Chap. 2: Initial Value Problems

§2.1: Introduction

MA385/530 – Numerical Analysis 1

October 2019

Emile Picard: his fundamental work on differential equations was only one of his many contributions to mathematics

Olga Ladyzhenskaya: her extensive achievements include providing the first proof of the convergence of finite difference methods for the Navier-Stokes equations

(See Chap 6 of Epperson for the introduction, and Chapter 12 of Süli and Mayers for the rest).

Motivation

The growth of some tumours can be modelled as

$$R'(t) = -\frac{1}{3}S_i R(t) + \frac{2\lambda\sigma}{\mu R + \sqrt{\mu^2 R^2 + 4\sigma}},$$

subject to the initial condition $R(t_0) = a$, where R is the radius of the tumour at time t.
Clearly, it would be useful to know the value of R as certain times in the future. Though it's essentially impossible to solve for R exactly, we can accurately estimate it. In this section, we'll study techniques for this.

Initial Value Problems (IVPs)

Initial Value Problems (IVPs) are differential equations of the form: *Find $y(t)$ such that*

$$\frac{\mathrm{d}\,y}{\mathrm{d}\,t} = f(t,y) \quad \text{for } t > t_0, \qquad \text{and } y(t_0) = y_0. \tag{1}$$

Here $y' = f(t,y)$ is the *differential equation* and $y(t_0) = y_0$ is the *initial value*.

Some IVPs are easy to solve. For example:

$$y' = t^2 \quad \text{with } y(1) = 1.$$

Most problems are much harder, and some don't have solutions at all. In many cases, it is possible to determine that a giving problem does indeed have a solution, even if we can't write it down. The idea is that the function f should be "Lipschitz", a notion closely related to that of a ***contraction***.

Definition 2.1 (Lipschitz Condition)

A function f satisfies a ***Lipschitz Condition*** (with respect to its second argument) in the rectangular region D if there is a positive real number L such that

$$|f(t,u) - f(t,v)| \leq L|u - v| \tag{2}$$

for all $(t,u) \in D$ and $(t,v) \in D$.

Example 2.2

For each of the following functions f, show that is satisfies a *Lipschitz condition*, and give an upper bound on the Lipschitz constant L.

(i) $f(t, y) = y/(1 + t)^2$ for $0 \leq t \leq \infty$.

(ii) $f(t, y) = 4y - e^{-t}$ for all t.

(iii) $f(t, y) = -(1 + t^2)y + \sin(t)$ for $1 \leq t \leq 2$.

Theorem 2.3 (Picard's)

Suppose that the real-valued function $f(t, y)$ is continuous for $t \in [t_0, t_M]$ and $y \in [y_0 - C, y_0 + C]$; that $|f(t, y_0)| \leq K$ for $t_0 \leq t \leq t_M$; and that f satisfies the *Lipschitz condition* (2). If

$$C \geq \frac{K}{L}\left(e^{L(t_M - t_0)} - 1 \right),$$

then (1) has a unique solution on $[t_0, t_M]$. Furthermore

$$|y(t) - y(t_0)| \leq C \qquad t_0 \leq t \leq t_M.$$

You are not required to know this theorem for this course. However, it's important to be able to determine when a given f satisfies a Lipschitz condition.

Exercise 2.1

For the following functions show that they satisfy a Lipschitz condition on the corresponding domain, and give an upper-bound for L:

(i) $f(t, y) = 2yt^{-4}$ for $t \in [1, \infty)$,

(ii) $f(t, y) = 1 + t \sin(ty)$ for $0 \leq t \leq 2$.

Exercise 2.2

Many text books, instead of giving the version of the Lipschitz condition we use, give the following: *There is a finite, positive, real number L such that*

$$\left|\frac{\partial}{\partial y} f(t, y)\right| \leq L \qquad \text{for all } (t, y) \in D.$$

Is this statement *stronger than* (i.e., more restrictive then), *equivalent to* or *weaker than* (i.e., less restrictive than) the usual Lipschitz condition? Justify your answer.

Hint: the Wikipedia article on Lipschitz continuity is very informative.

Initial Value Problems

§2.2: Euler's Method

MA385/530 – Numerical Analysis 1

October 2019 (Week 5)

Our goal is to generate numerical solutions to initial value differential equations. The solutions to such problems are functions (usually, of one variable that we'll denote t). Our approximation will give estimates of the values of this function at certain points.

We'll denote the points we at which we are seeking approximations as

$$t_0 < t_1 < \cdots < t_n.$$

The methods we'll use are all **one-step** methods, and the first example we'll consider is ***Euler's Method***.

Although it is not too important, we'll make the assumption that the points are equally spaced. So

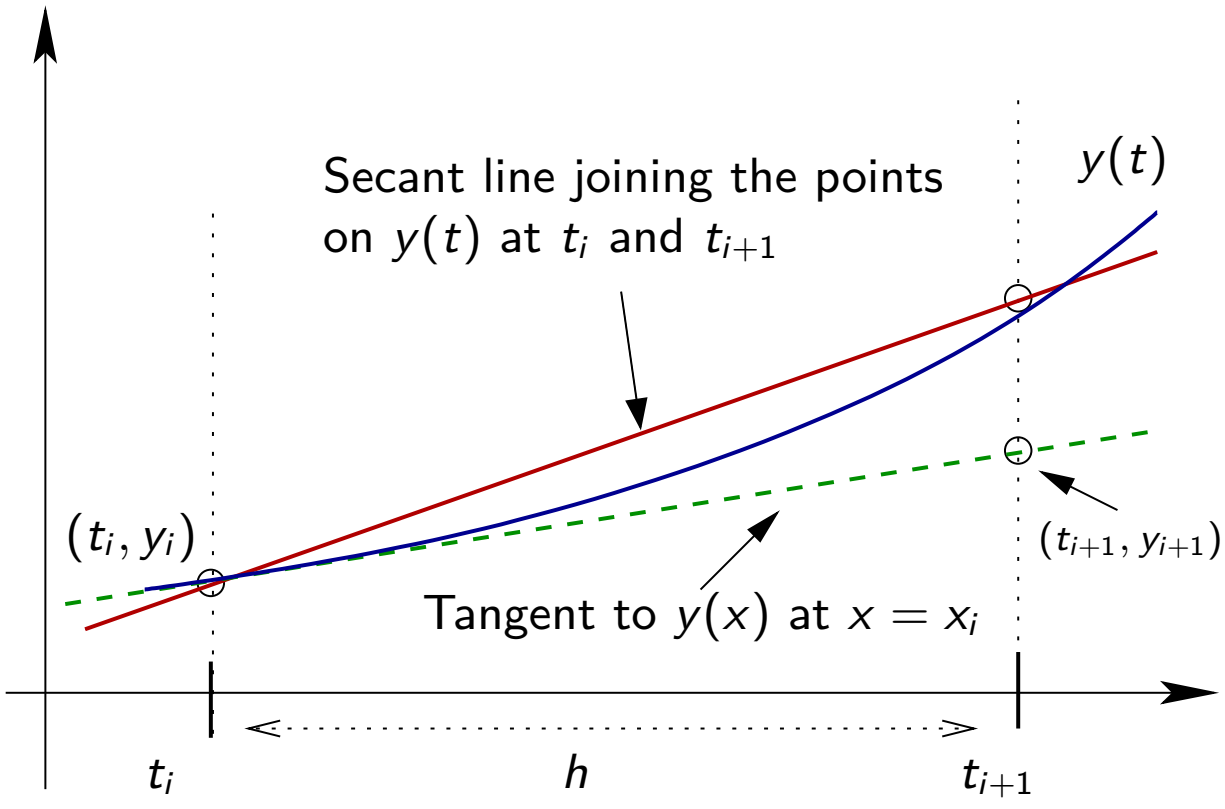
$$t_{i+1} - t_i = \frac{t_n - t_0}{n} = h.$$

The simplest method is ***Euler's Method***. We motivate it as follows.

Motivation

Suppose we know $y(t_i)$, and want to compute $y(t_{i+1})$. From the differential equation we can calculate the slope of the tangent to y at t_i. If this approximates the slope of the line joining $(t_i, y(t_i))$ and $(t_{i+1}, y(t_{i+1}))$, then

$$y'(t_i) = f(t_i, y(t_i)) \approx \frac{y_{i+1} - y_i}{t_{i+1} - t_i}.$$


Secant line joining the points
on $y(t)$ at t_i and t_{i+1}
$y(t)$
(t_i, y_i)
(t_{i+1}, y_{i+1})
Tangent to $y(x)$ at $x = x_i$
t_i
h
t_{i+1}

Euler's Method

Choose equally spaced points $t_0, t_1, \ldots, t_n$ so that

$$t_i - t_{i-1} = h = (t_n - t_0)/n \quad \text{for } i = 0, \ldots, n-1.$$

We call h the "time step". Let y_i denote the approximation for $y(t)$ at $t = t_i$. Set

