MOX–Report No. 16/2010

Positivity and conservation properties of some integration schemes for mass action kinetics

Luca Formaggia, Anna Scotti

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

Positivity and conservation properties of some integration schemes for mass action kinetics*

L. Formaggia and A. Scotti♯

March 12, 2010

♯ MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
luca.formaggia@polimi.it
anna.scotti@mail.polimi.it

March 15, 2010
Keywords: integration schemes, chemical reactions, unconditional positivity, Patakar-type methods.

AMS Subject Classification: 65L04, 65L20.

Abstract

The numerical schemes approximating chemical reactions according to the mass action law should reproduce at least two properties of the corresponding physical system: mass conservation and nonnegativity of the concentrations. This paper analyzes the equations of mass action kinetics providing a proof of the existence, uniqueness, and positivity of the solution under mild hypothesis on the reaction rate and the stoichiometric coefficients. We then consider some classic integration schemes in terms of conservation, positivity and accuracy compared to schemes tailored for production-destruction systems, and propose an original scheme which guarantees conservation, nonnegativity of the solution and has order of convergence between two and three.

Introduction

The evolution of a set of species which interact through chemical reactions is usually described by a system of ordinary differential equations. Although single elementary reactions are associated with simple, low-dimensional systems

*This work has been supported by Eni S.p.A.

of nonlinear ODEs, in realistic situations the number of reactions and involved species can be rather high and leads to complex high-dimensional systems. These are often difficult to solve numerically since usually different reaction rates coexist and the resulting system is rather stiff [10, 21]. Explicit methods need an unreasonably small step size to be stable, and one has usually to resort to implicit schemes. Furthermore the physical system has some properties which should be retained by the numerical method, notably mass conservation and non-negativity of the solution.

In this paper we will discuss and compare some numerical schemes for systems of ordinary differential equations of the form

$$\frac{d\boldsymbol{c}}{dt} = \boldsymbol{\sigma}\boldsymbol{r}(\boldsymbol{c}(t), t),$$

where $\boldsymbol{\sigma}$ is a matrix of integer coefficients, the stoichiometric matrix, and $\boldsymbol{r}$ the reaction rates.

We will first show that under rather general assumptions on the initial data, reaction rates and the stoichiometric matrix, the system admits a global non-negative solution and satisfies a conservation property. To this aim we will apply the technique proposed in [12]. We will then consider some numerical method for the approximation of this ODE system, focusing on their conservation and positivity-preserving properties, as well as their accuracy and cost-effectiveness.

In several computer code for the solution of this type of problems it is not uncommon to enforce the positivity of the solution by using a clipping procedure. This is however a poor choice, which cannot guarantee robust and conservative algorithms. Another common choice is to employ Runge Kutta or multi-step methods with adaptive time step, expecting that the same error estimate that controls accuracy is able to reduce the risk of negative solution components. This is rather empirical, since a-priori there is no direct link between accuracy and positivity.

More sophisticated and robust approaches have been presented in the literature. In [3, 22] the proposed method consists in a post-processing (a projection on the feasible set) of the solution resulting from a one-step integration method. The projection method is more effective when applied to schemes that favors positivity: in [22] several Rosenbrock-type methods are presented.

A well known result by Bolley and Crouzeix, [6], states that for linear dissipative problems the unconditional positivity of traditional multistep and Runge Kutta methods restricts the order of the method to one. Z. Horváth in [14, 15] discusses the maximal step size for positivity of Runge Kutta and diagonally-split Runge Kutta (DSRK) methods for dissipative problems, showing that the step size threshold depends on the radius of positivity of the scheme. He shows that it is possible to construct diagonally-split Runge Kutta methods which are unconditionally positive and have an order higher than 1. The $\mathrm{AN}_f(0)$-stable DSRK methods introduced in [1] are indeed unconditionally contractive and

they can break the order barrier. However, these schemes tend to be expensive when applied to large systems.

An analysis of the time step threshold for two steps methods in linear multistep and one-leg form is given in [23]. Moving to nonstandard schemes [18], ad-hoc forms of the derivative can be formulated as described in [19] to ensure the positivity of the numerical solution.

Being computationally efficient, more traditional linear one-step methods and predictor-corrector methods are still largely used in practise. In this paper we analyze some of these schemes to assess their conservation properties and the step threshold for positivity. When applied to nonlinear problems these methods imply the iterative solution of a nonlinear problem which can introduce a more restrictive constraint for positivity. An estimate of the limit step length will be given in the case of a particular fixed point iteration method.

We will also consider a diagonally implicit Runge Kutta scheme schemes proposed in [5] which ensure both positivity and mass conservation as well as second order convergence. Afterwards, we will focus on methods based on the Patankar trick, which are designed for production-destruction equations of the form

$$\frac{dc_i}{dt} = P_i(\boldsymbol{c}) - D_i(\boldsymbol{c})$$

where $P_i(\boldsymbol{c}) \geq 0$, $D_i(\boldsymbol{c}) \geq 0$. The Patankar trick, originally denoted by *source term linearisation*, yields a stable and unconditionally positive method. The modified Patankar type method proposed by [9] is also conservative and second order accurate. Finally, we will propose an improvement of the MPRK method [9] introducing, as a corrector step, a modified third order BDF method. It is worth mentioning that time step adaptivity is of fundamental importance when dealing with complex realistic problems: nevertheless, a in-depth analysis of adaptive schemes is beyond the scope of this work. However, most of the results here presented may be of use also in the adaptive setting.

All the methods have been applied to the integration of test cases of increasing complexity, starting from a synthetic low-dimensional case to a realistic case of about 200 equations. The purpose is to verify the theoretical findings and quantify the computational costs.

1 Governing equations

In the following, vectors of $\mathbb{R}^N$ will be indicated in boldface, e.g. $\boldsymbol{v}$ denotes a vector of components v_i, for $1 \leq i \leq N$. For a vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|_p = \left(\sum_{i=1}^N |v_i|^p\right)^{1/p}$ indicates the p-norm. Notable cases are that of the 2-norm or Euclidean norm, which we will indicate by simply $\|\boldsymbol{v}\|$ and the 1-norm $\|\boldsymbol{v}\|_1 = \sum_{i=1}^N |v_i|$.

For a matrix $A \in \mathbb{R}^{N\times M}$ we denote again by $\|\boldsymbol{A}\|_p$ the matrix norm subordinated to the corresponding vector norm,

$$\|\boldsymbol{A}\|_p = \sup_{\substack{\mathbf{v}\in\mathbb{R}^M \\ \|\mathbf{v}\|_p=1}} \|A\mathbf{v}\|_p.$$

With $\boldsymbol{v} > 0$ we indicate that all component of $\boldsymbol{v}$ are strictly positive. Given two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ of $\mathbb{R}^N$ we indicate by $\boldsymbol{v}^T$ the transpose of $\boldsymbol{v}$ and by $\boldsymbol{v}^T\boldsymbol{w}$ the scalar product between the two vectors (i.e. we use the convention that a vector is always a column vector). Finally $\mathbf{e} = [1, \ldots, 1]^T$ is a vector of all ones.

Let us consider a generic set of chemical reactions involving N components, whose molar concentration are indicated by $\boldsymbol{c} = [c_1, \ldots, c_N]^T$, and M reactions. The latter are governed by the reaction rates $\boldsymbol{r} = [r_1, \ldots, r_M]^T$, where $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{c}, t) : \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}^M$. We have set $\mathbb{R}_+ = (0, \infty)$ and assumed that $t = 0$ is the initial time for our chemical process. The concentrations are then governed by the following system

$$\begin{aligned} &\frac{d\boldsymbol{c}(t)}{dt} = \boldsymbol{\sigma}\boldsymbol{r}(\boldsymbol{c}(t), t), \quad t > 0 \\ &\boldsymbol{c}(0) = \boldsymbol{c}_0, \end{aligned} \tag{1}$$

where $\boldsymbol{\sigma}$ is the stoichiometric matrix of integers σ_{ij}, with $1 \le i \le N$ and $1 \le j \le M$.

We will make some assumptions throughout this chapter.

Assumption 1.1 *The stoichiometric matrix $\boldsymbol{\sigma}$ satisfies the following hypothesis,*

1. *non-triviality:*
$$\begin{cases} \forall j \ \exists i \ s.t. \ \sigma_{ij} \neq 0 \\ \forall i \ \exists j \ s.t. \ \sigma_{ij} \neq 0 \end{cases} \tag{2}$$

2. *conservation: $\boldsymbol{e} \in \mathrm{Ker}(\boldsymbol{\sigma}^T)$ or, equivalently, $\boldsymbol{e} \perp \mathrm{range}(\boldsymbol{\sigma})$.*

3. *$\mathrm{Rank}(\boldsymbol{\sigma}) = M$, i.e. the columns of $\boldsymbol{\sigma}$ are linearly independent.*

Assumption 1.2 *The reaction rates are such that*

1. *For each $j = 1, \ldots, M$, $r_j \in C^0(\overline{\mathbb{R}}^N_+, \overline{\mathbb{R}}_+)$, and $r_j(\cdot, t)$ is locally Lipschitz continuous in $\mathbb{R}^N$, uniformly in t. Furthermore, $r_j(\overline{\boldsymbol{c}}, \cdot) \in L^\infty(\mathbb{R}_+)$ for any $\overline{\boldsymbol{c}} \in \overline{\mathbb{R}}^N_+$;*

2. *$\boldsymbol{r} > 0$ if $\boldsymbol{c} > 0$ and $\boldsymbol{r} = \mathbf{0}$ if $\boldsymbol{c} = \mathbf{0}$, for all $t > 0$;*

3. *if $\sigma_{ij} < 0$ then there exists a $q_j \in C^0(\overline{\mathbb{R}}^N_+, \overline{\mathbb{R}_+})$, with $q_j > 0$ if $\boldsymbol{c} > 0$ and $q_j = 0$ if $\boldsymbol{c} = \mathbf{0}$, such that $r_j(\boldsymbol{c}, t) = q_j(\boldsymbol{c}, t)c_i$.*

A possible expression for the reaction rate results from the mass action law [4],

$$r_j = r_j^{MA} = k_j \prod_{s=1}^{N} c_s^{-min(\sigma_{sj},0)}, \tag{3}$$

where the reaction constants k_j normally depend on the temperature T by the Arrhenius law [17],

$$k_j(T) = A_j e^{\left(-\frac{E_{a_j}}{R_U T}\right)}. \tag{4}$$

Here $A_j > 0$ is the so-called pre-exponential factor, R_U the universal gas constant, T the absolute temperature and E_{a_j} the activation energy. If we take the temperature as a given continuous function of time, satisfying $T \geq T_0 > 0$ at all times, then (3) satisfies Assumption 1.2. The first condition is indeed verified thanks to the hypothesis on the temperature. Since σ_{ij} are integers, if $\sigma_{ij} < 0$ then $r_j^{MA} = c_i q_j(\boldsymbol{c}, t)$ with $q_j = k_j c_i^{-\sigma_{ij}-1} \prod_{s \neq i} c_s^{-min(\sigma_{sj},0)}$, which provides the second condition. The final condition is satisfied by observing that the reaction constant is strictly positive, thus if $\boldsymbol{c} \geq 0$ we have that $r_j^{MA} \geq 0$, and the equality $r_j^{MA} = 0$ is attained if $\boldsymbol{c} = 0$ for the null factor law.

