SPECTRAL ANALYSIS OF NONSYMMETRIC QUASI-TOEPLITZ MATRICES WITH APPLICATIONS TO PRECONDITIONED MULTISTEP FORMULAS

DANIELE BERTACCINI† AND FABIO DI BENEDETTO‡

Abstract. The eigenvalue spectrum of the nonsymmetric preconditioned matrices arising in time-dependent partial differential equations is analyzed and discussed. The matrices generated by the underlying numerical integrators are small rank perturbations of block Toeplitz matrices; preconditioners based on small rank approximations for those are considered. The eigenvalue distribution of the preconditioned matrix influences often crucially the convergence of Krylov iterative accelerators. Due to the lack of symmetry, the classical approach based on generating functions gives very little insight here. Therefore, to characterize the eigenvalues, a difference equation approach exploiting the band Toeplitz and circulant patterns generalizing well known Trench’s results is proposed.

Key words. circulant preconditioners, nonsymmetric Toeplitz matrices, eigenvalues, difference equations, linear systems of time-step integrators, linear multistep formulas in boundary value form, boundary value problems.

AMS subject classifications. 15A18, 65F10, 65F15, 65N22, 65Q05

1. Introduction. In this paper we investigate the spectrum of the eigenvalues and other properties for some classes of matrices. In particular, we focus on small rank perturbations of block nonsymmetric Toeplitz matrices preconditioned by circulant approximations introduced in [3, 4].

An \( n \times n \) matrix \( A_n = (a_{j,k}) \) is said to be Toeplitz if \( a_{j,k} = a_{j-k}, j, k = 1, ..., n \), i.e. \( A_n \) is constant along its diagonals, quasi Toeplitz if it is a small rank perturbation of a Toeplitz matrix. An \( n \times n \) matrix \( \tilde{A}_n \) is said to be circulant if it is Toeplitz and its diagonals satisfy \( \tilde{a}_{n-j} = \tilde{a}_{-j}, j = 1, ..., n - 1 \). The circulant matrices \( \tilde{A}_n \) are diagonalized by the Fourier matrix \( F = (F_{j,k}), F_{j,k} = e^{2\pi i j k / n} / \sqrt{n} \), \( j, k = 0, ..., n - 1 \), \( i \) is the imaginary unit, see [16]. From the previous arguments, it follows that such matrices are easily and efficiently invertible using the Fast Fourier Transform (FFT), see, e.g. [13].

Perturbations of block nonsymmetric Toeplitz matrices arise in the numerical approximation of time-dependent partial differential equations (PDEs) by means of generalizations of implicit linear multistep formulas used in boundary value form; see, e.g., [17, 1, 11]. However, the techniques considered here could be adapted to other discretization schemes based on finite differences for PDEs as well.

Other circulant-like matrices used in the PDE context can be found in [6].

As explained in §2.1, the matrices of the underlying linear systems can be written as follows:

\[
M = A \otimes I - h B \otimes J,
\]

(1.1)

where \( A \) and \( B \) are \( n \times n \) small rank perturbations of band Toeplitz matrices whose entries are given by the coefficients of the scheme involved, \( I \) is the identity and \( J \) is a \( m \times m \) matrix which can be large and sparse. More precisely, \( J \) is the Jacobian

---

* This work is partially supported by MIUR, grant numbers 2002014121 and 2004015437, and by grant no. 8111317 from Università di Roma “La Sapienza”.

† Università di Roma “La Sapienza”, Dipartimento di Matematica, P.le A. Moro 2, 00185 Roma, Italy. E-mail: bertaccini@mat.uniroma1.it.

‡ Università di Genova, Dipartimento di Matematica, via Dodecaneso 35, 16146, Genova, Italy. E-mail: dibenede@dima.unige.it.
matrix of a system of ordinary or partial differential equations discretized in space by finite differences, see [4] for details.

Unfortunately, when \( m \) and/or \( n \) are (even moderately) large, iterative solvers for (1.1), used without preconditioners or with general purpose preconditioners, such as those based on incomplete factorizations, often converge very slow or do not converge at all, see [4, Section 5]. Moreover, direct methods are not appropriate because they cannot exploit the block structure of (1.1).

Preconditioners introduced in [4] take into account this structure. More precisely, they are block-circulant and, in matrix form, can be written as

\[
P = \tilde{A} \otimes I - h \tilde{B} \otimes \tilde{J},
\]

where \( \tilde{A} \) and \( \tilde{B} \) are circulant-like approximations for \( A, B \), respectively and \( \tilde{J} \) is a suitable approximation for \( J \); their performance has been tested in several papers; see, e.g., [4, 5, 14].

