On modified Patankar schemes and oscillations: towards new stability definitions

Davide Torlo[1] & Philipp Öffner[2] & Hendrik Ranocha[3]

12th July 2021

Icosahom 2020

[1]INRIA Bordeaux - Sud Ouest, France
[2]Johannes Gutenberg-University Mainz, Germany
[3]Fachbereich Mathematik und Informatik der Universität Münster, Germany

Outline

1. Production–Destruction System
2. (Modified) Patankar Methods
3. Properties and Troubles
4. Oscillation Study
5. Inconsistency Study
6. Simulations
7. Conclusion

Outline

1. Production–Destruction System
2. (Modified) Patankar Methods
3. Properties and Troubles
4. Oscillation Study
5. Inconsistency Study
6. Simulations
7. Conclusion

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.
Simplest example: 2 x 2 linear system

$$\begin{cases} d_t u_1 = u_2 - u_1 \\ d_t u_2 = u_1 - u_2 \end{cases}$$

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

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

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 of the method design:

- One step method
- **Unconditionally positive** (for any Δt)
- **Unconditionally conservative**
- (High order accurate)

Outline

1. Production–Destruction System
2. (Modified) Patankar Methods
3. Properties and Troubles
4. Oscillation Study
5. Inconsistency Study
6. Simulations
7. Conclusion

Patankar Method

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

Patankar and Modified Patankar Methods

Modified Patankar (mP)
Burchard, Deleersnijder & Meister, APNUM 47.1 (2003)

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

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

Extensions of Modified Patankar Schemes

Similar ideas

- MPRK(2,2,α): Kopecz, Meister in APNUM 123 (2018)
- MPRK(4,3,α,β): Kopecz, Meister in BIT 58.3 (2018)
- MPRKSO(2,2,α,β): Huang, Shu in JSC 78.3 (2019)
- MPRKSO(4,3): Huang, Zhao, Shu in JSC 79.2 (2019)
- mPDeC: Öffner, Torlo in APNUM 153 (2020)
- SI-RK2, SI-RK3: Chertock, Cui, Kurganov, Wu in SIAM J. Numer. Anal. (2015). Patankar schemes, they are weighting only the destruction term $\Longrightarrow$ not conservative

Properites:

- **Unconditionally positive** (for any Δt)
- Unconditionally conservative (except SI-RK)
- High order

Extensions of Modified Patankar Schemes: MPRK(3,2)

A novel second-order modified Patankar–Runge–Kutta with three stages based on SSPRK(3,3) is

$$
\begin{aligned}
y^1 &= u^n, \\
y_i^2 &= u^n + \Delta t \sum_j \left(p_{ij}(y^1) \frac{y_j^2}{y_j^1} - d_{ij}(y^1) \frac{y_i^2}{y_i^1} \right), \\
y^3 &= u^n + \Delta t \sum_j \left(\frac{p_{ij}(y^1) + p_{ij}(y^2)}{4} \frac{y_j^3}{y_j^2} - \frac{d_{ij}(y^1) + d_{ij}(y^2)}{4} \frac{y_i^3}{y_i^2} \right), \\
u^{n+1} &= u^n + \Delta t \sum_j \left(\frac{p_{ij}(y^1) + p_{ij}(y^2) + 4p_{ij}(y^3)}{6} \frac{u_j^{n+1}}{y_j^2} \right. \\
&\qquad \left. - \frac{d_{ij}(y^1) + d_{ij}(y^2) + 4d_{ij}(y^3)}{6} \frac{u_i^{n+1}}{y_i^2} \right).
\end{aligned}
\tag{MPRK(3,2)}
$$

Outline

1. Production–Destruction System
2. (Modified) Patankar Methods
3. Properties and Troubles
4. Oscillation Study
5. Inconsistency Study
6. Simulations
7. Conclusion

Steady state and wrong steady state

Steady state preservation

All the Patankar schemes are steady state preserving.
A steady state is u^∞ such that $\sum_j p_{ij}(u^\infty) - d_{ij}(u^\infty) = 0$ and $d_t u^\infty = 0$.
A scheme *preserves* the steady state if, given u^n steady state, then $u^{n+1} = u^n$.

Wrong steady states

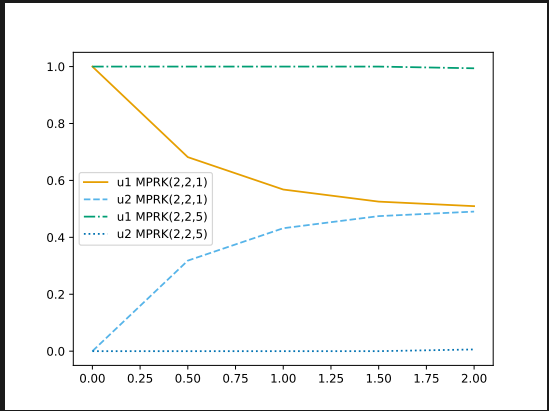
Some Patankar schemes preserve also some states which are not analytically steady.
This happen when the Lipschitz constant of the schemes tends to infinity, i.e., when some quantities $u_i \to 0$.

