ODE weak approximation. Problem formulation.

Problem: Consider a solution $X : [0, T] \to \mathbb{R}^d$ of a system of ordinary differential equations, with flux

$$a : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d,$$

i.e.

$$\begin{cases} \dfrac{dX}{dt}(t) = a(t, X(t)), \quad 0 < t \leq T, \\ X(0) = X_0, \end{cases} \tag{38}$$

and a given general function $g : \mathbb{R}^d \to \mathbb{R}$.

We want to approximate the value $g(X(T))$

To this end, consider an approximation $\overline{X}$ of (38) by *any given numerical method*, satisfying the same initial condition

$$\overline{X}(0) = X(0) = X_0 \tag{39}$$

with time steps

$$0 = t_0 < \cdots < t_N = T.$$

This will imply a *global error*

$$g(X(T)) - g(\overline{X}(T)) \tag{40}$$

that we intend to estimate and control.

Example: Let X solve, for a given $A \in \mathbb{R}^{d\times d}$,

$$\begin{aligned} \frac{dX}{dt}(t) &= A\,X(t), \quad 0 < t \leq T, \\ X(0) &= X_0, \end{aligned} \tag{41}$$

and $g(X) = g \cdot X = \sum_{i=1}^{d} g_i X_i$, for a given vector $g \in \mathbb{R}^d$.

Then if $\overline{X}$ is given by the Forward Euler method,

$$\overline{X}(t_{n+1}) = (I + A\Delta t_n)\overline{X}(t_n),\ n = 0, \ldots, N-1$$

the global error becomes

$$\begin{aligned} g(X(T)) - g(\overline{X}(T)) =& g \cdot (X(T) - \overline{X}(T)) \\ =& g \cdot \left(e^{AT} - \prod_{n=1}^{N} (I + A\Delta t_n) \right) X_0 \end{aligned}$$

Adaptive Algorithms

Given a desired accuracy, $TOL > 0$ then find a time mesh for $[0, T]$,

$$0 = t_0 < t_1 < \ldots < t_N = T,$$

s.t. we can bound the time discretization error by

$$\left| g(X(T)) - g(\overline{X}_N) \right| \approx E_T \leq TOL,$$

while the computational work, which is proportional to the number of time steps, N, is minimized.

Simpler problem: Adaptive Integration

Given a function $X : [0,1] \to \mathbb{R}$, approximate

$$\boxed{g(X) = \int_0^1 X(t)dt}$$

Discretization Error with Forward Euler:

$$\int_0^1 X(t)dt - \sum_{n=0}^{N-1} X(t_n)\Delta t_n \simeq \sum_{n=0}^{N-1} \rho_n(\Delta t_n)^2 \quad \simeq \quad \text{TOL},$$

$$\textbf{Error density: } \rho_n \quad \equiv \quad \frac{X'(t_n)}{2}.$$

To find the Optimal Number of steps:

Solve the problem

$$\begin{cases} \min \ \text{Work} \sim N \\ \text{s.t.} \\ error \approx \text{TOL} \end{cases}$$

Uniform time steps: $$N_u \simeq \frac{1}{\text{TOL}} \|\rho\|_{L^1(0,1)}$$

Adaptive time steps: $$N_a \simeq \frac{1}{\text{TOL}} \|\rho\|_{L^{\frac{1}{2}}(0,1)}$$

Observe that on the last line, the notation

$\|\rho\|_{L^{\frac{1}{2}}(0,1)} \equiv \left(\int_0^1 \sqrt{\rho}\right)^2$ does not imply a norm.

Optimal Number of time steps

$$\text{Solve the problem (I)} * \begin{cases} \min \displaystyle\int_0^1 \frac{ds}{\Delta t(s)} \\ s.t. \\ \displaystyle\int_0^1 \rho(s)\Delta t(s) ds = \text{TOL} \end{cases}$$

For any $\lambda \in \mathbb{R}$, consider the *Lagrangian function*

$$\mathcal{L}(\Delta t, \lambda) \equiv \int_0^1 \frac{ds}{\Delta t(s)} + \lambda \left(\int_0^1 \rho(s)\Delta t(s) ds - \text{TOL} \right)$$

The first order optimality conditions for (I) are expressed in terms of the first variations to the Lagrangian $\mathcal{L}$ wrt Δt and λ:

$$\lim_{\epsilon \to 0} \frac{\mathcal{L}(\Delta t + \epsilon v, \lambda) - \mathcal{L}(\Delta t, \lambda)}{\epsilon} = 0, \quad \text{for all smooth } v$$

$$\lim_{\epsilon \to 0} \frac{\mathcal{L}(\Delta t, \lambda + \epsilon) - \mathcal{L}(\Delta t, \lambda)}{\epsilon} = 0.$$

The first optimality condition yields

$$-\int_0^1 \frac{v(s)ds}{(\Delta t(s))^2} + \lambda \int_0^1 \rho(s)v(s)ds = 0, \text{ for all smooth } v$$

while the second gives back the constraint

$$\int_0^1 \rho(s)\Delta t(s)ds = \text{TOL}$$

By Lagrangian relax. we obtain $\begin{cases} \Delta t^* = 1/\sqrt{\lambda\rho}, \\ \sqrt{\lambda} = \dfrac{1}{TOL}\displaystyle\int_0^1 \sqrt{\rho}ds \end{cases}$

and finally the optimal work

$$N_a \simeq \frac{1}{\text{TOL}}\left(\int_0^1 \sqrt{\rho}\right)^2 = \frac{1}{\text{TOL}}\|\rho\|_{L^{\frac{1}{2}}(0,1)}$$

Obs: To read about Lagrangian relaxation see Section 3.6 in the book *Introduction to Applied Mathematics*, by G. Strang.

Obs: Once we determine the *error density* ρ we can find the optimal choice for $\Delta t(s)$.

Question: How can we determine an approximation of ρ for ODE problems?

Does it pay to use adaptivity?

Adaptivity for singular functions Let

$$X(t) \equiv \frac{1}{\sqrt{t}},\ 0 < t \leq 1 \text{ and take } X_\epsilon(t) \equiv \frac{1}{\sqrt{t+\epsilon}}$$

$$\text{We want: } \left| \int_0^1 (X - X_\epsilon) \right| = o(\text{TOL}) \quad \Rightarrow \quad \epsilon^{1/2} \lesssim o(\text{TOL})$$

Uniform time steps:

$$N_u = \frac{1/4}{\text{TOL}} \int_0^1 \frac{dt}{(t+\epsilon)^{3/2}} \sim \frac{1}{\text{TOL}} \frac{1}{\epsilon^{1/2}} \gtrsim \mathcal{O}\left(\frac{1}{TOL^2}\right)$$

Adaptive time steps:

$$N_a = \frac{1/4}{\text{TOL}} \left(\int_0^1 \frac{dt}{(t+\epsilon)^{3/4}} \right)^2 \sim \mathcal{O}\left(\frac{1}{\text{TOL}}\right)$$

Optimal convergence rates p-th order discretization error:

$$|g(X(T)) - g(\overline{X}(T))| \simeq \sum_n \underbrace{\rho_n \Delta t_n^{p+1}}_{=r_n} \leq \text{TOL}$$

Uniform Discretization: $$N_u \simeq \frac{T}{\text{TOL}^{\frac{1}{p}}} \|\rho\|_{L^1}^{\frac{1}{p}}$$

Adaptive Discretization: $$N_a \simeq \frac{1}{\text{TOL}^{\frac{1}{p}}} \|\rho\|_{L^{\frac{1}{p+1}}}^{\frac{1}{p}}$$

Question: **Is there an algorithm achieving this?**

How can we define and estimate the error density, ρ_n, for ODEs?

