1 Simple Euler method and its modifications

1.1 Simple Euler method for the 1st-order IVP

Consider the IVP

$$y'(x) = f(x, y), \qquad y(x_0) = y_0 \,. \tag{1.1}$$

Let: $x_i = x_0 + i\,h, \quad i = 0, 1, \ldots n$

$y_i = y(x_i)$ — true solution evaluated at points x_i

Y_i — the solution to be calculated numerically.

Replace

$$y'(x) \longrightarrow \frac{Y_{i+1} - Y_i}{h} \,.$$

Then Eq. (1.1) gets replaced with

$$Y_{i+1} = Y_i + h\, f(x_i, Y_i) \qquad Y_0 = y_0 \,. \tag{1.2}$$

1.2 Local error of the simple Euler method

The calculated solution satisfies Eq. (1.2). Next, assuming that the true solution of IVP (1.1) has (at least) a second derivative $y''(x)$, one can use the Taylor expansion to write:

$$\begin{aligned} y_{i+1} = \;\; y(x_i + h) = &\;\; y_i + y_i'\, h + y_i''(x_i^*)\, \frac{h^2}{2} \\ = &\;\; y_i + h\, f(x_i, y_i) + O(h^2) \,. \end{aligned} \tag{1.3}$$

Here x_i^* is some point between x_i and $x_{i+1} = x_i + h$, and we have used Eq. (0.5).

Notation $O(h^k)$ for any k means the following:

$$q = O(h^k) \qquad \text{whenever} \qquad \lim_{h \to 0} \frac{q}{h^k} = \text{const} < \infty \,, \;\; \text{const} \neq 0 \,.$$

For example,

$$5h^2 + 1000h^3 = O(h^2); \qquad \text{or} \qquad \frac{h}{1 + h\, \cos(3 + 2h)} = O(h) \,.$$

We now introduce a new notation. The *local truncation error* shows how well the solution Y_{i+1} of the finite-difference scheme approximates the exact solution y_{i+1} of the ODE *at point x_{i+1}, assuming that at x_i the two solutions were the same, i.e. $Y_i = y_i$*. Comparing the last line of Eq. (1.3) with Eq. (1.2), we see that the local truncation error of the simple Euler method is $O(h^2)$. It tends to zero when $h \to 0$.

Another useful notation is that of *discretization error*. It shows how well the finite-difference scheme approximates the ODE. Let us now estimate this error. First, we note from (1.2) and (1.3) that the computed and exact solutions satisfy:

$$\frac{Y_{i+1} - Y_i}{h} = f(x_i, Y_i) \qquad \text{and} \qquad \frac{y_{i+1} - y_i}{h} = f(x_i, y_i) + O(h),$$

whence the discretization error of the simple Euler method is seen to be $O(h)$.

1.3 Global error of the Euler method; Propagation of errors

As we have said above, the local truncation error shows how well the computed solution approximates the exact solution at one given point, assuming that these two solutions have been the same up to that point. However, as we compute the solution of the finite-difference scheme, the local truncation errors at each step accumulate. This results in that the difference between the computed solution Y_i and exact solution y_i at some point x_i down the line becomes *much greater* than the local truncation error.

Let $\epsilon_i = y_i - Y_i$ denote the error (the difference between the true and computed solutions) at $x = x_i$. This error (or, sometimes, its absolute value) is called the *global error* of the finite-difference method.

Our goal in this subsection will be to find an *upper bound* for this error. Let us emphasize that finding an upper bound for the error rather than the error itself is the best one can do. (Indeed, if one could have found the actual error ϵ_i, one would have then simply added it to the numerical solution Y_i and obtained the exact solution y_i.) The main purpose of finding the upper bound for the error is to determine how it depends on the step size h. We will do this now for the simple Euler method (1.2).

