



# Formal Solutions for Polarized Radiative Transfer.

## III. Stiffness and Instability

Gioele Janett<sup>1,2</sup> and Alberto Paganini<sup>3</sup>

<sup>1</sup> Istituto Ricerche Solari Locarno (IRSOL), 6605 Locarno-Monti, Switzerland; [gioele.janett@irsol.ch](mailto:gioele.janett@irsol.ch)

<sup>2</sup> Seminar for Applied Mathematics (SAM) ETHZ, 8093 Zurich, Switzerland

<sup>3</sup> University of Oxford, Mathematical Institute, OX2 6GG Oxford, UK

Received 2017 December 18; revised 2018 February 16; accepted 2018 March 2; published 2018 April 18

### Abstract

Efficient numerical approximation of the polarized radiative transfer equation is challenging because this system of ordinary differential equations exhibits stiff behavior, which potentially results in numerical instability. This negatively impacts the accuracy of formal solvers, and small step-sizes are often necessary to retrieve physical solutions. This work presents stability analyses of formal solvers for the radiative transfer equation of polarized light, identifies instability issues, and suggests practical remedies. In particular, the assumptions and the limitations of the stability analysis of Runge–Kutta methods play a crucial role. On this basis, a suitable and pragmatic formal solver is outlined and tested. An insightful comparison to the scalar radiative transfer equation is also presented.

**Key words:** methods: numerical – polarization – radiative transfer

### 1. Introduction

The transfer of partially polarized light is described by the following linear system of first-order coupled inhomogeneous ODEs

$$\frac{d}{ds}\mathbf{I}(s) = -\mathbf{K}(s)\mathbf{I}(s) + \boldsymbol{\epsilon}(s), \quad (1)$$

where  $s$  is the spatial coordinate measured along the ray under consideration,  $\mathbf{I}$  is the Stokes vector,  $\mathbf{K}$  is the propagation matrix, and  $\boldsymbol{\epsilon}$  is the emission vector. For notational simplicity, the frequency dependence of these quantities is not explicitly indicated.

It is common practice to solve Equation (1) by means of numerical methods, because its analytical solution is known for a few simple atmospheric models (which determine  $\mathbf{K}$  and  $\boldsymbol{\epsilon}$ ) only. However, Equation (1) exhibits stiff behavior, i.e., formal solvers may face instability issues. For instance, Murphy (1990) observed instability problems using the DELO-parabolic method. Thereafter, Bellot Rubio et al. (1998) encountered instability when using the cubic Hermitian method for the spectral synthesis of strong lines. De la Cruz Rodríguez & Piskunov (2013) underlined the importance of preserving stability when DELO methods are extended to high-order schemes in terms of quadratic and cubic Bézier interpolations. Štěpán & Trujillo Bueno (2013) use Bézier interpolants to control abrupt changes in the atmospheric quantities, which potentially lead to instabilities. Steiner et al. (2016) proposed a different approach to deal with strong gradients, using piecewise continuous reconstructions and slope limiters. Finally, Janett et al. (2017a, 2017b) provide a characterization of formal solvers in terms of their stability region paying particular attention to the eigenvalues of the propagation matrix.

The concept and the relevance of stability are ubiquitous in numerical analysis, and numerical methods for ODEs are not an exception (e.g., Dahlquist 1963; Deuflhard & Bornemann 2002). In particular, stability is a necessary condition for convergence. Indeed, to ensure that a numerical solution of an ODE converges, it is first necessary to show that the numerical scheme employed is consistent, that is, that the local error

introduced in one step decays superlinearly with respect to the step-size  $\Delta t$ . Unfortunately, this consistency condition is not sufficient to ensure convergence because the cumulative sum of local errors may grow exponentially. However, this exponential growth cannot happen if the numerical method is stable. In light of this, Hackbusch (2014) concludes that “whether consistency implies convergence depends on stability.” Stability analysis is employed to provide additional requirements to numerical methods (e.g., a limited step-size). However, these particular stability requirements are problem-dependent and often difficult to be determined.

This paper aims to give a deeper analysis on stability conditions, when facing the numerical integration of Equation (1). Section 2 focuses on the propagation matrix and on its eigenvalues. Section 3 presents the stability analysis of Runge–Kutta methods. Particular attention is paid to the assumptions and the limitations of this analysis, emphasizing their relevance in the formal solution for polarized light. Section 4 analyzes the effect of the conversion to optical depth on numerical stability, while Section 5 exposes the numerical approximation of this conversion. Section 6 describes the structure of a pragmatic numerical method for the numerical integration of Equation (1). Section 7 presents complementary considerations on this topic. Finally, Section 8 provides remarks and conclusions.

### 2. The Propagation Matrix

The propagation matrix  $\mathbf{K}$  that appears in Equation (1) can be written in the form (Landi Degl’Innocenti & Landolfi 2004)

$$\mathbf{K} = \begin{pmatrix} \eta_I & \eta_Q & \eta_U & \eta_V \\ \eta_Q & \eta_I & \rho_V & -\rho_U \\ \eta_U & -\rho_V & \eta_I & \rho_Q \\ \eta_V & \rho_U & -\rho_Q & \eta_I \end{pmatrix}, \quad (2)$$

where the seven independent coefficients are, in general, functions of the frequency, propagation direction, and of a series of physical parameters describing the atmosphere. The matrix  $\mathbf{K}$  can be decomposed into three different contributions,

namely,

$$\begin{pmatrix} \eta_I & 0 & 0 & 0 \\ 0 & \eta_I & 0 & 0 \\ 0 & 0 & \eta_I & 0 \\ 0 & 0 & 0 & \eta_I \end{pmatrix} + \begin{pmatrix} 0 & \eta_Q & \eta_U & \eta_V \\ \eta_Q & 0 & 0 & 0 \\ \eta_U & 0 & 0 & 0 \\ \eta_V & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \rho_V & -\rho_U \\ 0 & -\rho_V & 0 & \rho_Q \\ 0 & \rho_U & -\rho_Q & 0 \end{pmatrix}.$$

The first matrix is called the absorption matrix, it is diagonal, and it is responsible for the usual exponential decay of the whole Stokes vector. The second matrix is called the dichroism matrix, it is symmetric, and it is responsible for dichroism effects, i.e., the property of absorbing light to different extents depending on the polarization states. The third matrix is called the dispersion matrix, it is skew-symmetric, and it describes the coupling of the Stokes components due to anomalous dispersion effects.

The propagation matrix coefficients consist, in general, of two different kinds of contributions: continuum processes (due to bound-free and free-free transitions) and spectral lines (due to bound-bound transitions). In solar context, continuum processes do not introduce dichroism or anomalous dispersion effects.

This section describes the propagation matrix coefficients for an isolated spectral line originating from the atomic transition between two levels with total angular momentum  $J_u$  (upper level) and  $J_\ell$  (lower level), respectively. Each  $J$ -level is composed of  $2J + 1$  magnetic sublevels, which are degenerate in the absence of magnetic fields and are characterized by the magnetic quantum number  $M$  ( $M = -J, -J + 1, \dots, J$ ). The magnetic field removes the degeneracy among the various sublevels (Zeeman effect), inducing energy splitting, that is,

$$\Delta E = \nu_L g M,$$

where  $\nu_L$  is the Larmor frequency and  $g$  is the Landé factor. The spectral line takes into account the contribution of all the allowed transitions connecting an upper sublevel ( $J_u M_u$ ) and a lower sublevel ( $J_\ell M_\ell$ ). Atomic polarization is neglected.

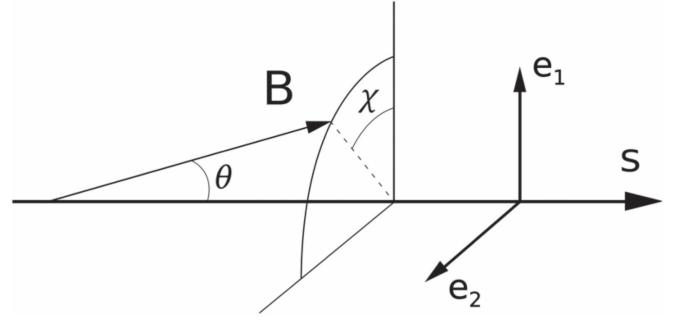
Coming back to the matrix  $\mathbf{K}$ , the total absorption coefficient  $\eta_I$  can be written as

$$\eta_I = k_c + k_L \phi_I,$$

where  $k_c$  is the local continuum absorption coefficient,  $k_L$  is the (frequency integrated) line absorption coefficient, and  $\phi_I$  is the intensity absorption profile. Note that  $\eta_I$  can always be assumed to be positive.<sup>4</sup> The dichroism coefficients and the anomalous dispersion coefficients read

$$\eta_i = k_L \phi_i, \quad \rho_i = k_L \psi_i,$$

respectively, where  $i = Q, U, V$ . When the orientation of the magnetic field  $\mathbf{B}$  with respect to the line of sight is described



**Figure 1.** Angles  $\theta$  and  $\chi$  specify the direction of the magnetic field  $\mathbf{B}$  with respect to the coordinate system of the line of sight  $s$ . The Stokes component  $Q$  is defined as the intensity difference of the linearly polarized light in the two orthogonal axes  $e_1$  and  $e_2$  in the plane perpendicular to the light beam.

with the inclination angle  $\theta$  and the azimuth angle  $\chi$  (as in Figure 1), one has

$$\begin{aligned} \phi_I &= \frac{1}{2} \left[ \phi_0 \sin^2 \theta + \frac{\phi_{-1} + \phi_1}{2} \right] (1 + \cos^2 \theta), \\ \phi_Q &= \frac{1}{2} \left[ \phi_0 - \frac{\phi_{-1} + \phi_1}{2} \right] \sin^2 \theta \cos 2\chi, \\ \phi_U &= \frac{1}{2} \left[ \phi_0 - \frac{\phi_{-1} + \phi_1}{2} \right] \sin^2 \theta \sin 2\chi, \\ \phi_V &= \frac{1}{2} [\phi_1 - \phi_{-1}] \cos \theta, \\ \psi_Q &= \frac{1}{2} \left[ \psi_0 - \frac{\psi_{-1} + \psi_1}{2} \right] \sin^2 \theta \cos 2\chi, \\ \psi_U &= \frac{1}{2} \left[ \psi_0 - \frac{\psi_{-1} + \psi_1}{2} \right] \sin^2 \theta \sin 2\chi, \\ \psi_V &= \frac{1}{2} [\psi_1 - \psi_{-1}] \cos \theta. \end{aligned} \quad (3)$$