In more general situations, the temperature could itself depend on $\frac{d\boldsymbol{c}}{dt}$, because of the heat produced or absorbed by the reactions. We will not consider this case here. However, we note that Assumption 1.2 may apply to expressions for the reactions rates different from (3).

If the chemical system is closed it must satisfy a mass conservation principle, namely $\sum_{k=1}^{N} c_k = \boldsymbol{e}^T \boldsymbol{c}$ must be constant. In other words, $g = \boldsymbol{e}^T \boldsymbol{c}$ is a first integral of (1). The following proposition holds [5].

Theorem 1.1 *The second condition of Assumption 1.1 implies that the solution of system (1) satisfies the mass conservation principle.*

Conversely, if $\boldsymbol{e}^T \boldsymbol{c}$ is constant, $\boldsymbol{c}$ being a non-trivial solution of (1), $\mathrm{Rank}(\boldsymbol{\sigma}) = M$ *and the reaction rates satisfy Assumption 1.2, then $\boldsymbol{e} \in \mathrm{Ker}(\boldsymbol{\sigma}^T)$.*

Proof. From $\mathbf{e}^T \boldsymbol{\sigma} = \mathbf{0}$ we get

$$\frac{dg}{dt} = \boldsymbol{e}^T \frac{d\boldsymbol{c}}{dt} = \boldsymbol{e}^T \boldsymbol{\sigma} \boldsymbol{r}(\boldsymbol{c}) = 0,$$

and thus g is a first integral.

Conversely, if g is a first integral then

$$0 = \frac{dg}{dt} = \boldsymbol{e}^T \frac{d\boldsymbol{c}}{dt} = \boldsymbol{e}^T \boldsymbol{\sigma} \boldsymbol{r}(\boldsymbol{c}),$$

and since $\boldsymbol{\sigma}$ has full column rank and $\boldsymbol{r} \neq 0$ (a consequence of the non-triviality of the solution), we must have $\boldsymbol{e}^T \boldsymbol{\sigma} = 0$. □

Remark 1.1 *The definition of conservation given above is a basic one. Indeed, it can be specialized for chemical and biochemical problems, where elemental mass conservation is required, i.e. the amount of any element, which is a constituent of one or more species, must not change. Schemes that guarantee this conservation properties are discussed in [7] and [8].*

Clearly the proof of Proposition (1.1) is merely formal, since it assumes the existence of a solution of system (1), an issue that we are going to address in the next section.

2 Existence, uniqueness and positivity of the solution

Let us set

$$\boldsymbol{f} = \boldsymbol{\sigma}\boldsymbol{r}, \tag{5}$$

define the constant $K = \|\boldsymbol{c}_0\|_1$ and the set

$$\Omega = \{\boldsymbol{x} \in \overline{\mathbb{R}}_+^N : \|\boldsymbol{x}\|_1 \leq K\}.$$

If $K = 0$, by the second of Assumption 1.2, $\dfrac{d\boldsymbol{c}}{dt} = 0$ and $\boldsymbol{c} = 0\ \forall t$. Therefore we assume that $K > 0$. We introduce a modified forcing term $\tilde{\boldsymbol{f}}$ as follows.

$$\tilde{\boldsymbol{f}}(\boldsymbol{c}, t) = \boldsymbol{f}(\boldsymbol{h}(\boldsymbol{c}), t), \tag{6}$$

where

$$\boldsymbol{h}(\boldsymbol{c}) = \begin{cases} \boldsymbol{c} & \text{if } \boldsymbol{c} \in \Omega \\ K\dfrac{\boldsymbol{c}}{\|\boldsymbol{c}\|_1} & \text{otherwise.} \end{cases} \tag{7}$$

It is immediate to verify that $\boldsymbol{h}(\boldsymbol{c})$ is nothing else than the projection of $\boldsymbol{c}$ on Ω in the 1-norm. Since Ω is convex $\boldsymbol{h}(\boldsymbol{c})$ is unique. Furthermore, $\|\boldsymbol{h}(\boldsymbol{c}_1) - \boldsymbol{h}(\boldsymbol{c}_2)\|_1 \leq \|\boldsymbol{c}_1 - \boldsymbol{c}_2\|_1$, for all $\boldsymbol{c}_1$ and $\boldsymbol{c}_1$ in $\mathbb{R}^N$.

We are now in the position of stating the following

Lemma 2.1 *The modified function $\tilde{\boldsymbol{f}}$ is bounded in $\mathbb{R}^N \times \mathbb{R}_+$ and is globally Lipschitz continuous w.r.t. the fist argument, uniformly in the time t.*

Proof. Since Ω is a compact set of R^N, a consequence of Assumption 1.2 and of the definition of $\tilde{\boldsymbol{f}}$ is that

$$\sup_{\substack{\boldsymbol{c}\in\mathbb{R}^N \\ t\in\mathbb{R}_+}} \|\tilde{\boldsymbol{f}}(\boldsymbol{c}, t)\|_1 = \sup_{\substack{\boldsymbol{c}\in\Omega \\ t\in\mathbb{R}_+}} \|\boldsymbol{f}(\boldsymbol{c}, t)\|_1 \leq \|\boldsymbol{\sigma}\|_1 \sup_{\substack{\boldsymbol{c}\in\Omega \\ t\in\mathbb{R}_+}} \|\boldsymbol{r}(\boldsymbol{c}, t)\|_1 = k_1 < \infty.$$

Furthermore, for any $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ in $\mathbb{R}^N$,

$$\|\tilde{\boldsymbol{f}}(\boldsymbol{c}_1, t) - \tilde{\boldsymbol{f}}(\boldsymbol{c}_2, t)\|_1 \leq \Lambda\|\boldsymbol{\sigma}\|_1\|\boldsymbol{h}(\boldsymbol{c}_1) - \boldsymbol{h}(\boldsymbol{c}_2)\|_1 \leq \Lambda\|\boldsymbol{\sigma}\|_1\|\boldsymbol{c}_1 - \boldsymbol{c}_2\|_1.$$

Here, the existence of

$$\Lambda = \sup_{\boldsymbol{c}_1,\boldsymbol{c}_2\in\Omega} \frac{\|\boldsymbol{r}(\boldsymbol{c}_1,t)-\boldsymbol{r}(\boldsymbol{c}_2,t)\|_1}{\|\boldsymbol{c}_1-\boldsymbol{c}_2\|_1}$$

is assured by the local uniform Lipschitz continuity of each component of $\boldsymbol{r}$. □

We new recall a classical global existence result ([13]).

Theorem 2.1 *Let $X = \Omega \times (\tau_1, \tau_2)$ with $\Omega \subset \mathbb{R}^N$, and assume that $\boldsymbol{f}(\boldsymbol{y}, t)$ is defined in $\overline{X}$ with the following properties:*

a) $\boldsymbol{f}$ is continuous in an open subset $D \subset X$;

b) $\boldsymbol{f}$ is locally Lipschitz continuous in D with respect to $\boldsymbol{y}$ and uniformly in t;

c) there are two non-negative constants k_1 and k_2 such that

$$\|\boldsymbol{f}(t,\boldsymbol{y})\| \le k_1 + k_2\|\boldsymbol{y}\| \qquad \forall(\boldsymbol{y},t)\in\overline{X}.$$

Then for all $(\boldsymbol{\xi},\tau)\in X$ there is at least one solution $\boldsymbol{\phi}(t;\tau,\boldsymbol{\xi})$ of problem

$$\begin{cases} \frac{d\boldsymbol{\phi}}{dt}(t) = \boldsymbol{f}(\boldsymbol{\phi}(t),t), \quad t\in(\tau_1,\tau_2), \\ \phi(\tau) = \boldsymbol{\xi}, \end{cases}$$

defined in $[\tau_1, \tau_2]$ and with values in D.

We are now in the position of proving the following:

Theorem 2.2 *If Assumptions 1.1 and 1.2 hold, for any initial condition $\boldsymbol{c}(0) = \boldsymbol{c}_0 \ge 0$ system (1) has a unique non-negative solution $\boldsymbol{c} \in [C^1(\overline{\mathbb{R}}_+)]^N$. Furthermore, $c_i(t) > 0$ for all $t \ge 0$ whenever $c_{0,i} > 0$.*

Proof. Theorem 2.1 can be applied to the modified problem

$$\begin{cases} \dfrac{d\boldsymbol{\phi}}{dt} = \tilde{\boldsymbol{f}}(\boldsymbol{\phi},t), \\ \boldsymbol{\phi}(0) = \boldsymbol{c}_0 \end{cases} \tag{8}$$

since $\tilde{\boldsymbol{f}}$ satisfies all its hypothesis on $D = \mathbb{R}^N \times \mathbb{R}^+$. In particular, the last inequality in Theorem 2.1 is satisfied for $k_2 = 0$ and k_1 as defined in Lemma 2.1.

We now consider the solution of the modified problem for a $\boldsymbol{c}_0 \ge 0$, The trivial case $\boldsymbol{c}_0 = \boldsymbol{0}$, thanks to the properties of the reaction rates, would provide the solution $\boldsymbol{\phi} = \boldsymbol{0}$.

We will assume in the following that there is a i such that $c_{0,i} > 0$. To prove non-negativity of the solution we follow the technique proposed in [12]. Let us define, for each component ϕ_i, two sets of indexes: I_i^- containing those j such that $\sigma_{ij} < 0$, and I_i^+ containing the j such that $\sigma_{ij} \ge 0$. Thanks to the second condition of Assumption 1.2 (notice that it is satisfied by $\tilde{\boldsymbol{f}}$ as well) the i-th component of the solution of the modified problem satisfies for all $t > 0$ the equation

$$\frac{d\phi_i(t)}{dt} = \sum_{j\in I_i^-} \sigma_{ij} q_j(\boldsymbol{\phi}(t),t)\phi_i(t) + \sum_{j\in I_i^+} \sigma_{ij} r_j(\boldsymbol{\phi}(t),t), \tag{9}$$

which we rewrite in the form

$$\frac{d\phi_i(t)}{dt} = -a_i(t)\phi_i(t) + b_i(t), \quad t > 0, \quad i = 1, \ldots, N, \tag{10}$$

where a_i and b_i are continuous functions of time. Thus, we have that

$$\frac{d}{dt}\left(e^{\int_0^t a_i(\tau)d\tau}\phi_i(t)\right) = e^{\int_0^t a_i(\tau)d\tau}b_i(t) \quad t > 0, \tag{11}$$

and, by setting $y_i(t) = e^{\int_0^t a_i(\tau)d\tau}\phi_i(t)$, we obtain that for each $i = 1, \ldots, N$

$$y_i(t) = y_i(0) + \int_0^t e^{\int_0^\theta a_i(\tau)d\tau}b_i(\theta)d\theta, \quad t > 0. \tag{12}$$

Furthermore, because of the stated properties of the reaction rate $a_i \geq 0$ and $b_i \geq 0$ whenever $\boldsymbol{\phi} \geq 0$. Because of the continuity of $\boldsymbol{\phi}$ and the given initial condition, $\boldsymbol{\phi} \geq 0$ for $t \in [0, T)$ for a given $T > 0$.

