ERROR ESTIMATES FOR NUMERICAL METHODS: PROOFS.

The proofs of the basic error estimates are not hard, relying mainly on Taylor approximation. We consider time-dependent vector fields $f(t, y)$ with p components, defined for all $y \in \mathbb{R}^p$. The 'local Lipschitz condition' is important: f is 'locally Lipschitz' if for every $R > 0$, it is Lipschitz with constant $L > 0$ (depending on R) in the ball of radius R:

$$|z_1| \leq R, |z_2| \leq R \Rightarrow |f(t, z_1) - f(t, z_2)| \leq L|z_1 - z_2|.$$

It is a standard result that f is locally Lipschitz if it is continuously differentiable (in y) (i.e., of 'class C^1'.)

Theorem 1 (Error estimate for Euler's method). Consider the initial-value problem:

$$y' = f(t, y), \quad y(t) \in \mathbb{R}^p, \quad y(a) = y_0.$$

Assume $f(t, y)$ is continuously differentiable in y and that an exact solution $y(t)$ exists, defined for $t \in [a, b]$. Let $N \in \mathbb{N}$, $h = (b - a)/N$ (the step size) and consider the recursion:

$$t_{n+1} = t_n + h, \quad t_0 = a, \quad y_{n+1} = y_n + f(t_n, y_n), \quad n = 1, \ldots, N - 1.$$

Let $e_n = y(t_n) - y_n$ be the approximation error at t_n, $n = 0, \ldots, N$ ($e_0 = 0$). Then there exist constants $C > 0$ and $N_0 > 0$ so that, for $N > N_0$:

$$|e_n| \leq Ch, \quad n = 0, \ldots, N.$$

Proof. Choose $R > 0$ so that $\max_{t \in [a,b]} |y(t)| \leq R$. Let $L > 0$ be a Lipschitz constant for f (in the variable y) in the ball $\{|y| \leq 2R\}$. Then: By Taylor's theorem:

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + r_n, \quad |r_n| \leq ch^2,$$

where c depends on $y_{|[a,b]}$. Subtracting from this the recursion relation $y_{n+1} = y_n + hf(t_n, y_n)$, and using the fact that $y(t)$ is a solution of the ODE, we have for $n = 0, \ldots, N - 1$:

$$e_{n+1} - e_n = h(y'(t_n) - f(t_n, y_n)) + r_n = h(f(t_n, y(t_n)) - f(t_n, y_n)) + r_n.$$

Now assume (inductively) the following hold:

$$|y_n| \leq 2R, \quad |e_n| \leq \frac{ch}{L}[(1 + hL)^n - 1].$$

We claim these estimates still hold for y_{n+1} and e_{n+1}. Indeed, since $y(t_n)$ and y_n are both in the ball $\{|y| \leq 2R\}$, we may use the Lipschitz condition to conclude, for $n = 0, \ldots, N-1$:

$$|f(t_n, y(t_n)) - f(t_n, y_n)| \leq L|y(t_n) - y_n|,$$

and hence:

$$|e_{n+1} - e_n| \leq hL|e_n| + ch^2,$$

$$|e_{n+1}| \leq (1+Lh)|e_n| + ch^2 \leq (1+Lh)\frac{ch}{L}[(1+hL)^n - 1] + ch^2 = \frac{ch}{L}[(1+hL)^{n+1} - 1];$$

in particular, since:

$$\frac{c}{L}(1+hL)^{n+1} \leq \frac{c}{L}e^{(n+1)hL} \leq \frac{c}{L}e^{(b-a)L} := C,$$

we have $|e_{n+1}| \leq Ch$, so:

$$|y_{n+1}| \leq |y(t_{n+1})| + |e_{n+1}| \leq R + Ch \leq 2R,$$

provided $h < h_0$, or $N > N_0$, where h_0 (or N_0) depends on a, b, c, L and R. This completes the induction step. As just seen, this implies, for $n = 0, \ldots, N$:

$$|e_n| \leq Ch, \quad C = \frac{c}{L}e^{(b-a)L},$$

as we wished to show.

Remark 1. (*Dependence of C and N_0.*) There are good reasons (other than OCD personality) to keep track of what constants depend on. From the proof, we see that C and N_0 depend on $b - a$, L (Lipschitz constant of f over B_{2R}, a ball containing the solution) and c. It is not hard to see that c and L are controlled by the 'C^1 norm' of f over B_{2R}. Similar comments apply to the constants in theorems 2 and 3 that follow, and will be omitted. (It is a fact of life that mathematics, done properly, is full of little details like this- mostly left unwritten, since the people who know enough to care about them can usually fill in the gaps themselves. It does make things tricky for beginners, though.)

