Convergence of Numerical Methods for ODE’s

Hart Smith

Department of Mathematics
University of Washington, Seattle

Math 555, Winter 2014
ODE / IVP: \( x'(t) = f(t, x(t)), \quad a \leq t \leq b, \quad x(a) = x_a. \)

**General One-step Numerical Scheme:**

- Divide \([a, b]\) into \(N\) intervals length \(h = (b - a)/N\)
- evenly spaced tick marks: \(t_j = a + jh, \quad j = 0, \ldots, N\)
- recursively define \(x\) values: \(x_{j+1} = x_j + h \psi(h, t_j, x_j)\)

Euler’s method: \(\psi(h, t, x) = f(t, x) \colon x_{j+1} = x_j + hf(t_j, x_j)\)

*Allowing dependence on \(h\) gives higher order approximation...*

**Consistency condition:**

\(\psi(0, t, x) = f(t, x).\)
Example: \( x' = Kx, \ t \in [0, 1], \) solution \( x = e^{Kt}x(0). \)

Euler’s method: \( \psi(h, t_j, x_j) = f(x_j) = Kx_j \)

\[
x_{j+1} = x_j + hKx_j = (1 + hK)x_j
\]

\[
\vdots
\]

\[
= (1 + hK)^{j+1}x_0
\]

Express in terms of \( t_j = jh, \)

\[
x_j = (1 + hK)^jx_0 = (1 + Kh)^{t_j/h}x_0
\]

Recall: \( \lim_{h \to 0} (1 + hK)^{t/h} = e^{Kt}, \) so if set \( x_0 = x(0) \)

\[
\lim_{h \to 0} |x(t_j) - x_j| = 0 \quad \text{uniformly for} \quad t_j \in [0, 1]
\]
Euler’s method for \( x' = x, \ 0 \leq t \leq 1, \ x_0 = 1. \)
Modified Euler’s: \( \psi(h, x_j) = f(x_j + \frac{h}{2} f(x_j)) = K x_j + \frac{h}{2} K^2 x_j \)

\[
x_{j+1} = (1 + hK + \frac{1}{2} h^2 K^2) x_j \\
= (1 + hK + \frac{1}{2} h^2 K^2)^{j+1} x_0
\]

Express in terms of \( t_j = jh \),

\[
x_j = (1 + hK + \frac{1}{2} h^2 K^2)^{t_j/h} x_0
\]

Better approximation than Euler’s method:

\[
(1 + hK)^{t/h} < (1 + hK + \frac{1}{2} h^2 K^2)^{t/h} < (e^{hK})^{t/h} = e^{Kt}.
\]
Modified Euler’s vs. Euler’s for \( x' = x, \ x_0 = 1 \).
Convergence of numerical method for \( x' = f(t, x(t)) \)

**Goal:** show that, under conditions on \( \psi(h, t, x) \),

\[
\max_{j=1,\ldots,N} \left| x(t_j) - x_j \right| \to 0 \quad \text{as} \quad h \to 0.
\]

**Idea:** if each step in \( j \) adds an error of size \( h \cdot o(h) \),

maximum cumulative error is \( Nh \cdot o(h) = o(h) \).

Maximum deviation of \( \psi \) from exact differential for solution \( x \):

\[
\tau(h) = \sup_t \left| \frac{x(t + h) - x(t)}{h} - \psi(h, t, x(t)) \right|
\]

Necessary condition for \( \lim_{h \to 0} \tau(h) = 0 \):

\( \psi(0, t, x) = f(t, x) \).
Assume $\psi(h, t, x)$ continuous on $[0, h_0] \times [a, b] \times \mathbb{R}^n$

If $\psi(0, t, x) = f(t, x)$, then $\lim_{h \to 0} \tau(h) = 0$.

Proof. $\tau(h) = \sup_t \left| \frac{x(t + h) - x(t)}{h} - \psi(h, t, x(t)) \right|$  

$= \sup_t \left| \frac{1}{h} \int_t^{t+h} \psi(0, s, x(s)) - \psi(h, t, x(t)) \, ds \right|$  

By uniform continuity on compact sets:

$$\sup_{s \in [t, t+h], t \in [a, b]} \left| \psi(0, s, x(s)) - \psi(h, t, x(t)) \right| = o(h).$$
Assume that \( f(t, x) \) and \( \psi(h, t, x) \) are continuous, and uniformly Lipschitz in \( x \) with Lipschitz constant \( K \), and \( \psi(0, t, x) = f(t, x) \).

Suppose \( x'(t) = f(t, x(t)) \), \( x_{j+1} = x_j + h\psi(h, t_j, x_j) \). Then

\[
\max_{j=0, \ldots, N} |x(t_j) - x_j| \leq e^{K(b-a)}|x(t_0) - x_0| + \tau(h) \left( \frac{e^{K(b-a)} - 1}{K} \right)
\]

In particular, if \( |x(t_0) - x_0| \to 0 \) as \( h \to 0 \), then

\[
\max_{j=0, \ldots, N} |x(t_j) - x_j| \to 0 \quad \text{as} \quad h \to 0.
\]
Lemma (single-step error estimate)

\[|x(t_{j+1}) - x_{j+1}| \leq (1 + hK)|x(t_j) - x_j| + h\tau(h).\]

**Proof.** By definition of \(\tau(h)\), since \(t_{j+1} = t_j + h\),

\[|x(t_{j+1}) - x(t_j) - h\psi(h, t_j, x(t_j))| \leq h\tau(h).\]

By construction: \(x_{j+1} - x_j = h\psi(h, t_j, x_j)\).

