Error: Sources, Propagation & Analysis

S. Natesan

I.I.T. Guwahati

August 4, 2017

Outline

Floating-Point Numbers

Scientific calculations are usually carried out in floating-point arithmetic.

An *n*-digit floating point number in base β has the form

$$x = \pm(\cdot d_1 d_2 \cdots d_n)_\beta \, \beta^e$$

where $(\cdot d_1 d_2 \cdots d_n)$ is a β-fraction called the **mantissa**, and e is an integer called the **exponent**. Such a floating-point number is said to be **normalized** in case $d_1 \neq 0$, or else $d_1 = d_2 = \cdots = d_n = 0$.

For most computers, $\beta = 2$, although on some, $\beta = 16$, and in hand calculations and on calculators $\beta = 10$.

Floating-Point Numbers

The precision or length n of floating-point numbers on any particular computer is usually determined by the word length of the computer and may therefore vary widely. They are **single precision, double precision**.

Calculations in double precision usually doubles the storage requirements and more than doubles running time as compared with single precision.

The exponent e is limited to a range

$$m < e < M,$$

for certain integers m and M. In general, $m = -M$.

Floating-Point Numbers

There are two commonly used ways of translating a given real number x into an n β-digit floating-point number $fl(x)$, rounding and chopping.

In **rounding** $fl(x)$ is chosen as the normalized floating-point number nearest x; some special rule, such as symmetric rounding (rounding to an even digit), is used in case of a tie.

In **chopping** $fl(x)$ is chosen as the nearest normalized floating-point number between x and 0.

If, for example, two-decimal digit floating-point numbers are used, then

$$fl\left(\frac{2}{3}\right) = \begin{cases} 0.67 \times 10^0, & \text{rounded} \\ 0.66 \times 10^0, & \text{chopped} \end{cases}$$

Floating-Point Numbers

And

$$fl(-838) = \begin{cases} -0.84 \times 10^3, & \text{rounded} \\ -0.83 \times 10^3, & \text{chopped} \end{cases}$$

On some computers, this definition of $fl(x)$ is modified in case

$$|x| \geq \beta^M \quad \textbf{(overflow)}$$

or

$$0 < |x| \leq \beta^{m-n}, \quad \textbf{(underflow)},$$

where m and M are the bounds on the exponents.

In this case, either $fl(x)$ is not defined causing a stop, or else $fl(x)$ is represented by a special number which is not subject to the usual rules of arithmetic when combined with ordinary floating-point numbers.

Round–off Error

The difference between x and $fl(x)$ is called the **round-off error**. The round-off error depends on the size of x and is therefore, best measured relative to x. For if we write

$$fl(x) = x(1+\delta), \tag{1.1}$$

where $\delta = \delta(x)$ is some number depending on x, then it is possible to *bound* δ *independently* of x, at least as long as x causes no overflow or underflow.

For such an x, it is not difficult to show that

$$|\delta| < \frac{1}{2}\beta^{1-n}, \quad \text{in rounding}$$

while

$$-\beta^{1-n} < \delta \leq 0, \quad \text{in chopping.}$$

The maximum possible value for $|\delta|$ is often called the **unit round-off** and is denoted by u.

Round–off Error

If x is **rounded** so that $\widetilde{x}$ is the n–digit approximation to it, then

$$|x - \widetilde{x}| \leq \frac{1}{2} \times 10^{-n}. \tag{1.2}$$

If the $(n+1)$st digit of x is $0, 1, 2, 3$ or 4, then $x = \widetilde{x} + \varepsilon$, with $\varepsilon < \frac{1}{2}10^{-n}$, and therefore, the inequality (1.2) follows.

If $(n+1)$st digit of x is $5, 6, 7, 8$ or 9, then $\widetilde{x} = \widehat{x} + 10^{-n}$, where $\widehat{x}$ is a number with the same n digits as x and all digits beyond the nth are 0.

Now, $x = \widehat{x} + \delta \times 10^{-n}$, with $\delta \geq \frac{1}{2}$ and $\widetilde{x} - x = (1 - \delta) \times 10^{-n}$.

