ImEx Stability with Applications to the Dispersive Shallow Water Equations

David Shirokoff

CNA at CMU

ICERM

January 10, 2022

In Collaboration with:

Rodolfo Rubén Rosales (MIT)
Benjamin Seibold (Temple)
Dong Zhou (CalState LA)
Wooyoung Choi (NJIT)
Linwan Feng (NJIT)

DMS-2012268
DMS-1719693

Theory paper: SINUM, 55:5 (2017), 2336-2360
Practice paper: JCP, 376:1 (2019), 295-321

Part I
The Stability Theory

ImEx schemes

Goal Time Step ODE: $$\boxed{\boldsymbol{u}_t = L\boldsymbol{u} + \boldsymbol{f}(t)}$$

where given initial data and: $L \in \mathbb{R}^{N\times N} \quad \boldsymbol{u}(t), \boldsymbol{f}(t) \in \mathbb{R}^N$

If, L stiff, try **IMEX**: $$\boldsymbol{u}_t = \underbrace{A\boldsymbol{u}}_{\textbf{IM}\text{plicit}} + \underbrace{B\boldsymbol{u}}_{\textbf{EX}\text{plicit}} + \boldsymbol{f}(t)$$

- where $L = A + B$ (NOT unique!)
- Convention: A stiff, B non-stiff,

Example:
$$\frac{1}{k}(\boldsymbol{u}_{n+1} - \boldsymbol{u}_n) = A\boldsymbol{u}_{n+1} + B\boldsymbol{u}_n$$

Difficulty: Both A, B are stiff

- Occurs when B is difficult to treat implicitly.

<u>Example 1</u>: Variable coefficient diffusion, try splitting

$$u_t = \big(d(x)u_x\big)_x = \alpha u_{xx} + \Big(\big(d(x) - \alpha\big)u_x\Big)_x$$

- $A\boldsymbol{u} = \alpha u_{xx}$ (**IM**plicit)
- $B\boldsymbol{u} = \Big(\big(d(x) - \alpha\big)u_x\Big)_x$ (**EX**plicit)

Not trivial to avoid diffusive time step.

- Old idea [Duglous, Dupont, 1971]
- For α large enough, simple Euler scheme has **NO** time step restriction. (Unconditionally stable); same type of approach as convex-concave splittings.

Difficulty: Both A, B are stiff

Example 2: (Original motivation) Incompressible Navier-Stokes

$$\mathbf{u}_t = \mu\nabla^2\mathbf{u} - \nabla p - (\mathbf{u}\cdot\nabla)\mathbf{u} + \mathbf{f}, \qquad \nabla\cdot\mathbf{u} = 0$$

Boundary conditions: $\mathbf{u} = 0$

To avoid saddle-point problem, split Stokes operator:

[S., Rosales 2011], see also [Henshaw, 1994], [Johnston, Liu 2002, 2004] and [Liu, Liu, Pego, 2010]

$$\mathbf{u}_t = \mu\nabla^2\mathbf{u} - \nabla p(\mathbf{u}) - (\mathbf{u}\cdot\nabla)\mathbf{u} + \mathbf{f}$$

BC: $\mathbf{n}\times\mathbf{u} = 0, \quad \nabla\cdot\mathbf{u} = 0.$

$$\nabla^2 p = \nabla\cdot\left(\mathbf{f} - (\mathbf{u}\cdot\nabla)\mathbf{u}\right)$$

BC: $\mathbf{n}\cdot\nabla p = \mathbf{n}\cdot\left(\mu\nabla^2\mathbf{u} + \mathbf{f} - (\mathbf{u}\cdot\nabla)\mathbf{u}\right) + \lambda\mathbf{n}\cdot\mathbf{u}$

Formulation allows for:

$\mu\nabla^2\mathbf{u}$ (**IM**plicit)

$\nabla p(\mathbf{u})$ (**EX**plicit)

$(\mathbf{u}\cdot\nabla)\mathbf{u}$ (**EX**plicit)

Assumptions and Outline:

Consider multistep **IMEX**: $$u_t = \underbrace{Au}_{\textbf{IM}\text{plicit}} + \underbrace{Bu}_{\textbf{EX}\text{plicit}} + f(t)$$

- Assume that BOTH A and B are (possibly) stiff.
- Assume A symmetric, negative definite. (such as previous examples)

$$A^T = A, \quad \langle x, Ax \rangle < 0 \text{, for all } x \neq 0$$

1) Sufficient conditions for ***Unconditional Stability***.
2) Necessary conditions for ***Unconditional Stability***.
3) Applications: Including the dispersive shallow water equations.

Done by defining an unconditional stability diagram.

ODE Stability:

Numerical stability : $\boldsymbol{u}_t = A\boldsymbol{u}$

Decouple!

Absolute stability region
(Property of time-stepping scheme ONLY)

$$(*) \quad \boldsymbol{u}_t = \lambda \boldsymbol{u}$$

k Time step

$$\mathcal{A} := \{k\lambda \in \mathbb{C} : (*) \text{ bounded}\}$$

Spectrum
(Property of matrix only)

$$\sigma(A) := \{\lambda : A\boldsymbol{u} = \lambda\boldsymbol{u}, \boldsymbol{u} \neq 0\}$$

Stability: **Necessary** and **sufficient** $k\sigma(A) \subseteq \mathcal{A}$

- Allows design of time-stepping for classes of problems (matrices).
- Unconditional stability is easy to analyze ($\mathcal{A}$ must contain a cone).

Difficulties for **IMEX** $\boldsymbol{u}_t = A\boldsymbol{u} + B\boldsymbol{u}$

- Matrices do not commute, necessary and sufficient conditions more difficult.
- No decoupling. Our approach introduces a stability diagram.

Multistep IMEX

Multistep **IMEX** takes form: [Crouzeix, 1980], [Ascher, Ruuth, Wetton, 1995]

$$\frac{1}{k}\sum_{j=0}^{r} a_j \boldsymbol{u}_{n+j} = \sum_{j=0}^{r}\left(c_j A\boldsymbol{u}_{n+j} + b_j B\boldsymbol{u}_{n+j} + b_j \boldsymbol{f}_{n+j}\right)$$

- r is the order, i.e. time stepping error scales $\mathcal{O}(k^r)$
- $b_r = 0$ so that the scheme is explicit in B
- Coefficients are not independent. Satisfy order conditions.

<u>Example</u>: Euler (SBDF1)

$$\frac{1}{k}(\boldsymbol{u}_{n+1} - \boldsymbol{u}_n) = A\boldsymbol{u}_{n+1} + B\boldsymbol{u}_n$$

<u>Example</u>: Semi-implicit backward differentiation 3

$$c_2 = 1,\ c_1 = 0,\ c_0 = 0$$
$$a_2 = 3/2,\ a_1 = -2,\ a_0 = 1/2$$
$$b_2 = 0,\ b_1 = 2,\ b_0 = -1$$

Unconditional stability

Multistep **IMEX** takes form:

$$\frac{1}{k}\sum_{j=0}^{r} a_j \boldsymbol{u}_{n+j} = \sum_{j=0}^{r}\left(c_j A\boldsymbol{u}_{n+j} + b_j B\boldsymbol{u}_{n+j} + b_j \boldsymbol{f}_{n+j}\right)$$

Unconditional stability:

Solutions $\boldsymbol{u}_n$ remain uniformly bounded for all n and $k > 0$.

- Not a trivial property due to the explicit term (demanding a lot!).
- Property depends on **BOTH** coefficients (a_j, b_j, c_j) **AND** (A, B)
- "Easier" to analyze when (A, B) commute (not assumed here)
- For a proposed splitting (A, B) one may to choose (a_j, b_j, c_j)

Unconditional stability diagram

Quick derivation: seek solutions of form $\boldsymbol{u}_n = z^n \boldsymbol{v} \qquad \boldsymbol{v} \in \mathbb{C}^N$

$$\frac{1}{k}\sum_{j=0}^{r} a_j \boldsymbol{u}_{n+j} = \sum_{j=0}^{r}\Big(c_j A \boldsymbol{u}_{n+j} + b_j B \boldsymbol{u}_{n+j}\Big)$$

Nonlinear eigenvalue problem: $\Big(\frac{1}{k}a(z) \;- c(z)\; A - b(z)\; B\Big)\boldsymbol{v} = 0 \quad (*)$

With polynomial coeff.: $a(z) = \sum_{j=0}^{r} a_j z^j \quad b(z) = \sum_{j=0}^{r} b_j z^j \quad c(z) = \sum_{j=0}^{r} c_j z^j$

(not independent – order conditions)

Dot $(*)$ through by $\overline{(-A)^{p-1}\boldsymbol{v}}$ yields

$$a(z) = y\big(c(z) - \mu b(z)\big)$$

where

$$\begin{cases} y = -k\dfrac{\langle \boldsymbol{v}, (-A)^p \boldsymbol{v}\rangle}{\langle \boldsymbol{v}, (-A)^{p-1}\boldsymbol{v}\rangle}; \\[2ex] \mu = \dfrac{\langle \boldsymbol{v}, (-A)^{p-1} B \boldsymbol{v}\rangle}{\langle \boldsymbol{v}, (-A)^{p}\boldsymbol{v}\rangle} \end{cases}$$

(p – a real number, some freedom to choose)

Unconditional stability diagram

Then, if knew eigenvector $\boldsymbol{v} \in \mathbb{C}^N$ for a fixed $k > 0$

Could compute $$y = -k\frac{\langle \boldsymbol{v}, (-A)^p \boldsymbol{v}\rangle}{\langle \boldsymbol{v}, (-A)^{p-1} \boldsymbol{v}\rangle}, \qquad \mu = \frac{\langle \boldsymbol{v}, (-A)^{p-1} B\boldsymbol{v}\rangle}{\langle \boldsymbol{v}, (-A)^p \boldsymbol{v}\rangle}$$

Then if all solutions to $$a(z) = y\big(c(z) - \mu b(z)\big) \qquad (**)$$

Have $|z| < 1$ then sufficient for stability.