The distribution of the eigenvalues of the (left) preconditioned matrix \( P^{-1}M \) influences the convergence of iterations of Krylov subspace methods (see [7]), as well as the spectrum of \( M \) does for nonpreconditioned iterations. Unfortunately, a theoretical investigation is hard because, in general, \( P \) and \( M \) are nonsymmetric and non symmetricizable. Moreover, an analysis of the eigenvalues based on the generating function of the underlying Toeplitz matrices is not feasible here, although very meaningful for Hermitian matrices; see, e.g., [13, 21] and references therein.

These difficulties motivate us to a “direct” analysis, based on the generalized eigenvalue problem

\[
M u = \lambda P u
\]

which is mandatory in order to gather information. The tools used here are completely different from those in previous works such as [8] or [23] and other works by the latter author. In particular, we cannot write anymore \( \tilde{A}, \tilde{B} \) as small rank perturbations of \( A, \ B \), respectively.

By using instead linear difference equation theory and generalizing Trench’s approach [30, 31], we derive closed formulas and first-order expansions for \( \lambda \) as a function of the stepsize (in time) \( h \) and of the eigenvalues of the Jacobian matrix \( J \); such characterization involves the roots of a sparse polynomial whose degree is related to the size of \( M \) and \( P \).

Our estimates are explicitly computed for some well-known 2-step integrators, and compared with the “true” eigenvalues approximated by Matlab. The approach seems very useful for spectrum localization, and is not too expensive provided that \( M \) and \( P \) have moderate size, or an efficient rootfinder is associated with our technique.

The paper is organized as follows. §2 introduces the problem and the main circulant preconditioning techniques. In §3 we discuss the relevant literature for spectral analysis, and we explain in more detail the motivation of our work. §4 is devoted to the spectral analysis, from the general case to the 2-step case study. In §5 we describe two classical PDE examples, representing the test problems for our experiments of §6.

2. Preliminaries. Let us consider a model problem based on a first-order initial-boundary value time-dependent partial differential equation

\[
\begin{aligned}
\frac{\partial u}{\partial t} &= \mathcal{L}(u) + f, \quad u \in D, \\
\mathcal{G}(u) &= g, \quad u \in \partial D, \\
u &= u_0, \quad t = t_0, \quad u \in D,
\end{aligned}
\]

(2.1)
where $D$ is an open domain in $\mathbb{R}^n$, $n \geq 1$, $L$ is a differential operator, nonlinear in $u$ in general, $G$ is the operator which provides the boundary conditions and $f$ is the forcing term. Sometimes equations (2.1) are called of evolution or evolutionary because they are used to describe evolving phenomena and combine differentiation with respect to both space and time. Here, for simplicity, we will focus on a linear model, i.e., operators $L$ and $G$ will be linear in (2.1) and provide some simple model problems in the sequel. However, most of the techniques considered here can be applied to a more general nonlinear framework by recalling that often numerical codes linearize the nonlinear algebraic equations by using a quasi-Newton step; see [18].

2.1. Linear multistep formulas in boundary value form. We are mainly interested in schemes which discretize (2.1) in space generating a (possibly very large) system of ordinary differential equations. Then, an approximate solution of (2.1) is sought by using a method for ordinary differential equations. In particular, we focus on linear multistep formulas applied in boundary value form. These methods are a generalization of implicit linear multistep formulas (see [1, 11]) where initial and boundary conditions are used even in presence of an initial value problem. Such schemes have a relatively long history (see, e.g., [17, 1]) and are of common use in some communities where “time” has no special orientation. The underlying algorithms approximate the solution of a continuous differential boundary value problem by means of a discrete boundary value problem (a system of difference equations). By discretizing with respect to the space variables in (2.1) we get the initial value problem (IVP)

\[
\begin{align*}
   \{ & y'(t) = f(t, y(t)) := J y(t) + g(t), \quad t \in (t_0, T] \\
   & y(t_0) = \eta,
\end{align*}
\]  

(2.2)

where $y(t), g(t) : \mathbb{R} \rightarrow \mathbb{R}^m$, $J \in \mathbb{R}^{m \times m}$, $\eta \in \mathbb{R}^m$, $j = 1, 2$. The continuous problem (2.2) can be reduced to a discrete boundary value problem, i.e., to a system of linear difference equations by the following $k$-step linear multistep formula of order $p$ used with $\nu > 0$ initial and $k - \nu > 0$ final conditions over a uniform mesh $t_j = t_0 + j h$, $j = 0, \ldots, s$:

\[
\sum_{i=0}^{k} \alpha_i y_{n+i} = h \sum_{i=0}^{k} \beta_i f_{n+i}, \quad n = 0, \ldots, s - k,
\]  