Steady state and wrong steady state

Example:

$$
\begin{cases} d_t u_1 = u_2 - u_1 \\ d_t u_2 = u_1 - u_2 \end{cases}
$$
$$
u^0 = (1, 10^{-15})^T,
$$
$$
u^\infty = (0.5, 0.5)^T.
$$

MPRK(2,2,1) catches the right behavior.
MPRK(2,2,5) get stuck at the 0 state.



Oscillations

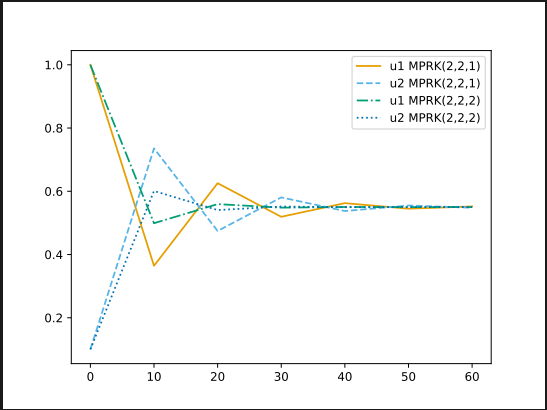
Δt large $\Longrightarrow$ oscillations

In the previous example choosing $\Delta t = 10$ we have that

- MPRK(2,2,1) oscillates around the steady state.
- MPRK(2,2,2) oscillates around the steady state.

Questions

- Unexpected from an unconditionally positive scheme
- Implicit Euler is provably not oscillating
- What is actually the problem?



Oscillations

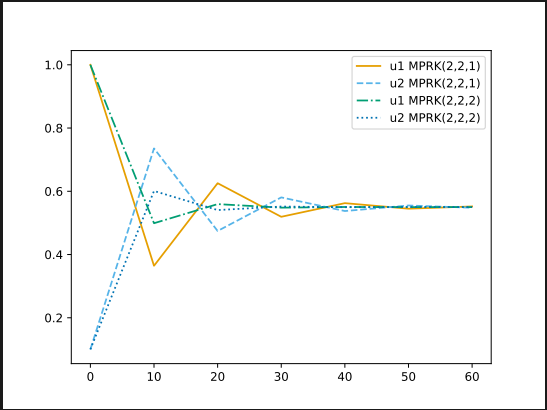
Δt large $\Longrightarrow$ oscillations

In the previous example choosing $\Delta t = 10$ we have that

- MPRK(2,2,1) oscillates around the steady state.
- MPRK(2,2,2) oscillates around the steady state.

Questions

- Unexpected from an unconditionally positive scheme
- Implicit Euler is provably not oscillating
- What is actually the problem?



Outline

1. Production–Destruction System
2. (Modified) Patankar Methods
3. Properties and Troubles
4. Oscillation Study
5. Inconsistency Study
6. Simulations
7. Conclusion

Oscillation definition and modus operandi

Oscillation

Analytically

$$|u_i(t) - u_i(0)| \leq |u_i(0) - u_i^{\infty}|. \quad (4)$$

A scheme **oscillates** when the previous inequality is not verified, i.e., $\exists t$:

$$|u_i(t) - u_i(0)| > |u_i(0) - u_i^{\infty}|. \quad (5)$$

How to detect when a scheme shows oscillations?

The schemes depends on

- Problem (and its parameters)
- IC
- Δt
- Parameters (α, β)

Analytical study or numerical study?

Oscillation definition and modus operandi

Problem: general 2×2 linear system

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} -\theta & (1-\theta) \\ \theta & -(1-\theta) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad (6)$$

$$\begin{pmatrix} u_1^0 \\ u_2^0 \end{pmatrix} = \begin{pmatrix} 1-\varepsilon \\ \varepsilon \end{pmatrix}, \quad (7)$$

with $0 < \theta, \varepsilon < 1$.
Other values obtainable rescaling time and variables.

Analytical study

- Doable for few methods
- Fix the problem ($\theta = 0.5$)

Numerical study

- Discretize $\theta \in [0, 1]$
- Discretize $\varepsilon \in [0, 0.5]$ (symmetry)
- Discretize $\Delta t \in [2^{-5}, 2^5]$
- Discretize schemes parameters (if any)
- Check only first timestep

MPRK(2,2,1)=MPDeC2

Theorem (Time restriction for MPRK(2,2,1) for 2×2 linear systems)

Consider the system (6) *with the initial conditions* (7)*. The MPRK(2,2,1) is not oscillating if the timestep restriction* $\Delta t \leq 2$ *holds, for any IC* $0 < \varepsilon < 1$ *and any system* $0 \leq \theta \leq 1$*.*

Sketch of the proof.

