"Socrates Dialectical Process: The first step is the separation of a subject into its elements. After this, by defining and discovering more about its parts, one better comprehends the entire subject "

Socrates (469-399) BCE[†]

Chapter 1

Basic Concepts

Numerical methods have been around for a long time. However, the usage of numerical methods was limited due to the lengthy hand calculations involved in their implementation. In our current society the application of numerical analysis and numerical methods occurs in just about every field of science and engineering. This is due in part to the rapidly changing digital computer industry. Digital computers have provided a fast computational device for the development and implementation of numerical methods which can handle a variety of difficult mathematical problems. To understand how numerical methods and numerical analysis techniques are developed the reader is required to have knowledge of certain background material from calculus and linear algebra. We begin by reviewing some fundamentals which are used extensively in this text.

Derivative of a Function

The derivative of a continuous function $y = f(x)$ is defined by the limiting process

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = f'(x) = \frac{dy}{dx},$$

if this limit exits. This limiting process can be represented using the alternative notations

$$\frac{dy}{dx} = \lim_{\Delta x\to 0} \frac{\Delta y}{\Delta x} \quad \text{or} \quad \left.\frac{dy}{dx}\right|_{x=x_0} = f'(x_0) = \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

The notation $m = f'(x_0)$ denotes the derivative evaluated at the point x_0. This derivative represents the slope m of the tangent line to the curve $y = f(x)$ which passes through the point $(x_0, f(x_0))$.

[†] BCE ("Before Common Era") replaces B.C. ("Before Christ") usage.

Fundamental Theorem of Calculus

Let $F(x)$ denote any function such that $\frac{dF(x)}{dx} = f(x)$, where $f(x)$ is a continuous function over the domain $a \leq x \leq b$. Divide the domain (a, b) into n equal subintervals of length $\Delta x = \frac{b-a}{n}$. This can be done by defining $a = x_0$ and $b = x_n$ with $x_i = x_0 + i\Delta x$ for $i = 0, 1, 2, \ldots, n$. The resulting numbers

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

are then said to partition the interval (a, b) into n equal subintervals. Let c_i denote a number in the ith subinterval, where $x_{i-1} \leq c_i \leq x_i$ for $i = 1, 2, \ldots, n$ and form the sum

$$S_n = \sum_{i=1}^{n} f(c_i)\Delta x = f(c_1)\Delta x + f(c_2)\Delta x + \cdots + f(c_n)\Delta x$$

The fundamental theorem of calculus states that

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{i=1}^{n} f(c_i)\Delta x = \int_a^b f(x)\, dx = F(x)\,]_a^b = F(b) - F(a)$$

represents the area under the curve $y = f(x)$, above the x-axis if $f(x) > 0$, and between the limits $x = a$ and $x = b$.

Note that if $G(x) = \int_a^x f(t)\, dt$, then $\frac{dG(x)}{dx} = f(x)$.

Taylor Series for Functions of a Single Variable

A function $f(x)$ of a single variable x is said to be analytic in the neighborhood of a point $x = x_0$ if it can be represented in a convergent power series of the form

$$\begin{aligned} f(x) =& f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^m(x_0)}{m!}(x - x_0)^m + \cdots \\ f(x) =& \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \end{aligned} \tag{1.1}$$

where by definition $0! = 1$ and the zero derivative denotes the function itself so that $f^{(0)}(x_0) = f(x_0)$. For a Taylor series to exist in the neighborhood of a point x_0, one must assume that the function $f(x)$ has continuous derivatives of all orders which can be evaluated at the point x_0.

Some well known Taylor series expansions are

$$\begin{aligned} e^x =& 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots \\ \sin x =& x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \\ \cos x =& 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \end{aligned}$$

If the Taylor series expansion given by equation (1.1) is truncated after the mth derivative term, then one can write

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^m(x_0)}{m!}(x - x_0)^m + Error \quad (1.2)$$

where the error term is given by

$$Error = \frac{f^{(m+1)}(\xi)}{(m+1)!}(x - x_0)^{m+1}, \qquad x_0 < \xi < x. \quad (1.3)$$

Taylor series expansions of the form

$$f(x_0 + h) = f(x_0) + f'(x_0)h + f''(x_0)\frac{h^2}{2!} + \cdots + f^{(m)}(x_0)\frac{h^m}{m!} + f^{(m+1)}(\xi_1)\frac{h^{m+1}}{(m+1)!}$$

$$f(x_0 - h) = f(x_0) - f'(x_0)h + f''(x_0)\frac{h^2}{2!} + \cdots + (-1)^m f^{(m)}(x_0)\frac{h^m}{m!} + (-1)^{m+1} f^{(m+1)}(\xi_2)\frac{h^{m+1}}{(m+1)!}$$

are used extensively in later chapters.

