ESAIM: M2AN 56 (2022) 1843–1870
https://doi.org/10.1051/m2an/2022042

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

POSITIVITY-PRESERVING METHODS FOR ORDINARY DIFFERENTIAL EQUATIONS

Sergio Blanes[1,*], Arieh Iserles[2] and Shev Macnamara[3]

Abstract. Many important applications are modelled by differential equations with positive solutions. However, it remains an outstanding open problem to develop numerical methods that are both (i) of a high order of accuracy and (ii) capable of preserving positivity. It is known that the two main families of numerical methods, Runge–Kutta methods and multistep methods, face an order barrier. If they preserve positivity, then they are constrained to low accuracy: they cannot be better than first order. We propose novel methods that overcome this barrier: second order methods that preserve positivity unconditionally and a third order method that preserves positivity under very mild conditions. Our methods apply to a large class of differential equations that have a special graph Laplacian structure, which we elucidate. The equations need be neither linear nor autonomous and the graph Laplacian need not be symmetric. This algebraic structure arises naturally in many important applications where positivity is required. We showcase our new methods on applications where standard high order methods fail to preserve positivity, including infectious diseases, Markov processes, master equations and chemical reactions.

Mathematics Subject Classification. 65L05, 65P99, 65L04.

Received October 14, 2021. Accepted April 22, 2022.

1. Introduction

Numerical integration of mathematical models is an essential step in the implementation and analysis of population models: chemical reactions (see *e.g.* [19, 49] or [24]), biochemical systems [12], and the evolution of epidemics [33] (see also [21] and references therein). Such models are usually formulated as a system of Ordinary Differential Equations (ODEs)

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(0) = \mathbf{y}_0 \in \mathbb{R}^d, \tag{1.1}$$

where $\mathbf{f}(t, \mathbf{y})$ is, in the context of this paper, consistent with two requirements of the application being modeled. First, if $y_i^0 \geq 0,\ i = 1, \ldots, d$ then we have *positivity preservation*:

$$y_i(t) \geq 0 \quad \forall t, \quad i = 1, \ldots, d.$$

Keywords and phrases. Positivity-preserving methods, graph Laplacian matrices, exponential integrators, Magnus integrators.

[1] Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, E-46022 Valencia, Spain.

[2] Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB4 1LE, UK.

[3] Australian Research Council Centre of Excellence, for Mathematical and Statistical Frontiers (ACEMS), School of Mathematical and Physical Sciences, University of Technology Sydney, NSW 2007, Australia.

*Corresponding author: serblaza@imm.upv.es

Second, there exist $\mathbf{w}_\ell = (w_{\ell,1}, \ldots, w_{\ell,d})^\top$, $\ell = 1, 2, \ldots, k$ such that $\mathbf{w}_\ell^\top \mathbf{f}(t, \mathbf{y}) = 0$ so, the solution satisfies the conditions (with $\mathbf{y} = (y_1, \ldots, y_d)^\top$, $\mathbf{y}_0 = (y_1^0, \ldots, y_d^0)^\top$)

$$\sum_{i=1}^{d} w_{\ell,i} y_i = \sum_{i=1}^{d} w_{\ell,i} y_i^0 = c_\ell, \qquad \ell = 1, 2, \ldots, k$$

with $w_{\ell,i} \geq 0$ and $c_\ell > 0$. The most important special case is $k = 1$ and $\mathbf{w}_1 = (1, \ldots, 1)^\top = \mathbf{1}$, which is referred to as *mass preservation*, and in this case we may assume without loss of generality that $c_1 = 1$.

Although the focus of this article is mainly on positivity and mass preservation ODEs, positivity preservation is a much wider challenge. For example, Lotka–Volterra models [3,16] preserve positivity but not mass as well as some parabolic problems [26]. The stochastic differential equation associated with the Nobel prize winning Black–Scholes model in finance has positive solutions, but standard numerical solvers, such as the Euler–Maruyama method, fail to preserve positivity. The Kolmogorov Lecture at the Ninth World Congress in Probability and Statistics concerned methods for preserving positivity in the setting of the stochastic Langevin equations [36].

We note in passing that even with these two requirements, equation (1.1) may display rich dynamical behaviour: some systems of this kind converge to a unique steady state, others have a number of steady states, yet others exhibit oscillatory behaviour.

The methods proposed in this work are constructed to preserve positivity, while keeping linear invariants preservation to high accuracy (symplectic integrators preserve the symplectic structure of Hamiltonian systems while not exactly preserving energy, but this gives good properties in regards to error propagation over long time intervals). However, in the case where only mass preservation is required there are well known mathematical results that allow us to adapt the methods to preserve exactly, and for this reason this case is now treated in more detail.

1.1. Graph Laplacians and ODEs

A useful way to envisage mass and positivity preservation is that for every $t \geq 0$ the state variable $\mathbf{y}(t)$ is a discrete probability distribution of d species. This corresponds to the case $k = 1$, $\mathbf{w}_1 = \mathbf{1}$ and, as we will show in Proposition 1.1, these properties can be preserved if the vector field in (1.1) can be written in the form (see also *e.g.* [5,15,20])

$$\mathbf{f}(t, \mathbf{y}) = A(t, \mathbf{y})\mathbf{y}$$

where the matrix $A : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is a graph Laplacian.

Definition. An $n \times n$ real matrix A is a graph Laplacian if it has the following properties:

Property 1 (Pattern of signs)**.** $A_{k,\ell} \geq 0$ for $k, \ell = 1, \ldots, n$, $k \neq \ell$, $A_{k,k} \leq 0$ for $k = 1, \ldots, n$ and
Property 2 (Zero column sum)**.** $\sum_{k=1}^{n} A_{k,\ell} = 0$ for $\ell = 1, \ldots, n$.

We denote the set of all $n \times n$ graph Laplacians by $\mathcal{L}_n$. The same term "graph Laplacian" is used with different meanings in the literature – in our work, we allow it to be non-symmetric.

For simplicity, we consider the autonomous case. (The general nonautonomous case can be considered similarly, as we will show latter.) We focus on the solution of the nonlinear ODE

$$\mathbf{y}' = A(\mathbf{y})\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{y}_0 \in \mathbb{R}^d, \tag{1.2}$$

where we assume throughout that $A(\mathbf{y})$ has the same pattern of signs as a graph Laplacian, *i.e.* we assume Property 1 of the definition above. (In some examples, such as the MAPK cascade example, we do not assume Property 2, *i.e.* we do not always assume $\mathbf{1}^\top A(\mathbf{y}) = \mathbf{0}^\top$, and we demonstrate that our methods can nevertheless work well.) We typically also assume that all components of the initial condition are nonnegative. Many applications fit this framework: Markov processes in continuous time on discrete states; master equations [39];

single molecule chemistry [51] (Fig. 4); studies of robustness of Turing pattern formation in stochastic settings [27, 43]; and lasers and quantum dots [53].

Given two compatible matrices P and Q we say that $P \succ Q$ if $P_{i,j} > Q_{i,j}$ for all i, j and $P \succeq Q$ if $P_{i,j} \geq Q_{i,j}$. We assume that $\mathbf{y}_0 \succeq \mathbf{0}$ and $\mathbf{1}^\top \mathbf{y}_0 = 1$. Then the solutions of (1.2) have the following desirable features.

Proposition 1.1. *Solutions of (1.2) with $\mathbf{y}_0 \succeq \mathbf{0}, \mathbf{1}^T\mathbf{y}_0 = 1$ have the following two properties:*

Positivity. $\mathbf{y}(t) \succeq \mathbf{0}$ *for all* $t \geq 0$, *and*
Conservation of mass. $\mathbf{1}^\top \mathbf{y}(t) = 1$, *for all* $t \geq 0$.

Proof. The statement about mass conservations is trivial, because

$$\mathbf{1}^\top \mathbf{y}'(t) = \mathbf{1}^\top A(\mathbf{y}(t))\mathbf{y}(t) = \mathbf{0}^\top \mathbf{y}(t) = 0$$

implies that $\mathbf{1}^\top \mathbf{y}(t) \equiv \text{const} = \mathbf{1}^\top \mathbf{y}_0 = 1$.

To prove the statement about positivity, we consider any $t^* \geq 0$ such that there exists $k^* \in \{1, 2, \ldots, d\}$ with $y_{k^*}(t^*) = 0$ and such that $\mathbf{y}(t^*) \succeq \mathbf{0}$ – clearly, unless such t^* exists, $\mathbf{y}(t)$ stays forever in the nonnegative cone. Note that it is perfectly possible for t^* to be zero, also it is possible that several components of $\mathbf{y}(t)$ vanish at $t = t^*$, this makes no difference to our argument. We note that, by (1.2),

$$y'_{k^*}(t^*) = \sum_{\ell=1}^{d} A_{k^*,\ell}(\mathbf{y}(t^*))y_\ell(t^*) \geq 0,$$

because A is a graph Laplacian, so off-diagonal entries are nonnegative. Therefore y_{k^*} cannot change sign at t^*, and it must stay in the nonnegative cone. □

Remark. Note in the proof of Proposition 1.1 that Property 1 alone of the definition of the Laplacian (pattern of signs) suffices to give positivity, and that, separately, Property 2 alone of the definition of the Laplacian suffices to give mass preservation. In particular, if the matrix $A(\mathbf{y})$ has the same pattern of $\pm$ signs as a Laplacian (but we make no assumption on the column sums of $A(\mathbf{y})$), then it is still true that solutions of $\mathbf{y}' = A(\mathbf{y})\mathbf{y}$, preserve positivity.

Let us now consider some properties of graph Laplacian matrices that allow us to deduce additional qualitative properties of the solution of (1.2).

Theorem 1.2. *Let $A \in \mathcal{L}_n$. Then it has an eigenvalue at the origin, which is simple if A is irreducible, and all its other eigenvalues reside in $\mathbb{C}^- = \{z \in \mathbb{C} : \operatorname{Re} z < 0\}$.*

Proof. Since $\mathbf{1}^\top A = \mathbf{0}^\top$, it follows that $0 \in \sigma(A)$. To locate the remaining eigenvalues we use the Gerschgorin theorem, applying it to columns (typically it is applied to rows, but this makes no difference). Thus, letting

$$\mathbb{S}_\ell = \left\{ z \in C : |z - A_{\ell,\ell}| \leq \sum_{k \neq \ell} |A_{k,\ell}| \right\}, \qquad \ell = 1, \ldots, n,$$

we have $\sigma(A) \subset \bigcup_{\ell=1}^n \mathbb{S}_\ell$. By the definition of graph Laplacian, all Gerschgorin discs live in $\mathbf{cl}\,\mathbb{C}^-$ and adjoin $\mathrm{i}\mathbb{R}$ only at the origin. Therefore $\sigma(A) \setminus \{0\} \in \mathbb{C}^-$.

It remains to prove that 0 is a simple eigenvalue. Let $\alpha = \min_{k=1,\ldots,n} A_{k,k}$, then the entries of $B = A - \alpha I \neq O$ are all nonnegative. Therefore, according to Frobenius–Perron theory [4], irreducibility implies that the largest in modulus eigenvalue of B is positive and simple. Since this is $-\alpha$, it follows that 0 is a simple eigenvalue of A. □

Incidentally, one of the less well-known formulations of the Gerschgorin theorem states that if A is irreducible then an eigenvalue might be on the boundary of one Gerschgorin disc only if it is on the boundary of all Gerschgorin discs – this is certainly the case with 0.

Proposition 1.3. *Assume the matrix $A \in \mathcal{L}_n$ is symmetric. Then* $\mathrm{d}\|\mathbf{y}(t)\|^2/\mathrm{d}t \leq 0$.

Proof. We compute

$$\frac{1}{2}\frac{\mathrm{d}\|\mathbf{y}(t)\|^2}{\mathrm{d}t} = \mathbf{y}^\top(t)\mathbf{y}'(t) = \mathbf{y}^\top(t)A(\mathbf{y}(t))\mathbf{y}(t) \leq \alpha_+(A(\mathbf{y}(t)))\|\mathbf{y}(t)\|^2,$$

where $\alpha_+(B)$ is the *spectral abscissa* – the eigenvalue of the matrix B with the largest real part (which in the case of A is real because of the Perron–Frobenius theory). This is true because $\alpha_+(B) \geq \mathbf{v}^\top B\mathbf{v}/\|\mathbf{v}\|^2$ for any square matrix B and a nonzero vector $\mathbf{v}$. Since our $A(\mathbf{y})$ is graph Laplacian, it follows at once from the Gerschgorin theorem that $\alpha_+(A(\mathbf{y}(t))) \leq 0$ and, since $0 \in \sigma(A(\mathbf{y}(t)))$, we deduce that $\mathrm{d}\|\mathbf{y}(t)\|^2/\mathrm{d}t \leq 0$. □

Let $\hat{\mathbf{y}}$ be the eigenvector corresponding to the simple eigenvalue 0. In the symmetric case, it is clear that $\|\mathbf{y}(t) - \hat{\mathbf{y}}\|^2$ is a monotonically decreasing function – using the fact that $A(\mathbf{y}(t))\hat{\mathbf{y}} = \mathbf{0}$,

$$[\mathbf{y}(t) - \hat{\mathbf{y}}]' = \mathbf{y}'(t) = A(\mathbf{y}(t))\mathbf{y}(t) = A(\mathbf{y}(t))[\mathbf{y}(t) - \hat{\mathbf{y}}]$$

and we continue as before.

In the nonsymmetric case, the issue of stability needs more discussion. The two defining properties of the graph Laplacian together ensure that the columns of the matrix exponential are probability vectors, so that, when A is a constant matrix, in the 1-norm we always have $\|\exp(tA)\| = 1$, $t \geq 0$. In the case of a constant matrix, these matrices are sometimes known as "W-matrices" in the statistical physics literature and, by studying the adjoint $\mathbf{z}'(t) = A^\top\mathbf{z}(t)$ – with arguments similar to those of our Proposition 1.1 – it is known that the minimum of the solution $\mathbf{z}$ is increasing, and that the maximum is decreasing. In the 2-norm, a sufficient condition for strong stability of $\mathbf{y}' = A\mathbf{y}(t)$ with solution $\mathbf{y}(t) = \exp(tA)\mathbf{y}(0)$, is that $(A + A^\top)$ be negative definite. Note that this condition is more restrictive than merely the assumption that the eigenvalues of A have negative real part (because then it would still be possible that $(A + A^\top)$ had a positive eigenvalue). This issue of stability is related to "*the hump*" in the classical literature on the numerical analysis of the matrix exponential, and to the *lognorm*, and also to the subject of *pseudospectra.* Nonsymmetric graph Laplacians exhibit significant pseudospectra, manifesting themselves in various ways, such as a more subtle stability analysis, and the failure of standard eigenvalue algorithms [30, 37, 41]. A sufficient condition for stability of operator splitting methods is that each part separately be strongly stable, although this may be too pessimistic in practice. For operator splitting methods, the graph Laplacian can sometimes be expressed as the sum of two matrices, each of which is separately a graph Laplacian with a physical interpretation [38]. In general, operator splitting does not preserve the steady state [52] – so it is worth pointing out that the novel splitting methods that we introduce in this work, for example later in (3.2), in our numerical experiments, do have the desirable property that they preserve the steady state. In the nonautonomous case, but still linear case, it can be shown under suitable assumptions that the difference of any two solutions is decreasing in the 1-norm, but the issue of stability is *much* more delicate. For instance, see the catalogue of counterexamples, and Theorem 3.1 described in [17].

