Positive and Conservative Schemes for Mass Action Kinetics

E. Bertolazzi
Laboratorio di Matematica Applicata, Dipartimento di Ingegneria Civile ed Ambientale
Università degli Studi di Trento, Mesiano, Trento, Italy
bertolazz@ing.unitn.it

(Received and accepted May 1996)

Abstract—Some implicit schemes for the discretization of mass action kinetics are presented and discussed. Although all the schemes are classical, the solution procedure takes great advantage of the peculiar structure of the ODE. All the schemes maintain linear first integrals such as atomic mass conservation. The main features of the first-order scheme is positivity preservation for arbitrarily large time steps, while for the higher order schemes, there is numerical evidence that this property is maintained with fairly large time steps. For the first-order scheme, the existence of a nonnegative solution of the algebraic system arising in the advancing step is proved for arbitrarily large time steps, while uniqueness is proved for limited time steps. Finally, an efficient procedure is presented for solving the nonlinear systems involved in the time step. These systems are solved without the use of Jacobian matrices or their approximations, but by the repeated inversion of $M$-matrices, a procedure which is both easy and fast, considerably simplifying and accelerating the computer implementation of the schemes.

Keywords—Chemical kinetics, ODE, $M$-matrices, Brouwer fixed point.

1. INTRODUCTION

The equations describing reacting systems may be quite difficult to solve numerically. The difficulties arise from the stiffness of the equations, and sometimes because these equations are part of a hyperbolic system where the speeds of the reactions are several orders of magnitude greater than the characteristic velocities of the system. The equations have analytic properties reflecting important physical laws that must be maintained in the numerical solution. Classical schemes have numerical analogs of these properties, but they are subject to severe time step restrictions. Such limitations can be excessive for accuracy considerations and computer time. For these reasons, the construction of schemes for the reaction terms that do not suffer time-step restrictions is important. The aim of this paper is to describe some numerical schemes that maintain numerical analogous of physical properties without time step restrictions.

In the literature, there are many papers devoted to the solution of stiff systems of ODE of the form

$$\dot{y} = f(y), \quad y(0) = y_0.$$  

The most popular one-step numerical schemes seem to be implicit or diagonally implicit Runge-Kutta methods (see, e.g., [1–6]). All these schemes require at each time step the solution of a nonlinear algebraic system, which is usually achieved by Newton or Newton-like techniques.
These, in turn, require the knowledge of the Jacobian matrix of $f$ or of a corresponding numerical approximation. As a consequence, the iteration procedure at each time step can be very time consuming. It is most important for the applications to find a robust and fast procedure for solving the nonlinear system at each time step. On the basis of numerical experimentation, we believe that the procedure we propose achieves these results for the stiff system related to mass action kinetics.

2. THE EQUATIONS

Consider a generic reacting system, and for simplicity, under constant environmental conditions (temperature, pressure, etc.) and without source or sink terms. Consider now a generic reaction

$$aA + bB + \cdots = mM + nN + \cdots$$

that can be written in the form

$$\sum_i \sigma_i A_i = 0,$$

(1)

where $A_i$ is a generic substance and $\sigma_i$ its stochiometric coefficient. These coefficients permit us to correlate the variation in molar concentration of the substances involved in the reaction as follows:

$$\frac{1}{\sigma_1} \frac{dn_1}{dt} = \frac{1}{\sigma_2} \frac{dn_2}{dt} = \cdots = \frac{1}{\sigma_s} \frac{dn_s}{dt} = r,$$

(2)

where $n_i$ is the molar concentration of the $i^{th}$ substance, and $r$, a nonnegative number, is the reaction velocity. Because the environmental conditions are assumed constant, it follows that, in general, $r = r(n_1, n_2, \ldots, n_s)$. Moreover, $r$ is defined only for $n_i \geq 0$ and is a continuous function. Introducing

$$R_+^s = \{ [x_1, x_2, \ldots, x_s] \in \mathbb{R}^s | x_i \geq 0, i = 1, 2, \ldots, s \},$$

it follows that $r \in C \left( R_+^s, R_+ \right)$. We can write equation (2) as

$$\frac{dn_i}{dt} = \sigma_i r(n_1, n_2, \ldots, n_s), \quad i = 1, 2, \ldots, s.$$  

(3)

If more reactions, for example $k$, are included in the system, then (1) will be modified as follows:

$$\sum_{i=1}^s \sum_{j=1}^k \sigma_{i,j} A_i = 0, \quad j = 1, 2, \ldots, k,$$

where $\sigma_{i,j}$ is the stochiometric coefficient of substance $i$ in the $j^{th}$ reaction and (3) becomes

$$\frac{dn_i}{dt} = \sum_{j=1}^k \sigma_{i,j} r_j(n_1, n_2, \ldots, n_s), \quad i = 1, 2, \ldots, s,$$

