Initial value problems: stiff IVPs

We now have the primary tools for solving initial value problems with finite difference methods. We can implement as well as characterize the error and stability of a variety of one-step and multi-step methods.

We close our investigations into finite difference methods for initial value problems by considering *stiff problems*. These problems provide a special set of difficulties that are worth discussing. These problems also highlight the importance of understanding absolute stability regions of numerical methods for IVPs.

In this lecture, we will characterize stiff problems, understand what makes them challenging to solve numerically, and discuss finite difference methods that are used for this special class of initial value problem.

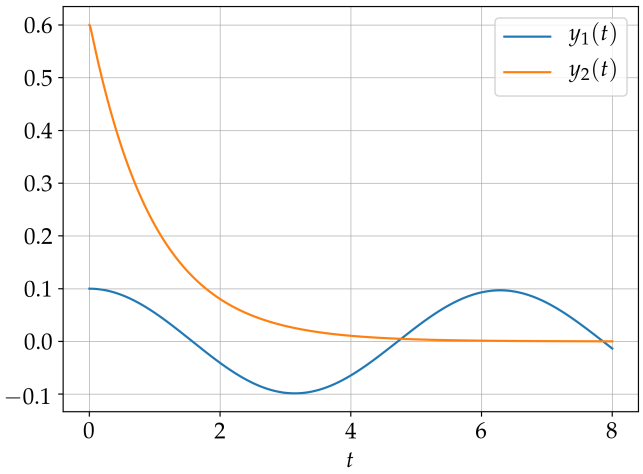
1 *A stiff problem*

Let us motivate our investigation into stiff IVPs by considering an example problem. We will study a system of two damped harmonic oscillators:

$$\begin{aligned} \ddot{y}_1 + c_1\dot{y}_1 + k_1 y_1 &= 0 \\ \ddot{y}_2 + c_2\dot{y}_2 + k_2 y_2 &= 0 \end{aligned} \tag{1}$$

with initial conditions $y_1(0) = \gamma_1$, $\dot{y}_1(0) = 0$, $y_2(0) = \gamma_2$, $\dot{y}_2(0) = 0$.

We will select $k_1 = 1$, $c_1 = 1 \times 10^{-2}$, $k_2 = 100$, $c_2 = 100$, $\gamma_1 = 0.1$, and $\gamma_2 = 0.6$. Figure 1 provides a plot of the solutions y_1 and y_2.



It is not obvious from the plot that we are in for a challenge when numerically solving this system: y_1 and y_2 both look smooth and

Note that we can write this system in first-order form by defining $z = [y_1, \dot{y}_1, y_2, \dot{y}_2]^T$ and writing

$$\begin{aligned} \dot{z} &= Az \\ z(0) &= z_0 \end{aligned} \tag{2}$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -k_1 & -c_1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -k_2 & -c_2 \end{bmatrix} \tag{3}$$

and $z_0 = [\gamma_1, 0, \gamma_2, 0]^T$.

Figure 1: Plot of the solutions to the damped harmonic oscillator problem.

decay at modest rates. Do not let this gentle-looking figure deceive you: here be dragons!

So what is the issue in solving these stiff problems numerically? Let us investigate this by applying Heun's method and the trapezoid method to solve this problem. Both are second order, but recall that the trapezoid method is absolutely stable over the entire left-half plane whereas Heun's method has a much more restricted stability region (we showed this fact in the last lecture).

Table 1 shows the error in the solution obtained via Heun's method and the trapezoid method.

Δt	Heun's method	trapezoid method
0.5	8.28×10^{48}	0.169
0.25	1.65×10^{78}	5.26×10^{-3}
0.1	1.08×10^{128}	6.40×10^{-4}
0.05	6.81×10^{146}	1.60×10^{-4}
0.025	1.023×10^{64}	4.00×10^{-5}
0.01	1.28×10^{-5}	6.41×10^{-6}

Table 1: Error at $t = 8$ for Heun's method and the trapezoid method applied to the stiff problem (1).