(2.3)

where $y_n$ is the discrete approximation to $y(t_n)$, $f_n = f(t_n, y_n) \equiv J y_n + g_n$, $g_n = g(t_n)$, while the values $y_0, \ldots, y_{\nu-1}$, $y_{s-k+\nu+1}, \ldots, y_s$ of the approximation computed in the mesh points $t_0, \ldots, t_{\nu-1}$, $t_{s-k+\nu+1}, \ldots, t_s$, respectively, are assumed to be given. We observe that the boundary value problem (2.2) provides only the initial value $y_0$. The missing values are supplied by coupling the method (2.3) with other difference schemes of order $p$, sometimes called additional methods, which provide an additional set of equations, independent of those in (2.3). For simplicity, we can assume that these formulas have the same number of steps of (2.3) but different coefficients $\alpha^{(r)}_j$, $\beta^{(r)}_j$, $r = 1, \ldots, \nu - 1, s - k + \nu + 1, \ldots, s - 1$, $j = 0, \ldots, k > \nu$, see [4] for details.

For practical implementation, we cast the above discrete problem in matrix form:

\[
MY = b, \quad Y = (y_0^T, y_1^T, \ldots, y_s^T)^T, \quad M = A \otimes I_m - hB \otimes J,
\]

\[
b = e_1 \otimes \eta + h(B \otimes I_m)g, \quad g = (g(t_0) \cdots g(t_s))^T
\]  

(2.4)
where \( e_i \in \mathbb{R}^{s+1}, i = 1, \ldots, s + 1 \), is the \( i \)-th column of the identity matrix and \( A, B \in \mathbb{R}^{(s+1) \times (s+1)} \) are quasi Toeplitz matrices. Examples of matrices \( A, B \) and \( M \) for 2-step formulas can be found in the sequel.

In the case of linear multistep formulas used in boundary value form, for the IVP (2.2), we have the following generic pattern:

\[
A = \begin{pmatrix}
1 & \cdots & 0 \\
\alpha_0^{(1)} & \cdots & \alpha_k^{(1)} \\
\vdots & \ddots & \vdots \\
\alpha_0^{(s-k+\nu+1)} & \cdots & \alpha_k^{(s-k+\nu+1)} \\
\alpha_0 & \cdots & \alpha_k \\
\alpha_0 & \cdots & \alpha_k \\
\vdots & \ddots & \vdots \\
\alpha_0 & \cdots & \alpha_k \\
\alpha_0 & \cdots & \alpha_k \\
\end{pmatrix}, \quad (2.5)
\]

where \( \alpha_j^{(r)}, j = 0, \ldots, k \), are the coefficients of the additional formulas. The matrix \( B \) is defined similarly, but with \( \beta_j \) instead of \( \alpha_j \) and with the entries of the first and last rows set to zero; see [4] for details.

We stress that the matrix \( M \) in (2.4) is usually nonsymmetric, non-diagonally dominant and large and sparse if, e.g., \( n \) or \( m \) are large and \( J \) is sparse.

2.2. A review of block circulant preconditioners. We noted in [4] that when \( u \in \mathbb{R}^d, d > 1 \), and a fine enough spatial discretization is used in (2.1), direct methods are often not feasible to solve linear systems (2.4). Iterative methods are mandatory when the discrete problem is generated by a three-dimensional or even two-dimensional differential model (2.1). In [3, 4] Krylov subspace methods were proposed to solve (2.4). However, without preconditioning, the convergence can be very slow or iterations do not converge at all. Therefore, in [3, 4] a preconditioning strategy based on circulant matrices was introduced (see also [13]). Thus, other approximations were introduced in [5, 6]; see [6] for a more comprehensive bibliography. By left preconditioning we mean solving the equivalent nonsymmetric linear system

\[
P^{-1}Mx = P^{-1}b
\]

instead of \( Mx = b \). Right preconditioning is obtained by considering

\[
MP^{-1}y = b, \quad x = P^{-1}y.
\]

Note that matrices \( MP^{-1} \) and \( P^{-1}M \) are similar and hence share the same eigenvalues. Since we are interested in the eigenvalues of (2.6), our analysis is entirely based on left preconditioning. We stress that convergence of Krylov iterative methods for non normal matrices cannot be described by eigenvalues only (see, e.g., [19]), and it is mandatory to consider more of the problem at hand like the matrix of eigenvectors or the pseudospectrum or the field of values of the transformed problem. See [7]
for a discussion related on problems (2.4), (2.6) where $P^{-1}M$ has tightly clustered spectrum.