To this end, we begin by considering Eq. (1.2) and the 1st line of Eq. (1.3):

$$\begin{aligned} Y_{i+1} &= Y_i + h\, f(x_i, Y_i) \\ y_{i+1} &= y_i + h\, f(x_i, y_i) + \frac{h^2}{2} y''(x_i^*) \end{aligned}$$

Subtract the 1st equation above from the 2nd to obtain the error at x_{i+1}:

$$\epsilon_{i+1} = \epsilon_i + h\, \left(f(x_i, y_i) - f(x_i, Y_i) \right) + \frac{h^2}{2} y''(x_i^*)\,. \tag{1.4}$$

Now apply the "triangle inequality", valid for any three numbers a, b, c:

$$a = b + c \qquad \Rightarrow \qquad |a| \le |b| + |c|, \tag{1.5}$$

to Eq. (1.4) and obtain:

$$\begin{aligned} |\epsilon_{i+1}| \le &\quad |\epsilon_i| + hL|\epsilon_i| + \frac{h^2}{2} |y''(x_i^*)| \\ = &\quad (1 + hL)|\epsilon_i| + \frac{h^2}{2} |y''(x_i^*)|\,. \end{aligned} \tag{1.6}$$

In writing the second term in the above formula, we used the fact that $f(x, y)$ satisfies the Lipschitz condition with respect to y (see Lecture 0).

To complete finding the upper bound for the error $|\epsilon_{i+1}|$, we need to estimate $y''(x_i^*)$. We use the Chain rule for a function of two variables (recall Calculus III) to obtain:

$$y''(x) = \frac{d^2 y(x)}{dx^2}\Big|_{\text{use the ODE}} = \frac{df(x, y)}{dx} = f_x \frac{dx}{dx} + f_y \frac{dy}{dx} = f_x + f_y f\,. \tag{1.7}$$

Considering the first term on the r.h.s. of (1.7), let us assume that

$$|f_x| \le M_1 \qquad \text{for some } M_1 < \infty. \tag{1.8}$$

In cases when this asumption does not hold (as, for example, for $f(x, y) = x^{1/3} \sin \frac{1}{x}$), the estimate obtained below (see (1.16)) is not valid, but a modified estimate can usually be found on a case-by-case basis. So here we proceed with assumption (1.8).

Considering the second term on the r.h.s. of (1.7), we first recall that f satisfies the Lipschitz condition with respect to y, which means that

$$|f_y| \le M_2 \qquad \text{for some } M_2 < \infty, \tag{1.9}$$

except possibly at a finite number of y-values where f_y does not exist (like at $y = 0$ for $f(y) = |y|$). Finally, the other factor of the second term on the r.h.s. of (1.7) is also bounded, because f is assumed to be continuous and on the closed region R (see the Existence and Uniqueness Theorem in Lecture 0). Thus,

$$|f| \le M_3 \qquad \text{for some } M_3 < \infty. \tag{1.10}$$

Combining Eqs. (1.7–1.10), we see that

$$|y''(x_i^*)| \le M_1 + M_2 M_3 \equiv M < \infty\,. \tag{1.11}$$

Now combining Eqs. (1.6) and (1.11), we obtain:

$$|\epsilon_{i+1}| \le (1 + hL)|\epsilon_i| + \frac{h^2}{2}M\,. \tag{1.12}$$

This last equation implies that $|\epsilon_{i+1}| \le z_{i+1}$, where z_{i+1} satisfies the following recurrence equation:

$$z_{i+1} = (1 + hL)z_i + \frac{h^2}{2}M\,, \qquad z_0 = 0\,. \tag{1.13}$$

(Condition $z_0 = 0$ follows from the fact that $\epsilon_0 = 0$; see the initial conditions in Eqs. (1.1) and (1.2).)

