$$\frac{dc_i}{dt} = \underbrace{P_i(c)}_{\sum_j p_{ij}(c)} - \underbrace{D_i(c)}_{\sum_j d_{ij}(c)} \qquad P_i,\ D_i \geq 0$$

$$p_{ij} = d_{ji} \quad \forall\ i \neq j \qquad p_{ii} = d_{ii} = 0$$

$S = c_1$

$I = c_2$

$$p_{21} = d_{12} = \beta \frac{SI}{N}$$

$$\underline{\sum c_i(t)} = \underline{\sum_i c_i(0)} \qquad \sum_i \frac{d}{dt} c_i = \left(\sum_i \sum_j p_{ij}(c) - d_{ij}(c) \right)$$

$$= \left(\sum_i \sum_j d_{ji}(c) - \sum_i \sum_j d_{ij}(c) \right) = 0$$

IF D_i ; $c_i \to 0 \Rightarrow D_i(c) \to 0$

$\Rightarrow$ POSITIVE

$$\partial_t c_i = -\underbrace{D_i(c)}_{\geq 0} + \underbrace{P_i(c)}_{\leq \underbrace{L c_i}_{\geq 0}} \geq 0$$

$c_i \to 0$

$c_i \geq 0$

$\sqrt{\geq 0}$

- HIGH ORDER
- POSITIVE
- CONSERVATIVE

$$c_i^{n+1} = c_i^n + \Delta t \left(P_i(c^n) - D_i(c^n) \right) + \Delta t \sum_j p_{ij}(c^n) - d_{ij}(c^n)$$

$$\sum_i c_i^{n+1} = \sum_i c_i^n + \Delta t \sum_{i,j} \underbrace{\underbrace{p_{ij}(c^n)}_{d_{ji}(c^n)} - d_{ij}(c^n)}_{=0}$$

$=$

CONSERVATIVE

NOT POSITIVE $\quad \exists\, i : D_i(c^n) > 0$

$$> P_i(c^n)$$

$c_i^{n+1} < 0 \Rightarrow$ NOT POSITIVE

$$c_i^{n+1} = c_i^n + \Delta t\left(\sum p_{ij}(c^n) - \sum d_{ij}(c^n)\,\frac{c_i^{n+1}}{c_i^n}\right)$$

$$\underbrace{\left[1 + \Delta t \sum_j \frac{d_{ij}(c^n)}{c_i^n}\right]}_{>0} c_i^{n+1} = \underbrace{c_i^n + \Delta t \sum_j p_{ij}(c^n)}_{\geq 0} \quad \forall i$$

$\Rightarrow c_i^{n+1} > 0$ ✓ POSITIVE

$$\frac{c_i^{n+1}}{c_i^n} = 1 + O(\Delta t)$$

NOT EXPENSIVE

SAME ACCURACY

$$c_i^{n+1} = c_i^n + \Delta t \left(\sum_{j=1} p_{ij}(c^n) \frac{c_j^{n+1}}{c_j^n} - \sum d_{ij}(c^n) \frac{c_i^{n+1}}{c_i^n} \right)$$

$$\sum_i c_i^{n+1} = \sum_i c_i^n + \Delta t \underbrace{\sum_{ij} \underbrace{p_{ij}(c^n)}_{d_{ji}} \frac{c_j^{n+1}}{c_j^n}} - \underbrace{\sum_{ij} d_{ij}(c^n) \frac{c_i^{n+1}}{c_i^n}}$$

• CONSERVATIVE

$$\underbrace{\underbrace{\sum_{i,j} d_{ij}(c^n) \frac{c_i^{n+1}}{c_i^n}} \qquad\qquad}_{=0}$$

• POSITIVE

$$c_i^{n+1} = c_i^n + \Delta t \left(\sum_{j=1} p_{ij}(c^n) \frac{c_j^{n+1}}{c_j^n} - \sum d_{ij}(c^n) \frac{c_i^{n+1}}{c_i^n} \right)$$

$$\underbrace{\left[1 + \Delta t \sum_j \frac{d_{ij}(c^n)}{c_i^n} \right]}_{\to\, M_{ii} =} c_i^{n+1} - \underbrace{\Delta t \sum_j \frac{p_{ij}(c^n)}{c_j^n}}_{\to\, M_{ij} =} \cdot c_j^{n+1} = c_i^n$$

$$\underline{\underline{M}}\, \underline{c}^{n+1} = \underline{c}^n \qquad M_{ii} = \qquad M_{ij} = \qquad i \neq j$$

GOAL $\quad \underline{c}^{n+1} = \underbrace{\underline{\underline{M}}^{-1}}_{>0} \underbrace{\underline{c}^n}_{>0} > 0 \qquad$ POSITIVE M^{-1}

IFF $\quad (M^{-1})_{ij} \geq 0$

$$M_{ii} > \sum_{j\neq i} |M_{ji}| = \Delta t \sum_{j\neq i} \frac{P_{ij}(c^n)}{c_j^n}$$

$d_{ii} = 0$

$p_{ii} = 0$

$$\mathbin{\rotatebox{90}{=}}$$

$$1 + \Delta t \sum_{\ell} \frac{d_{i\ell}(c^n)}{c_i^n} > \Delta t \sum_{\ell} \frac{d_{i\ell}(c^n)}{c_i^n} = \Delta t \sum_{\ell} \frac{P_{\ell i}(c^n)}{c_i^n}$$

$$= \sum_{j\neq i} |M_{ji}| \quad \checkmark$$

$$\overline{M x = \underbrace{b}_{\geq 0}}$$

$$M = \underbrace{D}_{\operatorname{diag}(M)} - \underbrace{L}_{= D - M}$$

$$x = D^{-1} L x + D^{-1} b$$

$$D x - L x = b \iff D x = L x + b$$

$$D x^{(k+1)} = L x^{(k)} + b$$

JACOBI

ITERATIVE METHOD

- CONVERGE TOWARDS x
- POSITIVE AT ALL (k)

$$x^{(k+1)} = D^{-1}Lx^{(k)} + D^{-1}b$$

$$x^{(k+1)} - x =: e^{(k+1)}$$

$$\|e^{(k+1)}\|_\infty = \|x^{(k+1)} - x\| = \|D^{-1}Lx^{(k)} + \cancel{D^{-1}b} - D^{-1}Lx - \cancel{D^{-1}b}\|$$

$$= \|D^{-1}L(x^{(k)} - x)\| \leq \underbrace{\|D^{-1}L\|_\infty}_{<1} \|x^{(k)} - x\|$$

$$\|A\|_\infty = \max_j \sum_i |A_{ij}|$$

$$\|D^{-1}L\|_\infty = \max_j \sum_{i \neq j} \frac{|-M_{ij}|}{|M_{jj}|} = \max_j \frac{\overbrace{\sum_{i \neq j} |M_{ij}|}}{|M_{jj}|} < 1$$

$$x^{(k+1)} = \underbrace{D^{-1}}_{\geq 0} \underbrace{L}_{\geq 0} \underbrace{x^{(k)}}_{\geq 0} + \underbrace{D^{-1}}_{\geq 0} \underbrace{b}_{\geq 0} \geq 0$$

$$x^{(k+1)} \rightarrow x \geq 0 .$$

Arbitrary high-order, conservative and positive preserving Patankar-type deferred correction schemes

