Error Inhibiting Schemes for Differential Equations

Adi Ditkowski

Department of Applied Mathematics
Tel Aviv University

Joint work with Sigal Gottlieb, Chi-Wang Shu and Paz Fink.

ICERM, August 2018.

Outline of the talk:

- Review of the classical theory.
  - Semi-discrete approximations for PDEs.
  - Fully-discrete approximations for PDEs or ODEs.
- Error Inhibiting Schemes for ODEs.
- Error Inhibiting Schemes for PDEs.
  - Block Finite Difference schemes for the Heat equation.

Review of the classical theory

Semi-discrete approximations for PDEs.

Consider the differential problem:

$$\begin{aligned} \frac{\partial u}{\partial t} &= P\left(\frac{\partial}{\partial x}\right) u \ , \qquad x \in \Omega \subset \mathbb{R}^d \ , t \geq 0 \\ u(t=0) &= f \ . \end{aligned}$$

It is assumed that this problem is well posed, In particular $\exists K(t) < \infty$ s.t. $||u(t)|| \leq K(t)||f||$. Typically $K(t) = Ke^{\alpha t}$.

Let Q be the discretization of $P\left(\frac{\partial}{\partial x}\right)$ where we assume:

Assumption 1: The discrete operator Q is based on the the grid points $\{x_j\}$, $j = 1, \ldots, N$.

Let Q be the discretization of $P\left(\frac{\partial}{\partial x}\right)$ where we assume:

Assumption 1: The discrete operator Q is based on the the grid points $\{x_j\}$, $j = 1, \ldots, N$.

Assumption 2: Q is semibound in some equivalent scalar product $(\cdot, \cdot)_H = (\cdot, H\cdot)$, i.e.

$$(\mathbf{w}, Q\mathbf{w})_H \leq \alpha (\mathbf{w}, \mathbf{w})_H = \alpha \|\mathbf{w}\|_H^2$$

Let Q be the discretization of $P\left(\frac{\partial}{\partial x}\right)$ where we assume:

Assumption 1: The discrete operator Q is based on the the grid points $\{x_j\}$, $j = 1, \ldots, N$.

Assumption 2: Q is semibound in some equivalent scalar product $(\cdot, \cdot)_H = (\cdot, H\cdot)$, i.e.

$$(\mathbf{w}, Q\mathbf{w})_H \leq \alpha\,(\mathbf{w}, \mathbf{w})_H = \alpha\,\|\mathbf{w}\|_H^2$$

Assumption 3: The local truncation error of Q is T_e and is defined by $\mathbf{T}_e = P\mathbf{w} - Q\mathbf{w}$,
where $w(x)$ is a smooth function and $\mathbf{w}$ is the projection of $w(x)$ onto the grid. $\mathbf{T}_e \xrightarrow{N\to\infty} 0$

Example:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + F(x,t) \ , \qquad x \in [0, 2\pi) \ , t \geq 0 \\ u(t=0) &= f(x) \end{aligned}$$

with periodic boundary conditions. Consider the approximation:

$$\begin{aligned} \mathbf{u}_{xx} &\approx \frac{1}{h^2} \begin{pmatrix} \ddots & \ddots & \ddots & & & 1 \\ & 1 & -\mathbf{2} & 1 & & \\ & & 1 & -\mathbf{2} & 1 & \\ 1 & & & \ddots & \ddots & \ddots \end{pmatrix} \mathbf{u} \\ &= D_+ D_- \mathbf{u} \ . \end{aligned}$$

Then
$(T_e)_j = \frac{h^2}{12} \left(u_j\right)_{xxxx} + O(h^4)$ and $(\mathbf{w}, D_+ D_- \mathbf{w}) \leq 0$

Consider the semi–discrete approximation:

$$\frac{\partial \mathbf{v}}{\partial t} = Q\mathbf{v} \ , \quad t \geq 0$$

$$\mathbf{v}(t=0) = \mathbf{f} \, .$$

Proposition: Under Assumptions 1–3 The semi–discrete approximation converges.

Proposition: Under Assumptions 1–3 The semi–discrete approximation converges.

Proof: Let $\mathbf{u}$ is the projection of $u(x,t)$ onto the grid. Then

$$\frac{\partial \mathbf{u}}{\partial t} = P\mathbf{u} = Q\mathbf{u} + \mathbf{T}_e$$

$$\frac{\partial \mathbf{v}}{\partial t} = Q\mathbf{v}$$

Let $\mathbf{E} = \mathbf{u} - \mathbf{v}$ then

$$\frac{\partial \mathbf{E}}{\partial t} = Q\mathbf{E} + \mathbf{T}_e$$

$$\frac{\partial \mathbf{E}}{\partial t} = Q\mathbf{E} + \mathbf{T}_e$$

By taking the H scalar product with $\mathbf{E}$:

$$\begin{aligned}\left(\mathbf{E}, \frac{\partial \mathbf{E}}{\partial t}\right)_H &= \frac{1}{2}\frac{\partial}{\partial t}(\mathbf{E},\mathbf{E})_H = \|\mathbf{E}\|_H \frac{\partial}{\partial t}\|\mathbf{E}\|_H \\ &= (\mathbf{E}, Q\mathbf{E})_H + (\mathbf{E}, \mathbf{T}_e)_H \\ &\leq \alpha \|\mathbf{E}\|_H^2 + \|\mathbf{E}\|_H \|\mathbf{T}_e\|_H\end{aligned}$$

Thus

$$\frac{\partial}{\partial t}\|\mathbf{E}\|_H \leq \alpha \|\mathbf{E}\|_H + \|\mathbf{T}_e\|_H$$

Therefore:

$$\|\mathbf{E}\|_H(t) \leq \|\mathbf{E}\|_H(0)e^{\alpha t} + \frac{e^{\alpha t}-1}{\alpha} \max_{0\leq\tau\leq t} \|\mathbf{T}_e\|_H \xrightarrow{N\to\infty} 0$$

Fully-discrete approximations for PDEs or ODEs.
Consider the differential problem:

$$\begin{aligned} \frac{\partial u}{\partial t} &= P u \\ u(t=0) &= f \, . \end{aligned}$$

It is assumed that this problem is well posed, In particular $\exists K(t) < \infty$ s.t. $||u(t)|| \leq K(t)||f||$. Typically $K(t) = Ke^{\alpha t}$.

Remark: in order to simplify the explanation we consider the constant coefficients P.

Consider the multistep approximation:

$$v_{n+1} = \sum_{j=0}^{p} Q_j v_{n-j}$$

where $t_n = n\Delta t$ and v_n is the approximation to $u(t_n)$.
Denoting:

$$\begin{aligned} U_n &= (u(t_n), u(t_{n-1}), ..., u(t_{n-p}))^T \\ V_n &= (v_n, v_{n-1}, ..., v_{n-p})^T \ . \end{aligned}$$

The scheme can be written as

$$V_{n+1} = \begin{pmatrix} Q_0 & Q_1 & \dots & & Q_{n-p} \\ I & & & & \\ 0 & I & & & \\ & & \ddots & & \\ 0 & \dots & & I & 0 \end{pmatrix} V_n = Q\, V_n$$

We assume:

Assumption 1: In some equivalent norm $\|\cdot\|_H$

$$\|Q\|_H \leq 1 + \alpha \Delta t$$

We assume:

Assumption 1: In some equivalent norm $\|\cdot\|_H$

$$\|Q\|_H \leq 1 + \alpha \Delta t$$

Assumption 2: The local truncation error of Q is T_n which is defined by

$$\Delta t T_n = W_{n+1} - QW_n$$

where W_{n+1} is the solution of the PDE/ODE whoe 'initial condition' is W_n at t_n. It is assumed that $T_n \xrightarrow{N\to\infty} 0$

