A New Perspective on the Wrapping Effect in Interval Methods for Initial Value Problems for Ordinary Differential Equations

Nedialko S. Nedialkov[1] and Kenneth R. Jackson[2]

[1] Department of Computing and Software, McMaster University,
Hamilton, Ontario, L8S 4L7, Canada.
nedialk@mcmaster.ca

[2] Department of Computer Science, University of Toronto,
Toronto, Ontario, M5S 3G4, Canada.
krj@cs.toronto.edu

Abstract. The problem of reducing the wrapping effect in interval methods for initial value problems for ordinary differential equations has usually been studied from a geometric point of view. We develop a new perspective on this problem by linking the wrapping effect to the stability of the interval method. Thus, reducing the wrapping effect is related to finding a more stable scheme for advancing the solution. This allows us to exploit eigenvalue techniques and to avoid the complicated geometric arguments used previously.

We study the stability of several anti-wrapping schemes, including Lohner's QR-factorization method, which we show can be viewed as a simultaneous iteration for computing the eigenvalues of an associated matrix. Using this connection, we explain how Lohner's method improves the stability of an interval method and show that, for a large class of problems, its global error is not much bigger than that of the corresponding point method.

1 Introduction

We consider interval methods (often called validated methods) for initial value problems (IVPs) for ordinary differential equations (ODEs) of the form

$$y' = f(y), \quad y(t_0) = y_0, \tag{1}$$

where $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is sufficiently differentiable. These methods [19,23] compute bounds that are guaranteed to contain the solution of (1) at points t_j, $j = 1, 2, \ldots, m$ in the interval $(t_0, t_m]$, for some $t_m > t_0$. They are usually based on Taylor series or our recent extension of Hermite-Obreschkoff schemes to interval methods [17,18]. Both types of methods use high-order derivatives; thus the requirement that f is sufficiently differentiable is essential. To simplify the notation, we apply our methods to autonomous systems throughout this paper. However, the methods can be extended easily to nonautonomous systems.

To date, only one-step methods have been used as the basis for interval methods for (1). Each step of these methods usually consists of two phases. We refer to them as Algorithm I and Algorithm II [19]. On an integration step from t_{j-1} to t_j,

Algorithm I validates existence and uniqueness of the solution of (1) for all $t \in [t_{j-1}, t_j]$ and computes a priori bounds for this solution for all $t \in [t_{j-1}, t_j]$, [19,20]; and

Algorithm II computes tight bounds for the solution of (1) at t_j.

A major problem in the second phase is the *wrapping effect* [16]. It occurs when a solution set that is not a box in $\mathbb{R}^n$, $n \geq 2$, is enclosed, or wrapped, by a box on each integration step. (By a box we mean a parallelepiped with edges that are parallel to the axes of an orthogonal coordinate system.) As a result of such a wrapping, an overestimation is often introduced on each integration step. These overestimations accumulate as the integration proceeds, and the computed bounds may soon become unacceptably large.

The problem of reducing the wrapping effect has usually been studied from a geometric perspective as finding an enclosing set that introduces as little overestimation of the enclosed set as possible. For example, parallelepipeds [5,13,15,16], ellipsoids [9,10,12,21], convex polygons [24], and zonotopes [14] have been employed in reducing the wrapping effect. Other, rather complicated techniques can be found in [3] and [7]. Recently, Berz and Makino [4] proposed a method for reducing the wrapping effect based on Taylor series expansions not only with respect to time, as in the methods discussed in [19] and [23], but also with respect to the initial conditions.

In this paper, we develop a new point of view on the wrapping effect by viewing it from the perspective of the stability of an interval method. To be more specific, we link the wrapping effect to the stability of an interval Taylor series (ITS) method for IVPs for ODEs and interpret the problem of reducing the wrapping effect as one of finding a more stable scheme for advancing the solution.

To study the stability of ITS methods, and thereby the wrapping effect, we employ eigenvalue techniques, which have proven so useful in the study of the stability of point methods (see for example [2] or [8]). To adopt such techniques, we do not investigate ITS methods for the general problem (1), but consider ITS methods for the simpler, linear, constant-coefficient problem

$$y' = By, \quad y(t_0) = y_0, \tag{2}$$

where $B \in \mathbb{R}^{n \times n}$ and $n \geq 2$. (The wrapping effect does not occur on scalar problems.) We also restrict our considerations to Taylor series methods with a constant number of terms and a constant stepsize.

We study the stability of several schemes for reducing the wrapping effect. In particular, we show that Lohner's QR-factorization method, which is currently the most successful general anti-wrapping scheme, can be viewed as a simultaneous iteration (see for example [25]) for computing the eigenvalues of an associated matrix. Furthermore, using some properties of this iteration, we explain how the QR-factorization method improves the stability of the associated ITS method.

The new perspective on the wrapping effect that we develop also suggests how to search for more stable schemes for advancing the solution using eigenvalue techniques. Hence, we can avoid complicated geometric arguments that have been used to date when constructing methods for reducing the wrapping effect.

This paper is organized as follows. In Sect. 2, we introduce notation and definitions that we use later. In Sect. 3, we present a traditional explanation of the wrapping effect in terms of wrapping a parallelepiped, which is not a box in $\mathbb{R}^n$, by a box. Then, in Sect. 4, we view the wrapping effect as a source of instability in an ITS method and describe how the problem of reducing the wrapping effect can be viewed as the problem of finding a more stable scheme for advancing the solution. In Sect. 5, we outline two methods for reducing the wrapping effect, namely the parallelepiped method and Lohner's QR-factorization method [15]. Section 6 shows that the parallelepiped method can be viewed as the power method for computing the largest eigenvalue of the associated matrix for advancing the solution. This section also explains why the parallelepiped method often fails due to nearly-singular matrices and large errors, while Sect. 7 identifies problems for which this method works well. Section 8 explains how Lohner's QR-factorization method can be interpreted as a simultaneous iteration for computing the eigenvalues of the associated matrix for advancing the solution. Using some properties of this iteration and the QR algorithm [6] for computing eigenvalues, we show that for a wide class of problems, the global error of Lohner's QR-factorization method is not much bigger than an upper bound for the global error of the corresponding point Taylor series (PTS) method, for the same stepsize and order. We also propose a simplified QR-factorization method and show that its global error is not much bigger than a bound on the global error of the corresponding PTS method. Finally, in Sect. 9, we summarize the importance of this work.

2 Preliminaries

In subsection 2.1, we introduce interval arithmetic and present some of its properties that we use later. For a more detailed exposition on interval arithmetic, see for example [1] or [16]. In subsection 2.2, we introduce some "non-interval" notation and definitions.

2.1 Interval Arithmetic

The set of intervals on the real line $\mathbb{R}$ is defined by

$$\mathbb{IR} = \left\{ [a] = [\underline{a}, \bar{a}] \mid \underline{a}, \bar{a} \in \mathbb{R},\ \underline{a} \leq \bar{a} \right\}.$$

If $\underline{a} = \bar{a}$ then $[a]$ is a *point* interval; if $\underline{a} \geq 0$ then $[a]$ is *nonnegative* ($[a] \geq 0$); and if $\underline{a} = -\bar{a}$ then $[a]$ is *symmetric*. Two intervals $[a]$ and $[b]$ are equal if $\underline{a} = \underline{b}$ and $\bar{a} = \bar{b}$.

Let $[a]$ and $[b] \in \mathbb{IR}$, and $\circ \in \{+, -, *, /\}$. The interval-arithmetic operations are defined [16] by

$$[a] \circ [b] = \{ x \circ y \mid x \in [a],\ y \in [b] \}, \qquad 0 \notin [b] \text{ when } \circ = /,$$

which can be written in the following equivalent form (we omit $*$ in the notation):

$$[a] + [b] = [\underline{a} + \underline{b}, \bar{a} + \bar{b}], \tag{3}$$

$$[a] - [b] = [\underline{a} - \bar{b}, \bar{a} - \underline{b}], \tag{4}$$

$$[a][b] = [\min\left\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\right\}, \max\left\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\right\}], \tag{5}$$

$$[a]/[b] = [\underline{a}, \bar{a}][1/\bar{b}, 1/\underline{b}], \qquad 0 \notin [b]. \tag{6}$$

We have an inclusion of intervals

$$[a] \subseteq [b] \iff \underline{a} \geq \underline{b} \text{ and } \bar{a} \leq \bar{b}.$$

The interval-arithmetic operations are *inclusion monotone*. That is, for real intervals $[a]$, $[a_1]$, $[b]$, and $[b_1]$ such that $[a] \subseteq [a_1]$ and $[b] \subseteq [b_1]$,

$$[a] \circ [b] \subseteq [a_1] \circ [b_1], \qquad \circ \in \{ +, -, *, / \}.$$

Although interval addition and multiplication are associative, the distributive law does not hold in general [1]. That is, we can easily find three intervals $[a]$, $[b]$, and $[c]$, for which

$$[a]([b] + [c]) \neq [a][b] + [a][c].$$

However, for any three intervals $[a]$, $[b]$, and $[c]$, the subdistributive law

$$[a]([b] + [c]) \subseteq [a][b] + [a][c],$$

does hold. Moreover, there are important cases in which the distributive law

$$[a]([b] + [c]) = [a][b] + [a][c]$$

does hold. For example, it holds if $[b][c] \geq 0$, if $[a]$ is a point interval, or if $[b]$ and $[c]$ are symmetric. In particular, for $\alpha \in \mathbb{R}$, which can be interpreted as the point interval $[\alpha, \alpha]$, and intervals $[a]$ and $[b]$, we have

$$\alpha([a] + [b]) = \alpha[a] + \alpha[b]. \tag{7}$$

For an interval $[a]$, we define the width and the midpoint of $[a]$ as

$$\mathrm{w}([a]) = \bar{a} - \underline{a} \quad \text{and} \tag{8}$$

$$\mathrm{m}([a]) = (\bar{a} + \underline{a})/2, \tag{9}$$

respectively [16]. Using (3–5) and (8), one can easily show that

$$\mathrm{w}([a] \pm [b]) = \mathrm{w}([a]) + \mathrm{w}([b]) \quad \text{and} \tag{10}$$

$$\mathrm{w}(\alpha[a]) = |\alpha|\, \mathrm{w}([a]), \tag{11}$$

for any $\alpha \in \mathbb{R}$.

By an *interval vector* we mean a vector with interval components. We denote the set of n-dimensional real interval vectors by $\mathbb{IR}^n$. For $[a] \in \mathbb{IR}^n$, we denote its ith component by $[a_i]$.

Addition of interval vectors is defined component-wise. That is, for $[a] \in \mathbb{IR}^n$ and $[b] \in \mathbb{IR}^n$, the ith component of $[c] = [a] \pm [b]$ is

$$[c_i] = [a_i] \pm [b_i], \quad \text{for } i = 1, 2, \ldots, n. \tag{12}$$

For a real matrix $A \in \mathbb{R}^{n \times n}$ with components a_{ij} and an interval vector $[b] \in \mathbb{IR}^n$ with components $[b_j]$, the components of $[c] = A[b]$ are given by

$$[c_i] = \sum_{j=1}^{n} a_{ij}[b_j], \quad \text{for } i = 1, 2, \ldots, n. \tag{13}$$

For $A \in \mathbb{R}^{n \times n}$, $[b] \in \mathbb{IR}^n$, and $[c] \in \mathbb{IR}^n$, it follows from (7) and (12–13) that

$$A([b] + [c]) = A[b] + A[c]. \tag{14}$$

We define inclusion, width, and midpoint component-wise for interval vectors:

$$[a] \subseteq [b] \iff [a_i] \subseteq [b_i] \quad (\text{for all } i), \tag{15}$$

$$\mathrm{w}([a]) = \big(\mathrm{w}([a_1]), \mathrm{w}([a_2]), \ldots, \mathrm{w}([a_n])\big)^T, \quad \text{and} \tag{16}$$

$$\mathrm{m}([a]) = \big(\mathrm{m}([a_1]), \mathrm{m}([a_2]), \ldots, \mathrm{m}([a_n])\big)^T. \tag{17}$$

Let A be a real matrix. We denote by $|A|$ the matrix with components $|a_{ij}|$, for all i and j. Let $A \in \mathbb{R}^{n \times n}$, $[b] \in \mathbb{IR}^n$, and $[c] \in \mathbb{IR}^n$. We can easily obtain from (7–8), (10–13), and (16) that

$$\mathrm{w}([b] \pm [c]) = \mathrm{w}([b]) + \mathrm{w}([c]), \tag{18}$$

$$\mathrm{w}\big(A[b]\big) = |A|\mathrm{w}([b]), \quad \text{and} \tag{19}$$

$$\mathrm{w}\big(A([b] + [c])\big) = |A|\mathrm{w}([b]) + |A|\mathrm{w}([c]). \tag{20}$$

Definition 1. Let $A \in \mathbb{R}^{n\times n}$ and $[x] \in \mathbb{IR}^n$. The wrapping of the parallelepiped

$$\{\, Ax \mid x \in [x] \,\}$$

is the box specified by the interval vector $A[x]$, [11]; see Fig. 1.

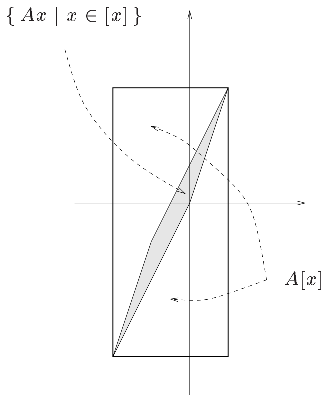


Fig. 1. The wrapping of the parallelepiped $\{\, Ax \mid x \in [x] \,\}$ by the box $A[x]$, where $A = \begin{pmatrix} 1 & -2 \\ 3 & -4 \end{pmatrix}$, $[x] = ([0,1],[0,1])^T$, and $A[x] = ([-2,1],[-4,3])^T$

2.2 Miscellanea

Throughout this paper, we use the infinity norm, which, to reduce the amount of notation, we denote simply by $\|\cdot\|$. By the norm equivalence

$$m\|x\|_Y \le \|x\| \le M\|x\|_Y, \tag{21}$$

where $\|\cdot\|_Y$ is any vector norm on $\mathbb{R}^n$, $x \in \mathbb{R}^n$, and m and M are constants that depend on $\|\cdot\|_Y$ but not x, the bounds that we derive in the infinity norm can be translated into any other norm.

Definition 2. For $A \in \mathbb{R}^{n\times n}$, the condition number of A is

$$\mathrm{cond}(A) = \|A\|\,\|A^{-1}\|.$$

Definition 3. The spectral radius of a matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\{\, \lambda_i \,:\, i = 1,2,\ldots,n \,\}$, is

$$\rho(A) = \max_{i=1,2,\ldots,n} \{\, |\lambda_i| \,\}.$$

It is not hard to show that

$$\rho(A) \le \rho(|A|). \tag{22}$$

Let $\mathrm{Re}(z)$ be the real part of a complex number z.

