Chapter 1

Introduction

1.1 Preliminaries

In this note we concern ourselves with the numerical methods only for the first order ordinary differential (ODE) system in the normal form

$$\frac{dy}{dx} = f(x, y) \tag{1.1}$$

where $x \in R$, $y \in R^n$ and $f : R \times R^n \longrightarrow R^n$. This discussion is adequate because a general m-th order ODE

$$\frac{d^m y}{dx^m} = f(x, y, \frac{dy}{dx}, \ldots, \frac{d^{m-1}y}{dx^{m-1}}) \tag{1.2}$$

with $x, y \in R$ can be reduced to the system (1.1) by defining

$$\begin{aligned} y_1 &:= y \\ y_2 &:= \frac{dy_1}{dx} = \frac{dy}{dx} \\ y_3 &:= \frac{dy_2}{dx} = \frac{d^2y}{dx^2} \\ &\vdots \\ y_m &:= \frac{dy_{m-1}}{dx} = \frac{d^{m-1}y}{dx^{m-1}} \end{aligned} \tag{1.3}$$

and thus

$$\frac{d}{dx}\begin{bmatrix} y_1 \\ \vdots \\ y_{m-1} \\ y_m \end{bmatrix} = \begin{bmatrix} y_2 \\ \vdots \\ y_m \\ f(x, y_1, \ldots, y_{m-1}) \end{bmatrix} \tag{1.4}$$

The system (1.1) may process infinitely many solutions. Each solution is call an *integral curve*. In general we can pick out a particular solution by prescribing certain additional conditions.

Definition 1.1.1 *The system (1.1) together with the initial condition*

$$y(x_0) = y_0 \tag{1.5}$$

is called an initial value problem (IVP).

Definition 1.1.2 *The system (1.1) together with the boundary condition*

$$g(y(a), y(b)) = 0) \tag{1.6}$$

where $a \leq x \leq b$ and $g : R^n \times R^n \longrightarrow R^n$ is called a two-point boundary value problem.

Example. The IVP

$$\begin{aligned} \frac{dy}{dx} &= -\sqrt{|1-y^2|} \\ y(0) &= 1 \end{aligned}$$

has at least two solutions $y(x) \equiv 1$ and $y(x) = \cos x$, and hence has infinitely many solutions. For instance, the function

$$y(x) = \begin{cases} 1 & 0 \leq x \leq \alpha \\ \cos x - \alpha & \alpha \leq x \leq \alpha + \pi \\ -1 & \alpha + \pi \leq x \end{cases}$$

is a solution for any α.

A general criterion that guarantees the existence of a solution to an IVP is the the following Cauchy-Peano Theorem.

Theorem 1.1.1 *Suppose $f(x, y)$ is continuous on an open set $D \subset R \times R^n$ containing the initial point (x_0, y_0). Then the IVP has a solution so long as $(x, y(x)) \in D$.*

Definition 1.1.3 *A function $f(x, y)$ is said to satisfy a Lipschitz condition in y with constant L in a region $D \subset R \times R^n$ if*

$$\|f(x, u) - f(x, v)\| \leq L\|u - v\| \tag{1.7}$$

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

Theorem 1.1.2 *Suppose that $f(x, y)$ and $g(x, y)$ are continuous on an open set $D \in R \times R^n$ and that $f(x, y)$ satisfies a Lipschitz condition with constant L. Suppose further that*

$$\|f(x, y) - g(x, y)\| \leq \epsilon$$

for all $(x, y) \in D$. If functions $u(x)$ and $v(x)$ satisfy, respectively, the systems

$$\begin{aligned} \frac{du}{dx} &= f(x, u) \\ \frac{dv}{dx} &= g(x, u) \end{aligned}$$

for $x \in [a, b]$, and $(x, u(x)), (x, v(x)) \in D$, then

$$\|u(x) - v(x)\| \leq \{\|u(a) - v(a)\| + (b - a)\epsilon\} e^{L(x-a)}. \tag{1.8}$$

Proof. By integration, we have

$$u(x) = u(a) + \int_a^x u'dt = u(a) + \int_a^x f(t, u(t))dt$$

and a similar expression for $v(x)$. It follows that

$$\begin{aligned} u(x) - v(x) = u(a) - v(a) \quad &+ \quad \int_a^x \{f(t, v(t)) - g(t, v(t))\}\, dt \\ &+ \quad \int_a^x \{f(t, u(t)) - f(t, v(t))\}\, dt. \end{aligned}$$

Upon taking the norm on both sides, we obtain

$$\|u(x) - v(x)\| \le \|u(a) - v(a)\| + (b-a)\epsilon + \int_a^x L\|u(t) - v(t)\|dt$$

for $a \le x \le b$. Define

$$\begin{aligned} \Delta(x) &:= \|u(x) - v(x)\| \\ R(x) &:= \|u(a) - v(a)\| + (b-a)\epsilon + \int_a^x L\|u(t) - v(t)\|dt. \end{aligned}$$

