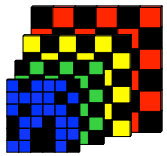
Numerical Integration

Professor Henry Arguello

Universidad Industrial de Santander
Colombia

March 26, 2014



High Dimensional Signal Processing Group
www.hdspgroup.com
henarfu@uis.edu.co

Outline: Chapter 9

Introduction to Differential Equations

Consider the equation

$$\frac{dy}{dt} = 1 - e^{-t}. \tag{1}$$

It is a differential equation because it involves the derivative dy/dt of the "unknown function" $y = y(t)$. Only the independent variable t appears on the right side of equation (1): hence a solution is an antiderivative of $1 - e^{-t}$. The rules of integration can be used to find $y(t)$:

$$y(t) = t + e^{-t} + C, \tag{2}$$

where C is the constant of integration. All the function in (2) are solutions of (1) because they satisfy the requirement that $y'(t) = 1 - e^{-t}$.They form the family of curves in Figure 9.2. Integration was the technique used to find the explicit formula for the functions in (2), and Figure 9.2 emphasize that there is one degree of freedom involved in the solution, that is the constant of integration C.

By varying the value of C, we "move the solution curve" up or down, and a particular curve can be found that will pass through any desired point.

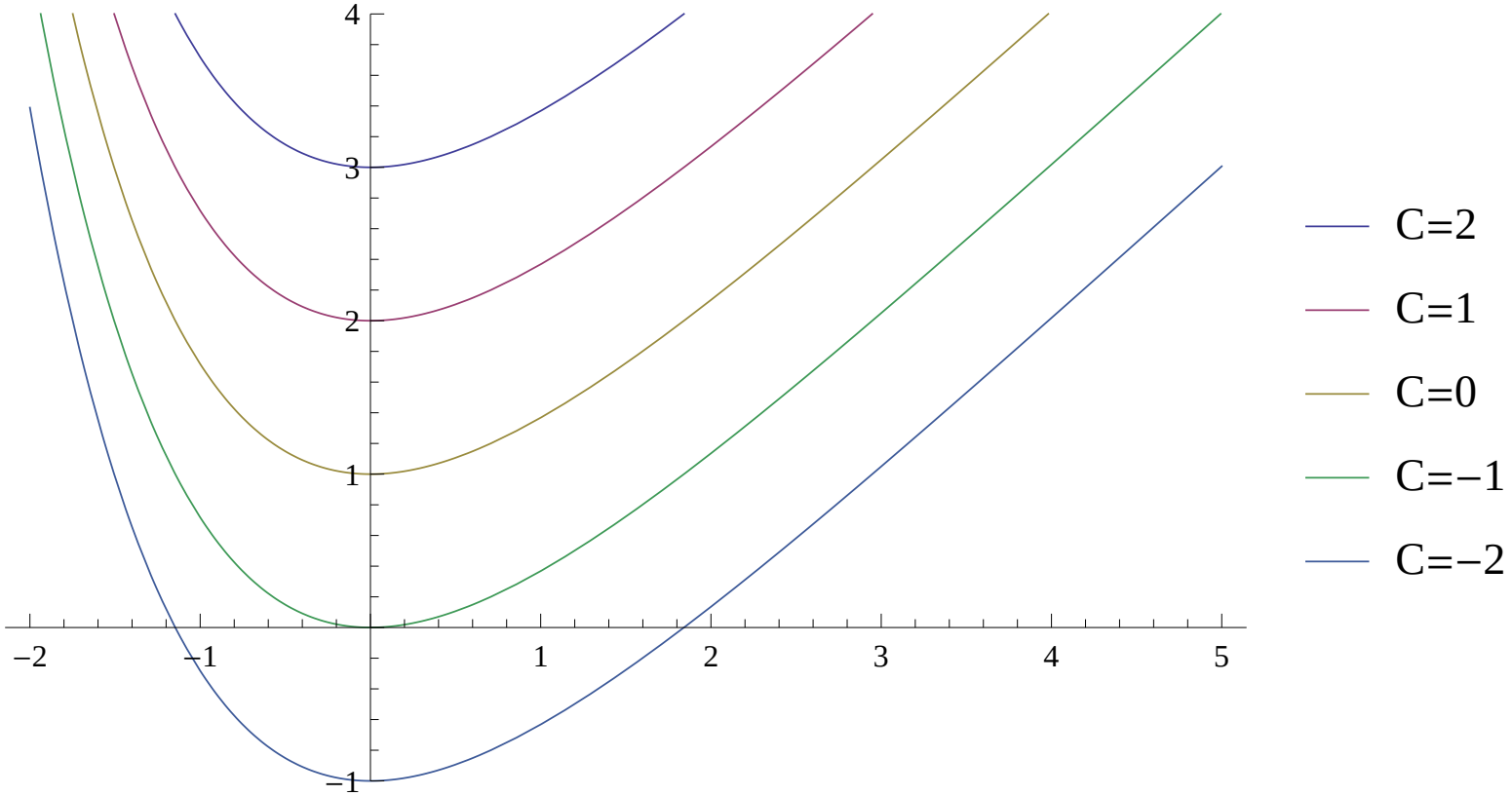


Figure 9.2: The solution curves $y(t) = t + e^{-t} + C$

Introduction to Differential Equations

we usually measure how a change in one variable affects another variable. When this is translated into a mathematical model, the result is an equation involving the rate of change of the unknown function and the independent and/or dependent variable.
Consider the temperature $y(t)$ of a cooling object. It might be conjectured that the rate of change of the temperature of the body is related to the temperature difference between its temperature and that of the surrounding medium. Experimental evidence verifies this conjecture. Newton's law of cooling asserts that the rate of change is directly proportional to the difference in these temperatures. If A is the temperature of the surrounding medium and $y(t)$ is the temperature of the body at time t, then

$$\frac{dy}{dt} = -k(y - A), \tag{3}$$

where k is a positive constant. The negative sign is required because dy/dt will be negative when the temperature of the body is greater than the temperature of the medium.

If the temperature of the object is known at time $t = 0$, we call this an initial condition and include this information in the statement of the problem. Usually, we are asked to solve

$$\frac{dy}{dt} = -k(y - A) \quad \text{with} \quad y(0) = y_0. \tag{4}$$

The technique of separation of variable can be used to find the solution

$$y = A + (y_0 - A)e^{-kt}. \tag{5}$$

For each choice of y_0, the solution curve will be different, and there is no simple way to move one curve around to get another one. The initial value is a point where the desired solution is "nailed down". Several solution curves are shown in Figure 9.3, and it can be observed that as t gets large the temperature of the object approaches room temperature. If $y_0 < A$, the body is warming instead of cooling.

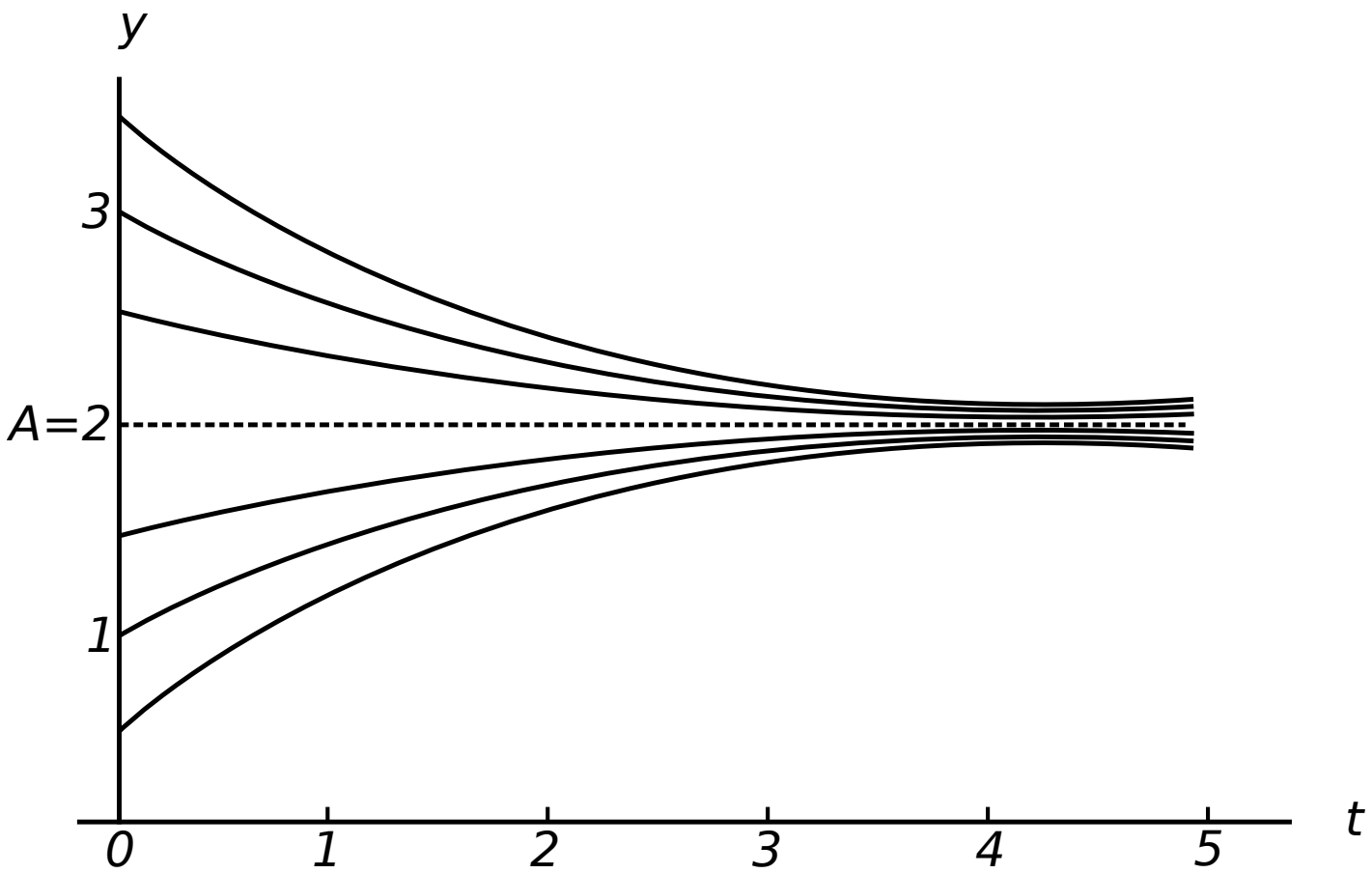


Figure 9.3: The solution curves $y = A + (y_0 - A)e^{-kt}$ for Newton's law of cooling (and warming)

Initial Value Problem

Definition 9.1. A Solution to the initial value problem (I.V.P)

$$y' = f(t,y) \quad \text{with} \quad y(t_0) = y_0 \tag{6}$$

on an interval $[t_0, b]$ is a differentiable function $y = y(t)$ such that

$$y(t_0) = y_0 \quad \text{and} \quad y'(t) = f(t, y(t)) \quad \text{for all } t \in [t_0, b]. \tag{7}$$

Notice that the solution curve $y = y(t)$ must pass through the initial point (t_0, y_0).

Geometric Interpretation

At each point (t, y) in the rectangular Region $R = \{(t, y) : a \leq t \leq b, c \leq y \leq d\}$, the slope of a solution curve $y = y(t)$ can be found using the implicit formula $m = f(t, y(t))$. Hence the values $m_{i,j} = f(t_i, y_j)$ can be computed throughout the rectangle, and each value $m_{i,j}$ represents the slope of the line tangent to a solution curve that passes through the point (t_i, y_j).

Introduction to Differential Equations

A slope field or direction field is a graph that indicates the slopes $\{m_{i,j}\}$ over the region. It can be used to visualize how a solution curve "fits" the slope constraint. To move along a solution curve, one must start at the initial point and check the slope field to determine in which direction to move. Then take a small step from t_0 to $t_0 + h$ horizontally and move the appropriate vertical distance $hf(t_0, y_0)$ so that the resulting displacement has the required slope. The next point on the solution curve is (t_1, y_1).Repeat the process to continue your journey along the curve. Since a finite number of steps will be used, the method will produce an approximation to the solution.

Example 9.1. The slope field for $y' = (t - y)/2$ over the rectangle $R = \{(t, y) : 0 \leq t \leq 5, 0 \leq y \leq 4\}$ is shown in Figure 9.4. The solution curves with the following initial values are shown:

1. For $y(0) = 1$, the solution is $y(t) = 3e^{-t/2} - 2 + t$.
2. For $y(0) = 4$, the solution is $y(t) = 6e^{-t/2} - 2 + t$.

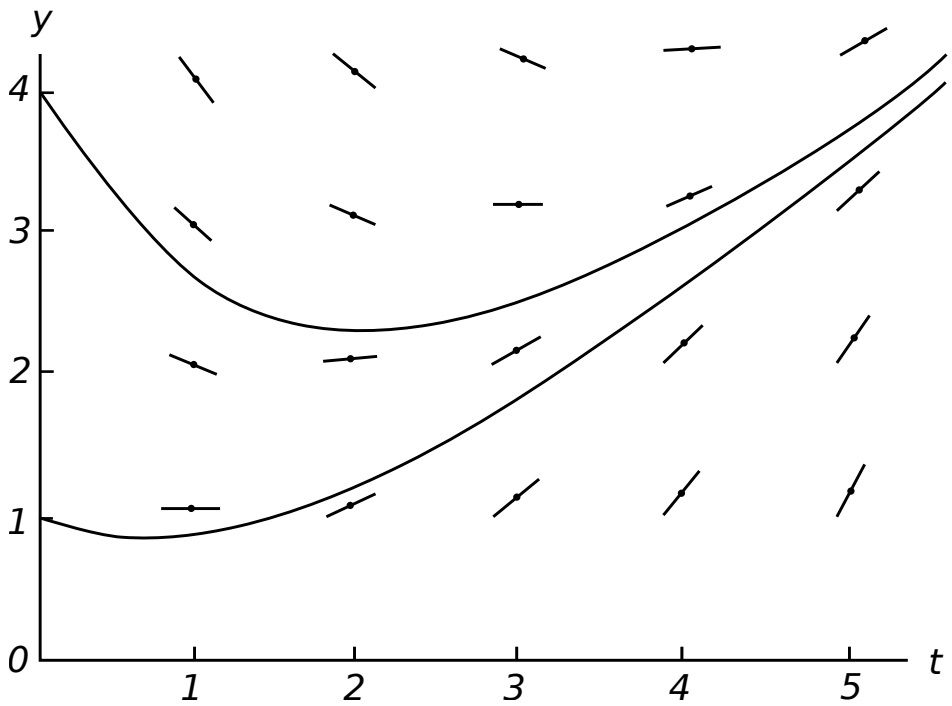


Figure 9.4: The slope field for the differential equation $y' = f(x, y) = (t - y)/2$.

Definition 9.2

Given the rectangle $R = \{(t, y) : a \leq t \leq b, c \leq y \leq d\}$, assume that $f(t, y)$ is continuous on R. The function f is said to satisfy a **Lipschitz condition** in the variable y on R provided that a constant $L > 0$ exists with the property that

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2| \tag{8}$$

whenever $(t, y_1), (t, y_2) \in R$. The constant L is called a **Lipschitz constant** for f.

Theorem 9.1

Suppose that $f(t,y)$ is defined on the region R. If there exists a constant $L > 0$ so that

$$|f_y(t,y)| \leq L \quad \text{for all } (t,y) \in R, \tag{9}$$

then f satisfies a Lipschitz condition in the variable y with Lipschitz constant L over the rectangle R.

Proof Fix t an use the mean value theorem to get c_1 with $y_1 < c_1 < y_2$ so that

$$\begin{aligned} |f(t,y_1) - f(t,y_2)| &= |f_y(t,c_1)(y_1 - y_2)| \\ &= |f_y(t,c_1)||y_1 - y_2| \leq L|y_1 - y_2|. \end{aligned}$$

Theorem 9.2 (Existence and Uniqueness).

Assume that $f(t,y)$ is continuous in a region $R = \{(t,y) : t_0 \leq t \leq b, c \leq y \leq d\}$. If f satisfies a Lipschitz condition on R in the variable y and $(t_0, y_0) \in R$, then the initial value problem (6), $y' = f(t,y)$ with $y(t_0) = y_0$, has a unique solution $y = y(t)$ on some subinterval $t_0 \leq t \leq t_0 + \delta$

Let us apply Theorems 9.1 and 9.2 to the function $f(t,y) = (t-y)/2$. The partial derivatives is $f_y(t,y) = -1/2$. Hence $|f_y(t,y)| \leq \frac{1}{2}$ and, according to the Theorem 9.1, the Lipschitz constant is $L = \frac{1}{2}$. Therefore, by Theorem 9.2 the **I.V.P.** Has a unique solution.

Euler's Method

The reader should be convinced that not all initial value problems can be solved explicitly, and often it is impossible to find a formula for the solution $y(t)$; for example, there is no "closed-form expression" for the solution to $y = t^3 + y^2$ with $y(0) = 0$. Hence for engineering and scientific purposes it is necessary to have methods for approximating the solution. If a solution with many significant digits is required, then more computing effort and a sophisticated algorithm must be used.

The first approach, called Euler's method, serves to illustrate the concepts involved in the advanced methods. It has limited use because of the larger error that is accumulated as the process proceeds. However, it is important to study because the error analysis is easier to understand.