Since $1 - \delta \leq \frac{1}{2}$. Inequality (1.2) follows. ■

Round–off Error

If x is a decimal number, the **chopped or truncated** n–digit approximation to it is the number $\widehat{x}$ obtained by simply discarding all digits beyond the nth. For it, we have

$$|x - \widetilde{x}| \ < \ 10^{-n}. \tag{1.3}$$

The relationship between x and $\widetilde{x}$ is such that $x - \widetilde{x}$ has 0 in the first n places and $x = \widetilde{x} + \delta \times 10^{-n}$, with $0 \leq \delta \leq 1$. Hence, we have

$$|x - \widetilde{x}| \ = |\delta| \times 10^{-n} \ < \ 10^{-n}.$$

The inequality (1.3) follows.

Normalized Scientific Notation

In the decimal system, any real number can be expressed in **normalized scientific notation**.

This means that the decimal point is shifted and appropriate powers of 10 are supplied so that all the digits are to the right of the decimal point and first digit displayed is not 0.

For example,

$$732.5051 = 0.7325051 \times 10^3 = 0.7325051e + 03$$
$$-0.005612 = -0.5612^{-2} = -0.5612e - 02.$$

In general, a nonzero real number x can be expressed as

$$x = \pm r \times 10^n = \pm r\, e + 0n,$$

where r is a number in the range $\frac{1}{10} \leq r < 1$, and n is an integer (positive, negative, or zero).

Floating–Point Error

When an arithmetic operation is applied to two floating-point numbers, the result usually fails to be a floating-point number of the same length.

If, for example, we deal with two-decimal digit numbers and

$$x = 0.20 \times 10^1 = 2, \quad y = 0.77 \times 10^{-6}, \quad z = 0.30 \times 10^1 = 3,$$

then

$$x+y = 0.200000077 \times 10^1, \quad x \cdot y = 0.154 \times 10^{-5}, \quad \frac{x}{z} = 0.666 \cdots \times 10^0.$$

Hence, if ω denotes one of the arithmetic operations (addition, subtraction, multiplication, or division) and ω^* denotes the floating-point operation of the same name provided by the computer, then, we can be sure that usually

$$x\omega^* y \neq x\omega y.$$

Backward Error Analysis

We can express the floating-point operation ω^* as

$$x\omega^* y = fl(x\omega y). \tag{2.1}$$

In words, the floating-point sum of two floating-point numbers usually equals the floating-point number which represents the exact sum of the two numbers. Hence (unless overflow or underflow occurs) we have

$$x\omega^* y = (x\omega y)(1+\delta), \qquad \text{for some } |\delta| \leq u \tag{2.2}$$

where u is the unit round-off.

In certain situation, it is more convenient to use the equivalent formula

$$x\omega^* y = (x\omega y)/(1+\delta), \qquad \text{for some } |\delta| \leq u. \tag{2.3}$$

Equations (2.2)-(2.3) express the basic idea of backward error analysis.

Backward Error Analysis

Example

For example, consider the value of the function

$$f(x) = x^{2^n}.$$

At a point x_0 can be calculated by n squaring, *i.e.,* by carrying out the sequence of steps

$$x_1 := x_0^2,\ x_2 := x_1^2,\ \cdots,\ x_n := x_{n-1}^2$$

with $f(x_0) = x_n$.

In floating-point arithmetic, we compute the sequence of numbers

$$\widehat{x}_1 = x_0^2(1+\delta_1),\ \widehat{x}_2 = (\widehat{x}_1)^2(1+\delta_2), \cdots, \widehat{x}_n = (\widehat{x}_{n-1})^2(1+\delta_n),$$

with $|\delta_i| \leq u$ for all i.

Backward Error Analysis

The computed answer is, therefore,

$$\widehat{x}_n = x_0^{2^n}(1+\delta_1)^{2^{n-1}} \cdots (1+\delta_{n-1})^2(1+\delta_n).$$

To simplify this expression, we observe that, if $|\delta_1|, \cdots, |\delta_r| \leq u$, then

$$(1+\delta_1)\cdots(1+\delta_r) = (1+\delta)^r, \text{ for some } \delta \text{ with } |\delta| \leq u.$$

Also, then

$$(1+\delta)^r = (1+\eta)^{r+1}, \quad \text{for some } |\eta| \leq u.$$

Consequently,

$$\widehat{x}_n = x_0^{2^n}(1+\delta)^{2^n} = f(x_0(1+\delta)), \text{ for some } |\delta| \leq u.$$

In words, the computed vale $\widehat{x}_n$ for $f(x_0)$ is the exact value of $f(x)$ at the perturbed argument $x = x_0(1+\delta)$.