The results for Heun's method are interesting in two ways:

i) Heun's method is unstable for all but the last value of Δt, which suggests that the small absolute stability region plays a role in the admissible time step.

ii) The error jumps from $O(10^{64})$ to $O(10^{-6})$. As numericists, we are often interested in methods that can provide a tolerable (but not overly small) error. The reason is that we are always seeking to develop the fastest possible methods, and having an excessively small error means that we have selected a time step that is too small, and have thus performed more computations than needed. How we define a 'tolerable error' is problem dependent, but the fact that we can not obtain an error between $O(10^{64})$ and $O(10^{-6})$ is troubling!

Let us explain i) and ii) in turn. Regarding the unstable computations obtained using Heun's method for certain values of Δt, point i): we can quantify the restrictions on Δt by converting the IVP $\dot{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{z}$ from equation (2) to the diagonal model problem $\dot{\boldsymbol{u}} = \boldsymbol{\Lambda}\boldsymbol{u}$ that we used to determine absolute stability regions in the last lecture. This conversion is performed via the eigendecomposition of $\boldsymbol{A}$, $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}$. Thus, the model problem associated with our stiff system is

$$\dot{\boldsymbol{u}} = \boldsymbol{\Lambda}\boldsymbol{u} \tag{4}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{A}$. The eigenvalues of $\boldsymbol{A}$ may be computed from your software of choice

The procedure for going from (2) to the diagonal IVP is as follows:

$$\begin{aligned} \dot{\boldsymbol{z}} &= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}\boldsymbol{z} \\ \implies \quad \boldsymbol{V}^{-1}\dot{\boldsymbol{z}} &= \boldsymbol{\Lambda}\boldsymbol{V}^{-1}\boldsymbol{z} \end{aligned} \tag{5}$$

Now we define the variable $\boldsymbol{u} := \boldsymbol{V}^{-1}\boldsymbol{z}$ and note that since $\boldsymbol{V}$ is not a function of time,

$$\boldsymbol{V}^{-1}\dot{\boldsymbol{z}} = \dot{(\boldsymbol{V}^{-1}\boldsymbol{z})} = \dot{\boldsymbol{u}} \tag{6}$$

from which we obtain

$$\dot{\boldsymbol{u}} = \boldsymbol{\Lambda}\boldsymbol{u} \tag{7}$$

Notice that the initial condition in terms of $\boldsymbol{u}$ is $\boldsymbol{V}^{-1}\boldsymbol{z}_0$, which we write as $[d_1, d_2, d_3, d_4]^T$ for simplicity.

to be $-0.005 \pm 1i$, -1.01, and -98.99. The first three eigenvalues all lie within the absolute stability region of Heun's method for all Δt values considered, but one could verify through direct computation that the last lies outside of the region for all but the last value of Δt.

The first point, i), is completely explicable using our framework of absolute stability. It is point ii), the fact that the stability restriction makes it impossible to compute modestly accurate solutions with Heun's method, that is the hallmark of stiff problems. What is the reason behind this undesirable behavior? The answer lies in the eigenvalues of the system.

Notice that the solution to the model problem $\dot{\boldsymbol{u}} = \boldsymbol{\Lambda u}$ is

$$\boldsymbol{u} = \begin{bmatrix} e^{(-0.005+1i)t} d_1 \\ e^{(-0.005-1i)t} d_2 \\ e^{-1.01t} d_3 \\ e^{-98.99t} d_4 \end{bmatrix} \tag{8}$$

Remember that the variables $d_1, \ldots, d_4$ denote the initial condition $\boldsymbol{z}_0$ expressed in terms of the eigenvectors.

Then we may obtain the solution in the physical $\boldsymbol{z}$ coordinates via $\boldsymbol{z} = \boldsymbol{V u}$. Writing out $\boldsymbol{V}$ in terms of its columns via

$$\boldsymbol{V} = \begin{bmatrix} | & | & | & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4 \\ | & | & | & | \end{bmatrix} \tag{9}$$

we have that

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{V u} \\ &= \left(e^{(-0.005+1i)t} d_1 \right) \boldsymbol{v}_1 + \left(e^{(-0.005-1i)t} d_2 \right) \boldsymbol{v}_2 \\ &\quad \left(e^{-1.01t} d_3 \right) \boldsymbol{v}_3 + \left(e^{-98.99t} d_4 \right) \boldsymbol{v}_4 \end{aligned} \tag{10}$$