Don't know $\boldsymbol{v} \in \mathbb{C}^N$. Instead define **unconditional stability** region

$$\left.\begin{aligned} \mathcal{D} &:= \{\mu \in \mathbb{C} : (**) \text{ stable } \forall y < 0\} \\ &= \{\mu \in \mathbb{C} : \mu b(z) = c(z) \text{ stable}\} \end{aligned}\right\}$$

Worst case when $y \to -\infty$. Big simplification.

Some similarity to diagrams defined in [Frank, Hundsdorfer, Verwer, 1997], [Koto, 2009]

Shape of Unconditional Stability Diagram:

Theorem 1 *(SBDF) The set $\mathcal{D}$ is simply connected, contains the origin $0 \in \mathcal{D}$, and has a boundary parameterized by the curve*

$$\partial\mathcal{D} = \left\{ \frac{z^r}{z^r - (z-1)^r} : |z| = 1, \ \arg z_0 \le \arg z \le \ 2\pi - \arg z_0 \right\},$$

where: $z_0 = 1,$ *for order* $r = 1$*, and*

$$z_0 = \frac{1}{1 - 2\cos(\pi/r)e^{\imath\pi/r}}, \quad \textit{for orders } 2 \le r \le 5.$$

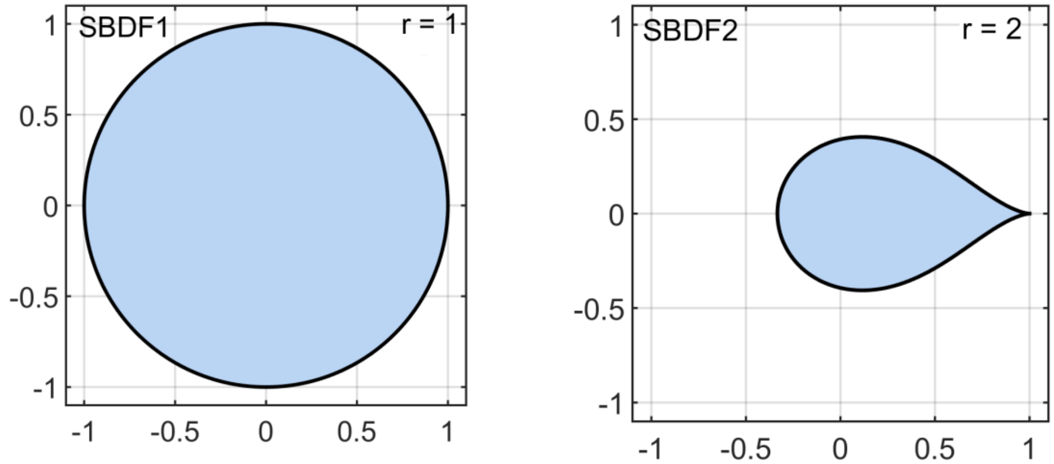
The right-most m_r and left-most m_l points of $\partial\mathcal{D}$ are on the real axis where:

$$\begin{array}{llll} \textit{for } r = 1, & m_l = -1 & \textit{and} & m_r = 1, \\ \textit{for } 2 \le r \le 5, & m_l = -(2^r - 1)^{-1} & \textit{and} & m_r = \big(1 + 2^r \cos^r(\pi/r)\big)^{-1}. \end{array}$$

Proof 1 *$\mathcal{D} = \varphi^{-1}(\mathcal{T})$ is the image of a triangle $\mathcal{T}$ under the conformal mapping $\varphi(z) = (\frac{z}{z-1})^{1/r}$ and a correctly chosen branch cut.*

U. Stability: Nec. & suff. Conditions

Right: $\mathcal{D}$ for two popular **IMEX** schemes



Sufficient condition:

Let $W_p := \left\{ \langle \boldsymbol{v}, (-A)^{p-1} B \boldsymbol{v} \rangle : \langle \boldsymbol{v}, (-A)^{p} \boldsymbol{v} \rangle = 1 \right\}$ all allowable μ by A, B.

If $W_p \subseteq \mathcal{D}$ then **IMEX** scheme is unconditionally stable.

Necessary condition:

$\sigma((-A)^{-1}B) = \{\mu \in \mathbb{C} : \mu(-A)\boldsymbol{u} = B\boldsymbol{u}, \boldsymbol{u} \neq \boldsymbol{0}\}$ Generalized eigenvalues

$\sigma((-A)^{-1}B) \subseteq \overline{\mathcal{D}} \cup \{1\}$ Is necessary for unconditional stability.

Remarks:

1) The set $W_p := \left\{ \langle \boldsymbol{v}, (-A)^{p-1} B \boldsymbol{v} \rangle : \langle \boldsymbol{v}, (-A)^p \boldsymbol{v} \rangle = 1 \right\}$

For computations can be written as a *numerical range (chebfun):*

$W(X) := \{ \langle \boldsymbol{x}, X \boldsymbol{x} \rangle : \|\boldsymbol{x}\| = 1, \boldsymbol{x} \in \mathbb{C}^n \}$ where $X = (-A)^{\frac{p}{2}-1} B \, (-A)^{-\frac{p}{2}}$

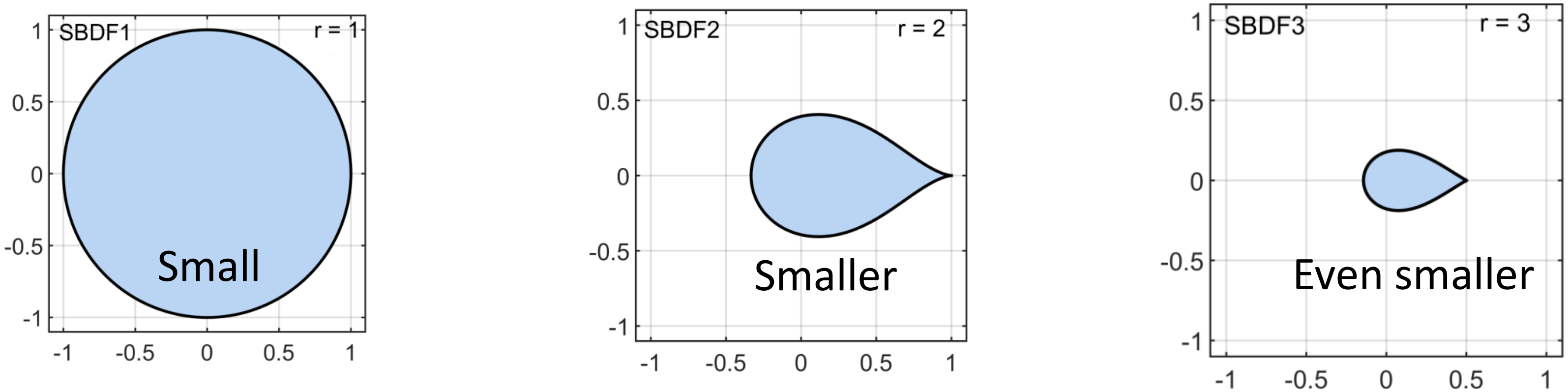
2) If matrices are normal and commute, then W_p is the convex hull of $\sigma((-A)^{-1} B)$, i.e. necessary and sufficient are *almost* the same.

3) The sufficient condition using $\mathcal{D}$ is weaker than other unconditional stability criteria, i.e. [Akrivis et. al, 1998, 1999, 2003].

4) Choose time-stepping coefficients for a fixed set of matrices.

5) Everything is "easily" computable.

Time-stepping coefficients with big $\mathcal{D}$



Small regions limit the unconditionally stable matrix splittings (A, B)

We would like to have **BIG** regions (if possible).

How? General idea for new coefficients:

1. Use $\mathcal{D} = \{\mu \in \mathbb{C} : \mu b(z) = c(z) \text{ stable}\}$
2. Implies want $b(z)$ small when z on (or in) unit circle.
3. Using order conditions, implies need roots of $c(z)$ *close* to 1.

New IMEX coefficients

- One parameter family, $0 < \delta \leq 1$, reduce to SBDF $\delta = 1$.
- Orders $1 \leq r \leq 5$
- Defined by polynomial coefficients.

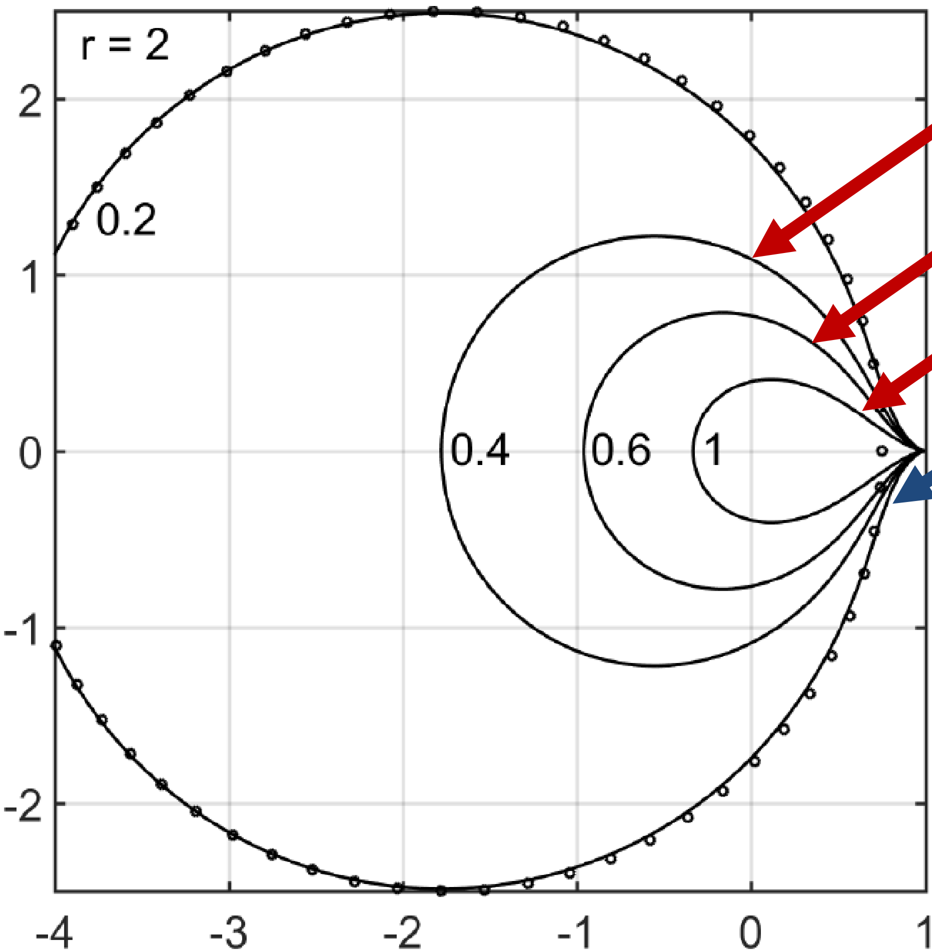
$$\begin{aligned}
&(Implicit\ coefficients) && c(z) = (z - 1 + \delta)^r \\
&(Explicit\ coefficients) && b(z) = (z - 1 + \delta)^r - (z-1)^r \\
&(Derivative\ coefficients) && a(z) = \sum_{j=1}^{r} \frac{f^{(j)}(1)}{j!}(z-1)^j
\end{aligned}$$

where $f(z) = (\ln z)(z - 1 + \delta)^r$

(Coefficients are zero stable)

