximating cv is equal to $|\epsilon_t|$, the absolute relative true error for a approximating v.

14. Prove that if $|\epsilon_t| \leq 0.5 \times 10^{-m}$, then the true value v and the approximate value a are equal at the first m significant digits. Hint 1: you may assume $v > 0$ and $a > 0$. Hint 2: use the previous exercise.

15. Prove or disprove the converse of the statement in the previous exercise. In other words, if the true value and approximation agree in the first m digits, is it necessarily true that $|\epsilon_t| \leq 0.5 \times 10^{-m}$?

16. Suppose your approximation yields an absolute relative true error of 0.003%. How many significant digits of your approximation are guranteed to be accurate.

17. Suppose $x \geq 0$ and $y \geq 0$ are approximate values with respective true errors δ_1 and δ_2. Determine the true error inherent in the product xy. Compare your answer with the answer

$$\frac{\partial f}{\partial x}\delta_1 + \frac{\partial f}{\partial y}\delta_2,$$

where $f(x, y) = xy$.

18. Consider a sequence of numbers x_n, $n \geq 0$, that satisfies the equation $ax_n + bx_{n-1} = 0$, where $a, b \neq 0$, for all $n \geq 1$. Show that this equation is satisfied by $x_n = (-b/a)^n$.

19. Suppose x_n is an inceasing sequence of numbers with the property that $(x_n - x_{n-1})/x_n = c$, where $0 < c < 1$ is a constant. In other words, the relative approximation error is constant. Show that x_n does not converge. Hint: use the previous exercise.

20. The formula for strain S on a longitudinal bar is given by $S = F/(AE)$, where F is the applied force, A is the cross-sectiional area, and E is Young's modulus. If $F = 50 \pm 0.50$ N, $A = 0.2 \pm 0.002$ m^2, and $E = 210 \times 10^9 \pm 1 \times 10^9$ Pa, determine a first-order approximation of the maximum error in measuring S.Compare your approximation with the actual maximum true error.

Exercise Hints and Answers

1. $|E_t| = 0.3$, $|\epsilon_t| = 0.3/6 = 0.05$

2. $|E_t| = h$, $|\epsilon_t| = h/2a$.

3. Given that $a = 4$, We need $h/(2)(4) \leq 0.01$ which implies $h \leq 0.08$. Therefore, $h = 0.08$ is the least h-value that can obtain a relative error of at most 1%.

4. We have
$$\int_0^3 2x\mathrm{d}x = x^2\,|_0^3 = 9 - 0 = 9,$$
where as the apprixmation is
$$0 + 2 + 4 = 6,$$
yielding an absolute true error of $|E_t| = 3$ and absolute relative error of $|\epsilon_t| = 3/9 = 0.333333$.

5. The approximation is
$$\sum_{i=0}^{n-1} \frac{3}{n}(2i\frac{3}{n}) = \frac{18}{n^2}\sum_{i=0}^{n-1} i = \frac{18}{n^2}\frac{n(n-1)}{2} = 9(n-1)/n = 9(1-1/n).$$
The absolute true error is thus equal to $|E_t| = 9 - (9 - 9/n) = 9/n$, while the absolute relative error equals $|\epsilon_t| = 1/n$.

6. Since $|\epsilon_t| = 1/n$, $n = 100$ gives a relative error of 0.01, or 1%.

7. For $n \geq 1$, we have
$$f^{[n]}(x) = \frac{(-1)^{n+1}(n-1)!}{x^n}.$$
Therefore, the n th term is
$$f^{[n]}(x) = \begin{cases} 0 & \text{if } n = 0 \\ \frac{(-1)^{n+1}(x-1)^n}{n} & n \geq 1 \end{cases}$$

8. For $n \geq 2$, we have
$$f^{[n]}(x) = \frac{(-1)^{n+1}(1 \cdot 3 \cdot \cdots \cdot (2n-1))}{2^n x^{\frac{1}{2}-n}}.$$
Therefore, the n th term is
$$f^{[n]}(x) = \begin{cases} 1 & \text{if } n = 0 \\ \frac{1}{2}(x-1) & \text{if } n = 1 \\ \frac{(-1)^{n+1}(1\cdots(2n-3))(x-1)^n}{n!2^n} & n \geq 2 \end{cases}$$

9. We have
$$1 + x/1 + x^2/2 + x^3/6 + x^4/24 + x^5/120 = 0.3666667,$$
while
$$1 + x/1 + x^2/2 + x^3/6 + x^4/24 + x^5/120 + x^6/720 = 0.3680556,$$
which yields $\epsilon_a = (0.3680556 - 0.3666667)/0.3680556 = 0.003773615$ which yields an error of less than 1%.

