EVALUATION OF POWER SERIES

TSOGTGEREL GANTUMUR

Abstract. We consider the evaluation of elementary transcendental functions such as e^x, $\log x$, $\sin x$, $\arctan x$, with the help of power series. We also discuss power series algorithms for computing the digits of π. Some notes on historically important algorithms are included.

Contents

1. Taylor polynomials

After the basic arithmetic and comparison operations, the next important operations for real number computations are the evaluation of elementary functions such as $\sqrt{x}$, $\exp x$, $\log x$, $\sin x$, and $\arctan x$. Each of these functions can be represented as a power series, which yields an efficient way to approximately evaluate the function. In fact, the discovery or power series in the 17-th century led to a huge leap in the computational capability of humans.

Let us start by fixing some terminology.

Definition 1. A function $f : (a, b) \to \mathbb{R}$ is called *analytic at* $c \in (a, b)$ if it is developable into a power series around c, i.e, if there are coefficients $a_n \in \mathbb{R}$ and $r > 0$ such that

$$f(x) = \sum_{n=0}^{\infty} a_n (x - c)^n, \qquad \text{for all} \quad x \in (c - r, c + r). \tag{1}$$

Moreover, f is said to be *analytic in* (a, b) if it is analytic at each $c \in (a, b)$.

Remark 2. (a) This definition can be extended to complex valued functions $f : \Omega \subset \mathbb{C} \to \mathbb{C}$ in a straightforward way.

(b) Power series can be differentiated term-wise, implying that the coefficients of the power series of f about c are given by $a_n = f^{(n)}(c)/n!$. In other words, if f is analytic at c, then the following *Taylor series* converges in a neighbourhood of c.

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(c)}{n!} (x - c)^n. \tag{2}$$

This formula was first published by Brook Taylor in 1715, although it was known previously to several mathematicians, including James Gregory as early as 1671.

Date: April 2, 2018.

Example 3. (a) Arguably the most important power series is the geometric series

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, \tag{3}$$

which converges for all x satisfying $|x| < 1$.

(b) The next in line is perhaps

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \tag{4}$$

which converges for all $x \in \mathbb{R}$. This was discovered by Leonhard Euler in 1748.

Assume that (1) converges, and let

$$T_n(x) = \sum_{k=0}^{n} a_k (x-c)^k, \qquad R_n(x) = f(x) - T_n(x), \tag{5}$$

where T_n is called the *n-th degree Taylor polynomial of f*, and R_n is called the *remainder*. The idea now is that in order to approximate $f(x)$, we compute $T_n(x)$ for some large n, such that the remainder $R_n(x)$ is small.

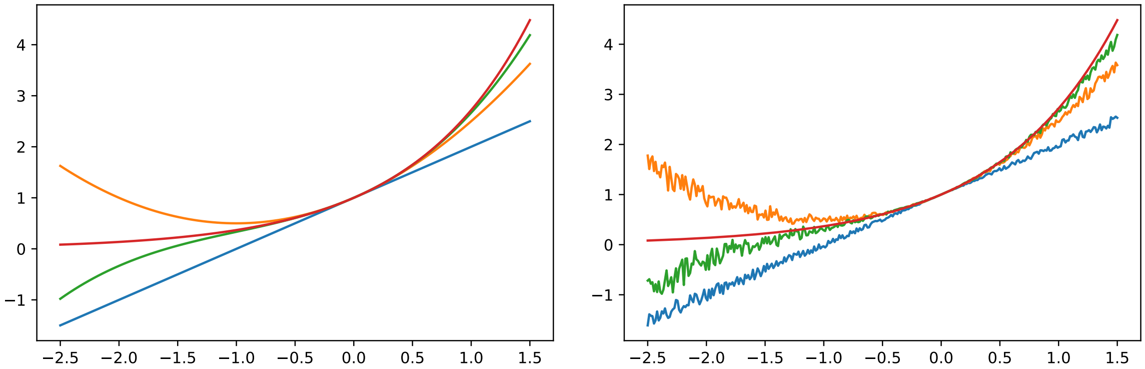


Figure 1. The exponential function e^x and its Taylor polynomials T_1, T_2, and T_3. To model roundoff error, the graph on the right is generated by introducing random error in the computation of the Taylor polynomials.

We need a way to estimate the remainder term. The simplest case is when the series (1) is alternating, in which case we have

$$|R_n(x)| \leq |a_{n+1}(x-c)^{n+1}|. \tag{6}$$

This is in fact approximately true in general, as the following theorem shows.

Theorem 4 (Lagrange 1797)**.** *Let $f \in \mathscr{C}([c,x])$ be $n+1$ times differentiable in (c,x), with the n-th derivative $f^{(n)}$ continuous in $[c,x)$. Then there exists $\xi \in (c,x)$, such that*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}. \tag{7}$$

Proof. The case $n = 0$ is simply the mean value theorem. We will give a proof only for the case $n = 1$, which contains all essential ideas of the general case.

We look for a quadratic polynomial $q(z) = \alpha + \beta(z-c) + \gamma(z-c)^2$ satisfying

$$q(c) = f(c), \qquad q(x) = f(x), \qquad \text{and} \qquad q'(c) = f'(c), \tag{8}$$

which turns out to be the following unique polynomial

$$q(z) = f(c) + f'(c)(z-c) + \big[f(x) - f(c) - f'(c)(x-c)\big]\frac{(z-c)^2}{(x-c)^2}. \tag{9}$$

Let $g(z) = f(z) - q(z)$, so that $g(c) = g(x) = 0$ and $g'(c) = 0$. Then g is twice differentiable in (c, x), with

$$g'(z) = f'(z) - f'(c) - \big[f(x) - f(c) - f'(c)(x-c)\big]\frac{2(z-c)}{(x-c)^2}, \tag{10}$$

and

$$g''(z) = f''(z) - \frac{2[f(x) - f(c) - f'(c)(x-c)]}{(x-c)^2}. \tag{11}$$

Moreover, $g'(c)$ exists and $g' \in \mathscr{C}([c,x))$. Since $g(c) = g(x)$, by Rolle's theorem, there is $\eta \in (c, x)$ such that $g'(\eta) = 0$. Now recalling that $g'(c) = 0$ and $g' \in \mathscr{C}([c,x))$, another application of Rolle's theorem gives the existence of $\xi \in (c, \eta)$ such that $g''(\xi) = 0$. In other words, we have

$$f(x) = f(c) + f'(c)(x-c) + \frac{1}{2}f''(\xi)(x-c)^2, \tag{12}$$

for some $\xi \in (c, x)$. □

Example 5 (exp)**.** As $(e^x)' = e^x$, the exponential series (4) has the remainder term

$$|R_n(x)| = \frac{e^{\xi}|x|^{n+1}}{(n+1)!}, \tag{13}$$

for some $\xi \in (0, x)$ or $\xi \in (x, 0)$, depending on whether $x > 0$ or $x < 0$. In case $x > 0$, we have $e^{\xi} < e^x$, which yields the following estimate on the relative error of $T_n(x)$:

$$\frac{|R_n(x)|}{e^x} < \frac{x^{n+1}}{(n+1)!}. \tag{14}$$

In case $x < 0$, the best we can do is $e^{\xi} \le 1$, and so

$$|R_n(x)| < \frac{|x|^{n+1}}{(n+1)!}. \tag{15}$$

Note that the latter estimate is identical to (6), which comes from the alternating character of the series (4) for $x < 0$.

Exercise 1. (a) Show that

$$1 + x + \ldots + x^n \to \frac{1}{1-x} \quad \text{as} \quad n \to \infty, \tag{16}$$

for $|x| < 1$.

(b) Look for a function f satisfying $f'(x) = f(x)$ and $f(0) = 1$ in the form

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \tag{17}$$

and arrive at the exponential series (4). Show that the series converges for all $x \in \mathbb{R}$.

Exercise 2. Show that the *binomial series*

$$(1+x)^a = 1 + ax + \frac{a(a-1)}{2!}x^2 + \frac{a(a-1)(a-2)}{3!}x^3 + \ldots, \tag{18}$$

converges for $|x| < 1$, where $a \in \mathbb{R}$ is a constant.