ODE global error estimation goal

We want to derive estimates of the form

$$g(X(T)) - g(\overline{X}(T)) = \sum_{\text{time steps}} \text{local error} \cdot \text{weight} + \text{higher order error}. \tag{42}$$

with a given general function $g : \mathbb{R}^d \to \mathbb{R}$.

Obs: Ideally, the weight should not depend on the size of the time steps. This allows us to choose the time grid optimally (remember the previous integration example). The weight is the sensitivity of $g(X(T))$ wrt perturbations in the differential equation.

ODE local error

The estimates (42) will use the local error e defined by

$$e(t_n) \equiv \widetilde{X}(t_n) - \overline{X}(t_n), \tag{43}$$

where the local exact solution $\widetilde{X}$ satisfies, for each time step $(t_{n-1}, t_n]$,

$$\begin{aligned} \frac{d\widetilde{X}}{dt}(t) &= a(t, \widetilde{X}(t)), \quad t_{n-1} < t \leq t_n, \\ \widetilde{X}(t_{n-1}) &= \overline{X}(t_{n-1}). \end{aligned} \tag{44}$$

Definition 10 *We say that a method has order p if for sufficiently smooth problems we have $|e(t_n)| \leq C\Delta t_n^{p+1}$, with C not depending on Δt_n, nor on p.*

Example:

Consider again a Forward Euler approximation for (41).

Then $\widetilde{X}$ satisfies, for each time step $(t_{n-1}, t_n]$,

$$\begin{aligned} \frac{d\widetilde{X}}{dt}(t) &= A\widetilde{X}(t), \quad t_{n-1} < t \leq t_n, \\ \widetilde{X}(t_{n-1}) &= \overline{X}(t_{n-1}), \end{aligned} \tag{45}$$

which has the exact solution
$\widetilde{X}(t) = e^{A(t-t_{n-1})}\overline{X}(t_{n-1}), \quad t_{n-1} < t \leq t_n.$

The local error is therefore

$$
\begin{aligned}
& e(t_n) = \widetilde{X}(t_n) - \overline{X}(t_n) \\
=& (e^{A\Delta t_n)} - (I + \Delta t_n A))\overline{X}(t_{n-1}) \\
=& ((I + \Delta t_n A + \frac{(\Delta t_n A)^2}{2} + \mathcal{O}(\Delta t_n^3)) - (I + \Delta t_n A))\overline{X}(t_{n-1}) \\
=& (\Delta t_n)^2 \frac{A^2}{2}\overline{X}(t_{n-1}) + \mathcal{O}(\Delta t_n^3) \\
=& \mathcal{O}(\Delta t_n^2)
\end{aligned}
$$

and we see that the Forward Euler method has order 1.

ODE, The Variational Principle

Let $X(s;t,y)$ denote the solution of (38) at time s, which at time t takes the value y, i.e.

$$\begin{aligned} &\frac{dX}{ds}(s;t,y) = a(s,X(s;t,y)), \quad t < s \leq T, \\ &X(t;t,y) = y. \end{aligned} \tag{46}$$

Define, for the given function $g : \mathbb{R}^d \to \mathbb{R}$ in (40), the *cost to go function* $u : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ by

$$u(t,y) \equiv g(X(T;t,y)), \quad t < T, \tag{47}$$

provided the differential equation (46) on (t,T) has a unique solution for all initial data y in $\mathbb{R}^d$ and all $t \in (0,T)$.

ODE differentiability wrt initial condition

The following theorem ensures the differentiability of $X(s;t,y)$ wrt the initial condition y.

Theorem 15 (Differentiability wrt initial condition) *Let $X(s;t,y)$ solve (46) with flux $a(\cdot,\cdot)$ that is continuously differentiable wrt x in a neighborhood of (t,y).*

Then $X'(s;t,y) = \frac{\partial X}{\partial y}(s;t,y)$ *exists in a neighborhood of* (t,y) *and follows the* linear ODE

$$\begin{cases} \dfrac{dX'}{ds}(s;t,y) = \dfrac{\partial a}{\partial x}(s, X(s;t,y))\, X'(s;t,y) \\ X'(t;t,y) = I \end{cases}$$

Obs (math grads): If interested in the proof of the previous theorem, see E. Harrier, S.P. Norsett and G. Wanner, Solving Ordinary Differential Equations I, p. 95 or Arnold, Ordinary Differential Equations, p. 279.

Example

Consider again the linear ODE (41).

Then $a(x) = Ax$ and the Jacobian matrix is just $\frac{\partial a}{\partial x} = A$.

Therefore, the first variation solves

$$\begin{aligned} \frac{dX'}{ds}(s;t,y) =& AX'(s;t,y),\ s > t \\ X'(t;t,y) =& I \end{aligned}$$

i.e.

$$X'(s;t,y) = e^{A(s-t)},\ s > t$$

Obs: Due to the linearity of this example's ODE the variation X' does not depend on the initial condition y.

In the following theorem, the global approximation error for differential equations is represented in terms of the local errors and their weights, depending on the first variation of the cost to go function u.

ODE. Error representation

Theorem 16 *Assume that (46) has a unique continuous solution X for all initial data $y \in \mathbb{R}^d$ and that the flux $a(t,x)$ is continuously differentiable in x, for all $t \in (0,T)$. Let $e(t_n) \equiv \widetilde{X}(t_n) - \overline{X}(t_n)$ denote the local error of an approximation, $\overline{X}$, of (38), satisfying (39) and (44). Then, for any differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, the cost to go function u is well defined by (47) and the global error is a weighted sum of the local*

error with the representation

$$g(X(T))-g(\overline{X}(T)) = \sum_{n=1}^{N} \left(e(t_n), \int_0^1 \Psi\Big(t_n, \overline{X}(t_n) + se(t_n)\Big) ds \right), \tag{48}$$

where $(\cdot,\cdot)$ *is the standard scalar product on* $\mathbb{R}^d$ *and* $\Psi(t,y) \equiv \Psi_X(t) \in \mathbb{R}^d$ *is the first variation of* u *in the sense that for all* $w \in \mathbb{R}^d$ *and all sufficiently small* $\delta > 0$

$$u(t, y+\delta w) - u(t,y) = (\Psi_X(t), \delta w) + o(\delta).$$

The weight function Ψ_X *satisfies, for* $t < s < T$, *the dual*

equation

$$
\begin{aligned}
-\frac{d\Psi_X}{ds}(s) &= (a')^*(s, X(s;t,y))\ \Psi_X(s), \\
\Psi_X(T) &= g'(X(T;t,y)),
\end{aligned}
\tag{49}
$$

where $(a')^*(s,x)$ *is the transpose of the Jacobian matrix* $a'(s,x) \equiv \{\frac{\partial a_i}{\partial x_j}(s,x)\} \in \mathbb{R}^{d\times d}$, *and* X *solves* (46).