In the observer's frame, the explicit expressions of the absorption profiles  $\phi_q$  and the dispersion profiles  $\psi_q$  ( $q = -1, 0, 1$ ) read, respectively,

$$\phi_q = \sum_{M_\ell, M_u} S_q^{J_\ell J_u}(M_\ell, M_u) \frac{1}{\sqrt{\pi}} H(\omega, a), \quad (4)$$

$$\psi_q = \sum_{M_\ell, M_u} S_q^{J_\ell J_u}(M_\ell, M_u) \frac{1}{\sqrt{\pi}} L(\omega, a), \quad (5)$$

where  $S_q^{J_\ell J_u}(M_\ell, M_u)$  is the relative strength of the Zeeman component  $q$  connecting the upper sublevel ( $J_u M_u$ ) and the lower sublevel ( $J_\ell M_\ell$ ). Using Wigner 3- $j$  symbols, its explicit expression is given by

$$S_q^{J_\ell J_u}(M_\ell, M_u) = 3 \begin{pmatrix} J_u & J_\ell & 1 \\ -M_u & M_\ell & -q \end{pmatrix}^2.$$

The functions  $H$  and  $L$  appearing in Formulas (4) and (5) correspond to the Voigt and Faraday-Voigt profiles defined by

$$\begin{aligned} H(\omega, a) &= \frac{a}{\pi} \int_{-\infty}^{\infty} e^{-x^2} \frac{1}{(\omega - x)^2 + a^2} dx, \\ L(\omega, a) &= \frac{1}{\pi} \int_{-\infty}^{\infty} e^{-x^2} \frac{\omega - x}{(\omega - x)^2 + a^2} dx, \end{aligned}$$

<sup>4</sup> Stimulated emission (which enters  $k_L$ ) is capable of producing an inversion of populations between two atomic levels. This could lead to a negative total absorption coefficient that yields an amplification of the radiation during the propagation. This phenomenon, which is at the basis of the devices such as lasers and masers, is completely negligible in solar applications and is not considered in this work.

respectively. Denoting with  $g_u$  and  $g_\ell$  the Landé factors associated to the upper and lower levels, respectively, the quantity  $\omega$  is defined as

$$\omega = \nu - \nu_A + \nu_B(g_u M_u - g_\ell M_\ell),$$

where the reduced frequency  $\nu$  is defined by

$$\nu = \frac{\nu_0 - \nu}{\Delta\nu_D},$$

with  $\nu$  and  $\nu_0$  being the frequency under consideration and line-center frequency, respectively. The Doppler width of the line  $\Delta\nu_D$  is given by

$$\Delta\nu_D = \frac{\nu_0 w_T}{c}$$

where  $w_T$  denotes the random velocity of the atoms due to thermal and microturbulent motions, and  $c$  is the speed of light. The quantity

$$\nu_A = \frac{w_A}{w_T},$$

is the normalized frequency shift due to a bulk motion of velocity  $w_A$  in the medium. The normalized Zeeman splitting  $\nu_B$  is given by

$$\nu_B = \frac{\nu_L}{\Delta\nu_D}$$

The damping constant  $a$  is given by

$$a = \frac{\Gamma}{\Delta\nu_D},$$

where  $\Gamma$  takes into account the natural width of the line  $\Gamma_n$  (due to the finite life-time of the upper and lower level) and the collisional width  $\Gamma_c$  (due to collisions of the atom under consideration with other atoms and ions in the plasma) and it reads

$$\Gamma = \Gamma_n + \Gamma_c.$$

### 2.1. Eigenvalues of the Propagation Matrix

Let

$$\boldsymbol{\eta} = (\eta_Q, \eta_U, \eta_V)^T \text{ and } \boldsymbol{\rho} = (\rho_Q, \rho_U, \rho_V)^T$$

denote the dichroism and the anomalous dispersion vectors, respectively. The four eigenvalues of the propagation matrix  $\mathbf{K}$  read (Landi Degl'Innocenti & Landolfi 2004)

$$\begin{aligned} \lambda^{(1)} &= \eta_I + \Lambda_+(\boldsymbol{\eta}, \boldsymbol{\rho}), \\ \lambda^{(2)} &= \eta_I - \Lambda_+(\boldsymbol{\eta}, \boldsymbol{\rho}), \\ \lambda^{(3)} &= \eta_I + i\Lambda_-(\boldsymbol{\eta}, \boldsymbol{\rho}), \\ \lambda^{(4)} &= \eta_I - i\Lambda_-(\boldsymbol{\eta}, \boldsymbol{\rho}), \end{aligned} \quad (6)$$

where

$$\begin{aligned} \Lambda_+(\boldsymbol{\eta}, \boldsymbol{\rho}) &= \sqrt{\sqrt{(\eta^2 - \rho^2)^2/4 + (\boldsymbol{\eta} \cdot \boldsymbol{\rho})^2} + (\eta^2 - \rho^2)/2}, \\ \Lambda_-(\boldsymbol{\eta}, \boldsymbol{\rho}) &= \sqrt{\sqrt{(\eta^2 - \rho^2)^2/4 + (\boldsymbol{\eta} \cdot \boldsymbol{\rho})^2} - (\eta^2 - \rho^2)/2}, \end{aligned}$$

**Table 1**  
Factors  $\Lambda_+$  and  $\Lambda_-$  for Different Values of  $\boldsymbol{\eta}$  and  $\boldsymbol{\rho}$

Special Cases	$\Lambda_+$	$\Lambda_-$
$\boldsymbol{\eta} = \boldsymbol{\rho} = 0$	0	0
$\rho = 0$	$\eta$	0
$\eta = 0$	0	$\rho$
$\boldsymbol{\eta} \parallel \boldsymbol{\rho}$	$\eta$	$\rho$
$\boldsymbol{\eta} \perp \boldsymbol{\rho}$ and $\eta = \rho$	0	0
$\boldsymbol{\eta} \perp \boldsymbol{\rho}$ and $\eta > \rho$	$\sqrt{\eta^2 - \rho^2}$	0
$\boldsymbol{\eta} \perp \boldsymbol{\rho}$ and $\rho > \eta$	0	$\sqrt{\rho^2 - \eta^2}$

and

$$\eta^2 = \eta_Q^2 + \eta_U^2 + \eta_V^2, \quad \rho^2 = \rho_Q^2 + \rho_U^2 + \rho_V^2.$$

The module of the dichroism vector satisfies

$$\eta \leq \eta_I, \quad (7)$$

but no similar relation holds for  $\rho$ . The comprehension of these expressions is facilitated by Table 1, where the factors  $\Lambda_+$  and  $\Lambda_-$  are given for certain special cases. Note that  $\Lambda_+$  and  $\Lambda_-$  do not depend on the azimuth angle  $\chi$  of the magnetic field vector and they always assume real positive values limited by

$$0 \leq \Lambda_+ \leq \eta, \quad 0 \leq \Lambda_- \leq \rho. \quad (8)$$

The combination of conditions (7) and (8) guarantees that the real part of the eigenvalues in Equation (6) is always positive. Therefore, the spectral radius  $r(\mathbf{K})$  of the propagation matrix  $\mathbf{K}$  satisfies

$$\eta_I \leq r(\mathbf{K}) = \eta_I \cdot \max \{1 + \Lambda_+/\eta_I, \sqrt{1 + \Lambda_-^2/\eta_I^2}\}. \quad (9)$$

Finally, knowing if the propagation matrix  $\mathbf{K}$  is diagonalizable is relevant information, because stability analysis is notably simpler in this case. If  $\eta = 0$  or  $\rho = 0$ , the propagation matrix is normal (see Appendix A) and, consequently, diagonalizable in  $\mathbb{R}$ . If  $\boldsymbol{\eta} \cdot \boldsymbol{\rho} \neq 0$ , then both  $\Lambda_+ > 0$  and  $\Lambda_- > 0$ . This implies that  $\mathbf{K}$  has four distinct eigenvalues and can be thus diagonalized in  $\mathbb{C}$ . On the other hand, if  $\boldsymbol{\eta} \perp \boldsymbol{\rho}$  and neither  $\eta = 0$  nor  $\rho = 0$ ,  $\mathbf{K}$  may not be diagonalizable because its eigenvalues are not distinct (see Table 1).

### 3. Stability Analysis

Performing stability analysis of numerical methods for ODEs is often quite involved. A gentle introduction to stability analysis of numerical methods for ODEs can be found in Higham & Trefethen (1993).

This section is dedicated to the study of the stability properties of Runge–Kutta methods applied to Equation (1). In this equation, the Stokes vector  $\mathbf{I}$  is the only quantity that can propagate or amplify errors introduced in previous steps. Consequently, the emission term  $\epsilon$  can be omitted in the stability analysis, because it does not explicitly depend on  $\mathbf{I}$ .

Moreover, Equation (1) is linear in the variable  $\mathbf{I}$  and the propagation matrix  $\mathbf{K}$  depends on the space variable  $s$ . In this case, it is common to analyze the dynamics of the system assuming that  $\mathbf{K}$  is constant around each position  $s_0$  of interest. Denoting by  $\mathbf{A} = -\mathbf{K}(s_0)$  the propagation matrix with “frozen” coefficients, one easily performs the stability analysis

on the simpler initial value problem (IVP)

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t), \quad \mathbf{y}(0) = \mathbf{y}_0. \quad (10)$$

The remainder of this section is structured as follows: Section 3.1 presents the stability analysis further assuming that the “frozen” matrix  $\mathbf{A}$  is diagonalizable. This particular case is notably simpler, because the linear system of ODEs is reduced to a set of scalar problems via diagonalization. Section 3.2 analyzes the case of a more general (non-diagonalizable) “frozen” matrix. Finally, Section 3.3 addresses the limits due to the “frozen” matrix assumption, by investigating how spatial variations in matrix  $\mathbf{A}$  affect the stability of numerical methods.