Euler's Method

Let $[a, b]$ be the interval over which we want to find the solution to the well-posed I.V.P. $y' = f(t, y)$ with $y(a) = y_0$. In actuality, we will not find a differentiable function that satisfies the I.V.P. Instead, a set of points$\{(t_k, y_k)\}$ is generated, and the points are used for an approximation $(i.e., y(t_k) \approx y_k)$. How can we proceed to construct a "set of points" that will "satisfy a differential equation approximately"? First we choose the abscissas for the points. For convenience we subdivide the interval $[a, b]$ into M equal subintervals and select the mesh points

$$(1) \qquad t_k = a + kh \qquad \text{for} \quad k = 0, 1, ..., M \qquad \text{where} \quad h = \frac{b - a}{M}.$$

Euler's Method

The value h is called the ***step size***. We now proceed to solve approximately

$$(2) \qquad y' = f(t,y) \qquad \text{over} \quad [t_0, t_M] \quad \text{with} \quad y(t_0) = y_0.$$

Assume that $y(t)$, $y'(t)$, and $y''(t)$ are continuous and use Taylor's theorem to expand $y(t)$ about $t = t_0$. For each value t there exists a value c_1 that lies between t_0 and t so that

$$(3) \qquad y(t) = y(t_0) + y'(t_0)(t - t_0) + \frac{y''(c_1)(t - t_0)^2}{2}.$$

When $y'(t_0) = f(t_0, y(t_0))$ and $h = t_1 - t_0$ are substituted in equation (3), the result is an expression for $y(t_1)$:

$$(4) \qquad y(t_1) = y(t_0) + hf(t_0, y(t_0)) + y''(c1)\frac{h^2}{2}.$$

Euler's Method

If the step size h is chosen small enough, then we may neglect the second-order term (involving h^2) and get

$$(5) \qquad y_1 = y_0 + hf(t_0, y_0),$$

which is Euler's approximation. The process is repeated and generates a sequence of points that approximates the solution curve $y = y(t)$. The general step for Euler's method is

$$(6) \qquad t_{k+1} = t_k + h, \quad y_{k+1} = y_k + hf(t_k, y_k) \qquad \text{for} \quad k = 0, 1, ..., M-1.$$

Euler's Method

If the step size h is chosen small enough, then we may neglect the second-order term (involving h^2) and get

$$y_1 = y_0 + hf(t_0, y_0), \tag{5}$$

which is Euler's approximation. The process is repeated and generates a sequence of points that approximates the solution curve $y = y(t)$. The general step for Euler's method is

$$t_{k+1} = t_k + h, \quad y_{k+1} = y_k + hf(t_k, y_k) \quad \text{for} \quad k = 0, 1, ..., M-1. \tag{6}$$

Example 9.2. Use Euler's method to solve approximately the initial value problem

$$y' = Ry \quad \text{over} \quad [0, 1] \quad \text{with} \quad y(0) = y_0 \quad \text{and} \quad R \quad \text{constant.} \tag{7}$$

Euler's Method

The step size must be chosen, and then the second formula in (6) can be determined for computing the ordinates. This formula is sometimes called a *difference equation*, and in this case it is

$$(8) \qquad y_{k+1} = y_k(1 + hR) \quad \text{for} \quad k = 0, 1, ..., M - 1.$$

If we trace the solution values recursively, we see that

$$y_1 = y_0(1 + hR)$$

$$(9) \qquad y_2 = y_1(1 + hR) = y_0(1 + hR)^2$$

$$\vdots$$

$$y_M = y_{M-1}(1 + hR) = y_0(1 + hR)^M.$$

Euler's Method

Table 9.1 Compound Interest in Example 9.3

Step size, h	Number of iterations, M	Approximation to $y(5)$, y_M
1	5	$1000\left(1+\frac{0.1}{1}\right)^{5}=1610.51$
$\frac{1}{12}$	60	$1000\left(1+\frac{0.1}{12}\right)^{60}=1645.31$
$\frac{1}{360}$	1800	$1000\left(1+\frac{0.1}{360}\right)^{1800}=1648.61$

For most problems there is no explicit formula for determining the solution points, and each new point must be computed successively from the previous point. However, for the initial value problem (7) we are fortunate; Euler's method has the explicit solution

Euler's Method

$$(10) \qquad t_k = kh \qquad y_k = y_0(1+hR)^k \qquad \text{for} \quad k = 0, 1, ..., M.$$

Formula (10) can be viewed as the "compound interest" formula, and the Euler approximation gives the future value of a deposit.

Example 9.3. Suppose that \$1000 is deposited and earns 10% interest compounded continuously over 5 years. What is the value at the end of 5 years?

We choose to use Euler approximations with $h = 1, \frac{1}{12}$, and $\frac{1}{360}$ to approximate $y(5)$ for the I.V.P.:

$$y' = 0.1y \quad \text{over} \quad [0, 5] \quad \text{with} \quad y(0) = 1000.$$

Formula (10) with $R = 0.1$ produces Table 9.1.

Euler's Method

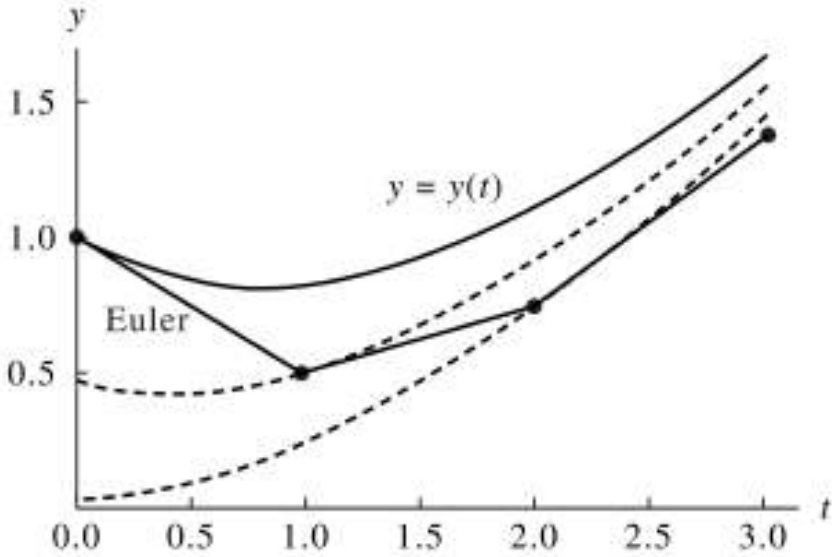
Think about the different values y_5, y_{60}, and $y1800$ that are used to determine the future value after 5 years. These values are obtained using different step sizes and reflect different amounts of computing effort to obtain an approximation to $y(5)$. The solution to the I.V.P. is $y(5) = 1000e^{0.5} = 1648.72$. If we did not use the closed-form solution (10), then it would have required 1800 iterations of Euler's method to obtain y_{1800}, and we still have only five digits of accuracy in the answer!

If bankers had to approximate the solution to the I.V.P. (7), they would choose Euler's method because of the explicit formula in (10). The more sophisticated methods for approximating solutions do not have an explicit formula for finding y_k , but they will require less computing effort.

Euler's Method

Figure 9.5 Euler's approximations

$$y_{k-1} = y_k + hf(t_k, y_k).$$



Geometric Description

If you start at the point (t_0, y_0) and compute the value of the slope $m_0 = f(t_0, y_0)$ and move horizontally the amount h and vertically $hf(t_0, y_0)$, then you are moving along the tangent line to $y(t)$ and will end up at the point (t_1, y_1) (see Figure 9.5). Notice that (t_1, y_1) is not on the desired solution curve! But this is the approximation that we are generating.

Hence we must use (t_1, y_1) as though it were correct and proceed by computing the slope $m_1 = f(t_1, y_1)$ and using it to obtain the next vertical displacement $hf(t_1, y_1)$ to locate (t_2, y_2), and so on.

Step Size versus Error

The methods we introduce for approximating the solution of an initial value problem are called ***difference methods or discrete variable methods***. The solution is approximated at a set of discrete points called a grid (or mesh) of points. An elementary single-step method has the form $y_{k+1} = y_k + h\Phi(t_k, y_k)$ for some function Φ called an ***increment function***.

When using any discrete variable method to solve an initial value problem approximately, there are two sources of error: discretization and round off.

Definition 9.3.
Assume that $\{(t_k, y_k)\}_{k=0}^{M}$ is the set of discrete approximations and that $y = y(t)$ is the unique solution to the initial value problem. The ***global discretization error*** e_k is defined by

$$(11) \qquad e_k = y(t_k) - y_k \qquad \text{for} \quad k = 0, 1, ..., M.$$

It is the difference between the unique solution and the solution obtained by the discrete variable method. The ***local discretization error*** ϵ_{k+1} is defined by

$$(12) \qquad \epsilon_{k+1} = y(t_{k+1}) - y_k - h\Phi(t_k, y_k) \qquad \text{for} \quad k = 0, 1, ..., M-1.$$

It is the error committed in the single step from t_k to t_{k+1}.

When we obtained equation (6) for Euler's method, the neglected term for each step was $y^{(2)}(c_k)(h^2/2)$. If this was the only error at each step, then at the end of the interval $[a, b]$, after M steps have been made, the accumulated error would be

Euler's Method

$$\sum_{k=1}^{M} y^{(2)}(c_k)\frac{h^2}{2} \approx My^{(2)}(c)\frac{h^2}{2} = \frac{hM}{2}y^{(2)}(c)h = \frac{(b-a)y^{(2)}(c)}{2}h = O(h^1).$$

There could be more error, but this estimate predominates. A detailed discussion on this topic can be found in advanced texts on numerical methods for differential equations.

Theorem 9.3 (Precision of Euler's Method).

Assume that $y(t)$ is the solution to the I.V.P. given in (2). If $y(t) \in C^2[t_0, b]$ and $\{(t_k, y_k)\}_{k=0}^{M}$ is the sequence of approximations generated by Euler's method, then

$$|e_k| = |y(t_k) - y_k| = O(h), \tag{13}$$

$$|\epsilon_{k+1}| = |y(t_k + 1) - y_k - hf(t_k, y_k)| = O(h^2).$$

Euler's Method

The error at the end of the interval is called the final global error (F.G.E.):

$$E(y(b), h) = |y(b) - y_M| = O(h). \tag{14}$$

Remark. The final global error $E(y(b), h)$ is used to study the behavior of the error for various step sizes. It can be used to give us an idea of how much computing effort must be done to obtain an accurate approximation.
Examples 9.4 and 9.5 illustrate the concepts in Theorem 9.3. If approximations are computed using the step sizes h and $h/2$, we should have

$$E(y(b), h) \approx Ch \tag{15}$$

Euler's Method

for the larger step size, and

$$E\left(y(b), \frac{h}{2}\right) \approx C\frac{h}{2} = \frac{1}{2}Ch \approx \frac{1}{2}E(y(b), h). \tag{16}$$

Hence the idea in Theorem 9.3 is that if the step size in Euler's method is reduced by a factor of $\frac{1}{2}$, we can expect that the overall F.G.E. will be reduced by a factor of $\frac{1}{2}$.

for the larger step size, and

$$E\left(y(b), \frac{h}{2}\right) \approx C\frac{h}{2} = \frac{1}{2}Ch \approx \frac{1}{2}E(y(b), h). \tag{16}$$

Hence the idea in Theorem 9.3 is that if the step size in Euler's method is reduced by a factor of $\frac{1}{2}$, we can expect that the overall F.G.E. will be reduced by a factor of $\frac{1}{2}$.

Example 9.4. Use Euler's method to solve the I.V.P.

$$y' = \frac{t-y}{2} \quad \text{on} \quad [0,3] \quad \text{with} \quad y(0) = 1.$$

Compare solutions for $h = 1, 1/2, 1/4, 1/8$.

Figure 9.6 Comparison of Euler solutions with different step sizes for $y' = (t - y)/2$ over $[0, 3]$ with the initial condition $y(0) = 1$.

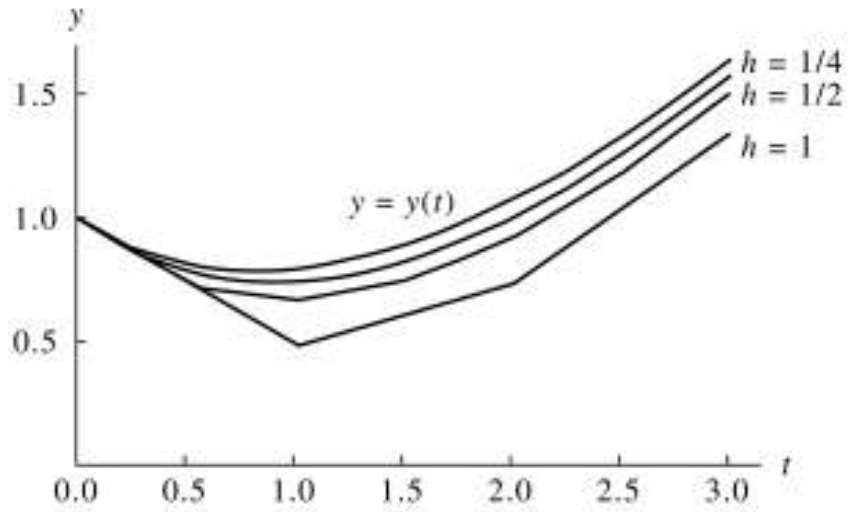


Figure 9.6 shows graphs of the four Euler solutions and the exact solution curve $y(t) = 3e^{-t/2} - 2 + t$. Table 9.2 gives the values for the four solutions at selected abscissas. For the step size h = 0.25, the calculations are

$$y_1 = 1.0 + 0.25\left(\frac{0.0 - 1.0}{2}\right) = 0.875,$$

Euler's Method

$$y_2 = 0.875 + 0.25\left(\frac{0.25 - 0.875}{2}\right) = 0.796875, \quad etc.$$

This iteration continues until we arrive at the last step:

$$y(3) \approx y_{12} = 1.440573 + 0.25\left(\frac{2.75 - 1.440573}{2}\right) = 1.604252.$$

Euler's Method

$$y_2 = 0.875 + 0.25\left(\frac{0.25 - 0.875}{2}\right) = 0.796875, \quad etc.$$

This iteration continues until we arrive at the last step:

$$y(3) \approx y_{12} = 1.440573 + 0.25\left(\frac{2.75 - 1.440573}{2}\right) = 1.604252.$$

Example 9.5. Compare the F.G.E. when Euler's method is used to solve the I.V.P.

$$y' = \frac{t - y}{2} \quad \text{over} \quad [0, 3] \quad \text{with} \quad y(0) = 1,$$

using step sizes $1, 1/2, ..., 1/64$. Table 9.3 gives the F.G.E. for several step sizes and shows that the error in the approximation to $y(3)$ decreases by about $1/2$ when the step size is reduced by a factor of $1/2$.

Euler's Method

For the smaller step sizes the conclusion of Theorem 9.3 is easy to see:

$$E(y(3), h) = y(3) - y_M = O(h^1) \approx Ch, \quad \text{where} \quad C = 0.256.$$

Table 9.2 Comparison of Euler Solutions with Different Step Sizes for $y' = (t - y)/2$ over $[0, 3]$ with $y(0) = 1$

	y_k				
t_k	$h = 1$	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$y(t_k)$ Exact
0	1.0	1.0	1.0	1.0	1.0
0.125				0.9375	0.943239
0.25			0.875	0.886719	0.897491
0.375				0.846924	0.862087
0.50		0.75	0.796875	0.817429	0.836402
0.75			0.759766	0.786802	0.811868
1.00	0.5	0.6875	0.758545	0.790158	0.819592
1.50		0.765625	0.846386	0.882855	0.917100
2.00	0.75	0.949219	1.030827	1.068222	1.103638
2.50		1.211914	1.289227	1.325176	1.359514
3.00	1.375	1.533936	1.604252	1.637429	1.669390

Table 9.3 Relation between Step Size and F.G.E. for Euler Solutions to $y' = (t - y)/2$ over $[0, 3]$ with $y(0) = 1$

Step size, h	Number of steps, M	Approximation to $y(3)$, y_M	F.G.E. Error at $t = 3$, $y(3) - y_M$	$O(h) \approx Ch$ where $C = 0.256$
1	3	1.375	0.294390	0.256
$\frac{1}{2}$	6	1.533936	0.135454	0.128
$\frac{1}{4}$	12	1.604252	0.065138	0.064
$\frac{1}{8}$	24	1.637429	0.031961	0.032
$\frac{1}{16}$	48	1.653557	0.015833	0.016
$\frac{1}{32}$	96	1.661510	0.007880	0.008
$\frac{1}{64}$	192	1.665459	0.003931	0.004

Heun's Method

The next approach, Heun's method, introduces a new idea for constructing an algorithm to solve the I.V.P.

$$y'(t) = f(t, y(t)) \quad \text{over} \quad [a, b] \quad \text{with} \quad y(t_0) = y_0. \tag{1}$$

To obtain the solution point (t_1, y_1), we can use the fundamental theorem of calculus and integrate $y'(t)$ over $[t_0, t_1]$ to get

$$\int_{t_0}^{t_1} f(t, y(t))\, dt = \int_{t_0}^{t_1} y'(t)\, dt = y(t_1) - y(t_0), \tag{2}$$

where the antiderivative of $y'(t)$ is the desired function $y(t)$. When the equation (2) is solved for $y(t_1)$, the result is

$$y(t_1) = y(t_0) + \int_{t_0}^{t_1} f(t, y(t))\, dt. \tag{3}$$

Heun's Method

Now a numerical integration method can be used to approximate the definite integral in (3). If the trapezoidal rule is used with the step size $h = t_1 - t_0$, then the result is

$$y(t_1) \approx y(t_0) + \frac{h}{2}(f(t_0, y(t_0)) + f(t_1, y(t_1))). \tag{4}$$

Notice that the formula on the right-hand side of (4) involves the yet to be determined value $y(t_1)$. To proceed, we use an estimate for $y(t_1)$. Euler's solution will suffice for this purpose. After it is substituted into (4), the resulting formula for finding (t_1, y_1) is called **Heun's method:**

$$y_1 = y(t_0) + \frac{h}{2}(f(t_0, y_0) + f(t_1, y_0 + hf(t_0, y_0))). \tag{5}$$

Heun's Method

The process is repeated and generates a sequence of points that approximates the solution curve $y = y(t)$. At each step, Euler's method is used as a prediction, and then the trapezoidal rule is used to make a correction to obtain the final value. The general step for Heun's method is

$$\begin{aligned} P_{k+1} &= y_k + hf(t_k, y_k), \quad t_{k+1} = t_k + h, \\ y_{k+1} &= y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, p_{k+1})) \end{aligned} \tag{6}$$

Notice the role played by differentiation and integration in Heun's method. Draw the line tangent to be the solution curve $y = y(t)$ at the point (t_0, y_0) and use it to find the predicted point (t_1, p_1). Now look at the graph $z = f(t, y(t))$ and consider the points (t_0, f_0) and (t_1, f_1), where $f_0 = f(t_0, y_0)$ and $f_1 = f(t_1, p_1)$. The area of the trapezoid with vertices (t_0, f_0) and (t_1, f_1) is an approximation to the integral in (3), which is used to obtain the final value in equation (5). The graphs are shown in Figure 9.7.