For order r = 2, some similarity to [Akrivis, Karakatsani, 2003]

$\mathcal{D}$ For New IMEX coefficients



Bigger

Big

SBDF2

Huge (asymptotic circle)

Circle:

$$\left| z + \frac{1}{r\delta} - \frac{r+1}{2r} \right| \leq \frac{1}{r\delta}$$

Centre $\sim -(r\delta)^{-1}$

Radius $\sim (r\delta)^{-1}$

Regions can be made arbitrarily large (Real part < 1)

Part II
Idealized Examples

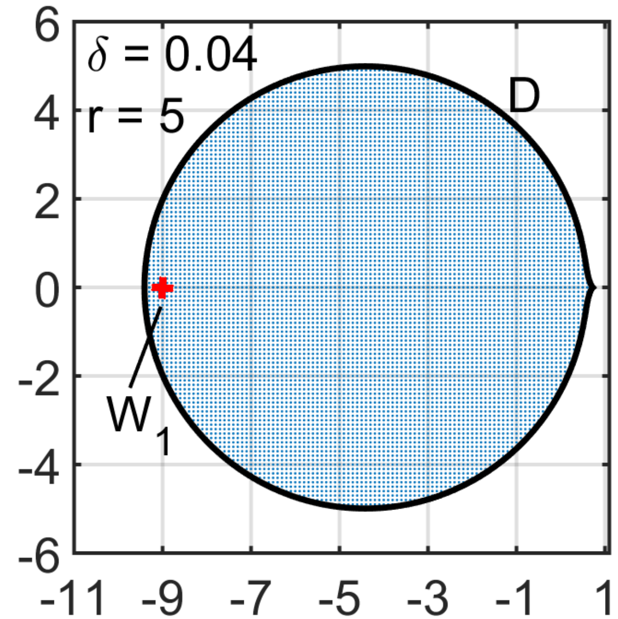
Implications/Examples

Consider: $$u_t = -10u = -u - 9u$$

Where: $$Au = -u \quad Bu = -9u$$

Numerical range: $$W_p := \left\{ \langle v, (-A)^{p-1} B v \rangle : \langle v, (-A)^p v \rangle = 1 \right\}$$

$$W_1 = \{-9\}$$



Choice of δ :

$$\delta < 2(1 - (0.9)^{1/r})$$

For order, $r = 5$ use any value

$\delta \leq 0.04$ guarantees $W_p \subseteq \mathcal{D}$

Get unconditional stability.

Example: New IMEX, variable diffusion

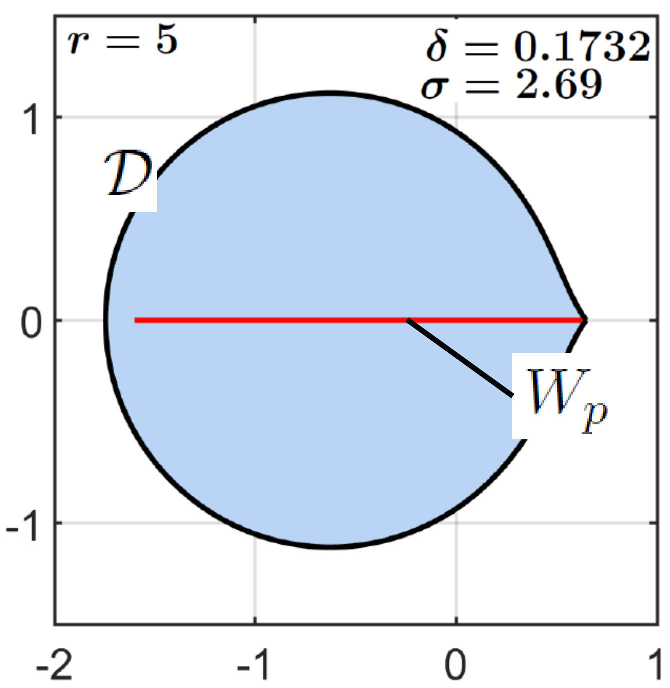
$$u_t = \big(d(x)u_x\big)_x + f(x,t) \qquad d(x) = 4 + 3\cos(2\pi x).$$

$$\boldsymbol{A}_h\boldsymbol{u} \approx \sigma u_{xx} \qquad \boldsymbol{B}_h\boldsymbol{u} \approx \big((d(x)-\sigma)u_x\big)_x$$

Spectral discretization in space N = 64 modes.

Satisfies the sufficient condition for unconditional stability.

$$W_p \subseteq \mathcal{D}$$



Variable diffusion: SBDF Limited 2nd Order

Given $$u_t = \big(d(x)u_x\big)_x + f(x,t)$$

Splitting $$\boldsymbol{A}_h\boldsymbol{u} \approx \sigma u_{xx} \qquad \boldsymbol{B}_h\boldsymbol{u} \approx \big((d(x)-\sigma)u_x\big)_x$$

Either spectral or finite difference spatial discretization.

Theorem 2 *(Limitations on SBDF) Fix $d(x) > 0$. First and second order: there always exists σ that guarantees unconditional stability. Third order (or higher): SBDF is* **NOT** *in general stable for any σ. However, a modified scheme (δ, σ) is* **ALWAYS** *unconditionally stable.*

New IMEX coefficients, 3-5th order schemes unconditionally stable:

$$\sigma = d_{min} \qquad \delta < 2\Big(1 - \big(1 - \frac{d_{min}}{d_{max}}\big)^{\frac{1}{r}}\Big), \quad \textit{for } 1 \le r \le 5.$$

$$d_{min} = \min_{x\in\Omega} d(x), \quad d_{max} = \max_{x\in\Omega} d(x).$$

Example: New IMEX, Variable diffusion

Num. Steps	k	Error $r=1$	Rate	Error $r=2$	Rate	Error $r=3$	Rate	Error $r=4$	Rate	Error $r=5$	Rate
5	1	7.9e+00	-	4.7e+01	-	3.4e+02	-	9.0e+02	-	8.8e+02	-
10	2^{-1}	3.4e+00	1.2	6.7e+01	-0.5	4.9e+02	-0.5	2.2e+03	-1.3	4.3e+03	-2.3
20	2^{-2}	4.3e+00	-0.4	2.4e+01	1.5	5.6e+02	-0.2	3.7e+03	-0.7	6.3e+03	-0.5
40	2^{-3}	1.3e+00	1.7	3.5e+01	-0.5	5.4e+02	0.1	6.3e+03	-0.8	5.8e+04	-3.2
80	2^{-4}	6.9e-01	1.0	7.1e+00	2.3	1.3e+01	5.4	7.4e+02	3.1	6.0e+03	3.3
160	2^{-5}	2.7e-01	1.4	1.0e+00	2.8	1.1e+01	0.2	5.3e+01	3.8	5.7e+01	6.7
320	2^{-6}	2.2e-01	0.3	6.0e-01	0.8	2.5e+00	2.2	2.8e+00	4.2	7.1e+00	3.0
640	2^{-7}	2.9e-01	-0.4	5.3e-01	0.2	6.3e-01	2.0	1.5e-01	4.3	4.0e-01	4.1
1280	2^{-8}	2.5e-01	0.2	2.2e-01	1.3	5.0e-02	3.7	3.6e-02	2.1	2.5e-02	4.0
2560	2^{-9}	1.6e-01	0.6	5.6e-02	2.0	4.9e-03	3.4	3.5e-03	3.4	2.8e-04	6.4
5120	2^{-10}	9.1e-02	0.8	1.2e-02	2.2	8.5e-04	2.5	2.0e-04	4.1	1.0e-05	4.8
1.0e+04	2^{-11}	4.8e-02	0.9	2.8e-03	2.1	1.3e-04	2.7	1.1e-05	4.2	3.8e-07	4.7
2.0e+04	2^{-12}	2.5e-02	1.0	6.7e-04	2.1	1.8e-05	2.9	6.1e-07	4.2	1.3e-08	4.9
4.1e+04	2^{-13}	1.2e-02	1.0	1.6e-04	2.0	2.4e-06	2.9	3.6e-08	4.1	1.1e-09	3.5
8.2e+04	2^{-14}	6.3e-03	1.0	4.0e-05	2.0	3.0e-07	3.0	2.2e-09	4.0	1.4e-09	-
1.6e+05	2^{-15}	3.1e-03	1.0	9.8e-06	2.0	3.8e-08	3.0	2.3e-10	3.3	2.8e-09	-

BLUE REGION NOT Possible with SBDF! Regardless of how one chooses σ.