Thus, the error $|\epsilon_i|$ is bounded by z_i, and we need to solve Eq. (1.13) to find that bound. The way to do so is analogous to solving a linear inhomogeneous equation (see Section 0.4). However, before we obtain the solution, let us develop an intuitive understanding of what kind of answer we should expect. To this end, let us assume for the moment that $L = 0$ in Eq. (1.13). Then we have:

$$\begin{aligned} z_{i+1} = \quad & z_i + \frac{h^2}{2}M = \quad (z_{i-1} + \frac{h^2}{2}M) + \frac{h^2}{2}M \; = \; \dots \\ = \quad & z_0 + \frac{h^2}{2}M \cdot i = \quad 0 + \frac{h^2}{2}M \cdot \frac{x_i - x_0}{h} \; = \; h \cdot \frac{M(x_i - x_0)}{2} \; = \; O(h)\,. \end{aligned}$$

That is, the *global error* $|\epsilon_i|$ should have the size $O(h)$. In other words,

$$\boxed{\begin{array}{rcl} \text{Global error} & = & \text{Number of steps} \ \times \ \text{Local error} \\ & \text{or} & \\ O(h) & = & O\left(\frac{1}{h}\right) \ \times \ O(h^2) \end{array}}$$

Now let us show that a similar estimate also holds for $L \ne 0$. First, solve the homogeneous version of (1.13):

$$z_{i+1} = (1 + hL)\, z_i \qquad \Rightarrow \qquad z_{i,\text{hom}} = (1 + hL)^i\,. \tag{1.14}$$

Note that this is an analogue of $e^{a(x_i - x_0)}$ in Section 0.4, because

$$(1 + hL)^i = (1 + hL)^{(x_i - x_0)/h}|_{h \to 0} \approx e^{L(x_i - x_0)},$$

where we have used the definition of x_i, found after (1.1), and also the results of Section 0.5.

In close analogy to the method used in Section 0.4, we seek the solution of (1.13) in the form $z_i = c_i z_{i,\text{hom}}$ (with $c_0 = 0$). Substituting this form into (1.13) and using Eq. (1.14), we obtain:

$$c_{i+1}(1+hL)^{i+1} = (1+hL)\cdot c_i(1+hL)^i + \frac{h^2}{2}M \qquad \Rightarrow$$

$$\begin{aligned} c_{i+1} &= c_i + \frac{h^2M}{2\,(1+hL)^{i+1}} = c_{i-1} + \frac{h^2M}{2\,(1+hL)^{i+1-1}} + \frac{h^2M}{2\,(1+hL)^{i+1}} \\ &= \ldots = c_0 + \sum_{k=1}^{i+1} \frac{h^2M}{2}\frac{1}{(1+hL)^k} \\ &= \big|_{\text{geometric series}} \quad \frac{h^2M}{2(1+hL)}\frac{1-\frac{1}{(1+hL)^{i+1}}}{1-\frac{1}{(1+hL)}} = \frac{hM}{2L}\left(1-\frac{1}{(1+hL)^{i+1}}\right). \end{aligned} \tag{1.15}$$

Combining (1.14) and (1.15), and using (0.16), we finally obtain:

$$z_{i+1} = \frac{hM}{2L}\left((1+hL)^{i+1}-1\right) = \frac{hM}{2L}\left((1+hL)^{(x_{i+1}-x_0)/h}-1\right) \approx \frac{hM}{2L}\left(e^{L(x_{i+1}-x_0)}-1\right) = O(h),$$

$$\Rightarrow$$

$$|\epsilon_{i+1}| \le \frac{hM}{2L}\left(e^{L(x-x_0)}-1\right) = O(h)\,. \tag{1.16}$$

This is the upper bound for the global error of the simple Euler method (1.2).

Thus, in the last two subsections, we have shown that for the simple Euler method:

- Local truncation error $=$ $O(h^2)$;
- Discretization error $=$ $O(h)$;
- Global error $=$ $O(h)$.

The exponent of h in the global error is often referred to as *the order* of the finite-difference method. Thus, the simpler Euler method is the 1st-order method.

Question: How does the above bound for the error change when we include the machine round-off error (which occurs because numbers are computed with finite accuracy, usually 10^{-16})?

