S0168-9274(05)00224-2/FLA AID:1854 Vol.●●●(●●●)
APNUM:m3SC+ v 1.50 Prn:12/01/2006; 12:16
apnum1854
[+model] P.1(1-23)
by:R.M. p. 1

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Applied Numerical Mathematics ●●● (●●●●) ●●●–●●●

APPLIED NUMERICAL MATHEMATICS

www.elsevier.com/locate/apnum

A second-order, unconditionally positive, mass-conserving integration scheme for biochemical systems

Jorn Bruggeman [a,*], Hans Burchard [b], Bob W. Kooi [a], Ben Sommeijer [c]

[a] *Department of Theoretical Biology, Faculty of Earth & Life Sciences, Vrije Universiteit, de Boelelaan 1087, 1081 HV Amsterdam, The Netherlands*

[b] *Baltic Sea Research Institute Warnemünde, Seestraße 15, D-18119 Rostock-Warnemünde, Germany*

[c] *CWI, National Research Institute for Mathematics and Computer Science, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands*

Abstract

Biochemical systems are bound by two mathematically-relevant restrictions. First, state variables in such systems represent non-negative quantities, such as concentrations of chemical compounds. Second, biochemical systems conserve mass and energy. Both properties must be reflected in results of an integration scheme applied to biochemical models. This paper first presents a mathematical framework for biochemical problems, which includes an exact definition of biochemical conservation: elements and energy, rather than state variable units, are conserved. We then analyze various fixed-step integration schemes, including traditional Euler-based schemes and the recently published modified Patankar schemes, and conclude that none of these deliver unconditional positivity and biochemical conservation in combination with higher-order accuracy. Finally, we present two new fixed-step integration schemes, one first-order and one second-order accurate, which do guarantee positivity and (biochemical) conservation.

Keywords: Patankar-type schemes; Runge–Kutta methods; Unconditional positivity; Mass conservation

1. Introduction

Biochemistry holds an interesting niche for numerical mathematics, as it combines a great need for numerical techniques with model systems that are constrained by mathematically-relevant, real-world restrictions. Such restrictions are a direct consequence of the nature of biochemical state variables: these represent combinations of mass and/or energy, both quantities that cannot become negative, nor created or destroyed (as specified by the first law of thermodynamics) [20]. The former property implies state variables in biochemical systems are unconditionally positive. The latter imposes a type of conservation. These properties represent two of the few cornerstones in biochemistry, and are explicitly respected by any consistent biochemical modelling approach [12]. Integration schemes must not violate either positivity or conservation, if they are to produce results that are meaningful in biochemical context.

* Corresponding author.
E-mail addresses: jorn.bruggeman@falw.vu.nl (J. Bruggeman), hans.burchard@io-warnemuende.de (H. Burchard), bob.kooi@falw.vu.nl (B.W. Kooi), b.p.sommeijer@cwi.nl (B. Sommeijer).

0168-9274/$30.00
doi:10.1016/j.apnum.2005.12.001

The exact mathematical implications of biochemical restrictions are often not well understood. In particular, treatments of biochemical conservation in integration schemes abound [22,21,3], but are often limited at best. Various proposed definitions of conservation are tailored to simple biochemical systems [3], and would be demonstrably incorrect for many more advanced systems. Most other definitions forgo – as argued in this paper – the exact meaning of biochemical conservation [21], and should be considered too lenient for application to biochemical problems. In Section 2, we introduce a mathematical framework for biochemical systems that combines ideas from metabolic control analysis [19] and theoretical biology [10,7]. We treat a minimal set of biochemical concepts necessary to derive an exact definition of biochemical conservation. This ultimately renders conditions for conservation as well as positivity, against which any integration scheme can be tested.

The need for unconditionally positive schemes may not be obvious, as non-positive integration schemes can render solutions that – though negative – approach the true solution well. Additionally, these schemes recover from negative values in some cases. For many biochemical systems, however, this argument is not valid. Such systems include ODEs such as $\mathrm{d}y/\mathrm{d}t = -y/(y+k)$ with $k > 0$ (for substrate consumed in enzymatic reactions, or prey taken by predators), or $\mathrm{d}z/\mathrm{d}t = -z^2$ (when two molecules of the same compound react, or species mortality is density-dependent). Obviously, $y = 0$ and $z = 0$ are system invariants. Given positive initial values, both y and z converge to 0. For negative initial values, this is not guaranteed: for $y_0 < -k$ and $z_0 < 0$, state variables will approach negative infinity. In other words, crossing the t-axis can cause convergence to an unrealistic attractor that would never be approached by the true solution. Hence, integration schemes that allow negative values cannot guarantee consistency with the original system of ODEs.

Many schemes obtain conditional positivity through adaptive time stepping. For instance, in fluid flow dynamics the use of Courant-type conditions to ensure positivity (as well as stability) is prevalent [17]. However, we aim to apply integration schemes to biochemical systems hosted in an existing biogeochemical modeling framework for water columns [2]. This comprehensive framework imposes a global time step; within a step, splitting schemes are applied in order to solve different parts of the problem – advection, diffusion, biochemistry – with different numerical methods. The framework was not designed with an adaptive time step in mind, and would, not surprisingly, require substantial modification to deal with such. Therefore, we will in this paper consider the framework as given, and exclusively deal with the biochemistry part of the problem. We are thus confronted with the task of solving the biochemical system for some predetermined time step. For such a scenario, non-adaptive schemes represent the most straightforward and easily implemented solution; adaptive time stepping – though possible if the adaptive step is an integer fraction of the global step, or interpolation is used – is somewhat of a hassle. Therefore, we restrict ourselves in this paper to integration schemes that are *un*conditionally positive and conservative.

Few integration schemes offer unconditional positivity without caveats. In their fundamental paper [1], Bolley and Crouzeix have shown that, within the class of traditional methods like linear multistep and Runge–Kutta methods, unconditional positivity restricts the order of the method to one. In [8] much attention is paid to this topic and it is surveyed how to arrive at conditionally positive methods by taking special starting values. This applies in particular to higher order BDF methods (also called Gear methods) which lack unconditional positivity due to negative coefficients (in spite of the excellent stability properties of these methods). Additionally, the first-order methods that satisfy the condition of unconditional positivity are often computationally expensive (e.g. backward Euler), and hence unsuitable when one values computational efficiency.

Mickens initiated the development of non-standard integration schemes [14], designed to preserve the physical properties of the original systems (in particular stability properties). For several systems, efficient, non-standard first-order schemes have been proposed that guarantee positivity of the solution [18,9]. However, for other systems, such schemes have not been constructed, although great effort has been put into their development. Therefore, non-standard schemes do not represent a definitive, generic solution for the condition of positivity. Another approach has been suggested by Sandu [21], and involves a projection method to get around the first-order barrier; however, its projection technique is founded on a common, 'macroscopic' definition of conservation that we show in Section 2 (Definition 7) to be insufficient for biochemical systems.

Burchard et al. [3] presented a collection of unconditionally positive schemes that are inspired by the so-called Patankar trick [16]. The collection includes the first-order accurate Modified Patankar (MP) scheme, and the second-order accurate Modified Patankar–Runge–Kutta (MPRK) scheme. Both the original Patankar scheme and the MP/MPRK schemes obtain unconditional positivity by treating the positive terms (sink fluxes) in the right-hand sides of ODEs differently from the negative terms (source fluxes). Unlike the original Patankar scheme, the modi-

fied Patankar schemes could be shown to satisfy a minimal definition of conservation [3]. In Section 2 of this paper, however, we demonstrate that this basic definition of conservation is unsuitable for many biochemical problems, and show that the modified Patankar schemes are conservative in the strict biochemical sense, only if certain, restrictive conditions are met.

In Section 4, we systematically analyze a selection of integration schemes, both traditional (Euler schemes) and recent (Modified Patankar schemes), and prove that none of these satisfy the requirements of unconditional positivity and (biochemical) conservation. In the same section, we propose two new fixed-step integration schemes, inspired by the Patankar trick [16] and the work of Burchard et al. [3], and prove that these schemes are unconditionally positive, and conservative in the biochemical sense. Finally, in Section 5, the accuracy, order, and computational cost of the new schemes is analyzed empirically with the two simple test cases described in Section 3.

2. Biochemical concepts

A generic system of I ordinary differential equations will be denoted by:

$$\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t}(t) = \mathbf{f}\big(t, \mathbf{c}(t)\big),$$

$\mathbf{c}(t)$ denoting the vector of length I with state variable values at time t, and $\mathbf{f}(t, \mathbf{c}(t))$ denoting the vector with ODE right-hand sides. Elements of $\mathbf{c}(t)$ and $\mathbf{f}(t, \mathbf{c}(t))$ will be denoted by $c_i(t)$ and $f_i(t, \mathbf{c}(t))$, respectively, $i \in \{0, \ldots, I\}$. For any vector $\mathbf{c}$, $\mathbf{c} > 0$ will be used to denote $c_i > 0,\ \forall i$.

For numerical schemes, the time at integration step n will be denoted by t^n. The time step will be denoted by Δt. As this paper deals only with schemes using a fixed time step, we have $t^{n+1} = t^n + \Delta t$ for any $n \in \mathbb{N}$. The numerical approximation of the solution vector $\mathbf{c}(t^n)$ will be denoted by $\mathbf{c}^n$.

2.1. A framework for biochemical systems

The typical biochemical system of I ordinary differential equations can be completely described by a set of R *reactions*. A reaction describes the conversion of a set of source compounds into a set of sink compounds. Compounds appear in ODE-based system definitions as state variables.

Take for instance the four-dimensional system of which the dynamic behavior is described by the following set of ODEs:

$$\begin{aligned}
\frac{\mathrm{d}C}{\mathrm{d}t} &= -ar_{\max}\frac{C}{K_C + C}\frac{N}{K_N + N}P,\\
\frac{\mathrm{d}N}{\mathrm{d}t} &= -br_{\max}\frac{C}{K_C + C}\frac{N}{K_N + N}P,\\
\frac{\mathrm{d}P}{\mathrm{d}t} &= r_{\max}\frac{C}{K_C + C}\frac{N}{K_N + N}P - eP,\\
\frac{\mathrm{d}D}{\mathrm{d}t} &= eP.
\end{aligned} \tag{1}$$

This system describes the growth of phytoplankton (P) on two nutrients C and N (e.g. a carbon source and a nitrogen source), and the death of phytoplankton, resulting in formation of detritus (D). The system contains six parameters: C requirement a (dimension: C/P), N requirement b (dimension: N/P), maximum specific growth rate $r_{\max}$ (dimension: $time^{-1}$), C half-saturation K_C (dimension: C), N half-saturation K_N (dimension: N), and phytoplankton mortality e (dimension: $time^{-1}$). Note that system (1) was chosen for notational simplicity rather than realistic kinetics: the product of hyperbolae suffers from several problems regarding interpretation; better, mechanistic replacements have been suggested [11,12].

This system can also be represented by two reactions:

$$\begin{aligned}
&aC + bN \xrightarrow{r_1(C,N,P)} 1P && r_1(C,N,P) = r_{\max}\frac{C}{K_C + C}\frac{N}{K_N + N}P,\\
&\qquad\qquad \text{with}\\
&1P \xrightarrow{r_2(P)} 1D && r_2(P) = eP.
\end{aligned} \tag{2}$$

Each reaction distinguishes source compounds (left of the arrow), and sink compounds (right of the arrow).

For each reaction, the *reaction rate* $r_j(t, \mathbf{c})$ represents the rate at which the reaction progresses. These may depend both on time and the state of the system. The dimension (dim) of a reaction rate typically equals that of one of the participating variables, per time. In (2), both reaction rates have been arbitrarily normalized on the reference variable P: $\dim(r_1) = P$ produced $\cdot$ *time*$^{-1}$ and $\dim(r_2) = P$ destroyed $\cdot$ *time*$^{-1}$. Reaction rates may be negative, in which case the reaction is reversed, and the roles of sink and source interchange.