Remark 2. The one 'tricky' point in the proof is guessing the form of M_n in the bound $|e_n| \leq M_n h$, which is proved inductively based on: $|e_{n+1}| \leq (1+hL)|e_n| + ch^2$. How did we guess that $M_n = [(1+hL)^n - 1]\frac{c}{L}$ would work?

The estimate for $|e_{n+1}|$ in terms of $|e_n|$ suggests the recursion relation:

$$M_{n+1} = (1+hL)M_n + ch, \quad M_0 = 0,$$

and this leads to the 'linear difference equation':

$$M_{n+1} - M_n = hLM_n + ch, \quad M_0 = 0,$$

which is completely analogous to a (non-homogeneous) linear first-order DE. It has a constant solution $M_n \equiv -c/L$, and the 'general solution of the homogeneous equation' is $M_n = C(1+hL)^n$, so the solution of the difference equation with IC $M_0 = 0$ is:

$$M_n = (1 + hL)^n \frac{c}{L} - \frac{c}{L},$$

which is exactly the expression 'guessed'. This observation will be useful when we repeat the trick in the next proof.

Theorem 2. (Error estimate for the midpoint Euler method.) Let $f(t, y)$, $y \in \mathbb{R}^p$, be twice continuously differentiable in (t, y) (in particular the partial derivatives f_t and $d_y f = f_y$ are locally Lipschitz in y). Assume the initial-value problem:

$$y' = f(t, y), \quad y(a) = y_0$$

has a solution defined in the interval $[a, b]$. Let $N \in \mathbb{N}$, $h = (b - a)/N$ and consider the recurrence relation ('discrete evolution'): $t_0 = a$,

$$t_{n+1} = t_n + h, \quad y_{n+1} = y_n + hf(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)), \quad n = 0, \ldots, N - 1.$$

Let $e_n = y(t_n) - y_n$ be the approximation error $(n = 0, \ldots, N, e_0 = 0.)$ Then there exist constants $C > 0$, $N_0 > 0$, so that for N sufficiently large we have:

$$|e_n| \leq Ch^2, \quad n = 0, \ldots, N.$$

Proof. Taylor's theorem applied to the solution $y(t)$ gives:

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \frac{h^2}{2}(f_t + f_y[f])_{|(t_n, y(t_n))} + c_1 h^3.$$

We also have the first-order Taylor approximation:

$$f(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)) = f(t_n, y_n) + \frac{h}{2}(f_t + f_y[f])_{|(t_n, y_n)} + c_2 h^2.$$

(Note that the meaning of $f_y[f]$ is $d_y f[f]$, for a time-dependent vector field $f(t, y)$ in $\mathbb{R}^p$.) Thus:

$$y_{n+1} = y_n + hf(t_n, y_n) + \frac{h^2}{2}(f_t + f_y[f])_{|(t_n,y_n)} + c_2h^3.$$

Subtracting the expression for y_{n+1} from that for $y(t_{n+1})$, we obtain:

$$e_{n+1} = e_n + h[f(t_n, y(t_n)) - f(t_n, y_n)] + \frac{h^2}{2}(f_t + f_y[f])_{|(t_n,y(t_n))} - \frac{h^2}{2}(f_t + f_y[f])_{|(t_n,y_n)} + (c_1 - c_2)h^3.$$

Denoting by $L_{f_t}, L_{f_y[f]}$ the respective Lipschitz constants (in a ball of radius $2R$) of $f_t, f_y[f]$, we conclude (assuming, of course, $y(t_n)$ and y_n are in this ball):

$$|e_{n+1} - e_n| \leq hL_f|e_n| + \frac{h^2}{2}(L_{f_t} + L_{f_y[f]})|e_n| + ch^3,$$

and with L denoting the sum of the Lipschitz constants:

$$|e_{n+1}| \leq (1 + hL)|e_n| + ch^3,$$

assuming $h < 1$. Now we're in the same situation as in the previous proof. Choose $R > 0$ so that $|y(t)| \leq R$ for $t \in [a, b]$. Assuming, for a given $n = 0, \ldots, N - 1$:

$$|y_n| \leq 2R, \quad |e_n| \leq M_n h^2, \quad M_n = [(1 + hL)^n - 1]\frac{c}{L},$$

we show the same bounds hold for $n + 1$. Indeed the bound for $|e_n|$ implies:

$$|e_{n+1}| \leq (1+Lh)|e_n| + ch^3 \leq (1+Lh)\frac{ch^2}{L}[(1+hL)^n - 1] + ch^3 = \frac{ch^2}{L}[(1+hL)^{n+1} - 1];$$

in particular:

$$|e_{n+1}| \leq Ch^2, \quad C = \frac{c}{L}e^{(b-a)L},$$

and this implies:

$$|y_{n+1}| \leq |y(t_{n+1})| + |e_{n+1}| \leq R + Ch^2,$$

which is smaller than $2R$ if the step size h is chosen small enough (or N is chosen large enough). We conclude:

$$|e_n| \leq Ch^2, \quad n = 0, \ldots, N,$$

as desired.

