Chapter 2
Convergence of Numerical Methods

In the last chapter we derived the forward Euler method from a Taylor series expansion of u^{n+1} and we utilized the method on some simple example problems without any supporting analysis. This chapter on *convergence* will introduce our first analysis tool in numerical methods for the solution of ODEs.

6 Self-Assessment

Before reading this chapter, you may wish to review...

- limits [18.01 Lecture 2: Video]

After reading this chapter you should be able to...

- list the two primary sources of error in the numerical solution of ODEs
- describe the concept of convergence of a numerical method
- discuss the relationship between global order of accuracy and the rate of convergence of a numerical method
- evaluate the global order of accuracy of a method from experimental data

7 Types of errors in the numerical solution of ODEs

When we approximate the solution of ODEs numerically, there are two primary sources of error: rounding (or floating point) errors and truncation errors. Rounding errors are associated to the floating-point arithmetic that our computers use to perform calculations. Truncation errors, on the other hand, are errors we incur based on the numerical method; these errors would exist even in the absence of rounding errors. To some extent, we cannot control rounding errors (they are determined by the precision of the machine), but we can control truncation errors.

Exercise 1. How are truncation errors introduced in the forward Euler method (7)?

(a) Taylor series approximation of u^n
(b) Taylor series approximation of u^{n+1}
(c) Taylor series approximation of u_t^n
(d) none of the above

8 Convergence

One important property of numerical methods related to truncation errors is *convergence*. In Exercise 3 you experimented with the forward Euler method for various time steps Δt and observed the error at time $t = 1$. You may have noticed that as you decreased Δt, the errors also decreased. In the limit as $\Delta t \to 0$, this behavior is representative of convergence.

Definition 1 (Convergence). A numerical scheme for solving

$$u_t = f(u,t), \qquad u(0) = u^0, \;\; 0 < t \leq T \tag{9}$$

is *convergent* if

$$\max_{n \in \{0,1,\ldots,T/\Delta t\}} |v^n - u^n| \to 0 \quad \text{as} \quad \Delta t \to 0. \tag{10}$$

This is a mathematically precise definition of what you observed for the forward Euler method in Exercise 3. Let's take a moment to investigate Definition 1 a little more deeply. The first part of the statement is familiar by now; we are solving a first-order ODE of the form $u_t = f(u,t)$ with given initial conditions. We assume that a method is defined to produce our numerical solution v^n for $n = 1, 2, \ldots, T/\Delta t$ (we assign $v^0 = u^0$); e.g., the forward Euler method.

In words, the convergence statement is as follows: If we shrink the time step smaller and smaller, the largest absolute error between the numerical solution and the exact solution will also get smaller and smaller. For any numerical solution, the absolute error can be measured at each of the time indices $n = 0, 1, \ldots, T/\Delta t$: that is given by $|v^n - u^n|$. Thus, for any numerical solution, we can also define the maximum absolute error over those time indices: that is written $\max_{n \in \{0,1,\ldots,T/\Delta t\}} |v^n - u^n|$. This is the worst case error for the numerical solution. If we drive the largest absolute error to zero, it is implied that the error at each time index will also be driven to zero.

To summarize, our convergence statement says that in the limit $\Delta t \to 0$, the numerical solution collapses onto the exact solution and the error goes to zero at all time indices. It is important to note that in the limit $\Delta t \to 0$, the last time index $T/\Delta t \to \infty$ even for finite T; the time interval between adjacent numerical solution points (t^n, v^n) and (t^{n+1}, v^{n+1}) is also shrinking to zero. So not only does our numerical solution approach zero error at the time steps $t^0, t^1, \ldots$, these time steps are getting closer and closer together so that the ordered pairs $(t^0, v^0); (t^1, v^1); \ldots$ make up the entirety of the exact continuous solution $u(t)$.

Video explaining definition of convergence

9 Rate of Convergence (Global Order of Accuracy)

In addition to knowing whether a numerical method will converge, we are also interested to know at what *rate* will it converge. The *rate of convergence* is known as the *global order of accuracy* and describes the decrease in error $\max_{n \in \{0,1,\ldots,T/\Delta t\}} |v^n - u^n|$ one can expect for a given decrease in time step Δt in the limit $\Delta t \to 0$.

Definition 2 (Global order of accuracy). Assume that the forcing function $f(u,t)$ is sufficiently smooth. (In particular, we need $f(u,t)$ to have p continuous derivatives, i.e., up to and including $\partial^p f/\partial t^p$ and $\partial^p f/\partial u^p$.) A numerical method has a global order of accuracy p if

$$\max_{n \in \{0,1,\ldots,T/\Delta t\}} |v^n - u^n| \leq \mathscr{O}(\Delta t^p) \quad \text{as} \quad \Delta t \to 0. \tag{11}$$

This is the second time we've come across the $\mathscr{O}(\cdot)$ notation (see (4)). Recall that this means that the terms on the right side of the expression $\max_{n \in \{0,1,\ldots,T/\Delta t\}} |v^n - u^n| \leq \mathscr{O}(\Delta t^p)$ can be accurately approximated by $c\Delta t^p$ for some constant c in the limit $\Delta t \to 0$. That is, terms of the form Δt^q for $q > p$ can be safely ignored since they are much smaller in magnitude than Δt^p in the limit $\Delta t \to 0$.