Constants preceding sink- and source variables in reactions refer to the amount of the variable destroyed or produced per reaction, and are commonly referred to as *stoichiometric coefficients*. These coefficients are *independent* of both time and the state of the system. The dimension of stoichiometric coefficients equals the dimension of their associated variable, divided by the dimension of the reference variable on which the reaction rate was normalized, e.g. $\dim(a) = \dim(C) \cdot \dim(P)^{-1}$.

When dealing with reactions and stoichiometric coefficients, it is appropriate to define the system using matrix–vector notation. This allows us to split the system in a time- and state-dependent part that describes reaction rates, and a constant part that describes stoichiometric coefficients.

Definition 1. For a given (biochemical) system of R reactions, define the time- and state-dependent reaction rate vector $\mathbf{r} \in \mathbb{R}^R$, such that every element r_j equals the rate at which reaction j progresses [20,19,21].

For system (2), the reaction rate vector is given by:

$$\mathbf{r}(C, N, P) = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} r_{\max} \frac{C}{K_C + C} \frac{N}{K_N + N} P \\ e P \end{pmatrix}. \tag{3}$$

Definition 2. For a given (biochemical) system of R reactions, define the time- and state-independent stoichiometry matrix $\mathbf{S} \in \mathbb{R}^{I \times R}$ [20,19,21]. Let every element S_{ij} represent a stoichiometric coefficient, which describes the amount of state variable i produced per reaction j. If a compound acts as source in a reaction, it is consumed rather than produced, and the corresponding S_{ij} is negative. If a compound acts as sink, it is produced and the corresponding S_{ij} is positive.

For system (2), the stoichiometry matrix is:

$$\mathbf{S} = \begin{pmatrix} -a & 0 \\ -b & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix}. \tag{4}$$

Rows correspond with state variables C, N, P and D, respectively, columns with reactions.

Definition 3. The product of the stoichiometry matrix and the reaction rate vector renders the net change in state variable values, i.e. the right-hand side of the ODEs:

$$\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t} = \mathbf{S}\mathbf{r}(t, \mathbf{c}). \tag{5}$$

For example (2), one can easily check that the product of (4) and (3) indeed renders the ODEs given in (1).

2.2. Positivity

Definition 4. A system of ODEs is called unconditionally positive if $\mathbf{f}(t, \mathbf{c})$ is such that $\mathbf{c}(t) > 0$ for all $t > 0$, given $\mathbf{c}(0) > 0$.

Definition 5. An integration scheme Φ is called unconditionally positive if $\mathbf{c}^{n+1} > 0$ for any given $\mathbf{c}^n > 0$ and any arbitrary time step $\Delta t > 0$.

It is important to note that an unconditionally positive integration scheme only makes sense if applied to systems that are themselves unconditionally positive.

2.3. Conservation

In biochemical context, conservation refers to the fact that atoms and energy are conserved. State variables represent compounds, which are time-invariant compositions of atoms of various element species, and a given amount of chemical energy (e.g. enthalpy, or Gibbs free energy). Compounds are not conserved, but their constituents are. Therefore, the common closed-system-based definition of conservation that states the sum of all state variables is constant does not cover the biochemical concept of conservation.

Definition 6. For a given (biochemical) system, define the time- and state-independent compound composition matrix $\mathbf{E} \in \mathbb{R}^{E\times I}$, such that every element E_{ij} equals the amount of compound constituent i (energy or some element species) per compound j. Every row in $\mathbf{E}$ corresponds with a compound constituent; the number of rows E (i.e. the number of constituents monitored) depends solely on the interest of the modeler.

It is worth noting that many authors distinguish an elemental composition matrix that defines element counts per compound, and an energy vector that defines energy (or enthalpy) per compound [20,10,7]. For the sake of simplicity, we combine both in the compound composition matrix $\mathbf{E}$: the energy vector is represented by one row of $\mathbf{E}$.

Typically, one includes in $\mathbf{E}$ only a selection of constituents tuned to the modeler's interest. The selection may exclude many element species, and even energy, if the modeler takes no interest in energetics. For system (2), one could focus on the element species carbon (C) and nitrogen (N), and write:

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & a & a \\ 0 & 1 & b & b \end{pmatrix}, \tag{6}$$

where rows correspond with compound constituents (the elements C and N, respectively) and columns with compounds (C, N, P and D, respectively).

In other words, state variable C contains 1 atom of carbon and no nitrogen, state variable N contains no carbon and 1 atom of nitrogen, and both variable P and D contain a atoms of carbon and b atoms of nitrogen.

The product of compound composition matrix $\mathbf{E}$ and stoichiometry matrix $\mathbf{S}$ renders a matrix that defines the change in elements and energy (rows) for the different reactions (columns). For the system to be conservative in the biochemical sense, *the total amount of any element species, and the total amount of energy must not be affected by any reaction*. For a conservative system, this implies that the product of $\mathbf{E}$ and $\mathbf{S}$ must render a zero matrix.

Definition 7. A system is called conservative if it can be written as a set of reactions such as (2), with an associated compound composition matrix $\mathbf{E}$, that multiplied with stoichiometry matrix $\mathbf{S}$, renders an $E \times R$ zero matrix, that is: $\mathbf{ES} = \mathbf{0}$ [20,10].

Above implies the columns of $\mathbf{S}$ are part of the null space of $\mathbf{E}$. Equivalently, range($\mathbf{S}$) $\subset$ nullspace($\mathbf{E}$), and since $\mathbf{E} \in \mathbb{R}^{E\times I}$, we have rank($\mathbf{S}$) $\leqslant I - E$ if we assume rows of $\mathbf{E}$ are linearly independent [21]. In other words, the number of linearly independent rows in $\mathbf{S}$ cannot be greater than the dimension of the system (I) diminished with the number of linearly independent conservation laws (rows of $\mathbf{E}$).

If we multiply (6) and (4) for system (2):

$$\mathbf{ES} = \begin{pmatrix} 1 & 0 & a & a \\ 0 & 1 & b & b \end{pmatrix} \begin{pmatrix} -a & 0 \\ -b & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

i.e. none of the reactions changes the total amount of any compound constituent (C, N).

In effect, Definition 7 defines conservation on the *microscopic* level: the level of individual reactions. One commonly encounters a *macroscopic* definition of conservation that directly follows from Definition 7: $\mathbf{Ef}(t, \mathbf{c}) = \mathbf{ESr}(t, \mathbf{c}) = \mathbf{0}$, which in turn implies $\mathbf{Ec}(t)$ is constant [21]. However, macroscopic conservation is a necessary but not sufficient condition for microscopic conservation. If rank($\mathbf{S}$) $< I - E =$ dim nullspace($\mathbf{E}$) (assuming the rows of $\mathbf{E}$ are linearly independent), there exists a subspace $H \subset \mathbb{R}^I$, $H \subset$ nullspace($\mathbf{E}$) for which macroscopic but not microscopic conservation is met (Andreas Meister, personal communication). Definition 7 is thus notably more strict than the common macroscopic definition of conservation [21], and deserves additional motivation. The main difference between

macro- and microscopic conservation is this: systems that are conservative on the macroscopic level might preserve total element- and energy counts in the system by redistributing mass and energy along routes unaccounted for by any reaction. Specifically, this would allow for apparent stoichiometric ratios to deviate from the imposed ratios in **S**; as these ratios are key determinants in various problems in biochemistry (e.g. the famous carbon:nitrogen:phosphorus 'Redfield' ratio in marine systems), such deviations are better avoided.

Remark 8. Many biochemical systems are open to mass and energy, in the sense that one or more compounds participating in reactions are external to the system. Such external compounds appear in the system as (environmental) parameters, rather than as state variables, and are therefore not included in **S**. For the purpose of demonstrating conservation, however, one can easily define a complete stoichiometry matrix that does include external compounds; this matrix, rather than its subset **S**, should meet Definition 7.

Numerical integration involves a scheme-specific approximation of the change in state variable values within one time step. To ensure meaningful results, the state variable change must foremost be consistent with the original system. For biochemical systems, this obviously requires conservation of mass and energy. Arguably, consistence with biochemistry not only implies conservation on the macroscopic level (no elements or energy leave the system), but also on the microscopic level (no elements or energy are destroyed or created within one reaction). Analytically, conservation across individual reactions is ensured by constructing **S** such that it meets Definition 7. Microscopic conservation for any arbitrary biochemical system is then only guaranteed if state variables change directly and exclusively according to the original stoichiometry matrix **S**, using a scheme-specific approximation of reaction rates.

Definition 9. An integration scheme Φ is called consistent with respect to biochemical systems, and conservative with respect to mass and energy, if for every integration step $n+1$, there exists a vector $\mathbf{r}^n$ that satisfies:

$$\mathbf{c}^{n+1} - \mathbf{c}^n = \mathbf{S}\mathbf{r}^n \Delta t, \tag{7}$$

S denoting the system-specific stoichiometry matrix (Definition 2). Vector $\mathbf{r}^n$ may be thought of as to represent a scheme-specific approximation of the average reaction rate vector $\mathbf{r}(t, \mathbf{c})$ between t^n and t^{n+1}.

Obviously, for biochemical systems that are conservative in the sense of Definition 7, (7) implies conservation at the macroscopic level:

$$\mathbf{E}\big(\mathbf{c}^{n+1} - \mathbf{c}^n\big) = \mathbf{E}\big(\mathbf{S}\mathbf{r}^n \Delta t\big) = \mathbf{E}\mathbf{S}\big(\mathbf{r}^n \Delta t\big) = \mathbf{0}, \tag{8}$$

i.e. the total amount of any element species, and of energy, remains constant. Condition (7) implies (8), but the opposite is not necessarily true. This again reflects the difference between macroscopic and microscopic conservation as described under Definition 7.

In the special case where all compounds are of equal composition, the columns of compound composition matrix **E** are identical, i.e. $E_{ij} = E_{i1}, \forall j \in \{2, \ldots, I\}$. Then, (8) implies that the sum of all c_i is constant; this is often used as definition of conservation [3], but obviously falls short for biochemical purposes, except for simple cases such as the nitrogen-based NPZ-type models like the one of Fasham et al. [5].

2.4. Order of accuracy

It is common ODE practice to define the order of accuracy of a method by means of its local truncation error:

Definition 10. Let $\mathbf{c}^{n+1}$ denote the numerical approximation obtained by applying the method Φ, starting at time t^n on the exact solution, i.e., $\mathbf{c}^n = \mathbf{c}(t^n)$. Then, $\mathbf{e} := \mathbf{c}(t^{n+1}) - \mathbf{c}^{n+1}$, the error made in one step, is called the local truncation error.

The method Φ is said to be of order p if **e** behaves as [6]:

$$\mathbf{e} = \mathrm{O}\big((\Delta t)^{p+1}\big).$$

To visualize the order of the methods to be discussed, we will employ a relative mean square error, taken over all time steps and averaged over all state variables:

$$ERR = \frac{1}{I}\sum_{i=1}^{I}\sqrt{\frac{\sum_{n=1}^{N}(c_i(t^n)-c_i^n)^2}{\sum_{n=1}^{N}(c_i(t^n))^2}}. \tag{9}$$

3. Model problems

To analyze the performance of the new integration schemes at realistic Δt, we apply these schemes to two test cases: a simple linear system for which an analytical solution is available, and the simple non-linear system (1).

3.1. Simple linear system

The simple linear system used in Burchard et al. [3] is given by:

$$\frac{\mathrm{d}c_1}{\mathrm{d}t} = c_2 - ac_1, \qquad \frac{\mathrm{d}c_2}{\mathrm{d}t} = ac_1 - c_2, \tag{10}$$

with non-dimensional time, non-dimensional parameter $a \geqslant 0$ and initial values $c_1(0) = c_1^0 > 0$ and $c_2(0) = c_2^0 > 0$.