$$y_{i+1} = y_i + hf(t_i, y_i), \quad i = 0, 1, \ldots, n-1. \tag{3}$$

Example 2.4

Taking $h = 1$, estimate $y(4)$ where

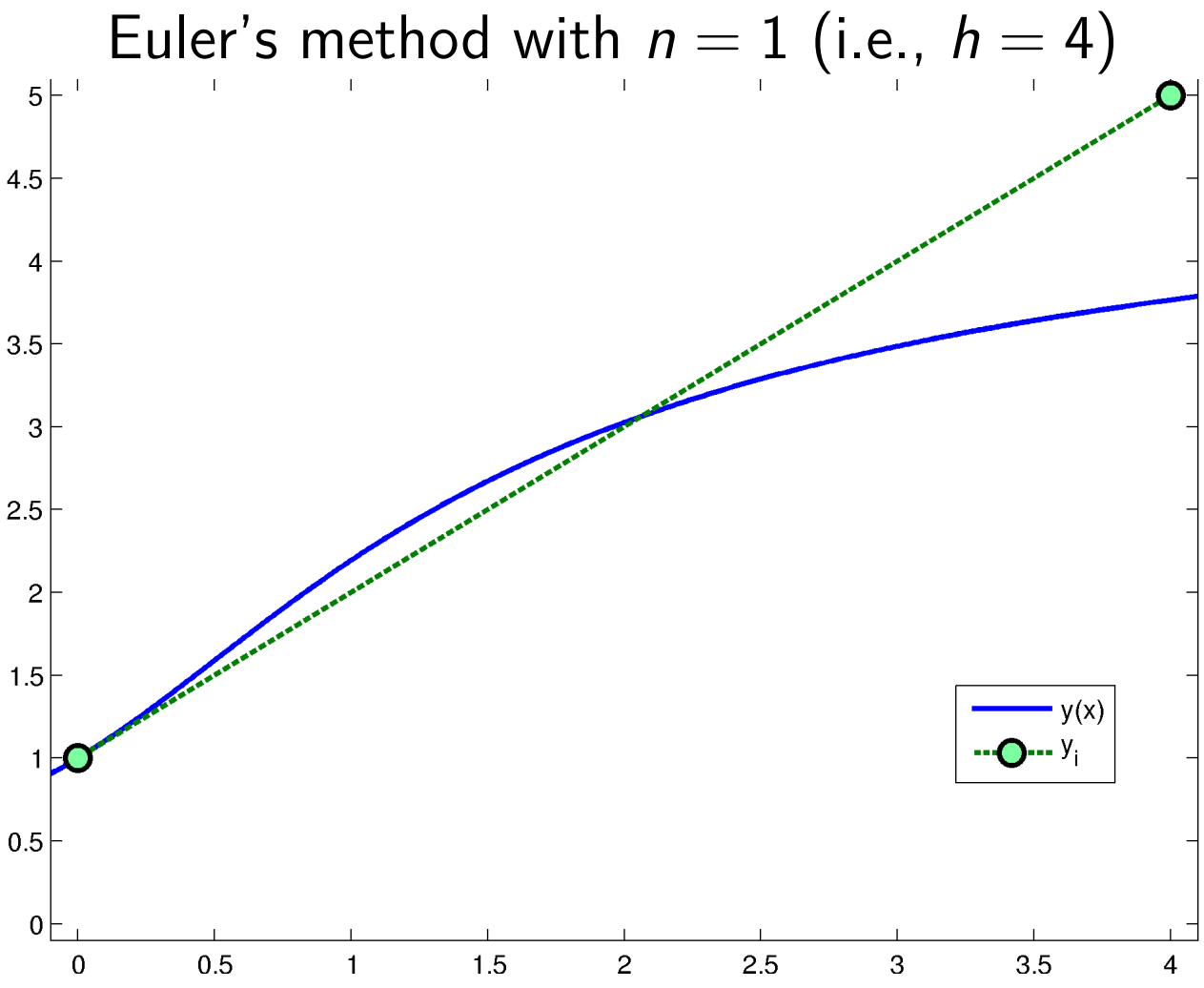
$$y'(t) = y/(1 + t^2), \qquad y(0) = 1.$$

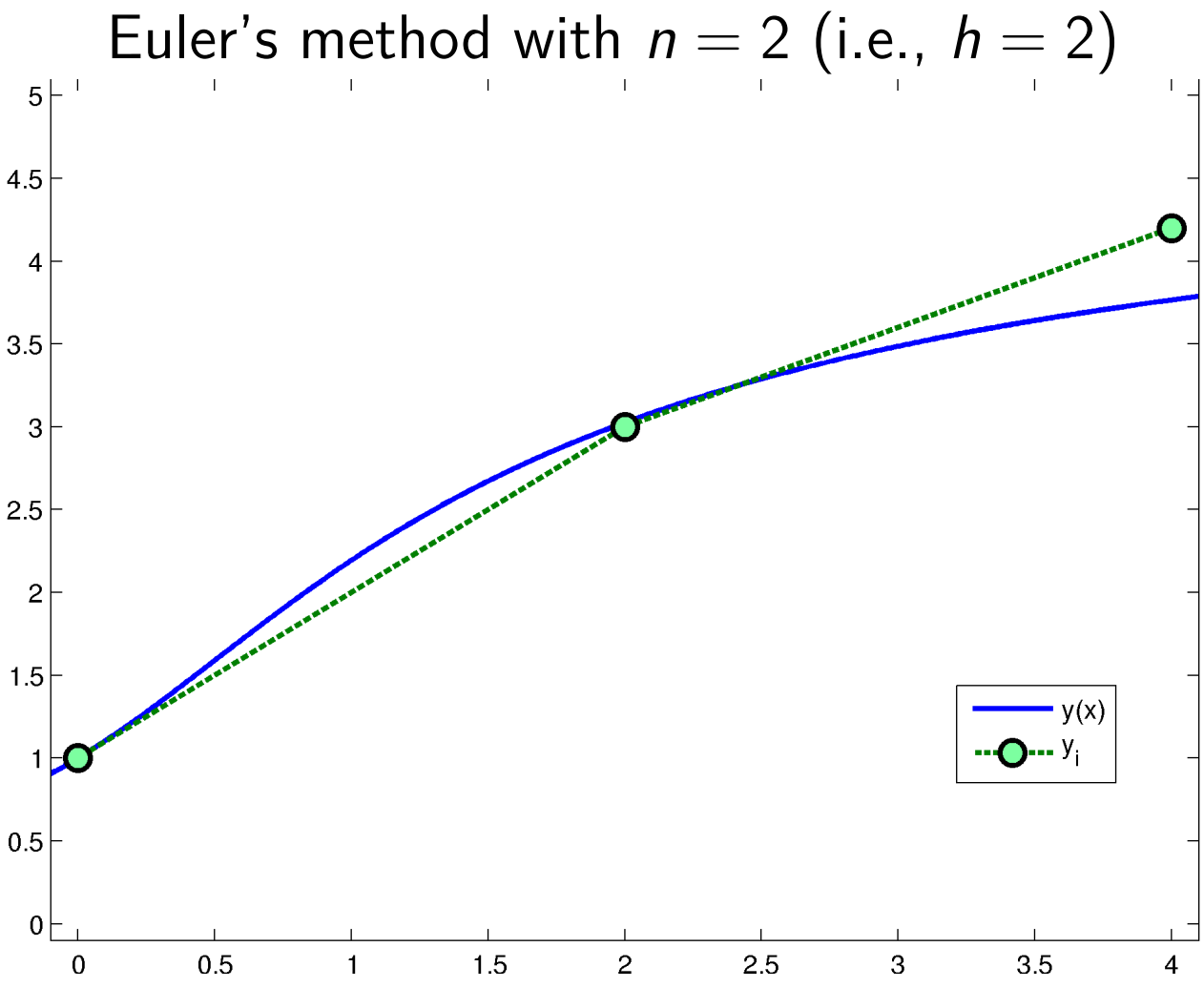
Choosing $h = 1$ we get

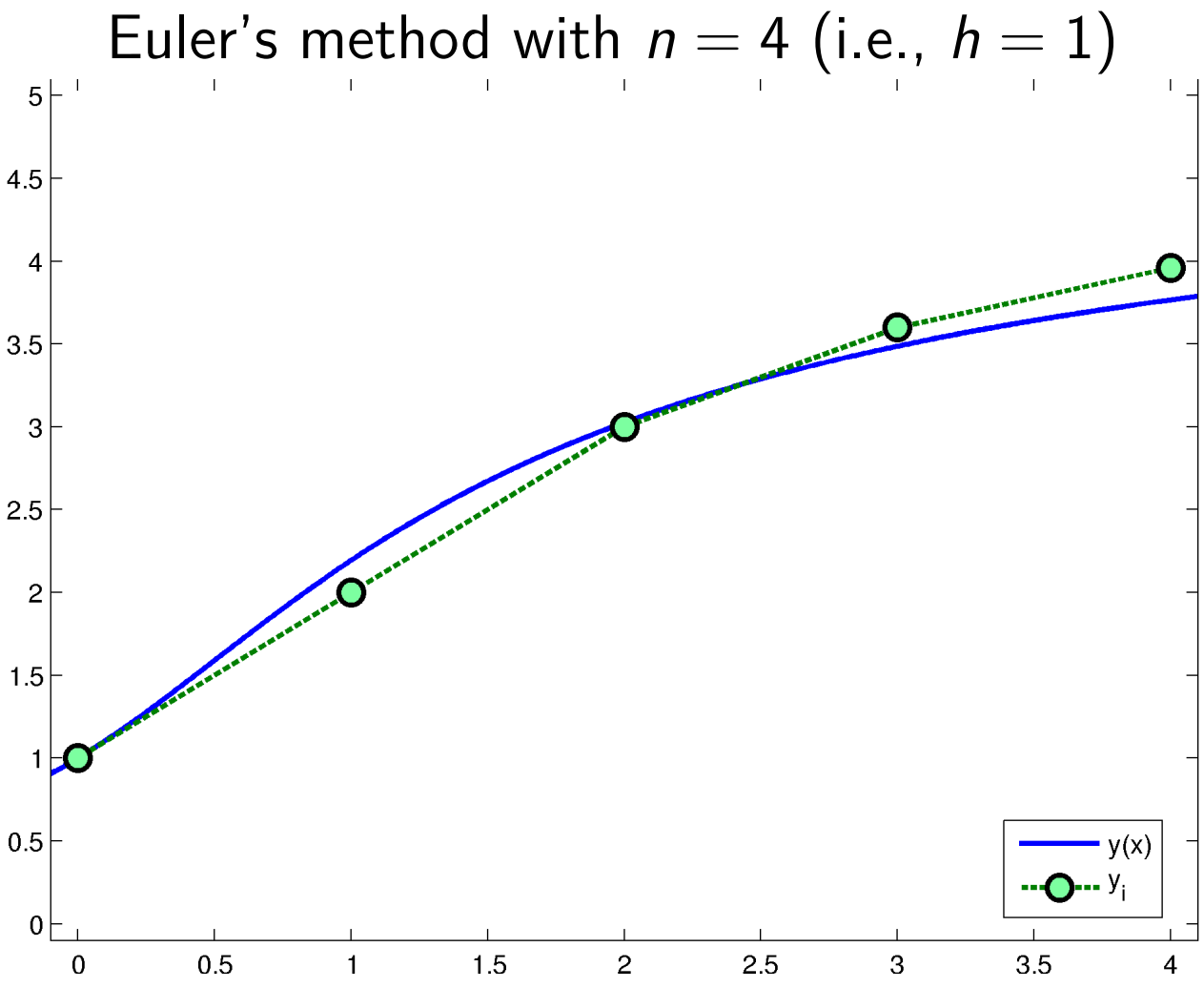
- $i = 0$: $t_0 = 0$, $y_0 = 1$.
- $i = 1$: $t_1 = t_0 + h = 1$.
  $y_1 = y_0 + hf(t_0, y_0) = 1 + \frac{1}{1+0^2} = 2$.
- $i = 2$: $t_2 = t_0 + 2h = 2$.
  $y_2 = y_1 + hf(t_1, y_1) = 2 + 1\frac{2}{1+1^2} = 3$.
- $i = 3$: $t_3 = t_0 + 3h = 3$.
  $y_3 = y_2 + hf(t_2, y_2) = 3 + 1\frac{3}{1+2^2} = 3.6$
- $i = 4$: $t_n = t_4 = t_0 + 4h = 4$.
  $y_n = y_4 = y_3 + hf(t_3, y_3) = 3.6 + \frac{3.6}{1+3^2} = \mathbf{3.96}$