Heun's Method

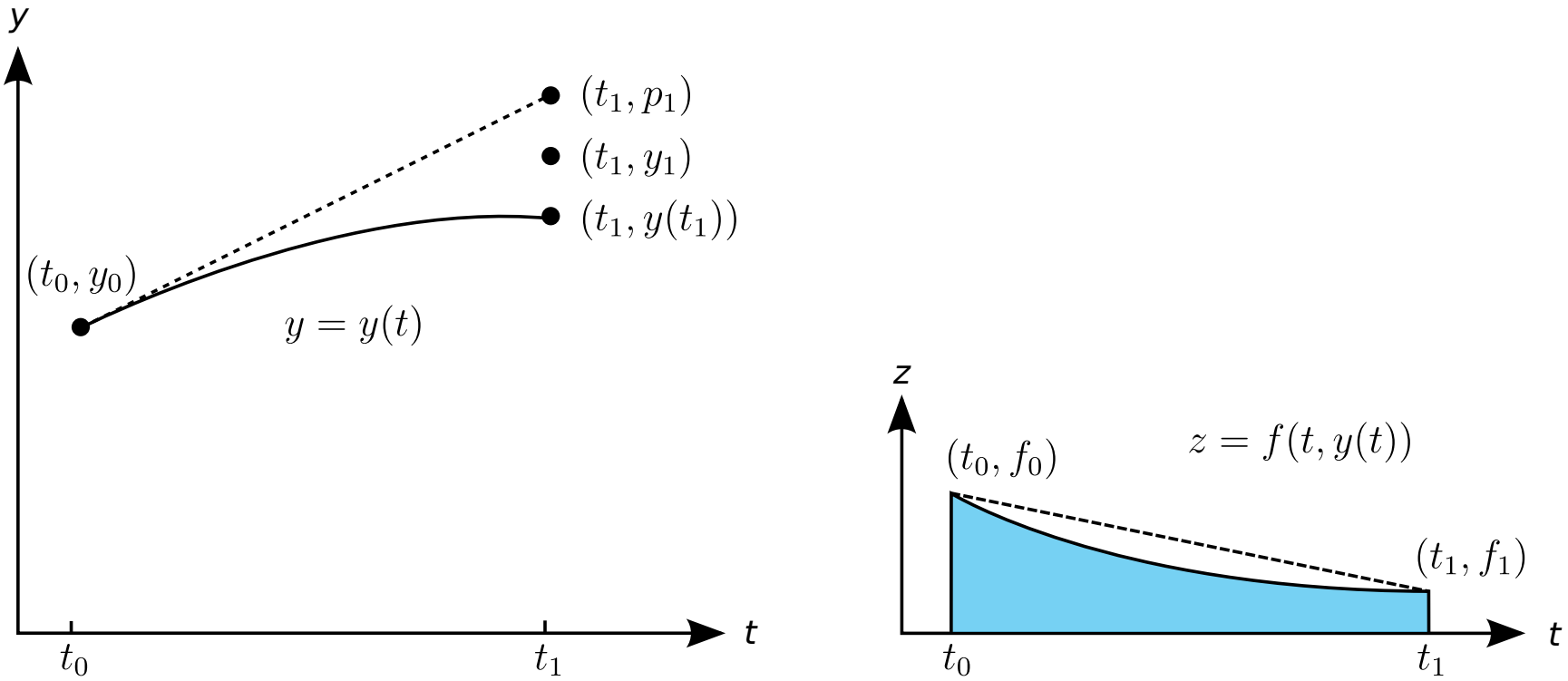


Figure 9.7: The graphs $y = y(t)$ and $z = f(t, y(t))$ in the derivation of Heun's method.

Step Size versus Error

The error term for the trapezoidal rule used to approximate the integral in (3) is

$$-y^{(2)}(c_k)\frac{h^3}{12}. \tag{7}$$

If the only error at each step is that given in (7), after M steps the accumulated error for Heun's method would be

$$-\sum_{k=1}^{M} y^{(2)}(c_k)\frac{h^3}{12} \approx \frac{b-a}{12}y^{(2)}(c)h^2 = \boldsymbol{O}(h^2). \tag{8}$$

The next theorem is important, because it states the relationship between F.G.E. and the step size. It is used to give us an idea of how much computing effort must be done to obtain an accurate approximation using Heun's method.

Heun's Method

Theorem 9.4 (Precision of Heun's Method).

Assume that $y(t)$ is the solution to the I.V.P. (1). If $y(t) \in C^3[t_0, b]$ and $\{(t_k, y_k)\}_0^M$ is the sequence of approximations generated by Heun's method, then

$$\begin{aligned} |e_k| = & \quad |y(t_k) - y_k| = \boldsymbol{O}(h^2), \\ |\epsilon_{k+1}| = & \quad |y(t_{k+1}) - y_k - h\Phi(t_k, y_k)| = \boldsymbol{O}(h^3), \end{aligned} \tag{9}$$

where $\Phi(t_k, y_k) = y_k + (h/2)(f(t_k, y_k) + f(t_{k+1}, y_k + hf(t_k, y_k)))$.
In particular, the final global error (F.G.E.) at the end of the interval will satisfy

$$E(y(b), h) = |y(b) - y_M| = \boldsymbol{O}(h^2). \tag{10}$$

Heun's method

Examples 9.6 and 9.7 illustrate Theorem 9.4. If approximations are computed using the step sizes h and $h/2$, we should have

$$E(y(b), h) \approx Ch^2 \tag{11}$$

for the larger step size, and

$$E\left(y(b), \frac{h}{2}\right) \approx C\frac{h^2}{4} = \frac{1}{4}Ch^2 \approx \frac{1}{4}E(y(b), h). \tag{12}$$

Hence the idea in Theorem 9.4 is that if the step size in Heun's method is reduced by a factor of $\frac{1}{2}$ we can expect that the overall F.G.E. will be reduced by a factor of $\frac{1}{4}$.

Heun's Method

Example 9.6. Use Heun's method to solve the I.V.P.

$$y' = \frac{t-y}{2} \quad \text{on} \quad [0,3] \quad \text{with} \quad y(0) = 1.$$

compare solution for $h = 1, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$.

Figure 9.8 shows the graphs of the first two Heun solutions and the exact solution curve $y(t) = 3e^{-t/2} - 2 + t$. Table 9.4 gives the values for the four solutions at selected abscissas. For the step size $h = 0.25$, a sample calculation is

$$\begin{aligned} f(t_0, y_0) &= \frac{0-1}{2} = -0.5 \\ p_1 &= 1.0 + 0.25(-0.5) = 0.875, \\ f(t_1, p_1) &= \frac{0.25 - 0.875}{2} = -0.3125, \\ y_1 &= 1.0 + 0.125(-0.5 - 0.3125) = 0.8984375. \end{aligned}$$

This iteration continues until we arrive at the last step:

$$y(3) \approx y_{12} = 1.511508 + 0.125(0.619246 + 0.666840) = 1.672269. \tag{13}$$

Heun's Method

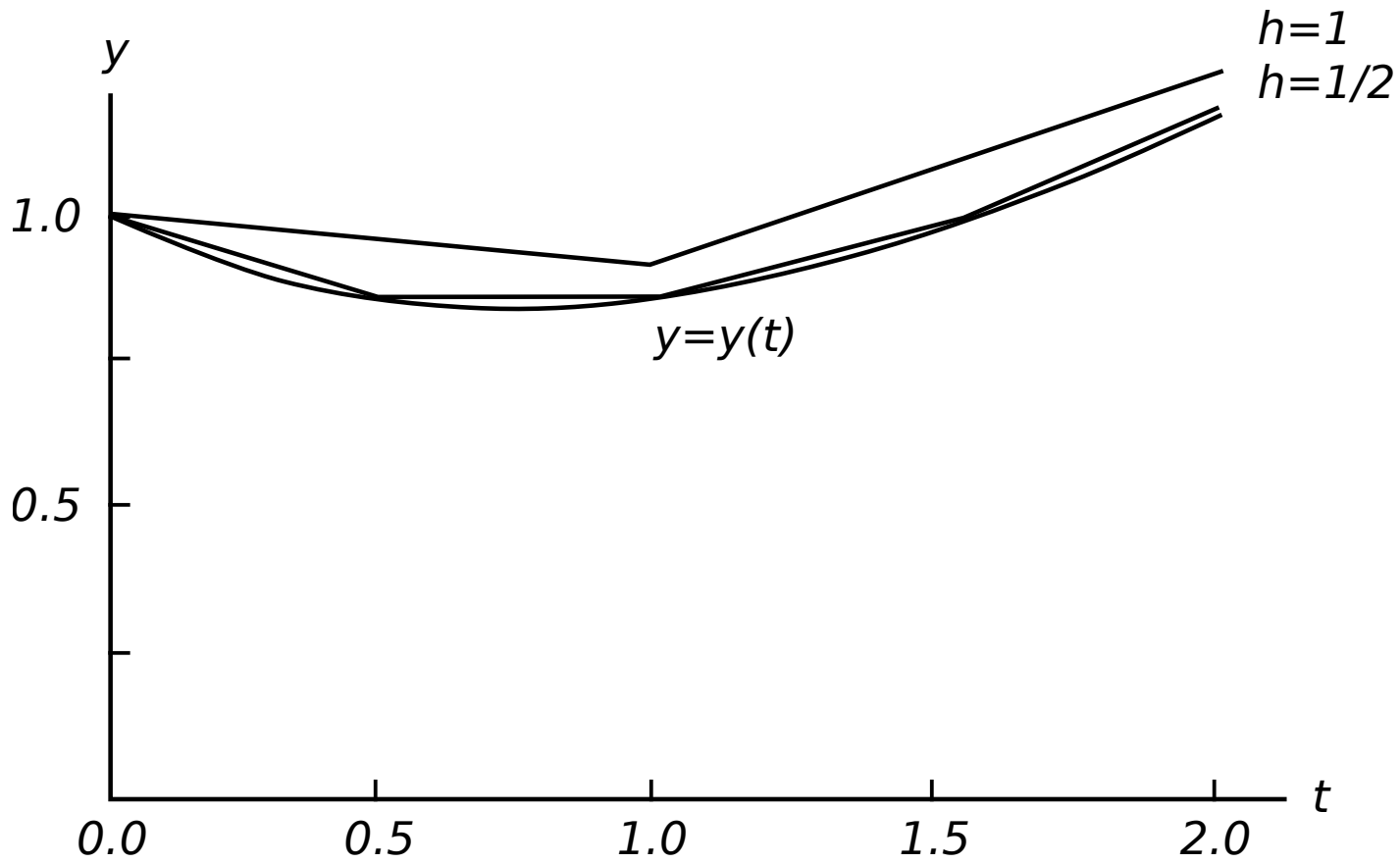


Figure 9.8: Comparison of Heun solutions with different step sizes for $y' = (t - y)/2$ over $[0, 2]$ with the initial condition $y(0) = 1$.

Heun's Method

Comparison of Heun Solutions with Different Step Sizes for $y' = (t - y)/2$ over $[0, 3]$ with $y(0) = 1$

	y_k				
t_k	$h = 1$	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$y(t_k)$ Exact
0	1.0	1.0	1.0	1.0	1.0
0.125				0.943359	0.943239
0.25			0.898438	0.897717	0.897491
0.375				0.862406	0.862087
0.50		0.84375	0.838074	0.836801	0.836402
0.75			0.814081	0.812395	0.811868
1.00	0.875	0.831055	0.822196	0.820213	0.819592
1.50		0.930511	0.920143	0.917825	0.917100
2.00	1.171875	1.117587	1.106800	1.104392	1.103638
2.50		1.373115	1.362593	1.360248	1.359514
3.00	1.732422	1.682121	1.672269	1.670076	1.669390

Heun's Method

Example 9.7 Compare the F.G.E. when Heun's method is used to solve

$$y' = \frac{t-y}{2} \quad \text{over} \quad [0,3] \quad \text{with} \quad y(0) = 1,$$

using step size $1, \frac{1}{2}, ..., \frac{1}{64}$.

Table 9.5 gives the F.G.E. and shows that the error in the approximation to $y(3)$ decreases by about $\frac{1}{4}$ when the step size is reduced by a factor of $\frac{1}{2}$:

$$E(y(3),h) = y(3) - y_M = \boldsymbol{O}(h^2) \approx Ch^2, where C = -0.0432.$$

Heun's Method

Table 9.5 Relation between Step Size and F.G.E. for Heun Solution to $y' = (t - y)/2$ over $[0, 3] with y(0) = 1$.

Step size, h	Number of steps, M	Approximation to $y(3)$, y_M	F.G.E Error at $t = 3$, $y(3) - y_M$	$O(h^2) \approx Ch^2$ where $C = -0.0432$
1	3	1.732422	−0.063032	−0.043200
$\frac{1}{2}$	6	1.682121	−0.012731	−0.010800
$\frac{1}{4}$	12	1.672269	−0.002879	−0.002700
$\frac{1}{8}$	24	1.670076	−0.000686	−0.000675
$\frac{1}{16}$	48	1.669558	−0.000168	−0.000169
$\frac{1}{32}$	96	1.669432	−0.000042	−0.000042
$\frac{1}{64}$	192	1.669401	−0.000011	−0.000011

Taylor Series Method

The Taylor series method is of general applicability, and it is the standard to which we compare the accuracy of the various other numerical methods for solving an I.V.P. It can be devised to have any specified degree of accuracy. We start by reformulating Taylor's theorem in a form that is suitable for solving differential equations.

Theorem 9.5 (Taylor's Theorem).

Assume that $y(t) \in C^{N+1}[t_0, b]$ and that $y(t)$ has a Taylor series expansion of order N about the fixed value $t = t_k \in [t_0, b]$:

$$(1) \qquad y(t_k + h) = y(t_k) + hT_N(t_k, y(t_k)) + O(h^{N+1})$$

,
where

$$(2) \qquad T_N(t_k, y(t_k)) = \sum_{j=1}^{N} \frac{y^{(j)}(t_k)}{j!} h^{j-1}$$

Taylor Series Method

and $y^{(j)}(t) = f^{(j-1)}(t, y(t))$ denotes the $(j-1)$st total derivative of the function f with respect to t. The formulas for the derivatives can be computed recursively:

$$
\begin{aligned}
&y'(t) = f \\
&y''(t) = f_t + f_y y' = f_t + f_y f \\
&y^{(3)}(t) = f_{tt} + 2f_{ty}y' + f_y y'' + f_{yy}(y')^2 = f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_y(f_t + f_y f) \\
&y^{(4)}(t) = f_{ttt} + 3f_{tty}y' + 3f_{tyy}(y')^2 + 3f_{ty}y'' + f_y y''' + 3f_{yy}y'y'' + f_{yyy}(y')^3 \\
&= (f_{ttt} + 3ftty f + 3f_{tyy}f^2 + f_{yyy}f^3) + f_y(f_{tt} + 2f_{ty}f + f_{yy}f^2) + 3(f_t + f_y f)(f_{ty} \\
&+ f_{yy}f) + f_y^2(f_t + f_y f)
\end{aligned} \tag{3}
$$

Taylor Series Method

and, in general,

$$(4) \qquad y^{(N)}(t) = P^{(N-1)}f(t, y(t)),$$

where P is the derivative operator

$$P = \left(\frac{\partial}{\partial t} + f\frac{\partial}{\partial y}\right).$$

The approximate numerical solution to the I.V.P. $y'(t) = f(t, y)$ over $[t_0, t_M]$ is derived by using formula (1) on each subinterval $[t_k, t_{k+1}]$. The general step for Taylor's method of order N is

$$(5) \qquad y_{k+1} = y_k + d_1 h + \frac{d_2 h^2}{2!} + \frac{d_3 h^3}{3!} + \cdots + \frac{d_N h^N}{N!},$$

where $d_j = y^{(j)}(t_k)$ for $j = 1, 2, ..., N$ at each step $k = 0, 1, ..., M - 1$.

Taylor Series Method

The Taylor method of order N has the property that the final global error (F.G.E.) is of the order $O(h^{N+1})$; hence N can be chosen as large as necessary to make this error as small as desired. If the order N is fixed, it is theoretically possible to a priori determine the step size h so that the F.G.E. will be as small as desired. However, in practice we usually compute two sets of approximations using step sizes h and $h/2$ and compare the results.

Theorem 9.6 (Precision of Taylor's Method of Order N).

Assume that $y(t)$ is the solution to the I.V.P.. If $y(t) \in C^{N+1}[t_0, b]$ and$(t_k, y_k)_{k=0}^{M}$ is the sequence of approximations generated by Taylor's method of order N, then

$$|e_k| = |y(t_k) - y_k| = O(h^N), \tag{6}$$

$$|\epsilon_{k+1}| = |y(t_{k+1}) - y_k - hT_N(t_k, y_k)| = O(h^{N+1}).$$

Taylor Series Method

In particular, the F.G.E. at the end of the interval will satisfy

$$E(y(b),h) = |y(b) - y_M| = O(h^N). \tag{7}$$

Examples 9.8 and 9.9 illustrate Theorem 9.6 for the case $N = 4$. If approximations are computed using the step sizes h and $h/2$, we should have

$$E(y(b),h) \approx Ch^4 \tag{8}$$

for the larger step size, and

$$E\left(y(b),\frac{h}{2}\right) \approx Ch^4 16 = \frac{1}{16}Ch^4 \approx \frac{1}{16}E(y(b),h). \tag{9}$$

Hence the idea in Theorem 9.6 is that if the step size in the Taylor method of order 4 is reduced by a factor of $\frac{1}{2}$, the overall F.G.E. will be reduced by about $\frac{1}{16}$.

Example 9.8. Use the Taylor method of order $N = 4$ to solve $y' = (t - y)/2$ on $[0, 3]$ with $y(0) = 1$. Compare solutions for $h = 1, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$.

The derivatives of $y(t)$ must first be determined. Recall that the solution $y(t)$ is a function of t, and differentiate the formula $y'(t) = f(t, y(t))$ with respect to t to get $y^{(2)}(t)$. Then continue the process to obtain the higher derivatives.

$$y'(t) = \frac{t-y}{2},$$

$$y^{(2)}(t) = \frac{d}{dt}\left(\frac{t-y}{2}\right) = \frac{1-y'}{2} = \frac{1-(t-y)/2}{2} = \frac{2-t+y}{4},$$

$$y^{(3)}(t) = \frac{d}{dt}\left(\frac{2-t+y}{4}\right) = \frac{0-1+y'}{4} = \frac{-1+(t-y)/2}{4} = \frac{-2+-y}{8},$$

$$y^{(4)}(t) = \frac{d}{dt}\left(\frac{-2+t-y}{8}\right) = \frac{-0+1-y'}{8} = \frac{1-(t-y)/2}{8} = \frac{2-t+y}{16}.$$

Taylor Series Method

To find y_1, the derivatives given above must be evaluated at the point $(t_0, y_0) = (0, 1)$. Calculation reveals that

$$d_1 = y'(0) = \frac{0.0-1.0}{2} = -0.5,$$

$$d_2 = y^{(2)}(0) = \frac{2.0-0.0+1.0}{4} = 0.75,$$

$$d_3 = y^{(3)}(0) = \frac{-2.0+0.0-1.0}{8} = -0.375,$$

$$d_4 = y^{(4)}(0) = \frac{2.0-0.0+1.0}{16} = 0.1875.$$

Next the derivatives d_j are substituted into (5) with $h = 0.25$, and nested multiplication is used to compute the value y_1:

$$y_1 = 1.0 + 0.25\left(-0.5 + 0.25\left(\frac{0.75}{2} + 0.25\left(\frac{-0.375}{6} + 0.25\left(\frac{0.1875}{24} \right) \right) \right) \right) = 0.8974915.$$

Taylor Series Method

The computed solution point is $(t_1, y_1) = (0.25, 0.8974915)$.

To determine y_2, the derivatives d_j must now be evaluated at the point $(t_1, y_1) = (0.25, 0.8974915)$. The calculations are starting to require a considerable amount of computational effort and are tedious to do by hand. Calculation reveals that

$$d_1 = y'(0.25) = \frac{0.25 - 0.8974915}{2} = -0.3237458,$$

$$d_2 = y^{(2)}(0.25) = \frac{2.0 - 0.25 + 0.8974915}{4} = 0.6618729,$$

$$d_3 = y^{(3)}(0.25) = \frac{-2.0 + 0.25 - 0.8974915}{8} = -0.3309364,$$

$$d_4 = y^{(4)}(0.25) = \frac{2.0 - 0.25 + 0.8974915}{16} = 0.1654682.$$

Taylor Series Method

Now these derivatives d_j are substituted into (5) with $h = 0.25$, and nested multiplication is used to compute the value y_2:

$$y_2 = 0.8974915 + 0.25\left(-0.3237458 + 0.25\left(\frac{0.6618729}{2} + 0.25\left(\frac{-0.3309364}{6} + 0.25\left(\frac{0.1654682}{24}\right)\right)\right)\right) = 0.8364037.$$

The solution point is $(t_2, y_2) = (0.50, 0.8364037)$. Table 9.6 gives solution values at selected abscissas using various step sizes.