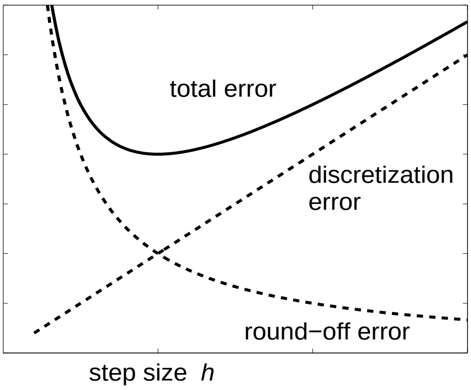
Answer: In the above formulae, replace $h^2M/2$ by $h^2M/2 + r$, where r is the maximum value of the round-off error. Then Eq. (1.16) gets replaced with

$$|\epsilon_{i+1}| \le \left(\frac{h^2M}{2}+r\right)\frac{1}{hL}\left(e^{L(x-x_0)}-1\right) = \left(\frac{hM}{2L}+\frac{r}{hL}\right)\left(e^{L(x-x_0)}-1\right) \tag{1.17}$$

The r.h.s. of the above bound is schematically plotted in the figure below. We see that for *very small* h, the term r/h can be dominant.

Moral:

Decreasing the step size of the difference equation does not always result in the increased accuracy of the obtained solution.



1.4 Modifications of the Euler method

In this subsection, our goal is to find finite-difference schemes which are more accurate than the simple Euler method (i.e., the global error of the sought methods should be $O(h^2)$ or better).

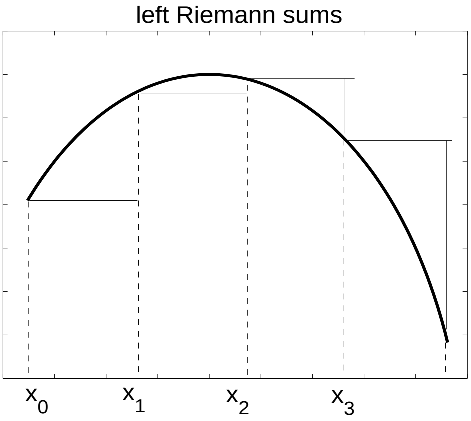
Again, we first want to develop an intuitive understanding of how this can be done, and then actually do it. So, to begin, we notice an obvious fact that the ODE $y' = f(x, y)$ is just a more general case of $y' = f(x)$. The solution of the latter equation is $y = \int f(x)dx$. Whenever we cannot evaluate the integral analytically in closed form, we resort to approximating the integral by the Riemann sums.

A very crude approximation to $\int_a^b f(x)dx$ is provided by the left Riemann sums:

$$Y_{i+1} = Y_i + h\, f(x_i)\,.$$

This is the analogue of the simple Euler method (1.2):

$$Y_{i+1} = Y_i + h\, f(x_i, Y_i)\,.$$



Approximations of the integral $\int_a^b f(x)dx$ that are known to be more accurate than the left Riemann sums are the Trapezoidal Rule and the Midpoint Rule:

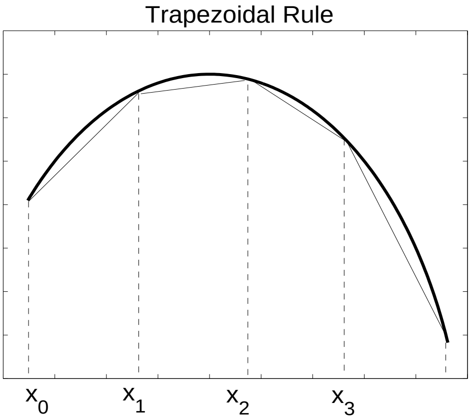
Trapezoidal Rule:

$$Y_{i+1} = Y_i + h\frac{f(x_i) + f(x_{i+1})}{2}\,.$$

Its analogue for the ODE is to look like this:

$$Y_{i+1} = Y_i + \frac{h}{2}\left(f(x_i, Y_i) + f(x_{i+1}, Y_i + Ah)\right)\,, \qquad (1.18)$$

where the coefficient A is to be determined. Method (1.18) is called the *Modified Euler method.*



