1 Error in Euler's Method

Experience with Euler's[1] method raises some interesting questions about numerical approximations for the solutions of differential equations.

1. What determines the amount of numerical error in an approximation?
2. Why is that halving the step size tends to decrease the numerical error in Euler's method by one-half and the numerical error in the modified Euler method by one-quarter?
3. Are some differential equations more difficult to approximate numerically than others? If so, can this be predicted without doing numerical experiments?

None of these questions has a simple answer, but we will at least be able to offer a partial answer for each.

MODEL PROBLEM 1

The initial value problem

$$\frac{dy}{dt} = \frac{2y - 18t}{1+t}, \quad y(0) = 4$$

has solution $y = 4 + 8t - 5t^2$. Determine how many subdivisions are needed to assure that the Euler's method approximation on the interval $[0, 2]$ has error no greater than 0.001.

INSTANT EXERCISE 1
Verify the solution formula for Model Problem 1.

The Meaning of ***Error***

Any approximation of a function necessarily allows a possibility of deviation from the correct value of the function. **Error** is the term used to denote the amount by which an approximation fails to equal the exact solution (exact solution minus approximation).[2] Error occurs in an approximation for several reasons.

[1] pronounced "Oy-ler"

[2] In everyday vocabulary, we tend to use the term *error* to denote an avoidable human failing, as for example when a baseball player mishandles the ball. In numerical analysis, error is a characteristic of an approximation correctly performed. Unacceptable error dictates a need for greater refinement or a better method.

Truncation error in a numerical method is error that is caused by using simple approximations to represent exact mathematical formulas. The only way to completely avoid truncation error is to use exact calculations. However, truncation error can be reduced by applying the same approximation to a larger number of smaller intervals or by switching to a better approximation. Analysis of truncation error is the single most important source of information about the theoretical characteristics that distinguish better methods from poorer ones. With a combination of theoretical analysis and numerical experiments, it is possible to estimate truncation error accurately.

Round-off error in a numerical method is error that is caused by using a discrete number of significant digits to represent real numbers on a computer. Since computers can retain a large number of digits in a computation, round-off error is problematic only when the approximation requires that the computer subtract two numbers that are nearly identical. This is exactly what happens if we apply an approximation to intervals that are too small. Thus, the effort to decrease truncation error can have the unintended consequence of introducing significant round-off error.

Practitioners of numerical approximation are most concerned with truncation error, but they also try to restrict their efforts at decreasing truncation error to improvements that do not introduce significant round-off error. In our study of approximation error, we consider only truncation error. We seek information about error on both a local and global scale. *Local truncation error* is the amount of truncation error that occurs in one step of a numerical approximation. *Global truncation error* is the amount of truncation error that occurs in the use of a numerical approximation to solve a problem.

Taylor's Theorem and Approximations

The principal tool in the determination of truncation error is Taylor's theorem, a key result in calculus that provides a formula for the amount of error in using a truncated Taylor series to represent the corresponding function. Taylor's theorem in its most general form applies to approximations of any degree; here we present only the information specifically needed to analyze the error in Euler's method.

Theorem 1 *Let y be a function of one variable having a continuous second derivative on some interval $I = [0, t_f]$ and let f be a function of two variables having continuous first partial derivatives on the rectangular region $R = I \times [y_0 - r, y_0 + s]$, with $r, s > 0$.*[3]

1. *Given T and t in I, there exists some point $\tau \in I$ such that*

$$y(t) = y(T) + y'(T)(t-T) + \frac{y''(\tau)}{2}(t-T)^2. \tag{1}$$

2. *Given points (T, Y) and (T, y) in R, there exists some point $(T, \eta) \in R$ such that*

$$f(T, y) = f(T, Y) + f_y(T, \eta)(y - Y), \tag{2}$$

where f_y is the partial derivative of f with respect to its second argument.

[3] This notation means that the first variable is in the interval I and the second is in the interval $[y_0 - r, y_0 + s]$.

Theorem 1 serves to quantify the idea that the difference in function values for a smooth function should vanish as the evaluation points become closer.

