Chapter 17

Backward Error Analysis

Main concepts: Example with forward euler and arbitrary f. Forward, backward, symplectic Euler and pendulum. Symplectic integrators.

Up to this point, the error analysis in this course has considered the difference between the numerical sequence y_n, $n = 0, 1, \ldots$ and the exact solution $y(t_n)$ of the differential equation (1.3) passing through either the initial point $y(0) = y_0$ (global error) or through an earlier point on the numerical solution, $y(t_m) = y_m$, $m < n$ (residual or local error). This approach to error analysis yields quantitative statements about the solution, but does not tell us much about the qualitative behavior of a method.

17.1 A Modified Differential Equation

In this section we introduce an alternative approach to error analysis, in which we attempt to find a different differential equation

$$\tilde{y}' = \tilde{f}(\tilde{y}) \tag{17.1}$$

for which the numerical solution is exact: $y_n = \tilde{y}(t_n)$. This approach is called *backward error analysis*, and the differential equation (17.1) is called the *modified equation*.

17.1.1 A modified equation

Let us procede with an example. Since for any convergent method, we should have $\tilde{f} \to f$ as $h \to 0$, let us try to find a vector function $f_1(y)$ such that

$$\tilde{f}(\tilde{y}) := f(\tilde{y}) + hf_1(\tilde{y}) \tag{17.2}$$

more accurately 'approximates' the numerical solution generated by forward Euler(2.1).

The solution to the modified equation (17.1) for this $\tilde{f}$ over a step of size h, can be expanded in a Taylor series. Denoting $\tilde{y}_n \equiv \tilde{y}(t_n)$ and the Jacobian of f by $f'(y)$,

$$\begin{aligned}
\tilde{y}_{n+1} &= \tilde{y}_n + h\left(f(\tilde{y}_n) + hf_1(\tilde{y}_n)\right) + \frac{h^2}{2}\left(f'(\tilde{y}_n) + hf_1'(\tilde{y}_n)\right)\tilde{f}(\tilde{y}_n) + \mathcal{O}(h^3) \\
\tilde{y}_{n+1} &= \tilde{y}_n + h\left(f(\tilde{y}_n) + hf_1(\tilde{y}_n)\right) + \frac{h^2}{2}\left(f'(\tilde{y}_n) + hf_1'(\tilde{y}_n)\right)\left(f(\tilde{y}_n) + hf_1(\tilde{y}_n)\right) + \mathcal{O}(h^3) \\
\tilde{y}_{n+1} &= \tilde{y}_n + hf(\tilde{y}_n) + h^2\left[f_1(\tilde{y}_n) + \frac{1}{2}f'(\tilde{y}_n)f(\tilde{y}_n)\right] + \mathcal{O}(h^3)
\end{aligned} \tag{17.3}$$

For forward Euler, we have

$$y_{n+1} = y_n + hf(y_n),$$

so, by choosing the term in brackets in (17.3) to be zero, i.e.

$$f_1(\tilde{y}) = -\frac{1}{2}f'(\tilde{y})f(\tilde{y}), \tag{17.4}$$

we see that forward Euler applied to (1.3), while being only a first order accurate approximation to the original ODE, is actually producing a *second order* accurate approximation to the modified system (17.1), (17.2), (17.4).

We can expect therefore, that statements made about the modified equation (17.1) with this choice of $\tilde{f}$ will give us additional insight into the qualitative behavior of the forward Euler method, more so than statements made about the original problem (1.3).

Example. Consider the initial value problem

$$y' = y^2, \quad y(0) = 1$$

with exact solution

$$y(t) = \frac{1}{1-t}.$$

The exact solution exists only for $t < 1$. On the other hand, if we apply forward Euler to this problem, the sequence

$$y_{n+1} = y_n + hy_n^2$$

does exist for all time (On a computer, one would eventually get an overflow error).

The modified differential equation is

$$\tilde{y}' = f(\tilde{y}) - \frac{h}{2}f'(\tilde{y})f(\tilde{y}) = \tilde{y}^2 - h\tilde{y}^3.$$

This system has an unstable equilibrium at $y = 0$ and an asymptotically stable equilibrium at $y = 1/h$. In particular, a solution exists for all time.

Figure 17.1 shows the exact solution, the forward Euler solution and the modified equation solution for $h = 0.1$. The modified equation is not much closer to the numerical solution than the exact solution is, but it does exist for all time.