10. The table is shown below.

n	$P_n(1)$	E_a	ϵ_a
0	1	na	na
2	0.5	-0.5	-1.00
4	0.5416667	0.0416667	0.07692313
6	0.5402778	-0.0013889	-0.002570715

11. Function $\cos^{[8]}(x) = \cos(x)$ has an upper bound of 1. Thus $R_7(1)$ has an upper bound of $1/8! = 0.0000248015$.

12. We have

    a. 185000 has 6

    b. 0.0185 has 3

    c. 1.0185 has 5

    d. 1.85×10^3 has 3

    e. 1.850×10^2 has 4

    f. 1.8500×10^{-2} has 5

    g. 100.00 has 5

    h. 100.001 has 6

13. Letting ϵ'_t denote the absolute relative true error for ca approximating cv, we have

$$\epsilon'_t = |cv - ca|/|cv| = c|v - a|/cv = |v - a|/v = \epsilon_t.$$

14. By the previous exercise, we may scale both v and a so that v has the form $0.d_1d_2\cdots$, where $d_1 \in \{1, \ldots, 9\}$. Now suppose that a differs from v at digit d_j, where $1 \leq j \leq m$. Then

$$|\epsilon_t| \geq 10^{-j} > 0.5 \times 10^{-m},$$

    a contradiction. Therefore v and a agree in the first m digits.

15. Consider $v = 0.19$ and $a = 0.10$, These numbers agree in the first $m = 1$ significant digits. However,

$$\epsilon_t = 0.09/0.19 = 0.47 > 0.5 \times 10^{-1} = 0.05.$$

    Therefore, the statement is not always true.

16. Dividing 0.003 by 100 yields $0.00003 = 0.3 \times 10^{-4}$. Therefore, the first four significant digits are guaranteed accurate.

17. The true value is

$$(x + \delta_1)(y + \delta_2) = (xy + x\delta_2 + y\delta_1 + \delta_1\delta_2,$$

    which gives a true error of

$$x\delta_2 + y\delta_1 + \delta_1\delta_2,$$

    Moreover, $\frac{\partial f}{\partial x} = y$ and $\frac{\partial f}{\partial y} = x$, which yields a first-order approximation of

$$y\delta_1 + x\delta_2,$$

    and so the first-order true-error approximation differs from the actual true error by $\delta_1\delta_2$.

18. We have

$$a(-b/a)^n + b(-b/a)^{n-1} = (-b/a)^{n-1}(a(-b/a) + b) = (-b/a)^{n-1}(-b + b) = (-b/a)^{n-1}(0) = 0.$$

19. We have $(x_n - x_{n-1})/x_n = c$, which implies that

$$x_n - x_{n-1} = cx_n,$$

which yields the equation

$$(1 - c)x_n - x_{n-1} = 0.$$

But from the previous exercise we kow that $(1/(1-c))^n$ is a solution to this equation Moreover, $1/(1 - c) > 1$ which implies

$$\lim_{n\to\infty} (1/(1 - c))^n = \infty.,$$

and so x_n diverges.

20. We have

$$\frac{\partial S}{\partial F} = 1/(AE),\ \frac{\partial S}{\partial A} = -F/(A^2E),\ \text{and}\ \frac{\partial S}{\partial E} = -F/(AE^2).$$

Moreover, evaluating these derivatives at $(50, 0.2, 210\times10^9)$ yields 2.380952×10^{-11}, -5.952381×10^{-9}, and $-5.668934 \times 10^{-21}$, respectively. Therefore, the first order approximation of the maximum error is

$$(2.380952\times10^{-11})(0.50)+(5.95238\times10^{-9})(0.002)+(5.668934\times10^{-21})(1\times10^9) = 1.366213\times10^{-10}.$$

Finally, the actual maximum true error is

$$50.5/((0.198)209 \times 10^9) - 50/((.2)(210 \times 10^9)) = 2.986115 \times 10^{-11}.$$