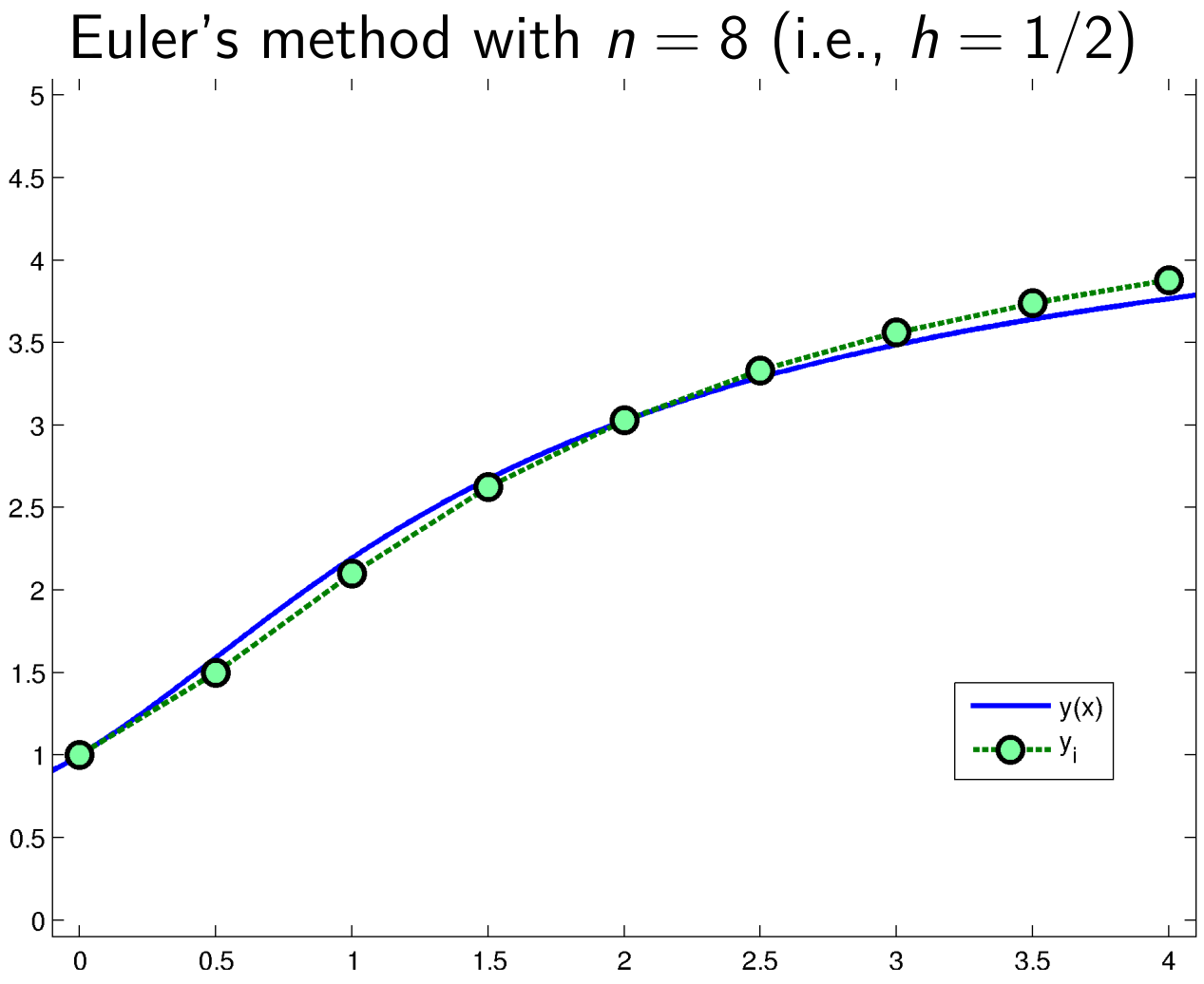
If we had chosen $h = 4$ we would have only required one step: $y_n = y_0 + 4f(t_0, y_0) = \mathbf{5}$. However, this would not be very accurate.

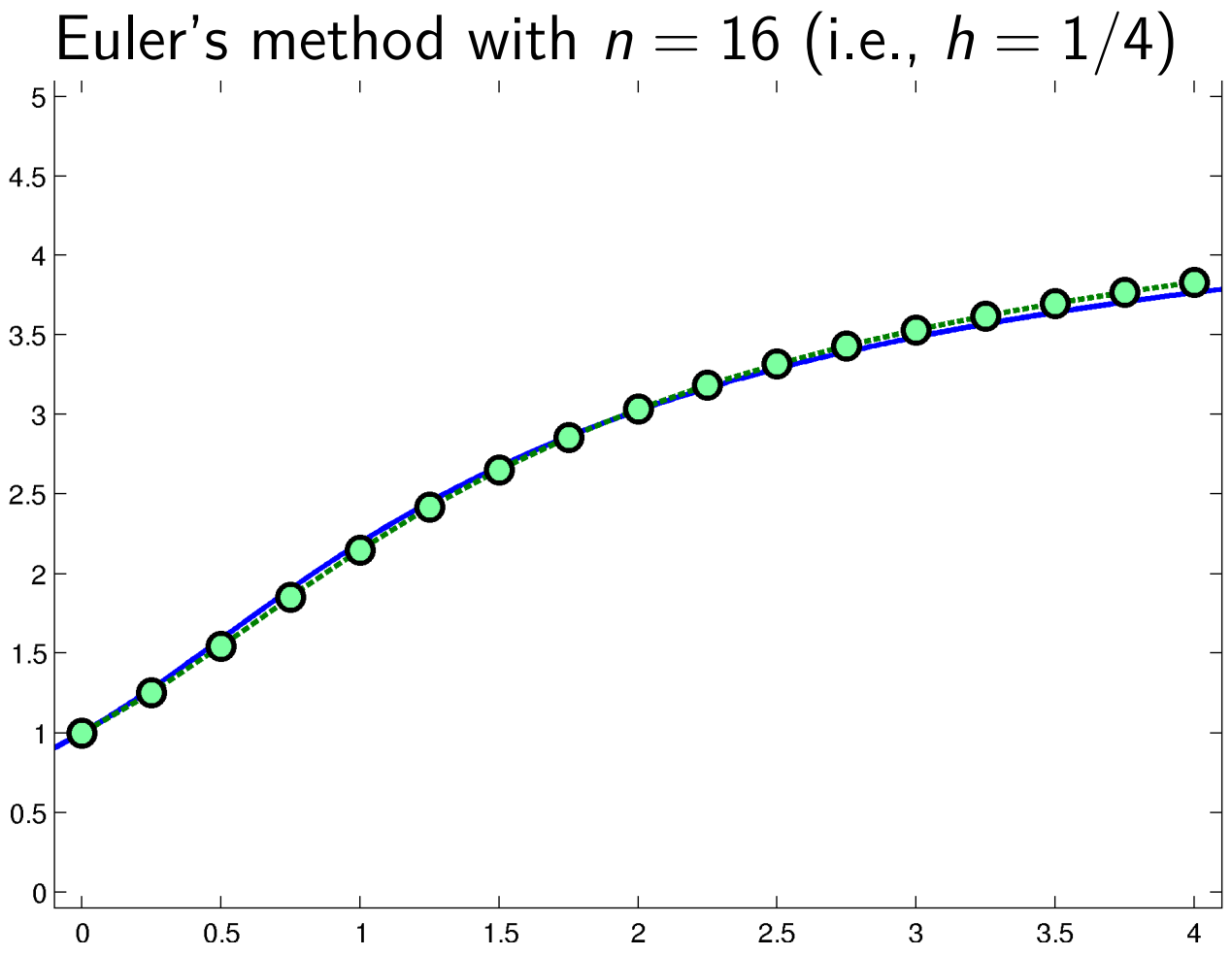
With a little work one can show that the solution to this problem is $y(t) = e^{\tan^{-1}(t)}$ and so $y(4) = 3.7652$. Hence the computed solution with $h = 1$ is much more accurate than the computed solution when $h = 4$. This is also demonstrated in next figure below, and in the follow table, where we see that the error seems to be proportional to h.

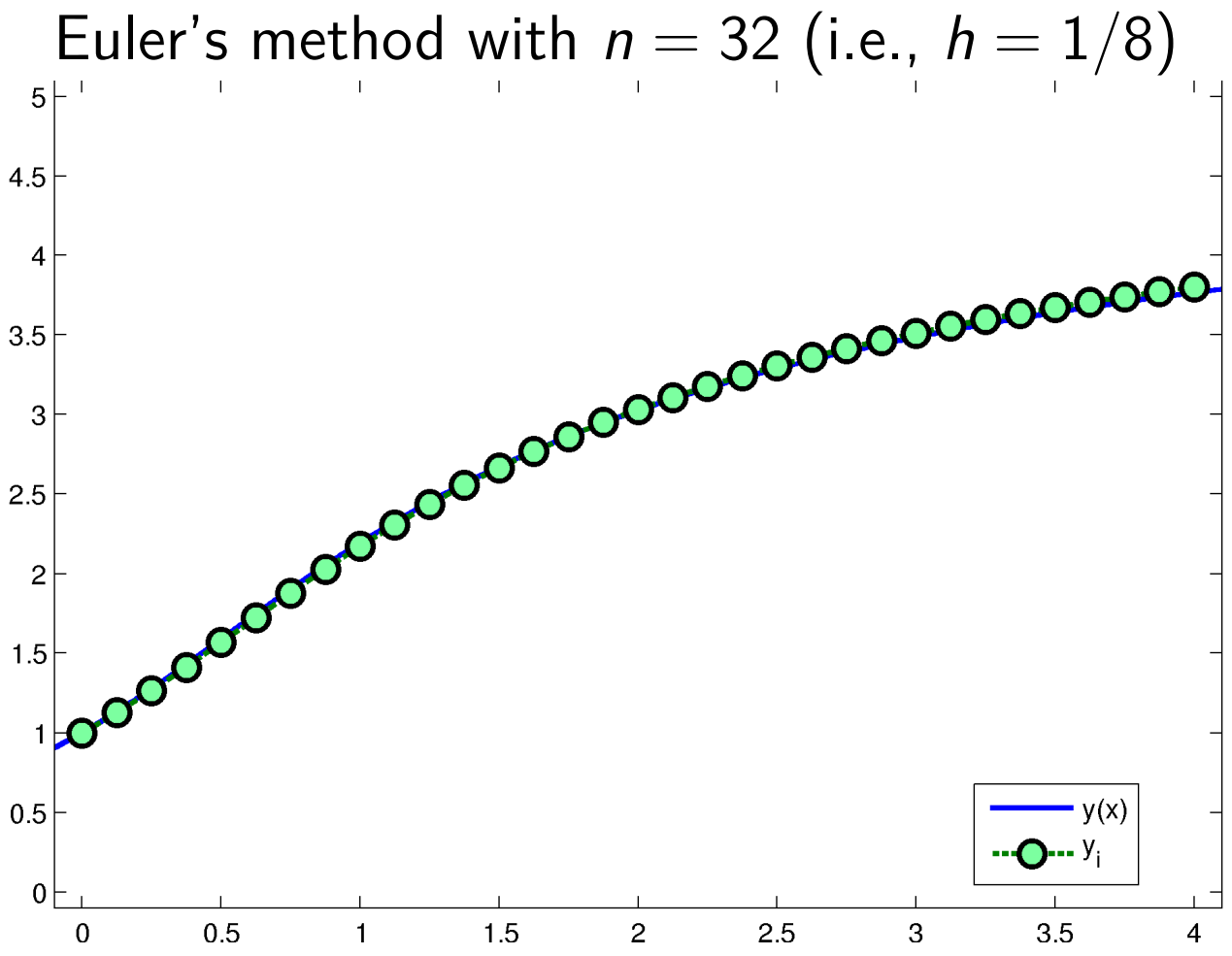

Euler's method with $n = 1$ (i.e., $h = 4$)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
y(x)
y_i