To sum up, the solution of a mathematical model given by (1.2) with $\mathbf{y}_0 \succeq \mathbf{0}$ and where $A(\mathbf{y})$ is a graph Laplacian matrix (assuming $\mathbf{y} \succeq \mathbf{0}$) always preserves mass and always preserves positivity. Often, the model (1.2) is also stable and converges to a steady state. These features correspond to the phenomenological desiderata in for example epidemiological models.

In theory, there are always exact formulae for the right eigenvector corresponding to the zero eigenvalue of a nonsymmetric graph Laplacian matrix A, *via* the Matrix-Tree Theorem [22]. This is the steady state of the corresponding linear Laplacian dynamical system, and in special cases, there are also formulae for the dynamical solutions [17, 18, 30].

Unfortunately, in general, the exact solution of these dynamical systems is unknown, so we need to resort to numerical algorithms. Using backward error analysis, we can envisage a numerical method as the exact solution of a perturbed model. While this is typically adequate across a single step, unless the method is chosen carefully,

a numerical solution is highly unlikely to respect the important special structure of (1.2) across the entire time interval of interest.

The mathematical models we are considering in this paper are based on differential equations whose solutions preserve some underlying geometric structure. The design and analysis of numerical integrators that preserve the qualitative features of the underlying differential equations is the subject of *Geometric Numerical Integration* [7, 25, 29, 50]. We are not only concerned with the accuracy and stability of numerical schemes but also with their geometric properties, which reflect important features of the phenomena being modelled. This endows the integrators with an improved qualitative behaviour, but also typically leads to significantly more accurate results.

For example, in [21] the authors consider a mathematical model for the COVID-19 epidemic in Italy, while paying much attention so that the proposed model has the structure of (1.2), but then numerically solve it using the first order explicit Euler method

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n)$$

where h is the time step and $\mathbf{y}_n \simeq \mathbf{y}(t_n)$ with $t_n = t_0 + nh$. We easily see that

$$\mathbf{1}^\top \mathbf{y}_{n+1} = \mathbf{1}^\top \mathbf{y}_n + h\mathbf{1}^\top \mathbf{f}(\mathbf{y}_n) = \mathbf{1}^\top \mathbf{y}_n = \ldots = \mathbf{1}^\top \mathbf{y}_0$$

and then the mass is preserved (this is also the case for most standard methods like Runge–Kutta or multistep methods). However, it is well known that, in general, this method does not preserve positivity unconditionally.

This inadequate behaviour cannot be rectified by a standard higher-order method: in [10] it is shown that within the class of linear multistep and Runge–Kutta methods unconditional positivity restricts the order of the method to just one[1].

For non-stiff problems and for relatively short time integration, an Euler method, or any other standard method, can provide sufficiently accurate, satisfactory results. However, if a mathematical model is stiff (this is typical to equations of chemical kinetics) or need be solved for long time intervals, standard methods may produce negative solutions or become unstable. While the stiffness in chemical kinetics equations can be dealt with using implicit methods and mass is preserved by most numerical methods, positivity remains an outstanding challenge.

The most efficient solvers considered in [24] for low to medium accuracy in the numerical solution of stiff kinetic equations are Rosenbrock methods. In addition, they are among the simplest implicit schemes to be implemented in a code, yet they fail to preserve positivity. Note that there exist exponential Rosenbrock-type methods [28] that involve the computation of the exponential of the Jacobian. However, in general, this Jacobian is not a graph Laplacian and positivity cannot be guaranteed.

The objective of preserving mass and positivity in numerical integration, in particular within the context of chemical kinetics, received a measure of attention, although perhaps less than it deserves given its importance in applications. An obvious device to avoid the solution from becoming negative is *clipping:* the practice of converting a negative component to zero. This, of course, interferes with the preservation of mass but the latter can be recovered using laborious optimization procedure in every time step [49]. The effects of this costly algorithm on long-term accuracy and stability are unknown.

Another approach toward preservation of mass and positivity are Runge–Kutta–Patankar methods [5, 14, 34, 47]. The idea is to adapt Runge–Kutta-like methods for *production–destruction systems* in chemical kinetics. We will show that this class of methods can be seen as particular approximation to the methods proposed in the present work.

2. Illustrative examples

To illustrate our analysis we consider several simple population models from the literature.

[1]This is a necessary condition which, alas, is not sufficient: the above explicit Euler method is of order one but does not preserve positivity.

Example 2.1 (The Robertson reaction)**.** Let us consider the following example of chemical reactions, $A \longrightarrow B$ and $B + B \longrightarrow B + C \longrightarrow A + C$, leading to the stiff differential equations for concentrations $\mathbf{y} = (y_1, y_2, y_3)$ of A, B, C [24] (p. 157):

$$\begin{aligned} y_1' &= -0.04y_1 + 10^4 y_2 y_3, & y_1(0) &= 1 \\ y_2' &= 0.04y_1 - 10^4 y_2 y_3 - 3 \cdot 10^7 y_2^2, & y_2(0) &= 0 \\ y_3' &= 3 \cdot 10^7 y_2^2, & y_3(0) &= 0, \end{aligned} \tag{2.1}$$

that, rewritten in a vector form, read

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -0.04 & 10^4 y_3 & 0 \\ 0.04 & -3 \cdot 10^7 y_2 - 10^4 y_3 & 0 \\ 0 & 3 \cdot 10^7 y_2 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \tag{2.2}$$

where the matrix is graph Laplacian. This example fits into the framework of Theorem 4.2, which comes later.

Note that the system can also be written in many different ways, for example

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -0.04 & 0 & 10^4 y_2 \\ 0.04 & -3 \cdot 10^7 y_2 & -10^4 y_2 \\ 0 & 3 \cdot 10^7 y_2 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \tag{2.3}$$

where now the matrix is no longer a graph Laplacian. As we will see, it is crucial to write properly the equations for the numerical solutions to preserve their qualitative properties.

Example 2.2 (The SIR model)**.** The *Susceptible–Infected–Recovered (SIR) model* describes the temporal epidemic evolution in terms of three variables for the population: $S(t)$: (Susceptible), $I(t)$ (Infected) and $R(t)$ (Recovered). It is usually asssumed that the total population does not change during the infection period. S, I and R denote the fractions with respect to the total population: $S(t)+I(t)+R(t) \equiv 1$. This model was proposed in [33]

$$\begin{aligned} S' &= -R_0 SI\ , \\ I' &= R_0 SI - I\ , \\ R' &= I\ , \end{aligned} \tag{2.4}$$

where $R_0 > 0$ is the basic reproduction number, and the system can be written in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} S \\ I \\ R \end{bmatrix} = \begin{bmatrix} -R_0 I & 0 & 0 \\ R_0 I & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} S \\ I \\ R \end{bmatrix} \tag{2.5}$$

which is like (1.2) with $\mathbf{y} = [S, I, R]^\top$ and the matrix is evidently a graph Laplacian.

Example 2.3 (Laplacian dynamics on graphs (autonomous and linear))**.** Graph Laplacian dynamics, $\mathbf{y}' = \mathcal{L}(G)\mathbf{y}$, where $\mathcal{L}(G)$ is a constant matrix, representing the Laplacian of a directed graph G, gives rise to a large class of applications in biochemical kinetics, including Michaelis–Menten enzyme kinetics, allosteric enzymes, G-protein coupled receptors, ion channels, and gene regulation [22] (Eq. (3)). Discussion of conditions under which such linear systems always converge to a steady state, and discussion of the sense in which that might be considered unique is given in [45]. That linear setting $\mathbf{y}' = \mathcal{L}(G)\mathbf{y}$ is a special case of the more general framework here where we focus on the exact nonlinear model in (1.2).

Example 2.4 (Cardiac ion channels (nonautonomous and linear))**.** Nonautonomous Laplacian systems, $\mathbf{y}' = A(t)\mathbf{y}$, have many important applications, including cardiac ion channel kinetics [17,18]. In special cases, there are also exact solutions for the dynamical solutions, such as the explicit Magnus formulæ in [30], and closely related invariant manifolds of binomial-like solutions.

Example 2.5 (MAPK cascade (autonomous and nonlinear))**.** The mitogen-activated protein kinase (MAPK) cascade is fundamental in cell signalling biology and cancer biology, and it is modelled by eighteen differential equations with rates given by the Law of Mass Action, together with some linear conservation laws [48]. By our Theorem 4.2, in the sequel, this MAPK model fits our framework of (1.2), subject to the remarks we make following Proposition 1.1. The Laplacian dynamics mentioned in the above constant coefficient and linear examples, where convergence to a steady state is common [45], makes it tempting to conjecture that the model we focus on here in (1.2), likewise always converges to a steady state. However, a counterexample is provided by the MAPK cascade, which can be modelled by our nonlinear Laplacian dynamics (1.2), and which has been shown by numerical simulations to exhibit both bistability and oscillations [48].

We have taken the model of [23] (Tab. 3, Fig. 3, Eqs. (12)–(17)), which is closely related to the MAPK cascade, and rewritten it here in the form of our model (1.2), to show that it is clearly an example of the Laplacian dynamics that we study in this paper:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} -k_7 - k_1 y_2 & 0 & 0 & k_2 & 0 & k_6 \\ 0 & -k_1 y_1 & k_5 & 0 & 0 & 0 \\ 0 & 0 & -k_3 y_1 - k_5 & k_2 & k_4 & 0 \\ (1-\alpha)k_1 y_2 & \alpha k_1 y_1 & 0 & -k_2 & 0 & 0 \\ 0 & 0 & k_3 y_1 & 0 & -k_4 & 0 \\ k_7 & 0 & 0 & 0 & 0 & -k_6 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix}. \tag{2.6}$$

We take the same rate constants $k_1 = \frac{100}{3}$, $k_2 = \frac{1}{3}$, $k_3 = 50$, $k_4 = \frac{1}{2}$, $k_5 = \frac{10}{3}$, $k_6 = \frac{1}{10}$, $k_7 = \frac{7}{10}$, and initial state $y(0) = [0.1, 0.175, 0.15, 1.15, 0.81, 0.5]^\top$. Note that we have $\mathbf{y}' = A(\alpha, \mathbf{y})\mathbf{y}$, where $\alpha \in [0,1]$ is a parameter we can freely choose in this interval and the matrix $A(\alpha, \mathbf{y})$ has the same pattern of signs as a Laplacian, but that a column of $A(\alpha, \mathbf{y})$ does not always sum to zero, so this fits our framework of (1.2), subject to the remarks we make following Proposition 1.1, and this is also an example of our later Theorem 4.2. This model possesses two conservation of mass laws, namely both $y_1 + y_4 + y_6$ and $y_2 + y_3 + y_4 + y_5$ are constants, which have physical interpretation in terms of the total enzyme of two types of kinases. Those two conservation laws correspond to $\mathbf{w}_1 = [1,0,0,1,0,1]^\top$, and $\mathbf{w}_2 = [0,1,1,1,1,0]^\top$, respectively. Note that we have

$$\mathbf{w}_1^\top A(\alpha, \mathbf{y})\mathbf{y} = 0, \qquad \mathbf{w}_2^\top A(\alpha, \mathbf{y})\mathbf{y} = 0,$$

and this is irrespective of the value of $\alpha \in (0,1)$. However, if we take $\alpha = 0$ we have that

$$\mathbf{w}_1^\top A(0, \mathbf{y}) = 0, \qquad \mathbf{w}_2^\top A(0, \mathbf{y}) \neq 0,$$

while for $\alpha = 1$

$$\mathbf{w}_1^\top A(1, \mathbf{y}) \neq 0, \qquad \mathbf{w}_2^\top A(1, \mathbf{y}) = 0.$$

It should be possible to use methods based on matrix exponentials (such as the methods proposed in this paper) to respect *e.g.* the second conservation law, if we take $\alpha = 1$ because $\mathbf{w}_2^\top A(\mathbf{y}) = 0$, so $\mathbf{w}_2^\top \exp(tA(\mathbf{y})) = \mathbf{w}_2^\top$. However, because $\mathbf{w}_1^\top A(\mathbf{y}) \neq 0$, it will be difficult (and probably impossible) to maintain exactly the first conservation law by methods that compute matrix exponentials[2]. This is typical of applications in chemical

[2]The situation whereby it is impossible to satisfy several conservation laws under discretisation – except, of course, by the exact solution – is familiar in Geometric Numerical Integration [25].

kinetics, and for example, the famous Michaelis–Menten enzyme kinetics model (which always converges to a unique and simple steady state) also fits the framework, with a matrix that has the same pattern of signs as a Laplacian, but that does not have zero column sum, and the model still has two simple well-known linear conservation laws. Significantly, by numerical simulation, it has been shown that solutions of this model (2.6) show oscillations [23] (Fig. 5).

3. Positivity preserving second-order methods

Let us first consider the particular case in which the matrix A is constant. Then the exact solution is given *via* the exponential:

$$\mathbf{y}(t) = \mathrm{e}^{tA}\mathbf{y}_0.$$

If A is a graph Laplacian matrix it is a consequence of Theorem 1.2 that $\sigma(\mathrm{e}^{tA}) \subset \{z \in \mathbb{C} : |z| \leq 1\}$, hence the solution is stable (subject to the discussion of stability we gave earlier, in the nonsymmetric case).

The exponential of a graph Laplacian matrix is fundamental to the work of this paper, and this calls for a more detailed study of its qualitative properties.

3.1. The exponential of a graph Laplacian matrix

We begin with column sums for an arbitrary square matrix.

Proposition 3.1. *Suppose that* $\mathbf{1}^\top A = \mathbf{0}^\top$. *Then* $\mathbf{1}^\top \mathrm{e}^A = \mathbf{1}^\top$.