Definition 4. Let the eigenvalues of B in (2) be $\{\, \lambda_i \,:\, i = 1, 2, \ldots, n \,\}$ and assume $\mathrm{Re}(\lambda_i) < 0$ for all $i = 1, 2, \ldots, n$. An interval method for computing enclosures $[y_j]$ of the true solution of (2) is *asymptotically unstable* if, when applied with constant stepsizes and order to (2),

$$\lim_{j \to \infty} \|\mathrm{w}([y_j])\| = \infty.$$

3 How the Wrapping Effect Arises in Interval Methods for IVPs for ODEs: A Traditional Explanation

In this section, we derive an ITS method for implementing Algorithm II for (2) and show how the wrapping effect arises in this method.

Let

$$T_r(z) = \sum_{i=0}^{r} \frac{z^i}{i!}$$

be the Taylor polynomial of degree r. Since, in this paper, we consider only Taylor series methods with a constant number of terms k and a constant stepsize $h = t_j - t_{j-1}$, we denote

$$T = T_{k-1}(hB) = \sum_{i=0}^{k-1} \frac{(hB)^i}{i!}. \tag{23}$$

If $y(t)$ is the solution of (2), then

$$y(t_j) = T y(t_{j-1}) + \check{z}_j, \tag{24}$$

where

$$\check{z}_j = \frac{(hB)^k}{k!} \check{y}_{j-1},$$

and the ith component of $\check{y}_{j-1}$ is the ith component of $y(t)$ evaluated at some $\xi_{j-1,i} \in [t_{j-1}, t_j]$.

Suppose that we have computed a tight enclosure $[y_{j-1}]$ on the $(j-1)$st step such that

$$y(t_{j-1}) \in [y_{j-1}]. \tag{25}$$

Suppose also that on the next step, in Algorithm I, we have verified existence and uniqueness of the solution of

$$u' = Bu, \quad u(t_{j-1}) = y_{j-1} \in [y_{j-1}], \tag{26}$$

for all $y_{j-1} \in [y_{j-1}]$ and all $t \in [t_{j-1}, t_j]$, and have computed an a priori enclosure $[\tilde{y}_{j-1}]$ for the solution $u(t)$ of (26) such that

$$u(t) \in [\tilde{y}_{j-1}],$$

for all $t \in [t_{j-1}, t_j]$ and all $y_{j-1} \in [y_{j-1}]$, [19,20]. Then, $y(t) \in [\tilde{y}_{j-1}]$ for all $t \in [t_{j-1}, t_j]$. Hence, $\check{y}_{j-1} \in [\tilde{y}_{j-1}]$, and

$$\check{z}_j = \frac{(hB)^k}{k!}\check{y}_{j-1} \in \frac{(hB)^k}{k!}[\tilde{y}_{j-1}] \equiv [z_j]. \tag{27}$$

Using (24–25) and (27), we derive

$$\begin{aligned} y(t_j) &= Ty(t_{j-1}) + \check{z}_j \\ &\in \{\, Ty_{j-1} + [z_j] \mid y_{j-1} \in [y_{j-1}] \,\} &(28)\\ &\subseteq T[y_{j-1}] + [z_j] \equiv [y_j]. &(29) \end{aligned}$$

We wish to compute the set in (28), but, in practice, we compute instead the more tractable interval vector $[y_j]$ in (29). Thus, we obtain the simple interval Taylor series scheme

$$[y_j] = T[y_{j-1}] + [z_j] \tag{30}$$

for implementing Algorithm II.

We refer to this scheme as the *direct method* [17]. Here, $[z_j]$ can be viewed as an enclosure of the local error of the direct method, and $\mathrm{w}([z_j])$ can be interpreted as the size of the bound on this error. We can keep it small by decreasing the stepsize or increasing the order. However, $T[y_{j-1}]$ can be a large overestimation of the set $\{\, Ty_{j-1} \mid y_{j-1} \in [y_{j-1}] \,\}$. The reason for this is that the set $\{\, Ty_{j-1} \mid y_{j-1} \in [y_{j-1}] \,\}$ is not generally a box, but it is enclosed, or wrapped, by the box $T[y_{j-1}]$; see Fig. 1. Therefore, if we compute on each step products of the form $T[y_{j-1}]$, we may incur a wrapping on each step, resulting in unacceptably wide bounds for the solution of (2).

This can also be seen from the following considerations (see also [17] or [23]). Using (14), we derive from (30) that

$$\begin{aligned} [y_j] &= T[y_{j-1}] + [z_j] = T(T[y_{j-2}] + [z_{j-1}]) + [z_j] \\ &= T(T[y_{j-2}]) + T[z_{j-1}] + [z_j] \\ &\;\;\vdots \\ &= T^j y_0 + \underbrace{T\big(\cdots T(T}_{(j-1)\ T\text{'s}}[z_1]) \cdots\big) + \cdots \\ &\qquad + T(T[z_{j-2}]) + T[z_{j-1}] + [z_j]. \end{aligned} \tag{31}$$

Consider how $[z_1]$ propagates in the direct method. That is, consider the term $T\big(\cdots T(T[z_1]) \cdots\big)$ in (31). There is usually one wrapping in computing $T[z_1]$, two wrappings in computing $T(T[z_1])$, and $(j-1)$ wrappings in computing $T\big(\cdots T(T[z_1])\cdots\big)$. Thus, the excess caused by a wrapping on each step accumulates, possibly causing an unacceptably large wrapping effect in propagating $[z_1]$. Figure 2 illustrates the overestimation of the set $\left\{\, A^2x \mid x \in [x] \,\right\}$, when it is wrapped twice in computing $A(A[x])$, where A and $[x]$ are given in Fig. 1.

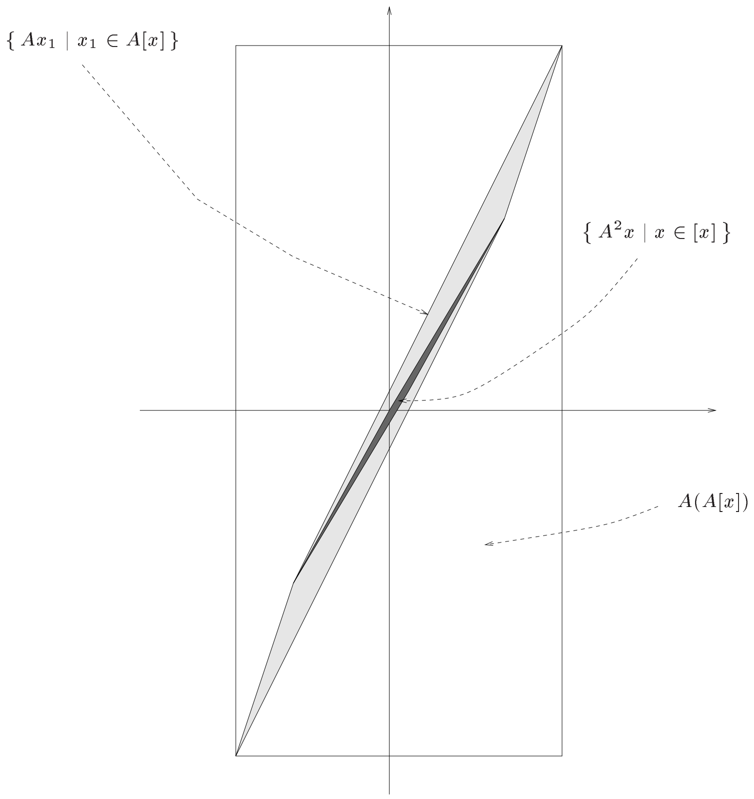


Fig. 2. The box $A[x]$ from Fig. 1 is mapped into $\{\, Ax_1 \mid x_1 \in A[x] \,\}$, which is wrapped by $A(A[x])$

Below, we present further insights into the wrapping effect in the direct method. From (24) and (27),

$$\begin{aligned} y(t_j) &= Ty(t_{j-1}) + \check{z}_j \\ &= T^j y_0 + T^{j-1}\check{z}_1 + \cdots + T\check{z}_{j-1} + \check{z}_j \\ &\in T^j y_0 + T^{j-1}[z_1] + \cdots + T[z_{j-1}] + [z_j] \equiv [v_j]. \end{aligned} \tag{32}$$

Note that we are interested in computing an easily-representable enclosure of the set

$$\left\{\, T^{j-1} z_1 \mid z_1 \in [z_1] \,\right\} \tag{33}$$

such that this enclosure contains as little overestimation of (33) as possible. If we enclose (33) by $T^{j-1}[z_1]$, we introduce at most one wrapping, while in (31), we can have $(j-1)$ wrappings when enclosing (33) by $T\big(\cdots T(T[z_1])\cdots\big)$. Note also that if $|T|^{j-1} \gg |T^{j-1}|$, as often happens, then typically

$$\begin{aligned} \mathrm{w}\Big(T\big(\cdots T(T[z_1])\cdots\big)\Big) &= |T|^{j-1}\mathrm{w}([z_1]) \\ &\gg |T^{j-1}|\mathrm{w}([z_1]) = \mathrm{w}\big(T^{j-1}[z_1]\big). \end{aligned}$$

Since in this paper we consider only problems with point initial conditions, we can interpret $\mathrm{w}([y_j])$ as the size of the bound on the global error of the direct method on the jth step. Taking widths in (31) and (32) and using (18–19), we derive

$$\begin{aligned} \mathrm{w}([y_j]) &= |T|\mathrm{w}([y_{j-1}]) + \mathrm{w}([z_j]) \\ &= \sum_{i=1}^{j} |T|^{j-i}\, \mathrm{w}([z_i]), \end{aligned} \tag{34}$$

while

$$\mathrm{w}([v_j]) = \sum_{i=1}^{j} |T^{j-i}|\, \mathrm{w}([z_i]). \tag{35}$$

Obviously, if $|T|^{j-i} \gg |T^{j-i}|$ for some $i \in \{1,\ldots,j-1\}$, then $\mathrm{w}([y_j])$ may be much larger than $\mathrm{w}([v_j])$. Therefore, because of the wrapping effect, or $|T|^{j-i} \gg |T^{j-i}|$, the global error $\mathrm{w}([y_j])$ of the direct method often becomes unacceptably large.

The reason why we do not use a scheme of the form (32) in practice is that it can be significantly more expensive than (30), because we have to store the necessary powers of T or recompute them on each step. Moreover, if the ITS method uses a stepsize control, it follows from (23) that the $T_j = T_{k-1}(h_j B)$ are normally not all equal. Hence, we have to store all matrix products of the form T_1, T_1T_2, …, $T_1T_2\cdots T_j$.

4 The Wrapping Effect as a Source of Instability in Interval Methods for IVPs for ODEs

Consider now the point Taylor series (PTS) method given by

$$u_j = T u_{j-1}, \quad u_0 = y_0. \tag{36}$$

Denote by

$$\delta_j = y(t_j) - u_j \tag{37}$$

the global error of this method at t_j, $j = 1, 2, \ldots,$. Then, from (24) and (36–37), we can express the global error of the PTS method at t_j using the local truncation errors $\check{z}_i$, $i = 1, 2, \ldots, j$, as

$$\begin{aligned} \delta_j &= T\delta_{j-1} + \check{z}_j \\ &= \sum_{i=1}^{j} T^{j-i}\, \check{z}_i. \end{aligned} \tag{38}$$

In this section, we investigate how $\|\mathrm{w}([y_j])\|$ compares to $\|\delta_j\|$, where $\mathrm{w}([y_j])$ is given in (34). An important observation is that the global error in the PTS method propagates with T, while the global error in the direct method propagates with $|T|$. Assuming that T and $|T|$ are diagonalizable (to simplify the analysis), we show that

- if $\rho(T) = \rho(|T|)$, then the bounds on $\|\mathrm{w}([y_j])\|$ and $\|\delta_j\|$ may not be much different, but
- if $\rho(T) < \rho(|T|)$, then $\|\mathrm{w}([y_j])\|$ may be much larger than $\|\delta_j\|$.

Recall that $\rho(T) \le \rho(|T|)$; see subsection 2.2.

Let V_T and $V_{|T|}$ be the matrices that diagonalize T and $|T|$, respectively. Let D_T and $D_{|T|}$ be the associated diagonal matrices. Then

$$T = V_T^{-1} D_T V_T \quad \text{and} \tag{39}$$

$$|T| = V_{|T|}^{-1} D_{|T|} V_{|T|}. \tag{40}$$

Assume also that

$$\|\check{z}_i\| \le \delta \quad \text{and} \tag{41}$$

$$\|\mathrm{w}([z_i])\| \le \beta, \quad i = 1, 2, \ldots, j, \tag{42}$$

for some constants δ and β, respectively.

We derive from (38–39) and (41) that

$$\left\|\delta_j\right\| \le \delta \left(\mathrm{cond}(V_T) \sum_{i=1}^{j-1} \big(\rho(T)\big)^i + 1 \right), \tag{43}$$

and we derive from (34), (40), and (42) that

$$\left\|\mathrm{w}([y_j])\right\| \le \beta \left(\mathrm{cond}(V_{|T|}) \sum_{i=1}^{j-1} \big(\rho(|T|)\big)^i + 1 \right). \tag{44}$$

Obviously, if $\rho(T) = \rho(|T|)$, the upper bound for $\|\mathrm{w}([y_j])\|$ is approximately a constant times the upper bound for $\|\delta_j\|$.

Example 1. Consider

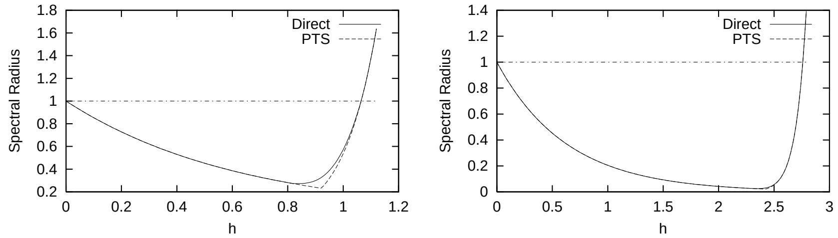
$$y' = By = \begin{pmatrix} -2 & 1 \\ 1 & -4 \end{pmatrix} y, \quad y(0) = (1,1)^T. \tag{45}$$

Since this is a quasi-monotone problem, the wrapping effect should not cause a significant overestimation in propagating the error in the direct method on this problem [22].

In Figs. 3(a) and 3(b), we plot $\rho\big(T_{k-1}(hB)\big)$ and $\rho\big(|T_{k-1}(hB)|\big)$ versus h for $k = 10$ and $k = 30$.[1] For the ranges of stepsize considered,

$$\rho\big(T_{k-1}(hB)\big) \approx \rho\big(|T_{k-1}(hB)|\big). \tag{46}$$

Note that the eigenvalues of B are (approximately) -1.5858 and -4.4142.



(**a**) $k = 10$ (**b**) $k = 30$

Fig. 3. Graphs of Direct $\equiv \rho\big(|T_{k-1}(hB)|\big)$ and PTS $\equiv \rho\big(T_{k-1}(hB)\big)$ versus h for $k = 10$ and $k = 30$ for Example 1

Using (46) together with (43) and (44), we see that the bound on $\|\mathrm{w}([y_j])\|$ cannot be much bigger than the bound on $\|\delta_j\|$. Therefore, if these bounds are fairly tight, the wrapping effect does not cause a significant overestimation of the global error in the direct method on this problem.