Backward Error Analysis

Example

consider calculation of the number s from the equation

$$a_1 b_1 + \cdots + a_r b_r + a_{r+1} s = c. \tag{2.4}$$

We use the formula

$$s = \frac{1}{a_{r+1}} \left(c - \sum_{k=1}^{r} a_k b_k \right).$$

If we obtain s through the steps

$$\begin{aligned} s_0 &:= c \\ s_i &:= s_{i-1} - a_i b_i, \quad i = 1, \cdots, r \\ s &:= s_r / a_{r+1} \end{aligned}$$

Backward Error Analysis

The corresponding numbers computed in floating-point arithmetic satisfy

$$\widehat{s}_0 = c$$

$$\widehat{s}_i = [\widehat{s}_{i-1} - a_i b_i(1+\delta)](1+\delta), \quad i = 1, \cdots, r$$

$$\widehat{s} = \widehat{s}_r/[a_{r+1}(1+\delta)]$$

Here, we have used Eqns. (2.2) and (2.3), and did not use the subscripts for δ. Consequently,

$$\begin{aligned} a_{r+1}(1+\delta)\widehat{s} &= \widehat{s}_r \\ &= \widehat{s}_{r-1}(1+\delta) - a_r b_r(1+\delta)^2 \\ &= \widehat{s}_{r-2}(1+\delta)^2 - a_{r-1}b_{r-1}(1+\delta)^3 - a_r b_r(1+\delta)^2 \\ &\vdots \\ &= \widehat{s}_0(1+\delta)^r - a_1 b_1(1+\delta)^{r+1} - \cdots - a_r b_r(1+\delta)^2. \end{aligned}$$

Backward Error Analysis

That is,

$$a_1 b_1 (1+\delta)^{r+1} + \cdots + a_r b_r (1+\delta)^2 + a_{r+1}(1+\delta)\widehat{s} = \widehat{s}_0 (1+\delta)^r.$$

This shows that the computed value $\widehat{s}$ for s satisfies the perturbed equation

$$a_1 b_1 (1+\delta)^{r+1} + \cdots + a_r b_r (1+\delta)^2 + a_{r+1}(1+\delta)\widehat{s} = c(1+\delta)^r.$$

Note that we can reduce all exponents by 1 in case $a_{r+1} = 1$, that is, in case the last division need not be carried out.

Absolute and Relative Errors

If the number x^* is an approximation to the exact answer x, then we call the difference as **error**,

$$\text{Error } = x - x^*,$$

and therefore,

$$\text{Exact} = \text{approximation} + \text{error}.$$

The **absolute error** is defined as

$$|x - x^*|.$$

The **relative error** is defined as

$$\left|\frac{x - x^*}{x}\right|.$$

Precisely, if $\alpha = (x - x^*)/x$, then $(x - x^*)/x^* = \alpha/(1 - \alpha)$.

Loss of Significant Digits

If x^* is an approximation to x, then we say that x^* approximates x to r significant β-digits provided the absolute error $|x - x^*|$ is at most $1/2$ in the rth significant β-digit of x.

This can be expressed in a formula as

$$|x - x^*| \leq \frac{1}{2}\beta^{s-r+1}$$

with s the largest integer such that $\beta^s \leq |x|$.

For instance,

- $x^* = 3$ agrees with $x = \pi$ to one significant (decimal) digit,
- $x^* = 22/7 = 3.1428\cdots$ is correct to the three significant digits (as an approximation to π).

Loss of Significant Digits

Suppose, we have to calculate the number

$$z = x - y$$

and that we have approximations x^* and y^* for x and y, respectively, available, each of which is good to r digits.