Table 9.6 Comparison of the Taylor Solutions of Order $N = 4$ for $y' = (t - y)/2$ over $[0, 3]$ with $y(0) = 1$

	y_k				
t_k	$h = 1$	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$y(t_k)$ Exact
0	1.0	1.0	1.0	1.0	1.0
0.125				0.9432392	0.9432392
0.25			0.8974915	0.8974908	0.8974917
0.375				0.8620874	0.8620874
0.50		0.8364258	0.8364037	0.8364024	0.8364023
0.75			0.8118696	0.8118679	0.8118678
1.00	0.8203125	0.8196285	0.8195940	0.8195921	0.8195920
1.50		0.9171423	0.9171021	0.9170998	0.9170997
2.00	1.1045125	1.1036826	1.1036408	1.1036385	1.1036383
2.50		1.3595575	1.3595168	1.3595145	1.3595144
3.00	1.6701860	1.6694308	1.6693928	1.6693906	1.6693905

Example 9.9. Compare the F.G.E. for the Taylor solutions to $y' = (t - y)/2$ over $[0, 3]$ with $y(0) = 1$ given in Example 9.8.
Table 9.7 gives the F.G.E. for these step sizes and shows that the error in the approximation $y(3)$ decreases by about $\frac{1}{16}$ when the step size is reduced by a factor of $\frac{1}{2}$:

$$E(y(3), h) = y(3) - y_M = O(h^4) \approx Ch^4, \quad \text{where} \quad C = -0.000614.$$

Table 9.7 Relation between Step Size and F.G.E. for the Taylor Solutions to$y = (t - y)/2$ over $[0, 3]$

Step size, h	Number of steps, M	Approximation to $y(3)$, y_M	F.G.E. Error at $t = 3$, $y(3) - y_M$	$O(h^2) \approx Ch^4$ where $C = -0.000614$
1	3	1.6701860	−0.0007955	−0.0006140
$\frac{1}{2}$	6	1.6694308	−0.0000403	−0.0000384
$\frac{1}{4}$	12	1.6693928	−0.0000023	−0.0000024
$\frac{1}{8}$	24	1.6693906	−0.0000001	−0.0000001

Runge-Kutta Methods

The Taylor methods in the preceding section have the desirable feature that the F.G.E. is of order $\boldsymbol{O}(h^N)$, and N can be chosen large so that this error is small. However, the shortcomings of the Taylor methods are the a priori determination of N and the computation of the higher derivatives, which can be very complicated. Each Runge-Kutta method is derived from an appropriate Taylor method in such a way that the F.G.E. is of order $\boldsymbol{O}(h^N)$. A trade-off is made to perform several function evaluations at each step and eliminate the necessity to compute the higher derivatives. These methods can be constructed for any order N . The Runge-Kutta method of order $N = 4$ is most popular. It is a good choice for common purposes because it is quite accurate, stable, and easy to program. Most authorities proclaim that it is not necessary to go to a higher-order method because the increased accuracy is offset by additional computational effort. If more accuracy is required, then either a smaller step size or an adaptive method should be used.

Runge-Kutta Methods

The fourth-order Runge-Kutta method (**RK4**) simulates the accuracy of the Taylor series method of order $N = 4$. The method is based on computing y_{k+1} as follows:

$$y_{k+1} = y_k + w_1 k_1 + w_2 k_2 + w_3 k_3 + w_4 k_4, \tag{1}$$

where k_1, k_2 , k_3, and k_4 have the form

$$\begin{aligned} k1 &= hf(t_k, y_k), \\ k2 &= hf(t_k + a_1 h, y_k + b_1 k_1), \\ k3 &= hf(t_k + a_2 h, y_k + b_2 k_1 + b_3 k_2), \\ k4 &= hf(t_k + a_3 h, y_k + b_4 k_1 + b_5 k_2 + b_6 k_3). \end{aligned} \tag{2}$$

By matching coefficients with those of the Taylor series method of order $N = 4$ so that the local truncation error is of order $\boldsymbol{O}(h^5)$, Runge and Kutta were able to obtain the following system of equations:

Runge-Kutta Methods

$$
\begin{aligned}
b_1 &= a_1, \\
b_2 + b_3 &= a_2, \\
b_4 + b_5 + b_6 &= a_3, \\
w_1 + w_2 + w_3 + w_4 &= 1, \\
w_2 a_1 + w_3 a_2 + w_4 a_3 &= \frac{1}{2}, \\
w_2 a_1^2 + w_3 a_2^2 + w_4 a_3^2 &= \frac{1}{3}, \\
w_2 a_1^3 + w_3 a_2^3 + w_4 a_3^3 &= \frac{1}{4}, \\
w_3 a_1 b_3 + w_4 (a_1 b_5 + a_2 b_6) &= \frac{1}{6}, \\
w_3 a_1 a_2 b_3 + w_4 a_3 (a_1 b_5 + a_2 b_6) &= \frac{1}{8}, \\
w_3 a_1^2 b3 + w_4 (a_1^2 b_5 + a_2^2 b_6) &= \frac{1}{12}, \\
w_4 a_1 b_3 b_6 &= \frac{1}{24}
\end{aligned}
\tag{3}
$$

Runge-Kutta Methods

The system involves 11 equations in 13 unknowns. Two additional conditions must be supplied to solve the system. The most useful choice is

$$a_1 = \frac{1}{2} \quad \text{and} \quad b_2 = 0. \tag{4}$$

Then the solution for the remaining variables is

$$\begin{aligned} &a_2 = \frac{1}{2}, \quad a_3 = 1, \quad b_1 = \frac{1}{2}, \quad b_3 = \frac{1}{2}, \quad b_4 = 0, \quad b_5 = 0, \quad b_6 = 1 \\ &w_1 = \frac{1}{6}, \quad w_2 = \frac{1}{3}, \quad w_3 = \frac{1}{3}, \quad w_4 = \frac{1}{6} \end{aligned} \tag{5}$$

Runge-Kutta Methods

The values in (4) and (5) are substituted into (2) and (1) to obtain the formula for the standard Runge-Kutta method of order $N = 4$, which is stated as follows. Start with the initial point (t_0, y_0) and generate the sequence of approximations using

$$y_{k+1} = y_k + \frac{h(f_1 + 2f_2 + 2f_3 + f_4)}{6}, \tag{6}$$

where

$$\begin{aligned} f_1 &= f(t_k, y_k), \\ f_2 &= f\left(t_k + \frac{h}{2}, y_k + \frac{h}{2} f_1\right), \\ f_3 &= f\left(t_k + \frac{h}{2}, y_k + \frac{h}{2} f_2\right), \\ f_4 &= f(t_k + h, y_k + h f_3). \end{aligned} \tag{7}$$

Discussion about the Method

The complete development of the equations in (7) is beyond the scope of this book and can be found in advanced texts, but we can get some insights. Consider the graph of the solution curve $y = y(t)$ over the first subinterval $[t_0, t_1]$. The function values in (7) are approximations for slopes to this curve. Here f_1 is the slope at the left, f_2 and f_3 are two estimates for the slope in the middle, and f_4 is the slope at the right (see Figure 9.9(a)). The next point (t_1, y_1) is obtained by integrating the slope function

$$y(t_1) - y(t_0) = \int_{t_0}^{t_1} f(t, y(t))\, dt. \tag{8}$$

Runge-Kutta Methods

If Simpson's rule is applied with step size h/2, the approximation to the integral in (8) is

$$\int_{t_0}^{t_1} f(t, y(t))\, dt \approx \frac{h}{6}(f(t_0, y(t_0)) + 4f(t_{1/2}, y(t_{1/2})) + f(t_1 + y(t_1))), \tag{9}$$

where $t_{1/2}$ is the midpoint of the interval.

Runge-Kutta Methods

Three function values are needed; hence we make the obvious choice $f(t_0, y(t_0)) = f_1$ and $f(t_1, y(t_1)) \approx f_4$. For the value in the middle we chose the average of f_2 and f_3 :

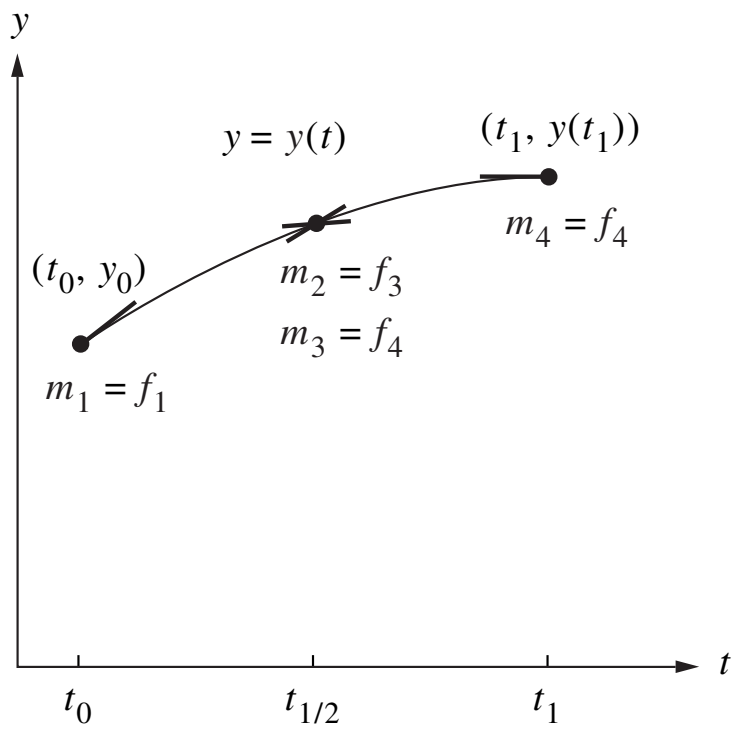
$$f(t_{1/2}, y(t_{1/2})) \approx \frac{f_2 + f_3}{2}.$$

These values are substituted into (9), which is used in equation (8) to get y_1:

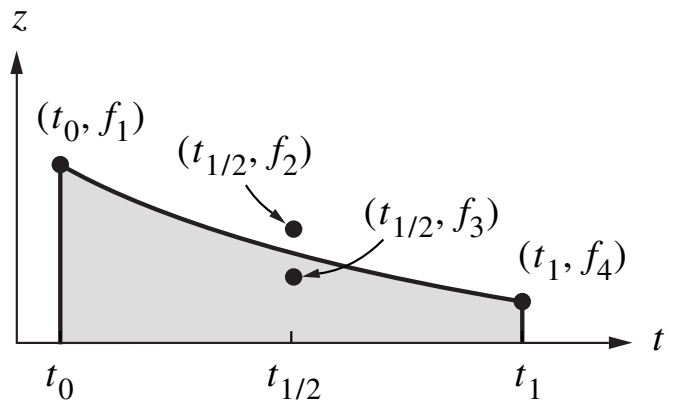
$$y_1 = y_0 + \frac{h}{6}\left(f_1 + \frac{4(f_2 + f_3)}{2} + f_4\right). \tag{10}$$

When this formula is simplified, it is seen to be equation (6) with $k = 0$. The graph for the integral in (9) is shown in Figure 9.9(b).

Runge-Kutta Methods



(a) Predicted slopes m_j to the solution curve $y = y(t)$



(b) Integral approximation:

$$y(t_1) - y_0 = \frac{h}{6}(f_1 + 2f_2 + 2f_3 + f_4)$$

Figure 9.9: The graphs $y = y(t)$ and $z = f(t, y(t))$ in the discussion of the Runge-Kutta method of order $N = 4$.

Step Size versus Error

The error term for Simpson's rule with step size $h/2$ is

$$-y^{(4)}(c_1)\frac{h^5}{2880}. \tag{11}$$

If the only error at each step is that given in (11), after M steps the accumulated error for the RK4 method would be

$$-\sum_{k=1}^{M} y^{(4)}(c_k)\frac{h^5}{2880} \approx \frac{b-a}{5760}y^{(4)}(c)h^4 \approx \boldsymbol{O}(h^4). \tag{12}$$

The next theorem states the relationship between F.G.E. and step size. It is used to give us an idea of how much computing effort must be done when using the RK4 method.

Theorem 9.7 (Precision of the Runge-Kutta Method)

Assume that y(t) is the solution to the I.V.P. If $y(t) \in C^5[t_0, b]$ and $\{(t_k, y_k)\}_{k=0}^{M}$ is the sequence of approximations generated by the Runge-Kutta method of order 4, then

$$\begin{aligned} |e_k| &= |y(t_k) - y_k| = \boldsymbol{O}(h^4), \\ |\epsilon_{k+1}| &= |y(t_{k+1}) - y_k - h\boldsymbol{T}_N(t_k, y_k)| = \boldsymbol{O}(h^5). \end{aligned} \tag{13}$$

In particular, the F.G.E. at the end of the interval will satisfy

$$E(y(b), h) = |y(b) - y_M| = \boldsymbol{O}(h^4). \tag{14}$$

Runge-Kutta Methods

Examples 9.10 and 9.11 illustrate Theorem 9.7. If approximations are computed using the step sizes h and h/2, we should have

$$E(y(b), h) \approx Ch^4 \tag{15}$$

for the larger step size, and

$$E\left(y(b), \frac{h}{2}\right) \approx C\frac{h^4}{16} = \frac{1}{16}Ch^4 \approx \frac{1}{16}E(y(b), h). \tag{16}$$

Hence the idea in Theorem 9.7 is that if the step size in the RK4 method is reduced by a factor of $\frac{1}{2}$ we can expect that the overall F.G.E. will be reduced by a factor of $\frac{1}{16}$.

Example 9.10. Use the RK4 method to solve the I.V.P. $y' = (t - y)/2$ on $[0, 3]$ with $y(0) = 1$. Compare solutions for $h = 1, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$.

Example 9.10. Use the RK4 method to solve the I.V.P. $y' = (t - y)/2$ on $[0, 3]$ with $y(0) = 1$. Compare solutions for $h = 1, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$.

Table 9.8 gives the solution values at selected abscissas. For the step size $h = 0.25$, a sample calculation is

$$
\begin{aligned}
f_1 &= \frac{0.0 - 1.0}{2} = -0.5, \\
f_2 &= \frac{0.125 - (1 + 0.25(0.5)(-0.5))}{2} = -0.40625, \\
f_3 &= \frac{0.125 - (1 + 0.25(0.5)(-0.40625))}{2} = -0.4121094, \\
f_4 &= \frac{0.25 - (1 + 0.25(-0.4121094))}{2} = -0.3234863, \\
y_1 &= 1.0 + 0.25\left(\frac{-0.5 + 2(-0.40625) + 2(-0.4121094) - 0.3234863}{6}\right) \\
&= 0.8974915.
\end{aligned}
$$

Table 9.8 Comparison of the RK4 Solutions with Different Step Sizes for $y = (t - y)/2$ over $[0, 3]$ with $y(0) = 1$

	y_k				
t_k	$h = 1$	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$y(t_k)$ Exact
0	1.0	1.0	1.0	1.0	1.0
0.125				0.9432392	0.9432392
0.25			0.8974915	0.8974908	0.8974917
0.375				0.8620874	0.8620874
0.50		0.8364258	0.8364037	0.8364024	0.8364023
0.75			0.8118696	0.8118679	0.8118678
1.00	0.8203125	0.8196285	0.8195940	0.8195921	0.8195920
1.50		0.9171423	0.9171021	0.9170998	0.9170997
2.00	1.1045125	1.1036826	1.1036408	1.1036385	1.1036383
2.50		1.3595575	1.3595168	1.3595145	1.3595144
3.00	1.6701860	1.6694308	1.6693928	1.6693906	1.6693905

Example 9.11. Compare the F.G.E. when the RK4 method is used to solve $y' = (t - y)/2$ over $[0, 3]$ with $y(0) = 1$ using step sizes $1, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$.

Example 9.11. Compare the F.G.E. when the RK4 method is used to solve $y' = (t - y)/2$ over $[0, 3]$ with $y(0) = 1$ using step sizes $1, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$.

Table 9.9 gives the F.G.E. for the various step sizes and shows that the error in the approximation to $y(3)$ decreases by about $\frac{1}{16}$ when the step size is reduced by a factor of $1/2$.

$$E(y(3), h) = y(3) - y_M = \boldsymbol{O}(h^4) \approx Ch^4 \quad \text{where} \quad C = -0.000614.$$

Table 9.9 Relation between Step Size and F.G.E. for the RK4 Solutions to $y = (t - y)/2$ over $[0, 3]$ with $y(0) = 1$

Step size, h	Number of steps, M	Approximation to $y(3)$, y_M	F.G.E. Error at $t = 3$, $y(3) - y_M$	$O(h^4) \approx Ch^4$ where $C = -0.000614$
1	3	1.6701860	−0.0007955	−0.0006140
$\frac{1}{2}$	6	1.6694308	−0.0000403	−0.0000384
$\frac{1}{4}$	12	1.6693928	−0.0000023	−0.0000024
$\frac{1}{8}$	24	1.6693906	−0.0000001	−0.0000001

Runge-Kutta Methods

A comparison of Examples 9.10 and 9.11 and Examples 9.8 and 9.9 shows what is meant by the statement “The RK4 method simulates the Taylor series method of order N = 4.” For these examples, the two methods generate identical solution sets $\{(tk, yk)\}$ over the given interval. The advantage of the RK4 method is obvious; no formulas for the higher derivatives need to be computed nor do they have to be in the program. It is not easy to determine the accuracy to which a Runge-Kutta solution has been computed. We could estimate the size of $y^{(4)}(c)$ and use formula (12). Another way is to repeat the algorithm using a smaller step size and compare results. A third way is to adaptively determine the step size, which is done in Program 9.5. In Section 9.6 we will see how to change the step size for a multistep method.

Runge-Kutta Methods of Order $N = 2$

The second-order Runge-Kutta method (denoted RK2) simulates the accuracy of the Taylor series method of order 2. Although this method is not as good to use as the RK4 method, its proof is easier to understand and illustrates the principles involved. To start, we write down the Taylor series formula for $y(t+h)$:

$$y(t+h) = y(t) + hy'(t) + \frac{1}{2}h^2 y''(t) + C_T h^3 + \cdots, \tag{17}$$

where C_T is a constant involving the third derivative of $y(t)$ and the other terms in the series involve powers of h^j for $j > 3$.