Davide Torlo

MathLab, Mathematics Area, SISSA International School for Advanced Studies, Trieste, Italy
davidetorlo.it

Based on: Öffner, P. & Torlo, D. *Arbitrary high-order, conservative and positivity preserving Patankar-type deferred correction schemes.* APNUM 153, 15–34 (2020).
https://doi.org/10.1016/j.apnum.2020.01.025

Outline

1. Production–Destruction system
2. Deferred Correction
3. Modified Patankar DeC (mPDeC)
4. Numerics

Outline

1. Production–Destruction system
2. Deferred Correction
3. Modified Patankar DeC (mPDeC)
4. Numerics

Production–Destruction system

Consider production-destruction systems (PDS)

$$\begin{cases} d_t c_i = P_i(\mathbf{c}) - D_i(\mathbf{c}), \quad i = 1, \dots, I, & P_i(\mathbf{c}) = \sum_{j=1}^{I} p_{i,j}(\mathbf{c}), \\ \mathbf{c}(t=0) = \mathbf{c}_0, & D_i(\mathbf{c}) = \sum_{j=1}^{I} d_{i,j}(\mathbf{c}), \end{cases} \tag{1}$$

where

$$p_{i,j}(\mathbf{c}), d_{i,j}(\mathbf{c}) \geq 0, \qquad \forall i, j \in I, \quad \forall \mathbf{c} \in \mathbb{R}^{+,I}.$$

Applications: Chemical reactions, biological systems, population evolutions and PDEs.
Example: SIRD

$$\begin{cases} d_t S = -\beta \frac{SI}{N} \\ d_t I = \beta \frac{SI}{N} - \gamma I - \delta I \\ d_t R = \gamma I \\ d_t D = \delta I \end{cases}$$

Production–Destruction system

Consider production-destruction systems (PDS)

$$\begin{cases} d_t c_i = P_i(\mathbf{c}) - D_i(\mathbf{c}), \quad i = 1, \dots, I, & P_i(\mathbf{c}) = \sum_{j=1}^{I} p_{i,j}(\mathbf{c}), \\ \mathbf{c}(t=0) = \mathbf{c}_0, & D_i(\mathbf{c}) = \sum_{j=1}^{I} d_{i,j}(\mathbf{c}), \end{cases} \tag{1}$$

where

$$p_{i,j}(\mathbf{c}), d_{i,j}(\mathbf{c}) \geq 0, \qquad \forall i,j \in I, \quad \forall \mathbf{c} \in \mathbb{R}^{+,I}.$$

Property 1: Conservation

$$\sum_{i=1}^{I} c_i(0) = \sum_{i=1}^{I} c_i(t), \quad \forall t \geq 0$$

$$\iff \quad p_{i,j}(\mathbf{c}) = d_{j,i}(\mathbf{c}), \qquad \forall i,j \in I, \quad \forall \mathbf{c} \in \mathbb{R}^{+,I}.$$

Production–Destruction system

Consider production-destruction systems (PDS)

$$\begin{cases} d_t c_i = P_i(\mathbf{c}) - D_i(\mathbf{c}), \quad i = 1, \dots, I, & P_i(\mathbf{c}) = \sum_{j=1}^{I} p_{i,j}(\mathbf{c}), \\ \mathbf{c}(t=0) = \mathbf{c}_0, & D_i(\mathbf{c}) = \sum_{j=1}^{I} d_{i,j}(\mathbf{c}), \end{cases} \tag{1}$$

where

$$p_{i,j}(\mathbf{c}), d_{i,j}(\mathbf{c}) \geq 0, \qquad \forall i, j \in I, \quad \forall \mathbf{c} \in \mathbb{R}^{+,I}.$$

Property 2: Positivity

If P_i, D_i Lipschitz, and if when $c_i \to 0 \Rightarrow D_i(\mathbf{c}) \to 0 \Longrightarrow$
$c_i(0) > 0\, \forall i \in I \Longrightarrow c_i(t) > 0\, \forall i \in I\; \forall t > 0.$

Production–Destruction system

Consider production-destruction systems (PDS)

$$\begin{cases} d_t c_i = P_i(\mathbf{c}) - D_i(\mathbf{c}), \quad i = 1, \dots, I, & P_i(\mathbf{c}) = \sum_{j=1}^{I} p_{i,j}(\mathbf{c}), \\ \mathbf{c}(t=0) = \mathbf{c}_0, & D_i(\mathbf{c}) = \sum_{j=1}^{I} d_{i,j}(\mathbf{c}), \end{cases} \tag{1}$$

where

$$p_{i,j}(\mathbf{c}), d_{i,j}(\mathbf{c}) \geq 0, \qquad \forall i,j \in I, \quad \forall \mathbf{c} \in \mathbb{R}^{+,I}.$$

Goal:

- One step method
- Unconditionally positive
- Unconditionally conservative
- High order accurate

Solvers

Explicit Euler

- $\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t \left(\mathbf{P}(\mathbf{c}^n) - \mathbf{D}(\mathbf{c}^n) \right)$
- Conservative
- First order
- Not unconditionally positive, if Δt is too big... CFL conditions

Implicit Euler

- $\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t \left(\mathbf{P}(\mathbf{c}^{n+1}) - \mathbf{D}(\mathbf{c}^{n+1}) \right)$
- Conservative & positive
- First order
- Expensive to be solved/not unique solution: Nonlinear solvers!!!