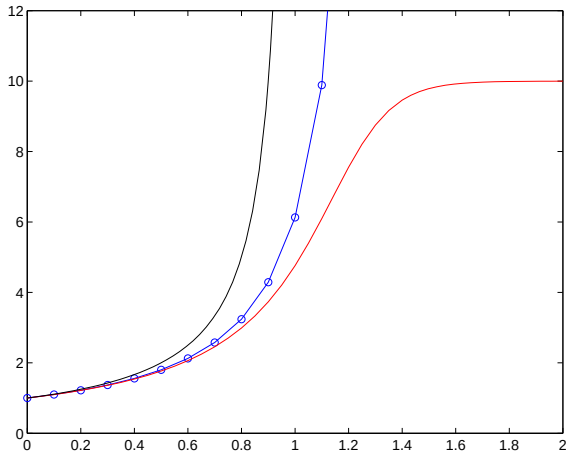

Figure 17.1: Exact (black), forward Euler (blue), and 1-term modified equation solutions to $y' = y^2$, $y(0) = 1$.

17.1.2 More modified equations

To get even more accurate modified equations, we can include higher order perturbations in (17.2) such that the numerical method becomes a better and better approximation. The m-term modified equation is

$$\tilde{f}(\tilde{y}) := f(\tilde{y}) + hf_1(\tilde{y}) + h^2 f_2(\tilde{y}) + \cdots + h^2 f_m(\tilde{y}) \tag{17.5}$$

We may also talk about the modified equation for $m \to \infty$, formally, without any specification of whether or not the series converges. In fact, we have a hierarchy of modified equations, for $m = 1, 2, \ldots$.

Let us introduce some notation. We denote the Jacobian of $f(y)$ by f', the (symmetric) bilinear form obtained by differentiating $f'(y)$ with respect to y by $f''(\cdot,\cdot)$, etc. Then,

$$\begin{aligned}
\frac{d}{dt} f(y(t)) &= f'f \\
\frac{d^2}{dt^2} f(y(t)) &= f''(f,f) + f'f'f \\
\frac{d^3}{dt^3} f(y(t)) &= f'''(f,f,f) + 3f''(f'f,f) + f'f''(f,f) + f'f'f'f. \\
&\vdots
\end{aligned}$$

We procede to determine f_2. We write

$$\tilde{y}(t_{n+1}) = \tilde{y}(t_n) + h\tilde{f} + \frac{h^2}{2}\frac{d}{dt}\tilde{f} + \frac{h^3}{3!}\frac{d^2}{dt^2}\tilde{f}, + \mathcal{O}(h^4), \tag{17.6}$$

where all terms on the right side are evaluated at t_n. Then, since

$$\begin{aligned}
\frac{d}{dt}\tilde{f} &= f'\tilde{f} + hf_1'\tilde{f} + \mathcal{O}(h^2) \\
&= f'f + hf'f_1 + hf_1'f + \mathcal{O}(h^2)
\end{aligned}$$

and

$$\begin{aligned}
\frac{d^2}{dt^2}\tilde{f} &= f''(\tilde{f},\tilde{f}) + f'\tilde{f}'\tilde{f} + \mathcal{O}(h) \\
&= f''(f,f) + f'f'f + \mathcal{O}(h),
\end{aligned}$$

(17.6) becomes

$$\begin{aligned}
\tilde{y}(t_{n+1}) &= \tilde{y}(t_n) + h\left[f + hf_1 + h^2 f_2\right] + \frac{h^2}{2}\left[f'f + hf'f_1 + hf_1'f\right] + \frac{h^3}{3!}\left[f''(f,f) + f'f'f\right] + \mathcal{O}(h^4) \\
&= \tilde{y}(t_n) + hf + h^2\left[f_1 + \frac{1}{2}f'f\right] + h^3\left[f_2 + \frac{1}{2}f'f_1 + \frac{1}{2}f_1'f + \frac{1}{6}f''(f,f) + \frac{1}{6}f'f'f\right] + \mathcal{O}(h^4).
\end{aligned} \tag{17.7}$$

The modified equation agrees with the iterates of forward Euler if both terms in brackets above are equal to zero. The first term just gives (17.4). The second term can be solved for f_2 to give

$$f_2 = -\frac{1}{2}(f'f_1 + f_1'f) - \frac{1}{6}\left[f''(f,f) + f'f'f\right]. \tag{17.8}$$