Now notice that in the true solution, the $e^{-98.99t} d_4 \boldsymbol{v}_4$ term will decay extremely quickly, so that at any instant in time the solution is practically described by the first three terms. This is borne out in figure 1—there are no rapid transients in sight, and the solution is instead dominated by the smooth behavior contained in the other three eigenvectors. Yet, it is precisely this pesky fourth term that is imposing our stability restriction, and thereby requiring for Heun's method that we use a time step well below what is needed to capture the essential behavior of the system. Of course, the trapezoid method does not contain these issues, as its stability region encompasses the entire left-half plane. Indeed, we may surmise from this example problem that methods with large stability regions are better suited to stiff problems.

2 *Characterizing stiff problems*

With this example under our belts to ground us, let us characterize stiff systems more generally.

Characterization of stiffness for IVPs

The stiffness of an IVP $\dot{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{z}$ is characterized by its "stiffness ratio",

$$\mathcal{R}_s = \frac{\max_{j=1,\dots,n}(|\lambda_j|)}{\min_{j=1,\dots,n}(|\lambda_j)|} \tag{11}$$

where n is the dimension of the IVP. The IVP is called stiff if $\mathcal{R}_s \gg 1$.

To extend this definition to nonlinear IVPs, we may consider a Jacobian matrix $\boldsymbol{A}$ that arises from linearizing the nonlinear term about some reference state.

Notice that it is the ratio of the largest to the smallest eigenvalues that is important. In the motivating example at the beginning of this lecture, if all of the eigenvalues had been -100, then we would have required a small time step regardless of our choice of method for *accuracy reasons*. This is because in this case the dynamics of the system are actually occurring quickly and must be resolved.

The crux of a stiff system is that the dynamics of interest are evolving slowly relative to some fast time scale. It is for these problems that methods with small absolute stability regions have severe time step restrictions that are not coincident with the time step size one would want to use to accurately resolve the relatively slow dynamics that govern the actual system response.

3 *Finite difference methods for stiff problems*

As we already discovered, a key feature to efficiently simulating stiff IVPs is to select a method with a large stability region. The backwards Euler method and trapezoid method are therefore good choices for solving stiff problems.

There are another class of methods, backwards differentiation formula (BDF) methods, that are well-suited to stiff problems. BDF methods are multi-step methods that are derived directly from the differential form of the IVP $\dot{\boldsymbol{u}} = \boldsymbol{f}(\boldsymbol{u}, t)$, rather than from the integrated form $\int_{t_k}^{t_{k+1}} \dot{\boldsymbol{u}}(t)\ dt = \int_{t_k}^{t_{k+1}} \boldsymbol{f}(\boldsymbol{u}(t), t)\ dt$. Specifically, an r-step BDF may be derived by writing

$$\dot{\boldsymbol{u}}(t_{k+1}) = \boldsymbol{f}(\boldsymbol{u}(t_{k+1}), t_{k+1}) \tag{12}$$

and expressing $\boldsymbol{u}(t)$ as a degree $r+1$ polynomial using the Lagrange basis functions that we have come to know and love. Evaluating the polynomial approximation of $\boldsymbol{u}(t)$ at $t = t_{k+1}$ leads to a finite

difference formula

$$\sum_{j=k-r+1}^{k+1} \alpha_{j-(k-r+1)} \boldsymbol{u}_j = \Delta t \beta_r \boldsymbol{f}(\boldsymbol{u}_{k+1}, t_{k+1}) \tag{13}$$

That is, $\beta_0 = \beta_1 = \cdots = \beta_{r-1} = 0$.

We will discuss why these methods are valuable for stiff problems in a moment, but let us first consider some examples:

1-STEP BDF METHOD (BACKWARDS EULER):

$$\boldsymbol{u}_{k+1} - \boldsymbol{u}_k = \Delta t \boldsymbol{f}(\boldsymbol{u}_{k+1}, t_{k+1}) \tag{14}$$

2-STEP BDF METHOD:

$$3\boldsymbol{u}_{k+1} - 4\boldsymbol{u}_k + \boldsymbol{u}_{k-1} = 2\Delta t \boldsymbol{f}(\boldsymbol{u}_{k+1}, t_{k+1}) \tag{15}$$