Table 1. Approximation of e^x by its Taylor polynomials. The last row shows the exact value e^x. For $x = 1$, we see that each additional term brings about 1 correct decimal digit. For $x = \frac{1}{4}$ and $x = \frac{1}{16}$, each additional term gives approximately 1.5 and 2.5 correct decimal digits, respectively. This type of convergence is called a *linear convergence*. In the last column, we never get the last digit correct, because of roundoff errors. The use guard digits in the intermediate computations would be needed to settle this issue, cf. Example 6.

n	$T_n(1) \approx e$	$T_n(\frac{1}{4}) \approx \sqrt[4]{e}$	$T_n(\frac{1}{16}) \approx \sqrt[16]{e}$
1	**2**.0000000000000000	1.**2**500000000000000	1.0**6**25000000000000
2	**2**.5000000000000000	1.2**8**12500000000000	1.064**4**531250000000
3	**2**.6666666666666665	1.2**8**38541666666667	1.0644**9**38151041667
4	2.**7**083333333333330	1.284**0**169270833335	1.0644944**5**08870444
5	2.7**1**66666666666663	1.28402**5**0651041669	1.06449445**8**8343304
6	2.71**8**0555555555554	1.284025**4**041883683	1.06449445891**7**1146
7	2.718**2**539682539684	1.28402541**6**2985183	1.064494458917**8**538
8	2.718**2**787698412700	1.284025416**6**769605	1.06449445891785**9**5
9	2.71828**1**5255731922	1.2840254166**8**74727	1.06449445891785**9**5
10	2.718281**8**011463845	1.2840254166**8**77356	1.06449445891785**9**5
∞	2.7182818284590452	1.2840254166877415	1.0644944589178594

Exercise 3 (Cauchy's form of the Taylor remainder)**.** In the setting of Theorem 4, show that

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k + \frac{f^{(n+1)}(\eta)}{n!}(x-\eta)^n(x-c). \tag{19}$$

for some $\eta \in (0, x)$ or $\eta \in (x, 0)$, depending on whether $x > 0$ or $x < 0$.

2. Roundoff error analysis

We shall consider the computation of the Taylor polynomial (5) in inexact arithmetic. Let us denote the k-th term of the Taylor polynomial by b_k, and its computed value by $\tilde{b}_k$, as

$$b_k = a_k(x-c)^k, \qquad \tilde{b}_k = (1+\beta_k)b_k, \qquad k = 0, \ldots, n, \tag{20}$$

where β_k accounts for the error made during the computation of b_k. If we assume that the coefficients a_k are given, we need at most n multiplications in $a_k(x-c)^k$, and hence we can estimate $|\beta_k| \leq \rho_\times(n, \varepsilon)$, with ε being the machine precision, and $\rho_\times(n,\varepsilon) = \frac{n\varepsilon}{1-n\varepsilon}$. In general, $\rho_\times(n,\varepsilon)$ depends on how the coefficients a_k are computed, on how the product $a_k(x-c)^k$ is computed, and on the error in the input value x. Proceeding further, introduce the notations

$$\begin{aligned} y_n &= b_0 + b_1 + \ldots + b_n, \\ y'_n &= \tilde{b}_0 + \tilde{b}_1 + \ldots + \tilde{b}_n, \\ \tilde{y}_n &= \tilde{b}_0 \oplus \tilde{b}_1 \oplus \ldots \oplus \tilde{b}_n, \end{aligned} \tag{21}$$

where $y_n = T_n(x)$ is the true value, and our goal is to estimate $y_n - \tilde{y}_n$. Assuming the "naive summation" algorithm, we get the intermediate estimate

$$\begin{aligned} |\tilde{y}_n - y'_n| &\leq \big((1+\varepsilon)^n - 1\big)|\tilde{b}_0 + \tilde{b}_1| + \big((1+\varepsilon)^{n-1} - 1\big)|\tilde{b}_2| + \ldots + \varepsilon|\tilde{b}_n| \\ &\leq \rho_+(n,\varepsilon) \sum_{k=0}^{n} \big(1 + |\beta_k|\big)|b_k| \leq \rho_+(n,\varepsilon)\big(1 + \rho_\times(n,\varepsilon)\big) \sum_{k=0}^{n} |b_k|, \end{aligned} \tag{22}$$

with $\rho_+(n,\varepsilon) = \frac{n\varepsilon}{1-n\varepsilon}$. In general, $\rho_+(n,\varepsilon)$ should depend on how the summation in (21) is carried out. Finally, an application of the triangle inequality yields

$$\begin{aligned}|\tilde{y}_n - y_n| &\leq |\tilde{y}_n - y_n'| + |y_n' - y_n| \leq \rho_+(n,\varepsilon)\big(1+\rho_\times(n,\varepsilon)\big)\sum_{k=0}^{n}|b_k| + \sum_{k=0}^{n}|\beta_k||b_k| \\ &\leq \big(\rho_+ + \rho_\times + \rho_+\rho_\times\big)\sum_{k=0}^{n}|b_k|,\end{aligned} \tag{23}$$

with $\rho_+ = \rho_+(n,\varepsilon)$ and $\rho_\times = \rho_\times(n,\varepsilon)$, and hence

$$\frac{|\tilde{y}_n - y_n|}{|y_n|} \leq (\rho_+ + \rho_\times + \rho_+\rho_\times)\frac{|b_0| + \ldots + |b_n|}{|b_0 + \ldots + b_n|} = (\rho_+ + \rho_\times + \rho_+\rho_\times)\kappa_+(b_0,\ldots,b_n), \tag{24}$$

where $\kappa_+(b_0,\ldots,b_n)$ is the condition number of the summation $b_0 + \ldots + b_n$.

Example 6 (exp)**.** With the notations of the preceding paragraph, for the exponential series (4), we have $b_k = x^k/k!$, which may be computed by using $2k-2$ multiplications and divisions, with the relative error

$$|\beta_k| = \frac{|\tilde{b}_k - b_k|}{|b_k|} \leq 4k\varepsilon \leq 4n\varepsilon =: \rho_\times(n,\varepsilon), \tag{25}$$

assuming that $k \leq n \leq \frac{1}{4\varepsilon}$. From the latter assumption, we infer $\rho_+(n,\varepsilon) \leq 2n\varepsilon$, and hence

$$\frac{|\tilde{y}_n - y_n|}{|y_n|} \leq (2n\varepsilon + 4n\varepsilon + 8n^2\varepsilon^2)\kappa_+(b_0,\ldots,b_n) \leq 8n\varepsilon\kappa_+(b_0,\ldots,b_n), \tag{26}$$

where we have used $n\varepsilon \leq \frac{1}{4}$ once again in the last step. Furthermore, we have

$$\kappa_+(b_0,\ldots,b_n) = \frac{e^{|x|}}{e^x} = e^{|x|-x} = e^{\max\{0,-2x\}}, \tag{27}$$

indicating a potentially catastrophic cancellation for x large negative. Thus keeping in mind that e^x for $x < 0$ can be computed by $e^x = 1/e^{|x|}$, in the following, we assume that $x \geq 0$. Taking into account that $y_n \leq e^x$, we get

$$|\tilde{y}_n - y_n| \leq 8|y_n|n\varepsilon \leq 8e^x n\varepsilon, \tag{28}$$

and invoking (14), we infer

$$\frac{|\tilde{y}_n - e^x|}{e^x} \leq \frac{|\tilde{y}_n - y_n|}{e^x} + \frac{|y_n - e^x|}{e^x} \leq 8n\varepsilon + \frac{x^{n+1}}{(n+1)!}. \tag{29}$$

To simplify it a bit, assuming $x \leq b$ and $n+1 \geq m$ for some constants $0 < b < 1$ and $m \geq 2$, we can replace (29) by

$$\frac{|\tilde{y}_n - e^x|}{e^x} \leq 8n\varepsilon + \frac{b^{n+1}}{m!}. \tag{30}$$

If ε is given, then even though the second term decays with n, the first term grows unboundedly. The right hand side is minimized when n is such that

$$\frac{b^{n+1}}{m!} \approx \frac{8\varepsilon}{\log(\frac{1}{b})}. \tag{31}$$

For example, taking $\varepsilon = 2^{-52}$, $b = e^{-2}$, and $m = 9$, we get $n \approx 10$, which suggests that $80\varepsilon + 4\varepsilon = 84\varepsilon$ is the best possible relative error, if we do all computations in double precision. This is consistent with (especially the results for $x = \frac{1}{16}$ in) Table 1.

On the other hand, if a target accuracy, say, $\delta > 0$ is specified, then one could choose n so large that $\frac{x^{n+1}}{(n+1)!} \leq \frac{\delta}{2}$, and then choose $\varepsilon > 0$ so that $8n\varepsilon \leq \frac{\delta}{2}$. For example, if want $\delta = 2^{-52}$ for $x \leq e^{-1}$, then $n = 13$ would guarantee $\frac{x^{n+1}}{(n+1)!} \leq \frac{\delta}{2}$. This implies that all computations must be performed with relative error $\varepsilon \leq \frac{\delta}{16n} \approx 2^{-60}$, i.e., in order to have the value e^x in double precision, one must compute with 8 guard bits, and sum the first 14 terms of the power series.

Exercise 4. There is a way to efficiently evaluate polynomials, known as *Horner's scheme*:

$$a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 = ((\cdots((a_n x + a_{n-1})x + a_{n-2})x\cdots)x + a_1)x + a_0. \tag{32}$$

Since this requires n to be known beforehand, it is not a practical method for power series. However, we can rescue the method by writing

$$a_0 + a_1 x + \ldots + a_{n-1}x^{n-1} + a_n x^n = [((\cdots((a_0 y + a_1)y + a_2)y\cdots)y + a_{n-1})y + a_n]x^n, \tag{33}$$

with $y = 1/x$. Perform a roundoff error analysis for the modified Horner scheme (33).