Then

$$z^* = x^* - y^*$$

is an approximation for z, which is also good to r digits unless x^* and y^* agree to one or more digits.

In this latter case, there will be cancelation of digits during the subtraction, and consequently z^* will be accurate to fewer than r digits.

Loss of Significant Digits

Consider, for example

$$x^* = 0.76545421 \times 10^1, \qquad y^* = 0.76544200 \times 10^1$$

and assume each to be an approximation to x and y, respectively, correct to seven significant digits.

Then, in eight-digit floating-point arithmetic

$$z^* = x^* - y^* = 0.12210000 \times 10^{-3}$$

is that *exact* difference between x^* and y^*.

But as an approximation to $z = x - y$, z^* is good only to three digits, since the fourth significant digit of z^* is derived from the eighth digits of x^* and y^*, both possibly in error.

Loss of Significant Digits

Hence, while the error in z^* (as an approximation to $z = x - y$) is at most the sum of the errors in x^* and y^*, the relative error in z^* is possibly $10,000$ times the relative error in x^* or y^*.

Loss of significant digits is therefore dangerous only if we wish to keep the relative error small.

Such loss can often be avoided by anticipating its occurrence.

Example

Consider, for example, the evaluation of the function

$$f(x) = 1 - \cos(x)$$

in six decimal digit arithmetic.

Loss of Significant Digits

Since $\cos(x) \approx 1$ for x near zero, there will be loss of significant digits for x near zero.

For we cannot calculate $\cos(x)$ to more than six digits, so that the error in the calculated value may be as large as 5.10^{-7}, hence as large as, or larger than $f(x)$ for x near zero. One can use an alternative formula for $f(x)$, such as

$$f(x) = 1 - \cos(x) = \frac{1 - \cos^2(x)}{1 + \cos(x)} = \frac{\sin^2(x)}{1 + \cos(x)}$$

which can be evaluated quite accurately for small x; else, one could make use of the Taylor expansion for $f(x)$,

$$f(x) = \frac{x^2}{2} - \frac{x^4}{24} + \cdots$$

which shows, for example, that for $|x| \leq 10^{-3}$, $x^2/2$ agrees with $f(x)$ to at least six significant digits.

Loss of Significant Digits

Example

Find the roots of the quadratic equation

$$ax^2 + bx + c = 0.$$

We know that the roots are given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{3.1}$$

Let us assume that $b^2 - 4ac > 0$, that $b > 0$,and that we wish to find the root of smaller absolute value using (3.1); i.e.,

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{3.2}$$

If $4ac$ is small compared with b^2, then $\sqrt{b^2 - 4ac}$ will agree with b to several places. Hence, given that $\sqrt{b^2 - 4ac}$ will be calculated correctly only to as many places as are used in the calculations.

Loss of Significant Digits

To be specific, consider the equation

$$x^2 + 111.11x + 1.2121 = 0.$$

Using (3.2) and five-decimal floating-point chopped arithmetic, we calculate

$$\begin{aligned} b^2 &= 12,345 \\ b^2 - 4ac &= 12,340 \\ \sqrt{b^2 - 4ac} &= 111.09 \\ x_1 &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} = -0.01000 \\ \text{while in fact,} \quad x_1 &= -0.010910 \end{aligned}$$

which is accurate to five digits.

Loss of Significant Digits

Once an error is committed, it contaminates subsequent results. This **error propagation** through subsequent calculations is conveniently studied in terms of the two related concepts of *condition* and **instability**.

The word **condition** is used to describe the sensitivity of the function value $f(x)$ to changes in the argument x. The condition is usually measured by the maximum relative change in the function value $f(x)$ caused by a unit relative change in the argument.

In a somewhat informal formula

$$\max\left\{\left|\frac{f(x)-f(x^*)}{f(x)}\right| \Big/ \left|\frac{x-x^*}{x}\right| : |x-x^*| \text{ "small"}\right\} \approx \left|\frac{f'(x)x}{f(x)}\right|$$

Loss of Significant Digits

The larger the condition, the more ill-conditioned the function is said to be. Here we have made use of the fact that

$$f(x) - f(x^*) \approx f'(x)(x - x^*),$$

i.e., the change in argument from x to x^* changes the function value by approximately $f'(x)(x - x^*)$.