### 3.1. Reduction to the Scalar Case

A matrix  $\mathbf{A}$  is called diagonalizable if there is an invertible matrix  $\mathbf{U}$  such that

$$\mathbf{A} = \mathbf{U}^{-1}\mathbf{D}\mathbf{U},$$

where  $\mathbf{D}$  is a diagonal matrix whose entries are the eigenvalues of  $\mathbf{A}$ . From Equation (10), it is easy to see that  $\mathbf{x} = \mathbf{U}\mathbf{y}$  satisfies

$$\mathbf{x}'(t) = \mathbf{D}\mathbf{x}(t), \quad \mathbf{x}(0) = \mathbf{U}\mathbf{y}_0. \quad (11)$$

Runge–Kutta methods are affine covariant. This means that the very same approximation of  $\mathbf{y}$  that is obtained by applying a Runge–Kutta method to the IVP (10) can be computed applying the Runge–Kutta method to the IVP (11) first and multiplying the result with the matrix  $\mathbf{U}^{-1}$  at the end. For this reason, the IVP (10) can be replaced by the IVP (11), and since the latter is a system of decoupled differential equations, it is sufficient to consider the scalar case

$$x'(t) = \lambda x(t), \quad x(0) = x_0, \quad (12)$$

where  $\lambda$  represents any of the eigenvalues of  $\mathbf{A}$ .

The solution of the IVP (12) is given by

$$x(t) = x_0 e^{\lambda t}.$$

When  $\text{Re}(\lambda) < 0$ ,  $x(t)$  converges to zero as  $t \rightarrow \infty$ . The imaginary part of  $\lambda$  only introduces an oscillatory behavior of the solution.

Let  $\{t_k\}$  be a discrete grid, and let  $x_k \approx x(t_k)$  be a numerical solution computed with a Runge–Kutta method. Then,  $x_k$  and  $x_{k+1}$  satisfy

$$x_{k+1} = \phi(\lambda \Delta t) x_k, \quad (13)$$

where  $\Delta t = t_{k+1} - t_k$ , and  $\phi$  is the stability function of the numerical method (Frank 2012).

A numerical solution of an IVP is said to be asymptotically stable if the sequence  $\{x_k\}$  converges to zero for  $k \rightarrow \infty$ . Intuitively, this guarantees that any perturbation in the solution is attenuated with the recursive numerical integration. In light of Equation (13), asymptotic stability is equivalent to

$$|\phi(\lambda \Delta t)| < 1. \quad (14)$$

The stability of a numerical solution is therefore related to both the step-size  $\Delta t$  and the eigenvalue  $\lambda$ . More precisely, it depends on the product  $\lambda \Delta t$ . The stability region  $S$  of a Runge–Kutta method is defined as the set of complex values  $z = \lambda \Delta t$  for which Equation (14) is satisfied, that is,

$$S = \{z \in \mathbb{C} : |\phi(z)| < 1\}.$$

To give an example, the stability function of the explicit Runge–Kutta 4 method is given by

$$\phi_{\text{RK4}}(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24},$$

and it is displayed in yellow in Figure 2.

If  $\text{Re}(\lambda) < 0$ , asymptotic stability is guaranteed when  $\lambda \Delta t$  lies inside the stability region of the numerical method. For the particular case of Equation (1), Section 2.1 shows that the real part of the eigenvalues of the propagation operator  $-\mathbf{K}$  is always nonpositive. Since its eigenvalues are known explicitly, Equation (9) can be used to derive a sharp upper bound on the step-size  $\Delta s$  that ensures asymptotic stability of the numerical solution.

If the Runge–Kutta method is consistent, complex numbers  $z$  with negative real part and sufficiently small absolute value lie in the stability region  $S$ . Therefore, for consistent methods, instabilities can be prevented by choosing a sufficiently small step-size  $\Delta t$ . However, the downside of small step-sizes is that, for a fixed integration interval, the number of integration steps increases.

To overcome the need of choosing very small step-sizes, the stability region of the numerical method employed should be as large as possible. In particular, to ensure that the numerical solution remains asymptotically stable independently of the choice of  $\Delta t$ , the stability region should comprise the complex left half-plane  $\mathbb{C}^-$ . Runge–Kutta methods that satisfy this condition are called *A-stable*, and one of the simplest *A-stable* Runge–Kutta methods is the (implicit) trapezoidal method.

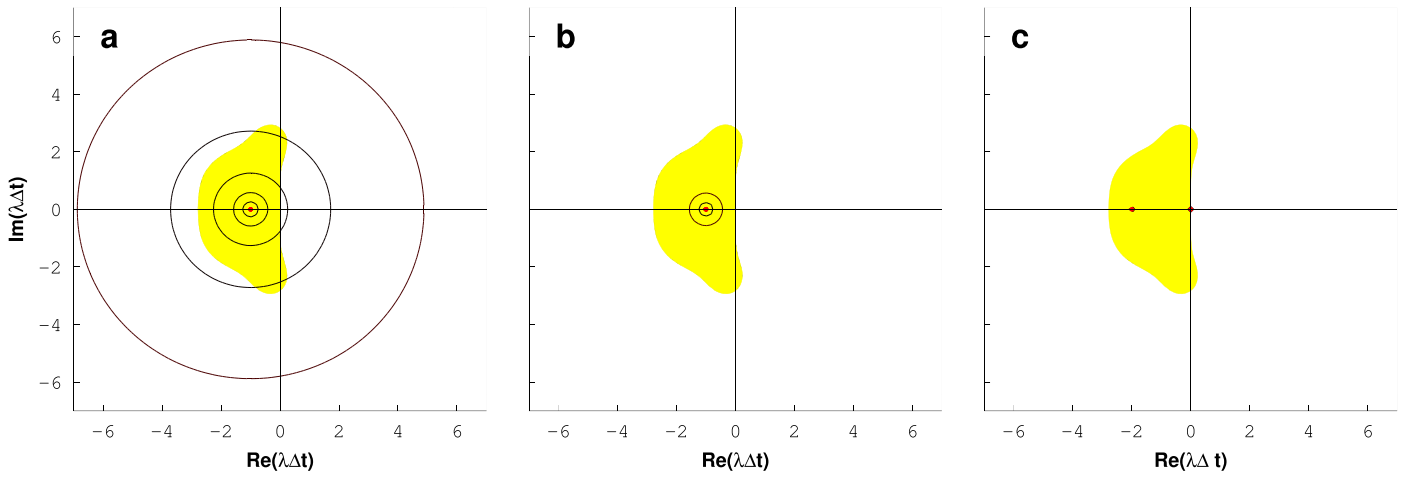
*A-stability* guarantees that the numerical solution is stable if  $\text{Re}(\lambda) < 0$ . However, if  $\text{Re}(\lambda \Delta t)$  is a large negative value, *A-stability* may not be sufficient to replicate the exponential decay of the sequence  $\{x(k \Delta t)\}$ , because *A-stability* does not guarantee that

$$\lim_{\text{Re}(z) \rightarrow -\infty} \phi(z) = 0. \quad (15)$$

For instance, the stability function of the trapezoidal method  $\phi_T$  satisfies  $\lim_{\text{Re}(z) \rightarrow -\infty} \phi_T(z) = -1$ . In this case, the numerical solution  $\{x_k\}$  will still converge to zero, but the decay becomes arbitrarily slow as  $\text{Re}(\lambda \Delta t) \rightarrow -\infty$ . *A-stable* Runge–Kutta methods that further satisfy condition (15) are called *L-stable*, and they correctly replicate exponential attenuations even when the step-size is large. The simplest *L-stable* Runge–Kutta method is the implicit (or backward) Euler scheme.

Runge–Kutta methods are also classified into explicit and implicit methods. A Runge–Kutta method that does not require solving a system of equations to update the solution is called explicit; otherwise, it is called implicit. Clearly, explicit methods are computationally less expensive. However, explicit Runge–Kutta methods cannot be *A-stable*, and therefore also not *L-stable*. Diagonally implicit Runge–Kutta methods (e.g., Kennedy & Carpenter 2016) offer a good compromise between stability, order of accuracy, and computational complexity. For instance, second-order *L-stable* diagonally implicit Runge–Kutta methods are available. When applied to linear ODEs like Equation (1), their computational cost is particularly competitive because it grows only linearly with respect to the number of Runge–Kutta stages. However, the same computational cost may grow as  $d^3$ , where  $d$  is the dimension of the ODE. Note that  $d = 4$  in Equation (1).





**Figure 2.** Pseudospectra for the matrices (a)  $\mathbf{K}_1$ , (b)  $\mathbf{K}_2$ , and (c)  $\mathbf{K}_3$  given in Equation (16). The black circles (which are almost invisible in (c)) are the boundaries of the  $\epsilon$ -pseudospectrum  $\Lambda_\epsilon$  for  $\epsilon = 10^{-1}, 10^{-2}, 10^{-3}, \dots$ , where the outermost curve corresponds to  $\epsilon = 10^{-1}$ . The red dots represent the eigenvalues of the matrix, which are four times degenerate in (a) and (b) and twice degenerate in (c). In yellow, the stability region for the explicit Runge-Kutta 4 method.

### 3.2. Analysis of Pseudospectra

The stability analysis presented in the previous section hinges on assuming that the matrix  $\mathbf{A}$  in Equation (10) is diagonalizable and only considers the scalar IVP (12) with  $\lambda$  representing any of the eigenvalues of  $\mathbf{A}$ . However, Section 2.1 shows that, in general, the propagation matrix  $\mathbf{K}$  may not be diagonalizable.