Proof. By the series definition of the exponential

$$\mathbf{1}^\top \mathrm{e}^A = \mathbf{1}^\top \left(\sum_{n=0}^{\infty} \frac{A^n}{n!} \right) = (\mathbf{1}^\top I) + \sum_{n=1}^{\infty} (\mathbf{1}^\top A) \frac{A^{n-1}}{n!} = \mathbf{1}^\top + \sum_{n=1}^{\infty} (\mathbf{0}^\top) \frac{A^{n-1}}{n!} = \mathbf{1}^\top.$$

□

Remark. Replace A by tA in the proposition to see that, as a corollary, if $\mathbf{1}^\top A = \mathbf{0}^\top$, then $\mathbf{1}^\top \mathrm{e}^{tA} = \mathbf{1}^\top$.

Remark. Graph Laplacians have the property $\mathbf{1}^\top A = \mathbf{0}^\top$ by definition, so for graph Laplacians it is also true that $\mathbf{1}^\top \mathrm{e}^{tA} = \mathbf{1}^\top$.

We need the following elements of the *Perron–Frobenius theory* [4] (pp. 26 and 27). Let $B \in \mathbb{R}^{d\times d}$, $B \succeq O$. Then $\rho(B)$ is an eigenvalue of B and we can choose the corresponding eigenvector $\mathbf{v}$ such that $\mathbf{v} \succeq \mathbf{0}$. Moreover, if in addition B is *irreducible* then $\rho(B)$ is a simple eigenvalue and $\mathbf{v}$ is the only eigenvector of B with nonnegative entries.

Let $a^* = \min_{i=1,\ldots,d} A_{i,i} < 0$ and set $\tilde{A} = A - a^* I$. Then

$$\mathrm{e}^{tA} = \mathrm{e}^{ta^* I + t\tilde{A}} = \mathrm{e}^{ta^*} \mathrm{e}^{t\tilde{A}}.$$

Since $\tilde{A} \succeq O$, all its nonnegative powers are also nonnegative and we deduce that $\mathrm{e}^{t\tilde{A}} \succeq O$. Therefore $\mathrm{e}^{tA} \succeq O$. Indeed, the Mittag–Leffler matrix function of a graph Laplacian, $E_\alpha(At^\alpha)$, is likewise a stochastic matrix, *i.e.* $E_\alpha(At^\alpha) \succeq O$, and all entries are positive, and columns sum to unity [40]. Here the Mittag–Leffler function $E_\alpha(z) = \sum_{k=0}^{\infty} z^k/\Gamma(\alpha k + 1)$ is a one-parameter generalisation of the exponential, and the exponential is recovered as the special case once $\alpha \to 1$. Furthermore, when A is Laplacian then the pattern of signs in the resolvent, and the properties of M-matrices, show that for large n, all entries of the matrix $\left(I - \frac{t}{n}A\right)^{-n}$ are nonnegative. This suggests the results we derive here may be extended to more general settings.

Additionally, once A is irreducible and we denote by $\mathbf{w} \succeq \mathbf{0}$ the left eigenvector of A corresponding to the zero eigenvalue, then $\mathbf{w}^\top \tilde{A} = -a^* \mathbf{w}^\top$. Since $a^* < 0$, we deduce that $|a^*| = \rho(\tilde{A})$ and $\mathbf{w} = \mathbf{1}$.

Proposition 3.2.

$$\sigma(\tilde{A}) \subset \{z \in \mathbb{C} : |z| \leq |a^*|\}.$$

Proof. By the Gerschgorin theorem applied to the columns of $\tilde{A}$ (or the standard Gerschgorin theorem applied to $\tilde{A}^\top$) and because $A_{j,i} \geq 0$ for $i \neq j$, we have

$$\sigma(\tilde{A}) \subset \bigcup_{i=1}^{d} \left\{ z \in \mathbb{C} : |z - A_{i,i} + a^*| \leq \sum_{\substack{j=1 \\ j \neq i}}^{d} A_{j,i} \right\}$$

and the proof follows. □

Proposition 3.3. *Let* $\mathbb{R}^d \ni \mathbf{p} \succeq \mathbf{0}$ *be such that* $\mathbf{1}^\top \mathbf{p} = 1$ *and set* $\mathbf{q} = \mathrm{e}^{tA}\mathbf{p}$. *Then for every* $t \geq 0$ $\mathbf{q} \succeq \mathbf{0}$ *and* $\mathbf{1}^\top \mathbf{q} = 1$.

Proof. We deduce at once that $\mathbf{q} \succeq \mathbf{0}$ because $\mathrm{e}^{tA} \succ O$. Moreover, $\mathbf{1}^\top \mathbf{q} = \mathbf{1}^\top \mathrm{e}^{tA}\mathbf{p} = \mathbf{1}^\top \mathbf{p} = 1$, concluding the proof. □

Remark. Note that all previously stated results apply to maps of the form $\mathbf{z} = \mathrm{e}^{\sigma S}\mathbf{x}$ where $\mathbf{x} \succeq \mathbf{0}$, S is a graph Laplacian and σ is a non-negative constant. Since S can have large negative eigenvalues, taking negative values of σ is likely to lead to a poorly conditioned problem where negative solutions can occur and this compels us to avoid this choice. In the sequel we propose several methods that involve maps of the form $\mathrm{e}^{\sigma S}$ with S being graph Laplacian, and we will see that condition $\sigma > 0$ limits the order of the methods to two in the time step, an order barrier.

Since $\mathrm{e}^{tA} = [I - tA]^{-1} + \mathcal{O}(t^2)$, the following well known result will be useful in the sequel.

Proposition 3.4. *If* A *is graph-Laplacian then* $[I - tA]^{-1} \succeq O$.

Proof. Given $B = I - tA$ we have that $B_{ii} > 0, \forall i$ and $B_{i,j} \leq 0,\ i \neq j$. Since $\sigma(A) \setminus \{0\} \in \mathbb{C}^-$ then we have $\sigma(I - tA) \in \mathbb{C}^+$, which is an M-matrix whose inverse has only non-negative elements. □

3.2. Splitting methods

Splitting methods are frequently used to solve differential equations that are separable into solvable parts. However, for stiff as well as for non-separable problems it is more convenient to proceed as follows [6]. Let us consider the following system in the extended space

$$\begin{aligned} \mathbf{x}' &= A(\mathbf{z})\mathbf{x}, \qquad & \mathbf{x}(0) &= \mathbf{x}_0 = \mathbf{y}_0, \\ \mathbf{z}' &= A(\mathbf{x})\mathbf{z}, \qquad & \mathbf{z}(0) &= \mathbf{z}_0 = \mathbf{y}_0, \end{aligned}$$

where $\mathbf{x}(t) = \mathbf{y}(t) = \mathbf{z}(t)$. The system is separable into two solvable parts

$$\mathcal{A}: \quad \begin{cases} \mathbf{x}' = A(\mathbf{z})\mathbf{x}, \\ \mathbf{z}' = 0 \end{cases} \qquad \Rightarrow \qquad \begin{cases} \mathbf{x}(t) = \mathrm{e}^{tA(\mathbf{z}_0)}\mathbf{x}_0, \\ \mathbf{z}(t) = \mathbf{z}_0 \end{cases}$$

and

$$\mathcal{B}: \quad \begin{cases} \mathbf{x}' = 0, \\ \mathbf{z}' = A(\mathbf{x})\mathbf{z} \end{cases} \qquad \Rightarrow \qquad \begin{cases} \mathbf{x}(t) = \mathbf{x}_0, \\ \mathbf{z}(t) = \mathrm{e}^{tA(\mathbf{x}_0)}\mathbf{z}_0. \end{cases}$$

We solve the system with the symmetric second order Strang splitting method, *i.e.* advance half a step with $\mathcal{A}$ followed by a step with $\mathcal{B}$ and conclude with another half a step with $\mathcal{A}$:

$$\begin{aligned} \mathbf{x}_{1/2} &= \mathrm{e}^{\frac{1}{2}hA(\mathbf{z}_0)}\mathbf{x}_0, \\ \mathbf{z}_1 &= \mathrm{e}^{hA(\mathbf{x}_{1/2})}\mathbf{z}_0, \\ \mathbf{x}_1 &= \mathrm{e}^{\frac{1}{2}hA(\mathbf{z}_1)}\mathbf{x}_{1/2}. \end{aligned} \tag{3.1}$$

Since $\mathbf{z}_0 \succeq 0$, the frozen matrix $A(\mathbf{z}_0)$ is a graph Laplacian, therefore $\mathbf{x}_{1/2} \succeq 0$ and preserves the 1-norm, and similarly for $\mathbf{z}_1$ and $\mathbf{x}_1$.

In addition, $\mathbf{x}_1$ and $\mathbf{z}_1$ correspond to symmetric second order approximations: $\mathbf{z}_1$ can be seen as the exponential midpoint and $\mathbf{x}_1$ as the exponential trapezoidal rule. Then, we can advance the solution either with $\mathbf{z}_1$ or with $\mathbf{x}_1$ but, in general, more accurate results are obtained with the smoothing technique, *i.e.* taking the solution for the next step as the average

$$\mathbf{y}_1 = \frac{1}{2}(\mathbf{x}_1 + \mathbf{z}_1) \tag{3.2}$$

where again the 1-norm is preserved and all components of $\mathbf{y}_1$ are nonnegative. The Lie group structure is not preserved by this linear combination, but this is not a property that concerns us in the present context. In addition, the difference $\mathbf{x}_1 - \mathbf{z}_1$ can be taken as an estimate of local error, using the scheme as a variable time-step algorithm in order to get more accurate results.

Remark. If A is graph Laplacian and irreducible then (3.1) is a time-symmetric second order method that preserves mass and positivity unconditionally (the average (3.2) breaks time symmetry) and converges to the steady state solution. Let $\mathbf{y}_f$ be the steady state solution, then $\mathbf{f}(\mathbf{y}_f) = A(\mathbf{y}_f)\mathbf{y}_f = \mathbf{0}$. Since $\sigma(A) \setminus \{0\} \in \mathbb{C}^-$ where 0 is a simple eigenvalue and the method is a composition of exponentials of A, it must converge to a steady state solution, say $\hat{\mathbf{y}}_f$. However, we observe that if we take $\mathbf{y}_0 = \mathbf{y}_f$ then it is trivial to check that $\mathbf{x}_{1/2} = \mathbf{z}_1 = \mathbf{x}_1 = \mathbf{y}_f$, so $\mathbf{y}_1 = \mathbf{y}_f$ and then $\hat{\mathbf{y}}_f = \mathbf{y}_f$.

Note also that

$$\mathbf{u} = \left[I - \frac{1}{2}hA(\mathbf{z}_0)\right]^{-1}\mathbf{x}_0 = \mathrm{e}^{\frac{1}{2}hA(\mathbf{z}_0)}\mathbf{x}_0 + \mathcal{O}(h^2)$$

which is a first order approximation to the exponential that, by our earlier Proposition 3.4, still preserves positivity. We can replace $A(\mathbf{x}_{1/2})$ by $A(\mathbf{u})$ in (3.1) and, since this matrix is multiplied by h, the method retains second order of accuracy.

The non-autonomous case. Let us now consider the non-autonomous system

$$\mathbf{y}' = A(t, \mathbf{y})\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{y}_0.$$

This occurs, for example, when a chemical reaction takes place at variable temperature and the coefficients $k_i(t)$ are time dependent or when the parameter R_0 in the SIR model changes due to political decisions, variations in behaviour or the evolution of a pathogen.

In this case we duplicate the system, but taking the time as two dependent variables

$$\begin{aligned} \mathbf{x}' &= A(z_t, \mathbf{z})\mathbf{x}, & \mathbf{x}(0) &= \mathbf{x}_0 = \mathbf{y}_0 \\ x_t' &= 1, & x_t(0) &= t_0 \\ \mathbf{z}' &= A(x_t, \mathbf{x})\mathbf{z}, & \mathbf{z}(0) &= \mathbf{z}_0 = \mathbf{y}_0 \\ z_t' &= 1, & z_t(0) &= t_0 \end{aligned}$$

where $\mathbf{x}(t) = \mathbf{y}(t) = \mathbf{z}(t)$ and $x_t(t) = z_t(t) = t$: the system is now autonomous and separable into solvable parts: the outcome is an algorithm similar to (3.1),

$$\mathbf{x}_{1/2} = \mathrm{e}^{\frac{h}{2}A(t_0,\mathbf{z}_0)}\mathbf{x}_0,$$

$$\mathbf{z}_1 = \mathrm{e}^{hA\left(t_0+h/2,\mathbf{x}_{1/2}\right)}\mathbf{z}_0,$$
$$\mathbf{x}_1 = \mathrm{e}^{\frac{h}{2}A(t_0+h,\mathbf{z}_1)}\mathbf{x}_{1/2},$$

and finally

$$\mathbf{y}_1 = \frac{1}{2}(\mathbf{x}_1 + \mathbf{z}_1). \tag{3.3}$$

3.3. Magnus integrators

A more general procedure to construct higher-order methods is to consider Magnus integrators.

Let $A : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$. We consider the equation

$$\mathbf{y}' = A(t,\mathbf{y})\mathbf{y}, \quad t \geq 0, \qquad \mathbf{y}(0) = \mathbf{y}_0, \tag{3.4}$$

and suppose that $\mathbf{y}_n \simeq \mathbf{y}(t_n)$. One approach toward the solution of (3.4) is

$$\begin{aligned} \mathbf{y}^{[0]} &\equiv \mathbf{y}_n, \\ \mathbf{y}^{[m+1]'} &= A\Big(t,\mathbf{y}^{[m]}(t)\Big)\mathbf{y}^{[m+1]}, \quad \mathbf{y}^{[m+1]}(t_n) = \mathbf{y}_n, \qquad m = 0,1,\ldots,m^*-1, \\ \mathbf{y}_{n+1} &= \mathbf{y}^{[m^*]}(t_{n+1}), \qquad \text{where} \qquad t_{n+1} = t_n + h_n, \end{aligned} \tag{3.5}$$

that corresponds to an approximation to the exact solution to order m^*. The linear ODE in (3.5) can be solved *e.g.* by Magnus series expansion [42] (see also [8,31,32] and references therein). For simplicity, we first consider the autonomous case, with $A(\mathbf{y})$, and next we show the results for the non-autonomous problem.