One can be a little more restrictive when specifying the range of possible values for τ and η; however, nothing is gained by doing so. We cannot use Theorem 1 to *compute* the error in an approximation. The theorem provides formulas for the error, but the catch is that there is no way to determine τ and η without knowing the exact solution. It may seem that this catch makes the formulas useless, but this is not the case. We do know that τ and η are confined to a given closed interval, and therefore we can compute worst-case values for the quantities $y''(\tau)$ and $f_y(T, \eta)$.

Example 1

Suppose we want to approximate the function $\ln(0.5 + t)$ near the point $T = 0.5$. We have

$$y(t) = \ln(0.5 + t), \quad y(T) = 0, \quad y'(t) = \frac{1}{0.5 + t}, \quad y'(T) = 1, \quad y''(t) = \frac{-1}{(0.5 + t)^2}.$$

Let's assume that our goal[4] is to find an upper bound for the largest possible error in approximating $y(t)$ by $t - 0.5$ with $0 \leq t \leq 1$. The approximation formula of Equation (1) yields

$$y(t) = 0 + (t - 0.5) + \frac{y''(\tau)}{2}(t - 0.5)^2$$

and the approximation error E is defined by

$$E = (t - 0.5) - y(t).$$

We therefore have

$$|E| = \frac{|y''(\tau)|}{2}(t - 0.5)^2,$$

with $0 \leq \tau \leq 1$. Now let

$$M = \max_{0 \leq t \leq 1} |y''(t)| = \max_{0 \leq t \leq 1} \left| \frac{-1}{(0.5 + t)^2} \right| = 4.$$

Given the range of possible τ values, the worst case is $|y''(\tau)| = M$; thus,

$$|E| \leq 2(t - 0.5)^2, \quad t \in [0, 1].$$

◇

Local Truncation Error for Euler's Method

Consider an initial value problem

$$y' = f(t, y(t)), \quad y(0) = y_0, \tag{3}$$

where f has continuous first partial derivatives on some region R defined by $0 \leq t \leq t_f$ and $y_0 - r \leq y \leq y_0 + s$. Euler's method is a scheme for obtaining an approximate value y_{n+1} for

[4] This problem is an artificial one because we know a formula for y and can therefore calculate the error exactly. However, the point is to illustrate how information about derivatives of y can be used to generate error estimates for cases where we *don't* have a formula for y.

$y(t_{n+1})$ using only the approximation y_n for $y(t_n)$ and the function f that calculates the slope of the solution curve through any point. Specifically, the method is defined by the formula

$$y_{n+1} = y_n + hf(t_n, y_n), \quad \text{where} \quad h = t_{n+1} - t_n. \tag{4}$$

We define the **global truncation error** at step n in any numerical approximation of (3) by

$$E_n = y(t_n) - y_n. \tag{5}$$

Our aim is to find a worst-case estimate of the global truncation error in the numerical scheme for Euler's method, assuming that the correct solution and numerical approximation stay within R. This is a difficult task because we have so little to work with. We'll start by trying to determine the relationship between the error at time t_{n+1} and the error at time t_n. Figure 1 shows the relationships among the relevant quantities involved in one step of the approximation. The approximation and exact solution at each of the two time steps are related by the error definitions. The approximations at the two time steps are related by Euler's method. We still need to have a relationship between the exact solution values at the two time steps, and this is where Theorem 1 is needed.

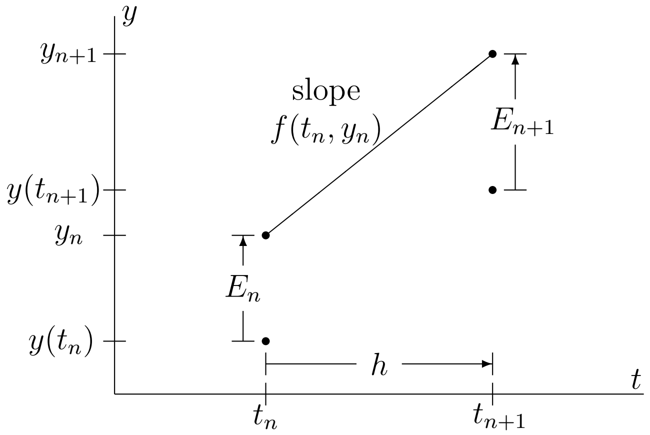

Figure 1: The relationships between the solution values and Euler approximations

The comparison of $y(t_{n+1})$ and $y(t_n)$ must begin with Equation (1). Combining Equation (1), with $T = t_n$, and Equation (4) yields