Instead of using eigenvalues, Higham & Trefethen (1993) suggest to perform stability analyses focusing on pseudospectra (more details on pseudospectra are given in Appendix B). They point out that the analysis based on eigenvalues can lead to too liberal conditions for the absence of stiffness, because eigenvalues describe the asymptotic behavior only, whereas instability and stiffness are transient phenomena that depend on how the effects compound over few integration steps. For instance, if the spectrum  $\sigma(\mathbf{A}) \subset \mathbb{C}^-$ , then the solution  $\mathbf{y}(t)$  to the IVP (10) satisfies  $\lim_{t \rightarrow \infty} \|\mathbf{y}(t)\| = 0$ . However, the decay of  $\|\mathbf{y}(t)\|$  may not be monotone. Higham & Trefethen (1993) conclude that numerical instability around  $t$  in Equation (10) occurs when the pseudospectra of the frozen coefficient matrix  $\Delta t \mathbf{A}$  fail to fit within the stability region  $S$  of the numerical method. This alternative stability analysis is particularly insightful when the pseudospectra of  $\mathbf{A}$  are highly dispersed. Unfortunately, computing pseudospectra is a computationally demanding task that is not affordable in real-time computations. Nevertheless, one can rely on two generic observations. First, if  $\mathbf{A}$  is normal, its pseudospectrum is tightly clustered around the spectrum and the difference between transient and asymptotic stability behaviors is irrelevant. Second, pseudospectra tend to be particularly dispersed if  $\mathbf{A}$  is both non-normal and “close” to a non-diagonalizable matrix.

Appendix A shows by direct calculation that the propagation matrix  $\mathbf{K}$  is normal if and only if

$$\boldsymbol{\eta} \times \boldsymbol{\rho} = 0.$$

In particular, if  $\boldsymbol{\eta} \perp \boldsymbol{\rho}$  and neither  $\boldsymbol{\eta} = 0$  nor  $\boldsymbol{\rho} = 0$ ,  $\mathbf{K}$  could be both non-normal and non-diagonalizable. Moreover, numerical tests show that an additional requirement to produce largely dispersed pseudospectra is given by  $\boldsymbol{\eta} \approx \boldsymbol{\rho} \gg \boldsymbol{\eta}_\parallel$ . This empirical condition assures that the four eigenvalues are degenerate (see Table 1).

However, the entries of the propagation matrix  $\mathbf{K}$  satisfy the dichroism condition (7). This condition guarantees that the “empirical condition” above is never satisfied and it prevents the pseudospectra from being dispersed. For this reason, the diagonalization step performed in Section 3.1 does not pose relevant problems in the stability analysis.

Some numerical evidence is presented in Figure 2, which displays the pseudospectra (for  $\Delta s = 1$ ) of the three matrices

$$\mathbf{K}_1 = \begin{pmatrix} 1 & 30 & 10 & 0 \\ 30 & 1 & 0 & 30 \\ 10 & 0 & 1 & 10 \\ 0 & -30 & -10 & 1 \end{pmatrix}, \quad \mathbf{K}_2 = \begin{pmatrix} 1 & \frac{8}{10} & \frac{3}{10} & 0 \\ \frac{8}{10} & 1 & 0 & -\frac{8}{10} \\ \frac{3}{10} & 0 & 1 & -\frac{3}{10} \\ 0 & \frac{8}{10} & \frac{3}{10} & 1 \end{pmatrix},$$

$$\mathbf{K}_3 = \begin{pmatrix} 1 & \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 1 & -20 & 0 \\ 0 & 20 & 1 & -20 \\ \frac{\sqrt{2}}{2} & 0 & 20 & 1 \end{pmatrix}. \quad (16)$$

The matrices  $\mathbf{K}_1$  and  $\mathbf{K}_2$  satisfy  $\boldsymbol{\eta} \perp \boldsymbol{\rho}$  and  $\boldsymbol{\eta} = \boldsymbol{\rho} \neq 0$ , showing dispersed pseudospectra in Figures 2(a) and (b), respectively. Due to its (unphysically) large  $\boldsymbol{\eta}$  and  $\boldsymbol{\rho}$ ,  $\mathbf{K}_1$  presents a remarkably scattered pseudospectrum, while  $\mathbf{K}_2$ , which satisfies condition (7), shows a tighter pseudospectrum. Figure 2(c) shows that the pseudospectrum of the matrix  $\mathbf{K}_3$ , which does not satisfy the condition  $\boldsymbol{\eta} \approx \boldsymbol{\rho}$ , is not dispersed. Experiments performed for different values of  $\Delta s$  lead to similar results.

### 3.3. Variation of Eigenvalues along the Integration Path

The stability analyses presented in Sections 3.1 and 3.2 neglect the dependence of the propagation matrix  $\mathbf{K}$  on the spatial variable  $s$ . The following example shows that, in principle, stability issues may arise even when numerical integrations are based on A-stable methods. Janett et al. (2017a)

give graphical illustrations of this phenomenon in terms of modifications of the stability region of the trapezoidal method.

Consider the scalar test case

$$y' = \lambda(t)y, \quad y(t_0) = y_0, \quad (17)$$

where  $\lambda(t)$  is a real function. The numerical approximation  $y_1$  of  $y(t_0 + \Delta t)$  computed with the trapezoidal method, which is A-stable, reads

$$y_1 = \left( \frac{1 + \lambda(t_0)\Delta t/2}{1 - \lambda(t_0 + \Delta t)\Delta t/2} \right) y_0. \quad (18)$$

If  $\lambda(t)$  is negative for  $t \in [t_0, t_0 + \Delta t]$ , then  $|y(t_0 + \Delta t)| < |y_0|$ . For the numerical method to be stable, it is necessary that  $|y_1| < |y_0|$  as well. However, this is not always guaranteed if  $-\lambda(t_0)\Delta t > 2$ . For instance, if  $\Delta t = -12/\lambda(t_0)$  and  $\lambda(t_0 + \Delta t) = \lambda(t_0)/2 < 0$ , then  $y_1 = (-5/4)y_0$ . This means that stability can be lost for sufficiently large variations in  $\lambda(t)$  and  $\Delta t$  big enough, and this despite the trapezoidal method being A-stable. This example shows that large variations in the coefficient  $\lambda$  affect the stability of the trapezoidal scheme. In particular, the stability region depends on the values  $\lambda(t_0)\Delta t$  and  $\lambda(t_0 + \Delta t)\Delta t$ .

More generally, the stability region of a Runge–Kutta method depends on how much the quantity  $\lambda(t_0 + x\Delta t)\Delta t$  varies for  $x \in [0, 1]$ . Let us assume that the function  $\lambda(t)$  is continuous.<sup>5</sup> Continuous dependence on  $t$  implies that variations of  $\lambda(t_0 + x\Delta t)$ , for  $x \in [0, 1]$ , can be controlled by choosing a sufficiently small  $\Delta t$ . In turn, this implies that the smaller  $\Delta t$ , the tighter the bound on the variations of the quantity  $\lambda(t_0 + x\Delta t)\Delta t$ , so that numerical stability can be recovered.

For the sake of completeness, one should mention that in the nonscalar case there are examples of non-constant matrices  $\mathbf{B}(t)$  that satisfy  $\sigma(\mathbf{B}(t)) \subset \mathbb{C}^-$  for every  $t$ , and for which the solution to the IVP

$$\mathbf{y}'(t) = \mathbf{B}(t)\mathbf{y}(t), \quad \mathbf{y}(0) = \mathbf{y}_0,$$

satisfies  $\lim_{t \rightarrow \infty} \|\mathbf{y}(t)\| = \infty$  (Josic & Rosenbaum 2008). In general, this happens when the matrix is non-normal and is related to its pseudospectra being dispersed. However, in light of the discussion presented in Section 3.2, it is unlikely that Equation (1) supports this kind of unstable solution.

#### 4. Conversion to Optical Depth

To reduce variations of the propagation matrix  $\mathbf{K}$  along the ray path, several authors (e.g., Rees et al. 1989; De la Cruz Rodríguez & Piskunov 2013) suggest to rewrite Equation (1) in terms of the optical depth  $\tau$ . This idea was first exploited to devise numerical schemes for the transfer of unpolarized light, providing significant stability enhancements (see Appendix C).

To transplant Equation (1) to the optical depth regime, one can consider the map  $g: [s_0, s_f] \rightarrow [\tau_0, \tau_f] \subset \mathbb{R}^+$ , defined as the solution of the IVP<sup>6</sup>

$$g'(s) = \eta_I(s), \quad g(s_0) = \tau_0, \quad (19)$$

where<sup>7</sup>  $\tau_f = g(s_f)$ . Since  $\eta_I > 0$ ,  $g$  is strictly monotone increasing and thus a (differentiable) bijection from  $[s_0, s_f]$  to  $[\tau_0, \tau_f]$ . Let  $\mathbf{Z}: [\tau_0, \tau_f] \rightarrow \mathbb{R}^4$  be defined by

$$\mathbf{I}(s) = \mathbf{Z}(g(s)) \quad \text{for every } s \text{ in } [s_0, s_f].$$

On the one hand, by direct differentiation,

$$\mathbf{I}(s)' = \mathbf{Z}'(g(s)) \cdot g'(s) = \eta_I(s)\mathbf{Z}'(g(s)).$$

On the other hand, from Equation (1),

$$\mathbf{I}(s)' = -\mathbf{K}(s)\mathbf{Z}(g(s)) + \epsilon(s).$$

By combining the two equations above, one obtains

$$\begin{aligned} \frac{d}{d\tau}\mathbf{Z}(\tau) &= -\frac{\mathbf{K}(g^{-1}(\tau))}{\eta_I(g^{-1}(\tau))}\mathbf{Z}(\tau) + \frac{\epsilon(g^{-1}(\tau))}{\eta_I(g^{-1}(\tau))} \\ &= -\tilde{\mathbf{K}}(g^{-1}(\tau))\mathbf{Z}(\tau) + \tilde{\epsilon}(g^{-1}(\tau)), \end{aligned} \quad (20)$$

which is formally equivalent to Equation (1).