Similar to the semi-discrete case

$$U_{n+1} = QU_n + \Delta t T_n$$

$$V_{n+1} = QV_n$$

Let $E_n = U_n - V_n$ then

$$E_{n+1} = QE_n + \Delta t T_n$$

Denoting by

$$V_n \;=\; S_{\Delta t}(t_n, t_\nu)\, V_\nu \quad (=\; Q^{n-\nu}\, V_\nu \text{ for constant coefficients})$$

Then, using the discrete Duhamel's principle

$$E_n \;=\; S_{\Delta t}\,(t_n, 0)\, E_0 \;+\; \Delta t \sum_{\nu=0}^{n-1} S_{\Delta t}\,(t_n, t_{\nu+1})\, T_\nu \;,$$

or, equivalently

$$E_n \;=\; Q^n E_0 \;+\; \Delta t \sum_{\nu=0}^{n-1} Q^{n-\nu-1}\, T_\nu \;.$$

Therefore, using $\|Q^\mu\|_H \leq (1 + \alpha \Delta t)^\mu \approx e^{\alpha t_\mu}$:

$$\|E_n\|_H \;\leq\; \|E_0\|_H\, e^{\alpha t} + \frac{e^{\alpha t} - 1}{\alpha} \max_{0 \leq \mu \leq 0} \|T_\mu\|_H \xrightarrow{N \to \infty} 0$$

Indeed, for all the classical schemes, e.g.

ODE	PDE
Euler	Forward Euler
Backward Euler	Backward Euler
Trapezoid	Lax–Friedrichs
Multistep methods	Lax–Wendroff
Runge–Kutta methods	Crank–Nicholson
	Leap–Frog
	Compact schemes
	Deferred–correction methods
	FE (see Strang and Fix)

$$\|E_n\|_H = O\left(\|T_\mu\|_H\right).$$

Observation

$$\frac{\partial \mathbf{E}}{\partial t} = Q\mathbf{E} + \Delta t \mathbf{T}_e \text{ and } E_{n+1} = QE_n + T_n$$

are exact while

$$\|\mathbf{E}\|_H(t) \leq \|\mathbf{E}\|_H(0)e^{\alpha t} + \frac{e^{\alpha t} - 1}{\alpha} \max_{0 \leq \tau \leq t} \|\mathbf{T}_e\|_H \xrightarrow{N \to \infty} 0$$

and

$$\|E_n\|_H \leq \|E_0\|_H e^{\alpha t} + \frac{e^{\alpha t} - 1}{\alpha} \max_{0 \leq \mu \leq 0} \|T_\mu\|_H \xrightarrow{N \to \infty} 0$$

are estimates!

Preliminary example:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + F(x,t) \quad , \qquad x \in [0, 2\pi) \, , t \geq 0 \\ u(t=0) &= f(x) \end{aligned}$$

with periodic boundary conditions.

Preliminary example:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + F(x,t) \ , \qquad x \in [0, 2\pi) \ , t \geq 0 \\ u(t=0) &= f(x) \end{aligned}$$

with periodic boundary conditions.
Consider the scheme

$$\begin{aligned} \frac{\partial v_j}{\partial t} &= D_+ D_- v_j + (-1)^j c\, v_j + F(x_j, t) \ ; \ \ x_j = jh, \ \ h = 2\pi/N, \\ v_j(t=0) &= f_j \end{aligned}$$

where N is even.

Preliminary example:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + F(x,t) \ , \qquad x \in [0, 2\pi) \ , t \geq 0 \\ u(t=0) &= f(x) \end{aligned}$$

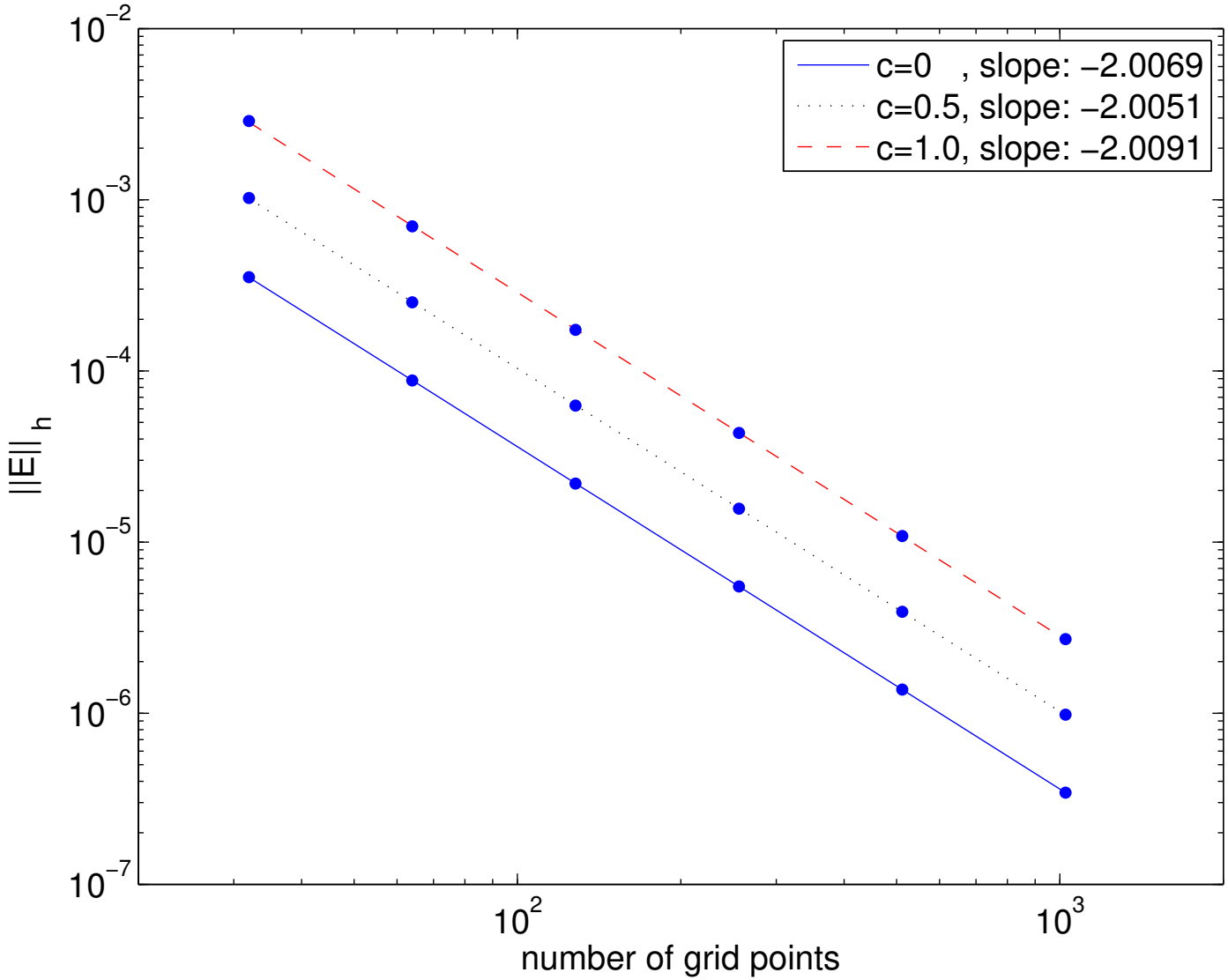
with periodic boundary conditions.
Consider the scheme

$$\begin{aligned} \frac{\partial v_j}{\partial t} &= D_+ D_- v_j + (-1)^j c\, v_j + F(x_j, t) \ ; \ \ x_j = jh, \ \ h = 2\pi/N, \\ v_j(t=0) &= f_j \end{aligned}$$

where N is even.
The truncation error is

$$(T_e)_j = \frac{h^2}{12} \left(u_j\right)_{xxxx} - (-1)^j c\, v_j = O(1)$$

However:


c=0 , slope: –2.0069
c=0.5, slope: –2.0051
c=1.0, slope: –2.0091
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
$||E||_h$
10^2
10^3
number of grid points

Error inhibiting schemes for ODEs

Ordinary Differential Equations.

