Parabola Volume 28, Issue 2 (1992)

CALCULATING EXPONENTIALS AND LOGARITHMS

Bill McLean*

The exponential function $x \mapsto e^x$ and its inverse, the (natural) logarithm function $x \mapsto \ln x$ $(x > 0)$, are amongst the most important in mathematics, arising as they do in many different applications. At school, we learn some of the properties of these functions, such as the identities

$$e^{x+y} = e^x e^y \tag{1}$$

and

$$\ln(xy) = \ln x + \ln y, \tag{2}$$

but are not usually told how numerical values of e^x and $\ln x$ can be found. Consequently, you just have to trust your calculator when it tells you, say, that $e^{2{\cdot}045} = 7{\cdot}729\ldots$, and in bygone days you would have had to trust your table of mathematical functions or your slide rule. What if, however, you were perverse enough *not* to trust these? How could you calculate exponentials and logarithms using only a pencil and paper? Here, the term "pencil and paper" is intended metaphorically: the calculation should involve only simple arithmetic operations — additions, subtractions, multiplications and divisions — that could in principle be done by hand.

I am going to describe methods for calculating e^x and $\ln x$ based on the identities (1) and (2). Similar methods have in fact been used by some electronic calculators and computers, although manufacturers of these devices usually do not reveal details of the algorithms employed.

Unfortunately, the methods involve a certain amount of cheating: it is necessary to know the values of $\ln(10)$ and of the numbers $a_0, a_1, a_2, \ldots$ defined by

$$a_j = \ln(1 + 10^{-j}) \quad \text{for } j \geq 0. \tag{3}$$

In a future article, I will explain how to calculate these numbers by pencil and paper; for the moment you can just use your calculator (!) to check that

$$\ln(10) \doteq 2{\cdot}3025\,8509.$$

* Bill is an applied mathematician at the University of New South Wales

Here, I have used the symbol $\doteq$ to indicate that the decimal expansion has been rounded to the number of digits shown. The first few of the numbers a_j are

$$a_0 \doteq 6931\,4718,$$

$$a_1 \doteq 0953\,1018,$$

$$a_2 \doteq 0099\,5033,$$

$$a_3 \doteq 0009\,9950,$$

$$a_4 \doteq 0001\,0000,$$

and you will notice that

$$a_0 > a_1 > a_2 > \cdots > 0$$

because $\ln x$ decreases when x decreases, and because $\ln x > 0$ for $x > 1$. It is also possible to prove that

$$a_j < 10^{-j} \qquad \text{and} \qquad a_j < 10a_{j+1} \quad \text{for all } j \geq 0; \tag{4}$$

in fact, $a_j \approx 10^{-j}$ once j is larger than 1, and the approximation gets better and better as j increases.

Let $f(x)$ be either e^x or $\ln x$. Our overall strategy is to approximate x by a number x_d with the property that $f(x_d)$ can be evaluated using only simple arithmetic operations. Since $x \approx x_d$, the continuity of f ensures that $f(x) \approx f(x_d)$, but it is important to quantify the errors in these approximations, a task for which the following terminology is useful. In general, if $x \approx \hat{x}$, then the number $|\hat{x} - x|$ is called the *absolute error* in $\hat{x}$ as an approximation to x, whereas $|\hat{x} - x|/|x|$ is called the *relative error*, because it tells you the size of the error relative to the size of x. (If $x = 0$, then the relative error is not defined.) For example, in the approximation $752{\cdot}351 \approx 750$, the absolute error is

$$|750 - 752{\cdot}351| = 2{\cdot}351 \doteq 2,$$

but the relative error is only

$$\frac{|750 - 752{\cdot}351|}{752{\cdot}351} = \frac{2{\cdot}351}{752{\cdot}351} \doteq 0{\cdot}003,$$

or, in other words, only 0·3%. For numbers of moderate size it does not matter much whether you look at absolute or relative errors, but for numbers that are either very large or very small it is usually more meaningful to consider relative errors.

EXPONENTIALS

We are now ready to discuss the calculation of e^x, where x may be any real number, positive or negative (or zero). The first step is to find the unique integer N such that

$$N \ln(10) \leq x < (N+1) \ln(10), \tag{5}$$

and to compute the quantity

$$r_0 = x - N \ln(10),$$

so that

$$x = N \ln(10) + r_0 \quad \text{with} \quad 0 \leq r_0 < \ln(10). \tag{6}$$

A moment's thought shows that our pencil and paper criterion is satisfied, given the value of $\ln(10)$. All we have done, in effect, is to divide x by $\ln(10)$ to obtain a quotient N and a remainder r_0, something that can be achieved, if $x > 0$, by repeated subtraction of $\ln(10)$ from x until the result is less than $\ln(10)$. If $x < 0$, we instead use repeated addition of $\ln(10)$ until the result is greater than or equal to zero. Also, it is worth pointing out that (5) is equivalent to

$$10^N \leq e^x < 10^{N+1},$$

so already at this point in the calculation we know the order of magnitude of e^x, i.e., we know its value to the nearest power of 10.