- $\theta = 0$ and $\theta = 1$ are trivial, because positive preserving and conservative.
- $\varepsilon = \theta$ trivial, IC are steady state.
- We check for $u_1^1 =$ ratio of polynomials 1st degree in Δt and θ and 2nd degree in ε.
- `Mathematica` ... Stability (no oscillations) if $\Delta t \leq z$, where z is the only positive zero of the polynomial $p_{\varepsilon,\theta}(x)$

$$p_{\varepsilon,\theta}(x) = x^3 - x^2 - 2\left(\frac{\varepsilon}{\theta} + \frac{1-\varepsilon}{1-\theta}\right)x - 2\frac{\varepsilon(1-\varepsilon)}{\theta(1-\theta)}.$$

□

MPRK(2,2,1)=MPDeC2

Theorem (Time restriction for MPRK(2,2,1) for 2×2 linear systems)

Consider the system (6) *with the initial conditions* (7)*. The MPRK(2,2,1) is not oscillating if the timestep restriction* $\Delta t \leq 2$ *holds, for any IC* $0 < \varepsilon < 1$ *and any system* $0 \leq \theta \leq 1$*.*

Sketch of the proof.

- $$p_{\varepsilon,\theta}(x) = x^3 - x^2 - 2\left(\frac{\varepsilon}{\theta} + \frac{1-\varepsilon}{1-\theta}\right)x - 2\frac{\varepsilon(1-\varepsilon)}{\theta(1-\theta)}.$$
- Algebraic estimations on the zeros $y \leq w \leq z$ using $\varepsilon, \theta \in (0,1)$
- ywz positive, $yz + wz + yw$ negative, $y \leq w < 0 < z$
- $w + y < -\frac{2}{z}, \quad 0 = z + y + w - 1 < z - \frac{2}{z} - 1, \quad 0 < z^2 - z - 2, \Longrightarrow z > 2.$
- $\Delta t \leq 2 \Longrightarrow$ No oscillations.

□

Other schemes analytically

- MPRK(3,2) too many stages $\Longrightarrow$ Polynomial of degree 6 in $\Delta t, \theta, \varepsilon$.
- MPRK(2,2,α) too many parameters.
- MPDeC3 on too many stages.
- MPRK(4,3,α, β), MPRKSO(2,2,α, β), MPRKSO(4,3) too many stages and parameters.

If too many parameters $\Longrightarrow$ we can remove parameters from the problem (later).
Numerical study!

Oscillations: find time step restriction numerically

Numerical study

- Discretize $\theta \in [0, 1]$
- Discretize $\varepsilon \in [0, 0.5]$ (symmetry)
- Discretize $\Delta t \in [2^{-5}, 2^{5}]$
- Discretize schemes parameters (if any)
- Check only first timestep

Goal: find the largest Δt such that the scheme is stable

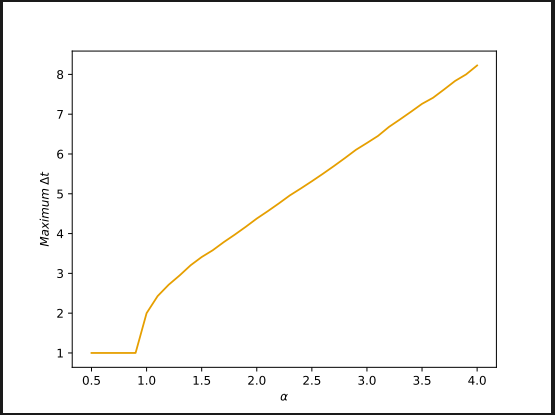


Figure: Δt restriction for MPRK(2,2,α)

Oscillations: find time step restriction numerically

Numerical study

- Discretize $\theta \in [0, 1]$
- Discretize $\varepsilon \in [0, 0.5]$ (symmetry)
- Discretize $\Delta t \in [2^{-5}, 2^5]$
- Discretize schemes parameters (if any)
- Check only first timestep

Goal: find the largest Δt such that the scheme is stable

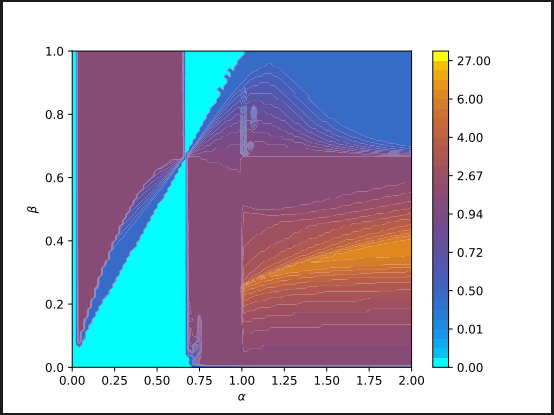


Figure: Δt restriction for MPRK(4,3,α, β)

Oscillations: find time step restriction numerically

Numerical study

- Discretize $\theta \in [0, 1]$
- Discretize $\varepsilon \in [0, 0.5]$ (symmetry)
- Discretize $\Delta t \in [2^{-5}, 2^{5}]$
- Discretize schemes parameters (if any)
- Check only first timestep