Then

$$\begin{aligned} \Delta(x) &\le R(x) \\ R' &= L\Delta(x) \le LR(x) \end{aligned}$$

It follows that

$$\frac{d}{dx}(R(x)e^{-L(x-a)} = e^{-L(x-a)}(R'(x) - LR(x)) \le 0.$$

The results

$$\begin{aligned} R(x)e^{-L(x-a)} &\le R(a) \\ \Delta(x) \le R(a)e^{L(x-a)} & \end{aligned}$$

follows form the non-increasing property. □

Corollary 1.1.3 *Suppose $f(x, y)$ satisfies condition in Theorem 1.1.1. Then the IVP has a unique solution.*

Remark. Theorem 1.1.1 states that Lipschitz problems are *well-posed*. That is, small perturbations in the stated problem only leads to small changes in the solution. This is the basis of numerical ODE.

1.2 The Euler Method

Consider the IVP

$$y' = f(x, y),\ y(x_0) = y_0 \tag{1.9}$$

with $x, y \in R$ and $x \in [a, b]$. In this section we want to use so called Euler method, the simplest numerical method,

$$y_{n+1} := y_n + hf(x_n, y_n) \tag{1.10}$$

where $x_n = x_0 + nh$ is the n-th nodal point and h is the step size, to understand the following questions:

1. What is the magnitude of the *global error*

$$e_n := y_n - y(x_n) \tag{1.11}$$

   at the n-th step? How does the error propogate?

2. How does the step size h affect the accuracy?

3. What kinds of errors are involved in the calculation? How do they affect the overall accuracy? How to control the error to get the best possible accuracy?

Definition 1.2.1 *The local truncation error T_{n+1} at x_{n+1} for the method (1.10) is defined to be*

$$T_{n+1} := y(x_{n+1}) - y(x_n) - hf(x_n, y(x_n)). \tag{1.12}$$

That is, the local truncation error is the difference between the exact solution $y(x_{n+1})$ and the approximate solution y_{n+1} provided no previous errors have been introduced into the numerical scheme.

Lemma 1.2.1 *Assume that $f(x, y)$ satisfies Lipschitz conditions in y with constant L and in x with constant K. Then $T_{n+1} = O(h^2)$.*

Proof. By the mean value theorem, there exists $0 \le \theta \le 1$ such that

$$\begin{aligned} |T_{n+1}| &= |hf(x_n + \theta h, y(x_n + \theta h)) - hf(x_n, y(x_n))| \\ &\le h|f(x_n + \theta h, y(x_n)) - f(x_n, y(x_n))| \\ &\quad + h|f(x_n + \theta h, y(x_n + \theta h)) - f(x_n + \theta h, y(x_n))| \\ &\le \theta h^2 K + hL|y(x_n + \theta h) - y(x_n)| \\ &\le \theta(K + LZ)h^2 \end{aligned}$$

where $Z := \max_{x\in[a,b]} |y'(x)|$. □

Remark. If y'' is continuous and bounded by C, then by Taylor's Theorem we know

$$|T_{n+1}| = \frac{h^2}{2}|y''(\xi)| \le \frac{C}{2}h^2 \tag{1.13}$$

where $\xi \in (x_n, x_{n+1})$.

Subtracting (1.12) from (1.10), we can now express the global error e_{n+1} at x_{n+1} as

$$e_{n+1} = e_n + h[f(x_n, y_n) - f(x_n, y(x_n))] - T_{n+1}. \tag{1.14}$$

It follows that

$$|e_{n+1}| \leq (1 + hL)|e_n| + T \tag{1.15}$$

where $T = \max |T_n| = O(h^2)$. This inequality (1.15) can be applied repeatedly. Hence we obtain the following theorem.

Theorem 1.2.2 *Given any n and h, so long as $a \leq x_0 + nh \leq b$, we have the estimate*

$$\begin{aligned} |e_n| &\leq T\frac{(1+hL)^n - 1}{hL} + (1+hL)^n|e_0| \\ &\leq \frac{T}{hL}(e^{L(b-a)} - 1) + e^{L(b-a)}|e_0|. \end{aligned} \tag{1.16}$$

Proof. Obvious the theorem is true when $n = 0$. Assume now that the theorem is true for n. We have from (1.15) that

$$\begin{aligned} |e_{n+1}| &\leq (1+hL)T\frac{(1+hL)^n - 1}{hL} + (1+hL)^{n+1}|e_0| + T \\ &= T\frac{(1+hL)^{n+1} - 1}{hL} + (1+hL)^{n+1}|e_0|. \end{aligned}$$

The assertion follows from the fact that $1 + hL \leq e^{hL}$. □

Remark. It is clear from the inequality (1.16) that if e_0 is at least $O(h^2)$, then e_n is of order $O(h)$. In particular, this shows that as $h \longrightarrow 0$ (Consequently, $n \longrightarrow \infty$ with $x = x_0 + nh$ fixed), $y_n \longrightarrow y(x)$.