ODE Error representation proof

By the construction (47), the cost to go function u is constant along the characteristics $\widetilde{X}$, i.e. for all t and τ in $[t_{n-1}, t_n]$

$$u(t, \widetilde{X}(t)) = u(\tau, \widetilde{X}(\tau)). \tag{50}$$

Therefore, the initial condition for the local problem (44) shows that for $n = 1, \ldots, N$,

$$u(t_n, \widetilde{X}(t_n)) = u(t_{n-1}, \widetilde{X}(t_{n-1})) = u(t_{n-1}, \overline{X}(t_{n-1})),$$

Consequently, the initial condition (39) and (47) imply that

$$\begin{aligned}\sum_{n=1}^{N}(u(t_n,\widetilde{X}(t_n))-u(t_n,\overline{X}(t_n))) &= u(0,\overline{X}(0))-u(T,\overline{X}(T))\\ &= u(0,X(0))-u(T,\overline{X}(T))\\ &= g(X(T))-g(\overline{X}(T)).\end{aligned} \tag{51}$$

Introduce the auxiliary function $U : [0,1] \to \mathbb{R}$, defined by

$$U(s) = u(t_n, s\widetilde{X}(t_n) + (1-s)\overline{X}(t_n)).$$

The equality

$$U(1) - U(0) = \int_0^1 U'(s)ds$$

shows that each term in the sum (51) can be written

$$u(t_n, \widetilde{X}(t_n)) - u(t_n, \overline{X}(t_n)) = \left(e(t_n), \int_0^1 \Psi(t_n, \overline{X}(t_n) + se(t_n))ds \right),$$

which proves (48).

The first variation

$$\partial X(s;t,y)/\partial y$$

exists, by the differentiability assumptions on $a(s,x)$ and Theorem 15.

The combination of the existence of the first variation of X and the assumption that g is differentiable, implies by (46)-(47) that Ψ exists.

Finally, to verify that Ψ satisfies the dual equation (49), observe that, for any $w \in \mathbb{R}^d$ and $\delta \to 0$, two solutions X^1 and X^2 of (46), with initial data $X^1(t) = y \in \mathbb{R}^d$ and

$X^2(t) = y + \delta w$ satisfy

$$
\begin{aligned}
0 &= \frac{d}{dt}\left(u(t, X^2(t)) - u(t, X^1(t))\right) \\
&= \frac{d}{dt}\left(\Psi_{X^1}, X^2(t) - X^1(t)\right) + o(\delta) \\
&= \left(\frac{d}{dt}\Psi_{X^1}, X^2(t) - X^1(t)\right) + \left(\Psi_{X^1}, \frac{d}{dt}X^2(t) - \frac{d}{dt}X^1(t)\right) + o(\delta) \\
&= \left(\delta\frac{d}{dt}\Psi_{X^1}, w\right) + \left(\delta\Psi_{X^1}, a'(t, X^1(t))w\right) + o(\delta).
\end{aligned}
$$

Thus, dividing by δ we have

$$
0 = \left(\frac{d}{dt}\Psi_{X^1}, w\right) + \left(\Psi_{X^1}, a'(t, X^1(t))w\right) + o(1).
$$

Since $w \in \mathbb{R}^d$ is abitrary, this proves (49) as $\delta \to 0$.

ODE Error representation

Example 8 *Consider again the linear ODE* (41) *approximated by Forward Euler. We want to compute* $g(X) = g \cdot X$.

Then $a(x) = Ax$ *and the Jacobian matrix is just* $\frac{\partial a}{\partial x} = A$. *The dual* Ψ_X *solves*

$$\begin{cases} -\dfrac{d\Psi_X}{ds}(s) = A^* \, \Psi_X(s), \\ \Psi_X(T) = g, \end{cases} \Rightarrow \Psi_X(s) = e^{A^*(T-s)} g, \; s \leq T$$

As we saw before, the local error is

$$e(t_n) = (\Delta t_n)^2 \frac{A^2}{2} \overline{X}(t_{n-1}) + \mathcal{O}(\Delta t_n^3)$$

Therefore, the global error representation yields

$$
\begin{aligned}
g(X(T)) - g(\overline{X}(T)) & \\
&= \sum_{n=1}^{N} \left((\Delta t_n)^2 \frac{A^2}{2} \overline{X}(t_{n-1}), e^{A^*(T-t_n)} g \right) \\
&+ \sum_{n=1}^{N} \mathcal{O}(\Delta t_n^3) \\
&= \sum_{n=1}^{N} (\Delta t_n)^2 \left(e^{A(T-t_n)} \frac{A^2}{2} \overline{X}(t_{n-1}), g \right) \\
&+ \sum_{n=1}^{N} \mathcal{O}(\Delta t_n^3)
\end{aligned}
$$

Observations:

– Since we can further estimate

$$\|\overline{X}(t_{n-1})\| \leq C_T$$

the previous estimate implies

$$g(X(T)) - g(\overline{X}(T)) = \sum_{n=1}^{N} \mathcal{O}(\Delta t_n^2) \leq \mathcal{O}(\Delta t_{\max})T$$

– The only non computable term in the leading order term of the error representation is $e^{A(T-t)}$, which comes from the dual Ψ_X. We will use approximate it in general to produce useful error expansions that depend only on computed values.

Exercise 13 *Recall the global error representation,* (48)

$$g(X(T))-g(\overline{X}(T)) = \sum_{n=1}^{N} \left(e(t_n), \int_0^1 \nabla u\Big(t_n, \overline{X}(t_n) + se(t_n)\Big) ds \right), \tag{52}$$

Show *that another way to get the dual equations for* $\Psi_X(t) \equiv \nabla u(t, X(t))$ *is to take spatial derivatives in the equation*

$$0 = \frac{du}{dt}(X(t)) = \frac{\partial u}{\partial t} + \nabla u(t, X(t)) \cdot a(t, X(t)).$$

ODE Weight approximation, dual equations

Our next goal is to construct error expansions useful for adaptive methods, namely an asymptotic expansion of the representation (48) with leading order term in computable form.

Approximation of the Weight

The averaged weight function Ψ in Theorem 16, which is needed to determine the optimal step size in an adaptive method, can be computed by approximating (46) and (49).

Therefore, any p-th order accurate approximation $(\overline{X}, \overline{\Psi})$ of (X, Ψ), which solves the systems of differential equations (38) and (49), satisfies

$$\left|\overline{\Psi}(t_n) - \Psi(t_n, \overline{X}(t_n))\right| = \mathcal{O}((\max \Delta t)^p) \qquad (53)$$

where $\Delta t_n = t_n - t_{n-1}$ with $\max \Delta t \equiv \max_n \Delta t_n$.