Patankar trick

$$c_i^{n+1} = c_i^n + \Delta t \left(P_i(\mathbf{c}^n) - D_i(\mathbf{c}^n) \frac{c_i^{n+1}}{c_i^n} \right)$$

$$\left(1 + \Delta t \frac{D_i(\mathbf{c}^n)}{c_i^n} \right) c_i^{n+1} = c_i^n + \Delta t P_i(\mathbf{c}^n)$$

- Not conservative
- First order
- Positive
- Implicit, but easy

Solvers

Modified Patankar (mP)
Burchard, Deleersnijder & Meister

$$c_i^{n+1} = c_i^n + \Delta t \left(\sum_j p_{i,j}(\mathbf{c}^n) \frac{c_j^{n+1}}{c_j^n} - \sum_j d_{i,j}(\mathbf{c}^n) \frac{c_i^{n+1}}{c_i^n} \right) \tag{2}$$

$\mathrm{M}(\mathbf{c}^n)\mathbf{c}^{n+1} = \mathbf{c}^n$ where M is

$$\begin{cases} m_{i,i}(\mathbf{c}^n) = 1 + \Delta t \sum_{k=1}^{I} \frac{d_{i,k}(\mathbf{c}^n)}{c_i^n}, & i = 1, \dots, I, \\ m_{i,j}(\mathbf{c}^n) = -\Delta t \frac{p_{i,j}(\mathbf{c}^n)}{c_j^n}, & i,j = 1, \dots, I,\ i \neq j. \end{cases} \tag{3}$$

- Conservative
- First order
- Positive
- Linear system at each timestep

- Extension to RK2 and RK3 (Burchard, Deleersnijder, Meister, Kopecz)
- Extension to PDEs (Huang, Zhao, Shu)

Outline

1. Production–Destruction system
2. Deferred Correction
3. Modified Patankar DeC (mPDeC)
4. Numerics

Deferred Correction discretization

We should discretize our variable on $[t^n, t^{n+1}]$ in M substeps ($\mathbf{c}^{n,m}$).

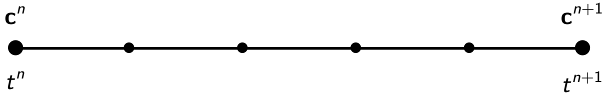


Figure: Subtimeintervals

Then, we can rewrite $\mathbf{c}^m = \mathbf{c}^0 + \int_{t^0}^{t^m} \mathbf{P}(\mathbf{c}(s)) - \mathbf{D}(\mathbf{c}(s))\, ds$.
Equispaced points $\Rightarrow$ order $= M + 1$.

$$\underline{\mathbf{c}} := (\mathbf{c}^0, \dots, \mathbf{c}^M) \in \mathbb{R}^{M \times I} \tag{4}$$

Deferred Correction discretization

We should discretize our variable on $[t^n, t^{n+1}]$ in M substeps ($\mathbf{c}^{n,m}$).

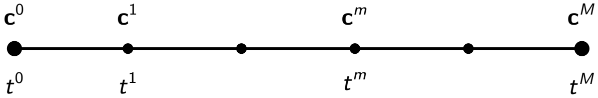


Figure: Subtimeintervals

Then, we can rewrite $\mathbf{c}^m = \mathbf{c}^0 + \int_{t^0}^{t^m} \mathbf{P}(\mathbf{c}(s)) - \mathbf{D}(\mathbf{c}(s))\, ds$.
Equispaced points $\Rightarrow$ order $= M + 1$.

$$\underline{\mathbf{c}} := (\mathbf{c}^0, \dots, \mathbf{c}^M) \in \mathbb{R}^{M \times I} \tag{4}$$

DeC operators

$\mathcal{L}^2$ operator

$$\mathbf{E} := \mathbf{P} - \mathbf{D}$$

$$\mathcal{L}^2(\mathbf{c}^0, \dots, \mathbf{c}^M) = \mathcal{L}^2(\underline{\mathbf{c}}) :=$$

$$\begin{cases} \mathbf{c}^M - \mathbf{c}^0 - \int_{t^0}^{t^M} \mathbf{E}(\mathbf{c}(s))ds \\ \vdots \\ \mathbf{c}^1 - \mathbf{c}^0 - \int_{t^0}^{t^1} \mathbf{E}(\mathbf{c}(s))ds \end{cases}$$

- Implicit RK
- Order of accuracy $\geq M+1$
- Difficult to solve directly

DeC operators

$\mathcal{L}^2$ operator

$$\mathcal{L}^2(\mathbf{c}^0,\dots,\mathbf{c}^M)=\mathcal{L}^2(\underline{\mathbf{c}}):=$$
$$\begin{cases} \mathbf{c}^M-\mathbf{c}^0-\Delta t\sum_{r=0}^M \theta_r^M \mathbf{E}(\mathbf{c}^r)\\ \dots\\ \mathbf{c}^1-\mathbf{c}^0-\Delta t\sum_{r=0}^M \theta_r^1 \mathbf{E}(\mathbf{c}^r)\end{cases}$$

- Implicit RK
- Order of accuracy $\geq M+1$
- Difficult to solve directly

DeC operators

$\mathcal{L}^2$ operator

$$\mathcal{L}^2(\mathbf{c}^0,\dots,\mathbf{c}^M)=\mathcal{L}^2(\underline{\mathbf{c}}):=$$
$$\begin{cases}\mathbf{c}^M-\mathbf{c}^0-\Delta t\sum_{r=0}^{M}\theta_r^M\mathbf{E}(\mathbf{c}^r)\\ \dots\\ \mathbf{c}^1-\mathbf{c}^0-\Delta t\sum_{r=0}^{M}\theta_r^1\mathbf{E}(\mathbf{c}^r)\end{cases}$$

- Implicit RK
- Order of accuracy $\geq M+1$
- Difficult to solve directly

$\mathcal{L}^1$ operator

$$\mathcal{L}^1(\mathbf{c}^0,\dots,\mathbf{c}^M)=\mathcal{L}^1(\underline{\mathbf{c}}):=$$
$$\begin{cases}\mathbf{c}^M-\mathbf{c}^0-\Delta t\beta^M\mathbf{E}(\mathbf{c}^0)\\ \dots\\ \mathbf{c}^1-\mathbf{c}^0-\Delta t\beta^1\mathbf{E}(\mathbf{c}^0)\end{cases}$$

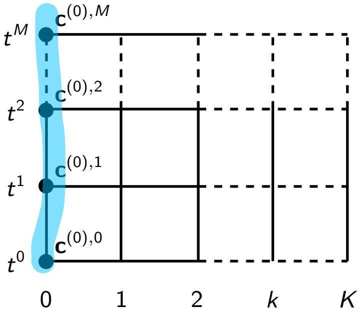
- First order accurate
- Explicit or easy to solve

Deferred Correction

How to combine two methods keeping the accuracy of the second and the stability and simplicity of the first one?

$$\mathbf{c}^{0,(k)} := \mathbf{c}(t^n), \quad k = 0, \dots, K,$$
$$\mathbf{c}^{m,(0)} := \mathbf{c}(t^n), \quad m = 1, \dots, M$$
$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$

- $\mathcal{L}^1(\underline{\mathbf{c}}) = 0$, first order accuracy, easily invertible.
- $\mathcal{L}^2(\underline{\mathbf{c}}) = 0$, high order $M+1$.