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \frac{h^2}{2}y''(\tau). \tag{6}$$

This result is a step in the right direction, but it is not yet satisfactory. A useful comparison of $y_(t_{n+1})$ with $y(t_n)$ can have terms consisting entirely of known quantities and error terms, but not quantities, such as $f(t_n, y(t_n))$, that need to be evaluated at points that are not known exactly. These quantities have to be approximated by known quantities. This is where Equation (2) comes into the picture. Substituting

$$f(t_n, y(t_n)) = f(t_n, y_n) + f_y(t_n, \eta)[y(t_n) - y_n] = f(t_n, y_n) - f_y(t_n, \eta)E_n$$

into Equation (6) gives us

$$y(t_{n+1}) = y(t_n) + hf(t_n, y_n) - hf_y(t_n, \eta)E_n + \frac{h^2}{2}y''(\tau). \tag{7}$$

Equation (7) meets our needs because every term is either a quantity of interest, a quantity that can be evaluated, or an error term. Subtracting this equation from the Euler formula (4) yields

$$y_{n+1} - y(t_{n+1}) = y_n - y(t_n) + hf_y(t_n, \eta)E_n - \frac{h^2}{2}y''(\tau),$$

or

$$E_{n+1} = [1 + hf_y(t_n, \eta)]E_n - \frac{h^2}{2}y''(\tau). \tag{8}$$

This result indicates the relationship between the errors at successive steps. The **local truncation error** is defined to be the error in step $n + 1$ when there is no error in step n; hence, the local truncation error for Euler's method is $-h^2y''(\tau)/2$. The local truncation error has two factors of h, and we say that it is $\mathcal{O}(h^2)$.[5] The quantity $[1 + hf_y(t_n, \eta)]E_n$ represents the error at step $n + 1$ caused by the error at step n. This propagated error is larger than E_n if $f_y > 0$ and smaller than E_n if $f_y < 0$. Having $f_y < 0$ is generally a good thing because it causes truncation errors to diminish as they propagate.

Global Truncation Error for Euler's Method

As we see by Equation (8), the error at a given step consists of error propagated from the previous steps along with error created in the current step. These errors might be of opposite signs, but the quality of the method is defined by the worst case rather than the best case. The worst case occurs when y'' and E_n have opposite algebraic signs and $f_y > 0$. Thus, we can write

$$|E_{n+1}| \leq (1 + h|f_y(t_n, \eta)|)\ |E_n| + \frac{|y''(\tau)|}{2}h^2. \tag{9}$$

The quantity $f_y(t_n, \eta)$ cannot be evaluated, but we do at least know that f_y is continuous. We can also show that y'' is continuous. By using the chain rule, we can differentiate the differential equation (3) to obtain the formula

$$y'' = f_t + f_y y' = f_t + ff_y. \tag{10}$$

The problem statement requires f to have continuous first derivatives, so y'' must also be continuous. Equation (10) also gives us a way to calculate $y''(\tau)$ at any point $(\tau, y(\tau))$ in R. Given that the region R is closed and bounded, we can apply the theorem from multivariable calculus that says that a continuous function on a closed and bounded region has a maximum and a minimum value. Hence, we can define positive numbers K and M by

$$K = \max_{(t,y)\in R} |f_y(t, y)| < \infty, \quad M = \max_{(t,y)\in R} |(f_t + ff_y)(t, y)| < \infty. \tag{11}$$

For the worst case, we have to assume $f_y(t_n, \eta) = K$ and $|y''(\tau)| = |(f_t + ff_y)(\tau, y(\tau))| = M$, even though neither of these is likely to be true. Substituting these bounds into the error

[5] This is read as "big oh of h^2."

estimate of Equation (9) puts an upper bound on the size of the truncation error at step $n+1$ in terms of the size of the truncation error at step n and global properties of f:

$$|E_{n+1}| \leq (1+Kh)|E_n| + \frac{Mh^2}{2}. \tag{12}$$

In all probability, there is a noticeable gap between actual values of $|E_{n+1}|$ and the upper bound in Equation (12). The upper bound represents not the worst case for the error, but the error for the unlikely case where each quantity has its worst possible value.