Let assume that there exists a i such that $\phi_i(0) > 0$ while $\phi_i(T) = 0$. We would necessarily have

$$0 = y_i(T) = y_i(0) + \int_0^T e^{\int_0^\theta a_i(\tau)d\tau}b_i(\theta)d\theta, \tag{13}$$

with $a_i \geq 0$, $b_i \geq 0$ and $y_i(0) > 0$. This is clearly a contradiction, by which we prove that in fact $\phi_i(t)$ must be positive for all $t \in [0, T]$. If instead $\phi_i(0) = 0$, again the non-negativity of a_i and b_i in expression (12) leads to the conclusion that $y_i(T) \geq 0$, and thus $\phi_i(T)$ is non-negative. Therefore, given a non-negative initial condition, we have that $\boldsymbol{\phi}(t) \geq 0$ for $t \in [0, T]$, and all components initially positive remain so. Since T is now arbitrary we may conclude that $\boldsymbol{\phi}(T) \geq 0$ for all $t > 0$ and that no component initially positive goes to zero in a finite time.

Using the fact that $\tilde{\boldsymbol{f}}(\boldsymbol{c}, t) = \boldsymbol{\sigma r}(\boldsymbol{h}(\boldsymbol{c}), t)$, by exploiting Assumption 1.1 and Theorem 1.1 we may deduce that $\sum_{i=1}^N \phi_i(t) = K$, with $K = \sum_{i=1}^N \phi_i(0) \geq 0$. Furthermore, the non-negativity of the solution allows us to write that $\|\boldsymbol{\phi}(t)\|_1 = K$, for all $t \geq 0$. Consequently, along the solution trajectory $\tilde{\boldsymbol{f}} = \boldsymbol{f}$ and then $\boldsymbol{\phi}$ coincides with a solution $\boldsymbol{c}$ of the original problem (1). Since the solution is contained in a compact of $\mathbb{R}^N$, it is the only solution. □

3 Properties of some numerical schemes

We want to investigate the properties of some numerical schemes with respect to conservation and positivity. We will start from some well known one-step schemes and then we will present methods more specialized for the problem at hand.

We first recall, without giving the proof, a result given in [5].

Lemma 3.1 *Under Assumptions 1.1 and 1.2, for any $\boldsymbol{c} \in \overline{\mathbb{R}}_+^N$ and $t \geq 0$ there exists a $N \times N$ matrix $\mathrm{R} = \mathrm{R}(\boldsymbol{c}, t)$ whose entries satisfy, for $1 \leq i, j \leq N$*

$$a)\ \mathrm{R}_{ii} \leq 0, \quad b)\ \mathrm{R}_{ij} \geq 0 \quad \text{for } i \neq j \text{ and } \ c)\ \sum_{i=1}^N \mathrm{R}_{ij} = 0,$$

and such that $\boldsymbol{\sigma r}(\boldsymbol{c}, t) = \mathrm{R}(\boldsymbol{c}, t)\boldsymbol{c}$.

3.1 Theta-methods

A generic step of a theta-method from time t^n to $t^{n+1} = t^n + \Delta t$ reads

$$\boldsymbol{c}^{n+1} = \boldsymbol{c}^n + \Delta t \boldsymbol{\sigma}(\theta \boldsymbol{r}^{n+1} + (1-\theta)\boldsymbol{r}^n), \tag{14}$$

where $\boldsymbol{c}^n \in \mathbb{R}^N$ is the approximated solution a time t^n, $\boldsymbol{r}^n = \boldsymbol{r}(\boldsymbol{c}^n, t^n)$ and $\theta \in [0,1]$. For $\theta = 0$ and $\theta = 1$ we have the explicit and implicit Euler scheme, respectively, while $\theta = 1/2$ gives the second-order Crank-Nicholson method. For the properties of convergence and absolute stability of the scheme the reader may refer to [21].

For $\theta \neq 0$, at each step we need to solve the nonlinear system

$$\boldsymbol{c}^{n+1} - \theta \Delta t \boldsymbol{\sigma} \boldsymbol{r}(\boldsymbol{c}^{n+1}, t^{n+1}) = \boldsymbol{c}^n + (1-\theta)\Delta t \boldsymbol{\sigma} \boldsymbol{r}^n, \tag{15}$$

which can also be written as

$$[I - \theta \Delta t \mathrm{R}^{n+1}(\boldsymbol{c}^{n+1})]\boldsymbol{c}^{n+1} = \boldsymbol{c}^n + (1-\theta)\Delta t \mathrm{R}^n(\boldsymbol{c}^n)\boldsymbol{c}^n, \tag{16}$$

where $\mathrm{R}^n(\boldsymbol{c}^n) = \mathrm{R}(\boldsymbol{c}^n, t^n)$.

Before going any further we state the following

Lemma 3.2 *If* R *is a* $N \times N$ *matrix satisfying all properties stated in Lemma 3.1, then matrix* $\mathrm{A} = I - \kappa \mathrm{R}$*, is a non-singular* M*-matrix for all* $\kappa > 0$*,* I *being the identity matrix.*

Proof. Indeed, A is a Z-matrix, since all off-diagonal terms are non positive. In addition, it is diagonally dominant by column, with positive diagonal elements. Thus the real part of all eigenvalues is strictly positive. This two conditions imply that A is a non-singular M-matrix. □

Theorem 3.1 *Under Assumptions 1.1 and 1.2 a theta-method is conservative for any* $\theta \in [0,1]$*. Furthermore, it is conditionally positive whenever* $\theta \neq 1$*, while for* $\theta = 1$ *is positive for any value of* Δt*.*

Proof. Conservation is easily found by pre-multiplying both members of (14) by $\boldsymbol{e}^T$ and exploiting the properties of the stoichiometric matrix to obtain that $\boldsymbol{e}^T\boldsymbol{c}^{n+1} = \boldsymbol{e}^T\boldsymbol{c}^n = \ldots = \boldsymbol{e}^T\boldsymbol{c}_0$.

As for positivity, we first consider the case $\theta = 0$. Let k be such that

$$f_k = \sum_j \sigma_{kj} r_j^n = \min_{1 \leq i \leq N} \sum_j \sigma_{ij} r_j^n, \tag{17}$$

then $\boldsymbol{c}^{n+1} \geq 0$ if $\boldsymbol{c}^n \geq 0$ under the condition $\Delta t \leq \frac{c_k}{|f_k|}$ if $f_k < 0$, while positivity is obtained with no condition on Δt if $f_k \geq 0$. We recall that in any case Δt cannot be arbitrarily large for stability reasons.

We now consider the case $\theta \neq 0$. In this case $\boldsymbol{c}^{n+1}$ is the solution $\boldsymbol{x}$ of the non linear problem

$$\left[I - \theta \Delta t \mathrm{R}^{n+1}(\boldsymbol{x})\right]^{-1} \boldsymbol{x} = \left[I + (1-\theta)\Delta t \mathrm{R}^n(\boldsymbol{c}^n)\right] \boldsymbol{c}^n. \tag{18}$$

To prove existence of a positive solution we set $K = \|\boldsymbol{c}^n\|_1$ and consider the compact set

$$\Omega = \{\boldsymbol{y} \in \mathbb{R}^N : \quad \|\boldsymbol{y}\|_1 \leq K \quad \boldsymbol{y} \geq 0\}. \tag{19}$$

Furthermore, we assume that $\boldsymbol{c}^n \geq 0$.

Being Ω a convex compact of $\overline{\mathbb{R}}_+^N$ the matrix $\mathrm{S}_\theta(\boldsymbol{x}) = I - \theta\Delta t \mathrm{R}^{n+1}(\boldsymbol{x})$ is well defined for all $\mathbf{x} \in \Omega$. We look now for a fixed point of

$$\boldsymbol{\varphi}(\boldsymbol{x}) = \mathrm{S}_\theta^{-1}(\boldsymbol{x}) \left[I + (1-\theta)\Delta t \mathrm{R}^n(\boldsymbol{c}^n)\right] \boldsymbol{c}^n \tag{20}$$

in Ω. If it exists it is clearly a solution of (18).

Thanks to the properties of R and Lemma 3.2 matrix $\mathrm{S}_\theta(\boldsymbol{x})$ is a a non-singular M-matrix for any $\boldsymbol{x} \in \Omega$. Furthermore, the fixed point function satisfies

$$\boldsymbol{\varphi}(\boldsymbol{x}) - \theta\Delta t \mathrm{R}^{n+1}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{c}^n + (1-\theta)\Delta t \mathrm{R}^n(\boldsymbol{c}^n)\boldsymbol{c}^n,$$

thus $\boldsymbol{e}^T\boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{e}^T\boldsymbol{c}^n = K$.

We define k and f_k as in (17) and set

$$\tau_1 = \begin{cases} \infty, & \text{if } \theta = 1 \quad \text{or } f_k \geq 0 \\ \frac{c_k}{(1-\theta)|f_k|}, & \text{in all other cases.} \end{cases} \tag{21}$$

If $\Delta t < \tau_1$ we have that $\left[I + (1-\theta)\Delta t \mathrm{R}^n(\boldsymbol{c}^n)\right] \boldsymbol{c}^n \geq 0$. We conclude that under this condition $\boldsymbol{\varphi}(\boldsymbol{x}) \geq 0$, and consequently $\|\boldsymbol{\varphi}(\boldsymbol{x})\|_1 = K$.

Remark 3.1 *Note that this result is consistent with that presented in [14] for Runge Kutta methods.*

Thus, $\boldsymbol{\varphi} : \Omega \to \Omega$ and, by the Brouwer fixed point theorem, we conclude that there is a fixed point $\boldsymbol{x} \in \Omega$ of $\boldsymbol{\varphi}$. It satisfies $\|\boldsymbol{x}\|_1 = K$, so in fact $\boldsymbol{x} \in \partial\Omega$.

Applying the argument to all iterations k we can infer that, possibly under restrictions on the time step, a theta-method is positive and its solution satisfies $\|\boldsymbol{c}^n\|_1 = \|\boldsymbol{c}_0\|_1$ for all $n \geq 0$. □

In general the non-linear problem (15) is solved approximatively by using an iterative procedure that provides a sequence $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots\}$ converging to the solution of (15), and we choose a $\boldsymbol{x}_M$ as approximation of $\boldsymbol{c}^{n+1}$, according to a suitable stopping criterion.

If we want that the previous result be of practical use we need to extend it to the elements of the approximating sequence.

Corollary 3.2 *Let us consider $\boldsymbol{c}^n \geq 0$, $\|\boldsymbol{c}^n\|_1 = K > 0$, and set $\boldsymbol{x}_0 = \boldsymbol{c}^n$. Furthermore we assume that $\mathrm{R}^{n+1}(\boldsymbol{x})$ has continuous derivatives for all $\boldsymbol{x} \in \Gamma$, where $\Gamma = \{\boldsymbol{y} \in \mathbb{R}^N : \quad \|\boldsymbol{y}\|_1 = K \quad \boldsymbol{y} \geq 0\}$. Then, if Δt is sufficiently small, the sequence generated by the fixed point iteration*

$$\boldsymbol{x}_{k+1} = \boldsymbol{\varphi}(\boldsymbol{x}_k), \quad k = 0, 1, \ldots,$$

where $\boldsymbol{\varphi}$ is the function defined in (20), converges to a fixed point $\boldsymbol{c}^{n+1} \geq 0$ with $\|\boldsymbol{c}^{n+1}\|_1 = K$ if $\boldsymbol{x}_0 \in \Gamma$. Furthermore, any element of the sequence satisfies $\|\boldsymbol{x}_k\|_1 = K$ and $\boldsymbol{x}_k \geq 0$.

Proof. The fact that $\boldsymbol{x}_k \geq 0$ and $\|\boldsymbol{x}_k\|_1 = K$ is immediately inferred from the demonstration of Theorem 3.1: it suffices to consider $\Delta t < \tau_1$, since we have just demostrated that under that condition $\boldsymbol{\varphi} : \Gamma \to \Gamma$.