Consider the differential problem:

$$\begin{aligned} \frac{\partial u}{\partial t} &= f\,u\,, \quad f = const \\ u(t=0) &= u_0\,. \end{aligned}$$

It can be solved using a s steps multystep method such as the Adams-Bashforth scheme.

The first Dahlquist barrier states that any explicit, s step, linear multistep method can be of order less or equal to s.

Constructing an error inhibiting method

Define vectors of length s that contains the exact and numerical solutions at times $(t_n + j\Delta t/s)$ for $j = 0, \ldots, s-1$

$$U_n = \left(u(t_{n+(s-1)/s}), \ldots, u(t_{n+1/s}), u(t_n)\right)^T, \tag{1}$$

$$V_n = \left(v_{n+(s-1)/s}, \ldots, v_{n+1/s}, v_n\right)^T. \tag{2}$$

This scheme uses s terms for generating the next s terms, unlike explicit linear multistep methods which use s terms to generate one term.
The block one-step method can be written as:

$$V_{n+1} = QV_n \quad \text{where} \quad Q = A + \Delta tBf$$

This particular formulation is called a Type 3 DIMSIM in Butcher's 1993 paper. Implicit one-step: Shampine, L.F., Watts, H.A 1969.

Constructing an error inhibiting method

Suppose we construct a method such that:

C1. $\mathbf{rank}(A) = 1$.

C2. Its non-zero eigenvalue is equal to 1 and its corresponding eigenvector is

$$(1, \ldots, 1)^T .$$

C3. A can be diagonalized.

C4. The matrices A and B are constructed such that:

$$\|Q\tau_\nu\| \leq O(\Delta t)\,\|\tau_\nu\| \quad (\textbf{note} : \Delta t\,\tau_n = U_{n+1} - Q_n U_n)$$

This is accomplished by requiring the local truncation error to live in the null space of A.

Constructing an error inhibiting method

Property [C2] assures that the method produces the exact solution for the trivial case $u_t = 0$, i.e. $f = 0$.

Note that the term $\Delta t B f$ is only an $O(\Delta t)$ perturbation to A, so the matrix Q will have one eigenvalue, $z_1 = 1 + O(\Delta t)$ whose eigenvector has the form

$$\psi_1 = (1 + O(\Delta t), \ldots, 1 + O(\Delta t))^T$$

and the rest of the eigenvalues satisfy $z_j = O(\Delta t)$ for $j = 2, \ldots, s$. Since the $\|Q\| = 1 + O(\Delta t)$ we can conclude that this scheme is stable.

Property [C4] makes the error inhibiting magic happen.

Constructing an error inhibiting method

Recall that we defined the truncation error
$\Delta t \tau_n = U_{n+1} - Q_n U_n$.
The global error is $E_n = U_n - V_n$, and its evolution can be described, in the linear constant coefficient case, by

$$E_n = Q^n E_0 + \Delta t \sum_{\nu=0}^{n-1} Q^{n-\mu-1} \tau_\nu \ .$$

- The initial error E_0, which is assumed to very small.
- The last term in the sum, $\Delta t \tau_{n-1}$, is by definition $O(\Delta t)\|\tau_{n-1}\|$.
- The rest of the sum, $\Delta t \sum_{\nu=0}^{n-2} Q^{n-\nu-1}\tau_\nu$, has the potential of accumulating – this is the term we need to bound!

Constructing an error inhibiting method:

Recall that [C4] ensures that

$$\|Q\tau_\nu\| \leq O(\Delta t)\,\|\tau_\nu\|\,.$$

$$\begin{aligned}
\left\|\Delta t \sum_{\nu=0}^{n-2} Q^{n-\nu-1}\tau_\nu\right\| &\leq \Delta t \sum_{\nu=0}^{n-2} \left\|Q^{n-\nu-2}\right\| \|Q\tau_\nu\| \\
&\leq \Delta t \sum_{\nu=0}^{n-2} \|Q\|^{n-\nu-2}\, O(\Delta t)\|\tau_\nu\| \quad \text{due to [C4]} \\
&\leq O(\Delta t)\left(\max_{\nu=0,\ldots,n-2}\|\tau_\nu\|\right)\Delta t \sum_{\nu=0}^{n-2}(1+c\Delta t)^{n-\nu-2} \\
&\leq O(\Delta t)\left(\max_{\nu=0,\ldots,n-2}\|\tau_\nu\|\right).
\end{aligned}$$

This is one order higher than you would normally get!

Constructing an error inhibiting method: extension to non-constant coefficient and nonlinear

Is this still true for the more general, nonlinear case?

Yes! We proved this in the paper:
Adi Ditkowski, and Sigal Gottlieb. "Error Inhibiting Block One-step Schemes for Ordinary Differential Equations." Journal of Scientific Computing (2017): 1-21.

You can read the ArXiv version at
https://arxiv.org/abs/1701.08568

The numerical results that follow demonstrate that this works.

A third order error inhibiting method with $s = 2$.

Our first error inhibiting scheme takes the values of the solution at the times t_n and $t_{n+\frac{1}{2}}$ and obtains the solution at the time-level t_{n+1} and $t_{n+\frac{3}{2}}$.

The exact solution vector for this problem is $U_n = \left(u(t_{n+1/2}), u(t_n)\right)^T$ and the vector of numerical approximations is $V_n = \left(v_{n+1/2}, v_n\right)^T$.

The scheme is given by:

$$V_{n+1} = \frac{1}{6}\begin{pmatrix} -1 & 7 \\ -1 & 7 \end{pmatrix} V_n + \frac{\Delta t}{24}\begin{pmatrix} 55 & -17 \\ 25 & 1 \end{pmatrix}\begin{pmatrix} f\left(v_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}\right) \\ f\left(v_n, t_n\right) \end{pmatrix},$$

A third order error inhibiting method with $s = 2$.

This method has truncation error

$$\tau_n = \frac{23}{576} \begin{pmatrix} 7 \\ 1 \end{pmatrix} \frac{d^3}{dt^3} u(t_n) \Delta t^2 + O(\Delta t^3) .$$

The matrix A can be diagonalized as follows:

$$A = \frac{1}{6} \begin{pmatrix} -1 & 7 \\ -1 & 7 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 1 & 7 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & \\ & 0 \end{pmatrix} \begin{pmatrix} -1 & 7 \\ 1 & -1 \end{pmatrix}$$

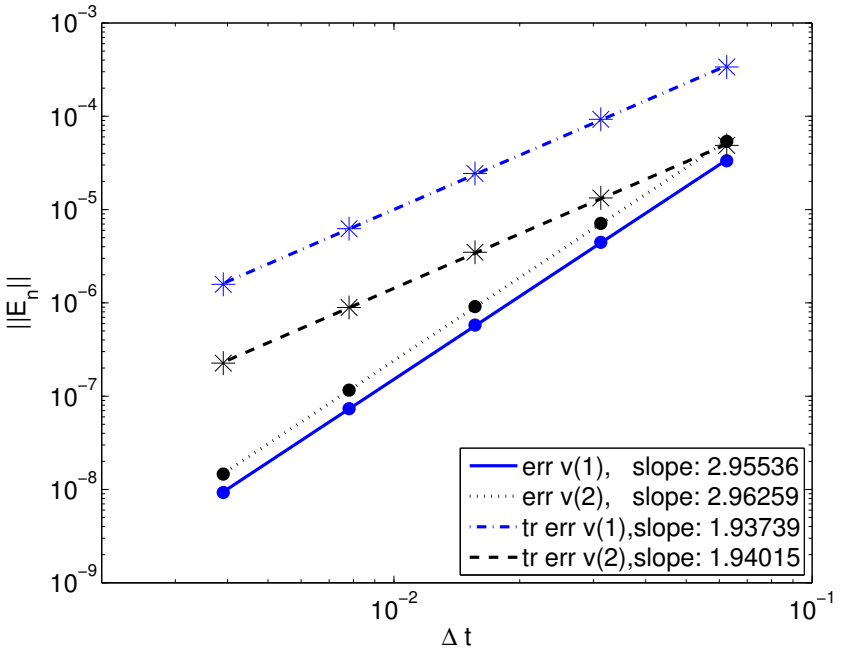
Note that the leading order of the truncation error is in the space of the second eigenvector of A, the one that corresponds to the zero eigenvalue. This is what gives the error inhibiting property.