Euler's method with $n = 2$ (i.e., $h = 2$)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
y(x)
y_i


Euler's method with $n = 4$ (i.e., $h = 1$)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
y(x)
y_i


Euler's method with $n = 8$ (i.e., $h = 1/2$)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
y(x)
y_i


Euler's method with $n = 16$ (i.e., $h = 1/4$)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4


Euler's method with $n = 32$ (i.e., $h = 1/8$)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
y(x)
y_i

n	h	y_n	$\lvert y(t_n) - y_n \rvert$
1	4	5.0	1.235
2	2	4.2	0.435
4	1	3.960	0.195
8	1/2	3.881	0.115
16	1/4	3.831	0.065
32	1/8	3.800	0.035

Table: Error in Euler's method for Example 2.4

Exercise 2.3

As a special case in which the error of Euler's method can be analysed directly, consider Euler's method applied to

$$y'(t) = y(t), \quad y(0) = 1.$$

The true solution is $y(t) = e^t$.

(i) Show that the solution to Euler's method can be written as

$$y_i = (1 + h)^{t_i/h}, \quad i \geq 0.$$

(ii) Show that

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

This then shows that, if we denote by $y_n(T)$ the approximation for $y(T)$ obtained using Euler's method with n intervals between t_0 and T, then

$$\lim_{n \to \infty} y_n(T) = e^T.$$

Hint: Let $w = (1 + h)^{1/h}$, so that $\log w = (1/h) \log(1 + h)$. Now use l'Hospital's rule to find $\lim_{h \to 0} w$.

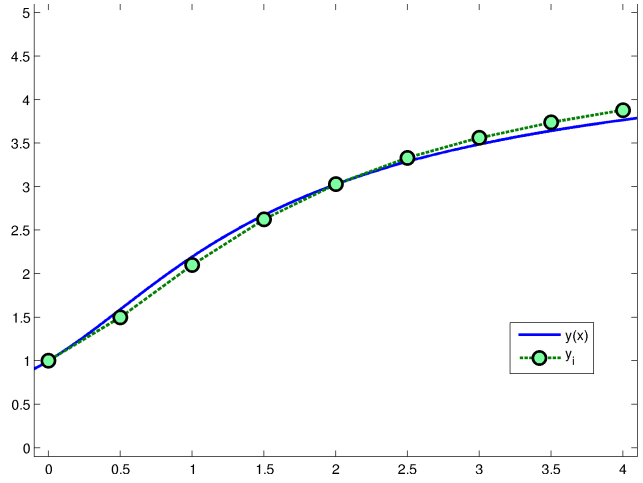
Initial Value Problems

§2.3: Error Analysis of one-step methods

(but mainly of Euler's Method)

MA385/530 – Numerical Analysis 1

October 2019



Euler's method is an example of a **one-step methods**, which have the *general* form:

$$y_{i+1} = y_i + h\Phi(t_i, y_i; h). \tag{4}$$

To get Euler's method, just take $\Phi(t_i, y_i; h) = f(t_i, y_i)$.

In the introduction, we motivated Euler's method with a geometrical argument. An alternative, more mathematical way of deriving Euler's Method is to use a *Truncated Taylor Series*:

This not only motivates Euler's formula, but also suggests that at each step the method introduces a (local) error of $h^2 y''(\eta)/2$. (More of this later).

Definition 2.5

Global Error: $\mathcal{E}_i = y(t_i) - y_i$.

Definition 2.6

Truncation Error:

$$T_i := \frac{y(t_{i+1}) - y(t_i)}{h} - \Phi(t_i, y(t_i); h). \tag{5}$$

It can be helpful to think of T_i as representing how much the difference equation differs from the differential equation. For Euler's method, it can be determined using a Taylor Series.

The relationship between the global error and truncation errors is explained in the following (important!) result (also, compare with Picard's Theorem).

Theorem 2.7 (Thm 12.1 in Süli & Mayers)

Let $\Phi()$ be *Lipschitz* with constant L. Then

$$|\mathcal{E}_n| \leq T\left(\frac{\mathrm{e}^{L(t_n - t_0)} - 1}{L}\right), \tag{6}$$

where $T = \max_{i=0,1,\ldots n} |T_i|$.

For Euler's method, we get

$$T = \max_{0 \le j \le n} |T_j| \le \frac{h}{2} \max_{t_0 \le t \le t_n} |y''(t)|.$$

Example 2.8

Given the problem:

$$y' = 1 + t + \frac{y}{t} \quad \text{for } t > 1; \quad y(1) = 1,$$

find an approximation for $y(2)$.

(i) Give an upper bound for the global error taking $n = 4$ (i.e., $h = 1/4$)

(ii) What n should you take to ensure that the global error is no more that 0.1?

To answer these questions we need to use (6), which requires that we find L and an upper bound for T. In this instance, L is easy:

To find T we need an upper bound for $|y''(t)|$ on $[1, 2]$, even though we don't know $y(t)$...

With these values of L and T, using (6) we find $\mathcal{E}_n \leq 0.644$. In fact, the true answer is 0.43, so we see that (6) is somewhat pessimistic.

. . . . . .

To answer (ii): *What n should you take to ensure that the global error is no more that 0.1?* (We should get $n = 26$. This is not that sharp: $n = 19$ will do).

We are often interested in the *convergence* of a method. That is, is it true that

$$\lim_{h \to 0} y_n = y(t_n)?$$

Or equivalently that,

$$\lim_{h \to 0} \mathcal{E}_n = 0?$$

Given that the global error for Euler's method can be bounded:

$$|\mathcal{E}_n| \leq h \frac{\max |y''(t)|}{2L} \left(\mathrm{e}^{L(t_n - t_0)} - 1 \right) = hK, \tag{7}$$

we can say it converges.

So now we know, for Euler's method, that $y_n \to y(t_n)$ as $n \to \infty$, but how quickly?

Definition 2.9

The **order of accuracy** of a numerical method is p if there is a constant K so that

$$|\mathcal{E}_n| \leq Kh^p.$$

So Euler's method is first-order.

The term **order of convergence** is often use instead of **order of accuracy**.

One of the requirements for convergence is *Consistency*:

Definition 2.10

A one-step method $y_{n+1} = y_n + h\Phi(t_n, y_n; h)$ is *consistent* with the differential equation $y'(t) = f(t, y(t))$ if $f(t, y) \equiv \Phi(t, y; 0)$.

Next we'll try to develop methods that are of higher order than Euler's method; that is that we can show

$$|\mathcal{E}_n| \leq Kh^p \quad \text{for some} \quad p > 1.$$

Suppose we numerically solve some differential equation and estimate the error. If we think this error is too large we could redo the calculation with a smaller value of h. Or we could use a better method, for example ***Runge-Kutta*** methods. These are high-order methods that rely on evaluating $f(t, y)$ a number of times at each step in order to improve accuracy.

We'll first motivate one such method and then later look at the general framework.

The goal will be to develop some techniques to help us derive our own methods for accurately solving IVPs. Rather than using formal theory, we will reason based on carefully chosen examples.

Exercise 2.4

An important step in the proof of Theorem 2.3.3, but which we didn't do in class, requires the observation that if $|\mathcal{E}_{i+1}| \leq |\mathcal{E}_i|(1 + hL) + h|T_i|$, then

$$|\mathcal{E}_i| \leq \frac{T}{L}\left[(1 + hL)^i - 1\right] \qquad i = 0, 1, \ldots, N.$$

Use induction to show that is indeed the case.

Exercise 2.5

Suppose we use Euler's method to find an approximation for $y(2)$, where y solves

$$y(1) = 1, \qquad y' = (t - 1)\sin(y).$$

(i) Give an upper bound for the global error taking $n = 4$ (i.e., $h = 1/4$).

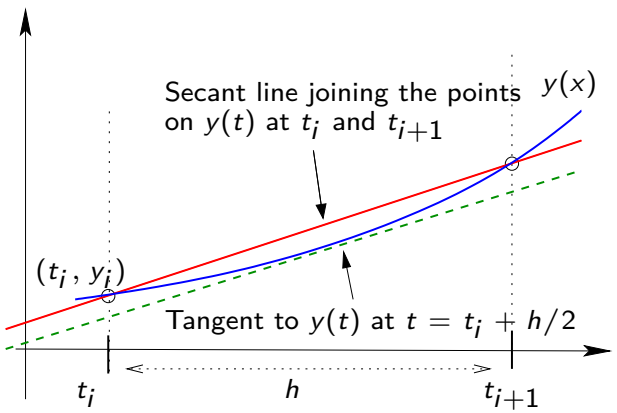
(ii) What n should you take to ensure that the global error is no more that 10^{-3}?

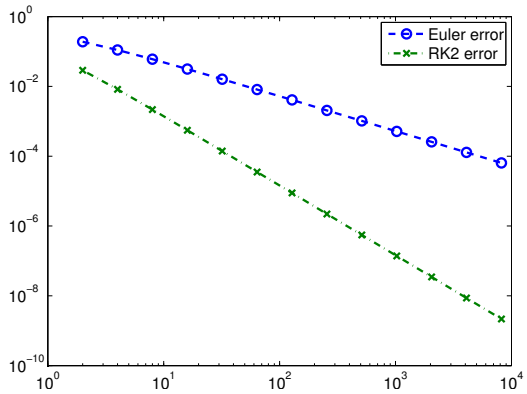
Initial Value Problems

§2.4 Runge-Kutta 2 (RK2)

MA385/530 – Numerical Analysis 1

October 2019





Recall our original motivation of Euler's method: use the slope of the tangent to y at t_i as an approximation for the slope of the secant line joining the points $(t_i, y(t_i))$ and $(t_{i+1}, y(t_{i+1}))$.