For example, for $m^* = 1$ we have $\mathbf{y}^{[1]'} = A(\mathbf{y}_n)\mathbf{y}^{[1]}$, therefore $\mathbf{y}^{[1]}(t) = \mathrm{e}^{(t-t_n)A(\mathbf{y}_n)}\mathbf{y}_n$ and we obtain the first-order method

$$\mathbf{y}_{n+1} = \mathrm{e}^{h_n A(\mathbf{y}_n)}\mathbf{y}_n. \tag{3.6}$$

Letting $m^* = 2$ leads to a second-order method $\mathbf{y}^{[2]'} = A\big(\mathrm{e}^{(t-t_n)A(\mathbf{y}_n)}\mathbf{y}_n\big)\mathbf{y}^{[2]}$, whose Magnus solution truncated to the first term that provides second order approximations in the time step is

$$\begin{aligned} \mathbf{y}^{[2]}(t) &= \exp\left(\int_{t_n}^{t} A\Big(\mathrm{e}^{(\tau-t_n)A(\mathbf{y}_n)}\mathbf{y}_n\Big)\mathrm{d}\tau\right)\mathbf{y}_n \\ &\approx \exp\left(\frac{t-t_n}{2}\Big(A(\mathbf{y}_n) + A\Big(\mathrm{e}^{(t-t_n)A(\mathbf{y}_n)}\mathbf{y}_n\Big)\Big)\right)\mathbf{y}_n \end{aligned}$$

(note that the approximation of the integral with the trapezoidal rule is fully consistent with second order). This results in the second-order method

$$\mathbf{y}_{n+1} = \exp\left(\frac{h_n}{2}\Big(A(\mathbf{y}_n) + A\Big(\mathrm{e}^{h_n A(\mathbf{y}_n)}\mathbf{y}_n\Big)\Big)\right)\mathbf{y}_n. \tag{3.7}$$

If we consider instead the midpoint rule we have

$$\mathbf{y}_{n+1} = \exp\Big(h_n A\Big(\mathrm{e}^{\frac{1}{2}h_n A(\mathbf{y}_n)}\mathbf{y}_n\Big)\Big)\mathbf{y}_n. \tag{3.8}$$

Note that (3.8) coincides with $\mathbf{z}_1$ in (3.1) for the first step. This method requires only two exponentials but it is not time symmetric. If it is important to preserve time symmetry, the three-exponential method (3.1) should be used, otherwise this simple and cheaper scheme suffices.

Remark. If A is graph Laplacian and irreducible then the first-order methods (3.6) as well as the second order methods (3.7) and (3.8) preserve mass and positivity unconditionally and converge to the steady state solution similarly to the previous splitting methods.

We can easily apply these Magnus integrators to non-autonomous problems. The first-order method is, obviously

$$\mathbf{y}_{n+1} = \mathrm{e}^{h_n A(t_n,\mathbf{y}_n)}\mathbf{y}_n. \tag{3.9}$$

The trapezoidal second-order method is given by

$$\mathbf{y}_{n+1} = \exp\left(\frac{h_n}{2}\left(A(t_n,\mathbf{y}_n) + A\left(t_{n+1}, \mathrm{e}^{h_n A(t_n,\mathbf{y}_n)}\mathbf{y}_n\right)\right)\right)\mathbf{y}_n, \tag{3.10}$$

while the corresponding second-order midpoint rule method is

$$\mathbf{y}_{n+1} = \exp\left(h_n A\left(t_n + \frac{h_n}{2}, \mathrm{e}^{\frac{1}{2}h_n A(t_n\mathbf{y}_n)}\mathbf{y}_n\right)\right)\mathbf{y}_n. \tag{3.11}$$

Note that if we consider the first order approximation to $\mathrm{e}^{h_n A(t_n,\mathbf{y}_n)}$ or $\mathrm{e}^{\frac{1}{2}h_n A(t_n\mathbf{y}_n)}$ given by

$$\mathbf{u}_1 = [I - h_n A(t_n,\mathbf{y}_n)]^{-1}\mathbf{y}_n \qquad \text{or} \qquad \mathbf{u}_2 = \left[I - \frac{h_n}{2}A(t_n,\mathbf{y}_n)\right]^{-1}\mathbf{y}_n$$

as the internal stages in (3.10) or (3.11) then the new and cheaper schemes read

$$\mathbf{y}_{n+1} = \exp\left(\frac{h_n}{2}(A(t_n,\mathbf{y}_n) + A(t_{n+1},\mathbf{u}_1))\right)\mathbf{y}_n, \tag{3.12}$$

or

$$\mathbf{y}_{n+1} = \exp\left(h_n A\left(t_n + \frac{h_n}{2}, \mathbf{u}_2\right)\right)\mathbf{y}_n. \tag{3.13}$$

and still preserve mass and positivity as well as the second order accuracy. We can either compute the exponentials to high accuracy, or to look for cheaper approximations that preserve both mass and positivity, this being a problem to be studied further in the future.

3.4. Patankar methods

A well-known approach toward preservation of mass and positivity are Runge–Kutta–Patankar methods [14, 34, 35, 46, 47]. The idea is to use Runge–Kutta-like methods for *production–destruction systems* in chemical kinetics, of the form

$$y_k' = \sum_{j=1}^{d} p_{k,j}(t,\mathbf{y}) - \sum_{j=1}^{d} d_{k,j}(t,\mathbf{y}), \qquad k = 1,\dots,d, \tag{3.14}$$

where $p_{k,j}(t,\mathbf{y}), d_{k,j}(t,\mathbf{y}) \geq 0$. The first order Patankar method is given by

$$y_{n+1,k} = y_{n,k} + h\sum_{j=1}^{d}\left[p_{k,j}(t_n,\mathbf{y}_n) - d_{k,j}(t_n,\mathbf{y}_n)\frac{y_{n+1,k}}{y_{n,k}}\right], \qquad k = 1,\dots,d.$$

This method preserves positivity but does not preserve mass. In [14] the authors propose a Modified Patankar Euler scheme (MPE) given by

$$y_{n+1,k} = y_{n,k} + h\sum_{j=1}^{d}\left[p_{k,j}(t_n,\mathbf{y}_n)\frac{y_{n+1,j}}{y_{n,j}} - d_{k,j}(t_n,\mathbf{y}_n)\frac{y_{n+1,k}}{y_{n,k}}\right], \qquad k = 1,\dots,d, \tag{3.15}$$

which preserves both mass and positivity unconditionally. Note that we can write (3.14) in our graph-Laplacian notation as

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}) = A(t,\mathbf{y})\mathbf{y}$$

with

$$A_{k,j}(t,\mathbf{y}) = p_{k,j}(t,\mathbf{y})\frac{1}{y_j} - d_{k,j}(t,\mathbf{y})\frac{\delta_{k,j}}{y_k}$$

where $\delta_{k,j} = 0$ if $k \neq j$ and $\delta_{k,k} = 1$. Then, equation (3.15) can be written in matrix form as

$$\mathbf{y}_{n+1} = \mathbf{y}_n + hA(t_n,\mathbf{y}_n)\mathbf{y}_{n+1} \tag{3.16}$$

that preserves mass (because $A(t_n,\mathbf{y}_n)$ is graph Laplacian) and as already shown preserves positivity.

Note that

$$\mathbf{y}_{n+1} = [I - hA(t_n,\mathbf{y}_n)]^{-1}\mathbf{y}_n = e^{hA(t_n,\mathbf{y}_n)}\mathbf{y}_n + \mathcal{O}(h^2),$$

corresponding to a first order rational approximation to the first order exponential Magnus integrator.

The second-order Modified Patankar–Runge–Kutta scheme (MPRK) is given by

$$u_k = y_{n,k} + h\sum_{j=1}^{d}\left[p_{k,j}(t_n,\mathbf{y}_n)\frac{u_j}{y_{n,j}} - d_{k,j}(t_n,\mathbf{y}_n)\frac{u_k}{y_{n,k}}\right], \qquad k = 1,\dots,d,$$

$$y_{n+1,k} = y_{n,k} + \frac{h}{2}\sum_{j=1}^{d}\left\{[p_{k,j}(t_n,\mathbf{y}_n) + p_{k,j}(t_{n+1},\mathbf{u})]\frac{y_{n+1,j}}{u_j} - [d_{k,j}(t_n,\mathbf{y}_n) + d_{k,j}(t_{n+1},\mathbf{u})]\frac{y_{n+1,k}}{u_k}\right\}$$

[14] (Eq. (27)), which can be written in matrix form as

$$\begin{aligned}\mathbf{u} &= [I - hA(t_n,\mathbf{y}_n)]^{-1}\mathbf{y}_n\\ \mathbf{y}_{n+1} &= \left[I - \frac{h}{2}\Big(A(t_n,\mathbf{y}_n)D(\mathbf{y}_n,\mathbf{u}) + A(t_{n+1},\mathbf{u})\Big)\right]^{-1}\mathbf{y}_n\end{aligned} \tag{3.17}$$

where

$$D(\mathbf{y}_n,\mathbf{u}) = \operatorname{diag}\left(\frac{y_{n,1}}{u_1},\dots,\frac{y_{n,d}}{u_d}\right)$$

and $\mathbf{y}_n = (y_{n,1},\dots,y_{n,d})^\top$, $\mathbf{u} = (u_1,\dots,u_d)^\top$.

Note that $\mathbf{u}$ coincides with $\mathbf{u}_1$ in (3.10) and then the modified Patankar method can be considered as a particular second order approximation to the second order Magnus method (3.10). Obviously, different second order approximations to this exponential or to the method using the midpoint rule (3.11) would lead to different modified second order Patankar methods.

Note that during the integration some of the values u_i may approach zero. In this case one may take, for example, $\frac{y_{n,i}}{u_i} = 0$ when u_i is smaller than a given tolerance. Some caution is required if any component of the solution is very close to zero and suddenly grows, as it happens with some of the numerical examples we will consider.

This method has shown a good performance on stiff problems [13]. Higher order modified Patankar methods have also been recently obtained in the literature [20,35,46] and it would be interesting to find if there is any connection with our exponential integrators.

There are other families of methods which consider some kind of adaptive time steps which depend on the phase space and the time step which allows to preserve positivity as well as the linear invariants [2,11,44] but they are not considered in this work and a proper study of their performance with respect to the new methods is left for future research.

3.5. Higher order methods

Continuing in this vain,

$$\mathbf{y}^{[3]\prime} = A\left(\exp\left(\int_{t_n}^{t} A\left(\mathrm{e}^{(\tau-t_n)A(\mathbf{y}_n)}\mathbf{y}_n\right)\mathrm{d}\tau\right)\mathbf{y}_n\right)\mathbf{y}^{[3]} = A_2(t)\mathbf{y}^{[3]}$$

and a fourth-order Magnus reads (this is a 4th-order approximation to $\mathbf{y}^{[3]}$ which is a third order approximation to the exact solution, so the methods will be of order three)

$$\mathbf{y}_{n+1} = \exp\left(\int_{t_n}^{t_{n+1}} A_2(\tau)\mathrm{d}\tau - \frac{1}{2}\int_{t_n}^{t_{n+1}}\int_{t_n}^{\tau}[A_2(\tau), A_2(\eta)]\mathrm{d}\eta\mathrm{d}\tau\right)\mathbf{y}_n.$$

The temptation is now to discretise using standard Magnus quadrature at Gauss–Legendre points but this does not work because the definition of A_2 itself contains an integral. Moreover, the critical issue is the dependence of A_2 on t, not on $\mathbf{y}_n$.

We approximate

$$\mathbf{y}_{n+1} \approx \exp\left(\frac{h_n}{2}(\mathcal{B}_1 + \mathcal{B}_2) + \frac{\sqrt{3}}{12}h_n^2[\mathcal{B}_1, \mathcal{B}_2]\right)\mathbf{y}_n,$$

where

$$\mathcal{B}_1 = A_2\Big(t_n + \Big(\tfrac{1}{2} - \tfrac{\sqrt{3}}{6}\Big)h\Big), \qquad \mathcal{B}_2 = A_2\Big(t_n + \Big(\tfrac{1}{2} + \tfrac{\sqrt{3}}{6}\Big)h\Big)$$

– except that A_2 itself has a built-in integral,

$$A_2(t) = A\left(\exp\left(\int_{t_n}^{t} A_1(\eta)\mathrm{d}\eta\right)\mathbf{y}_n\right).$$

The simplest solution is to approximate that integral also by two-point Gauss–Legendre (note that the interval of integration in the inner integral is of length $\left(\frac{1}{2} - \frac{\sqrt{3}}{6}\right)h_n$ and we need to adjust quadrature points), whereby

$$\begin{aligned}
\mathcal{B}_1 &\approx A\Big(\exp\Big(\Big(\tfrac{1}{4} - \tfrac{\sqrt{3}}{12}\Big)h_n\Big[A_1\Big(t_n + \Big(\tfrac{1}{3} - \tfrac{\sqrt{3}}{6}\Big)h_n\Big) + A_1\big(t_n + \tfrac{1}{6}h_n\big)\Big]\Big)\mathbf{y}_n\Big),\\
\mathcal{B}_2 &\approx A\Big(\exp\Big(\Big(\tfrac{1}{4} + \tfrac{\sqrt{3}}{12}\Big)h_n\Big[A_1\big(t_n + \tfrac{1}{6}h_n\big) + A_1\Big(t_n + \Big(\tfrac{1}{3} + \tfrac{\sqrt{3}}{6}\Big)h_n\Big)\Big]\Big)\mathbf{y}_n\Big).
\end{aligned}$$

Brief explanation: the first integral is in the interval $\left[t_n, t_n + \left(\frac{1}{2} - \frac{\sqrt{3}}{6}\right)\right]$ and the Gauss–Legendre nodes $\frac{1}{2} \pm \frac{\sqrt{3}}{6}$ need be multiplied by the length of the interval. Ditto in the second interval, $\left[t_n, t_n + \left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right)\right]$ and we are saved a single function evaluation because, by happy coincidence, real numbers commute and $\left(\frac{1}{2} - \frac{\sqrt{3}}{6}\right)\left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right) = \left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right)\left(\frac{1}{2} - \frac{\sqrt{3}}{6}\right) = \frac{1}{6}$.

Thus, altogether we need three function evaluations, one more than standard Magnus. Note moreover that the integration in A_1 is explicit,

$$A_1(t) = A\Big(\mathrm{e}^{(t-t_n)A(\mathbf{y}_n)}\Big).$$

We observe that $\mathcal{B}_i,\ i = 1, 2$, are graph Laplacians, but this need not be the case for their commutator $[\mathcal{B}_1, \mathcal{B}_2]$. This problem can be bypassed using commutator-free Magnus integrators [1,9].