The next step is to divide r_0 by a_0 to obtain an integer quotient k_0 and a remainder r_1, so that

$$r_0 = k_0 a_0 + r_1 \qquad \text{with} \qquad 0 \leq r_1 < a_0.$$

We repeat this procedure d times:

$$\begin{aligned} r_1 &= k_1 a_1 + r_2 \qquad \text{with} \qquad 0 \leq r_2 < a_1, \\ &\vdots \\ r_d &= k_d a_d + r_{d+1} \qquad \text{with} \qquad 0 \leq r_{d+1} < a_d. \end{aligned}$$

With the help of (4), we find that $k_j a_j = r_j - r_{j+1} \leq r_j < a_{j-1} < 10 a_j$ and so

$$0 \leq k_j < 10 \quad \text{for } j \geq 0.$$

Substituting the expressions for $r_0, r_1, \ldots, r_d$ into (6), one after another, we find that

$$\begin{aligned} x &= N\ln(10) + k_0a_0 + r_1 \\ &= N\ln(10) + k_0a_0 + k_1a_1 + r_2 \\ &\vdots \\ &= N\ln(10) + k_0a_0 + k_1a_1 + \cdots + k_da_d + r_{d+1}, \end{aligned}$$

and therefore

$$x = x_d + r_{d+1} \qquad \text{with} \qquad 0 \le r_{d+1} < a_d, \tag{7}$$

where x_d is defined by

$$x_d = N\ln(10) + k_0a_0 + k_1a_1 + \cdots + k_da_d.$$

The identity (1) now plays its crucial role, and at the same time the reason for the mysterious definition (3) of a_j is revealed. Indeed,

$$e^{a_j} = 1 + 10^{-j},$$

so

$$e^{x_d} = 10^N(1+1)^{k_0}(1+10^{-1})^{k_1}\cdots(1+10^{-d})^{k_d}.$$

How accurate is e^{x_d} as an approximation to e^x? In view of (7), we have

$$\frac{|e^{x_d} - e^x|}{|e^x|} = \frac{|e^{x_d} - e^{x_d+r_{d+1}}|}{e^{x_d+r_{d+1}}} = e^{-r_{d+1}}(e^{r_{d+1}} - 1) < e^{a_d} - 1 = 10^{-d},$$

which means that

$$e^x \approx e^{x_d} \quad \text{with relative error less than } 10^{-d}.$$

You might like to check that you understand the method by trying it out in the case when $x = 2.045$ and $d = 3$. You should find that $N = 0$, $k_0 = 2$, $k_1 = 6$, $k_2 = 8$ and $k_3 = 7$, so that

$$x_d = 2a_0 + 6a_1 + 8a_2 + 7a_3 \doteq 2{\cdot}04475$$

and

$$e^{x_d} = (1+1)^2(1+10^{-1})^6(1+10^{-2})^8(1+10^{-3})^7 \doteq 7{\cdot}72726.$$

By comparison, $e^x \doteq 7{\cdot}72915$, giving an absolute error of about $0{\cdot}2 \times 10^{-2}$ and a relative error of about $0{\cdot}2 \times 10^{-3}$.

Natural Logarithms

Now consider the calculation of $\ln x$, where x may be any positive real number. (Remember that $\ln x$ is only defined for $x > 0$.) This time, the first step is to find the unique integer N satisfying

$$10^{N-1} < x \leq 10^N,$$

and then to let $t_0 = 10^{-N}x$, so that

$$x = 10^N t_0 \qquad \text{with} \qquad 10^{-1} < t_0 \leq 1. \tag{8}$$

The second step is to find the unique integer k_0 such that

$$2^{-k_0-1} < t_0 \leq 2^{-k_0}.$$

In other words, $2^{k_0}t_0 \leq 1 < 2^{k_0+1}t_0$, so we just have to calculate $2t_0$, 2^2t_0, 2^3t_0, ... until arriving at a number, namely 2^{k_0+1}, that exceeds 1. Having found k_0, we let $t_1 = 2^{k_0}t_0$, so that

$$t_0 = (1+1)^{-k_0}t_1 \qquad \text{with} \qquad (1+1)^{-1} < t_1 \leq 1.$$

The next d steps give us, in the same way,

$$\begin{aligned} t_1 &= (1+10^{-1})^{-k_1}t_2 \qquad \text{with} \qquad (1+10^{-1})^{-1} < t_2 \leq 1, \\ &\vdots \\ t_d &= (1+10^{-d})^{-k_d}t_{d+1} \qquad \text{with} \qquad (1+10^{-d})^{-1} < t_{d+1} \leq 1. \end{aligned}$$

Notice that $1 \geq t_{j+1} = (1+10^{-j})^{k_j}t_j > (1+10^{-j})^{k_j}(1+10^{1-j})^{-1}$, so

$$(1+10^{-j})^{k_j} < 1+10^{1-j}.$$

Taking logarithms, $k_j a_j < a_{j-1}$, and it then follows from the second inequality in (4) that

$$0 \leq k_j < 10 \qquad \text{for } 0 \leq j \leq d,$$

just as was the case before in the calculation of e^x.