CFL = $k \leq 2^{-18}$ (Explicit schemes)

Exact solution: $u^* = \sin(20t)e^{\sin(2\pi x)}$ $t_f = 5$

Nonlinear problem

Nonlinear Diffusion Example

$$\begin{cases} \rho_t + \nabla \cdot (\boldsymbol{V} \rho) = 0 & \text{(Conservation of mass)} \\ \boldsymbol{V} = -\dfrac{\tilde{\kappa}}{\tilde{\mu}} \nabla p & \text{(Darcy's law)} \\ p = p_0 \rho^\gamma & \text{(Eqn. of state)} \end{cases}$$

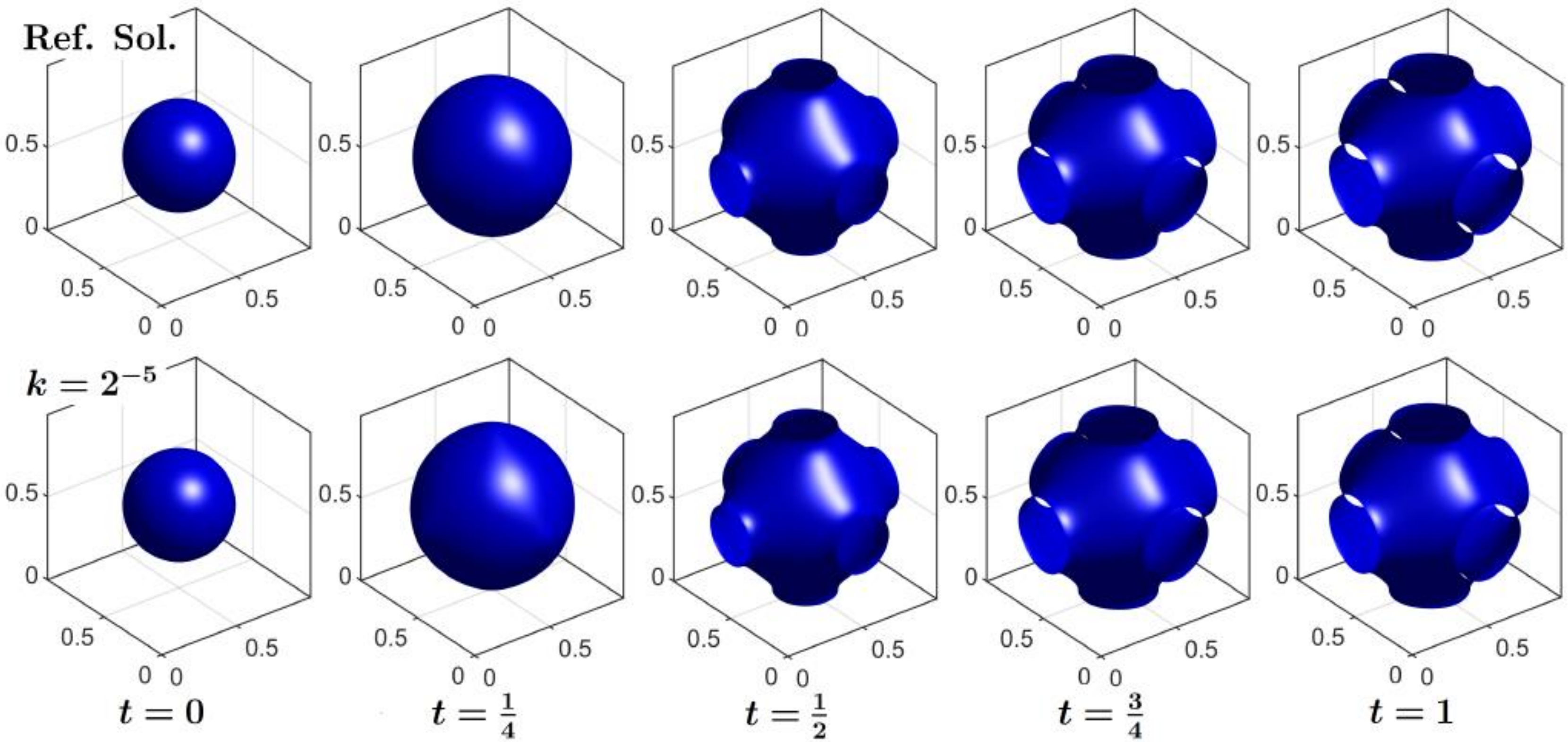
Combine:
$$\rho_t = \nabla \cdot \Big(\rho^\gamma \nabla \rho \Big)$$

Splitting:
$$\boldsymbol{A}_h \boldsymbol{u} \approx \sigma u_{xx} \qquad \mathbf{B}\boldsymbol{u} \approx \nabla \cdot \Big(\big(\rho^\gamma(\boldsymbol{x}) - \sigma \big) \nabla \boldsymbol{u} \Big)$$

Use new formulas for parameters.

Avoids nonlinear implicit terms – with a constant in time linear implicit term!
- Similarity in flavor to Rosenbrock methods (but no Jacobian here)
- Avoiding implicit nonlinear terms also see: [Duchemin, Eggers, 2014], [Bruno, Cubillos, 2016] and [Bruno, Cubillos, 2017] on quasi-unconditional stability, for compressible Euler.

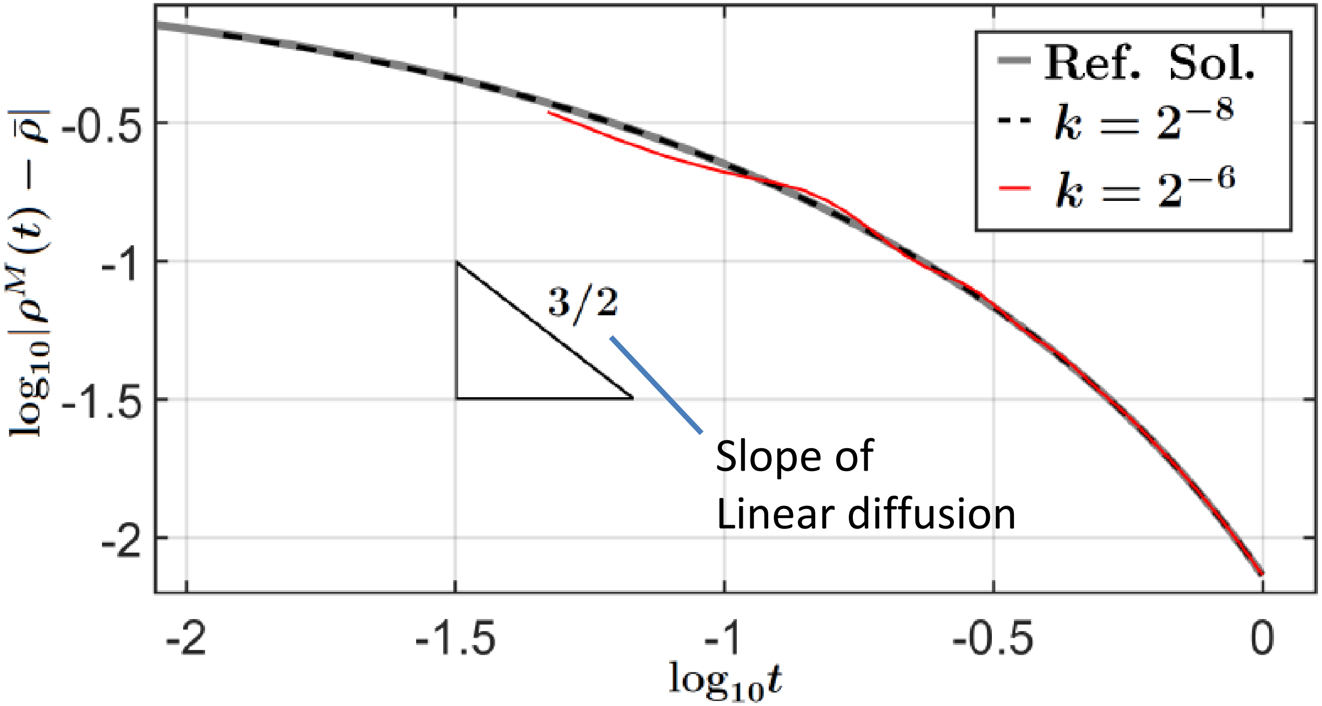
Nonlinear diffusion in a periodic domain



Visual inspection
(N = 128^3 Fourier modes)

Top: Ref. Sol. BDF2 65,000 time steps (run overnight)
Bottom: 32 time steps (a few seconds)

Anomalous diffusion: decay of peak



Gray: Ref. Sol. BDF2 65,000 time steps (run overnight).
Dashed: 256 time steps (< minute), several digits of accuracy.
Red: 64 time steps (seconds), a few digits of accuracy.

Nonlinear diffusion: convergence test

Manufactured Solution Test

Num. Steps	k	Error $r=1$	Rate	Error $r=2$	Rate	Error $r=3$	Rate	Error $r=4$	Rate	Error $r=5$	Rate
8	2^{-3}	1.0e+00	-	8.3e-01	-	6.4e-02	-	9.7e-02	-	3.4e-04	-
16	2^{-4}	7.7e-01	0.4	3.8e-01	1.1	3.4e-02	0.9	2.2e-02	2.1	8.1e-04	-1.2
32	2^{-5}	5.0e-01	0.6	8.3e-02	2.2	8.6e-03	2.0	1.9e-03	3.6	1.2e-04	2.7
64	2^{-6}	2.6e-01	0.9	1.5e-02	2.4	1.4e-03	2.6	1.2e-04	4.0	7.6e-06	4.0
128	2^{-7}	1.3e-01	1.0	3.6e-03	2.1	1.9e-04	2.8	6.6e-06	4.2	3.0e-07	4.7
256	2^{-8}	6.4e-02	1.0	8.6e-04	2.1	2.5e-05	2.9	3.8e-07	4.1	1.3e-08	4.5
512	2^{-9}	3.2e-02	1.0	2.1e-04	2.0	3.2e-06	3.0	2.2e-08	4.1	6.2e-09	-
1024	2^{-10}	1.6e-02	1.0	5.2e-05	2.0	4.0e-07	3.0	9.8e-10	4.5	1.2e-08	-
2048	2^{-11}	7.8e-03	1.0	1.3e-05	2.0	5.0e-08	3.0	6.9e-10	-	2.4e-08	-