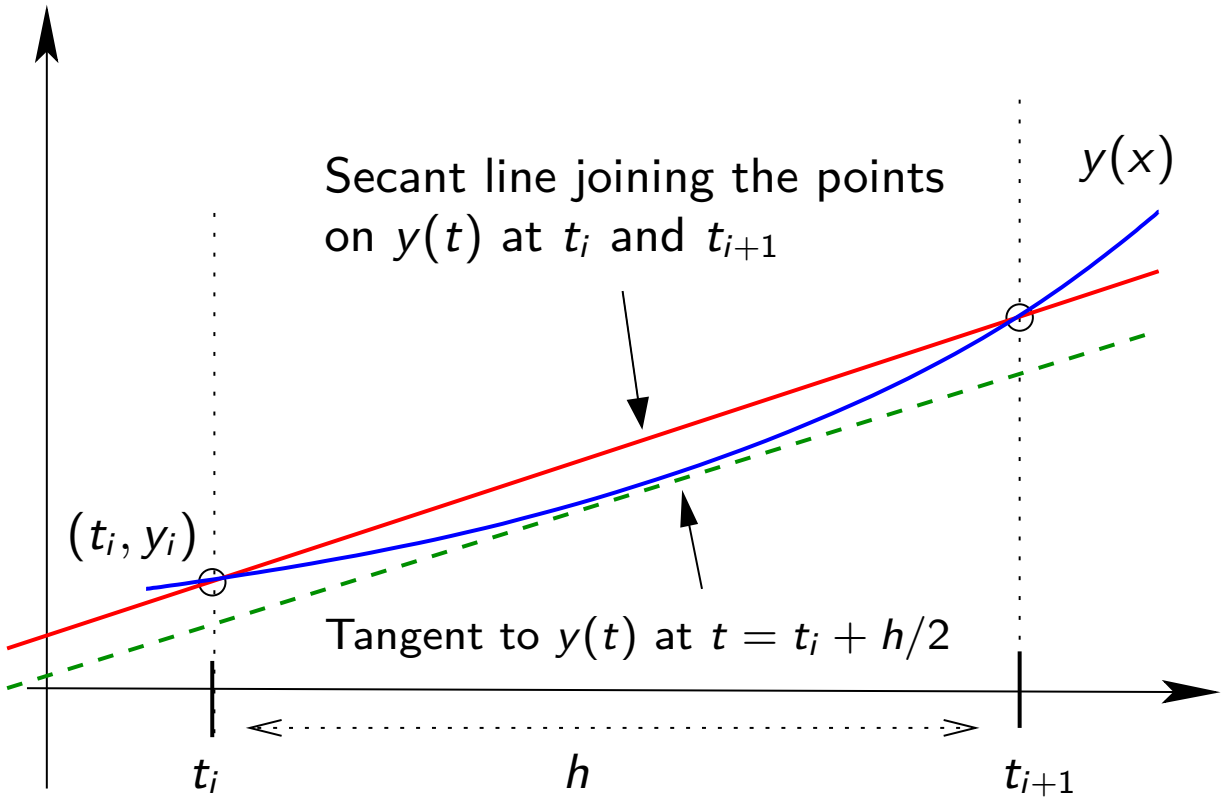
One could argue, given the diagram on the next slide, that the slope of the tangent to y at $t = (t_i + t_{i+1})/2 = t_i + h/2$ would be a better approximation. This would give

$$y(t_{i+1}) \approx y_i + hf\left(t_i + \frac{h}{2}, y(t_i + \frac{h}{2})\right). \tag{8}$$

However, we don't know $y(t_i + h/2)$, but can approximate it using Euler's Method: $y(t_i + h/2) \approx y_i + (h/2)f(t_i, y_i)$.

Modified (Midpoint) Euler's Method

$$y_{i+1} = y_i + hf\left(t_i + \frac{h}{2}, y_i + \frac{h}{2} f(t_i, y_i)\right). \tag{9}$$



Example 2.11

Use the Modified Euler Method to approximate $y(1)$ where

$$y(0) = 1, \qquad y'(t) = y \log(1 + t^2).$$

This has the solution $y(t) = (1 + t^2)^t \exp(-2t + 2 \tan^{-1} t)$.

	Euler		Modified	
n	$\mathcal{E}_n$	$\mathcal{E}_n/\mathcal{E}_{n-1}$	$\mathcal{E}_n$	$\mathcal{E}_n/\mathcal{E}_{n-1}$
1	3.02e-01		7.89e-02	
2	1.90e-01	1.59	2.90e-02	2.72
4	1.11e-01	1.72	8.20e-03	3.54
8	6.02e-02	1.84	2.16e-03	3.79
16	3.14e-02	1.91	5.55e-04	3.90
32	1.61e-02	1.95	1.40e-04	3.95
64	8.13e-03	1.98	3.53e-05	3.98
128	4.09e-03	1.99	8.84e-06	3.99

Clearly we get a much more accurate result using the Modified Euler Method. Even more importantly, we get a higher *order of accuracy*: if h is reduced by a factor of **two**, the error in the Modified method is reduced by a factor of **four**.

We can also make a direct comparison of the two methods by using a log-log plot of the errors.

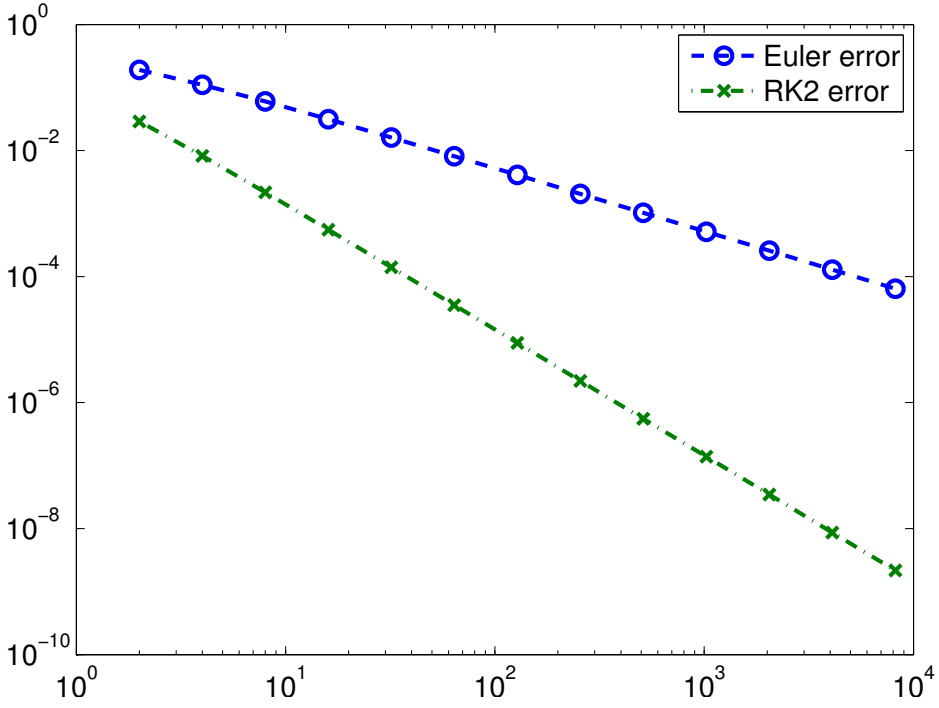


Figure: Log-log plot of the errors when Euler's and Modified Euler's methods are used to solve Example 2.11

The "*Modified Euler Method*" is an example of one of the (large) family of 2^{nd}-*order Runge-Kutta* (RK2). Recall that that one-step methods are written as $y_{i+1} = y_i + h\Phi(t_i, y_i; h)$

The general RK2 method is

$$\begin{aligned} k_1 &= f(t_i, y_i) \qquad k_2 = f(t_i + \alpha h, y_i + \beta h k_1). \\ \Phi(t_i, y_i; h) &= (ak_1 + bk_2) \end{aligned} \tag{10}$$

Example: take $a = 1, b = 0$.

The general RK2 method is

$$k_1 = f(t_i, y_i) \qquad k_2 = f(t_i + \alpha h, y_i + \beta h k_1).$$
$$y_{i+1} = y_i + h(a k_1 + b k_2)$$

Example 2: take $\alpha = \beta = 1/2, a = 0, b = 1$.

Our aim now is to deduce general rules for choosing a, b, α and β. We'll see that if we pick any one of these four parameters, then the requirement that the method be consistent and second-order determines the other three.

By demanding that RK2 be ***consistent*** we get that $a + b = 1$.

Next we need to know how to choose α and β. The formal way is to use a two-dimensional Taylor series expansion. It is quite technical. (FYI, detailed will be posted as an appendix to these notes). Instead we'll take a less rigorous, ***heuristic*** approach.

(From Wikipedia: "A heuristic technique (Ancient Greek: 'find' or 'discover'), often called simply a heuristic, is any approach to problem solving, learning, or discovery that employs a practical method, not guaranteed to be optimal, perfect, logical, or rational, but instead sufficient for reaching an immediate goal.")

Because we expect that, for a second order accurate method, $|\mathcal{E}_n| \leq Kh^2$ where K depends on $y'''(t)$, if we choose a problem for which $y'''(t) \equiv 0$, we expect no error...

In the above example, the right-hand side of the differential equation, $f(t, y)$, depended only on t. Now we'll try the same trick: using a problem with a simple known solution (and zero error), but for which f depends explicitly on y.

Consider the DE $y(1) = 1, y'(t) = y(t)/t$. It has a simple solution: $y(t) = t$. We now use that any RK2 method should be exact for this problem to deduce that $\alpha = \beta$.

Now we collect the above results all together and show that the second-order Runge-Kutta (RK2) methods are:

$$y_{i+1} = y_i + h(ak_1 + bk_2)$$

$$k_1 = f(t_i, y_i), \qquad k_2 = f(t_i + \alpha h, y_i + \beta h k_1),$$

where we choose any $b \neq 0$ and then set

$$a = 1 - b, \qquad \alpha = \frac{1}{2b}, \qquad \beta = \alpha.$$

It is easy to verify that the Modified method satisfies these criteria.

Exercise 2.6

A popular RK2 method, called the *Improved Euler Method*, is obtained by choosing $\alpha = 1$.

