

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

Philipp Öffner & Davide Torlo

Universität Zürich
Institut für Mathematik

8th July 2020

Icosahom MS

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

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

Production–Destruction system

Consider **production-destruction** systems (PDS)

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

where

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

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)

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

where

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

Property 1: **Conservation**

$$\begin{aligned} \sum_{i=1}^I c_i(0) &= \sum_{i=1}^I c_i(t), \quad \forall t \geq 0 \\ \iff p_{i,j}(\mathbf{c}) &= d_{j,i}(\mathbf{c}), \quad \forall i, j \in I, \quad \forall \mathbf{c} \in \mathbb{R}^{+,I}. \end{aligned}$$

Consider **production-destruction** systems (PDS)

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

where

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

Property 2: **Positivity**

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

Consider **production-destruction** systems (PDS)

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

where

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

Goal:

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

Explicit Euler

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

Implicit Euler

- $\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t (\mathbf{P}(\mathbf{c}^{n+1}) - \mathbf{D}(\mathbf{c}^{n+1}))$
- 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

Modified Patankar (mP)

Burchard, Deleersnijder & Meister

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

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

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

- Conservative
- First order
- Positive
- Linear system at each timestep
- Extension to RK2 and RK3 (Burchard, Deleersnijder, Meister, Kopecz)
- Extension to PDEs (Huang, Zhao, Shu)

- 1 Production–Destruction system
- 2 Deferred Correction**
- 3 Modified Patankar DeC (mPDeC)
- 4 Numerics
- 5 Outlook

Deferred Correction discretization

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

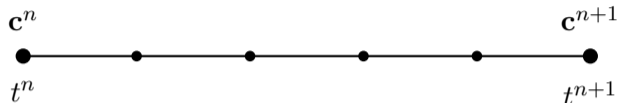


Figure: Subtimeintervals

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 = $M + 1$.

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

Deferred Correction discretization

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

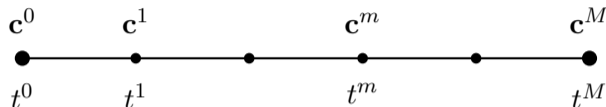


Figure: Subtimeintervals

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 = $M + 1$.

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

\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

\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}^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?

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

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

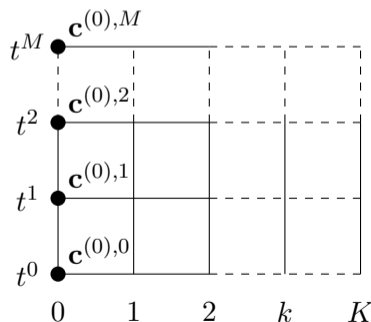
$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$

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.

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



Deferred Correction

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

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

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

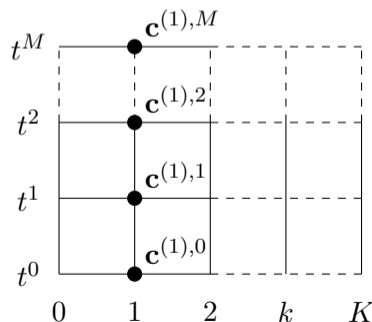
$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$

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.

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



Deferred Correction

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

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

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

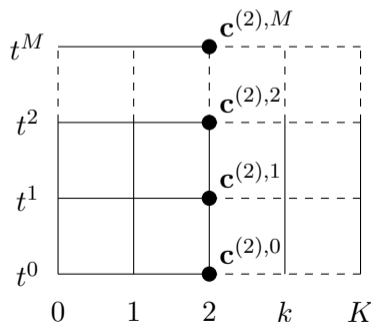
$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$

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.

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



Deferred Correction

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

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

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

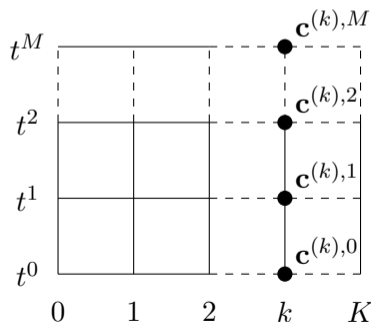
$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$

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.