CFL = $k \leq 2^{-18}$
(Explicit schemes)

k	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}	2^{-11}	2^{-12}
$r=1,\ R_k$	1.22	1.58	1.52	1.28	1.14	1.07	1.03	1.02	1.01
$r=2,\ R_k$	0.39	5.12	2.06	2.02	1.96	1.95	1.96	1.97	1.98
$r=3,\ R_k$	1.38	3.54	1.63	8.52	2.83	2.74	2.81	2.88	2.93

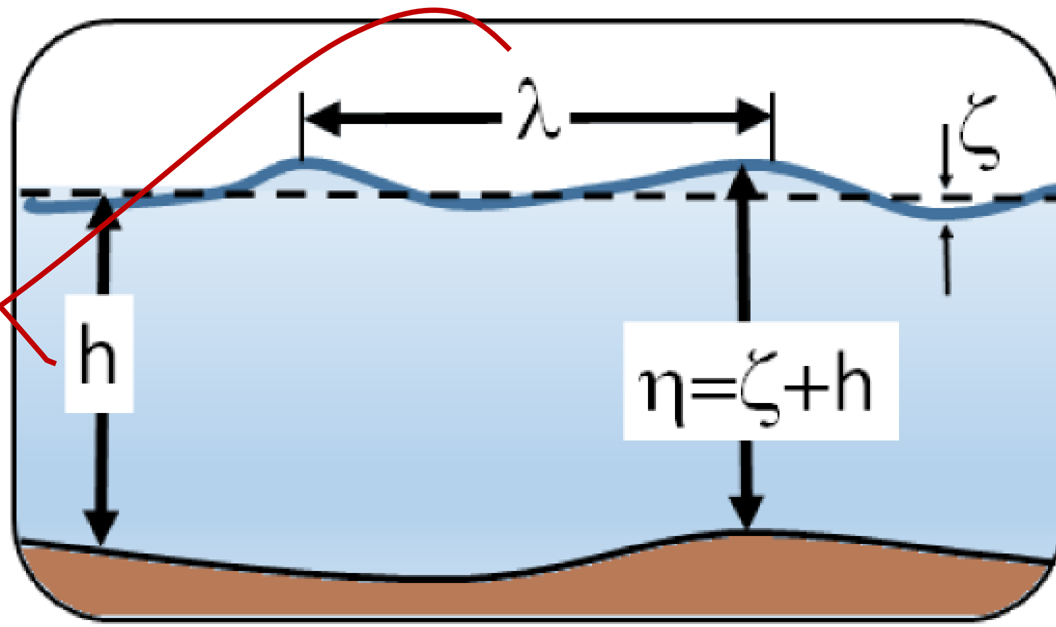
$$R_k := \log_2\left(\|\rho_{4k}(t_f) - \rho_{2k}(t_f)\|_{\infty,h} / \|\rho_{2k}(t_f) - \rho_k(t_f)\|_{\infty,h}\right)$$

Part III
Dispersive Shallow Water Equations

Shallow water equations

Shallow water equation regime of validity:

$$\beta \equiv \frac{\bar{h}}{\lambda} \ll 1$$



Expand Euler equations in (formal) powers of β, introduce:

$h(\boldsymbol{x})$ – bathymetry

$\zeta(\boldsymbol{x}, t)$ – free surface height

$\boldsymbol{u}(\boldsymbol{x}, t)$ – depth avg. velocity (2d)

$\eta(\boldsymbol{x}, t) = h + \zeta$ – water depth (> 0)

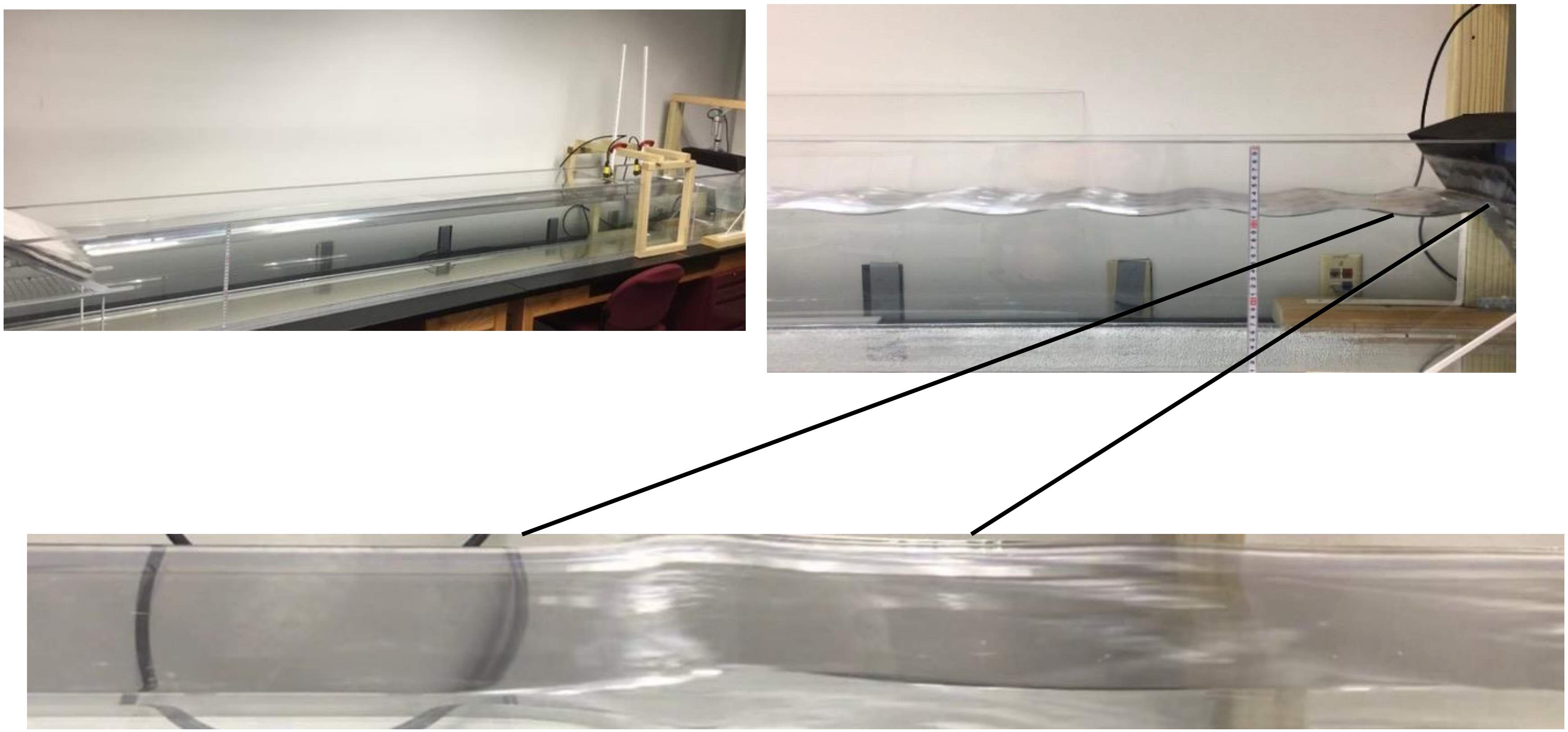
$$\eta_t + \nabla \cdot (\eta \boldsymbol{u}) = 0$$

$$\underbrace{\boldsymbol{u}_t + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + \nabla \zeta}_{\text{SWE terms}} = \boxed{O(\beta^2)}$$

Workhorse for ocean modeling.

Higher corrections = more physics.

Wave tank experiments



Courtesy of Wooyoung Choi's Lab – NJIT,
Department of mathematical science

Dispersive shallow water equations (DSWE)

See: [Rayleigh 1876]; [Serre 1953];
[Su, Gardner 1969]; [Green, Naghdi 1976]

Shallow water equation regime of validity: $\beta \equiv \frac{\bar{h}}{\lambda} \ll 1$

$$\eta_t + \nabla \cdot (\eta \boldsymbol{u}) = 0$$

$$\underbrace{\boldsymbol{u}_t + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + \nabla \zeta}_{\text{SWE terms}} = \underbrace{\frac{1}{\eta} \nabla \left\{ \frac{1}{3} \eta^3 \left(\nabla \cdot \boldsymbol{u}_t + \boldsymbol{u} \cdot \nabla (\nabla \cdot \boldsymbol{u}) - (\nabla \cdot \boldsymbol{u})^2 \right) + \frac{1}{2} \eta^2 \left(\boldsymbol{u}_t \cdot \nabla h + (\boldsymbol{u} \cdot \nabla)^2 h \right) \right\} - \left\{ \frac{1}{2} \eta \left(\nabla \cdot \boldsymbol{u}_t + \boldsymbol{u} \cdot \nabla (\nabla \cdot \boldsymbol{u}) - (\nabla \cdot \boldsymbol{u})^2 \right) + \left(\boldsymbol{u}_t \cdot \nabla h + (\boldsymbol{u} \cdot \nabla)^2 h \right) \right\} \nabla h}_{O(\beta^2)}$$

$O(\beta^2)$ (truncated to order 4)

Additional physics:

- Shallow water equations (SWE) hyperbolic – contain shocks.
- DSWE – add dispersion effects to SWE (which can regularize shock) can create additional oscillations, e.g., more waves; valid for O(1) nonlinearities (not just weakly nonlinear).