A continuous function $f(x)$ is said to have a root of multiplicity m if

$$f(x_0) = f'(x_0) = f''(x_0) = \cdots = f^{(m-1)}(x_0) = 0 \quad \text{but} \quad f^{(m)}(x_0) \neq 0 \quad (1.4)$$

That is, a root of multiplicity m is such that the function and its first $(m-1)$ derivatives are zero at $x = x_0$. If $m = 1$, then the root is called a simple root. For example, if, $f(x_0) = 0$, and $f'(x_0) \neq 0$, then x_0 is a simple root. In contrast, the conditions $f(x_0) = 0$, $f'(x_0) = 0$, and $f''(x_0) \neq 0$, imply x_0 is a root of multiplicity 2. Note that a function $f(x)$ which has a root x_0 of multiplicity m has the Taylor series expansion about x_0 of the form

$$f(x_0 + h) = f^{(m)}(x_0)\frac{h^m}{m!} f^{(m+1)}(x_0)\frac{h^{m+1}}{(m+1)!} + \cdots$$

since the function and its first $(m-1)$ derivatives are zero at $x = x_0$.

The Landau Symbol $\mathcal{O}$

The Landau symbol $\mathcal{O}$, sometimes referred to as "big Oh", is used to compare the behavior of one function $f(h)$ with another function $g(h)$ as $h \to 0$. One writes

$$f(h) = \mathcal{O}(g(h)) \quad \text{if} \quad |f(h)| \leq C|g(h)|, \quad C \text{ is a positive constant,}$$

for all h sufficiently small such that $\lim_{h\to 0} \frac{|f(h)|}{|g(h)|} \leq C < \infty$. For example, consider the Taylor series expansion for $\sin x$. One can write

$$\sin x = x - \frac{x^3}{3!} + \mathcal{O}(x^5)$$

since

$$\lim_{x\to 0}\frac{\sin x - x - \frac{x^3}{3!}}{x^5} = \frac{1}{5!} = Constant$$

The Landau symbol $\mathcal{O}$ is used in perturbational methods and numerical methods and is sometimes referred to as an order relation. It will be used throughout this text in the truncation of infinite series to denote the order of the error terms. For example, the Taylor series expansion for $f(x_0+h)$ when truncated after the second term can be expressed

$$f(x_0+h) = f(x_0) + f'(x_0)h + \mathcal{O}(h^2)$$

to indicate that the error term is proportional to h^2. One can write

$$Error \le C|h|^2 \quad \text{for any constant } C \text{ satisfying} \quad \frac{f''(\xi)}{2!} \le C.$$

The notation $\mathcal{O}(h^n)$ is used to denote the error being small and behaving like Ch^n, as h gets small, where C is a constant. The statement $Error = \mathcal{O}(h^n)$ is read "the error is of order h^n" and means $\lim_{h\to 0}(Error) = Ch^n$ for some positive constant C.

Taylor Series for Functions of Two Variables

A Taylor series expansion of a function of two variables $f(x,y)$ in the neighborhood of a point (x_0,y_0) can be written in the form

$$\begin{aligned} f(x,y) =& f(x_0,y_0) + \frac{\partial f}{\partial x}(x-x_0) + \frac{\partial f}{\partial y}(y-y_0) \\ &+ \frac{1}{2!}\left[\frac{\partial^2 f}{\partial x^2}(x-x_0)^2 + 2\frac{\partial^2 f}{\partial x \partial y}(x-x_0)(y-y_0) + \frac{\partial^2 f}{\partial y^2}(y-y_0)^2\right] + \cdots \end{aligned} \tag{1.5}$$

where all partial derivatives are to be evaluated at the point (x_0,y_0). The above Taylor series expansion can also be written in an operator notation. Define the partial derivative operators

$$D_x f = \frac{\partial f}{\partial x}, \quad D_y f = \frac{\partial f}{\partial y}, \quad D_x^2 = \frac{\partial^2 f}{\partial x^2}, \quad D_x D_y = \frac{\partial^2 f}{\partial x \partial y}, \quad D_y^2 f = \frac{\partial^2 f}{\partial y^2}, \quad \text{etc.}$$