Commutator-free Magnus integrators. We describe briefly, using an example, the construction of commutation-free integrators, based upon the work of [9]. We approximate the solution across a single time step by

$$\mathbf{y}_{n+1} \approx \exp\left(\frac{h_n}{2}(\beta\mathcal{B}_2 + \alpha\mathcal{B}_1)\right)\exp\left(\frac{h_n}{2}(\alpha\mathcal{B}_2 + \beta\mathcal{B}_1)\right)\mathbf{y}_n,$$

where

$$\alpha = \frac{1}{2} + \frac{\sqrt{3}}{3}, \qquad \beta = \frac{1}{2} - \frac{\sqrt{3}}{3}$$

and the algorithm is given by

$$\begin{aligned}
\mathbf{x}_1 &= \exp\Big(\Big(\tfrac{1}{3}-\tfrac{\sqrt{3}}{6}\Big)h_nA(\mathbf{y}_n)\Big)\mathbf{y}_n, && A_{1,1}=A(\mathbf{x}_1),\\
\mathbf{x}_2 &= \exp\big(\tfrac{1}{6}h_nA(\mathbf{y}_n)\big)\mathbf{y}_n, && A_{1,2}=A(\mathbf{x}_2),\\
\mathbf{x}_3 &= \exp\Big(\Big(\tfrac{1}{3}+\tfrac{\sqrt{3}}{6}\Big)h_nA(\mathbf{y}_n)\Big)\mathbf{y}_n, && A_{1,3}=A(\mathbf{x}_3),\\
\mathbf{x}_4 &= \exp\Big(\Big(\tfrac{1}{4}-\tfrac{\sqrt{3}}{12}\Big)h_n(A_{1,1}+A_{1,2})\Big)\mathbf{y}_n, && \mathcal{B}_1=A(\mathbf{x}_4),\\
\mathbf{x}_5 &= \exp\Big(\Big(\tfrac{1}{4}+\tfrac{\sqrt{3}}{12}\Big)h_n(A_{1,2}+A_{1,3})\Big)\mathbf{y}_n, && \mathcal{B}_2=A(\mathbf{x}_5),\\
\mathbf{x}_6 &= \exp\big(\tfrac{1}{2}h_n(\alpha\mathcal{B}_2+\beta\mathcal{B}_1)\big)\mathbf{y}_n,\\
\mathbf{y}_{n+1} &= \exp\big(\tfrac{1}{2}h_n(\beta\mathcal{B}_2+\alpha\mathcal{B}_1)\big)\mathbf{x}_6.
\end{aligned} \tag{3.18}$$

This is a seven-exponential method that might be useful when highly accurate results are desired and the cost of each exponential is not excessive. It preserves positivity for moderately stiff problems since it is conditionally positivity preserving. If $\mathbf{y}_n \succeq \mathbf{0}$ then it is easy to see that $\mathbf{x}_i \succeq \mathbf{0},\ i=1,2,3,4,5$ since $A_{1,1}, A_{1,2}, A_{1,3}, \mathcal{B}_1, \mathcal{B}_2 \in \mathcal{L}_d$. However, positivity is guaranteed as long as the matrices

$$\alpha\mathcal{B}_2+\beta\mathcal{B}_1, \qquad \beta\mathcal{B}_2+\alpha\mathcal{B}_1$$

are graph Laplacians [41]. Unfortunately, $\beta<0$ but $\alpha/|\beta| = 7+4\sqrt{3} \simeq 14$, and unless $\mathcal{B}_1, \mathcal{B}_2$ drastically change in a short time interval or their sparsity structure is "unlucky", their linear combinations are likely to inherit graph-Laplacian structure. In other words, while preservation of graph Laplacians for this third-order method is not assured, it is highly likely in practice.

Finally, we present this Magnus integrator to be used on non-autonomous problems. A third-order commutator-free method can be obtained following the same approximations as previously and taking, for example, the midpoint rule when approximating the intermediate integrals that ensure the third order of accuracy for the method, resulting in the following algorithm

$$\begin{aligned}
A_1 &= A\Big(t_n+\Big(\tfrac{1}{6}-\tfrac{\sqrt{3}}{12}\Big)h_n, \mathbf{y}_n\Big), \quad A_2 = A\big(t_n+\tfrac{1}{12}h_n, \mathbf{y}_n\big), \quad A_3 = A\Big(t_n-\Big(\tfrac{1}{6}-\tfrac{\sqrt{3}}{12}\Big)h_n, \mathbf{y}_n\Big),\\
\mathbf{x}_1 &= \exp\Big(\Big(\tfrac{1}{3}-\tfrac{\sqrt{3}}{6}\Big)h_nA_1\Big)\mathbf{y}_n, && A_{1,1} = A\Big(t_n+\Big(\tfrac{1}{3}-\tfrac{\sqrt{3}}{6}\Big)h_n, \mathbf{x}_1\Big)\\
\mathbf{x}_2 &= \exp\big(\tfrac{1}{6}h_nA_2\big)\mathbf{y}_n, && A_{1,2} = A\big(t_n+\tfrac{1}{6}h_n, \mathbf{x}_2\big)\\
\mathbf{x}_3 &= \exp\Big(\Big(\tfrac{1}{3}+\tfrac{\sqrt{3}}{6}\Big)h_nA_3\Big)\mathbf{y}_n, && A_{1,3} = A\Big(t_n+\Big(\tfrac{1}{3}+\tfrac{\sqrt{3}}{6}\Big)h_n, \mathbf{x}_3\Big)\\
\mathbf{x}_4 &= \exp\Big(\Big(\tfrac{1}{4}-\tfrac{\sqrt{3}}{12}\Big)h_n(A_{1,1}+A_{1,2})\Big)\mathbf{y}_n, && \mathcal{B}_1 = A\Big(t_n+\Big(\tfrac{1}{2}-\tfrac{\sqrt{3}}{6}\Big)h_n, \mathbf{x}_4\Big)\\
\mathbf{x}_5 &= \exp\Big(\Big(\tfrac{1}{4}+\tfrac{\sqrt{3}}{12}\Big)h_n(A_{1,2}+A_{1,3})\Big)\mathbf{y}_n, && \mathcal{B}_2 = A\Big(t_n+\Big(\tfrac{1}{2}+\tfrac{\sqrt{3}}{6}\Big)h_n, \mathbf{x}_5\Big)\\
\mathbf{x}_6 &= \exp\big(\tfrac{1}{2}h_n(\alpha\mathcal{B}_2+\beta\mathcal{B}_1)\big)\mathbf{y}_n,\\
\mathbf{y}_{n+1} &= \exp\big(\tfrac{1}{2}h_n(\beta\mathcal{B}_2+\alpha\mathcal{B}_1)\big)\mathbf{x}_6.
\end{aligned} \tag{3.19}$$

4. Hidden graph Laplacian structures for polynomial ODEs

4.1. The recovery of graph Laplacian structure

Given an ODE system of the form

$$y_k' = \sum_{\ell=1}^{d} b_k^{\ell} y_\ell + \sum_{\ell=1}^{d}\sum_{i=1}^{d} a_k^{\ell,i} y_\ell y_i, \qquad k=1,\dots,d, \tag{4.1}$$

with suitable initial conditions $\mathbf{y}(0) \succeq \mathbf{0}$, $\mathbf{1}^\top \mathbf{y}(0) = 1$, we seek conditions so that it can be written in the form (1.2), where the matrix $A(\mathbf{y})$ is a graph Laplacian, namely that for every nonnegative $\mathbf{y}$ such that $\mathbf{1}^\top \mathbf{y} = 1$ it is true that $A_{k,k}(\mathbf{y}) \le 0$ and $A_{k,\ell}(\mathbf{y}) \ge 0$, $\ell \ne k$. Moreover, we seek constructive means of deriving such a matrix A, given (4.1).

Our first observation is that the representation of (4.1) in the form (1.2) is additive, in the sense that if we can do so for two different right-hand sides of (4.1), we can do so for their sum. By the same token, if we can do so separately for the first sum and the second, double sum in (4.1), all we need is simply add the two representations. The first sum is trivial and corresponds to the constant-matrix representation $\mathbf{y}' = B\mathbf{y}$, where $B = (b_k^\ell)$ is a graph Laplacian. Consequently, the task at hand reduces to the derivation of a representation (1.2) of the system

$$y_k' = \sum_{\ell=1}^d \sum_{i=1}^d a_k^{\ell,i} y_\ell y_i, \qquad k = 1, \dots, d.$$

With greater generality, we may just as well consider the multinomial ODE system

$$y_k' = \sum_{j=1}^m \sum_{\substack{\ell_1+\cdots+\ell_d=j \\ \ell_1,\dots,\ell_d \ge 0}} a_k^{\ell_1,\dots,\ell_d} y_1^{\ell_1} \cdots y_d^{\ell_d} = \sum_{j=1}^m \sum_{|\ell|=j} a_k^\ell \mathbf{y}^\ell, \qquad k = 1, \dots, d,$$

with initial conditions $\mathbf{y}(0) = \mathbf{y}_0 \succeq \mathbf{0}$, $\mathbf{1}^\top \mathbf{y}_0 = 1$. Again, the challenge is to write it in the form (1.2) with a graph Laplacian $A(\mathbf{y})$ and, again, we can use the same argument to split the task at hand into a sum of homogeneous problems of the form

$$y_k' = \sum_{|\ell|=j} a_k^\ell \mathbf{y}^\ell, \qquad k = 1, \dots, d \tag{4.2}$$

for $j = 2, \dots, m$ – the case $j = 1$ is trivial.

The problem, though, is that (4.2) can be written in the form (1.2) in a multitude of ways – indeed, even the coefficients a_k^ℓ are not unique. This can be seen in the simplest nontrivial case, $d = 2$ and $j = 2$:

$$\begin{aligned} y_1' &= a_1^{1,1} y_1^2 + \left(a_1^{1,2} + a_1^{2,1}\right) y_1 y_2 + a_1^{2,2} y_2^2, \\ y_2' &= a_2^{1,1} y_1^2 + \left(a_2^{1,2} + a_2^{2,1}\right) y_1 y_2 + a_2^{2,2} y_2^2. \end{aligned}$$

Therefore

$$A(\mathbf{y}) = \begin{bmatrix} a_1^{1,1} y_1 + \beta_{1,1} y_2 & \beta_{1,2} y_1 + a_1^{2,2} y_2 \\ a_2^{1,1} y_1 + \beta_{2,1} y_2 & \beta_{2,2} y_1 + a_2^{2,2} y_2 \end{bmatrix},$$

where

$$\beta_{1,1} + \beta_{1,2} = a_1^{1,2} + a_1^{2,1}, \qquad \beta_{2,1} + \beta_{2,2} = a_2^{1,2} + a_2^{2,1}. \tag{4.3}$$

We deduce that in this case the graph-Laplacian conditions (which must hold for all $\mathbf{y} \succeq \mathbf{0}$) are

$$\begin{aligned} &a_1^{1,1}, a_2^{2,2}, \beta_{1,1}, \beta_{2,2} \le 0, \\ &a_1^{1,1} + a_2^{1,1} = a_1^{2,2} + a_2^{2,2} = \beta_{1,1} + \beta_{2,1} = \beta_{1,2} + \beta_{2,2} = 0. \end{aligned}$$

Six equalities (inclusive of (4.3)) and four inequalities for eight variables: impossible in some configurations, while other configurations lead to an infinity of solutions.

Henceforth we let $\mathbf{e}_i$ stand for the ith unit vector.

Theorem 4.1. *The ODE system*

$$y_k' = \sum_{\substack{\ell_1+\cdots+\ell_d=2 \\ \ell_1,\ldots,\ell_d\geq 0}} a_k^{\ell} y_1^{\ell_1} y_2^{\ell_2} \cdots y_d^{\ell_d}, \qquad k = 1,\ldots,d \tag{4.4}$$

admits the graph Laplacian representation (1.2) *subject to the assumptions*

$$a_k^{2\mathbf{e}_k} \leq 0, \quad a_k^{2\mathbf{e}_i} \geq 0, \qquad k,i = 1,\ldots,d, \quad i \neq k, \tag{4.5}$$

$$a_k^{\mathbf{e}_i+\mathbf{e}_k} \leq 0, \quad a_k^{\mathbf{e}_i+\mathbf{e}_j} \geq 0, \quad k,i,j = 1,\ldots,d, \quad i \neq j, \quad k \neq i,j, \tag{4.6}$$

$$\sum_{k=1}^{d} a_k^{\mathbf{e}_k+\mathbf{e}_i} = 0, \qquad i = 1,\ldots,d. \tag{4.7}$$

Proof. We prove the theorem by constructing explicitly a graph Laplacian $A(\mathbf{y})$, letting

$$A_{k,\ell}(\mathbf{y}) = a_k^{2\mathbf{e}_\ell} y_\ell + a_k^{\mathbf{e}_\ell+\mathbf{e}_{\ell+1}} y_{\ell+1}, \qquad k,\ell = 1,\ldots,d \qquad (\text{mod } d). \tag{4.8}$$

All that remains is to prove that $A(\mathbf{y})$, as defined in (4.8), is indeed a graph Laplacian. Thus, recalling that $y_1,\ldots,y_d \geq 0$ and that k is computed *modulo* d,

$$A_{k,k}(\mathbf{y}) = a_k^{2\mathbf{e}_k} y_k + a_k^{\mathbf{e}_k+\mathbf{e}_{k+1}} y_{k+1} \leq 0$$

because of (4.5) and (4.6). These two conditions also imply that

$$A_{k,\ell}(\mathbf{y}) = a_k^{2\mathbf{e}_\ell} y_\ell + a_k^{\mathbf{e}_\ell+\mathbf{e}_{\ell+1}} y_{\ell+1} \geq 0, \qquad k \neq \ell.$$

Finally, it follows from (4.7) that

$$\sum_{k=1}^{d} A_{k,\ell}(\mathbf{y}) = \left(\sum_{k=1}^{d} a_k^{2\mathbf{e}_\ell}\right) y_\ell + \left(\sum_{k=1}^{d} a_k^{\mathbf{e}_\ell+\mathbf{e}_{\ell+1}}\right) y_{\ell+1} = 0$$

and we are done. □

As an example, we revisit (2.1), focussing on the quadratic part. Now

$$a_1^{\mathbf{e}_2+\mathbf{e}_3} = 10^4, \quad a_2^{\mathbf{e}_2+\mathbf{e}_3} = -10^4, \quad a_2^{2\mathbf{e}_2} = -3\cdot 10^7, \quad a_3^{2\mathbf{e}_2} = 3\cdot 10^7$$

and the remaining coefficients are zero: it is easy to verify that the conditions of Theorem 4.1 are satisfied. The representation (4.8), incidentally, corresponds to (2.2), the graph-Laplacian form of of the Robertson reaction.