A natural choice of approximation $\overline{\Psi}$, for a p-th order one step method $\overline{X}$ written in the form

$$\overline{X}(t_n) = A(\overline{X}(t_{n-1}), \Delta t_n), \tag{54}$$

is the *exact discrete dual*

$$\begin{aligned} \overline{\Psi}(t_{n-1}) &= \left(\partial_x A(\overline{X}(t_{n-1}), \Delta t_n)\right)^* \overline{\Psi}(t_n), \\ \overline{\Psi}(T) &= \partial_x g(\overline{X}(T)), \end{aligned} \tag{55}$$

i.e.

$$\begin{aligned} \overline{\Psi}_i(t_{n-1}) &= \sum_{j=1}^{d} \partial_{x_i} A_j(\overline{X}(t_{n-1}), \Delta t_n) \overline{\Psi}_j(t_n), \\ \overline{\Psi}_i(T) &= \partial_{x_i} g(\overline{X}(T)), \end{aligned} \tag{56}$$

which yields a p-th order accurate approximation $(\overline{X}, \overline{\Psi})$ of (X, Ψ) and satisfies

$$\overline{\Psi}_i(t_{n-1}) = \partial_{x_i} g\left(\overline{X}(T; \overline{X}(t_{n-1}) = x)\right). \qquad (57)$$

Example 9 *Consider again the linear ODE* (41) *approximated by Forward Euler. We want to compute* $g(X) = g \cdot X$.

Then we have (abusing the notation a bit)

$$A(x, \Delta t) = x + \Delta t A x = (I + \Delta t A)x$$

and the Jacobian of the method becomes $A'(x, \Delta t) = \frac{\partial A(x, \Delta t)}{\partial x} = (I + \Delta t A)$. *Thus the discrete dual solves* (56), *i.e.*

$$\begin{cases} \overline{\Psi}(t_{n-1}) = (I + \Delta t_n A^*)\overline{\Psi}(t_n),\ n = N-1, N-2, \dots \\ \quad \overline{\Psi}(T) = g. \end{cases}$$

which is again a Forward Euler discretization!

Obs:

– The relation (57) is the discrete version of the fact that $\Psi(t)$ is the first variation of $g(X(T))$ with respect to variation in the location of the path $X(t)$ at time t, and (57) holds precisely when $\overline{\Psi}$ is defined by (56).

– The Jacobian matrix $\partial_{x_i} A_j(\overline{X}(t_{n-1}), \Delta t_n) = \frac{\partial \overline{X}_j(t_n)}{\partial \overline{X}_i(t_{n-1})}$ can be approximated by numerical differentiation of $\overline{X}(t_n)$ with respect to $\overline{X}(t_{n-1})$, or alternatively the Jacobian can be evaluated explicitly for each method, e.g. to preserve a sparse structure.

Exercise 14 *Consider again the linear ODE* (41) *approximated by a numerical method of the form*

$$\overline{X}(t_n) = \overline{X}(t_{n-1}) + P(A\Delta t_n)\overline{X}(t_{n-1}).$$

Verify that the exact discrete dual solves

$$\begin{cases} \overline{\Psi}(t_{n-1}) = P(A^*\Delta t_n)\overline{\Psi}(t_n),\ n = N-1, N-2, \dots \\ \quad \overline{\Psi}(T) = g. \end{cases}$$

Use this result to derive the exact discrete dual equations for the trapezoidal method. Can you identify the resulting discretization method for the dual equation?

The error expansion is based on an approximation $\overline{\Psi}$ of the weight Ψ and an approximation $\bar{e}$ of the local error e

$$\sum_{n=1}^{N} \underbrace{\big(e(t_n), \Psi\big)}_{\text{exact error}} - \sum_{n=1}^{N} \underbrace{\big(\bar{e}(t_n), \overline{\Psi}\big)}_{\text{error approximation}}$$

$$= \sum_{n=1}^{N} \underbrace{\big(e(t_n) - \bar{e}(t_n), \overline{\Psi}\big)}_{\text{local error approximation}} + \sum_{n=1}^{N} \underbrace{\big(e(t_n), \Psi - \overline{\Psi}\big)}_{\text{weight approximation}} . \tag{58}$$

We have to analyze the last two terms and conclude that they are higher order wrt the error approximation we use.

ODE Weight approximation, dual equations.

Theorem 17 *Suppose that (53) and the assumptions of Theorem 16 hold. Let $\partial_{xx}u(t,x)$ in (47) be uniformly bounded for $(t,x) \in [0,T] \times \mathbb{R}^d$. Then the global approximation error for the ODE (38) satisfies*

$$g(X(T)) - g(\overline{X}(T)) = \sum_{n=1}^{N} \left(e(t_n), \overline{\Psi}(t_n) + \mathcal{O}(|e(t_n)|) + \mathcal{O}((\max \Delta t)^p) \right) \tag{59}$$

where $e(t_n) \equiv \widetilde{X}(t_n) - \overline{X}(t_n)$ is the local error and $(\overline{X}, \overline{\Psi})$ is a p-th order accurate approximation of the system (38) and (49).

Example Consider again the linear ODE (41) approximated by Forward Euler. We want to compute $g(X) = g \cdot X$.

Then the global approximation error for the ODE (38) satisfies

$$g(X(T)) - g(\overline{X}(T))$$
$$= \sum_{n=1}^{N} \Big(e(t_n), \overline{\Psi}(t_n) \Big) + \sum_{n=1}^{N} \{ \mathcal{O}(\Delta t_n^4) + \mathcal{O}(\Delta t_{\max} \Delta t_n^2) \} \tag{60}$$

ODE local error approximation

The next step necessary to derive an error estimate based on computable quantities is to approximate the local error

$$e = \widetilde{X} - \overline{X}$$

by replacing the exact local solution $\widetilde{X}$ by an approximation $\overline{\overline{X}}$ of higher accuracy than $\overline{X}$, i.e., with smaller time steps or a higher order method in a higher precision.

For smooth solutions X, the existence of the limits

$$\begin{aligned} &\lim_{\Delta t \to 0} (\Delta t_n)^{-(p+1)} \left(\widetilde{X}(t_n) - \overline{X}(t_n) \right), \\ &\lim_{\Delta t \to 0} (\Delta t_n)^{-(q+1)} \left(\widetilde{X}(t_n) - \overline{\overline{X}}(t_n) \right), \end{aligned} \tag{61}$$

determines by Richardson extrapolation a constant γ, for $q \geq p$[a], s.t.

$$e(t_n) = \widetilde{X}(t_n) - \overline{X}(t_n) = \gamma \left(\overline{\overline{X}}(t_n) - \overline{X}(t_n) \right) + o(\Delta t_n^{p+1}). \tag{62}$$

[a] cf. G. Dahlquist and Å. Björk, Numerical Methods

For instance there holds:

$$\gamma = 2^p/(2^p - 1)$$

for $\overline{\overline{X}}$ computed with the half mesh size and $q = p$; and $\gamma = 1$ for $\overline{\overline{X}}$ computed with a higher order method $q > p$,[a].