Example

Consider the evaluation of the function

$$f(x) = \sqrt{x}.$$

Here, $f'(x) = 1/(2\sqrt{x})$, hence the condition of f is, approximately

$$\left|\frac{f'(x)x}{f(x)}\right| = \frac{[1/(2\sqrt{x})]x}{\sqrt{x}} = \frac{1}{2},$$ which is well-codintioned.

Loss of Significant Digits

By contrast, if

$$f(x) = \frac{10}{1 - x^2}$$

then $f'(x) = 20x/(1 - x^2)^2$, so that

$$\left| \frac{f'(x)x}{f(x)} \right| = \left| \frac{[20x/(1 - x^2)^2]x}{10/(1 - x^2)} \right| = \frac{2x^2}{|1 - x^2|}$$

and this number can be quite large for $|x|$ near to 1. Thus, for x near to 1 or -1, this function is quite ill-conditioned. It very much magnifies relative errors in the argument there.

Loss of Significant Digits

Consider the function

$$f(x) = \sqrt{x+1} - \sqrt{x}$$

for "large", say for $x \approx 10^4$. Its condition there is

$$\left| \frac{f'(x)x}{f(x)} \right| = \frac{1}{2} \frac{|1/\sqrt{x+1} - 1/\sqrt{x}|x}{\sqrt{x+1} - \sqrt{x}} = \frac{1}{2} \frac{x}{\sqrt{x+1}\sqrt{x}} \approx \frac{1}{2}$$

which is quite good. But if we calculate $f(12345)$ in six-decimal arithmetic, we find that

$$\begin{aligned} f(12345) &= \sqrt{12346} - \sqrt{12345} \\ &= 111.113 - 111.108 = 0.005 \end{aligned}$$

while, actually $f(12345) = 0.00450003262627751\cdots$.

Loss of Significant Digits

So the calculated answer is in error by 10 percent. We analyze the computational process. It consists of the following four steps:

$$x_0 := 12345, \quad x_1 := x_0+1, \quad x_2 := \sqrt{x_1}, \quad x_3 := \sqrt{x_0}, \quad x_4 := x_2-x_3 \tag{4.1}$$

Now, consider for example, the function f_3, i.e., the function which describes how the final answer x_4 depends on x_3. We have

$$f_3(t) = x_2 - t$$

hence its condition is, approximately,

$$\left| \frac{f_3'(t)t}{f_3(t)} \right| = \left| \frac{t}{x_2 - t} \right|.$$

Loss of Significant Digits

This number is usually near to 1, i.e., f_3 is usually well-conditioned *except* when t is near x_2. In this latter case, f_3 can be quite badly conditioned.

For example, in our particular case, $t \approx 111.11$ while $x_2 - t \approx 0.005$, so the condition is $\sim 22,222$, or more than 40,000 times as big as the condition of f itself.

We conclude that the process defined in (4.1) is an unstable way to evaluate f. A stable is the following:

$$f(x) = \frac{1}{\sqrt{x+1} + \sqrt{x}}.$$

In six-decimal arithmetic, this gives

$$f(1234) = \frac{1}{\sqrt{12346} + \sqrt{12345}} = \frac{1}{222.221} = 0.00450002$$

which is in error by only 0.0003 percent.

Loss of Significant Digits

The computational process is

$$x_0 := 12345, \quad x_1 := x_0+1, \quad x_2 := \sqrt{x_1}, \quad x_3 := \sqrt{x_0}, \quad x_4 := x_2+x_3,$$

Here, for example, $f_3(t) = 1/(x_2 + t)$, and the condition of this function is, approximately

$$\left|\frac{f'(t)t}{f(t)}\right| = \left|\frac{t}{x_2 + t}\right| \approx \frac{1}{2}$$

for $t \approx x_2$, which is the case here. Thus, the condition of f_3 is quite good; it is as good as that of f itself.

Convergence of Iterative Algorithms

Many of the algorithms that will be developed here will be iterative in nature. These algorithms will generate a sequence of approximations that converge toward the desired solution.