Remark. *What makes this proof work?* Consider the Taylor expansions of order two for the exact solution:

$$y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}y''(t) + O(h^3)$$

$$= y(t) + hf(t,y) + \frac{h^2}{2}(f_t + f_y[f])_{|(t,y)} + O(h^3),$$

and for the approximate solution:

$$\tilde{y}(t+h) = y(t) + hf(t + \frac{h}{2}, y + \frac{h}{2}f(t,y))$$

$$= y(t) + h[f(t,y) + \frac{h}{2}(f_t + f_y[f])_{|(t,y)}] + O(h^3).$$

Note that $y(t+h)$ and $\tilde{y}(t+h)$ coincide up to second order in h! (The approximation ends up being second order, and not third, since the errors potentially accumulate as we move right along the interval $[a,b]$, leading us to concede a factor of $N = \frac{b-a}{h}$.)

In general terms, the idea behind Runge-Kutta methods for ODE (which date back to the early 1900s) is to devise an approximate solution in which terms involving partial derivatives of f in the Taylor expansion (in h) of the exact solution are replaced (in the approximate solution) by 'nested evaluations' of f, in such a way that the Taylor expansion of the approximate solution coincides with that of the exact one, up to a given order.

Classical 4th. order Runge-Kutta.

The fourth-order Runge-Kutta method for the general first-order IVP is based on the recursion involving an average of four 'slopes' m_i, themselves obtained recursively:

$$t_{n+1} = t_n + h, \quad y_{n+1} = y_n + hm.$$

$$\begin{aligned} m &= \tfrac{1}{6}(m_1 + 2m_2 + 2m_3 + m_4) \\ m_1 &= f(t_n, y_n) \\ m_2 &= f(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2}m_1) \\ m_3 &= f(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2}m_2) \\ m_4 &= f(t_n + h, y_n + hm_3) \end{aligned}$$

Prior to proving a theorem on the error estimate for RK4, we examine how the 'Taylor series heuristic' of the last remark extends to suggest this method is fourth-order. For simplicity, we deal only with the autonomous case, $y' = f(y)$. Consider the fourth-order Taylor expansion of the solution:

$$y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \frac{h^3}{6}y^{(3)}(t) + \frac{h^4}{24}y^{(4)}(t) + O(h^5)$$

$$:= y(t) + (T_4 f)(y(t), h) + O(h^5).$$

This defines $(T_4 f)(y, h)$, whose coefficients are easily found in terms of f:

$$\begin{aligned}
&y' = f(y)\\
&y'' = f_y[f]\\
&y^{(3)} = f_{yy}(f, f) + f_y[f_y[f]]\\
&y^{(4)} = f_{yyy}(f, f, f) + 3f_{yy}(f_y[f], f) + f_y[f_{yy}(f, f)] + f_y[f_y[f_y[f]]]
\end{aligned}$$

We wish to compare this with the Taylor expansion of the approximate solution, where for the moment we assume the coefficient c_i of m_i is to be determined:

$$\tilde{y}(t+h) = y(t) + h(c_1 m_1 + c_2 m_2 + c_3 m_3 + c_4 m_4)$$

$$:= y(t) + T(y(t), h) + O(h^5).$$

This defines implicitly the notation $T(y, h)$, and to compute $T(y, h)$ explicitly in terms of f we obtain the Taylor approximations of the 'slopes' $m_i(h, y)$:

$$m_1 = f(y)$$

$$m_2 = f(y + \tfrac{h}{2}m_1) = f(y) + \tfrac{h}{2}f_y[f] + (\tfrac{h}{2})^2\tfrac{1}{2}f_{yy}(f, f) + (\tfrac{h}{2})^3\tfrac{1}{6}f_{yyy}(f, f, f) + O(h^4)$$

$$\begin{aligned}
&m_3 = f(y + \tfrac{h}{2}m_2) = f(y) + \tfrac{h}{2}f_y[f] + (\tfrac{h}{2})^2\tfrac{1}{2}f_{yy}(f, f) + \tfrac{h^2}{4}f_y[f_y[f]]\\
&+\tfrac{h^3}{16}f_y[f_{yy}(f, f)] + (\tfrac{h}{2})^2\tfrac{h}{2}f_{yy}(f, f_y[f]) + (\tfrac{h}{2})^3\tfrac{1}{6}f_{yyy}(f, f, f) + O(h^4)
\end{aligned}$$