[a] see E. Harrier, S.P. Norsett and G. Wanner, Solving Ordinary Differential Equations I

Exercise 15 (Local error estimation for F. Euler)
Let X solve (38) and $\overline{X}$ be a Forward Euler approximation. Let

$$a_n \equiv a(t_n, \overline{X}(t_n)).$$

Show using Richardson's extrapolation that we have the approximation

$$\begin{aligned} e(t_n) =& \Delta t_n \left(a\left(t_n + \frac{\Delta t_n}{2}, \overline{X}(t_n) + a_n \frac{\Delta t_n}{2} \right) - a_n \right) \\ &+ o(\Delta t_n^2) \end{aligned}$$

Let $\Delta t(t) \equiv \Delta t_n$, $t_{n-1} < t \leq t_n$. The replacement of the exact local error with this approximate local error yields

Theorem 18 *Suppose that the limits* (61) *exist and that the assumptions of Theorems 16 and 17 hold. Then the global approximation error for the differential equation* (38) *satisfies the estimate*

$$g(X(T)) - g(\overline{X}(T)) = \sum_{n=1}^{N} \left(\bar{e}(t_n), \overline{\Psi}(t_n)\right) + \int_0^T o(\Delta t^p(t))dt \tag{63}$$

where $\bar{e}(t_n) \equiv \gamma \left(\overline{\overline{X}}(t_n) - \overline{X}(t_n)\right)$ *is the approximation of the local error in* (62) *and* $(\overline{X}, \overline{\Psi})$ *is a p-th order accurate approximation of the system* (38) *and* (49).

Recall Adaptive Algorithms Given a desired accuracy, $TOL > 0$ then find

$$0 = t_0 < t_1 < \ldots < t_N = T, \text{ a partition of } [0, T]$$

s.t. we can bound the time discretization error by

$$\left| g(X(T)) - g(\overline{X}_N) \right| \approx E_T \leq TOL,$$

while the computational work, which is proportional to the number of time steps, N, is minimized.

Question: What is the best way to choose the time steps?

Let us now motivate the optimal choice of steps

$$\Big|\text{local error} \cdot \text{weight}\Big| = \text{constant},$$

for approximation methods which have no essential constraint on the step sizes, such as one step methods (54). For the time steps $0 = t_0 < \cdots < t_N = T$, let the piecewise constant mesh function Δt be defined by

$$\Delta t(\tau) \equiv \Delta t_n \equiv t_n - t_{n-1} \quad \text{for} \quad \tau \in (t_{n-1}, t_n], \quad n = 1, \ldots, N$$

Then the number of time steps that corresponds to a mesh Δt, for the interval $[0, T]$, can be defined by

$$N(\Delta t) \equiv \int_0^T \frac{1}{\Delta t(\tau)} d\tau. \tag{64}$$

Consider, for $\tau \in (t_{n-1}, t_n]$ and $n = 1, \ldots, N$, the piecewise constant function ρ, which measures the density of the global error from (63)

$$\rho(\tau) \equiv \rho_n \equiv \frac{(\bar{e}(t_n), \overline{\Psi}(t_n))}{\Delta t_n^{p+1}} + \mathcal{O}(\Delta t_n) \tag{65}$$

and its approximate counterpart $\bar{\rho}$, obtained from (63) with

$$\bar{\rho}(\tau) \equiv \bar{\rho}_n \equiv \text{sign}\left(\bar{e}(t_n), \overline{\Psi}(t_n)\right) \max\left(\frac{|\left(\bar{e}(t_n), \overline{\Psi}(t_n)\right)|}{\Delta t_n^{p+1}}, \ \delta\right) \tag{66}$$

where we choose the lower bound

$$\delta \equiv TOL^{\bar{\gamma}}, \quad 0 < \bar{\gamma} < \frac{1}{p+1}. \tag{67}$$

As usual, we define the sign function by

$$\text{sign}(x) = \begin{cases} 1 \text{ for } x \geq 0, \\ -1 \text{ for } x < 0. \end{cases}$$

Remark 15 *The convergence of the approximate error density towards the exact one requires that* $\max \Delta t \to 0$ *as* $TOL \to 0$. *The choice for the constant* $\delta = TOL^{\gamma} > 0$ *guarantees that* $\max \Delta t \to 0$.

We choose to minimize the number of steps N in (64) under the more stringent constraint

$$\sum_{i=1}^{N} |\bar{\rho}_i| \Delta t_i^{p+1} = \int_0^T |\bar{\rho}(\tau)| \Delta t^p(\tau) d\tau = TOL. \tag{68}$$

This yields, with a standard application of a Lagrange multiplier, the optimal time steps Δt^* satisfying

$$|\bar{\rho}_i| \Delta t_i^{p+1} = \ constant \tag{69}$$

and

$$\Delta t^* \equiv \frac{TOL^{\frac{1}{p}}}{|\bar{\rho}|^{\frac{1}{p+1}}} \left(\int_0^T |\bar{\rho}(\tau)|^{\frac{1}{p+1}} d\tau \right)^{-\frac{1}{p}}. \tag{70}$$

This optimal choice gives $TOL = |\,\mathtt{E}_T\,|$, where

$$\mathtt{E}_T \equiv \sum_{i=1}^{N} \bar{\rho}_i \Delta t_i^{p+1}, \tag{71}$$

only for problems with positive density functions $\bar{\rho}$, since otherwise (71) and (68) may give $TOL \gg |\,\mathtt{E}_T\,|$.

The goal of the adaptive algorithm described below is to construct a partition Δt of $[0, T]$ such that

$$|\bar{\rho}_n| \Delta t_n^{p+1} \approx \frac{TOL}{N}, \quad \forall\, n = 1, \ldots, N, \tag{72}$$

which is an approximation of the optimal (69).

To achieve (72) let $s_1 \approx 1$ be a given constant, start with an initial partition $\Delta t[1]$ and then specify iteratively a new partition $\Delta t[k+1]$, from $\Delta t[k]$, using the following division strategy: for $n = 1, 2, \ldots, N[k]$, let the error indicator from interval n at refinement level k be

$$\bar{r}_n[k] \equiv |\bar{\rho}_n[k]|(\Delta t_n[k])^{p+1}, \tag{73}$$

and

$$\textbf{if} \quad \bar{r}_n[k] > s_1 \frac{TOL}{N[k]} \quad \textbf{then}$$
$$\text{divide } \Delta t_n[k] \text{ into 2 uniform substeps}$$
$$\textbf{else} \tag{74}$$
$$\text{let the new step be the same as the old}$$
$$\textbf{endif}$$

With this division strategy, it is natural to use the stopping criterion:

$$\textbf{if} \quad \left(\max_{1 \le n \le N[k]} \bar{r}_n[k] \le S_1 \frac{TOL}{N[k]} \right) \quad \textbf{then} \quad \text{stop.} \tag{75}$$

Here S_1 is a given constant, with $S_1 > s_1 \approx 1$

A consequence of the uniform convergence of $\bar{\rho}$ as $TOL \to 0+$, and (66) is that for all time steps n and all refinement levels k there exists positive functions $\hat{c}$ and $\hat{C}$ close to 1 for sufficiently refined meshes, such that the error density satisfies

$$\begin{aligned} \hat{c} \leq \left| \frac{\bar{\rho}(t_n)[\text{parent}(n,k)]}{\bar{\rho}(t_n)[k]} \right| \leq \hat{C}, \\ \hat{c} \leq \left| \frac{\bar{\rho}(t_n)[k-1]}{\bar{\rho}(t_n)[k]} \right| \leq \hat{C}, \end{aligned} \tag{76}$$

provided $\max_{n,k} \Delta t_n[k]$ is sufficiently small.

In other words, (76) holds with e.g. $\hat{c} = 2^{-1} = \hat{C}^{-1}$ for sufficiently small $\max_{n,k} \Delta t_n[k]$.