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



Deferred Correction

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

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

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

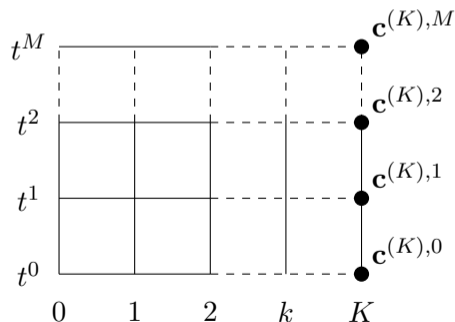
$$\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) = \mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}) \text{ with } k = 1, \dots, K.$$

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.

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



If we write explicitly the DeC step we see that

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

If we write explicitly the DeC step we see that

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

- 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

- 1 Production–Destruction system
- 2 Deferred Correction
- 3 Modified Patankar DeC (mPDeC)**
- 4 Numerics
- 5 Outlook

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

$$\mathcal{L}_i^2(\mathbf{c}^{0,(k-1)}, \dots, \mathbf{c}^{M,(k-1)}) = \mathcal{L}_i^2(\underline{\mathbf{c}}^{(k-1)}) := \begin{pmatrix} c_i^{M,(k-1)} - c_i^{0,(k-1)} - \Delta t \sum_{r=0}^M \theta_r^M \sum_{j=1}^I \left(p_{i,j}(\mathbf{c}^{r,(k-1)}) - d_{i,j}(\mathbf{c}^{r,(k-1)}) \right) \\ \vdots \\ c_i^{1,(k-1)} - c_i^{0,(k-1)} - \Delta t \sum_{r=0}^M \theta_r^1 \sum_{j=1}^I \left(p_{i,j}(\mathbf{c}^{r,(k-1)}) - d_{i,j}(\mathbf{c}^{r,(k-1)}) \right) \end{pmatrix},$$

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

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

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

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

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

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

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

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

- 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 $\forall k, m$

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

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 $\forall k, m$

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

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 $\forall k, m$

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

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

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

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

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 $\mathcal{L}^1 - \mathcal{L}^2$:
 $\|\mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}, \underline{\mathbf{c}}^{(k)}) - \mathcal{L}^1(\underline{\mathbf{c}}^*) + \mathcal{L}^2(\underline{\mathbf{c}}^*, \underline{\mathbf{c}}^*)\| \leq C_L \Delta t \|\underline{\mathbf{c}}^{(k-1)} - \underline{\mathbf{c}}^*\|.$

Intermediate steps for Lipschitz continuity

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

$$\|\underline{\mathbf{c}}^{(k)} - \underline{\mathbf{c}}^*\| \leq C_0 \|\mathcal{L}^1(\underline{\mathbf{c}}^{(k)}) - \mathcal{L}^1(\underline{\mathbf{c}}^*)\| = \tag{7}$$

$$= C_0 \|\mathcal{L}^1(\underline{\mathbf{c}}^{(k-1)}) - \mathcal{L}^2(\underline{\mathbf{c}}^{(k-1)}, \underline{\mathbf{c}}^{(k)}) - \mathcal{L}^1(\underline{\mathbf{c}}^*) + \mathcal{L}^2(\underline{\mathbf{c}}^*, \underline{\mathbf{c}}^*)\| \leq \tag{8}$$

$$\leq C \Delta t \|\underline{\mathbf{c}}^{(k-1)} - \underline{\mathbf{c}}^*\| \tag{9}$$

After K iterations

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

- 1 Production–Destruction system
- 2 Deferred Correction
- 3 Modified Patankar DeC (mPDeC)
- 4 Numerics**
- 5 Outlook