3. The exponential function

Recall the power series

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots \qquad (x \in \mathbb{R}), \tag{34}$$

for the exponential function. The first thing to notice here is that since $e^x \approx 0$ for $x << 0$, we expect cancellation of digits in this regime, cf. Example 6. On the other hand, from the condition number $\kappa(x) = |x|$ of the exponential, we should expect the computations of e^x and of e^{-x} to have roughly the same difficulty. Indeed, this expectation can be realized by taking advantage of the relation

$$e^{-x} = \frac{1}{e^x}, \tag{35}$$

to flip the sign of x.

From now on, we assume that $x > 0$. If x is large, the terms of the series (34) will grow with n, until around the point where $n! \approx x^n$. This is undesirable, as it would inflate the number of terms needed to sum to achieve a desired accuracy. To deal with the problem, we perform an *argument reduction* before utilizing any power series, e.g., by expressing e^x in terms of e^y with small y. In this regard, the law of addition comes in handy. Thus let $b > 1$ be a constant, and let $m \in \mathbb{N}$ and $y \in \mathbb{R}$ be such that

$$x = y + m \log b, \tag{36}$$

where the idea is of course that we choose y small. For instance, we can ensure $0 \leq y < \log b$, or $-\frac{1}{2}\log b < y \leq \frac{1}{2}\log b$. Then in light of

$$e^x = e^{y + m\log b} = b^m e^y, \tag{37}$$

the power series computation can now be done only for e^y. Here, the choice $b = e$ simplifies (36), but we would need to compute the power e^m in (37). On the other hand, the choice $b = 2$ would lead to the simple power 2^m in (37), but we would need a value of $\log 2$ in (36).

The preceding method may be called an *additive* argument reduction. We can also approach the argument reduction problem *multiplicatively*. Let $r = x/n$, where n is some large integer, preferably of the form $n = 2^k$. Then we have

$$e^x = e^{nr} = (e^r)^n, \tag{38}$$

where e^r is to be computed with power series. If $n = 2^k$, then the power $(e^r)^n$ can be computed by repeated squaring, as in, e.g., $(e^r)^8 = (((e^r)^2)^2)^2$.

Remark 7. Let us look at the roundoff error. We assume $n = 2^k$, and that the division $r = x/n$ can be done exactly. Furthermore, suppose that $z = e^r$ is computed with relative precision $\eta > 0$, as

$$\tilde{z} = e^r(1 + \delta_0), \qquad |\delta_0| \leq \eta. \tag{39}$$

We perform the repeated squaring

$$\begin{aligned}\tilde{y} &= [\dots[\tilde{z}^2(1+\delta_1)]^2(1+\delta_2)\dots]^2(1+\delta_k)\\ &= e^x(1+\delta_0)^{2^k}(1+\delta_1)^{2^{k-1}}(1+\delta_2)^{2^{k-2}}\cdots(1+\delta_k),\end{aligned}\tag{40}$$

with the same relative precision: $|\delta_j| \le \eta$. Noting that

$$(1+\delta_j)^m \le (1+\eta)^m \le 1+2m\eta, \qquad \text{for} \quad 2m\eta \le 1,\tag{41}$$

we infer

$$\begin{aligned}\frac{|\tilde{y}-e^x|}{e^x} &\le 2^{k+1}\eta(1+2^{k+1}\eta) + 2^k\eta(1+2^k\eta) + \dots + \eta\\ &\le 2^{k+2}\eta + 2^{2k+3}\eta^2 = 4n\eta(1+2n\eta) \le 8n\eta.\end{aligned}\tag{42}$$

This means that we lose approximately k correct significant bits, and in particular, if we want, say, $\varepsilon > 0$ as the accuracy of the final computation, then the intermediate computations must be done with the relative precision $\eta = 2^{-k-3}\varepsilon$.

Remark 8. We can do a bit better if $x > 0$ is small, by working with $E(x) = e^x - 1$ instead of e^x. To be more precise, note that

$$E(x) = x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots, \qquad \text{and} \qquad E(x) \le \frac{x}{1-x},\tag{43}$$

with the latter being true for $0 \le x < 1$. This function has the following "doubling formula"

$$E(2r) = e^{2r} - 1 = (e^r-1)(e^r+1) = (e^r-1)(2+e^r-1) = E(r)\big(2+E(r)\big).\tag{44}$$

As before, we assume that $n = 2^k$, and that the division $r = x/n$ is done exactly. Suppose that each application of the doubling formula is performed with relative precision η, that is,

$$\tilde{E}(2r) = \tilde{E}(r)\big(2+\tilde{E}(r)\big)(1+\delta),\tag{45}$$

with $|\delta| \le \eta$. Writing $\tilde{E}(r) = E(r)(1+\xi)$, we have

$$\tilde{E}(2r) - E(2r) = E(r)\big(2[(1+\xi)(1+\delta)-1] + E(r)[(1+\xi)^2(1+\delta)-1]\big),\tag{46}$$

which yields

$$\begin{aligned}|\tilde{E}(2r) - E(2r)| &\le E(r)\big(2(|\xi|+|\delta|) + E(r)(2|\xi|+|\delta|)\big) + O(\eta^2)\\ &= E(2r)(|\xi|+|\delta|) + E(r)^2|\xi| + O(\eta^2),\end{aligned}\tag{47}$$

and

$$\frac{|\tilde{E}(2r)-E(2r)|}{E(2r)} \le |\xi| + \eta + \frac{E(r)|\xi|}{2+E(r)} + O(\eta^2) \le (1+x)|\xi| + \eta + O(\eta^2),\tag{48}$$

where we have used the estimate $E(r) \le 2r \le 2x$, under the assumption that $x \le \frac{1}{2}$. We then apply the latter formula k times, to arrive at

$$\begin{aligned}\frac{|\tilde{E}(x)-E(x)|}{E(x)} &\le \eta[1 + (1+x) + (1+x)^2 + \dots + (1+x)^k] + O(\eta^2)\\ &\le \eta\frac{(1+x)^{k+1}-1}{x} + O(\eta^2)\\ &\le \frac{(k+1)\eta}{1-(1+k)x} + O(\eta^2).\end{aligned}\tag{49}$$

This shows that the loss of accuracy is logarithmic in n. Writing $\tilde{E}(x) = E(x)(1+\xi)$, and assuming $x \le \frac{k-1}{2(k+1)}$ for convenience, we have

$$|\xi| \le k\eta + O(\eta^2).\tag{50}$$

Finally, the computation of $e^x = 1 + E(x)$ is

$$\tilde{y} = 1 \oplus \tilde{E}(x) = \big(1 + \tilde{E}(x)\big)(1+\delta) = \big(1 + E(x)(1+\xi)\big)(1+\delta), \tag{51}$$

with $|\delta| \le \varepsilon$, giving the error estimate

$$\frac{|\tilde{y} - e^x|}{e^x} \le \varepsilon + \frac{|\xi|(1+\varepsilon)E(x)}{e^x} \le \varepsilon + \frac{1}{2}|\xi|(1+\varepsilon) \le \varepsilon + k\eta + O(\eta^2 + \eta\varepsilon). \tag{52}$$

Remark 9. In view of the relation

$$e^x = \sinh x + \sqrt{1 + \sinh^2 x}, \tag{53}$$

it is possible to replace the exponential series by the series

$$\sinh x = \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \dots \qquad (x \in \mathbb{R}). \tag{54}$$

The latter series converges about twice as fast, because the terms have the "stepsize" x^2 as opposed to x. However, this reduction can only be used once, meaning that the usual argument reduction must still be performed beforehand.

Exercise 5. From the definitions

$$\sinh x = \frac{e^x - e^{-x}}{2}, \qquad \text{and} \qquad \cosh x = \frac{e^x + e^{-x}}{2}, \tag{55}$$

derive the power series (54), and

$$\cosh x = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!} = 1 + \frac{x^2}{2} + \frac{x^4}{4!} + \frac{x^6}{6!} + \dots \qquad (x \in \mathbb{R}). \tag{56}$$

Prove the relation (53).

Exercise 6. Perform a roundoff error analysis of the additive argument reduction (36).

4. Logarithmic functions

For logarithms, the basic power series is

$$\log(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \qquad (-1 < x \le 1), \tag{57}$$

which was discovered independently by Gerardus Mercator and Isaac Newton around 1667. The series only converges for $-1 < x \le 1$, and so we need argument reduction to compute $\log y$ for arbitrary $y > 0$. To this end, let us perform the multiplicative reduction

$$y = r2^n, \qquad \text{so that} \quad \log y = n \log 2 + \log r, \tag{58}$$

where we require $n \in \mathbb{Z}$ and $1 \le r < 2$, ensuring $r = 1 + x$ with $0 \le x < 1$. Note that we can replace the condition $1 \le r < 2$ by the more general $\rho \le r < 2\rho$, where $\rho < 1$ but $\rho \approx 1$.