The analytical solution of this system is given by:

$$c_1(t) = \left(1 + c\mathrm{e}^{-(a+1)t}\right)c_1^\infty,$$

with the asymptotic solution

$$c_1^\infty = \frac{c_1^0 + c_2^0}{a+1} \quad \text{and} \quad c = \frac{c_1^0}{c_1^\infty} - 1.$$

Given that the system is closed and conservative (which implies $c_1(t)$ and $c_2(t)$ represent compounds with the same composition and unit), $c_1(t) + c_2(t) = c_1^0 + c_2^0$ for all $t \geqslant 0$. Thus, $c_2(t)$ is defined by:

$$c_2(t) = c_1^0 + c_2^0 - c_1(t).$$

In sample simulations, $a = 5$, $c_1^0 = 0.9$ and $c_2^0 = 0.1$ are used. Obviously, $c_1(t) + c_2(t) = 1$ for all t. All values correspond with those used by Burchard et al. [3]. The analytical solution of the system for these values is shown in Fig. 1.

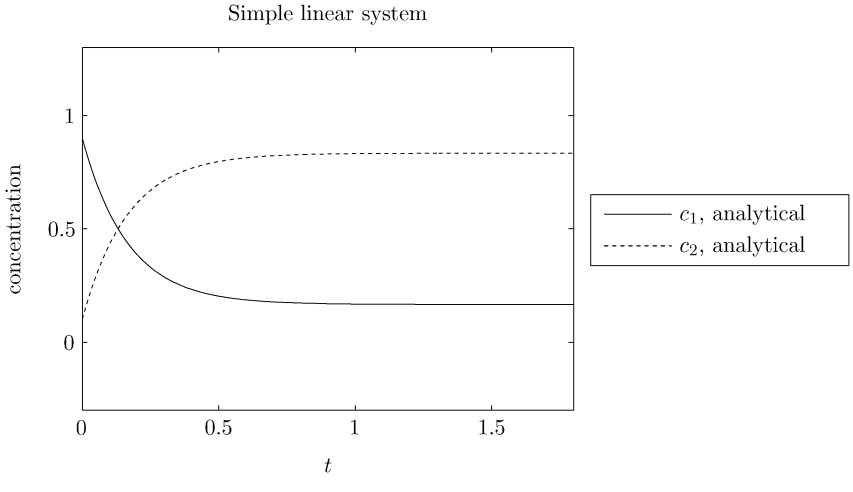

Fig. 1. Analytical reference solution for the simple linear system (10).

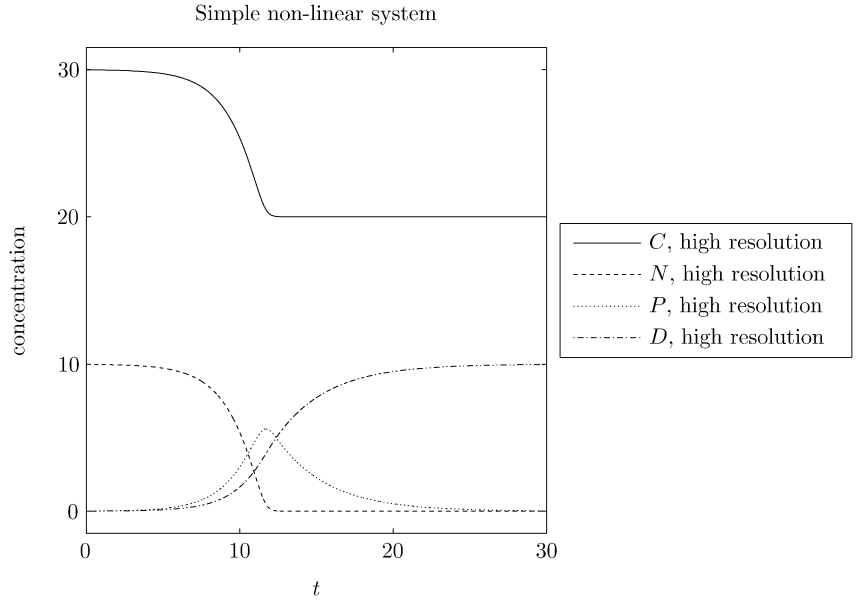

Fig. 2. High-resolution numerical approximation of the solution of the simple non-linear system, as given by (1). This approximation was obtained with a fourth-order accurate Runge–Kutta scheme [6, p. 138], with $\Delta t = 0.01$.

3.2. Simple non-linear system

As second test case, we use the simple biochemical system (1). This system may be interpreted as describing the ecosystem in the upper-mixed layer of the ocean in spring. It is similar to the simple non-linear system presented in Burchard et al. [3], but includes an additional nutrient C. This change has been made to demonstrate conservation problems of the modified Patankar scheme proposed in Burchard et al. [3] when reactions contain more than one source compound. Note that for the limiting case $C \to \infty$, system (1) reduces to the simple non-linear system of Burchard et al. [3] if $b = K_C = K_N = r_{\max} = 1$.

In sample simulations, we use $a = b = K_C = K_N = r_{\max} = 1$ and $e = 0.3$. Initial state variable values were set to $C^0 = 29.98$, $N^0 = 9.98$, and $P^0 = D^0 = 0.01$. Since phytoplankton P requires equal amounts of C and N for growth ($a = b = 1$) while C is available in much higher amount than N, C represents in effect a non-limiting nutrient.

Total initial amounts of compound constituents are given by the product of the compound composition matrix as defined in (6), and the vector of initial state variable values, i.e.:

$$\mathbf{E}\mathbf{c}^0 = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 29.98 \\ 9.98 \\ 0.01 \\ 0.01 \end{pmatrix} = \begin{pmatrix} 30 \\ 10 \end{pmatrix}.$$

Note that this implies $C + P + D = 30$ and $N + P + D = 10$ for all t, since the system is conservative (see also Definition 7).

An analytical solution for the system cannot be obtained. Hence, we resort to a high-resolution approximation of the solution to compare the results of various schemes against. This reference solution is shown in Fig. 2.

4. Numerical schemes

4.1. Forward Euler, Runge–Kutta

To familiarize the reader with biochemical conservation in the sense of (7), we first consider the well-known forward Euler scheme:

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t \mathbf{f}\left(t^n, \mathbf{c}^n\right).$$

This well-known scheme is first-order accurate. It is obviously not unconditionally positive, even given $\mathbf{c}^n > 0$: for any $f_i(t^n, \mathbf{c}^n) < 0$, there exists a time step Δt that results in $c_i^{n+1} < 0$.

The forward Euler scheme is conservative with respect to mass and energy. Applying the scheme to (5), one obtains:

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t \mathbf{S}\mathbf{r}\big(t^n, \mathbf{c}^n\big),$$

and it is easily seen that the forward Euler scheme satisfies (7), with $\mathbf{r}^n = \mathbf{r}(t^n, \mathbf{c}^n)$.

Second- and higher order Runge–Kutta schemes are, like the forward Euler scheme, derived using Taylor series expansion of $\mathbf{c}(t)$. For these schemes, one can also easily show that they are conservative, but not unconditionally positive.

4.2. Backward Euler

The backward Euler (or implicit Euler) scheme is given by:

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t \mathbf{f}\big(t^{n+1}, \mathbf{c}^{n+1}\big).$$

This scheme is known to be unconditionally positive [1,8]. Also, it is conservative in the sense of Definition 9 with $\mathbf{r}^n = \mathbf{r}(t^{n+1}, \mathbf{c}^{n+1})$. In addition, the backward Euler scheme is also, unlike forward Euler, well suited for solving stiff systems. Thus, it would seem well suitable for biochemical systems. However, the main drawback of the backward Euler scheme is the high computational cost in each step: it requires the solution of a system of non-linear equations. The Newton method typically applied to this end requires (approximation of) the matrix $\partial \mathbf{f}(t, \mathbf{c})/\partial \mathbf{c}$, which for most biochemical systems cannot be calculated analytically. Thus, one has to resort to numerical approximation of $\partial \mathbf{f}(t, \mathbf{c})/\partial \mathbf{c}$, for instance by finite differences. This involves repeated, costly evaluations of ODE right-hand sides.

Aside of associated computational costs, backward Euler is only first-order accurate, and higher-order implicit schemes of this type (e.g. Gear schemes) cannot be constructed without sacrificing positivity [1,8].

4.3. Modified Patankar

Using the reaction-based system definition presented in Section 2, the modified Patankar (MP) scheme [3] is given by:

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t \mathbf{S}'\big(\mathbf{c}^n, \mathbf{c}^{n+1}\big)\mathbf{r}\big(t^n, \mathbf{c}^n\big),$$

matrix $\mathbf{S}'(\mathbf{c}^n, \mathbf{c}^{n+1})$ being of the same size as the stoichiometry matrix $\mathbf{S}$, with elements:

$$S'_{ij} = \begin{cases} S_{ij} \frac{c_i^{n+1}}{c_i^n} & \text{for } S_{ij} r_j < 0, \\ 0 & \text{for } S_{ij} r_j = 0, \\ S_{ij} \sum_{k \in K_j} \big(\rho_{ijk} \frac{c_k^{n+1}}{c_k^n}\big) & \text{for } S_{ij} r_j > 0,\ K_j \neq \emptyset, \\ S_{ij} & \text{for } S_{ij} r_j > 0,\ K_j = \emptyset. \end{cases} \tag{11}$$

The set K_j represents the set of indices of state variables that act as source in reaction j:

$$K_j = \big\{ i\colon\ S_{ij} r_j < 0,\ i \in \{1, \ldots, I\} \big\}. \tag{12}$$

Constants $0 \leqslant \rho_{ijk} \leqslant 1$ in (11) are constrained by the condition

$$\sum_{k \in K_j} \rho_{ijk} = 1.$$

In words, source fluxes are multiplied with an associated 'relative change': the ratio between the approximated source value at t^{n+1} and its value at t^n. Sink fluxes are multiplied with a weighed sum of relative changes in the corresponding sources.

It is worth noting that the original Patankar approach [16] is very similar, the only difference being that $S'_{ij} = S_{ij}$ for $S_{ij} r_j > 0$ in (11).

The MP scheme has been shown to be unconditionally positive [3]. In the same paper, this scheme has been shown to be conservative in the sense that for a closed system, the sum of state variables is constant. This, however, does not imply that the scheme is conservative in the sense of Definition 9.

For our example (2), matrix $\mathbf{S}'(\mathbf{c}^n, \mathbf{c}^{n+1})$ would be given by:

$$\mathbf{S}'\left(\mathbf{c}^n, \mathbf{c}^{n+1}\right) = \begin{pmatrix} -a\frac{c_C^{n+1}}{c_C^n} & 0 \\ -b\frac{c_N^{n+1}}{c_N^n} & 0 \\ 1\left(\rho_{3,1,C}\frac{c_C^{n+1}}{c_C^n} + \rho_{3,1,N}\frac{c_N^{n+1}}{c_N^n}\right) & -1\frac{c_P^{n+1}}{c_P^n} \\ 0 & 1\frac{c_P^{n+1}}{c_P^n} \end{pmatrix},$$

with $\rho_{3,1,C} + \rho_{3,1,N} = 1$.