Example. Continuing the previous example, by substituing $f = \tilde{y}^2$ and $f_1 = -\tilde{y}^3$ into (17.8) the second term is

$$f_2 = \frac{3}{2}\tilde{y}^4.$$

Continuing the procedure outlined above to determine higher order terms (and making use of symbolic mathematics software) gives the five-term modified equation

$$\tilde{y}' = \tilde{y}^2 - h\tilde{y}^3 + h^2\frac{3}{2}\tilde{y}^4 - h^3\frac{8}{3}\tilde{y}^5 + h^4\frac{31}{6}\tilde{y}^6 - h^4\frac{157}{15}\tilde{y}^7.$$

The m-term modified equations for $m = 1, \ldots, 5$ are plotted in Figure 17.2 for $h = 0.05$ and $h = 0.02$. The modified equations with m odd each have an attracting fixed point proportional to $1/h$. Note in the figure that the modified equation solutions 'converge' to the numerical solution very quickly as h become smaller.

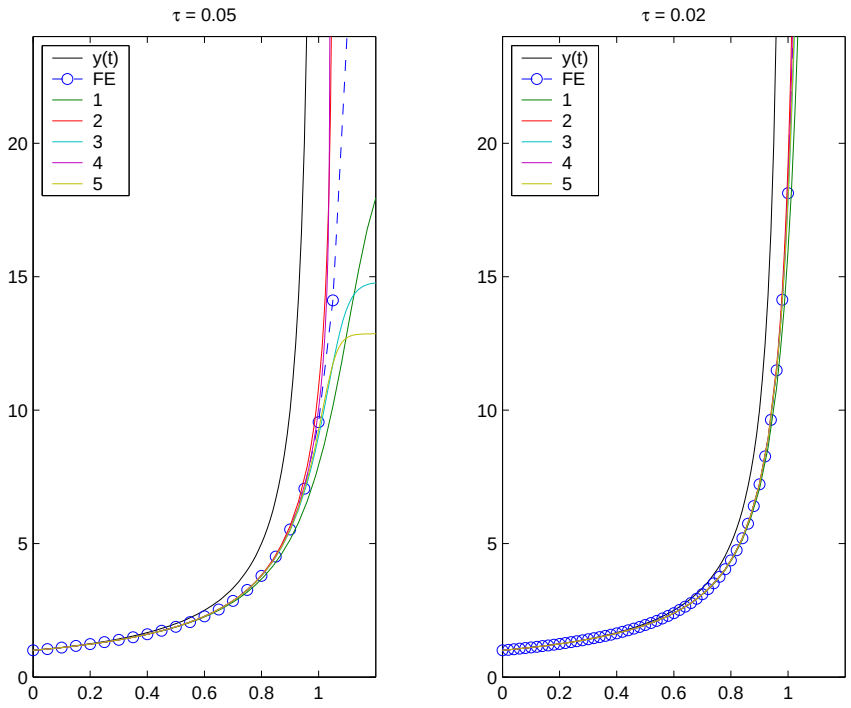

Figure 17.2: Forward Euler and m-term modified equation solutions to $y' = y^2$, $y(0) = 1$. On the left $h = 0.05$ and on the right $h = 0.02$.

Example. As a second example consider the nonlinear pendulum. Let q be the angle, measured clockwise from the vertical at the lowest point, and let $p = \dot{q}$. We assume a scaling of variables such that mass and gravitational acceleration are equal to one. The equations of motion are

$$\begin{aligned} \dot{q} &= p \\ \dot{p} &= -\sin q. \end{aligned}$$

And the system conserves the total energy

$$H = \frac{p^2}{2} - \cos q. \tag{17.9}$$

We will compare three methods for this problem, forward Euler, backward Euler, and the following method, which we refer to as *symplectic Euler*

$$\begin{aligned} q_{n+1} &= q_n + hp_n \\ p_{n+1} &= p_n - h\sin q_{n+1}. \end{aligned}$$

Note that this method can be computed explicitly even though q is evaluated at time level $n+1$ on the right. The method is first order accurate.

The solutions obtained with all three methods are plotted in Figure 17.3. The black curves are level curves of the energy (17.9). Most solutions are periodic; the pendulum either swings back and forth or around and around. There is a center equilibrium point at the bottom of its arc and a saddle point at the top. The saddle points are connected by separatrices.

The numerical solutions with all three methods were computing using stepsize $h = 0.1$ for 200 steps, with initial condition $q(0) = -3\pi/4$, $p(0) = 0$. The exact solution through this is point is a closed energy level set, the pendulum swings back and forth. The forward Euler solution is seen to spiral outwards, gaining speed until it crosses the separatrix. The backward Euler solution spirals inward towards the center equilibrium point. The symplectic Euler solution appears to be a closed preiodic orbit. What does backward error analysis tell us about this problem?