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

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

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

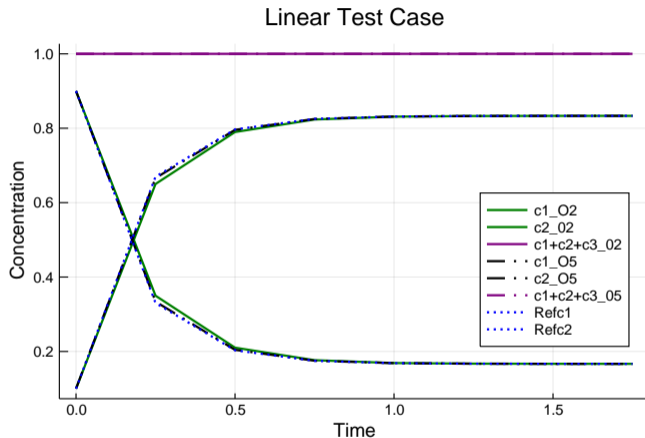


Figure: Second and fifth order methods together with the reference solution (12)

Linear test: Convergence

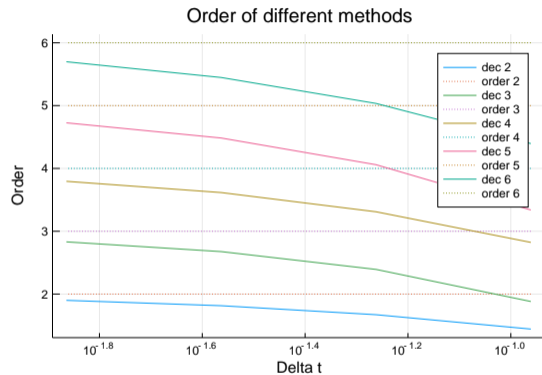
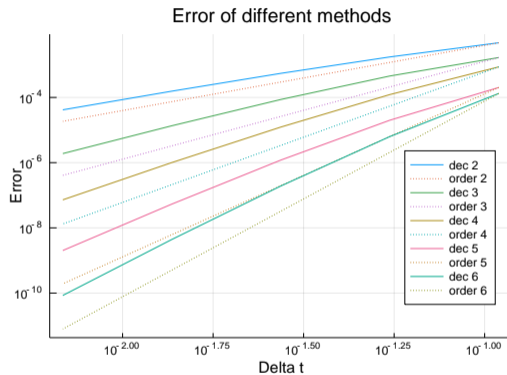


Figure: Second to sixth order error decay and slope of the errors

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

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

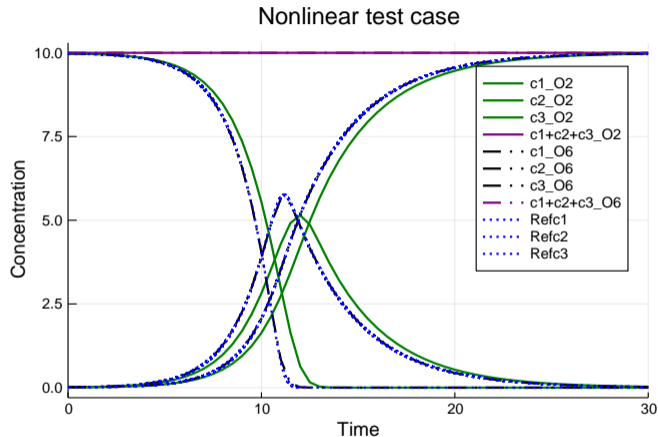


Figure: Second order and sixth order methods together with the reference solution (SSPRK104)