EIS on a nonlinear scalar equation

Given the nonlinear scalar equation of the form:

$$u_t = -u^2 = f(u) \ , \qquad t \geq 0$$
$$u(t=0) = 1 \ . \tag{3}$$

We see the truncation error is only second order but the global error is third order:

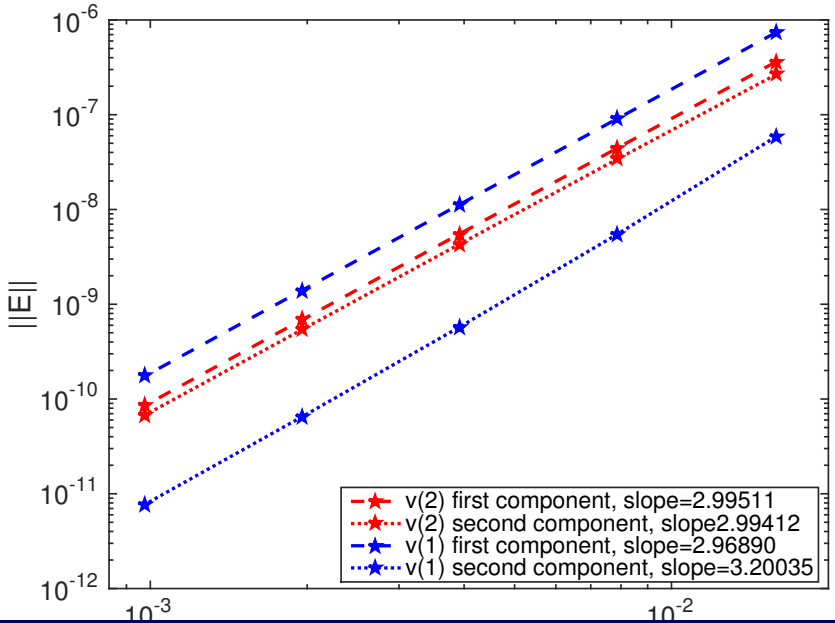


EIS on a nonlinear system

This works on a nonlinear system as well! Consider the van der Pol system

$$\begin{aligned} u_t^{(1)} &= u^{(2)} \\ u_t^{(2)} &= 0.1[1-(u^{(1)})^2]u^{(2)} - u^{(1)} \end{aligned} \tag{4}$$

Once again, we see that the convergence rate is indeed third order:



Not all Type 3 DIMSIM methods are error inhibiting!

It is important to note that not all Type 3 DIMSIM methods are error inhibiting!

The property that the local truncation error lives in the space spanned by the eigenvectors of A that correspond to the zero eigenvalues is needed for the error inhibiting behavior to occur, and this property is not generally satisfied.

To observe this, we study the DIMSIM scheme of Type 3 presented by J. C. Butcher in his 1993 paper.

Not all Type 3 DIMSIM methods are error inhibiting!

The scheme

$$\begin{pmatrix} v_{n+3} \\ v_{n+2} \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 7 & -3 \\ 7 & -3 \end{pmatrix}\begin{pmatrix} v_{n+1} \\ v_n \end{pmatrix} + \frac{\Delta t}{8}\begin{pmatrix} 9 & -7 \\ -3 & -3 \end{pmatrix}\begin{pmatrix} f(v_{n+1}, t_{n+1}) \\ f(v_n, t_n) \end{pmatrix}$$

was given by Butcher in his 1993 paper on DIMSIM methods.

This scheme has truncation error

$$\tau_n = \frac{1}{48}\begin{pmatrix} 23 \\ 3 \end{pmatrix}\frac{d^3}{dt^3}u(t_n)\,\Delta t^2 + O(\Delta t^3)\,.$$

Not all Type 3 DIMSIM methods are error inhibiting!

The matrix A can be diagonalized as follows:

$$A = \frac{1}{4}\begin{pmatrix} 7 & -3 \\ 7 & -3 \end{pmatrix} \tag{5}$$

$$= \begin{pmatrix} 1 & 3/7 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & \\ & 0 \end{pmatrix}\frac{1}{4}\begin{pmatrix} 7 & -3 \\ -7 & 7 \end{pmatrix}. \tag{6}$$

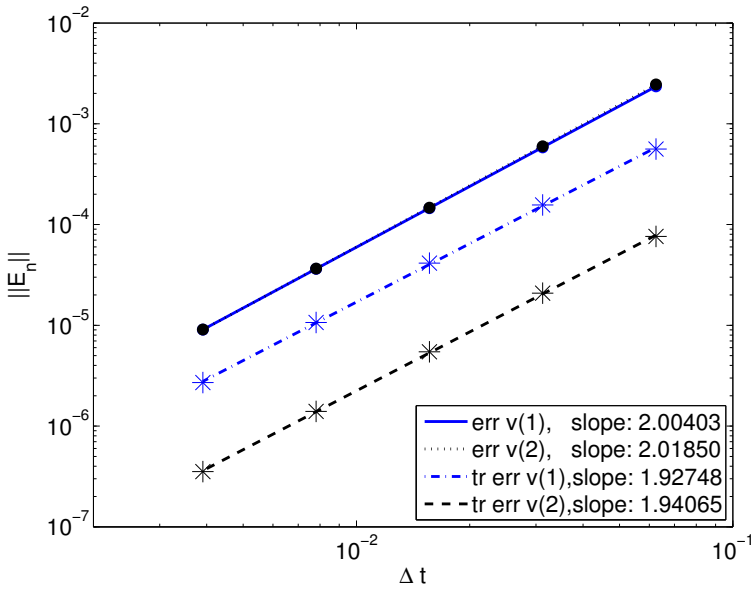
The truncation error τ_n can be written as a linear combination of the two eigenvectors of A as follows:

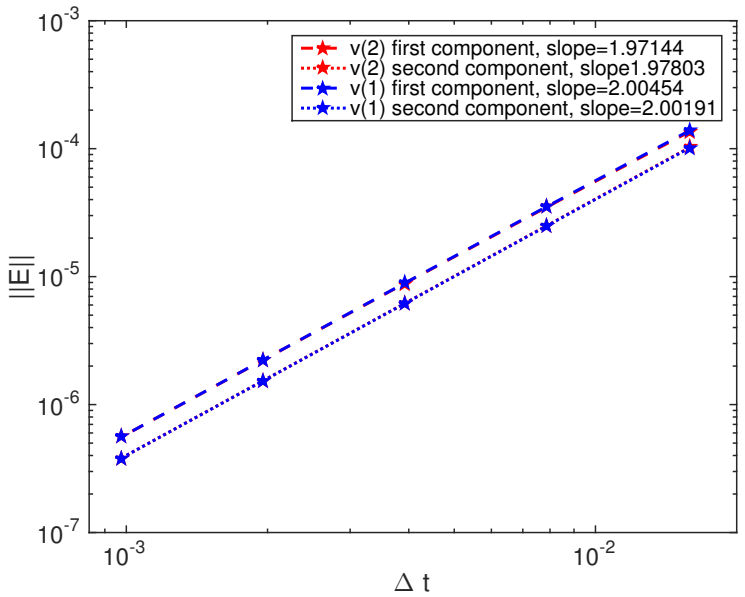
$$\tau_n = \left[\frac{19}{24}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{35}{48}\begin{pmatrix} 3/7 \\ 1 \end{pmatrix}\right]\frac{d^3}{dt^3}u(t_n)\,\Delta t^2 + O(\Delta t^3)\,. \tag{7}$$

Unlike the error inhibiting scheme, here the first term in this expansion is of the order of $O(\tau_n) = O(\Delta t^2)$ so a term of order $\Delta t O(\tau_n) = O(\Delta t^3)$ is accumulated at each time step, and the global error is second order.