and write the Taylor series expansion given by equation (1.5) in the special case where $x = x_0 + h$ and $y = y_0 + k$. One can then write equation (1.5) in the operator form

$$\begin{aligned} f(x_0+h, y_0+k) =& f(x_0,y_0) + \sum_{n=1}^{\infty}\frac{1}{n!}(hD_x + kD_y)^n f(x,y) \\ f(x_0+h, y_0+k) =& f(x_0,y_0) + (hD_x + kD_y)f + \frac{1}{2!}(hD_x + kD_y)^2 f \\ &+ \frac{1}{3!}(hD_x + kD_y)^3 f + \frac{1}{4!}(hD_x + kD_y)^4 f + \cdots \end{aligned} \tag{1.6}$$

where all partial derivatives are to be evaluated at the point (x_0, y_0). Note that the operator terms $(hD_x + kD_y)^m f$ can be evaluated by using the binomial expansion.

Example 2-1. **(Taylor series.)** If $f(x,y)$ and all of its partial derivatives through the nth order are defined and continuous over the rectangular region R defined by $R = \{\, a \le x \le b,\ c \le y \le d\}$ and the Taylor series is truncated after the nth derivative terms, then the error term can be calculated from knowledge of Taylor series expansions of a single variable. That is, one can replace x by $x_0 + t(x - x_0)$ and y by $y_0 + t(y - t_0)$ in $f(x,y)$ to obtain a function of the single variable t. One can define

$$\phi(t) = f(x_0 + t(x - x_0), y_0 + t(y - t_0))$$

so that with (x,y) and (x_0, y_0) fixed, the function $\phi(t)$ is a function of a single variable t. Expanding $\phi(t)$ about $t = 0$ gives

$$\phi(t) = \phi(0) + \phi'(0)t + \phi''(0)\frac{t^2}{2!} + \cdots + \phi^{(n)}(0)\frac{t^n}{n!} + Error$$

where

$$Error \ = \ \phi^{(n+1)}(t^*)\frac{t^{n+1}}{(n+1)!}, \qquad 0 < t^* < t.$$

Note that at $t = 1$ we have $\phi(1) = f(x,y)$ and when $t = 0$ we have $\phi(0) = f(x_0, y_0)$ so that one can write

$$\phi(1) = f(x,y) = \phi(0) + \phi'(0) + \frac{\phi''(0)}{2!} + \cdots + \frac{\phi^{(n)}(0)}{n!} + Error$$

where for $x - x_0 = h$ and $y - y_0 = k$ we have

$$\begin{aligned}
\phi(0) &= f(x_0, y_0) \\
\phi'(0) &= \left(\frac{\partial f}{\partial x}(x - x_0) + \frac{\partial f}{\partial y}(y - y_0)\right) = (hD_x + kD_y)f \\
\phi''(0) &= \left(\frac{\partial^2 f}{\partial x^2}(x - x_0)^2 + 2\frac{\partial^2 f}{\partial x \partial y}(x - x_0)(y - y_0) + \frac{\partial^2 f}{\partial y^2}(y - y_0)^2\right) = (hD_x + kD_y)^2 f \\
&\vdots \\
\phi^{(n)}(0) &= (hD_x + kD_y)^n f
\end{aligned}$$

where all partial derivatives are to be evaluated at the point (x_0, y_0). The error term is given by

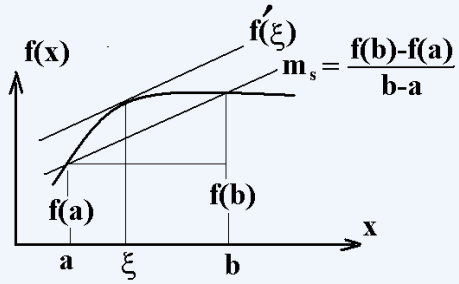
$$Error\Big|_{t=1} = \frac{\phi^{(n+1)}(t^*)}{(n+1)!} = (hD_x + kD_y)^{n+1} f\Big|_{x=\xi, y=\eta} \qquad \text{for} \quad 0 < t^* < 1,$$

where $\xi = x_0 + t^* h,\ \eta = y_0 + t^* k$ represent some point within the region R.