The matrix  $\tilde{\mathbf{K}} = \mathbf{K}/\eta_I$  satisfies

$$\tilde{\mathbf{K}} = \begin{pmatrix} 1 & h_Q & h_U & h_V \\ h_Q & 1 & r_V & -r_U \\ h_U & -r_V & 1 & r_Q \\ h_V & r_U & -r_Q & 1 \end{pmatrix},$$

where, for  $i = Q, U, V$ ,

$$h_i = \frac{\eta_i}{\eta_I} = \frac{k_L \phi_i}{k_c + k_L \phi_I}, \quad r_i = \frac{\rho_i}{\eta_I} = \frac{k_L \psi_i}{k_c + k_L \phi_I},$$

and the modified emission vector is given by  $\tilde{\epsilon} = \epsilon/\eta_I$ .

The eigenvalues of the propagation matrix  $\tilde{\mathbf{K}}$  can be directly expressed in terms of the eigenvalues of  $\mathbf{K}$ , namely

$$\tilde{\lambda}^{(i)}(\tau) = \frac{\lambda^{(i)}(g^{-1}(\tau))}{\eta_I(g^{-1}(\tau))}, \quad \text{for } i = 1, 2, 3, 4. \quad (21)$$

In light of Equations (6)–(8), the spectral radius  $r(\tilde{\mathbf{K}})$  of  $\tilde{\mathbf{K}}$  satisfies

$$\begin{aligned} 1 \leq r(\tilde{\mathbf{K}}) &= \max \{1 + \Lambda_+/\eta_I, \sqrt{1 + \Lambda_-^2/\eta_I^2}\} \\ &\leq \max \{2, \sqrt{1 + \rho^2/\eta_I^2}\}, \end{aligned}$$

and the real part of the eigenvalues of the propagation operator  $-\tilde{\mathbf{K}}$  is always nonpositive.

The conversion to optical depth given by Equation (19) freezes the diagonal elements of  $\tilde{\mathbf{K}}$  to 1. The variations of its off-diagonal coefficients can be estimated in the following two limiting cases.

Case 1: the continuum absorption is much larger than line processes, i.e.,  $k_c \gg k_L \phi_I$ , and  $k_c \gg \eta_i$ ,  $\rho_i$  for  $i = Q, U, V$ . In this case, the off-diagonal coefficients of  $\tilde{\mathbf{K}}$ , and in turn the absolute value of their variations, tend to zero. Equations (21) and (6) imply that the eigenvalues of  $\tilde{\mathbf{K}}$  are close to 1.

Case 2: the line absorption dominates over the continuum absorption, i.e.,  $k_L \phi_I \gg k_c$ . In this case, the variation along the ray path of the off-diagonal coefficients of  $\tilde{\mathbf{K}}$  is basically independent of  $k_c$  and  $k_L$ . Equations (21) and (6) imply that the eigenvalues of  $\tilde{\mathbf{K}}$  are also basically independent of  $k_c$  and  $k_L$ .

<sup>5</sup> This assumption is required by the Picard–Lindelöf Theorem to ensure that the IVP (17) has a solution, and that this solution is unique.

<sup>6</sup> In the literature,  $g$  is usually defined as the solution of  $g'(s) = -\eta_I(s)$ . However, the negative sign induces an unnecessary and possibly confusing change in the integration direction.

<sup>7</sup> The subscript “f” stands for “final.”

and their variations are only due to variations in the profiles given by Equation (3).

However, if the conditions of the two cases presented above are not met, it is not straightforward to infer conclusions on the values of the off-diagonal entries of  $\mathbf{K}$ , and strong variations in the propagation matrix may still be present (in particular, due to the dependence of the coefficients given by Equation (3) on variations of the magnetic field and of the bulk motions).

In conclusion, the conversion to optical depth usually reduces the amount of fluctuations of the propagation operator  $-\mathbf{K}$  along the ray path, but this is not guaranteed in general.

### 5. Numerical Conversion to Optical Depth

To apply a numerical scheme to Equation (20) it is necessary to have a certain knowledge of the function  $g$ . From Equation (19), one has

$$g(s) = \tau_0 + \int_{s_0}^s \eta_I(x) dx, \quad (22)$$

and numerical approximations of  $g$  can be obtained by replacing the integral with a numerical quadrature. It is absolutely crucial that this numerical approximation is strictly monotone increasing because one needs to access the values of its inverse  $g^{-1}$ . Replacing  $g$  with a numerical approximation could negatively affect the order of the method employed to solve the IVP (20). Janett et al. (2017a) explain that high-order solvers require a corresponding high-order numerical approximation of the integral in Equation (22). A very common approach to devise numerical quadrature rules is to replace integrands with interpolants that are successively integrated exactly. For instance, Gauss, Radau, Hermite, and Clenshaw–Curtis quadratures are based on this idea.

Here are some more concrete examples. The trapezoidal rule applied to Equation (22) reads

$$g(s) \approx \tau_0 + (s - s_0) \frac{\eta_I(s_0) + \eta_I(s)}{2}.$$

This quadrature is based on a linear interpolation of  $\eta_I$  through the points  $\{s_0, s\}$  and is second-order accurate. Higher-order monotone quadrature schemes can be obtained by replacing linear interpolation with higher-order monotone interpolants.

A concrete high-order example is given by the monotone cubic Hermite quadrature, that, applied to Equation (22), leads to

$$g(s) \approx \tau_0 + (s - s_0) \frac{\eta_I(s_0) + \eta_I(s)}{2} + (s - s_0)^2 \frac{\tilde{\eta}'_I(s_0) - \tilde{\eta}'_I(s)}{12},$$

where  $\tilde{\eta}'_I$  are suitable numerical approximations (of the first derivative  $\eta'_I$ ) that guarantee monotonicity. The approximation above is fourth-order accurate provided that the approximation  $\tilde{\eta}'_I \approx \eta'_I$  is at least of second order (Dougherty et al. 1989). The approximation  $\tilde{\eta}'_I$  described by Steffen (1990) satisfies both conditions, whereas the one described by Fritsch & Butland (1984) guarantees monotonicity, but it is second-order accurate on uniform grids only (it drops to first-order on non-uniform grids).

Hermite interpolation is not the only option for higher-order monotone interpolation schemes. For instance, Auer (2003) and De la Cruz Rodríguez & Piskunov (2013) prefer to employ monotonic quadratic Bézier splines. However, the high-order convergence of Bézier interpolations is achieved only when the Bézier interpolants are forced to be identical to the corresponding degree Hermite interpolants, which do not guarantee monotonicity.

Finally, when the atmospheric model is exponentially stratified along the ray path, Mihalas (1978) suggests to replace  $\eta_I$  with the exponential function<sup>8</sup>

$$\eta_I(s_0) e^{(x-s_0)/\alpha}, \quad \text{for } x \in [s_0, s],$$

with

$$\alpha = \frac{(s - s_0)}{\log \eta_I(s) - \log \eta_I(s_0)}.$$

After such a substitution, Equation (22) can be integrated exactly. The resulting map reads

$$g_M(s) = \tau_0 + (s - s_0) \frac{\eta_I(s) - \eta_I(s_0)}{\log \eta_I(s) - \log \eta_I(s_0)}.$$

The error introduced by this substitution is bounded by

$$\begin{aligned} |g_M(s) - g(s)| &\leq \int_{s_0}^s |\eta_I(s) - \eta_I(s_0) e^{(x-s_0)/\alpha}| dx, \\ &\leq (s - s_0) \max_{x \in [s_0, s]} |\eta_I(s) - \eta_I(s_0) e^{(x-s_0)/\alpha}|, \end{aligned}$$

and its accuracy clearly depends on the suitability of the exponential modeling.

### 6. Pragmatic Formal Solver

In practical applications, the propagation matrix  $\mathbf{K}$  and the emission vector  $\epsilon$  are known only at a discrete set of usually nonequidistant grid points  $\{s_i\}_{i=1}^N$ . In such an instance, one aims at computing a numerical solution of Equation (1) that is first of all physically meaningful (i.e., stable) and that, second, has a good ratio between accuracy and computational cost. Ideally, the numerical method should rely as much as possible on the provided values of  $\mathbf{K}$  and  $\epsilon$ , although these functions can be evaluated at other depths  $s$  via interpolation, if necessary. This interpolation must be sufficiently accurate to preserve the order of convergence of the numerical scheme used for numerical integration (Janett et al. 2017b).

Since stiffness is a transient behavior, the analysis presented in the previous sections suggests to consider each interval  $[s_i, s_{i+1}]$  at a time and sequentially. In each interval, depending on the cell width  $\Delta s$  and on the magnitude of the eigenvalues of  $\mathbf{K}$  at  $s_i$  and  $s_{i+1}$ , the approximation of  $\mathbf{I}(s_{i+1})$  is computed with either an explicit method  $\Psi^E$  (which is computationally inexpensive) or an  $A$ -stable method  $\Psi^A$ , or an  $L$ -stable method  $\Psi^L$ . Preferably, the methods  $\Psi^E$  and  $\Psi^A$  should be of the same order, while the method  $\Psi^L$  could be of lower order, because large attenuations usually prevent any propagation of information.

The following criteria can help in choosing between  $\Psi^E$ ,  $\Psi^A$ , or  $\Psi^L$ : (i) if the absolute value of the real part of the eigenvalues multiplied by the cell width is large, one should

<sup>8</sup> If  $\eta_I$  is known at the grid points  $\{s_i\}_{i=1}^N$ , it is natural to employ the piecewise exponential model by adapting the parameter  $\alpha$  to every interval  $[s_i, s_{i+1}]$ .

use  $\Psi^L$  to guarantee the correct exponential attenuation of the Stokes vector. Otherwise, (ii) the method  $\Psi^E$  is used whenever stable (to reduce computational cost), and (iii) if  $\Psi^E$  is not stable, one uses  $\Psi^A$  with the optional conversion to optical depth if  $\Psi^A$  loses stability due to the variations of the eigenvalues in the interval  $[s_i, s_{i+1}]$ .