(4)

where $r_j$ is the speed of the $j^{th}$ reaction. Introducing

$$n = \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_s \end{bmatrix}, \quad r(n) = \begin{bmatrix} r_1(n) \\ r_2(n) \\ \vdots \\ r_k(n) \end{bmatrix}, \quad S = \begin{bmatrix} \sigma_1, 1, \sigma_1, 2, \ldots, \sigma_1, k \\ \sigma_2, 1, \sigma_2, 2, \ldots, \sigma_2, k \\ \vdots \\ \sigma_s, 1, \sigma_s, 2, \ldots, \sigma_s, k \end{bmatrix},$$

system (4) can be rewritten as

$$\frac{dn}{dt} = Sr(n).$$  

(5)
If $M_i$ is the atomic mass of the $i^{th}$ species, then $\rho_i = M_i n_i$ is its density. Defining

$$D = \text{diag}(M_1, M_2, \ldots, M_s),$$
$$\rho = D n,$$
$$f(\rho) = r(D^{-1} \rho),$$
$$\mathcal{V} = DS,$$

chemical system (5) can be written in term of densities as follows:

$$\frac{d\rho}{dt} = \mathcal{V} f(\rho).$$  \hspace{1cm} (6)

In the following sections, we will study this system.

2.1. Reaction Rate Constraints

Consider a single reaction system, for which, we have

$$f(\rho) = [f(\rho)], \quad \mathcal{V} = \begin{bmatrix}
\nu_1 \\
\nu_2 \\
\vdots \\
\nu_s
\end{bmatrix}.$$ If $\nu_i < 0$, it means that the species $\rho_i$ is a reactant consumed in the reaction. Obviously, if a reactant is absent ($\rho_i = 0$), then the reaction is stopped and the reaction rate $f$ must be zero. This constraint can be formulated as follows:

$$\text{if } \nu_i < 0 \text{ and } \rho_i = 0 \implies f(\rho) = 0.$$  \hspace{1cm} (7)

Constraint (7) is too weak to be useful, so we need some stronger analytic requirement for the reaction rates. In mass action, kinetics reaction rates often have the form

$$f(\rho) = \alpha_1 \rho_1^{\alpha_2} \cdots \rho_s^{\alpha_s},$$

where $c$ is a constant and $\alpha_i \geq 0$. The requirement (7) for (8) becomes

$$\text{if } \nu_i < 0 \implies \alpha_i \geq 1.$$ 

or equivalently, if $\nu_i < 0 \implies f(\rho)/\rho_i \in C \left(\mathbb{R}^+_* , \mathbb{R}_+^* \right)$. This suggests the following general requirement for the reaction rate $f$ with its stoechiometric vector $\mathcal{V}$:

(i) $f \in C \left(\mathbb{R}^+_* , \mathbb{R}_+^* \right)$, and

(ii) if $\nu_i < 0$, then the function $q(\rho) := f(\rho)/\rho_i$ is such that $q \in C \left(\mathbb{R}^+_* , \mathbb{R}_+^* \right)$.

2.2. Nonnegativity of the Solutions

The density vector $\rho$ is defined only for nonnegative values of its components, so that equation (6) is applicable if nonnegativity of its solution is assured when the initial data are nonnegative. Denoting by $\rho(t; \rho_0)$, the solution of system (6) with $\rho_0$ as initial value, then this requirement becomes

$$\text{if } \rho_0 \geq 0 \implies \forall t > 0, \quad \rho(t; \rho_0) \geq 0.$$  \hspace{1cm} (9)

Assuming constraints (i) and (ii) for all the reaction rates, then it is easy to show that (9) holds (see [7]).
2.3. First Integrals

We consider the kernel of $\mathcal{V}^T$, i.e.,

$$\text{Ker} (\mathcal{V}^T) = \{ z | z^T \mathcal{V} = 0 \}$$

if $z \in \text{Ker} (\mathcal{V}^T)$ then, from (6), it follows that

$$z^T \frac{d \rho}{dt} = z^T \mathcal{V} f(\rho) = 0,$$

so that the function

$$g_0(\rho) = z^T \rho, \quad z \in \text{Ker} (\mathcal{V}^T)$$

is a first integral of (6). In the case of a conservative system, we have $\|\rho\|_1 = \text{constant}$ so that $g_0(\rho) = e^T \rho$ is a first integral of (6), where $e = [1, 1, \ldots, 1]^T$. From now on, it is assumed that $g_0$ is a first integral of the system (6), or equivalently, that $e \in \text{Ker} (\mathcal{V}^T)$.

