Numerically Stable Formulas for a Material Point Based Explicit Exponential Integrator

P. Nadukandi

Publication CIMNE Nº-408, July 2014

Numerically Stable Formulas for a Material Point Based Explicit Exponential Integrator

P. Nadukandi

Publication CIMNE Nº-408, July 2014

International Center for Numerical Methods in Engineering
Gran Capitán s/n, 08034 Barcelona, Spain

Numerically stable formulas for a material point based explicit exponential integrator

Prashanth Nadukandi
Centre Internacional de Mètodos Numèrics en Enginyeria (CIMNE),
Edifici C1, Gran Capitan s/n, 08034 Barcelona, Spain.
Email: npras@cimne.upc.edu, Tel: +34934010795, Fax: +34934016517

July 9, 2014

Abstract

We present numerically stable formulas for the analytical solution in the closed form of the so-called X-IVAS scheme in 3D. The X-IVAS scheme is a material point based explicit exponential integrator. An intermediate step in the X-IVAS scheme is the solution of tangent curves for piecewise linear vector fields defined on simplicial meshes. This is what we refer to as particle tracing of streamlines and independent formulas for the same can be easily distilled from the ones presented for the X-IVAS scheme. The formulas involve functions of matrices which are defined using the corresponding Newton interpolating polynomial. The evaluation of these formulas is stable, i.e. a certain number of significant digits in the computed values are guaranteed to be exact. Using the double-precision floating-point arithmetic specified by the IEEE 754 standard, we obtain at least 10 significant decimal digits in the worst case scenarios. These scenarios involve fourth-order divided differences of the exponential function. Additionally, an optimal series approximation of divided differences is presented which is an essential part of the exposition.

1 Introduction

The particle finite element method (PFEM) [1, 2] is a versatile particle based numerical method. It is shown to successfully simulate a wide variety of engineering problems [3, 4, 5, 6]. A recent development within the framework of the PFEM is the X-IVAS (eXplicit Integration along the Velocity and Acceleration Streamlines) scheme [7, 8]. The development of the X-IVAS scheme is motivated in the quest to attain enhanced accuracy, stability and efficiency in the numerical simulations. Efficiency is sought by formulating an explicit method that admits large time steps (algorithmic efficiency) and which can make the best out of the available computational capacity (resource efficiency) via parallel computations on multicore CPUs, GPGPUs etc. Accuracy and stability are sought based on the notion that the streamlines are a good approximation to the pathlines and time integration of position and velocity along the streamlines yields a better and a more stable approximation than doing so via standard finite difference time integrators.

In this article we focus on the analytical solution of the X-IVAS scheme and present numerically stable formulas for the same in 2D and 3D. The analytical solution of the X-IVAS scheme in 2D was given by Idelsohn et.al. [7]. The solution in 3D was omitted therein pointing out that the extension to 3D is straightforward. The functions of matrices appearing therein are defined using the Jordan canonical form of the same. In 2D, the analytical procedure to express matrices in the Jordan form is straightforward. Unfortunately in 3D (and for matrices of larger dimensions) the analytical procedure to arrive at the Jordan form is arduous as repeated eigenvalues with different Jordan blocks might exist. As this procedure is not described in [7], we infer that the use of a numerical library, e.g. LAPACK, is suggested for this purpose. As the Jordan structure of a 3D matrix involves multiple cases, we beg to differ with [7] that it is not trivial to derive and implement (code) the approach suggested therein to evaluate the analytical solution of the X-IVAS scheme in 3D.

A similar procedure to compute analytically the streamlines on a linear tetrahedra was presented by Diachin and Herzog [9]. Therein, the functions of matrices were computed using a procedure based on matrix decomposition methods. The singular value decomposition was used to determine the matrix rank which in turn was used to classify the calculation procedure into four cases in 3D. Using the matrix Schur decomposition the functions of matrices were transformed to equivalent functions of upper triangular matrices. The evaluation of the latter was done following a recursive relation proposed by Parlett [10].

Despite being analytical in nature, we do not consider the calculation procedures in [9, 7] as formulas as they are not expressed in the closed form. Formulas for particle tracing of streamlines in 3D were first presented by Nielson and Jung [11]. To be precise, formulas were given in 2D and 3D for the analytical solution of tangent curves for linearly varying vector fields over tetrahedral domains. These formulas are algebraically elegant and are expressed in the closed form. We infer from the algebraic structure of these formulas that the functions of matrices appearing therein are defined using the corresponding Lagrange interpolating polynomial. In 2D, the analytical solution of the tangent curves are classified into five cases which depend on the eigenvalues of the system matrix. In 3D, nine cases are contemplated for the same.

Unfortunately, using finite precision arithmetic the *as is* evaluation of the expressions that appear in the analytical solution procedure of [7] and in the formulas of [11] are only conditionally stable. Numerical instabilities occur in a neighbourhood of the removable singularities present in these expressions. These instabilities are caused by cancellations in finite precision arithmetic leading to a *gradual* loss of significant digits. This introduces evaluation errors that gradually build up as we approach the points of removable singularities. To make matters worse, these instabilities are not evident unless of course when they are catastrophic. This is best explained using an example. Consider the following formula defined in a piecewise manner,

$$\varphi(\lambda, t) := \begin{cases} \dfrac{\mathrm{e}^{\lambda t} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ t & \text{if } \lambda = 0. \end{cases} \tag{1}$$

It is clear that when $\lambda = 0$ we have a singularity in the expression $(\exp(\lambda t) - 1)/\lambda$. This singularity is removed by defining $\varphi(0, t) = t$, i.e. by assigning the value $(\exp(\lambda t) - 1)/\lambda$ takes in the limit $\lambda \to 0$. It is in this piecewise manner that all the cases of removable singularities present in the formulas were defined in [7, 11]. Following the examples presented in [11], demonstrating the formula evaluation for $\lambda = 0$ (the singular case) and for some $\lambda = O(1)$ (the case $\lambda \neq 0$) is not sufficient to claim that these formulas are stable when working with finite precision. In fact, this is a typical example used to demonstrate numerical instabilities in formula evaluations [12, 13].

In practice, the values which λ takes could be a result of a previous computation which might include a negligible roundoff error. So the singular case $\lambda = 0$ might actually occur as $\lambda = 0 \pm \varepsilon$ where $0 < \varepsilon \ll 1$. It follows that the case $\lambda \neq 0$ will be used in the evaluation of $\varphi(0 \pm \varepsilon, t)$ were we might loose all the significant digits due to cancellation. It is not unusual to find a relative evaluation error greater than 100% whenever λ takes a value in a sufficiently close neighbourhood of 0. Errors of this magnitude render the evaluation meaningless.

The conditional stability issue also extends to analytical solution procedure presented in [9]. Although the matrix decompositions are generally not unique, they are robust/stable with respect to rounding errors. However, the recursive algorithm used to compute the exponential function of block upper triangular matrices breaks down in certain situations. Moreover implementations in finite precision arithmetic can be expected to give inaccurate results in other particular situations, cf. [10, page 199]. The message is clear: irrespective of the choice of the solution procedure, issues related to numerical instability exist and they need to be addressed.

Further, once such instabilities are identified, it is often not trivial to localize the terms in these formulas that participate to obtain a finite limit at the removable singularities. Identifying such terms is crucial to control numerical instabilities and bound the loss of significant digits. We discuss these issues here and present algebraically equivalent yet numerically stable formulas for particle tracing of streamlines and the X-IVAS scheme in both 2D and 3D.

This paper is organized as follows. The convention used in the kinematic description of the flow is briefly described in Section 2. Section 3 describes the X-IVAS scheme and its analytical solution with sufficient detail. Although we could have referred to the original paper [7] for the same, we redo this part for the sake of completeness and to use a uniform convention throughout the exposition. As the analytical solution of the X-IVAS scheme involves functions of matrices, we dedicate Section 4 to this topic. Here we explain why we choose to define functions of matrices using its Newton interpolation polynomial. Using this definition we present formulas for exponential functions of 2×2 and 3×3 matrices in Section 4.2. We briefly summarize the formulas for the eigenvalues of 2×2 and 3×3 matrices in Section 4.3. In Section 5 we discuss the stable evaluation of formulas using finite precision arithmetic. The issue with unstable evaluations near removable singularities present in the formulas are explained using an example. In Section 5.2 we briefly summarize how double-precision floating-point numbers are stored as per the IEEE 754 standard. The optimal series approximation of divided differences are presented in Section 5.3. In Section 5.4, the stable piecewise evaluation technique used in this work is explained in detail via an example. Following this technique we present in Section 5.5 the stable piecewise definition of all the expressions in the solution of the X-IVAS scheme which otherwise yield unstable evaluations. A proposal for the implementation of the stable formulas is summarized in Section 6 and some examples are presented in Section 7.

2 Preliminaries

In this section we describe briefly the convention used in the description of the flow. The independent variables in Lagrangian kinematics are (χ, t), where χ represents a label to identify particles (material points) and t represents the time elapsed after labeling. The primary dependent variable is the fluid particle trajectory denoted as $\boldsymbol{X}(\chi, t)$. The initial particle positions denoted by $\boldsymbol{X}^0 := \boldsymbol{X}(\chi, 0)$ are assumed to be given. A natural choice for the label χ is the ordered triple $\boldsymbol{X}^0$. The Lagrangian velocity and acceleration, denoted as

$\dot{\boldsymbol{X}}(\chi,t)$ and $\ddot{\boldsymbol{X}}(\chi,t)$, respectively are defined as follows.

$$\dot{\boldsymbol{X}}(\chi,t) := \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X}(\chi,t), \quad \ddot{\boldsymbol{X}}(\chi,t) := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{X}(\chi,t) \tag{2}$$

On the other hand, the independent variables in Eulerian kinematics are $(\boldsymbol{x},t)$. Here $\boldsymbol{x}$ denotes the spatial coordinate. The primary dependent variable is the fluid velocity $\boldsymbol{u}(\boldsymbol{x},t)$. The so-called fundamental principle of kinematics [14] states that the velocity $\boldsymbol{u}(\boldsymbol{x},t)$ and acceleration $\boldsymbol{a}(\boldsymbol{x},t)$ at a given time t and fixed position $\boldsymbol{x}$ (Eulerian description) is equal to the velocity $\dot{\boldsymbol{X}}(\chi,t)$ and acceleration $\ddot{\boldsymbol{X}}(\chi,t)$ of a particle that is present at that position and at that instant (Lagrangian description). Thus,

$$\boldsymbol{u}(\boldsymbol{x},t) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X}(\chi,t)\right|_{\boldsymbol{X}(\chi,t)=\boldsymbol{x}}, \quad \boldsymbol{a}(\boldsymbol{x},t) = \left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{X}(\chi,t)\right|_{\boldsymbol{X}(\chi,t)=\boldsymbol{x}} \tag{3}$$

As a corollary we have the following exact but implicit equations of particle motion.

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X}(\chi,t) = \boldsymbol{u}(\boldsymbol{X}(\chi,t),t), \quad \frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{X}(\chi,t) = \frac{\mathrm{d}}{\mathrm{d}t}\dot{\boldsymbol{X}}(\chi,t) = \boldsymbol{a}(\boldsymbol{X}(\chi,t),t) \tag{4}$$

3 The X-IVAS scheme

3.1 Introduction

The X-IVAS scheme is a material point based explicit exponential integrator; cf. [15, 16, 17] for an overview of exponential integrators. The idea is to perform the time integration not along the pathlines which are yet unknown but along the streamlines obtained for the latest known configuration. The latter choice makes the method explicit and permits one to use very large time steps (e.g. 20-25 times the classical CFL limit [7]) which reduce significantly the computational time. Following this idea the equations of motion are given by,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X}(\chi,t) = \boldsymbol{u}(\boldsymbol{X}(\chi,t),t^n), \quad \frac{\mathrm{d}}{\mathrm{d}t}\dot{\boldsymbol{X}}(\chi,t) = \boldsymbol{a}(\boldsymbol{X}(\chi,t),t^n) \tag{5}$$

We remark that the X-IVAS scheme is just based on the above idea and that Eq. (5) will be subjected to further simplifications before we fix the approximate equations of motion. This is because the data corresponding to the dependent variables is stored with the particles which form a sufficiently large yet finite set. It implies that the data at any given time is available as discrete samples at the spatial locations occupied by the particles. Data interpolation is inevitable to have spatially continuous vector fields and to solve for the particle motion. Hence in the equations of motion $\boldsymbol{u}(\boldsymbol{x},t^n)$ and $\boldsymbol{a}(\boldsymbol{x},t^n)$, which are unknown for an arbitrary $\boldsymbol{x}$ are replaced by the interpolated counterparts $\boldsymbol{u}^h(\boldsymbol{x},t^n)$ and $\boldsymbol{a}^h(\boldsymbol{x},t^n)$, respectively. The superscript h represents the discretization size associated to the interpolation. It follows that the trajectory obtained from these interpolated vector fields needs to be represented as $\boldsymbol{X}^h(\chi,t)$.

In the following we describe the X-IVAS scheme to integrate the equations of particle motion from time t^n to t^{n+1} as a four step process.

Step 1: Projection. This step involves the projection of vector fields stored with the particles onto a simplicial mesh. Consider a simplicial mesh over the problem domain and a set of characteristic domains corresponding to every node of the mesh. Let $\mathcal{P}^i$ be an operator that projects data onto a mesh node with index i from a set of sample points in the corresponding characteristic domain. Using this projection operator we calculate the velocity $\bar{\boldsymbol{u}}^i(t^n)$ and acceleration $\bar{\boldsymbol{a}}^i(t^n)$ vector fields at the mesh nodes as follows.

$$\bar{\boldsymbol{u}}^i(t^n) := \mathcal{P}^i[\dot{\boldsymbol{X}}(\chi, t^n)], \quad \bar{\boldsymbol{a}}^i(t^n) := \mathcal{P}^i[\ddot{\boldsymbol{X}}(\chi, t^n)] \tag{6}$$

This projection step is unnecessary when $t^n = 0$ where one can obtain $\bar{\boldsymbol{u}}^i(0)$ and $\bar{\boldsymbol{a}}^i(0)$ directly from the prescribed initial conditions.

Step 2: Interpolation. In this step we do a piecewise linear interpolation of vector fields projected onto the mesh nodes. Using the velocity $\bar{\boldsymbol{u}}^i(t^n)$ and acceleration $\bar{\boldsymbol{a}}^i(t^n)$ vector fields at the mesh nodes we construct a piecewise linear interpolation of these vector fields as follows.

$$\boldsymbol{u}^h(\boldsymbol{x}, t^n) := \mathrm{N}^i(\boldsymbol{x})\bar{\boldsymbol{u}}^i(t^n), \quad \boldsymbol{a}^h(\boldsymbol{x}, t^n) := \mathrm{N}^i(\boldsymbol{x})\bar{\boldsymbol{a}}^i(t^n) \tag{7}$$

In the above equation $\mathrm{N}^i(\boldsymbol{x})$ represents the piecewise linear shape function corresponding to the node i. Let $\boldsymbol{x}^j$ denote the spatial coordinate of node j, $\langle\ \rangle_j$ denote the average operator over the index j and δ^{ij} denote the Kronecker delta. For a given simplex, we can express $\mathrm{N}^i(\boldsymbol{x})$ in terms of its gradient $\boldsymbol{\nabla}\,\mathrm{N}^i$ (which is constant within the simplex) and the spatial coordinate $\boldsymbol{x}$ as follows.

$$\mathrm{N}^i(\boldsymbol{x}) := \boldsymbol{\nabla}\,\mathrm{N}^i \cdot (\boldsymbol{x} - \langle \boldsymbol{x}^j \rangle_j) + \frac{1^i}{\delta^{kk}} \tag{8}$$

Using the above equation $\boldsymbol{u}^h(\boldsymbol{x}, t^n)$ and $\boldsymbol{a}^h(\boldsymbol{x}, t^n)$ can be expressed within each simplex as follows.

$$\boldsymbol{u}^h(\boldsymbol{x}, t^n) = [\bar{\boldsymbol{u}}^i(t^n) \otimes \boldsymbol{\nabla}\,\mathrm{N}^i] \cdot (\boldsymbol{x} - \langle \boldsymbol{x}^j \rangle_j) + \langle \bar{\boldsymbol{u}}^j(t^n) \rangle_j = \mathbf{A}^n \cdot \boldsymbol{x} + \mathbf{b}^n \tag{9}$$

$$\boldsymbol{a}^h(\boldsymbol{x}, t^n) = [\bar{\boldsymbol{a}}^i(t^n) \otimes \boldsymbol{\nabla}\,\mathrm{N}^i] \cdot (\boldsymbol{x} - \langle \boldsymbol{x}^j \rangle_j) + \langle \bar{\boldsymbol{a}}^j(t^n) \rangle_j = \mathbf{C}^n \cdot \boldsymbol{x} + \mathbf{d}^n \tag{10}$$

Here $\otimes$ denotes the tensor product. Further, $\mathbf{A}^n, \mathbf{b}^n, \mathbf{C}^n$ and $\mathbf{d}^n$ are constant tensors evaluated for each simplex at time t^n and are defined as follows.

$$\mathbf{A}^n := [\bar{\boldsymbol{u}}^i(t^n) \otimes \boldsymbol{\nabla}\,\mathrm{N}^i], \quad \mathbf{b}^n := \langle \bar{\boldsymbol{u}}^i(t^n) \rangle_i - \mathbf{A}^n \cdot \langle \boldsymbol{x}^i \rangle_i \tag{11}$$

$$\mathbf{C}^n := [\bar{\boldsymbol{a}}^i(t^n) \otimes \boldsymbol{\nabla}\,\mathrm{N}^i], \quad \mathbf{d}^n := \langle \bar{\boldsymbol{a}}^i(t^n) \rangle_i - \mathbf{C}^n \cdot \langle \boldsymbol{x}^i \rangle_i \tag{12}$$

Step 3: Integration. Here we describe the time integration of the approximate equations of particle motion. The approximate equations of motion for the particles in the X-IVAS scheme can be written as follows.

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X}^h(\chi, t) = \boldsymbol{u}^h(\boldsymbol{X}^h(\chi, t), t^n), \quad \frac{\mathrm{d}}{\mathrm{d}t}\dot{\boldsymbol{X}}^h(\chi, t) = \boldsymbol{a}^h(\boldsymbol{X}^h(\chi, t), t^n) \tag{13}$$

Recall that the above equations are expressed in a piecewise manner as both $\boldsymbol{u}^h$ and $\boldsymbol{a}^h$ are defined in this manner. To be precise, within each simplex the particle motion is driven by the following equations which vary from one simplex to another.

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X}^h(\chi, t) = \mathbf{A}^n \cdot \boldsymbol{X}^h(\chi, t) + \mathbf{b}^n, \quad \frac{\mathrm{d}}{\mathrm{d}t}\dot{\boldsymbol{X}}^h(\chi, t) = \mathbf{C}^n \cdot \boldsymbol{X}^h(\chi, t) + \mathbf{d}^n \tag{14}$$

Likewise, the time integration of these equations should also be done in a piecewise manner. In other words, if a particle tends to exit the current simplex prior to the end of the time step, its subsequent motion is driven by the equations written for the simplex in which it tends to enter and so forth until the end of the time step.

Further, certain relationships that existed between the dependent variables in the exact equations of motions no longer hold for the corresponding variables in the approximate equations of motion. That is,

$$\dot{\boldsymbol{X}}^h(\chi,t) \neq \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X}^h(\chi,t), \quad \ddot{\boldsymbol{X}}^h(\chi,t) := \frac{\mathrm{d}}{\mathrm{d}t}\dot{\boldsymbol{X}}^h(\chi,t) \neq \frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{X}(\chi,t), \tag{15}$$

Nevertheless, the X-IVAS scheme is consistent in the sense that these relations are recovered as the discretization size $h \to 0$. The analytical solution to the pair of equations given in Eq. (14) can be written within each simplex as follows.

$$\boldsymbol{X}^h(\chi,t) = \mathrm{e}^{(t-t^n)\mathbf{A}^n} \cdot \boldsymbol{X}^h(\chi,t^n) + \left[\int_{t^n}^{t} \mathrm{e}^{(t-\tau)\mathbf{A}^n}\,\mathrm{d}\tau\right] \cdot \mathbf{b}^n \tag{16}$$

$$\dot{\boldsymbol{X}}^h(\chi,t) = \dot{\boldsymbol{X}}^h(\chi,t^n) + \mathbf{C}^n \cdot \left[\int_{t^n}^{t} \boldsymbol{X}^h(\chi,\tau)\,\mathrm{d}\tau\right] + (t-t^n)\mathbf{d}^n \tag{17}$$

Note that the particle motion is restricted to the tangent curve of $\boldsymbol{u}^h(\boldsymbol{x},t^n)$ (i.e. the streamline) on which it was located at time t^n and is accelerated along this curve up to time t^{n+1}. Note that the solution for the particle velocity $\dot{\boldsymbol{X}}^h(\chi,t)$ is given as an integral of the particle position $\boldsymbol{X}^h(\chi,t)$. This integral is left here *as is* for compactness and its evaluated form will be given in the following section.

Step 4: Update. In this step we update the dependent variables at time t^{n+1} and repeat the process. At the end of the time step we obtain $\boldsymbol{X}^h(\chi,t^{n+1})$ and $\dot{\boldsymbol{X}}^h(\chi,t^{n+1})$ which are governed by the kinematics of the flow. The state of $\ddot{\boldsymbol{X}}^h(\chi,t^{n+1})$ is governed by the dynamics of the internal and the external force terms that appear in the momentum balance equation of the flow.