We need to show that the fixed point iteration converges. To this purpose we note that, given two points $\boldsymbol{x}$ and $\boldsymbol{y}$ of Γ, and setting $\boldsymbol{h} = \boldsymbol{y} - \boldsymbol{x}$, all points $\boldsymbol{x(t)} = \boldsymbol{x} + \boldsymbol{h}t$ for $t \in [0, 1]$ belong to Γ. Since

$$\boldsymbol{\varphi}(\boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x}) = \int_0^1 \nabla\boldsymbol{\varphi}(\boldsymbol{x} + \boldsymbol{h}t) \cdot \boldsymbol{h}dt,$$

we obtain that

$$\|\boldsymbol{\varphi}(\boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x})\|_1 \leq \sup_{\boldsymbol{x}\in\Gamma} \|\nabla\boldsymbol{\varphi}(\boldsymbol{x})\|_1 \|\boldsymbol{y} - \boldsymbol{x}\|_1,$$

where we have set

$$\|\nabla\boldsymbol{\varphi}(\boldsymbol{x})\|_1 = \max_{1\leq k\leq N} \sum_{i=1}^{N} |\partial_k \varphi_i|. \tag{22}$$

Consequently, the iterations converge if $\Lambda = \sup_{\boldsymbol{x}\in\Gamma} \|\nabla\boldsymbol{\varphi}(\boldsymbol{x})\|_1 < 1$.

We have that

$$\nabla\boldsymbol{\varphi}(\boldsymbol{x}) = \theta\Delta t(I - \theta\Delta t\mathrm{R}(\boldsymbol{x}))^{-1}\nabla\mathrm{R}(\boldsymbol{x})(I - \theta\Delta t\mathrm{R}(\boldsymbol{x}))^{-1}(I + \Delta t(1-\theta)\mathrm{R}(\boldsymbol{c}^n))\boldsymbol{c}^n,$$

and then,

$$\begin{aligned}\|\nabla\boldsymbol{\varphi}(\boldsymbol{x})\|_1 &= \theta\Delta t\|(I - \theta\Delta t\mathrm{R}(\boldsymbol{x}))^{-1}\nabla\mathrm{R}(\boldsymbol{x})(I - \theta\Delta t\mathrm{R}(\boldsymbol{x}))^{-1}(I + \Delta t(1-\theta)\mathrm{R}(\boldsymbol{c}^n))\boldsymbol{c}^n\|_1 \\ &\leq \theta\Delta t\|(I - \theta\Delta t\mathrm{R}(\boldsymbol{x}))^{-1}\|_1^2\|\nabla\mathrm{R}(\boldsymbol{x})\|_1\|I + \Delta t(1-\theta)\mathrm{R}(\boldsymbol{c}^n)\|_1 K,\end{aligned}$$

where

$$\|\nabla\mathrm{R}(\boldsymbol{x})\|_1 = \max_{1\leq j,k\leq N} \sum_{1=1}^{N} |\partial_k R_{ij}(\boldsymbol{x})|.$$

Under the condition $\Delta t < \tau_2 = (\theta \max_{\boldsymbol{x}\in\Gamma} \|\mathrm{R}(\boldsymbol{x})\|_1)^{-1}$, we have that

$$\|[I - \theta\Delta t\mathrm{R}(\boldsymbol{x})]^{-1}\|_1 \leq \frac{1}{1 - \theta\Delta t\|\mathrm{R}(\boldsymbol{x})\|_1}$$

for all $\boldsymbol{x} \in \Gamma$. Furthermore, the condition $\Delta t < \tau_2$ implies that

$$\|I + (1-\theta)\Delta t\mathrm{R}(\boldsymbol{c}^n)\|_1 < 1 + (1-\theta)\Delta t\|\mathrm{R}(\boldsymbol{c}^n)\|_1 \leq \frac{1}{\theta}$$

and therefore

$$\|\nabla\boldsymbol{\varphi}(\boldsymbol{x})\|_1 < \frac{K\Delta t\|\nabla\mathrm{R}(\boldsymbol{x})\|_1}{(1 - \theta\Delta t\|\mathrm{R}(\boldsymbol{x})\|_1)^2}.$$

For a $\boldsymbol{x} \in \Gamma$ we set $a = 1 - \theta\Delta t\|\mathrm{R}(\boldsymbol{x})\|_1$. Then $0 < a < 1$, and $\|\nabla\boldsymbol{\varphi}(\boldsymbol{x})\|_1 < 1$ is satisfied if $\dfrac{\Delta t K\|\nabla\mathrm{R}(\boldsymbol{x})\|_1}{a^2} < 1$, that is if

$$\frac{1-a}{\theta a^2} < \frac{\|\mathrm{R}(\boldsymbol{x}^k)\|_1}{K\|\nabla\mathrm{R}(\boldsymbol{x}^k)\|_1}.$$

Setting $z = z(\boldsymbol{x}) = \dfrac{\|\mathrm{R}(\boldsymbol{x})\|_1}{K\|\nabla \mathrm{R}(\boldsymbol{x})\|_1}$ the condition may be rewritten as $\theta z a^2 + a - 1 > 0$. The quadric has roots $a_{1,2} = \dfrac{-1 \pm \sqrt{1+4\theta z}}{2\theta z}$. One root is negative, the other, say a_2, is positive and smaller than one. Then the polynomial is positive for $a > a_2 = \dfrac{\sqrt{1+4\theta z}-1}{2\theta z}$. To have the most restrictive condition we have to take $z = \min_{\boldsymbol{x}\in\Gamma} z(\boldsymbol{x})$ in the previous expression and then use a time step satisfying $\Delta t < \tau_3 = \dfrac{1-a_2}{\theta \max_{\boldsymbol{x}\in\Gamma} \|\mathrm{R}(\boldsymbol{x})\|_1}$.

It follows that the fixed point iteration $\boldsymbol{x}_{k+1} = \boldsymbol{\varphi}(\boldsymbol{x}_k)$ converges if $\Delta t < \min(\tau_1, \tau_2, \tau_3)$. □

Remark 3.2 *We can conclude from the proof of the previous theorem that if Δt is sufficiently small, function $\boldsymbol{\varphi}$ has a single fixed point which lays on Γ.*

Remark 3.3 *If we use a Newton iteration for problem 14 we can easily show that the scheme is conservative, i.e. $\boldsymbol{e}^T \boldsymbol{x}^k = 0$. Yet, positivity is not guaranteed in general.*

3.2 A second order diagonally implicit Runge Kutta

Let us consider a two-stage diagonally implicit Runge Kutta method [2], also called semi-implicit Runge Kutta method, defined by the following Butcher array

$$\begin{array}{cc|c} \alpha & 0 & \alpha \\ 1-\alpha & \alpha & 1 \\ \hline 1-\alpha & \alpha & \end{array}$$

with $\alpha = 1 \pm \sqrt{2}/2$. This method is A-stable and S-stable, thus suitable for stiff problems. With respect to a generic implicit Runge Kutta, diagonally implicit Runge Kutta (DIRK) have a significant computational advantage, since the stages can be solved in cascade. The scheme can be written, after some manipulations, as

$$\begin{cases} \boldsymbol{x} - \alpha \Delta t \boldsymbol{\sigma} \boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{c}^n \\ \boldsymbol{y} - \alpha \Delta t \boldsymbol{\sigma} \boldsymbol{r}(\boldsymbol{y}) = \dfrac{(2\alpha-1)\boldsymbol{c}^n + (1-\alpha)\boldsymbol{x}}{\alpha} \\ \boldsymbol{c}^{n+1} = \boldsymbol{y}, \end{cases} \tag{23}$$

Theorem 3.3 *Under Assumptions 1.1 and 1.2, method 23 is conservative for any Δt. Furthermore, it is unconditionally positive.*

Proof. Let us set $\boldsymbol{b} = \dfrac{(2\alpha-1)\boldsymbol{c}^n + (1-\alpha)\boldsymbol{x}}{\alpha}$. We first prove that the scheme is positivity preserving for any Δt. Given $\boldsymbol{c}^n \geq 0$ the positivity of $\boldsymbol{x}$ is ensured because $(I - \alpha \Delta t \mathrm{R}(\boldsymbol{x}))$ is an M-matrix for any α and Δt. It follows that $\boldsymbol{b} \geq 0$ because $\frac{2\alpha-1}{\alpha}$ and $\frac{1-\alpha}{\alpha}$ are positive. The positivity of $\boldsymbol{y}$, and thus, the positivity of $\boldsymbol{c}^{n+1}$ is again ensured by the fact that $(I - \alpha \Delta t \mathrm{R}(\boldsymbol{y}))$ is an M-matrix for any α and Δt.

To prove that the method is conservative we first show that $\|\boldsymbol{x}\|_1 = \|\boldsymbol{c}^n\|_1 = K$ pre-multiplying the first equation of (23) by $\boldsymbol{e}^T$ and exploiting the fact that $\boldsymbol{e} \in \mathrm{Ker}(\boldsymbol{\sigma}^T)$, and that both $\boldsymbol{x}$ and $\boldsymbol{c}^n$ are nonnegative. It follows that $\|\boldsymbol{b}\|_1 = K$, indeed

$$\|\boldsymbol{b}\|_1 = \frac{(2\alpha-1)\|\boldsymbol{c}_n\|_1 + (1-\alpha)\|\boldsymbol{x}\|}{\alpha} = \frac{2\alpha - 1 + 1 - \alpha}{\alpha} K = K.$$

Pre-multiplying the second equation in (23) by $\boldsymbol{e}^T$ we obtain that $\|\boldsymbol{y}\|_1 = K$, and, therefore, $\|\boldsymbol{c}^{n+1}\|_1 = K$. □

The two stages of the method involve the solution of non-linear problems that can be approximated by the iterative procedure proposed in [5]:

1. set $\boldsymbol{x}_0 = \boldsymbol{c}^n$
2. repeat $\boldsymbol{x}_{k+1} = (I - \alpha\Delta t \mathrm{R}(\boldsymbol{x}_k))^{-1}\boldsymbol{c}^n$ until convergence is achieved up to a given tolerance
3. set $\boldsymbol{y}_0 = \boldsymbol{c}^n$ and $\boldsymbol{b} = [(2\alpha-1)\boldsymbol{c}^n + (1-\alpha)\boldsymbol{x}_{k+1}]/\alpha$
4. repeat $\boldsymbol{y}_{j+1} = (I - \alpha\Delta t \mathrm{R}(\boldsymbol{y}_j))^{-1}\boldsymbol{b}$ until convergence is achieved up to a given tolerance
5. set $\boldsymbol{c}^{n+1} = \boldsymbol{y}_{j+1}$.

The advantage of employing the iterative procedure for the solution of (23) is that we avoid the solution of the two nonlinear problems.