■

Mean Value Theorem

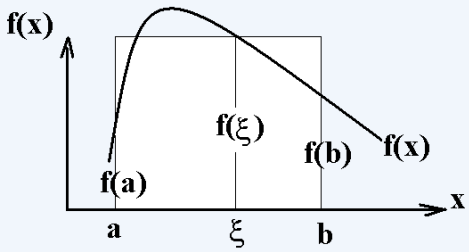
The mean value theorem states that if $f(x)$ is a continuous function on the closed interval $[a, b]$, then there exists a point ξ, satisfying $a < \xi < b$, such that the slope m_s of the secant line through the points $(a, f(a))$ and $(b, f(b))$ equals the slope of the curve $f(x)$ at $x = \xi$. This can be written and illustrated as follows.



$$m_s = \frac{f(b) - f(a)}{b - a} = f'(\xi), \qquad a < \xi < b. \tag{1.7}$$

Mean Value Theorem for Integrals

The mean value theorem for integrals states the if $f(x)$ is a continuous function and integrable over an interval $[a, b]$, then there exists a value ξ satisfying $a < \xi < b$ such that the average value of the function times the length of the interval from a to b must equal the area under the curve $f(x)$ between a and b. This can be written and illustrated as follows.



$$f(\xi)(b - a) = \int_a^b f(x)\, dx, \quad a < \xi < b \tag{1.8}$$

The extended mean value theorem for integrals states that if $f(x)$ and $g(x)$ are continuous functions on the closed interval $[a, b]$ and $g(x)$ does not change sign throughout the interval, then there exits a point ξ such that

$$\int_a^b f(x)g(x)\, dx = f(\xi) \int_a^b g(x)\, dx, \qquad a < \xi < b. \tag{1.9}$$

Other forms of this mean value theorem are for the conditions $f(x)$ is positive and monotonic over the interval (a,b) and $g(x)$ is integrable, then one can say there exists at least one value for ξ such that

$$\int_a^b f(x)g(x)\, dx = f(a) \int_a^\xi g(x)\, dx, \qquad a \le \xi \le b.$$

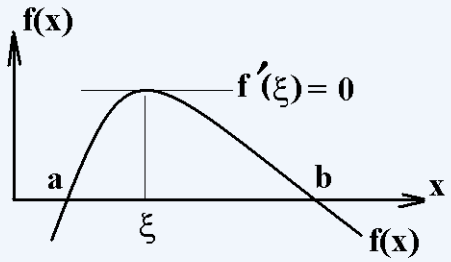
Extreme Value Theorem

The extreme value theorem states that if $f(x)$ is a continuous function over the closed interval $[a,b]$, then there will exist points ξ and η such that $f(\xi)$ is a maximum value of $f(x)$ over the interval and $f(\eta)$ is a minimum value of $f(x)$ over the interval. One can then write

$$minimum = f(\eta) \leq f(x) \leq f(\xi) = maximum, \qquad \text{for} \quad x \in [a,b].$$

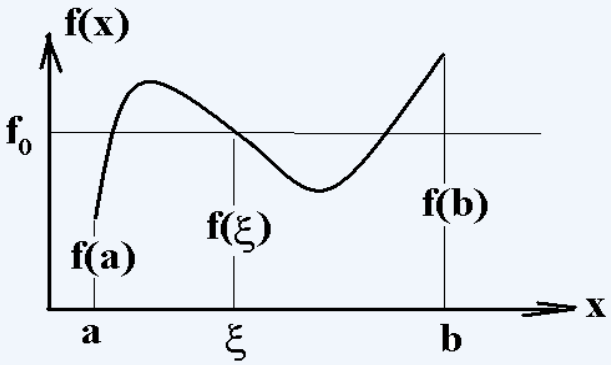
Rolle's Theorem

The Rolle's theorem assumes that $f(x)$ is continuous and differentiable on the closed interval $[a,b]$. One form of Rolle's theorem states that if $f(a) = 0$ and $f(b) = 0$, then there must exist at least one point ξ in the interval such that $f'(\xi) = 0, \quad a < \xi < b$.



Intermediate Value Theorem

The intermediate value theorem states that if $f(x)$ is continuous on the closed interval $[a,b]$ and there exists a value f_0 such that $f(a) < f_0 < f(b)$, then there exists at least one value ξ such that $f(\xi) = f_0$. In the accompanying figure note that for the f_0 selected there exists more than one value for ξ such that $f(\xi) = f_0$.