3.2 Remarks on the analytical solution

In this section we simplify the analytical solution given in Eq. (16) and Eq. (17) and identify relationships among the terms that appear therein, if any. Consider three matrices $\mathbf{P}, \mathbf{Q}$ and $\mathbf{R}$ which in turn are defined as functions of a given matrix $\mathbf{A}$ and a scalar τ as follows.

$$\mathbf{P}(\tau,\mathbf{A}) := \mathrm{e}^{\tau\mathbf{A}}, \quad \mathbf{Q}(\tau,\mathbf{A}) := \int_0^\tau \mathrm{e}^{\xi\mathbf{A}}\,\mathrm{d}\xi, \quad \mathbf{R}(\tau,\mathbf{A}) := \int_0^\tau \int_0^\eta \mathrm{e}^{\xi\mathbf{A}}\,\mathrm{d}\xi\,\mathrm{d}\eta \tag{18}$$

As it can be seen from the above equation, the considered matrices are exponential functions of the given matrix $\mathbf{A}$. The matrix $\mathbf{P}$ is usually called the propagator [18]. The following relationships can be identified between the matrices $\mathbf{P}$ and $\mathbf{Q}$.

$$\mathbf{Q}(\tau,\mathbf{A}) = \int_0^\tau \mathbf{P}(\xi,\mathbf{A})\,\mathrm{d}\xi = \left[\mathrm{e}^{\tau\mathbf{A}} - \mathbf{I}\right] \cdot \mathrm{inv}(\mathbf{A}) = \left[\mathbf{P}(\tau,\mathbf{A}) - \mathbf{I}\right] \cdot \mathrm{inv}(\mathbf{A}) \tag{19}$$

$$\Rightarrow \quad \mathbf{P}(\tau,\mathbf{A}) = \mathbf{Q}(\tau,\mathbf{A}) \cdot \mathbf{A} + \mathbf{I} \tag{20}$$

Likewise, the matrices $\mathbf{Q}$ and $\mathbf{R}$ satisfy the following relationships.

$$\mathbf{R}(\tau,\mathbf{A}) = \int_0^\tau \mathbf{Q}(\xi,\mathbf{A})\,\mathrm{d}\xi = \left[\left(\mathrm{e}^{\tau\mathbf{A}} - \mathbf{I}\right) \cdot \mathrm{inv}(\mathbf{A}) - \tau\mathbf{I}\right] \cdot \mathrm{inv}(\mathbf{A}) \tag{21}$$

$$\mathbf{R}(\tau,\mathbf{A}) = \left[\mathbf{Q}(\tau,\mathbf{A}) - \tau\mathbf{I}\right] \cdot \mathrm{inv}(\mathbf{A}), \quad \Rightarrow \quad \mathbf{Q}(\tau,\mathbf{A}) = \mathbf{R}(\tau,\mathbf{A}) \cdot \mathbf{A} + \tau\mathbf{I} \tag{22}$$

In the above equations inv($\mathbf{A}$) denotes the matrix inverse of $\mathbf{A}$. Further, the products involving inv($\mathbf{A}$) and $\mathbf{A}$ in these equations are commutative, i.e. the order in which they appear are irrelevant. Using these definitions, we can express the analytical solution of the equations of motion in the X-IVAS scheme as follows.

$$\boldsymbol{X}^h(\chi,t) = \mathbf{P}(t-t^n,\mathbf{A}^n)\cdot\boldsymbol{X}^h(\chi,t^n) + \mathbf{Q}(t-t^n,\mathbf{A}^n)\cdot\mathbf{b}^n \tag{23}$$

$$\dot{\boldsymbol{X}}^h(\chi,t) = \dot{\boldsymbol{X}}^h(\chi,t^n) + \mathbf{C}^n\cdot[\mathbf{Q}(t-t^n,\mathbf{A}^n)\cdot\boldsymbol{X}^h(\chi,t^n) + \mathbf{R}(t-t^n,\mathbf{A}^n)\cdot\mathbf{b}^n] + (t-t^n)\mathbf{d}^n \tag{24}$$

Recall that a nodal projection of the data carried by the particles onto the background mesh is done after every time step and the tensors $\mathbf{A}^n, \mathbf{b}^n, \mathbf{C}^n$ and $\mathbf{d}^n$ have to be recalculated for each element using the projected data. It follows that the matrices $\mathbf{P}, \mathbf{Q}$ and $\mathbf{R}$ also need to be recalculated for each element after every time step.

3.3 Piecewise integration using Newton linearisation

In this section we explain an approach to perform the piecewise integration of particle motions described earlier in the paragraph following Eq. (14). The task effectively reduces to finding the exit points of the particles on the simplex boundary. To find the exit points we need to solve the intersection of its trajectory with the simplex boundary. The procedure followed here to solve for the exit points was presented earlier by Kipfer et.al. [19].

Recall that for any simplex, the boundary consists of three straight edges in 2D and four flat faces in 3D. Let $\widetilde{\boldsymbol{x}}$ denote the centroid of an edge/face of a 2D/3D simplex and $\boldsymbol{\Lambda}$ denote the normal to the considered edge/face. The equation of the line/plane containing the edge/face is given by,

$$(\boldsymbol{x}-\widetilde{\boldsymbol{x}})\cdot\boldsymbol{\Lambda} = 0 \tag{25}$$

Substituting $\boldsymbol{x} = \boldsymbol{X}^h(\chi,t)$ in the above equation, we get the equation for the intersection of the particle trajectory with the considered edge/face of the simplex. Thus the intersection point satisfies,

$$\boldsymbol{\Lambda}\cdot\Big[\widetilde{\boldsymbol{x}} - \mathbf{P}(t-t^n,\mathbf{A}^n)\cdot\boldsymbol{X}^h(\chi,t^n) - \mathbf{Q}(t-t^n,\mathbf{A}^n)\cdot\mathbf{b}^n\Big] = 0 \tag{26}$$

$$\Rightarrow \boldsymbol{\Lambda}\cdot\Big[\widetilde{\boldsymbol{x}} - \boldsymbol{X}^h(\chi,t^n - \mathbf{Q}(t-t^n,\mathbf{A}^n)\cdot[\mathbf{A}^n\cdot\boldsymbol{X}^h(\chi,t^n)+\mathbf{b}^n]\Big] = 0 \tag{27}$$

In the above equation we solve for the time when the particle trajectory intersects the considered edge/face of the simplex. This is an implicit equation and we solve it using the Newton's linearisation method. Let δt^i and r^i denote the time increment and the residual at the i^{th} iteration, respectively. The iterative solution procedure can be described as follows.

$$r^i := \boldsymbol{\Lambda}\cdot\Big[\widetilde{\boldsymbol{x}} - \boldsymbol{X}^h(\chi,t^n - \mathbf{Q}(t^i-t^n,\mathbf{A}^n)\cdot[\mathbf{A}^n\cdot\boldsymbol{X}^h(\chi,t^n)+\mathbf{b}^n]\Big] \tag{28}$$

$$\delta t^i := t^{i+1} - t^i, \quad r^i + \frac{\mathrm{d}r^i}{\mathrm{d}t}\delta t^i = 0 \tag{29}$$

Simplifying the above equation, we can express the time increment δt^i as follows.

$$\delta t^i = \frac{\boldsymbol{\Lambda}\cdot[\widetilde{\boldsymbol{x}} - \boldsymbol{X}^h(\chi,t^n - \mathbf{Q}(t^i-t^n,\mathbf{A}^n)\cdot[\mathbf{A}^n\cdot\boldsymbol{X}^h(\chi,t^n)+\mathbf{b}^n]]}{\boldsymbol{\Lambda}\cdot[\mathbf{P}(t^i-t^n,\mathbf{A}^n)\cdot[\mathbf{A}^n\cdot\boldsymbol{X}^h(\chi,t^n)+\mathbf{b}^n]]} \tag{30}$$

Clearly any acceptable solution for an exit point should be such that $t \geq t^n$. This process should be done for all the edges/faces of the simplex and the smallest of all acceptable solutions should be chosen.

In the Newton's linearisation method the matrices $\mathbf{P}, \mathbf{Q}$ and $\mathbf{R}$ have to be evaluated at every iteration as the time increments need not be uniform. Should one decide to use *analytical* sub-stepping procedures to arrive at the exit point and if a constant *sub* time step is used for all the particles throughout the sub-stepping procedure, then we need to compute these matrices for each element just once. The incremental method for computing tangent curves [11] and the analytical time stepping algorithm called ANTS [9] are based on this idea.

4 Formulas for evaluating functions of matrices

4.1 Introduction

In this section we describe the extension of a scalar function $\mathrm{f}(\lambda)$ of a scalar argument λ to the case when the argument is a square matrix $\mathbf{A}$. Let the size of $\mathbf{A}$ be $n \times n$ and assume that $\mathrm{f}(\lambda)$ takes well-defined values (including values associated with derivatives where appropriate) at the eigenvalues of $\mathbf{A}$ denoted by the sequence $Z := \{\lambda_1,\, \lambda_2, \cdots,\, \lambda_n\}$. In particular we will focus on the exponential function $\exp(\lambda)$ as it plays a major role in the terms of the analytical solution described in the earlier section.

Functions of matrices can be defined in various yet equivalent ways; cf. [20] for a comprehensive presentation of the same. One such definition is a direct extension to matrix arguments of the Taylor series of $\mathrm{f}(\lambda)$. Specifically for $\exp(\lambda)$, the series converges absolutely and ultimately very quickly ($n!$ grows at a much faster rate than λ^n). However, representing $\mathrm{f}(\mathbf{A})$ by Taylor series is seldom done in practice as often a large number of terms must be added up until every subsequent term is smaller than the sum of the series.

Despite popular choices to represent $\mathrm{f}(\mathbf{A})$ either via the Jordan canonical form of $\mathbf{A}$ or by the Hermite interpolating[1] polynomial, in this work we will use the definition that represents $\mathrm{f}(\mathbf{A})$ by the Newton interpolating polynomial. The coefficients of the Newton interpolating polynomial have the algebraic structure of divided differences. On the one hand, it warns us about the *gradual* loss of significant digits in some limit cases[2] due to cancellations in floating point arithmetic; divided differences are known to suffer from cancellation errors near removable singularities. On the other hand, it paves way to systematically design procedures for the stable evaluation of $\mathrm{f}(\mathbf{A})$ near removable singularities. By stable evaluation we mean that a certain number[3] of significant digits in the computed values are guaranteed to be exact. Further, unlike the Hermite interpolating polynomial, the Newton interpolating polynomial is independent of the Jordan structure of $\mathbf{A}$ which makes it convenient to implement in a computer program.

We denote the k^{th}-order divided difference of $\mathrm{f}(\lambda)$ on the subsequence $Z_i^k := \{\lambda_i,\, \lambda_{i+1}, \cdots,$

[1] The use of the phrase *interpolating polynomial* does not imply that $\mathrm{f}(\mathbf{A})$ defined in this way is an approximation

[2] For instance, when some of the eigenvalues are zero or are sufficiently close to each other leading to nearly confluent divided differences.

[3] In the worst case scenarios, for the terms involving the second, third and fourth order exponential divided differences, this number is usually above $14, 12$ and 10 digits, respectively

$\lambda_{i+k}\}$ as $\mathrm{f}[\lambda_i;\ \lambda_{i+1};\cdots;\ \lambda_{i+k}]$ and define it using the following recurrence equations.

$$\mathrm{f}[\lambda_i] := \mathrm{f}(\lambda_i) \tag{31}$$

$$\mathrm{f}[\lambda_i;\lambda_{i+1};\cdots;\lambda_{i+k}] := \frac{\mathrm{f}[\lambda_{i+1};\lambda_{i+2};\cdots;\lambda_{i+k}] - \mathrm{f}[\lambda_i;\lambda_{i+1};\cdots;\lambda_{i+k-1}]}{\lambda_{i+k}-\lambda_i} \tag{32}$$

$$\lambda_i = \lambda_{i+1} = \cdots = \lambda_{i+k} \Rightarrow \mathrm{f}[\lambda_i;\lambda_{i+1};\cdots;\lambda_{i+k}] := \frac{1}{k!}\,\frac{\partial^k}{\partial\lambda^k}\,\mathrm{f}(\lambda)\bigg|_{\lambda=\lambda_i} \tag{33}$$

It is a well-known fact that the value of $\mathrm{f}[\lambda_i;\ \lambda_{i+1};\cdots;\ \lambda_{i+k}]$ does not depend on the order of λ_i, $\lambda_{i+1},\cdots$, λ_{i+k} in Z_i^k. In other words $\mathrm{f}[\lambda_i;\ \lambda_{i+1};\cdots;\ \lambda_{i+k}]$ is a symmetric function of λ_i, $\lambda_{i+1},\cdots$, λ_{i+k}. The Newton interpolating polynomial $\mathrm{p}(\lambda)$ for $\mathrm{f}(\lambda)$ is expressed as follows.

$$\mathrm{p}(\lambda) := \mathrm{f}[\lambda_1] + \sum_{k=1}^{n-1} \mathrm{f}[\lambda_1;\lambda_2;\cdots;\lambda_{1+k}](\lambda-\lambda_1)(\lambda-\lambda_2)\cdots(\lambda-\lambda_k) \tag{34}$$

$$\mathrm{p}(\lambda_i) = \mathrm{f}(\lambda_i), \quad \forall i \in \{1,2,\cdots,n\} \tag{35}$$

A fundamental result in matrix theory which can be used to evaluate $\mathrm{f}(\mathbf{A})$ is that,

$$\mathrm{f}(\mathbf{A}) = \mathrm{p}(\mathbf{A}) = \mathrm{f}[\lambda_1]\mathbf{I} + \sum_{k=1}^{n-1} \mathrm{f}[\lambda_1;\lambda_2;\cdots;\lambda_{1+k}](\mathbf{A}-\lambda_1\mathbf{I})(\mathbf{A}-\lambda_2\mathbf{I})\cdots(\mathbf{A}-\lambda_k\mathbf{I}) \tag{36}$$

It is possible to write the k^{th} divided difference $\mathrm{f}[\lambda_1;\ \lambda_2;\cdots;\ \lambda_{1+k}]$ as follows[4].

$$\mathrm{f}[\lambda_1;\lambda_2;\cdots;\lambda_{1+k}] := \sum_{i=1}^{1+k} \frac{\mathrm{f}(\lambda_i)}{\prod_{j\neq i}(\lambda_i-\lambda_j)}, \quad j \in \{1,2,\cdots,1+k\} \tag{37}$$

We do not prefer to make the above simplification to arrive at an elegant form as it does not avoid existing issues related to cancellation errors and makes matters worse by obscuring them[5].

Henceforth we restrict the exposition to the cases where $n \leq 3$. These are the only cases to be considered to express the solution of the X-IVAS scheme in closed form. For $n = 3$, we can evaluate $\mathrm{f}(\mathbf{A})$ as follows.

$$\mathrm{f}(\mathbf{A}) = \mathrm{f}(\lambda_1)\mathbf{I} + \mathrm{f}[\lambda_1;\lambda_2](\mathbf{A}-\lambda_1\mathbf{I}) + \mathrm{f}[\lambda_1;\lambda_2;\lambda_3](\mathbf{A}-\lambda_1\mathbf{I})(\mathbf{A}-\lambda_2\mathbf{I}) \tag{38}$$

Without loss of generality we assume that the eigenvalue λ_3 is a real number and the eigenvalues λ_1 and λ_2 might be complex numbers. Complex eigenvalues will always occur in conjugate pairs, i.e. $\{\lambda_1,\lambda_2\} = \{\lambda_\mathrm{c},\lambda_\mathrm{c}^*\}$. The subscript c indicates that it is a complex number and the superscript $*$ indicates that it is a complex conjugate.

Although Eq. (38) holds for all eigenvalues, this form is convenient to implement in a computer program when the eigenvalues are real numbers. In the case of complex eigenvalues Eq. (38) can be simplified to evaluate $\mathrm{f}(\mathbf{A})$ as follows.

$$\begin{aligned}\mathrm{f}(\mathbf{A}) = {} & \frac{\mathrm{Im}[f^*(\lambda_\mathrm{c})\lambda_\mathrm{c}]}{\mathrm{Im}[\lambda_\mathrm{c}]}\mathbf{I} + \frac{\mathrm{Im}[\mathrm{f}(\lambda_\mathrm{c})]}{\mathrm{Im}[\lambda_\mathrm{c}]}\mathbf{A} \\ & + \left[\frac{\mathrm{f}(\lambda_3)\,\mathrm{Im}(\lambda_\mathrm{c}) - \lambda_3\,\mathrm{Im}[\mathrm{f}(\lambda_\mathrm{c})] - \mathrm{Im}[f^*(\lambda_\mathrm{c})\lambda_\mathrm{c}]}{\mathrm{Im}[\lambda_\mathrm{c}]}\right]\left[\frac{\mathbf{A}^2 - 2\,\mathrm{Re}(\lambda_\mathrm{c})\mathbf{A} + |\lambda_\mathrm{c}|^2\mathbf{I}}{\lambda_3^2 - 2\,\mathrm{Re}(\lambda_\mathrm{c})\lambda_3 + |\lambda_\mathrm{c}|^2}\right]\end{aligned} \tag{39}$$

[4] Using this identity we can transform the formulas given here to the ones presented by Nielson and Jung [11]

[5] The example in Section 5.1 drives the point home

Here, the functions $\mathrm{Re}(\lambda_\mathrm{c})$ and $\mathrm{Im}(\lambda_\mathrm{c})$ return the real and imaginary parts of a complex argument λ_c, respectively.

4.2 Formulas for exponential functions of 2×2 and 3×3 matrices

In this section we consider the case when $\mathrm{f}(\lambda) := \exp(\tau\lambda)$ and write the expressions for the matrices $\mathbf{P}(\tau, \mathbf{A})$, $\mathbf{Q}(\tau, \mathbf{A})$ and $\mathbf{R}(\tau, \mathbf{A})$ which were defined earlier in Eq. (18). It is straightforward to verify the following results.

$$\int_0^\tau \mathrm{e}^{\xi\lambda}\,\mathrm{d}\xi = \frac{\mathrm{e}^{\tau\lambda} - 1}{\lambda} = \tau \exp[0; \tau\lambda] \tag{40}$$

$$\int_0^\tau \int_0^\eta \mathrm{e}^{\xi\lambda}\,\mathrm{d}\xi\,\mathrm{d}\eta = \frac{\mathrm{e}^{\tau\lambda} - 1 - \tau\lambda}{\lambda^2} = \tau^2 \exp[0; 0; \tau\lambda] \tag{41}$$

Following this line we define two auxiliary functions $\mathrm{q}(x)$ and $\mathrm{r}(x)$ which are divided differences of the exponential function.

$$\mathrm{q}(x) := \exp[0; x] = \begin{cases} \dfrac{\mathrm{e}^x - 1}{x} & \text{if } x \neq 0, \\ 1 & \text{if } x = 0. \end{cases} \tag{42}$$

$$\mathrm{r}(x) := \exp[0; 0; x] = \mathrm{q}[0; x] = \begin{cases} \dfrac{\mathrm{e}^x - 1 - x}{x^2} & \text{if } x \neq 0, \\ \dfrac{1}{2} & \text{if } x = 0. \end{cases} \tag{43}$$

Using these auxiliary functions we can express $\mathbf{P}(\tau, \mathbf{A}), \mathbf{Q}(\tau, \mathbf{A})$ and $\mathbf{R}(\tau, \mathbf{A})$ as follows.

$$\mathbf{P}(\tau, \mathbf{A}) = \mathrm{e}^{\tau\lambda_1}\mathbf{I} + \tau \exp[\tau\lambda_1; \tau\lambda_2](\mathbf{A} - \lambda_1\mathbf{I}) + \tau^2 \exp[\tau\lambda_1; \tau\lambda_2; \tau\lambda_3](\mathbf{A} - \lambda_1\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I}) \tag{44}$$

$$\mathbf{Q}(\tau, \mathbf{A}) = \tau\,\mathrm{q}(\tau\lambda_1)\mathbf{I} + \tau^2\,\mathrm{q}[\tau\lambda_1; \tau\lambda_2](\mathbf{A} - \lambda_1\mathbf{I}) + \tau^3\,\mathrm{q}[\tau\lambda_1; \tau\lambda_2; \tau\lambda_3](\mathbf{A} - \lambda_1\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I}) \tag{45}$$

$$\mathbf{R}(\tau, \mathbf{A}) = \tau^2\,\mathrm{r}(\tau\lambda_1)\mathbf{I} + \tau^3\,\mathrm{r}[\tau\lambda_1; \tau\lambda_2](\mathbf{A} - \lambda_1\mathbf{I}) + \tau^4\,\mathrm{r}[\tau\lambda_1; \tau\lambda_2; \tau\lambda_3](\mathbf{A} - \lambda_1\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I}) \tag{46}$$

The expression for $\mathbf{P}(\tau, \mathbf{A})$ is a trivial extension of Eq. (38). We obtain $\mathbf{Q}(\tau, \mathbf{A})$ by integrating the terms in $\mathbf{P}(\xi, \mathbf{A})$ with respect to ξ; cf. Eq. (19). Likewise, $\mathbf{R}(\tau, \mathbf{A})$ is obtained by integrating the terms in $\mathbf{Q}(\xi, \mathbf{A})$ with respect to ξ; cf. Eq. (21). We have used the following results to arrive at these equations.