3. DISCRETIZATION

The aim of the paper is to develop a numerical scheme for which the numerical solution possesses a counterpart of the properties (9) and (10) previously outlined, without stability restrictions. A general one-step scheme can be written as

$$\rho^{n+1} = G(\rho^n), \quad n = 1, 2, \ldots,$$

where $G$ is the time-step map. For this scheme, the conservation property (10) can be written as

$$z^T \rho^{n+1} = z^T \rho^n, \quad \forall z \in \text{Ker} (\mathcal{V}^T),$$

and nonnegativity property (9) for the scheme (10) can be written as

$$\text{if} \quad \rho^n \geq 0 \quad \Rightarrow \quad \rho^{n+1} = G(\rho^n) \geq 0.$$

Properties (12) and (13) are equivalent to the following properties of the map $G$:

$$G(\rho) \geq 0, \quad \forall \rho \geq 0,$$

$$z^T(\rho - G(\rho)) = 0, \quad \forall z \in \text{Ker} (\mathcal{V}^T), \quad \forall \rho \geq 0.$$

In the following sections, it is shown that classical schemes do not share both properties (14a) and (14b).

4. AN EXAMPLE

Consider a single reaction system

$$\frac{d \rho}{dt} = \mathcal{V} f(\rho),$$

where the reaction rate $f$ is an homogeneous function of degree 1, for example

$$f(\rho) = c \rho_1^{\alpha_1} \rho_2^{\alpha_2} \cdots \rho_s^{\alpha_s},$$

with

$$\alpha_i = \begin{cases} 1, & \text{if } \nu_i < 0, \\ 0, & \text{if } \nu_i \geq 0, \end{cases}$$

and discretize (15) with some classical schemes.
4.1. Explicit Euler Scheme

This scheme can be written as

\[ \frac{\rho^{n+1} - \rho^n}{\Delta t} = \nabla f(\rho^n), \]

for which the map \( G \) becomes

\[ G(\rho) = \rho + \Delta t \nabla f(\rho), \]

and conditions (14a) and (14b) become

- for the nonnegativity (14a), it is necessary and sufficient that

\[ \Delta t \leq \frac{\rho_i^n}{-\nu_i \nabla f(\rho^n)}, \quad \text{if} \quad \nu_i < 0, \]

- the conservation (14b) is always satisfied, in fact

\[ z^T (\rho - G(\rho)) = -\Delta t z^T \nabla f(\rho^n) = 0, \quad \forall z \in \text{Ker} (\nabla^T). \]

4.2. Linearized Implicit Euler Scheme

Consider the implicit Euler scheme

\[ \frac{\rho^{n+1} - \rho^n}{\Delta t} = \nabla f(\rho^{n+1}), \]

and use a single Newton-Raphson step to approximate \( \rho^{n+1} \)

\[ \rho^{n+1} \approx \rho^{n+1} = \rho^n - (I - \Delta t \nabla \nabla f(\rho^n))^{-1} (\rho^n - \Delta t \nabla f(\rho^n) - \rho^n) \]

\[ = (I - \Delta t \nabla \nabla f(\rho^n))^{-1} (\rho^n + \Delta t \nabla f(\rho^n) - \nabla f(\rho^n) \rho^n)) \]

\[ = (I - \Delta t \nabla \nabla f(\rho^n))^{-1} \rho^n, \]

where homogeneity of \( f \) was used. Using \( \rho^{n+1} \) instead of \( \rho^{n+1} \), we obtain the linearized implicit Euler scheme, for which the solution step is

\[ \rho^{n+1} = (I - \Delta t \nabla \nabla f(\rho^n))^{-1} \rho^n, \]

and using the Sherman-Morrison formula [8]

\[ (A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}, \]

then the map \( G \) becomes

\[ G(\rho) = \rho + \frac{\Delta t \nabla \nabla f(\rho) \rho}{1 - \Delta t \nabla \nabla f(\rho) \nabla}, \]

(17)

and conditions (14a) and (14b) become

- for the nonnegativity (14a), it is necessary and sufficient that

\[ \Delta t \leq \frac{\rho_i^n (1 - \Delta t \nabla \nabla f(\rho^n) \nabla)}{-\nu_i \nabla \nabla f(\rho^n) \rho^n}, \quad \text{if} \quad \nu_i < 0, \]

- the conservation (14b) is always satisfied, in fact from (17), it follows that

\[ z^T (\rho - G(\rho)) = -\Delta t z^T \nabla \nabla f(\rho) \rho = 0, \quad \forall z \in \text{Ker} (\nabla^T). \]
4.3. Runge-Kutta's Fourth-Order Scheme

The Runge-Kutta's Fourth-Order Scheme for System (6) takes the form

\[ K_0 = \Delta t \, f(\rho^n) \, \nu, \]
\[ K_1 = \Delta t \, f\left(\rho^n + \frac{K_0}{2}\right) \, \nu, \]
\[ K_2 = \Delta t \, f\left(\rho^n + \frac{K_1}{2}\right) \, \nu, \]
\[ K_3 = \Delta t \, f(\rho^n + K_2) \, \nu, \]
\[ \rho^{n+1} = \rho^n + \frac{K_0 + 2K_1 + 2K_2 + K_3}{6}. \]

- A necessary condition for the applicability is \( \rho^n + K_0/2 \geq 0 \), so that

\[ \Delta t \leq \frac{2\rho^n}{-\nu_i f(\rho^n)}, \quad \text{if} \quad \nu_i < 0. \]

- The conservation (14b) is always satisfied, in fact

\[ z^T K_0 = \Delta t f(\rho^n) z^T \nu = 0, \quad \forall z \in \text{Ker}(\nu^T), \]

and analogously for \( K_1, K_2, K_3 \).