Not all Type 3 DIMSIM methods are error inhibiting!

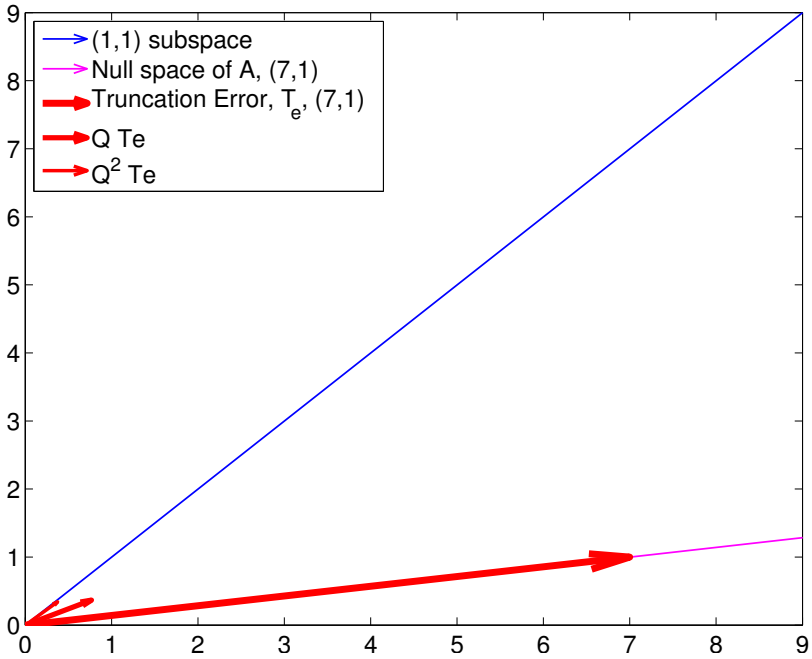
Both this method and our error inhibiting method satisfy the order conditions in Theorem 3.1 of Butcher's paper only up to second order ($p = 2$). But this method gives second order accuracy, while our error inhibiting method gave third order accuracy in the same example.



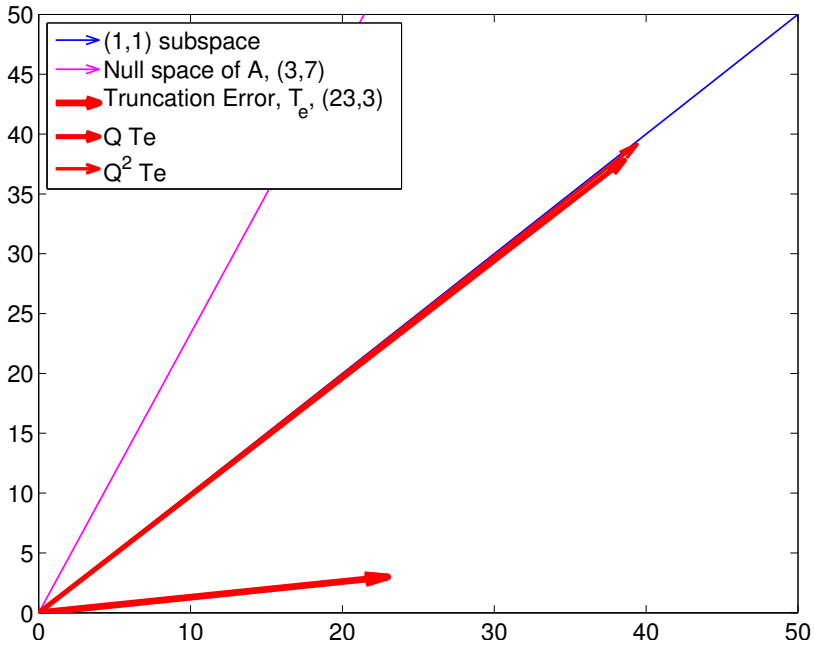


EIS ev Type 3 DIMSIM

EIS scheme



Type 3 DIMSIM



τ, $Q\tau$ and $Q^2\tau$, $Q = A + \Delta t B$ for both schemes ($\Delta t = 1/20$).

A fourth order error inhibiting methods with $s = 3$.

This method takes the values of the solution at the times t_n, $t_{n+\frac{1}{3}}$, and $t_{n+\frac{2}{3}}$
and uses these three values to obtain the solution at the time-level
t_{n+1}, $t_{n+\frac{4}{3}}$, and $t_{n+\frac{5}{3}}$.

The exact solution vector is given by
$U_n = \left(u(t_{n+2/3}), u(t_{n+1/3}), u(t_n)\right)^T$, and the vector of numerical approximations is $V_n = \left(v_{n+2/3}, v_{n+1/3}, v_n\right)^T$.

A fourth order error inhibiting methods with $s = 3$.

This method is given by:

$$V_{n+1} = \frac{1}{768}\begin{pmatrix} 467 & -1996 & 2297 \\ 467 & -1996 & 2297 \\ 467 & -1996 & 2297 \end{pmatrix} V_n + \frac{\Delta t}{1152}\begin{pmatrix} 5439 & -6046 & 3058 \\ 2399 & -1694 & 1362 \\ 703 & 354 & 626 \end{pmatrix}\begin{pmatrix} f\left(v_{n+2/3}, t_{n+2/3}\right) \\ f\left(v_{n+1/3}, t_{n+1/3}\right) \\ f\left(v_n, t_n\right) \end{pmatrix}$$

which has a local truncation error of third order,

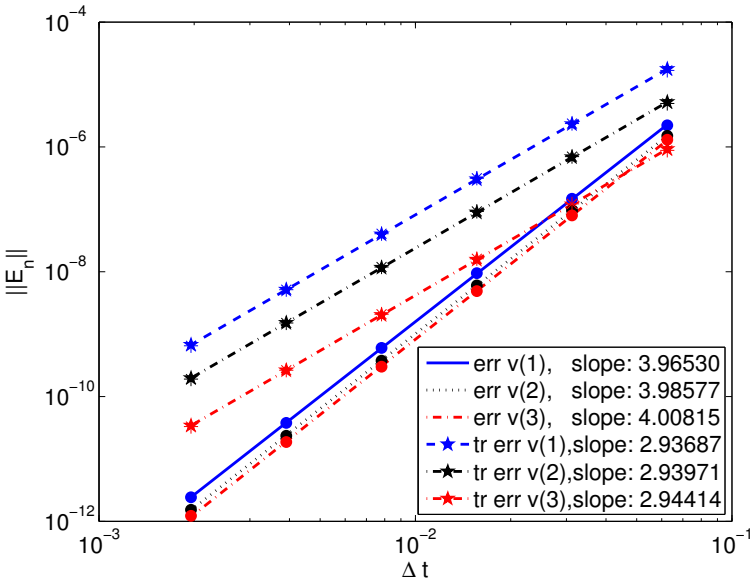
$$\tau_n = \frac{1}{373248}\begin{pmatrix} 43699 \\ 12787 \\ 2227 \end{pmatrix}\frac{d^4}{dt^4}u(t_n)\,\Delta t^3 + O(\Delta t^4)$$

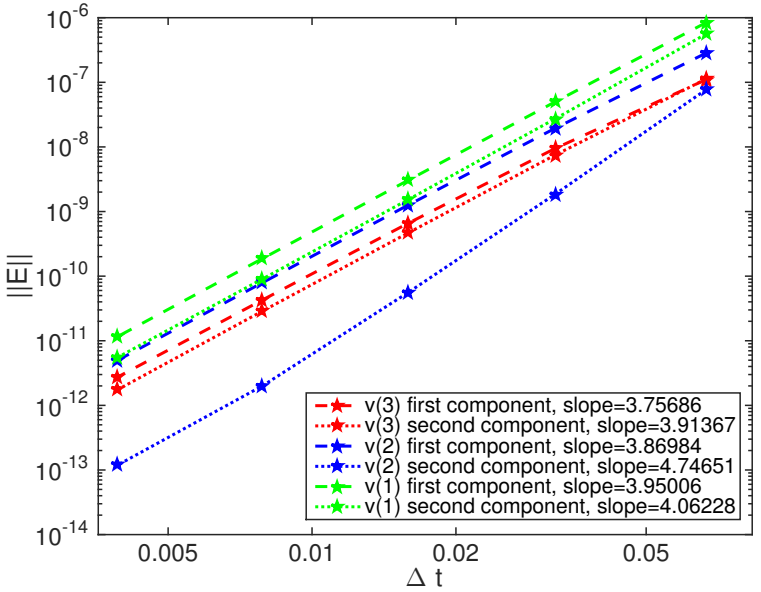
It can be verified that

$$Q_n\tau_n = O(\Delta t\tau_n) = O(\Delta t^4)\,.$$

A fourth order error inhibiting methods with $s = 3$.