Goal: find the largest Δt such that the scheme is stable

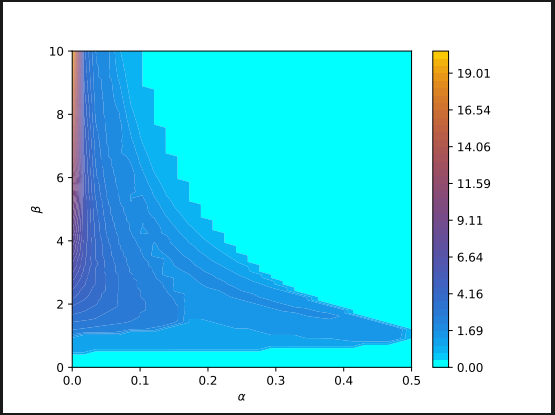


Figure: Δt restriction for MPRKSO(2,2,α, β)

Oscillations: find time step restriction numerically

Numerical study

- Discretize $\theta \in [0, 1]$
- Discretize $\varepsilon \in [0, 0.5]$ (symmetry)
- Discretize $\Delta t \in [2^{-5}, 2^{5}]$
- Discretize schemes parameters (if any)
- Check only first timestep

Goal: find the largest Δt such that the scheme is stable

Method	Δt bound
MPRKSO(4,3)	1.31
SI-RK2	1.41
SI-RK3	1.27
MPRK(3,2)	16.56

Oscillations: find time step restriction numerically

Numerical study

- Discretize $\theta \in [0, 1]$
- Discretize $\varepsilon \in [0, 0.5]$ (symmetry)
- Discretize $\Delta t \in [2^{-5}, 2^{5}]$
- Discretize schemes parameters (if any)
- Check only first timestep

Goal: find the largest Δt such that the scheme is stable

Equispaced points

Order	Δt bound
2	2.0
3	1.19
4	1.11
5	1.07
6	1.04
7	1.04
8	1.37
9	6.96
10	1.0
11	16.0
12	1.0
13	40.79
14	1.07
15	27.85
16	1.80

Gauss-Lobatto points

Order	Δt bound
2	2.0
3	1.19
4	1.07
5	1.04
6	1.0
7	1.0
8	1.0
9	1.0
10	1.0
11	1.0
12	1.0
13	1.0
14	1.0
15	1.0
16	1.0

Figure: Δt bound for mPDeC with equispaced and Gauss–Lobatto points.

Outline

1. Production–Destruction System
2. (Modified) Patankar Methods
3. Properties and Troubles
4. Oscillation Study
5. Inconsistency Study
6. Simulations
7. Conclusion

Wrong steady state: inconsistency

Problem: general 2×2 linear system

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \qquad (8)$$

$$\begin{pmatrix} u_1^0 \\ u_2^0 \end{pmatrix} = \begin{pmatrix} 1-\varepsilon \\ \varepsilon \end{pmatrix}, \qquad (9)$$

with $0 < \theta, \varepsilon < 1$.
Other values obtainable rescaling time and variables.

When $u_2 = \varepsilon \approx 0$ some schemes do not move from this wrong steady state. Analytically, we observe that

$$\lim_{\varepsilon \to 0} \lim_{\Delta t \to 0} u_2^1 \neq \lim_{\Delta t \to 0} \lim_{\varepsilon \to 0} u_2^1. \qquad (10)$$

When $\lim_{\varepsilon \to 0} u_2^1 = \mathcal{O}(\Delta t)$ we are inconsistent with the problem.
We fix the system at $\theta = 0.5$ as we observe no differences with other values in this study.

Wrong steady state: inconsistency

MPRK(2,2,α)

$$\lim_{\varepsilon\to 0} u_2^1(\varepsilon) = \begin{cases} 0 & \alpha > 1 \\ (1-\varepsilon)\frac{2\Delta t+3\Delta t^2}{2+5\Delta t+4\Delta t^2} & \alpha = 1 \\ (1-\varepsilon)\frac{2\Delta t+(4\alpha-1)\Delta t^2}{2+(2+4\alpha)\Delta t+(4\alpha-1)\Delta t^2} & \alpha < 1 \end{cases} \tag{11}$$

MPRK(2,2,$\alpha = 1$) order 2

$$\lim_{\varepsilon\to 0}\lim_{\Delta t\to 0} u_1^1 - u_1(\Delta t) = \left(\frac{2}{3} - \frac{4\varepsilon}{3} + \mathcal{O}(\varepsilon^3)\right)\Delta t^3 + \mathcal{O}(\Delta t^4)$$

$$\lim_{\Delta t\to 0}\lim_{\varepsilon\to 0} u_1^1 - u_1(\Delta t) = \left(\frac{\Delta t^3}{6} + \mathcal{O}(\Delta t^4)\right) + \left(\frac{\Delta t^2}{2} - \frac{11\Delta t^3}{6} + \mathcal{O}(\Delta t^4)\right)\varepsilon + \left(-\frac{\Delta t}{2} + \frac{\Delta t^2}{4} + \frac{3\Delta t^3}{4} + \mathcal{O}(\Delta t^4)\right)\varepsilon^2 + \mathcal{O}(\varepsilon^3).$$

Wrong steady state: inconsistency

MPRK(2,2,α)

$$\lim_{\varepsilon\to 0} u_2^1(\varepsilon) = \begin{cases} 0 & \alpha > 1 \\ (1-\varepsilon)\frac{2\Delta t+3\Delta t^2}{2+5\Delta t+4\Delta t^2} & \alpha = 1 \\ (1-\varepsilon)\frac{2\Delta t+(4\alpha-1)\Delta t^2}{2+(2+4\alpha)\Delta t+(4\alpha-1)\Delta t^2} & \alpha < 1 \end{cases} \tag{11}$$