For example, this strategy can be implemented using Heun's method (which is also known as the explicit trapezoidal rule and has order 2) as  $\Psi^E$ , the implicit trapezoidal rule (which also has order 2) as  $\Psi^A$ , and the implicit Euler method (which has order 1) as  $\Psi^L$ . These methods employ  $\mathbf{K}$  and  $\epsilon$  at grid points only, avoiding the use of interpolated off-grid points' quantities. Computing the eigenvalues of  $\mathbf{K}$  at a point  $s$  is roughly one-third as expensive<sup>9</sup> as one step of  $\Psi^E$ , whereas  $\Psi^A$  is roughly twice as expensive as  $\Psi^E$ . The implicit Euler method is less expensive than  $\Psi^A$ , but more than  $\Psi^E$ . A second-order  $L$ -stable method would be at least as expensive as  $\Psi^A$ , but since  $L$ -stability is only required when large exponential attenuations are present, one can opt for a lower-order scheme.

To assess the stability of Heun's method  $\Psi^E$ , one should verify that

$$|\phi_{\Psi^E}| = \left| 1 + \Delta s \frac{\lambda(s_i) + \lambda(s_{i+1}) + \Delta s \lambda(s_i) \lambda(s_{i+1})}{2} \right| < 1.$$

However, it is worth distinguishing the cases when  $|\phi_{\Psi^E}|$  is close to 1: if  $\lambda \Delta s$  is close to 0,  $\Psi^E$  can be trusted; however, if  $\lambda \Delta s$  is close to the boundary of the stability domain away from 0, it is advisable to switch to  $\Psi^A$ , because  $\Psi^E$  may suffer from instability. To verify the stability of  $\Psi^A$  and decide whether to opt for the conversion to optical depth, one can repeat the same argument used for  $\Psi^E$  but using Formula (18) instead of  $\phi_{\Psi^E}$ .

A practical example is given by Figure 3, which shows the evolution of the approximate Stokes vector for the Fe I line at 6301.50 Å computed with a FALC atmospheric model (Fontenla et al. 1993) supplemented with a constant magnetic field.<sup>10</sup> The different rows refer to computational grids of increasing refinements and the approximate solution is calculated by the pragmatic numerical scheme suggested above.

The method  $\Psi^L$  is used if there is an eigenvalue whose real part is  $< -7/\Delta s$  (at  $s_i$  or  $s_{i+1}$ ). The method  $\Psi^E$  is used if  $|\phi_{\Psi^E}| < 0.6$  or if the real part of both eigenvalues (at  $s_i$  and  $s_{i+1}$ ) is  $> -10^{-3}$ . The method  $\Psi^A$  is converted to optical depth if  $|\phi_{\Psi^A}| > 0.8$ . These parameters should not be considered as an ultimate choice, but they provide a concrete example. However, repeating the experiments with similar choices of parameters delivers similar results. The reference solution is computed using the implicit Euler method on a grid that contains 9999 points.

The experiments show that the pragmatic strategy effectively switches among the methods, delivering physically meaningful approximations independently from the coarseness of the grid. As predicted by the analysis, the use of  $\Psi^L$  (purple dots) decreases with the refinement of the grid: it is replaced by  $\Psi^A$  (yellow and orange dots), which is in turn replaced by  $\Psi^E$  (blue dots). Table 2 summarizes the use (in percentage) of  $\Psi^E$ ,  $\Psi^A$  (without and with conversion to optical depth), and  $\Psi^L$  for

<sup>9</sup> This fraction decreases if Heun's method is replaced by a higher-order explicit Runge–Kutta scheme, because the latter inevitably requires the computation of more stages.

<sup>10</sup> The values of  $\mathbf{K}$  and  $\epsilon$  have been computed with the RH code of Uitenbroek (2001).

each grid. These values have been approximated to the second digit.

Although not shown, one must point out that the use of  $\Psi^L$  is necessary in order to deal with the stiffness of optically thick cells. This is partly visible in the fourth row, where  $U$  shows overshoots. A similar numerical experiment based on  $\Psi^E$  and  $\Psi^A$  only presents oscillations in the spatial region  $[-0.12 \times 10^5, 2.5 \times 10^5]$  if the grid is too coarse.

For comparison, Figure 4 shows the numerical evolution of the Stokes vector when this is computed, relying solely on  $\Psi^E$ . With 140 points, this numerical solution is completely spurious because of numerical instability. With 200 points, the result is physically correct only after a certain depth. In particular, in the depth region  $[-0.12 \times 10^5, 2.5 \times 10^5]$ , this numerical solution oscillates wildly and the relative error with respect to the reference solution is of the order of  $10^6$ .

Finally, using the bound (9) on the spectral radius instead of computing the eigenvalues to decide which method to employ delivers similar results and is computationally (slightly) cheaper.

## 7. Supplemental Remarks

This section provides two additional considerations concerning the stability of the formal solution of the polarized radiative transfer.

### 7.1. Stability of DELO Methods

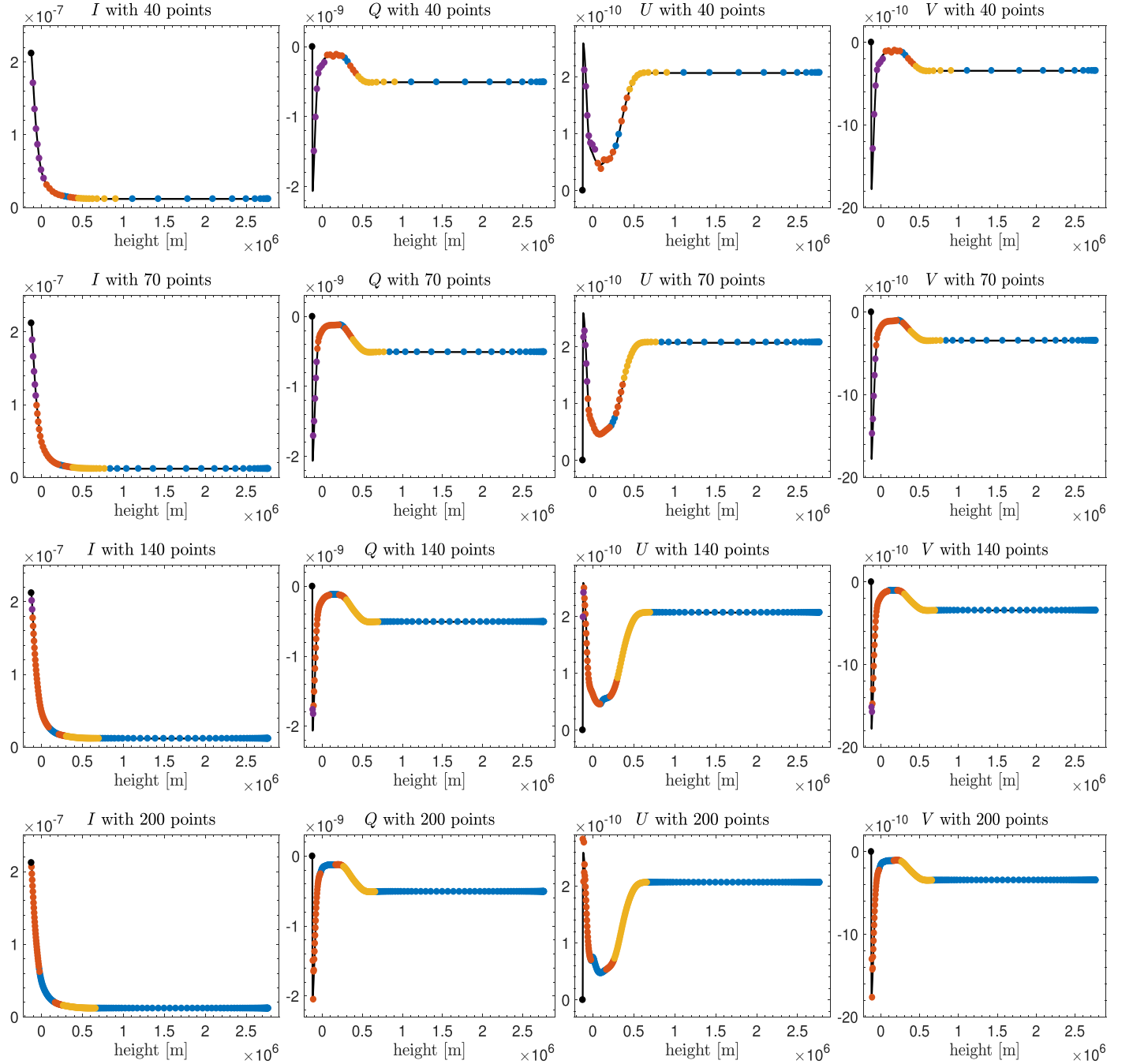
DELO methods belong to the class of exponential integrators: aiming at removing stiffness from the problem, the DELO strategy analytically integrates the diagonal elements of the propagation matrix (Guderley & Hsu 1972). Rees et al. (1989) first proposed the application of this technique to Equation (1), which has been very successful thanks to its stability properties. For this reason, the DELO strategy has since been chosen to develop higher-order methods: e.g., the DELO-parabolic (Murphy 1990; Janett et al. 2017a) and the DELO-Bézier (De la Cruz Rodríguez & Piskunov 2013) methods. DELO methods are currently widespread for the numerical evaluation of Equation (1).

The DELO strategy relies on the spatial scale conversion given by Equation (22) (which potentially introduces numerical errors) and it deals with the modified propagation matrix  $\mathcal{K} = \mathbf{K}/\eta_l - \mathbf{1}$ , where  $\mathbf{1}$  represents the  $4 \times 4$  identity matrix. The stability functions of the DELO-linear and the DELO-parabolic methods satisfy condition (15). When the norm of the matrix  $\mathcal{K}$  tends to zero (e.g., for a diagonal matrix  $\mathbf{K}$ ), DELO methods tend to  $A$ -stability (Janett et al. 2017a) and, consequently, to  $L$ -stability. This fact explains the usual good performance of the DELO-linear method when dealing with very coarse grids and suggests its suitability as the  $L$ -stable method  $\Psi^L$  in the pragmatic formal solver described in the previous section.