Runge-Kutta Methods

The derivatives $y'(t)$ and $y''(t)$ in equation (17) must be expressed in terms of $f(t,y)$ and its partial derivatives. Recall that

$$y'(t) = f(t,y). \tag{18}$$

The chain rule for differentiating a function of two variables can be used to differentiate (18) with respect to t, and the result is

$$y''(t) = f_t(t,y) + f_y(t,y)y'(t).$$

Using (18), this can be written

$$y''(t) = f_t(t,y) + f_y(t,y)f(t,y). \tag{19}$$

The derivatives (18) and (19) are substituted in (17) to give the Taylor expression for $y(t+h)$:

$$\begin{aligned} y(t+h) = y(t) + hf(t,y) + \frac{1}{2}h^2 f_t(t,y) \\ + \frac{1}{2}h^2 f_y(t,y)f(t,y) + C_T h^3 + \cdots . \end{aligned} \tag{20}$$

Runge-Kutta Methods

Now consider the Runge-Kutta method of order $N = 2$, which uses a linear combination of two function values to express $y(t+h)$:

$$y(t+h) = y(t) + Ahf_0 + Bhf_1, \tag{21}$$

where

$$\begin{aligned} f_0 &= f(t,y), \\ f_1 &= f(t+Ph, y+Qhf_0). \end{aligned} \tag{22}$$

Next the Taylor polynomial approximation for a function of two independent variables is used to expand $f(t,y)$ (see the Exercises). This gives the following representation for f_1:

$$f_1 = f(t,y) + Phf_t(t,y) + Qhf_y(t,y)f(t,y) + C_P h^2 + \cdots, \tag{23}$$

where C_P involves the second-order partial derivatives of $f(t,y)$. Then (23) is used in (21) to get the RK2 expression for $y(t+h)$:

$$\begin{aligned} y(t+h) = y(t) &+ (A+B)hf(t,y) + BPh^2 f_t(t,y) \\ &+ BQh^2 f_y(t,y)f(t,y) + BC_P h^3 + \cdots. \end{aligned} \tag{24}$$

Runge-Kutta Methods

A comparison of similar terms in equations (20) and (24) will produce the following conclusions:

$$hf(t,y) = (A+B)hf(t,y) \quad \text{implies that } 1 = A+B,$$
$$\frac{1}{2}h^2 f_t(t,y) = BPh^2 f_t(t,y) \quad \text{implies that } \frac{1}{2} = BP,$$
$$\frac{1}{2}h^2 f_y(t,y)f(t,y) = BQh^2 f_y(t,y)f(t,y) \quad \text{implies that } \frac{1}{2} = BQ.$$

Hence, if we require that A, B, P, and Q satisfy the relations

$$A+B=1 \quad BP = \frac{1}{2} \quad BQ = \frac{1}{2}, \tag{25}$$

then the RK2 method in (24) will have the same order of accuracy as the Taylor's method in (20).

Runge-Kutta Methods

Since there are only three equations in four unknowns, the system of equations (25) is underdetermined, and we are permitted to choose one of the coefficients. There are several special choices that have been studied in the literature; we mention two of them.

Case(i): Choose $A = \frac{1}{2}$. This choice leads to $B = \frac{1}{2}$, $P = 1$, and $Q = 1$. If equation (21) is written with these parameters, the formula is

$$y(t+h) = y(t) + \frac{h}{2}(f(t,y) + f(t+h, y+hf(t,y))). \tag{26}$$

When this scheme is used to generate $\{(t_k, y_k)\}$, the result is Heun's method.

Case(ii): Choose $A = 0$. This choice leads to $B = 1$, $P = \frac{1}{2}$, and $Q = \frac{1}{2}$. If equation (21) is written with these parameters, the formula is

$$y(t+h) = y(t) + hf\left(t + \frac{h}{2}, y + \frac{h}{2}f(t,y)\right). \tag{27}$$

When this scheme is used to generate $\{(t_k, y_k)\}$, it is called the **modified Euler-Cauchy method.**

Runge-Kutta-Fehlberg Method (RKF45)

One way to guarantee accuracy in the solution of an I.V.P. is to solve the problem twice using step sizes h and $h/2$ and compare answer at the mesh points corresponding to the larger step size. But this require a significant amount of computation for the smaller step size and must be repeated if it is determined that the agreement is not good enough.

The Runge-Kutta-Fehlberg method (denoted RKF45) is one way to try to solve this problem. It has a procedure to determine if the proper step size h is being used. At each step, two different approximations for the solution are made and compared. If the two answers are in close agreement, the approximation is accepted. If the two answers do not agree to a specified accuracy, the step size is reduced. If the answers agree to more significant digits than required, the step size is increased.

Runge-Kutta Methods

Each step requires the use of the following six values:

$$
\begin{aligned}
k_1 &= hf(t_k, y_k), \\
k_2 &= hf\left(t_k + \frac{1}{4}h, y_k + \frac{1}{4}k_1\right), \\
k_3 &= hf\left(t_k + \frac{3}{8}h, y_k + \frac{3}{32}k_1 + \frac{9}{32}k_2\right), \\
k_4 &= hf\left(t_k + \frac{12}{13}h, y_k + \frac{1932}{2197}k_1 - \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3\right), \\
k_5 &= hf\left(t_k + h, y_k + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{512}k_3 - \frac{845}{4104}k_4\right), \\
k_6 &= hf\left(t_k + \frac{1}{2}h, y_k - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2565}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5\right).
\end{aligned}
\tag{28}
$$

Runge-Kutta Methods

Then an approximation to the solution of the I.V.P. is made using a Runge-Kutta method of order 4:

$$y_{k+1} = y_k + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4101}k_4 - \frac{1}{5}k_5, \tag{29}$$

where the four function values $f_1, f_3, f_4,$ and f_5 are used. Notice that f_2 is not used in formula (29). A better value for the solution is determined using a Runge-Kutta method of order 5:

$$z_{k+1} = y_k + \frac{16}{135}k_1 + \frac{6656}{12,825}k_3 + \frac{28,561}{56,430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6. \tag{30}$$

The optimal step size sh can be determined by multiplying the scalar s times the current step size h. The scalar s is

$$s = \left(\frac{tol\, h}{2|z_{k+1} - y_{k+1}|}\right)^{1/4} \approx 0.84\left(\frac{tol\, h}{|z_{k+1} - y_{k+1}}\right)^{1/4} \tag{31}$$

where tol is the specified error control tolerance.

Runge-Kutta Methods

The derivation of formula (31) can be found in advanced books on numerical analysis. It is important to learn that a fixed step size is not the best strategy even though it would give a nicer-appearing table of values. If values are needed that are not in the table, polynomial interpolation should be used.

Example 9.12. Compare RKF45 and RK4 solutions to the I.V.P

$$y' = 1 + y^2 \quad \text{with} \quad y(0) = 0 \quad \text{on} \quad [0, 1.4].$$

Example 9.12. Compare RKF45 and RK4 solutions to the I.V.P

$$y' = 1 + y^2 \quad \text{with} \quad y(0) = 0 \quad \text{on} \quad [0, 1.4].$$

An RKF45 program was used with the value $tol = 2 \times 10^{-5}$ for the error control tolerance. It changed the step size automatically and generated the 10 approximations to the solution in Table 9.10. And RK4 program was used with the a priori step size of $h = 0.1$, which required the computer to generate 14 approximations at the equally spaced points in Table 9.11. The approximations at the right endpoint are

$$y(1.4) \approx y_{10} = 5.7985045 \quad \text{and} \quad y(1.4) \approx y_{14} = 5.7919748$$

and the errors are

$$E_{10} = -0.0006208 \quad \textit{and} \quad E_{14} = 0.0059089$$

for the RKF45 and RK4 methods, respectively. The RKF45 method has the smaller error.

Table 9.10 RKF45 Solution to $y' = 1 + y^2, y(0) = 0$

k	t_k	RK45 approximation y_k	True solution $y(t_k) = tan(t_k)$	Error $y(t_k) - y_k$
0	0.0	0.0000000	0.0000000	0.0000000
1	0.2	0.2027100	0.2027100	0.0000000
2	0.4	0.4227933	0.4227931	−0.0000002
3	0.6	0.6841376	0.6841368	−0.0000008
4	0.8	1.0296434	1.0296386	−0.0000048
5	1.0	1.5574398	1.5774077	−0.0000321
6	1.1	1.9648085	1.9647597	−0.0000488
7	1.2	2.5722408	2.5721516	−0.0000892
8	1.3	3.6023295	3.6021024	−0.0002271
9	1.35	4.4555714	4.4552218	−0.0003496
10	1.4	5.7985045	5.7978837	−0.0006208

Runge-Kutta Methods

Table 9.11 RK4 Solution to $y' = 1 + y^2, y(0) = 0$

k	t_k	RK4 approximation y_k	True solution $y(t_k) = tan(t_k)$	Error $y(t_k) - y_k$
0	0.0	0.0000000	0.0000000	0.0000000
1	0.1	0.1003346	0.1003347	0.0000001
2	0.2	0.2027099	0.2027100	0.0000001
3	0.3	0.3093360	0.3093362	0.0000002
4	0.4	0.4227930	0.4227932	0.0000002
5	0.5	0.5463023	0.5463025	0.0000002
6	0.6	0.6841368	0.6841368	0.0000000
7	0.7	0.8422886	0.8422884	−0.0000002
8	0.8	1.0296391	1.0296386	−0.0000005
9	0.9	1.2601588	1.2601582	−0.0000006
10	1.0	1.5574064	1.5574077	0.0000013
11	1.1	1.9647466	1.9647597	0.0000131
12	1.2	2.5720718	2.5721516	0.0000798
13	1.3	3.6015634	3.6021024	0.0005390
14	1.4	5.7919748	5.7978837	0.0059089

Predictor-Corrector Methods

The methods of Euler, Heun, Taylor, and Runge-Kutta are called single-step methods because they use only the information from one previous point to compute the successive point; that is, only the initial point (t_0, y_0) is used to compute (t_1, y_1), and in general, y_k is needed to compute y_{k+1}. After several points have been found, it is feasible to use several prior points in the calculation. For illustration, we develop the Adams-Bashforth four-step method, which requires y_{k-3}, y_{k-2}, y_{k-1}, and y_k in the calculation of $y_k + 1$. This method is not self-starting; four initial points (t_0, y_0), (t_1, y_1), (t_2, y_2), and (t_3, y_3), must be given in advance in order to generate the points $\{(tk, yk) : k \geq 4\}$ (this can be done with one of the methods from the previous sections).

Predictor-Corrector Methods

A desirable feature of a multistep method is that the local truncation error (L.T.E.) can be determined and a correction term can be included, which improves the accuracy of the answer at each step. Also, it is possible to determine if the step size is small enough to obtain an accurate value for y_{k+1}, yet large enough so that unnecessary and time-consuming calculations are eliminated. Using the combinations of a predictor and corrector requires only two function evaluations of $f(t, y)$ per step.

Adams-Bashforth-Moulton Method

The Adams-Bashforth-Moulton predictor-corrector method is a multi-step method derived from the fundamental theorem of calculus:

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))dt. \tag{1}$$

The predictor uses the Lagrange polynomial approximation for $f(t, y(t))$ based on the points (t_{k-3}, f_{k-3}), (t_{k-2}, f_{k-2}), (t_{k-1}, f_{k-1}), and (t_k, f_k). It is integrated over the interval $[t_k, t_{k+1}]$ in (1). This process produces the Adams-Bashforth predictor:

$$p_{k+1} = y_k + \frac{h}{24}(-9f_{k-3} + 37f_{k-2} - 59f_{k-1} + 55f_k). \tag{2}$$

Adams-Bashforth-Moulton Method

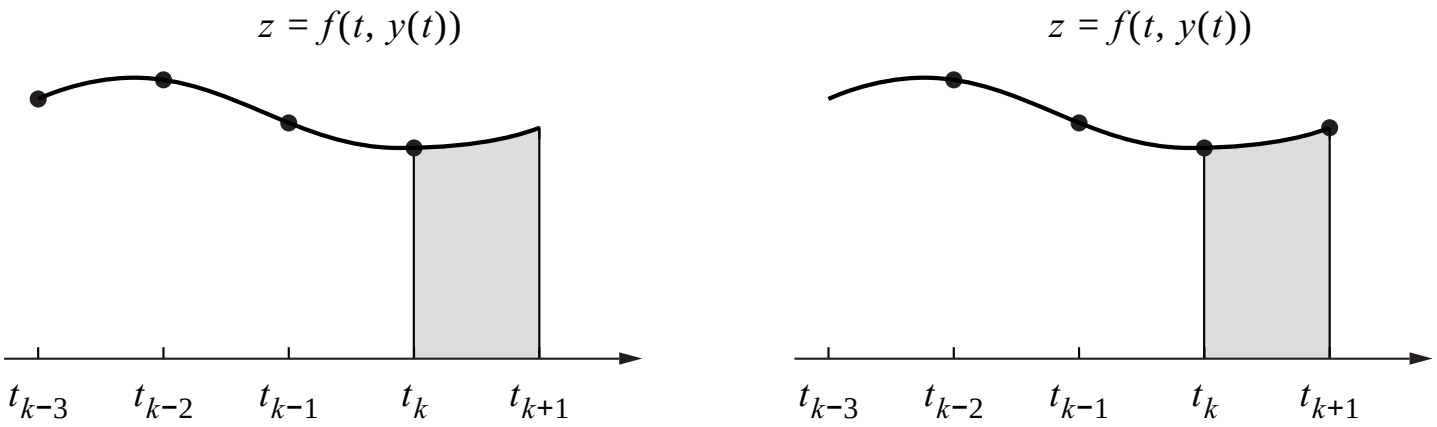


Figure 9.10 Integration over $[t_k, t_{k-1}]$ in the Adams-Bashforth method.
(a) The four nodes for the Adams-Bashforth predictor (extrapolation is used),
(b) The four nodes for the Adams-Moulton corrector (interpolation is used)

The corrector is developed similarly. The value p_{k+1} just computed can now be used. A second Lagrange polynomial for $f(t, y(t))$ is constructed, which is based on the points (t_{k-2}, f_{k-2}), (t_{k-1}, f_{k-1}), (t_k, f_k), and the new point $(t_{k+1}, f_{k+1}) = (t_{k+1}, f(t_{k+1}, p_{k+1}))$. This polynomial is then integrated over $[t_k, t_{k+1}]$, producing the Adams-Moulton corrector:

Adams-Bashforth-Moulton Method

$$y_{k+1} = y_k + \frac{h}{24}(f_{k-2} - 5f_{k-1} + 19f_k + 9f_{k+1}). \tag{3}$$

Figure 9.10 shows the nodes for the Lagrange polynomials that are used in developing formulas (2) and (3), respectively.

Error Estimation and Correction

The error terms for the numerical integration formulas used to obtain both the predictor and corrector are of the order $O(h^5)$. The L.T.E. for formulas (2) and (3) are

$$y(t_{k+1}) - p_{k+1} = \frac{251}{720}y^{(5)}(c_{k+1})h^5 \qquad \text{(L.T.E. for the predictor)}, \tag{4}$$

$$y(t_{k+1}) - y_{k+1} = \frac{-19}{720}y^{(5)}(d_{k+1})h^5 \qquad \text{(L.T.E. for the corrector)}. \tag{5}$$

Adams-Bashforth-Moulton Method

Suppose that h is small and $y^{(5)}(t)$ is nearly constant over the interval; then the terms involving the fifth derivative in (4) and (5) can be eliminated, and the result is

$$(6) \qquad y(t_{k+1}) - y_{k+1} \approx \frac{-19}{270}(y_{k+1} - p_{k+1}).$$

The importance of the predictor-corrector method should now be evident. Formula (6) gives an approximate error estimate based on the two computed values p_{k+1} and y_{k+1} and does not use $y^{(5)}(t)$.

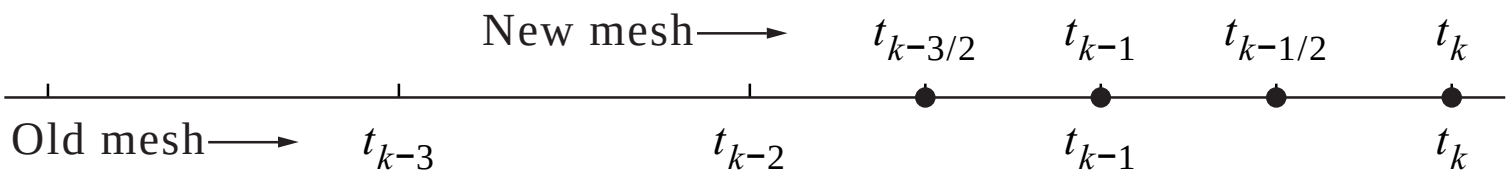


Figure 9.11 Reduction of the step size to $h/2$ in an adaptive method.

Practical Considerations

The corrector (3) used the approximation $f_{k+1} \approx f(t_{k+1}, p_{k+1})$ in the calculation of y_{k+1}. Since y_{k+1} is also an estimate for $y(t_{k+1})$, it could be used in the corrector (3) to generate a new approximation for f_{k+1}, which in turn will generate a new value for y_{k+1}. However, when this iteration on the corrector is continued, it will converge to a fixed point of (3) rather than the differential equation. It is more efficient to reduce the step size if more accuracy is needed.

Formula (6) can be used to determine when to change the step size. Although elaborate methods are available, we show how to reduce the step size to $h/2$ or increase it to $2h$. Let $RelErr = 5 * 10^{-6}$ be our relative error criterion, and let $Small = 10^{-5}$.

Adams-Bashforth-Moulton Method

$$(7) \qquad \text{If} \quad \frac{19}{270}\frac{|y_{k+1}-p_{k+1}|}{|y_{k+1}|+Small} > RelErr, \quad \text{then set} \quad h=\frac{h}{2}.$$

$$(8) \qquad \text{If} \quad \frac{19}{270}\frac{|y_{k+1}-p_{k+1}|}{|y_{k+1}|+Small} < \frac{RelErr}{100}, \quad \text{then set} \quad h=2h.$$

When the predicted and corrected values do not agree to five significant digits, then (7) reduces the step size. If they agree to seven or more significant digits, then (8) increases the step size. Fine-tuning of these parameters should be made to suit your particular computer.
Reducing the step size requires four new starting values. Interpolation of $f(t,y(t))$ with a fourth-degree polynomial is used to supply the missing values that bisect the intervals $[t_{k-2},t_{k-1}]$ and $[t_{k-1},t_k]$. The four mesh points $t_{k-3/2}$, t_{k-1}, $t_{k-1/2}$, and t_k used in the successive calculations are shown in Figure 9.11.

Adams-Bashforth-Moulton Method

The interpolation formulas needed to obtain the new starting values for the step size $h/2$ are

$$f_{k-1/2} = \frac{-5f_{k-4} + 28f_{k-3} - 70f_{k-2} + 140f_{k-1} + 35f_k}{128},$$

(9)

$$f_{k-3/2} = \frac{3f_{k-4} - 20f_{k-3} + 90f_{k-2} + 60f_{k-1} - 5f_k}{128}.$$

Increasing the step size is an easier task. Seven prior points are needed to double the step size. The four new points are obtained by omitting every second one, as shown in Figure 9.12.