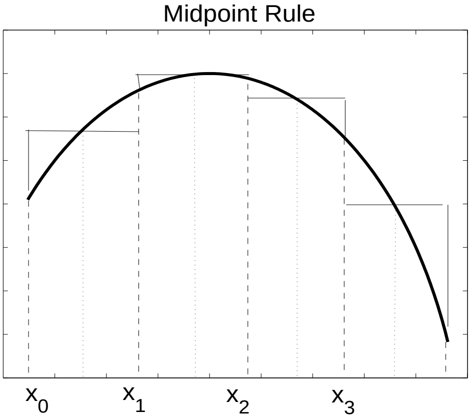
Midpoint Rule:

$$Y_{i+1} = Y_i + h\,f\left(x_i + \frac{h}{2}\right)\,.$$

Its analogue for the ODE is to look like this:

$$Y_{i+1} = Y_i + h\,f\left(x_i + \frac{h}{2},\, Y_i + Bh\right)\,, \quad (1.19)$$

where the coefficient B is to be determined. We will refer to method (1.19) as the *Midpoint method.*



The coefficients A in (1.18) and B in (1.19) are determined from the requirement that the corresponding finite-difference scheme have the global error $O(h^2)$ (as opposed to the simple Euler's $O(h)$), or equivalently, the local truncation error $O(h^3)$. Below we will determine the value of A. You will be asked to compute B along similar lines in one of the homework problems.

To determine the coeficient A in the Modified Euler method (1.18), let us rewrite that equation while Taylor-expanding its r.h.s. using Eq. (0.6) with $\Delta x = h$ and $\Delta y = Ah$:

$$\begin{aligned} Y_{i+1} = &\; Y_i + \frac{h}{2}f(x_i, Y_i) + \frac{h}{2}\left(f(x_i, Y_i) + [hf_x(x_i, Y_i) + (Ah)f_y(x_i, Y_i)] + O(h^2)\right) \\ = &\; Y_i + hf(x_i, Y_i) + \frac{h^2}{2}\left(f_x(x_i, Y_i) + Af_y(x_i, Y_i)\right) + O(h^3)\,. \end{aligned} \qquad (1.20)$$

Equation (1.20) yields the Taylor expansion of the *computed* solution Y_{i+1}. Let us compare it with the Taylor expansion of the *exact solution* $y(x_{i+1})$. To simplify the notations, we will denote $y'_i = y'(x_i)$, etc. Then, using Eq. (1.7):

$$\begin{aligned} y_{i+1} = &\; y_i + hy'_i + \frac{h^2}{2}y''_i + O(h^3) \\ = &\; y_i + hf(x_i, y_i) + \frac{h^2}{2}\left(f_x(x_i, y_i) + f(x_i, y_i)f_y(x_i, y_i)\right) + O(h^3)\,. \end{aligned} \qquad (1.21)$$

Upon comparing the last lines of Eqs. (1.20) and (1.21), we see that in order for method (1.18) to have the local truncation error of $O(h^3)$, one should take $A = f(x_i, Y_i)$.

Thus, the Modified Euler method can be programmed into a computer code as follows:

$$\begin{cases} Y_0 = y_0, \\ \bar{Y}_{i+1} = Y_i + hf(x_i, Y_i), \\ Y_{i+1} = Y_i + \dfrac{h}{2}\left(f(x_i, Y_i) + f(x_{i+1}, \bar{Y}_{i+1})\right) . \end{cases} \tag{1.22}$$

Remark: An alternative way to code in the last line of the above equation is

$$Y_{i+1} = \frac{1}{2}\left(Y_i + \bar{Y}_{i+1} + hf(x_{i+1}, \bar{Y}_{i+1})\right) . \tag{1.23}$$

This way is more efficient, because it requires only one evaluation of function f, which is usually the most time-consuming operation, while the last line of (1.22) requires two function evaluations.