$$\int_0^\tau \xi \exp[\xi\lambda_1; \xi\lambda_2]\,\mathrm{d}\xi = \tau \frac{\exp[0; \tau\lambda_2] - \exp[0; \tau\lambda_1]}{\lambda_2 - \lambda_1} = \tau^2\,\mathrm{q}[\tau\lambda_1; \tau\lambda_2] \tag{47}$$

$$\begin{aligned} \int_0^\tau \xi^2 \exp[\xi\lambda_1; \xi\lambda_2; \xi\lambda_3]\,\mathrm{d}\xi &= \int_0^\tau \xi \frac{\exp[\xi\lambda_2; \xi\lambda_3] - \exp[\xi\lambda_1; \xi\lambda_2]}{\lambda_3 - \lambda_1}\,\mathrm{d}\xi \\ &= \tau^3\,\mathrm{q}[\tau\lambda_1; \tau\lambda_2; \tau\lambda_3] \end{aligned} \tag{48}$$

$$\int_0^\tau \int_0^\eta \xi \exp[\xi\lambda_1; \xi\lambda_2]\,\mathrm{d}\xi\,\mathrm{d}\eta = \tau^2 \frac{\exp[0; 0; \tau\lambda_2] - \exp[0; 0; \tau\lambda_1]}{\lambda_2 - \lambda_1} = \tau^3\,\mathrm{r}[\tau\lambda_1; \tau\lambda_2] \tag{49}$$

$$\begin{aligned} \int_0^\tau \int_0^\eta \xi^2 \exp[\xi\lambda_1; \xi\lambda_2; \xi\lambda_3]\,\mathrm{d}\xi\,\mathrm{d}\eta &= \int_0^\tau \int_0^\eta \xi \frac{\exp[\xi\lambda_2; \xi\lambda_3] - \exp[\xi\lambda_1; \xi\lambda_2]}{\lambda_3 - \lambda_1}\,\mathrm{d}\xi\,\mathrm{d}\eta \\ &= \tau^4\,\mathrm{r}[\tau\lambda_1; \tau\lambda_2; \tau\lambda_3] \end{aligned} \tag{50}$$

Let α, β be real numbers and consider a complex number λ_c as defined below.

$$\mathrm{i} := \sqrt{-1}, \quad \lambda_\mathrm{c} := \alpha + \mathrm{i}\beta, \quad \Rightarrow \lambda_\mathrm{c}^* = \alpha - \mathrm{i}\beta \tag{51}$$

The cardinal sine function $\operatorname{sinc}(x)$ is defined as follows.

$$\operatorname{sinc}(x) := \begin{cases} \dfrac{\sin(x)}{x} & \text{if } x \neq 0, \\ 1 & \text{if } x = 0. \end{cases} \tag{52}$$

Further, define two auxiliary functions $\Psi(x, y)$ and $\Phi(x, y)$ as follows.

$$\Psi(x, y) := \cos(y) - x \operatorname{sinc}(y) \tag{53}$$

$$\Phi(x, y) := \exp[-\mathrm{i}y; x; \mathrm{i}y] = \begin{cases} \dfrac{\mathrm{e}^x - \Psi(-x, y)}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0), \\ \dfrac{1}{2} & \text{if } (x, y) = (0, 0). \end{cases} \tag{54}$$

In the case of complex eigenvalues, i.e. $\{\lambda_1, \lambda_2\} = \{\lambda_\mathrm{c}, \lambda_\mathrm{c}^*\}$ we can evaluate $\mathbf{P}(\tau, \mathbf{A}), \mathbf{Q}(\tau, \mathbf{A})$ and $\mathbf{R}(\tau, \mathbf{A})$ as follows.

$$\mathbf{P}(\tau, \mathbf{A}) = \mathrm{e}^{\tau\alpha} \left[\Psi(\tau\alpha, \tau\beta)\mathbf{I} + \tau \operatorname{sinc}(\tau\beta)\mathbf{A} + \tau^2\, \Phi(\tau\lambda_3 - \tau\alpha, \tau\beta)[(\mathbf{A} - \alpha\mathbf{I})^2 + \beta^2\mathbf{I}]\right] \tag{55}$$

$$\begin{aligned} \mathbf{Q}(\tau, \mathbf{A}) = \tau\mathrm{e}^{\tau\alpha}\big[&\operatorname{sinc}(\tau\beta)\mathbf{I} + \tau\, \Phi(-\tau\alpha, \tau\beta)(\mathbf{A} - 2\alpha\mathbf{I}) \\ &+ \tau^2\, \Phi(\star, \tau\beta)[-\tau\alpha; \tau\lambda_3 - \tau\alpha][(\mathbf{A} - \alpha\mathbf{I})^2 + \beta^2\mathbf{I}]\big] \end{aligned} \tag{56}$$

$$\begin{aligned} \mathbf{R}(\tau, \mathbf{A}) = \tau^2\mathrm{e}^{\tau\alpha}\big[&\Phi(-\tau\alpha, \tau\beta)\mathbf{I} + \tau\, \Phi(\star, \tau\beta)[-\tau\alpha; -\tau\alpha](\mathbf{A} - 2\alpha\mathbf{I}) \\ &+ \tau^2\, \Phi(\star, \tau\beta)[-\tau\alpha; -\tau\alpha; \tau\lambda_3 - \tau\alpha][(\mathbf{A} - \alpha\mathbf{I})^2 + \beta^2\mathbf{I}]\big] \end{aligned} \tag{57}$$

The expression for $\mathbf{P}(\tau, \mathbf{A})$ is a trivial extension of Eq. (39) and the choice of the auxiliary functions $\Psi(x, y)$ and $\Phi(x, y)$ is motivated by the structure of the same. The notation $\Phi(\star, y)[x_1; x_2]$ means that the divided differences are to be taken with respect to the variable in whose place the symbol $\star$ appears. The rest of this section describes some results which were used to arrive at the expressions for $\mathbf{Q}(\tau, \mathbf{A})$ and $\mathbf{R}(\tau, \mathbf{A})$ from the expression for $\mathbf{P}(\tau, \mathbf{A})$.

The following integrals are straightforward.

$$\int_0^\tau \mathrm{e}^{\xi\alpha} \cos(\xi\beta)\, \mathrm{d}\xi = \frac{\mathrm{e}^{\tau\alpha}[\alpha\cos(\tau\beta) + \beta\sin(\tau\beta)] - \alpha}{\alpha^2 + \beta^2} = \tau\mathrm{e}^{\tau\alpha}[\operatorname{sinc}(\tau\beta) - \tau\alpha\, \Phi(-\tau\alpha, \tau\beta)] \tag{58}$$

$$\int_0^\tau \mathrm{e}^{\xi\alpha} \sin(\xi\beta)\, \mathrm{d}\xi = \frac{\mathrm{e}^{\tau\alpha}[\alpha\sin(\tau\beta) - \beta\cos(\tau\beta)] + \beta}{\alpha^2 + \beta^2} = \tau^2\beta\mathrm{e}^{\tau\beta}\, \Phi(-\tau\alpha, \tau\beta) \tag{59}$$

$$\Rightarrow \int_0^\tau \mathrm{e}^{\xi\alpha}\, \Psi(\xi\alpha, \xi\beta)\, \mathrm{d}\xi = \tau\mathrm{e}^{\tau\alpha}[\operatorname{sinc}(\tau\beta) - 2\tau\alpha\, \Phi(-\tau\alpha, \tau\beta)] \tag{60}$$

Using Eqs. (59) and (60) we obtain the following result.

$$\int_0^\tau \mathrm{e}^{\xi\alpha}\xi^2\,\Phi(\xi\lambda_3-\xi\alpha,\xi\beta)\,\mathrm{d}\xi = \int_0^\tau \frac{\mathrm{e}^{\xi\lambda_3}-\mathrm{e}^{\xi\alpha}[\Psi(\xi\alpha,\xi\beta)+\xi\lambda_3\,\mathrm{sinc}(\xi\beta)]}{(\lambda_3-\alpha)^2+\beta^2}\,\mathrm{d}\xi \tag{61}$$

$$= \tau\left[\frac{\mathrm{q}(\tau\lambda_3)-\mathrm{e}^{\tau\alpha}[\mathrm{sinc}(\tau\beta)+\tau(\lambda_3-2\alpha)\,\Phi(-\tau\alpha,\tau\beta)]}{(\lambda_3-\alpha)^2+\beta^2}\right] \tag{62}$$

$$= \tau\mathrm{e}^{\tau\alpha}\left[\frac{[\mathrm{e}^{\tau\lambda_3}-1]\mathrm{e}^{-\tau\alpha}-\tau\lambda_3[\mathrm{sinc}(\tau\beta)+\tau(\lambda_3-2\alpha)\,\Phi(-\tau\alpha,\tau\beta)]}{\tau\lambda_3[(\lambda_3-\alpha)^2+\beta^2]}\right] \tag{63}$$

$$= \tau\mathrm{e}^{\tau\alpha}\left[\frac{\mathrm{e}^{\tau\lambda_3-\tau\alpha}-\Psi(\tau\alpha-\tau\lambda_3,\tau\beta)-\tau^2[(\lambda_3-\alpha)^2+\beta^2]\,\Phi(-\tau\alpha,\tau\beta)}{\tau\lambda_3[(\lambda_3-\alpha)^2+\beta^2]}\right] \tag{64}$$

$$= \tau^3\mathrm{e}^{\tau\alpha}\left[\frac{\Phi(\tau\lambda_3-\tau\alpha,\tau\beta)-\Phi(-\tau\alpha,\tau\beta)}{(\tau\lambda_3-\tau\alpha)-(-\tau\alpha)}\right] = \tau^3\mathrm{e}^{\tau\alpha}\,\Phi(\star,\tau\beta)[-\tau\alpha,\tau\lambda_3-\tau\alpha] \tag{65}$$

The results given in Eqs. (59), (60) and (65) are used to obtain $\mathbf{Q}(\tau,\mathbf{A})$ from $\mathbf{P}(\tau,\mathbf{A})$. Substituting $\lambda_3=0$ in Eq. (65) we get the following result.

$$\int_0^\tau \mathrm{e}^{\xi\alpha}\xi^2\,\Phi(-\xi\alpha,\xi\beta)\,\mathrm{d}\xi = \tau^3\mathrm{e}^{\tau\alpha}\,\Phi(\star,\tau\beta)[-\tau\alpha;-\tau\alpha] \tag{66}$$

Using Eqs. (65) and (66) we arrive at the following result.

$$\int_0^\tau \mathrm{e}^{\xi\alpha}\xi^3\,\Phi(\star,\xi\beta)[-\xi\alpha;\xi\lambda_3-\xi\alpha]\,\mathrm{d}\xi = \int_0^\tau \mathrm{e}^{\xi\alpha}\xi^2\frac{\Phi(\xi\lambda_3-\xi\alpha,\xi\beta)-\Phi(-\xi\alpha,\xi\beta)}{\lambda_3}\,\mathrm{d}\xi \tag{67}$$

$$= \tau^3\mathrm{e}^{\tau\alpha}\left[\frac{\Phi(\star,\tau\beta)[-\tau\alpha;\tau\lambda_3-\tau\alpha]-\Phi(\star,\tau\beta)[-\tau\alpha;-\tau\alpha]}{\lambda_3}\right] \tag{68}$$

$$= \tau^4\mathrm{e}^{\tau\alpha}\,\Phi(\star,\tau\beta)[-\tau\alpha;-\tau\alpha;\tau\lambda_3-\tau\alpha] \tag{69}$$

The results given in Eqs. (59), (66) and (69) are used to obtain $\mathbf{R}(\tau,\mathbf{A})$ from $\mathbf{Q}(\tau,\mathbf{A})$.

Recall that we have considered the case when $n=3$ (3D problems) and note that the equations for $\mathbf{P}(\tau,\mathbf{A})$, $\mathbf{Q}(\tau,\mathbf{A})$ and $\mathbf{R}(\tau,\mathbf{A})$ are expressed as the sum of three terms. Due to the properties of the polynomial in the Newton's form, the corresponding equations for $n=2$ (2D problems) can be obtained from the equations for $n=3$ by dropping out the third term.

4.3 Formulas for the eigenvalues of 2×2 and 3×3 matrices

The eigenvalues of a given matrix are found by solving its characteristic equation. Let $\det(\mathbf{A})$ and $\mathrm{tr}(\mathbf{A})$ denote the determinant and trace of the matrix $\mathbf{A}$, respectively. The characteristic equation of the matrix $\mathbf{A}$ is given by the following.

$$\det(\mathbf{A}-\lambda\mathbf{I})=0 \tag{70}$$

When $n=2$, the characteristic equation can be expressed as,

$$\lambda^2-\mathrm{tr}(\mathbf{A})\lambda+\det(\mathbf{A})=0 \tag{71}$$

The solution of the above quadratic equation is straight-forward.

$$\lambda_1=\frac{\mathrm{tr}(\mathbf{A})-\sqrt{\mathrm{tr}(\mathbf{A})^2-4\det(\mathbf{A})}}{2},\quad \lambda_2=\frac{\mathrm{tr}(\mathbf{A})+\sqrt{\mathrm{tr}(\mathbf{A})^2-4\det(\mathbf{A})}}{2} \tag{72}$$

Clearly, we get complex eigenvalues when $\operatorname{tr}(\mathbf{A})^2 < 4\det(\mathbf{A})$ and on the contrary we get real eigenvalues. In this section we follow the convention that the square roots are single-valued and positive. Note that the two admissible solutions to the square root function are already taken into consideration in the above formula. In the case of real eigenvalues, it is straightforward to verify that the above formula guarantees $\lambda_1 \leq \lambda_2$.

When $n = 3$, the characteristic equation can be expressed as,

$$\lambda^3 - \operatorname{tr}(\mathbf{A})\lambda^2 + \frac{\operatorname{tr}(\mathbf{A})^2 - \operatorname{tr}(\mathbf{A}^2)}{2}\lambda - \det(\mathbf{A}) = 0 \tag{73}$$

The solution of the above cubic equation can be found by the classical method published by Gerolamo Cardano, cf. [21]. The calculation steps of the same are summarized below.

$$\mathbf{B} := \mathbf{A} - \frac{\operatorname{tr}(\mathbf{A})}{3}\mathbf{I}, \quad Q := \frac{\operatorname{tr}(\mathbf{B}^2)}{6}, \quad R := \frac{\det(\mathbf{B})}{2} \tag{74}$$

$$\lambda_1 = \frac{\operatorname{tr}(\mathbf{A})}{3} + \sqrt[3]{R - \sqrt{R^2 - Q^3}}\,\mathrm{e}^{-\mathrm{i}(2\pi/3)} + \sqrt[3]{R + \sqrt{R^2 - Q^3}}\,\mathrm{e}^{\mathrm{i}(2\pi/3)} \tag{75}$$

$$\lambda_2 = \frac{\operatorname{tr}(\mathbf{A})}{3} + \sqrt[3]{R - \sqrt{R^2 - Q^3}}\,\mathrm{e}^{-\mathrm{i}(4\pi/3)} + \sqrt[3]{R + \sqrt{R^2 - Q^3}}\,\mathrm{e}^{\mathrm{i}(4\pi/3)} \tag{76}$$

$$\lambda_3 = \frac{\operatorname{tr}(\mathbf{A})}{3} + \sqrt[3]{R - \sqrt{R^2 - Q^3}} + \sqrt[3]{R + \sqrt{R^2 - Q^3}} \tag{77}$$

We follow the convention that the cube roots that appear in the above expressions are real and single valued. The three admissible solutions to the cube root function are already taken into consideration in the above formula.

Note that when the discriminant $(R^2 - Q^3) > 0$, we obtain complex eigenvalues. In this case, the formulas are already in a suitable format for implementation. When $(R^2 - Q^3) \leq 0$ we obtain real eigenvalues and the formulas for the same can be written in a form better suited for implementation as follows.

$$\theta := \arccos(\frac{R}{\sqrt{Q^3}}), \quad \lambda_n = \frac{\operatorname{tr}(\mathbf{A})}{3} + 2\sqrt{Q}\cos(\frac{2\pi n + \theta}{3}) \tag{78}$$

where arccos() denotes the inverse cosine function whose range is defined to be the closed interval $[0, \pi]$. The formula for the real eigenvalues given in Eq. (78) guarantees $\lambda_1 \leq \lambda_2 \leq \lambda_3$. This can be verified using the following results.

$$0 \leq \theta \leq \pi \Rightarrow \begin{array}{c} -1 \leq \cos(\frac{2\pi + \theta}{3}) \leq \frac{-1}{2} \\ \frac{-1}{2} \leq \cos(\frac{4\pi + \theta}{3}) \leq \frac{1}{2} \\ \frac{1}{2} \leq \cos(\frac{6\pi + \theta}{3}) \leq 1 \end{array} \Rightarrow \begin{array}{c} \frac{\operatorname{tr}(\mathbf{A})}{3} - 2\sqrt{Q} \leq \lambda_1 \leq \frac{\operatorname{tr}(\mathbf{A})}{3} - \sqrt{Q} \\ \frac{\operatorname{tr}(\mathbf{A})}{3} - \sqrt{Q} \leq \lambda_2 \leq \frac{\operatorname{tr}(\mathbf{A})}{3} + \sqrt{Q} \\ \frac{\operatorname{tr}(\mathbf{A})}{3} + \sqrt{Q} \leq \lambda_3 \leq \frac{\operatorname{tr}(\mathbf{A})}{3} + 2\sqrt{Q} \end{array} \tag{79}$$

Note that in the case of two equal eigenvalues, it will be either $\lambda_1 = \lambda_2$ or $\lambda_2 = \lambda_3$. In all the situations the eigenvalue λ_3 is always a real number.

5 Stable evaluation of formulas under finite precision

5.1 Introduction

The issue with stable evaluation of formulas is best explained by an example. The example consists in the naïve evaluation of a second-order divided difference given by the expression

h	Formula1 evaluation	Exact 16 digits	Formula2 evaluation
10^{-01}	1.503 335 165 136 320	1.503 335 165 136 325	1.503 335 165 136 292
10^{-02}	1.372 811 947 550 877	1.372 811 947 550 820	1.372 811 947 550 871
10^{-03}	1.360 500 848 424 467	1.360 500 848 315 854	1.360 500 848 386 436
10^{-04}	1.359 276 824 430 971	1.359 276 836 249 607	1.359 276 831 150 054
10^{-05}	1.359 152 790 283 402	1.359 154 505 717 948	1.359 151 840 209 960
10^{-06}	1.359 135 026 857 928	1.359 142 273 371 229	1.359 375
10^{-07}	1.332 267 628 772 320	1.359 141 050 143 621	1.359 375
10^{-08}	2.220 446 084 949 470	1.359 140 927 820 931	4
10^{-09}	0	1.359 140 915 588 663	256
10^{-10}	0	1.359 140 914 365 436	0
10^{-11}	−2 220 445.681 810 107	1.359 140 914 243 114	−4 194 304
10^{-12}	−222 005 130.399 6447	1.359 140 914 230 881	−268 435 456
10^{-13}	0	1.359 140 914 229 658	17 179 869 184
10^{-14}	2 223 999 815 985.422	1.359 140 914 229 536	4 398 046 511 104
10^{-15}	0	1.359 140 914 229 523	0

Table 1: Loss of significant digits in the naïve evaluations of $\exp[1; 1+h; 1+2h]$.

$\exp[1; 1+h; 1+2h]$. We denote by Formula1 the "as is" expression of the second-order divided difference.

$$x_1 = 1, \quad x_2 = 1+h, \quad x_3 = 1+2h \tag{80}$$

$$\exp[x_1; x_2; x_3] = \frac{1}{x_3 - x_1}\left[\frac{e^{x_3} - e^{x_2}}{x_3 - x_2} - \frac{e^{x_2} - e^{x_1}}{x_2 - x_1}\right] \tag{81}$$

Using Eq. (37) the above equation can be rearranged in an algebraically equivalent form which we denote as Formula2.

$$\exp[x_1; x_2; x_3] = \frac{e^{x_1}}{(x_1 - x_2)(x_1 - x_3)} + \frac{e^{x_2}}{(x_2 - x_1)(x_2 - x_3)} + \frac{e^{x_3}}{(x_3 - x_1)(x_3 - x_2)} \tag{82}$$

All the expressions that appear in the formulas given in [11] are expressed in the above simplified form.

Table 1 illustrates the results of the naïve evaluation of both formulas for values of h gradually tending to zero and using double precision floating point arithmetic. The exact values up to 16 digits of precision are given in the third column. The significant digits in both the formula evaluations that coincide with the exact values are highlighted in green colour. We observe a gradual loss of significant digits in both the formula evaluations which deteriorates as $h \to 0$. In fact, for $h \leq 10^{-8}$ we lose all the significant digits in both the formula evaluations making these computations useless. These numerical instabilities worsen when we switch the computations to single precision floating point arithmetic.

In these evaluations which have removable singularities when $h = 0$, it is critical to first identify terms which suffer from cancellation errors. This identification serves two purposes: a) to foresee numerical instability in formula evaluations and b) to assist in the design of procedures to control the loss of significant digits in these evaluations. In this sense, Formula1 has an algebraic structure which facilitates the identification of such terms. Removable singularities in Formula1 are localized to terms which appear as divided differences which in turn are known to suffer from cancellation errors. Thus, the identification of cancellations errors in Formula1 is trivial. On the other hand, removable singularities in Formula2 are not localized, i.e. the terms that participate to obtain a finite limit at the removable singularities are dispersed within the formula. This obscures the a priori identification of possible cancellation errors as they are not evident, i.e. one might not foresee numerical instability.