To support this claim, we integrated (45) for $t \in [0, 20]$ with the PTS method and the direct method.[2] In both cases, we used $h = 0.1$, and $k = 30$. In Fig. 4, we plot the base-10 logarithm of the global error in these two methods versus $t_j = jh = 0.1j$.

Now, we show that, if $\rho(T) < \rho(|T|)$, $\|\mathrm{w}([y_j])\|$ may be much bigger than $\|\delta_j\|$. Denote the eigenvalues of T by $\gamma_1, \gamma_2, \ldots, \gamma_n$ and assume

$$|\gamma_1| > |\gamma_2| \geq |\gamma_3| \geq \cdots \geq |\gamma_{n-1}| \geq |\gamma_n|. \tag{47}$$

[1] The computations are performed in Matlab. The plots in this paper are produced with Gnuplot.

[2] The PTS method is implemented in Matlab. The interval methods in this paper are implemented in the VNODE [17] package.

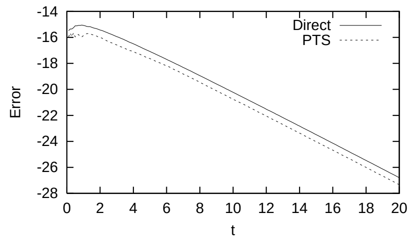


Fig. 4. $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ versus t_j for Example 1

Assume also for simplicity that T has a full set of linearly independent eigenvectors $v_1, v_2, \ldots, v_n$, where $Tv_i = \gamma_i v_i$ for $i = 1, 2, \ldots, n$. We can express $\check{z}_1$ as a linear combination of the eigenvectors of T,

$$\check{z}_1 = \sum_{i=1}^{n} c_i v_i,$$

where $c_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$. If $c_1 \neq 0$, then

$$\begin{aligned} T^{j-1}\check{z}_1 &= \gamma_1^{j-1} \left(c_1 v_1 + \left(\frac{\gamma_2}{\gamma_1}\right)^{j-1} c_2 v_2 + \cdots + \left(\frac{\gamma_n}{\gamma_1}\right)^{j-1} c_n v_n \right) \\ &\to \gamma_1^{j-1} c_1 v_1, \quad \text{as } j \to \infty. \end{aligned}$$

Thus, for sufficiently large j,

$$T^{j-1}\check{z}_1 \approx \big(\rho(T)\big)^{j-1} c_1 v_1. \tag{48}$$

(In practice, $T^{j-1}\check{z}_1$ is computed in floating-point arithmetic with an iteration of the form $p_l = Tp_{l-1}$, $p_1 = \check{z}_1$, $l = 2, 3, \ldots, j$. In general, if $c_1 = 0$, then roundoff errors in this iteration eventually introduce a nonzero component in the direction of v_1.)

Similarly, denote the eigenvalues of $|T|$ by $\nu_1, \nu_2, \ldots, \nu_n$ and assume

$$|\nu_1| > |\nu_2| \geq |\nu_3| \geq \cdots \geq |\nu_{n-1}| \geq |\nu_n|.$$

Assume also for simplicity that $|T|$ has a full set of linearly independent eigenvectors $w_1, w_2, \ldots, w_n$, where $|T|w_i = \nu_i w_i$ for $i = 1, 2, \ldots, n$. We can express $\mathrm{w}([z_1])$ as a linear combination of the eigenvectors of $|T|$,

$$\mathrm{w}([z_1]) = \sum_{i=1}^{n} d_i w_i,$$

where $d_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$. If $d_1 \neq 0$,

$$\begin{aligned} T^{j-1}\mathrm{w}([z_1]) &= \nu_1^{j-1} \left(d_1 w_1 + \left(\frac{\nu_2}{\nu_1}\right)^{j-1} d_2 w_2 + \cdots + \left(\frac{\nu_n}{\nu_1}\right)^{j-1} d_n w_n \right) \\ &\to \nu_1^{j-1} d_1 w_1, \quad \text{as } j \to \infty. \end{aligned}$$

Thus, for sufficiently large j,

$$|T|^{j-1}\mathrm{w}([z_1]) \approx \big(\rho(|T|)\big)^{j-1} d_1 w_1. \tag{49}$$

From (48) and (49), we see that, for sufficiently large j,

$$\frac{\big\||T|^{j-1}\mathrm{w}([z_1])\big\|}{\big\|T^{j-1}\check{z}_1\big\|} \approx \left(\frac{\rho(|T|)}{\rho(T)}\right)^{j-1} \frac{\|d_1 w_1\|}{\|c_1 v_1\|}. \tag{50}$$

Hence, if $\rho(T) < \rho(|T|)$, $\big\||T|^{j-1}\mathrm{w}([z_1])\big\|$ may be much bigger than $\|T^{j-1}z_1\|$. This implies that the global error bound for the direct method may be significantly larger than the global error for the PTS method, as j becomes sufficiently large.

For example, assume $\rho(T) < 1$. Then the PTS method computes approximations for which $\lim_{j\to\infty}\|u_j\| = 0$. However, $\rho(T) < 1$ does not imply $\rho(|T|) < 1$. Therefore, it is possible to have $\rho(T) < 1$, but $\rho(|T|) > 1$. Then from (49),

$$\big\||T|^{j-1}\mathrm{w}([z_1])\big\| \approx \big(\rho(|T|)\big)^{j-1}\|d_1 v_1\| \to \infty, \quad \text{as } j \to \infty,$$

and our bound on $\|\mathrm{w}([y_j])\| \to \infty$, as $j \to \infty$. That is, the direct method may be asymptotically unstable.

Example 2. The problem

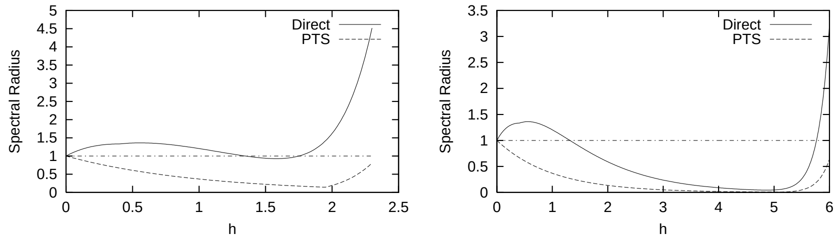
$$y' = By = \begin{pmatrix} 1 & -2 \\ 3 & -4 \end{pmatrix} y, \quad y(0) = (1, -1)^T \tag{51}$$

is an example for which the true solution components converge rapidly to zero, but the wrapping effect can cause significant overestimation of the error [15,17]. Here, the eigenvalues of B are -1 and -2.

Let us look at how our analysis of the direct method applies to (51). For this purpose, we plot in Figs. 5(a) and 5(b) $\rho\big(T_{k-1}(hB)\big)$ and $\rho\big(|T_{k-1}(hB)|\big)$ versus h for $k = 10$ and $k = 30$.

Consider the case $k = 30$. For $h \lessapprox 6.09$, $\rho\big(T_{29}(hB)\big) < 1$, and the PTS method computes approximations for which $\lim_{j\to\infty}\|u_j\| = 0$. However, for $h \lessapprox 1.32$ and $h \gtrapprox 5.7$, $\rho\big(|T_{29}(hB)|\big) > 1$, and the bounds for the direct method blow-up, suggesting that $\lim_{j\to\infty}\|\mathrm{w}([y_j])\| = \infty$. To support this conjecture, we used the direct method with $h = 0.2$ and $k = 30$ to compute $\|\mathrm{w}([y_j])\| \approx 5.2 \times 10^5$ at $t = 40$.

It is interesting to note that for $1.32 \lessapprox h \lessapprox 5.7$, the direct method may not suffer significantly from the wrapping effect. For example, with $h = 1.5$, we computed with the direct method $\|\mathrm{w}([y_j])\| \approx 2.6 \times 10^{-14}$ at $t = 300$. Therefore, even if the direct method suffers from the wrapping effect for a particular problem and range of stepsizes, there might be another range of stepsizes for which the wrapping effect is not a significant difficulty for this method.



(**a**) $k = 10$ (**b**) $k = 30$

Fig. 5. Graphs of Direct $\equiv \rho\big(|T_{k-1}(hB)|\big)$ and PTS $\equiv \rho\big(T_{k-1}(hB)\big)$ versus h for $k = 10$ and $k = 30$ for Example 2

Intuitively, we have associated the wrapping effect in an ITS method with its global error being significantly larger than the global error of the corresponding PTS method. We showed that when $\rho(T) = \rho(|T|)$, the wrapping effect is not a serious difficulty for the direct method. However, when $\rho(T) < \rho(|T|)$, the direct method may suffer significantly from the wrapping effect. In particular, if $\rho(T) < 1$ and $\rho(|T|) > 1$, the PTS method is stable, but the associated ITS method is likely asymptotically unstable. To reduce the wrapping effect, or improve the stability of an ITS method, we have to find a more stable scheme for advancing the solution.

5 The Parallelepiped and Lohner's QR-Factorization Methods

In this section, we describe and briefly discuss two methods, namely the parallelepiped and Lohner's QR-factorization methods, for reducing the wrapping effect.

Let

$$\begin{aligned} A_0 &= I, \\ \hat{y}_0 &= y_0, \\ s_j &= \mathrm{m}([z_j]), \qquad \text{and} \\ \hat{y}_j &= T\hat{y}_{j-1} + s_j \quad (j \geq 1), \end{aligned}$$

where I is the identity matrix, $[z_j]$ is defined in (27), and T is defined in (23). Let also $[r_0]$ be an interval vector with each component equal to $[0, 0]$. We compute (for $j \geq 1$)

$$[y_j] = T\hat{y}_{j-1} + (TA_{j-1})\,[r_{j-1}] + [z_j] \tag{52}$$

and propagate for the next step

$$[r_j] = \left(A_j^{-1} T A_{j-1}\right)[r_{j-1}] + A_j^{-1}\left([z_j] - s_j\right), \tag{53}$$

where $A_j \in \mathbb{R}^{n\times n}$ is nonsingular. We discuss in the next two subsections how A_j may be chosen.

5.1 The Parallelepiped Method

If $A_j = TA_{j-1}$, we obtain the parallelepiped method [15]. This choice for A_j has also been suggested in [5,13,16]. As Lohner points out [15], the parallelepiped method works particularly well for pure rotations, but breaks down when B (in (2)) has two eigenvalues with different real components. In this case, after a sufficiently long integration, A_j becomes nearly singular.

Using eigenvalue techniques, we show in Sect. 6 that the matrices A_j often become nearly singular and that the global error of this method is often much larger than the global error of the corresponding PTS method. In Sect. 7, we show that the parallelepiped method works well when all the eigenvalues of T are equal in magnitude, as observed by Lohner. In this case, the global error associated with the parallelepiped method is not much bigger than the bound for the global error of the corresponding PTS method.

5.2 Lohner's QR-Factorization Method

Lohner proposes [15] the following two schemes for selecting the transformation matrices A_j, $j \geq 1$ (in both schemes $A_0 = Q_0 = I$):

$$\begin{aligned} TA_{j-1} &= Q_j R_j, \\ A_j &= Q_j \end{aligned} \tag{54}$$

and

$$\begin{aligned} TA_{j-1}P_j &= Q_j R_j, \\ A_j &= Q_j . \end{aligned} \tag{55}$$

Here, Q_j is orthogonal, R_j is upper triangular, and P_j is a permutation matrix chosen as follows.

Let $l_{j-1,i}$ be the length (in the Euclidean norm) of the ith column of TA_{j-1} in (55) and $\left(\mathrm{w}([r_{j-1}])\right)_i$ be the ith component of $\mathrm{w}([r_{j-1}])$. Then P_j is such that the components of

$$\Big(l_{j-1,1}\left(\mathrm{w}([r_{j-1}])\right)_1,\, l_{j-1,2}\left(\mathrm{w}([r_{j-1}])\right)_2,\, \ldots,\, l_{j-1,n}\left(\mathrm{w}([r_{j-1}])\right)_n\Big)P_j$$

are in non-increasing order.

We refer to (52–53) with A_j's chosen from either (54) or (55) as Lohner's *QR-factorization method*. However, to distinguish between the methods, we also refer to (52–53) with A_j's from (54) as the *QR method* and to (52–53) with A_j's from (55) as the *QR-P method*.

In this section, we outline a geometric explanation of why the QR-factorization method is successful at reducing the wrapping effect. For more details see [15] or [19]. In Sect. 8, we use eigenvalue techniques to show how the QR method improves the stability of the direct and parallelepiped methods.

The interval vector $[r_j]$ in (53) can be interpreted as an enclosure of the global error (at t_j) that is propagated to the next step. Obviously, we are interested in keeping the overestimation in $[r_j]$ as small as possible.

Since $A_j = Q_j$ is orthogonal, $A_j^{-1} = Q_j^T$. Thus, we avoid the problem of having to compute the inverse of a nearly singular matrix. In addition, since Q_j is orthogonal,

$$|A_j^{-1}|\mathrm{w}([z_j]) = |Q_j^T|\mathrm{w}([z_j])$$

is not much bigger than $\mathrm{w}([z_j])$, which we keep sufficiently small (by reducing the stepsize or increasing the order).

The set

$$\{\, (TA_{j-1})r_{j-1} = (TQ_{j-1})r_{j-1} \mid r_{j-1} \in [r_j] \,\} \tag{56}$$

is generally a parallelepiped in $\mathbb{R}^n$. The set

$$\{\, (A_j^{-1}TA_{j-1})r_{j-1} = (Q_j^T TQ_{j-1})r_{j-1} \mid r_{j-1} \in [r_j] \,\} \tag{57}$$

is the parallelepiped in (56) rotated such that one of its edges is parallel to an axis in the standard coordinate system for $\mathbb{R}^n$. Hence, the box $(A_j^{-1}TA_{j-1})[r_{j-1}]$ that encloses the parallelepiped (57) is such that one of the edges of $(A_j^{-1}TA_{j-1})[r_{j-1}]$ is parallel to an edge of (57). Intuitively, if we wrap a parallelepiped by a box that matches one of the edges of this parallelepiped, we should normally introduce a smaller overestimation than if we do not match any of its edges. Furthermore, if we match the longest edge of the enclosed parallelepiped, as we do in the QR-P method, we should have a smaller overestimation than if we match an edge that is not the longest one.

6 Why the Parallelepiped Method Often Fails

Since $A_0 = I$ and $A_j = TA_{j-1}$ for $j \geq 1$, $A_j = T^j$. Substituting T^j for A_j in (52) and (53), we write the parallelepiped method as

$$[y_j] = T\hat{y}_{j-1} + T^j[r_{j-1}] + [z_j] \quad \text{and} \tag{58}$$

$$[r_j] = [r_{j-1}] + T^{-j}\left([z_j] - s_j\right). \tag{59}$$

Obviously, we require that T is nonsingular, which is the case when the stepsize, h, is sufficiently small.