Substituting the expressions for $t_0, t_1, \ldots, t_d$ into (8), one after another, gives

$$\begin{aligned} x &= 10^N(1+1)^{-k_0}t_1 \\ &= 10^N(1+1)^{-k_0}(1+10^{-1})^{-k_1}t_2 \\ &\vdots \\ &= 10^N(1+1)^{-k_0}(1+10^{-1})^{-k_1}\cdots(1+10^{-d})^{-k_d}t_{d+1}, \end{aligned}$$

and so

$$x = x_d t_{d+1} \qquad \text{with} \qquad (1+10^{-d})^{-1} < t_{d+1} \leq 1,$$

where

$$x_d = 10^N(1+1)^{-k_0}(1+10^{-1})^{-k_1}\cdots(1+10^{-d})^{-k_d}.$$

Using the functional identity (2), and recalling the definition of a_j in (3), we see that

$$\ln x_d = N\ln(10) - k_0 a_0 - k_1 a_1 - \cdots - k_d a_d.$$

Futhermore,

$$|\ln x_d - \ln x| = |\ln x_d - \ln(x_d t_{d+1})| = \ln(t_{d+1}^{-1}) < \ln(1+10^{-d}) = a_d < 10^{-d},$$

which means that

$$\ln x \approx \ln x_d \quad \text{with absolute error less than } 10^{-d}.$$

As an example, try taking $x = 13.412$ and $d = 3$. You should find that $N = 2$, $k_0 = 2$, $k_1 = 6$, $k_2 = 5$ and $k_3 = 1$, so that

$$x_d = 10^2(1+1)^{-2}(1+10^{-1})^{-6}(1+10^{-2})^{-5}(1+10^{-3})^{-1} \doteq 13.41353$$

and

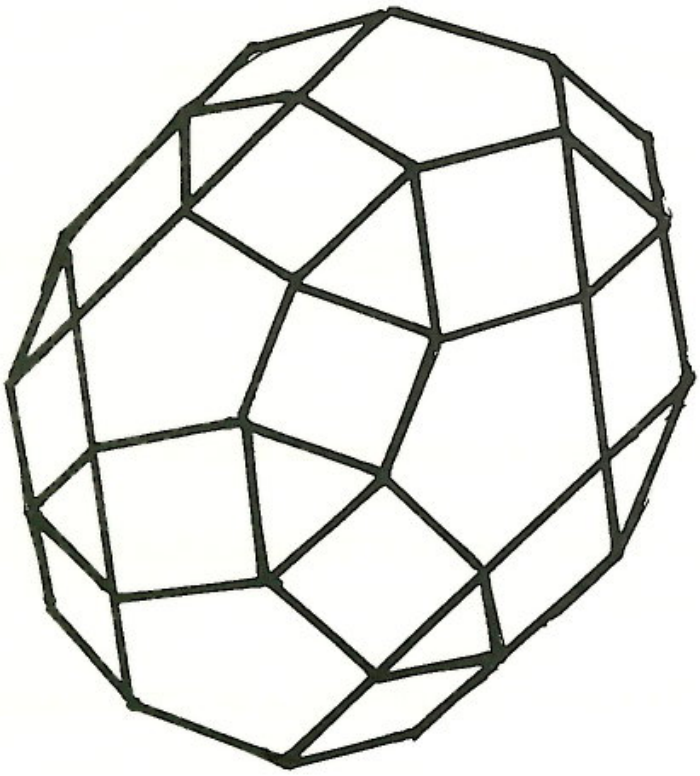
$$\ln x_d = 2\ln(10) - 2a_0 - 6a_1 - 5a_2 - a_3 \doteq 2{\cdot}59626.$$

By comparison, $\ln x \doteq 2.59615$, giving an absolute error of about $0{\cdot}1 \times 10^{-3}$.

A feature of the calculations described above is that they involve repeated multiplication by $1 + 10^{-j}$, an operation that only requires a shift of the decimal point and an addition: for example,

$$\begin{aligned} . \quad (1 + 10^{-3}) \times 7.4625 = \quad & 7{\cdot}4625 \\ & + 0{\cdot}0074\,625 \\ = \quad & 7{\cdot}4699\,625 \doteq 7{\cdot}4700. \end{aligned}$$

Contrast this with the multiplication of two general d-digit numbers a and b. Each of the d digits of a must be multiplied by b, and the d results added with appropriate shifting of decimal points to yield $a \times b$. In the interests of computational efficiency, it is desirable to minimise the use of such long multiplications, and likewise of long divisions.



Rhombitruncated Dodecahedron