In the analytical solution of the X-IVAS scheme, the following expressions might suffer from cancellation errors in a straight-forward (naïve) evaluation of the same using finite precision arithmetic.

$$\begin{gathered}\exp[\tau\lambda_1;\tau\lambda_2],\quad \exp[\tau\lambda_1;\tau\lambda_2;\tau\lambda_3],\quad \mathrm{q}(\tau\lambda_1),\quad \mathrm{q}[\tau\lambda_1;\tau\lambda_2],\quad \mathrm{q}[\tau\lambda_1;\tau\lambda_2;\tau\lambda_3],\\ \mathrm{r}(\tau\lambda_1),\quad \mathrm{r}[\tau\lambda_1;\tau\lambda_2],\quad \mathrm{r}[\tau\lambda_1;\tau\lambda_2;\tau\lambda_3],\quad \Phi(-\tau\alpha,\tau\beta),\quad \Phi(\tau\lambda_3-\tau\alpha,\tau\beta),\\ \Phi(\star,\tau\beta)[-\tau\alpha;-\tau\alpha],\quad \Phi(\star,\tau\beta)[-\tau\alpha;\tau\lambda_3-\tau\alpha],\quad \Phi(\star,\tau\beta)[-\tau\alpha;-\tau\alpha;\tau\lambda_3-\tau\alpha]\end{gathered} \tag{83}$$

The above expressions can be identified as the elements of the following nested set of divided differences.

$$\begin{aligned}\Big\{&\big\{\exp[x_1;x_2],\mathrm{q}(x)\big\},\big\{\exp[x_1;x_2;x_3],\mathrm{q}[x_1;x_2],\mathrm{r}(x),\Phi(x,y)\big\},\\ &\big\{\mathrm{q}[x_1;x_2;x_3],\mathrm{r}[x_1;x_2],\Phi(\star,y)[x_1;x_2]\big\},\big\{\mathrm{r}[x_1;x_2;x_3],\Phi(\star,y)[x_1;x_1;x_2]\big\}\Big\}\end{aligned} \tag{84}$$

The order of the divided differences gradually increase from first-order in the first subset to fourth-order in the last subset. All elements of a subset are particular cases of the first element of that subset. For instance,

$$\mathrm{q}[x_1;x_2]=\exp[0;x_1;x_2],\quad \mathrm{r}(x)=\exp[0;0;x],\quad \Phi(x,y)=\exp[-\mathrm{i}y;x;\mathrm{i}y] \tag{85}$$

Following this line, it is possible to express all the divided differences in Eq. (84) as the divided differences of the exponential function; The details of the same are given in Section 5.4. As $h\to 0$, the rate of loss of significant digits in a naïve evaluation of divided differences is generally equal to the order of the same. In the considered example, i.e. $\exp[x_1;x_2;x_3]$ we loose significant digits at a second order rate. Following this line, naïve evaluations of the third and the fourth subsets in Eq. (84) are meaningless for $h\leq 10^{-5}$ and $h\leq 10^{-4}$, respectively.

An algorithm for the accurate computation of divided differences of the exponential function was already presented by McCurdy et.al [22]. Following this line, a similar algorithm for the accurate computation of divided differences of the auxiliary functions q() and r(), cf. Eq. (42) and Eq. (43), was presented by Caliari [23]. These algorithms have a wider scope, i.e. they were designed to evaluate functions of $n\times n$ matrices appearing in exponential integrators for large systems of equations (ordinary or differential). A user who already has these algorithms implemented, might just invoke them to evaluate the divided differences listed in Eq. (83) and use them in the formulas for $\mathbf{P}(\tau,\mathbf{A})$, $\mathbf{Q}(\tau,\mathbf{A})$ and $\mathbf{R}(\tau,\mathbf{A})$ given in Section 4.2. This would address the numerical stability issues in the formula evaluations.

In what follows, we discuss a simple yet stable piecewise evaluation technique for divided differences of limited scope. In other words, this technique should be understood as a *specialization* for the at most fourth-order divided differences found in the formulas for $\mathbf{P}(\tau,\mathbf{A})$, $\mathbf{Q}(\tau,\mathbf{A})$ and $\mathbf{R}(\tau,\mathbf{A})$.

5.2 Double precision floating point numbers

It pays to understand how floating point numbers are stored in a computer. In this section we briefly describe how double precision floating point numbers are stored as per the IEEE 754 standard. Any decimal floating point number within the range of the double can be written in the normalized form as follows.

$$\begin{array}{rcccl} \text{Decimal form} \rightarrow (-1)^s\, 2^e\, 1.f \approx & \multicolumn{1}{|c|}{s} & \multicolumn{1}{c|}{(e+1023)_\mathrm{b}} & \multicolumn{1}{c|}{0.f_\mathrm{b}} & \leftarrow \text{Binary form} \\ \cline{2-4} & 1 & 11 & 52 & \leftarrow \text{No. of bits stored} \end{array} \tag{86}$$

In the above equation the boolean $s \in \{0, 1\}$ is called the *sign bit*, the integer e is called the *exponent*; $-1022 \leq e \leq 1023$ and the fraction f is called the *significand.* The numbers with a subscript b are expressed in the binary format. For instance,

$$0.1 = (-1)^0\, 2^{-4}\, 1.6, \quad \Rightarrow \quad s = 0, \quad (e + 1023) = 1019, \quad 0.f = 0.6 \tag{87}$$

$$1019_\mathrm{b} = 11\,1111\,1011, \quad 0.6_\mathrm{b} = 0.\{1001\}_\infty \tag{88}$$

The binary expression for 0.6 is a nonterminating fraction. The notation $\{1001\}_n$ means that the bits 1001 are repeated n times. Thus, using finite precision (52 bits to store the significand) the fraction 0.6 cannot be represented exactly. We get a fraction which is exactly representable as a double by truncating the nonterminating binary fraction after 52 bits. Further adjacent doubles are obtained by adding and subtracting a unit in the least significant position (`ulp`), respectively. These three exactly representable numbers are,

$$\begin{array}{|c|c|c|}\hline 0 & 011\,1111\,1011 & \{1001\}_{12}\,1000 \\ \hline \end{array} = \frac{1}{10} - \frac{8}{5} 2^{-56} \tag{89}$$

$$\begin{array}{|c|c|c|}\hline 0 & 011\,1111\,1011 & \{1001\}_{12}\,1001 \\ \hline \end{array} = \frac{1}{10} - \frac{3}{5} 2^{-56} \tag{90}$$

$$\begin{array}{|c|c|c|}\hline 0 & 011\,1111\,1011 & \{1001\}_{12}\,1010 \\ \hline \end{array} = \frac{1}{10} + \frac{2}{5} 2^{-56} \tag{91}$$

Hence, when 0.1 is stored as a double, it is rounded to the nearest representable number.

$$0.1_\mathrm{b} = 0.0001\{1001\}_\infty \approx \begin{array}{|c|c|c|}\hline 0 & 011\,1111\,1011 & \{1001\}_{12}\,1010 \\ \hline \end{array} = \frac{1}{10} + \frac{2}{5} 2^{-56} \tag{92}$$

Due to finite precision, the error in representing/storing a given decimal number as a double is at most half `ulp`. For a given number, the `ulp` depends on its exponent as shown below.

$$1\,\texttt{ulp} = 2^e\, 2^{-52} \tag{93}$$

In other words, the gap between two adjacent doubles is nonuniform. This allows 52 bits of precision throughout the range of double when expressed in the normalized form.

5.3 Optimal series approximation of divided differences

In this section we establish optimal series approximation of divided differences of a given function $\mathrm{f}(x)$. Consider the sequence $\{x_1, x_2, \cdots, x_n\}$ and some definitions related to this sequence.

$$x_\mathrm{a} := \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \widetilde{x}_i := x_i - x_\mathrm{a}, \quad \mathcal{X} := \{\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_n\} \tag{94}$$

$$\mathcal{X}_\mathrm{p} := \texttt{choose}(\mathcal{X}, 2), \quad \binom{n}{k} := \frac{n!}{k!(n-k)!}, \quad x_\mathrm{p}^2 := \sum_{i=1}^{\binom{n}{2}} \prod_{j=1}^{2} \mathcal{X}_\mathrm{p}(i,j) \tag{95}$$

where x_a is the mean value of the sequence and $\widetilde{x}_i$ is the fluctuation of x_i about the mean. The function $\texttt{choose}(\mathcal{X}, 2)$ returns a sequence $\mathcal{X}_\mathrm{p}$ consisting of pair-combinations ($2-$combinations) of elements from $\mathcal{X}$. The sum of the product of the pairs in $\mathcal{X}_\mathrm{p}$ is stored as the square of the auxiliary variable x_p. The result is stored as x_p^2 to highlight the fact that it is a second order term. Likewise, the triple, quadruple and quintuple combinations of $\mathcal{X}$ are denoted as $\mathcal{X}_\mathrm{t}$, $\mathcal{X}_\mathrm{q}$ and $\mathcal{X}_\mathrm{v}$, respectively. Further, the sum of the product of the triples, quadruples and quintuples are stored in x_t^3, x_q^4 and x_v^5, respectively. Following x_p^2, the superscripts (which are ordinary powers) in x_t^3, x_q^4 and x_v^5 highlight the fact that they are third, fourth and fifth order terms, respectively. Thus,

$$\mathcal{X}_\mathrm{t} := \texttt{choose}(\mathcal{X}, 3), \quad \mathcal{X}_\mathrm{q} := \texttt{choose}(\mathcal{X}, 4), \quad \mathcal{X}_\mathrm{v} := \texttt{choose}(\mathcal{X}, 5) \tag{96}$$

$$x_\mathrm{t}^3 := \sum_{i=1}^{\binom{n}{3}} \prod_{j=1}^{3} \mathcal{X}_\mathrm{t}(i,j), \quad x_\mathrm{q}^4 := \sum_{i=1}^{\binom{n}{4}} \prod_{j=1}^{4} \mathcal{X}_\mathrm{q}(i,j), \quad x_\mathrm{v}^5 := \sum_{i=1}^{\binom{n}{5}} \prod_{j=1}^{5} \mathcal{X}_\mathrm{v}(i,j) \tag{97}$$

Using the above definitions, we can derive[6] the following identity for the divided differences of $\mathrm{f}(x)$. The mean value theorem guarantees the existence of a ξ in the smallest interval containing $\{x_1, x_2, \cdots, x_n\}$ such that,

$$\mathrm{f}^{(n)}(\xi) := \frac{\partial^n}{\partial \lambda^n}\,\mathrm{f}(\lambda)\bigg|_{\lambda=\xi}, \quad \mathrm{f}(x_n) = \mathrm{f}(\xi + x_n - \xi) = \mathrm{f}(\xi) + \sum_{n=1}^{\infty}(x_n - \xi)^n\,\frac{\mathrm{f}^{(n)}(\xi)}{n!} \tag{98}$$

$$\begin{aligned}\mathrm{f}[x_1; x_2; \cdots; x_n] = {} & \frac{\mathrm{f}^{(n-1)}(x_\mathrm{a})}{(n-1)!} - x_\mathrm{p}^2\,\frac{\mathrm{f}^{(n+1)}(x_\mathrm{a})}{(n+1)!} + x_\mathrm{t}^3\,\frac{\mathrm{f}^{(n+2)}(x_\mathrm{a})}{(n+2)!} + (x_\mathrm{p}^4 - x_\mathrm{q}^4)\,\frac{\mathrm{f}^{(n+3)}(x_\mathrm{a})}{(n+3)!} \\ & + (x_\mathrm{v}^5 - 2x_\mathrm{p}^2 x_\mathrm{t}^3)\,\frac{\mathrm{f}^{(n+4)}(\xi)}{(n+4)!}\end{aligned} \tag{99}$$

Note that the first term in the above equation provides a second-order approximation to $\mathrm{f}[x_1; x_2; \cdots; x_n]$. If the series is expanded with respect to any point other than x_a, the first-order terms are resurrected. Thus, the approximation is optimal for the choice x_a. For the first-order divided difference $\mathrm{f}[x_1; x_2]$, the above equation can be simplified and easily extended to any number of terms as shown below.

$$h := \frac{x_2 - x_1}{2}, \quad x_\mathrm{p}^2 = -h^2, \quad x_\mathrm{t}^2 = 0, \quad x_\mathrm{q}^2 = 0, \quad x_\mathrm{v}^2 = 0 \tag{100}$$

$$\mathrm{f}[x_1; x_2] = \mathrm{f}^{(1)}(x_\mathrm{a}) + h^2\frac{\mathrm{f}^{(3)}(x_\mathrm{a})}{3!} + \cdots + h^{2n-2}\frac{\mathrm{f}^{(2n-1)}(x_\mathrm{a})}{(2n-1)!} + h^{2n}\frac{\mathrm{f}^{(2n+1)}(\xi)}{(2n+1)!} \tag{101}$$

[6] As the algebra involved is overwhelming and error-prone, we have used the computer algebra system `Maple` to perform the simplifications and verifications. Thus, human intervention is dedicated to identify patterns and to discover abstract expressions such as x_p, x_t, x_q, etc.

Likewise, for the second-order divided difference $\mathrm{f}[x_1; x_2; x_3]$, Eq. (99) can be simplified to the following.

$$x_\mathrm{p}^2 = -\frac{3}{2} x_\sigma^2, \quad x_\sigma^2 := \frac{\widetilde{x}_1^2 + \widetilde{x}_2^2 + \widetilde{x}_3^2}{3}, \quad x_\mathrm{t}^3 = \widetilde{x}_1 \widetilde{x}_2 \widetilde{x}_3, \quad x_\mathrm{q}^2 = 0, \quad x_\mathrm{v}^2 = 0 \tag{102}$$

$$\mathrm{f}[x_1; x_2; x_3] = \frac{1}{2}\,\mathrm{f}^{(2)}(x_\mathrm{a}) + \frac{x_\sigma^2}{16}\,\mathrm{f}^{(4)}(x_\mathrm{a}) + \frac{x_\mathrm{t}^3}{120}\,\mathrm{f}^{(5)}(x_\mathrm{a}) + \frac{x_\sigma^4}{320}\,\mathrm{f}^{(6)}(x_\mathrm{a}) + 3 x_\sigma^2 x_\mathrm{t}^3 \frac{\mathrm{f}^{(7)}(\xi)}{7!} \tag{103}$$

where x_σ is the standard deviation of the considered sequence. It is possible to relate x_p and x_σ for all n and in this work we exploit this relationship as it reduces the number of arithmetic operations.

$$x_\mathrm{p}^2 = -\frac{n}{2} x_\sigma^2 \tag{104}$$

5.4 Stable piecewise evaluation technique

To control (bound) the loss of significant digits in the evaluations of the subsets in Eq. (84), we resort to a piecewise evaluation of the same. In other words, we switch the evaluations to the corresponding series expansions of the same should the difference of the independent variables be less than some threshold. These threshold values are chosen such that we retain as many significant digits as possible. Although this technique is elementary, it is systematic. Nevertheless, to put it in the words of Kahan and Darcy [12], before this technique can be used, three messy questions need tidy answers: 1) What value should be assigned to the threshold in this technique? 2) How many terms in the series approximation should this technique retain? and 3) How accurate is this technique? In this section we describe the piecewise evaluation technique in full detail and answer the above three questions.

The first subset in Eq. (84) can be evaluated to machine precision by rearranging them to the following functional form[7].

$$\mathrm{sinhc}(x) := \begin{cases} \frac{\sinh(x)}{x} & \text{if } x \neq 0, \\ 1 & \text{if } x = 0. \end{cases} \tag{105}$$

$$\exp[x_1; x_2] = \mathrm{e}^{(x_1+x_2)/2}\,\mathrm{sinhc}\left(\frac{x_2 - x_1}{2}\right), \quad \mathrm{q}(x) = \mathrm{e}^{x/2}\,\mathrm{sinhc}\left(\frac{x}{2}\right) \tag{106}$$

We now describe the details of the piecewise evaluation technique using the evaluation of $\exp[x_1; x_2; x_3]$ as an example. This term can be written as,

$$\exp[x_1; x_2; x_3] = \mathrm{e}^{x_2} \exp[x_1 - x_2; 0; x_3 - x_2] = \mathrm{e}^{x_2} \frac{\mathrm{q}(x_3 - x_2) - \mathrm{q}(x_1 - x_2)}{x_3 - x_1} \tag{107}$$

where the function $\mathrm{q}(x)$ is evaluated as shown in Eq. (106). Without loss of generality, we assume $x_1 \le x_2 \le x_3$. Consequently we have,

$$\forall\, \xi \in [x_1, x_3], \quad |\xi - x_2| \le (x_3 - x_1) \tag{108}$$

$$x_\sigma^2 \le (x_3 - x_1)^2, \quad |x_\mathrm{t}^3| \le (x_3 - x_1)^3 \tag{109}$$

In the evaluations of divided differences, the loss of significant digits is essentially due to the cancellations that occur in the dependent variables. Particularly, in Eq. (107) the loss of

[7] In this form the difference of the independent variables appear symbolically as input to a function that could be evaluated to machine precision

significance is due to the cancellations that occur in the term $\mathrm{q}(x_3 - x_2) - \mathrm{q}(x_1 - x_2)$. This term admits the following series expansion.

$$\mathrm{q}(x_3 - x_2) - \mathrm{q}(x_1 - x_2) = \frac{x_3 - x_1}{2}[1 + (x_\mathrm{a} - x_2) + \cdots] \tag{110}$$

Let $x_3 - x_1 = 2^{-m}$ where $m \geq 1$ is an integer. Then, Eq. (108) implies that the higher order terms in Eq. (110) gradually tend to zero. Thus,

$$\left.\begin{array}{l} x_3 - x_1 = 2^{-m} \\ x_1 \leq x_2 \leq x_3 \end{array}\right\} \quad \Rightarrow \quad \mathrm{q}(x_3 - x_2) - \mathrm{q}(x_1 - x_2) = O(2^{-(m+1)}) \tag{111}$$

When written in the normalized decimal form (cf. Eq. (86)), the exponent of $\mathrm{q}(x_3 - x_2)$ and $\mathrm{q}(x_1 - x_2)$ will be 0 and -1, respectively. This can be inferred using Eq. (108) as shown below.

$$0 \leq x \leq 2^{-m} \Rightarrow 2^0 \leq \mathrm{q}(x) < 2^1, \quad -2^m \leq x \leq 0 \Rightarrow 2^{-1} \leq \mathrm{q}(x) < 2^0 \tag{112}$$

$$0 \leq (x_3 - x_2) \leq 2^{-m} \Rightarrow \mathrm{q}(x_3 - x_2) = (-1)^0\, 2^0\, 1.\widehat{f} \approx \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.\widehat{f}_\mathrm{b} \\ \hline \end{array} \tag{113}$$

$$-2^{-m} \leq (x_1 - x_2) \leq 0 \Rightarrow \mathrm{q}(x_1 - x_2) = (-1)^0\, 2^{-1}\, 1.\widetilde{f} \approx \begin{array}{|c|c|c|}\hline 0 & 1022_\mathrm{b} & 0.\widetilde{f}_\mathrm{b} \\ \hline \end{array} \tag{114}$$

where $\widehat{f}$ and $\widetilde{f}$ denote the significands of $\mathrm{q}(x_3 - x_2)$ and $\mathrm{q}(x_1 - x_2)$, respectively. The subtraction $\mathrm{q}(x_3 - x_2) - \mathrm{q}(x_1 - x_2)$ can be described schematically as follows.

$$\begin{aligned} \mathrm{q}(x_3 - x_2) - \mathrm{q}(x_1 - x_2) &= (-1)^0\, 2^0\, 1.\widehat{f} - (-1)^0\, 2^{-1}\, 1.\widetilde{f} \\ &\approx \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.\widehat{f}_\mathrm{b} \\ \hline \end{array} - \begin{array}{|c|c|c|}\hline 0 & 1022_\mathrm{b} & 0.\widetilde{f}_\mathrm{b} \\ \hline \end{array} && \text{normalized form} \\ &= \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.\widehat{f}_\mathrm{b} \\ \hline \end{array} - \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.1\widetilde{f}_\mathrm{b} \\ \hline \end{array} && \text{align radix points} \\ &= \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.\{0\}_m 1 f_\mathrm{b} \\ \hline \end{array} && O(2^{-(m+1)}) \\ &= \begin{array}{|c|c|c|}\hline 0 & (1022 - m)_\mathrm{b} & 0.f_\mathrm{b} \\ \hline \end{array} && \text{normalized form} \end{aligned} \tag{115}$$

We see that among the stored 52 bits of $\widehat{f}_\mathrm{b}$ and $\widetilde{f}_\mathrm{b}$, the first m bits are lost due to cancellation. After subtraction, the unit bit at the $m + 1^{\mathrm{th}}$ place will become the implicit bit of the result which is not stored, cf. Eq. (86). The exponent of the result will become $-(m + 1)$. The significand of the result will become the remaining bits denoted in Eq. (115) as f_b of which only $51 - m$ bits are significant.

We see that if $\exp[x_1; x_2; x_3]$ is evaluated as shown in Eq. (107) we loose significant bits at a first order rate. We will call this form of evaluation as the *direct evaluation.* If $x_1 \leq x_2 \leq x_3$ and $x_3 - x_1 = 2^{-m}$ then in the direct evaluation of $\exp[x_1; x_2; x_3]$ we are left with $51 - m$ significant bits in the significand.