Further reduction is possible, by the recipe

$$\begin{aligned} \log(1+x) &= 2\log\sqrt{1+x} = 2\log\frac{(\sqrt{1+x})(1+\sqrt{1+x})}{1+\sqrt{1+x}} = 2\log\frac{\sqrt{1+x}+1+x}{1+\sqrt{1+x}} \\ &= 2\log\Big(1 + \frac{x}{1+\sqrt{1+x}}\Big) =: 2\log(1+z), \end{aligned} \tag{59}$$

which can be applied repeatedly. Note that

$$z = \frac{x}{1+\sqrt{1+x}} < \frac{x}{2} \qquad \text{for} \quad x > 0, \tag{60}$$

and also that this form is numerically better behaved than the alternative $z = \sqrt{1+x} - 1$.

Finally, once the problem is reduced to evaluating $\log(1+z)$ with z small, we can resort to the series

$$\frac{1}{2}\log\frac{1+x}{1-x} = \sum_{n=0}^{\infty}\frac{x^{2n+1}}{2n+1} = x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \dots \qquad (|x| < 1), \tag{61}$$

discovered by James Gregory in 1668. Solving $\frac{1+x}{1-x} = 1+z$ for x gives $x = \frac{z}{2+z} \approx \frac{z}{2}$, so not only the the "stepsize" of Gregory's series is x^2 as in Remark 9, but also the argument is about twice as small as the argument of the corresponding Mercator series. However, note that as in Remark 9, this reduction can only be used once.

Example 10. Note that we need an accurate value of $\log 2$ for the reduction (58).

(a) Mercator's series for $\log 2$ is

$$\log 2 = 2 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots, \tag{62}$$

whose convergence is extremely slow. We can also use

$$\log 2 = -\log\frac{1}{2} = \frac{1}{2} + \frac{1}{2\cdot 2^2} + \frac{1}{3\cdot 2^3} + \frac{1}{4\cdot 2^4} + \dots, \tag{63}$$

which is much better.

(b) On the other hand, for Grogory's series, $2 = \frac{1+x}{1-x}$ gives $x = \frac{1}{3}$, and hence we have

$$\log 2 = \frac{2}{3} + \frac{2}{3\cdot 3^3} + \frac{2}{5\cdot 3^5} + \frac{2}{7\cdot 3^7} + \dots, \tag{64}$$

which is way faster.

(c) We do the reduction (59) once, to get

$$\log 2 = 2\log\left(1 + \frac{1}{1+\sqrt{2}}\right), \tag{65}$$

and compute the latter logarithm by Gregory's series, with

$$x = \frac{1}{3+2\sqrt{2}}. \tag{66}$$

Compared to b), the magnitude of x has almost been halved.

Table 2. Approximation of $\log 2$ by the methods presented in Example 10. The last row shows the exact value $\log 2$. In the last column, the maximum possible precision under the floating point arithmetic has been attained.

n	Mercator (62)	Mercator (63)	Gregory (64)	Gregory (66)
1	1.0000000000000000	**0**.5000000000000000	0.6**9**13580246913580	0.69**3**0256795263684
2	**0**.5000000000000000	0.**6**250000000000000	0.69**3**0041152263374	0.6931**4**46209503476
3	**0**.8333333333333333	0.**6**666666666666666	0.693**1**347573322881	0.693147**1**218850720
4	**0**.5833333333333333	0.**6**822916666666666	0.6931**4**60473908271	0.693147**1**791455733
5	**0**.7833333333333332	0.**6**885416666666666	0.69314**7**0737597851	0.69314718**0**5246939
6	**0**.6166666666666666	0.6**9**11458333333332	0.693147**1**702560119	0.69314718055**9**0457
7	**0**.7595238095238095	0.6**9**22619047619046	0.693147**1**795482411	0.69314718055**9**9221
8	**0**.6345238095238095	0.6**9**27501860119046	0.69314718**0**4592440	0.693147180559**9**448
9	**0**.7456349206349207	0.6**9**29671999007935	0.693147180**5**498115	0.6931471805599**4**54
10	**0**.6456349206349207	0.69**3**0648561507935	0.6931471805**5**89162	0.6931471805599**4**54
∞	0.6931471805599453	0.6931471805599453	0.6931471805599453	0.6931471805599453

Remark 11 (Briggs' method)**.** Logarithms were independently discovered by John Napier and Jost Bürgi around 1614. The property $\ell(x\cdot y)=\ell(x)+\ell(y)$ was the main focus of their investigations, and what they called logarithms were instances of

$$\ell(x)=\alpha\log(x/\beta), \tag{67}$$

where α and β are constants. Specifically, Napier's logarithm is set up so that $\ell(10^7)=0$ and $\ell(10^7-1)=1$, while Bürgi's logarithm satisfies $\ell(10^8)=0$ and $\ell(10^8+10^4)=10$. Upon consultation with Napier, Henry Briggs compiled large tables of logarithms around 1620, with the normalization $\ell(1)=0$ and $\ell(10)=1$. This is of course the common logarithm with base 10. Essentially, the method they used is to pick a large number n, and compute the powers

$$x_k=(1+\tfrac{1}{n})^k, \qquad k=0,\ldots, \tag{68}$$

after which, they set

$$\ell(x_k)=k. \tag{69}$$

Hence the name *logarithm* (logos – ratio, arithmos – number). It is easy to see that

$$k=\frac{\log x_k}{\log(1+\frac{1}{n})}=\frac{n\log x_k}{\log(1+\frac{1}{n})^n}, \tag{70}$$

and that for large n,

$$\frac{k}{n}=\frac{\log x_k}{\log(1+\frac{1}{n})^n}\approx\log x_k, \tag{71}$$

which would be one way to arrive at the natural logarithm. Of course, n does not have to be an integer. Thus, for instance, in order to have $\ell(x_k)=1$ for $x_k=10$, we take repeated square roots of 10, to find, e.g., $\alpha=\sqrt[N]{10}$ with $N=2^{54}$, as Briggs did. Then we use $1+\frac{1}{n}=\alpha$ in (68), which ensures that $x_N=10$. Finally, to set $\ell(10)=1$, we do a scaling in (69), so that

$$x_k=\alpha^k \qquad\Longleftrightarrow\qquad \log_{10}x_k=\ell(x_k)=\frac{k}{N}. \tag{72}$$

Obviously, computing all powers α^k for $k=1,2,\ldots,N$ is an impossible task. To find the logarithm of a specific number, say $x=2$, we need k such that $\alpha^k=x$. Taking the N-th root from both sides, we get $\alpha^{k/N}=\sqrt[N]{x}$. The right hand side can be computed by repeated square roots, and since $\alpha\approx1$, the left hand side can be written as

$$\alpha^{k/N}\approx1+\frac{k}{N}(\alpha-1), \tag{73}$$

yielding the following formula

$$\log_{10}x_k=\frac{k}{N}\approx\frac{\sqrt[N]{x}-1}{\alpha-1}=\frac{\sqrt[N]{x}-1}{\sqrt[N]{10}-1}. \tag{74}$$

In fact, from our perspective, Briggs' method corresponds to the reduction (59) taken to its extreme: It is equivalent to applying the reduction 54 times, which makes $\sqrt[N]{x}-1$ very small, and finally invoking the "power series" approximation $\log\sqrt[N]{x}\approx\sqrt[N]{x}-1$. The main work involved here is the square root operation performed 54 times. In order to get 14 correct decimals in the final result, Briggs used 32 to 40 decimals in the intermediate calculations. The improvements brought about by Mercator's and Gregory's series were clearly phenomenal, especially when all computations were performed by hand. It is a general observation that before power series, square root extractions were the workhorse of heavy computations.

Exercise 7. (a) By integrating

$$\frac{1}{1+x}=1-x+x^2-x^3+\ldots, \tag{75}$$

derive the Mercator series (57).

(b) Expanding each of the logarithms in $\log(1+x)-\log(1-x)$ into Mercator's series, derive the series (61). Show that

$$\operatorname{arctanh} x = \frac{1}{2}\log\frac{1+x}{1-x}, \tag{76}$$

where $\operatorname{arctanh} x$ is the inverse function of the hyperbolic tangent

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{77}$$

Exercise 8. Perform a roundoff error analysis of the argument reduction (59) applied k times.

Exercise 9. (a) Suppose that we want to compute $\log 2$ by writing $\log 2 = \log\frac{2}{3} + \log\frac{4}{3}$, and employing Gregory's series to the resulting two logarithms. Would this method be faster than that described in Example 10(c)?

(b) Suggest fast methods to compute $\log 5$ and $\log 7$, in the spirit of (a).

(c) Suggest several methods to compute $\log 3$, in the spirit of Example 10.