In the next two subsections, we show that, as we integrate with this method,

- the matrices $A_j = T^j$ often become nearly singular, and
- $\mathrm{w}([y_j])$ often grows substantially, even when $\rho(T) < 1$. In this case, the PTS method computes approximations for which $\lim_{j\to\infty} \|u_j\| = 0$.

6.1 Nearly Singular Matrices

As in Sect. 4, we denote the eigenvalues of T by $\gamma_1, \gamma_2, \ldots, \gamma_n$ and assume again (to simplify the analysis) that T has a set of linearly independent eigenvectors $v_1, v_2, \ldots, v_n$, where $Tv_i = \gamma_i v_i$. Assume also that

$$|\gamma_1| > |\gamma_2| \geq |\gamma_3| \geq \cdots \geq |\gamma_{n-1}| \geq |\gamma_n| > 0, \tag{60}$$

where the last inequality, $|\gamma_n| > 0$, follows from our assumption that T is nonsingular.

We can view the ith column of T^j as the vector obtained from the power method applied to the unit vector e_i (for which the ith component is one and the rest are zero). For each $i = 1, 2, \ldots, n$, we can express e_i as a linear combination of the eigenvectors of T,

$$e_i = \sum_{j=1}^{n} c_{i,j} v_j, \quad c_{i,j} \in \mathbb{R}. \tag{61}$$

If $c_{i,1} \neq 0$ for all i's, then

$$\begin{aligned} T^j e_i &= \gamma_1^j \left(c_{i,1} v_1 + \left(\frac{\gamma_2}{\gamma_1}\right)^j c_{i,2} v_2 + \cdots + \left(\frac{\gamma_n}{\gamma_1}\right)^j c_{i,n} v_n \right) \\ &\to \gamma_1^j c_{i,1} v_1, \quad \text{as } j \to \infty. \end{aligned} \tag{62}$$

That is, for sufficiently large j, each column of T^j becomes nearly parallel to the eigenvector v_1.

In practice, we compute T^j in floating-point arithmetic using the iteration $A_j = TA_{j-1}$, $A_0 = I$. If $c_{i,1} = 0$ for some $i = 1, 2, \ldots, n$, the ith column of the computed T^j (in floating-point arithmetic) usually becomes nearly parallel to v_1 since roundoff errors eventually introduce a nonzero component in the direction of v_1. Note, though, that this is not always the case. For example, if T is diagonal, then roundoff errors do not introduce nonzero components in the direction of v_1 for the columns $T^j e_i$, where $i = 2, 3, \ldots, n$.

Since the computed T^j matrices become closer and closer to singular, as j increases, computing tight enclosures for T^{-j} becomes impossible, and the parallelepiped method breaks down. In addition, in floating-point arithmetic, the columns of T^j may overflow, if $|\gamma_1| > 1$, or underflow to zero, if $|\gamma_1| < 1$.

6.2 Large Errors

In this subsection, we assume, in addition to (60), that

$$|\gamma_{n-1}| > |\gamma_n|$$

holds. Let

$$G = \frac{|hB|^k}{k!}.$$

Since $[\tilde{y}_{j-1}]$ must contain $[y_{j-1}]$,

$$[z_j] = \frac{(hB)^k}{k!}[\tilde{y}_{j-1}] \supseteq \frac{(hB)^k}{k!}[y_{j-1}],$$

and

$$\mathrm{w}([z_j]) = \frac{|hB|^k}{k!}\mathrm{w}([\tilde{y}_{j-1}]) \geq \frac{|hB|^k}{k!}\mathrm{w}([y_{j-1}]) = G\,\mathrm{w}([y_{j-1}]). \tag{63}$$

Taking widths in (58–59) and using (18–19), we obtain[3] that

$$\mathrm{w}([y_j]) = |T^j|\mathrm{w}([r_{j-1}]) + \mathrm{w}([z_j]) \geq |T^j|\mathrm{w}([r_{j-1}]) \quad \text{and} \tag{64}$$
$$\mathrm{w}([r_j]) = \mathrm{w}([r_{j-1}]) + |T^{-j}|\mathrm{w}([z_j]) \geq |T^{-j}|\mathrm{w}([z_j]). \tag{65}$$

Let j be an odd number. Using (63–65), we derive

$$\begin{aligned}
\mathrm{w}([y_j]) &\geq |T^j|\mathrm{w}([r_{j-1}]) \geq |T^j||T^{-(j-1)}|\,\mathrm{w}([z_{j-1}]) \\
&\geq |T^j||T^{-(j-1)}|\,G\,\mathrm{w}([y_{j-2}]) \\
&\;\;\vdots \\
&\geq \left(|T^j||T^{-(j-1)}|\,G\right)\left(|T^{j-2}||T^{-(j-3)}|\,G\right)\cdots \\
&\qquad \left(|T^3||T^{-2}|\,G\right)\mathrm{w}([z_1]).
\end{aligned} \tag{66}$$

Denote by V_1 the $n \times n$ matrix with columns v_1 and by V_n the $n \times n$ matrix with columns v_n. Let

$$C_1 = \mathrm{diag}(c_{1,1}, c_{2,1}, \cdots, c_{n,1}) \quad \text{and} \quad C_n = \mathrm{diag}(c_{1,n}, c_{2,n}, \cdots, c_{n,n}),$$

where $c_{i,1}$ and $c_{i,n}$, $i = 1, 2, \ldots, n$, are the coefficients from (61). We now assume not only that $c_{i,1} \neq 0$ for all $i = 1, 2, \ldots, n$ (as in the previous subsection) but also that $c_{i,n} \neq 0$ for all $i = 1, 2, \ldots, n$. Now let

$$F = |\gamma_n||V_1 C_1||V_n C_n|. \tag{67}$$

We show that $|T^j||T^{-(j-1)}|$ in (66) becomes closer and closer to $|\gamma_1/\gamma_n|^j\,F$, as we integrate with the parallelepiped method.

For sufficiently large j, it follows from (62) that

$$|T^j| \approx |\gamma_1^j||V_1 C_1|. \tag{68}$$

[3] Throughout the rest of this paper, we shall not refer to properties from Sect. 2 when we use them.

Similarly,

$$|T^{-(j-1)}| \approx \frac{1}{|\gamma_n^{j-1}|}|V_n C_n|. \tag{69}$$

From (67–69),

$$|T^j||T^{-(j-1)}| \approx \left|\frac{\gamma_1}{\gamma_n}\right|^j |\gamma_n||V_1 C_1||V_n C_n| = \left|\frac{\gamma_1}{\gamma_n}\right|^j F. \tag{70}$$

From (66) and (70), we obtain for sufficiently large j that

$$\begin{aligned} \mathrm{w}([y_j]) &\geq |T^j||T^{-(j-1)}|\, G\, \mathrm{w}([y_{j-2}]) \\ &\approx \left|\frac{\gamma_1}{\gamma_n}\right|^j F\, G\, \mathrm{w}([y_{j-2}]). \end{aligned} \tag{71}$$

Since we assumed $|\gamma_1| > |\gamma_n|$, $|\gamma_1/\gamma_n|^j \gg 1$ for j sufficiently large. Therefore, $\mathrm{w}([y_j])$ can be much larger than $\mathrm{w}([y_{j-2}])$. Hence, as we integrate, we obtain enclosures with increasing widths. Thus, when $|\gamma_1| > |\gamma_2| \geq \cdots \geq |\gamma_{n-1} > |\gamma_n| > 0$, the parallelepiped method is likely asymptotically unstable. Note that this happens even when $\rho(T) = |\gamma_1| < 1$, in which case the PTS method is stable.

Example 3. Consider

$$y' = By = \begin{pmatrix} 1 & -2 \\ 3 & -4 \end{pmatrix} y, \quad y(0) = (1, -1)^T. \tag{72}$$

As noted in Example 2, the eigenvalues of B are -1 and -2. Let $h = 0.1$ and $k = 16$. The eigenvalues of $T = T_{15}(0.1B)$ are $\gamma_1 \approx 0.9048$ and $\gamma_2 \approx 0.8187$. Since $\gamma_1 > \gamma_2 > 0$, from the analysis in this subsection, we expect, for a sufficiently long integration of this problem, the parallelepiped to produce a large error and to break down.

We integrated (72) with the parallelepiped method with $h = 0.1$ and $k = 16$. After computing $[y_{134}]$, our solver could not continue since it could not enclose the inverse of $A_{134} = T^{134}$. Indeed, for $j = 134$,

$$T^{134} \approx 10^{-6} \begin{pmatrix} 4.545428 & -3.030284 \\ 4.545425 & -3.030281 \end{pmatrix}$$

is close to singular with $\mathrm{cond}\big(T^{134}\big) \approx 1.98 \times 10^7$.

Furthermore, the bound on the global error of the parallelepiped method at $t = 13.4$ is $\|\mathrm{w}([y_{134}])\| \approx 1.98 \times 10^2$, while the bound on the global error of the PTS method at $t = 13.4$ is $\|\delta_{134}\| \approx 1.8 \times 10^{-18}$. Note that, for the exact solution, $\|y(13.4)\| \approx 7.6 \times 10^{-6}$. To illustrate the growth in the error bounds in the parallelepiped method on this problem, we plot in Fig. 6 $\log_{10} \|\mathrm{w}([y_j])\|$ versus $t_j = jh = 0.1j$.

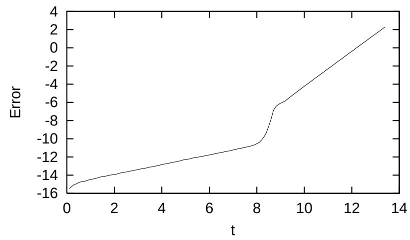


Fig. 6. Plot of $\log_{10} \|\mathrm{w}([y_j])\|$ versus t_j for the parallelepiped method on Example 3

7 When Does the Parallelepiped Method Work Well

Assume that the eigenvalues of T satisfy

$$|\gamma_1| \geq |\gamma_2| \geq |\gamma_3| \geq \cdots \geq |\gamma_{n-1}| \geq |\gamma_n| > 0$$

and that T can be diagonalized. Let V_T be the matrix that diagonalizes T and $D_T = \mathrm{diag}(\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_n)$. We derive an upper bound for the global error of the parallelepiped method and show that, if $|\gamma_1| = |\gamma_n|$, this bound is not much bigger than the bound in (43) for the global error of the corresponding PTS method.

Using (59) and then (58), we derive

$$\begin{aligned} \|\mathrm{w}([r_{j-1}])\| &\leq \|\mathrm{w}([r_{j-2}])\| + \big\||T^{-(j-1)}|\big\| \|\mathrm{w}([z_{j-1}])\| \\ &\leq \sum_{i=1}^{j-1} \big\||T^{-i}|\big\| \|\mathrm{w}([z_i])\| \end{aligned} \tag{73}$$

and

$$\begin{aligned} \|\mathrm{w}([y_j])\| &\leq \big\||T^{j}|\big\| \|\mathrm{w}([r_{j-1}])\| + \|\mathrm{w}([z_j])\| \\ &\leq \big\||T^{j}|\big\| \sum_{i=1}^{j-1} \big\||T^{-i}|\big\| \|\mathrm{w}([z_i])\| + \|\mathrm{w}([z_j])\|. \end{aligned} \tag{74}$$

Substituting

$$\begin{aligned} \big\||T^{j}|\big\| &= \big\|T^{j}\big\| \leq \mathrm{cond}(V_T)\, |\gamma_1|^j \quad \text{and} \\ \big\||T^{-i}|\big\| &= \big\|T^{-i}\big\| \leq \mathrm{cond}(V_T)\, \frac{1}{|\gamma_n|^i} \end{aligned}$$

into (74) and assuming that

$$\|\mathrm{w}([z_i])\| \leq \beta, \quad \text{for } i = 1, 2, \ldots, j,$$

we obtain that

$$\|\mathrm{w}([y_j])\| \leq \beta \left(\big(\mathrm{cond}(V_T)\big)^2 \sum_{i=1}^{j-1} \frac{|\gamma_1|^j}{|\gamma_n|^i} + 1 \right).$$

If $\rho(T) = |\gamma_1| = |\gamma_n|$, then

$$\|\mathrm{w}([y_j])\| \leq \beta \left(\left(\mathrm{cond}(V_T)\right)^2 \sum_{i=1}^{j-1} \left(\rho(T)\right)^i + 1 \right). \tag{75}$$

Comparing the bounds in (75) and (43), we see that, if $|\gamma_1| = |\gamma_n|$, the global error of the parallelepiped method is not much bigger than the bound for the global error of the corresponding point method. Therefore, the parallelepiped method works well for problems for which the eigenvalues of T are of the same magnitude. Such problems arise, for example, when the matrix B is skew symmetric, $B = -B^T$. Then, for a given h,

$$e^{hB} \left(e^{hB}\right)^T = e^{hB} e^{hB^T} = e^{hB} e^{-hB} = I,$$

which implies that e^{hB} is orthogonal. Hence, each eigenvalue of e^{hB} is of magnitude one. Since $T = T_{k-1}(hB)$ is normally a good approximation to e^{hB}, the eigenvalues of T are approximately of magnitude one.

Example 4. Consider the IVP problem

$$y' = By = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} y, \quad y(0) = (1, 0)^T.$$

Moore [16] used this problem to illustrate the wrapping effect. Lohner [15] reports that the parallelepiped method with $h = 0.32$ and $k = 10$ produces tight bounds for this problem at $t_m = 10000$.

Since B is skew symmetric, it is easy to see from our analysis why the parallelepiped method produces tight bounds for this problem: the eigenvalues of $T_9(0.32B)$ are $\approx 0.9492 \pm 0.3146i$, which are within 5.0×10^{-5} of magnitude one.

However, the eigenvalues of $|T|$ are ≈ 1.2638 and ≈ 0.6347. It follows from (49) that $|T|^{j-1}\mathrm{w}([z_1])$ grows like $(1.2638)^{j-1}$. Therefore, the global error in the direct method on this problem blows-up quickly.

8 How the QR Method Improves Stability

A key observation in the derivations of the results that follow is that Lohner's QR method, when applied with a constant stepsize and order to problem (2), is closely related to Francis' QR algorithm [6] for finding the eigenvalues of T. To see this, recall that in Lohner's QR method,

$$T A_{j-1} = Q_j R_j \tag{76}$$

$$A_j = Q_j, \tag{77}$$

where $A_0 = Q_0 = I$, Q_j is orthogonal, and R_j is upper triangular. Let

$$S_j = Q_j^T T Q_j \quad \text{and} \tag{78}$$

$$\widehat{Q}_j = Q_{j-1}^T Q_j \quad (j \geq 1). \tag{79}$$

Note that $S_0 = T$ since $Q_0 = I$. It follows from (78–79) that

$$S_{j-1} = \widehat{Q}_j R_j \quad \text{and} \tag{80}$$

$$S_j = R_j \widehat{Q}_j. \tag{81}$$

Now observe that the iteration (80–81) is just the *unshifted QR algorithm* for finding the eigenvalues of T. Observe also that (76–77) is the *simultaneous iteration* (see for example [25]) for computing the eigenvalues of T.