DeC Theorem

- $\mathcal{L}^1$ coercive
- $\mathcal{L}^1 - \mathcal{L}^2$ Lipschitz

DeC converges and $\min(K, M+1)$ is the order of accuracy.

Deferred Correction

How to combine two methods keeping the accuracy of the second and the stability and simplicity of the first one?

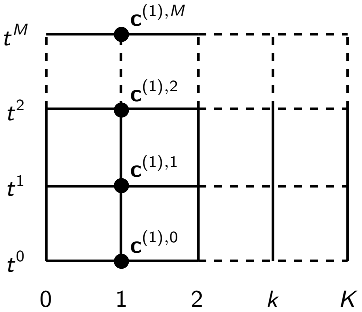
- $\mathcal{L}^1(\underline{\mathbf{c}}) = 0$, first order accuracy, easily invertible.
- $\mathcal{L}^2(\underline{\mathbf{c}}) = 0$, high order $M+1$.

$$\mathbf{c}^{0,(k)} := \mathbf{c}(t^n), \quad k = 0, \dots, K,$$
$$\mathbf{c}^{m,(0)} := \mathbf{c}(t^n), \quad m = 1, \dots, M$$
$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$

DeC Theorem

- $\mathcal{L}^1$ coercive
- $\mathcal{L}^1 - \mathcal{L}^2$ Lipschitz

DeC converges and $\min(K, M+1)$ is the order of accuracy.

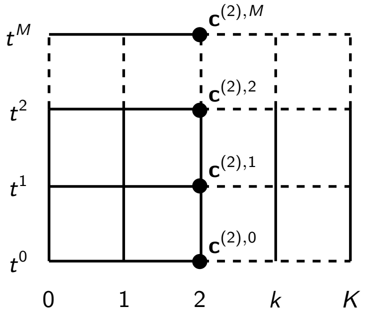


Deferred Correction

How to combine two methods keeping the accuracy of the second and the stability and simplicity of the first one?

$$\mathbf{c}^{0,(k)} := \mathbf{c}(t^n), \quad k = 0, \dots, K,$$
$$\mathbf{c}^{m,(0)} := \mathbf{c}(t^n), \quad m = 1, \dots, M$$
$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$

- $\mathcal{L}^1(\underline{\mathbf{c}}) = 0$, first order accuracy, easily invertible.
- $\mathcal{L}^2(\underline{\mathbf{c}}) = 0$, high order $M+1$.



DeC Theorem

- $\mathcal{L}^1$ coercive
- $\mathcal{L}^1 - \mathcal{L}^2$ Lipschitz

DeC converges and $\min(K, M+1)$ is the order of accuracy.

Deferred Correction

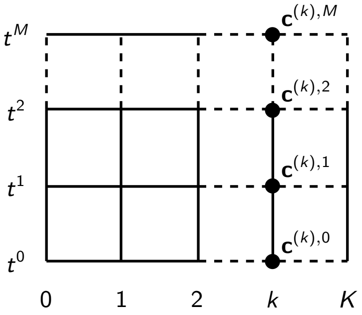
How to combine two methods keeping the accuracy of the second and the stability and simplicity of the first one?

- $\mathcal{L}^1(\underline{\mathbf{c}}) = 0$, first order accuracy, easily invertible.
- $\mathcal{L}^2(\underline{\mathbf{c}}) = 0$, high order $M+1$.

$$\mathbf{c}^{0,(k)} := \mathbf{c}(t^n), \quad k = 0, \dots, K,$$

$$\mathbf{c}^{m,(0)} := \mathbf{c}(t^n), \quad m = 1, \dots, M$$

$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$



DeC Theorem

- $\mathcal{L}^1$ coercive
- $\mathcal{L}^1 - \mathcal{L}^2$ Lipschitz

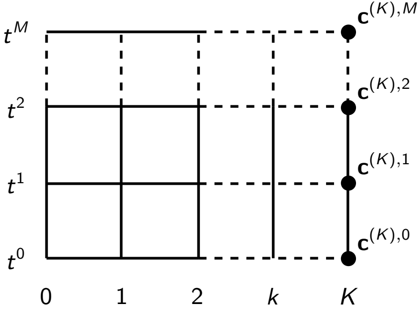
DeC converges and $\min(K, M+1)$ is the order of accuracy.

Deferred Correction

How to combine two methods keeping the accuracy of the second and the stability and simplicity of the first one?

$$\mathbf{c}^{0,(k)} := \mathbf{c}(t^n), \quad k = 0, \dots, K,$$
$$\mathbf{c}^{m,(0)} := \mathbf{c}(t^n), \quad m = 1, \dots, M$$
$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$

- $\mathcal{L}^1(\underline{\mathbf{c}}) = 0$, first order accuracy, easily invertible.
- $\mathcal{L}^2(\underline{\mathbf{c}}) = 0$, high order $M+1$.



DeC Theorem

- $\mathcal{L}^1$ coercive
- $\mathcal{L}^1 - \mathcal{L}^2$ Lipschitz

DeC converges and $\min(K, M+1)$ is the order of accuracy.

Explicit DeC

If we write explicitly the DeC step we see that

$$
\begin{aligned}
\mathcal{L}_i^{1,m}(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}_i^{1,m}(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}_i^{2,m}(\underline{\mathbf{c}}^{(k-1)}) \iff \\
c_i^{(k),m} - c_i^0 - \Delta t \beta^m E_i(\mathbf{c}^0) = c_i^{(k-1),m} - c_i^0 - \Delta t \beta^m E_i(\mathbf{c}^0) \\
- c_i^{(k-1),m} + c_i^0 + \Delta t \sum_{r=0}^{M} \theta_r^m E_i(\mathbf{c}^{(k-1),r}) \iff \\
c_i^{(k),m} = c_i^0 + \Delta t \sum_{r=0}^{M} \theta_r^m E_i(\mathbf{c}^{(k-1),r}) \iff \\
c_i^{(k),m} = c_i(t^n) + \Delta t \sum_{r=0}^{M} \theta_r^m E_i(\mathbf{c}^{(k-1),r})
\end{aligned}
\tag{5}
$$

