

Mathematical Tripos Part IB: Lent Term 2024

Numerical Analysis – Lecture 6

6 Ordinary differential equations - basics

Problem 6.1 We wish to approximate the exact solution of the *ordinary differential equation (ODE)*

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad 0 \leq t \leq T, \quad (6.1)$$

where $\mathbf{y} \in \mathbb{R}^d$ and the function $\mathbf{f} : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is sufficiently ‘smooth’. In principle, it is enough for $\mathbf{f}$ to satisfy the *Lipschitz condition* with respect to the second argument to ensure that the solution exists and is unique, namely there exists $\lambda > 0$ such that

$$\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})\| \leq \lambda \|\mathbf{x} - \mathbf{y}\|, \quad t \in [0, T], \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \quad (6.2)$$

Yet, for simplicity, we henceforth assume that $\mathbf{f}$ is analytic: in other words, we are always able to expand locally into Taylor series.

Equation (6.1) is accompanied by the initial condition $\mathbf{y}(0) = \mathbf{y}_0$.

For a small *time step* $h > 0$, we let $t_n = nh$ and our purpose is to approximate $\mathbf{y}_{n+1} \approx \mathbf{y}(t_{n+1})$, from $\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_n$ and equation (6.1).

Definition 6.2 (One-step method) A one-step method for solving (6.1) is a map $\mathbf{y}_{n+1} = \varphi(t_n, \mathbf{y}_n)$, i.e. an algorithm which allows $\mathbf{y}_{n+1}$ to depend only on $t_n, \mathbf{y}_n, h$ and $\mathbf{f}$.

Method 6.3 (Euler’s method) We approximate $\mathbf{y}'$ by the first difference $\mathbf{y}'(t) \approx \frac{1}{h}(\mathbf{y}(t+h) - \mathbf{y}(t))$, thus obtaining the (forward) *Euler method*

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(t_n, \mathbf{y}_n), \quad n = 0, 1, \dots \quad (6.3)$$

On $[t_n, t_{n+1}]$, it approximates $\mathbf{y}(t)$ with a straight line with slope $\mathbf{f}(t_n, \mathbf{y}_n)$.

Definition 6.4 Let $T > 0$ be given, and suppose that, for every $h > 0$, a method produces the sequence $\mathbf{y}_n = \mathbf{y}_{n,h}$, where $0 \leq n \leq \lfloor T/h \rfloor$. We say that the method *converges*, if $\max_n \|\mathbf{y}_n - \mathbf{y}(t_n)\| \rightarrow 0$ as $h \rightarrow 0$. It follows that if $nh \rightarrow t$ as $h \rightarrow 0$ and $n \rightarrow \infty$, then we have convergence $\mathbf{y}_{n,h} \rightarrow \mathbf{y}(t)$ to the exact solution of (6.1), uniformly for $t \in [0, T]$.

Theorem 6.5 Suppose that $\mathbf{f}$ satisfies the Lipschitz condition (6.2). Then the Euler method (6.3) converges, and the error $\mathbf{e}_n = \mathbf{y}(t_n) - \mathbf{y}_n$ satisfies

$$\|\mathbf{e}_n\| \leq c_0 h. \quad (6.4)$$

Proof. The Taylor series for the exact solution $\mathbf{y}$ provides $\mathbf{y}(t_{n+1}) = \mathbf{y}(t_n) + h\mathbf{y}'(t_n) + \frac{1}{2}h^2\mathbf{y}''(\tau_n)$, where τ_n is a point in the interval $[t_n, t_{n+1}]$ (different points for different components of $\mathbf{y}$ if $\mathbf{y} \in \mathbb{R}^d$). Therefore, using the identity $\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t))$, we may write

$$\mathbf{y}(t_{n+1}) = \mathbf{y}(t_n) + h\mathbf{f}(t_n, \mathbf{y}(t_n)) + \frac{1}{2}h^2\mathbf{y}''(\tau_n). \quad (6.5)$$

By subtracting formula (6.3) from this equation, and assuming that $\frac{1}{2}\|\mathbf{y}''\|_\infty = c$, we find that the errors $\mathbf{e}_n = \mathbf{y}(t_n) - \mathbf{y}_n$ satisfy the following relations

$$\begin{aligned} \|\mathbf{e}_{n+1}\| &\leq \|\mathbf{e}_n\| + h\|\mathbf{f}(t_n, \mathbf{y}(t_n)) - \mathbf{f}(t_n, \mathbf{y}_n)\| + ch^2 \\ &\stackrel{(*)}{\leq} \|\mathbf{e}_n\| + \lambda h\|\mathbf{y}(t_n) - \mathbf{y}_n\| + ch^2 = (1 + \lambda h)\|\mathbf{e}_n\| + ch^2, \end{aligned}$$

where in (*) we used the Lipschitz condition on $\mathbf{f}$. Consequently, by induction,

$$\|\mathbf{e}_{n+1}\| \leq (1 + \lambda h)^m \|\mathbf{e}_{n+1-m}\| + ch^2 \sum_{i=0}^{m-1} (1 + \lambda h)^i, \quad m = 1, 2, \dots, n+1.$$