Linearization (about constant state):

$$\boldsymbol{u}(\boldsymbol{x}, t) = \hat{\boldsymbol{u}} e^{\imath\omega t - \imath \boldsymbol{k} \cdot \boldsymbol{x}} \qquad \eta(\boldsymbol{x}, t) = h_0 + \hat{\zeta} e^{\imath\omega t - \imath \boldsymbol{k} \cdot \boldsymbol{x}}$$

Dispersion relation $\Rightarrow$ $\omega = \dfrac{|\boldsymbol{k}|}{\sqrt{1 + |\boldsymbol{k}|^2}}$

Dispersive shallow water equations (DSWE)

See: [Rayleigh 1876]; [Serre 1953]; [Su, Gardner 1969]; [Green, Naghdi 1976]

Shallow water equation regime of validity: $\beta \equiv \frac{\bar{h}}{\lambda} \ll 1$

$$\eta_t + \nabla \cdot (\eta \boldsymbol{u}) = 0$$

$$\underbrace{\boldsymbol{u}_t + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + \nabla \zeta}_{\text{SWE terms}} = \underbrace{\boxed{\frac{1}{\eta} \nabla \left\{ \frac{1}{3} \eta^3 \left(\nabla \cdot \boldsymbol{u}_t \right.\right.} + \boxed{\boldsymbol{u} \cdot \nabla (\nabla \cdot \boldsymbol{u})} - (\nabla \cdot \boldsymbol{u})^2 \Big) \boxed{+ \frac{1}{2} \eta^2 \left(\boldsymbol{u}_t \cdot \nabla h \right.} + (\boldsymbol{u} \cdot \nabla)^2 h \Big) \Big\} \boxed{- \left\{ \frac{1}{2} \eta \left(\nabla \cdot \boldsymbol{u}_t \right.\right.} + \boldsymbol{u} \cdot \nabla (\nabla \cdot \boldsymbol{u}) - (\nabla \cdot \boldsymbol{u})^2 \Big) + \boxed{\left(\boldsymbol{u}_t \cdot \nabla h \right.} + (\boldsymbol{u} \cdot \nabla)^2 h \Big) \Big\} \nabla h}_{O(\beta^2)}$$

$O(\beta^2)$ (truncated to order 4)

Numerics

<u>Difficulty 1:</u>

Nonlinear, and **``Mixed'' space-time** derivative terms, e.g., $\sim \eta^{-1} \nabla (\eta^3 \nabla \cdot \boldsymbol{u}_t)$

<u>Difficulty 2?</u>

RHS has term $\eta^{-1} \nabla (\boldsymbol{u} \cdot \nabla (\nabla \cdot \boldsymbol{u}))$, explicit treatment could be very stiff?

DSWE: Structure of the time derivative

$$\eta_t + \nabla\cdot(\eta \boldsymbol{u}) = 0,$$

$$\mathcal{G}\boldsymbol{u}_t + \eta\boldsymbol{u}\cdot\nabla\boldsymbol{u} + \eta\nabla\zeta = \nabla\left\{\frac{1}{3}\eta^3\left(\boldsymbol{u}\cdot\nabla(\nabla\cdot\boldsymbol{u}) - (\nabla\cdot\boldsymbol{u})^2\right) + \frac{1}{2}\eta^2(\boldsymbol{u}\cdot\nabla)^2 h\right\}$$
$$- \eta\left\{\frac{1}{2}\eta\left(\boldsymbol{u}\cdot\nabla(\nabla\cdot\boldsymbol{u}) - (\nabla\cdot\boldsymbol{u})^2\right) + (\boldsymbol{u}\cdot\nabla)^2 h\right\}\nabla h,$$

Bring all time derivatives to left hand side, where

$$\mathcal{G} \equiv \mathcal{N}_\eta + \mathcal{K}_{\eta,h}: \begin{cases} \mathcal{N}_\eta \equiv \eta\boldsymbol{I} - \nabla\frac{1}{3}\eta^3\nabla\cdot, & \text{(dispersion due to the surface)} \\ \mathcal{K}_{\eta,h} \equiv -\nabla\left(\frac{1}{2}\eta^2\nabla h^T\cdot\right) + \frac{1}{2}\eta^2\nabla h(\nabla\cdot\,) + \eta\nabla h(\nabla h^T\cdot\,) & \text{(bathymetry)} \end{cases}$$

Nonlinear and time-dependent operator.

Goal – adopt matrix-free methods where we avoid constructing $\mathcal{G}$.
e.g., address "Difficulty" #1 (in previous slide)

Numerical work for SWE with dispersion:
[Li, Choi, Hyman 2004]; [Choi, Goullet, Jo 2011]; [Khakimzyanov Dutykh Fedotova Mitsotakis, 2020] (effectively 1d); [Patel Kumar Rajni 2020].

DSWE Constraint form

Evolution. $\eta_t + \nabla \cdot (\eta \boldsymbol{u}) = 0,$

$$\boldsymbol{U}_t + \eta \nabla \zeta + \nabla(\eta \boldsymbol{u}\boldsymbol{u}) = \boldsymbol{F}(\eta, \boldsymbol{u}, h).$$

Constraint. $\boldsymbol{U} - \mathcal{G}\boldsymbol{u} = 0,$

$$\mathcal{G} \equiv \mathcal{N}_\eta + \mathcal{K}_{\eta,h}.$$

[Ignore this (for now)]

$$\begin{aligned}\boldsymbol{F}(\eta, \boldsymbol{u}, h) \equiv\ & \nabla \Big\{ \frac{1}{3}\eta^3 \left(\boldsymbol{u} \cdot \nabla (\nabla \cdot \boldsymbol{u}) - (\nabla \cdot \boldsymbol{u})^2 \right) + \\ & \frac{1}{2}\eta^2 (\boldsymbol{u} \cdot \nabla)^2 h + \eta^2 \nabla \cdot (\eta \boldsymbol{u}) \nabla \cdot \boldsymbol{u} + \eta \nabla \cdot (\eta \boldsymbol{u}) \nabla h \cdot \boldsymbol{u} \Big\} \\ & - \nabla h \Big\{ \frac{1}{2}\eta^2 \left(\boldsymbol{u} \cdot \nabla (\nabla \cdot \boldsymbol{u}) - (\nabla \cdot \boldsymbol{u})^2 \right) + \\ & \eta (\boldsymbol{u} \cdot \nabla)^2 h + \eta \nabla \cdot (\eta \boldsymbol{u}) (\nabla \cdot \boldsymbol{u}) + \nabla \cdot (\eta \boldsymbol{u}) (\nabla h \cdot \boldsymbol{u}) \Big\}\end{aligned}$$

Consider two time stepping schemes. Although unrelated, their efficacy will be related:

1) ``standard'' conceptual approach:

$$\eta_t + \nabla \cdot (\eta \boldsymbol{u}) = 0,$$
$$\boldsymbol{U}_t + \eta \nabla \zeta + \nabla(\eta \boldsymbol{u}\boldsymbol{u}) = \boldsymbol{F}(\eta, \boldsymbol{u}, h).$$

Explicit time-stepping (e.g., Runge-Kutta)

$$\boldsymbol{U} - \mathcal{G}\boldsymbol{u} = 0,$$

Fully implicit solve via matrix-free method (e.g., PCG).

Key challenge: Find a preconditioner

2) Implicit-explicit (ImEx) multistep methods – avoid a fully implicit treatment of G.

$$\boldsymbol{w}_t = \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{u}, t)$$
$$\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{u}) = 0.$$

$$\boldsymbol{w} = (\eta, \boldsymbol{U})^T$$
$$\boldsymbol{f} = (-\nabla \cdot (\eta \boldsymbol{u}),\ -\eta \nabla \zeta - \nabla(\eta \boldsymbol{u}\breve{\boldsymbol{u}}) + \boldsymbol{F})^T$$
$$\boldsymbol{g} = \boldsymbol{U} - \mathcal{N}_\eta \boldsymbol{u} - \mathcal{K}_{\eta,h} \boldsymbol{u},$$

Spatial discretizations (MOL) → Differential algebraic equation (index-1).

Key challenge: Zero stability (linear implicit)

ImEx linear multistep for index-1 DAEs

Theoretical tool to formulate schemes for DAEs (e.g., [Hairer Wanner, Vol. II], see also [Constantinescu, Sandu 2010]; ImEx LMMs [Crouzeix, 1980], [Ascher, Ruuth, Wetton, 1995]):

$$w_t = f(w,u,t), \qquad \epsilon u_t = U - \mathcal{G}u$$

Discretize, then take: $\epsilon \to 0$.

However, in our case G is stiff (and non-linear). We differ in that we look at ImEx schemes applied to the constraint equation:

$$w_t = f\overbrace{(w,u,t)}^{\textbf{EX}\text{plicit}},$$
$$\epsilon u_t = \underbrace{-\mathcal{A}u}_{\textbf{IM}\text{plicit}} \underbrace{- \mathcal{B}u}_{\textbf{EX}\text{plicit}} + U$$

- Where $\mathcal{G} = \mathcal{A} + \mathcal{B}$, Choice is NOT unique!
- Convention: $\mathcal{A}$ stiff, $\mathcal{B}$ non-stiff. Generally not conventional to apply ImEx to the constraint.

What we want is to take $\mathcal{A}$ <u>linear</u> and <u>constant coefficient</u>

ImEx linear multistep for index-1 DAEs

Linear multistep ImEx scheme (after taking $\epsilon \to 0$)

$$\text{(ImEX DAE)} \quad \begin{aligned} \frac{1}{\Delta t}\sum_{j=0}^{s} a_j \boldsymbol{w}^{n+j} &= \sum_{j=0}^{s} b_j \boldsymbol{f}(\boldsymbol{w}^{n+j}, \boldsymbol{u}^{n+j}, t^{n+j}), \\ 0 &= \sum_{j=0}^{s} c_j (\boldsymbol{U}^{n+j} - \boldsymbol{A}\boldsymbol{u}^{n+j}) - b_j \boldsymbol{B}\boldsymbol{u}^{n+j}. \end{aligned}$$

Where $\mathcal{G} = \mathcal{A} + \mathcal{B}$,

So what can go wrong?

Answer: Zero-stability, e.g., stability with $\Delta t = 0$

$$0 = \sum_{j=0}^{s} c_j (\boldsymbol{U}^{n+j} - \boldsymbol{A}\boldsymbol{u}^{n+j}) - b_j \boldsymbol{B}\boldsymbol{u}^{n+j}.$$

Require: solutions $\boldsymbol{u}_n$ uniformly bounded for all n.

- Not a trivial property due to the explicit stiff terms
- Property depends on **BOTH** coefficients **AND** splitting.