Using Eq. (103) we can write the series expansion for $\exp[x_1; x_2; x_3]$ as follows.

$$\exp[x_1; x_2; x_3] = \frac{\mathrm{e}^{x_\mathrm{a}}}{2}\mathcal{S}, \quad \mathcal{S} := \left[1 + \frac{x_\sigma^2}{8} + \frac{x_\mathrm{t}^3}{60} + \frac{x_\sigma^4}{160} + \frac{x_\sigma^2 x_\mathrm{t}^3}{840} + \cdots\right] \tag{116}$$

The above form to evaluate $\exp[x_1; x_2; x_3]$ will be called as the *series evaluation.* Clearly, in the series evaluation we do not find removable singularities which imply that there are no instances of cancellation errors. However, the truncation of the series will introduce an error which will limit the number of significant digits in the series evaluation that match those in an exact evaluation. When the series $\mathcal{S}$ is truncated after the first n terms it will be denoted as $\mathcal{S}_n$. As $\exp(x_\mathrm{a})/2$ can be evaluated to machine precision, the number of significant digits in the series evaluation is essentially limited by the term $\mathcal{S}_n$. The series $\mathcal{S}$ when written in the normalized decimal form has a zero exponent when $x_1 \leq x_2 \leq x_3$ and $x_3 - x_1 = 2^{-m}$. This can be inferred using Eqs. (103) and (109) as follows.

$$\exists\, \xi \in [x_1, x_3] \text{ such that, } \frac{\mathrm{e}^{x_\mathrm{a}}}{2}\left[1 + \frac{x_\sigma^2}{8} + \frac{x_\mathrm{t}^3}{60} + \cdots\right] = \frac{\mathrm{e}^{x_\mathrm{a}}}{2} + \frac{x_\sigma^2}{16}\mathrm{e}^{\xi} \tag{117}$$

$$1 + \frac{x_\sigma^2}{8} + \frac{x_\mathrm{t}^3}{60} + \cdots = 1 + \frac{x_\sigma^2}{8}\mathrm{e}^{\xi - x_\mathrm{a}}, \quad 2^0 \leq 1 + \frac{x_\sigma^2}{8}\mathrm{e}^{\xi - x_\mathrm{a}} \leq 1 + 2^{-(2m+3)}\mathrm{e}^{2^{-m}} < 2^1 \tag{118}$$

It follows that all terms except the first one contribute to the significand of $\mathcal{S}$. Thus,

$$\mathcal{S} = (-1)^0\, 2^0\, 1.f \approx \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.f_\mathrm{b} \\ \hline\end{array} \tag{119}$$

Hence, when $\mathcal{S}$ is replaced by $\mathcal{S}_n$, the associated truncation error can be understood as to limit the number of significant digits in the series evaluation. The truncation error associated to $\mathcal{S}_n$ is denoted as $\mathcal{E}_n$. From Eqs. (109) and (116) we infer,

$$\mathcal{E}_1 = O\left(\frac{x_\sigma^2}{8}\right) \leq O(2^{-(2m+3)}), \quad \mathcal{E}_2 = O\left(\frac{x_\mathrm{t}^3}{60}\right) \leq O(2^{-(3m+6)}) \tag{120}$$

$$\mathcal{E}_3 = O\left(\frac{x_\sigma^4}{160}\right) \leq O(2^{-(4m+8)}), \quad \mathcal{E}_4 = O\left(\frac{x_\sigma^2 x_\mathrm{t}^3}{840}\right) \leq O(2^{-(5m+10)}) \tag{121}$$

Expressing $\mathcal{S}_n = \mathcal{S} - \mathcal{E}_n$ in the double storage format we get,

$$\mathcal{S}_1 \approx \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.f_\mathrm{b} \\ \hline\end{array} - \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.\{0\}_{2m+2}1\cdots \\ \hline\end{array} \tag{122}$$

$$\mathcal{S}_2 \approx \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.f_\mathrm{b} \\ \hline\end{array} - \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.\{0\}_{3m+5}1\cdots \\ \hline\end{array} \tag{123}$$

$$\mathcal{S}_3 \approx \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.f_\mathrm{b} \\ \hline\end{array} - \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.\{0\}_{4m+7}1\cdots \\ \hline\end{array} \tag{124}$$

$$\mathcal{S}_4 \approx \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.f_\mathrm{b} \\ \hline\end{array} - \begin{array}{|c|c|c|}\hline 0 & 1023_\mathrm{b} & 0.\{0\}_{5m+9}1\cdots \\ \hline\end{array} \tag{125}$$

where $\mathcal{E}_n$ is written after the alignment of radix points and the remaining digits in the significands are denoted as $\cdots$. This implies that we have $(2m+2), (3m+5), (4m+7)$ and $(5m+9)$ significant digits in $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ and $\mathcal{S}_4$, respectively.

For each $\mathcal{S}_n$ we solve for m by matching the accuracy of the series evaluation with the one obtained in the direct evaluation. In this way, we obtain the threshold value of $(x_3 - x_1) = 2^{-m}$ and the lower bound for the number of significant digits `nsd` in a piecewise evaluation of $\exp[x_1; x_2; x_3]$. Thus,

$$\mathcal{S}_1: \quad 51 - m = 2m + 2 \quad \Rightarrow \quad m = 16, \quad \texttt{nsd} = 34 \text{ bits} \approx 11 \text{ decimal digits} \tag{126}$$

$$\mathcal{S}_2: \quad 51 - m = 3m + 5 \quad \Rightarrow \quad m = 12, \quad \texttt{nsd} = 39 \text{ bits} \approx 12 \text{ decimal digits} \tag{127}$$

$$\mathcal{S}_3: \quad 51 - m = 4m + 7 \quad \Rightarrow \quad m = 9, \quad \texttt{nsd} = 42 \text{ bits} \approx 13 \text{ decimal digits} \tag{128}$$

$$\mathcal{S}_4: \quad 51 - m = 5m + 9 \quad \Rightarrow \quad m = 7, \quad \texttt{nsd} = 44 \text{ bits} \approx 14 \text{ decimal digits} \tag{129}$$

h	Formula3 evaluation	Exact 16 digits	Formula4 evaluation
10^{-01}	1.503 335 165 136 323	1.503 335 165 136 325	1.503 335 165 136 323
10^{-02}	1.372 811 947 550 791	1.372 811 947 550 820	1.372 811 947 550 791
10^{-03}	1.360 500 848 316 010	1.360 500 848 315 854	1.360 500 848 315 854
10^{-04}	1.359 276 836 253 229	1.359 276 836 249 607	1.359 276 836 249 607
10^{-05}	1.359 154 505 691 532	1.359 154 505 717 948	1.359 154 505 717 948
10^{-06}	1.359 142 273 371 116	1.359 142 273 371 229	1.359 142 273 371 229
10^{-07}	1.359 141 050 143 620	1.359 141 050 143 621	1.359 141 050 143 622
10^{-08}	1.359 140 927 820 931	1.359 140 927 820 931	1.359 140 927 820 931
10^{-09}	1.359 140 915 588 663	1.359 140 915 588 663	1.359 140 915 588 663
10^{-10}	1.359 140 914 365 436	1.359 140 914 365 436	1.359 140 914 365 436
10^{-11}	1.359 140 914 243 114	1.359 140 914 243 114	1.359 140 914 243 114
10^{-12}	1.359 140 914 230 881	1.359 140 914 230 881	1.359 140 914 230 881
10^{-13}	1.359 140 914 229 658	1.359 140 914 229 658	1.359 140 914 229 658
10^{-14}	1.359 140 914 229 536	1.359 140 914 229 536	1.359 140 914 229 536
10^{-15}	1.359 140 914 229 524	1.359 140 914 229 523	1.359 140 914 229 524

Table 2: Loss of significant digits controlled in the stable piecewise evaluations of $\exp[1; 1+h; 1+2h]$.

In the above equations, the solution for m is rounded to the nearest integer. Using this rounded m we estimate `nsd` as the minimum of the number of significant digits found in the direct and the series evaluations. As the loss of significant digits is bounded from below, the piecewise evaluation of $\exp[x_1; x_2; x_3]$ is stable.

The numerical test presented in Section 5.1 is repeated here with $x_1 = 1$, $x_2 = 1 + h$, $x_3 = 1+2h$. Table 2 illustrated the results of the piecewise evaluation of the same using double precision floating point arithmetic and for values h gradually tending to zero. The exact values up to 16 digits of precision are given in the third column. We denote by Formula3 and Formula4 the piecewise evaluations considering $\mathcal{S}_1$ and $\mathcal{S}_4$ for the series evaluations, respectively. The significant digits in both formula evaluations that differ from the exact values are highlighted in green color. The lower bounds for the number of significant digits given in Eqs. (126) and (129) are reproduced in this test for Formula3 and Formula4, respectively.

5.5 Stable formulas for exponential divided differences

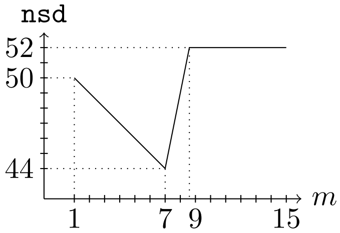
In this section we present stable piecewise definitions of all the expressions that belong to the subsets in Eq. (84). In the series evaluation part of every piecewise definition, we will consider the first four terms in the corresponding series expansion. Recall that each of these expressions can be written as some divided difference of the exponential function, cf. Eq. (85). The exponential function is its own derivative. This feature along with the abstraction (e.g. $x_\mathrm{p}, x_\mathrm{t}$ etc.) in the optimal series expansion permits us to use multiple terms in the series expansion without incurring substantial computational cost.

The first subset in Eq. (84) can be evaluated to machine precision without resorting to a series evaluation, cf. Eq. (106). The first element in the second subset, i.e. $\exp[x_1; x_2; x_3]$ was used as an example to describe the details of the piecewise evaluation technique in the previous section. The stable piecewise definition of the same when $x_1 \leq x_2 \leq x_3$ can be summarized as follows.

$$\exp[x_1; x_2; x_3] = \begin{cases} \mathrm{e}^{x_2} \dfrac{\mathrm{q}(x_3 - x_2) - \mathrm{q}(x_1 - x_2)}{x_3 - x_1} & \text{if } (x_3 - x_1) > 2^{-7} \\ \dfrac{\mathrm{e}^{x_\mathrm{a}}}{2} \left[1 + \dfrac{x_\sigma^2}{8} + \dfrac{x_\mathrm{t}^3}{60} + \dfrac{x_\sigma^4}{160} \right] & \text{else} \end{cases} \tag{130}$$

It is essential to sort the arguments lest the series evaluation should incur a significant truncation error. The variation in the number of significant digits `nsd` in $\exp[x_1; x_2; x_3]$ with respect to m, where $(x_3 - x_1) = 2^{-m}$, is denoted as $\texttt{nsd}(\exp[x_1; x_2; x_3], m)$. Using the above stable formula for $\exp[x_1; x_2; x_3]$ we obtain,

$$\texttt{nsd}(\exp[x_1; x_2; x_3], m) = \begin{cases} 51 - m & \text{if } m < 7 \\ 5m + 9 & \text{if } 7 \leq m < 8.6 \\ 52 & \text{if } m \geq 8.6 \end{cases}, \tag{131}$$



Recall that $\mathrm{q}[x_1; x_2] = \exp[x_1; 0; x_2]$, $\mathrm{r}(x) = \exp[0; 0; x]$ and $\Phi(x, y) = \exp[-\mathrm{i}y; \mathrm{i}y; x]$. Let `sort` be a sorting function and $\widehat{x}_1 \leq \widehat{x}_2 \leq \widehat{x}_3$. Then, using Eq. (130) a stable formula for $\mathrm{q}[x_1; x_2]$ is,

$$\{\widehat{x}_1, \widehat{x}_2, \widehat{x}_3\} = \texttt{sort}(\{x_1, 0, x_2\}), \quad \mathrm{q}[x_1; x_2] = \exp[\widehat{x}_1; \widehat{x}_2; \widehat{x}_3]; \tag{132}$$

Likewise a stable formula for $\mathrm{r}(x)$ is,

$$\{\widehat{x}_1, 0, \widehat{x}_3\} = \texttt{sort}(\{0, 0, x\}), \quad \mathrm{r}(x) = \exp[\widehat{x}_1; 0; \widehat{x}_3]; \tag{133}$$

As $\exp[-\mathrm{i}y; \mathrm{i}y; x]$ involves complex numbers it needs special attention. Recall that the exponential function is holomorphic, i.e. it is complex differentiable in a neighbourhood of every point in its domain. This implies that it is infinitely differentiable and equals to its own Taylor series. Thus, the optimal series approximation of divided differences presented in Section 5.3 naturally extends to $\exp[-\mathrm{i}y; \mathrm{i}y; x]$. Following this line, a stable formula for $\Phi(x, y)$ can be obtained as shown below.

$$z := x + \mathrm{i}y, \quad z_\mathrm{a} := \frac{x}{3}, \quad z_\sigma^2 := 2z_\mathrm{a}^2 - \frac{2}{3}y^2, \quad z_\mathrm{t}^3 := 2z_\mathrm{a}(z_\mathrm{a}^2 + y^2) \tag{134}$$

$$\Phi(x, y) = \exp[-\mathrm{i}y; \mathrm{i}y; x] = \begin{cases} \dfrac{\mathrm{e}^{z/2} \operatorname{sinhc}(z^*/2) - \operatorname{sinc}(y)}{z} & \text{if } |z| > 2^{-7} \\ \dfrac{\mathrm{e}^{z_\mathrm{a}}}{2} \left[1 + \dfrac{z_\sigma^2}{8} + \dfrac{z_\mathrm{t}^3}{60} + \dfrac{z_\sigma^4}{160} \right] & \text{else} \end{cases} \tag{135}$$

The above definition assumes the availability of a complex math library which provides an interface for a stable evaluation of common arithmetic operations, elementary and transcendental functions. This assumption holds for the `C++` programming language which is equipped with the standard math library `<complex>`.

We now discuss the stable evaluation of $\exp[x_1; x_2; x_3; x_4]$, which is a template for the elements of the third subset in Eq. (84). Following Eq. (99), the series expansion of this term can be written as,

$$\exp[x_1;x_2;x_3;x_4] = \frac{\mathrm{e}^{x_\mathrm{a}}}{3!}\left[1 - \frac{x_\mathrm{p}^2}{20} + \frac{x_\mathrm{t}^3}{120} + \frac{x_\mathrm{p}^4 - x_\mathrm{q}^4}{840} - \frac{x_\mathrm{p}^2 x_\mathrm{t}^3}{3360} + \cdots\right] \tag{136}$$

Let $x_1 \leq x_2 \leq x_3 \leq x_4$. Truncating the above series after four terms and following a procedure similar to the one described in Section 5.4 we infer,

$$\mathcal{E}_4 = O\left(\frac{x_\mathrm{p}^2 x_\mathrm{t}^3}{3360}\right), \quad x_4 - x_1 = 2^{-m} \Rightarrow \mathcal{E}_4 \leq O(2^{-(5m+11)}) \Rightarrow \mathtt{nsd}(\mathcal{S}_4, m) = 5m + 10 \tag{137}$$

Thus, in the series evaluation of $\exp[x_1;x_2;x_3;x_4]$ we have $5m+10$ significant digits when $(x_4 - x_1) = 2^{-m}$. The direct evaluation of the same is written as,

$$\exp[x_1;x_2;x_3;x_4] = \frac{\exp[x_2;x_3;x_4] - \exp[x_1;x_2;x_3]}{x_4 - x_1} \tag{138}$$

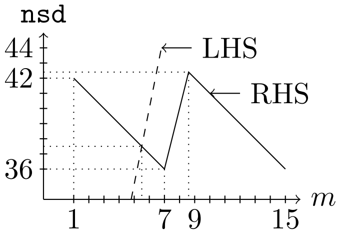
wherein the second order divided differences are evaluated using the stable formula given in Eq. (130). The subtraction $\exp[x_2;x_3;x_4] - \exp[x_1;x_2;x_3]$ yields a $O(x_4 - x_1)$ term which results in a further loss of of $m+1$ significant bits. Using Eq. (131) we infer,

$$x_4 - x_1 = 2^{-m} \Rightarrow 0 \leq x_3 - x_1 \leq 2^{-m} \tag{139}$$

$$\Rightarrow \mathtt{nsd}(\exp[x_1;x_2;x_3], m) \geq \begin{cases} 44 & \text{if } m < 7 \\ 5m+9 & \text{if } 7 \leq m < 8.6 \\ 52 & \text{if } m \geq 8.6 \end{cases} \tag{140}$$

An identical lower bound can be found for the term $\exp[x_2;x_3;x_4]$. Assuming this lower bound to be the `nsd` of the terms in the numerator we estimate the `nsd` in the direct evaluation of $\exp[x_1;x_2;x_3;x_4]$ as the result of subtracting $m+1$ from the former. We solve for m by matching the `nsd` in the series evaluation with the `nsd` in the direct evaluation.

$$5m + 10 = -(m+1) + \begin{cases} 44 & \text{if } m < 7 \\ 5m+9 & \text{if } 7 \leq m < 8.6 \\ 52 & \text{if } m \geq 8.6 \end{cases}, \tag{141}$$

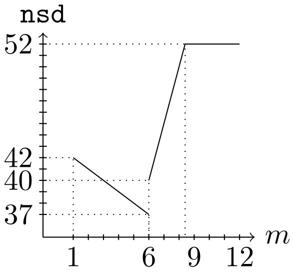


The solution to the above equation is $m = 5.5$. When $m = 5$, the `nsd` in the direct and series evaluations are 38 and 35, respectively. When $m = 6$, we obtain for the same 37 and 40, respectively. As the lower bound for the `nsd` is maximized for the integer value $m = 6$, we choose it as the rounded solution. Thus, a stable piecewise definition of $\exp[x_1;x_2;x_3;x_4]$ when $x_1 \leq x_2 \leq x_3 \leq x_4$ can be summarized as follows.

$$\exp[x_1;x_2;x_3;x_4] = \begin{cases} \dfrac{\exp[x_2;x_3;x_4] - \exp[x_1;x_2;x_3]}{x_4 - x_1} & \text{if } (x_4 - x_1) > 2^{-6} \\ \dfrac{\mathrm{e}^{x_\mathrm{a}}}{3!}\left[1 - \dfrac{x_\mathrm{p}^2}{20} + \dfrac{x_\mathrm{t}^3}{120} + \dfrac{x_\mathrm{p}^4 - x_\mathrm{q}^4}{840}\right] & \text{else} \end{cases} \tag{142}$$

Using the above piecewise definition for $\exp[x_1;x_2;x_3;x_4]$ we obtain,

$$\texttt{nsd}(\exp[x_1;x_2;x_3;x_4],m)=\begin{cases}43-m & \text{if } m<6\\ 5m+10 & \text{if } 6\le m<8.4\\ 52 & \text{if } m\ge 8.4\end{cases}, \tag{143}$$



It is straightforward to verify that $\mathrm{q}[x_1;x_2;x_3]=\exp[0;x_1;x_2;x_3]$, $\mathrm{r}[x_1;x_2]=\exp[0;0;x_1;x_2]$ and $\Phi(\star,y)[x_1;x_2]=\exp[-\mathrm{i}y;\mathrm{i}y;x_1;x_2]$. Using Eq. (142) a stable formula for $\mathrm{q}[x_1;x_2;x_3]$ is,

$$\{\widehat{x}_1,\widehat{x}_2,\widehat{x}_3,\widehat{x}_4\}=\texttt{sort}(\{0,x_1,x_2,x_3\}),\quad \mathrm{q}[x_1;x_2;x_3]=\exp[\widehat{x}_1;\widehat{x}_2;\widehat{x}_3;\widehat{x}_4]; \tag{144}$$

Likewise a stable formula for $\mathrm{r}[x_1;x_2]$ is,

$$\{\widehat{x}_1,\widehat{x}_2,\widehat{x}_3,\widehat{x}_4\}=\texttt{sort}(\{0,0,x_1,x_2\}),\quad \mathrm{r}[x_1;x_2]=\exp[\widehat{x}_1;\widehat{x}_2;\widehat{x}_3;\widehat{x}_4]; \tag{145}$$

Again $\exp[-\mathrm{i}y;\mathrm{i}y;x_1;x_2]$ deserves special attention. Following the approach taken to define $\Phi(x,y)=\exp[-\mathrm{i}y;\mathrm{i}y;x]$ in Eq. (135), a stable formula for $\Phi(\star,y)[x_1;x_2]$ can be obtained as follows.

$$\{\widehat{x}_1,\widehat{x}_2\}=\texttt{sortabs}(\{x_1,x_2\}),\quad z_1:=\widehat{x}_1+\mathrm{i}y,\quad z_2:=\widehat{x}_2+\mathrm{i}y,\quad z_{\mathrm{a}}:=\frac{\widehat{x}_1+\widehat{x}_2}{4} \tag{146}$$

$$z_{\mathrm{p}}^2:=y^2+\widehat{x}_1\widehat{x}_2-6z_{\mathrm{a}}^2,\quad z_{\mathrm{t}}^3:=2z_{\mathrm{a}}(y^2-\widehat{x}_1\widehat{x}_2+4z_{\mathrm{a}}^2),\quad z_{\mathrm{q}}^4:=(y^2+z_{\mathrm{a}}^2)(\widehat{x}_1\widehat{x}_2-3z_{\mathrm{a}}^2) \tag{147}$$

$$\Phi(\star,y)[x_1;x_2]=\begin{cases}\left[\dfrac{\exp[\widehat{x}_1;\widehat{x}_2]-\mathrm{e}^{z_1/2}\,\mathrm{sinhc}(z_1^*/2)}{z_2^*}-\Phi(\widehat{x}_1,y)\right]\dfrac{1}{z_2} & \text{if } |z_2|>2^{-6}\\ \dfrac{\mathrm{e}^{z_{\mathrm{a}}}}{3!}\left[1-\dfrac{z_{\mathrm{p}}^2}{20}+\dfrac{z_{\mathrm{t}}^3}{120}+\dfrac{z_{\mathrm{p}}^4-z_{\mathrm{q}}^4}{840}\right] & \text{else}\end{cases} \tag{148}$$

where `sortabs` is a function that sorts its arguments with respect to its absolute value, i.e. $|\widehat{x}_1|\le|\widehat{x}_2|$. The term $\Phi(\widehat{x}_1,y)$ is evaluated using the stable formula given in Eq. (135).