5. Trigonometric functions

Recall the well known Maclaurin series for the sine and cosine functions

$$\begin{aligned}\sin x &= \sum_{k=0}^{\infty}(-1)^k\frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots \qquad (x\in\mathbb{R}),\\ \cos x &= \sum_{k=0}^{\infty}(-1)^k\frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots \qquad (x\in\mathbb{R}).\end{aligned} \tag{78}$$

Both series were discovered independently by Isaac Newton and Gottfried Leibniz during the period 1669–1676. It turns out that this was a rediscovery, in that the series (78) appeared in the writings of early 16-th century Indian mathematicians, who attributed the discovery to Madhava of Sangamagramma. Unfortunately, their knowledge was not widespread, and apparently had no influence on the development of calculus in Europe.

The fundamental properties of these functions are *periodicity*

$$\sin(x+2\pi n) = \sin x, \qquad \cos(x+2\pi n) = \cos x, \qquad n\in\mathbb{Z}, \tag{79}$$

symmetry

$$\sin(-x) = -\sin x, \qquad \cos(-x) = \cos x, \tag{80}$$

and the *law of addition*

$$\begin{aligned}\sin(x+y) &= \sin x\cos y + \cos x \sin y,\\ \cos(x+y) &= \cos x\cos y - \sin x\sin y.\end{aligned} \tag{81}$$

Their interrelations are summed up in

$$\sin^2 x + \cos^2 x = 1, \qquad \text{and} \qquad \cos x = \sin(\tfrac{\pi}{2} - x). \tag{82}$$

Thus by periodicity and symmetry, the arguments of both functions $\sin x$ and $\cos x$ can be reduced to the case $0 \le x < \frac{\pi}{2}$. This makes the series alternating, which means straightforward remainder estimates. Moreover, the law of addition implies the double angle formulas

$$\sin 2x = 2\sin x\cos x, \qquad \cos 2x = \cos^2 x - \sin^2 x, \tag{83}$$

which in turn give us the following argument reduction recipes

$$\begin{aligned}\sin x &= 2\sin(\tfrac{x}{2})\sqrt{1-\sin^2(\tfrac{x}{2})},\\ \cos x &= 2\cos^2(\tfrac{x}{2}) - 1.\end{aligned} \tag{84}$$

Remark 12 (Tangent series)**.** In view of (82), we only need to be able to compute either $\sin x$ or $\cos x$, and all the other functions $\tan x$, $\cot x$, $\sec x$, and $\csc x$ can be expressed in terms of a single function. However, other series may be of independent interest. For instance, take the tangent series

$$\tan x = \sum_{k=0}^{\infty} \frac{t_{2k+1}x^{2k+1}}{(2k+1)!} = t_1 x + \frac{t_3x^3}{3!} + \frac{t_5x^5}{5!} + \frac{t_7x^7}{7!} + \ldots \qquad (|x| < \tfrac{\pi}{2}), \tag{85}$$

where $t_1 = 1$, $t_3 = 2$, $t_5 = 16$, $t_7 = 272$, etc., are positive integers, called the *tangent numbers*. These numbers are related to the *Bernoulli numbers* by

$$B_{2n} = (-1)^n \frac{2n}{4^{2n} - 2^{2n}} t_{2n-1}, \qquad n = 1, 2, \ldots. \tag{86}$$

To compute the tangent numbers, note that

$$t_{2k+1} = (\tan x)^{(2k+1)}\big|_{x=0}. \tag{87}$$

It is immediate from

$$(\tan x)' = 1 + \tan^2 x, \tag{88}$$

that $(\tan x)^{(n)}$ is a polynomial in $\tan x$, of degree not exceeding $n+1$. Then by writing

$$\begin{aligned} (\tan x)^{(n)} &= a_{n,0} + a_{n,1}\tan x + \ldots + a_{n,n+1}\tan^{n+1} x \\ &= a_{n-1,1}(\tan x)' + \ldots + a_{n-1,n}(\tan^n x)' \\ &= a_{n-1,1}(1+\tan^2 x) + \ldots + n a_{n-1,n}\tan^{n-1} x(1+\tan^2 x), \end{aligned} \tag{89}$$

we derive the recurrence relation

$$a_{n,j} = (j-1)a_{n-1,j-1} + (j+1)a_{n-1,j+1}, \tag{90}$$

with the understanding that $a_{n-1,j} = 0$ for all $j < 0$ and $j > n$. Thus starting with $a_{0,j} = \delta_{1,j}$, we can compute the coefficients $a_{n,j}$, and then set $t_{2k+1} = a_{2k+1,0}$. An attractive feature of this algorithm is that all operations are over positive integers. Table 3 illustrates the algorithm.

Table 3. Computation of the tangent numbers, cf. (90).

$n=0$	0	1							
$n=1$	**1**	0	1						
$n=2$	0	2	0	2					
$n=3$	**2**	0	8	0	6				
$n=4$	0	16	0	40	0	24			
$n=5$	**16**	0	136	0	240	0	120		
$n=6$	0	272	0	136	0	1680	0	720	
$n=7$	**272**	0	$272+3\cdot 136$	0	$3\cdot 136+5\cdot 1680$	0	$5\cdot 1680+7\cdot 720$	0	$7\cdot 720$

We end this section with a couple of historical remarks, summarizing the computational methods for trigonometric functions that existed *before* power series.

Remark 13 (Using square roots only)**.** Trigonometric functions have been around much longer than logarithms, since the times of ancient Greeks. While the Greeks only used the *chord function*

$$\operatorname{chord} x = 2\sin\tfrac{x}{2}, \tag{91}$$

explicitly, 5-th century Indian mathematicians introduced $\sin x$, $\cos x$, $\arcsin x$, and others. Let us try to compute the exact value of $\cos x$ for as many values of x as possible. Thus, the sine series (78), in combination with (82), implies

$$\cos\tfrac{\pi}{2} = \sin 0 = 0, \qquad \sin\tfrac{\pi}{2} = \cos 0 = 1. \tag{92}$$

The *double angle* formula for the cosine

$$\cos 2x = 2\cos^2 x - 1, \tag{93}$$

cf. (83), can be solved to yield the *angle bisection* formula

$$2\cos x = \sqrt{2 + 2\cos 2x}, \qquad (0 \le 2x \le \pi). \tag{94}$$

Starting with $2x = \frac{\pi}{2}$, this gives

$$2\cos\tfrac{\pi}{4} = \sqrt{2}, \qquad 2\cos\tfrac{\pi}{8} = \sqrt{2+\sqrt{2}}, \qquad 2\cos\tfrac{\pi}{16} = \sqrt{2+\sqrt{2+\sqrt{2}}}, \quad \ldots \tag{95}$$

and then by using the law of addition, we can work with the angles $\frac{3\pi}{16}$, $\frac{5\pi}{8}$, etc.

To generate more angles, we *trisect* angles by finding $\cos x$ from the *triple angle* formula

$$\cos 3x = 4\cos^3 x - 3\cos x. \tag{96}$$

The special case $3x = \frac{\pi}{2}$ is the cubic equation

$$(4t^2 - 3)t = 4t^3 - 3t = \cos\tfrac{\pi}{2} = 0, \tag{97}$$

which gives

$$\cos\tfrac{\pi}{6} = \tfrac{\sqrt{3}}{2}, \qquad \text{and so} \qquad \sin\tfrac{\pi}{6} = \tfrac{1}{2}. \tag{98}$$

Repeated bisections then yield

$$2\cos\tfrac{\pi}{12} = \sqrt{2+\sqrt{3}}, \qquad 2\cos\tfrac{\pi}{24} = \sqrt{2+\sqrt{2+\sqrt{3}}}, \quad \ldots \tag{99}$$

On the other hand, $\frac{\pi}{6}$ cannot be trisected any further by using square roots, i.e., by using only straightedge and compass. This impossibility was proved only in the 19-th century, which means that countless ancient and medieval mathematicians tried to solve the problem in vain. They were extremely reluctant to accept solutions involving tools more general than straightedge and compass. In this light, the bisection formula (94) is a recipe to *bisect an angle*, and the special cubic equation (97) corresponds to a problem of constructing an *equilateral triangle* (or equivalently, a regular *hexagon*), all with only straightedge and compass.