Using some properties of the QR algorithm, we show how Lohner's QR method improves the stability of an interval method for IVPs for ODEs. We consider first the case that T has eigenvalues of different magnitudes and then the case that T has eigenvalues of the same magnitude.

8.1 Eigenvalues of Different Magnitudes

Assuming that T is nonsingular with eigenvalues of different magnitudes, we derive an upper bound for the global error of Lohner's QR method. Then, we show that this bound is not much bigger than the bound (43) for the global error of the PTS method (36).

Based on our analysis of the QR method, we propose a simplified QR method, which we refer to as the *QR-S method*. We derive a simpler bound for its global error than the bound for the global error of the QR method. We also show that the global error of the QR-S method is not much bigger than the bound for the global error of the corresponding PTS method. Then, we compare empirically the global errors in the QR, QR-P, and QR-S methods to the global error in the PTS method.

Using (76–77), we write (52–53) in the following equivalent form:

$$[y_j] = T\hat{y}_{j-1} + (Q_j R_j)[r_{j-1}] + [z_j] \tag{82}$$

$$[r_j] = R_j[r_{j-1}] + Q_j^T([z_j] - s_j). \tag{83}$$

Obviously, this method does not require that we enclose the inverse of a nearly singular matrix. Therefore, unlike the parallelepiped method, the QR method cannot break down for that reason. (In practice, we have to enclose the inverse of a floating-point approximation to an orthogonal matrix.)

We can view $[r_j]$, $j = 1, 2, \dots$, as an enclosure of the global error at t_j that we propagate to the next step. Since

$$\mathrm{w}([r_j]) = |R_j|\mathrm{w}([r_{j-1}]) + |Q_j^T|\mathrm{w}([z_j]),$$

we can consider $|R_j|$ as the matrix for propagating the global error in the QR method. As we show in this section, the nature of the R_j matrices ensures better stability of this method, compared to the previous two methods.

From (82) and (83), we have for the width of $[y_j]$,

$$\begin{aligned} \mathrm{w}([y_j]) &= |Q_jR_j|\mathrm{w}([r_{j-1}]) + \mathrm{w}([z_j]) \\ &= |Q_jR_j||R_{j-1}|\cdots|R_2||Q_1^T|\mathrm{w}([z_1]) \\ &+ |Q_jR_j||R_{j-1}|\cdots|R_3||Q_2^T|\mathrm{w}([z_2]) \\ &\vdots \\ &+ |Q_jR_j||Q_{j-1}^T|\mathrm{w}([z_{j-1}]) \\ &+ \mathrm{w}([z_j]). \end{aligned} \tag{84}$$

To bound $\|\mathrm{w}([y_j])\|$, we shall bound the norm of each of the terms (after the second equal sign) in (84).

For any orthogonal $Q \in \mathbb{R}^{n\times n}$,

$$\big\|Q\big\| = \big\||Q|\big\| \le \sqrt{n}. \tag{85}$$

Using (42) and (85) and assuming again that $\|\mathrm{w}([z_{i-1}])\| \le \beta$ for all i, we bound the norm of

$$|Q_jR_j||R_{j-1}|\cdots|R_i||Q_{i-1}^T|\mathrm{w}([z_{i-1}]) \qquad (i = 2, 3, \ldots, j)$$

in (84) as

$$\big\||Q_jR_j||R_{j-1}|\cdots|R_i||Q_{i-1}^T|\mathrm{w}([z_{i-1}])\big\| \le \beta\, n \,\big\||R_j||R_{j-1}|\cdots|R_i|\big\|. \tag{86}$$

Below, we derive a bound for $\big\||R_j||R_{j-1}|\cdots|R_i|\big\|$.

We show first that

$$\lim_{j\to\infty}|R_j| = R \tag{87}$$

for some upper-triangular matrix R with eigenvalues $\{\,|\gamma_i| \,:\, i = 1, 2, \ldots, n\,\}$, where $\{\,\gamma_i \,:\, i = 1, 2, \ldots, n\,\}$ are the eigenvalues of T. Thus,

$$\lim_{j\to\infty}\rho(|R_j|) = \rho(R) = \rho(T). \tag{88}$$

From Theorem 3 in [6], under the QR iteration (80–81), the elements below the principal diagonal of S_j tend to zero, the moduli of those above the diagonal tend to fixed values, and the elements on the principal diagonal tend to the eigenvalues of T. Therefore,

$$\lim_{j\to\infty}|S_j| = R,$$

where R is upper-triangular with eigenvalues $\{\, |\gamma_i| \,:\, i = 1, 2, \ldots, n \,\}$. Since $S_{j-1} = \widehat{Q}_j R_j$ approaches an upper-triangular form, $\widehat{Q}_j$ approaches a diagonal form, where the magnitude of each diagonal entry tends to one. Hence,

$$\lim_{j\to\infty} |R_j| = \lim_{j\to\infty} |S_j \widehat{Q}_j^T| = \lim_{j\to\infty} |S_j| = R. \tag{89}$$

Let $E_l = |R_l| - R \in \mathbb{R}^{n\times n}$, whence

$$|R_l| = R + E_l. \tag{90}$$

From (89) and (90), for any $\epsilon > 0$, there exists $d \in \mathbb{N}$, $d > 0$, and $E \in \mathbb{R}^{n\times n}$ with $\|E\| \le \epsilon$, such that

$$|E_l| \le E, \qquad \text{for } \ l \ge d. \tag{91}$$

Let $i \ge d$ in (86). By (90) and (91),

$$\begin{aligned} \big\| |R_j||R_{j-1}| \cdots |R_i| \big\| &= \big\| |R + E_j||R + E_{j-1}| \cdots |R + E_i| \big\| \\ &\le \big\| (R + E)(R + E) \cdots (R + E) \big\| \\ &= \big\| (R + E)^{j-i+1} \big\|. \end{aligned} \tag{92}$$

Since R is nonsingular with distinct eigenvalues, there exists a matrix V_R that diagonalizes R. Then, from (92), Lemma 1, and (88),

$$\begin{aligned} \big\| |R_j||R_{j-1}| \cdots |R_i| \big\| &\le \operatorname{cond}(V_R)\big(\rho(R) + \operatorname{cond}(V_R)\,\epsilon\big)^{j-i+1} \\ &= \operatorname{cond}(V_R)\big(\rho(T) + \operatorname{cond}(V_R)\,\epsilon\big)^{j-i+1}. \end{aligned} \tag{93}$$

Substituting (93) into (86), we obtain for $i \ge d$ that

$$\begin{aligned} &\big\| |Q_j R_j||R_{j-1}| \cdots |R_i||Q_{i-1}^T| \mathrm{w}([z_{i-1}]) \big\| \\ &\qquad \le \beta\, n \operatorname{cond}(V_R)\big(\rho(T) + \operatorname{cond}(V_R)\,\epsilon\big)^{j-i+1}. \end{aligned} \tag{94}$$

Let $r \in \mathbb{R}$ be such that

$$\big\| |R_l| \big\| \le r \quad \text{for } \ l = 2, 3, \ldots, (d-1), \quad d \ge 3. \tag{95}$$

Such an r always exists. For example, we can chose $r = n\big\| |T| \big\| = n\big\|T\big\|$, because

$$\big\| |R_l| \big\| = \big\| |Q_l^T T Q_{l-1}| \big\| \le n \big\| |T| \big\| = n \big\| T \big\|.$$

If $i < d$, then from (93) with $i = d$ and (95), we obtain

$$\begin{aligned} &\big\| |Q_j R_j||R_{j-1}| \cdots |R_i||Q_{i-1}^T| \mathrm{w}([z_{i-1}]) \big\| \\ &\qquad \le \beta\, n \operatorname{cond}(V_R)\big(\rho(T) + \operatorname{cond}(V_R)\,\epsilon\big)^{j-d+1}\, r^{d-i}. \end{aligned} \tag{96}$$

Taking norms on both sides of (84) and using (94) and (96), we finally derive

$$\begin{aligned}\|\mathrm{w}([y_j])\| \le \beta\, n \operatorname{cond}(V_R) \Bigg(\big(\rho(T) + \operatorname{cond}(V_R)\,\epsilon\big)^{j-d+1} \sum_{i=1}^{d-2} r^i \\ + \sum_{i=1}^{j-d+1} \big(\rho(T) + \operatorname{cond}(V_R)\,\epsilon\big)^i \Bigg) + \beta,\end{aligned} \tag{97}$$

which gives an upper bound for the global error of the QR method at t_j. Recall that we used the infinity norm to derive (97). However, by the equivalence of vector norms (21), we have essentially the same result for any other norm.

From (43),

$$\begin{aligned}\|\delta_j\| &\le \delta \left(\operatorname{cond}(V_T) \sum_{i=1}^{j-1} \big(\rho(T)\big)^i + 1 \right) \\ &= \delta \operatorname{cond}(V_T) \left(\big(\rho(T)\big)^{j-d+1} \sum_{i=1}^{d-2} \big(\rho(T)\big)^i + \sum_{i=1}^{j-d+1} \big(\rho(T)\big)^i \right) + \delta.\end{aligned} \tag{98}$$

Let us compare the bounds in (97) and (98).

- We can assume that $\epsilon \ll 1$. Then,

$$\begin{aligned}\big(\rho(T) + \operatorname{cond}(V_R)\,\epsilon\big)^{j-d+1} \sum_{i=1}^{d-2} r^i &\approx \big(\rho(T)\big)^{j-d+1} \sum_{i=1}^{d-2} r^i \\ &= \frac{\sum_{i=1}^{d-2} r^i}{\sum_{i=1}^{d-2} \big(\rho(T)\big)^i} \left(\big(\rho(T)\big)^{j-d+1} \sum_{i=1}^{d-2} \big(\rho(T)\big)^i \right).\end{aligned}$$

  That is, the first term in (97) is approximately a constant times the first term in (98).
- The second sums in (97) and (98) are almost the same. That is,

$$\sum_{i=1}^{j-d+1} \big(\rho(T) + \operatorname{cond}(V_R)\,\epsilon\big)^i \approx \sum_{i=1}^{j-d+1} \big(\rho(T)\big)^i.$$

- Obviously, $\beta\, n \operatorname{cond}(V_R)$ from (97) is equal to some constant times $\delta \operatorname{cond}(V_T)$ from (98).

Thus, the bound for the global error of the QR method is not much bigger than the bound for the global error of the corresponding PTS method.

A Simplified QR Method Let Q be an $n \times n$ orthogonal matrix such that

$$S = Q^T T Q \tag{99}$$

is upper triangular. As we have assumed throughout this subsection that T has eigenvalues of different magnitudes, such a Q always exists, since (99) is the real Schur decomposition of T.

Consider now the QR method, but instead of $A_0 = Q_0 = I$, we set $A_0 = Q_0 = Q$. Then, we can choose

$$A_j = Q_j = Q, \quad \text{for } j \geq 1$$

in (52–53). With this choice for the transformation matrices, (52–53) becomes

$$[y_j] = T\hat{y}_{j-1} + (QS)[r_{j-1}] + [z_j] \quad \text{and} \tag{100}$$

$$[r_j] = S[r_{j-1}] + Q^T([z_j] - s_j), \tag{101}$$

for $j \geq 1$. We refer to (100–101) as the *QR-S method.*

From (100–101),

$$\begin{aligned}
\mathrm{w}([y_j]) &= |QS|\mathrm{w}([r_{j-1}]) + \mathrm{w}([z_j]) \\
&= |QS| \sum_{i=1}^{j-1} |S|^{j-i-1}\, |Q^T|\, \mathrm{w}([z_i]) + \mathrm{w}([z_j]) \\
&\leq |Q| \sum_{i=1}^{j-1} |S|^{j-i}\, |Q^T|\, \mathrm{w}([z_i]) + \mathrm{w}([z_j]).
\end{aligned} \tag{102}$$

Taking norms on both sides of (102) and employing (42) and (85), we derive that

$$\begin{aligned}
\|\mathrm{w}([y_j])\| &\leq \||Q|\| \sum_{i=1}^{j-1} \||S|^{j-i}\|\, \||Q^T|\|\beta + \beta \\
&\leq \beta \left(n \sum_{i=1}^{j-1} \||S|^i\| + 1 \right).
\end{aligned} \tag{103}$$

Since T has eigenvalues of distinct magnitudes, and S is upper triangular with the same eigenvalues as T, $|S|$ has distinct eigenvalues. Therefore, it can be diagonalized. Let $|S| = V_{|S|}^{-1}\, |D_S|\, V_{|S|}$, where D_S is a diagonal matrix with the eigenvalues of T on its diagonal. Then,

$$\||S|^i\| = \|V_{|S|}^{-1}\, |D_S|^i\, V_{|S|}\| \leq \mathrm{cond}(V_{|S|})\big(\rho(T)\big)^i. \tag{104}$$

Substituting (104) into (103), we derive that

$$\|\mathrm{w}([y_j])\| \leq \beta \left(n\, \mathrm{cond}(V_{|S|}) \sum_{i=1}^{j-1} \big(\rho(T)\big)^i + 1 \right). \tag{105}$$

Recall that for the PTS method,

$$\|\delta_j\| \leq \delta \left(\operatorname{cond}(V_T) \sum_{i=1}^{j-1} \big(\rho(T)\big)^i + 1 \right). \tag{106}$$

Comparing (105) and (106), we make the following conclusions.

- The global error in the QR-S method cannot be much bigger than $\approx cn$ times the upper bound for the global error of the PTS method, where c is some constant.
- If T is a normal matrix, then S is diagonal. In this case, $V_{|S|} = I$, and V_T is orthogonal. Therefore,

$$\|\mathrm{w}([y_j])\| \leq \beta \left(n \sum_{i=1}^{j-1} \big(\rho(T)\big)^i + 1 \right) \quad \text{and}$$
$$\|\delta_j\| \leq \delta \left(n \sum_{i=1}^{j-1} \big(\rho(T)\big)^i + 1 \right).$$

  That is, the bounds are almost the same.

Numerical Experiments We use the numerical experiments described below to compare

- the error propagation in the QR, QR-S, and QR-P methods to the error propagation in the PTS method, and
- the global errors in these methods,

assuming T is nonsingular with real eigenvalues of distinct magnitudes.

We consider the cases in which B is

1. symmetric,
2. non-normal, and
3. highly non-normal.

For each of these cases, we consider the subcases in which B has

1. negative eigenvalues,
2. positive eigenvalues, and
3. positive and negative eigenvalues.

In each of the examples below, we construct a 5×5 matrix B as follows. We compute an orthogonal matrix Q from the QR-factorization of a 5×5 Hilbert matrix, choose an upper-triangular or diagonal matrix R, as described in each example, and set $B = Q^T RQ$. Then we choose h, which can be different in the various examples, set $k = 30$, and compute $T = T_{29}(hB)$. For each problem, we use an initial condition

$$y(0) = (1, 1, 1, 1, 1)^T$$

and choose an appropriate t_m.