Finally we discuss the stable evaluation of $\exp[x_1;x_2;x_3;x_4;x_5]$, which is a template for the elements of the third subset in Eq. (84). To arrive at a stable formula for $\exp[x_1;x_2;x_3;x_4;x_5]$ we follow the same approach as was taken to define $\exp[x_1;x_2;x_3;x_4]$. Hence we summarize just the salient features. Assuming $x_1\le x_2\le x_3\le x_4\le x_5$ and considering a four term series evaluation we can infer,

$$x_5-x_1=2^{-m}\quad\Rightarrow\quad\forall x_1\le\xi\le\eta\le x_5,\quad 0\le\eta-\xi\le 2^{-m} \tag{149}$$

$$\mathcal{E}_4=O\left(\frac{x_{\mathrm{v}}^5-x_{\mathrm{p}}^2x_{\mathrm{t}}^3}{15120}\right)\quad\Rightarrow\mathcal{E}_4\le O(2^{-(5m+13)})\quad\Rightarrow\texttt{nsd}(\mathcal{S}_4,m)=5m+12 \tag{150}$$

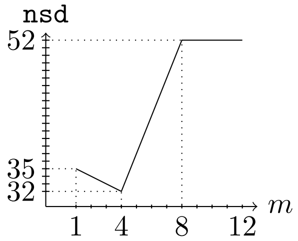
$$\Rightarrow\texttt{nsd}(\exp[x_1;x_2;x_3;x_4],m)\ge\begin{cases}37 & \text{if } m<6\\ 5m+10 & \text{if } 6\le m<8.4\\ 52 & \text{if } m\ge 8.4\end{cases} \tag{151}$$

An identical lower bound for nsd can be found for the term $\exp[x_2;x_3;x_4;x_5]$. The subtraction $\exp[x_2;x_3;x_4;x_5]-\exp[x_1;x_2;x_3;x_4]$ will result in a further loss of $m+1$ significant bits. Matching the accuracy of the series evaluation with that of the direct evaluation we obtain $m=4$. Thus a stable piecewise definition of $\exp[x_1;x_2;x_3;x_4;x_5]$ when $x_1\le x_2\le x_3\le x_4\le x_5$ can be summarized as follows.

$$\exp[x_1;x_2;x_3;x_4;x_5]=\begin{cases}\dfrac{\exp[x_2;x_3;x_4;x_5]-\exp[x_1;x_2;x_3;x_4]}{x_5-x_1} & \text{if } (x_5-x_1)>2^{-4}\\[2ex] \dfrac{\mathrm{e}^{x_{\mathrm{a}}}}{4!}\left[1-\dfrac{x_{\mathrm{p}}^2}{30}+\dfrac{x_{\mathrm{t}}^3}{210}+\dfrac{x_{\mathrm{p}}^4-x_{\mathrm{q}}^4}{1680}\right] & \text{else}\end{cases}\tag{152}$$

Using the above piecewise definition for $\exp[x_1;x_2;x_3;x_4;x_5]$ we obtain,

$$\texttt{nsd}(\exp[x_1;x_2;x_3;x_4;x_5],m)=\begin{cases}36-m & \text{if } m<4\\ 5m+12 & \text{if } 4\le m<8\\ 52 & \text{if } m\ge 8\end{cases},\tag{153}$$



It is straightforward to verify that $\mathrm{r}[x_1;x_2;x_3]=\exp[0;0;x_1;x_2;x_3]$ and $\Phi(\star,y)[x_1;x_2;x_3]=\exp[-\mathrm{i}y;\mathrm{i}y;x_1;x_2;x_3]$. Using Eq. (152) a stable formula for $\mathrm{r}[x_1;x_2;x_3]$ can be written as,

$$\{\widehat{x}_1,\widehat{x}_2,\widehat{x}_3,\widehat{x}_4,\widehat{x}_5\}=\texttt{sort}(\{0,0,x_1,x_2,x_3\}),\quad \mathrm{r}[x_1;x_2;x_3]=\exp[\widehat{x}_1;\widehat{x}_2;\widehat{x}_3;\widehat{x}_4;\widehat{x}_5];\tag{154}$$

As expected $\exp[-\mathrm{i}y;\mathrm{i}y;x_1;x_2;x_3]$ deserves special attention. Following the approach taken to define $\Phi(\star,y)[x_1;x_2]$ in Eq. (148), a stable formula for $\Phi(\star,y)[x_1;x_2;x_3]$ can be obtained as follows.

$$\{\widehat{x}_1,\widehat{x}_2,\widehat{x}_3\}=\texttt{sortabs}(\{x_1,x_2,x_3\}),\quad z_3:=\widehat{x}_3+\mathrm{i}y\tag{155}$$

$$z_{\mathrm{a}}:=\frac{\widehat{x}_1+\widehat{x}_2+\widehat{x}_3}{5},\quad z_{\mathrm{q}}^4:=6z_{\mathrm{a}}^4+\frac{1}{2}[11z_{\mathrm{a}}^2-(\widehat{x}_1^2+\widehat{x}_2^2+\widehat{x}_3^2)](y^2+3z_{\mathrm{a}}^2)-2z_{\mathrm{a}}\widehat{x}_1\widehat{x}_2\widehat{x}_3\tag{156}$$

$$z_{\mathrm{p}}^2:=\frac{1}{2}[2y^2+5z_{\mathrm{a}}^2-(\widehat{x}_1^2+\widehat{x}_2^2+\widehat{x}_3^2)],\quad z_{\mathrm{t}}^3:=\frac{z_{\mathrm{a}}}{2}[4y^2-35z_{\mathrm{a}}^2+3(\widehat{x}_1^2+\widehat{x}_2^2+\widehat{x}_3^2)]+\widehat{x}_1\widehat{x}_2\widehat{x}_3\tag{157}$$

$$\Phi(\star,y)[x_1;x_2;x_3]=\begin{cases}\left[\dfrac{\exp[\widehat{x}_1;\widehat{x}_2;\widehat{x}_3]-\exp[\mathrm{i}y;\widehat{x}_1;\widehat{x}_2]}{z_3^*}-\Phi(\star,y)[\widehat{x}_1;\widehat{x}_2]\right]\dfrac{1}{z_3} & \text{if } |z_3|>2^{-4}\\[2ex] \dfrac{\mathrm{e}^{z_{\mathrm{a}}}}{4!}\left[1-\dfrac{z_{\mathrm{p}}^2}{30}+\dfrac{z_{\mathrm{t}}^3}{210}+\dfrac{z_{\mathrm{p}}^4-z_{\mathrm{q}}^4}{1680}\right] & \text{else}\end{cases}\tag{158}$$

where the term $\Phi(\star,y)[\widehat{x}_1;\widehat{x}_2]$ is evaluated using the stable formula given in Eq. (148). Note that in the stable formula for $\Phi(\star,y)[\widehat{x}_1;\widehat{x}_2]$, just the direct evaluation of $\exp[\mathrm{i}y;\widehat{x}_1;\widehat{x}_2]$ is sufficient as the threshold value of $|z_2|$ to switch to a series evaluation is larger for the former than the latter. This means that the series evaluation of $\exp[\mathrm{i}y;\widehat{x}_1;\widehat{x}_2]$ will never be used in the stable evaluation of $\Phi(\star,y)[\widehat{x}_1;\widehat{x}_2]$. On the contrary, in the stable evaluation of

$\Phi(\star, y)[x_1; x_2; x_3]$ the switch to the series evaluation is governed by some threshold value of $|z_3|$ which includes the possibility $|z_2| \to 0$. Therefore, in Eq. (158) it is necessary to evaluate the term $\exp[\mathrm{i}y; \widehat{x}_1; \widehat{x}_2]$ in a piecewise manner.

Following Eqs. (130) and (135), a stable formula for $\exp[\mathrm{i}y; \widehat{x}_1; \widehat{x}_2]$ can be obtained as follows.

$$z_1 := \widehat{x}_1 + \mathrm{i}y, \quad z_2 := \widehat{x}_2 + \mathrm{i}y, \quad z_\mathrm{a} := \frac{\widehat{x}_1 + \widehat{x}_2 + \mathrm{i}y}{3} \tag{159}$$

$$\widetilde{z}_1 := \mathrm{i}y - z_\mathrm{a}, \quad \widetilde{z}_2 := \widehat{x}_1 - z_\mathrm{a}, \quad \widetilde{z}_3 := \widehat{x}_2 - z_\mathrm{a}, \quad z_\sigma^2 := \frac{\widetilde{z}_1^2 + \widetilde{z}_2^2 + \widetilde{z}_3^2}{3}, \quad z_\mathrm{t}^3 := \widetilde{z}_1 \widetilde{z}_2 \widetilde{z}_3 \tag{160}$$

$$\exp[\mathrm{i}y; \widehat{x}_1; \widehat{x}_2] = \begin{cases} \dfrac{\exp[\widehat{x}_1; \widehat{x}_2] - \mathrm{e}^{z_1/2} \operatorname{sinhc}(z_1^*/2)}{z_2^*} & \text{if } |z_2| > 2^{-7} \\ \dfrac{\mathrm{e}^{z_\mathrm{a}}}{2} \left[1 + \dfrac{z_\sigma^2}{8} + \dfrac{z_\mathrm{t}^3}{60} + \dfrac{z_\sigma^4}{160}\right] & \text{else} \end{cases} \tag{161}$$

6 Summary

In this section we propose a procedure to evaluate the analytical solution of the X-IVAS scheme in 3D and summarize the stable formulas associated to this procedure. First we compute the matrix $\mathbf{R}(\tau, \mathbf{A})$. If $\mathbf{A}$ has complex eigenvalues, then the divided difference coefficients (cf. Eq. (57)) in the definition of $\mathbf{R}(\tau, \mathbf{A})$ have the following structure: $\Phi(x_1, y)$, $\Phi(\star, y)[x_1; x_1]$ and $\Phi(\star, y)[x_1; x_1; x_3]$, respectively. Recall that latter three expressions are evaluated as $\exp[-\mathrm{i}y; \mathrm{i}y; x_1]$, $\exp[-\mathrm{i}y; \mathrm{i}y; x_1; x_1]$ and $\exp[-\mathrm{i}y; \mathrm{i}y; x_1; x_1; x_3]$, respectively. Using the general stable formulas given in the previous section, we summarize in seven steps the specializations[8] of the same for $\Phi(x_1, y)$, $\Phi(\star, y)[x_1; x_1]$ and $\Phi(\star, y)[x_1; x_1; x_3]$.

Step 1: Evaluate $\exp[x_1; x_1; x_3]$ as follows.

$$h := (x_3 - x_1), \quad x_\mathrm{a} = x_1 + \frac{h}{3} \tag{162}$$

$$\exp[x_1; x_1; x_3] = \begin{cases} \mathrm{e}^{x_1} \dfrac{\mathrm{q}(h) - 1}{h} & \text{if } |h| > 2^{-7} \\ \dfrac{\mathrm{e}^{x_\mathrm{a}}}{2} \left[1 + \dfrac{h^2}{36} + \dfrac{h^3}{810} + \dfrac{h^4}{3240}\right] & \text{else} \end{cases} \tag{163}$$

Step 2: Evaluate $\Phi(x_1, y) = \exp[-\mathrm{i}y; \mathrm{i}y; x_1]$ as follows.

$$z_1 := x_1 + \mathrm{i}y, \quad z_\mathrm{a} := \frac{x_1}{3}, \quad z_\sigma^2 := 2 z_\mathrm{a}^2 - \frac{2}{3} y^2, \quad z_\mathrm{t}^3 := 2 z_\mathrm{a}(z_\mathrm{a}^2 + y^2) \tag{164}$$

$$\Phi(x_1, y) = \exp[-\mathrm{i}y; \mathrm{i}y; x_1] = \begin{cases} \dfrac{\mathrm{e}^{z_1/2} \operatorname{sinhc}(z_1^*/2) - \operatorname{sinc}(y)}{z_1} & \text{if } |z_1| > 2^{-7} \\ \dfrac{\mathrm{e}^{z_\mathrm{a}}}{2} \left[1 + \dfrac{z_\sigma^2}{8} + \dfrac{z_\mathrm{t}^3}{60} + \dfrac{z_\sigma^4}{160}\right] & \text{else} \end{cases} \tag{165}$$

[8] Note that repeated values appear here (symbolically) as input for the divided differences. This facilitates further simplifications in the piecewise evaluations without compromising its accuracy/stability.

Step 3: Store $\widehat{x}_1, \widehat{x}_3$ such that $\{\widehat{x}_1, \widehat{x}_3\} = \texttt{sortabs}(\{x_1, x_3\})$ and evaluate $\exp[\mathrm{i}y; \widehat{x}_1; x_1]$ as follows.

$$\widehat{z}_1 := \widehat{x}_1 + \mathrm{i}y, \quad z_1 := x_1 + \mathrm{i}y, \quad z_\mathrm{a} := \frac{\widehat{x}_1 + x_1 + \mathrm{i}y}{3} \tag{166}$$

$$\widetilde{z}_1 := \mathrm{i}y - z_\mathrm{a}, \quad \widetilde{z}_2 := \widehat{x}_1 - z_\mathrm{a}, \quad \widetilde{z}_3 := x_1 - z_\mathrm{a}, \quad z_\sigma^2 := \frac{\widetilde{z}_1^2 + \widetilde{z}_2^2 + \widetilde{z}_3^2}{3}, \quad z_\mathrm{t}^3 := \widetilde{z}_1 \widetilde{z}_2 \widetilde{z}_3 \tag{167}$$

$$\exp[\mathrm{i}y; \widehat{x}_1; x_1] = \begin{cases} \dfrac{\exp[\widehat{x}_1; x_1] - \mathrm{e}^{\widehat{z}_1/2} \operatorname{sinhc}(\widehat{z}_1^*/2)}{z_1^*} & \text{if } |z_1| > 2^{-7} \\ \dfrac{\mathrm{e}^{z_\mathrm{a}}}{2} \left[1 + \dfrac{z_\sigma^2}{8} + \dfrac{z_\mathrm{t}^3}{60} + \dfrac{z_\sigma^4}{160}\right] & \text{else} \end{cases} \tag{168}$$

Note that threshold condition to switch from the direct evalaution to the series evaluation is the same as in Step 2. Hence, in a computer program Step 2 and Step 3 be implemented in the same conditional scope.

Step 4: Evaluate $\Phi(\widehat{x}_1, y) = \exp[-\mathrm{i}y; \mathrm{i}y; \widehat{x}_1]$ as follows.

$$\widehat{z}_1 := \widehat{x}_1 + \mathrm{i}y, \quad z_\mathrm{a} := \frac{\widehat{x}_1}{3}, \quad z_\sigma^2 := 2z_\mathrm{a}^2 - \frac{2}{3}y^2, \quad z_\mathrm{t}^3 := 2z_\mathrm{a}(z_\mathrm{a}^2 + y^2) \tag{169}$$

$$\Phi(\widehat{x}_1, y) = \exp[-\mathrm{i}y; \mathrm{i}y; \widehat{x}_1] = \begin{cases} \dfrac{\mathrm{e}^{\widehat{z}_1/2} \operatorname{sinhc}(\widehat{z}_1^*/2) - \operatorname{sinc}(y)}{\widehat{z}_1} & \text{if } |\widehat{z}_1| > 2^{-7} \\ \dfrac{\mathrm{e}^{z_\mathrm{a}}}{2} \left[1 + \dfrac{z_\sigma^2}{8} + \dfrac{z_\mathrm{t}^3}{60} + \dfrac{z_\sigma^4}{160}\right] & \text{else} \end{cases} \tag{170}$$

Step 5: Evaluate $\Phi(\star, y)[x_1; x_1] = \exp[-\mathrm{i}y; \mathrm{i}y; x_1; x_1]$ as follows.

$$z_1 := x_1 + \mathrm{i}y, \quad z_\mathrm{a} := \frac{x_1}{2}, \quad z_\mathrm{p}^2 := y^2 - 2z_\mathrm{a}^2, \quad z_\mathrm{t}^3 := 2z_\mathrm{a}y^2, \quad z_\mathrm{q}^4 := z_\mathrm{a}^2(z_\mathrm{a}^2 + y^2) \tag{171}$$

$$\Phi(\star, y)[x_1; x_1] = \begin{cases} \left[\dfrac{\mathrm{e}^{x_1} - \mathrm{e}^{z_1/2} \operatorname{sinhc}(z_1^*/2)}{z_1^*} - \Phi(x_1, y)\right] \dfrac{1}{z_1} & \text{if } |z_1| > 2^{-6} \\ \dfrac{\mathrm{e}^{z_\mathrm{a}}}{3!} \left[1 - \dfrac{z_\mathrm{p}^2}{20} + \dfrac{z_\mathrm{t}^3}{120} + \dfrac{z_\mathrm{p}^4 - z_\mathrm{q}^4}{840}\right] & \text{else} \end{cases} \tag{172}$$

Note that the evaluation of $\Phi(x_1, y)$ is already done in Step 2.

Step 6: Evaluate $\Phi(\star, y)[\widehat{x}_1; x_1] = \exp[-\mathrm{i}y; \mathrm{i}y; \widehat{x}_1; x_1]$ as follows.

$$\widehat{z}_1 := \widehat{x}_1 + \mathrm{i}y, \quad z_1 := x_1 + \mathrm{i}y, \quad z_\mathrm{a} := \frac{\widehat{x}_1 + x_1}{4} \tag{173}$$

$$z_\mathrm{p}^2 := y^2 + \widehat{x}_1 x_1 - 6z_\mathrm{a}^2, \quad z_\mathrm{t}^3 := 2z_\mathrm{a}(y^2 - \widehat{x}_1 x_1 + 4z_\mathrm{a}^2), \quad z_\mathrm{q}^4 := (y^2 + z_\mathrm{a}^2)(\widehat{x}_1 x_1 - 3z_\mathrm{a}^2) \tag{174}$$

$$\Phi(\star, y)[\widehat{x}_1; x_1] = \begin{cases} \left[\dfrac{\exp[\widehat{x}_1; x_1] - \mathrm{e}^{\widehat{z}_1/2} \operatorname{sinhc}(\widehat{z}_1^*/2)}{z_1^*} - \Phi(\widehat{x}_1, y)\right] \dfrac{1}{z_1} & \text{if } |z_1| > 2^{-6} \\ \dfrac{\mathrm{e}^{z_\mathrm{a}}}{3!} \left[1 - \dfrac{z_\mathrm{p}^2}{20} + \dfrac{z_\mathrm{t}^3}{120} + \dfrac{z_\mathrm{p}^4 - z_\mathrm{q}^4}{840}\right] & \text{else} \end{cases} \tag{175}$$

Note that the evaluation of $\Phi(\widehat{x}_1, y)$ is already done in Step 4. Further, the condition to switch from the direct evaluation to the series evaluation is the same as in Step 5. Hence, in a computer program Step 5 and Step 6 can be implemented in the same conditional scope.

Step 7: Evaluate $\Phi(\star, y)[x_1; x_1; x_3] = \exp[-\mathrm{i}y; \mathrm{i}y; x_1; x_1; x_3] = \exp[-\mathrm{i}y; \mathrm{i}y; \widehat{x}_1; x_1; \widehat{x}_3]$ as follows.

$$\widehat{z}_3 := \widehat{x}_3 + \mathrm{i}y, \quad z_\mathrm{a} := \frac{\widehat{x}_1 + x_1 + \widehat{x}_3}{5}, \quad z_\mathrm{p}^2 := \frac{1}{2}[2y^2 + 5z_\mathrm{a}^2 - (\widehat{x}_1^2 + x_1^2 + \widehat{x}_3^2)] \tag{176}$$

$$z_\mathrm{t}^3 := \frac{z_\mathrm{a}}{2}[4y^2 - 35z_\mathrm{a}^2 + 3(\widehat{x}_1^2 + x_1^2 + \widehat{x}_3^2)] + \widehat{x}_1 x_1 \widehat{x}_3 \tag{177}$$

$$z_\mathrm{q}^4 := 6z_\mathrm{a}^4 + \frac{1}{2}[11z_\mathrm{a}^2 - (\widehat{x}_1^2 + x_1^2 + \widehat{x}_3^2)](y^2 + 3z_\mathrm{a}^2) - 2z_\mathrm{a}\widehat{x}_1 x_1 \widehat{x}_3 \tag{178}$$

$$\Phi(\star, y)[x_1; x_1; x_3] = \begin{cases} \left[\dfrac{\exp[x_1; x_1; x_3] - \exp[\mathrm{i}y; \widehat{x}_1; x_1]}{\widehat{z}_3^*} - \Phi(\star, y)[\widehat{x}_1; x_1]\right]\dfrac{1}{\widehat{z}_3} & \text{if } |\widehat{z}_3| > 2^{-4} \\ \dfrac{\mathrm{e}^{z_\mathrm{a}}}{4!}\left[1 - \dfrac{z_\mathrm{p}^2}{30} + \dfrac{z_\mathrm{t}^3}{210} + \dfrac{z_\mathrm{p}^4 - z_\mathrm{q}^4}{1680}\right] & \text{else} \end{cases} \tag{179}$$

Note that the evaluation of $\exp[x_1; x_1; x_3]$, $\exp[\mathrm{i}y; \widehat{x}_1; x_1]$ and $\Phi(\star, y)[\widehat{x}_1; x_1]$ is already done in Step 1, Step 3 and Step 4, respectively.

If $\mathbf{A}$ has real eigenvalues, then the divided difference coefficients (cf. Eq. (46)) in the definition of $\mathbf{R}(\tau, \mathbf{A})$ have the following structure: $\mathrm{r}(x_1)$, $\mathrm{r}[x_1; x_2]$ and $\mathrm{r}[x_1; x_2; x_3]$, respectively. Recall that the formulas to compute the eigenvalues guarantee a sorted input data, i.e. $x_1 \leq x_2 \leq x_3\}$. Further, $\mathrm{r}(x_1)$, $\mathrm{r}[x_1; x_2]$ and $\mathrm{r}[x_1; x_2; x_3]$ is evaluated as $\exp[0; 0; x_1]$, $\exp[0; 0; x_1; x_2]$ and $\exp[0; 0; x_1; x_2; x_3]$, respectively. To avoid sorting the input data augmented with zeros in the latter three expressions, we propose an alternate yet stable approach [9]. Recall that the definition of $\mathbf{R}(\tau, \mathbf{A})$, cf. Eq. (46), is independent of the ordering of the eigenvalues. So we may sort the eigenvalues with respect to their modulus and define $\mathbf{R}(\tau, \mathbf{A})$ using these sorted eigenvalues. In this alternate definition, the divided difference coefficients will have the following structure: $\mathrm{r}(\widehat{x}_1)$, $\mathrm{r}[\widehat{x}_1; \widehat{x}_2]$ and $\mathrm{r}[\widehat{x}_1; \widehat{x}_2; \widehat{x}_3]$, respectively and where $\{\widehat{x}_1, \widehat{x}_2, \widehat{x}_3\} = \texttt{sortabs}(\{x_1, x_2, x_3\})$. Then we may follow an approach similar to the one taken in the case of complex eigenvalues and fixing the imaginary parts to be zero. We summarize the evaluation of $\mathrm{r}(\widehat{x}_1)$, $\mathrm{r}[\widehat{x}_1; \widehat{x}_2]$ and $\mathrm{r}[\widehat{x}_1; \widehat{x}_2; \widehat{x}_3]$ in five steps.