Explicit DeC

If we write explicitly the DeC step we see that

$$
\begin{aligned}
&\mathcal{L}^{1,m}_i(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^{1,m}_i(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^{2,m}_i(\underline{\mathbf{c}}^{(k-1)}) \iff \\
&c_i^{(k),m} - c_i^0 - \Delta t \beta^m E_i(\mathbf{c}^0) = c_i^{(k-1),m} - c_i^0 - \Delta t \beta^m E_i(\mathbf{c}^0) \\
&\qquad - c_i^{(k-1),m} + c_i^0 + \Delta t \sum_{r=0}^{M} \theta_r^m E_i(\mathbf{c}^{(k-1),r}) \iff \\
&\boxed{c_i^{(k),m} = c_i^0 + \Delta t \sum_{r=0}^{M} \theta_r^m E_i(\mathbf{c}^{(k-1),r})} \iff \\
&c_i^{(k),m} = c_i(t^n) + \Delta t \sum_{r=0}^{M} \theta_r^m E_i(\mathbf{c}^{(k-1),r})
\end{aligned}
\tag{5}
$$

Ingredients

- We want to use the DeC for high order accuracy
- We want to recast positivity and conservation
- We will use the Patankar trick
- We want an implicit method (to get positivity), but only linearly implicit (no nonlinear solvers)
- We have to modify $\mathcal{L}^2$ using the trick

Outline

1. Production–Destruction system
2. Deferred Correction
3. Modified Patankar DeC (mPDeC)
4. Numerics

Modified Patankar $\mathcal{L}^2$

Modify the operator $\mathcal{L}^2$ according to the Patankar trick!

$$\mathcal{L}^2_i(\mathbf{c}^{0,(k-1)},\dots,\mathbf{c}^{M,(k-1)} \qquad\qquad\qquad) = \mathcal{L}^2_i(\underline{\mathbf{c}}^{(k-1)} \qquad) :=$$

$$\begin{cases} c_i^{M,(k-1)} - c_i^{0,(k-1)} - \Delta t \sum_{r=0}^{M} \theta_r^M \sum_{j=1}^{I} \Big(p_{i,j}(\mathbf{c}^{r,(k-1)}) \qquad\qquad - d_{i,j}(\mathbf{c}^{r,(k-1)}) \qquad\qquad \Big), \\ \vdots \\ c_i^{1,(k-1)} - c_i^{0,(k-1)} - \Delta t \sum_{r=0}^{M} \theta_r^1 \sum_{j=1}^{I} \Big(p_{i,j}(\mathbf{c}^{r,(k-1)}) \qquad\qquad - d_{i,j}(\mathbf{c}^{r,(k-1)}) \qquad\qquad \Big), \end{cases}$$

Modified Patankar $\mathcal{L}^2$

Modify the operator $\mathcal{L}^2$ according to the Patankar trick!

$$\mathcal{L}^2_i(\mathbf{c}^{0,(k-1)},\dots,\mathbf{c}^{M,(k-1)},\mathbf{c}^{0,(k)},\dots,\mathbf{c}^{M,(k)}) = \mathcal{L}^2_i(\underline{\mathbf{c}}^{(k-1)},\underline{\mathbf{c}}^{(k)}) :=$$

$$\begin{cases} c_i^{M,(k-1)} - c_i^{0,(k-1)} - \Delta t \sum_{r=0}^{M} {\color{red}\theta_r^M} \sum_{j=1}^{I} \left(p_{i,j}(\mathbf{c}^{r,(k-1)}) {\color{blue}\frac{c_j^{M,(k)}}{c_j^{M,(k-1)}}} - d_{i,j}(\mathbf{c}^{r,(k-1)}) {\color{green}\frac{c_i^{M,(k)}}{c_i^{M,(k-1)}}} \right), \\ \vdots \\ c_i^{1,(k-1)} - c_i^{0,(k-1)} - \Delta t \sum_{r=0}^{M} {\color{red}\theta_r^1} \sum_{j=1}^{I} \left(p_{i,j}(\mathbf{c}^{r,(k-1)}) {\color{blue}\frac{c_j^{1,(k)}}{c_j^{1,(k-1)}}} - d_{i,j}(\mathbf{c}^{r,(k-1)}) {\color{green}\frac{c_i^{1,(k)}}{c_i^{1,(k-1)}}} \right), \end{cases}$$

Modified Patankar $\mathcal{L}^2$

Modify the operator $\mathcal{L}^2$ according to the Patankar trick!

$$\mathcal{L}^2_i(\mathbf{c}^{0,(k-1)},\dots,\mathbf{c}^{M,(k-1)},\mathbf{c}^{0,(k)},\dots,\mathbf{c}^{M,(k)}) = \mathcal{L}^2_i(\underline{\mathbf{c}}^{(k-1)},\underline{\mathbf{c}}^{(k)}) :=$$

$$\begin{cases} c_i^{M,(k-1)} - c_i^{0,(k-1)} - \Delta t \sum_{r=0}^{M} \theta_r^M \sum_{j=1}^{I} \left(p_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{M,(k)}_{\gamma(j,i,\theta_r^M)}}{c^{M,(k-1)}_{\gamma(j,i,\theta_r^M)}} - d_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{M,(k)}_{\gamma(i,j,\theta_r^M)}}{c^{M,(k-1)}_{\gamma(i,j,\theta_r^M)}} \right), \\ \vdots \\ c_i^{1,(k-1)} - c_i^{0,(k-1)} - \Delta t \sum_{r=0}^{M} \theta_r^1 \sum_{j=1}^{I} \left(p_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{1,(k)}_{\gamma(j,i,\theta_r^1)}}{c^{1,(k-1)}_{\gamma(j,i,\theta_r^1)}} - d_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{1,(k)}_{\gamma(i,j,\theta_r^1)}}{c^{1,(k-1)}_{\gamma(i,j,\theta_r^1)}} \right), \end{cases}$$

where $\gamma(a,b,\theta) = a$ if $\theta > 0$ and $\gamma(a,b,\theta) = b$ if $\theta < 0$.