The Jacobian for the pendulum is

$$f'(q,p) = \begin{bmatrix} 0 & 1 \\ -\cos q & 0 \end{bmatrix}.$$

The $\mathcal{O}(h)$ perturbation in the modified vector field for the forward Euler method is

$$f_1 = -\frac{1}{2} f' f = \begin{pmatrix} -\frac{1}{2}\sin q \\ -\frac{1}{2} p \cos q \end{pmatrix}.$$

This vector field is plotted in the leftmost frame of Figure 17.4. Note that the component of the vector field that is normal to the level sets of constant energy is everywhere pointing outwards. The center point becomes an unstable equilbrium. The effect of this perturbation is an increase in energy along each trajectory, which is consistent with the behavior exibited by forward Euler.

It may come as no surprise that the $\mathcal{O}(h)$ perturbation of backward Euler is just the negative of that of forward Euler, as was also the case for the principle term in the local error of backward Euler. We have

$$f_1 = \frac{1}{2} f' f = \begin{pmatrix} \frac{1}{2}\sin q \\ \frac{1}{2} p \cos q \end{pmatrix},$$

and this vector field is plotted in the second frame of Figure 17.4. In this case the perturbation causes a decrease in energy along each trajectory, and the center point becomes a stable equilibrium.

For the symplectic method, we expand the numerical solution for p_{n+1} about time t_n

$$\begin{aligned} q_{n+1} &= q_n = h p_n \\ p_{n+1} &= p_n - h(\sin q_n + h\dot{q}_n \cos q_n + \mathcal{O}(h^2)) \\ &= p_n - h \sin q_n - h^2 p_n \cos q_n + \mathcal{O}(h^3). \end{aligned}$$

The term f_1 must to be chosen such that the first term in brackets in (17.7) agrees with this:

$$f_1(q_n, p_n) + \frac{1}{2} f'(q_n, p_n) f(q_n, p_n) = \begin{pmatrix} 0 \\ -p_n \cos q_n \end{pmatrix},$$

that is,

$$f_1(q,p) = \begin{pmatrix} \frac{1}{2}\sin q_n \\ -\frac{1}{2} p_n \cos q_n \end{pmatrix}.$$

The pendulum equations can be written in the form of a skew-symmetric matrix times a gradient, as in the previous section, to make it more explicit that (17.9) is conserved. Let $y = (q,p)^T$; then

$$y' = J\nabla H(y), \quad J := \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}.$$

It is interesting that the modified vector field f_1 can also be written as a product of J and a gradient, namely,

$$f_1 = J\nabla H_1(y), \quad H_1(q,p) = \frac{1}{2} p \sin q$$

It follows that the 1-term modified equation is

$$\tilde{y}' = J\nabla \left(H(\tilde{y}) + hH_1(\tilde{y}) \right),$$

which also has a conserved quantity $\tilde{H} = H + hH_1$. In particular, this means that the 1-term modified equation also has periodic solutions. The level sets of $\tilde{H}$ are plotted in the third frame of Figure 17.4, using a larger stepsize perturbation $h = 0.2$ to exaggerate the effect.

Note that $\tilde{H}$ is *not* a conserved quantitity of the numerical method, since we have only used a single term in the modified equation expansion. However, since the symplectic Euler method is a second order accurate approximation to this modified equation, we would expect that $\tilde{H}$ to be better preserved than H.

We could continue to develop higher and higher order modified equations for the nonlinear pendulum, and we would find that *each of them can be written as a product of J and a gradient.* In other words, the symplectic Euler method has a modified equation of the form

$$\tilde{y}' = J\nabla \left(H(\tilde{y}) + hH_1(\tilde{y}) + h^2 H_2(\tilde{y}) + \cdots \right).$$

For this problem, this result suggests that the numerical solution lies on a closed loop in phase space.