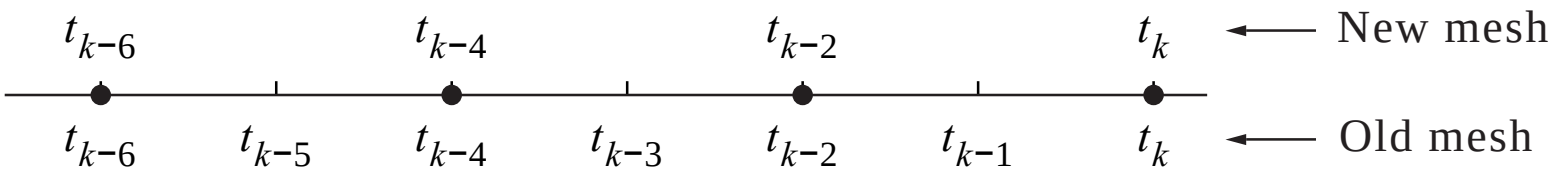


Figure 9.12 Increasing the step size to $2h$ in an adaptive method.

Milne-Simpson Method

Another popular predictor-corrector scheme is known as the Milne-Simpso method. Its predictor is based on integration of $f(t, y(t))$ over the interval $[t_{k-3}, t_{k+1}]$:

$$y(t_{k+1}) = y(t_{k-3}) + \int_{t_{k-3}}^{t_{k+1}} f(t, y(t))dt. \tag{10}$$

The predictor uses the Lagrange polynomial approximation for $f(t, y(t))$ based on the points (t_{k-3}, f_{k-3}), (t_{k-2}, f_{k-2}), (t_{k-1}, f_{k-1}), and (t_k, f_k). It is integrated over the interval $[t_{k-3}, t_{k+1}]$. This produces the Milne predictor:

$$p_{k+1} = y_{k-3} + \frac{4h}{3}(2f_{k-2} - f_{k-1} + 2f_k). \tag{11}$$

Milne-Simpson Method

The corrector is developed similarly. The value p_{k+1} can now be used. A second Lagrange polynomial for $f(t,y(t))$ is constructed, which is based on the points (t_{k-1},f_{k-1}), (t_k,f_k), and the new point $(t_{k+1},f_{k+1}) = (t_{k+1},f(t_{k+1},p_{k+1}))$. The polynomial is integrated over $[t_{k-1},t_{k+1}]$, and the result is the familiar Simpson's rule:

$$y_{k+1} = y_{k-1} + \frac{h}{3}(f_{k-1} + 4f_k + f_{k+1}). \tag{12}$$

Error Estimation and Correction

The error terms for the numerical integration formulas used to obtain both the predictor and corrector are of the order $O(h^5)$. The L.T.E. for the formulas in (11) and (12) are

$$y(t_{k+1}) - p_{k+1} = \frac{28}{90}y^{(5)}(c_{k+1})h^5 \qquad \text{(L.T.E. for the predictor)}, \tag{13}$$

Milne-Simpson Method

$$y(t_{k+1}) - y_{k+1} = \frac{-1}{90} y^{(5)}(d_{k+1}) h^5 \qquad \text{(L.T.E. for the corrector).} \tag{14}$$

Suppose that h is small enough so that $y^{(5)}(t)$ is nearly constant over the interval $[t_{k-3}, t_{k+1}]$. Then the terms involving the fifth derivative can be eliminated in (13) and (14) and the result is

$$y(t_{k+1}) - p_{k+1} \approx \frac{28}{29}(y_{k+1} - p_{k+1}). \tag{15}$$

Formula (15) gives an error estimate for the predictor that is based on the two computed values p_{k+1} and y_{k+1} and does not use $y^{(5)}(t)$. It can be used to improve the predicted value. Under the assumption that the difference between the predicted and corrected values at each step changes slowly, we can substitute p_k and y_k for p_{k+1} and y_{k+1} in (15) and get the following modifier:

Milne-Simpson Method

$$m_{k+1} = p_{k+1} + 28\frac{y_k - p_k}{29}. \tag{16}$$

This modified value is used in place of p_{k+1} in the correction step, and equation (12) becomes

$$y_{k+1} = y_{k-1} + \frac{h}{3}(f_{k-1} + 4f_k + f(t_{k+1}, m_{k+1})). \tag{17}$$

Therefore, the improved (modified) Milne-Simpson method is

$$p_{k+1} = y_{k-3} + \frac{4h}{3}(2f_{k-2} - f_{k-1} + 2f_k) \qquad \text{(predictor)}$$

$$m_{k+1} = p_{k+1} + 28\frac{y_k - p_k}{29} \qquad \text{(modifier)}$$

(18)

$$f_{k+1} = f(t_{k+1}, m_{k+1})$$

$$y_{k+1} = y_{k-1} + \frac{h}{3}(f_{k-1} + 4f_k + f_{k+1}) \qquad \text{(corrector).}$$

Milne-Simpson Method

Hamming's method is another important method. We shall omit its derivation, but furnish a program at the end of the section. As a final precaution we mention that all the predictor-corrector methods have stability problems. Stability is an advanced topic and the serious reader should research this subject.

Example 9.13. Use the Adams-Bashforth-Moulton, Milne-Simpson, and Hamming methods with $h = 1/8$ and compute approximations for the solution of the I.V.P.

$$y' = \frac{t - y}{2}, \quad y(0) = 1 \quad \text{over} \quad [0, 3].$$

A Runge-Kutta method was used to obtain the starting values $y_1 = 0.94323919$, $y_2 = 0.89749071$, and $y_3 = 0.86208736$.

Milne-Simpson Method

Then a computer implementation of Programs 9.6 through 9.8 produced the values in Table 9.12. The error for each entry in the table is given as a multiple of 10^{-8}. In all entries there are at least six digits of accuracy. In this example, the best answers were produced by Hamming's method.

Table 9.12 Comparison of the Adams-Bashforth-Moulton, Milne-Simpson, and Hamming Methods for Solving $y' = (t - y)/2$, $y(0) = 1$

k	Adams-Bashforth-Moulton	Error	Milne-Simpson	Error	Hamming's method	Error
0.0	1.00000000	$0E-8$	1.00000000	$0E-8$	1.00000000	$0E-8$
0.5	0.83640227	$8E-8$	0.83640231	$4E-8$	0.83640234	$1E-8$
0.625	0.81984673	$16E-8$	0.81984687	$2E-8$	0.81984688	$1E-8$
0.75	0.81186762	$22E-8$	0.81186778	$6E-8$	0.81186783	$1E-8$
0.875	0.81194530	$28E-8$	0.81194555	$3E-8$	0.81194558	$0E-8$
1.0	0.81959166	$32E-8$	0.81959190	$8E-8$	0.81959198	$0E-8$
1.5	0.91709920	$46E-8$	0.91709957	$9E-8$	0.91709967	$-1E-8$
2.0	1.10363781	$51E-8$	1.10363822	$10E-8$	1.10363834	$-2E-8$
2.5	1.35951387	$52E-8$	1.35951429	$10E-8$	1.35951441	$-2E-8$
2.625	1.43243853	$52E-8$	1.43243899	$6E-8$	1.43243907	$-2E-8$
2.75	1.50851827	$52E-8$	1.50851869	$10E-8$	1.50851881	$-2E-8$
2.875	1.58756195	$51E-8$	1.58756240	$6E-8$	1.58756248	$-2E-8$
3.0	1.66938998	$50E-8$	1.66939038	$10E-8$	1.66939050	$-2E-8$

The Right Step

Our selection of methods has a purpose first, their development is easy enough for a first course; second, more advanced methods have a similar development; third, most undergraduate problems can be solved by one of these methods. However, when a predictor-Corrector method is used to solve the I.V.P. y $f(t, y)$, where $y(t_0)$ over a large interval, difficulties sometimes occur.

If $f_y(t, y) < 0$ and the step size is too large, a predictor-corrector method might be unstable. As a rule of thumb, stability exists when a small error is propagated as a decreasing error, and instability exists when a .small error is propagated as an increasing error. When too large a step size is used over a large interval, instability will result and is sometimes manifest by oscillations in the computed solution. They can be attenuated by changing to a smaller step size. Formula (7) through (9) suggest hove to modify the algorithm(s).

The Right Step

When step-size control is included, the following error estimate(s) should be used:

$$y(t_k) - y_k \approx 19\frac{p_k - y_k}{270} \quad \text{(Adams-Bashforth-Moulton)}, \tag{19}$$

$$y(t_k) - y_k \approx \frac{p_k - y_k}{29} \quad \text{(Milne-Simpson)} \tag{20}$$

$$y(t_k) - Y_k \approx \frac{p_k - y_k}{121} \quad \text{(Hamming)}. \tag{21}$$

In all methods, the corrector step is a type of fixed-point iteration. It can be proved that the step size h for the methods must satisfy the following conditions:

$$h << \frac{2.66667}{|f_y(t,y)|} \quad \text{(Adams-Bashforth-Moulton)}, \tag{22}$$

$$h << \frac{3.00000}{|f_y(t,y)|} \quad \text{(Milne-Simpson)} \tag{23}$$

$$h << \frac{2.66667}{|f_y(t,y)|} \quad \text{(Hamming)}. \tag{24}$$

The Right Step

The notation $<<$ in (22) through (24) means "much smaller than" The next example shows that more stringent inequalities should be used:

$$h < \frac{0.75}{|f_y(t,y)|} \quad \text{(Adams-Bashforth-Moulton)}, \tag{25}$$

$$h < \frac{0.45}{|f_y(t,y)|} \quad \text{(Milne-Simpson)} \tag{26}$$

$$h < \frac{0.69}{|f_y(t,y)|} \quad \text{(Hamming)}. \tag{27}$$

Inequality (27) is found in advanced books on numerical analysis. The other two inequalities seem appropriate for the example.

Example 9.1 Use the Adams-Bashforth-Moulton, Milne-Simpson, and Hamming methods and compute approximations for the solution of

$$y' = 30 - 5y, \quad y(0) = 1 \quad \text{over the interval} \quad [0, 10].$$

The Right Step

All three methods are of the order $O(h^4)$. When $N = 120$ steps was used for all three methods, the maximum error for each method occurred at a different place:

$y(0.41666667) - y_5 \approx -0.00277037$ (Adams-Bashforth-Moulton),
$y(0.33333333) - y_4 \approx -00139255$ (Milne-Simpson),
$y(0.33333333) - y_4 \approx -0.00104982$ (Hamming).

At the right end points $t = 10$, the error was

$y(10) - Y_{120} \approx 0.00000000$ (Adams-Bashforth-Moulton),
$y(10) - Y_{120} \approx 0.00001015$ (Milne-Simpson),
$y(10) - Y_{120} \approx 0.00000000$ (Hamming).

Both the Adams-Bashforth-Moulton and Hamming methods gave approximate solution with eight digits of accuracy at the right end point.

The Right Step

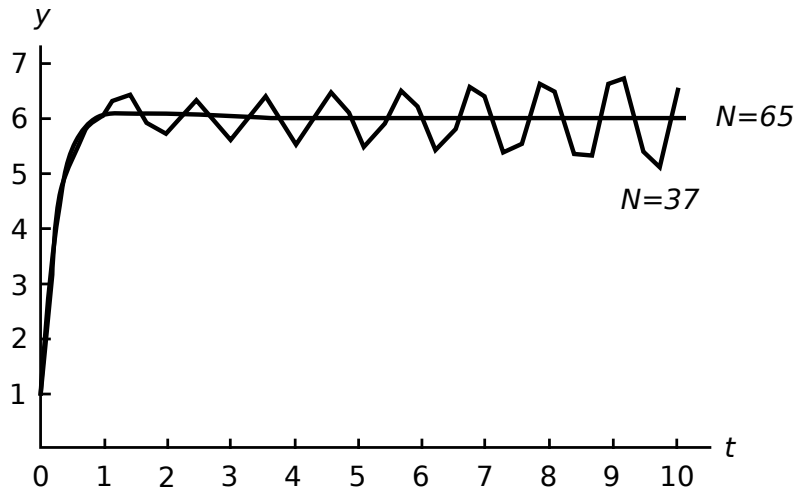


Figure 9.13 (a) The Adams-Bashforth-Moulton solution to $y' = 30 - 5y$ with $N = 37$ steps produces oscillation. It is stabilized when $N = 65$ because $h = 10/65 = 0.1538 \approx 0.15 = 0.75/5 = 0.75/|fy(t, y)|$.

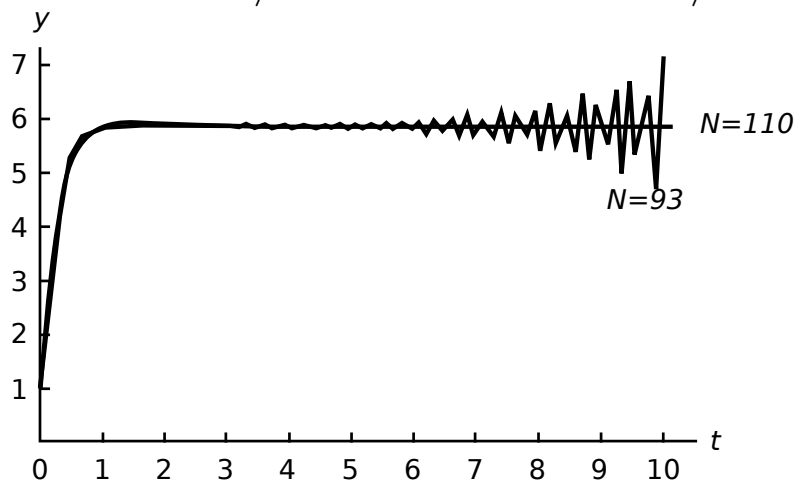


Figure 9.13 (b) The Milne-Simpson solution to $y' = 30 - 5y$ with $N = 93$ steps produces oscillation. It is stabilized when $N = 110$ because $h = -10/110 = 0.0909 \approx 0.09 = 0.45/5 = 0.45/|f_y(t, y)|$.

The Right Step

It is instructive to see that if the step size is too large the computed solution oscillates about the true solution. Figure 9.13 illustrates this phenomenon. The small number of steps was determined experimentally so that the oscillations was about the same magnitude. The large number of steps required to attenuate the oscillations were determined with equations (25) through (27).

Each of the following three programs requires that the first four coordinates of T and Y be initial starting values obtained by another method. Consider Example 9.13, where the step size was $h = \frac{1}{8}$ and the interval was $[0, 3]$.

The Right Step

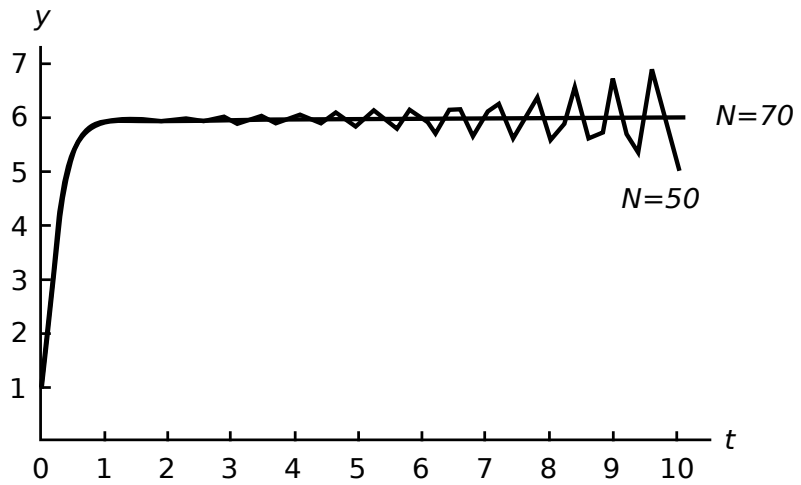


Figure 9.13 (c) Hamming's solation to $y' = 30 - 5y$ with $N = 50$ steps produces oscillation. It is stabilized when $N = 70$ because $h = 10/70 = 0.1428 \approx 0.138 = 0.69/5 = 0.695/|f_y(t, y)|$.

Systems of Differential Equations

This section is an introduction to systems of differential equations. To illustrate the concepts, we consider the initial value problem

$$\begin{aligned} \frac{dx}{dt} &= f(t,x,y) \\ \frac{dy}{dt} &= g(t,x,y) \end{aligned} \qquad \text{with} \qquad \begin{cases} x(t_0) = x_0, \\ y(t_0) = y_0. \end{cases} \tag{1}$$

A solution to (1) is a pair of differentiable functions $x(t)$ and $y(t)$ with the property that when $t, x(t)$, and $y(t)$ are substituted in $f(t,x,y)$ and $g(t,x,y)$, the result is equal to the derivative $x'(t)$ and $y'(t)$, respectively; that is,

$$\begin{aligned} x'(t) &= f(t,x(t),y(t)) \\ y'(t) &= g(t,x(t),y(t)) \end{aligned} \qquad \text{with} \qquad \begin{cases} x(t_0) = x_0, \\ y(t_0) = y_0. \end{cases} \tag{2}$$

Systems of Differential Equations

For example, consider the system of differential equations

$$\begin{aligned} \frac{dx}{dt} &= x + 2y \\ \frac{dy}{dt} &= 3x + 2y \end{aligned} \qquad \text{with} \qquad \begin{cases} x(0) = 6, \\ y(0) = 4. \end{cases} \tag{3}$$

The solution to the I.V.P. (3) is

$$\begin{aligned} x(t) &= 4e^{4t} + 2e^{-t}, \\ y(t) &= 6e^{4t} - 2e^{-t}. \end{aligned} \tag{4}$$

This is verified by directly substituting $x(t)$ and y(t) into the right-hand side of (3), computing the derivatives of (4), and substituting them in the left side of (3) to get

$$\begin{aligned} 16e^{4t} - 2e^{-t} &= (4e^{4t} + 2e^{-t}) + 2(6e^{4t} - 2e^{-t}), \\ 24e^{4t} + 2e^{-t} &= 3(4e^{4t} + 2e^{-t}) + 2(6e^{4t} - 2e^{-t}). \end{aligned}$$

Numerical Solutions

A numerical solution to (1) over the interval $a \leq t \leq b$ is found by considering the differentials

$$dx = f(t,x,y)dt \quad \text{and} \quad dy = g(t,x,y)dt. \tag{5}$$

Euler's method for solving the system is easy to formulate. The differentials $dt = t_{k+1} - t_k, dx = x_{k+1} - x_k$, and $dy = y_{k+1} - y_k$ are substituted into (5) to get

$$\begin{aligned} x_{k+1} - x_k &\approx f(t_k, x_k, y_k)(t_{k+1} - t_k), \\ y_{k+1} - y_k &\approx g(t_k, x_k, y_k)(t_{k+1} - t_k). \end{aligned} \tag{6}$$

The interval is divided into M subintervals of width $h = (b-a)/M$, and the mesh points are $t_{k+1} = t_k + h$. This is used in (6) to get the recursive formulas for Euler's method:

$$\begin{aligned} t_{k+1} &= t_k + h, \\ x_{k+1} &= x_k + hf(t_k, x_k, y_k), \\ y_{k+1} &= y_k + hg(t_k, x_k, y_k) \quad \textit{for } k = 0, 1, ..., M-1. \end{aligned} \tag{7}$$

Systems of Differential Equations

A higher-order method should be used to achieve a reasonable amount of accuracy. For example, the Runge-Kutta formulas of order 4 are

$$\begin{aligned} x_{k+1} &= x_k + \frac{h}{6}(f_1 + 2f_2 + 2f_3 + f_4), \\ y_{k+1} &= y_k + \frac{h}{6}(g_1 + 2g_2 + 2g_3 + g_4), \end{aligned} \tag{8}$$

where

$$\begin{aligned} f_1 &= f(t_k, x_k, y_k), & g_1 &= g(t_k, x_k, y_k), \\ f_2 &= f\left(t_k + \frac{h}{2}, x_k + \frac{h}{2}f_1, y_k + \frac{h}{2}g_1\right), & g_2 &= g\left(t_k + \frac{h}{2}, x_k + \frac{h}{2}f_1, y_k + \frac{h}{2}g_1\right), \\ f_3 &= f\left(t_k + \frac{h}{2}, x_k + \frac{h}{2}f_2, y_k + \frac{h}{2}g_2\right), & g_3 &= g\left(t_k + \frac{h}{2}, x_k + \frac{h}{2}f_2, y_k + \frac{h}{2}g_2\right), \\ f_4 &= f(t_k + h, x_k + hf_3, y_k + hg_3), & g_4 &= g(t_k + h, x_k + hf_3, y_k + hg_3). \end{aligned}$$

Example 9.15. Use the Runge-Kutta method in (8) and compute the numerical solution to (3) over the interval $[0.0, 0.2]$ using 10 subintervals and the step size $h = 0.02$.