Nonlinear test: Convergence

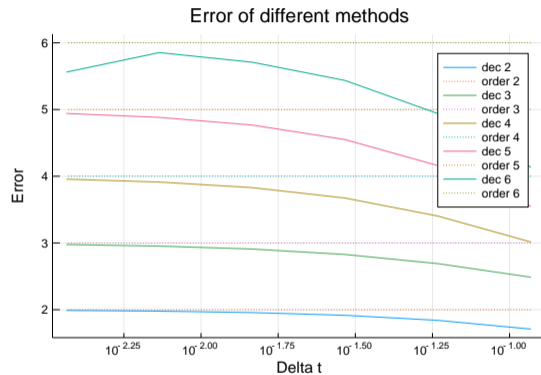
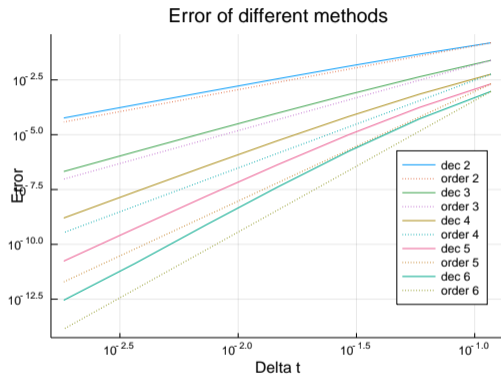
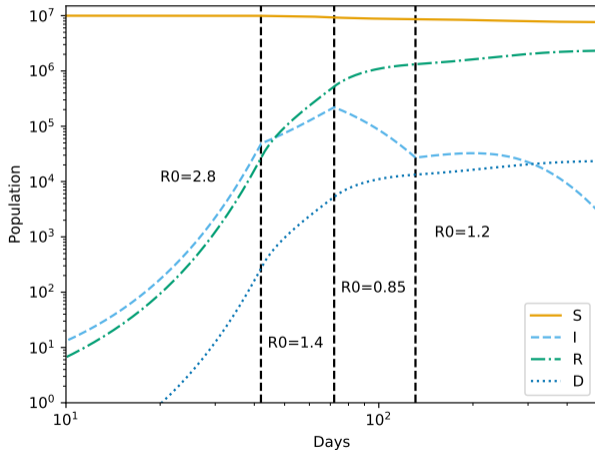


Figure: Second to sixth order error behaviors and slopes of the errors

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

Solved with mPDeC5



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

with initial conditions $\mathbf{c}^0 = (1, 0, 0)$.

The time interval of interest is $[10^{-6}, 10^{10}]$. The PDS for (14) reads

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

and zero for the other combinations.

We use exponential timesteps to better catch the behaviour of the solution

$$\Delta t^n = 2 \cdot \Delta t^{n-1}.$$

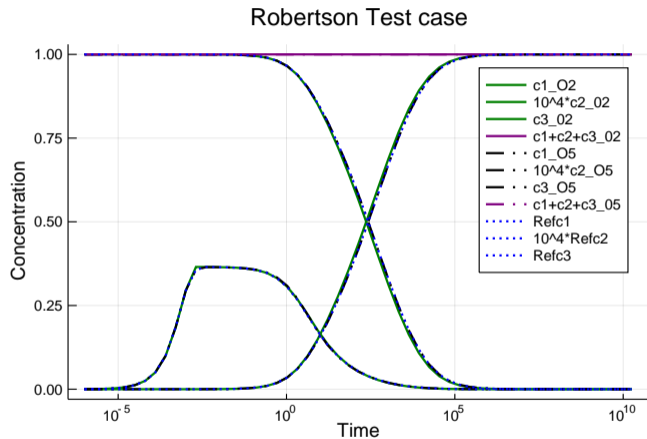


Figure: Second and fifth order solutions and references

- 1 Production–Destruction system
- 2 Deferred Correction
- 3 Modified Patankar DeC (mPDeC)
- 4 Numerics
- 5 Outlook**

Extensions:

- Shu-Osher formulation of RK
- Stiff source terms
- Stability analysis
- Shallow water equation
- Euler equations

If you want to check out the code, it's really easy (~ 150 lines), in Julia, on git.

```
https://git.math.uzh.ch/abgrall_group/  
deferred-correction-patankar-scheme
```

Thank you!