5. A SEMI-IMPLICIT SCHEME

All the schemes previously considered have discrete analogs of the first integrals (10), but they suffer a time-step limitation for nonnegativity. To avoid these limitations, a semi-implicit scheme is developed. First, a technical lemma is needed.

**Lemma 1.** Let \( v \in \mathbb{R}^s \) be a vector such that

\[ \sum_{i=1}^{s} v_i = 0, \]

and let \( f : \mathbb{R}^s_+ \mapsto \mathbb{R}^s_+ \) be a continuous function, such that, for those indices \( i \) for which \( v_i < 0 \),

\[ \frac{f(x)}{x_i} \] is also continuous on \( \mathbb{R}^n_+ \).

Then \( f(x)v \) can be written as

\[ C(x,v)x = f(x)v, \]

where \( C(x,v) \) is a \( s \times s \) matrix with continuous entries such that

\[ C_{i,i}(x,v) \leq 0, \quad i = 1, 2, \ldots, s, \]
\[ C_{i,j}(x,v) \geq 0, \quad i \neq j, \]
\[ \sum_{i=1}^{s} C_{i,j}(x,v) = 0, \quad j = 1, 2, \ldots, s. \]

**Proof.** Let \( v = v^+ + v^- \) where

\[ v^+ = [\max(0,v_1), \max(0,v_2), \ldots, \max(0,v_s)]^T, \]
\[ v^- = [\min(0,v_1), \min(0,v_2), \ldots, \min(0,v_s)]^T. \]

By (19), it is possible to define the matrix \( C(x,v) \) as follows:

\[ C(x,v) = \sum_{i=1}^{s} \left( \frac{f(x)}{x_i} v_i^- \right) \left[ e_i - \frac{v^+}{\|v^+\|_1} \right] e_i^T, \]

where \( e_i \) are the vectors of the canonical base in \( \mathbb{R}^s \), i.e., the \( i^{th} \) component is one, while all the others are zero. By a straightforward computation, it is easy to verify that (18) holds.
THEOREM 2. Let $\mathcal{V}$ be a $s \times k$ matrix and $f : \mathbb{R}_+^s \mapsto \mathbb{R}_+^k$ be a continuous map; then if for those indices $i, j$ for which $\nu_{i,j} < 0$,

$$\frac{f_i(x)}{x_j},$$

is also continuous on $\mathbb{R}_+^n$, and

$$\sum_{i=1}^s \nu_{i,j} = 0, \quad j = 1, 2, \ldots, k,$$

then $\mathcal{V}f(\rho)$ can be written in the form $C(\rho, \mathcal{V})\rho$, where $C(\rho, \mathcal{V})$ is a $s \times s$ matrix with continuous entries, such that

- (a) $C_{i,i}(\rho, \mathcal{V}) \leq 0$, $i = 1, 2, \ldots, s$,
- (b) $C_{i,j}(\rho, \mathcal{V}) \geq 0$, $i \neq j$,
- (c) $\sum_{i=1}^s C_{i,j}(\rho, \mathcal{V}) = 0$, $j = 1, 2, \ldots, s$.

PROOF. The source term $\mathcal{V}f(\rho)$ can be written as

$$\mathcal{V}f(\rho) = \sum_{j=1}^k f_j(\rho)\mathcal{V}_{i,j}.$$ 

The pair $f_j, \mathcal{V}_{i,j}$ ($j$th column of $\mathcal{V}$) satisfies the hypothesis of Lemma 1, so that we can assert

$$C^{(j)}(\rho, \mathcal{V}_{i,j})\rho = f_j(\rho)\mathcal{V}_{i,j}, \quad j = 1, 2, \ldots, k,$$

with $C^{(j)}(\rho, \mathcal{V}_{i,j})$ that satisfies properties (a)–(c). Then, it is possible to define

$$C(\rho, \mathcal{V}) = \sum_{j=1}^k C^{(j)}(\rho, \mathcal{V}_{i,j}),$$

and $C(\rho, \mathcal{V})$ obviously has the properties (a)–(c).

With the matrix $C$, it is possible to define a simple semi-implicit numerical scheme for (6) as follows:

$$\frac{\rho^{n+1} - \rho^n}{\Delta t} = C(\rho^n, \mathcal{V})\rho^{n+1},$$

which results in the following advancing step:

$$\rho^{n+1} = (I - \Delta t C(\rho^n, \mathcal{V}))^{-1} \rho^n.$$

The matrix $I - \Delta t C(\rho^n, \mathcal{V})$ is strictly diagonally dominant with elements positive on the diagonal and nonnegative elsewhere, consequently, it is a $M$-matrix. By definition of $M$-matrix, it follows that $(I - \Delta t C(\rho^n, \mathcal{V}))^{-1} \geq 0$, and consequently, $\rho^{n+1} \geq 0$ for arbitrarily large $\Delta t$. Unfortunately, this scheme has not a numerical analogs of first integral (10).