Assumption 1 *Assume that $\hat{c}$ satisfies (76), for the elements or time steps corresponding to the maximal error indicator, $\max_n r_n[k]$, on each refinement level, and that*

$$S_1 \geq \frac{2}{\hat{c}} s_1, \quad 1 > \frac{\hat{c}^{-1}}{2^{1+p}}. \tag{77}$$

Observe that this assumption may imply a condition on the initial mesh size.

Observe also that the above assumption can be checked out during computations because it only involves computable quantities.

The following adaptive algorithm, [MSTZ], performs successive mesh refinements with the following property:

it either reduces the maximal error indicator by a factor

or

stops with the error asymptotically bounded by the prescribed accuracy requirement TOL.

Furthermore, the algorithm stops using the optimal number of degrees of freedom, up to a problem independent factor.

Obs (math grads) Theorems 19, 20 and 21 give precise statements for the above claims.

[MSTZ] Adaptive algorithm for error control

Initialization. The user chooses

1. an initial error tolerance, TOL,
2. a number, $N[1]$, of initial uniform steps $\Delta t[1]$ for $[0, T]$,
3. a number, $s_1 = 2$, in the refinement criterion and a rough estimate of c in (76) to compute S_1 using (79).

Set the iteration number k to 0.

[MSTZ] Adaptive algorithm for error control

Step I. Increment the iteration number k by 1. For $n = 1, \ldots, N[k]$, compute the approximation $\overline{X}(t_n)$ of (38) using a p-th order accurate numerical method (54), and to obtain the local error, compute the approximate local exact solution $\overline{\overline{X}}(t_n)$ of (44) using a higher accuracy than for $\overline{X}(t_n)$. Compute the approximation of the local error $\bar{e}(t_n)$ by (62) and the approximate weight $\overline{\Psi}(t_n)$, for $n = N[k], \ldots, 1$, using the p-th order accurate method (56).

Step II. If $\left(\max_{1 \le i \le N[k]} \bar{r}_i[k] \le \frac{S_1 \mathrm{TOL}}{N[k]} \right)$ **then** stop the program

else

do for all time steps $i = 1, \ldots, N[k]$

if $\left(\bar{r}_i[k] > s_1 \dfrac{TOL}{N[k]} \right)$ **then**

$$\Delta t(t)[k+1] = \frac{\Delta t_i[k]}{2}, \quad t_{i-1}[k] < t \le t_i[k],$$

else

$$\Delta t(t)[k+1] = \Delta t_i[k], \quad t_{i-1}[k] < t \le t_i[k],$$

endif

enddo

go to Step I.

endif

Numerical example

Here we present numerical experiments with the adaptive algorithm `MSTZ`, using the `MATLAB` version 5.3 software package.

To study its performance, we choose the Lorenz problem and a problem with a singularity and we compare the results to the adaptive algorithm `ODE45` in `MATLAB` and to a constant step size algorithm, denoted `Uniform`.

Lorenz Numerical example

Example 10 *Consider the well-known Lorenz system, which is the three dimensional system of ordinary differential equations,*

$$\begin{aligned} a_1(t,x) &= -\sigma x_1 + \sigma x_2, \\ a_2(t,x) &= r x_1 - x_2 - x_1 x_3, \quad 0 \le t \le T, \quad x \in \mathbb{R}^3, \\ a_3(t,x) &= x_1 x_2 - b x_3 \end{aligned} \tag{78}$$

where σ , r and b are given positive constants.

The Lorenz system was introduced to describe the flow of fluid in a box which is heated along the bottom. This model was intended to simulate medium-scale atmospheric convection.

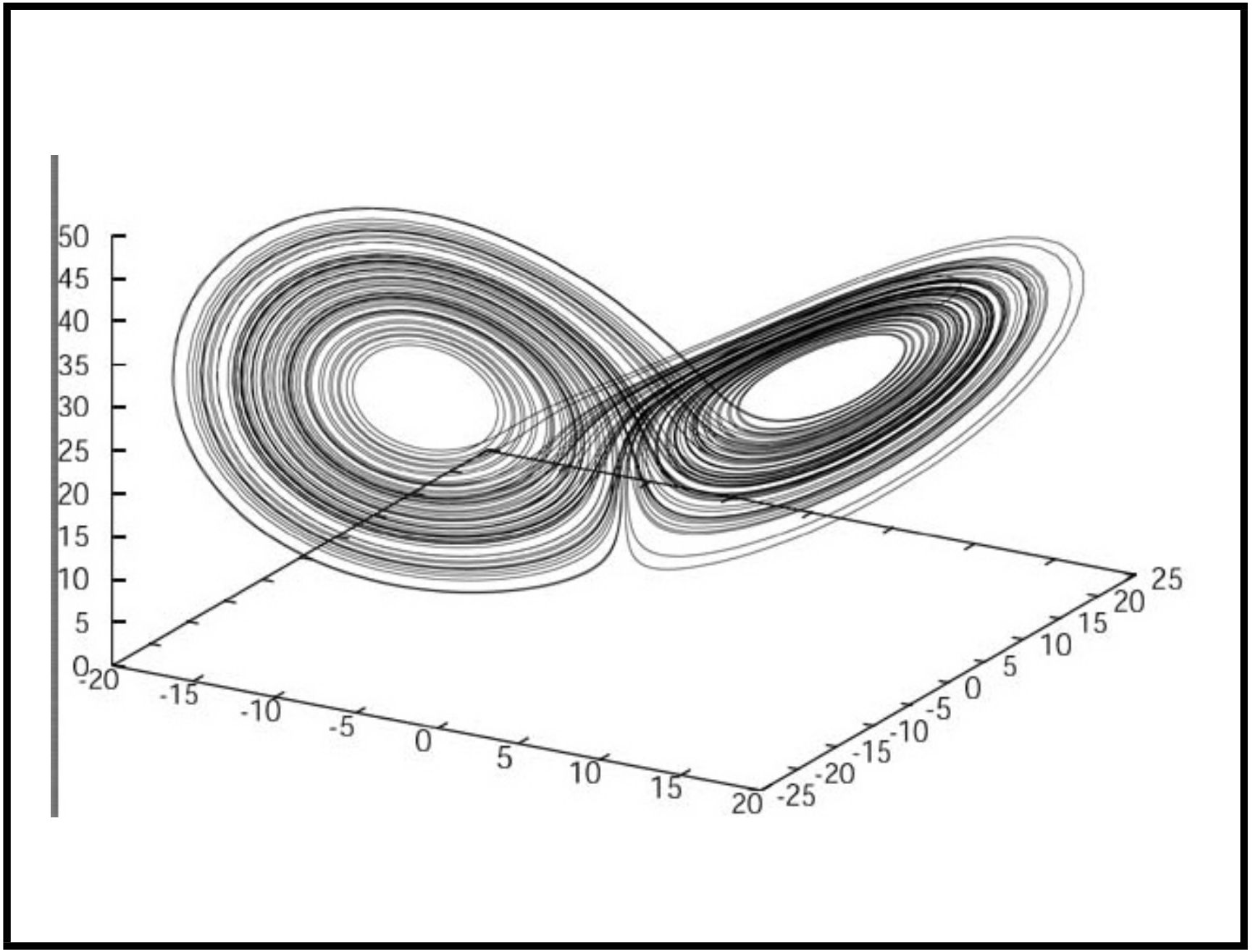
Lorenz simplified some of the Navier-Stokes equations in the area of fluid dynamics and obtained three ordinary differential equations.