Step 1: Evaluate $\exp[x_1; x_2; x_3]$ using the definition given in Eq. (130) prior to sorting the input data using `sortabs`.

Step 2: Evaluate $\exp[0; 0; \widehat{x}_1]$ as follows.

$$x_\mathrm{a} := \frac{\widehat{x}_1}{3}, \quad \exp[0; 0; \widehat{x}_1] = \begin{cases} \dfrac{\exp[0; \widehat{x}_1] - 1}{\widehat{x}_1} & \text{if } |\widehat{x}_1| > 2^{-7} \\ \dfrac{\mathrm{e}^{x_\mathrm{a}}}{2}\left[1 + \dfrac{x_\mathrm{a}^2}{4} + \dfrac{x_\mathrm{a}^3}{30} + \dfrac{x_\mathrm{a}^4}{40}\right] & \text{else} \end{cases} \tag{180}$$

[9] This approach is stable only when at least one of the input data is zero. Otherwise, we loose control of the truncation error in the series approximation.

Step 3: Evaluate $\exp[0;\widehat{x}_1;\widehat{x}_2]$ as follows.

$$x_{\mathrm{a}} := \frac{\widehat{x}_1+\widehat{x}_2}{3}, \quad \widetilde{x}_1 := -x_{\mathrm{a}}, \quad \widetilde{x}_2 := \widehat{x}_1 - x_{\mathrm{a}}, \quad \widetilde{x}_3 := \widehat{x}_2 - x_{\mathrm{a}} \tag{181}$$

$$x_\sigma^2 := \frac{\widetilde{x}_1^2+\widetilde{x}_2^2+\widetilde{x}_3^2}{3}, \quad x_{\mathrm{t}}^3 := \widetilde{x}_1\widetilde{x}_2\widetilde{x}_3 \tag{182}$$

$$\exp[0;\widehat{x}_1;\widehat{x}_2] = \begin{cases} \dfrac{\exp[\widehat{x}_1;\widehat{x}_2]-\exp[0;\widehat{x}_1]}{\widehat{x}_2} & \text{if } |\widehat{x}_2| > 2^{-7} \\ \dfrac{\mathrm{e}^{x_{\mathrm{a}}}}{2}\left[1+\dfrac{x_\sigma^2}{8}+\dfrac{x_{\mathrm{t}}^3}{60}+\dfrac{x_\sigma^4}{160}\right] & \text{else} \end{cases} \tag{183}$$

Step 4: Evaluate $\exp[0;0;\widehat{x}_1;\widehat{x}_2]$ as follows.

$$x_{\mathrm{a}} = \frac{\widehat{x}_1+\widehat{x}_2}{4}, \quad x_{\mathrm{p}}^2 = \widehat{x}_1\widehat{x}_2 - 6x_{\mathrm{a}}^2, \quad x_{\mathrm{t}}^3 = 2x_{\mathrm{a}}(4x_{\mathrm{a}}^2 - \widehat{x}_1\widehat{x}_2), \quad x_{\mathrm{q}}^4 = x_{\mathrm{a}}^2(\widehat{x}_1\widehat{x}_2 - 3x_{\mathrm{a}}^2) \tag{184}$$

$$\exp[0;0;\widehat{x}_1;\widehat{x}_2] = \begin{cases} \dfrac{\exp[0;\widehat{x}_1;\widehat{x}_2]-\exp[0;0;\widehat{x}_1]}{\widehat{x}_2} & \text{if } |\widehat{x}_2| > 2^{-6} \\ \dfrac{\mathrm{e}^{x_{\mathrm{a}}}}{3!}\left[1-\dfrac{x_{\mathrm{p}}^2}{20}+\dfrac{x_{\mathrm{t}}^3}{120}+\dfrac{x_{\mathrm{p}}^4-x_{\mathrm{q}}^4}{840}\right] & \text{else} \end{cases} \tag{185}$$

Note that the evaluation of $\exp[0;\widehat{x}_1;\widehat{x}_2]$ and $\exp[0;0;\widehat{x}_1]$ is already done in Step 3 and Step 2, respectively.

Step 5: Evaluate $\exp[0;0;\widehat{x}_1;\widehat{x}_2;\widehat{x}_3]$ as follows.

$$x_{\mathrm{a}} := \frac{\widehat{x}_1+\widehat{x}_2+\widehat{x}_3}{5}, \quad x_{\mathrm{q}}^4 := 6x_{\mathrm{a}}^4 + \frac{3x_{\mathrm{a}}^2}{2}[11x_{\mathrm{a}}^2 - (\widehat{x}_1^2+\widehat{x}_2^2+\widehat{x}_3^2)] - 2x_{\mathrm{a}}\widehat{x}_1\widehat{x}_2\widehat{x}_3 \tag{186}$$

$$x_{\mathrm{p}}^2 := \frac{1}{2}[5x_{\mathrm{a}}^2 - (\widehat{x}_1^2+\widehat{x}_2^2+\widehat{x}_3^2)], \quad x_{\mathrm{t}}^3 := \frac{x_{\mathrm{a}}}{2}[3(\widehat{x}_1^2+\widehat{x}_2^2+\widehat{x}_3^2) - 35x_{\mathrm{a}}^2] + \widehat{x}_1\widehat{x}_2\widehat{x}_3 \tag{187}$$

$$\exp[0;0;\widehat{x}_1;\widehat{x}_2;\widehat{x}_3] = \begin{cases} \left[\dfrac{\exp[x_1;x_2;x_3]-\exp[0;\widehat{x}_1;\widehat{x}_2]}{\widehat{x}_3} - \exp[0;0;\widehat{x}_1;\widehat{x}_2]\right]\dfrac{1}{\widehat{x}_3} & \text{if } |\widehat{x}_3| > 2^{-4} \\ \dfrac{\mathrm{e}^{x_{\mathrm{a}}}}{4!}\left[1-\dfrac{x_{\mathrm{p}}^2}{30}+\dfrac{x_{\mathrm{t}}^3}{210}+\dfrac{x_{\mathrm{p}}^4-x_{\mathrm{q}}^4}{1680}\right] & \text{else} \end{cases} \tag{188}$$

Note that the evaluation of $\exp[x_1;x_2;x_3]$, $\exp[0;\widehat{x}_1;\widehat{x}_2]$ and $\exp[0;0;\widehat{x}_1;\widehat{x}_2]$ is already done in Step 1, Step 3 and Step 4, respectively.

Using the evaluation of $\mathbf{R}(\tau,\mathbf{A})$, we can evaluate $\mathbf{Q}(\tau,\mathbf{A})$ and $\mathbf{P}(\tau,\mathbf{A})$ using the relationships given in Eqs. (20) and (22).

7 Examples

We present two examples to validate the numerical stability in the evaluation of the proposed formulas for the X-IVAS scheme. In these examples the eigenvalues of the matrix $\mathbf{A}$ and the gap between them gradually tends to zero. The symbolic evaluation of the formulas for the chosen eigenvalues are done using Maple and the first 16 significant decimal digits are stored as reference solutions. These reference solutions are used to measure the relative error in the formula evaluations using double precision floating point arithmetic. In this way, we study the stability of the formulas in the neighbourhood of the removable singularities. These examples might not be representative of typical situations which one encounters in practice. However particular instances of the considered situations do occur occasionally[10].

7.1 Example 1

In this example we consider the case when two of the eigenvalues of the matrix $\mathbf{A}$ are complex numbers. We define $h := 10^{-n}$ and choose $n \in \{1, 2, 3, \cdots, 15\}$. For each h, we define the matrix $\mathbf{A}$ as follows.

$$\mathbf{A} := \begin{bmatrix} a+h & -2h & b \\ 2h & a+h & c \\ 0 & 0 & a+dh \end{bmatrix} \Rightarrow \texttt{eigs}(\mathbf{A}) = \{a+h \pm \mathrm{i}2h, a+dh\} \tag{189}$$

where $\texttt{eigs}(\mathbf{A})$ represents the eigenvalues of $\mathbf{A}$. As per the chosen notation we identify $\alpha = a+h$, $\beta = 2h$ and $\lambda_3 = a+dh$. We can drive all the eigenvalues and/or the gap betwen them to zero by appropriately choosing the parameters a and d. For each $\mathbf{A}$ we evaluate $\mathbf{P}(\tau, \mathbf{A})$, $\mathbf{Q}(\tau, \mathbf{A})$ and $\mathbf{R}(\tau, \mathbf{A})$ using the stable formulas summarized in the previous section.

For comparison, we also evaluate the formula obtained for $\mathbf{R}(\tau, \mathbf{A})$ by integrating twice the formula for $\mathbf{P}(\tau, \mathbf{A})$ given in the Eq. (55). However, unlike in Eqs. (56) and (57), here we perform the integrals without rearranging the coefficients in the obtained integrals as divided differences. Following this line, we obtain the following alternate but algebraically equivalent formula[11] for $\mathbf{R}(\tau, \mathbf{A})$.

$$\begin{aligned} \mathbf{R}(\tau, \mathbf{A}) &= \tau^2 \mathrm{e}^{\tau\alpha}\, \Phi(-\tau\alpha, \tau\beta)\mathbf{I} + \tau^3\, \Upsilon(-\tau\alpha, \tau\beta)(\mathbf{A} - 2\alpha\mathbf{I}) \\ &\quad + \tau^4 \left[\frac{\mathrm{r}(\tau\lambda_3) - \tau(\lambda_3 - 2\alpha)\, \Upsilon(-\tau\alpha, \tau\beta) - \mathrm{e}^{\tau\alpha}\, \Phi(-\tau\alpha, \tau\beta)}{\tau^2[(\lambda_3 - \alpha)^2 + \beta^2]} \right] [(\mathbf{A} - \alpha\mathbf{I})^2 + \beta^2\mathbf{I}] \end{aligned} \tag{190}$$

where the auxiliary function $\Upsilon(x, y)$ is defined as follows.

$$\Upsilon(x, y) := \frac{1 - \mathrm{e}^x[\mathrm{sinc}(y) + 2x\, \Phi(x, y)]}{x^2 + y^2} \tag{191}$$

The formula for $\mathbf{R}(\tau, \mathbf{A})$ given in Eq. (190) and the *as is* evaluation of the same are denoted as the *usual formula* and the *usual evaluation*, respectively. The matrices $\mathbf{Q}(\tau, \mathbf{A})$ and $\mathbf{P}(\tau, \mathbf{A})$ are evaluated using the relationships $\mathbf{Q}(\tau, \mathbf{A}) = \mathbf{R}(\tau, \mathbf{A}) \cdot \mathbf{A} + \tau\mathbf{I}$ and $\mathbf{P}(\tau, \mathbf{A}) = \mathbf{Q}(\tau, \mathbf{A}) \cdot \mathbf{A} + \mathbf{I}$, respectively.

First we examine the evaluation of the coefficient $\Phi(\star, \tau\beta)[-\tau\alpha; -\tau\alpha; \tau\lambda_3 - \tau\alpha]$ using both the usual and the stable formulas. This coefficient in the usual formula is,

$$\Phi(\star, \tau\beta)[-\tau\alpha; -\tau\alpha; \tau\lambda_3 - \tau\alpha] = \left[\frac{\mathrm{r}(\tau\lambda_3) - \tau(\lambda_3 - 2\alpha)\, \Upsilon(-\tau\alpha, \tau\beta) - \mathrm{e}^{\tau\alpha}\, \Phi(-\tau\alpha, \tau\beta)}{[(\lambda_3 - \alpha)^2 + \beta^2]\mathrm{e}^{\tau\alpha}} \right] \tag{192}$$

Choosing $\tau = 1$, $a = 0$ and $d = 4$ the coefficient simplifies to $\Phi(\star, 2h)[-h; -h; 3h]$ and its evaluation using both the usual and stable formulas is shown in Table 3. Therein, the first

[10] This study originated due to the numerical instabilities found in the unit testing phase

[11] We assume that one would have derived this algebraically equivalent formula should he/she be unaware of the numerical instabilities in their evaluation

h	Usual evaluation	Exact 16 significant digits $\times 10^2$	Stable evaluation $\times 10^2$
10^{-01}	$+4.252\,986\,132\,164\,403 \times 10^{-02}$	$4.252\,986\,132\,162\,584$	$4.252\,986\,132\,163\,262$
10^{-02}	$+4.175\,027\,955\,863\,544 \times 10^{-02}$	$4.175\,027\,977\,160\,363$	$4.175\,027\,977\,159\,309$
10^{-03}	$+4.167\,319\,156\,849\,885 \times 10^{-02}$	$4.167\,500\,277\,976\,287$	$4.167\,500\,277\,976\,286$
10^{-04}	$+5.724\,100\,125\,516\,537 \times 10^{-02}$	$4.166\,750\,002\,777\,976$	$4.166\,750\,002\,777\,976$
10^{-05}	$-2.993\,217\,938\,464\,562 \times 10^{+01}$	$4.166\,675\,000\,027\,777$	$4.166\,675\,000\,027\,778$
10^{-06}	$+7.859\,720\,003\,935\,763 \times 10^{+05}$	$4.166\,667\,500\,000\,277$	$4.166\,667\,500\,000\,277$
10^{-07}	$-9.042\,591\,902\,568\,542 \times 10^{+09}$	$4.166\,666\,750\,000\,002$	$4.166\,666\,750\,000\,002$
10^{-08}	$-9.088\,426\,027\,317\,757 \times 10^{+13}$	$4.166\,666\,675\,000\,000$	$4.166\,666\,674\,999\,999$
10^{-09}	$-7.139\,611\,827\,083\,238 \times 10^{+17}$	$4.166\,666\,667\,500\,000$	$4.166\,666\,667\,499\,999$
10^{-10}	$+2.298\,194\,888\,769\,120 \times 10^{+21}$	$4.166\,666\,666\,750\,000$	$4.166\,666\,666\,749\,999$
10^{-11}	$+2.291\,964\,121\,707\,557 \times 10^{+24}$	$4.166\,666\,666\,675\,000$	$4.166\,666\,666\,674\,999$
10^{-12}	$-6.126\,014\,110\,580\,096 \times 10^{+29}$	$4.166\,666\,666\,667\,500$	$4.166\,666\,666\,667\,499$
10^{-13}	$+8.610\,789\,808\,137\,885 \times 10^{+33}$	$4.166\,666\,666\,666\,750$	$4.166\,666\,666\,666\,749$
10^{-14}	$-2.213\,384\,780\,309\,904 \times 10^{+37}$	$4.166\,666\,666\,666\,675$	$4.166\,666\,666\,666\,674$
10^{-15}	$-2.213\,384\,780\,372\,381 \times 10^{+40}$	$4.166\,666\,666\,666\,667$	$4.166\,666\,666\,666\,667$

Table 3: Significant digits that match the exact evaluation of $\Phi(\star, 2h)[-h; -h; 3h]$.

16 significant decimal digits of the usual evaluation and the stable evaluation are compared to those of an exact evaluation. The evaluations are done using double precision floating point arithmetic. The significant digits in both formula evaluations that differ from the exact values are highlighted in green colour. Note that we loose significant digits at a fourth-order rate in the usual evaluation. This is expected as $\Phi(\star, 2h)[-h; -h; 3h]$ can be expressed as a fourth-order divided difference. The established lower bound of at least 10 significant digits is reproduced in the stable evaluations. We have not considered the effect of carry-over digits while highlighting those that differ from the exact evaluation. This explains why for smaller values of h in the stable evaluation we have outliers in the highlighting pattern despite those being accurate up to machine precision.

Table 4 illustrates the relative errors in the evaluations of the matrices $\mathbf{P}(\tau, \mathbf{A})$, $\mathbf{Q}(\tau, \mathbf{A})$ and $\mathbf{R}(\tau, \mathbf{A})$ choosing $\tau = 1$, $a = 0$, $b = c = 1$ and $d = 4$. The notation $\mathbf{R}^{\mathrm{u}}$ and $\mathbf{R}^{\mathrm{s}}$ denotes the usual and stable evaluations of the matrix $\mathbf{R}$, respectively and the norm used in $||\mathbf{R}||$ is the Frobenius norm. The relative errors $(||\mathbf{P}^{\mathrm{s}} - \mathbf{P}||/||\mathbf{P}||)$, $(||\mathbf{Q}^{\mathrm{s}} - \mathbf{Q}||/||\mathbf{Q}||)$ and $(||\mathbf{R}^{\mathrm{s}} - \mathbf{R}||/||\mathbf{R}||)$ are found to be within the guaranteed evaluation accuracies established for the same and reflect the robustness of the stable formulas. The gradual loss of significance as $h \to 0$ is reflected as a gradual increase in the relative error (from machine epsilon to values intolerably high) in the usual evaluations of the considered matrices. The maximum relative error in the usual evaluations of $\mathbf{P}^{\mathrm{u}}$, $\mathbf{Q}^{\mathrm{u}}$ and $\mathbf{R}^{\mathrm{u}}$ are of the order of 10^{-4}, 10^{10} and 10^{25}, respectively. In other words, as $h \to 0$ we observe $(||\mathbf{P}^{\mathrm{u}} - \mathbf{P}||/||\mathbf{P}||)$ is $O(h)$ times smaller than $(||\mathbf{Q}^{\mathrm{u}} - \mathbf{Q}||/||\mathbf{Q}||)$ which in turn is $O(h)$ times smaller than $(||\mathbf{R}^{\mathrm{u}} - \mathbf{R}||/||\mathbf{R}||)$ which in turn is $O(h)$ times smaller than $\Phi(\star, 2h)[-h; -h; 3h]$. The following results explain this

h	$\frac{\lvert\lvert\mathbf{R}^{\mathrm{u}}-\mathbf{R}\rvert\rvert}{\lvert\lvert\mathbf{R}\rvert\rvert}$	$\frac{\lvert\lvert\mathbf{R}^{\mathrm{s}}-\mathbf{R}\rvert\rvert}{\lvert\lvert\mathbf{R}\rvert\rvert}$	$\frac{\lvert\lvert\mathbf{Q}^{\mathrm{u}}-\mathbf{Q}\rvert\rvert}{\lvert\lvert\mathbf{Q}\rvert\rvert}$	$\frac{\lvert\lvert\mathbf{Q}^{\mathrm{s}}-\mathbf{Q}\rvert\rvert}{\lvert\lvert\mathbf{Q}\rvert\rvert}$	$\frac{\lvert\lvert\mathbf{P}^{\mathrm{u}}-\mathbf{P}\rvert\rvert}{\lvert\lvert\mathbf{P}\rvert\rvert}$	$\frac{\lvert\lvert\mathbf{P}^{\mathrm{s}}-\mathbf{P}\rvert\rvert}{\lvert\lvert\mathbf{P}\rvert\rvert}$
10^{-01}	6.1×10^{-15}	7.4×10^{-15}	5.3×10^{-16}	1.1×10^{-15}	1.1×10^{-16}	2.6×10^{-16}
10^{-02}	6.7×10^{-12}	4.8×10^{-13}	6.9×10^{-14}	7.5×10^{-15}	1.1×10^{-15}	9.7×10^{-17}
10^{-03}	5.7×10^{-09}	4.3×10^{-17}	6.1×10^{-12}	$0.0 \times 10^{+00}$	1.1×10^{-14}	9.9×10^{-17}
10^{-04}	4.9×10^{-06}	6.9×10^{-17}	5.2×10^{-10}	$0.0 \times 10^{+00}$	9.8×10^{-14}	$0.0 \times 10^{+00}$
10^{-05}	9.4×10^{-04}	1.2×10^{-16}	1.0×10^{-08}	$0.0 \times 10^{+00}$	1.8×10^{-13}	2.1×10^{-21}
10^{-06}	$2.4 \times 10^{+00}$	2.1×10^{-16}	2.6×10^{-06}	5.9×10^{-17}	4.9×10^{-12}	9.9×10^{-17}
10^{-07}	$2.8 \times 10^{+03}$	3.0×10^{-17}	3.0×10^{-04}	8.3×10^{-17}	5.7×10^{-11}	9.9×10^{-17}
10^{-08}	$2.8 \times 10^{+06}$	$0.0 \times 10^{+00}$	3.0×10^{-02}	8.3×10^{-17}	5.7×10^{-10}	9.9×10^{-17}
10^{-09}	$2.2 \times 10^{+09}$	3.0×10^{-17}	$2.4 \times 10^{+00}$	5.9×10^{-17}	4.5×10^{-09}	9.9×10^{-17}
10^{-10}	$7.2 \times 10^{+11}$	$0.0 \times 10^{+00}$	$7.7 \times 10^{+01}$	$0.0 \times 10^{+00}$	1.4×10^{-08}	$0.0 \times 10^{+00}$
10^{-11}	$7.2 \times 10^{+13}$	$0.0 \times 10^{+00}$	$7.7 \times 10^{+02}$	1.2×10^{-27}	1.4×10^{-08}	2.0×10^{-27}
10^{-12}	$1.9 \times 10^{+18}$	2.1×10^{-16}	$2.0 \times 10^{+06}$	5.9×10^{-17}	3.8×10^{-06}	9.9×10^{-17}
10^{-13}	$2.7 \times 10^{+21}$	2.1×10^{-16}	$2.9 \times 10^{+08}$	5.9×10^{-17}	5.4×10^{-05}	9.9×10^{-17}
10^{-14}	$6.9 \times 10^{+23}$	6.2×10^{-31}	$7.4 \times 10^{+09}$	$0.0 \times 10^{+00}$	1.4×10^{-04}	1.9×10^{-30}
10^{-15}	$6.9 \times 10^{+25}$	3.0×10^{-17}	$7.4 \times 10^{+10}$	1.4×10^{-31}	1.4×10^{-04}	$0.0 \times 10^{+00}$

Table 4: Relative errors in the usual and stable evaluation when $\lambda = \{h \pm \mathrm{i}2h, 4h\}$.

behaviour.

$$\mathbf{Z} := (\mathbf{A} - \alpha\mathbf{I})^2 + \beta^2\mathbf{I} = h \begin{bmatrix} 0 & 0 & b(d-1)+ch \\ 0 & 0 & c(d-1)+bh \\ 0 & 0 & dh(d-2)+5h \end{bmatrix} \Rightarrow \begin{array}{c} \mathbf{ZA} = (a+dh)h\mathbf{Z} \\ \mathbf{ZA}^2 = (a+dh)^2h\mathbf{Z} \end{array} \tag{193}$$

$$\mathbf{R}(\tau,\mathbf{A}) \approx \Phi(\star,\tau\beta)[-\tau\alpha;-\tau\alpha;\tau\lambda_3-\tau\alpha]\mathbf{Z}, \quad \mathbf{Q}(\tau,\mathbf{A}) \approx \mathbf{RA}, \quad \mathbf{P}(\tau,\mathbf{A}) \approx \mathbf{RA}^2 \tag{194}$$

Equation (194) holds when the evaluation error of $\Phi(\star,\tau\beta)[-\tau\alpha;-\tau\alpha;\tau\lambda_3-\tau\alpha]$ is large; Observe in Table 3 that $\Phi(\star,2h)[-h;-h;3h] \approx 10^{40}$ when $h \to 0$. Substituting $a = 0$, $b = c = 1$ and $d = 4$ in the above equations we can arrive at the maximum relative errors found in the usual evaluations of the matrices. Recall that the matrices $\mathbf{P}(\tau,\mathbf{A})$ and $\mathbf{Q}(\tau,\mathbf{A})$ govern the evolution of the particle positions. Likewise, the matrices $\mathbf{Q}(\tau,\mathbf{A})$ and $\mathbf{R}(\tau,\mathbf{A})$ govern the evolution of the particle velocities.