Modified Patankar DeC (mPDeC)

Reminder: initial states $c_i^{0,(k)}$ are identical for any correction (k)
DeC Patankar can be rewritten for $k=1,\dots,K$, $m=1,\dots,M$ and $\forall i \in I$ into

$$
\begin{split}
&\mathcal{L}_i^{1,m}(\underline{\mathbf{c}}^{(k)}) - \mathcal{L}_i^{1,m}(\underline{\mathbf{c}}^{(k-1)}) + \mathcal{L}_i^{2,m}(\underline{\mathbf{c}}^{(k)}, \underline{\mathbf{c}}^{(k-1)}) = 0\\
&c_i^{m,(k)} - c_i^0 - \Delta t \sum_{r=0}^{M} \theta_r^m \sum_{j=1}^{I} \left(p_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{m,(k)}_{\gamma(j,i,\theta_r^m)}}{c^{m,(k-1)}_{\gamma(j,i,\theta_r^m)}} - d_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{m,(k)}_{\gamma(i,j,\theta_r^m)}}{c^{m,(k-1)}_{\gamma(i,j,\theta_r^m)}} \right) = 0.
\end{split}
\tag{6}
$$

- Conservation
- Positivity
- High order accuracy

Conservation

The mPDeC scheme is unconditionally conservative for all substages, i.e.,

$$\sum_{i=1}^{I} c_i^{m,(k)} = \sum_{i=1}^{I} c_i^0,$$

for all $k = 1, \dots, K$ and $m = 0, \dots, M$.
Using formulation (6), we can easily see that $\forall k, m$

$$0 = \sum_{i \in I} c_i^{m,(k)} - \sum_{i \in I} c_i^0 =$$

d_{ji}

$$= \Delta t \sum_{i,j=1}^{I} \sum_{r=0}^{M} \theta_r^m \left(p_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{m,(k)}_{\gamma(j,i,\theta_r^m)}}{c^{m,(k-1)}_{\gamma(j,i,\theta_r^m)}} - d_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{m,(k)}_{\gamma(i,j,\theta_r^m)}}{c^{m,(k-1)}_{\gamma(i,j,\theta_r^m)}} \right) =$$

Conservation

The mPDeC scheme is unconditionally conservative for all substages, i.e.,

$$\sum_{i=1}^{I} c_i^{m,(k)} = \sum_{i=1}^{I} c_i^0,$$

for all $k = 1, \dots, K$ and $m = 0, \dots, M$.
Using formulation (6), we can easily see that $\forall k, m$

$$\sum_{i\in I} c_i^{m,(k)} - \sum_{i \in I} c_i^0 =$$

$$=\Delta t \sum_{i,j=1}^{I} \sum_{r=0}^{M} \theta_r^m \left(\underbrace{d_{j,i}(\mathbf{c}^{r,(k-1)}) \frac{c^{m,(k)}_{\gamma(j,i,\theta_r^m)}}{c^{m,(k-1)}_{\gamma(j,i,\theta_r^m)}}} - \underbrace{d_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{m,(k)}_{\gamma(i,j,\theta_r^m)}}{c^{m,(k-1)}_{\gamma(i,j,\theta_r^m)}}} \right) =$$

Conservation

The mPDeC scheme is unconditionally conservative for all substages, i.e.,

$$\sum_{i=1}^{I} c_i^{m,(k)} = \sum_{i=1}^{I} c_i^0,$$

for all $k = 1, \dots, K$ and $m = 0, \dots, M$.
Using formulation (6), we can easily see that $\forall k, m$

$$\begin{aligned}
&\sum_{i\in I} c_i^{m,(k)} - \sum_{i \in I} c_i^0 = \\
=&\Delta t \sum_{i,j=1}^{I} \sum_{r=0}^{M} \theta_r^m \left(d_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{m,(k)}_{\gamma(j,i,\theta_r^m)}}{c^{m,(k-1)}_{\gamma(j,i,\theta_r^m)}} - d_{i,j}(\mathbf{c}^{r,(k-1)}) \frac{c^{m,(k)}_{\gamma(i,j,\theta_r^m)}}{c^{m,(k-1)}_{\gamma(i,j,\theta_r^m)}} \right) = 0.
\end{aligned}$$

Positivity

At each step (m,k) implicit linear system with mass matrix

$$\mathrm{M}(\mathbf{c}^{m,(k-1)})_{ij} =
\begin{cases}
1 + \Delta t \sum_{r=0}^{M} \sum_{l=1}^{I} \frac{\theta_r^m}{c_i^{m,(k-1)}} \left(d_{i,l}(\mathbf{c}^{r,(k-1)}) \chi_{\{\theta_r^m>0\}} - p_{i,l}(\mathbf{c}^{r,(k-1)}) \chi_{\{\theta_r^m<0\}} \right) & \text{for } i=j \\
-\Delta t \sum_{r=0}^{M} \frac{\theta_r^m}{c_j^{m,(k-1)}} \left(p_{i,j}(\mathbf{c}^{r,(k-1)}) \chi_{\{\theta_r^m>0\}} - d_{i,j}(\mathbf{c}^{r,(k-1)}) \chi_{\{\theta_r^m<0\}} \right) & \text{for } i \neq j
\end{cases}$$

>0

<0

>0

<0

<0

- Diagonally dominant by columns
- Invertible
- $\mathrm{M}^{-1} > 0$

$D > 0$

$L > 0$

$L = D - M$

✓

High order accuracy

Let $\underline{\mathbf{c}}^*$ be the solution of the $\mathcal{L}^2$ operator, i.e., $\mathcal{L}^2(\underline{\mathbf{c}}^*, \underline{\mathbf{c}}^*) = 0$.

- Coercivity operator $\mathcal{L}^1$: $||\mathcal{L}^1(\underline{\mathbf{c}}) - \mathcal{L}^1(\underline{\mathbf{c}}^*)|| \geq C_1 ||\underline{\mathbf{c}} - \underline{\mathbf{c}}^*||$
- Lipschitz continuity operator $\mathcal{L}^1 - \mathcal{L}^2$:
  $||\mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \underbrace{\mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}, \underline{\mathbf{c}}^{(k)})} - \mathcal{L}^1(\underline{\mathbf{c}}^*) + \mathcal{L}^2(\underline{\mathbf{c}}^*, \underline{\mathbf{c}}^*)|| \leq C_L \Delta t ||\underline{\mathbf{c}}^{(k-1)} - \underline{\mathbf{c}}^*||$.

  Intermediate steps for Lipschitz continuity
  - $\mathbf{c}^{m,(k)} = \mathbf{c}^0 + \Delta t G(\mathbf{c}^{m,(k-1)})\mathbf{c}^0$
  - $\boxed{\frac{c_i^{(k)}}{c_i^{(k-1)}}} = 1 + \Delta t^{k-1} g_i + \mathcal{O}(\Delta t^k)$

$c_i^{(k),m}$ K-TH ORDER ACCURATE APPROX OF $c_i(t^m)$

$$\frac{c_i(t^m) + O(\Delta t^k)}{c_i(t^m) + O(\Delta t^{k-1})} = \underline{\underline{1}} + \underline{O(\Delta t^{k-1})}$$