This problem shows the limitation of large time prediction that already appear in the simplified model.

Try the command

```
lorenz
```

in matlab.


50
45
40
35
30
25
20
15
10
5
0
-20
-15
-10
-5
0
5
10
15
20
-25
-20
-15
-10
-5
0
5
10
15
20
25

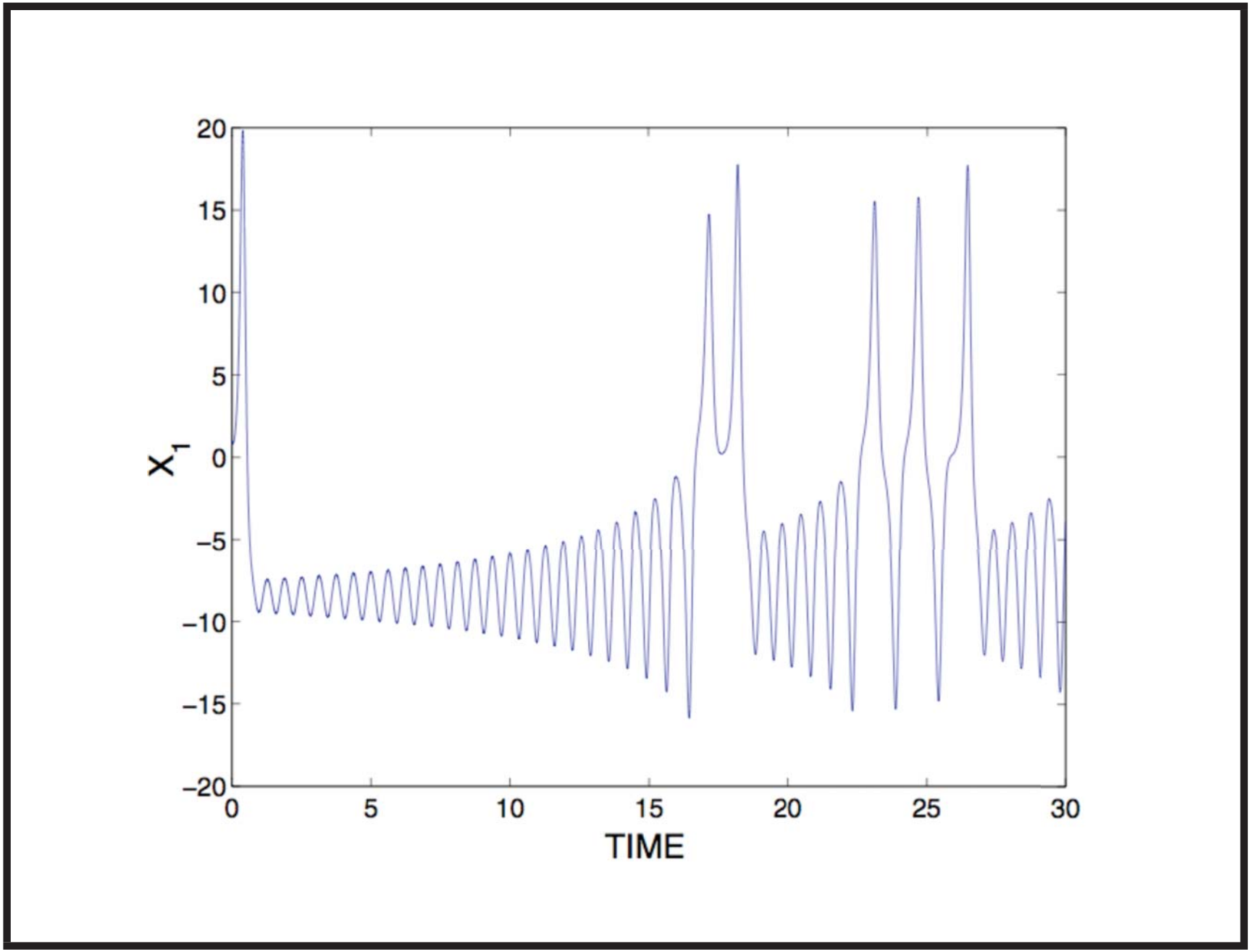
In our experiments, the coefficient values are $\sigma = 10$, $b = 8/3$ and $r = 28$ and the initial value is

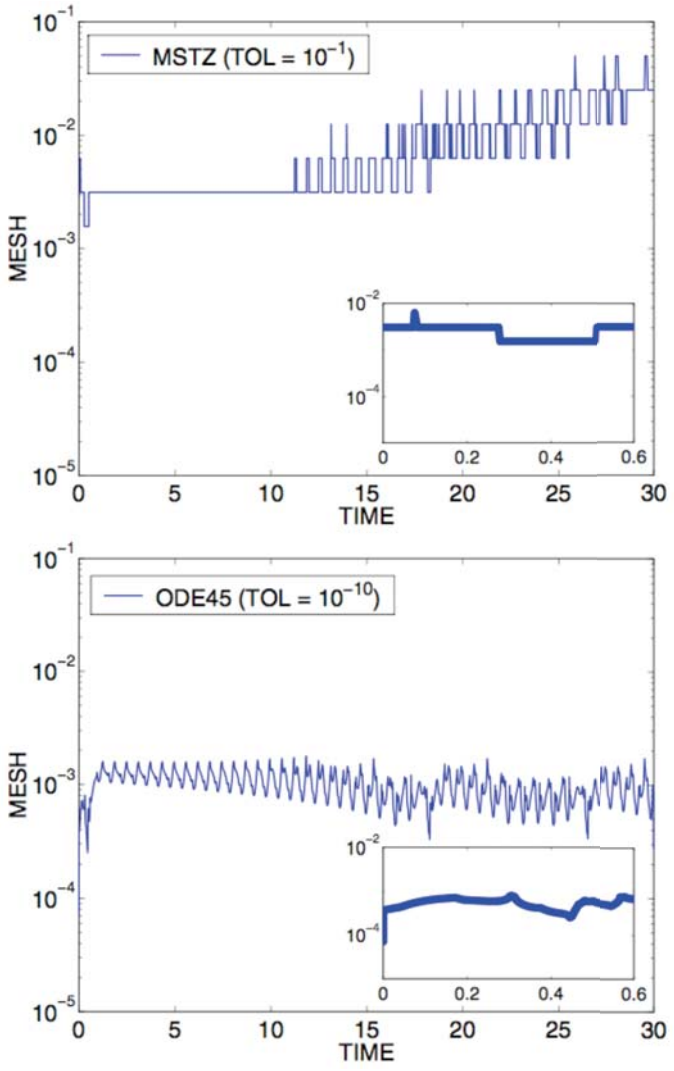
$$X(0) = (1, 0, 0).$$

The computed function is $g(x) = x_1$, i.e., we study the global error $|X_1(T) - \overline{X}_1(T)|$ at the final time $T = 30$. A reference computation with a `Fortran` implementation of `MSTZ` in quadruple precision gives the approximate value

$$X_1(30) \simeq -3.892637 \equiv g_1,$$

with $TOL = 10^{-7}$.