Lemma 3.3 *Let us consider $\boldsymbol{c}^n > 0$, $\|\boldsymbol{c}^n\|_1 = K$. We also assume that $\mathrm{R}^{n+1}(\boldsymbol{x})$ has continuous derivatives for all $\boldsymbol{x} \in \Gamma$, where $\Gamma = \{\boldsymbol{y} \in \mathbb{R}^N : \quad \|\boldsymbol{y}\|_1 = K \quad \boldsymbol{y} \geq 0\}$. Then, if Δt is sufficiently small, for any $\boldsymbol{x}_0 \in \Gamma$ the sequence generated by the fixed point iteration*

$$\boldsymbol{x}_{k+1} = \boldsymbol{\varphi}_1(\boldsymbol{x}_k), \quad k = 0, 1, \ldots,$$

where $\boldsymbol{\varphi}_1 = (I - \alpha\Delta t R(\boldsymbol{x}_k))^{-1}\boldsymbol{c}^n$, converges to a fixed point $\overline{\boldsymbol{x}} \geq 0$ with $\|\overline{\boldsymbol{x}}\|_1 = K$. Furthermore, any element of the sequence satisfies $\|\boldsymbol{x}_k\|_1 = K$ and $\boldsymbol{x}_k \geq 0$. Moreover, for any $\boldsymbol{y}_0 \in \Gamma$ the sequence generated by the iteration

$$\boldsymbol{y}_{j+1} = \boldsymbol{\varphi}_2(\boldsymbol{y}_j), \quad j = 0, 1, \ldots,$$

where $\boldsymbol{\varphi}_2 = (I - \alpha\Delta t R(\boldsymbol{y}_j))^{-1}\boldsymbol{b}$ with $\boldsymbol{b} = \dfrac{(2\alpha-1)\boldsymbol{c}^n + (1-\alpha)\overline{\boldsymbol{x}}}{\alpha}$, converges to a fixed point $\overline{\boldsymbol{y}} \geq 0$ with $\|\overline{\boldsymbol{y}}\|_1 = K$, and any element of the sequence satisfies $\|\boldsymbol{y}_j\|_1 = K$ and $\boldsymbol{y}_j \geq 0$.

Proof. The fact that $\boldsymbol{x}_k \geq 0$ and $\|\boldsymbol{x}_k\|_1 = K$ is immediately inferred from the proof of Theorem 3.3.

To show that the fixed point iteration $\boldsymbol{x}_{k+1} = \varphi(\boldsymbol{x}_k)$ converges we define $\|\nabla\varphi_1\|_1$ as in (22). We have that

$$\nabla\varphi_1 = \alpha\Delta t(I - \alpha\Delta t R(\boldsymbol{x}))^{-1}\nabla R(\boldsymbol{x})$$

Under the condition $\Delta t < (\alpha \max_{\in\Gamma} \|R(\boldsymbol{x})\|_1)^{-1}$,

$$\|\nabla\varphi_1\|_1 \leq \alpha\Delta t \frac{\|\nabla R(\boldsymbol{x})\|_1}{1 - \alpha\Delta t\|R(\boldsymbol{x})\|}.$$

The first iterative procedure converges if $\sup_{x\in\Gamma} \|\nabla\varphi_1(\boldsymbol{x})\|_1 < 1$ i.e. if

$$\Delta t < \frac{1}{\alpha(\|\nabla R(\boldsymbol{x}_k)\|_1 + \|R(\boldsymbol{x}_k)\|_1)} \qquad \forall k.$$

As concerns the second fixed point iteration we have that

$$\nabla\varphi_2 = \alpha\Delta t(I - \alpha\Delta t R(\boldsymbol{y}))^{-1}\nabla R(\boldsymbol{y}),$$

thus the sequence converges to the fixed point $\boldsymbol{c}^{n+1}$ if

$$\Delta t < \frac{1}{\alpha(\|\nabla R(\boldsymbol{y}_j)\|_1 + \|R(\boldsymbol{y}_j)\|_1)} \qquad \forall j.$$

□

3.3 Patankar type methods

The method proposed by Patankar in [20] is suitable for those dynamic systems that can be described by production-destruction equations, that is

$$\frac{dc_i}{dt} = p_i(t, \boldsymbol{c}) - d_i(t, \boldsymbol{c}), \quad i = 1, ..., N,$$

where p_i and d_i are the rates of production and destruction of the i-th component, respectively.

In the case of our interest the set of chemical reactions can be re-written as

$$\frac{dc_i}{dt} = \sum_{\substack{j=1 \\ \sigma_{ij}>0}}^{M} \sigma_{ij} r_j(t, \boldsymbol{c}) - \sum_{\substack{j=1 \\ \sigma_{ij}<0}}^{M} |\sigma_{ij}| r_j(t, \boldsymbol{c}), \quad i = 1, ..., N. \tag{24}$$

which allows to identify p_i and d_i. If we split the stoichiometric coefficients matrix $\boldsymbol{\sigma}$ into $\boldsymbol{\sigma}^+$ and $\boldsymbol{\sigma}^-$, containing respectively the positive coefficients and the modulus of the negative ones, and indicate with $\boldsymbol{p}$ and $\boldsymbol{d}$ the vectors of components p_i and d_i respectively, we have

$$\boldsymbol{p}(\boldsymbol{c}) = \boldsymbol{\sigma}^+\boldsymbol{r}(\boldsymbol{c}), \qquad \boldsymbol{d}(\boldsymbol{c}) = \boldsymbol{\sigma}^-\boldsymbol{r}(\boldsymbol{c}).$$

Being the system is conservative, the global production rate must equal the global destruction rate,

$$\sum_{i=1}^{N} (p_i(\boldsymbol{c}) - d_i(\boldsymbol{c})) = 0.$$

Indeed,

$$\sum_{i=1}^{N}(p_i(\boldsymbol{c})-d_i(\boldsymbol{c}))=\boldsymbol{e}^T(p_i(\boldsymbol{c})-d_i(\boldsymbol{c}))=\boldsymbol{e}^T(\boldsymbol{\sigma}^+-\boldsymbol{\sigma}^-)\boldsymbol{r}(\boldsymbol{c})=\boldsymbol{e}^T\boldsymbol{\sigma}\boldsymbol{r}(\boldsymbol{c})=0.$$

Moreover, p_i and d_i can be written as the sum of N contributions:

$$p_i=\sum_{j=1}^{N}p_{ij},\qquad d_i=\sum_{j=1}^{N}d_{ij},$$

where in a conservative system p_{ij} and d_{ij} satisfy $p_{ij}=d_{ji}$ for $i\neq j$. If we define $\tilde{\boldsymbol{\sigma}}^+$, $\tilde{\boldsymbol{\sigma}}^-$ as follows

$$\tilde{\boldsymbol{\sigma}}_{ij}^{\pm}=\begin{cases}\dfrac{\boldsymbol{\sigma}_{ij}^{\pm}}{\sigma_{ij}} & \text{if } \sigma_{ij}\neq 0,\\ 0 & \text{else}\end{cases}$$

the matrices P and D, of components p_{ij} and d_{ij} respectively, can be easily computed as

$$P=\tilde{\boldsymbol{\sigma}}^+\mathrm{diag}(\boldsymbol{r})\tilde{\boldsymbol{\sigma}}^{-^T}\text{and}\quad D=\tilde{\boldsymbol{\sigma}}^-\mathrm{diag}(\boldsymbol{r})\tilde{\boldsymbol{\sigma}}^{+^T}.$$

The methods presented here exploit the so called Patankar trick, called by Patankar *source term linearisation*. It ensures unconditional positivity by weighting the destruction terms d_i with a factor $\frac{c_i^{n+1}}{c_i^n}$. To obtain a conservative method the original Patankar scheme has to be modified, as proposed in [9], introducing a weighting of the production terms as well.

A second order method based on the application of the Patankar trick to a two-stage Runge Kutta method has been proposed in [9]. The Modified Patankar Runge Kutta (MPRK) scheme reads

$$\begin{cases} y_i-\Delta t\left(\displaystyle\sum_{j=1}^{N}p_{ij}(t^n,\boldsymbol{c}^n)\frac{y_j}{c_j^n}-\sum_{j=1}^{N}d_{ij}(t^n,\boldsymbol{c}^n)\frac{y_i}{c_i^n}\right)=c_i^n,\quad i=1,...,N\\ c_i^{n+1}-\dfrac{\Delta t}{2}\left(\displaystyle\sum_{j=1}^{N}\left(p_{ij}(t^n,\boldsymbol{c}^n)+p_{ij}(t^{n+1},\boldsymbol{y})\right)\frac{c_j^{n+1}}{y_j}\right.\\ \qquad\left.-\displaystyle\sum_{j=1}^{N}\left(d_{ij}(t^n,\boldsymbol{c}^n)+d_{ij}(t^{n+1},\boldsymbol{y})\right)\frac{c_i^{n+1}}{y_i}\right)=c_i^n\quad i=1,...,N.\end{cases}\tag{25}$$

Each of the two stages involve the solution of a linear system. The method is second order accurate, conservative and unconditionally positive. For the proof we refer to [9].

3.4 A third order predictor-corrector method

We suggest here an original integration scheme based on the predictor-corrector strategy to improve the convergence of the MPRK method while preserving its positivity and conservation properties. The scheme, denoted by PCMP (Predictor-Corrector Modified Patankar) computes an approximation c_i^* by using the MPRK method

$$\begin{cases} y_i - \Delta t \left(\displaystyle\sum_{j=1}^{N} p_{ij}(t^n, \boldsymbol{c}^n) \frac{y_j}{c_j^n} - \sum_{j=1}^{N} d_{ij}(t^n, \boldsymbol{c}^n) \frac{y_i}{c_i^n} \right) = c_i^n, \quad i = 1, ..., N \\ c_i^* - \dfrac{\Delta t}{2} \left(\displaystyle\sum_{j=1}^{N} \left(p_{ij}(t^n, \boldsymbol{c}^n) + p_{ij}(t^{n+1}, \boldsymbol{y}) \right) \frac{c_j^*}{y_j} \right. \\ \qquad \left. - \displaystyle\sum_{j=1}^{N} \left(d_{ij}(t^n, \boldsymbol{c}^n) + d_{ij}(t^{n+1}, \boldsymbol{y}) \right) \frac{c_i^*}{y_i} \right) = c_i^n, \quad i = 1, ..., N \end{cases} \tag{26}$$

and then it sets

$$\text{(a)} \quad c_i^{(n+1)} = c_i^* \qquad \text{if } n < 2 \quad \text{or} \quad z_i^n < 0 \tag{27a}$$

$$\text{(b)} \quad c_i^{(n+1)} = z_i^n + \frac{6}{11} \Delta t \left(\sum_{j=1}^{N} p_{ij}(t^{n+1}, \boldsymbol{c}^*) \frac{c_j^{(n+1)}}{c_j^*} - \sum_{j=1}^{N} d_{ij}(t^{n+1}, \boldsymbol{c}^*) \frac{c_i^{(n+1)}}{c_i^*} \right), \quad \text{otherwise,} \tag{27b}$$

where $\boldsymbol{z} = \beta_1 \boldsymbol{c}^n + \beta_2 \boldsymbol{c}^{n-1} + \beta_3 \boldsymbol{c}^{n-2}$, and $\beta_1 = \frac{18}{11}$, $\beta_2 = -\frac{9}{11}$, $\beta_3 = \frac{2}{11}$. The PCMP scheme uses the MPRK scheme as a predictor, while the corrector step is based on the well know three-steps BDF method, where the forcing term $\boldsymbol{f}^{n+1}$ is substituted by $\boldsymbol{f}(\boldsymbol{c}^*)$ and the production and destruction terms in $\boldsymbol{f}$ are weighted as proposed by [9] to make the corrector step positive and conservative.