Some messy calculations are needed to obtain an upper bound for the global truncation error from the error bound of Equation (12). First, we need an expression for the error bound that does not depend on knowledge of earlier error bounds. We begin Euler's method with the initial condition; hence, there is no error at step 0. From the error bound,

$$|E_1| \leq \frac{Mh^2}{2}.$$

Continuing, we have

$$|E_2| \leq (1+Kh)\frac{Mh^2}{2} + \frac{Mh^2}{2}$$

and

$$|E_3| \leq (1+Kh)^2\frac{Mh^2}{2} + (1+Kh)\frac{Mh^2}{2} + \frac{Mh^2}{2} = \frac{Mh^2}{2}\sum_{i=0}^{2}(1+Kh)^i.$$

Clearly this result generalizes to

$$|E_n| \leq \frac{Mh^2}{2}S(h), \quad S(h) = \sum_{i=0}^{n-1}(1+Kh)^i.$$

The quantity $S(h)$ is a finite geometric series, and these can be calculated by a clever trick. We have

$$KhS(h) = (1+Kh)S(h)-S(h) = \sum_{i=1}^{n}(1+Kh)^i - \sum_{i=0}^{n-1}(1+Kh)^i = (1+Kh)^n - 1 = (1+Kh)^t_n/h-1.$$

This result simplifies the global error bound to

$$|E_n| \leq \frac{Mh}{2K}[(1+Kh)^{t_n/h}-1].$$

We can use calculus techniques to compute

$$\lim_{h\to 0}(1+Kh)^{t_n/h} = e^{Kt_n}$$

and to show that $(1+Kh)^{t_n/h}$ is a decreasing function of h for $Kh < 1$. Thus, for h small enough, we have the result

$$|E_n| \leq \frac{Mh}{2K}(e^{Kt_n}-1).$$

Theorem 2 summarizes the global truncation error result.

Theorem 2 *Let I and R be defined as in Theorem 1 and let K and M be defined as in Equation (11). Let $y_1, y_2, \ldots, y_N$ be the Euler approximations at the points $t_n = nh$. If the points (t_n, y_n) and $(t_n, y(t_n))$ are in R, and $Kh < 1$, then the error in the approximation of $y(t)$ is bounded by*

$$|E| \leq \frac{M}{2K}\left(e^{Kt} - 1\right)h.$$

There are two important conclusions to draw from Theorem 2. The first is that the error vanishes as $h \to 0$. Thus, the truncation error can be made arbitrarily small by reducing h, although doing so eventually causes large round-off errors. The second conclusion is that the error in the method is approximately proportional to h. The **order** of a numerical method is the number of factors of h in the global truncation error estimate for the method. Euler's method is therefore a **first-order** method. Halving the step-size in a first-order method reduces the error by approximately a factor of 2. As with Euler's method, the order of most methods is 1 less than the power of h in the local truncation error.

Returning to Model Problem 1, we can compute the global error bound by finding the values of K and M. Given the function f from Model Problem 1, we have

$$f_y(t,y) = \frac{2}{1+t}, \quad (f_t + ff_y)(t,y) = \frac{2y - 36t - 18}{(1+t)^2}.$$

The interval I is $[0,2]$, as prescribed by the problem statement. The largest and smallest values of y on I can be determined by graphing the exact solution, finding the vertex algebraically, or applying the standard techniques of calculus. By any of these methods,[6] we see that $0 \leq y \leq 7.2$. Clearly, f_y is always positive and achieves its largest value in R on the boundary $t = 0$; hence, $K = 2$. It can be shown that $f_t + ff_y$ is negative throughout R. Since $f_t + ff_y$ is increasing in y, its minimum (maximum in absolute value) must occur on the line $y = 0$. In fact, the minimum occurs at the origin, so $M = 18$.

INSTANT EXERCISE 2

Derive the formula for $f_t + ff_y$ for Model Problem 1. Also verify that $f_t + ff_y$ is negative throughout R and that the minimum of $f_t + ff_y$ on R occurs at the origin.