To demonstrate this result we revisit the two examples above:





Although the local truncation errors are only third order, the global errors are fourth order.

Error inhibiting schemes for PDEs

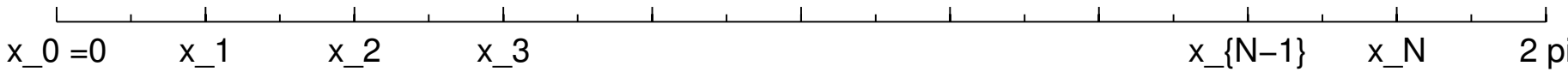
Block Finite Difference, EIS schemes ,for the Heat equation: 2 points block, 3rd order scheme

Consider the Heat equation

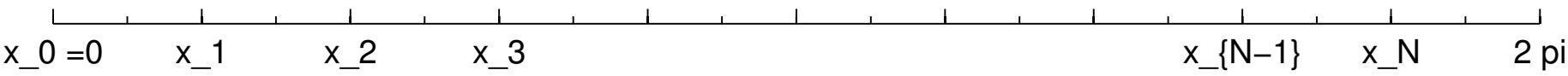
$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} \quad , \qquad x \in [0, 2\pi) \, , t \geq 0 \\ u(t=0) &= f(x) \end{aligned}$$

with periodic boundary conditions.

we use the grid, $x_j = j\,h$, $x_{j+1/2} = j\,h + h/2$, $h = 2\pi/(N+1)$ (altogether $2(N+1)$ points with spacing of $h/2$).



$$x_j = j\,h,\; x_{j+1/2} = j\,h + h/2,\; h = 2\pi/(N+1)\;.$$



and the approximation:

$$\mathbf{u}_{xx} \approx \frac{1}{(h/2)^2}\left[\begin{pmatrix} \ddots & \ddots & \ddots & & \\ & 1 & \mathbf{-2} & 1 & \\ & & 1 & \mathbf{-2} & 1 \\ & & & \ddots & \ddots \end{pmatrix} + c\begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \\ & -1 & \mathbf{3} & -3 & 1 \\ & 1 & -3 & \mathbf{3} & -1 \\ & & \ddots & \ddots & \ddots \end{pmatrix}\right]\mathbf{u}$$

The truncation error is

$$(T_e)_j = \frac{1}{12}\left(\frac{h}{2}\right)^2 (u_j)_{xxxx} + c\left[\left(\frac{h}{2}\right)(u_j)_{xxx} + \frac{1}{2}\left(\frac{h}{2}\right)^2 (u_j)_{xxxx}\right] + O(h^3) = O(h)$$

$$\begin{aligned}(T_e)_{j+\frac{1}{2}} &= \frac{1}{12}\left(\frac{h}{2}\right)^2 \left(u_{j+\frac{1}{2}}\right)_{xxxx} + c\left[-\left(\frac{h}{2}\right)\left(u_{j+\frac{1}{2}}\right)_{xxx} + \frac{1}{2}\left(\frac{h}{2}\right)^2 \left(u_{j+\frac{1}{2}}\right)_{xxxx}\right] + O(h^3) \\ &= O(h)\end{aligned}$$

$$T_e = O(h)$$

Analysis

Note that $x_j = j\,h$, $h = 2\pi/(N+1)$ then for $\omega \in \{-N/2, \ldots, N/2\}$

$$\mathrm{e}^{i\omega x_j} = \mathrm{e}^{-i(\omega - \mathrm{sign}(\omega)(N+1))x_j} \quad \text{and} \quad \mathrm{e}^{i\omega x_{j+1/2}} = -\mathrm{e}^{-i(\omega - \mathrm{sign}(\omega)(N+1))x_{j+1/2}}$$

Therefore we can look for eigenvectors in the form of:

$$\psi_k(\omega) = \frac{\alpha_k}{\sqrt{2\pi}} \begin{pmatrix} \vdots \\ \mathrm{e}^{i\omega x_j} \\ \mathrm{e}^{i\omega x_{j+1/2}} \\ \vdots \end{pmatrix} + \frac{\beta_k}{\sqrt{2\pi}} \begin{pmatrix} \vdots \\ \mathrm{e}^{-i(\omega - \mathrm{sign}(\omega)(N/2))x_j} \\ \mathrm{e}^{-i(\omega - \mathrm{sign}(\omega)(N/2))x_{j+1/2}} \\ \vdots \end{pmatrix}$$

where $|\alpha_k|^2 + |\beta_k|^2 = 1$, $k = 1, 2$.

The hairy expressions for α_k, β_k and the eigenvalues can be found. The eigenvalues are real and non positive for $|c| < 1/2$!

For $\omega h \ll 1$ the eigenvalues and eigenvectors are:

$$\hat{Q}_1(\omega) = -\omega^2 + \frac{(1+4c)\omega^4}{12-24c}\left(\frac{h}{2}\right)^2 + O(h^4)$$

$$\alpha_1 = 1 - \frac{c^2}{32(1-2c)^2}\left(\frac{\omega h}{2}\right)^6 + O\left(\mathrm{h}^7\right), \beta_1 = -\frac{ic}{4-8c}\left(\frac{\omega h}{2}\right)^3 + O(h^5)$$

and

$$\hat{Q}_2(\omega) = -\frac{4-8c}{(\mathrm{h}/2)^2} + (1-4c)\omega^2 + O(h^2)$$

$$\alpha_2 = \frac{ic}{2c-1}\left(\frac{\omega h}{2}\right) + O(h^3) \quad , \quad \beta_2 = 1 + O(h^2)$$