To simplify our study of the error propagation in the interval methods, we investigate only the propagation of the first error in these methods. Consider the QR method and the PTS method. Since $\mathrm{w}([z_1])$ in the QR method propagates to

$$|Q_j R_j||R_{j-1}| \cdots |R_2||Q_1^T|\mathrm{w}([z_1])$$

on the jth step (see (84)), and $\check{z}_1$ in the PTS method propagates to $T^{j-1}\check{z}_1$ on the jth step (see (38)), the ratio

$$\kappa_1(j) = \frac{\left\| |Q_j R_j||R_{j-1}| \cdots |R_2||Q_1^T| \right\|}{\left\| T^{j-1} \right\|} \tag{107}$$

($j \geq 2$) is a reasonable measure of the overestimation in propagating the errors in the QR method compared to the error propagation in the PTS method.

Similarly, we write

$$\kappa_2(j) = \frac{\left\| \left|Q_j^{(s)} R_j^{(s)}\right| \left|R_{j-1}^{(s)}\right| \cdots \left|R_2^{(s)}\right| \left|Q_1^{(s)T}\right| \right\|}{\left\| T^{j-1} \right\|} \tag{108}$$

for the QR-S method, where

$$Q_1^{(s)} = Q \quad \text{from a Schur form } S = Q^T T Q \quad \text{and} \tag{109}$$

$$T Q_{j-1}^{(s)} = Q_j^{(s)} R_j^{(s)}, \quad \text{for } j \geq 2. \tag{110}$$

Note that, if the eigenvalues of T are not in decreasing order of magnitudes on the diagonal of S, then this iteration will normally arrange them in such an order.

For the QR-P method, we write

$$\kappa_3(j) = \frac{\left\| \left|Q_j^{(p)} R_j^{(p)} {P^{(p)}}_j^T\right| \left|R_{j-1}^{(p)} {P^{(p)}}_{j-1}^T\right| \cdots \left|R_2^{(p)} {P^{(p)}}_2^T\right| \left|Q_1^{(p)T}\right| \right\|}{\left\| T^{j-1} \right\|}, \tag{111}$$

where $Q_l^{(p)}$ and $R_l^{(p)}$ are determined from the iteration

$$T = Q_1^{(p)} R_1^{(p)}, \tag{112}$$

$$T Q_{j-1}^{(p)} P_j^{(p)} = Q_j^{(p)} R_j^{(p)}, \quad \text{for } j \geq 2. \tag{113}$$

We compute $P_j^{(p)}$ as described in subsection 5.2, except that we assume that we have introduced only one local error $\mathrm{w}([z_1])$ with all components of the same size, and all subsequent local errors are zero in this method.

In Figs. 7–15, we plot $\kappa_1(j)$, $\kappa_2(j)$, and $\kappa_3(j)$ versus $t_j = jh$ as well as the base-10 logarithm of the global errors, $\log_{10} \|\mathrm{w}([y_j])\|$, of the QR, QR-S, and QR-P methods versus $t_j = jh$, and the base-10 logarithm of the global error of the PTS method, $\log_{10} \|\delta_j\|$, versus $t_j = jh$.

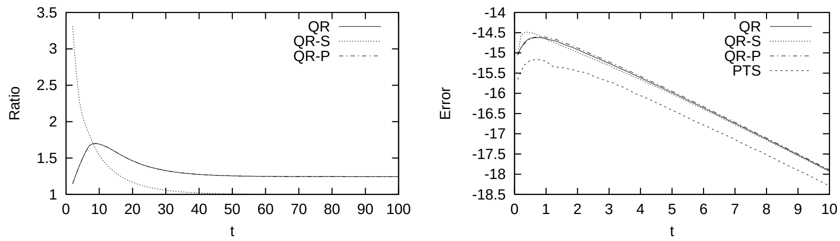
The computations of $\kappa_1(j)$, $\kappa_2(j)$, $\kappa_3(j)$, and $\|\delta_j\|$ are performed in Matlab, and the computations of $\|\mathrm{w}([y_j])\|$ are performed with the VNODE package [17]. The plots are produced with Gnuplot.

Example 5. B is symmetric with negative eigenvalues; see Fig. 7.

$$R = \mathrm{diag}(-1, -2, -3, -4, -5),$$

$h = 0.1$, and the eigenvalues of T are

$$\gamma(T) \approx \{\, 0.9048,\ 0.8187,\ 0.7408,\ 0.6703,\ 0.6065 \,\}.$$



(**a**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j

(**b**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

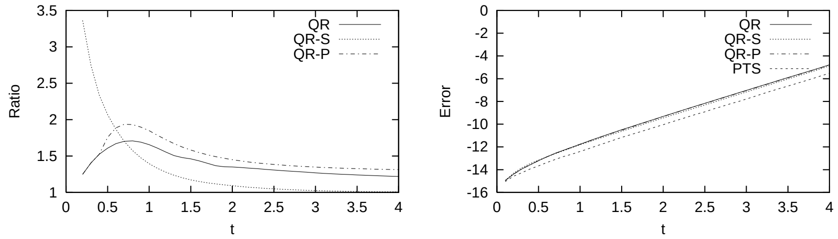
Fig. 7. B is symmetric with negative eigenvalues

Example 6. B is symmetric with positive eigenvalues; see Fig. 8.

$$R = \mathrm{diag}(1, 2, 3, 4, 5),$$

$h = 0.1$, and

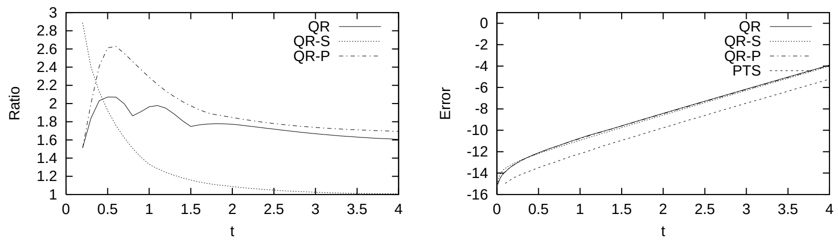
$$\gamma(T) \approx \{\, 1.1052,\ 1.2214,\ 1.3499,\ 1.4918,\ 1.6487 \,\}.$$



(**a**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j

(**b**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

Fig. 8. B is symmetric with positive eigenvalues



(**a**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j

(**b**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

Fig. 9. B is symmetric with positive and negative eigenvalues

Example 7. B is symmetric with positive and negative eigenvalues; see Fig. 9.

$$R = \mathrm{diag}(-1, -2, 3, 4, 5),$$

$h = 0.1$, and

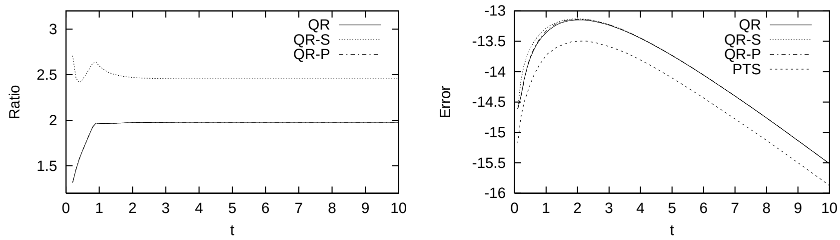
$$\gamma(T) \approx \{\, 0.8187,\ 0.9048,\ 1.3499,\ 1.6487,\ 1.4918 \,\}.$$

Example 8. B is non-normal with negative eigenvalues; see Fig. 10.

$$R = \begin{pmatrix} -1 & 6 & 0 & 0 & 0 \\ 0 & -3 & 6 & 0 & 0 \\ 0 & 0 & -5 & 6 & 0 \\ 0 & 0 & 0 & -7 & 6 \\ 0 & 0 & 0 & 0 & -9 \end{pmatrix},$$

$h = 0.1$, and

$$\gamma(T) \approx \{\, 0.9048,\, 0.7408,\, 0.6065,\, 0.4966,\, 0.4066 \,\}.$$



(**a**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j (**b**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

Fig. 10. B is non-normal with negative eigenvalues

Example 9. B is non-normal with positive eigenvalues; see Fig. 11.

$$R = \begin{pmatrix} 1 & 6 & 0 & 0 & 0 \\ 0 & 3 & 6 & 0 & 0 \\ 0 & 0 & 5 & 6 & 0 \\ 0 & 0 & 0 & 7 & 6 \\ 0 & 0 & 0 & 0 & 9 \end{pmatrix},$$

$h = 0.01$, and

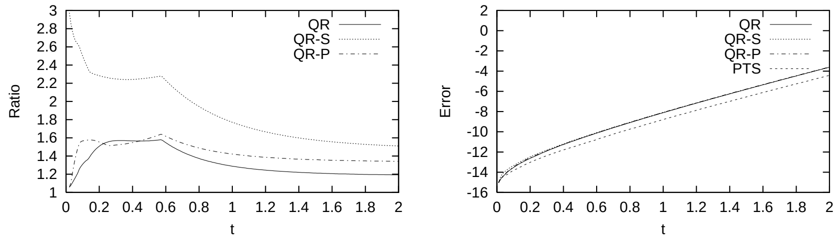
$$\gamma(T) \approx \{\, 1.0101,\, 1.0305\, 1.0513,\, 1.0725,\, 1.0942 \,\}.$$

Example 10. B is non-normal with positive and negative eigenvalues; see Fig. 12.

$$R = \begin{pmatrix} -1 & 6 & 0 & 0 & 0 \\ 0 & -3 & 6 & 0 & 0 \\ 0 & 0 & 5 & 6 & 0 \\ 0 & 0 & 0 & 7 & 6 \\ 0 & 0 & 0 & 0 & 9 \end{pmatrix},$$

$h = 0.01$, and

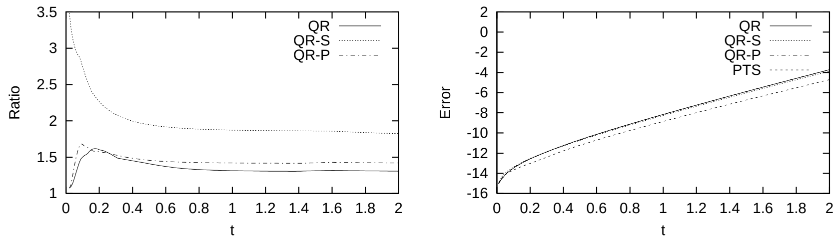
$$\gamma(T) \approx \{\, 0.9704,\, 0.9900,\, 1.0513,\; 1.0725,\; 1.0942 \,\}.$$



(a) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j

(b) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

Fig. 11. B is non-normal with positive eigenvalues



(a) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j

(b) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

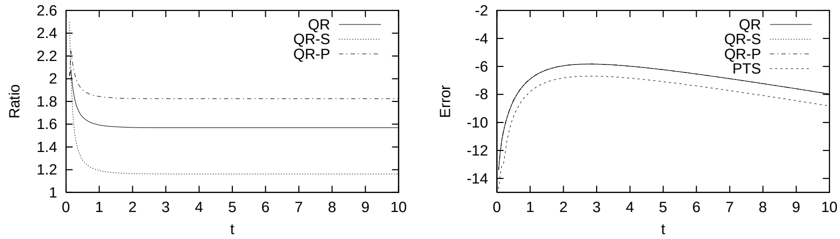
Fig. 12. B is non-normal with positive and negative eigenvalues

Example 11. B is highly non-normal with negative eigenvalues; see Fig. 13.

$$R = \begin{pmatrix} -1 & 50 & 0 & 0 & 0 \\ 0 & -3 & 50 & 0 & 0 \\ 0 & 0 & -5 & 50 & 0 \\ 0 & 0 & 0 & -7 & 50 \\ 0 & 0 & 0 & 0 & -9 \end{pmatrix},$$

$h = 0.05$, and

$$\gamma(T) \approx \{\, 0.9512,\, 0.8607,\, 0.7788,\, 0.7047,\, 0.6376 \,\}.$$



(**a**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j

(**b**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

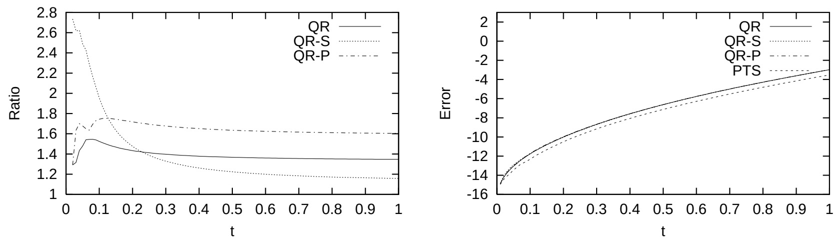
Fig. 13. B is highly non-normal with negative eigenvalues

Example 12. B is highly non-normal with positive eigenvalues; see Fig. 14.

$$R = \begin{pmatrix} 1 & 50 & 0 & 0 & 0 \\ 0 & 3 & 50 & 0 & 0 \\ 0 & 0 & 5 & 50 & 0 \\ 0 & 0 & 0 & 7 & 50 \\ 0 & 0 & 0 & 0 & 9 \end{pmatrix},$$

$h = 0.01$, and

$$\gamma(T) \approx \{\, 1.0101,\ 1.0305,\ 1.0513,\ 1.0725,\ 1.0942 \,\}.$$



(**a**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j

(**b**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

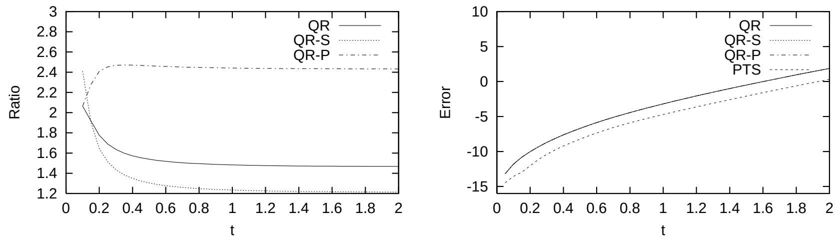
Fig. 14. B is highly non-normal with positive eigenvalues

Example 13. B is highly non-normal with positive and negative eigenvalues; see Fig. 15.

$$R = \begin{pmatrix} -1 & 50 & 0 & 0 & 0 \\ 0 & -3 & 50 & 0 & 0 \\ 0 & 0 & 5 & 50 & 0 \\ 0 & 0 & 0 & 7 & 50 \\ 0 & 0 & 0 & 0 & 9 \end{pmatrix},$$

$h = 0.05$, and

$$\gamma(T) \approx \{ 0.8607, 0.9512, 1.2840, 1.4191, 1.5683 \}.$$



(**a**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j

(**b**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

Fig. 15. B is highly non-normal with positive and negative eigenvalues

We have performed significantly more experiments than the ones reported in this subsection. Here, we summarize our observations.

As j becomes sufficiently large, the ratios $\kappa_i(j)$, $i = 1, 2, 3$, are usually not much bigger than one. Often, but not always, $\kappa_2(j)$ becomes smaller than $\kappa_1(j)$ and $\kappa_3(j)$, and $\kappa_3(j)$ generally becomes bigger than $\kappa_1(j)$.