To be conservative in the sense of Definition 9, there must exist a vector $\mathbf{r}^n$, such that matrix $\mathbf{S}'(\mathbf{c}^n, \mathbf{c}^{n+1})$ satisfies:

$$\mathbf{S}'\left(\mathbf{c}^n, \mathbf{c}^{n+1}\right)\mathbf{r}\left(t^n, \mathbf{c}^n\right) = \mathbf{S}\mathbf{r}^n.$$

For the example, this implies there must exist constants r_1^n and r_2^n, such that:

$$\begin{aligned}
&-a\frac{c_C^{n+1}}{c_C^n} r_1 = -a r_1^n, \\
&-b\frac{c_N^{n+1}}{c_N^n} r_1 = -b r_1^n, \\
&1\left(\rho_{3,1,C}\frac{c_C^{n+1}}{c_C^n} + \rho_{3,1,N}\frac{c_N^{n+1}}{c_N^n}\right) r_1 - 1\frac{c_P^{n+1}}{c_P^n} r_2 = 1 r_1^n - 1 r_2^n, \\
&1\frac{c_P^{n+1}}{c_P^n} r_2 = 1 r_2^n.
\end{aligned}$$

From the first two equations, one can derive that there only exists a valid r_1^n in the special case where $c_C^{n+1}/c_C^n = c_N^{n+1}/c_N^n$: the relative decreases of the nitrogen- and the carbon source must be equal. Obviously, this condition will only be met in rare, temporary states. Thus, for system (2), the MP scheme is typically not conservative in the sense of Definition 9. In fact, even the common, more lenient definition of conservation (8) does not hold for the MP scheme applied to the example: Figs. 3 and 5 show that the total amount of carbon (i.e. $C + P + D$) decreases over time.

The above result can be generalized: if for every set K_j, all elements i have equal $c_i^{n+1}/c_i^n = p_j$ ratios, the modified Patankar scheme can be written as

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t \mathbf{S}\mathbf{r}^n\left(t^n, \mathbf{c}^n, \mathbf{c}^{n+1}\right) \quad \text{with } r_j^n\left(t^n, \mathbf{c}^n, \mathbf{c}^{n+1}\right) = r_j\left(t^n, \mathbf{c}^n\right) p_j, \tag{13}$$

implying conservation in the sense of Definition 9. Consequently, if every K_j contains at most one element, the MP scheme is conservative with $p_j = c_1^{n+1}/c_1^n$ for non-empty K_j, and $p_j = 1$ for $K_j = \emptyset$.

We can conclude that the MP scheme is not conservative in the sense of Definition 9 for any arbitrary system (with arbitrary **E** and **S**). It *is* conservative if (1) all system reactions contain at most one source compound (all sets K_j contain at most one element), or (2) the relative changes in all sources are equal. However, for the many biochemical systems that do not satisfy either requirement, the MP approach is clearly not suitable.

The second-order Modified Patankar–Runge–Kutta (MPRK) scheme may be considered to consist of two consecutive MP steps. Therefore, it suffers from the problems with conservation as the MP approach. Although we do not present the mathematical proof showing that the MPRK scheme is not conservative, we do show in Figs. 4 and 5 that the MPRK scheme applied to the simple non-linear system (1) violates Eq. (8): the total amount of carbon (i.e.

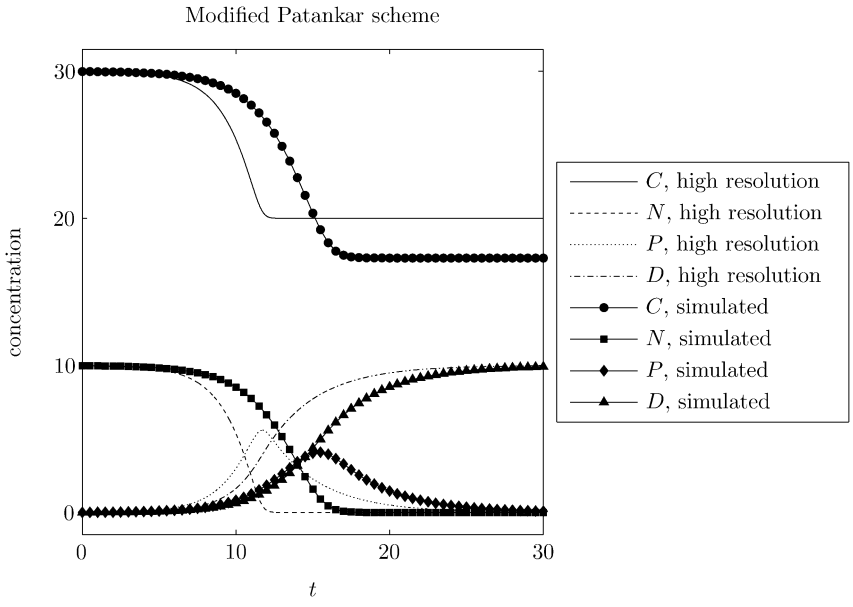

Fig. 3. Coarse numerical approximation ($\Delta t = 0.5$) with the Modified Patankar (MP) scheme [3] and high-resolution numerical approximation ($\Delta t = 0.01$) with a Runge–Kutta 4 scheme [6, p. 138] for the simple non-linear system (1).

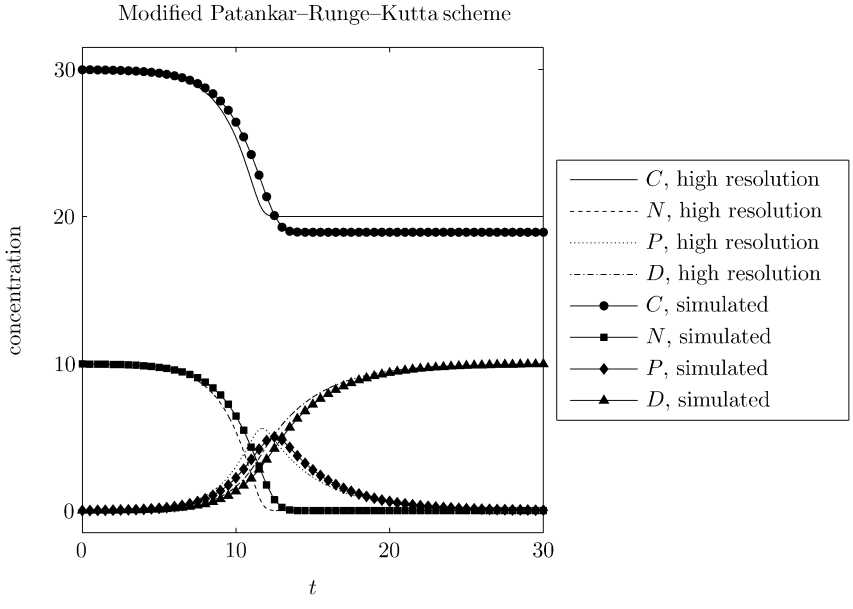

Fig. 4. Coarse numerical approximation ($\Delta t = 0.5$) with the Modified Patankar–Runge–Kutta (MPRK) scheme [3] and high-resolution numerical approximation ($\Delta t = 0.01$) with a Runge–Kutta 4 scheme [6, p. 138] for the simple non-linear system (1).

$C + P + D$) decreases over time, due to the inordinate decrease of C over time. Therefore, the MPRK scheme cannot be conservative in the sense of Definition 9.

4.4. New scheme: first-order accuracy

We propose an integration scheme that is based on the (forward) Euler scheme, but guarantees $\mathbf{c}^n > 0$, $n \in \mathbb{N}$, given $\mathbf{c}^0 > 0$.

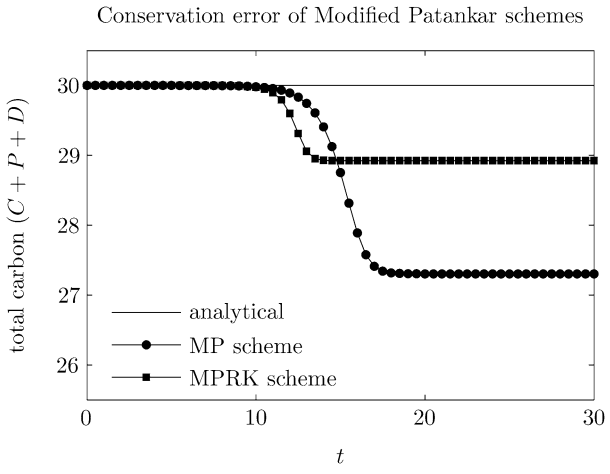

Fig. 5. The total amount of carbon (i.e. $C + P + D$) with the Modified Patankar (MP) and Modified Patankar–Runge–Kutta (MPRK) schemes [3] for the simple non-linear system (1).

The new scheme is given by:

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t \mathbf{f}\big(t^n, \mathbf{c}^n\big) \prod_{j \in J^n} \frac{c_j^{n+1}}{c_j^n} \quad \text{with } J^n = \big\{i\colon\ f_i\big(t^n, \mathbf{c}^n\big) < 0,\ i \in \{1, \ldots, I\}\big\}, \tag{14}$$

where J^n represents the set of indices for state variables with negative derivative. Note that for $J^n = \emptyset$, the product term drops out, and we are left with the Euler scheme (however, for a closed conservative system, $J^n \neq \emptyset$ per definition).

The scheme as given in (14) renders a system of I non-linear implicit equations, which must be solved to arrive at $\mathbf{c}^{n+1}$.

4.4.1. System reduction

Let us start with a positive solution vector $\mathbf{c}^n > 0$. Writing (14) component-wise and dividing the ith equation by c_i^n, we arrive at:

$$\frac{c_i^{n+1}}{c_i^n} = 1 + \frac{\Delta t f_i(t^n, \mathbf{c}^n)}{c_i^n} \prod_{j \in J^n} \frac{c_j^{n+1}}{c_j^n}, \quad i \in \{1, \ldots, I\}.$$

Calculating the product of all c_i^{n+1}/c_i^n, $\forall i \in J^n$, we find

$$\prod_{i \in J^n} \frac{c_i^{n+1}}{c_i^n} = \prod_{i \in J^n} \left(1 + \frac{\Delta t f_i(t^n, \mathbf{c}^n)}{c_i^n} \prod_{j \in J^n} \frac{c_j^{n+1}}{c_j^n}\right).$$

Defining $p := \prod_{j \in J^n} c_j^{n+1}/c_j^n$, this can be simplified to:

$$p = \prod_{j \in J^n} \left(1 + \frac{\Delta t f_j(t^n, \mathbf{c}^n)}{c_j^n} p\right).$$

To find p, it is convenient to define a function $g(p)$:

$$g(p) = \prod_{j \in J^n} \left(1 + \frac{\Delta t f_j(t^n, \mathbf{c}^n)}{c_j^n} p\right) - p = 0, \tag{15}$$

or, briefly:

$$g(p) = \prod_{j \in J^n} (1 + a_j p) - p = 0 \quad \text{with } a_j = \Delta t f_j\left(t^n, \mathbf{c}^n\right)/c_j^n.$$

The function $g(p)$ is a polynomial of a degree equal to the number of elements in J^n, and therefore has as many roots, which may be real or complex. Thus, the problem to solve has been reduced from a set of I non-linear implicit equations to a polynomial equation in one single unknown.

4.4.2. Restrictions on the p-domain

From Eq. (14) and the fact that we require $\mathbf{c}^{n+1} > 0$, we obtain the condition

$$c_i^n + \Delta t f_i\left(t^n, \mathbf{c}^n\right) p > 0 \quad \text{for all } i \in \{1, \ldots, I\}.$$

As time step Δt is positive by definition, parameter p has to satisfy the following I inequalities:

$$p > -\frac{c_i^n}{\Delta t f_i(t^n, \mathbf{c}^n)} \quad \text{if } f_i\left(t^n, \mathbf{c}^n\right) > 0, \tag{16}$$

$$p < -\frac{c_i^n}{\Delta t f_i(t^n, \mathbf{c}^n)} \quad \text{if } f_i\left(t^n, \mathbf{c}^n\right) < 0. \tag{17}$$

Note that if $f_i(t^n, \mathbf{c}^n) = 0$ for some $i \in \{1, \ldots, I\}$, no further restrictions on p are posed.

Since we require $\mathbf{c}^{n+1} > 0$, given $\mathbf{c}^n > 0$, we obtain:

$$p = \prod_{j \in J^n} \frac{c_j^{n+1}}{c_j^n} > 0, \tag{18}$$

which gives another lower bound for p. This lower bound exceeds, and thus replaces, the (negative) bound defined by (16).