(i) Use the Improved Euler Method to find an approximation for $y(4)$ when

$$y(0) = 1, \qquad y' = y/(1 + t^2),$$

taking $n = 2$. (If you wish, use MATLAB.)

(ii) Using a diagram similar to the one in Figure 1 for the Modified Euler Method, justify the assertion that the Improved Euler Method is more accurate than the basic Euler Method.

(iii) Show that the method is consistent.

(iv) Write out what this method would be for the problem: $y'(t) = \lambda y$ for a constant λ. How does this relate to the Taylor series expansion for $y(t_{i+1})$ about the point t_i?

Exercise 2.7

In his seminal paper of 1901, Carl Runge gave an example of what we now call a *Runge-Kutta 2 method*, where

$$\Phi(t_i, y_i; h) = \frac{1}{4} f(t_i, y_i) + \frac{3}{4} f\left(t_i + \frac{2}{3}h, y_i + \frac{2}{3}hf(t_i, y_i)\right).$$

(i) Show that it is consistent.

(ii) Show how this method fits into the general framework of RK2 methods. That is, what are a, b, α, and β? Do they satisfy the following conditions?

$$\beta = \alpha, \qquad b = \frac{1}{2\alpha}, \qquad a = 1 - b. \tag{11}$$

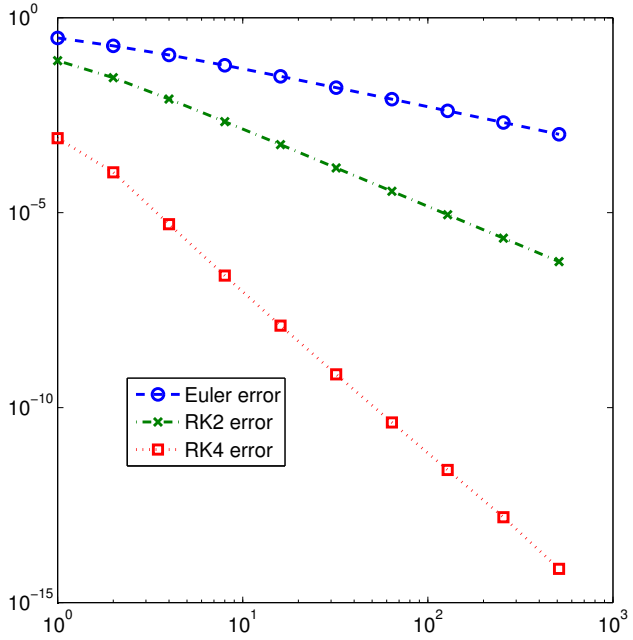
(iii) Use it to estimate the solution at the point $t = 2$ to $y(1) = 1$, $y' = 1 + t + y/t$ taking $n = 2$ time steps.

Initial Value Problems

§2.5 Runge-Kutta 4

MA385/530 – Numerical Analysis 1

October 2019



It is possible to construct methods that have higher orders of accuracy than **RK2** methods. Of these, the most used are probably those that belong to the **Runge-Kutta 4 (RK4)** family, and have the property that

$$|y(t_n) - y_n| \leq Ch^4.$$

However, even writing down the general form of the RK4 method, and then deriving conditions on the parameters is rather complicated. Therefore, we'll focus on just one RK4 method, and use examples, rather than theory, to demonstrate that it is 4th-order.

"The RK4 Method"

$$
\begin{aligned}
k_1 &= f(t_i, y_i), \\
k_2 &= f(t_i + \frac{h}{2}, y_i + \frac{h}{2}k_1), \\
k_3 &= f(t_i + \frac{h}{2}, y_i + \frac{h}{2}k_2), \\
k_4 &= f(t_i + h, y_i + hk_3), \\
y_{i+1} &= y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4).
\end{aligned}
$$

The RK4 method can be interpreted as follows :

As the following example shows, RK4 can be much more accurate than the Euler or RK2 methods for small h (i.e., large n). For the RK4, doubling n reduces the error by a factor of 8 (compared with 2 and 4 for the Euler and RK2 methods, respectively).

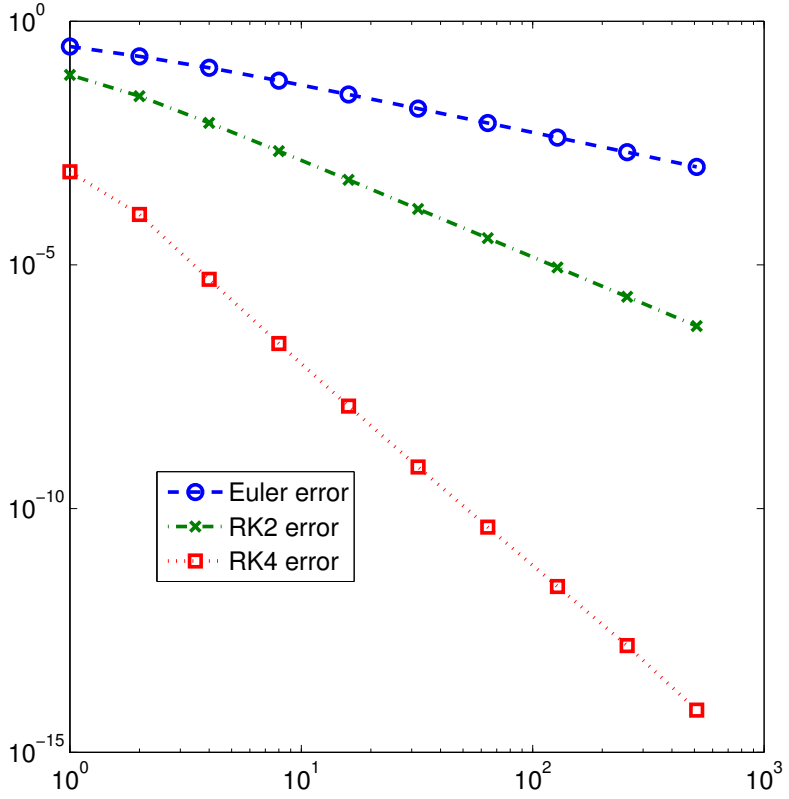
Example 2.12 (2.11 (again))

Compare Euler, Modified Euler, and RK4 for approximating $y(1)$ *where:* $y(0) = 1, \quad y'(t) = y \log(1 + t^2)$.

	Error: $\lvert y(t_n) - y_n \rvert$		
n	Euler	Modified	RK4
1	3.02e-01	7.89e-02	8.14e-04
2	1.90e-01	2.90e-02	1.08e-04
4	1.11e-01	8.20e-03	5.07e-06
8	6.02e-02	2.16e-03	2.44e-07
16	3.14e-02	5.55e-04	1.27e-08
32	1.61e-02	1.40e-04	7.11e-10
64	8.13e-03	3.53e-05	4.18e-11
128	4.09e-03	8.84e-06	2.53e-12
256	2.05e-03	2.21e-06	1.54e-13
512	1.03e-03	5.54e-07	7.33e-15

Example 2.12 (2.11 (again))

Compare Euler, Modified Euler, and RK4 for approximating $y(1)$ *where:* $y(0) = 1, \quad y'(t) = y \log(1 + t^2)$.



Although we won't do a detailed analysis of RK4, we can do a little. In particular, we would like to show it is

(i) consistent,

(ii) convergent and fourth-order, at least for some examples.

Example 2.13

It is easy to see that RK4 is consistent:

Example 2.14

In general, showing the rate of convergence is tricky. Instead, we'll demonstrate how the method relates to a Taylor Series expansion for the problem $y' = \lambda y$ where λ is a constant.

§2.5.2 Consistency and convergence of RK4 (59/84)

Many (seemingly different) RK have been proposed and studied. A unified approach of representing them was developed by John Butcher: write an s-stage method as

$$\Phi(t_i, y_i; h) = \sum_{j=1}^{s} b_j k_j, \qquad \text{where}$$

$$\begin{aligned}
k_1 &= f(t_i + \alpha_1 h, y_i),\\
k_2 &= f(t_i + \alpha_2 h, y_i + \beta_{21} h k_1),\\
k_3 &= f(t_i + \alpha_3 h, y_i + \beta_{31} h k_1 + \beta_{32} h k_2),\\
&\vdots\\
k_s &= f(t_i + \alpha_s h, y_i + \beta_{s1} h k_1 + \ldots \beta_{s,s-1} h k_{s-1}),
\end{aligned}$$

The most convenient way to represent the coefficients is in a tableau:

$$\begin{array}{c|cccccc} \alpha_1 & & & & & & \\ \alpha_2 & \beta_{21} & & & & & \\ \alpha_3 & \beta_{31} & \beta_{32} & & & & \\ \vdots & & & & & & \\ \alpha_s & \beta_{s1} & \beta_{s2} & \cdots & \beta_{s,s-1} & & \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array}$$

The tableaux for basic Euler, Modified Euler, and RK4 are:

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 2/6 & 2/6 & 1/6 \end{array}$$

A Runge Kutta method has s stages if it involves s evaluations of the function f. (That it, its formula features $k_1, k_2, \ldots, k_s$).

We've seen a 1-stage method that is 1^{st}-order.

We studied 2-stage methods that are 2^{nd}-order.

In an exercise, you'll construct a 3-stage method that is 3rd order.

And, of course, we have just considered a four-stage method that is 4^{th}-order.

It is tempting to think that for any s we can get a method of order s using s stages. However, it can be shown that, for example, to get a 5^{th}-order method, you need at least 6 stages; for a 7^{th}-order method, you need at least 9 stages. The theory involved is both intricate and intriguing, and involves aspects of group theory, graph theory, and differential equations. Students in third year might consider this as a topic for their final year project.

Exercise 2.8

We claim that, for $RK4$:

$$|\mathcal{E}_N| = |y(t_N) - y_N| \leq Kh^4.$$

for some constant K. How could you verify that the statement is true using the data of Table 2.3, at least for test problem in Example 2.4.2? Give an estimate for K.

Exercise 2.9

Recall the problem in Example 2.2.2: *Estimate* $y(2)$ *given that*

$$y(1) = 1, \qquad y' = f(t, y) := 1 + t + \frac{y}{t},$$

(i) Show that $f(t, y)$ satisfies a Lipschitz condition and give an upper bound for L.

(ii) Use Euler's method with $h = 1/4$ to estimate $y(2)$. Using the true solution, calculate the error.

(iii) Repeat this for the RK2 method of your choice (with $a \neq 0$) taking $h = 1/2$.

(iv) Use $RK4$ with $h = 1$ to estimate $y(2)$.