We can go further, and consider angle quadrisection, quintisection, etc. Quadrisection is simply repeated bisections, so it would offer nothing new. The next possibility is to consider the *quintuple angle* formula

$$\cos 5x = 16\cos^5 x - 20\cos^3 x + 5\cos x. \tag{100}$$

In general, this is worse than the trisection problem, but the special case $5x = \frac{\pi}{2}$ leads to

$$16t^4 - 20t^2 + 5 = 0, \tag{101}$$

where we have already factored out the monomial t. This is a biquadratic equation, which can easily be solved to yield

$$\cos\tfrac{\pi}{10} = \tfrac{\sqrt{5+\sqrt{5}}}{2\sqrt{2}}. \tag{102}$$

Note that the corresponding classical Greek mathematics is a construction of a regular *pentagon* with straightedge and compass. Again, $\frac{\pi}{10}$ cannot be quintisected any further, but it can be trisected once (In fact, we can simply apply the law of addition to $\frac{\pi}{60} = \frac{\pi}{10} - \frac{\pi}{12}$). Thus we get some new angles, such as $\frac{\pi}{60}$, $\frac{\pi}{30}$, $\frac{\pi}{20}$, $\frac{\pi}{120}$, $\frac{\pi}{240}$, and the repertoire of these angles remained the same for over 2000 years, until Carl Friedrich Gauss discovered that the regular 17-gon can be constructed with straightedge and compass. This fact can be expressed as

$$16\cos\tfrac{2\pi}{17} = -1 + \sqrt{17} + \sqrt{34 - 2\sqrt{17}} + 2\sqrt{17 + 3\sqrt{17} - \sqrt{170 + 38\sqrt{17}}}. \tag{103}$$

It was Gauss' first major discovery, and was so special to Gauss that he wanted a regular 17-gon to be engraved on his tombstone.

Remark 14 (More general methods)**.** In view of the preceding remark, the classical Greeks could not get an acceptable exact answer for chord $\frac{\pi}{360} = 2\sin\frac{\pi}{720}$. The value chord $\frac{\pi}{480}$ was accessible, and so Claudius Ptolemy invoked the approximation chord $x \approx x$ for $x \approx 0$, to find

$$\operatorname{chord}\tfrac{\pi}{360} \approx \tfrac{4}{3}\operatorname{chord}\tfrac{\pi}{480}, \tag{104}$$

and used bisection and the law of addition to compute chord x for $x = \frac{\pi}{720}, \frac{3\pi}{720}, \frac{4\pi}{720}, \ldots$, resulting in a table with 3 hexagesimal digits of accuracy. From our perspective, what Ptolemy does is an argument reduction that ensures $|x| \leq \frac{\pi}{720}$, followed by the "power series" approximation $\sin x \approx x$. Notice the similarity with Briggs' method, cf. Remark 11, both in the repeated use of square root extractions in the argument reduction, and in the simplicity of the final approximation. Ptolemy's method was taken to the extreme by the early renaissance scholars, such as Regiomontanus (1436–1476), who computed the sines for every minute, with 7 decimals, and Rheticus (1514–1574), who produced, e.g., a table of sines for every 10", with the accuracy of 10 decimals.

On the other hand, around 650, Bhaskara I gave the excellent approximation

$$\cos x = \frac{\pi^2 - 4x^2}{\pi^2 + x^2} + E(x), \tag{105}$$

which (we now know) satisfies the error bound $|E(x)| \leq 0.002$ for $|x| \leq \frac{\pi}{2}$. It appears that the continued attempts to improve this formula by the medieval Indian mathematicians culminated in the discovery of power series by Madhava.

Starting with digit-by-digit algorithms for extracting square and cubic roots, medieval Indian, Chinese, and Islamic mathematicians developed iterative algorithms for solving polynomial equations. Thus, Al-Kashi (c. 1420) solved the cubic (96) for $\cos x$, to trisect $\frac{\pi}{120}$, and computed $\cos\frac{\pi}{360}$ with 9 hexagesimal digits of accuracy. In an apparently independent development, François Viète (c. 1600) discovered iterative procedures to solve polynomial equations associated to dividing an angle into 3, 5, and 7 equal pieces. Note that the solution of general cubics, exposed in 1545 by Gerolamo Cardano, means that angle trisection was possible by that time, if one allows cubic root extractions. Further research in iterative methods eventually led to the discovery of the Newton-Raphson method by Henry Briggs, Isaac Newton, and Joseph Raphson. Briggs applied his iterative method to compute trigonometric functions of very small angles, which served as the basis for his famous trigonometric tables. It appears that Briggs discovered the Newton-Raphson method well before the birth of either Newton or Raphson. Finally, we must mention Jost Bürgi, who discovered (c. 1592) an ingenious iterative method that produces an entire table of sine at once.

Exercise 10. Look for a function f satisfying $f''(x) + f(x) = 0$ in the form

$$f(x) = \sum_{n=0}^{\infty} a_n x^n. \tag{106}$$

(a) Putting $f(0) = 0$ and $f'(0) = 1$, arrive at the sine series in (78). Show that the series converges for all $x \in \mathbb{R}$.
(b) Putting $f(0) = 1$ and $f'(0) = 0$, arrive at the cosine series in (78). Show that the series converges for all $x \in \mathbb{R}$.

Exercise 11. Perform a roundoff error analysis of the argument reduction (84) applied k times.

Exercise 12. Come up with an argument reduction formula for $\tan x$, and design an algorithm to compute $\tan x$ in the Ptolemy-Briggs style, that performs the argument reduction sufficiently many times, before invoking $\tan x \approx x$.

6. Inverse trigonometric functions

The basic series for inverse trigonometric functions are the *arctangent series*

$$\arctan x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{2n+1} = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots \qquad (|x| \leq 1), \tag{107}$$

which was discovered independently by Madhava in the early 15th century, James Gregory in 1671, and Gottfried Leibniz in 1673, and the *arcsine series*

$$\arcsin x = x + \frac{1}{2} \cdot \frac{x^3}{3} + \frac{1 \cdot 3}{2 \cdot 4} \cdot \frac{x^5}{5} + \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6} \cdot \frac{x^7}{7} + \ldots \qquad (-1 \leq x < 1), \tag{108}$$

which was discovered independently by Isaac Newton and Gottfried Leibniz, during the period 1669–1676. The law of addition for the tangent implies the corresponding law

$$\arctan a + \arctan b = \arctan \frac{a+b}{1-ab} \qquad (ab < 1). \tag{109}$$

Putting $a = b$ in turn gives the doubling formula

$$\arctan x = 2 \arctan \frac{x}{1 + \sqrt{1+x^2}}, \tag{110}$$

which can be used to reduce the argument of (107). Moreover, we have relations among the inverse trigonometric functions, such as

$$\arcsin x = \arctan \frac{x}{\sqrt{1-x^2}}, \qquad \arccos x = \arctan \frac{\sqrt{1-x^2}}{x}, \tag{111}$$

meaning that the ability to compute either (107) or (108) would be sufficient.

Exercise 13. (a) By integrating

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \ldots, \tag{112}$$

derive the arctangent series (107).

(b) By expanding $(1-x^2)^{1/2}$ into a binomial series, cf. Exercise 2, and integrating term by term, derive the arcsine series (108).

(c) Derive the sine series (78), by inverting (108).

Exercise 14. Perform a roundoff error analysis of the argument reduction (110) applied k times. Design an algorithm to compute $\arctan x$ in the Ptolemy-Briggs style, that carries out the reduction (110) sufficiently many times, before invoking $\arctan x \approx x$.

7. Computation of π

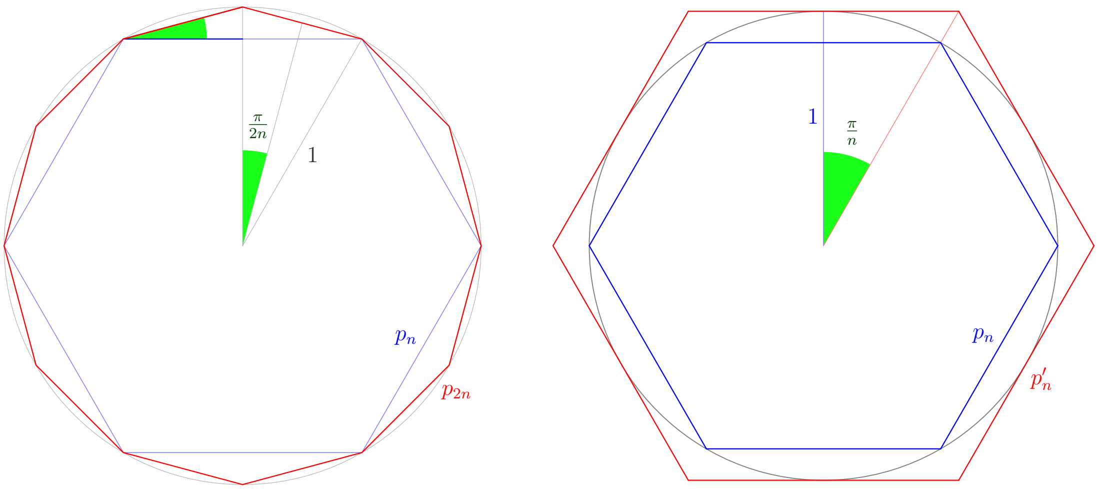
At the risk of deviating a bit from the main subject, as an interesting case study, we include here a brief historical review on methods to compute the digits of π, up to and including the point when power series dominated the field.