In this paper we focus only on equations (4.4). The situation is more subtle for higher-order equations. For example, consider the case $d = 2$, $m = 3$ and

$$\begin{aligned} y_1' &= -\alpha_2^{3,0} y_1^3 + \alpha_1^{2,1} y_1^2 y_2 + \alpha_1^{1,2} y_1 y_2^2 + \alpha_1^{0,3} y_2^3, \\ y_2' &= \alpha_2^{3,0} y_1^3 - \alpha_1^{2,1} y_1^2 y_2 - \alpha_1^{1,2} y_1 y_2^2 - \alpha_1^{0,3} y_2^3. \end{aligned}$$

The most general way of writing it in the form (1.2) is with the matrix

$$A(\mathbf{y}) = \begin{bmatrix} -\alpha_2^{3,0} y_1^2 - \beta_{2,1}^{2,1} y_1 y_2 - \beta_{2,1}^{1,2} y_2^2 & \left(\alpha_1^{2,1} + \beta_{2,1}^{2,1}\right) y_1^2 + \left(\alpha_1^{1,2} + \beta_{2,1}^{1,2}\right) y_1 y_2 + \alpha_1^{0,3} y_2^2 \\ \alpha_2^{3,0} y_1^2 + \beta_{2,1}^{2,1} y_1 y_2 + \beta_{2,1}^{1,2} y_2^2 & -\left(\alpha_1^{2,1} + \beta_{2,1}^{2,1}\right) y_1^2 - \left(\alpha_1^{1,2} + \beta_{2,1}^{1,2}\right) y_1 y_2 - \alpha_1^{0,3} y_2^3 \end{bmatrix},$$

where $\beta_{2,1}^{2,1}$ and $\beta_{2,1}^{1,2}$ are constants. Clearly, to have a graph Laplacian for all $\mathbf{y} \succeq \mathbf{0}$ we require $\alpha_1^{0,3}, \alpha_2^{3,0} \geq 0$ and the two parameters need to satisfy

$$\beta_{2,1}^{2,1} \geq \max\left\{0, -\alpha_1^{2,1}\right\}, \qquad \beta_{2,1}^{1,2} \geq \max\left\{0, -\alpha_1^{1,2}\right\}.$$

Note that it is possible for $\beta_{2,1}^{2,1} < 0$, say, and yet $A_{2,1} \geq 0$, provided that $\beta_{2,1}^{1,2} \geq 0$ and $\beta_{2,1}^{2,1} \geq -2\sqrt{\alpha_2^{3,0}\beta_{2,1}^{1,2}}$. As an example, we can write

$$y_1' = -y_1^3 + y_1^2 y_2 + y_2^3, \qquad y_2' = y_1^3 - y_1^2 y_2 - y_2^3$$

in the form (1.2) with

$$A(\mathbf{y}) = \begin{bmatrix} -y_1^2 & y_1^2 + y_2^2 \\ y_1^2 & -\left(y_1^2 + y_2^2\right) \end{bmatrix}$$

but it can also be written as

$$A(\mathbf{y}) = \begin{bmatrix} -y_1^2 + \frac{1}{2}y_1 y_2 - y_2^2 & \frac{1}{2}y_1^2 + y_1 y_2 + y_2^2 \\ y_1^2 - \frac{1}{2}y_1 y_2 + y_2^2 & -\frac{1}{2}y_1^2 - y_1 y_2 - y_2^2 \end{bmatrix} = \begin{bmatrix} -(y_1 - y_2)^2 & \frac{1}{2}\left(y_1^2 + y_2\right)^2 \\ (y_1 - y_2)^2 & -\frac{1}{2}\left(y_1^2 + y_2\right)^2 \end{bmatrix}.$$

Note that this cannot occur for quadratic equations (4.4) because, once $A_{k,\ell}(\mathbf{y})$ is a multilinear function of $\mathbf{y} \succeq \mathbf{0}$, it is a graph Laplacian only if all off-diagonal coefficients are nonnegative.

4.2. Chemical reactions by the Law of Mass Action

An important application are chemical reactions, where the rate of reaction is modelled by the Law of Mass Action. Then the model is a first-order ODE with a multivariate polynomial for the right hand side, so it can be considered an important special case of our framework. Suppose there are N reactions, where the j-th reaction is written in the form

$$r_{j,1}G_1 + r_{j,2}G_2 + \ldots + r_{j,M}G_M \quad \xrightarrow{k_j} \quad q_{j,1}G_1 + q_{j,2}G_2 + \ldots + q_{j,M}G_M,$$

$j = 1 \ldots, N$. Here $r_{i,j}, q_{i,j}$ are integer coefficients, $G_i,\ i = 1, 2, \ldots, M$ are symbols for the chemical species, y_i denotes the concentration of species i, and k_j is the rate constant. The model is the ODE

$$\mathbf{y}' = S\mathbf{p}, \qquad \mathbf{y}(0) = \mathbf{y}_0 \tag{4.9}$$

where $\mathbf{y} \in \mathbb{R}^M, S \in \mathbb{R}^{M\times N},\ \mathbf{p} \in \mathbb{R}^N$, and $S_{i,j} = q_{i,j} - r_{i,j}$ is the matrix of *stoichiometric vectors*, while

$$p_j = k_j \prod_{i=1}^{M} y_i^{r_{i,j}}$$

is the Law of Mass Action to model the rates of reaction. This is a nonlinear and autonomous differential equation (it would be non-autonomous if the rates $k_j = k_j(t)$ were time-varying, for example to model fluctuating temperatures). The following theorem shows this model can always be written in the form

$$\mathbf{y}' = \mathcal{L}(\mathbf{y})\mathbf{y}, \tag{4.10}$$

where the matrix $\mathcal{L}(\mathbf{y})$ has the same pattern of signs as a Laplacian, *i.e.* off-diagonal entries are nonnegative, and negative entries can only appear on the diagonal.

Theorem 4.2. *The nonlinear ODE (4.9) can be written in the quasi-linearised form (4.10) where the negative elements of the matrix $\mathcal{L}(\mathbf{y})$ only appear on the diagonal.*

Proof. We assume that $q_{i,j}, r_{i,j}$ are non-negative integers, k_j is a non-negative real number and $y_j \geq 0$. Then, a negative coefficient could only appear in the stoichiometric matrix $S_{i,j}$ if $r_{i,j} \geq 1$, and this happens in the equation for y_i'. Since we have $p_j = k_j\Big(\prod_{k=1,k\neq i}^{M} y_k^{r_{k,j}}\Big) y_i^{r_{i,j}}$ with $r_{i,j} \geq 1$, we may allocate this term to the diagonal of the matrix $\mathcal{L}(\mathbf{y})$. All other components where $r_{i,j} = 0$ in the right hand side of the equation for y_i' have positive coefficients and can be allocated outside the diagonal. □

Remark. Note that the matrix $\mathcal{L}(\mathbf{y})$ of (4.10) need not be unique, as we previously showed by the example of the Robertsons reaction in (2.2). The theorem shows that we may form $\mathcal{L}(\mathbf{y})$ so that it has the right pattern of signs to be a graph Laplacian. Similarly to the remarks following Proposition 1.1, this ensures positivity of the solutions, and the point we are making here is that the new numerical methods proposed in this paper can be applied, *via* (4.10), to this big class of important applications. The only difference between (4.10) and the primary focus of this paper in (1.2), is that in (1.2) we additionally assume that $\mathbf{1}^\top$ is in the left null space of $A(\mathbf{y})$, but that does not prevent us from applying the numerical schemes proposed in this paper, and they will preserve positivity as required. (Although there may be issues with other conservation laws, as we show in the autonomous oscillations example (2.6), and our atmospheric chemistry example (5.1).)

5. Numerical experiments

In this section we present some numerical experiments to illustrate the performance of the new methods on a number of examples from the literature. We denote:

- ES2: The symmetric second order 3-exponential splitting method (3.2) or (3.3);
- EM1: The first order 1-exponential Magnus integrator (3.6) or (3.9);
- EM2: The second order 2-exponential Magnus integrator (3.8) or (3.11);
- EM3: The third order 7-exponential Magnus integrator (3.18) or (3.19).

We will also consider, for comparison, the following more conventional numerical solvers:

- Euler: The first-order explicit Euler method;
- RK4: The 4-stage fourth-order explicit RK method (as a reference method to compare);
- ROS4: The 4-stage fourth-order Rosenbrock method with coefficients used by default in [24].
- MP2: The second order Modified Patankar method.

5.1. Example 1: The SIDARTHE mathematical model

We first consider a generalised SIR model (SIDARTHE) that has been used to model the evolution of the Cov-SARS-2 epidemic in Italy [21]. That model can also be used for any other country with appropriate data or it can be even extended *e.g.* to age-dependent variables.

The SIDARTHE dynamical system [21] consists of eight ordinary differential equations, describing the evolution of the population in each stage over time. The equations can be written in the form

$$\mathbf{y}' = A(\mathbf{b}(t), \mathbf{y})\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{y}_0 \in \mathbb{R}^8,$$

where $\mathbf{b} : \mathbb{R} \to \mathbb{R}^{15}$ is a vector function depending on 15 time-dependent parameters. The vector $\mathbf{b}$ was taken as a piecewise constant function, and the authors estimate the model parameters based on data from 20 February 2020 (day 1) to 5 April 2020 (day 46) and show the impact of progressive restrictions on the spread of the epidemic. For example, $\mathbf{b}$ is constant from day 1 to 4 (with a value of $R_0 = 2.28$), and changes to new constant values for the period 4 to 12 (with a value of $R_0 = 1.66$), and so on.

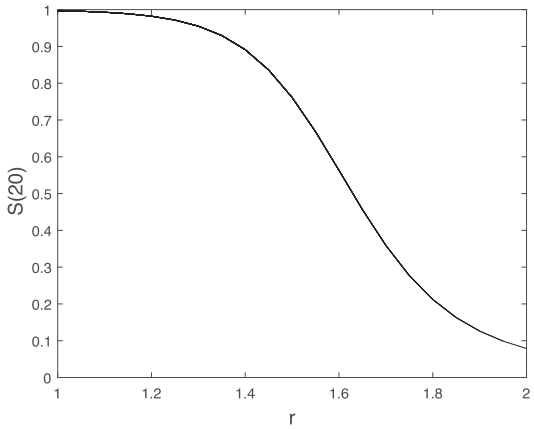

FIGURE 1. Solution for the susceptible population at $t = 20$ for different values of the parameter r where $\alpha = 0.57 \cdot r$.

Notice that since the vector field is not a smooth function (it is piecewise constant) the numerical methods deteriorate down to order one. However, a more realistic model should consider $\mathbf{b}(t)$ as a smooth time-dependent function, and in this case the order of the methods is recovered.

For simplicity, we take the same initial values for $\mathbf{b}$ and the same initial conditions $\mathbf{y}_0$ as in [21], but we take $\mathbf{b}$ constant for a longer period, from day 1 to 20.

We observed that the model is very sensitive to the parameter associated to the first component of $\mathbf{b}$, $b_1 = \alpha$. That parameter was taken initially as $\alpha = 0.57$, and we have analysed the solution for the first component of $\mathbf{y}$ (*i.e.* $y_1(t) = S(t)$, the susceptible (uninfected) population at day 20) for different values of $\alpha = 0.57 \cdot r$ with $r \in [1, 2]$. The results are shown in Figure 1.

Next, we take $r = 1.5$, corresponding to a moderately stiff problem ($S(20)$ still has not dramatically decreased) and we compute the 2-norm error of the solution $\mathbf{y}(20)$ *versus* the time step for the new methods as well as for the explicit Euler method that was used in [21]. The results are displayed in Figure 2 (left) where the order of the methods is clearly visible from the slopes of the curves.

The new methods require to compute matrix exponentials and this can be computationally costly in some cases. It is thus interesting to study if it is possible to replace the exact exponential of matrices by cheaper approximations while still preserving positivity.

This is not a very stiff problem and we have repeated the same numerical experiments while replacing each exponential by the second-order diagonal Padé approximation. In order to preserve positivity, we proceed as follows, given $A = \tilde{A} + a^* I$ where $\tilde{A} \succeq O$, we consider the following approximation to the exponential

$$\mathrm{e}^{tA} = \mathrm{e}^{ta^*} \mathrm{e}^{t\tilde{A}} \simeq \mathrm{e}^{ta^*} \frac{1 + \frac{1}{2} t\tilde{A}}{1 - \frac{1}{2} t\tilde{A}}.$$

Note that, since $\mathbf{1}^\top \tilde{A} = -a^*$, we have

$$\mathbf{1}^\top \mathrm{e}^{ta^*} \frac{1 + \frac{1}{2} t\tilde{A}}{1 - \frac{1}{2} t\tilde{A}} = \mathrm{e}^{ta^*} \frac{1 - \frac{1}{2} ta^*}{1 + \frac{1}{2} ta^*} \neq 1$$

and mass is not preserved. This can be fixed, for example, if we also approximate the scalar function e^{ta^*} by the second-order diagonal Padé approximation, so

$$\mathbf{1}^\top \frac{1 + \frac{1}{2} ta^*}{1 - \frac{1}{2} ta^*} \frac{1 + \frac{1}{2} t\tilde{A}}{1 - \frac{1}{2} t\tilde{A}} = 1$$

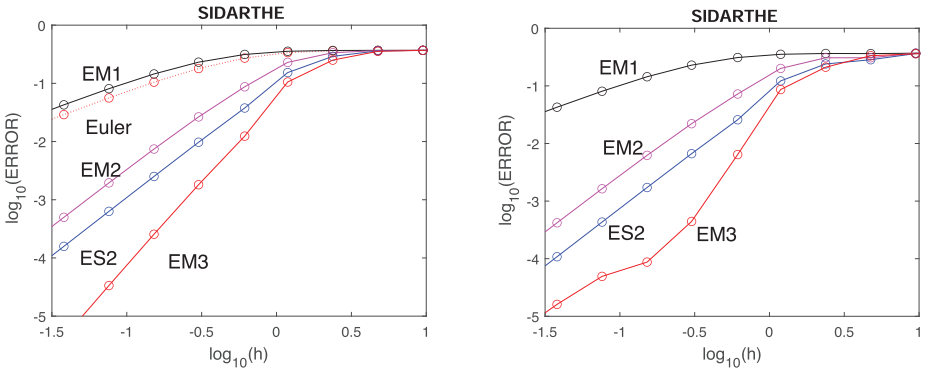

FIGURE 2. The 2-norm error of the solution of the SIDARTHE mathematical model at the final time *versus* the time step in double logarithmic scale: (*left*) the new methods compute the exponential of matrices to round-off accuracy (this, of course, is irrelevant to the explicit Euler method); and (*right*) the same new methods with each exponential replaced by the second order Padé approximation.

and this approach preserves norm and positivity in the stability region.