Exercise 2.10

Here is the tableau for a three stage Runge-Kutta method:

$$\begin{array}{c|ccc} 0 & & & \\ \alpha_2 & 1/2 & & \\ 1 & \beta_{31} & 2 & \\ \hline & 1/6 & b_2 & 1/6 \end{array}$$

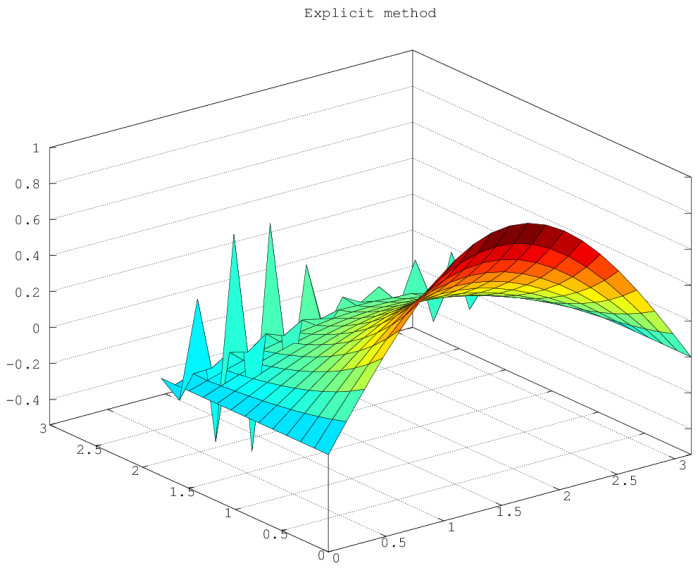
(i) Use that the method is consistent to determine b_2.

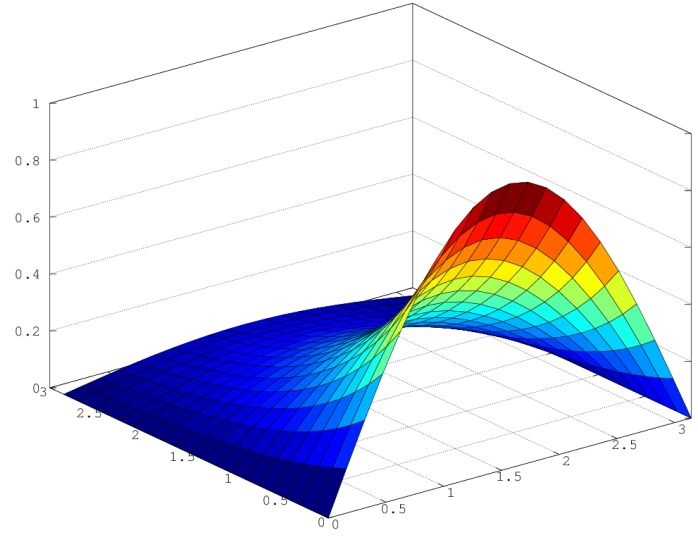
(ii) The method is exact when used to compute the solution to

$$y(0) = 0, \quad y'(t) = 2t, \ t > 0.$$

Use this to determine α_2.

(iii) The method should agree with an appropriate Taylor series for the solution to $y'(t) = \lambda y(t)$, up to terms that are $\mathcal{O}(h^3)$. Use this to determine β_{31}.





Initial Value Problems

§2.6 From IVPs to linear systems

MA385 – Numerical Analysis 1

November 2019

In this final section, we highlight some of the many important aspects of the numerical solution of IVPs that are *not* covered in detail in this course:

- Systems of ODEs;
- Higher-order equations;
- Implicit methods; and
- Problems in two dimensions.

We have the additional goal of seeing how these methods related to the earlier section of the course (nonlinear problems) and next section (linear equation solving).

So far we have solved only single IVPs. However, must interesting problems are coupled systems: find functions y and z such that

$$y'(t) = f_1(t, y, z),$$

$$z'(t) = f_2(t, y, z).$$

This does not present much of a problem to us. For example the Euler Method is extended to

$$y_{i+1} = y_i + hf_1(t, y_i, z_i),$$

$$z_{i+1} = z_i + hf_2(t, y_i, z_i).$$

Example 2.15

In pharmokinetics, the flow of drugs between the blood and major organs can be modelled

$$\frac{dy}{dt}(t) = k_{21}z(t) - (k_{12} + k_{\text{elim}})y(t).$$
$$\frac{dz}{dt}(t) = k_{12}y(t) - k_{21}z(t).$$
$$y(0) = d, \quad z(0) = 0.$$

where y is the concentration of a given drug in the blood-stream and z is its concentration in another organ. The parameters k_{21}, k_{12} and k_{elim} are determined from physical experiments.

Example 2.16

$$\frac{dy}{dt}(t) = k_{21}z(t) - (k_{12} + k_{\text{elim}})y(t).$$
$$\frac{dz}{dt}(t) = k_{12}y(t) - k_{21}z(t).$$
$$y(0) = d, \quad z(0) = 0.$$

Euler's method for this is:

$$y_{i+1} = y_i + h\big(- (k_{12} + k_{\text{elim}})y_i + k_{21}z_i\big),$$
$$z_{i+1} = z_i + h\big(k_{12}y_i + k_{21}z_i\big).$$

So far we've only considered first-order initial value problems. However, the methods can easily be extended to high-order problems:

$$y''(t) + a(t)y'(t) = f(t, y); \quad y(t_0) = y_0, y'(t_0) = y_1.$$

We do this by converting the problem to a system: set $z(t) = y'(t)$. Then:

$$\begin{aligned} z'(t) &= -a(t)z(t) + f(t, y), & z(t_0) &= y_1, \\ y'(t) &= z(t), & y(t_0) &= y_0. \end{aligned}$$

Now apply any one-step method to this system:

Example 2.17

Transform the following 2nd-order IVP as a system of 1st order problems, and write down the Euler method for the resulting problem:

$$y''(t) - 3y'(t) + 2y(t) + e^t = 0,$$

$$y(1) = e,\ y'(1) = 2e.$$

Although we won't dwell on the point, there are many problems for which the one-step methods we have seen will give a useful solution only when the step size, h, is small enough. For larger h, the solution can be very unstable.

Such problems are called "stiff" problems. They can be solved, but are best done with so-called "implicit methods", the simplest of which is the Implicit Euler Method:

$$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1}).$$

Note that y_{i+1} appears on both sizes of the equation. To implement this method, we need to be able to solve this non-linear problem. The most common method for doing this is Newton's method.

So far, in MA385/530, we've only considered *ordinary* differential equations: these are DEs which involve functions of just one variable. In our examples above, this variable was time.

However, many physical phenomena vary in space and time, and so the solutions to the differential equations the model them depend on two or more variables. The derivatives expressed in the equations are *partial derivatives* and so they are called *partial differential equations* (PDEs).

We will take a brief look at how to solve these (and how not to solve them). This will motivate the following section, on solving systems of linear equations.

Recall (again) the Black-Scholes equations for pricing an option:

$$\frac{\partial V}{\partial t} - \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} - rS\frac{\partial V}{\partial S} + rV = 0.$$

With a little effort, (see, e.g., Chapter 5 of "*The Mathematics of Financial Derivatives: a student introduction*", by Wilmott et al.) this can be transformed to the simpler-looking *heat equation*:

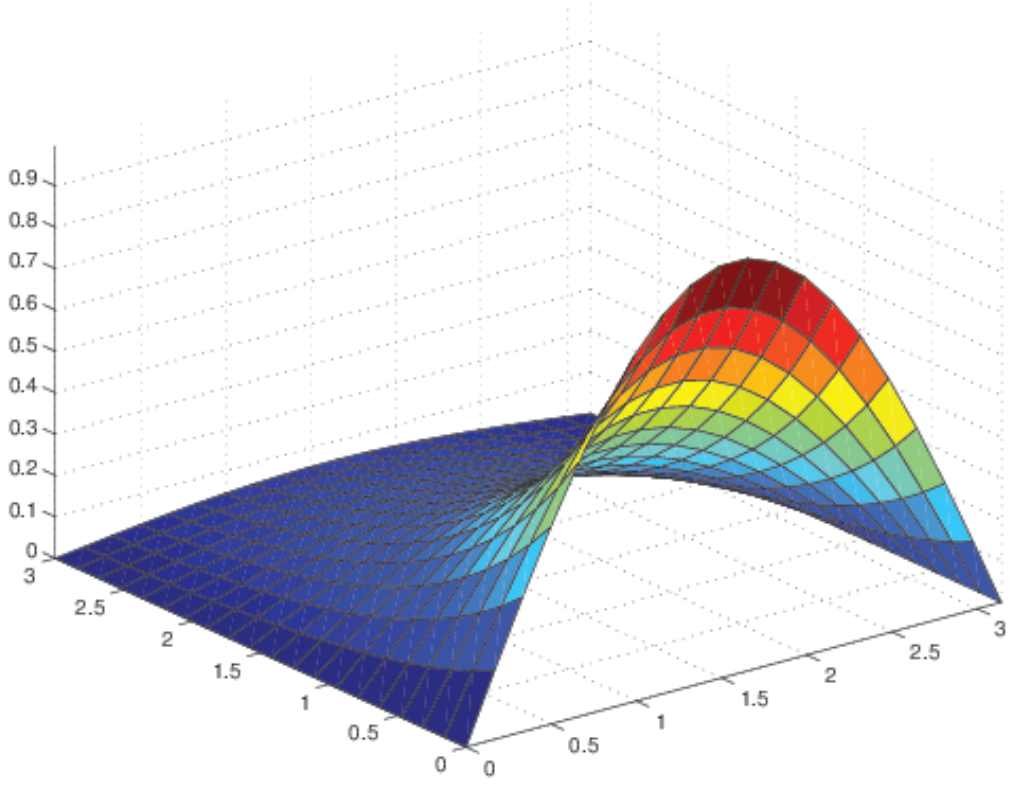
$$\frac{\partial u}{\partial t}(t,x) = \frac{\partial^2 u}{\partial x^2}(t,x), \quad \text{for} \quad (x,t) \in [0,L] \times [0,T],$$

and with the initial and boundary conditions

$$u(0,x) = g(x) \qquad \text{and } u(t,0) = a(t), u(t,L) = b(t).$$

Example 2.18

If $L = \pi$, $g(x) = \sin(x)$, $a(t) = b(t) \equiv 0$ then $u(t, x) = e^{-t} \sin(x)$.