6. A FULLY IMPLICIT SCHEME

All the previous considered schemes cannot have both a discrete analogs of first integral (10) and nonnegativity preservation for arbitrarily large time step. The implicit Euler scheme

$$\frac{\rho^{n+1} - \rho^n}{\Delta t} = \mathcal{V}f(\rho^{n+1})$$ (20)

can be used to avoid time-step restrictions, but the advancing steps involve the solution of a nonlinear system as follows:

$$\rho^{n+1} = \text{the solution of } x - \Delta t \mathcal{V}f(x) - \rho^n = 0.$$

For system (21), there is the question of existence and uniqueness of the solution and an iterative procedure is needed to find the solution.
6.1. Existence of a Solution

A nonnegative solution of the nonlinear system

\[ x - \Delta t Vf(x) - \rho^n = 0, \quad (22) \]

is also, by Theorem 2, a fixed point of the map

\[ \Phi(x) := (I - \Delta t C(x, V))^{-1} \rho^n. \quad (23) \]

**Theorem 3.** The map \( \Phi \) admits at least one nonnegative fixed point (i.e., with nonnegative components). Moreover if \( x^* \) is a nonnegative fixed point, then \( \|x^*\|_1 = \|\rho^n\|_1 \).

**Proof.** The map \( \Phi \) has the property \( \forall \rho \geq 0 \Rightarrow \Phi(\rho) \geq 0 \), and from (23), we can write

\[ \Phi(x) - \Delta t C(x, V)\Phi(x) = \rho^n, \quad (24) \]

multiplying (24) by \( e^T \) and using the fact \( e^T C(x, V) = 0 \), it follows

\[ e^T \Phi(x) = e^T \rho^n \Rightarrow \|\Phi(x)\|_1 = \|\rho^n\|_1, \quad \forall x \geq 0. \]

Consequently, the image \( \Phi(\mathbb{R}^+)^T \) is contained into the convex compact

\[ \mathcal{K} = \{ x \geq 0 \mid \|x\|_1 = \|\rho^n\|_1 \}, \quad (25) \]

so that, if \( x^* \) is a fixed point, it must be contained in \( \mathcal{K} \) and \( \|x^*\|_1 = \|\rho^n\|_1 \). Moreover, the map \( \Phi \) can be viewed as a continuous map from \( \mathcal{K} \) into \( \mathcal{K} \), and by the Brouwer fixed point theorem [9], it follows that \( \Phi \) has at least one fixed point.

Observe that the proof is independent of the magnitude of \( \Delta t \), so that the nonlinear system (22) has a nonnegative solution no matter how large the time step is.

6.2. The Question of Uniqueness

In general, it is not possible to see if system (22) has a unique solution for arbitrarily large \( \Delta t \). However, there are some special cases for which we have also uniqueness. It is the case, for example, when the system consists of only one reaction and the reaction rate \( f \) satisfies

\[ \nu_i \partial f \rho_i \leq 0, \quad i = 1, 2, \ldots, s, \]

which exclude auto-inhibition in the reaction. In this case, if \( x^* \) and \( y^* \) are two solutions of (22), it follows

\[ 0 = x^* - y^* - \Delta t V(f(x^*) - f(y^*)) = (I - \Delta t V\nabla f(\xi))(x^* - y^*), \quad (26) \]

and using the Sherman-Morrison formula (16)

\[ (I - \Delta t V\nabla f(\xi))^{-1} = I + \frac{\Delta t V\nabla f(\xi)}{1 - \Delta t \nabla f(\xi) V}, \quad (27) \]

so that by (26) and (27), it follows that \( x^* = y^* \). In the case of more than one reaction, it is possible to prove uniqueness for small \( \Delta t \).
THEOREM 4. Let \( \mathcal{V}(\rho) \) be a \( s \times k \) matrix and \( f \in C^1(\mathbb{R}^s_+, \mathbb{R}_+) \), then if for those indices \( i, j \) for which \( \nu_{i,j} < 0 \),

\[
\frac{f_i(x)}{x_j} \in C^1(\mathbb{R}^s_+, \mathbb{R}_+)
\]

then, for all \( \Delta t \) satisfying

\[
\Delta t < \frac{1}{s \| \rho^n \|_1 \max_{x \in K} \| \nabla C_i, j (z, \nu) \|_\infty},
\]

the map \( \Phi : \mathbb{R}^s_+ \rightarrow \mathbb{R}^s_+ \) defined in (23) is a contraction, where \( K \) is given by (25).