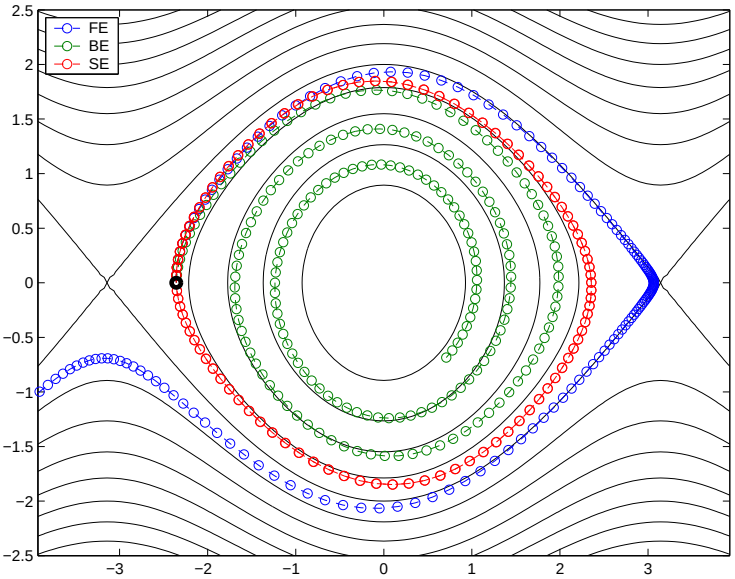

Figure 17.3: Numerical solutions of the nonlinear pendulum superimposed over the energy level sets.

17.2 Good News and Bad News

In this section we give some of the theory of backward error analysis, without proofs.

First the bad news: the sequence of modified equations (17.5) for $m = 1, 2, \ldots$ *diverges* in general. This is suggested by the left frame in Figure 17.2, where for m odd, the modified equations have asymptotic fixed points which are decreasing in magnitude.

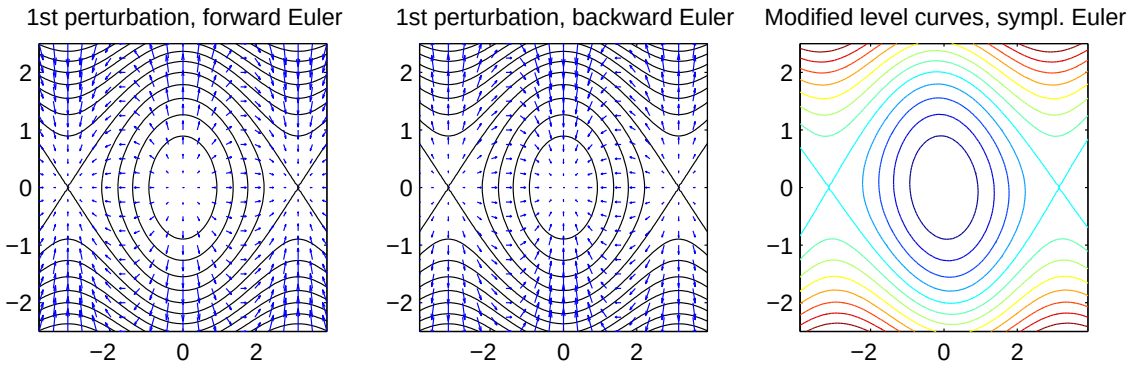

Figure 17.4: Vector field perturbation f_1 for forward Euler (left) and backward Euler (center). On the right are the energy level sets of the 1-term modified equation for the symplectic Euler method.

Next the good news: there exists an optimal value of m, dependent on h and denoted m_h^*, for which the difference between the m-term modified equation and the numerical solution is minimized. m_h^* increases like $1/h$ as h tends to zero, and is usually much larger than the order p of the numerical method. In other words, the modified equations are indeed a useful tool in understanding numerical methods.

Specifically, let $B_R(y_0) = \{y \in \mathbb{C}^d : \|y - y_0\| \leq r\}$.

Theorem 17.2.1 *Suppose f is analytic in $B_{2R}(y_0)$ and satisfies $\|f(y)\| < M$ there. Also suppose the coeffecients of the expansion of the numerical method about t_n is analytic and bounded on B_R. Then there exists a constant γ dependent only on the method, and a timestep $h_0 \propto R/M$ such that, for $h < h_o/4$, there is m_h^* equal to the largest integer m satisfying $hm \leq h_0$, satisfying*

$$\|\Psi_h y_0 - \Phi_{m_h^*}(h) y_0\| \leq h\gamma m_h^* e^{-h_0/h},$$

where Ψ_h is the numerical solution map for a step of size h, and $\Phi_{m_h^*}(h)$ is the exact solution of the m_h^*-term modified equation (17.5) over an interval of length h.

Backward error analysis has been applied successfully to numerical methods that preserve a Lie group, in which case it can be shown by induction that the modified equations preserve the Lie group as well.

A different version of backward error analysis introduces a non-autonomous modified differential equation (effectively increasing the problem dimension by one). This approach has been used to show under what conditions methods applied to dissipative problems and gradient flows will preserve the properties of the analogous continuous solutions. See the book by Stuart & Humphries (Cambridge, 1996).