With these values of K and M for Model Problem 1, Theorem 2 yields the estimate

$$|E_n| \leq 4.5\left(e^{2t_n} - 1\right)h.$$

In particular, at $t_N = 2$, the error estimate is approximately $241h$. Thus, setting $241h < 0.001$ is sufficient to achieve the desired error tolerance. Since $Nh = 2$, the conclusion is that $N = 482,000$ is large enough to guarantee the desired accuracy. This number is larger than what would be needed in actual practice, owing to our always assuming the worst in obtaining the error bound. The important part of Theorem 2 is the order of the method, not the error estimate. Knowing the order of the method allows us to obtain an accurate estimate of the error from numerical experiments. Figure 2 shows the error in the Euler approximations at $t = 2$ using various step sizes. The points fall very close to a straight line, indicating the typical first-order behavior. From the slope of the line, we estimate the actual numerical error to be approximately $e_N = 30h$, which is roughly one-eighth of that computed using the error bound of Theorem 2. This still indicates that 60,000 steps are required to obtain an error within the specified tolerance of 0.001. The desired accuracy ought better be attempted using a better method.

[6] We are admittedly cheating by using the solution formula, which we do not generally know. In practice, one can use a numerical solution to estimate the range of possible y values.

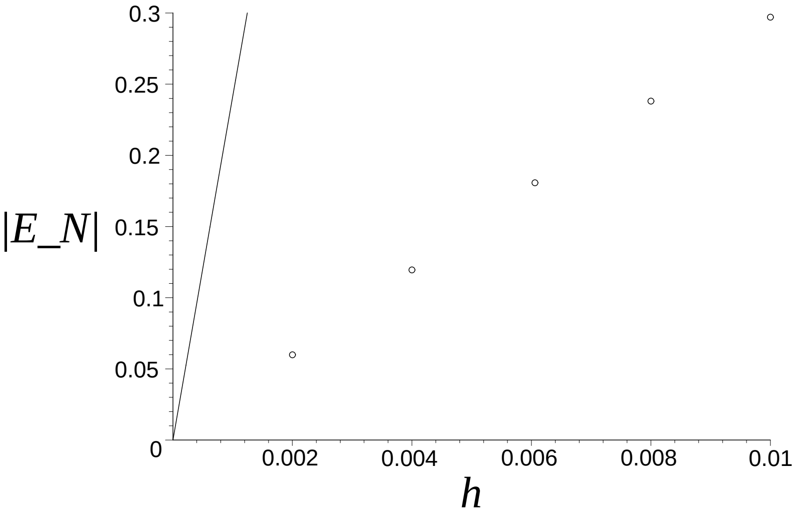

Figure 2: The Euler's method error in $y(2)$ for Model Problem 1, using several values of h, along with the theoretical error bound $241h$

Conditioning

Why does Euler's method do so poorly with Model Problem 1? Recall the interpretation of the local truncation error formula (8). Propagated error increases with n if $f_y > 0$ and decreases with n if $f_y < 0$. A differential equation is said to be **well-conditioned** if $\partial f/\partial y < 0$ and **ill-conditioned** if $\partial f/\partial y > 0$. Ill-conditioned equations are ones for which the truncation errors are magnified as they propagate. The reason that ill-conditioned problems magnify errors can be seen by an examination of the slope field for an ill-conditioned problem.

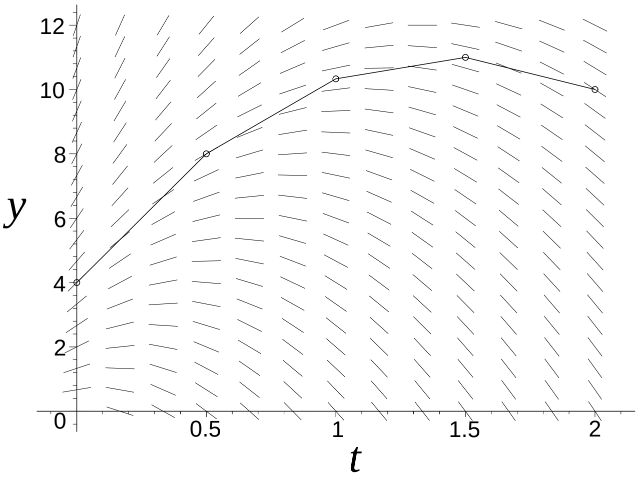

Figure 3: The Euler's method approximation for Model Problem 1 using $h = 0.5$, along with the slope field