In a homework problem, you will show that in Eq. (1.19), $B = \frac{1}{2}f(x_i, Y_i)$. Then the Midpoint method can be programmed as follows:

$$\begin{cases} Y_0 = y_0, \\ Y_{i+\frac{1}{2}} = Y_i + \dfrac{h}{2}f(x_i, Y_i), \\ Y_{i+1} = Y_i + hf\left(x_i + \dfrac{h}{2}, Y_{i+\frac{1}{2}}\right) . \end{cases} \tag{1.24}$$

Both the Modified Euler and the Midpoint methods have the local truncation error of $O(h^3)$ and the discretization and global errors of $O(h^2)$. Thus, these are the 2nd-order methods. The derivation of the local truncation error for the Modified Euler method is given in the Appendix to this section. This derivation will be needed for solving some of the homework problems.

Remark about notations: Different books use different names for the methods which we have called the Modified Euler and Midpoint methods.

1.5 An alternative way to improve the accuracy of a finite-difference method: Richardson method / Romberg extrapolation

We have shown that the global error of the simple Euler method is $O(h)$, which means that

$$Y_i^h = y_i + O(h) = y_i + (a\,h + b\,h^2 + \ldots) = y_i + a\,h + O(h^2) \tag{1.25}$$

where a, b, etc. are some constant coefficients that depend on the function f and its derivatives (as well as on the values of x), *but not on h*. The superscript h in Y_i^h means that this particular numerical solution has been computed with the step size h. We can now halve the step size and re-compute $Y_i^{h/2}$, which will satisfy

$$Y_i^{h/2} = y_i + \left(a\frac{h}{2} + b\left(\frac{h}{2}\right)^2 + \ldots\right) = y_i + \left(a\frac{h}{2} + O(h^2)\right) . \tag{1.26}$$

Let us clarify that $Y_i^{h/2}$ is *not* the numerical solution at $x_i + (h/2)$ but rather the numerical solution computed from x_0 up to x_i with the step size $(h/2)$.

Equations (1.25) and (1.26) form a system of linear equations for the unknowns a and y_i. Solving this system, we find

$$y_i = 2Y_i^{h/2} - Y_i^h + O(h^2)\,. \qquad (1.27)$$

Thus, a better approximation to the exact solution than either Y_i^h or $Y_i^{h/2}$ is $Y_i^{\text{improved}} = 2Y_i^{h/2} - Y_i^h$.

The above method of improving accuracy of the computed solution is called either the Romberg extrapolation or Richardson method. It works for any finite-difference scheme, not just for the simple Euler. However, it is not computationally efficient. For example, to compute Y^{improved} as per Eq. (1.27), one requires one function evaluation to compute Y_{i+1}^h from Y_i^h and two function evaluations to compute $Y_{i+1}^{h/2}$ from $Y_i^{h/2}$ (since we need to use two steps of size $h/2$ each). Thus, the total number of function evaluations to move from point x_i to point x_{i+1} is three, compared with two required for either the Modified Euler or Midpoint methods.

1.6 Appendix: Derivation of the local truncation error of the Modified Euler method

The idea of this derivation is the same as in Section 1.2, where we derived an estimate for the local truncation error of the simple Euler method. The details of the present derivation, however, are more involved. In particular, we will use the following formula, obtained similarly to (1.7):

$$\begin{aligned} y'''(x) &= \frac{d^3y(x)}{dx^3}\,\Big|_{\text{use the ODE}} = \frac{d^2f(x,y)}{dx^2}\,\Big|_{\text{use (1.7)}} \\ &= (f_x + f_y f)_x \frac{dx}{dx} + (f_x + f_y f)_y \frac{dy}{dx}\,\Big|_{\text{use the Product rule}} \\ &= f_{xx} + f_x f_y + 2ff_{xy} + f(f_y)^2 + f^2 f_{yy}\,. \end{aligned} \qquad (1.28)$$