Definition

The sequence $\{x_n\}$ **converges** to the value L provided that

$$\lim_{n\to\infty} x_n = L,$$

or, equivalently,

$$\lim_{n\to\infty} |x_n - L| = 0.$$

L is called the **limit** of the sequence. A sequence for which $\lim_{n\to\infty} x_n$ does not exist is said to **diverge**.

The two principal measures of convergence speed are known as rate of convergence, and order of convergence.

Rate of Convergence

Definition

Rate of Convergence Let $\{p_n\}$ be a sequence that converges to a number p. If there exists a sequence $\{\beta_n\}$ that converges to zero and a positive constant λ, independent of n, such that

$$|p_n - p| \leq \lambda|\beta_n|$$

for all sufficiently large values of n, then $\{p_n\}$ is said to converge to p with **rate of convergence** $O(\beta_n)$.

Example

Consider the following two sequences

$$\left\{\frac{n+3}{n+7}\right\}, \qquad \left\{\frac{2^n+3}{2^n+7}\right\}.$$

Rate of Convergence

Since

$$\lim_{n\to\infty} \frac{n+3}{n+7} = 1, \qquad \text{and} \qquad \lim_{n\to\infty} \frac{2^n+3}{2^n+7} = 1,$$

it follows that both sequences converge to the limit 1.

Table : *Corresponding terms in two sequences that converge to* 1.

n	$(n+3)/(n+7)$	$(2^n+3)/(2^n+7)$
1.0000e+000	5.0000e-001	5.5556e-001
2.0000e+000	5.5556e-001	6.3636e-001
3.0000e+000	6.0000e-001	7.3333e-001
8.0000e+000	7.3333e-001	9.8479e-001
9.0000e+000	7.5000e-001	9.9229e-001
1.0000e+001	7.6471e-001	9.9612e-001

Rate of Convergence

Now, we determine the rate of convergence of each sequence. After some simplification

$$\left|\frac{n+3}{n+7}-1\right| = \frac{4}{n+7} < 4\cdot\frac{1}{n}.$$

Hence, we may take $\lambda = 4$ and $\beta_n = 1/n$ in the definition of rate of convergence. It follows that the sequence

$$\left\{\frac{n+3}{n+7}\right\}$$

converges to 1 with rate of convergence $O(1/n)$.

Similarly, we can show that

$$\left|\frac{2^n+3}{2^n+7}-1\right| = \frac{4}{2^n+7} < 4\cdot\frac{1}{2^n}.$$

Rate of Convergence

Hence, we may take $\lambda = 4$, and $\beta_n = 1/2^n$ in the definition of rate of convergence, so the sequence

$$\left\{ \frac{2^n + 3}{2^n + 7} \right\}$$

converges to 1 with rate of convergence $O(1/2^n)$. These results confirm our numerical evidence since $1/2^n$ approaches zero faster than $1/n$ as $n \to \infty$.

Rate of Convergence

Definition

Let f be a function defined in the interval (a, b) that contains $x = 0$, and suppose that $\lim_{x \to 0} f(x) = L$. If there exists a function g for which $\lim_{x \to 0} g(x) = 0$ and a positive constant K such that

$$|f(x) - L| \leq K|g(x)|$$

for all sufficiently small values of x, then $f(x)$ is said to converge L with rate of convergence $O(g(x))$.

Rate of Convergence

Example

Determining rate of convergence for a function Consider the function

$$f(x) = \frac{\cos(x) - 1 + x^2/2}{x^4}.$$

What us the limit of f as $x \to 0$? Furthermore, at what rate does f converges to this limit?

Solution. From Taylor's theorem, we know that

$$\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720}\cos(\xi), \quad 0 < \xi < x.$$

Hence,

$$\frac{\cos(x) - 1 + x^2/2}{x^4} = \frac{1}{24} - \frac{1}{720}x^2\cos(\xi)$$

Rate of Convergence

Finally, because

$$\left|\frac{\cos(x) - 1 + x^2/2}{x^4} - \frac{1}{24}\right| = \frac{1}{720}|x^2 \cos(\xi)| \leq \frac{1}{720}|x^2|,$$

it follows that $\lim_{x\to 0} f(x) = 1/24$, and the rate of convergence is $O(x^2)$.