From Eq. (14) and conditions (17) and (18), we know that $0 < c_j^{n+1} < c_j^n$ for all $j \in J^n$, which implies there exists another upper bound for p:

$$p = \prod_{j \in J^n} \frac{c_j^{n+1}}{c_j^n} < 1. \tag{19}$$

Thus, the upper bound for p is given by conditions (17) and (19):

$$p_{\max} = \min\left(1, \min_{j \in J^n}\left(-\frac{c_j^n}{\Delta t f_j(t^n, \mathbf{c}^n)}\right)\right).$$

Now, the domain for p is given by $p \in (0, p_{\max})$.

For these bounds, one can calculate $g(p)$:

$$g(0) = 1,$$

$$g(p_{\max}) = \begin{cases} -p_{\max} & \text{for } p_{\max} = \min_{j \in J^n}\left(-\frac{c_j^n}{\Delta t f_j(t^n, \mathbf{c}^n)}\right), \\ \gamma - 1 \text{ with } \gamma < 1 & \text{for } p_{\max} = 1. \end{cases}$$

Since $g(0) > 0$ and $g(p_{\max}) < 0$ and $g(p)$ is continuous, $g(p)$ must cross the p-axis an uneven number of times within the p-domain. Hence, we know $g(p)$ has at least one real root in the domain $p \in (0, p_{\max})$, potentially more (the maximum depending on J^n).

4.4.3. Behavior of g(p) in the p-domain

The derivative of $g(p)$ equals:

$$\frac{\mathrm{d}g}{\mathrm{d}p} = \sum_{i \in J^n}\left(\frac{\Delta t f_i(t^n, \mathbf{c}^n)}{c_i^n} \prod_{j \in J^n,\, j \neq i}\left(1 + \frac{\Delta t f_j(t^n, \mathbf{c}^n)}{c_j^n} p\right)\right) - 1.$$

Replacing $\Delta t f_i(t^n, \mathbf{c}^n)/c_i^n$ by a_i yields

$$\frac{\mathrm{d}g}{\mathrm{d}p} = \sum_{i \in J^n} \left(a_i \prod_{j \in J^n,\, j \neq i} (1 + a_j p) \right) - 1.$$

Within the p-domain, we know $1 + a_i p > 0$ for all $i \in J^n$. Also, every a_i will be negative for all $i \in J^n$, as J^n by definition comprises only those state variables for which $f_i(t^n, \mathbf{c}^n) < 0$. Thus, it is easily verified that $\mathrm{d}g/\mathrm{d}p < 0$ within the p-domain.

Summarizing, within the p-domain of interest, $g(p)$ is a continuous decreasing function of p, starting at $g(0) > 0$, and ending at $g(p_{\max}) < 0$. Hence, there exists exactly one real root of $g(p)$ in this range.

To find this root, we may employ the relatively slow, but robust bisection iteration process, which is guaranteed to find a root of $g(p)$, as we know $g(p)$ changes sign within the p-domain$(0, p_{\max})$. While other schemes may find the root of $g(p)$ much faster than the bisection process, such other schemes are often not guaranteed to converge to the correct root (e.g. Newton–Raphson, Secant), or have a strongly problem-dependent convergence rate (e.g. Regula Falsi). More intelligent approaches, e.g. the use of a bisection-Newton–Raphson hybrid scheme, might be used to maximize performance, but such are not further explored in the present paper.

Theorem 11. *The scheme* (14) *is first-order accurate.*

Proof. Given $p := \prod_{j \in J^n} (c_j^{n+1}/c_j^n)$, the new scheme can be written as:

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \Delta t \mathbf{f}\big(t^n, \mathbf{c}^n\big) p.$$

For $\mathbf{c}^{n+1}$ to be a first-order approximation in $t = t^n + \Delta t$, we need $\lim_{\Delta t \to 0} p = 1$.

To prove this, we expand the product in Eq. (15):

$$g(p) = 1 + \sum_{i=1}^{k} b_i p^i (\Delta t)^i - p = 0, \tag{20}$$

with k denoting the cardinal number of set J^n, and every b_i denoting some product of $f_j(t^n, \mathbf{c}^n)/c_j^n$ ratios, independent of both p and Δt.

For (20), one can easily see that $p = 1$ is a solution for $\lim_{\Delta t \to 0} g(p) = 0$. Since the bisection technique is guaranteed to converge to the only valid real value of p, it will ultimately converge to $p = 1$, making the new scheme first-order accurate. However, the actual accuracy of the found $p^{(r)}$ depends on the number of bisection iterations r; only for $r \to \infty$, we find $p^{(r)} \to p$. Therefore, the scheme (14) is theoretically only first-order accurate if $r \to \infty$.

In practice, we stop bisection when the following condition is satisfied:

$$2 \frac{p_{\text{right}} - p_{\text{left}}}{p_{\text{right}} + p_{\text{left}}} < 10^{-9}, \tag{21}$$

p_{left} and p_{right} denoting the left- and right bounds of the bisection p-domain. Note that (21) implies that the first 9 digits of p are known accurately. We found that additional bisection iterations had no qualitative effect on the results. □

Theorem 12. *The scheme* (14) *is unconditionally positive.*

Proof. To enforce positive values for all c_i^{n+1}, we found above $p \in (0, p_{\max})$. The bisection technique will return a value for p from within this range, independent of the number of bisection iterations. Thus, the new first-order scheme is unconditionally positive.

This is perhaps even more obvious if one considers that (14) can be interpreted as a rescaling of the time step to preserve positivity, compared to forward Euler: $\Delta t \to p \Delta t$. The actual value of p as found by our scheme is restricted by (16) and (17): precisely the bounds required by conditionally positive schemes. □

Theorem 13. *The scheme* (14) *is conservative in the sense of Definition* 9.

Proof. Using Definition 3, the new scheme is given by

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \mathbf{Sr}\big(t^n, \mathbf{c}^n\big)(\Delta t p), \quad \text{with } p = \prod_{j\in J^n} \frac{c_j^{n+1}}{c_j^n}.$$

This clearly satisfies Definition 9, with

$$\mathbf{r}^n = p\mathbf{r}\big(t^n, \mathbf{c}^n\big).$$

Note that Definition 9 is satisfied for any arbitrary scalar p. Conservation of the new first-order scheme is therefore independent of the number of bisection iterations. □

4.5. New scheme: second-order accuracy

We suggest the following scheme for second-order accurate results that are unconditionally positive and conservative:

$$\begin{aligned}
\mathbf{c}^{(1)} &= \mathbf{c}^n + \Delta t\mathbf{f}\big(t^n, \mathbf{c}^n\big) \prod_{j\in J^n} \frac{c_j^{(1)}}{c_j^n},\\
\mathbf{c}^{n+1} &= \mathbf{c}^n + \frac{\Delta t}{2}\big(\mathbf{f}\big(t^n, \mathbf{c}^n\big) + \mathbf{f}\big(t^{n+1}, \mathbf{c}^{(1)}\big)\big) \prod_{k\in K^n} \frac{c_k^{n+1}}{c_k^{(1)}},
\end{aligned} \tag{22}$$

where

$$\begin{aligned}
J^n &= \big\{i\colon\ f_i\big(t^n, \mathbf{c}^n\big) < 0,\ i \in \{1, \ldots, I\}\big\},\\
K^n &= \big\{i\colon\ f_i\big(t^n, \mathbf{c}^n\big) + f_i\big(t^{n+1}, \mathbf{c}^{(1)}\big) < 0,\ i \in \{1, \ldots, I\}\big\}.
\end{aligned} \tag{23}$$

Clearly, the first part of this scheme is identical to (14), i.e. identical to one integration step of the first-order scheme. The second part in the scheme may be rewritten in the form:

$$\begin{aligned}
\mathbf{c}^{n+1} &= \mathbf{c}^n + \Delta t\mathbf{h}\big(t^n, t^{n+1}, \mathbf{c}^n, \mathbf{c}^{(1)}\big) \prod_{k\in K^n} \frac{c_k^{n+1}}{c_k^n}\\
&\text{with } \mathbf{h}\big(t^n, t^{n+1}, \mathbf{c}^n, \mathbf{c}^{(1)}\big) = \frac{1}{2}\big(\mathbf{f}\big(t^n, \mathbf{c}^n\big) + \mathbf{f}\big(t^{n+1}, \mathbf{c}^{(1)}\big)\big) \prod_{k\in K^n} \frac{c_k^n}{c_k^{(1)}},
\end{aligned} \tag{24}$$

which, since $\mathbf{h}(t^n, t^{n+1}, \mathbf{c}^n, \mathbf{c}^{(1)})$ is independent of $\mathbf{c}^{n+1}$, adequately demonstrates that also the second part is mathematically similar to a step with the first-order method.

Given that the new scheme consists of two consecutive parts that are each mathematically equivalent to a step with the first-order scheme (14), the scheme {(22), (23)} can be implemented in the same manner as described in Section 4.4. Thus, the new scheme will need to find the real root of two different polynomials, for which we will again apply the bisection technique.

Theorem 14. *The scheme* {(22), (23)} *is second-order accurate.*

Proof. As the first part of the new scheme is first-order accurate, we know:

$$\mathbf{c}^{(1)} = \mathbf{c}\big(t^{n+1}\big) + \mathrm{O}\big(\Delta t^2\big) = \mathbf{c}^n + \Delta t\mathbf{f}\big(t^n, \mathbf{c}^n\big) + \mathrm{O}\big(\Delta t^2\big).$$

This can be used in a Taylor expansion of $\mathbf{f}(t^{n+1}, \mathbf{c}^{(1)})$:

$$\begin{aligned}
\mathbf{f}\big(t^{n+1}, \mathbf{c}^{(1)}\big) &= \mathbf{f}\big(t^n, \mathbf{c}^n\big) + \frac{\partial \mathbf{f}}{\partial \mathbf{c}}\big(t^n, \mathbf{c}^n\big)\big(\mathbf{c}^{(1)} - \mathbf{c}^n\big) + \Delta t\frac{\partial \mathbf{f}}{\partial t}\big(t^n, \mathbf{c}^n\big) + \mathrm{O}\big(\Delta t^2\big)\\
&= \mathbf{f}\big(t^n, \mathbf{c}^n\big) + \Delta t\left(\frac{\partial \mathbf{f}}{\partial \mathbf{c}}\big(t^n, \mathbf{c}^n\big)\mathbf{f}\big(t^n, \mathbf{c}^n\big) + \frac{\partial \mathbf{f}}{\partial t}\big(t^n, \mathbf{c}^n\big)\right) + \mathrm{O}\big(\Delta t^2\big).
\end{aligned} \tag{25}$$

Using this in the second part of Eq. (22), we obtain:

$$\begin{aligned}\mathbf{c}^{n+1} &= \mathbf{c}^n + \frac{\Delta t}{2}\big(\mathbf{f}\big(t^n,\mathbf{c}^n\big)+\mathbf{f}\big(t^{n+1},\mathbf{c}^{(1)}\big)\big)\prod_{k\in K^n}\frac{c_k^{n+1}}{c_k^{(1)}}\\ &= \mathbf{c}^n + \frac{\Delta t}{2}\left(2\mathbf{f}\big(t^n,\mathbf{c}^n\big)+\Delta t\left(\frac{\partial \mathbf{f}}{\partial \mathbf{c}}\big(t^n,\mathbf{c}^n\big)\mathbf{f}\big(t^n,\mathbf{c}^n\big)+\frac{\partial \mathbf{f}}{\partial t}\big(t^n,\mathbf{c}^n\big)\right)+\mathrm{O}\big(\Delta t^2\big)\right)\prod_{k\in K^n}\frac{c_k^{n+1}}{c_k^{(1)}}\\ &= \mathbf{c}^n + \left(\Delta t\mathbf{f}\big(t^n,\mathbf{c}^n\big)+\frac{\Delta t^2}{2}\left(\frac{\partial \mathbf{f}}{\partial \mathbf{c}}\big(t^n,\mathbf{c}^n\big)\mathbf{f}\big(t^n,\mathbf{c}^n\big)+\frac{\partial \mathbf{f}}{\partial t}\big(t^n,\mathbf{c}^n\big)\right)+\mathrm{O}\big(\Delta t^3\big)\right)\prod_{k\in K^n}\frac{c_k^{n+1}}{c_k^{(1)}}.\end{aligned}$$

For the new scheme to be second-order accurate, the product term on the right must behave as $1+\mathrm{O}(\Delta t^2)$ for $\Delta t\to 0$. Notice that a behavior of this term as $1+\mathrm{O}(\Delta t)$ would destroy the second-order of the scheme.