Let's consider the meaning of this statement in more depth. If a method has global order of accuracy p, in the limit $\Delta t \to 0$, if we shrink the time step by a factor of two, we can expect the worst case error in the approximation to decrease by a factor of 2^p. Therefore, the higher the global order of accuracy, the faster the convergence rate of the numerical method. For a first-order accurate method ($p = 1$), we expect that decreasing the time step by a factor of two

(cutting it in half) would result in the halving of the worst case error. Likewise, for a second-order accurate method ($p = 2$), decreasing the time step by a factor of two should lead to the quartering of the worst case error. (It is important to remember that these statements only hold in the limit $\Delta t \to 0$. You will find that these results hold approximately for small Δt, but as you increase Δt you will likely not observe this behavior.)

We will now demonstrate this behavior numerically for the forward Euler method.

Example 1. To demonstrate the ideas of global accuracy, we will consider an ODE with $f = -u^2$ and an initial condition of $u(0) = 1$. The solution to this ODE is $u = (1+t)^{-1}$. Now, let us apply the forward Euler method to solving this problem for $t = 0$ to 10. The approximate solutions for a range of Δt are shown Figure 2 along with the exact solution. The forward Euler solutions are clearly approaching the exact solution as Δt decreases.

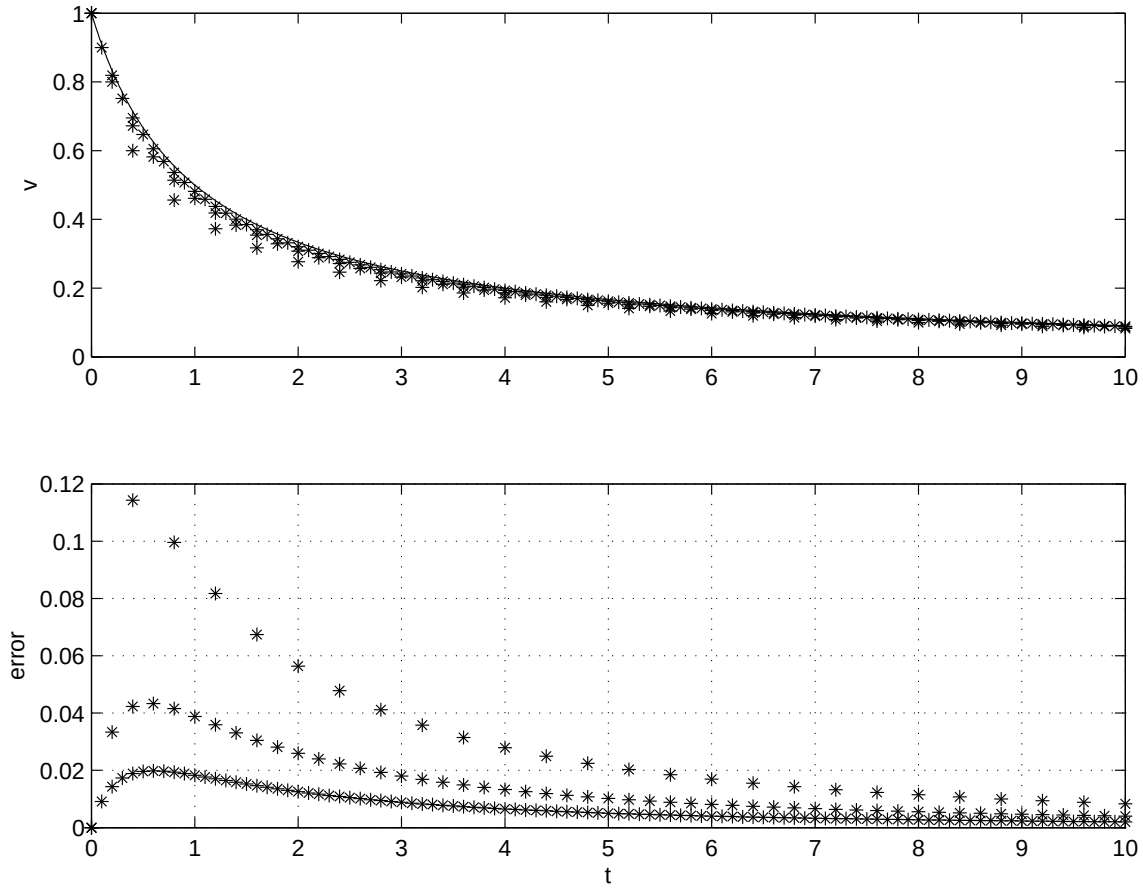

Fig. 2 Forward Euler solution for $u_t = -u^2$ with $u(0) = 1$ with $\Delta t = 0.1$, 0.2, and 0.4. Forward Euler (symbols) and exact solution (line) are shown in first plot. Error is shown in second plot.

Exercise 2. Use the plots above to determine the global order of accuracy for the forward Euler method.

(a) $p = 0$
(b) $p = 1$
(c) $p = 2$
(d) none of the above