MPRK(2,2,$\alpha = 0.5$) order 1

$$\lim_{\varepsilon\to 0}\lim_{\Delta t\to 0} u_1^1 - u_1(\Delta t) = \left(-\frac{1}{4\varepsilon}+\frac{5}{12}-\frac{5\varepsilon}{12}+\frac{\varepsilon^2}{4}+\mathcal{O}(\varepsilon^3)\right)\Delta t^3+\mathcal{O}(\Delta t^4)$$

$$\lim_{\Delta t\to 0}\lim_{\varepsilon\to 0} u_1^1 - u_1(\Delta t) = \left(-\frac{\Delta t^2}{2}+\frac{11\Delta t^3}{12}+\mathcal{O}(\Delta t^4)\right)+\left(\Delta t+\frac{\Delta t^2}{2}-\frac{19\Delta t^3}{6}+\mathcal{O}(\Delta t^4)\right)\varepsilon+$$
$$\left(-2-4\Delta t+4\Delta t^2+\frac{13\Delta t^3}{4}+\mathcal{O}(\Delta t^4)\right)\varepsilon^2+\mathcal{O}(\varepsilon^3).$$

Wrong steady state: inconsistency

MPRK(2,2,α)

$$\lim_{\varepsilon\to 0} u_2^1(\varepsilon) = \begin{cases} 0 & \alpha > 1 \\ (1-\varepsilon)\frac{2\Delta t+3\Delta t^2}{2+5\Delta t+4\Delta t^2} & \alpha = 1 \\ (1-\varepsilon)\frac{2\Delta t+(4\alpha-1)\Delta t^2}{2+(2+4\alpha)\Delta t+(4\alpha-1)\Delta t^2} & \alpha < 1 \end{cases} \tag{11}$$

MPRK(2,2,$\alpha = 2$) order 0

$$\lim_{\varepsilon\to 0}\lim_{\Delta t\to 0} u_1^1 - u_1(\Delta t) = \left(\frac{1}{2\varepsilon} + \frac{7}{6} - \frac{7\varepsilon}{6} + \frac{\varepsilon^2}{4} + \mathcal{O}(\varepsilon^3)\right)\Delta t^3 + \mathcal{O}(\Delta t^4)$$

$$\lim_{\Delta t\to 0}\lim_{\varepsilon\to 0} u_1^1 - u_1(\Delta t) = \left(\Delta t - \Delta t^2 + \frac{2\Delta t^3}{3} + \mathcal{O}(\Delta t^4)\right) + \mathcal{O}(\varepsilon) + \mathcal{O}(\Delta t).$$

Wrong steady state: inconsistency and order reduction

MPDeC3 order 1

$$\lim_{\varepsilon\to 0}\lim_{\Delta t\to 0} u_1^1 - u_1(\Delta t) = \left(-\frac{1}{864\varepsilon^2} - \frac{5}{72\varepsilon} + \frac{1789}{864} - \frac{1697\varepsilon}{432} + \frac{7\varepsilon^2}{96} + \mathcal{O}(\varepsilon^3)\right)\Delta t^4 + \mathcal{O}(\Delta t^5)$$

$$\lim_{\Delta t\to 0}\lim_{\varepsilon\to 0} u_1^1 - u_1(\Delta t) = \left(-\frac{2\Delta t^2}{3} + \mathcal{O}(\Delta t^3)\right) + \left(224\Delta t + \mathcal{O}(\Delta t^2)\right)\varepsilon - 74880\varepsilon^2 + \mathcal{O}(\Delta t^3).$$

MPRK(3,2) order 2

$$\lim_{\varepsilon\to 0}\lim_{\Delta t\to 0} u_1^{n=1} - u_1(\Delta t) = \left(2 - 4\varepsilon + \mathcal{O}(\varepsilon^3)\right)\Delta t^3 + \mathcal{O}(\Delta t^4)$$

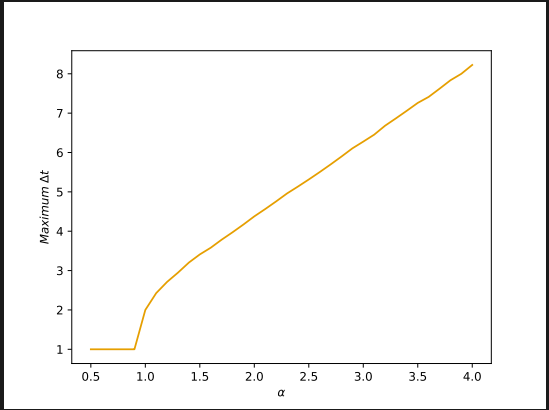
$$\lim_{\Delta t\to 0}\lim_{\varepsilon\to 0} u_1^{n=1} - u_1(\Delta t) = \left(\frac{4\Delta t^3}{3} + \mathcal{O}(\Delta t^4)\right) + \left(\frac{2\Delta t^2}{3} - \frac{71\Delta t^3}{12} + \mathcal{O}(\Delta t^4)\right)\varepsilon + \left(-\frac{2\Delta t}{3} + \frac{35\Delta t^2}{24} + \frac{139\Delta t^3}{144} + \mathcal{O}(\Delta t^4)\right)\varepsilon^2 + \mathcal{O}(\varepsilon^3).$$

Wrong steady state: numerical study

Instability Numerically

- $\varepsilon = 10^{-300}$
- $\theta = 0.5$
- $\Delta t = 1$
- The exact solution $u_2(t = 1) \approx 0.43$
- If $u_2^1 \gg 0$ consistent, else inconsistent.