The results are shown in Figure 2 (right). We observe that the schemes maintain their accuracy while being considerably cheaper. The third-order method EM3 exhibits second order accuracy (due to the second order Padé approximation) but this occurs only at higher accuracies.

For clarity in the presentation, the results for MP2 are not shown but, as expected they are slightly worse but close to the results given by EM2.

Unfortunately, this is not the case if we repeat the numerical experiment with the very stiff problem of Robertson's reaction. Once higher-order approximations to the exponential are used, positivity is not guaranteed. Not all higher-order Padé approximations preserve positivity, unlike the second order one, and this deserves further investigation.

5.2. Example 2: Robertson's reaction.

Let us now consider the Robertson's reaction written in the form (2.2) with initial conditions $\mathbf{y}_0 = [1, 0, 0]^\top$ and time interval $t \in [0, 0.3]$ as in [24] (p. 57). We numerically solve the problem repeatedly using different values for the time step and compute the 2-norm error of the solution at the final time. Here, we compare with the "exact" solution that is computed numerically with sufficiently high accuracy.

Notice that this is a very stiff problem that turns into a non-stiff problem if one applies an appropriate time transformation which can be integrated with a constant time step (in the fictitious time) by methods for non-stiff problems. This is basically the case studied in [14] with time step $h_n = 1.8^n \times h_0$ and initial time step $h_0 = 10^{-6}$ that allows to integrate for the interval $t \in [0, 10^{11}]$ with a very small number of time steps, but the details in the reaction at the very beginning can be lost.

Figure 3 (left) shows the error *versus* the time step in double logarithmic scale. The implicit Rosenbrock method, ROS4, outperforms the explicit RK methods, Euler and RK4, but also turns unstable for moderate values of the time step (and does not preserve positivity) while the new exponential methods preserve positivity and are unconditionally stable (the third order method, EM3, preserves positivity for all time steps considered). Note the relatively high accuracy provided by the new schemes even when considering large time steps. The best method among the proposed schemes depends on the desired accuracy where the computational cost has to be taken into account.

As in the previous example, the results for MP2, not shown, are slightly worse but close to the results given by EM2.

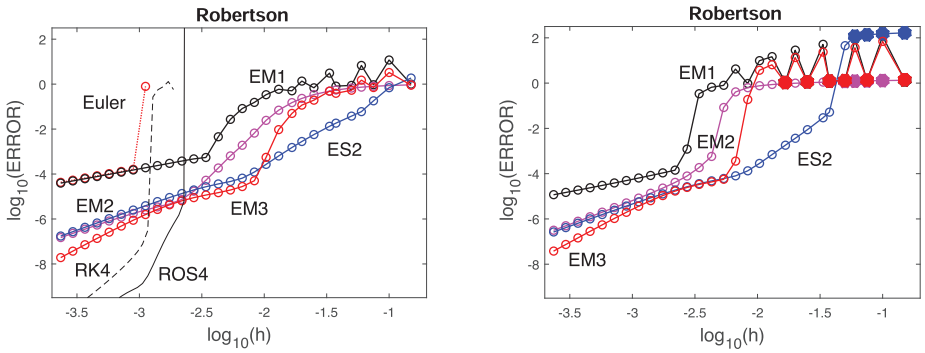

FIGURE 3. The 2-norm error of the solution of the Robertson's reaction at the final time *versus* the time step in double logarithmic scale: (*left*) the new methods are used to solve (2.2) where the matrix is graph Laplacian (the results of the standard methods Euler, RK4 and ROS4 are also included); and (*right*) the same new methods applied to solve the same problem, but written in the form (2.3), where the matrix is no longer a graph Laplacian (a filled circle indicates that a negative solution on any of the components has been obtained in the course of the time integration).

We have repeated the same numerical experiments using only the new exponential methods, but applied to the equations as given in (2.3), *i.e.* the same problem but written in a different way such that the matrix is no longer graph Laplacian. Figure 3 (right) shows the results obtained. We filled a relevant circle when, during the numerical integration, a negative solution was obtained on any of the components. For small time steps the performance is quite similar (and the performance for EM1 is actually somewhat better) but the errors grow faster for large time steps (lower accuracies) and, even worse, negative solutions do occur.

5.3. Example 3: The stratospheric reaction

Let us consider the basic stratospheric reaction mechanism studied in [49] that involves six species

$$\mathbf{y} = [\,[O^{1D}], [O], [O_3], [O_2], [NO], [NO_2]\,]^\top = [y_1, \dots, y_6]^\top$$

and whose model to obtain the evolution of the concentrations is given by the system of ODEs

$$\begin{aligned}
y_1' &= k_5 y_3 - k_6 y_1 - k_7 y_1 y_3 \\
y_2' &= 2k_1 y_4 - k_2 y_2 y_4 + k_3 y_3 - k_4 y_2 y_3 + k_6 y_1 - k_9 y_2 y_6 + k_{10} y_6 \\
y_3' &= k_2 y_2 y_4 - k_3 y_3 - k_4 y_2 y_3 - k_5 y_3 - k_7 y_1 y_3 - k_8 y_3 y_5 \\
y_4' &= -k_1 y_4 - k_2 y_2 y_4 + k_3 y_3 + 2k_4 y_2 y_3 + k_5 y_3 + 2k_7 y_1 y_3 + k_8 y_3 y_5 + k_9 y_2 y_6 \\
y_5' &= -k_8 y_3 y_5 + k_9 y_2 y_6 + k_{10} y_6 \\
y_6' &= k_8 y_3 y_5 - k_9 y_2 y_6 - k_{10} y_6
\end{aligned} \tag{5.1}$$

with

$$\begin{aligned}
&k_1 = 2.643 \cdot 10^{-10} \sigma^3(t), \quad && k_2 = 8.018 \cdot 10^{-17}, \quad && k_3 = 6.120 \cdot 10^{-4} \sigma(t), \\
&k_4 = 1.576 \cdot 10^{-15}, && k_5 = 1.070 \cdot 10^{-3} \sigma^2(t), && k_6 = 7.110 \cdot 10^{-11}, \\
&k_7 = 1.200 \cdot 10^{-10}, && k_8 = 6.062 \cdot 10^{-15}, && k_9 = 1.069 \cdot 10^{-11}, \\
&k_{10} = 1.289 \cdot 10^{-2} \sigma(t),
\end{aligned}$$

where

$$\sigma(t) = \begin{cases} \frac{1}{2} + \frac{1}{2} \cos\left(\pi \left| \frac{2T_L - T_R - T_S}{T_S - T_R} \right|^{\frac{2T_L - T_R - T_S}{T_S - T_R}}\right) & \text{if } T_R \le T_L \le T_S \\ 0 & \text{otherwise.} \end{cases}$$

The time is measured in seconds and it is taken as

$$T_L = \left(\frac{t}{3600}\right) \bmod 24, \qquad T_R = 4.5, \qquad T_S = 19.5.$$

The initial time is considered at noon, $t_0 = 12 \times 3600$, and it is integrated for three full days, until $t_f = t_0 + 72 \times 3600$ with initial conditions given by

$$\mathbf{y}_0 = \left[9.906 \cdot 10^1, 6.624 \cdot 10^8, 5.326 \cdot 10^{11}, 1.697 \cdot 10^{16}, 8.725 \cdot 10^8, 2.240 \cdot 10^8\right]^\top.$$

This is a non-autonomous systems that can be written in the form

$$\mathbf{y}' = A(t, \mathbf{y})\mathbf{y}$$

with $A(t, \mathbf{y})$ an explicitly time-dependent graph Laplacian matrix. We can write the vector field in terms of the production and destruction parts

$$A(t, \mathbf{y})\mathbf{y} = P(t, \mathbf{y}) - D(t, \mathbf{y})\mathbf{y}$$

where $P(t, \mathbf{y}), D(t, \mathbf{y})\mathbf{y}$ are non-negative. While the diagonal matrix D is unique in this case, we can write

$$P(t, \mathbf{y}) = A_P(t, \mathbf{y})\mathbf{y}$$

in many different ways for the matrix A_P. We have considered the following choice (other choices of A_P can be considered) for A,

$$\begin{bmatrix}
-(k_6 + k_7 y_3) & 0 & k_5 & 0 & 0 & 0 \\
k_6 & -(k_2 y_4 + k_4 y_3 + k_9 y_6) & k_3 & 2k_1 & 0 & k_{10} \\
0 & \frac{1}{3} k_2 y_4 & -\gamma & \frac{2}{3} k_2 y_2 & 0 & 0 \\
\frac{1}{2} k_7 y_3 & k_4 y_3 + \frac{1}{2} k_9 y_6 & \gamma + \frac{1}{2} k_7 y_1 & -(k_1 + k_2 y_2) & 0 & \frac{1}{2} k_9 y_2 \\
0 & 0 & 0 & 0 & -k_8 y_3 & k_{10} + k_9 y_2 \\
0 & 0 & 0 & 0 & k_8 y_3 & -(k_{10} + k_9 y_2)
\end{bmatrix}$$

with $\gamma = k_3 + k_5 + k_4 y_2 + k_7 y_1 + k_8 y_5$.

This problem has two linear mass conservation laws, the number of atoms of oxygen and nitrogen, respectively. Given

$$\mathbf{w}_1 = [1, 1, 3, 2, 1, 2]^\top, \qquad \mathbf{w}_2 = [0, 0, 0, 0, 1, 1]^\top$$

it is true that

$$\mathbf{w}_1^\top A(t, \mathbf{y})\mathbf{y} = \mathbf{w}_2^\top A(t, \mathbf{y})\mathbf{y} = \mathbf{0}.$$

Unfortunately, it is impossible to find a matrix A_P such that

$$\mathbf{w}_1^\top A(t, \mathbf{y}) = \mathbf{w}_2^\top A(t, \mathbf{y}) = \mathbf{0},$$

and both mass conservations cannot be simultaneously preserved by our schemes. We have to decide how to choose A_P to optimise the performance of our methods: this is typical to geometric numerical integration of differential equations with multiple invariants.

For this particular choice we have

$$\mathbf{w}_2^\top A(t, \mathbf{y}) = \mathbf{0}, \qquad \text{but} \qquad \mathbf{w}_1^\top A(t, \mathbf{y}) \neq \mathbf{0},$$

and then, in general, $\mathbf{w}_1^\top \mathbf{y}(t) \neq$ const. However, a good choice for A_P can provide solutions where this quantity is preserved to very high accuracy.

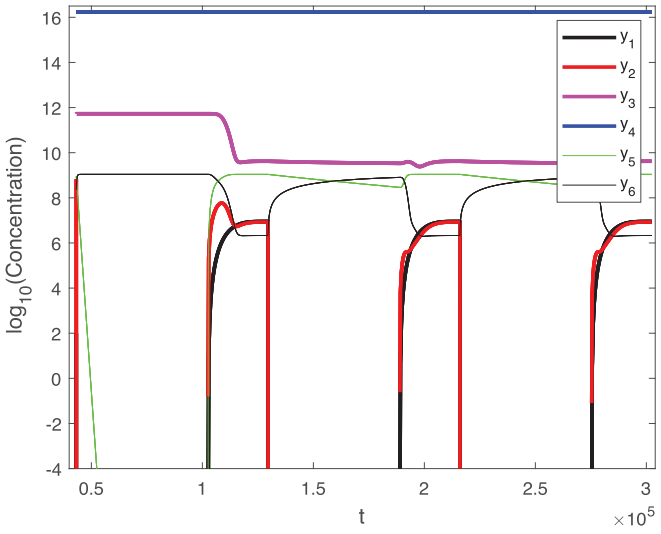

FIGURE 4. Solution for the concentrations for the stratospheric reaction in a logarithmic scale.

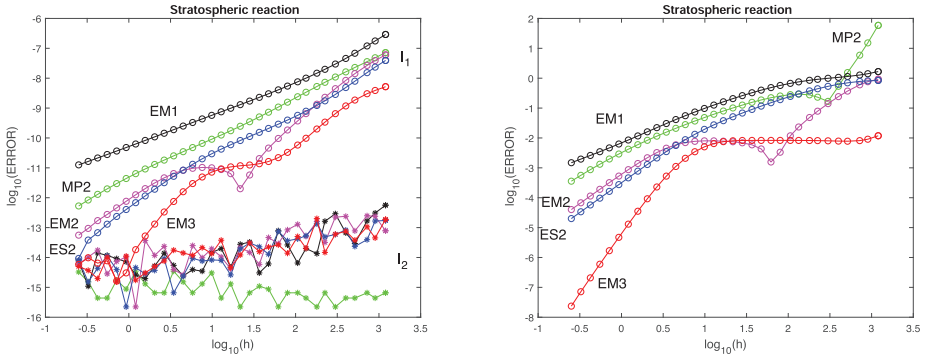

FIGURE 5. *Left*: error in the preserved quantities I_1 and I_2 for the stratospheric reaction model. *Right*: 2-norm error of the numerical solutions for the stratospheric reaction model for the components (y_3, y_4, y_5, y_6) at the final time $t_f = t_0 + 3600$ (one hour) *versus* the time step in double logarithmic scale.

We have observed that y_1, y_2 and y_5 take very small, but positive, values (say 10^{-200} or smaller) along the integration (standard methods usually provide negative values). In that case, measuring relative error is not appropriate for these components.

Figure 4 shows the evolution of the concentration of the different species in a logarithmic scale. Negative values in this plot correspond to having no particles.