Example: Semi-implicit backward differentiation 3

$$c_2 = 1,\ c_1 = 0,\ c_0 = 0$$

$$a_2 = 3/2,\ a_1 = -2,\ a_0 = 1/2$$

$$b_2 = 0,\ b_1 = 2,\ b_0 = -1$$

Zero-stability criteria

Simplified case: Assume $\mathcal{G}$ is time-independent (but can vary in space)

Substituting: $\mathcal{G}\boldsymbol{u} = \lambda \mathcal{A} \boldsymbol{u}$

and $\boldsymbol{u}^n = z^n \boldsymbol{v}$

Into: (*) $\sum_{j=0}^{r} c_j \boldsymbol{A} \boldsymbol{u}^{n+j} + b_j(\boldsymbol{G} - \boldsymbol{A})\boldsymbol{u}^{n+j} = 0$

Yields the following zero-stability criteria

Define: $\mathcal{D} := \{\mu \in \mathbb{C} : c(z) + (\mu - 1)b(z) \ \textit{has stable roots}\}$

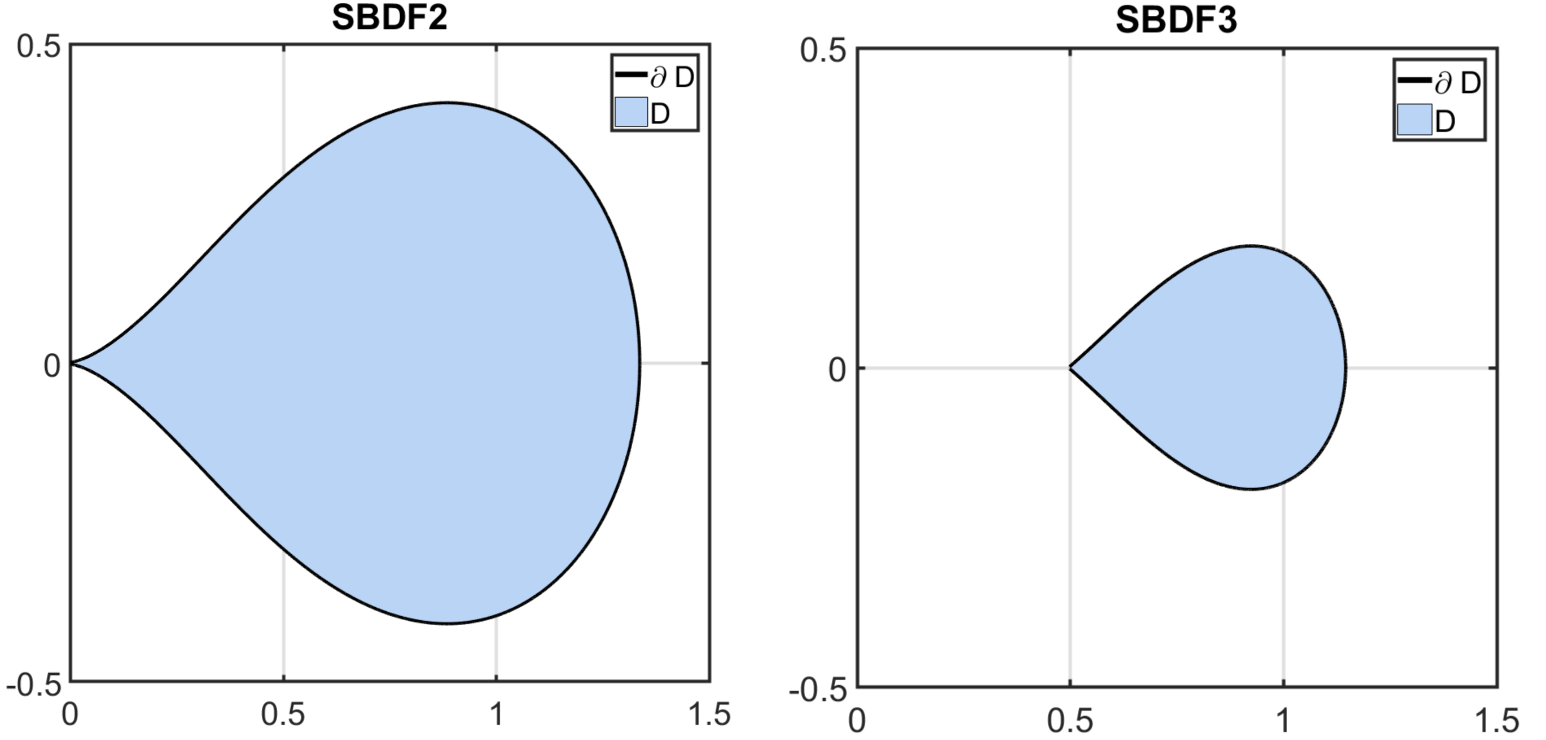
$$c(z) := z^r + c_{r-1} z^{r-1} + \ldots c_0$$

$$b(z) := b_{r-1} z^{r-1} + \ldots b_0.$$

Then (*) yields stable dynamics if: $\operatorname{eig}\left(\mathcal{A}^{-1}\mathcal{G}\right) \in \mathcal{D}$

Zero-stability criteria

Need generalized eigenvalues $\mathcal{G}u = \lambda \mathcal{A}u$ ➡ $\mathrm{eig}\left(\mathcal{A}^{-1}\mathcal{G}\right) \in \mathcal{D}$



You need the eigenvalues clustered near 1 → Just like preconditioning
Caveat: Need bounds (bounds become fundamentally worse/impossible at 3rd order without simultaneously choosing time-stepping schemes!)

Key Goal: Find $\mathcal{A}$

For both:
1) Runge-Kutta with preconditioner.
2) IMEX time-stepping for DAE

Need to find $\mathcal{A}$

- Easy to ``invert'';
- Ensure bound/clustering of generalized eigenvalues

$$\mathcal{G}u = \lambda \mathcal{A} u$$

Caveat: IMEX DAEs more stringent than RK with preconditioner (eig. val. must lie in region around 1).

Strategy: Calculus of variations approach to study and bound generalized eigenvalues.

Mathematical preliminaries

Algebraic constraint is: $\boldsymbol{U} - \mathcal{G}\boldsymbol{u} = 0,$ Goal: See an $\mathcal{A} = \sigma \boldsymbol{I} - \alpha \nabla \nabla \cdot$

Where: $\mathcal{G} \equiv \mathcal{N}_\eta + K_{\eta,h}$; consists of two contributions:

1) $\mathcal{N}_\eta \equiv \eta \boldsymbol{I} - \nabla \frac{1}{3} \eta^3 \nabla \cdot$ (depends only on water depth)

2) $\mathcal{K}_{\eta,h} \equiv -\nabla \left(\frac{1}{2} \eta^2 \nabla h^T \cdot \right) + \frac{1}{2} \eta^2 \nabla h (\nabla \cdot\ ) + \eta \nabla h (\nabla h^T \cdot\ )$ (bathymetry)

Operators have variational forms defined on the Hilbert space:

$$\|\boldsymbol{u}\|^2_{\mathrm{div}} := \int_\Omega \|\boldsymbol{u}\|^2 + (\nabla \cdot \boldsymbol{u})^2 \, \mathrm{d}\boldsymbol{x}, \quad H_{\mathrm{div}}(\Omega) := \{\boldsymbol{u} \mid \|\boldsymbol{u}\|_{\mathrm{div}} < \infty\}$$

(Include boundary conditions $(\boldsymbol{n} \cdot \boldsymbol{u} = 0 \text{ on } \partial\Omega)$ in definition for non-periodic case)

$$(\boldsymbol{f}, \boldsymbol{g}) := \int_\Omega \boldsymbol{f}(\boldsymbol{x})^T \boldsymbol{g}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \;\Rightarrow\; (\boldsymbol{v}, \mathcal{N}_\eta \boldsymbol{u}) = (\mathcal{N}_\eta \boldsymbol{v}, \boldsymbol{u}) \quad (\boldsymbol{u}, \mathcal{K}_{\eta,h} \boldsymbol{u}) = (\mathcal{K}_{\eta,h} \boldsymbol{u}, \boldsymbol{u})$$

Choice of A: Use variational techniques

Introduce new variables: $\phi := \nabla \cdot \boldsymbol{u}$ and $\boldsymbol{w}(\boldsymbol{x}) = \nabla h(\boldsymbol{x})$

$$(\boldsymbol{u}, \mathcal{G}\boldsymbol{u}) = \int_\Omega \underbrace{\eta\|\boldsymbol{u}\|^2 + \frac{1}{3}\eta^3\phi^2}_{(\boldsymbol{u}, \mathcal{N}_\eta \boldsymbol{u})} + \underbrace{\eta(\boldsymbol{w}^T\boldsymbol{u})^2 + \eta^2(\boldsymbol{w}^T\boldsymbol{u})\phi}_{(\boldsymbol{u}, \mathcal{K}_{\eta,h}\boldsymbol{u})} \, \mathrm{d}\boldsymbol{x}$$

Key observation (look for a sums of squares):

$$(\boldsymbol{u}, \mathcal{G}\boldsymbol{u}) = \int_\Omega \eta|\boldsymbol{u}|^2 + \boldsymbol{a}^T \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix} \boldsymbol{a} \, \mathrm{d}\boldsymbol{x}, \quad \text{where } \boldsymbol{a}(\boldsymbol{x}) = \begin{pmatrix} \eta^{1/2} w \\ \eta^{3/2}\phi \end{pmatrix}$$

Choice of A: Use variational techniques

Obtain a bound: $$\gamma(\boldsymbol{u}, \mathcal{A}\boldsymbol{u}) \leq (\boldsymbol{u}, \mathcal{G}\boldsymbol{u}) \leq (\boldsymbol{u}, \mathcal{A}\boldsymbol{u})$$

With implicit operator/preconditioner: $$\mathcal{A} = \sigma \boldsymbol{I} - \alpha \nabla\nabla\cdot$$

With coefficients: $$\alpha = \frac{4}{3}\eta_{\max}^{3} \qquad \sigma = \eta_{\max}\left(1 + \frac{4}{3}(\nabla h)_{\max}^{2}\right)$$

``Estimated'' quantities: $$\eta_{\max} := \max_{\boldsymbol{x}\in\Omega} \eta(\boldsymbol{x}) \qquad \eta_{\min} := \min_{\boldsymbol{x}\in\Omega} \eta(\boldsymbol{x}), \quad (\nabla h)_{\max} = \max_{\boldsymbol{x}\in\Omega} \|\nabla h(\boldsymbol{x})\|$$