In the sequel, some block-circulant and block-circulant-like preconditioners for (2.4) are briefly reviewed.

Let us consider the following approximation of the matrix $M$

$$P = \tilde{A} \otimes I_m - h\tilde{B} \otimes \tilde{J}$$

(2.7)

where $\tilde{J}$ is a suitable approximation of the Jacobian matrix or the Jacobian itself. $\tilde{A}$, $\tilde{B}$ are circulant matrices whose entries are derived from the coefficients of the main method (2.3) as follows:

$$\tilde{A} = \text{circ}(\tilde{a}_j), \quad \tilde{a}_j = c_{j,1}(s)\alpha_{j+\nu} + c_{j,2}(s)\alpha_{j+\nu-(s+1)},$$

$$\tilde{B} = \text{circ}(\tilde{b}_j), \quad \tilde{b}_j = c_{j,3}(s)\beta_{j+\nu} + c_{j,4}(s)\beta_{j+\nu-(s+1)}, \quad j = 0, \ldots, s,$$

(2.8)

where $\text{circ}(\cdot)$ denotes the circulant matrix having the first column specified in the argument, and the $c_{j,i}(s), i = 1, \ldots, 4, j = 0, \ldots, s$ are linear in $j$. It is understood that $\alpha_j (\beta_j)$ is zero for $j < 0$ or $j > k$ in (2.8), so that the sparsity of $A$, $B$ implies that of $\tilde{A}$, $\tilde{B}$. The coefficients $c_{j,i}(s)$ in (2.8) are chosen in such a way that $\tilde{A}$, $\tilde{B}$ are suitable approximations of $A$, $B$ in (2.4), respectively.

The approximation of $A$, $B$ with the T. Chan’s optimal circulant preconditioner (see [15]) requires that

$$c_{j,1}(s) = c_{j,3}(s) = 1 - \frac{j}{s+1}, \quad \text{and} \quad c_{j,2}(s) = c_{j,4}(s) = \frac{j}{s+1}, \quad j = 0, \ldots, s,$$

(2.9)

while for the Strang’s natural (or simple) circulant preconditioner (see [26])

$$c_{j,1}(s) = c_{j,3}(s) = 1, \quad j = 0, \ldots, \left\lfloor \frac{s+1}{2} \right\rfloor,$$

$$c_{j,2}(s) = c_{j,4}(s) = 1, \quad j = \left\lfloor \frac{s+1}{2} \right\rfloor + 1, \ldots, s, \quad c_{j,i}(s) = 0 \text{ otherwise.}$$

On the other hand, if we consider, instead of (2.9), the following definition of the coefficients $c_{j,i}(s)$ for $\tilde{A}$ and $\tilde{B}$

$$c_{j,1}(s) = c_{j,3}(s) = 1 + \frac{j}{s+1}, \quad c_{j,2}(s) = c_{j,4}(s) = \frac{j}{s+1}, \quad j = 0, \ldots, s,$$

(2.10)

we get the so-called P-circulant approximations which, used in (2.7), gives the P-circulant (block) preconditioner, introduced in [3, 4]. The latter definition avoids singularity problems which are sometimes typical of the former choices.

In [3, 4] and in [14] was shown that both the P-circulant and generalized Strang preconditioned systems can be effective to accelerate the convergence. Unfortunately, when the Jacobian matrix $J$ has some small (or zero) eigenvalues, the simple circulant or Strang preconditioner can be severely ill-conditioned or even singular (see [3, 4, 5]). An analysis of the spectrum for the preconditioned matrix based on simple circulant approximations for can be found in [8]. However, we stress that the tools used here are completely different from those in the former. In particular, we cannot write anymore $\tilde{A}$, $\tilde{B}$ as small rank perturbations of $A$, $B$, respectively.
Therefore, we will focus on preconditioners (2.7) based on T. Chan’s and on the $P$-circulant approximations in the following discussions. Practical examples for the matrices $A$, $B$, $\tilde{A}$, $\tilde{B}$, $M$, and $P$ can be found below.