If the initial condition is

$$\mathbf{v}_j(0) = \mathrm{e}^{i\omega x_j} \quad ; \quad \omega^2 h \ll 1$$

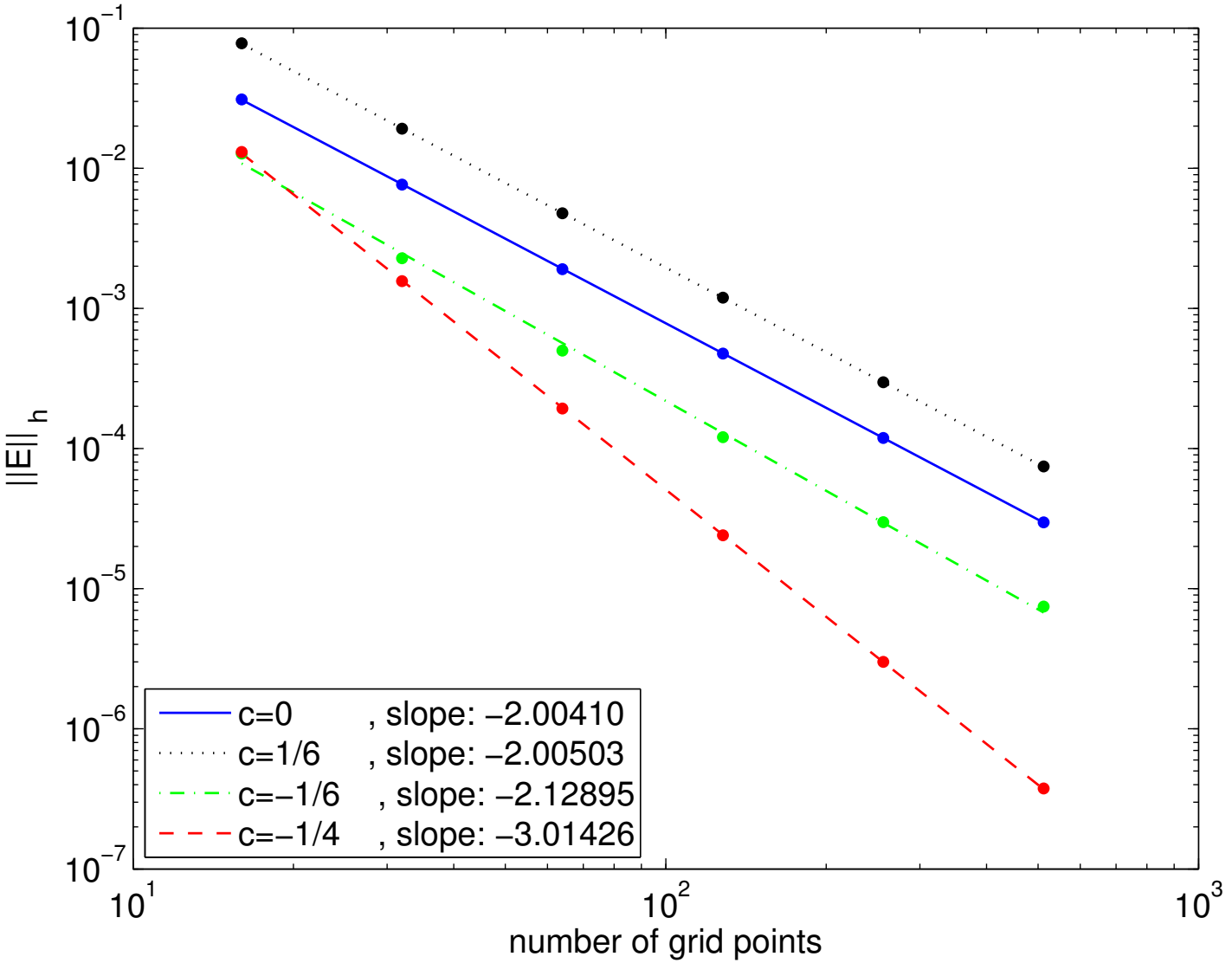
Then

$$\begin{aligned}(\mathbf{v})_j(t) &= \mathrm{e}^{-\omega^2 t}\left[\left(1-\frac{(1+4c)\omega^2 t}{12-24c}\left(\frac{\omega h}{2}\right)^2+O(h^4)\right)\mathrm{e}^{i\omega x_j}+\right.\\ &\left.\left(-\frac{ic}{4-8c}\left(\frac{\omega h}{2}\right)^3+O(h^5)\right)\mathrm{e}^{-i(\omega-\mathrm{sign}(\omega)(N/2))x_j}\right]\end{aligned}$$

The same expression hold for $x_{j+\frac{1}{2}}$.

Therefore the scheme is 2nd order. It is 3rd order if $c = -1/4$.

Indeed:


10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
$||E||_h$
10^1
10^2
10^3
number of grid points
c=0 , slope: –2.00410
c=1/6 , slope: –2.00503
c=–1/6 , slope: –2.12895
c=–1/4 , slope: –3.01426

Observation: for the 2 points block, the solution is:

$$(\mathbf{v})_j(t) = e^{-\omega^2 t}\left[\left(1-\frac{(1+4c)\omega^2 t}{12-24c}\left(\frac{\omega h}{2}\right)^2+O(h^4)\right)e^{i\omega x_j}+\left(-\frac{ic}{4-8c}\left(\frac{\omega h}{2}\right)^3+O(h^5)\right)e^{-i(\omega-\mathrm{sign}(\omega)(N/2))x_j}\right]$$

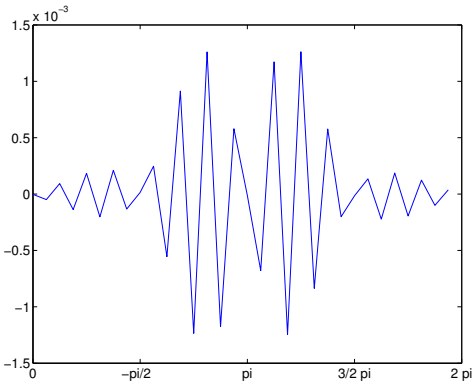
For the 3rd order scheme, $c=-1/4$, the error is highly oscillatory.

Observation: for the 2 points block, the solution is:

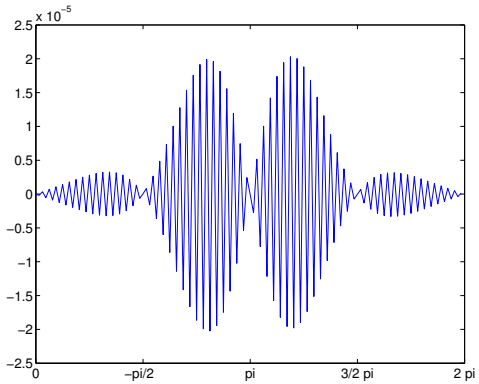
$$(\mathbf{v})_j(t) = \mathrm{e}^{-\omega^2 t}\left[\left(1-\frac{(1+4c)\omega^2 t}{12-24c}\left(\frac{\omega h}{2}\right)^2+O(h^4)\right)\mathrm{e}^{i\omega x_j}+\left(-\frac{ic}{4-8c}\left(\frac{\omega h}{2}\right)^3+O(h^5)\right)\mathrm{e}^{-i(\omega-\mathrm{sign}(\omega)(N/2))x_j}\right]$$

For the 3rd order scheme, $c=-1/4$, the error is highly oscillatory.