7.2 Example 2

In this example we consider the case when all the eigenvalues of the matrix $\mathbf{A}$ are real numbers. We consider the same sequence for h used in previous example. The matrix $\mathbf{A}$ is defined as follows.

$$\mathbf{A} := \begin{bmatrix} a+h & b & c \\ 0 & a+2h & d \\ 0 & 0 & a+3h \end{bmatrix} \Rightarrow \texttt{eigs}(\mathbf{A}) = \{a+h, a+2h, a+3h\} \tag{195}$$

As per the chosen notation we identify $\lambda_1 = a+h$, $\lambda_2 = a+2h$ and $\lambda_3 = a+3h$. By construction, all the eigenvalues and the gap between them can be driven to zero with decreasing values of h for appropriate choice of the parameter a. For each $\mathbf{A}$ we evaluate $\mathbf{P}(\tau,\mathbf{A})$, $\mathbf{Q}(\tau,\mathbf{A})$ and $\mathbf{R}(\tau,\mathbf{A})$ using the stable formulas summarized in the previous section.

h	Usual evaluation	Exact 16 significant digits $\times 10^2$	Stable evaluation $\times 10^2$
10^{-01}	$+4.703\,252\,003\,753\,182 \times 10^{-02}$	4.703 252 003 748 591	4.703 252 003 756 346
10^{-02}	$+4.217\,015\,681\,096\,254 \times 10^{-02}$	4.217 015 682 095 156	4.217 015 682 094 985
10^{-03}	$+4.171\,659\,312\,435\,238 \times 10^{-02}$	4.171 670 140 675 349	4.171 670 140 675 349
10^{-04}	$+4.170\,737\,644\,124\,122 \times 10^{-02}$	4.167 166 701 390 674	4.167 166 701 390 674
10^{-05}	$+7.401\,468\,326\,839\,742 \times 10^{-02}$	4.166 716 667 013 890	4.166 716 667 013 890
10^{-06}	$+7.401\,484\,980\,461\,280 \times 10^{+01}$	4.166 671 666 670 138	4.166 671 666 670 138
10^{-07}	$+0.000\,000\,000\,000\,000 \times 10^{+00}$	4.166 667 166 666 701	4.166 667 166 666 701
10^{-08}	$+3.700\,743\,378\,409\,752 \times 10^{+07}$	4.166 666 716 666 667	4.166 666 716 666 667
10^{-09}	$-3.700\,743\,420\,043\,115 \times 10^{+10}$	4.166 666 671 666 666	4.166 666 671 666 666
10^{-10}	$+0.000\,000\,000\,000\,000 \times 10^{+00}$	4.166 666 667 166 666	4.166 666 667 166 666
10^{-11}	$+0.000\,000\,000\,000\,000 \times 10^{+00}$	4.166 666 666 716 666	4.166 666 666 716 666
10^{-12}	$-3.700\,743\,415\,419\,038 \times 10^{+19}$	4.166 666 666 671 666	4.166 666 666 671 665
10^{-13}	$+3.700\,743\,415\,416\,816 \times 10^{+22}$	4.166 666 666 667 166	4.166 666 666 667 166
10^{-14}	$+7.401\,486\,830\,834\,415 \times 10^{+25}$	4.166 666 666 666 716	4.166 666 666 666 716
10^{-15}	$-7.401\,486\,830\,834\,381 \times 10^{+28}$	4.166 666 666 666 671	4.166 666 666 666 671

Table 5: Significant digits that match the exact evaluation of $\mathrm{r}[h; 2h; 3h]$.

For comparison, we also evaluate the formula for $\mathbf{R}(\tau, \mathbf{A})$ given in the Eq. (46) wherein the higher-order divided difference coefficients are evaluated *as is*. The first-order divided differences are evaluated using the functional form described in Eq. (106). This way of evaluating $\mathbf{R}(\tau, \mathbf{A})$ is what we refer to herein as the usual evaluation. The matrices $\mathbf{Q}(\tau, \mathbf{A})$ and $\mathbf{P}(\tau, \mathbf{A})$ are evaluated as in Example 1.

First we examine the evaluation of the coefficient $\mathrm{r}[\tau\lambda_1; \tau\lambda_2; \tau\lambda_3]$. Choosing $\tau = 1$ and $a = 0$, this coefficient simplifies to $\mathrm{r}[h; 2h; 3h]$ and its evaluation using both the usual and stable formulas is shown in Table 5. The evaluations are done using double precision floating point arithmetic and the first 16 significant decimal digits of the usual and the stable evaluations are compared to those of an exact evaluation. The significant digits in both formula evaluations that differ from the exact values are highlighted in green colour. As $\mathrm{r}[h; 2h; 3h]$ can be expressed as a fourth-order divided difference, we expect a fourth-order loss of significant digits. However, as the first-order divided differences are evaluated to machine precision, we just observe a third-order loss of significant digits in the usual evaluation. The established lower bound of at least 10 significant digits is reproduced in the stable evaluations.

Table 6 illustrates the relative errors in the evaluations of the matrices $\mathbf{P}(\tau, \mathbf{A})$, $\mathbf{Q}(\tau, \mathbf{A})$ and $\mathbf{R}(\tau, \mathbf{A})$ choosing $\tau = 1$, $a = 0$ and $b = c = d = 1$. The behaviour of the usual and stable evaluations are similar to what is observed in the previous example. The maximum relative error in the usual evaluations of $\mathbf{P}^{\mathrm{u}}$, $\mathbf{Q}^{\mathrm{u}}$ and $\mathbf{R}^{\mathrm{u}}$ are of the order of 10^{-2}, 10^{13} and 10^{28}, respectively. As before $(||\mathbf{P}^{\mathrm{u}} - \mathbf{P}||/||\mathbf{P}||)$ is $O(h)$ times smaller than $(||\mathbf{Q}^{\mathrm{u}} - \mathbf{Q}||/||\mathbf{Q}||)$ which in turn is $O(h)$ times smaller than $(||\mathbf{R}^{\mathrm{u}} - \mathbf{R}||/||\mathbf{R}||)$. The following results explain

h	$\frac{\|\|\mathbf{R}^{\mathrm{u}}-\mathbf{R}\|\|}{\|\|\mathbf{R}\|\|}$	$\frac{\|\|\mathbf{R}^{\mathrm{s}}-\mathbf{R}\|\|}{\|\|\mathbf{R}\|\|}$	$\frac{\|\|\mathbf{Q}^{\mathrm{u}}-\mathbf{Q}\|\|}{\|\|\mathbf{Q}\|\|}$	$\frac{\|\|\mathbf{Q}^{\mathrm{s}}-\mathbf{Q}\|\|}{\|\|\mathbf{Q}\|\|}$	$\frac{\|\|\mathbf{P}^{\mathrm{u}}-\mathbf{P}\|\|}{\|\|\mathbf{P}\|\|}$	$\frac{\|\|\mathbf{P}^{\mathrm{s}}-\mathbf{P}\|\|}{\|\|\mathbf{P}\|\|}$
10^{-01}	3.9×10^{-14}	7.1×10^{-14}	1.7×10^{-15}	3.2×10^{-15}	5.4×10^{-16}	4.0×10^{-16}
10^{-02}	1.0×10^{-11}	7.1×10^{-14}	1.2×10^{-13}	2.1×10^{-14}	1.0×10^{-15}	2.1×10^{-15}
10^{-03}	1.1×10^{-07}	1.3×10^{-16}	6.8×10^{-11}	7.8×10^{-17}	6.6×10^{-14}	1.1×10^{-16}
10^{-04}	3.8×10^{-05}	1.3×10^{-16}	1.4×10^{-09}	7.9×10^{-17}	9.3×10^{-14}	8.2×10^{-17}
10^{-05}	3.5×10^{-02}	5.2×10^{-17}	4.2×10^{-07}	$0.0 \times 10^{+00}$	8.1×10^{-12}	$0.0 \times 10^{+00}$
10^{-06}	$8.0 \times 10^{+01}$	$0.0 \times 10^{+00}$	4.4×10^{-05}	5.5×10^{-17}	5.8×10^{-11}	$0.0 \times 10^{+00}$
10^{-07}	4.5×10^{-02}	1.7×10^{-16}	6.7×10^{-05}	7.9×10^{-17}	7.8×10^{-11}	8.2×10^{-17}
10^{-08}	$4.0 \times 10^{+07}$	5.2×10^{-17}	4.7×10^{-01}	$0.0 \times 10^{+00}$	7.5×10^{-09}	8.2×10^{-17}
10^{-09}	$4.0 \times 10^{+10}$	$0.0 \times 10^{+00}$	$5.5 \times 10^{+01}$	5.5×10^{-17}	1.0×10^{-07}	8.2×10^{-17}
10^{-10}	3.4×10^{-01}	3.0×10^{-17}	8.3×10^{-02}	$0.0 \times 10^{+00}$	1.5×10^{-08}	$0.0 \times 10^{+00}$
10^{-11}	3.4×10^{-01}	3.0×10^{-17}	8.3×10^{-02}	$0.0 \times 10^{+00}$	1.5×10^{-08}	$0.0 \times 10^{+00}$
10^{-12}	$4.0 \times 10^{+19}$	3.0×10^{-17}	$5.5 \times 10^{+07}$	7.9×10^{-17}	1.0×10^{-04}	8.2×10^{-17}
10^{-13}	$4.0 \times 10^{+22}$	$0.0 \times 10^{+00}$	$5.5 \times 10^{+09}$	5.5×10^{-17}	1.0×10^{-03}	8.2×10^{-17}
10^{-14}	$8.0 \times 10^{+25}$	3.0×10^{-17}	$5.5 \times 10^{+11}$	$0.0 \times 10^{+00}$	8.0×10^{-03}	8.2×10^{-17}
10^{-15}	$8.0 \times 10^{+28}$	4.2×10^{-17}	$5.5 \times 10^{+13}$	$0.0 \times 10^{+00}$	6.1×10^{-02}	1.1×10^{-16}

Table 6: Relative errors in the usual and stable evaluation when $\lambda = \{h, 2h, 3h\}$.

this behaviour.

$$\mathbf{Z} := (\mathbf{A} - \lambda_1 \mathbf{I})(\mathbf{A} - \lambda_2 \mathbf{I}) = \begin{bmatrix} 0 & 0 & bd + ch \\ 0 & 0 & 2dh \\ 0 & 0 & 2h^2 \end{bmatrix} \Rightarrow \begin{matrix} \mathbf{ZA} = (a + 3h)\mathbf{Z} \\ \mathbf{ZA}^2 = (a + 3h)^2 \mathbf{Z} \end{matrix} \tag{196}$$

$$\mathbf{R}(\tau, \mathbf{A}) \approx \mathrm{r}[\tau\lambda_1; \tau\lambda_2; \tau\lambda_3]\mathbf{Z}, \quad \mathbf{Q}(\tau, \mathbf{A}) \approx \mathbf{RA}, \quad \mathbf{P}(\tau, \mathbf{A}) \approx \mathbf{RA}^2 \tag{197}$$

Equation (197) holds when the evaluation error of $\mathrm{r}[\tau\lambda_1; \tau\lambda_2; \tau\lambda_3]$ is large; Observe in Table 5 that $\mathrm{r}[h; 2h; 3h] \approx 10^{28}$ when $h \to 0$. Substituting $\tau = 1$, $a = 0$ and $b = c = d = 1$ in the above equations we can arrive at the maximum relative errors found in the usual evaluations of the matrices.

8 Conclusions

Formula evaluations in the neighbourhood of removable singularities suffer loss of significance when they are done using finite precision arithmetic. Formulas for particle tracing of streamlines and the solution of the X-IVAS scheme involve many removable singularities. Hence, the use numerically stable formulas for the same is a criteria for robustness.

We have proposed numerically stable formulas for the analytical solution in the closed form of the X-IVAS scheme. Therein, functions of matrices are defined using its Newton interpolating polynomial. In this form, removable singularities and the terms/expressions that participate to yield a finite limit at these points are grouped together as divided differences. In other algebraically equivalent formulas, say obtained by a simplification (expanding the divided differences and rearraging it to other elegant forms) of the proposed formulas, these terms/expressions get dispersed. The poor reputation of divided differences with respect to the loss of significance in a neighbourhood of removable singularities is a blessing in disguise; we get an *a priori* warning about a possible loss of significance. To control the loss of significance, we have presented piecewise definitions for these divided differences. To be precise, the piecewise definitions switch the evaluations to the respective series approximations of the divided differences should the gap between the independent variables be less than a specified threshold. These divided differences are expressible as the divided difference of the exponential function of an appropriate order less than or equal to four. For the terms involving the second, third and fourth order divided differences, the evaluation of the piecewise definitions of the same guarantee at least 14, 12 and 10 significant decimal digits to be exact, respectively. Thus, this piecewise evaluation technique is both simple and stable.

9 Acknowledgements

This study was partially supported by the SAFECON project of the European Research Council (European Commission) and the WAM-V project funded under the Navy Grant N62909-12-1-7101 issued by Office of Naval Research Global. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein.

References

[1] S. R. Idelsohn, E. Oñate, F. Del Pin, The particle finite element method: a powerful tool to solve incompressible flows with free-surfaces and breaking waves, International Journal for Numerical Methods in Engineering 61 (7) (2004) 964–989. doi:10.1002/nme.1096.
URL `http://doi.wiley.com/10.1002/nme.1096`

[2] E. Oñate, S. R. Idelsohn, F. Del Pin, R. Aubry, The particle finite element method. An overview., International Journal of Computational Methods 1 (2) (2004) 267–307. doi:10.1142/S0219876204000204.
URL `http://www.worldscinet.com/ijcm/01/0102/S0219876204000204.html`

[3] S. R. Idelsohn, J. Marti, A. Limache, E. Oñate, Unified Lagrangian formulation for elastic solids and incompressible fluids: Application to fluid–structure interaction problems via the PFEM, Computer Methods in Applied Mechanics and Engineering 197 (19-20) (2008) 1762–1776. doi:10.1016/j.cma.2007.06.004.
URL `http://linkinghub.elsevier.com/retrieve/pii/S004578250700237X`

[4] E. Oñate, S. R. Idelsohn, M. A. Celigueta, R. Rossi, Advances in the particle finite element method for the analysis of fluidmultibody interaction and bed erosion in free surface flows, Computer Methods in Applied Mechanics and Engineering 197 (19-20) (2008) 1777–1800. doi:10.1016/j.cma.2007.06.005.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0045782507002368`

[5] S. R. Idelsohn, M. de Mier-Torrecilla, E. Oñate, Multi-fluid flows with the Particle Finite Element Method, Computer Methods in Applied Mechanics and Engineering 198 (33-36) (2009) 2750–2767. doi:10.1016/j.cma.2009.04.002.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0045782509001534`

[6] M. de Mier-Torrecilla, Numerical Simulation of Multi-Fluid Flows with the Particle Finite Element Method, Phd thesis, Technical University of Catalonia (UPC) (2010).

[7] S. Idelsohn, N. Nigro, A. Limache, E. Oñate, Large time-step explicit integration method for solving problems with dominant convection, Computer Methods in Applied Mechanics and Engineering 217-220 (2012) 168–185. doi:10.1016/j.cma.2011.12.008.
URL http://linkinghub.elsevier.com/retrieve/pii/S0045782511003872

[8] S. R. Idelsohn, J. Marti, P. Becker, E. Oñate, Analysis of multifluid flows with large time steps using the particle finite element method, International Journal for Numerical Methods in Fluids 75 (9) (2014) 621–644. doi:10.1002/fld.3908.
URL http://doi.wiley.com/10.1002/fld.3908

[9] D. P. Diachin, J. A. Herzog, Analytic streamline calculations on linear tetrahedra, in: 13th Computational Fluid Dynamics Conference, American Institute of Aeronautics and Astronautics, Reston, Virigina, 1997, pp. 733–742. doi:10.2514/6.1997-1975.
URL http://arc.aiaa.org/doi/abs/10.2514/6.1997-1975

[10] B. Parlett, A recurrence among the elements of functions of triangular matrices, Linear Algebra and its Applications 14 (2) (1976) 117–121. doi:10.1016/0024-3795(76)90018-5.
URL http://linkinghub.elsevier.com/retrieve/pii/0024379576900185

[11] G. M. Nielson, I.-H. Jung, Tools for computing tangent curves for linearly varying vector fields over tetrahedral domains, IEEE Transactions on Visualization and Computer Graphics 5 (4) (1999) 360–372. doi:10.1109/2945.817352.
URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=817352

[12] W. Kahan, J. D. Darcy, How Javas Floating-Point Hurts Everyone Everywhere (1998).
URL http://www.cs.berkeley.edu/ wkahan/JAVAhurt.pdf

[13] N. J. Higham, Accuracy and Stability of Numerical Algorithms, Society for Industrial and Applied Mathematics, 2002. doi:10.1137/1.9780898718027.
URL http://epubs.siam.org/doi/book/10.1137/1.9780898718027

[14] J. F. Price, Lagrangian and Eulerian Representations of Fluid Flow: Kinematics and the Equations of Motion (2006).
URL http://www.whoi.edu/science/PO/people/jprice/class/ELreps.pdf

[15] M. Hochbruck, C. Lubich, H. Selhofer, Exponential Integrators for Large Systems of Differential Equations, SIAM Journal on Scientific Computing 19 (5) (1998) 1552–1574. doi:10.1137/S1064827595295337.
URL http://epubs.siam.org/doi/abs/10.1137/S1064827595295337

[16] A. Ostermann, M. Thalhammer, W. Wright, A Class of Explicit Exponential General Linear Methods, BIT Numerical Mathematics 46 (2) (2006) 409–431. doi:10.1007/s10543-006-0054-3.
URL http://link.springer.com/10.1007/s10543-006-0054-3

[17] M. Caliari, A. Ostermann, S. Rainer, Meshfree Exponential Integrators, SIAM Journal on Scientific Computing 35 (1) (2013) A431–A452. doi:10.1137/100818236.
URL http://epubs.siam.org/doi/abs/10.1137/100818236

[18] F. Gilbert, G. E. Backus, Propagator matrices in elastic wave and vibration problems, Geophysics 31 (2) (1966) 326–332. doi:10.1190/1.1439771.
URL http://library.seg.org/doi/abs/10.1190/1.1439771

[19] P. Kipfer, F. Reck, G. Greiner, Local Exact Particle Tracing on Unstructured Grids, Computer Graphics Forum 22 (2) (2003) 133–142. doi:10.1111/1467-8659.00655.
URL `http://doi.wiley.com/10.1111/1467-8659.00655`

[20] N. J. Higham, Functions of Matrices: Theory and Computation, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[21] E. W. Weisstein, Cubic Formula. From MathWorld–A Wolfram Web Resource.
URL `http://mathworld.wolfram.com/CubicFormula.html`

[22] A. McCurdy, K. C. Ng, B. N. Parlett, Accurate computation of divided differences of the exponential function, Mathematics of Computation 43 (168) (1984) 501–501. doi:10.1090/S0025-5718-1984-0758198-0.
URL `http://www.ams.org/jourcgi/jour-getitem?pii=S0025-5718-1984-0758198-0`

[23] M. Caliari, Accurate evaluation of divided differences for polynomial interpolation of exponential propagators, Computing 80 (2) (2007) 189–201. doi:10.1007/s00607-007-0227-1.
URL `http://link.springer.com/10.1007/s00607-007-0227-1`