Another approximation which was found effective (but is not considered here) is based on $\{\omega\}$-circulant approximations for matrices $A$ and $B$ in (2.4); see [6]. In particular, $\tilde{A}$, $\tilde{B}$ are $\{\omega\}$-circulant matrices approximating $A$ and $B$, respectively. The $\{\omega\}$-circulant matrices are Toeplitz matrices whose first entry of a row is given by multiplying the last entry of the preceding row by $\omega = \exp (i\theta)$; see [16] for more details. Notice that the $\{1\}$-circulant matrices ($\theta = 0$) are just circulant matrices (and therefore generate simple or Strang’s approximations for a given Toeplitz matrix), while $\{-1\}$-circulant matrices ($\theta = \pi$) are skew-circulant matrices.

We observe that various trigonometric approximations can be combined. For example, $\{\omega\}$–$P$–circulant preconditioners can be defined by using (2.10) to give the first row of the related $\{\omega\}$–circulant approximation. A similar combination can be made by using T. Chan’s optimal circulant matrices. Moreover, it is straightforward to observe that $P$–circulant approximations can be seen as $\{\omega\}$–circulant preconditioners with $\theta = 0$, whose entries are defined as in (2.10). More comments on these generalizations can be found in [6].

2.3. Dahlquist’s hypothesis. In the proposed eigenvalue analysis for the preconditioned linear systems (2.6), unless otherwise specified, we choose $J = \mu$ (i.e., a scalar) in (2.2), where $\mu \in \mathbb{C}^-$ (the “Dahlquist’s hypothesis”). It is customary to consider the above scalar problem in the linear stability theory of methods for ordinary differential equations (or ODEs); see e.g., [20, 11]. The parameter $\mu$ can be any one of the eigenvalues of the Jacobian matrix $J$ of the given partial differential equation, supposed diagonalizable. This framework is not restrictive for our analysis too. Indeed, notice that, supposing $J$ diagonalizable, we have

$$ J = VDV^{-1}, \quad D = \text{diag}(\mu_1, ..., \mu_m), $$

$$ M = A \otimes I_m - hB \otimes (VDV^{-1}) = (I_{s+1} \otimes V)(A \otimes I_m - hB \otimes D)(I_{s+1} \otimes V^{-1}), $$

where $I_{s+1}$ is the $(s+1) \times (s+1)$ identity matrix. A similar expression can be derived for $P$. Let us write

$$ M(q) = A - qB, \quad P(q) = \tilde{A} - q\tilde{B}, \quad q = h\mu. $$

It is straightforward to observe that the eigenvalues of the preconditioned/transformed linear system (2.6) are given by the union of the eigenvalues of the finite sequence of matrices

$$ \{P(q)^{-1}M(q)\}_q, \quad q = h\mu_i, \quad i = 1, ..., m. $$

Just to have an idea of the matrix structures, we sketch below the explicit expression of $M(q)$ and $P(q)$ for the particular example of a two-step generalized Adams-Moulton method in connection to a $P$-circulant preconditioner, with an additional final condition given by the implicit Euler method:

$$ M(q) = \begin{pmatrix} 1 & 0 & & & \\ -1 & 1 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 1 & 0 \\ & & & -1 & 1 \end{pmatrix} - q \cdot \begin{pmatrix} \frac{2}{12} & \frac{-1}{12} & \frac{-1}{12} & \frac{-1}{12} & \frac{-1}{12} \\ \frac{3}{12} & \frac{2}{12} & \frac{-1}{12} & \frac{-1}{12} & \frac{-1}{12} \\ \frac{5}{12} & \frac{3}{12} & \frac{2}{12} & \frac{-1}{12} & \frac{-1}{12} \\ \frac{5}{12} & \frac{3}{12} & \frac{2}{12} & \frac{1}{12} & \frac{-1}{12} \\ \frac{3}{12} & \frac{2}{12} & \frac{1}{12} & \frac{-1}{12} & \frac{-1}{12} \end{pmatrix}. $$
\[ P(q) = \begin{pmatrix}
1 & 0 & -\frac{s}{s+1} \\
-\frac{s}{s+1} & 1 & 0 \\
0 & -\frac{s}{s+1} & 1 \\
\end{pmatrix} - q \begin{pmatrix}
\tilde{b}_1 & \tilde{b}_s \\
\tilde{b}_s & \tilde{b}_1 \\
\end{pmatrix},
\]

where \( \tilde{b}_1 = -\frac{s + 2}{12(s + 1)}, \tilde{b}_s = \frac{5s}{12(s + 1)}. \)

3. Motivation of the work. From now on, we will assume Dahlquist’s hypothesis in order to simplify the theoretical analysis.

We have seen that a deep understanding of the iterative solution of (2.4) requires the knowledge of the following features.