This problem can't be solved explicitly for arbitrary g, a, b, and so a numerical scheme is used. Suppose we somehow know $\partial^2 u/\partial x^2$, then we could just use Euler's method:

$$u(t_{i+1}, x) = u(t_i, x) + h\frac{\partial^2 u}{\partial x^2}(t_i, x).$$

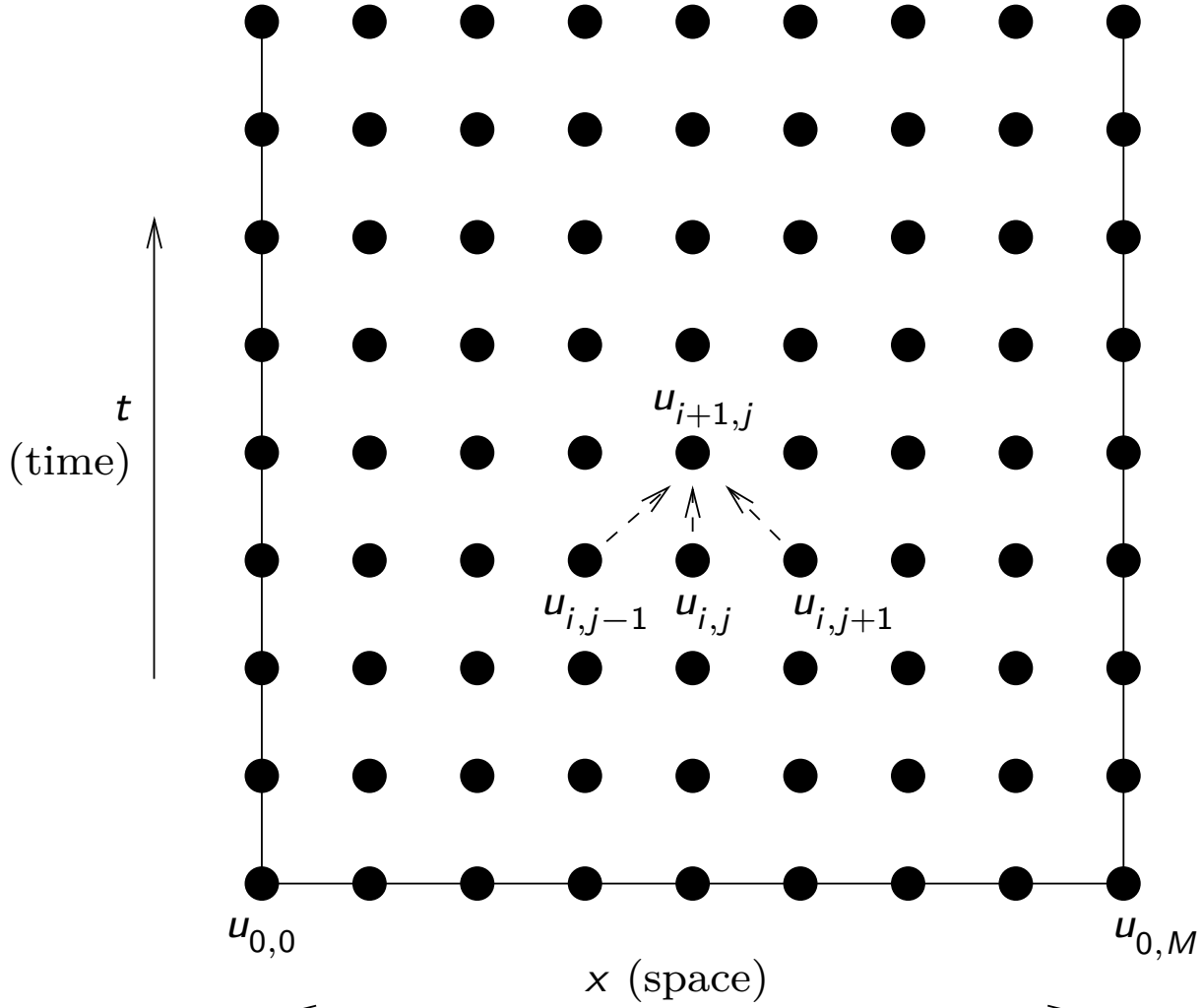
Although we don't know $\frac{\partial^2 u}{\partial x^2}(t_i, x)$ we can *approximate* it:

1. Divide $[0, T]$ into N intervals of width h, giving the grid $\{0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T\}$, with $t_i = t_0 + ih$.
2. Divide $[0, L]$ into M intervals of width H, giving the grid $\{0 = x_0 < x_1 < \cdots < x_M = L\}$ with $x_j = x_0 + jH$.
3. Denote by $u_{i,j}$ the approximation for $u(t, x)$ at (t_i, x_j).
4. For each $i = 0, 1, \ldots, N-1$, use the approximation:

$$\frac{\partial^2 u}{\partial x^2}(t_i, x_j) \approx \delta_x^2 u_{i,j} = \frac{1}{H^2}(u_{i,j-1} - 2u_{i,j} + u_{i,j+1}),$$

for $k = 1, 2, \ldots, M-1$.

Now set: $u_{i+1,j} := u_{i,j} - h\left[\delta_x^2 u_{i,j}\right]$.
This scheme is called an ***explicit method***: if we know $u_{i,j-1}$, $u_{i,j}$ and $u_{i,j+1}$ then we can explicitly calculate $u_{i+1,j}$.



Unfortunately, this method is not very stable: huge errors occur in the approximation. (Example).

Instead one might use an ***implicit method***: if we know $u_{i-1,j}$, we compute $u_{i,j-1}$, $u_{i,j}$ and $u_{i,j+1}$ simultaneously:

$$u_{i,j} - h\left[\delta_x^2 u_{i,j}\right] = u_{i-1,j}$$

This is actually a set of simultaneous equations:

$$\begin{gathered} u_{i,0} = a(t_i), \\ \alpha u_{i,j-1} + \beta u_{i,j} + \alpha u_{i,j+1} = u_{i-1,k}, \quad k = 1, 2, \ldots, M-1 \\ u_{i,M} = b(t_i), \end{gathered}$$

where $\alpha = -\frac{h}{H^2}$ and $\beta = \frac{2h}{H^2} + 1$.

This could be expressed more clearly as the matrix-vector equation:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \alpha & \beta & \alpha & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & \alpha & \beta & \alpha & \dots & 0 & 0 & 0 & 0 \\ & \vdots & & & \ddots & & & \vdots & \\ 0 & 0 & 0 & 0 & \dots & \alpha & \beta & \alpha & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & \alpha & \beta & \alpha \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_{i,0} \\ u_{i,1} \\ u_{i,2} \\ \vdots \\ u_{i,n-2} \\ u_{i,n-1} \\ u_{i,n} \end{pmatrix} = \begin{pmatrix} a(0) \\ u_{i-1,1} \\ u_{i-1,2} \\ \vdots \\ u_{i-1,n-2} \\ u_{i-1,n-1} \\ b(T) \end{pmatrix}.$$

So "all" we have to do now is solve this system of equations. That is what the next section of the course is about.

Exercise 2.11

Write down the Euler Method for the following 3rd-order IVP

$$y''' - y'' + 2y' + 2y = x^2 - 1,$$
$$y(0) = 1, y'(0) = 0, y''(0) = -1.$$

Exercise 2.12

Use a Taylor series to provide a derivation for the formula

$$\frac{\partial^2 u}{\partial x^2}(t_i, x_j) \approx \frac{1}{H^2}\left(u_{i,j-1} - 2u_{i,j} + u_{i,j+1}\right).$$

Exercise 2.13

Suppose that a 3-stage Runge-Kutta method tableaux has the following entries:

$$\alpha_2 = \frac{1}{3},\ \alpha_3 = \frac{1}{9},\ b_1 = 4,\ b_2 = \frac{15}{4},\ \beta_{32} = -\frac{2}{27}.$$

(i) Assuming that the method is *consistent*, determine the value of b_3.

(ii) Consider the initial value problem:

$$y(0) = 1,\ y'(t) = \lambda y(t).$$

Using that the solution is $y(t) = e^{\lambda t}$, write out a Taylor series for $y(t_{i+1})$ about $y(t_i)$ up to terms of order h^4 (use that $h = t_{i+1} - t_i$).
Using that your method should agree with the Taylor Series expansion up to terms of order h^3, determine β_{21} and β_{31}.

. . . . . .

Here are some entries for 3-stage Runge-Kutta method tableaux for Exercise 2.14.

Method 0: $\alpha_2 = 2/3,\ \alpha_3 = 0,\ b_1 = 1/12,\ b_2 = 3/4,\ \beta_{32} = 3/2$

Method 1: $\alpha_2 = 1/4,\ \alpha_3 = 1,\ b_1 = -1/6,\ b_2 = 8/9,\ \beta_{32} = 12/5$

Method 2: $\alpha_2 = 1/4,\ \alpha_3 = 1/2,\ b_1 = 2/3,\ b_2 = -4/3,\ \beta_{32} = 2/5$

Method 3: $\alpha_2 = 1/4,\ \alpha_3 = 1/3,\ b_1 = 3/2,\ b_2 = -8,\ \beta_{32} = 4/45$

Method 4: $\alpha_2 = 1,\ \alpha_3 = 1/4,\ b_1 = -1/6,\ b_2 = 5/18,\ \beta_{32} = 3/16$

Method 5: $\alpha_2 = 1,\ \alpha_3 = 1/5,\ b_1 = -1/3,\ b_2 = 7/24,\ \beta_{32} = 4/25$

Method 6: $\alpha_2 = 1,\ \alpha_3 = 1/6,\ b_1 = -1/2,\ b_2 = 3/10,\ \beta_{32} = 5/36$

Method 7: $\alpha_2 = 1/2,\ \alpha_3 = 1/7,\ b_1 = 7/6,\ b_2 = 22/15,\ \beta_{32} = -10/49$

Method 8: $\alpha_2 = 1/2,\ \alpha_3 = 1/8,\ b_1 = 4/3,\ b_2 = 13/9,\ \beta_{32} = -3/16$

Method 9: $\alpha_2 = 1/3,\ \alpha_3 = 1/9,\ b_1 = 4,\ b_2 = 15/4,\ \beta_{32} = -2/27$

Exercise 2.14 (Your own RK3 method)

Answer the following questions for Method K from the list above, where K is the last digit of your ID number. For example, if your ID number is 01234567, use Method 7.

(a) Assuming that the method is *consistent*, determine the value of b_3.

(b) Consider the initial value problem:

$$y(0) = 1,\ y'(t) = \lambda y(t).$$

Using that the solution is $y(t) = e^{\lambda t}$, write out a Taylor series for $y(t_{i+1})$ about $y(t_i)$ up to terms of order h^4 (use that $h = t_{i+1} - t_i$).
Using that your method should agree with the Taylor Series expansion up to terms of order h^3, determine β_{21} and β_{31}.

Exercise 2.15

(Attempt this exercises after completing Lab 3). Write a MATLAB program that implements your method from Exercise 2.14.
Use this program to check the order of convergence of the method. Have it compute the error for $n = 2$, $n = 4$, $\ldots$, $n = 1024$. Then produce a log-log plot of the errors as a function of n.