Inconsistency and oscillations: MPRK(2,2, α)



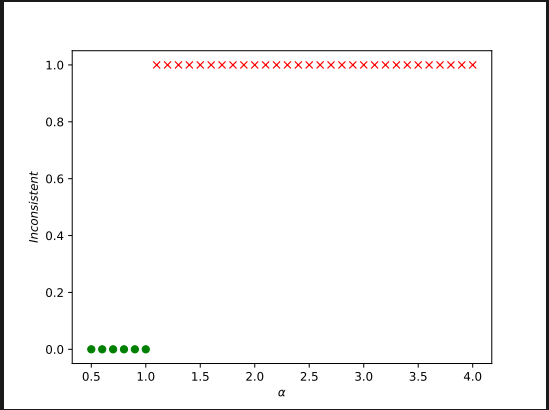
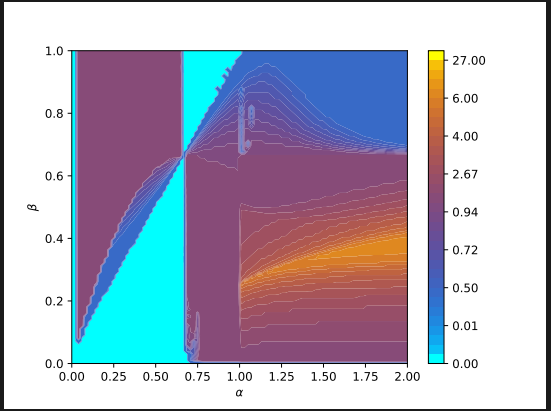


Figure: Δt restriction (left) vs inconsistency (right) for MPRK(2,2,α)

Inconsistency and oscillations: MPRK(4,3, α, β)



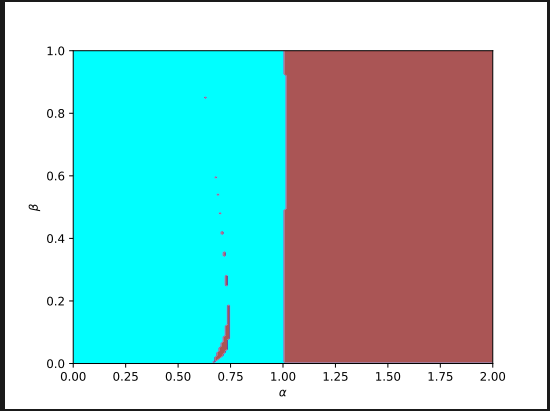
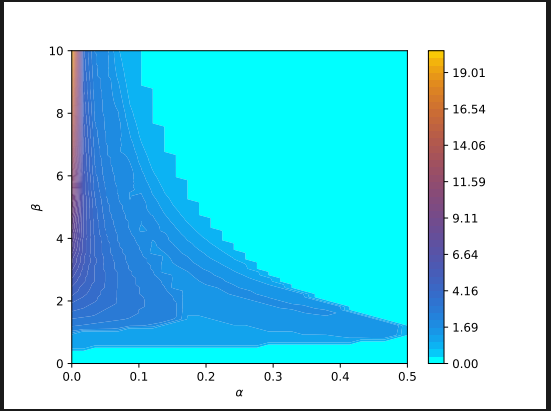


Figure: Δt restriction (left) vs inconsistency (right) for MPRK(4,3, α, β)

Inconsistency and oscillations: MPRKSO(2,2, α, β)



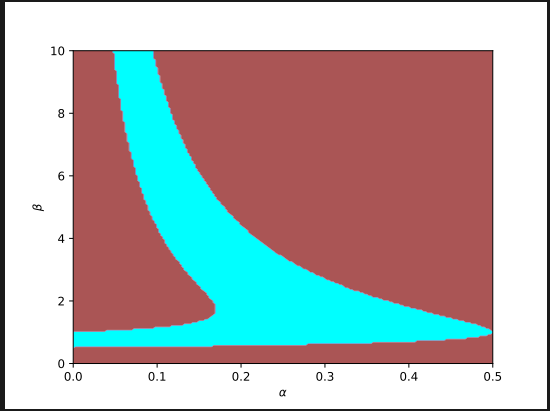


Figure: Δt restriction (left) vs inconsistency (right) for MPRK(4,3, α, β)

Inconsistency and oscillations: other methods

Scheme	Δt bound	Consistent
MPRKSO(4,3)	1.31	Yes
MPRK(3,2)	16.5	Yes
SI-RK2	1.41	Yes
SI-RK3	1.27	Yes