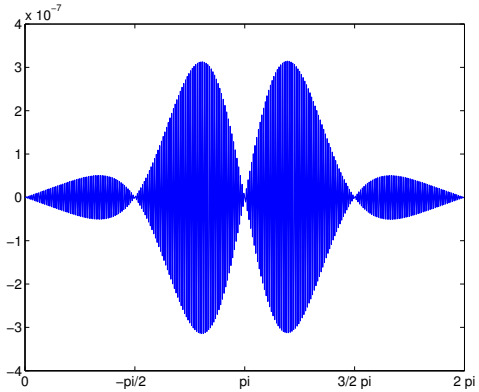
$N=16$



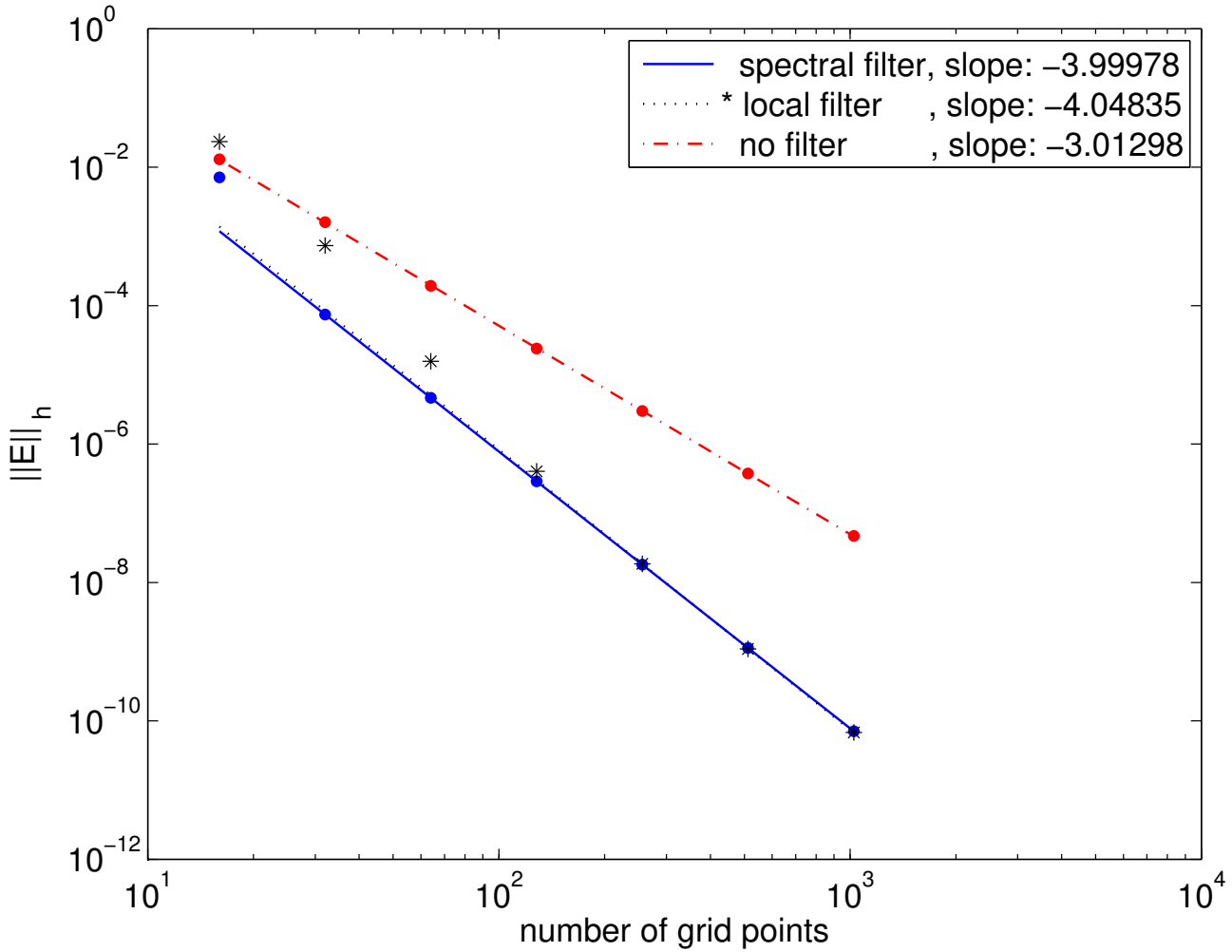
$N=64$



$N=256$



It was suggested by Jennifer k. Ryan that this term could be filtered at the final time. This method is called "post–processing"

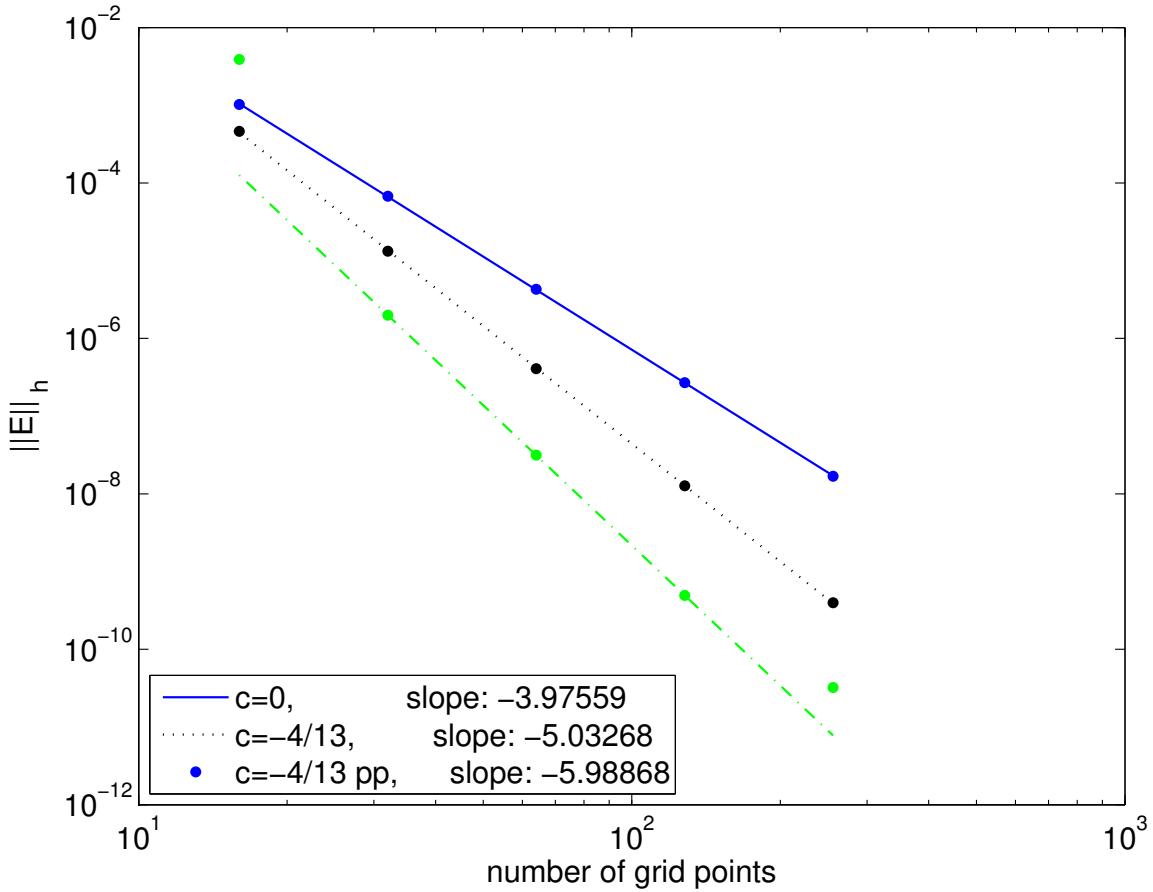


2 points block, 5th order scheme

Consider the approximation:

$$\mathbf{u}_{xx} \approx \frac{1}{12(h/2)^2}\left[\begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & -1 & 16 & \mathbf{-30} & 16 & -1 & \\ & & -1 & 16 & \mathbf{-30} & 16 & -1 \\ & & & \ddots & \ddots & \ddots & \ddots \end{pmatrix} + c\begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & 1 & -5 & \mathbf{10} & -10 & 5 & -1 \\ & -1 & 5 & -10 & \mathbf{10} & -5 & 1 \\ & & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}\right]\mathbf{u}$$

Using the same analysis as in the 3rd order scheme, It was shown that this is a 5th order scheme (6th order with post processing).



Is it a finite difference scheme?

- Note that we are talking on blocks rather than points.

Is it a finite difference scheme?

- Note that we are talking on blocks rather than points.

- In
  M. Zhang and C-W Shu, AN ANALYSIS OF THREE DIFFERENT FORMULATIONS OF THE DISCONTINUOUS GALERKIN METHOD FOR DIFFUSION EQUATIONS, Mathematical Models and Methods in Applied Sciences Vol. 13, No. 3 (2003) 595–413.
  The authors derived DCG schemes for the heat equations and observed the same phenomenon.

  This paper was motivating the current work.

Summary

- We showed that schemes can be constructed such that their convergence rates are higher than their truncation errors.

  It was done by having the truncation errors lies in a different subspace than the solution and constructing the numerical operators such that they attenuate the truncation errors and inhibit them from accumulating over time.

Summary

- We showed that schemes can be constructed such that their convergence rates are higher than their truncation errors.

  It was done by having the truncation errors lies in a different subspace than the solution and constructing the numerical operators such that they attenuate the truncation errors and inhibit them from accumulating over time.

- This methodology may be applied to other numerical methods, such as finite elements and Discontinuous Galerkin (DG).

THANK YOU !