1. How the spectrum of \( M \) depends on the discretization parameters? For instance, for which values of \( q \) (both involving the time step and the Jacobian of the PDE) we can ensure that the spectrum completely lies in \( \mathbb{C}^+ = \{ \lambda \in \mathbb{C} : \text{Re}\lambda > 0 \} \)? Can we exclude the pathological situation where \( M \) is singular?

2. When a suitable preconditioner is applied, we know that the spectrum of \( P^{-1}M \) is clustered. But which localization of the cluster and of the eventual outliers should be expected? Again, how that localization depends on \( q \)?

Concerning the first issue, the literature contains a plenty of spectral results (see e.g. \([10, 29]\)) involving Toeplitz matrices, even though the nonsymmetric case is more difficult to treat. Unfortunately, such results are of asymptotical type and require a critical assumption: the entries of \( M \) must not depend on the size (in other words, as the size varies we obtain a finite section of a fixed infinite matrix). This is not our case, since a variation of \( s \) gives a different value of \( q \) in many significant entries.

The only known results we can apply mainly concern algorithms for a quick computation of a few eigenvalues in \([2]\) or a theoretical analysis of the “pencil” \( A - qB \) (in the sense of generalized eigenvalues) in \([11]\): in the latter book we can find conditions on \( q \) for which \( M \) is nonsingular, that is a partial answer to our questions raised above.

In summary, to our knowledge a theoretical characterization of the eigenvalues of \( M = A - qB \) as functions of \( q \) is still missing; the underlying algebraic setting is the (standard) eigenvalue problem for nonsymmetric Toeplitz matrices with small rank corrections.

Concerning the second issue, some mathematical tools for the spectral analysis of \( P^{-1}M \) have been proposed in literature (see e.g. \([14, 23]\)), but they are all limited to the case where the difference \( M - P \) has small rank. This is true for some choices of \( P \) (like Strang’s preconditioner or its modifications), but several other important instances (like T. Chan’s or P-circulant approximations) give rise to matrices \( M - P \) much more involved.

Therefore appropriate analysis tools for the case where \( P \) is very different from \( M \) are still unknown (except for the Hermitian case, that is not of interest in this context): the underlying algebraic setting is the generalized eigenvalue problem for nonsymmetric Toeplitz matrices with small rank corrections.
The following sections will attempt to give answers to the open questions discussed so far.

4. Spectral analysis. From now on we focus on the generalized eigenvalue problem for nonsymmetric quasi Toeplitz matrices,

\[ M(q)u = \lambda P(q)u, \quad u \neq 0; \quad (4.1) \]

the standard eigenproblem falls into this notation by making the formal assumption \( P(q) = I \) (in this section we are interested just in the structure of the matrices involved).

The lack of symmetry implies that the classical approach based on generating functions gives here very little insight (see the partial results presented in [29]). The best way to characterize eigenvalues (and potentially eigenvectors) by exploiting the band Toeplitz pattern is given by the difference equation approach, proposed by Trench [30] for the standard, pure Toeplitz case.

Let the \( s+1 \) equations of (4.1), as well as the entries of \( u \), be indexed from 0 to \( s \); the indices from \( \nu \) to \( s-k+\nu \) correspond to the rows of \( M(q) \) and \( P(q) \) not affected by the low-rank correction, and containing all the coefficients of the main method. The resulting relations

\[
\sum_{i=0}^{k} (\alpha_i - q\beta_i)u_{i+j} = \lambda \sum_{i=0}^{k} (\bar{\alpha}_{i-\nu} - q\bar{\beta}_{i-\nu})u_{i+j}, \quad j = 0, \ldots, s-k
\]

(where we assume a periodic pattern for \( \bar{\alpha}_i \) and \( \bar{\beta}_i \), whenever a subscript is out of range), can be treated as a linear \( k \)-order homogeneous difference equation with constant coefficients; the first and last rows of (4.1) will provide us with initial and final conditions.

The eigenvector \( u \) is a nonzero solution of the difference problem and therefore can be characterized in terms of the algebraic characteristic equation of degree \( k \):

\[
\pi(z) - \lambda \tilde{\pi}(z) = 0, \quad \pi(z) := \sum_{i=0}^{k} (\alpha_i - q\beta_i)z^i, \quad \tilde{\pi}(z) := \sum_{i=0}^{k} (\bar{\alpha}_{i-\nu} - q\bar{\beta}_{i-\nu})z^i \quad (4.2)
\]

(notice that \( \tilde{\pi}(z) \) simplifies into \( z^\nu \) in the standard problem).