Order of Convergence

Order of convergence provides a different measure of convergence speed than rate of convergence.

Whereas rate of convergence examines individually the terms in the sequence of error values, $e_n = p_n - p$, order of convergence examines the relationship between successive error values, measuring the effectiveness with which each iteration reduces the approximation error.

Definition

Let $\{p_n\}$ be a sequence that converges to a number p. Let $e_n = p_n - p$ for $n \geq 0$. If there exist positive constants λ and α such that

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^\alpha} = \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^\alpha} = \lambda.$$

then $\{p_n\}$ is said to converge to p of **order α with asymptotic error constant λ**.

Order of Convergence

It follows that for a sequence that converges of order α, the error satisfies the asymptotic relation $|e_{n+1}| \approx \lambda |e_n|^{\alpha}$.

An iterative method is said to be of order α if the sequence it generates converges of order α. The most common values of α in practice are: $\alpha = 1$ (**known as linear convergence**), $\alpha = 2$ (**quadratic convergence**), and $\alpha = 3$ (**cubic convergence**).

Non-integer values for α are possible.

Note that when $\alpha = 1$, the sequences of error values satisfies

$$|e_{n+1}| \approx \lambda |e_n| \approx \lambda^2 |e_{n-1}| \approx \cdots \approx \lambda^n |e_0|.$$

Hence, a linearly convergent sequence converges with rate of convergence $O(\lambda^n)$.

Order of Convergence

Example

Consider the recursive scheme

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right). \tag{5.1}$$

(this is used to determine the square root of a positive real number a). Here, we are interested to determine the order of convergence of the generated sequence.

To accomplish this, we must be able to compute the error in the $(n+1)$st term in the sequence, $x_{n+1} - \sqrt{a}$, with the error in the nth term, $x_n - \sqrt{a}$.

Order of Convergence

We start by subtracting $\sqrt{a}$ from both sides of (5.1) and some simplification yields

$$\begin{aligned} x_{n+1} - \sqrt{a} &= \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) - \sqrt{a} \\ &= \frac{x_n^2 - 2x_n\sqrt{a} + a}{2x_n} \\ &= \frac{(x_n - \sqrt{a})^2}{2x_n}. \end{aligned}$$

Accordingly,

$$\lim_{n\to\infty} \frac{|x_{n+1} - \sqrt{a}|}{|x_n - \sqrt{a}|^2} = \lim_{n\to\infty} \frac{1}{2x_n} = \frac{1}{2\sqrt{a}}.$$

Hence, the sequence generated by this scheme has order of convergence equal to 2 and asymptotic error constant $1/(2\sqrt{a})$.

Order of Convergence

We have to check whether the sequence actually achieve quadratic convergence in practice? To answer to this question we need to select an a, generate the resulting sequence and examine the ratio $|e_n|/|e_{n-1}|^2$. If this ratio approaches a constant as n increases (the ratio should, in particular, approach the asymptotic error constant $1/(2\sqrt{a})$), then we have numerical evidence of quadratic convergence.

Table : *Quadratic Convergence for Example.*

n	x_n	$e_n = \|x_n - 3\|$	e_n/e_{n-1}^2
0.0000e+000	9.0000e+000	6.0000e+000	0
1.0000e+000	5.0000e+000	2.0000e+000	5.5556e-002
2.0000e+000	3.4000e+000	4.0000e-001	1.0000e-001
3.0000e+000	3.0235e+000	2.3529e-002	1.4706e-001
4.0000e+000	3.0001e+000	9.1554e-005	1.6537e-001
5.0000e+000	3.0000e+000	1.3970e-009	1.6666e-001

Order of Convergence

Note that the ration $|e_n|/|e_{n-1}|^2$ approaches a constant, thereby providing numerical confirmation of the quadratic convergence of the sequence. Further, the error ratio appears to be approaching $1/6 = 1/(2\sqrt{9})$, providing numerical confirmation that the asymptotic error constant for Eqn. (5.1) is $\lambda = 1/(2\sqrt{a})$.