Proof of DeC

$$\begin{aligned}
||\underline{\mathbf{c}}^{(k)}-\underline{\mathbf{c}}^*|| \leq & C_0||\mathcal{L}^1(\underline{\mathbf{c}}^{(k)})-\mathcal{L}^1(\underline{\mathbf{c}}^*)||= && (7)\\
= & C_0||\mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)})-\mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)},\underline{\mathbf{c}}^{(k)})-\mathcal{L}^1(\underline{\mathbf{c}}^*)+\mathcal{L}^2(\underline{\mathbf{c}}^*,\underline{\mathbf{c}}^*)|| \leq && (8)\\
\leq & C\Delta t||\underline{\mathbf{c}}^{(k-1)}-\underline{\mathbf{c}}^*|| && (9)
\end{aligned}$$

After K iterations

$$||\underline{\mathbf{c}}^{(K)}-\underline{\mathbf{c}}^*|| \leq C^K\Delta t^K||\underline{\mathbf{c}}^0-\underline{\mathbf{c}}^*||. \tag{10}$$

Outline

1. Production–Destruction system
2. Deferred Correction
3. Modified Patankar DeC (mPDeC)
4. Numerics

Linear test

$$\begin{aligned} c_1'(t) &= c_2(t) - 5c_1(t), & c_2'(t) &= 5c_1(t) - c_2(t), \\ c_1(0) &= c_1^0 = 0.9, & c_2(0) &= c_2^0 = 0.1\,. \end{aligned} \tag{11}$$

with

$$p_{1,2}(\mathbf{c}) = d_{2,1}(\mathbf{c}) = c_2, \quad p_{2,1}(\mathbf{c}) = d_{1,2}(\mathbf{c}) = 5c_1$$

and $p_{i,i}(\mathbf{c}) = d_{i,i}(\mathbf{c}) = 0$ for $i = 1, 2$.
Analytical solution is

$$c_1(t) = \frac{1}{6}\left(1 + \frac{22}{5}\exp(-6t)\right) \text{ and } c_2(t) = 1 - c_1(t). \tag{12}$$

Linear test

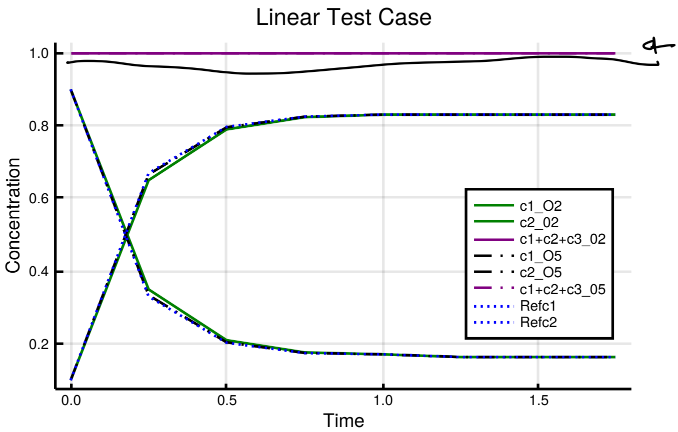


Figure: Second and fifth order methods together with the reference solution (12)

Linear test: Convergence

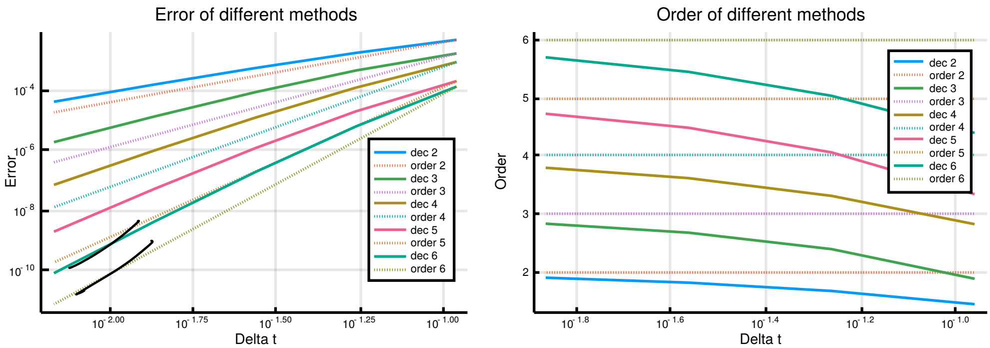


Figure: Second to sixth order error decay and slope of the errors

Nonlinear test

$$\begin{cases} c_1'(t) &= -\frac{c_1(t)c_2(t)}{c_1(t)+1}, \\ c_2'(t) &= \frac{c_1(t)c_2(t)}{c_1(t)+1} - 0.3c_2(t), \\ c_3'(t) &= 0.3c_2(t) \end{cases} \tag{13}$$

with initial condition $\mathbf{c}^0 = (9.98, 0.01, 0.01)^T$.
The PDS system in the matrix formulation can be expressed by

$$p_{2,1}(\mathbf{c}) = d_{1,2}(\mathbf{c}) = \frac{c_1(t)c_2(t)}{c_1(t)+1}, \quad p_{3,2}(\mathbf{c}) = d_{2,3}(\mathbf{c}) = 0.3c_2(t)$$

and $p_{i,j}(\mathbf{c}) = d_{i,j}(\mathbf{c}) = 0$ for all other combinations of i and j.

Nonlinear test

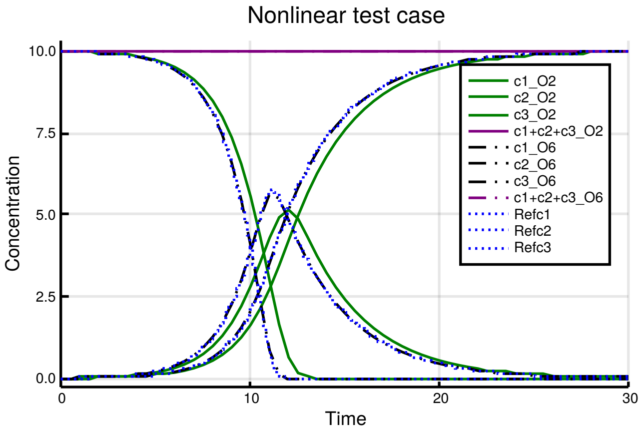


Figure: Second order and sixth order methods together with the reference solution (SSPRK104)