Number Representation

A base 10 (decimal) number system represents a number N in terms of various powers of 10 in a series having the form

$$\begin{aligned} N = &\cdots + \alpha_n(10)^n + \alpha_{n-1}(10)^{n-1} + \cdots + \alpha_3(10)^3 + \alpha_2(10)^2 + \alpha_1(10)^1 + \alpha_0(10)^0 \\ &+ \beta_1(10)^{-1} + \beta_2(10)^{-2} + \beta_3(10)^{-3} + \cdots \end{aligned} \tag{1.10}$$

where $\ldots, \alpha_n, \alpha_{n-1}, \ldots, \alpha_3, \alpha_2, \alpha_1, \alpha_0, \beta_1, \beta_2, \beta_3, \ldots$ are coefficients representing one of the digits $\{0,1,2,3,4,5,6,7,8,9\}$. Leading zeros and trailing zeros are not written.

For example, the number N=8326.432 in the base ten number system is really a shorthand representation for the number

$$N = 8(10)^3 + 3(10)^2 + 2(10)^1 + 6(10)^0 + 4(10)^{-1} + 3(10)^{-2} + 2(10)^{-3}.$$

A base 16 (hexadecimal) number system represents a number N in terms of various powers of 16 in a series having the form

$$\begin{aligned} N = &\cdots + \alpha_n(16)^n + \alpha_{n-1}(16)^{n-1} + \cdots + \alpha_3(16)^3 + \alpha_2(16)^2 + \alpha_1(16)^1 + \alpha_0(16)^0 \\ &+ \beta_1(16)^{-1} + \beta_2(16)^{-2} + \beta_3(16)^{-3} + \cdots \end{aligned} \tag{1.11}$$

where $\ldots, \alpha_n, \alpha_{n-1}, \ldots, \alpha_3, \alpha_2, \alpha_1, \alpha_0, \beta_1, \beta_2, \beta_3, \ldots$ are coefficients representing one of the digits $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$. (When the base is larger than 10 it is customary to use the letters $A - Z$ to represent the needed digits.) Numbers in the base b number system are represented using a subscript b. Some examples of base 10 numbers represented in the base 16 number system are listed for reference.

$10 = A_{16}$	$14 = E_{16}$	$30 = 1E_{16}$	$70 = 46_{16}$
$11 = B_{16}$	$15 = F_{16}$	$40 = 28_{16}$	$80 = 50_{16}$
$12 = C_{16}$	$16 = 10_{16}$	$50 = 32_{16}$	$90 = 5A_{16}$
$13 = D_{16}$	$20 = 14_{16}$	$60 = 3C_{16}$	$100 = 64_{16}$

A base 8 (octal) number system represents a number N in terms of various powers of 8 in a series having the form

$$\begin{aligned} N = &\cdots + \alpha_n(8)^n + \alpha_{n-1}(8)^{n-1} + \cdots + \alpha_3(8)^3 + \alpha_2(8)^2 + \alpha_1(8)^1 + \alpha_0(8)^0 \\ &+ \beta_1(8)^{-1} + \beta_2(8)^{-2} + \beta_3(8)^{-3} + \cdots \end{aligned} \tag{1.12}$$

where $\ldots, \alpha_n, \alpha_{n-1}, \ldots, \alpha_3, \alpha_2, \alpha_1, \alpha_0, \beta_1, \beta_2, \beta_3, \ldots$ are coefficients representing one of the digits $\{0, 1, 2, 3, 4, 5, 6, 7\}$.

A base 2 (binary) number system represents a number N in terms of various powers of 2 in a series having the form

$$\begin{aligned} N = &\cdots + \alpha_n(2)^n + \alpha_{n-1}(2)^{n-1} + \cdots + \alpha_3(2)^3 + \alpha_2(2)^2 + \alpha_1(2)^1 + \alpha_0(2)^0 \\ &+ \beta_1(2)^{-1} + \beta_2(2)^{-2} + \beta_3(2)^{-3} + \cdots \end{aligned} \tag{1.13}$$

where $\ldots, \alpha_n, \alpha_{n-1}, \ldots, \alpha_3, \alpha_2, \alpha_1, \alpha_0, \beta_1, \beta_2, \beta_3, \ldots$ are coefficients representing one of the digits $\{0, 1\}$.

Note that a base b number system requires b-digits to be used as coefficients in representing the numbers. For example, the Babylonians of long ago used a base 60 number system which requires 60 symbols to be used as digits. Some conventions from this system that have survived the many centuries is the fact that we have 60 seconds in a minute, 60 minutes in a hour, and 360 degrees in a circle.