Because a step with the new first-order scheme has been proven to be first order accurate, we can use the fact that we perform two first-order steps consecutively (see Eq. (24)). For $\Delta t\to 0$ we may write:

$$\prod_{k\in K^n}\frac{c_k^n}{c_k^{(1)}} = \prod_{k\in K^n}\frac{c_k^n}{c_k^n+\Delta t f_k(t^n,\mathbf{c}^n)+\mathrm{O}(\Delta t^2)},$$

$$\prod_{k\in K^n}\frac{c_k^{n+1}}{c_k^n} = \prod_{k\in K^n}\frac{c_k^n+\frac{\Delta t}{2}(f_k(t^n,\mathbf{c}^n)+f_k(t^{n+1},\mathbf{c}^{(1)}))\prod_{k\in K^n}\frac{c_k^n}{c_k^{(1)}}+\mathrm{O}(\Delta t^2)}{c_k^n},$$

$$\prod_{k\in K^n}\frac{c_k^{n+1}}{c_k^{(1)}} = \prod_{k\in K^n}\frac{c_k^n+\frac{\Delta t}{2}(f_k(t^n,\mathbf{c}^n)+f_k(t^{n+1},\mathbf{c}^{(1)}))\prod_{k\in K^n}\frac{c_k^n}{c_k^{(1)}}+\mathrm{O}(\Delta t^2)}{c_k^n+\Delta t f_k(t^n,\mathbf{c}^n)+\mathrm{O}(\Delta t^2)}.$$

Using the first term in the Taylor expansion of (25) for all $f_k(t^{n+1},\mathbf{c}^{(1)})$, we arrive at:

$$\prod_{k\in K^n}\frac{c_k^{n+1}}{c_k^{(1)}} = \prod_{k\in K^n}\frac{c_k^n+(\Delta t f_k(t^n,\mathbf{c}^n)+\mathrm{O}(\Delta t^2))\prod_{k\in K^n}\frac{c_k^n}{c_k^{(1)}}+\mathrm{O}(\Delta t^2)}{c_k^n+\Delta t f_k(t^n,\mathbf{c}^n)+\mathrm{O}(\Delta t^2)}.$$

Since $\prod_{k\in K^n} c_k^n/c_k^{(1)}$ behaves as $1+\mathrm{O}(\Delta t)$ for $\Delta t\to 0$, we see that

$$\prod_{k\in K^n}\frac{c_k^{n+1}}{c_k^{(1)}} = \prod_{k\in K^n}\frac{c_k^n+\Delta t f_k(t^n,\mathbf{c}^n)+\mathrm{O}(\Delta t^2)}{c_k^n+\Delta t f_k(t^n,\mathbf{c}^n)+\mathrm{O}(\Delta t^2)}$$

indeed behaves as $1+\mathrm{O}(\Delta t^2)$. Thus, the new scheme {(22), (23)} is second-order accurate. Notice that the above proof also demonstrates that the choice of the factor $\prod_{k\in K^n} c_k^{n+1}/c_k^{(1)}$ in (22) is not arbitrary. For instance, the alternative choice $\prod_{k\in K^n} c_k^{n+1}/c_k^n$ would lead to $1+\mathrm{O}(\Delta t)$ for $\Delta t\to 0$, making the scheme only first-order accurate.

Note that in theory, second-order accuracy is achieved only for an infinite number of bisection iterations. In practice, we stop bisection when condition (21) is satisfied; results did not further improve with more bisection iterations. □

Theorem 15. *The second-order scheme* {(22), (23)} *is unconditionally positive.*

Proof. As demonstrated in Eq. (24), the final step of the second-order scheme is mathematically equivalent to a first-order step, using the vector function $\mathbf{h}(t^n,t^{n+1},\mathbf{c}^n,\mathbf{c}^{(1)})$ rather than the typical $\mathbf{f}(t^n,\mathbf{c}^n)$. The first-order scheme has been shown to guarantee positive values, independent of $\mathbf{f}(t^n,\mathbf{c}^n)$ and of the number of bisection iterations. Therefore, the second-order scheme likewise guarantees positive values for all elements of any $\mathbf{c}^n$, given a positive starting vector $\mathbf{c}^0$. □

Theorem 16. *The second-order scheme* {(22), (23)} *is conservative in the sense of Definition* 9.

Proof. Using Definition 3, the final part of the new scheme is given by:

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \frac{\Delta t}{2}\big(\mathbf{Sr}\big(t^n, \mathbf{c}^n\big) + \mathbf{Sr}\big(t^{n+1}, \mathbf{c}^{(1)}\big)\big)p, \quad \text{with } p = \prod_{k\in K^n} \frac{c_k^{n+1}}{c_k^{(1)}}.$$

This may be written as:

$$\mathbf{c}^{n+1} = \mathbf{c}^n + \mathbf{S}\big(\mathbf{r}\big(t^n, \mathbf{c}^n\big) + \mathbf{r}\big(t^{n+1}, \mathbf{c}^{(1)}\big)\big)\left(\frac{\Delta t}{2}p\right),$$

which clearly shows that the second-order scheme satisfies Definition 9, with:

$$\mathbf{r}^n = \frac{p}{2}\big(\mathbf{r}\big(t^n, \mathbf{c}^n\big) + \mathbf{r}\big(t^{n+1}, \mathbf{c}^{(1)}\big)\big).$$

Note as with the first-order scheme, Definition 9 is satisfied for any arbitrary constant p. Conservation of the new second-order scheme is therefore independent of the number of bisection iterations. □

5. Discussion and conclusion

Mathematical analysis and numerical simulations have shown that the first- and second-order schemes we present in this paper are unconditionally positive, and conservative in the strict biochemical sense. The order of accuracy of both schemes has been proven mathematically, and is also well demonstrated by Fig. 14.

Figs. 6–13 demonstrate that the new schemes can deliver relatively accurate results, even at large Δt. Our numerical approximation of the solution of the linear system is clearly more accurate than the one provided by traditional schemes: the solution produced by the new first-order scheme is more accurate than that of the Euler scheme, and the solution provided by the new second-order scheme is more accurate than that of the Runge–Kutta 2 scheme. This is reflected by the local truncation error of the various schemes, as shown in Fig. 14. Conversely, the traditional forward Euler and Runge–Kutta schemes approximate the solution of the simple non-linear system more accurately than the new schemes.

As shown in Fig. 15, the new schemes require a computational cost that is substantially higher than that of the traditional schemes. This is a disadvantage, but not enough to discard the new schemes: biochemical problems require schemes that are unconditionally positive and conservative; results that do not satisfy these requirements are completely meaningless in biochemical context. It is also worth noting that the new schemes will scale more favorably to higher-dimensional systems than for instance the MP and MPRK schemes. This is due to the fact that the new schemes always solve a scalar polynomial equation, whereas the MP/MPRK schemes solve a linear system of order I.

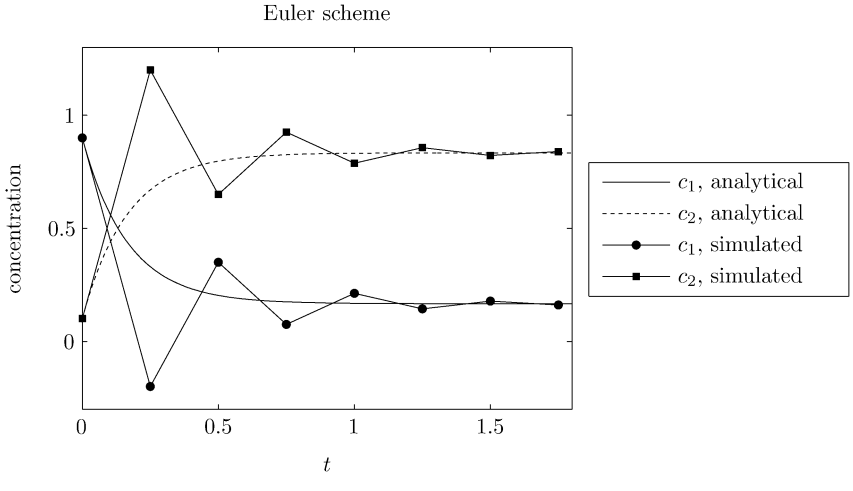

Fig. 6. Numerical approximation ($\Delta t = 0.25$) and analytical solution of the simple linear system (10) with the forward Euler scheme.

S0168-9274(05)00224-2/FLA AID:1854 Vol.●●●(●●●)
APNUM:m3SC+ v 1.50 Prn:12/01/2006; 12:16
apnum1854
[+model] P.18 (1-23)
by:R.M. p. 18

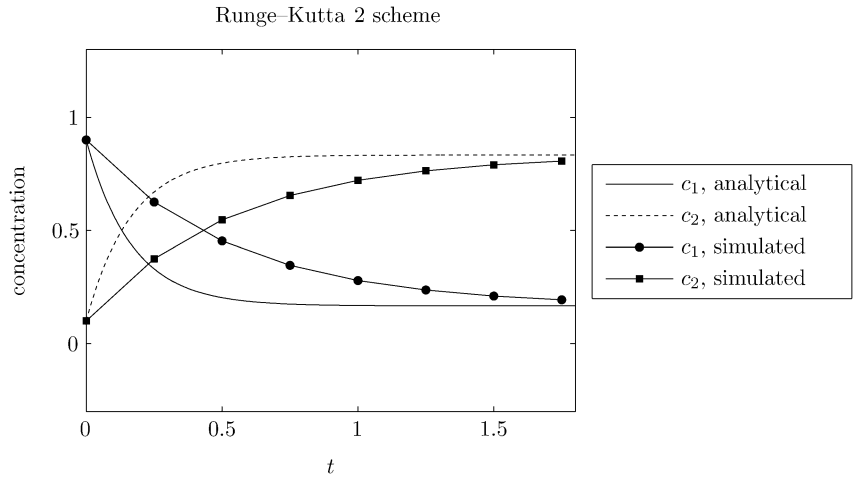

Fig. 7. Numerical approximation ($\Delta t = 0.25$) and analytical solution of the simple linear system (10) with the Runge–Kutta 2 scheme ((22) without product terms).

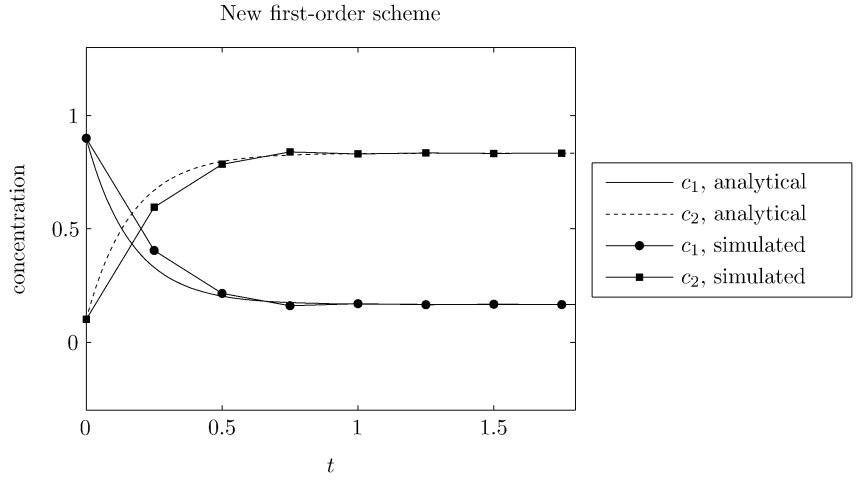

Fig. 8. Numerical approximation ($\Delta t = 0.25$) with the new first-order scheme and analytical solution for the simple linear system (10).