$$\begin{aligned}
&m_4 = f(y + hm_3) = f(y) + hf_y[f] + \tfrac{h^2}{2}f_y[f_y[f]] + \tfrac{h^2}{2}f_{yy}[f, f]\\
&+\tfrac{h^3}{8}f_y[f_{yy}(f, f)] + \tfrac{h^3}{4}f_y[f_y[f_y[f]]] + +\tfrac{h^3}{2}f_{yy}(f, f_y[f]) + \tfrac{h^3}{6}f_{yyy}(f, f, f) + O(h^4)
\end{aligned}$$

We would like to choose the c_i so that $(T_4 f)(y, h) = T(y, h)$. Comparing the coefficients of 'like terms' in the two expansions, we arrive at the following system (equations listed with corresponding term):

$$
\begin{cases}
f: & c_1 + c_2 + c_3 + c_4 = 1 \\
f_y[f]: & \frac{1}{2}c_2 + \frac{1}{2}c_3 + c_4 = \frac{1}{2} \\
f_{yy}(f,f): & \frac{1}{4}c_2 + \frac{1}{8}c_3 + \frac{1}{2}c_4 = \frac{1}{6} \\
f_y[f_y[f]]: & \frac{1}{4}c_3 + \frac{1}{2}c_4 = \frac{1}{6} \\
f_{yyy}(f,f,f): & \frac{1}{48}c_2 + \frac{1}{48}c_3 + \frac{1}{6}c_4 = \frac{1}{24} \\
f_{yy}(f_y[f],f): & \frac{1}{8}c_3 + \frac{1}{2}c_4 = \frac{1}{8} \\
f_y[f_{yy}(f,f)]: & \frac{1}{16}c_3 + \frac{1}{8}c_4 = \frac{1}{24} \\
f_y[f_y[f_y[f]]]: & \frac{1}{4}c_4 = \frac{1}{24}
\end{cases}
$$

This system is over-determined, but one readily checks that $c_1 = c_4 = 1/6, c_2 = c_3 = 1/3$ is a solution. That is, the 'Simpson rule' coefficients for the m_i in m are exactly what is needed for the two Taylor expansions to coincide. And then an argument similar to that previously used proves the following theorem.

Theorem 3. (Error estimate for 4th order Runge-Kutta, autonomous case.) Let f be a C^4 vector field in $\mathbb{R}^p$ (in particular, all partial derivatives of f up to third order are locally Lipschitz). Assume the initial-value problem: $y' = f(y), \quad y(t) \in \mathbb{R}^p, \quad y(a) = y_0$, has a solution defined for $t \in [a,b]$. Given $N \in \mathbb{N}$, $h = (b-a)/N$, let $y_n, n = 0, \ldots, N$, $y_N = b$, be generated by the Runge-Kutta recursion: $\quad y_{n+1} = y_n + hm$, with $m = m(h, y_n)$ as defined above. Let $e_n = y(a+nh) - y_n$ be the error at the nth. step. Then there exist $N_0 > 0$ and $C > 0$ (independent of n) so that for $N > N_0$, we have:

$$|e_n| \leq Ch^4, \qquad n = 0, \ldots, N.$$

Proof. Let $t_n = a + nh$. We have the Taylor expansions:

$$y(t_{n+1}) = y(t_n + h) = y(t_n) + (T_4 f)(y(t_n), h) + c_1 h^5;$$

$$y_{n+1} = y_n + hm(h, y_n) = y_n + T(y_n, h) + c_2 h^5.$$

Since, as seen above, $T_4 f(y,h) = T(y,h)$, we have for the error:

$$e_{n+1} = e_n + (T_4 f)(y(t_n), h) - (T_4 f)(y_n, h) + (c_1 - c_2)h^5.$$

Provided $h < 1$, and assuming $|y_n| \leq 2R$ (where the ball of radius R contains $y(t), t \in [a,b]$), this gives an estimate in terms of the Lipschitz constant L of $T_4 f$ (as a function of y) in the ball B_{2R}:

$$e_{n+1} \leq (1 + Lh)|e_n| + ch^5.$$

As before, this is used to show inductively that $|y_n| \leq 2R$ and:

$$|e_n| \leq \frac{ch^4}{L}[(1 + hL)^n - 1] \leq Ch^4,$$

with $C := (c/L)\exp((b - a)L)$.