We have repeated the numerical experiments, integrating for just one hour (instead of 72 h) and measured the two-norm relative error for the vector with components $\tilde{\mathbf{y}} = [y_3, y_4, y_5, y_6]$ since at the final time y_1 and y_2 vanish. The reference solution is obtained numerically using the third-order method and a sufficiently small time step. Figure 5 (right panel) shows the results obtained where we can observe the order of convergence of each method for this non-autonomous problem. Figure 5 (left panel) shows the error in the preservation of the quantities $I_1 = \mathbf{w}_1^\top \mathbf{y}(t_f)$ (curves with circles) and $I_2 = \mathbf{w}_2^\top \mathbf{y}(t_f)$ (curves with stars). Remarkably, the error

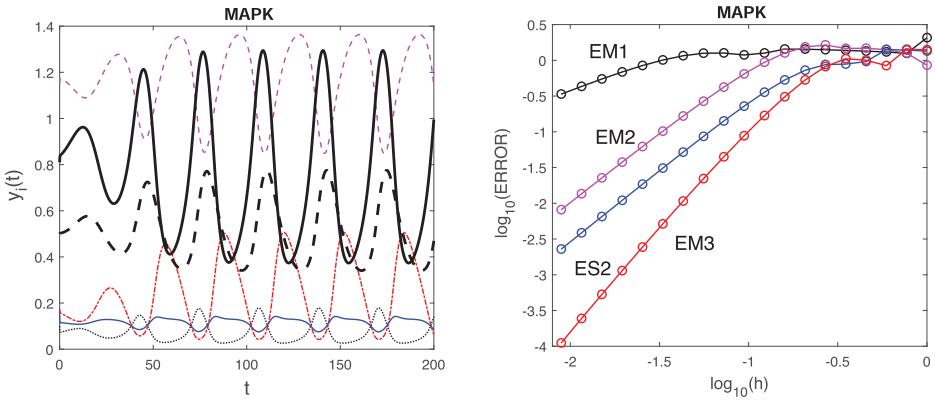

FIGURE 6. *Left*: solution of the MAPK cascade, given in (2.6), and *right*: two-norm relative error in the vector solution at the final time *versus* the time step in double logarithmic scale.

committed for I_1 is orders of magnitude smaller than the error in the actual solution, as seen the left panel in Fig. 5!

We observe that MP2, as in the previous examples, provide slightly worse results than EM2 when accurate results are desired, but the error considerably grows for large time steps (the approximation to the exponential in this case is not accurate). Surprisingly, it provides more accurate results in the exact preservation of I_2 and for all time steps positivity was preserved even if this property was not guaranteed for the method MP2 since A is not graph Laplacian. Were one to prove that the matrix A has no eigenvalues with positive real part then $I - tA$ would be an M-matrix and positivity would be guaranteed, and this deserves further investigation.

5.4. Example 4: The MAPK cascade

Finally, we consider the model of [23] (Tab. 3, Fig. 3, Eqs. (12)–(17)), which is closely related to the MAPK cascade, given in (2.6) with values therein for the parameters and initial conditions. The solution for each component is shown in the left panel of Figure 6 for the time interval $t \in [0, 200]$ (the initial conditions clearly identify each curve) where we observe that, after a transition period, the solution turns nearly periodic. Next, we have numerically solved the problem for $\alpha = 1$ using the new exponential methods using different values of the time step and measured the two-norm relative error in the vector solution at the final time. The right panel of Figure 6 shows the results obtained.

6. CONCLUSIONS

Preservation of inequalities is considerably more challenging than the recovery of "equality invariants" under discretisation. Thus, while numerous geometric numerical integration algorithms present us with a wide range of highly effective means to recover conservation laws, often of crucial importance in applications, this is not the case with inequalities and, of particular importance to us, nonnegativity of solutions. The importance of the latter in applications is clear – the number of chemical species cannot be negative, temperature cannot be less that $0\,^{\circ}$K, the number of infected people $I(t)$ cannot (sadly) be negative – yet we cannot be assured that computed ODE solutions remain nonnegative in this setting unless the order is unacceptably low. As aforementioned, the subject has already received significant attention and led to the development of Patankar-type methods [5,14,34,47]. In this paper we have developed a framework allowing us to use higher-order methods in this setting. While this framework is by no means final and many challenges remain, it represents in our view a useful contribution to a different kind of geometric numerical integration, one dealing with preservation of inequalities.

An outstanding challenge is to approximate the exponential of matrices by diagonal Padé approximants or by other means (*e.g.* Krylov-subspace methods) to reduce the cost of the algorithms for large ODE systems while still preserving positivity. Another is to explore the scope of methods, like the commutator-free Magnus integrators (3.18), which almost preserve positivity and formulate "almost preservation" in more precise terms.

Yet, perhaps the most interesting challenge is to explore the surprising success of "almost positivity-preserving" methods, *e.g.* the fourth-order commutator-free Magnus method, in the examples in this paper. Recall that classical ODE solvers that preserve positivity are restricted to order one [10], while in this paper we have introduced second-order positivity-preserving methods in the non-classical class of Magnus integrators, and other high-order methods have been introduced elsewhere, in particular modified Patankar methods. It is natural to formulate the conjecture that this is as much as can be done within the realm of such methods, but equally fascinating is the remarkable almost-preservation of positivity or mass (at any rate in the examples of this paper) by some higher-order methods. For example, Figure 5 (left) is concerned with two conservation laws in a stratospheric reaction: one is preserved correctly, up to roundoff error, while the other is preserved to much higher accuracy than the error committed (cf. Fig. 5 right) in the solution itself. We look forward to an explanation.

For the reader who simply wants a good method without studying all the details, we recommend to try the new splitting method we propose in equation (3.1), which we call 'ES2': it is simple, fast, easy to implement, second order accurate, and of course preserves positivity.

Acknowledgements. The authors thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme "Geometry, compatibility and structure preservation in computational differential equations" when work on this paper was undertaken. This work was supported by EPSRC grant EP/R014604/1. S.B. has been supported by project PID2019-104927GB-C21 (AEI/FEDER, UE).

References

[1] A. Alvermann and H. Fehske, High-order commutator-free exponential time-propagation of driven quantum systems. *J. Comput. Phys.* **230** (2011) 5930–5956.

[2] A.I. Ávila, S. Kopecz and A. Meister, A comprehensive theory on generalized BBKS schemes. *Appl. Numer. Math.* **157** (2020) 19–37.

[3] M. Beck and M.J. Gander, On the positivity of Poisson integrators for the Lotka–Volterra equations. *BIT Numer. Math.* **55** (2015) 319–340.

[4] A. Berman and R.J. Plemmons, Nonnegative Matrices in the Mathematical Sciences. *Computer Science and Applied Mathematics*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London (1979).

[5] E. Bertolazzi, Positive and conservative schemes for mass action kinetics. *Comput. Math. App.* **32** (1996) 29–43.

[6] S. Blanes, On the construction of symmetric second order methods for ODEs. *Appl. Math. Lett.* **98** (2019) 41–48.

[7] S. Blanes and F. Casas, A Concise Introduction to Geometric Numerical Integration. *Monographs and Research Notes in Mathematics*. CRC Press, Boca Raton, FL (2016).

[8] S. Blanes, F. Casas, J.A. Oteo and J. Ros, The Magnus expansion and some of its applications. *Phys. Rep.* **470** (2009) 151–238.

[9] S. Blanes, F. Casas and M. Thalhammer, High-order commutator-free quasi-Magnus exponential integrators for non-autonomous linear evolution equations. *Comput. Phys. Commun.* **220** (2017) 243–262.

[10] C. Bolley and M. Crouzeix, Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques. *RAIRO: Anal. Numér.* **12** (1978) 237–245.

[11] N. Broekhuizen, G.J. Rickard, J. Bruggeman and A. Meister, An improved and generalized second order, unconditionally positive, mass conserving integration scheme for biochemical systems. *Appl. Numer. Math.* **58** (2008) 319–340.

[12] J. Bruggeman, H. Burchard, B.W. Kooi and B. Sommeijer, A second-order, unconditionally positive, mass-conserving integration scheme for biochemical systems. *Appl. Numer. Math.* **57** (2007) 36–58.

[13] H. Burchard, E. Deleersnijder and A. Meister, Application of modified Patankar schemes to stiff biogeochemical models for the water column. *Ocean Dyn.* **55** (2005) 326–337.

[14] H. Burchard, E. Deleersnijder and A. Meister, A high-order conservative Patankar-type discretisation for stiff systems of production-destruction equations. *Appl. Numer. Math.* **47** (2003) 1–30.

[15] G. Colonna, On the relevance of superelastic collisions in argon and nitrogen discharges. *Plasma Sources Sci. Technol.* **29** (2020) 065008.

[16] F. Diele and C. Marangi, Geometric numerical integration in ecological modelling. *Mathematics* **8** (2020) 25.

[17] B.A. Earnshaw and J.P. Keener, Global asymptotic stability of solutions of nonautonomous master equations. *SIAM J. Appl. Dyn. Syst.* **9** (2010) 220–237.

[18] B.A. Earnshaw and J.P. Keener, Invariant manifolds of binomial-like nonautonomous master equations. *SIAM J. Appl. Dyn. Syst.* **9** (2010) 568–588.

[19] L. Edsberg, Integration package for chemical kinetics. In: Stiff Differential Systems (Proc. Internat. Sympos., Wildbad, 1973), edited by R.A. Willoughby. Springer, Boston, MA (1974) 81–95.

[20] L. Formaggia and A. Scotti, Positivity and conservation properties of some integration schemes for mass action kinetics. *SIAM J. Numer. Anal.* **49** (2011) 1267–1288.

[21] G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo and M Colaneri, Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat. Med.* **26** (2020) 855–860.

[22] J. Gunawardena, A linear framework for time-scale separation in nonlinear biochemical systems. *PloS One* **7** (2012) e36321.

[23] O. Hadač, F. Muzika, V. Nevoral, M. Přibyl and I. Schreiber, Minimal oscillating subnetwork in the Huang–Ferrell model of the MAPK cascade. *Plos One* **12** (2017) e0178457.

[24] E. Hairer and G. Wanner, Solving Ordinary Differential Equations. II. Stiff and Differential-Algebraic Problems, 2nd revised edition, paperback. Vol. 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin (2010).

[25] E. Hairer, C. Lubich and G. Wanner, Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations, Reprint of the second (2006) edition. Vol. 31 of *Springer Series in Computational Mathematics*. Springer, Heidelberg (2010).

[26] E. Hansen, F. Kramer and A. Ostermann, A second-order positivity preserving scheme for semilinear parabolic problems. *Appl. Numer. Math.* **62** (2012) 1428–1435.

[27] A. Hellander, J. Klosa, P. Lötstedt and S. MacNamara, Robustness analysis of spatiotemporal models in the presence of extrinsic fluctuations. *SIAM J. Appl. Math.* **77** (2017) 1157–1183.

[28] M. Hochbruck, A. Ostermann and J. Schweitzer, Exponential Rosenbrock-type methods. *SIAM J. Numer. Anal.* **47** (2008/2009) 786–803.

[29] A. Iserles, A First Course in the Numerical Analysis of Differential Equations, 2nd edition. *Cambridge Texts in Applied Mathematics*. Cambridge University Press, Cambridge (2009).

[30] A. Iserles and S. MacNamara, Applications of Magnus expansions and pseudospectra to Markov processes. *Eur. J. Appl. Maths* **30** (2019) 400–425.

[31] A. Iserles and S.P. Nørsett, On the solution of linear differential equations in Lie groups. *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.* **357** (1999) 983–1019.

[32] A. Iserles, H.Z. Munthe-Kaas, S.P. Nørsett and A. Zanna, Lie-Group Methods. In: Acta numerica, 2000. Vol. 9 of *Acta Numer.* Cambridge Univ. Press, Cambridge (2000) 215–365.

[33] W.O. Kermack and A.G. McKendrick, A contribution to the mathematical theory of epidemics. *Proc. R. Soc. London* **115** (1927) 700–721.

[34] S. Kopecz and A. Meister, On order conditions for modified Patankar–Runge–Kutta schemes. *Appl. Numer. Math.* **123** (2018) 159–179.

[35] S. Kopecz and A. Meister, Unconditionally positive and conservative third order modified Patankar–Runge–Kutta discretizations of production-destruction systems. *BIT Numer. Math.* **58** (2018) 691–728.

[36] S.C. Leite and R.J. Williams, A constrained Langevin approximation for chemical reaction networks. *Ann. Appl. Prob.* **29** (2019) 1541–1608.

[37] S. MacNamara, Cauchy integrals for computational solutions of master equations. *ANZIAM J.* **56** (2015) 32–51.

[38] S. MacNamara, A.M. Bersani, K. Burrage and R.B. Sidje, Stochastic chemical kinetics and the total quasi-steady-state assumption: application to the stochastic simulation algorithm and chemical master equation. *J. Chem. Phys.* **129** (2008) 095105.

[39] S. MacNamara, K. Burrage and R.B. Sidje, Multiscale modeling of chemical kinetics via the master equation. *Multiscale Modeling Simul.* **6** (2008) 1146–1168.

[40] S. MacNamara, B. Henry and W. Mclean, Fractional Euler limits and their applications. *SIAM J. Appl. Math.* **77** (2017) 447–469.

[41] S. MacNamara, S. Blanes and A. Iserles, Simulation of bimolecular reactions: numerical challenges with the graph Laplacian. *ANZIAM J.* **61** (2020) C59–C74.

[42] W. Magnus, On the exponential solution of differential equations for a linear operator. *Comm. Pure Appl. Math.* **7** (1954) 649–673.

[43] P.K. Maini, T.E. Woolley, R.E. Baker, E.A. Gaffney and S.S. Lee, Turing's model for biological pattern formation and the robustness problem. *J. R. Soc. Interface Focus* **2** (2012) 487–496.

[44] A. Martiradonna, G. Colonna and F. Diele, GeCo: Geometric Conservative nonstandard schemes for biochemical systems. *Appl. Numer. Math.* **155** (2020s) 38–57.

[45] I. Mirzaev and J. Gunawardena, Laplacian dynamics on general graphs. *Bull. Math. Biol.* **75** (2013) 2118–2149.

[46] P. Öffner and D. Torlo, Arbitrary high-order, conservative and positivity preserving Patankar-type deferred correction schemes. *Appl. Numer. Math.* **153** (2020) 15–34.

[47] S.V. Patankar, Numerical Heat Transfer and Fluid Flow. *Series in Computational Methods in Mechanics and Thermal Sciences*. Hemisphere Pub. Corp., New York (1980).

[48] L. Qiao, R.B. Nachbar, I.G. Kevrekidis and S.Y. Shvartsman, Bistability and oscillations in the Huang-Ferrell model of MAPK signaling. *PLoS Comput. Biol.* **3** (2007) e184.

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

[50] J.M. Sanz-Serna and M.P. Calvo, Numerical Hamiltonian Problems. Vol. 7 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London (1994).
[51] M.J. Shon and A.E. Cohen, Mass action at the single-molecule level. *J. Am. Chem. Soc.* **134** (2012) 14618–14623.
[52] R.L. Speth, W.H. Green, S. MacNamara and G. Strang, Balanced splitting and rebalanced splitting. *SIAM J. Numer. Anal.* **51** (2013) 3084–3105.
[53] C. Timm, Random transition-rate matrices for the master equation. *Phys. Rev. E* **80** (2009) 021140.