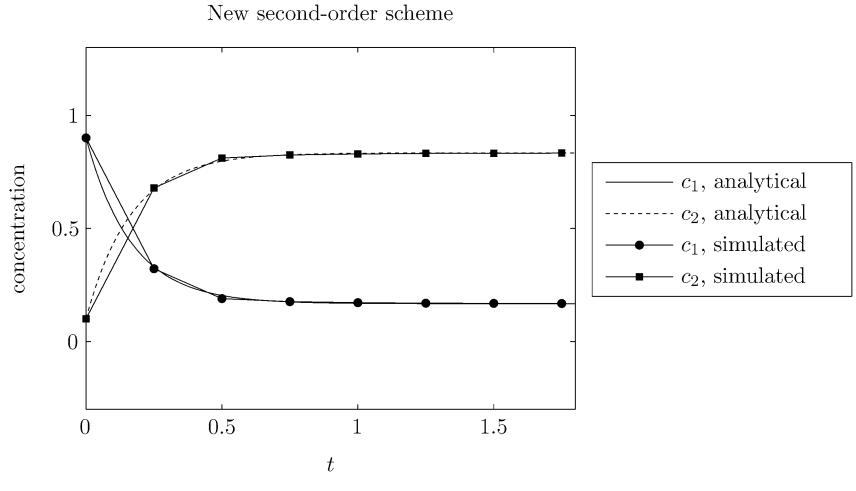

Fig. 9. Numerical approximation ($\Delta t = 0.25$) with the new second-order scheme and analytical solution for the simple linear system (10).

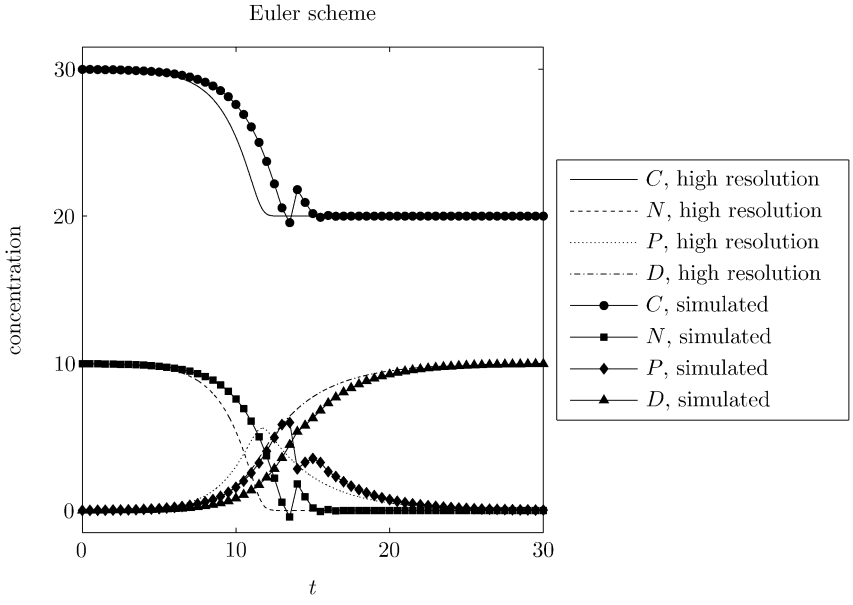

Fig. 10. Coarse numerical approximation ($\Delta t = 0.5$) with the forward Euler scheme and high-resolution numerical approximation ($\Delta t = 0.01$) with a Runge–Kutta 4 scheme [6, p. 138] for the simple non-linear system (1).

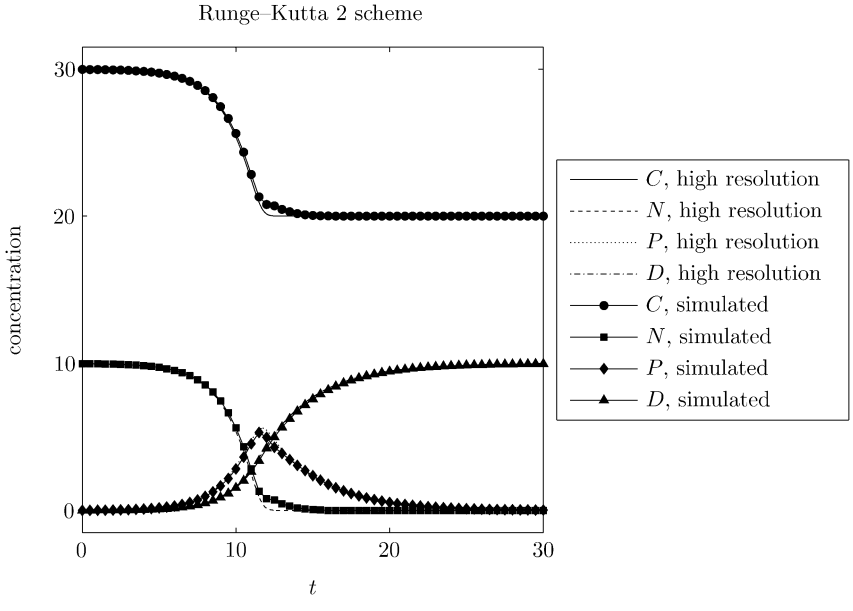

Fig. 11. Coarse numerical approximation ($\Delta t = 0.5$) with the Runge–Kutta 2 scheme ((22) without product terms) and high-resolution numerical approximation ($\Delta t = 0.01$) with a Runge–Kutta 4 scheme [6, p. 138] for the simple non-linear system (1).

Although unconditional positivity is the default in biochemistry, it is worth noting that some biochemical systems include one or more state variables that can become negative. For instance, the ERGOM model [15] includes a state variable 'oxygen concentration' that represents oxygen when positive, and hydrogen sulfide when negative. With the new schemes, this can easily be accounted for by excluding any such state variables from sets J^n and K^n.

Given the new schemes are unconditionally conservative and positive, the step size will be dictated by accuracy reasons only. To obtain maximum efficiency, one could combine the new schemes with techniques that dynamically adjust the time step based on estimated local error. While such approaches are beyond the scope of this paper, we may

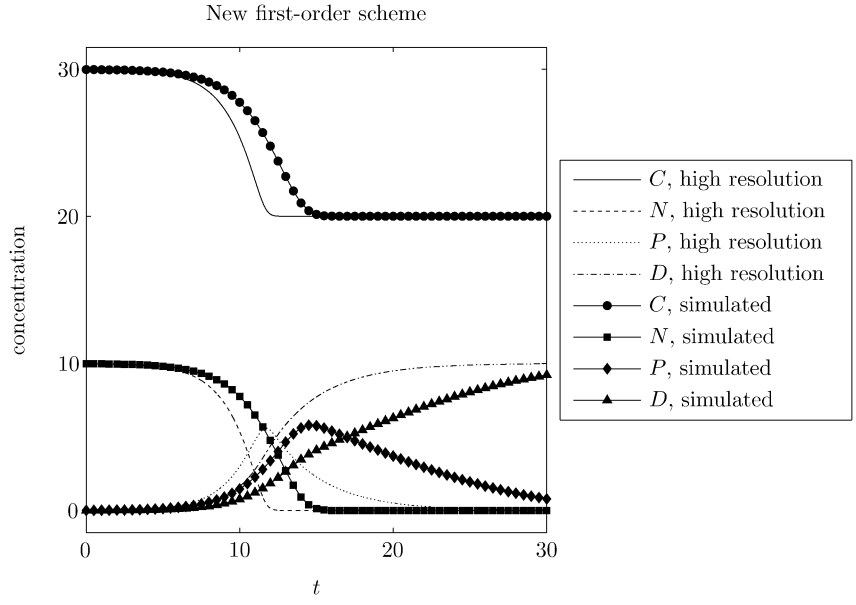

Fig. 12. Coarse numerical approximation ($\Delta t = 0.5$) with the new first-order scheme and high-resolution numerical approximation ($\Delta t = 0.01$) with a Runge–Kutta 4 scheme [6, p. 138] for the simple non-linear system (1).

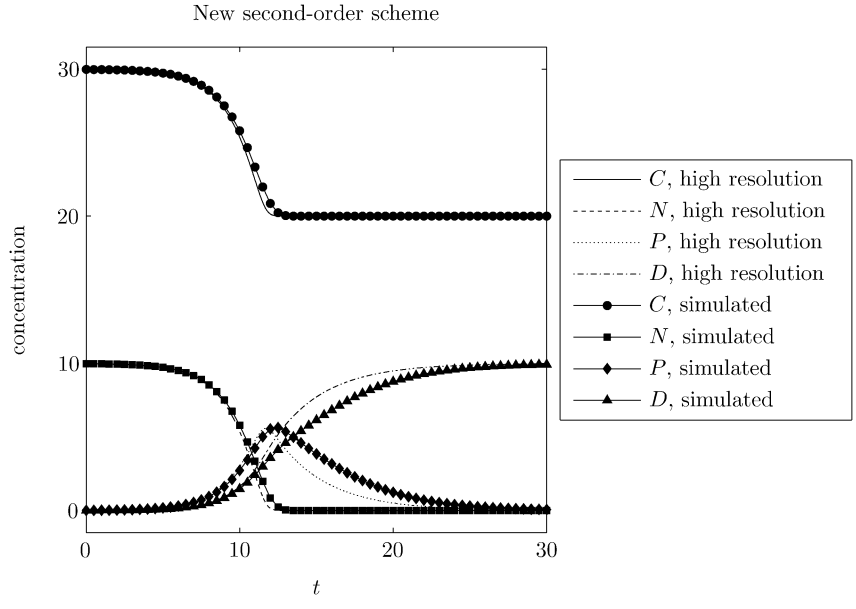

Fig. 13. Coarse numerical approximation ($\Delta t = 0.5$) with the new second-order scheme and high-resolution numerical approximation ($\Delta t = 0.01$) with a Runge–Kutta 4 scheme [6, p. 138] for the simple non-linear system (1).

remark that the second-order scheme (22) also provides us with a first-order approximation to the solution in the new time level. The difference between this value and the final result can be considered a (conservative) estimate of the local truncation error, and thus serve as the basis of an error estimator.

Further, we have shown that the modified Patankar and modified Patankar–Runge–Kutta schemes presented by Burchard et al. [3] are not conservative for any arbitrary biochemical system. The new schemes do conserve mass and energy in two special cases, namely if: (1) all reactions in the system contain one source compound, or (2) the relative change over time is the same for all source compounds.