The first rigorous method to compute π is due to Archimedes of Syracuse, c. 250 BC. He approximated the circumference of a circle by using inscribed and circumscribed regular polygons with 6, 12, 24, 48, and finally 96 sides, to conclude

$$3\tfrac{10}{71} < \pi < 3\tfrac{1}{7}, \tag{113}$$

where we note that $3\frac{10}{71} = 3.1408\ldots$ and $3\frac{1}{7} = 3.1428\ldots$. To add a bit more detail, it is easy to see from Figure 2 that the perimeter of the regular n-gon inscribed in the unit circle is $2n \sin \frac{\pi}{n}$, and the perimeter of the regular n-gon circumscribed about the unit circle is $2n \tan \frac{\pi}{n}$, yielding

$$n \sin \tfrac{\pi}{n} < \pi < n \tan \tfrac{\pi}{n}. \tag{114}$$



(A) $p_n = 2n \sin \frac{\pi}{n}$ and $p_{2n} = p_n / \cos \frac{\pi}{2n}$ (B) $p'_n = 2n \tan \frac{\pi}{n} = p_n / \cos \frac{\pi}{n}$

Figure 2. Archimedean bounds on π.

From our perspective, this is immediate because $\sin x < x < \tan x$ for $0 < x < \frac{\pi}{2}$. To compute $\sin \frac{\pi}{96}$, Archimedes used the recurrence formula displayed in Figure 2(a), or equivalently

$$\sin \alpha = \frac{\sin 2\alpha}{2 \cos \alpha} = \frac{\sin 4\alpha}{2^2 \cos \alpha \cos 2\alpha} = \ldots = \frac{\sin 16\alpha}{2^4 \cos \alpha \cos 2\alpha \cos 4\alpha \cos 8\alpha}, \tag{115}$$

with $\alpha = \frac{\pi}{96}$, in combination with the angle bisection technique (94) for the cosines, cf. the values (99). Note that $\sin 16\alpha = \sin \frac{\pi}{6}$ is available. In the end, we get bounds such as

$$\begin{aligned} &1 < \frac{\pi}{3} < \frac{2}{\sqrt{3}} = 1.1547\ldots \\ &1 < \frac{\pi}{3} \cdot \frac{\sqrt{2+\sqrt{3}}}{2} < \frac{2}{\sqrt{2+\sqrt{3}}} = 1.0352\ldots \\ &1 < \frac{\pi}{3} \cdot \frac{\sqrt{2+\sqrt{3}}}{2} \cdot \frac{\sqrt{2+\sqrt{2+\sqrt{3}}}}{2} < \frac{2}{\sqrt{2+\sqrt{2+\sqrt{3}}}} = 1.0086\ldots. \end{aligned} \tag{116}$$

This algorithm leads to the infinite product expansion

$$\frac{3}{\pi} = \frac{\sqrt{2+\sqrt{3}}}{2} \cdot \frac{\sqrt{2+\sqrt{2+\sqrt{3}}}}{2} \cdot \frac{\sqrt{2+\sqrt{2+\sqrt{2+\sqrt{3}}}}}{2} \cdots, \tag{117}$$

although Archimedes would be extremely reluctant to consider such things.

Until the invention of calculus, the Archimedes algorithm and its variations were the only methods available for computing the digits of π. Thus Claudius Ptolemy used his accurate trigonometric table to compute the perimeter of the inscribed 360-gon, and deduced

$$\pi \approx 3\tfrac{17}{120} = 3.14166\ldots. \tag{118}$$

Around 250, independently of Archimedes, Liu Hui developed an algorithm based on the *areas* of inscribed and circumscribed polygons, and replaced (114) by the slightly improved version

$$n \sin \tfrac{\pi}{n} < \pi < n \sin \tfrac{\pi}{n} + d_n, \qquad \text{where} \quad d_n = n \sin \tfrac{\pi}{n} - \tfrac{n}{2} \sin \tfrac{2\pi}{n}. \tag{119}$$

Note that

$$2n\sin\tfrac{\pi}{2n} = n\sin\tfrac{\pi}{n} + d_{2n}, \qquad 4n\sin\tfrac{\pi}{4n} = n\sin\tfrac{\pi}{n} + d_{2n} + d_{4n}, \tag{120}$$

etc. He then observed

$$d_{2n} \approx \tfrac{1}{4}d_n, \tag{121}$$

so that the right hand side of

$$\pi = n\sin\tfrac{\pi}{n} + d_{2n} + d_{4n} + \ldots \approx n\sin\tfrac{\pi}{n} + \big(1 + \tfrac{1}{4} + (\tfrac{1}{4})^2 + \ldots\big)d_{2n} = n\sin\tfrac{\pi}{n} + \tfrac{4}{3}d_{2n}, \tag{122}$$

would be more accurate than the simple update $2n\sin\frac{\pi}{2n} = n\sin\frac{\pi}{n} + d_{2n}$. Liu Hui tried his accelerated method on a 192-gon, which gave the same accuracy as that of the non-accelerated method on a 3072-gon. This is a precursor to the modern acceleration techniques.

In 1579, François Viète derived the formula

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2}\cdot\frac{\sqrt{2+\sqrt{2}}}{2}\cdot\frac{\sqrt{2+\sqrt{2+\sqrt{2}}}}{2}\cdot\frac{\sqrt{2+\sqrt{2+\sqrt{2+\sqrt{2}}}}}{2}\cdots, \tag{123}$$

by modifying the Archimedes algorithm to start with a *square*, instead of a hexagon, cf. (117). This beautiful formula is considered to be the dawn of modern mathematics, as it was the first ever explicit occurrence of an infinite process in mathematics.

Table 4. Approximation of π by Archimedean algorithms.

Date	Name	Number of sides	Decimal places
250 BC	Archimedes	96	3
150	Ptolemy	360	3-4
250	Liu Hui	$6\cdot 2^{9}$	5
480	Zu Chongzhi	$6\cdot 2^{12}$?	7
499	Aryabhata	384?	4
1424	Al-Kashi	$6\cdot 2^{27}$	14
1579	Viète	$6\cdot 2^{16}$	9
1593	van Roomen	2^{30}	15
1596	van Ceulen	$15\cdot 2^{31}$	20
1615	van Ceulen	2^{62}	33
1621	Snell	2^{30}	35
1630	Grienberger	2^{40}	38

Table 4 lists some of the notable progress achieved with the help of Archimedean algorithms. The last and most impressive of the more straightforward computations were done by Ludolph van Ceulen, when he used a polygon of 2^{62} sides, to derive

$$\pi = 3.14159265358979323846264338327950288\ldots. \tag{124}$$

In order to explain how the final accuracy depends on the number of sides of the polygon used, we note that

$$n\sin\tfrac{\pi}{n} = \pi - \frac{\pi^3}{6n^2} + O\Big(\frac{1}{n^4}\Big), \qquad n\tan\tfrac{\pi}{n} = \pi + \frac{\pi^3}{3n^2} + O\Big(\frac{1}{n^4}\Big). \tag{125}$$

Neglecting the higher order term, we see that doubling n would reduce the error roughly 4 times, i.e., going from n to $2n$ would add $\log_{10} 4 \approx 0.6$ correct decimal digits. For instance, from van Roomen's computation ($n = 2^{30}$) to van Ceulen's computation ($n = 2^{62}$), there are 32 doublings of n, which nicely explains why van Ceulen has $18 \approx 0.6\cdot 32$ additional correct decimal digits.

From Table 4, we see that Willebrord Snell had a success comparable to van Ceulen's by using a polygon with "only" 2^{30} sides. This is because he observed that the particular combination $\frac{2}{3}n\sin\frac{\pi}{n}+\frac{1}{3}n\tan\frac{\pi}{n}$ of the perimeters of the inscribed and circumscribed polygons converges faster than either of the perimeters. A mathematical explanation was given by Christiaan Huygens in 1654. From our perspective, the expansions (125) immediately yield

$$\tfrac{2}{3}n\sin\tfrac{\pi}{n}+\tfrac{1}{3}n\tan\tfrac{\pi}{n}=\pi+O\Big(\frac{1}{n^4}\Big). \tag{126}$$

Hence Snell's method converges twice as fast, in the sense that doubling n adds approximately 1.2 correct decimal digits. For instance, Snell was able to squeeze out 7 correct decimal digits from Archimedes' 96 sided polygon. It is also easy to see that Liu Hui's accelerated formula (122) has the same quality:

$$n\sin\tfrac{\pi}{n}+\tfrac{4}{3}d_{2n}=\pi+O\Big(\frac{1}{n^4}\Big),\qquad\text{because}\qquad d_{2n}=\frac{\pi^3}{8n^3}+O\Big(\frac{1}{n^4}\Big). \tag{127}$$

Table 5. Performance of Archimedean algorithms. Here n designates the number of doublings, i.e., $n=1$ corresponds to a hexagon or a square, and $n=2$ corresponds to a dodecagon or an octagon, etc.