From now on we assume that, for each eigenvalue \( \lambda \), all the roots \( z_1(\lambda), \ldots, z_k(\lambda) \) of the characteristic equation are distinct (otherwise, \( \lambda \) is called defective [31] but this pathological situation occurs just in isolated cases and for specific values of \( s \)). In this case, each component of the solution of the difference equation has the form

\[
u_j = \sum_{l=1}^{k} c_l z_l(\lambda)^j, \quad j = 0, \ldots, s, \quad (4.3)
\]

for suitable coefficients \( c_1, \ldots, c_k \) determined by the boundary conditions.

More specifically, the first \( \nu \) and the last \( k-\nu \) rows of (4.1) represent additionally conditions on the sequence \( u_j \); in the standard problem, we have \( \nu \) initial and \( k-\nu \) final conditions since just the first and last entries of \( u \) are respectively involved. In the generalized problem, the circulant structure of \( P(q) \) determines a mixing of initial and final entries in all these \( k \) equations, but for simplicity we keep the same terminology.
Substituting (4.3) into the mentioned equations, we obtain a total of \( k \) homogeneous relations involving the unknown coefficients \( c_1, \ldots, c_k \), which can be put in matrix form as follows:

\[
\begin{align*}
K_{\text{in}}(z_1(\lambda), \ldots, z_k(\lambda))c &= 0 \\
K_{\text{fin}}(z_1(\lambda), \ldots, z_k(\lambda))c &= 0,
\end{align*}
\]

(4.4)

where \( K_{\text{in}} \in \mathbb{C}^{\nu \times k} \), \( K_{\text{fin}} \in \mathbb{C}^{(k-\nu) \times k} \) and we have emphasized the dependence of these Vandermonde-like matrices on the roots of the characteristic equation. The trivial solution \( c = 0 \) would imply \( u = 0 \) and therefore must be discarded; hence the square matrix

\[
K(z_1(\lambda), \ldots, z_k(\lambda)) := \left( \begin{array}{c} K_{\text{in}} \\ K_{\text{fin}} \end{array} \right)
\]

(4.5)

must be singular, and its vanishing determinant can be regarded as a function of \( \lambda \) having the same zeros of the characteristic polynomial of (4.1).

Anyway, an alternative parameterization with respect to the roots \( z_j(\lambda) \) can be useful for a different characterization of \( \lambda \).

Let \( \zeta \) be one of the roots, say \( z_1(\lambda) \). From one point of view, \( \zeta \) is a function of \( \lambda \), but it is understood that \( \lambda \) can be retrieved as well from \( \zeta \) by means of the characteristic equation:

\[
\lambda(\zeta) = \frac{\pi_1(\zeta)}{\pi(\zeta)} \quad (\lambda(\zeta) = \zeta^{-\nu} \pi(\zeta) \text{ in the standard case});
\]

(4.6)

remark that any root gives the same value of \( \lambda \).

The other roots can be expressed in terms of \( \lambda \) by inverting some elementary symmetric functions.

For example, in the generalized problem with \( k = 2 \) and \( \nu = 1 \), the easiest way is considering the ratio between the constant term and the leading coefficient in (4.2):

\[
\frac{\alpha_0 - q\beta_0 - \lambda(\bar{a}_s - q\bar{b}_s)}{\alpha_2 - q\beta_2 - \lambda(\bar{a}_1 - q\bar{b}_1)} = z_1(\lambda)z_2(\lambda),
\]

whence, after the substitution \( \lambda = \lambda(\zeta) \) given in (4.6),

\[
z_2(\lambda) = \zeta^{-1} \frac{(\alpha_0 - q\beta_0)\pi(\zeta) - (\bar{a}_s - q\bar{b}_s)\pi(\zeta)}{(\alpha_2 - q\beta_2)\pi(\zeta) - (\bar{a}_1 - q\bar{b}_1)\pi(\zeta)} =: \zeta_2(\zeta).
\]

(4.7)

In general, we can assume to have explicit functions \( \zeta_2(\zeta), \ldots, \zeta_k(\zeta) \) that must replace \( z_2(\lambda), \ldots, z_k(\lambda) \) in the matrix \( K \) of (4.5). Thus

\[
\det K(\zeta, \zeta_2(\zeta), \ldots, \zeta_k(\zeta)) =: \det(\zeta; q)
\]

is a function of the single complex variable \( \zeta \), containing \( q \) as a parameter.

As we will see in the specific examples, the function \( \det(\zeta; q) \) can be sometimes reduced to the study of a sparse polynomial, which makes feasible a first-order analysis (perhaps a direct computation) of its roots \( \zeta(q) \). Finally, the relation (4.6) allows us to obtain a knowledge of \( \lambda = \lambda(q) \) from that of \( \zeta(q) \).