In particular, letting $m = n+1$ and bearing in mind that $\mathbf{e}_0 = \mathbf{0}$, we have

$$\|\mathbf{e}_{n+1}\| \leq ch^2 \sum_{i=0}^n (1 + \lambda h)^i = ch^2 \frac{(1 + \lambda h)^{n+1} - 1}{(1 + \lambda h) - 1} \leq \frac{ch}{\lambda} (1 + \lambda h)^{n+1}.$$

Since $1 + \lambda h \leq e^{\lambda h}$ and $(n+1)h \leq T$, we obtain that $(1 + \lambda h)^{n+1} \leq e^{\lambda T}$, therefore

$$\|\mathbf{e}_n\| \leq \frac{ce^{\lambda T}}{\lambda} h \rightarrow 0 \quad (h \rightarrow 0), \quad \text{uniformly for } 0 \leq nh \leq T. \quad \square$$

Definition 6.6 The *local truncation error* of a general numerical method $\mathbf{y}_{n+1} = \varphi_h(t_n, \mathbf{y}_0, \dots, \mathbf{y}_n)$ for solving (6.1) is the error of the method on the true solution, i.e., the value $\boldsymbol{\eta}_{n+1}$ such that

$$\mathbf{y}(t_{n+1}) = \varphi_h(t_n, \mathbf{y}(t_0), \mathbf{y}(t_1), \dots, \mathbf{y}(t_n)) + \boldsymbol{\eta}_{n+1}.$$

The *order* of the method is the largest integer $p \geq 0$ such that

$$\boldsymbol{\eta}_{n+1} = \mathcal{O}(h^{p+1})$$

for all $h > 0$, $n \geq 0$ and all sufficiently smooth functions $\mathbf{f}$ in (6.1). Note that $p \geq 1$ is necessary for convergence.

Remark 6.7 (The order of Euler's method) From (6.5), it follows that

$$\boldsymbol{\eta}_{n+1} = \frac{1}{2}h^2 \mathbf{y}''(\tau_n) = \mathcal{O}(h^2)$$

and we deduce that Euler's method is of order 1 (which is justified by (6.4))

Definition 6.8 (Theta methods) For $\theta \in [0, 1]$, we consider methods of the form

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h [\theta \mathbf{f}(t_n, \mathbf{y}_n) + (1 - \theta) \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})], \quad n = 0, 1, \dots \quad (6.6)$$

1) If $\theta = 1$, we recover Euler's method and the choices $\theta = 0$ and $\theta = \frac{1}{2}$ are known as

$$\begin{aligned} \text{Backward Euler:} \quad & \mathbf{y}_{n+1} = \mathbf{y}_n + h \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}), \\ \text{Trapezoidal rule:} \quad & \mathbf{y}_{n+1} = \mathbf{y}_n + \frac{1}{2}h [\mathbf{f}(t_n, \mathbf{y}_n) + \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})]. \end{aligned}$$

2) If $\theta \in [0, 1)$, then the theta method (6.6) is *implicit*: each time step requires a solution of a system of d (in general, nonlinear) algebraic equations for the unknown vector $\mathbf{y}_{n+1}$.

Denoting the right hand side of (6.6) by $\mathbf{g}(\mathbf{y}_{n+1})$, we thus need to solve $\mathbf{y} = \mathbf{g}(\mathbf{y})$ or $\mathbf{F}(\mathbf{y}) = \mathbf{0}$ where $\mathbf{F}(\mathbf{y}) = \mathbf{y} - \mathbf{g}(\mathbf{y})$. This can be done by iteration. For example,

$$\begin{aligned} \text{Fixed point (if contraction):} \quad & \mathbf{y}^{[k+1]} = \mathbf{g}(\mathbf{y}^{[k]}), \\ \text{Newton-Raphson (N-R):} \quad & \mathbf{y}^{[k+1]} = \mathbf{y}^{[k]} - [\mathbf{J}_{\mathbf{F}}(\mathbf{y}^{[k]})]^{-1} \mathbf{F}(\mathbf{y}^{[k]}), \\ \text{Modified N-R:} \quad & \mathbf{y}^{[k+1]} = \mathbf{y}^{[k]} - [\mathbf{J}_{\mathbf{F}}(\mathbf{y}^{[0]})]^{-1} \mathbf{F}(\mathbf{y}^{[k]}). \end{aligned}$$

Remark 6.9 (The order of the theta method) It follows from (6.6) and Taylor's theorem that the local truncation error of the theta method is

$$\begin{aligned} & \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h [\theta \mathbf{y}'(t_n) + (1 - \theta) \mathbf{y}'(t_{n+1})] \\ &= [h \mathbf{y}'(t_n) + \frac{1}{2}h^2 \mathbf{y}''(t_n) + \frac{1}{6}h^3 \mathbf{y}'''(t_n)] \\ & \quad - \theta h \mathbf{y}'(t_n) - (1 - \theta)h [\mathbf{y}'(t_n) + h \mathbf{y}''(t_n) + \frac{1}{2}h^2 \mathbf{y}'''(t_n)] + \mathcal{O}(h^4) \\ &= (\theta - \frac{1}{2})h^2 \mathbf{y}''(t_n) + (\frac{1}{2}\theta - \frac{1}{3})h^3 \mathbf{y}'''(t_n) + \mathcal{O}(h^4). \end{aligned}$$

Therefore the theta method is of order 1, except that the trapezoidal rule is of order 2.