n	Archimedes (117)	Liu Hui (127)	Viète (123)	Snell (126)
1	**3**.0000000000000000	**3**.0000000000000000	2.8284271247461903	**3**.0000000000000000
2	3.**1**058285412302489	3.14**1**1047216403318	**3**.0614674589207183	3.1**4**23491305446567
3	3.1326286132812382	3.141**5**619706315674	3.**1**214451522580529	3.14**1**6390562199918
4	3.1393502030468672	3.1415**9**07329687442	3.1365484905459398	3.1415**9**55404083902
5	3.14**1**0319508905098	3.14159**2**5335050572	3.1**4**03311569547530	3.14159**2**8338087959
6	3.1414524722854620	3.141592**6**460837791	3.14**1**2772509327729	3.141592**6**648502488
7	3.141**5**576079118575	3.14159265**3**1206555	3.141**5**138011443009	3.1415926**5**42935209
8	3.1415838921483181	3.141592653**5**604717	3.1415729403670913	3.14159265**3**6337748
9	3.1415**9**04632280500	3.1415926535**8**79608	3.1415877252771600	3.141592653**5**925420
10	3.14159**2**1059992713	3.14159265358**9**6785	3.1415**9**14215112002	3.14159265358**9**9645
11	3.1415925166921572	3.141592653589**7**856	3.14159**2**3455701176	3.1415926535898038

Just before the advent of power series, John Wallis discovered the noteworthy formula

$$\frac{2}{\pi}=\frac{1}{2}\cdot\frac{3}{2}\cdot\frac{3}{4}\cdot\frac{5}{4}\cdot\frac{5}{6}\cdot\frac{7}{6}\cdot\frac{7}{8}\cdot\frac{9}{8}\cdots, \tag{128}$$

which was modified by William Brouncker into the continued fraction

$$\frac{4}{\pi}=1+\cfrac{1}{2+\cfrac{9}{2+\cfrac{25}{2+\cfrac{49}{2+\cdots}}}}. \tag{129}$$

Both formulas were published in 1655, and π was computed to a few decimal places. However, these formulas converge too slowly to have any practical utility.

We now turn to power series. The arctangent series (107) gives the nice looking formula

$$\frac{\pi}{4}=\arctan 1=1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\dots, \tag{130}$$

but its convergence is again incredibly slow. For instance, to get an accuracy of 9 decimal digits, one needs to sum 1 billion terms. A better option is to use the relation $\tan\frac{\pi}{6} = \frac{1}{\sqrt{3}}$, which yields

$$\frac{\pi}{6} = \frac{1}{\sqrt{3}}\Big(1 - \frac{1}{3}\cdot\frac{1}{3^2} + \frac{1}{5}\cdot\frac{1}{3^4} - \frac{1}{7}\cdot\frac{1}{3^6} + \ldots\Big). \tag{131}$$

This was in fact used by Madhava to reach 11 decimal digits of accuracy.

In 1665, Isaac Newton used the expansion

$$\begin{aligned}\frac{\pi}{24} - \frac{\sqrt{3}}{32} &= \int_0^{1/4} \sqrt{x(1-x)}dx \\ &= \frac{1}{12} - \frac{1}{2^5\cdot 5} - \frac{1}{2^9\cdot 7} - \frac{1}{2^{13}\cdot 9} - \ldots - \frac{(2n-3)!!}{2^{3n+2}n!(2n+3)} - \ldots,\end{aligned} \tag{132}$$

to compute π to 16 decimals.

Then van Ceulen's record was finally broken by Abraham Sharp in 1699, when he computed 71 decimal digits of π. He used the series (131), as well as its variations based on other known values of $\tan x$, such as $\tan\frac{\pi}{8} = \sqrt{2} - 1$.

The next advance came in 1706, when John Machin used the formula

$$\frac{\pi}{4} = 4\arctan\frac{1}{5} - \arctan\frac{1}{239}, \tag{133}$$

to cross the 100 decimals mark. This formula is remarkable, because the powers of $\frac{1}{5}$ are easily computed in base 10, and the series for $\arctan\frac{1}{239}$ converges very rapidly. Since then, a wealth of similar formulas have been discovered. For example, we have

$$\frac{\pi}{4} = 5\arctan\frac{1}{7} + 2\arctan\frac{3}{79}, \tag{134}$$

due to Jurij Vega and Leonhard Euler (c. 1780), and

$$\frac{\pi}{4} = 12\arctan\frac{1}{18} + 8\arctan\frac{1}{57} - 5\arctan\frac{1}{239}, \tag{135}$$

due to Carl Friedrich Gauss (1863). Table 6 illustrates some of the aforementioned power series methods for computing π in action.

Table 6. Performance of power series formulas for π. Note that Gauss' formula saturates the machine arithmetic beyond 6 terms.

n	Madhava-Sharp (131)	Newton (132)	Machin (133)	Gauss (135)
1	**3**.4641016151377544	**3**.2990381056766580	3.**1**832635983263602	3.**1**443881670703955
2	3.0792014356780038	3.**1**490381056766577	3.1**4**05970293260603	3.141**5**875736078829
3	3.**1**561814715699539	3.1423416771052288	3.14**1**6210293250346	3.141592**6**647657662
4	3.1378528915956800	3.14**1**6906354385628	3.1415**9**17721821773	3.14159265**3**5629728 →
5	3.1**4**26047456630841	3.1416074056800398	3.141592**6**824043994	3.1415926535**8**98602
6	3.14**1**3087854628827	3.1415**9**50812734890	3.1415926**5**26153086	3.1415926535897**9**27
7	3.1416743126988367	3.1415930785574249	3.14159265**3**6235550	3.1415926535897927
8	3.141**5**687159417836	3.14159**2**7314480228	3.141592653**5**886025	3.1415926535897927
9	3.1415**9**97738115049	3.141592**6**683631725	3.1415926535**8**98362	3.1415926535897927
10	3.1415905109380793	3.1415926**5**64721790	3.1415926535897**9**22	3.1415926535897927
11	3.1415933045030808	3.1415926541650681	3.1415926535897940	3.1415926535897927

In Table 7, we list some of the historic computations. Perhaps the last great manual computation of π was that of William Shanks. He went back to Machin's formula (133), and computed 707 decimal digits, but it was later discovered that "only" 527 of them were correct. In the early days of the electronic computer era, it must have been difficult to resists the temptation to see how many more digits of π the newly invented machines can produce. In fact, computation of π became a standard test for new computers. As for the algorithms, almost all computations up to the late 1970's were done by employing Machin-like formulas. For instance, the 100,000 decimals mark was first crossed by Daniel Shanks and John Wrench, who used the Gauss formula (135), in combination with

$$\frac{\pi}{4} = 6 \arctan \frac{1}{8} + 2 \arctan \frac{1}{57} + \arctan \frac{1}{239}, \tag{136}$$

due to Carl Størmer (1896).

Table 7. Approximation of π by power series.

Date	Name	Series	Number of terms	Decimal places
1400	Madhava	(131)	21	11
1665	Isaac Newton	(132)	-	16
1699	Abraham Sharp	(131)	-	71
1706	John Machin	(133)	-	100
1789	Jurij Vega	(134)	-	136
1873	William Shanks	(133)	510	707 (527)
1961	Daniel Shanks team	(135)-(136)	-	100,000
2002	Yasumasa Kanada team	(137)-(138)	-	1.24 trillion

From the late 1970's, more sophisticated algorithms, based on ideas such as Ramanujan series and arithmetic-geometric mean iterations dominated the scene, but power series methods still remain competitive. This is evidenced by the record computation of Yasumasa Kanada and his team, performed in 2002 to find 1.24 trillion decimal digits of π. They used the Machin-like formulas

$$\frac{\pi}{4} = 44 \arctan \frac{1}{57} + 7 \arctan \frac{1}{239} - 12 \arctan \frac{1}{682} + 24 \arctan \frac{1}{12943}, \tag{137}$$

due to Størmer (1896), and

$$\frac{\pi}{4} = 12 \arctan \frac{1}{49} + 32 \arctan \frac{1}{57} - 5 \arctan \frac{1}{239} + 12 \arctan \frac{1}{110443}, \tag{138}$$

due to Kikuo Takano (1982).

Exercise 15. Prove Liu Hui's upper bound (119) by showing that

$$\sin x < 2 \sin x - \frac{\sin 2x}{2} \qquad \text{for} \quad 0 < x < \frac{\pi}{2}. \tag{139}$$

To compare this with the Archimedes' upper bound (114), show also that

$$2 \sin x - \frac{\sin 2x}{2} < \tan x - \sin x \qquad \text{for} \quad 0 < x < \frac{\pi}{2}. \tag{140}$$

Exercise 16. For $0 \leq a \leq 1$, show that

$$\int_0^a \sqrt{x(1-x)} dx = \frac{1}{8}\big(\tfrac{\pi}{2} - \arcsin b - b\sqrt{1-b^2}\big), \qquad \text{where} \quad b = 1 - 2a. \tag{141}$$

Then expanding the factor $\sqrt{1-x}$ in the integrand by the binomial theorem (18), termwise integrating the resulting series, and finally putting $a = \frac{1}{4}$, prove Newton's formula (132).