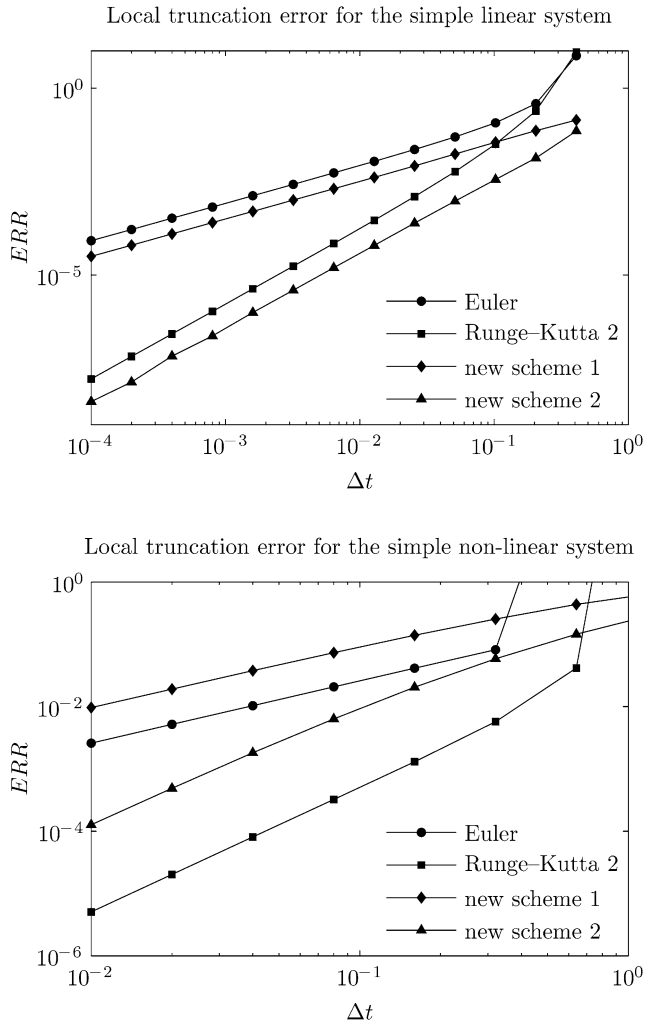

Fig. 14. The local truncation error summarized over all state variables, as defined in Eq. (9), for the reference forward Euler and Runge–Kutta 2 ((22) without product terms) schemes and the new first-order and second-order schemes. Above: results for the simple linear system (10), calculated with $\Delta t = 2^i \times 10^{-4}$ with $i = 0, \ldots, 12$, and $t_{\max} = 1.8$. Below: results for the simple non-linear system (1), calculated with $\Delta t = 2^i \times 10^{-2}$ with $i = 0, \ldots, 7$, and $t_{\max} = 30$. The strong increases in local truncation error at high Δt for the Euler and Runge–Kutta schemes are due to instability.

Condition (1) is only met if one deals with simple biochemical systems [5]; realistic systems readily include multiple source compounds per reaction. For instance, most marine ecosystems are modeled with phytoplankton growing on two or more nutrients (e.g. nitrate, phosphate, iron), and thus contain reactions with multiple source compounds [13]. It must be noted that in some cases, a system that does not satisfy condition (1) can be reduced to a system that does via elimination of variables (elimination is possible in any conservative system, as rows of **S** *cannot* be linearly independent; see Definition 7). This is demonstrated in Reder [19]. One could apply this technique to reduce the simple non-linear system (1) to a two-dimensional system that satisfies condition (1). However, for more complex and realistic systems, this is typically not feasible in practice.

Condition (2) is even more unlikely to be met, as it requires both specific initial state variable values (their ratio corresponding to the stoichiometric ratio of use), and highly simple system kinetics.

This does not imply that the modified Patankar schemes are without value: for biochemical systems that do meet one of both conditions, the modified Patankar schemes offer a relatively inexpensive approach, which incidentally is known to perform well on stiff systems [3,4].

In the present paper, we do not analyze the stability properties of the new schemes in detail. Application of the new schemes to the infamous, highly stiff Robertson test case proved that the schemes have at least some problems with

S0168-9274(05)00224-2/FLA AID:1854 Vol.●●●(●●●)
APNUM:m3SC+ v 1.50 Prn:12/01/2006; 12:16
apnum1854
[+model] P.22 (1-23)
by:R.M. p. 22

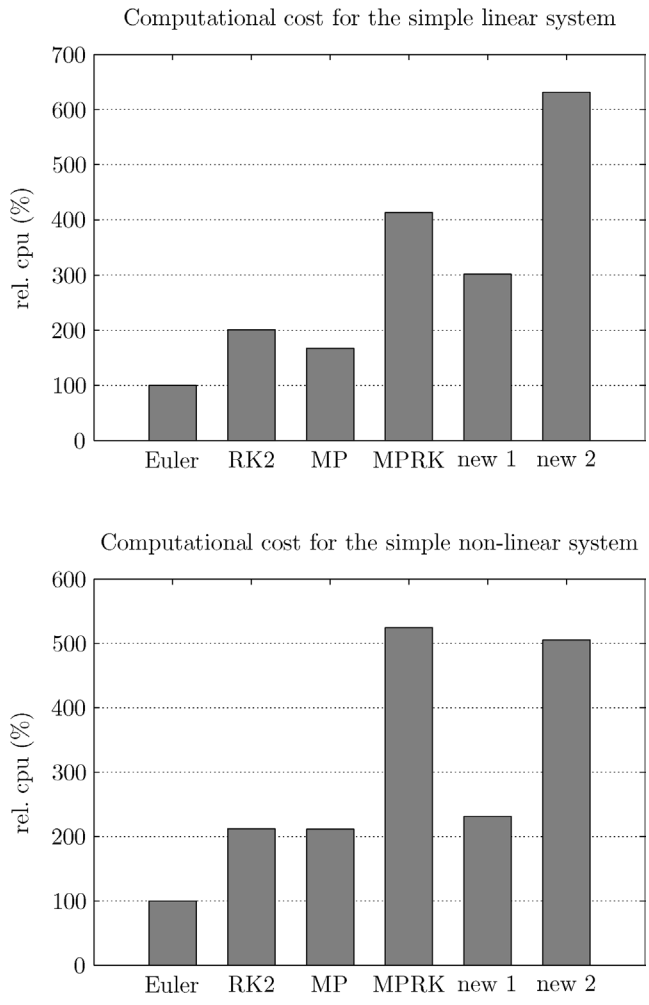

Fig. 15. Execution times of six integration schemes, relative to the forward Euler scheme, for the two sample systems. Shown respectively: the forward Euler scheme, the second-order Runge–Kutta scheme (RK2) ((22) without product terms), the Modified Patankar (MP) scheme [3], the Modified Patankar–Runge–Kutta (MPRK) scheme [3], the new first-order scheme (new 1) and the new second-order scheme (new 2). Computational costs for the backward Euler method depend strongly on the means chosen to solve the non-linear system of equations, and are therefore not shown.

(highly) stiff systems: the new first-order scheme rendered oscillating (but positive) solutions, whereas the second-order scheme stalled the system through extremely small p factors. This behavior differs notably from that of the modified Patankar schemes, which were capable of resolving the Robertson test case [3] and other stiff biochemical problems [4]. This difference between Patankar-inspired schemes can intuitively be explained as follows: MP schemes adjust the effective time step for different components of the system individually, whereas the new schemes adapt the effective time step for the whole system to the fastest component. Nevertheless, under no circumstances will the new schemes perform more poorly with respect to stability than the traditional schemes they were based upon, since the new schemes can be interpreted as traditional schemes with a downscaled time step. For many purposes they could suffice; preliminary results showed that the new schemes were capable of solving the realistic stiff test cases of [4]. This line of thought also suggests alternative approaches: one might selectively slow down individual reaction rates rather than the whole system, thus obtaining a scheme that performs well on stiff systems, and (unlike MP schemes) adheres to biochemical conservation.

In conclusion, this paper has presented a structured, mathematical approach to the biochemical concept of conservation. This approach integrates recurrent ideas in biochemistry [20,19,10,7], and, to our knowledge, has not before been used in the analysis of numerical schemes. It may provide a context of analysis for existing and future schemes, and will hopefully result in an increasing number of schemes known to be suited for biochemical problems.

Acknowledgements

The work of Jorn Bruggeman has been funded by the Computational Life Sciences program of the Netherlands Organisation for Scientific Research (NWO), project number 635.100.009. The work of Ben Sommeijer has been funded by the Dutch BSIK/BRICKS project. We thank Andreas Meister, Bas Kooijman, and two anonymous referees for their comments.

References

[1] C. Bolley, M. Crouzeix, Conservation de la positivité lors de la discrétisation des problèmes d'évolution parabolic, RAIRO Anal. Numer. 12 (3) (1978) 237–245.
[2] H. Burchard, K. Bolding, W. Kuhn, A. Meister, T. Neumann, L. Umlauf, Description of a flexible and extendable physical-biogeochemical model system for the water column, J. Marine Systems, in press.
[3] H. Burchard, E. Deleersnijder, A. Meister, A high-order conservative Patankar-type discretisation for stiff systems of production–destruction equations, Appl. Numer. Math. 47 (1) (2003) 1–30.
[4] H. Burchard, E. Deleersnijder, A. Meister, Application of Modified Patankar schemes of stiff biogeochemical models for the water column, Ocean Dynamics 55 (3–4) (2005) 326–337.
[5] M.J.R. Fasham, H.W. Ducklow, S.M. Mckelvie, A nitrogen-based model of plankton dynamics in the oceanic mixed layer, J. Marine Res. 48 (3) (1990) 591–639.
[6] E. Hairer, S.P. Nørsett, G. Wanner, Solving Ordinary Differential Equations I: Nonstiff Problems, second revised ed., Springer Ser. Comput. Math., vol. 8, Springer, Berlin, 1993.
[7] P.P.F. Hanegraaf, B.W. Kooi, The dynamics of a tri-trophic food chain with two-component populations from a biochemical perspective, Ecological Modelling 152 (1) (2002) 47–64.
[8] W.H. Hundsdorfer, J.G. Verwer, Numerical Solution of Time-Dependent Advection–Diffusion–Reaction Equations, Springer Ser. Comput. Math., vol. 33, Springer, Berlin, 2003.
[9] H. Jansen, E.H. Twizell, An unconditionally convergent discretization of the SEIR model, Math. Comput. Simulation 58 (2) (2002) 147–158.
[10] B.W. Kooi, P.P.F. Hanegraaf, Bi-trophic food chain dynamics with multiple component populations, Bull. Math. Biol. 63 (2) (2001) 271–299.
[11] S.A.L.M. Kooijman, The synthesizing unit as model for the stoichiometric fusion and branching of metabolic fluxes, Biophys. Chem. 73 (1–2) (1998) 179–188.
[12] S.A.L.M. Kooijman, Dynamic Energy and Mass Budgets in Biological Systems, second revised ed., Cambridge University Press, Cambridge, 2000.
[13] L.D.J. Kuijper, B.W. Kooi, T.R. Anderson, S.A.L.M. Kooijman, Stoichiometry and food-chain dynamics, Theoret. Population Biol. 66 (4) (2004) 323–339.
[14] R.E. Mickens, Applications of Nonstandard Finite Difference Schemes, World Scientific, Singapore, 2000.
[15] T. Neumann, W. Fennel, C. Kremp, Experimental simulations with an ecosystem model of the Baltic Sea: A nutrient load reduction experiment, Global Biogeochem. Cycles 16 (3) (2002).
[16] S.V. Patankar, Numerical Heat Transfer and Fluid Flow, Series in Computational Methods in Mechanics and Thermal Sciences, McGraw-Hill, New York, 1980.
[17] J. Pietrzak, The use of TVD limiters for forward-in-time upstream-biased advection schemes in ocean modeling, Monthly Weather Rev. 126 (3) (1998) 812–830.
[18] W. Piyawong, E.H. Twizell, A.B. Gumel, An unconditionally convergent finite-difference scheme for the SIR model, Appl. Math. Comput. 146 (2–3) (2003) 611–625.
[19] C. Reder, Metabolic control-theory—a structural approach, J. Theoret. Biol. 135 (2) (1988) 175–201.
[20] J.A. Roels, Energetics and Kinetics in Biotechnology, Elsevier Biomedical Press, Amsterdam, 1983.
[21] A. Sandu, Positive numerical integration methods for chemical kinetic systems, J. Comput. Phys. 170 (2) (2001) 589–602.
[22] L.F. Shampine, Conservations laws and the numerical solution of ODEs, Comput. Math. Appl. B 12 (5–6) (1986) 1287–1296.