PROOF. Observe that \( V \geq 0 \)

\[
e^T (I - \Delta t C(z, \nu)) = e^T \quad \Rightarrow \quad e^T = e^T (I - \Delta t C(z, \nu))^{-1},
\]

so that it follows \( \| (I - \Delta t C(z, \nu))^{-1} \|_1 = 1 \). Next

\[
\Phi(x) - \Phi(y) = (I - \Delta t C(x, \nu))^{-1} \rho^n - (I - \Delta t C(y, \nu))^{-1} \rho^n
\]

\[
= \Delta t (I - \Delta t C(x, \nu))^{-1} (C(x, \nu) - C(y, \nu)) (I - \Delta t C(y, \nu))^{-1} \rho^n,
\]

and taking the \( \| \cdot \|_1 \) norm on both sides

\[
\| \Phi(x) - \Phi(y) \|_1 \leq \Delta t \| C(x, \nu) - C(y, \nu) \|_1 \| \rho^n \|_1 \leq \Delta t L \| x - y \|_1,
\]

where

\[
L = s \| \rho^n \|_1 \max_{x \in K} \| \nabla C_i, j (z, \nu) \|_\infty,
\]

so that if \( \Delta t < 1/L \) the map \( \Phi \) becomes a contraction.

Obviously, if the map \( \Phi \) is a contraction, then it has a unique fixed point.

6.3. An Iterative Procedure

Theorem 4 suggests the use of \( \Phi \) as an iteration map to approximate the solution of the nonlinear system (22). The procedure is the following algorithm.

ALGORITHM 1.

Set \( x^0 = \rho^n \)

repeat

\[
x^{i+1} = (I - \Delta t C(x^i, \nu))^{-1} \rho^n
\]

until convergence

At convergence set \( \rho^{n+1} = x^{i+1} \).

Algorithm 1 was successfully used in [10] although \( \Delta t \) violates the condition of Theorem 4.

7. A SECOND-ORDER POSITIVE SCHEME

The implicit Euler scheme has no stability restriction, but it is only first-order accurate. A second-order scheme can be the following:

\[
\frac{\rho^{n+1} - \rho^n}{\Delta t} = \nu \frac{f(\rho^{n+1}) + f(\rho^n)}{2}.
\]

The solution step, as in (20), involves the solution of a nonlinear system in the unknown \( x \)

\[
x - \Delta t \frac{f(x)}{2} = \rho^n + \Delta t \frac{\mathcal{V}(\rho^n)}{2}.
\]
Following the same lines of Theorem (3), in order to guarantee the existence of a nonnegative solution, it is sufficient that
\[
\rho^n + \Delta t \frac{\nabla \cdot (\rho^n)}{2} \geq 0.
\]
This condition introduces a time-step bound. To avoid time-step restriction, an alternative approach is to switch to second-order accuracy when reactions are slow and first-order accuracy when reactions are fast. The scheme becomes
\[
\frac{\rho^{n+1} - \rho^n}{\Delta t} = \nabla \cdot \left( (1 - \alpha) f (\rho^{n+1}) + \alpha f (\rho^n) \right), \tag{28}
\]
where \(\alpha\) is such that
\[
\max_{\alpha \in [0,1/2]} (\rho^n + \alpha \Delta t \nabla \cdot (\rho^n)) \geq 0.
\]
If the reaction rates have very different orders of magnitudes, this scheme can lose too much accuracy for slow reactions. A better result can be obtained by using different \(\alpha\)'s for each reaction as follows:
\[
\frac{\rho^{n+1} - \rho^n}{\Delta t} = \sum_{j=1}^{k} \nabla_{ij} \left[ (1 - \alpha_i) f_j (\rho^{n+1}) + \alpha_j f_j (\rho^n) \right]. \tag{29}
\]
With (29), the solution step involves the solution of the nonlinear system in the unknown \(x\)
\[
x - \Delta t \sum_{j=1}^{k} (1 - \alpha_j) \nabla_{ij} f_j (x) = \rho^n + \Delta t \sum_{j=1}^{k} \alpha_j \nabla_{ij} f_j (\rho^n).
\]
For positivity preservation, it suffices that
\[
\rho^n + \sum_{j=1}^{k} \alpha_j \nabla_{ij} f_j (\rho^n) \geq 0. \tag{30}
\]
In order to satisfy condition (30), we set
\[
\alpha_j = \min \left\{ \frac{1}{2}, \left\{ \frac{\beta_j \rho^n}{\Delta t \nu_{ij} f_j (\rho^n)} \mid \nu_{ij} < 0, \ i = 1,2,\ldots,s \right\} \right\},
\]
where \(\beta_j > 0\) are weighting parameters with \(\beta_1 + \beta_2 + \cdots + \beta_k = 1\). The weighting factors are useful; if, for example, the first reaction is very slow, we can use very small \(\beta_1\) and maintain \(\alpha_1 = 1/2\). This permits us to increase the \(\alpha\)'s for the fast reactions. The weighting factors can be chosen in many ways. A simple choice could be \(\beta_i = 1/k\).

The final numerical procedure for system (6) is the following algorithm.