This suggests that the propagated first error in the interval methods (QR, QR-S, and QR-P) should not be much larger than the propagated first error in the PTS method. This also suggests that the QR-S method should generally produce tighter bounds than the QR and QR-P methods, while the QR-P should generally produce wider bounds than the QR method.

However, from the (b) parts of Figs. 7–15, the global errors in the interval methods at each t_j are nearly the same and not much bigger than the corresponding global error of the PTS method. Moreover, no interval method produces a substantially smaller global error than the other two at any t_j.

8.2 Eigenvalues of Equal Magnitude

If T is nonsingular with p eigenvalues of equal magnitude, then the matrices S_j defined in (78) tend to a block upper-triangular form with a $p \times p$ block on its main diagonal [26, pp. 520–521]. The eigenvalues of this block tend to these p eigenvalues.

In this subsection, we consider the case that T is nonsingular with at least one complex conjugate pair of eigenvalues and at most two eigenvalues of the same magnitude. In this case, S_j tends to a block upper-triangular form with 1×1 and 2×2 blocks on its diagonal. The eigenvalues of these blocks tend to the eigenvalues of T. Note that S_j does not converge to a fixed matrix, but the eigenvalues of its diagonal blocks converge to the eigenvalues of T.

When $S_{j-1} = \widehat{Q}_j R_j$ approaches such a block upper-triangular form, the orthogonal matrices $\widehat{Q}_j$ approach a block-diagonal form with 1×1 and 2×2 blocks on its main diagonal, where the 2×2 blocks are at the same locations as the 2×2 blocks of S_{j-1}.

For simplicity, assume that each sub-diagonal entry ("diagonal" includes 2×2 blocks) of S_{j-1} and each off-diagonal entry of $\widehat{Q}_j$ is zero, while in practice, they converge to zero. Consider a 2×2 block of S_{j-1},

$$\begin{pmatrix} (S_{j-1})_{l,l} & (S_{j-1})_{l,l+1} \\ (S_{j-1})_{l+1,l} & (S_{j-1})_{l+1,l+1} \end{pmatrix},$$

and denote its complex eigenvalues by γ_l and $\gamma_{l+1} = \overline{\gamma}_l$, which are also eigenvalues of T. From Lemma 2, the components of R_j that correspond to these eigenvalues are

$$\begin{aligned} (R_j)_{l,l} &= s_1 \alpha_{l,j-1} \quad \text{and} \\ (R_j)_{l+1,l+1} &= s_2 |\gamma_l|^2 / \alpha_{l,j-1}, \end{aligned}$$

where $s_1, s_2 = \pm 1$ and $\alpha_{l,j-1} = \sqrt{(S_{j-1})^2_{l,l} + (S_{j-1})^2_{l+1,l}}$.

Hence, as $j \to \infty$,

$$\begin{aligned} |(R_j)_{l,l}| &= \alpha_{l,j-1} \quad \text{and} \\ |(R_j)_{l+1,l+1}| &= |\gamma_l|^2 / \alpha_{l,j-1} \end{aligned}$$

oscillate around $|\gamma_l|$.

If $\rho(T) = |\gamma_l|$, then

$$\rho(|R_j|) \geq \rho(T), \quad \text{as } j \to \infty.$$

In fact, if $\alpha_{l,j-1} \ll |\gamma_l|$ or $\alpha_{l,j-1} \gg |\gamma_l|$, then $\rho(|R_j|)$ may be much larger than $\rho(T)$. However, we have not observed this in practice. The examples reported below are typical of our much larger class of test problems in that $\rho(|R_j|)$ is normally either equal to or just a little larger than $\rho(T)$.

On the other hand, if T has a dominant real eigenvalue, then we should generally expect that

$$\rho(|R_j|) \to \rho(T), \quad \text{as } j \to \infty.$$

Since the matrices $|R_j|$ do not approach a fixed matrix, as in the case with eigenvalues of distinct magnitudes, deriving a bound for the global error in this case is more difficult. However, we show in four examples, in which T has some eigenvalues of equal magnitude, that the global errors in the QR, QR-P, and QR-S methods are not much bigger than the global error of the PTS method.

Numerical Experiments We investigate the behavior of the QR, QR-S, and QR-P methods assuming T has at least one complex conjugate pair of eigenvalues. We consider the following four cases:

1. B has a dominant complex conjugate pair;
2. B has a dominant real eigenvalue;
3. B is normal, but neither symmetric nor skew symmetric; and
4. B is skew symmetric.

For each of the four examples, we describe below how B and the corresponding T matrices are constructed. Moreover, for each example, we plot

- $\rho(|R_j|)$, $\rho(|R_j^{(s)}|)$, $\rho(|R_j^{(p)} P_j^T|)$, and $\rho(T)$ versus $t_j = jh$, where $R_j^{(s)}$, $R_j^{(p)}$, and P_j are computed as described in the numerical experiments part of subsection 8.1;
- the ratios $\kappa_1(j)$, $\kappa_2(j)$, and $\kappa_3(j)$ defined in (107), (108), and (111) versus t_j;
- the norms

$$\begin{aligned}
\eta_1(j) &= \big\| |Q_j R_j| |R_{j-1}| \cdots |R_2| |Q_1^T| \big\|, \\
\eta_2(j) &= \Big\| |Q_j^{(s)} R_j^{(s)}| |R_{j-1}^{(s)}| \cdots |R_2^{(s)}| |Q_1^{(s)T}| \Big\|, \\
\eta_3(j) &= \Big\| |Q_j^{(p)} R_j^{(p)} P_j^T| |R_{j-1}^{(p)} P_{j-1}^T| \cdots |R_2^{(p)} P_2^T| |Q_1^{(p)T}| \Big\|, \quad \text{and} \\
\eta_4(j) &= \big\| T^{j-1} \big\|,
\end{aligned}$$

  which correspond to the QR, QR-S, QR-P, and PTS methods, versus t_j; and
- the base-10 logarithm of the global errors in the QR, QR-S, QR-P, and PTS methods versus t_j. The initial condition vector in each example has all components equal to one.

The computations of the spectral radii, $\kappa_i(j)$, $i = 1, 2, 3$, $\eta_i(j)$, $i = 1, 2, 3, 4$, and $\|\delta_j\|$ are performed in Matlab, and the computations of $\|\mathrm{w}([y_j])\|$ are performed with the VNODE package [17]. The plots are produced with Gnuplot.

Example 14. B has a dominant complex conjugate pair.

Let

$$B = \begin{pmatrix} -0.5 & 0.5 & 0 \\ -2.0 & -0.5 & 1 \\ 0 & 1 & -2 \end{pmatrix}.$$

For $k = 30$ and $h = 0.1$,

$$\gamma(T) \approx \{\, 0.9674 \pm 0.0836i,\, 0.7858 \,\}.$$

Here, $|\gamma_1| = |\gamma_2| \approx 0.9710 > |\gamma_3| \approx 0.7858$. That is, the complex conjugate pair of eigenvalues dominates.

In Fig. 16(a), we plot $\rho(|R_j|)$, $\rho(|R_j^{(s)}|)$, $\rho(|R_j^{(p)} P_j^T|)$, and $\rho(T)$ versus t_j. The oscillations in $\rho(|R_j|)$ are caused by the changes in the diagonal elements $(R_j)_{1,1}$ and $(R_j)_{2,2}$, which correspond to the complex eigenvalues $\gamma_{1,2} \approx 0.9674 \pm 0.0836i$. Since the QR and QR-S method differ in the starting matrix for the simultaneous iteration (identity in the QR method, and Q from a Schur form $S = Q^T T Q$ in the QR-S method), the oscillations in $\rho(|R_j^{(s)}|)$ can be explained by the same arguments as the oscillations in $\rho(|R_j|)$. However, we do not propose an explanation of why $\rho(|R_j^{(p)} P_j^T|)$ behaves similarly to $\rho(|R_j^{(s)}|)$ and $\rho(|R_j|)$.

In Fig. 16(b), we plot the ratios $\kappa_1(j)$, $\kappa_2(j)$, and $\kappa_3(j)$. Although $\kappa_2(j)$ can become much larger than one, the first error in the QR-S method is still damped to zero. This can be seen from Fig. 16(c), where we plot $\eta_1(j)$, $\eta_2(j)$, $\eta_3(j)$, and $\eta_4(j)$ versus t_j. In this figure, the above norms approach zero, as j increases, suggesting that $\mathrm{w}([z_1])$ is damped to zero, as j increases. Note that this happens even though the corresponding spectral radii in the interval methods are not always smaller than one.

From Fig. 16(d), we see that at each t_j, the global errors in the interval methods are not much larger than the corresponding global error in the PTS method. In addition, all these errors decrease in a similar manner, as j becomes sufficiently large.

Example 15. B has a dominant real eigenvalue.

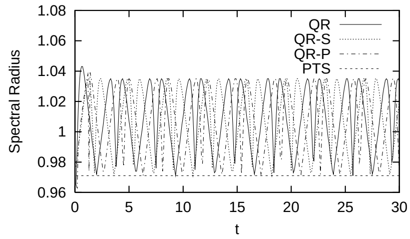
In this example, we choose

$$B = \begin{pmatrix} -2 & 2 & -3 \\ -1 & -3 & 1 \\ 1 & 1 & -1 \end{pmatrix},$$

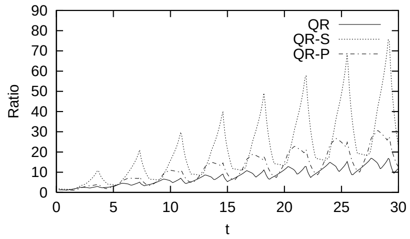
$k = 30$, and $h = 0.4$. The eigenvalues of $T = T_{29}(0.4B)$ are

$$\gamma(T) \approx \{\, 0.6703,\, 0.2629 \pm 0.2573i \,\}.$$

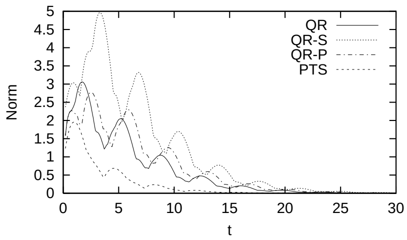
Here, $|\gamma_1| \approx 0.6703 > |\gamma_2| = |\gamma_3| \approx 0.3679$. That is, the real eigenvalue dominates.



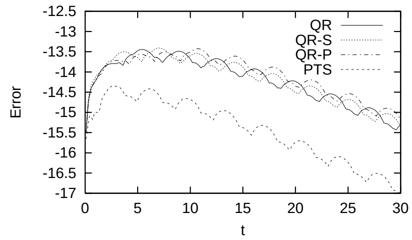
(**a**) $\rho(|R_j|)$, $\rho(|R_j^{(s)}|)$, $\rho(|R_j^{(p)} P_j^T|)$, and $\rho(T)$ vs. t_j



(**b**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j



(**c**) $\eta_i(j)$, $i = 1, 2, 3, 4$, vs. t_j

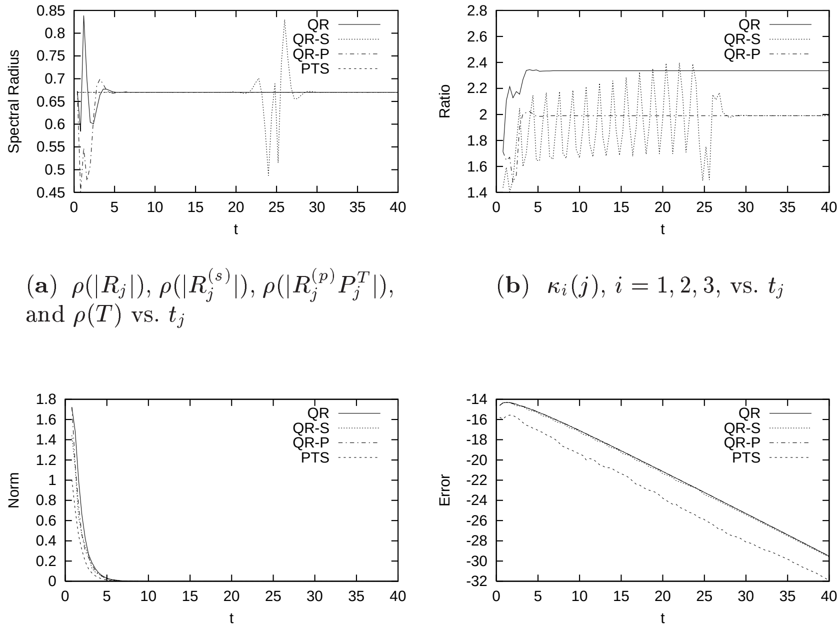


(**d**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

Fig. 16. B has a dominant complex conjugate pair

In Fig. 17(a), $\rho(|R_j|)$, $\rho(|R_j^{(s)}|)$, and $\rho(|R_j^{(p)} P_j^T|)$ settle to $\rho(T)$, but then $\rho(|R_j^{(s)}|)$ oscillates and settles again to $\rho(T)$. Thus, when the dominant eigenvalue is real, the methods behave like they do with real distinct eigenvalues. To support this point, compare Fig. 17(d) with the (b) parts of the figures in subsection 8.1. The reason for the "peculiar" oscillations in $\rho(|R_j^{(s)}|)$ is that the block upper-triangular matrix returned by the Matlab function `schur` (which we use to compute Q and S such that $S = Q^T T Q$ is block upper triangular) contains the dominant real eigenvalue in position $(3, 3)$ (that is, $S_{3,3} \approx 0.6703$), and the simultaneous iteration (110) moves this eigenvalue to position $(1, 1)$ in $R_j^{(s)}$, as j becomes sufficiently large.

Although the ratios $\kappa_1(j)$, $\kappa_2(j)$, and $\kappa_3(j)$ are bigger than one, the propagated error $\mathrm{w}([z_1])$ in the QR, QR-S, and QR-P is damped in these methods; see the plots in Figs. 17(c) and (d).



(**a**) $\rho(|R_j|)$, $\rho(|R_j^{(s)}|)$, $\rho(|R_j^{(p)}P_j^T|)$, and $\rho(T)$ vs. t_j

(**b**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j

(**c**) $\eta_i(j)$, $i = 1, 2, 3, 4$, vs. t_j

(**d**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

Fig. 17. B has a dominant real eigenvalue

Example 16. B is normal, but neither symmetric nor skew-symmetric.

We choose

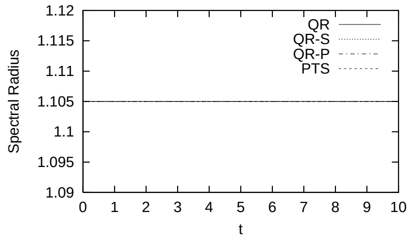
$$B = \begin{pmatrix} 1 & -3 & -4 & -5 & -6 \\ 3 & 1 & -5 & -6 & -7 \\ 4 & 5 & 1 & -7 & -8 \\ 5 & 6 & 7 & 1 & -9 \\ 6 & 7 & 8 & 9 & 1 \end{pmatrix},$$

$k = 30$, and $h = 0.1$. The eigenvalues of $T = T_{29}(0.1B)$ are

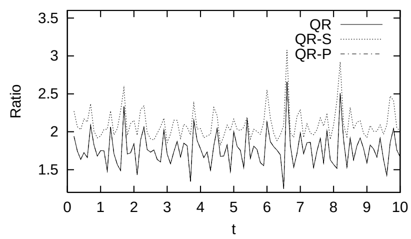
$$\gamma(T) \approx \{\, -0.3961 \pm 1.0317i,\ 1.0251 \pm 0.4129i,\ 1.1052 \,\},$$

all with a magnitude of ≈ 1.1052.