Let us recall that in deriving the local truncation error at point x_{i+1}, one always assumes that the exact solution y_i and the computed solution Y_i at the previous step (i.e. at point x_i) are equal: $y_i = Y_i$. Also, for brevity of notations, we will write f without arguments to mean either $f(x_i, y_i)$ or $f(x_i, Y_i)$:

$$f \;\equiv\; f(x_i, y_i) = f(x_i, Y_i).$$

By the definition, given in Section 1.2, the local truncation error of the Modified Euler method is computed as follows:

$$\epsilon_{i+1}^{\text{ME}} = y_{i+1} - Y_{i+1}^{\text{ME}}, \qquad (1.29)$$

where y_{i+1} and Y_{i+1} are the exact and computed solutions at point x_{i+1}, respectively (assuming that $y_i = Y_i$). We first find y_{i+1} using ODE (1.1):

$$\begin{aligned} y_{i+1} = &\quad y(x_i + h) \\ = &\quad y_i + hy_i' + \frac{h^2}{2}y_i'' + \frac{h^3}{6}y_i''' + O(h^4)\,\Big|_{\text{use (1.7) and (1.28)}} \\ = &\quad y_i + hf + \frac{h^2}{2}(f_x + ff_y) + \frac{h^3}{6}\left(f_{xx} + f_x f_y + 2ff_{xy} + f(f_y)^2 + f^2 f_{yy}\right) + O(h^4)\,. \end{aligned} \qquad (1.30)$$

We now find $Y_{i+1}^{\rm ME}$ from Eq. (1.22):

$$\begin{aligned} Y_{i+1}^{\rm ME} &= \qquad Y_i + \frac{h}{2}\left(f + f(x_i+h, Y_i+hf) \right) \Big|_{\text{for last term, use (0.6) with } \Delta x = h \text{ and } \Delta y = hf} \\ &= Y_i + \frac{h}{2}\left(f + \left\{ f + [hf_x + hff_y] + \frac{1}{2!}[h^2 f_{xx} + 2\cdot h\cdot hf \cdot f_{xy} + (hf)^2 f_{yy}] + O(h^3) \right\} \right) \\ &= \qquad Y_i + hf + \frac{h^2}{2}(f_x + ff_y) + \frac{h^3}{4}(f_{xx} + 2ff_{xy} + f^2 f_{yy}) + O(h^4)\,. \end{aligned} \tag{1.31}$$

Finally, subtracting (1.31) from (1.30), one obtains:

$$\begin{aligned} \epsilon_{i+1}^{\rm ME} &= h^3 \left[\left(\frac{1}{6} - \frac{1}{4} \right) (f_{xx} + 2ff_{xy} + f^2 f_{yy}) + \frac{1}{6}(f_x + ff_y)f_y \right] + O(h^4) \\ &= h^3 \left[-\frac{1}{12}(f_{xx} + 2ff_{xy} + f^2 f_{yy}) + \frac{1}{6}(f_x + ff_y)f_y \right] + O(h^4)\,. \end{aligned} \tag{1.32}$$

For example, let $f(x,y) = ay$, where $a = \text{const}$. Then

$$f_x = f_{xx} = f_{xy} = 0, \quad f_y = a, \quad \text{and} \quad f_{yy} = 0\,,$$

so that from (1.32) the local truncation error of the Modified Euler method, applied to the ODE $y' = ay$, is found to be

$$\epsilon_{i+1}^{\rm ME} = \frac{h^3}{6}\, a^3 y \,+\, O(h^4)\,.$$

1.7 Questions for self-assessment

1. What does the notation $O(h^k)$ mean?
2. What are the meanings of the local truncation error, discretization error, and global error?
3. Give an example when the triangle inequality (1.5) holds with the '$<$' sign.
4. Be able to explain all steps made in the derivations in Eqs. (1.15) and (1.16).
5. Why are the Modified Euler and Midpoint methods called 2nd-order methods?
6. Obtain (1.27) from (1.25) and (1.26).
7. Explain why the properly programmed Modified Euler method requires exactly two evaluations of f per step.
8. Why may one prefer the Modified Euler method over the Romberg extrapolation based on the simple Euler method?