**Algorithm 2.**

Set \(\rho^0\) to the initial condition
\[
\text{for } n = 0,1,2,\ldots \text{ do}
\]
Set \(\alpha_j\) with some procedure or set \(\alpha_j = 0\) for first-order implicit scheme.
Solve the nonlinear system in the unknown \(x\)
\[
x - \Delta t \sum_{j=1}^{k} (1 - \alpha_j) \nabla_{ij} f_j (x) = \rho^n + \Delta t \sum_{j=1}^{k} \alpha_j \nabla_{ij} f_j (\rho^n)
\]
by the Algorithm 1.
\[
\text{end for}\]
8. TWO BETTER SECOND-ORDER POSITIVE SCHEMES

The scheme (28) has the disadvantage of slowing down the accuracy for time steps that are large compared to the reaction rates. An alternative can be an implicit version of Collatz scheme [11] which takes the form

$$\frac{\rho^{n+1} - \rho^n}{\Delta t} = \nabla f \left( \rho^{n+1} - \frac{\Delta t}{2} \nabla f \left( \rho^{n+1} \right) \right).$$  \hspace{1cm} (31)

Equation (31) involves the solution of the nonlinear system in the unknown \( x \)

$$x - \Delta t \nabla f \left( x - \frac{\Delta t}{2} \nabla f(x) \right) = \rho^n. \hspace{1cm} (32)$$

To solve (32), it is convenient to introduce a new variable

$$y = x - \frac{\Delta t}{2} \nabla f(x),$$

so that system (32) is equivalent to the new system

$$x - \Delta t \nabla f(y) = \rho^n, \quad x - \frac{\Delta t}{2} \nabla f(x) = y. \hspace{1cm} (33)$$

This system, by Theorem 2, can be written in the following equivalent form:

$$\lambda y - \Delta t C(y, \nabla) y = \rho^n - x + \lambda y, \quad x - \frac{\Delta t}{2} C(x, \nabla)x = y,$$

which suggests the following iterative procedure to solve (33)

$$y^{l+1} = \left( I - \frac{\Delta t}{\lambda} C(y^l, \nabla) \right)^{-1} \left( \frac{\rho^n}{\lambda} - x^l + y^l \right), \hspace{1cm} (34)$$
$$x^{l+1} = \left( I - \frac{\Delta t}{2} C(x^l, \nabla) \right)^{-1} y^{l+1}.$$

REMARK 1. In this iterative scheme, there is a free parameter \( \lambda \), which can be used to avoid negative production in the components of \( y^{l+1} \). Experimentally, one can see that if \( \lambda \) is sufficiently large, there is no production of negative values and the procedure converges to the solution.

Another scheme can be an implicit version of Heun scheme [12] which takes the form

$$\frac{\rho^{n+1} - \rho^n}{\Delta t} = \frac{1}{2} \nabla \left[ f \left( \rho^{n+1} - \Delta t \nabla f \left( \rho^{n+1} \right) \right) + f \left( \rho^n \right) \right].$$  \hspace{1cm} (35)

Equation (35) involves the solution of the nonlinear system in the unknown \( x \)

$$x - \Delta t \nabla \left[ f \left(x - \Delta t \nabla f(x)\right) + f \left(x\right) \right] = \rho^n. \hspace{1cm} (36)$$

To solve (36), it is convenient to introduce a new variable

$$y = x - \Delta t \nabla f(x),$$

so that system (36) is equivalent to the new system

$$y - \Delta t \nabla f(y) = 2\rho^n - x, \quad x - \Delta t \nabla f(x) = y.$$

This system, by Theorem 2, can be written in the following equivalent form

$$(1 + \lambda) y - \Delta t C(y, \nabla) y = 2\rho^n - x + \lambda y, \quad x - \Delta t C(x, \nabla)x = y,$$

which suggests the following iterative procedure to solve (33)

$$y^{l+1} = \left( (1 + \lambda) I - \Delta t C(y^l, \nabla) \right)^{-1} \left( 2\rho^n - x^l + \lambda y^l \right), \hspace{1cm} (37)$$
$$x^{l+1} = \left( I - \Delta t C(x^l, \nabla) \right)^{-1} y^{l+1}.$$

REMARK. Also, in this iterative scheme, there is a free parameter \( \lambda \), which can be used to avoid negative production in the components of \( y^{l+1} \). Experimentally, one can see that if \( \lambda \approx 1 \), there is no production of negative values and the procedure converges to the solution.
9. A SECOND-ORDER DIAGONALLY IMPLICIT RUNGE-KUTTA SCHEME

The schemes (31) and (35) have the disadvantage of a slow rate of convergence of the procedures (34) and (37). An alternative can be the use of diagonally implicit Runge-Kutta schemes, so that, instead of solving a large nonlinear system, we solve a series of smaller nonlinear systems. In this paper, we use the S-stable two-stage diagonally implicit Runge-Kutta scheme of Alexander [1] that after some manipulation takes the form

\[ x - \alpha \Delta t V f(x) = \rho^n, \]
\[ y - \alpha \Delta t V f(y) = \frac{(2\alpha - 1)\rho^n + (1 - \alpha)x}{\alpha}, \]
\[ \rho^{n+1} = y, \]  

where \( \alpha = 1 \pm \sqrt{2}/2 \), so that the solution procedure is the following algorithm.