Equispaced points	
2-8,10 consistent	
Order	Δt bound
2	2.0
3	1.19
4	1.11
5	1.07
6	1.04
7	1.04
8	1.37
9	6.96
10	1.0
11	16.0
12	1.0
13	40.79
14	1.07
15	27.85
16	1.80

Gauss-Lobatto points	
All consistent	
Order	Δt bound
2	2.0
3	1.19
4	1.07
5	1.04
6	1.0
7	1.0
8	1.0
9	1.0
10	1.0
11	1.0
12	1.0
13	1.0
14	1.0
15	1.0
16	1.0

Outline

1. Production–Destruction System
2. (Modified) Patankar Methods
3. Properties and Troubles
4. Oscillation Study
5. Inconsistency Study
6. Simulations
7. Conclusion

Robertson Problem

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

$\mathbf{c}^0 = (1, 10^{-180}, 10^{-180})$
$t \in [10^{-6}, 10^{10}]$. The PDS for (30) reads

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

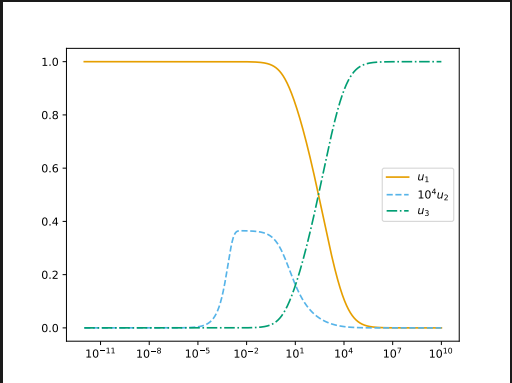


Figure: Robertson Problem

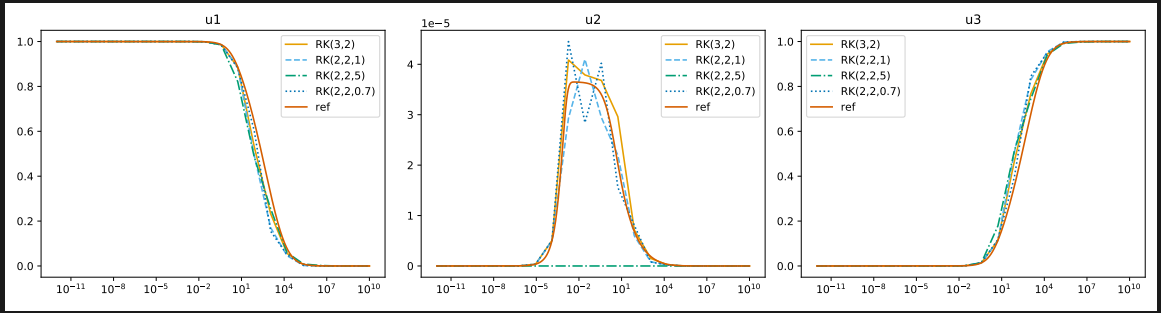
We use exponential timesteps to better catch the behaviour of the solution $\Delta t^n = 2 \cdot \Delta t^{n-1}$.

Robertson Problem

Test with $N = 20$ timesteps

MPRK(3,2) high Δt bound, consistent,
MPRK(2,2,1) $\Delta t = 2$ bound, consistent,

MPRK(2,2,5) $\Delta t \approx 10$ bound, inconsistent,
MPRK(2,2,0.7) $\Delta t = 1$ bound, consistent.

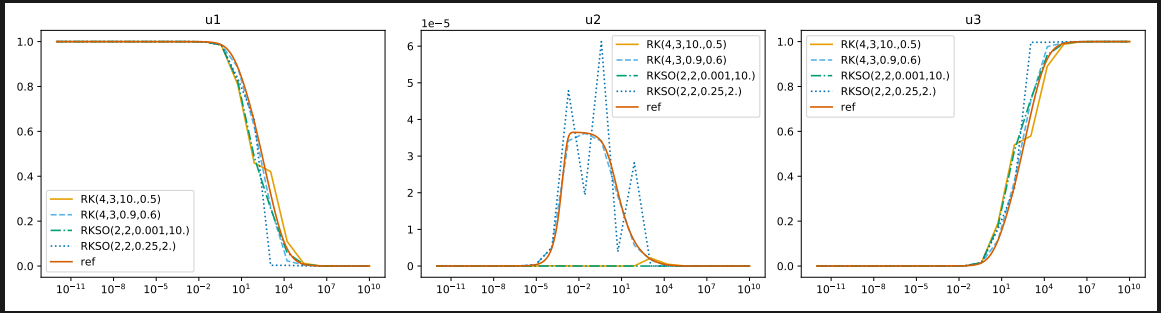


Robertson Problem

Test with $N = 20$ timesteps

MPRK(4,3,10,0.5) high Δt bound, inconsistent,
MPRK(4,3,0.9,0.6) $\Delta t \approx 1$ bound, consistent,