By Lipschitz condition on \(\psi\):

\[|h\psi(h, t_j, x(t_j)) - h\psi(h, t_j, x_j)| \leq hK|x(t_j) - x_j|.\]

Put this together:

\[|(x(t_{j+1}) - x_{j+1}) - (x(t_j) - x_j)| \leq hK|x(t_j) - x_j| + h\tau(h).\]
Cumulative error estimate. Notation: \( e_j = |x(t_j) - x_j| \geq 0 \).

If \( e_{j+1} \leq (1 + hK) e_j + h \tau(h) \), then

\[
e_j \leq (1 + hK)^j e_0 + \tau(h) \left( \frac{(1 + hK)^j - 1}{K} \right)
\]

\[
\leq e^{K(b-a)} e_0 + \tau(h) \left( \frac{e^{K(b-a)} - 1}{K} \right)
\]

Proof by Induction: true for \( e_0, e_1 \), assume true for \( e_j \),

\[
(1 + hK) \left( (1 + hK)^j e_0 + \tau(h) \left( \frac{(1 + hK)^j - 1}{K} \right) \right) + h \tau(h)
\]

\[
= (1 + hK)^{j+1} e_0 + \tau(h) \left( \frac{(1 + hK)^{j+1} - 1}{K} \right).
\]
Corollary: If \( e_0 = 0 \), i.e. \( x_0 = x(t_0) \), then

\[
\max_{j=0,\ldots,N} |x(t_j) - x_j| \leq \tau(h) \left( \frac{e^{K(b-a)} - 1}{K} \right)
\]

Goal: choose \( \psi(h, t, x) \) so that \( \tau(h) \) vanishes quickly as \( h \to 0 \).

Definition

The method \( \psi(h, t, x) \) is accurate to order \( p \) if \( \tau(h) = O(h^p) \) for all solutions \( x(t) \) to \( x'(t) = f(t, x(t)) \).

Since \( \tau(h) = \sup_t \left| \frac{x(t+h) - x(t)}{h} - \psi(h, t, x(t)) \right| \)

\( \psi(h, t, x) \) is accurate to order \( p \) if

\[
|x(t + h) - x(t) - h\psi(h, t, x(t))| \leq C h^{p+1}.
\]
Example: If $f \in C^1$, $\psi(h, t, x) = f(t, x)$ is accurate to order 1.

Proof: If $f \in C^1$, then $x \in C^2([a, b])$, and $x'(t) = f(t, x(t))$, so

$$x(t + h) - x(t) - hf(t, x(t)) = \frac{1}{2} h^2 x''(s), \quad s \in [t, t + h].$$

Let $C = \frac{1}{2} \sup_{s \in [a,b]} |x''(s)|$, then

$$|x(t + h) - x(t) - h\psi(h, t, x(t))| \leq C h^2.$$

Observation: assume $f \in C^p$, so $x \in C^{p+1}([a, b])$

$\psi(h, t, x)$ is accurate to order $p$ if and only if

$$h\psi(h, t, x(t)) = \sum_{j=1}^{p} \frac{h^j}{j!} D^j_t x(t) + O(h^{p+1})$$

that is, $x(t) + h\psi(h, t, x(t))$ differs from the order-$p$ Taylor expansion of $x(t)$ by a term of size $O(h^{p+1})$. 
Taylor Method: assume \( f \in C^p \), so \( x \in C^{p+1}([a, b]) \)

**Observation**

If \( D_t^j x(t) = f_j(t, x(t)) \), then

\[
D_t^{j+1} x(t) = (D_t f_j)(t, x(t)) + (D_x f_j)(t, x(t)) \cdot f(t, x(t))
\]

Set \( f_1(t, x) = f(t, x) \), and recursively define:

\[
f_{j+1}(t, x) = D_t f_j(t, x) + D_x f_j(t, x) \cdot f(t, x)
\]

**Taylor Method of order \( p \):** set \( \psi = T_p \), where

\[
T_p(h, t, x) = f(t, x) + \sum_{j=2}^{p} h^{j-1} f_j(t, x)
\]

\[
x(t + h) - x(t) - h T_p(h, t, x(t)) = O(h^{p+1}).
\]
First few $T_p$'s

\begin{align*}
T_1(t, x) & = f(t, x) \\
T_2(t, x) & = D_t f(t, x) + D_x f(t, x) \cdot f(t, x) \\
T_3(t, x) & = D_t^2 f(t, x) + 2D_x D_t f(t, x) \cdot f(t, x) + D_x f(t, x) \cdot D_t f(t, x) \\
& \quad + D_x^2 f(t, x) \cdot f(t, x)^2 + D_x f(t, x) \cdot D_x f(t, x) \cdot f(t, x)
\end{align*}

Not practical in general since requires evaluating order $p - 1$ derivatives of $f$. In practice, achieve higher order by nested evaluation.
\[ T_2 \text{ versus } T_1 \text{ for } x' = x, \ x_0 = x(0) = 1, \ h = \frac{1}{5} \]

\[ T_1(h, x) = x, \ \tau(h)\left(\frac{e^1-1}{1}\right) \leq \frac{1}{2}x''(s)h(e - 1) \lesssim .467 \]

\[ T_2(h, x) = (1 + \frac{h}{2})x, \ \tau(h)\left(\frac{e^1-1}{1}\right) \leq \frac{1}{6}x'''(s)h^2(e - 1) \lesssim .031 \]