### 7.2. Oscillations in the Evolution Operator

Here, a preliminary remark is required. When presenting the fourth-order  $A$ -stable cubic Hermitian method, Bellot Rubio et al. (1998) points to the improper sampling of the oscillations in the evolution operator elements as a reason for instability and inaccuracy. In particular, they investigate the case of strong lines, where the cubic Hermitian method flagrantly fails to reliably reproduce the emergent  $Q$  and  $U$  Stokes components when dealing with coarse spatial grids. In light of the stability analysis of Section 3, some considerations can be done.





**Figure 3.** Each row displays the evolution of the Stokes components along the vertical direction for the Fe I line at 6301.50 Å in the proximity of the line core frequency. The Stokes profiles have been computed using a FALC model atmosphere on a sequence of increasingly refined grids. The approximated solution is calculated using the pragmatic approach described in Section 6. The black line depicts the reference solution, which has been computed with  $\Psi^L$  on a very fine grid. The initial condition is  $I = (2.121 \times 10^{-7}, 0, 0, 0)^T$ . Different dot colors correspond to integrations with a different method. Blue dots indicate the use of  $\Psi^E$ , yellow dots of  $\Psi^A$ , orange dots of  $\Psi^A$  with conversion to optical depth, and purple dots of  $\Psi^L$ . The algorithm switches between the different methods depending on the stability criteria. The use of  $\Psi^E$  increases with the refinement of the grid.

One starts identifying the origin of the oscillations in the evolution operator elements. The formal solution of Equation (1) in terms of the evolution operator reads

$$I(s) = \mathbf{O}(s, s_0)I(s_0) + \int_{s_0}^s \mathbf{O}(s, x)\epsilon(x)dx,$$

where  $\mathbf{O}$  is a  $4 \times 4$  matrix. Under the assumption of a constant propagation matrix  $\mathbf{K}$  in the layer  $[s_0, s]$ , the evolution operator

can be written as

$$\mathbf{O}(s, s_0) = e^{-(s-s_0)\mathbf{K}}.$$

If either  $\Lambda_+ \neq 0$  or  $\Lambda_- \neq 0$ , the evolution operator can be decomposed as

$$\mathbf{O}(s, s_0) = \sum_{i=1}^4 e^{-(s-s_0)\lambda^{(i)}} \mathbf{N}_i(\eta, \rho),$$

**Table 2**  
Statistic of the Pragmatic Strategy

# of Grid Points	$\Psi^E$	$^a\Psi^A$	$^b\Psi^A$	$\Psi^L$
40	36%	23%	23%	18%
70	38%	30%	25%	7%
140	43%	27%	28%	2%
200	51%	20%	29%	0%

**Notes.**

<sup>a</sup> Without conversion to optical depth.

<sup>b</sup> With conversion to optical depth.

where  $\lambda^{(i)}$  are given by Equation (6) and  $\mathbf{N}_i$  are known  $4 \times 4$  matrices (see Appendix 5 of Landi Degl’Innocenti & Landolfi 2004). From Equation (6), one recognizes that

$$\lambda^{(1)}, \lambda^{(2)} \in \mathbb{R}, \text{ and } \lambda^{(3)}, \lambda^{(4)} \in \mathbb{C}.$$

The imaginary part of the eigenvalues  $\lambda^{(3)}$  and  $\lambda^{(4)}$  induces sinusoidal oscillations in the evolution operator elements, which correspond to the radiative transfer phenomena known as Faraday rotation and Faraday pulsation. These oscillations have spatial frequency  $\Lambda_-$ , a factor dominated by the anomalous dispersion coefficients (see Table 1). In fact, a strong anomalous dispersion vector  $\rho$  induces high-frequency oscillations in the evolution operator elements. In particular, the component  $\rho_V$  causes a rotation of the direction of maximum linear polarization, whereas the components  $\rho_Q$  and  $\rho_U$  induce a transformation from linear (circular) to circular (linear) polarization (Landi Degl’Innocenti & Landolfi 2004).

The presence of high-frequency oscillations alone is not a sufficient condition for instability when using an *A*-stable method. The additional requirement is the variation of the propagation matrix  $\mathbf{K}$  along the integration path. Moreover, the stability improvement when using denser spatial grids noted by Bellot Rubio et al. (1998) is not due to the proper sampling of the oscillations in the evolution operator, but to the reduction of large variations of the propagation matrix  $\mathbf{K}$  between consecutive grid points.

An illustrative example (not shown here) is given by the numerical evaluation of the formal solution when considering a constant propagation matrix with strong anomalous dispersion coefficients. In this case, an *A*-stable formal solver (e.g., the trapezoidal method) integrates Equation (1) without any instability issue.

## 8. Conclusions

This paper exposes the stability analysis of the numerical integration of the radiative transfer equation for polarized light. The main aim is to better understand the specific situations where instability issues appear and how to deal with them, rather than prescribing the ultimate stability criterion. This knowledge can be used to devise more robust formal solvers.

The first part focuses on the propagation matrix, identifying different structural properties, such as normality, diagonalizability, spectrum, and spectral radius.

The second part studies the stability properties of Runge–Kutta methods applied to Equation (1). Particular attention is paid to the assumptions and the limitations of the stability analysis, emphasizing their relevance in the formal solution for

polarized light. Special care is paid to better understand the role of spatial variations in the propagation matrix.

It is shown that the conversion to the optical depth spatial scale, defined by Equation (19), usually mitigates variation of the propagation matrix elements along the integration path. Appendix C shows that numerical instabilities due to variations in the eigenvalues are a concrete problem when dealing with polarized light only. In the scalar case, the conversion to optical depth cancels the variation of the unique eigenvalue along the ray path. An entire section is dedicated to the numerical conversion to optical depth based on Equation (22). This approximation introduces numerical errors that could lead to a reduced order of accuracy of the formal solver. In practice, high-order formal solvers require a corresponding high-order numerical evaluation of the integral in Equation (22).

Finally, the structure of a paradigmatic pragmatic formal solver is given in terms of a switching technique. This numerical scheme chooses between different numerical methods at each step of the integration. It uses an inexpensive explicit method as long as the integration of the ODE is not limited by stability requirements and it switches to an implicit method when stiffness appears. In optically thick cells, the method switches to an *L*-stable method to correctly replicate exponential attenuations and to avoid numerical oscillations. The criterion for the switching is based either on the eigenvalues or on the spectral radius of the propagation matrix  $\mathbf{K}$ . The numerical tests are promising: the pragmatic strategy effectively switches among the methods and it delivers physically meaningful approximations independently from the coarseness of the grid.

It is important to point out that the stability results presented in this work rely on assuming that, prior to discretization, the propagation matrix  $\mathbf{K}$  and the emission vector  $\epsilon$  are continuous functions. The effective performance of the pragmatic method on discontinuous atmospheric models remains to be explored. However, a switching technique based on choosing numerical methods depending on the local smoothness of the input data might be suitable to face discontinuities and high gradients.

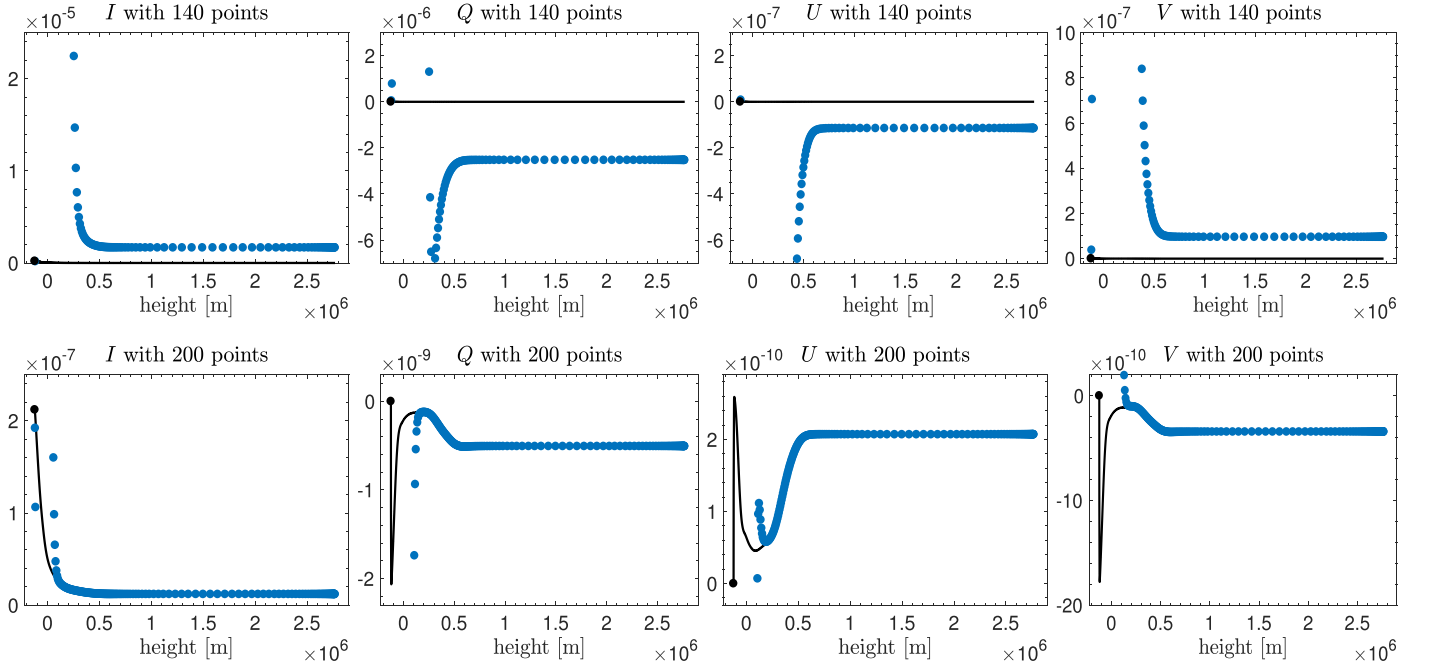
The financial support by the Swiss National Science Foundation (SNSF) through grant ID 200021\_159206/1 is gratefully acknowledged. The work of Alberto Paganini was partly supported by the EPSRC Grant EP/M011151/1. Special thanks are extended to L. Belluzzi and O. Steiner for reading and commenting on the paper. The authors also are grateful to the anonymous referee for providing valuable comments that helped to improve the article.