Some other names associated with number systems are the following. A base 3 number system is called a ternary system, a base 4 system is called a quaternary system, a base 5 system is called a quinary system, a base 6 number system is called a senary system, a base 7 number system is called a septenary system, a base 9 number system is called a nonary system, a base 11 number system is called a undenary number system, and a base 12 number system is called a duodenary number system.

Number Conversion

To convert a number N from a decimal base to base b it is necessary to calculate the coefficients $\alpha_n, \alpha_{n-1}, \ldots, \alpha_1, \alpha_0, \beta_1, \beta_2, \ldots$ in the base b number system, where $N = \alpha_n b^n + \alpha_{n-1} b^{n-1} + \cdots + \alpha_2 b^2 + \alpha_1 b^1 + \alpha_0 b^0 + \beta_1 b^{-1} + \beta_2 b^{-2} + \cdots$. The integer part of N is denoted $I[N]$ and can be expressed in the factored form

$$I[N] = \alpha_0 + b(\alpha_1 + b(\alpha_2 + b(\alpha_3 + \cdots + b\alpha_n)\cdots))$$

from which one can observe that if the integer part of N is divided by b, then α_0 is the remainder and the quotient is $Q_1 = \alpha_1 + b(\alpha_2 + b(\alpha_3 + \cdots b\alpha_n)\cdots))$. If Q_1 is divided by b, then the remainder is α_1 and the new quotient is $Q_2 = \alpha_2 + b(\alpha_3 + \cdots \alpha_n b)\cdots))$. Continuing this process and saving the remainders $\alpha_0, \alpha_1, \ldots, \alpha_n$ the coefficients for the integer part of N can be determined. The fractional part of N is denoted $F[N]$ and can be expressed in the form

$$F[N] = \beta_1 b^{-1} + \beta_2 b^{-2} + \beta_3 b^{-3} + \cdots$$

from which one can observe that if $F[N]$ is multiplied by b, then there results

$$bF[N] = \beta_1 + \beta_2 b^{-1} + \beta_3 b^{-2} + \cdots$$

so that β_1 is the integer part of $bF[N]$ and the term $\beta_2 b^{-1} + \beta_3 b^{-2} + \cdots$ represents the fractional part of $bF[N]$. Hence if one continues to multiply the resulting fractional parts by b, then one can calculate the coefficients $\beta_1, \beta_2, \beta_3, \ldots$ associated with the fractional part of the base b representation of N.

Example 2-2. (Number conversion.)
Convert the number $N = 123.640625$ to a base 2 representation.

Solution: We start with the integer part of N and write $I[N] = 123 = \sum_{i=0}^{n} \alpha_i (2)^i$. Now divide 123 by 2 and save the remainder R. Continue to divide the resulting quotients and save the remainders as the remainders give us the coefficients $\alpha_0, \alpha_1, \ldots, \alpha_n$ in the base 2 representation. One can construct the following table to find the coefficients

N	I[N/2] =Q	R
123	I[123/2]=61	$1 = \alpha_0$
61	I[61/2]=30	$1 = \alpha_1$
30	I[30/2]=15	$0 = \alpha_2$
15	I[15/2]=7	$1 = \alpha_3$
7	I[7/2]=3	$1 = \alpha_4$
3	I[3/2]=1	$1 = \alpha_5$
1	I[1/2]=0	$1 = \alpha_6$

The integer part of N can now be represented $I[N] = 123 = 1111011_2$ The fractional part of N is written in the form $F[N] = 0.640625 = \sum_{i=1}^{\infty} \beta_i (2)^{-i}$. Now continue to multiply the fractional part by 2 and save the integer part each time. These integer parts represent the coefficients $\beta_1, \beta_2, \ldots$. One can construct the following table for determining the coefficients

N	2N	F[2N]	I[2N]
0.640625	1.28125	0.28125	$1 = \beta_1$
0.28125	0.5625	0.5625	$0 = \beta_2$
0.5625	1.125	0.125	$1 = \beta_3$
0.125	0.25	0.25	$0 = \beta_4$
0.25	0.50	0.50	$0 = \beta_5$
0.50	1.00	0.00	$1 = \beta_6$

The fractional part of N can be represented $F[N] = 0.640625 = 0.101001_2$ and the original number N has the base 2 representation $N = 123.640625 = 1111011.101001_2$.

■