Theorem 3.4 *The PCMP method is unconditionally positive, conservative and has at least order 2 and at the best order 3.*

Proof. The positivity of the MPRK method ensures that $\boldsymbol{c}^* \geq 0$. In the case (a) we have that $\boldsymbol{c}^{n+1} = \boldsymbol{c}^* \geq 0$. To examine case (b) it is useful to write system (27) in matrix form

$$A \boldsymbol{c}^{n+1} = \boldsymbol{z}. \tag{28}$$

where A is a matrix whose entries are

- $a_{ii} = 1 + \dfrac{6}{11} \Delta t \dfrac{d_i(t^{n+1}, \boldsymbol{c}^*)}{c_i^*} > 0 \quad i = 1, ..., N$
- $a_{ij} = -\dfrac{6}{11} \Delta t \dfrac{p_{ij}(t^{n+1}, \boldsymbol{c}^*)}{c_j^*} \leq 0 \quad i, j = 1, ..., N, \; i \neq j.$

Thus, A has nonpositive off-diagonal entries and positive column sum, indeed

$$1+\frac{6}{11}\Delta t(\frac{d_j}{c_j^*}-\sum_{i=1,\,i\neq j}^{N}\frac{p_{ij}}{c_j^*})=1+\frac{6}{11}\Delta t(\frac{d_j}{c_j^*}-\sum_{i=1,\,i\neq j}^{N}\frac{d_{ji}}{c_j^*})>0.$$

Therefore A^T is a M-matrix, while the right hand side of the system is positive by construction, therefore $\boldsymbol{c}^{n+1}\geq 0$.

The conservation properties of the MPRK method ensure that $\|\boldsymbol{c}^*\|_1=\|\boldsymbol{c}_0\|_1$. To prove the conservativity of the PCMP method it is enough to show that $\|\boldsymbol{c}_{n+1}\|_1=\|\boldsymbol{c}^*\|_1$.

The proof is trivial in case (a). For case (b), let us sum both members of (27) over $i=1,...,N$ obtaining

$$\begin{aligned}\|\boldsymbol{c}^{(n+1)}\|_1=&\beta_1\|\boldsymbol{c}^n\|_1+\beta_2\|\boldsymbol{c}^{n-1}\|_1+\beta_3\|\boldsymbol{c}^{n-2}\|_1+\\&\frac{6}{11}\Delta t\sum_{i=1}^{N}\left(\sum_{j=1}^{N}p_{ij}(\boldsymbol{c}^*)\frac{c_j^{(n+1)}}{c_j^*}-\sum_{j=1}^{N}d_{ij}(\boldsymbol{c}^*)\frac{c_i^{(n+1)}}{c_i^*}\right)\end{aligned}\tag{29}$$

Since we have initialized the multistep method with the MPRK method which is conservative

$$\|\boldsymbol{c}^{(n+1)}\|_1-\|\boldsymbol{c}_0\|_1=\frac{6}{11}\Delta t\sum_{i=1}^{N}\left(\sum_{j=1}^{N}p_{ij}(\boldsymbol{c}^*)\frac{c_j^{(n+1)}}{c_j^*}-\sum_{j=1}^{N}d_{ij}(\boldsymbol{c}^*)\frac{c_i^{(n+1)}}{c_i^*}\right),\tag{30}$$

and with the same argument used in [9], thanks to symmetry of the production and destruction terms, the right hand side vanishes, thus the corrector step is conservative.

As concerns the start up of the method it is formally sufficient to use a second order method to compute $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$ to keep the global third order accuracy. However, if the reactions are particularly fast at the start, it has been found that accurate results are obtained in practise by adopting a third order method for the start up.

□

In the proposed formulation the start-up of the method is only of order two. In spite of this the method shows better accuracy with respect to the MPRK method when applied to problems that admit a slowly varying solution at the initial time. In cases with a fast evolution in a neighborhood of t_0 an initialization of order two can result in a loss of accuracy of the PCMC method. The difficulty can be circumvented computing the first three steps with a third order method, under a restriction on the step size for positivity. It has been shown [6] that a standard Runge Kutta method can be unconditionally positive only if it has order 1. The size step thresholds for the positivity of Runge Kutta methods are discussed in [14, 15], where the limit step amplitude is expressed as a function of the absolute monotonicity radius $R(A,b)$ of the method, and of the right hand side of the problem at hand. It is advisable to employ a Runge Kutta method with $R(A,b)>0$, for instance the RK3 defined by Butcher array

$$\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} \\ \hline \frac{1}{6} & \frac{1}{6} & \frac{2}{3} & \end{array}$$

has $R(A, b) = 1$. We will show by numerical experiments that for adequate values of Δt the PCMP method initialized with a third order RK has better accuracy for practical values of Δt than the standard PCMP.

4 Numerical tests

We want to test the performances of the methods listed in the previous section. We will first consider a synthetic case to assess experimentally the order and computational cost of the methods, then we will solve realistic cases with increasing complexity.

The synthetic case is given by:

$$\begin{aligned} 2A_1 &\rightarrow A_2 + A_3 \\ A_2 &\rightarrow A_3 \\ 2A_2 &\rightarrow A_1 + A_3 \end{aligned}$$

described by the following system of ODEs.

$$\frac{d\boldsymbol{c}}{dt} = \boldsymbol{\sigma}\boldsymbol{r} \tag{32}$$

with

$$\boldsymbol{\sigma} = \begin{bmatrix} -2 & 0 & 1 \\ 1 & -1 & -2 \\ 1 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{r} = \begin{bmatrix} c_1^2 \\ c_2 \\ c_2^2 \end{bmatrix} \tag{33}$$

To compare the performances of different integration schemes we consider as the exact solution the numerical solution obtained with the MPRK method and with $\Delta t = 0.0002$ (figure 12).

The error is defined as follows

$$e = \frac{1}{N_{step}} \sum_{k=1}^{N_{step}} \|c_i(t^k) - c_i^k\|_\infty,$$

and the nonlinear problems are solved with fixed point iterations, arrested when a residual norm of the order of 10^{-10} is achieved.

Figure 2 shows the experimental order of convergence of the methods discussed above. The Explicit Euler method and the Implicit Euler method are

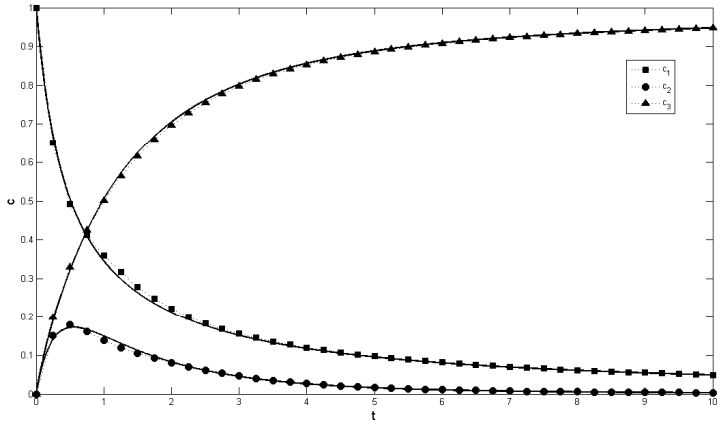


Figure 1: Reference solution of the test case (solid line), and PCMP method (40 time steps).

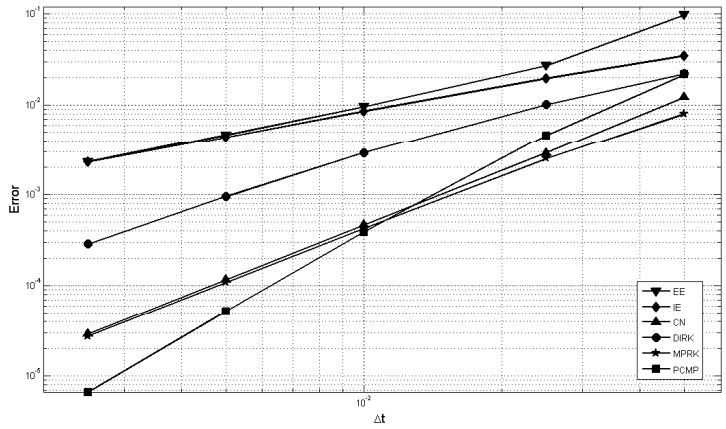


Figure 2: Convergence on a synthetic case a)

	Time $[ms]$	RHS evaluations
EI	33.5	924
CN	33.3	716
DIRK	80.8	1326
MPRK	12.8	200
PCMC	20.9	298

Table 1: Computational cost of the numerical methods

approximately of the first order, while the Crank Nicholson method, the DIRK method proposed in [5], and MPRK methods are of order two. In this simple case the introduction of the corrector step (PCMP method) enhances the convergence rate nearly up to three. As concerns the computational cost the results relative to a simulation with $\Delta t = 0.01$, summarized in table 1, show that the DIRK method is significantly more expensive with respect to the others. The MPRK method is instead very convenient, because it does not involve the solution of a nonlinear problem and thus requires a fixed number - two - of right hand side evaluations for each time step. The PCMP scheme requires at most three RHS evaluations per time step, so it is more expensive, yet it compares favorably with respect to the other methods such as Crank Nicholson and Implicit Euler.

Finally, we have observed that all methods conserve the mass up to machine error.

Based on these results we have selected the most efficient methods and applied them to more demanding test cases with realistic activation energies and time dependent temperature as an input. In the first case we are considering the following set of reactions, with the same activation energy but different frequency factors A_j,

$$A_1 \rightarrow 0.45A_2 + 0.5A_3 + 0.05A_4 \tag{34a}$$
$$A_3 \rightarrow 0.36A_2 + 0.54A_4 + 0.1A_5. \tag{34b}$$

at three different temperatures, $T_1 = 130°C$, $T_2 = 140°C$, $T_3 = 150°C$ on the time interval $[0, 500s]$. To compute the error we consider as the exact solution the numerical solution with $\Delta t = 0.0002$. In this case the reactions are of the first order and lead to a linear system, i.e. the matrix $R(t, \boldsymbol{c})$ is constant w.r.t. $\boldsymbol{c}$ and reads

$$R = \begin{bmatrix} -k_1 & 0 & 0 & 0 & 0 \\ 0.45k_1 & 0 & 0.36k_2 & 0 & 0 \\ 0.50k_1 & 0 & -k_2 & 0 & 0 \\ 0.05k_1 & 0 & 0.54k_2 & 0 & 0 \\ 0 & 0 & 0.1k_2 & 0 & 0 \end{bmatrix}. \tag{35}$$

For $T = T_1$ the reactions are both slow and a good accuracy is obtained even with large time steps: figure 3 compares the reference solution with the results of PCMP with 20 time steps. Figure 4 shows the convergence of the implicit Euler method, with order 1, the Crank Nicholson and MPRK method, both of order 2, and the PCMP method which in this case is nearly of order 3.

For $T = T_2$ the two reactions, and the first in particular, are faster, thus there is a loss of accuracy for large time steps in the first part of the simulations (see figure 5). The error graph (figure 6) shows that the PCMP method is better than the MPRK method only for small time steps.