Applying the scheme (1.10) to the differential equation

$$y' = \lambda y, \tag{1.17}$$

we obtain a difference equation

$$y_{n+1} = (1 + \lambda h)y_n. \tag{1.18}$$

In this case, the global error becomes

$$\begin{aligned} e_{n+1} &= (1+\lambda h)e_n + [-y(x_{n+1} + (1+\lambda h)y(x_n) \\ &= \text{Propagated Error} + \text{Local Truncation Error}. \end{aligned} \tag{1.19}$$

Obviously the propagated error is amplied unless $|1 + \lambda h| \leq 1$. In the complex λh-plane, this inequality implies a unit disc centered at the point $(-1, 0)$ which is called *the region of absolute stability*. When $\lambda \leq\leq 0$ (the so called *stiff equation*), the region imposes severe restriction on the step size h in order to maintain reasonable answers.

Example. The effect of absolute stability can be seen from the example:

$$\begin{aligned} y' &= -10000(y - t^2) + 2t \\ y(0) &= 0 \\ \text{Calculate } y(1) &= ? \end{aligned}$$

Suppose the Euler method was used with step size $h = 10^{-m}$ on a machine with 10 significant digits of accuracy. We obtain the following results:

h	n	$y(1)$
1	1	0
0.1	10	$0.90438207503 \times 10^{16}$
0.01	100	overflow
0.001	1000	0.99999900001
0.0001	10000	0.99999990000
0.00001	100000	0.99999998997

In general, errors when solving ODEs numerically are attributed to the following sources:

1. Truncation Error: errors in discretizing the differential equation to a difference equation.
2. Calculation Error: errors accumulated from previous steps.
3. Round-off Error: errors due to float-point arithmetic.

Most calculation on large electronic computers will have considerably larger truncation errors than roung-off errors. However, when the computation requires smaller step sizes and hence larger number of iterations, it is possible that round-off errors eventually are built up to affect the solution.

In practice, due to the floating-point arithmetic, an Euler step should be of the form

$$y_{n+1} = y_n + hf(x_n, y_n) + r_{n+1}. \tag{1.20}$$

We now analyze the effect of the round-off errors. Observe that

$$\begin{aligned} e_{n+1} &= e_n + h[f(x_n, y_n) - f(x_n, y(x_n)] - T_{n+1} + r_{n+1} \\ &= e_n + h\left[e_n \frac{\partial f}{\partial y}(x_n, y(x_n)) + \frac{1}{2}e_n^2 \frac{\partial^2 f}{\partial y^2}(x_n, \eta_n)\right] \\ &\quad -\frac{1}{2}h^2 y''(x_n) - \frac{1}{6}h^3 y'''(\xi_n) + r_{n+1}. \end{aligned} \tag{1.21}$$

Define $\delta_n := \frac{e_n}{h}$. Then (1.21) leads to the difference equation

$$\begin{aligned} \delta_{n+1} &= \delta_n + h\left[\delta_n \frac{\partial f}{\partial y} - \frac{1}{2}y''\right] + \frac{1}{2}e_n^2 \frac{\partial^2 f}{\partial y^2} - \frac{1}{6}h^2 y''' + \frac{r_{n+1}}{h} \\ &= \delta_n + h\left[\delta_n \frac{\partial f}{\partial y} - \frac{1}{2}y''\right] + O(h^2) + \frac{r_{n+1}}{h} \end{aligned} \tag{1.22}$$

This amounts to an Euler step applied to the IVP

$$\frac{d\delta}{dx} = \frac{\partial f}{\partial y}\delta - \frac{1}{2}y'' + O(h) + \frac{r}{h^2} \tag{1.23}$$

$$\delta(x_0) = \frac{e_0}{h}. \tag{1.24}$$

By Theorem 1.2.2, we know $\delta_n = \delta(x_n) + O(h)$. If follows that

$$e_n = h[\tilde{\delta}(x_n) + O(h)] \tag{1.25}$$

where $\tilde{\delta}(x)$ solves the IVP

$$\frac{d\delta}{dx} = \frac{\partial f}{\partial y}\delta - \frac{1}{2}y'' + O(h) + \frac{r}{h^2} \tag{1.26}$$

$$\delta(x_0) = \frac{e_0}{h} \tag{1.27}$$

since $\delta(x) - \tilde{d}elta(x) = O(h)$ by Theorem 1.1.2.