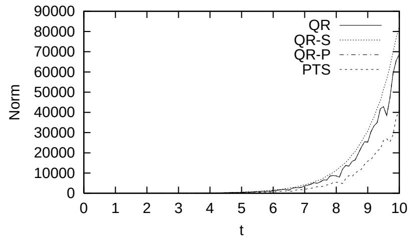
The plots are shown in Fig. 18. From Figs. 18(b) and 18(c), we expect the global errors in the interval methods to be a little bigger than in the PTS method. This is confirmed in Fig. 18(d), where we plot the base-10 logarithm of the global errors in these methods versus t_j.



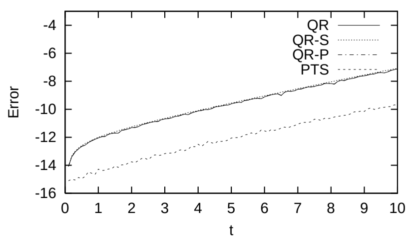
(**a**) $\rho(|R_j|)$, $\rho(|R_j^{(s)}|)$, $\rho(|R_j^{(p)}P_j^T|)$, and $\rho(T)$ vs. t_j



(**b**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j



(**c**) $\eta_i(j)$, $i = 1, 2, 3, 4$, vs. t_j



(**d**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

Fig. 18. B is normal, but neither symmetric nor skew symmetric

Example 17. B is skew symmetric.

In this example,

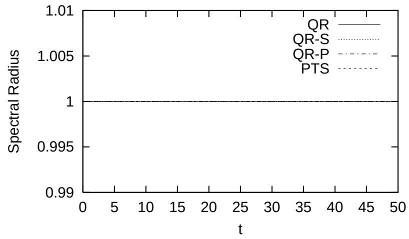
$$B = \begin{pmatrix} 0 & -3 & -4 & -5 & -6 \\ 3 & 0 & -5 & -6 & -7 \\ 4 & 5 & 0 & -7 & -8 \\ 5 & 6 & 7 & 0 & -9 \\ 6 & 7 & 8 & 9 & 0 \end{pmatrix},$$

$k = 30$, and $h = 0.1$. The eigenvalues of $T = T_{29}(0.1B)$ are

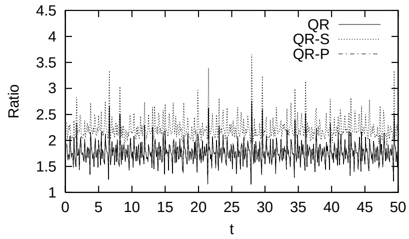
$$\gamma(T) = \{ -0.3584 \pm 0.9336i,\ 0.9276 \pm 0.3736i,\ 1.0000 \},$$

all with a magnitude of ≈ 1; see Fig. 19.

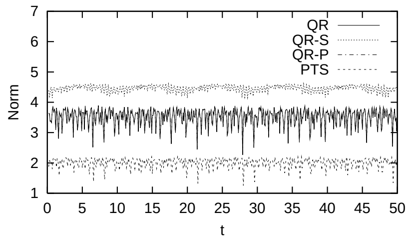
For this problem, the global error of the PTS method increases slowly, as j increases, as do the global errors in the interval methods. Note that the errors in the interval methods do not become much bigger than the global error of the PTS method; see Fig. 19(d).



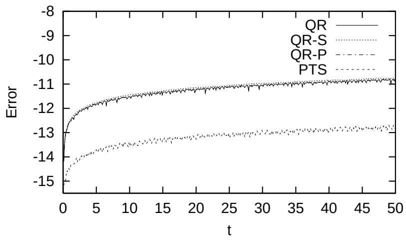
(**a**) $\rho(|R_j|)$, $\rho(|R_j^{(s)}|)$, $\rho(|R_j^{(p)}P_j^T|)$, and $\rho(T)$ vs. t_j



(**b**) $\kappa_i(j)$, $i = 1, 2, 3$, vs. t_j



(**c**) $\eta_i(j)$, $i = 1, 2, 3, 4$, vs. t_j



(**d**) $\log_{10} \|\mathrm{w}([y_j])\|$ and $\log_{10} \|\delta_j\|$ vs. t_j

Fig. 19. B is skew symmetric

The numerical results in this subsection suggest that, if T is nonsingular and has some eigenvalues of the same magnitude, then the global errors in the QR, QR-S, and QR-P methods (at each t_j) are nearly the same and not much bigger than the corresponding global error of the PTS method. In addition, the global errors in the interval methods decrease or increase similarly to the global error of the PTS method.

If T has a dominant real eigenvalue, then the spectral radii in the interval methods usually settle to $\rho(T)$, as j becomes sufficiently large. Thus, this case is similar to the case that T is nonsingular with eigenvalues of distinct magnitudes.

9 Conclusions

We used eigenvalue techniques to develop a new perspective on the wrapping effect in interval methods for IVPs for ODEs. In our approach, we view the wrapping effect as a source of instability in interval methods and the problem of reducing the wrapping effect as one of finding a more stable scheme for advancing the solution. Eigenvalue techniques have proven useful in studying the stability of standard (point) methods; it seems natural to employ such techniques in the study of interval methods.

Our approach complements the geometric arguments used to date to study the wrapping effect in interval methods for IVPs for ODEs, providing a much more quantitative approach to the problem. If only geometric arguments are used in studying the wrapping effect, it is often difficult to draw sound conclusions about how an interval method will perform.

For example, the earlier geometric explanations of why Lohner's QR-factorization method works well [15,17] do give us an intuitive understanding of the scheme, but we believe that they do not provide as good an explanation as that developed in Sect. 8. To be more specific, we showed that, for linear constant-coefficient problems, the global error of an ITS method using Lohner's QR-factorization scheme is generally not much larger than the global error of the corresponding point method. It is hard to see how one could draw this conclusion from a geometric approach alone.

It does not appear that our stability analysis of linear constant-coefficient problems can be easily extended to more general IVPs. However, given an anti-wrapping scheme, we can study its stability for this restricted class of problems. If such a scheme does not perform well for these simple problems, it almost certainly will not work well in general. Conversely, if it performs well for linear constant-coefficient problems, we believe it is quite likely to work well in general.

A Lemmas

Lemma 1. *Suppose that $A \in \mathbb{R}^{n\times n}$ is diagonalizable. That is, there exists a nonsingular matrix V and a diagonal matrix D such that $D = VAV^{-1}$. Let $E \in \mathbb{R}^{n\times n}$ be any matrix such that $\|E\| \le \beta$. Then, for any integer $p \ge 0$,*

$$\|(A+E)^p\| \le \operatorname{cond}(V)\big(\rho(A) + \beta \operatorname{cond}(V)\big)^p.$$

Proof.

$$\begin{aligned}
\big\|(A+E)^p\big\| &= \big\|V^{-1}V(A+E)^pV^{-1}V\big\| \\
&\le \big\|V^{-1}\big\|\big\|V(A+E)^pV^{-1}\big\|\big\|V\big\| \\
&= \operatorname{cond}(V)\big\|\big(V(A+E)V^{-1}\big)^p\big\| \\
&= \operatorname{cond}(V)\big\|(D+VEV^{-1})^p\big\| \\
&\le \operatorname{cond}(V)\big\|D+VEV^{-1}\big\|^p \\
&\le \operatorname{cond}(V)\left(\big\|D\big\| + \big\|VEV^{-1}\big\|\right)^p \\
&\le \operatorname{cond}(V)\big(\rho(A) + \beta \operatorname{cond}(V)\big)^p,
\end{aligned}$$

where we have used in the last line $\|D\| = \rho(D) = \rho(A)$.

Lemma 2. *Let*

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

have eigenvalues $\lambda_{1,2} = a \pm ib$ *with* $b \neq 0$ *and let* $C = QR$ *be its QR decomposition. Then*

$$R = \begin{pmatrix} s_1\alpha & s_1(c_{11}c_{12} + c_{21}c_{22})/\alpha \\ 0 & s_2|\lambda_1|^2/\alpha \end{pmatrix},$$

where

$$\alpha = \sqrt{c_{11}^2 + c_{21}^2} \neq 0,$$

s_1 *can be either* $+1$ *or* -1, *and* s_2 *can be either* $+1$ *or* -1.

Proof. First note that, since $C \in R^{2\times 2}$ and $\lambda_{1,2} \notin \mathbb{R}$, $c_{21} \neq 0$. Therefore, $\alpha = \sqrt{c_{11}^2 + c_{21}^2} \neq 0$.

It is straightforward to show that any orthogonal $Q \in \mathbb{R}^{2\times 2}$ satisfying $C = QR$, or equivalently $Q^T C = R$, must be of the form

$$Q^T = \frac{1}{\sqrt{c_{11}^2 + c_{21}^2}} \begin{pmatrix} s_1 c_{11} & s_1 c_{21} \\ -s_2 c_{21} & s_2 c_{11} \end{pmatrix},$$

where s_1 can be either $+1$ or -1, and s_2 can be either $+1$ or -1. Hence,

$$R = Q^T C = \frac{1}{\sqrt{c_{11}^2 + c_{21}^2}} \begin{pmatrix} s_1(c_{11}^2 + c_{21}^2) & s_1(c_{11}c_{12} + c_{21}c_{22}) \\ 0 & s_2(c_{11}c_{22} - c_{12}c_{21}) \end{pmatrix}.$$

The eigenvalues of C are the roots of the quadratic equation

$$\lambda^2 - (c_{11} + c_{22})\lambda + c_{11}c_{22} - c_{12}c_{21} = 0.$$

Hence,

$$|\lambda_1|^2 = \lambda_1\lambda_2 = c_{11}c_{22} - c_{12}c_{21}.$$

Therefore,

$$R = \frac{1}{\sqrt{c_{11}^2 + c_{21}^2}} \begin{pmatrix} s_1(c_{11}^2 + c_{21}^2) & s_1(c_{11}c_{12} + c_{21}c_{22}) \\ 0 & s_2|\lambda_1|^2 \end{pmatrix}$$

$$= \begin{pmatrix} s_1\alpha & s_1(c_{11}c_{12} + c_{21}c_{22})/\alpha \\ 0 & s_2|\lambda_1|^2/\alpha \end{pmatrix}.$$

References

1. G. Alefeld and J. Herzberger. *Introduction to Interval Computations*. Academic Press, New York, 1983.
2. U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM, Philadelphia, 1998.
3. C. Barbăroşie. Reducing the wrapping effect. *Computing*, 54(4):347–357, 1995.
4. M. Berz and K. Makino. Verified integration of ODEs and flows using differential algebraic methods on high-order Taylor models. *Reliable Computing*, 4:361–369, 1998.
5. P. Eijgenraam. *The Solution of Initial Value Problems Using Interval Arithmetic*. Mathematical Centre Tracts No. 144. Stichting Mathematisch Centrum, Amsterdam, 1981.
6. J. G. F. Francis. The QR transformation: A unitary analogue to the LR transformation—part 1. *The Computer Journal*, 4:265–271, 1961/1962.
7. T. Gambill and R. Skeel. Logarithmic reduction of the wrapping effect with application to ordinary differential equations. *SIAM J. Numer. Anal.*, 25(1):153–162, 1988.
8. E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer-Verlag, 2nd revised edition, 1991.
9. L. W. Jackson. A comparison of ellipsoidal and interval arithmetic error bounds. In *Studies in Numerical Analysis, vol. 2, Proc. Fall Meeting of the Society for Industrial and Applied Mathematics*, Philadelphia, 1968. SIAM.
10. L. W. Jackson. Automatic error analysis for the solution of ordinary differential equations. Technical Report 28, Dept. of Computer Science, University of Toronto, 1971.
11. L. W. Jackson. Interval arithmetic error–bounding algorithms. *SIAM J. Numer. Anal.*, 12(2):223–238, 1975.
12. W. Kahan. A computable error bound for systems of ordinary differential equations. *Abstract in SIAM Review*, 8:568–569, 1966.
13. F. Krückeberg. Ordinary differential equations. In E. Hansen, editor, *Topics in Interval Analysis*, pages 91–97. Clarendon Press, Oxford, 1969.
14. W. Kühn. Rigorously computed orbits of dynamical systems without the wrapping effect. *Computing*, 61(1):47–67, 1998.
15. R. J. Lohner. *Einschließung der Lösung gewöhnlicher Anfangs– und Randwertaufgaben und Anwendungen*. PhD thesis, Universität Karlsruhe, 1988.
16. R. E. Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, N.J., 1966.
17. N. S. Nedialkov. *Computing Rigorous Bounds on the Solution of an Initial Value Problem for an Ordinary Differential Equation*. PhD thesis, Department of Computer Science, University of Toronto, Toronto, Canada, M5S 3G4, February 1999.
18. N. S. Nedialkov and K. R. Jackson. An interval Hermite-Obreschkoff method for computing rigorous bounds on the solution of an initial value problem for an ordinary differential equation. *Reliable Computing*, 5(3):289–310, 1999. Also in T. Csendes, editor, *Developments in Reliable Computing*, pp. 289–310, Kluwer, Dordrecht, Netherlands, 1999.
19. N. S. Nedialkov, K. R. Jackson, and G. F. Corliss. Validated solutions of initial value problems for ordinary differential equations. *Applied Mathematics and Computation*, 105(1):21–68, 1999.

20. N. S. Nedialkov, K. R. Jackson, and J. D. Pryce. An effective high-order interval method for validating existence and uniqueness of the solution of an IVP for an ODE, 2000. Accepted for publication in Reliable Computing, 17 pages.
21. A. Neumaier. Global, rigorous and realistic bounds for the solution of dissipative differential equations. Part I : Theory. *Computing*, 52(4):315–336, 1994.
22. K. Nickel. How to fight the wrapping effect. In K. Nickel, editor, *Interval Analysis 1985*, Lecture Notes in Computer Science No. 212, pages 121–132. Springer, Berlin, 1985.
23. R. Rihm. Interval methods for initial value problems in ODEs. In J. Herzberger, editor, *Topics in Validated Computations: Proceedings of the IMACS-GAMM International Workshop on Validated Computations, University of Oldenburg*, Elsevier Studies in Computational Mathematics, pages 173–207. Elsevier, Amsterdam, New York, 1994.
24. N. Stewart. A heuristic to reduce the wrapping effect in the numerical solution of $x' = f(t, x)$. *BIT*, 11:328–337, 1971.
25. D. S. Watkins. Understanding the QR algorithm. *SIAM Review*, 24(4):427–440, October 1982.
26. J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Oxford Science Publications, Oxford, England, 1965.