**Algorithm 3.**
- Set \( x^0 = \rho^n \)
- repeat
  - \( x^{k+1} = (I - \alpha \Delta t C(x^k, V))^{-1} \rho^n \)
  - until convergence.
- Set \( y^0 = \rho^n \).
- Set \( b = \frac{[(2\alpha - 1)\rho^n + (1 - \alpha)x^{k+1}]}{\alpha} \)
- repeat
  - \( y^{i+1} = (I - \alpha \Delta t C(y^i, V))^{-1} b \)
  - until convergence
- At convergence set \( \rho^{n+1} = y^{i+1} \).

10. NUMERICAL TESTS

The numerical schemes (20), (28), (31), (35), and (38) are applied to the following differential system:

\[ \frac{dp}{dt} = f_1(p) z_1 + f_2(p) z_2 + f_3(p) z_3, \quad p(0) = [0, 1, 2]^T, \]

with

\[ f_1(p) = \frac{5p_2 p_3}{10^{-2} + (p_2 p_3)^2}, \quad f_2(p) = \frac{p_2 p_3}{10^{-18} + p_2 p_3 (10^{-8} + p_2 p_3)}, \]
\[ f_3(p) = 0.1(p_3 - p_2 - 2.5)^2 p_1 p_2, \]
\[ z_1 = [2, -1, -1]^T, \quad z_2 = [-1, 2, -1]^T, \quad z_3 = [-1, -1, 2]^T. \]

The solution is calculated in the time interval \([0, 1]\). Figure 1 shows the results of schemes (20) and (35) with time step \( \Delta t = 0.025 \). If \( p(t) \) is the exact solution and \( p^i \) is the computed approximation at \( t_i \), then we define the error as

\[ \text{Error} = \frac{1}{m} \sum_{i=1}^{m} \| p(t_i) - p^i \|_\infty, \quad t_i = \frac{i}{m}. \]

In practical computations, instead of exact solution, we use the solution shown in Figure 2, computed with Heun (35) scheme with time step \( \Delta t = 10^{-4} \). Tables 1 and 2 summarize the results obtained with the schemes presented in the paper with different time steps and also summarize the computational cost in terms of number of matrix inversions. The "****" in the
Table 1. Comparison table with $\Delta t = 0.025$.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Error</th>
<th>Inversions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euler (20)</td>
<td>0.0888</td>
<td>872</td>
</tr>
<tr>
<td>$\alpha$ (28)</td>
<td>0.0208</td>
<td>1486</td>
</tr>
<tr>
<td>Collatz (31)</td>
<td>0.0088</td>
<td>3868</td>
</tr>
<tr>
<td>Heun (35)</td>
<td>0.0189</td>
<td>3192</td>
</tr>
<tr>
<td>RK 2$^0$ order (38)</td>
<td>****</td>
<td>****</td>
</tr>
</tbody>
</table>
Table 2. Comparison table with $\Delta t = 0.01$.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Error</th>
<th>Inversions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euler (20)</td>
<td>0.0264</td>
<td>1394</td>
</tr>
<tr>
<td>$\alpha$ (28)</td>
<td>0.0037</td>
<td>1535</td>
</tr>
<tr>
<td>Collatz (31)</td>
<td>0.0029</td>
<td>7678</td>
</tr>
<tr>
<td>Heun (35)</td>
<td>0.0010</td>
<td>4090</td>
</tr>
<tr>
<td>RK 2$^\text{nd}$ order (38)</td>
<td>****</td>
<td>****</td>
</tr>
</tbody>
</table>

Figure 2. Numerical solution with scheme (35) and time step $\Delta t = 10^{-4}$.

Figure 3. Instability in the computation with classical explicit fourth order Runge-Kutta method with time step of $\Delta t = 10^{-6}$.

11. CONCLUSIONS

In the solution of systems of ODE arising in mass action kinetics, it is essential to maintain physical properties such as nonnegativity of the solution and atomic mass conservation. Classical explicit schemes maintain these properties with too strong time-step restrictions to be useful. Classical implicit schemes maintain these properties with weaker time-step restrictions, but require the solution of an algebraic system at each time step.
In this paper, we propose and discuss the solution of these algebraic systems without the use of Jacobian matrices, but by the repeated inversion of $M$-matrices that can be easily constructed, thus considerably simplifying and accelerating the computer implementation of the schemes.

In the case of first-order schemes, the existence of the solution of the algebraic system and the nonnegativity of each iteration is proved. Numerical evidence shows that the existence of the solution and the nonnegativity of each iteration is maintained for Heun implicit schemes with fairly large time steps.

REFERENCES