Nonlinear test: Convergence

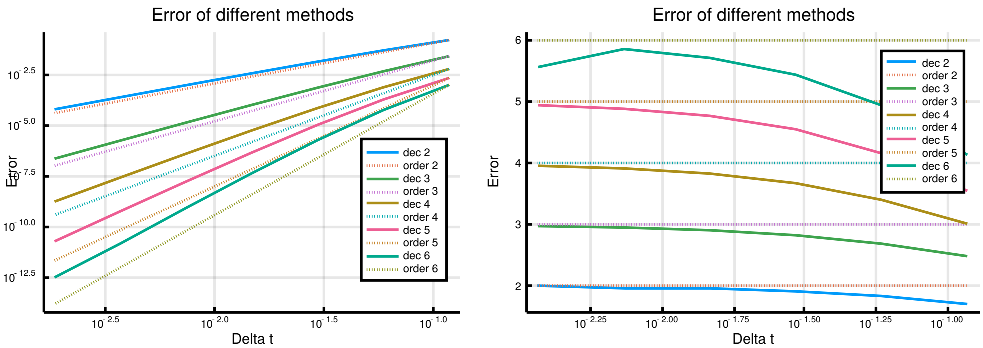
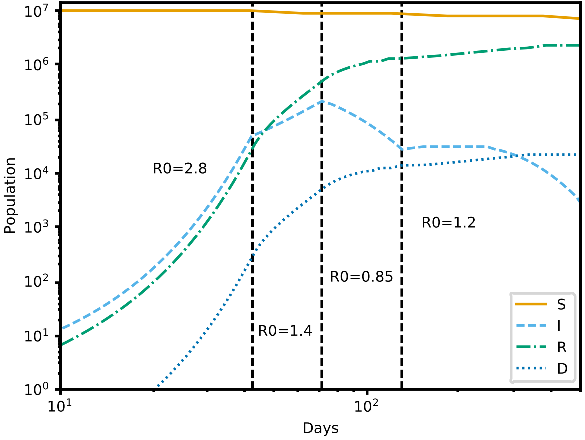


Figure: Second to sixth order error behaviors and slopes of the errors

SIRD

$$\begin{cases} d_t S = -\beta \frac{SI}{N} \\ d_t I = \beta \frac{SI}{N} - \gamma I - \delta I \\ d_t R = \gamma I \\ d_t D = \delta I \end{cases}$$

Solved with mPDeC5



Robertson test

$$\begin{aligned}
c_1'(t) &= 10^4 c_2(t)c_3(t) - 0.04c_1(t) \\
c_2'(t) &= 0.04c_1(t) - 10^4 c_2(t)c_3(t) - 3\cdot 10^7 c_2(t)^2 \\
c_3'(t) &= 3\cdot 10^7 c_2(t)^2
\end{aligned} \tag{14}$$

with initial conditions $\mathbf{c}^0 = (1,0,0)$.
The time interval of interest is $[10^{-6}, 10^{10}]$. The PDS for (14) reads

$$\begin{aligned}
&p_{1,2}(\mathbf{c}) = d_{2,1}(\mathbf{c}) = 10^4 c_2(t)c_3(t), \quad p_{2,1}(\mathbf{c}) = d_{1,2}(\mathbf{c}) = 0.04c_1(t), \\
&p_{3,2}(\mathbf{c}) = d_{2,3}(\mathbf{c}) = 3\cdot 10^7 c_2(t)
\end{aligned}$$

and zero for the other combinations.
We use exponential timesteps to better catch the behaviour of the solution $\Delta t^n = 2\cdot \Delta t^{n-1}$.

Robertson test

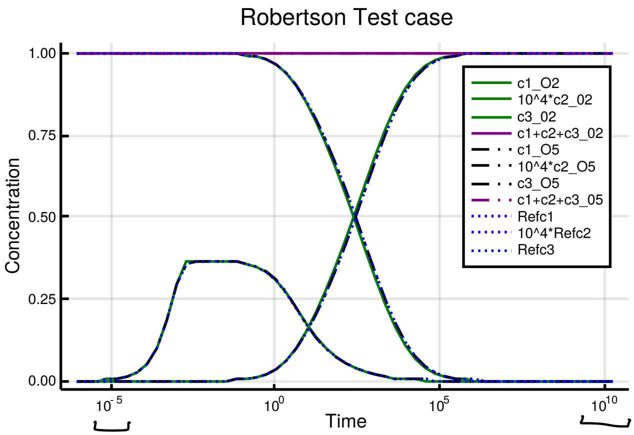


Figure: Second and fifth order solutions and references

Application to Shallow Water equations

$$\begin{cases} \partial_t h + \nabla \cdot (h\mathbf{u}) = 0 \\ \partial_t \mathbf{u} + \nabla \cdot (h\mathbf{u} \otimes \mathbf{u} + g\frac{h^2}{2}\mathrm{I}) = -gh\nabla b(\mathbf{x}) \end{cases}$$

- Slides
- Article post

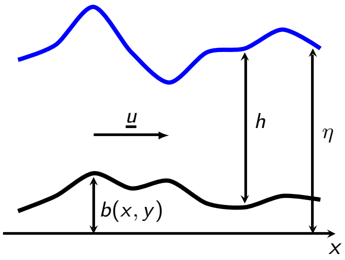


Figure: Shallow Water Equations: definition of the variables.

- MPDeC Code: If you want to check out the code, it's really easy ($\sim$ 150 lines), in Julia, on git. `https://git.math.uzh.ch/abgrall_group/deferred-correction-patankar-scheme`
- MPDeC Shallow Water code (Fortran) `https://github.com/accdavlo/sw-mpdec`

~~(COERCIVITY $\|\mathcal{L}^2(u)\| \geq C\|u\|$)~~

$$U = \begin{pmatrix} u^n \\ U^{n-1} \\ \vdots \\ U^1 \end{pmatrix} \quad \begin{pmatrix} u^n - U^0 + \Delta t\, \beta^n F(U^0) \\ \cdots \\ U^1 - U^0 + \Delta t\, \beta^1 F(U^0) \end{pmatrix}$$

"COERCIVITY": $\underbrace{\|\underbrace{\mathcal{L}^1(u)} - \underbrace{\mathcal{L}^1(v)}\|} \geq C \|u - v\|$

$$\mathcal{L}^2 = \left(u^n - U^0 + \Delta t \sum_r \Theta_r^n F(u^r) \right)$$

$\rightarrow$ $\mathcal{L}^1(u) = \left(\underline{u} - u^n - \Delta t\, \beta F(u^n) \right)$

$$\mathcal{L}^1(\underline{u}) - \mathcal{L}^1(\underline{v}) = \left(\underline{u} \cancel{- u^n} - \Delta t\, \cancel{F(u^n)} - \underline{v} \cancel{+ u^n} + \Delta t\, \cancel{F(u^n)} \right)$$

$$= (\underline{u} - \underline{v})$$

$C = 1$ $\quad \|\mathcal{L}^1(\underline{u}) - \mathcal{L}^1(\underline{v})\| = \|\underline{u} - \underline{v}\|$

$C = 1$

$F(u) = 0$ NOT A COERCIVE

$u \pm \sqrt{g \cdot h}$