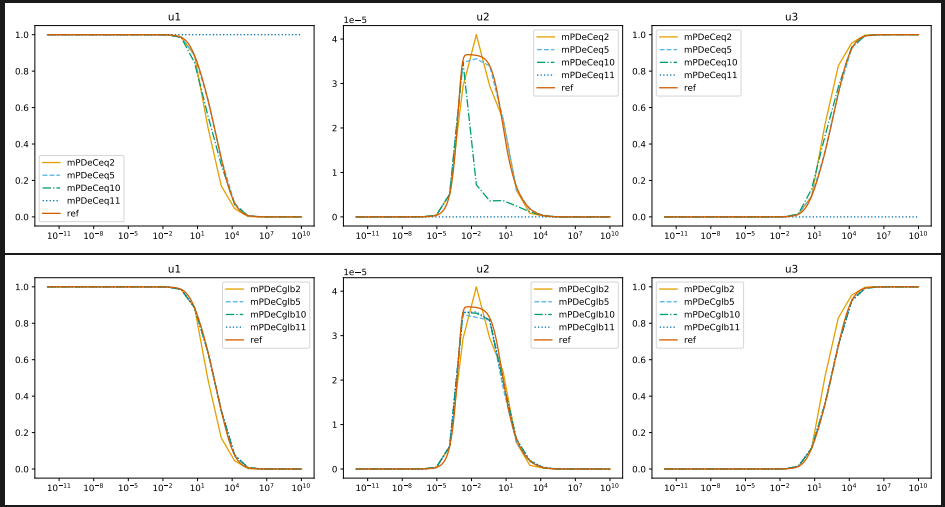
MPRKSO(2,2,0.001,10.) $\Delta t \approx 20$ bound, inconsistent,
MPRKSO(2,2,0.25,2.) $\Delta t \approx 1$ bound, consistent.



Robertson Problem: MPDeC

Equispaced
Order 2,5,10
$\Delta t \approx 1$, cons
Order 11
$\Delta t \approx 16$, inco

Gauss-Lobatto
Order
2,5,10,11
$\Delta t \approx 1$, cons



Outline

1. Production–Destruction System
2. (Modified) Patankar Methods
3. Properties and Troubles
4. Oscillation Study
5. Inconsistency Study
6. Simulations
7. Conclusion

Conclusion

Summary:

- (Modified) Patankar Schemes are **unconditionally** positive preserving
- Not **unconditionally stable** $\Longrightarrow$ oscillations
- Stability with some Δt bounds
- Inconsistency (order reduction and unphysical solution)
- No free meal: High Δt bounds, inconsistent.
- Best methods: low order, more stages.

Outlook:

- Find bound also in the nonlinear case
- More applied cases, same constraints?

Thank you!

Extensions of Modified Patankar Schemes MPRK(2,2,α)

The MPRK(2,2,α) of Kopecz and Meister (2018)

$$
\begin{aligned}
y^1 &= u^n, \\
y_i^2 &= u_i^n + \alpha \Delta t \sum_j \left(p_{ij}(y^1) \quad - d_{ij}(y^1) \quad \right), \\
u_i^{n+1} &= u_i^n + \Delta t \sum_j \Bigg(\left(\frac{2\alpha - 1}{2\alpha} p_{ij}(y^1) + \frac{1}{2\alpha} p_{ij}(y^2) \right) \\
&\qquad - \left(\frac{2\alpha - 1}{2\alpha} d_{ij}(y^1) + \frac{1}{2\alpha} d_{ij}(y^2) \right) \qquad \Bigg),
\end{aligned}
\qquad \text{(MPRK(2,2,}\alpha\text{))}
$$

$\alpha \in [1/2, \infty)$.
For $\alpha = 1$ based on Heun's method, i.e., SSPRK(2,2), already in Burchard, Deleersnijder, Meister (2003).

Extensions of Modified Patankar Schemes MPRK(2,2,α)

The MPRK(2,2,α) of Kopecz and Meister (2018)

$$
\begin{aligned}
y^1 &= u^n, \\
y_i^2 &= u_i^n + \alpha \Delta t \sum_j \left(p_{ij}(y^1) \frac{y_j^2}{y_j^1} - d_{ij}(y^1) \frac{y_i^2}{y_i^1} \right), \\
u_i^{n+1} &= u_i^n + \Delta t \sum_j \left(\left(\frac{2\alpha - 1}{2\alpha} p_{ij}(y^1) + \frac{1}{2\alpha} p_{ij}(y^2) \right) \frac{u_j^{n+1}}{(y_j^2)^{1/\alpha} (y_j^1)^{1-1/\alpha}} \right. \\
&\qquad \left. - \left(\frac{2\alpha - 1}{2\alpha} d_{ij}(y^1) + \frac{1}{2\alpha} d_{ij}(y^2) \right) \frac{u_i^{n+1}}{(y_i^2)^{1/\alpha} (y_i^1)^{1-1/\alpha}} \right),
\end{aligned}
\qquad \text{(MPRK(2,2,}\alpha\text{))}
$$

$\alpha \in [1/2, \infty)$.

For $\alpha = 1$ based on Heun's method, i.e., SSPRK(2,2), already in Burchard, Deleersnijder, Meister (2003).