20
15
10
5
0
−5
−10
−15
−20
X1
0
5
10
15
20
25
30
TIME



Comparison of the mesh functions of `MSTZ` and `ODE45`. The minimum value of Δt of `MSTZ` is 0.0016 which is 22 times larger than the minimum of `ODE45`.

Table 1. Example 3.1: Comparisons of the final number of steps, N_f, and the total number of steps, N^{tot}, with the global error, $\texttt{Error} \equiv |g_1 - g(\overline{X}(T))|$, using a 5-th order explicit Runge-Kutta method with adaptive steps for `MSTZ` or `ODE45` and uniform steps for `Uniform`

	Tolerance	`Error`	N_f	N^{tot}
`MSTZ`	$\text{TOL} = 10^{-1}$	0.01	6000	20000
	$\text{TOL} = 10^{-2}$	0.003	9000	34000
`Uniform`	$\text{TOL} = 10^{-1}$	0.06	10000	19000
	$\text{TOL} = 10^{-2}$	0.002	19000	38000
`ODE45`	$\text{TOL}= 10^{-10}$	0.04	34000	92000
	$\text{TOL}= 10^{-11}$	0.004	53000	144000

Table 3. Example 3.1 and 3.2: Comparisons of the ratio, Γ and $\overline{\Gamma}$, between the exact error and the approximate error using error density, $\bar{\rho}$ and $|\bar{\rho}|$ respectively, for `MSTZ`

	Lorenz (Example 3.1)		Singularity (Example 3.2)	
Tolerance	10^{-1}	10^{-2}	10^{-1}	10^{-4}
Γ	0.991	0.997	1.325	2.31
$\overline{\Gamma}$	1.707	1.220	1.325	2.66

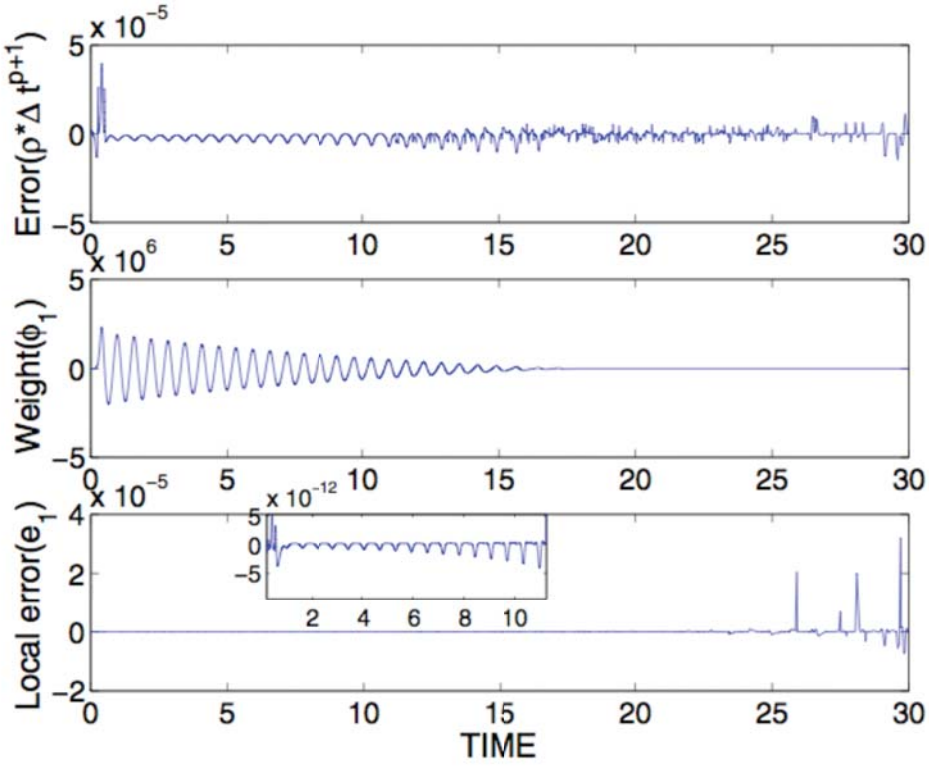


Fig. 3.6. Example 3.1: The error indicators and the first component of the weights and the local errors from MSTZ with TOL $= 10^{-1}$. The other two components have a similar behavior. Note that the local errors and the weights oscillate, but their product does not give a significant error cancellation, see Table 3

Theorem 19 (Stopping) *Suppose the adaptive algorithm uses the strategy (73)-(75). Assume that $\hat{c}$ satisfies (76) for the time steps corresponding to the maximal error indicator on each refinement level, and that*

$$S_1 \geq \frac{2}{c} s_1, \quad 1 > \frac{c^{-1}}{2^{p+1}}. \tag{79}$$

Then each refinement level either decreases the maximal error indicator with the factor

$$\max_{1 \leq n \leq N[k+1]} \bar{r}_n[k+1] \leq \frac{c^{-1}}{2^{p+1}} \max_{1 \leq n \leq N[k]} \bar{r}_n[k], \tag{80}$$

or stops the algorithm.

Theorem 20 (Accuracy) *Suppose that* (63), *(65-67) hold, and that the approximate error density converges pointwise to the exact one. Then the adaptive algorithm* (73)-(75) *satisfies*

$$\limsup_{\mathrm{TOL}\to 0+} \left(\mathrm{TOL}^{-1} \left| g(X(T)) - g(\overline{X}(T)) \right| \right) \leq S_1. \qquad (81)$$

Theorem 21 (Efficiency) *Assume that* $\hat{c} = \hat{c}(t)$ *satisfies* (76) *for all time steps at the final refinement level, that all initial time steps have been divided when the algorithm stops, and that* (63), *(65-67) hold. Then there exists a constant* $C > 0$, *bounded by* $(\frac{2^{p+1}}{s_1})^{\frac{1}{p}}$, *such that the final number of adaptive steps* N, *of the algorithm* (73)-(75), *satisfies*

$$TOL^{\frac{1}{p}} N \leq C \, \|\frac{\bar{\rho}}{\hat{c}}\|^{\frac{1}{p}}_{L^{\frac{1}{p+1}}} \leq C \left(\max_{0\leq t\leq T} \hat{c}(t)^{-\frac{1}{p}} \right) \|\bar{\rho}\|^{\frac{1}{p}}_{L^{\frac{1}{p+1}}}, \quad (82)$$

and $\|\bar{\rho}\|_{L^{\frac{1}{p+1}}} \to \|\widetilde{\rho}\|_{L^{\frac{1}{p+1}}}$, *asymptotically as* $TOL \to 0+$.

References:

"An Adaptive Algorithm for Ordinary Stochastic and Partial Differential Equations", K-S. Moon, E. Von Schwerin, A. Szepessy and R. Tempone, *Contemporary Mathematics, No. 383, 325-344, Jan 2006.*

"A variational principle for adaptive approximations of ordinary differential equations", K-S. Moon, A. Szepessy, R. Tempone and G. Zouraris, *Numerische Mathematik Vol. 96, No. 1, pp. 131 - 152 , 2003.*

"Convergence rates for adaptive approximation of ordinary differential equations", K.-S. Moon, A. Szepessy, R. Tempone and G. Zouraris, *Numerische Mathematik Vol. 96, No. 1, pp. 99 - 129 , 2003.*