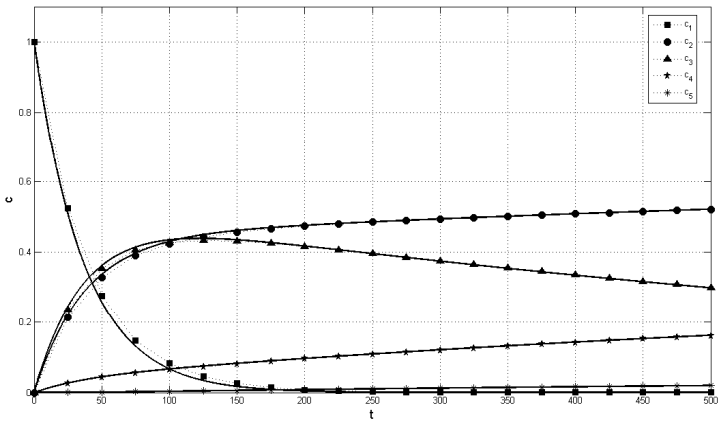


Figure 3: Problem (34), $T = T_1$, reference solution (solid line) and PCMP, 20 time steps

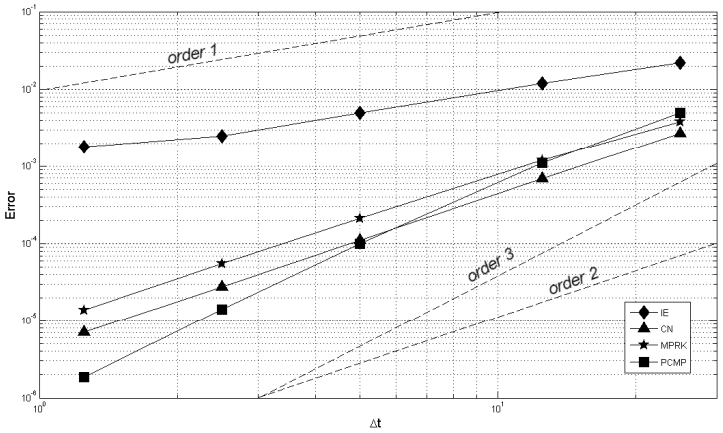


Figure 4: Convergence of some numerical schemes applied to problem (34) with $T = T_1$

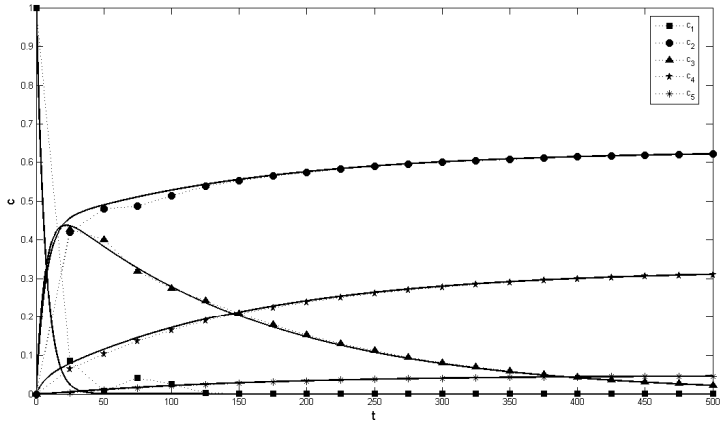


Figure 5: Problem (34), $T = T_2$, reference solution (solid line) and PCMP, 20 time steps

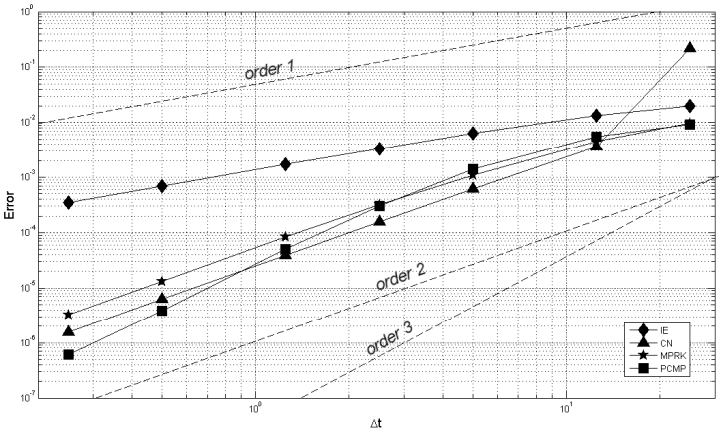


Figure 6: Convergence of some numerical schemes applied to problem (34) with $T = T_2$

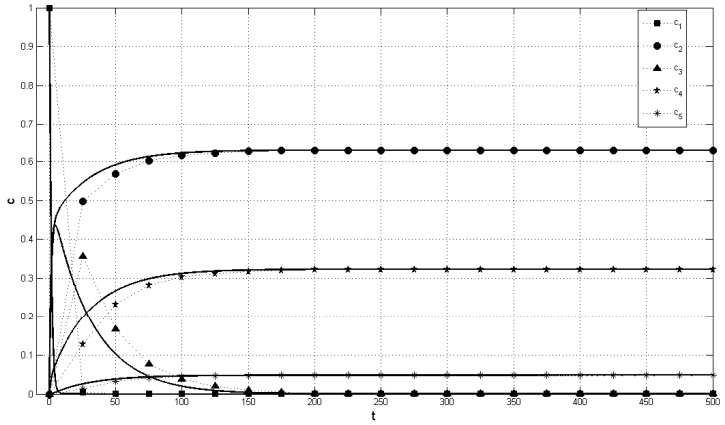


Figure 7: Problem (34), $T = T_3$, reference solution (solid line) and PCMP, 20 time steps

For $T = T_3$ the reactions are fast and the first in particular is exhausted after 1/10 of the simulation (see figure 7). It can be observed that the Crank Nicholson method fails for large time steps, producing artificial oscillations and negative values. As concerns the error, see figure 8, the order of convergence are those predicted by the theory only for Δt small enough. It should be stressed that the results could be further improved by time step adaptivity, and multirate techniques like those presented in [11, 16], are also advisable if very different reaction rates coexist. However, the analysis of these techniques is beyond the scope of this work.

We have also applied to this problem the PCMP method initialized with a third order Runge Kutta (we denote this modified scheme as PCMP*), limited to the time step amplitudes that ensure the positivity of the RK method. Figure 9 compares the error obtained by the original PCMP scheme and of the PCMP* scheme applied to problem (34) for the three values of the temperature, showing that a better initialization of the method yields a better accuracy for large Δt in particular for cases (b) and (c) where the solution evolves faster in the first part of the simulation. A detail of the solution for $T = T_2$ is shown, for the two methods above, in figure 10.

For this simple case it is possible to compute the maximum value of the time step τ_1 that is required to ensure the positivity of a theta-method. According to the definition 17 $f_k = -k_1$ with $k = 1$, $c_k = 1$, thus, for the Crank Nicholson method ($\theta = 0.5$) we have $\tau_1 = \frac{2}{k_1}$. For $T = T_1$ we have that $\tau_1 \simeq 74s$, therefore for practical values of Δt the positivity is ensured. For $T = T_2$ and $T = T_3$ τ_1 becomes an actual constraint, indeed the time step must be lower than $13.5s$ and $2.7s$ respectively to have a positive method. Figure 11 shows the positivity

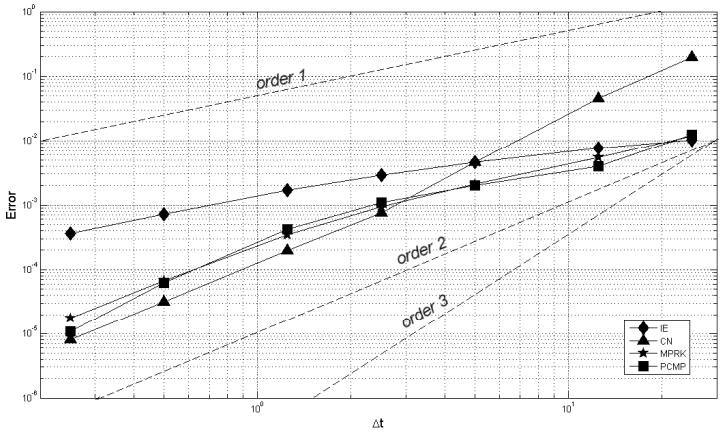
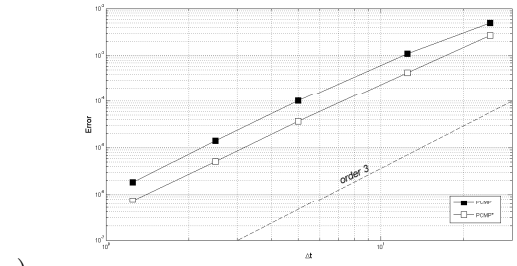
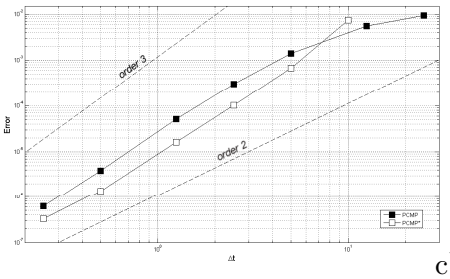


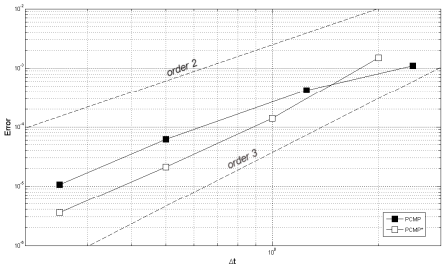
Figure 8: Convergence of some numerical schemes applied to problem (34) with $T = T_3$



a)



b)



c)

Figure 9: Convergence of the PCMP, initialized with a second order method, and the PCMP*, initialized with a third order method for $T = T_1$ (a), $T = T_2$ (b), $T = T_3$ (c)

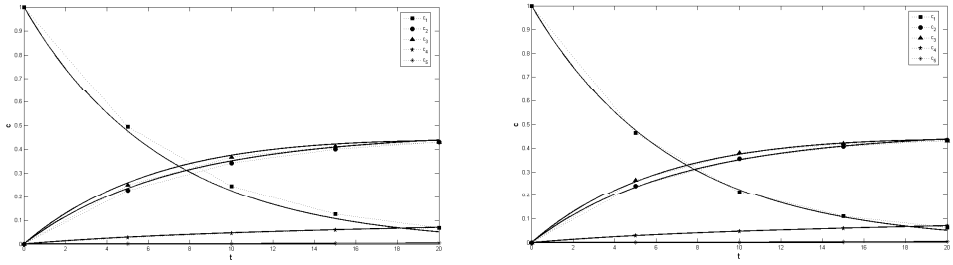


Figure 10: A detail of the numerical solution, $T = T_2$, for the PCMP scheme (left) and the PCMP* scheme (right).

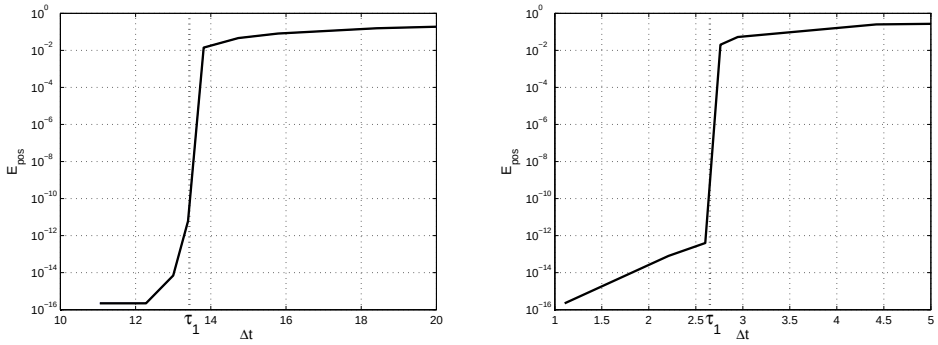


Figure 11: Positivity error for the Crank Nicholson method applied to the case 34 for $T = 140°C$ (left) and $T = 150°C$ (right) with different time-steps, highlighting the limit value τ_1.

error, defined as

$$e_{pos} = \min_{k=1}^{N_{step}} \min_{k=1}^{N} c_i^k \tag{36}$$

for different time step amplitudes, highlighting the change in the behavior of the method in correspondence of $\Delta t = \tau_1$. This confirms the theoretical analysis.