## Appendix A (Non-)Normality of the Propagation Matrix

A real square matrix  $\mathbf{A}$  is normal if

$$\mathbf{A}^T \mathbf{A} = \mathbf{A} \mathbf{A}^T, \quad (23)$$

and this is true, e.g., for symmetric or skew-symmetric matrices. Normal matrices are diagonalizable and their spectrum is stable with respect to small perturbations of the matrix components (Trefethen & Embree 2005).



**Figure 4.** Repetition of the experiment from Figure 3, but using solely Heun’s method to approximate the evolution of the Stokes components. Calculations are clearly affected by numerical instabilities.

The propagation matrix given by Equation (2) satisfies

$$\frac{\mathbf{K}^T \mathbf{K} - \mathbf{K} \mathbf{K}^T}{2} = \begin{pmatrix} 0 & \eta_U \rho_V - \eta_V \rho_U & \eta_Q \rho_V - \eta_V \rho_Q & \eta_Q \rho_U - \eta_U \rho_Q \\ \eta_U \rho_V - \eta_V \rho_U & 0 & 0 & 0 \\ \eta_Q \rho_V - \eta_V \rho_Q & 0 & 0 & 0 \\ \eta_Q \rho_U - \eta_U \rho_Q & 0 & 0 & 0 \end{pmatrix},$$

and hence it does not satisfy the normality condition in general. In particular,  $\mathbf{K}$  is normal if and only if

$$\eta \times \rho = 0.$$

In the cases of vanishing dichroism effects, i.e.,  $\eta = 0$ , or of vanishing anomalous dispersion effects, i.e.,  $\rho = 0$ , one recognizes that the matrix satisfies the normality condition.

## Appendix B Pseudospectrum

While the spectrum of a matrix is just the set of its eigenvalues, the pseudospectrum also depends on an additional parameter, a small number  $\epsilon > 0$ . Therefore, one usually refers to the  $\epsilon$ -pseudospectrum. The  $\epsilon$ -pseudospectrum  $\Lambda_\epsilon$  of a square matrix  $\mathbf{A}$  is defined by (Trefethen & Embree 2005)

$$\Lambda_\epsilon(\mathbf{A}) = \{z \in \mathbb{C} : \|(z\mathbf{1} - \mathbf{A})^{-1}\| \geq \epsilon^{-1}\},$$

i.e., it is the set of  $z \in \mathbb{C}$  that are eigenvalues of some matrix  $\mathbf{A} + \mathbf{E}$  with  $\|\mathbf{E}\| \leq \epsilon$ . Here, the matrix  $\mathbf{E}$  acts as a small perturbation to  $\mathbf{A}$ . Therefore, the pseudospectrum  $\Lambda_\epsilon(\mathbf{A})$  always contains the  $\epsilon$ -neighborhood of the spectrum  $\Lambda(\mathbf{A})$ . If one perturbs a normal matrix by a perturbation of operator norm at most  $\epsilon$ , then the spectrum moves by at most  $\epsilon$ . However, for a non-normal matrix, the pseudospectrum may be

widely dispersed and a small perturbation may induce the spectrum to change a lot.

In order to better understand the concept of pseudospectra, one usually produces approximate pictures of the boundaries of  $\Lambda_\epsilon(\mathbf{A})$  for various values of  $\epsilon$ , modifying  $\mathbf{A}$  by small random perturbations and looking at the spectra of these perturbations. The standard algorithm is to evaluate the smallest singular value<sup>11</sup> of the matrix  $\mathbf{B} = z\mathbf{1} - \mathbf{A} + \mathbf{E}$  for different values of  $z$  on a grid in the complex plane and then generate a contour plot from this data (Trefethen 1999). Different explicit examples are given in Figure 2.

These pictures are particularly useful for the stability analysis. Higham & Trefethen (1993) conclude that numerical instability around  $t$  in Equation (10) occurs when the pseudospectra of the frozen matrix  $\Delta t \mathbf{A}$  fail to fit the stability region of the numerical method. Therefore, one usually analyzes the stability of a numerical method by overplotting its stability region and the boundaries of  $\Lambda_\epsilon(\Delta t \mathbf{A})$  for different values of  $\epsilon$ , as presented in Figure 2. Note that the stability condition based on pseudospectra is more conservative, or less liberal, with respect to the one based on eigenvalues.

## Appendix C Stability for Scalar Formal Solutions

The transfer of unpolarized light is described by the first-order inhomogeneous scalar ODE (Mihalas 1978)

$$\frac{d}{ds} I(s) = -\eta_I(s) I(s) + \epsilon(s), \quad (24)$$

<sup>11</sup> The singular values of a matrix  $\mathbf{A}$  are the absolute values of the eigenvalues of the matrix  $\mathbf{A}^T \mathbf{A}$ .

where  $I$  is the specific intensity,  $\eta_l$  is the absorption coefficient, and  $\epsilon$  is the emissivity. To simplify the notation, the frequency dependence of these quantities is omitted.

In terms of the optical depth  $\tau$ , Equation (24) reads

$$\frac{d}{d\tau}Z(\tau) = -Z(\tau) + \frac{\epsilon(g^{-1}(\tau))}{\eta_l(g^{-1}(\tau))} = -Z(\tau) + S(g^{-1}(\tau)), \quad (25)$$

where  $S = \epsilon/\eta_l$  is the so-called source function,  $g$  is defined in Equation (19), and  $Z(\tau) = I(g^{-1}(\tau))$ .

The fundamental difference between Equations (24) and (25) is that in the latter the linear coefficient is constant (and equal to  $-1$ ), whereas the absorption coefficient  $\eta_l$  in Equation (24) depends on  $s$ . This implies that to devise a stable numerical scheme for (25) it is sufficient to follow the discussion presented in Section 3.1, whereas for (24) one needs to take into account the variations of  $\eta_l$  along the ray path, see Section 3.3. In particular, it is sufficient to employ an A-stable Runge–Kutta method to compute stable numerical solutions to Equation (25), whereas this may not be sufficient for Equation (24).

For the sake of completeness, the analytic solution to Equation (25) is given by

$$Z(\tau) = Z_0 e^{-(\tau-\tau_0)} + \int_{\tau_0}^{\tau} e^{-(x-\tau_0)} S(g^{-1}(x)) dx.$$

Therefore,

$$I(s) = I_0 e^{-(g(s)-g(s_0))} + \int_{s_0}^s e^{-(g(y)-g(s_0))} \epsilon(y) dy,$$

which can be approximated in a stable manner by replacing  $g$  and the integral on the right-hand side with numerical approximations. In this case, monotonicity of  $g$  guarantees  $L$ -stability.

## ORCID iDs

Gioele Janett  <https://orcid.org/0000-0003-3247-6612>

Alberto Paganini  <https://orcid.org/0000-0003-3309-7657>

## References

- Auer, L. 2003, in ASP Conf. Ser. 288, *Stellar Atmosphere Modeling*, ed. I. Hubeny, D. Mihalas, & K. Werner (San Francisco, CA: ASP), 3
- Bellot Rubio, L. R., Ruiz Cobo, B., & Collados, M. 1998, *ApJ*, **506**, 805
- Dahlquist, G. G. 1963, *BIT Numer. Math.*, **3**, 27
- De la Cruz Rodríguez, J., & Piskunov, N. 2013, *ApJ*, **764**, 33
- Deuflhard, P., & Bornemann, F. 2002, *Scientific Computing with Ordinary Differential Equations* (Berlin: Springer)
- Dougherty, R., Edelman, A., & Hyman, J. M. 1989, *MaCom*, **52**, 471
- Fontenla, J. M., Avrett, E. H., & Loeser, R. 1993, *ApJ*, **406**, 319
- Frank, J. 2012, *Computational Modelling and Dynamical Systems* (Amsterdam: Univ. Amsterdam and Univ. Edinburgh), <https://www.staff.science.uu.nl/~frank011/Articles/CMDS.pdf>
- Fritsch, F. N., & Butland, J. 1984, *SIAM J. Sci. Stat. Comput.*, **5**, 300
- Guderley, G., & Hsu, C.-C. 1972, *MaCom*, **26**, 51
- Hackbusch, W. 2014, *The Concept of Stability in Numerical Mathematics* (Berlin: Springer)
- Higham, D. J., & Trefethen, L. N. 1993, *Num. Math.*, **33**, 285
- Janett, G., Carlin, E. S., Steiner, O., & Belluzzi, L. 2017a, *ApJ*, **840**, 107
- Janett, G., Steiner, O., & Belluzzi, L. 2017b, *ApJ*, **845**, 104
- Josic, K., & Rosenbaum, R. 2008, *SIAMR*, **50**, 570
- Kennedy, C., & Carpenter, M. 2016, *Diagonally Implicit Runge–Kutta Methods for Ordinary Differential Equations, a Review*, NASA Technical Memorandum (Washington, DC: NASA)
- Landi Degl’Innocenti, E., & Landolfi, M. 2004, *Polarization in Spectral Lines* (Dordrecht: Kluwer)
- Mihalas, D. 1978, *Stellar Atmospheres* (2nd ed.; San Francisco, CA: Freeman)
- Murphy, G. A. 1990, PhD thesis, Univ. Sidney
- Rees, D. E., Durrant, C. J., & Murphy, G. A. 1989, *ApJ*, **339**, 1093
- Steffen, M. 1990, *A&A*, **239**, 443
- Steiner, O., Züger, F., & Belluzzi, L. 2016, *A&A*, **586**, A42
- Štěpán, J., & Trujillo Bueno, J. 2013, *A&A*, **557**, A143
- Trefethen, L. N. 1999, *Computation of Pseudospectra*, Technical Report NA-99/03 (Cambridge: Cambridge Univ. Press)
- Trefethen, L. N., & Embree, M. 2005, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators* (Princeton, NJ: Princeton Univ. Press)
- Uitenbroek, H. 2001, *ApJ*, **557**, 389