Figure 3 shows the slope field for Model Problem 1 along with the numerical solution using just 4 steps. The initial point (0,4) is on the correct solution curve, but the next approximation point (0.5,8) is not. This point is, however, on a different solution curve. If we could avoid making any further error beyond $t = 0.5$, the numerical approximation would follow *a* solution curve from then on, but it would not be the *correct* solution curve. The amount of error ultimately caused by the error in the first approximation step can become larger or smaller at later times, depending on the relationship between the different solution curves. Figure 4 shows our approximation again, along with the solution curves that pass through all the points used in the approximation. The solution curves are spreading apart with time, carrying the approximation farther away from the correct solution. The error introduced at each step moves the approximation onto a solution curve that is farther yet from the correct one.

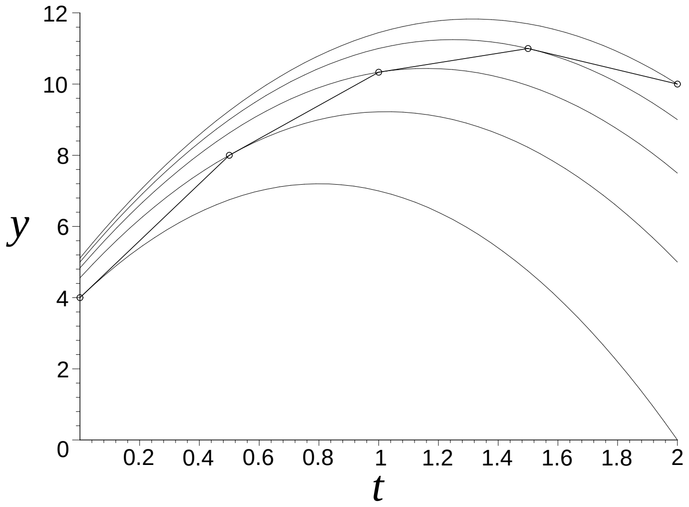

Figure 4: The Euler's method approximation for Model Problem 1 using $h = 0.5$, along with the solution curves that pass through the points used in the approximation

Recall that for Model Problem 1, we have

$$\frac{\partial f}{\partial y} = \frac{2}{1+t}.$$

The magnitude of f_y ranges from 2 at the beginning to 2/3 at $t = 2$, and yet this degree of ill-conditioning is sufficient to cause the large truncation error. The error propagation is worse for a differential equation that is more ill-conditioned than our model problem. In general, Euler's method is inadequate for any ill-conditioned problem. Well-conditioned problems are "forgiving" in the sense that numerical error moves the solution onto a different path that converges toward the correct path. Nevertheless, Euler's method is inadequate even for well-conditioned problems if a high degree of accuracy is required, owing to the slow first-order convergence. Commercial software generally uses fourth-order methods.

1 INSTANT EXERCISE SOLUTIONS

1. From $y = 4+8t-5t^2$, we have $y' = 8-10t$. Thus, $(1+t)y' = 8-2t-10t^2$ and $2y-18t = 8-2t-10t^2$.

2. We have

$$\frac{\partial f}{\partial t} = \frac{-18(1+t)+(18t-2y)}{(1+t)^2} = \frac{-2y-18}{(1+t)^2}$$

and

$$f\,\frac{\partial f}{\partial y} = \frac{2y-18t}{1+t}\,\frac{2}{1+t} = \frac{4y-36t}{1+t}.$$

Thus,

$$y'' = \frac{-2y-18}{(1+t)^2} + \frac{4y-36t}{(1+t)^2} = \frac{2y-36t-18}{(1+t)^2}.$$

Given $y=0$, $y'' = g(t) = -18(1+2t)/(1+t)^2$. The derivative of this function is

$$g' = -18\frac{2(1+t)^2 - 2(1+2t)(1+t)}{(1+t)^4} = -18\frac{2(1+t)-2(1+2t)}{(1+t)^3} = \frac{36t}{(1+t)^3} \geq 0;$$

hence, the minimum of g occurs at the smallest t, namely $t=0$.