Example 9.15. Use the Runge-Kutta method in (8) and compute the numerical solution to (3) over the interval $[0.0, 0.2]$ using 10 subintervals and the step size $h = 0.02$.

For the first point we have $t_1 = 0.02$, and the intermediate calculations required to compute x_1 and y_1 are

$$f_1 = f(0.00, 6.0, 4.0) = 14.0 \qquad g_1 = g(0.00, 6.0, 4.0) = 26.0$$

$$x_0 + \frac{h}{2} f_1 = 6.14 \qquad y_0 + \frac{h}{2} g_1 = 4.26$$

$$f_2 = f(0.01, 6.14, 4.26) = 14.66 \qquad g_2 = g(0.01, 6.14, 4.26) = 26.94$$

$$x_0 + \frac{h}{2} f_2 = 6.1466 \qquad y_0 + \frac{h}{2} g_2 = 4.2694$$

$$
\begin{aligned}
f_3 &= f(0.01, 6.1466, 4.2694) = 14.6854 \\
g_3 &= f(0.01, 6.1466, 4.2694) = 26.9786 \\
x_0 + hf_3 &= 6.293708 \quad y_0 + hg_3 = 4.539572 \\
f_4 &= f(0.02, 6.293708, 4.539572) = 15.372852 \\
g_4 &= f(0.02, 6.293708, 4.539572) = 27.960268.
\end{aligned}
$$

These values are used in the final computation:

$$
\begin{aligned}
x_1 &= 6 + \frac{0.02}{6}(14.0 + 2(14.66) + 2(14.6854) + 15.372852) = 6.29354551, \\
y_1 &= 4 + \frac{0.02}{6}(26.0 + 2(26.94) + 2(26.9786) + 27.960268) = 4.53932490.
\end{aligned}
$$

The calculations are summarized in Table 9.13.

Table 9.13 Runge-Kutta Solution to $x'(t) = x + 2y$, $y'(t) = 3x + 2y$ with the Initial Values $x(0) = 6$ and $y(0) = 4$

k	t_k	x_k	y_k
0	0.00	6.00000000	4.00000000
1	0.02	6.29354551	4.53932490
2	0.04	6.61562213	5.11948599
3	0.06	6.96852528	5.74396525
4	0.08	7.35474319	6.41653305
5	0.10	7.77697287	7.14127221
6	0.12	8.23813750	7.92260406
7	0.14	8.74140523	8.76531667
8	0.16	9.29020955	9.67459538
9	0.18	9.88827138	10.6560560
10	0.20	10.5396230	11.7157807

The numerical solutions contain a certain amount of error at each step. For the example above, the error grows, and at the right end point $t = 0.2$ it reaches its maximum:

$$x(0.2) - x_{10} = 10.5396252 - 10.5396230 = 0.0000022,$$
$$y(0.2) - y_{10} = 11.7157841 - 11.7157807 = 0.0000034.$$

Higher-Order Differential Equations

Higher-order differential equations involve the higher derivatives $x''(t), x'''(t)$, and so on. They arise in mathematical models for problems in physics and engineering. For example,

$$mx''(t) + cx'(t) + kx(t) = g(t)$$

represents a mechanical system in which a spring with spring constant k restores a displaced mass m. Damping is assumed to be proportional to the velocity, and the function $g(t)$ is an external force. It is often the case that the position $x(t_0)$ and velocity $x'(t_0)$ are known at a certain time t_0.

Systems of Differential Equations

By solving for the second derivative, we can write a second-order initial value problem in the form

$$x''(t) = f(t, x(t), x'(t)) \quad \text{with} \quad x(t_0) = x_0 \ \ \textit{and} \ \ x'(t_0) = y_0. \tag{9}$$

The second-order differential equation can be reformulated as a system of two first-order equations if we use the substitution

$$x'(t) = y(t). \tag{10}$$

Then $x''(t) = y'(t)$ and the differential equation in (9) becomes a system:

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= f(t, x, y) \end{aligned} \qquad \text{with} \qquad \begin{cases} x(t_0) = x_0, \\ y(t_0) = y_0. \end{cases} \tag{11}$$

A numerical procedure such as the Runge-Kutta method can be used to solve (11) and will generate two sequences $\{x_k\}$ and $\{y_k\}$. The first sequence is the numerical solution to (9). The next example can be interpreted as damped harmonic motion.

Example 9.16. Consider the second-order initial value problem

$$x''(t) + 4x'(t) + 5x(t) = 0 \quad \text{with} \quad x(0) = 3 \quad \text{and} \quad x'(0) = -5.$$

(a) Write down the equivalent system of two first-order equations.

(b) Use the Runge-Kutta method to solve the reformulated problem over $[0, 5]$ using $M = 50$ subintervals of width $h = 0.1$.

(c) Compare the numerical solution with the true solution:

$$x(t) = 3e^{-2t}cos(t) + e^{-2t}sin(t).$$

(a) The differential equation has the form

$$x''(t) = f(t, x(t), x'(t)) = -4x'(t) - 5x(t).$$

(b) Using the substitution in (10), we get the reformulated problem:

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= -5x - 4y \end{aligned} \qquad \text{with} \qquad \begin{cases} x(0) = 3, \\ y(0) = -5. \end{cases}$$

(c) Samples of the numerical computations are given in Table 9.14. The values $\{y_k\}$ are extraneous and are not included. Instead, the true solution values $\{x(t_k)\}$ are included for comparison.

Table 9.14 Runge-Kutta Solution to $x''(t) + 4x'(t) + 5x(t) = 0$ with the Initial Conditions $x(0) = 3$ and $x'(0) = -5$

k	t_k	x_k	$x(t_k)$
0	0.0	3.00000000	3.00000000
1	0.1	2.52564583	2.52565822
2	0.2	2.10402783	2.10404686
3	0.3	1.73506269	1.73508427
4	0.4	1.41653369	1.41655509
5	0.5	1.14488509	1.14490455
10	1.0	0.33324302	0.33324661
20	2.0	−0.00620684	−0.00621162
30	3.0	−0.00701079	−0.00701204
40	4.0	−0.00091163	−0.00091170
48	4.8	−0.00004972	−0.00004969
49	4.9	−0.00002348	−0.00002345
50	5.0	−0.00000493	−0.00000490

Boundary Value Problems

Another type of differential equation has the form

$$(1) \qquad x'' = f(t, x, x') \qquad \text{for} \qquad a \leq t \leq b,$$

with the boundary conditions

$$(2) \qquad x(a) = \alpha \qquad \text{and} \quad x(b) = \beta.$$

This is called a ***boundary value problem***.

The conditions that guarantee that a solution to (1) exists should be checked before any numerical scheme is applied; otherwise, a list of meaningless output may be generated. The general conditions are stated in the following theorem.

Theorem 9.8 (Boundary Value Problem).

Assume that $f(t,x,y)$ is continuous on the region $R = \{(t,x,y) : a \leq t \leq b, -\infty < x < \infty, -\infty < y < \infty\}$ and that $\partial f/\partial x = fx(t,x,y)$ and $\partial f/\partial y = fy(t,x,y)$ are continuous on R. If there exists a constant $M > 0$ for which f_x and f_y satisfy

$$(3) \qquad f_x(t,x,y) > 0 \quad \text{for all} \quad (t,x,y) \in R \quad \text{and}$$

$$(4) \qquad |f_y(t,x,y)| \leq M \quad \text{for all} \quad (t,x,y) \in R,$$

then the boundary value problem

$$(5) \qquad x" = f(t,x,x') \quad \text{with} \quad x(a) = \alpha \quad \textit{and} \quad x(b) = \beta$$

has a unique solution $x = x(t)$ for $a \leq t \leq b$.

Theorem 9.8 (Boundary Value Problem).

The notation $y = x(t)$ has been used to distinguish the third variable of the function $f(t, x, x)$. Finally, the special case of linear differential equations is worthy of mention.

Corollary 9.1 (Linear Boundary Value Problem).

Assume that f in Theorem 9.8 has the form $f(t, x, y) = p(t)y + q(t)x + r(t)$ and that f and its partial derivatives $\partial f/\partial x = q(t)$ and $\partial f/\partial y = p(t)$ are continuous on R. If there exists a constant $M > 0$ for which $p(t)$ and $q(t)$ satisfy

$$q(t) > 0 \quad \text{for all} \quad t \in [a, b] \tag{6}$$

and

$$|p(t)| \leq M = \max_{a \leq t \leq b} \{|p(t)|\} \tag{7}$$

Theorem 9.8 (Boundary Value Problem).

then the linear boundary value problem

$$x'' = p(t)x'(t) + q(t)x(t) + r(t) \quad \text{with} \quad x(a) = \alpha \quad \textit{and} \quad x(b) = \beta \tag{8}$$

has a unique solution $x = x(t)$ over $a \leq t \leq b$.

Reduction to Two I.V.P.s: Linear Shooting Method

Finding the solution of a linear boundary problem is assisted by the linear structure of the equation and the use of two special initial value problems. Suppose that $u(t)$ is the unique solution to the I.V.P.

$$u'' = p(t)u'(t) + q(t)u(t) + r(t) \quad \text{with} \quad u(a) = \alpha \quad \text{and} \quad u'(a) = 0. \tag{9}$$

Furthermore, suppose that v(t) is the unique solution to the I.V.P.

Boundary Value Problems

$$(10) \qquad v'' = p(t)v'(t) + q(t)v(t) \quad \text{with} \quad v(a) = 0 \quad \text{and} \quad v'(a) = 1.$$

Then the linear combination

$$(11) \qquad x(t) = u(t) + Cv(t)$$

is a solution to $x'' = p(t)x'(t) + q(t)x(t) + r(t)$ as seen by the computation

$$\begin{aligned} x'' = u'' + Cv'' &= p(t)u'(t) + q(t)u(t) + r(t) + p(t)Cv'(t) + q(t)Cv(t) \\ &= p(t)(u'(t) + Cv'(t)) + q(t)(u(t) + Cv(t)) + r(t) \\ &= p(t)x'(t) + q(t)x(t) + r(t). \end{aligned}$$

The solution $x(t)$ in equation (11) takes on the boundary values

$$(12) \quad \begin{aligned} x(a) &= u(a) + Cv(a) = \alpha + 0 = \alpha, \\ x(b) &= u(b) + Cv(b). \end{aligned}$$

Imposing the boundary condition$x(b) = \beta$ in (12) produces $C = (\beta - u(b))/v(b)$.

Therefore, if $v(b) \neq 0$, the unique solution to (8) is

$$x(t) = u(t) + \frac{\beta - u(b)}{v(b)} v(t). \tag{13}$$

Remark. If q fulfills the hypotheses of Corollary 9.1, this rules out the troublesome solution v(t) = 0, so that (13) is the form of the required solution. The details are left for the reader to investigate in the exercises.

Example 9.17. Solve the boundary value problem

$$x''(t) = \frac{2t}{1+t^2} x'(t) - \frac{2}{1+t^2} x(t) + 1$$

with $x(0) = 1.25$ and $x(4) = -0.95$ over the interval $[0, 4]$.

The functions p, q, and r are $p(t) = 2t/(1+t^2), q(t) = -2/(1+t^2)$, and $r(t) = 1$, respectively. The Runge-Kutta method of order 4 with step size $h = 0.2$ is used to construct numerical solutions $\{u_j\}$ and $\{v_j\}$ to equations (9) and (10), respectively.

The approximations $\{u_j\}$ for $u(t)$ are given in the first column of Table 9.15. Then $u(4) \approx u_{20} = -2.893535$ and $v(4) \approx v_{20} = 4$ are used with (13) to construct

$$w_j = \frac{b - u(4)}{v(4)} v_j = 0.485884 v_j.$$

Then the required approximate solution is $x_j = u_j + w_j$. Sample computations are given in Table 9.15, and Figure 9.24 shows their graphs. The reader can verify that $v(t) = t$ is the analytic solution for boundary value problem (10); that is,

$$v''(t) = \frac{2t}{1+t^2} v'(t) - \frac{2}{1+t^2} v(t)$$

with the initial conditions $v(0) = 0$ and $v'(0) = 1$.

Table 9.15 Approximate Solutions $x_j = u_j + w_j$ to the Equation $x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2} + 1$

t_j	u_j	w_j	$x_j = u_j + w_j$
0.0	1.250000	0.000000	1.250000
0.2	1.220131	0.097177	1.317308
0.4	1.132073	0.194353	1.326426
0.6	0.990122	0.291530	1.281652
0.8	0.800569	0.388707	1.189276
1.0	0.570844	0.485884	1.056728
1.2	0.308850	0.583061	0.891911
1.4	0.022522	0.680237	0.702759
1.6	−0.280424	0.777413	0.496989
1.8	−0.592609	0.874591	0.281982
2.0	−0.907039	0.971767	0.064728
2.2	−1.217121	1.068944	−0.148177
2.4	−1.516639	1.166121	−0.350518
2.6	−1.799740	1.263297	−0.536443
2.8	−2.060904	1.360474	−0.700430
3.0	−2.294916	1.457651	−0.837265
3.2	−2.496842	1.554828	−0.942014
3.4	−2.662004	1.652004	−1.010000
3.6	−2.785960	1.749181	−1.036779
3.8	−2.864481	1.846358	−1.018123
4.0	−2.893535	1.943535	−0.950000

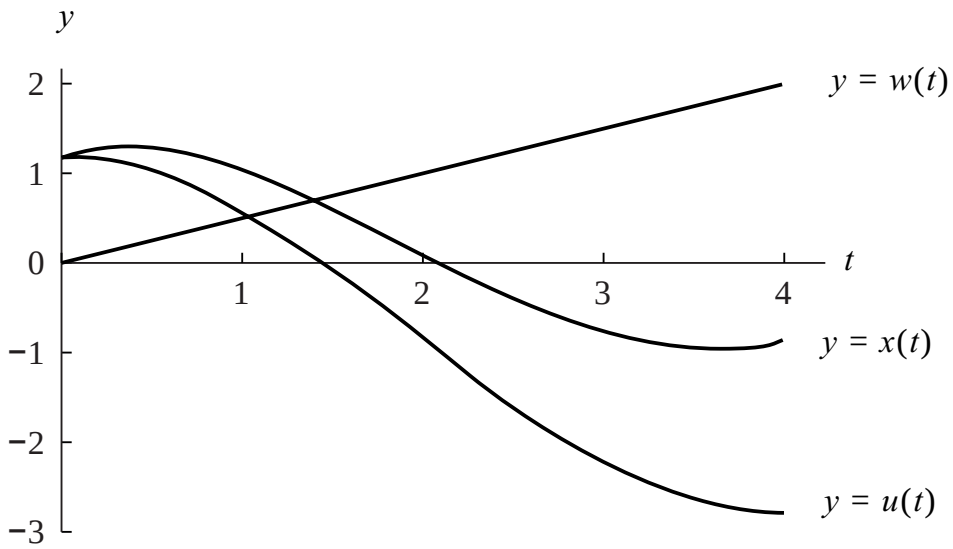


Figure 9.24 The numerical approximations $u(t)$ used to form $x(t) = u(t) + w(t)$, which is the solution to $x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1$

The approximations in Table 9.16 compare numerical solutions obtained with the linear shooting method with the step sizes $h = 0.2$ and $h = 0.1$ and the analytic solution

$$x(t) = 1.25 + 0.4860896526t - 2.25t^2 + 2t\ arctan(t) - \frac{1}{2}ln(1+t^2) + \frac{1}{2}t^2 ln(1+t^2).$$

Table 9.16 Numerical Approximations for $x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2} + 1$

t_j	x_j $h = 0.2$	$x(t_j)$ exact	$x(t_j) - x_j$ error	t_j	x_j $h = 0.1$	$x(t_j)$ exact	$x(t_j) - x_j$ error
0.0	1.250000	1.250000	0.000000	0.0	1.250000	1.250000	0.000000
				0.1	1.291116	1.291117	0.000001
0.2	1.317308	1.317350	0.000042	0.2	1.317348	1.317350	0.000002
				0.3	1.328986	1.328990	0.000004
0.4	1.326426	1.326505	0.000079	0.4	1.326500	1.326505	0.000005
				0.5	1.310508	1.310514	0.000006
0.6	1.281652	1.281762	0.000110	0.6	1.281756	1.281762	0.000006
0.8	1.189276	1.189412	0.000136	0.8	1.189404	1.189412	0.000008
1.0	1.056728	1.056886	0.000158	1.0	1.056876	1.056886	0.000010
1.2	0.891911	0.892086	0.000175	1.2	0.892076	0.892086	0.000010
1.6	0.496989	0.497187	0.000198	1.6	0.497175	0.497187	0.000012
2.0	0.064728	0.064931	0.000203	2.0	0.064919	0.064931	0.000012
2.4	− 0.350518	− 0.350325	0.000193	2.4	− 0.350337	− 0.350325	0.000012
2.8	− 0.700430	− 0.700262	0.000168	2.8	− 0.700273	− 0.700262	0.000011
3.2	− 0.942014	− 0.941888	0.000126	3.2	− 0.941895	− 0.941888	0.000007
3.6	− 1.036779	− 1.036708	0.000071	3.6	− 1.036713	− 1.036708	0.000005
4.0	− 0.950000	− 0.950000	0.000000	4.0	− 0.950000	− 0.950000	0.000000

A graph of the approximate solution when $h = 0.2$ is given in Figure 9.25. Included in the table are columns for the error. Since the Runge-Kutta solutions have error of order $O(h^4)$, the error in the solution with the smaller step size $h = 0.1$ is about $\frac{1}{16}$ the error of the solution with the large step size $h = 0.2$.