3-STEP BDF METHOD:

$$11\boldsymbol{u}_{k+1} - 18\boldsymbol{u}_k + 9\boldsymbol{u}_{k-1} - 2\boldsymbol{u}_{k-2} = 6\Delta t \boldsymbol{f}(\boldsymbol{u}_{k+1}, t_{k+1}) \tag{16}$$

4-STEP BDF METHOD:

$$25\boldsymbol{u}_{k+1} - 48\boldsymbol{u}_k + 36\boldsymbol{u}_{k-1} - 16\boldsymbol{u}_{k-2} + 3\boldsymbol{u}_{k-2} = 12\Delta t \boldsymbol{f}(\boldsymbol{u}_{k+1}, t_{k+1}) \tag{17}$$

Why are these methods useful for stiff problems? Figure 2 demonstrates the absolute stability regions for BDF-2, BDF-3, and BDF-4 (the plot for BDF-1, the backwards Euler method, is in the previous lecture).

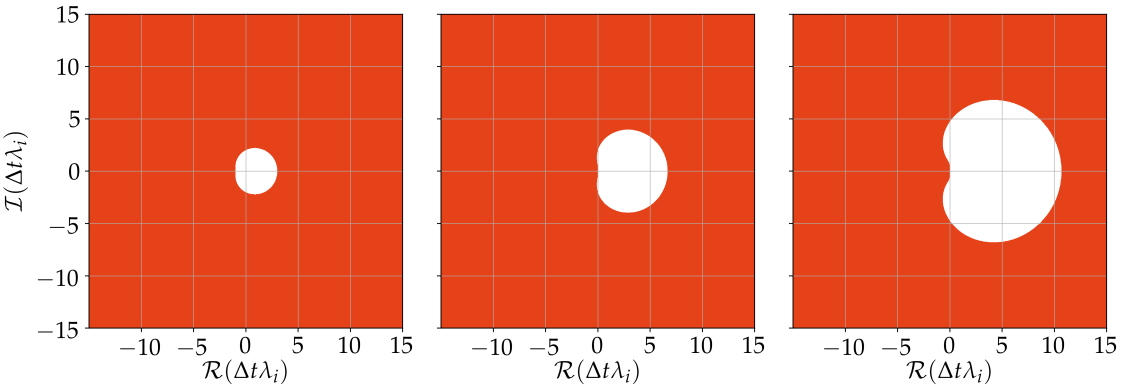


Figure 2: Stability region for the 2-step (a), 3-step (b), and 4-step (c) BDF methods. Note the larger scale of the axes compared with the absolute stability plots provided in previous lectures. This is done to more easily visualize the stability regions of these BDF methods.

The figure demonstrates that the core utility of BDF methods for stiff IVPs is in their expansive absolute stability regions. Why is it that these methods possess such desirable stability properties? We may intuit the answer by considering the stability of these methods in

the limit $\Delta t\lambda_i \to \infty$. If we plug in the various α and β coefficients for a BDF method into the absolute stability relation from the previous lecture, we arrive at

$$\sum_{j=0}^{r-1} \alpha_j \zeta^j + \Big[\alpha_r - \Delta t \beta_r \lambda_l\Big] \zeta^r = 0 \tag{18}$$

We can think of this limit as embodying dynamics with infinitely fast time scale. Thus, methods that are stable in this limit are ideal for handling stiff systems.

Dividing this expression by $\Delta t\lambda_l$ and taking the limit as $\Delta t\lambda_l \to \infty$, the expression reduces to $\beta_r \zeta^r = 0$. All roots of this expression are at zero. Thus, every BDF method is stable in the limit as $\Delta t\lambda_l \to \infty$, and is therefore a prime candidate for handling stiff IVPs.

It turns out that the r-step BDF method is $O(\Delta t^r)$ accurate (you may inspect this for yourself for BDF-1–BDF-4 by evaluating their truncation error). However, it is only possible to use up to a 6-step BDF method; the BDF methods are *not* convergent for $r > 6$ because $\Delta t\lambda = 0$ is not included in the absolute stability region for these cases. This serves as a testament to understanding the concept of convergence. Without this, one might blindly try to apply an 8-step BDF method, to catastrophic ends!