Finally, we consider a detailed system of 18 species (denoted by A1-A18) and 15 reactions,

$$
\begin{aligned}
&A_1 \to A_{18} && \\
&A_2 \to A_{17} && A_{11} \to 0.2A_9 + 0.8A_{17} \\
&A_3 \to A_{13} && A_{12} \to A_{11} \\
&A_4 \to A_{11} && A_{13} \to 0.15A_9 + 0.05A_{10} + 0.8A_{17} \\
&A_5 \to A_{14} && A_{14} \to A_{13} \\
&A_6 \to A_{12} && A_{15} \to A_{14} \\
&A_7 \to A_{15} && A_{16} \to A_{12} \\
&A_8 \to A_{16} && A_{17} \to 0.3A_9 + 0.7A_{18}
\end{aligned}
$$

For a fully realistic description each reaction is characterized not by a single value of the activation energy, but by a distribution on energies, to account for the fact that a fraction of the reactant may have an activation energy that is higher or lower than the average value. Each reaction is thus split in several parallel reactions, often denoted as "channels", which proceed with different speeds for a total of 221 equations. The numerical solution is reported in figure 12.

The wide range of reaction speeds makes the ODE system very stiff, thus the problem can be regarded as a hard test-bed for the integration schemes we have analyzed so far. For what concerns the conservation properties all the methods respect mass conservation with errors ranging from $10^{-7}\%$ to $10^{-5}\%$. Moreover, if we define the positivity error as in36 all the methods have identically null positivity errors, except the Crank-Nicholson method which is only conditionally positive. Figure 13 compares the time required to run the simulation with Crank Nicholson and PCMP the error being equal. The results show the advantage of using a method that is accurate but requires few matrix inversions on a large system of ODEs.

5 Conclusions

The behavior of a reacting system can be described by a set of ordinary differential equations whose characteristics depend on the stoichiometric coefficients of the reactions involved and on the function that models the reaction rates. We have provided an original proof of the global existence and uniqueness of the solution under fairly general regularity assumptions on the reaction rates r_i. Moreover we have shown that the solution is nonnegative for all $t > 0$. Numerical tests have highlighted that integration schemes can give unexpectedly inaccurate results if the positivity of the solution is not preserved by the numerical method, thus it is important to choose a method that is at the same time positive and conservative. Among the second order, unconditionally positive and conservative schemes we have analyzed the MPRK method is the most efficient for the applications of our interest. The PCMP method we have constructed

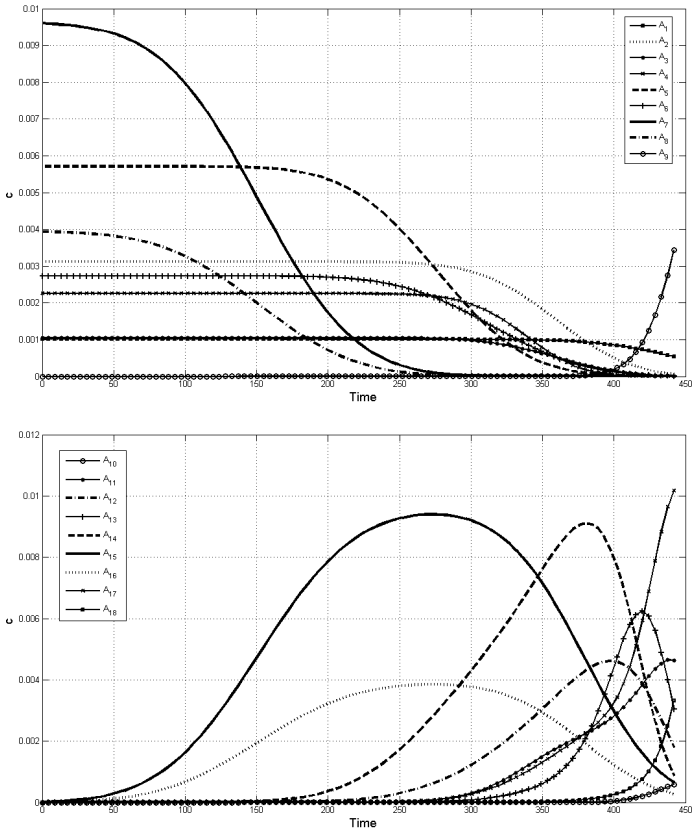


Figure 12: Numerical solution with 100 time steps, PCMP method

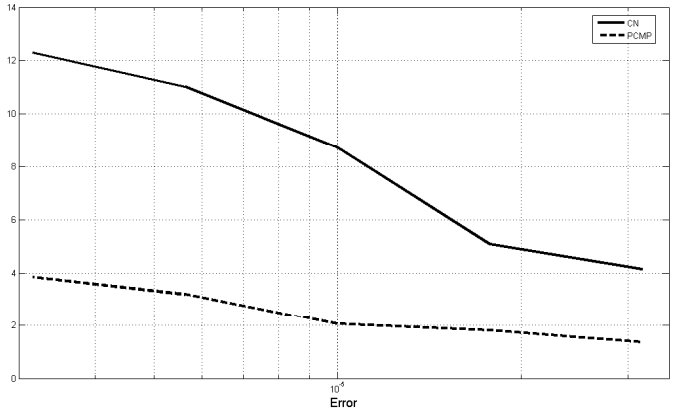


Figure 13: Computational time for the integration of the realistic system with the CN and PCMP methods, the error being the same.

combining the MPRK method with a corrector step is unconditionally positive and conservative, and can enhance the order of the method up to three at most. Moreover, it is computationally convenient on heavy realistic test cases.

6 Acknowledgement

This work has been supported by Eni S.p.A.

References

[1] L. Torelli A. Bellen, *Unconditional contractivity in the maximum norm of diagonally split Runge–Kutta methods*, Applied Numerical Mathematics **34** (1997), 528–543.

[2] R. Alexander, *Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s*, SIAM Journal on Numerical Analysis **14** (1977), no. 6, 1006–1021.

[3] A.Sandu, *Positive numerical methods for chemical kinetic systems*, J.Comput.Phys **170** (2001), 589 – 602.

[4] P. W. Atkins and J. De Paula, *Physical chemistry*, W. H. Freeman, 2002.

[5] E. Bertolazzi, *Positive and conservative schemes for mass action kinetics*, Computers Math. Applic. **32** (1996), no. 6, 29–43.

[6] C. Bolley and M. Crouzeix, *Conservation de la positivité lors de la discrétisation des problèmes d'evolution parabolic*, RAIRO Anal. Numer. **12** (1978), no. 3, 237–245.

[7] N. Broekhuizen, G. J. Rickard, J. Bruggeman, and A. Meister, *An improved and generalized second order, unconditionally positive, mass conserving integration scheme for biochemical systems*, Applied Numerical Mathematics **58** (2008), no. 3, 319–340.

[8] J. Bruggeman, H. Burchard, B. W. Kooi, and B. Sommeijer, *A second-order, unconditionally positive, mass-conserving integration scheme for biochemical systems*, Applied Numerical Mathematics **57** (2007), no. 1, 36–58.

[9] H. Burchard, E. Deleersnijder, and A. Meister, *A high-order conservative Patankar-type discretisation for stiff systems of production-destruction equations*, SIAM J. Numer. Anal. **47** (2003), no. 1, 1–30.

[10] P. Deuflhard and F.Bornemann, *Scientific computing with ordinary differential equations*, Springer, 2002.

[11] C. W. Gear and D. R. Wells, *Multirate linear multistep methods*, BIT **24** (1984), no. 4, 484–502.

[12] A. Halanay and V.Råsvan, *Application of Liapunov method in stability*, Kluver Academic Publishers, 1993.

[13] M. W. Hirsch and S. Smale, *Differential equations, dynamical systems and linear algebra*, Academic press, 1974.

[14] Z. Horváth, *Positivity of Runge-Kutta and diagonally split Runge-Kutta methods*, SIAM Journal on Numerical Analysis **28** (1998), 309 – 326.

[15] ———, *On the positivity step size threshold of Runge-Kutta methods*, Applied Numerical Mathematics **53** (2005), 341356.

[16] W. Hundsdorfer and V. Savcenco, *Analysis of a multirate theta-method for stiff odes*, Applied Numerical Mathematics **59** (2009), 693–706.

[17] A. D. McNaught and A. Wilkinson, *Iupac compendium of chemical terminology*, WileyBlackwell, 1997.

[18] R. E. Mickens, *Nonstandard finite difference models of differential equations*, World Scientific, Singapore, 1994.

[19] ———, *Calculation of denominator functions for nonstandard finite difference schemes for differential equations satisfying a positivity condition*, Numer. Methods Partial Differential Equations **23** (2007), no. 3, 672–691.

[20] S.V. Patankar, *Numerical heat transfer and fluid flow*, Series in Computational Methods in Mechanics and Thermal Sciences, McGraw-Hill, 1980.

[21] A. Quarteroni, R.Sacco, and F.Saleri, *Numerical mahematics*, Springer, 2006.

[22] A. Sandu, *Time-stepping methods that favor positivity for atmospheric chemistry modeling*, Atmospheric modeling (G. R. Carmichael D. P. Chock, ed.), IMA Vol. Math. Appl., 130, Springer New York, 2002, pp. 21–37.

[23] N.N. Pham Thi, W. Hundsdorfer, and P. Sommeijer, *Positivity for explicit two-step methods in linear multistep and one-leg form*, BIT Numerical Mathematics **46** (2006), no. 4, 875–882.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

16/2010 Luca Formaggia, Anna Scotti:
Positivity and conservation properties of some integration schemes for mass action kinetics

15/2010 Alfio Quarteroni, Luca Formaggia:
Domain Decomposition (DD) methods

14/2010 Paola F. Antonietti, Lourenco Beirão da Veiga, Marco Verani:
A Mimetic Discretization of Elliptic Obstacle Problems

13/2010 G.M. Porta, Simona Perotto, F. Ballio:
A Space-Time Adaptation Scheme for Unsteady Shallow Water Problems

12/2010 Riccardo Sacco, Paola Causin, Paolo Zunino, Manuela T. Raimondi:
A multiphysics/multiscale numerical simulation of scaffold-based cartilage regeneration under interstitial perfusion in a bioreactor

11/2010 Paolo Biscari, Sara Minisini, Dario Pierotti, Gianmaria Verzini, Paolo Zunino:
Controlled release with finite dissolution rate

10/2010 Alfio Quarteroni, Ricardo Ruiz Baier:
Analysis of a finite volume element method for the Stokes problem

09/2010 Laura M. Sangalli, Piercesare Secchi, Simone Vantini, Valeria Vitelli:
Joint Clustering and Alignment of Functional Data: an Application to Vascular Geometries

08/2010 Francesca Ieva, Anna Maria Paganoni:
Multilevel models for clinical registers concerning STEMI patients in a complex urban reality: a statistical analysis of MOMI^2 survey

07/2010 Laura M. Sangalli, Piercesare Secchi, Simone Vantini, Valeria Vitelli:
Functional clustering and alignment methods with applications