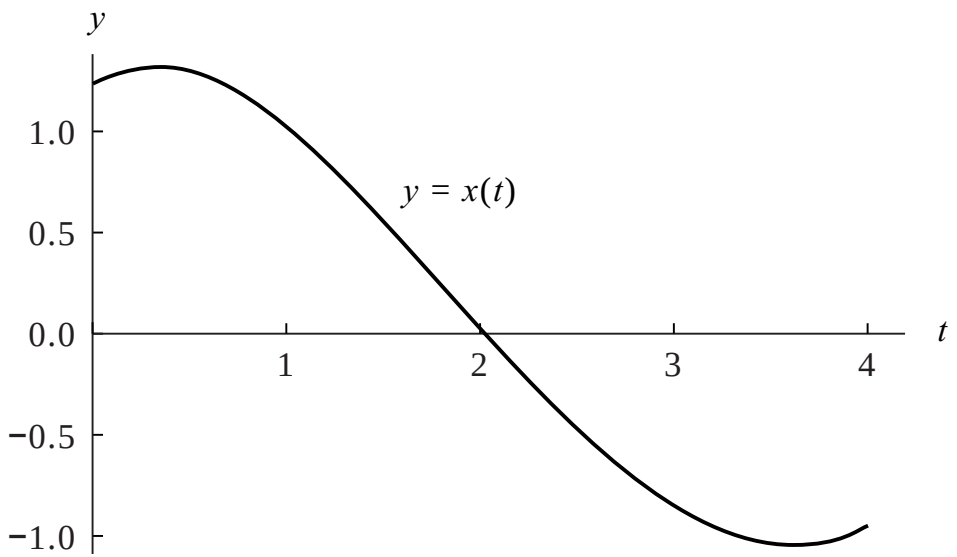


Figure 9.25 The graph of the numerical approximations for $x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1$

Finite-Difference Method

Methods involving difference quotient approximations for derivatives can be used for solving certain second-order boundary value problems. Consider the linear equation

$$x'' = p(t)x'(t) + q(t)x(t) + r(t) \tag{1}$$

over $[a, b]$ with $x(a) = \alpha$ and $x(b) = \beta$. Form a partition of $[a, b]$ using the points $a = t_0 < t_1 < \cdots < t_N = b$, where $h = (b - a)/N$ and $t_j = a + jh$ for $j = 0, 1, ..., N$. The central-difference formulas discussed in Chapter 6 are used to approximate the derivatives

$$x'(t_j) = \frac{x(t_{j+1}) - x(t_{j-1})}{2h} + \boldsymbol{O}(h^2) \tag{2}$$

and

$$x''(t_j) = \frac{x(t_{j+1}) - 2x(t_j) + x(t_{j-1})}{h^2} + \boldsymbol{O}(h^2). \tag{3}$$

Finite-Difference Method

To start the derivation, we replace each term $x(t_j)$ on the right side of (2) and (3) with x_j , and the resulting equations are substituted into (1) to obtain the relation

$$\frac{x_{j+1} - 2x_j + x_{j-1}}{h^2} + \boldsymbol{O}(h^2) = p(t_j)\left(\frac{x_{j+1} - x_{j-1}}{2h} + \boldsymbol{O}(h^2)\right) + q(t_j)x_j + r(t_j). \quad (4)$$

Next, we drop the two terms $\boldsymbol{O}(h^2)$ in (4) and introduce the notation $p_j = p(t_j)$, $q_j = q(t_j)$, and $r_j = r(t_j)$; this produces the difference equation

$$\frac{x_{j+1} - 2x_j + x_{j-1}}{h^2} = p_j \frac{x_{j+1} - x_{j-1}}{2h} + q_j x_j + r_j, \quad (5)$$

which is used to compute numerical approximations to the differential equation (1).

Finite-Difference Method

This is carried out by multiplying each side of (5) by h_2 and then collecting terms involving x_{j-1}, x_j , and x_{j+1} and arranging them in a system of linear equations:

$$\left(\frac{-h}{2}p_j - 1\right)x_{j-1} + (2 + h^2 q_j)x_j + \left(\frac{h}{2}p_j - 1\right)x_{j+1} = -h^2 r_j, \tag{6}$$

for $j = 1, 2, ..., N-1$, where $x_0 = \alpha$ and $x_N = \beta$. The system in (6) has the familiar tridiagonal form, which is more visible when displayed with matrix notation:

Finite-Difference Method

$$
\begin{bmatrix}
2+h^2q_1 & \frac{h}{2}p_1-1 & & & & \\
\frac{-h}{2}p_2-1 & 2+h^2q_2 & \frac{h}{2}p_2-1 & & & \boldsymbol{O} \\
 & \frac{-h}{2}p_j-1 & 2+h^2q_j & \frac{h}{2}p_j-1 & & \\
\boldsymbol{O} & & \frac{-h}{2}p_{N-2}-1 & 2+h^2q_{N-2} & \frac{h}{2}p_{N-2}-1 & \\
 & & & \frac{-h}{2}p_{N-1}-1 & 2+h^2q_{N-1}
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_j \\ x_{N-2} \\ x_{N-1} \end{bmatrix}
$$

$$
= \begin{bmatrix} -h^2r_1+e_0 \\ -h^2r_2 \\ -h^2r_j \\ -h^2r_{N-2} \\ -h^2r_{N-1}+e_N \end{bmatrix},
$$

Finite-Difference Method

where

$$e_0 = \left(\frac{h}{2}p_1 + 1\right)\alpha \quad \text{and} \quad e_N = \left(\frac{-h}{2}p_{N-1} + 1\right)\beta.$$

When computations with step size h are used, the numerical approximation to the solution is a set of discrete points $\{(t_j, x_j)\}$; if the analytic solution $x(t_j)$ is known, we can compare x_j and $x(t_j)$.

Example 9.18. Solve the boundary value problem

$$x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1$$

with $x(0) = 1.25$ and $x(4) = -0.95$ over the interval $[0, 4]$.

The functions p, q, and r are $p(t) = 2t/(1+t^2)$, $q(t) = -2/(1+t^2)$, and $r(t) = 1$, respectively. The finite-difference method is used to construct numerical solutions $\{x_j\}$ using the system of equations (6). Sample values of the approximations $\{x_{j,1}\}$, $\{x_{j,2}\}$, $\{x_{j,3}\}$, and $\{x_{j,4}\}$ corresponding to the step sizes $h_1 = 0.2$, $h_2 = 0.1$, $h_3 = 0.05$, and $h_4 = 0.025$ are given in Table 9.17. Figure 9.26 shows the graph of the polygonal path formed from $\{(t_j, x_{j,1})\}$ for the case $h_1 = 0.2$. There are 41 terms in the sequence generated with $h_2 = 0.1$, and the sequence $\{x_{j,2}\}$ only includes every other term from these computations; they correspond to the 21 values of $\{t_j\}$ given in Table 9.17.

Finite-Difference Method

Similarly, the sequences $\{x_{j,3}\}$ and $\{x_{j,4}\}$ are a portion of the values generated with step sizes $h_3 = 0.05$ and $h_4 = 0.025$, respectively, and they correspond to the 21 values of $\{t_j\}$ in Table 9.17.

Next we compare numerical solutions in Table 9.17 with the analytic solution: $x(t) = 1.25 + 0.486089652t - 2.25t^2 + 2t\ arctan(t) - \frac{1}{2}ln(1+t^2) + \frac{1}{2}t^2 ln(1+t^2)$. The numerical solutions can be shown to have error of order $\boldsymbol{O}(h^2)$. Hence reducing the step size by a factor of $\frac{1}{2}$ results in the error being reduced by about $\frac{1}{4}$.

A careful scrutiny of Table 9.18 will reveal that this is happening. For instance, at $t_j = 1.0$ the errors incurred with step sizes h_1, h_2, h_3, and h_4 are $e_{j,1} = 0.014780$, $e_{j,2} = 0.003660$, $e_{j,3} = 0.000913$, and $e_{j,4} = 0.000228$, respectively.

Their successive ratios $e_{j,2}/e_{j,1} = 0.003660/0.014780 = 0.2476$, $e_{j,3}/e_{j,2} = 0.000913/0.003660 = 0.2495$, and $e_{j,4}/e_{j,3} = 0.000228/0.000913 = 0.2497$ are approaching $\frac{1}{4}$.

Finite-Difference Method

Table 9.17 Numerical Approximations for $x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1$

t_j	$x_{j,1}$ $h = 0.2$	$x_{j,2}$ $h = 0.1$	$x_{j,3}$ $h = 0.05$	$x_{j,4}$ $h = 0.025$	$x(t_j)$ exact
0.0	1.250000	1.250000	1.250000	1.250000	1.250000
0.2	1.314503	1.316646	1.317174	1.317306	1.317350
0.4	1.320607	1.325045	1.326141	1.326414	1.326505
0.6	1.272755	1.279533	1.281206	1.281623	1.281762
0.8	1.177399	1.186438	1.188670	1.189227	1.189412
1.0	1.042106	1.053226	1.055973	1.056658	1.056886
1.2	0.874878	0.887823	0.891023	0.891821	0.892086
1.4	0.683712	0.698181	0.701758	0.702650	0.702947
1.6	0.476372	0.492027	0.495900	0.496865	0.497187
1.8	0.260264	0.276749	0.280828	0.281846	0.282184
2.0	0.042399	0.059343	0.063537	0.064583	0.064931
2.2	−0.170616	−0.153592	−0.149378	−0.148327	−0.147977
2.4	−0.372557	−0.355841	−0.351702	−0.350669	−0.350325
2.6	−0.557565	−0.541546	−0.537580	−0.536590	−0.536261
2.8	−0.720114	−0.705188	−0.701492	−0.700570	−0.700262
3.0	−0.854988	−0.841551	−0.838223	−0.837393	−0.837116
3.2	−0.957250	−0.945700	−0.942839	−0.942125	−0.941888
3.4	−1.022221	−1.012958	−1.010662	−1.010090	−1.009899
3.6	−1.045457	−1.038880	−1.037250	−1.036844	−1.036709
3.8	−1.022727	−1.019238	−1.018373	−1.018158	−1.018086
4.0	−0.950000	−0.950000	−0.950000	−0.950000	−0.950000

Finite-Difference Method

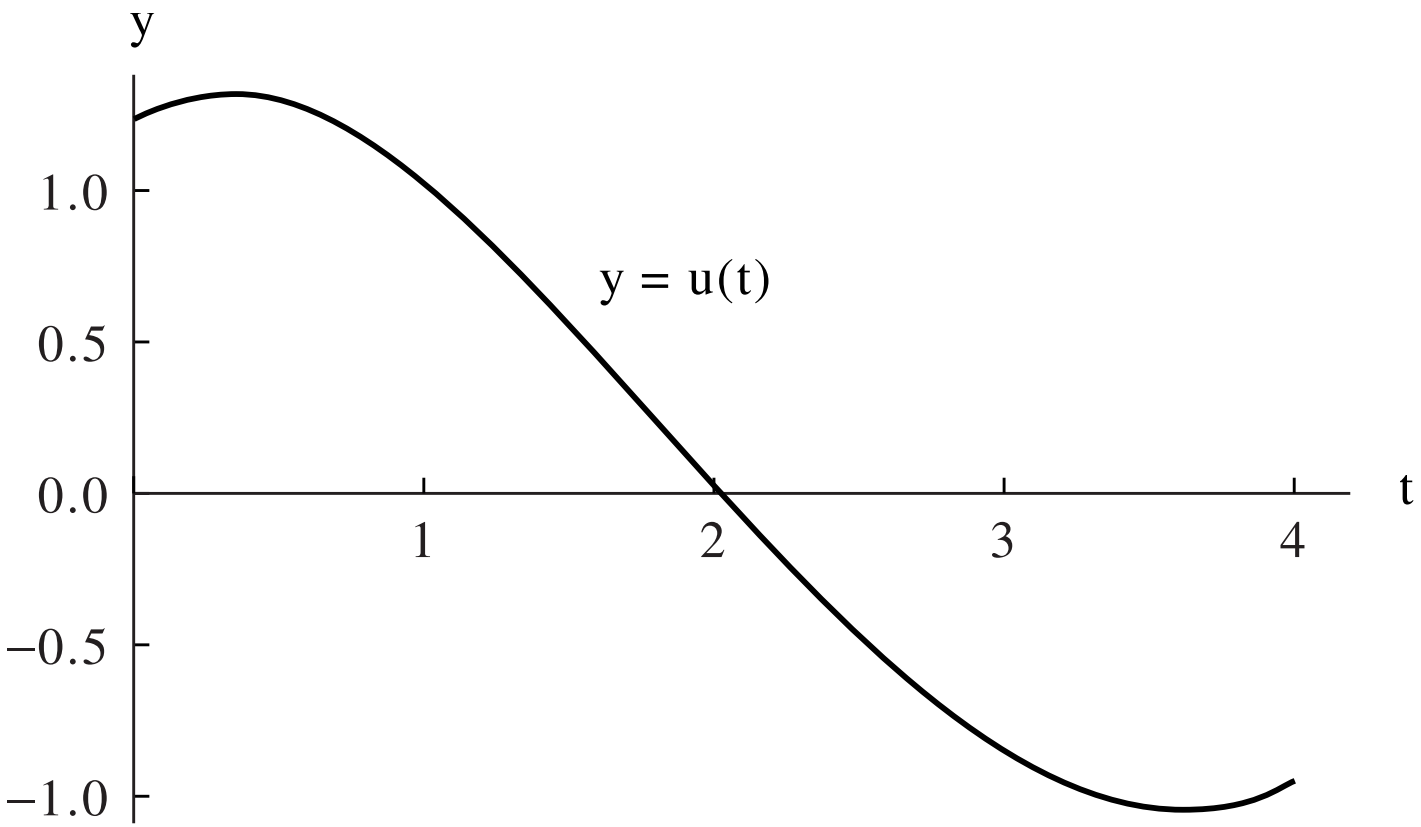


Figure 9.26 The graph of the numerical approximation for $x(t) = u(t) + w(t)$, which is the solution to

$$x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1$$

(using $h = 0.2$).

Finite-Difference Method

Finally, we show how Richardson's improvement scheme can be used to extrapolate the seemingly inaccurate sequences $\{x_{j,1}\}$, $\{x_{j,2}\}$, $\{x_{j,3}\}$, and $\{x_{j,4}\}$ and obtain six digits of precision. Eliminate the error terms $\boldsymbol{O}(h^2)$and $\boldsymbol{O}((h/2)^2)$ in the approximations $\{x_{j,1}\}$ and $\{x_{j,2}\}$ by generating the extrapolated sequence $\{z_{j,1}\} = \{(4x_{j,2} - x_{j,1})/3\}$. Similarly, the error terms $\boldsymbol{O}((h/2)^2)$ and $\boldsymbol{O}((h/4)^2)$ for $\{x_{j,2}\}$ and $\{x_{j,3}\}$ are eliminated by generating $\{z_{j,2}\} = \{(4x_{j,3} - x_{j,2})/3\}$.

Finite-Difference Method

It has been shown that the second level of Richardson's improvement scheme applies to the sequences $\{z_{j,1}\}$ and $\{z_{j,2}\}$, so the third improvement is $\{(16z_{j,2} - z_{j,1})/15\}$.
Let us illustrate the situation by finding the extrapolated values that correspond to $t_j = 1.0$. The first extrapolated value is

$$\frac{4x_{j,2} - x_{j,1}}{3} = \frac{4(1.053226) - 1.042106}{3} = 1.056932 = z_{j,1}.$$

The second extrapolated value is

$$\frac{4x_{j,3} - x_{j,2}}{3} = \frac{4(1.055973) - 1.053226}{3} = 1.056889 = z_{j,2}.$$

Finally, the third extrapolation involves the terms $z_{j,1}$ and $z_{j,2}$:

$$\frac{16z_{j,2} - zj,1}{15} = \frac{16(1.056889) - 1.056932}{15} = 1.056886.$$

This last computation contains six decimal places of accuracy. The values at the other points are given in Table 9.19.

Finite-Difference Method

Table 9.18 Errors in Numerical Approximations Using the Finite-Difference Method

t_j	$x(t_j) - x_{j,1} = e_{j,1}$	$x(t_j) - x_{j,2} = e_{j,2}$	$x(t_j) - x_{j,3} = e_{j,3}$	$x(t_j) - x_{j,4} = e_{j,4}$
	$h_1 = 0.2$	$h_2 = 0.1$	$h_3 = 0.05$	$h_4 = 0.025$
0.0	0.000000	0.000000	0.000000	0.000000
0.2	0.002847	0.000704	0.000176	0.000044
0.4	0.005898	0.001460	0.000364	0.000091
0.6	0.009007	0.002229	0.000556	0.000139
0.8	0.012013	0.002974	0.000742	0.000185
1.0	0.014780	0.003660	0.000913	0.000228
1.2	0.017208	0.004263	0.001063	0.000265
1.4	0.019235	0.004766	0.001189	0.000297
1.6	0.020815	0.005160	0.001287	0.000322
1.8	0.021920	0.005435	0.001356	0.000338
2.0	0.022533	0.005588	0.001394	0.000348
2.2	0.022639	0.005615	0.001401	0.000350
2.4	0.022232	0.005516	0.001377	0.000344
2.6	0.021304	0.005285	0.001319	0.000329
2.8	0.019852	0.004926	0.001230	0.000308
3.0	0.017872	0.004435	0.001107	0.000277
3.2	0.015362	0.003812	0.000951	0.000237
3.4	0.012322	0.003059	0.000763	0.000191
3.6	0.008749	0.002171	0.000541	0.000135
3.8	0.004641	0.001152	0.000287	0.000072
4.0	0.000000	0.000000	0.000000	0.000000

Table 9.19 Extrapolation of the Numerical Approximations $\{x_{j,1}\}$, $\{x_{j,2}\}$, $\{x_{j,3}\}$ Obtained with the Finite-Difference Method

t_j	$\frac{4x_{j,2}-x_{j,1}}{3} = z_{j,1}$	$\frac{4x_{j,3}-x_{j,2}}{3} = z_{j,2}$	$\frac{16z_{j,2}-z_{j,1}}{3}$	$x(t_j)$ Exact solution
0.0	1.250000	1.250000	1.250000	1.250000
0.2	1.317360	1.317351	1.317350	1.317350
0.4	1.326524	1.326506	1.326504	1.326505
0.6	1.281792	1.281764	1.281762	1.281762
0.8	1.189451	1.189414	1.189412	1.189412
1.0	1.056932	1.056889	1.056886	1.056886
1.2	0.892138	0.892090	0.892086	0.892086
1.4	0.703003	0.702951	0.702947	0.702948
1.6	0.497246	0.497191	0.497187	0.497187
1.8	0.282244	0.282188	0.282184	0.282184
2.0	0.064991	0.064935	0.064931	0.064931
2.2	−0.147918	−0.147973	−0.147977	−0.147977
2.4	−0.350268	−0.350322	−0.350325	−0.350325
2.6	−0.536207	−0.536258	−0.536261	−0.536261
2.8	−0.700213	−0.700259	−0.700263	−0.700262
3.0	−0.837072	−0.837113	−0.837116	−0.837116
3.2	−0.941850	−0.941885	−0.941888	−0.941888
3.4	−1.009870	−1.009898	−1.009899	−1.009899
3.6	−1.036688	−1.036707	−1.036708	−1.036708
3.8	−1.018075	−1.018085	−1.018086	−1.018086
4.0	−0.950000	−0.950000	−0.950000	−0.950000