Formula: $$\gamma = \min\left\{ \left(\frac{\eta_{\min}}{\eta_{\max}}\right)^{3}, \left(\frac{\eta_{\min}}{\eta_{\max}}\right)\left(1 + \frac{4}{3}(\nabla h)_{\max}^{2}\right)^{-1} \right\}$$

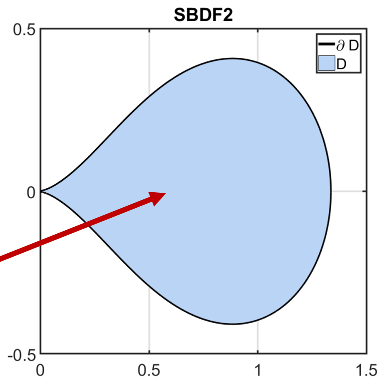
Significance of the result

Consider the generalized eigenvalues: $\mathcal{G}u = \lambda \mathcal{A}u$

Substitute into: $\gamma(u, \mathcal{A}u) \leq (u, \mathcal{G}u) \leq (u, \mathcal{A}u)$

1) Eigenvalue bound (are real in the right half plane)

$$\gamma \leq \mathrm{eig}\left(\mathcal{A}^{-1}\mathcal{G}\right) \leq 1$$



Significance --> the bounds are sufficient to make SBDF2 zero-stable (e.g., SBDF2 will work); For higher order SBDF methods, need to simultaneously splitting and time stepping schemes [cf. Seibold, S, Zhou 2019]

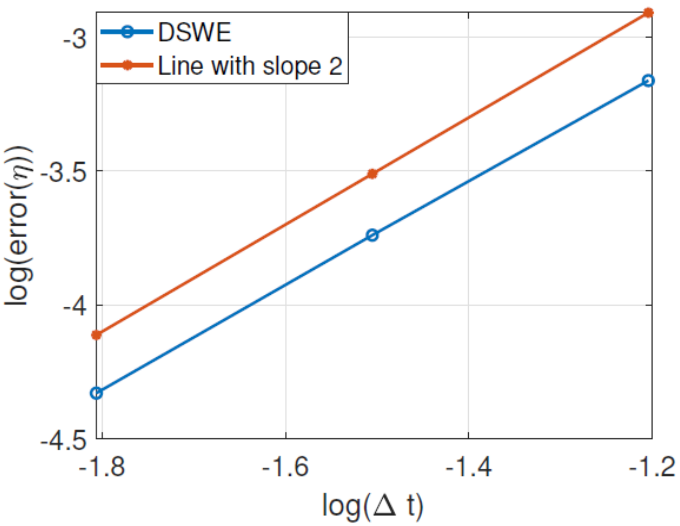
2) Eigenvalue bound implies the conditioning number bound:

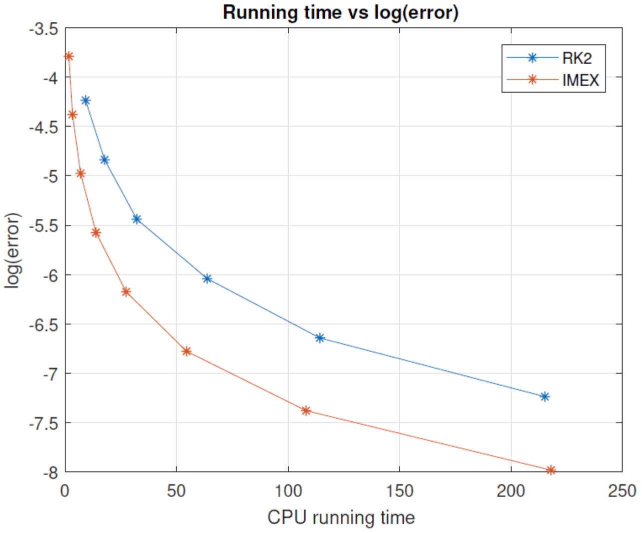
$$\kappa\left(\mathcal{A}^{-1}\mathcal{G}\right) \leq \gamma^{-1} = \max\left\{ \left(\frac{\eta_{\max}}{\eta_{\min}}\right)^3, \left(\frac{\eta_{\max}}{\eta_{\min}}\right)\left(1 + \frac{4}{3}(\nabla h)^2_{\max}\right) \right\}$$

No discretization in space (hence is mesh independent for suitable discrete operators)
Quantities can be ``estimated''; est. # iterations on preconditioned CG $U - \mathcal{G}u = 0$.

Some numerical results:

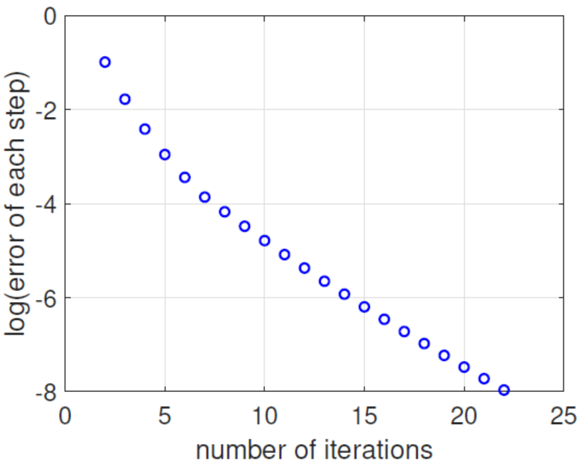
ImEx – convergence (2d test problem)



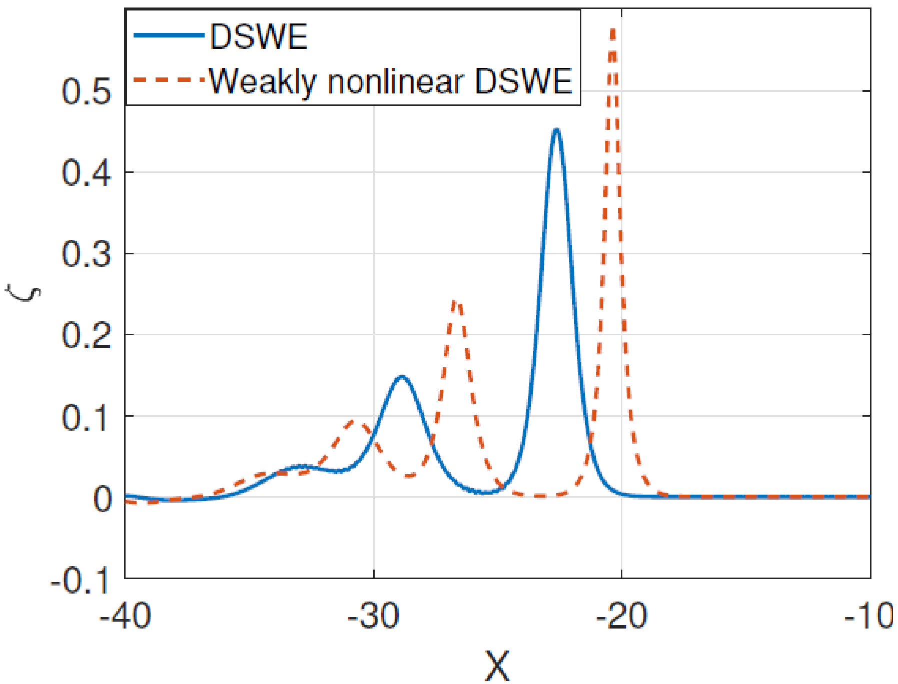


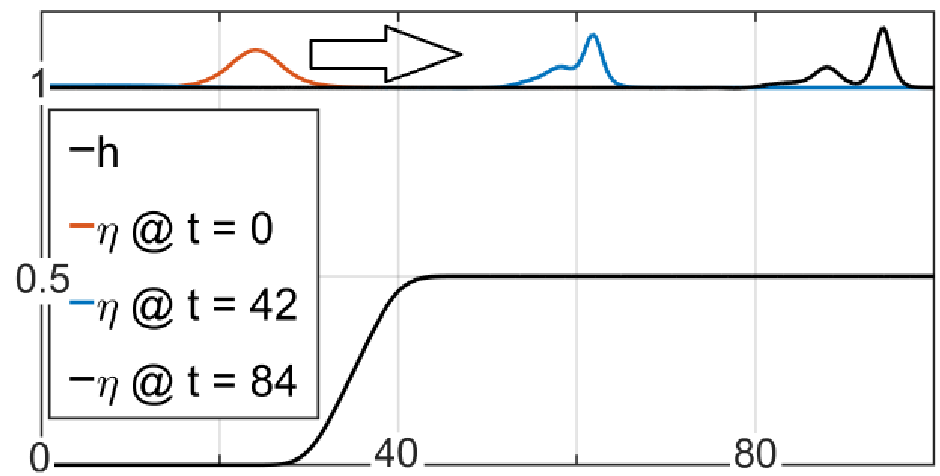
CPU run time for RK2 vs ImEx.

Iteration method: Convergence vs. iterations.



Some physical results:





Conclusions and Outlook:

#1. Model (scalar) ODEs: $$u_t = \lambda u + \mu u$$

Are useful for analyzing stability of: $$\boldsymbol{u}_t = A\boldsymbol{u} + B\boldsymbol{u} + \boldsymbol{f}(t)$$

Still provide necessary conditions when the matrices **do not** commute.

#2. Characterization of multistep methods. Fundamental barriers to getting 3rd order (elliptic problems); 2nd order (advection/wave problems); similar in spirit but (fundamentally different) Dahlquist barriers.

#3. Approaches avoid stiff implicit nonlinear terms (beyond 2nd order).

#4. Applications: New approaches for DAEs, and preconditioners for DSWE (avoid full treatment of mixed space-time derivatives).

Ref. (i) Theory paper: SINUM, 55:5 (2017), 2336-2360
(ii) Practice paper: JCP, 376:1 (2019), 295-321
(iii) DSWE manuscript in late stages of preparation.

Thank You!