**Remark 4.1.** \( \det(\zeta; q) \) has a number of redundant roots that we may discard in order to simplify the analysis. Some of them are “spurious” values for which
ζ = ζ_j(ζ) or ζ_j(ζ) = G(ζ) with j ≠ l (the matrix K turns out to have two equal columns), violating the assumption of distinct roots. Furthermore, if ζ is a root of det(ζ; q) then ζ_2(ζ), ..., ζ_k(ζ) are roots as well, and they all give the same eigenvalue λ. In summary, since there are s + 1 eigenvalues we expect to find k(s + 1) roots of det(ζ; q), plus the spurious roots (whose number cannot be estimated a priori, in general).

**Remark 4.2.** Once the behavior of ζ(q) has been obtained, in principle this can be used also for the study of eigenvectors: the key relation is (4.3) and the main issue would be the behavior (in terms of q) of the coefficients c_1, ..., c_k. This problem is not treated in the present paper, where we are only interested in the eigenvalues λ(q).

**4.1. A case study: 2-step formulas.** In this paper, we focus on two-step methods as the principal (or main) scheme (2.3) for a linear multistep formula in boundary value form, with one initial condition and one final condition provided by an implicit Euler scheme. For those methods, we have

\[ k = 2, \ \nu = 1 \]

and the Dahlquist’s hypothesis of §2.3 allows us to assume that

\[
A = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & 0 \\ \alpha_0 & \alpha_1 & \alpha_2 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ -1 & 1 & \end{pmatrix}, \quad B = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 \\ \beta_0 & \beta_1 & \beta_2 \\ \vdots & \vdots & \vdots \\ 0 & 1 & \end{pmatrix}
\]

where parameters are given in Table 4.1 for the most common cases.

<table>
<thead>
<tr>
<th>Type</th>
<th>α_0</th>
<th>α_1</th>
<th>α_2</th>
<th>β_0</th>
<th>β_1</th>
<th>β_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midpoint (MP)</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Simpson (S)</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>1/3</td>
<td>2/3</td>
<td>1/3</td>
</tr>
<tr>
<td>Adams Moulton (AM)</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>5/12</td>
<td>3/12</td>
<td>-1/12</td>
</tr>
</tbody>
</table>

*Table 4.1: Coefficients for some two-steps formulas.*

Circulant approximations for A and B are given by

\[
\tilde{A} = \begin{pmatrix} \tilde{a}_0 & \tilde{a}_1 & \tilde{a}_s \\ \tilde{a}_s & \tilde{a}_0 & \tilde{a}_1 \\ \vdots & \vdots & \vdots \\ \tilde{a}_1 & \tilde{a}_s & \tilde{a}_0 \\ \tilde{a}_s & \tilde{a}_0 & \tilde{a}_1 \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} \tilde{b}_0 & \tilde{b}_1 & \tilde{b}_s \\ \tilde{b}_s & \tilde{b}_0 & \tilde{b}_1 \\ \vdots & \vdots & \vdots \\ \tilde{b}_1 & \tilde{b}_s & \tilde{b}_0 \end{pmatrix}
\]

where the examples for its entries considered here are shown in Table 4.2.

We are not interested in T. Chan’s approximation for the Midpoint and Simpson methods, since it becomes singular in such cases.

The characteristic equation (4.2) has a quadratic form, with

\[
\pi(z) = \gamma_0 + \gamma_1 z + \gamma_2 z^2 \quad (\gamma_i := \alpha_i - q\beta_i)
\]
Hence we define \( \zeta \) the numerator vanishes for \( \zeta = 0 \), whereas the denominator looses its quadratic term. In the light of Remark 4.1, we know that the function \( \det(\zeta; q) \) has two spurious roots for which \( \zeta = \zeta_2 \), satisfying the quadratic equation \( N(\zeta) - \zeta D(\zeta) = 0 \).

In order to form the explicit expression of \( \det(\zeta; q) \), first we must compute the \( 2 \times 2 \) matrix

\[
K(\zeta, \zeta_2) = \begin{pmatrix}
\gamma_{\text{in}}(\zeta) & \gamma_{\text{in}}(\zeta_2) \\
\gamma_{\text{fin}}(\zeta) & \gamma_{\text{fin}}(\zeta_2)
\end{pmatrix},
\]

where \( \gamma_{\text{in}}(\cdot) \) and \( \gamma_{\text{fin}}(\cdot) \) are suitable polynomials obtained by imposing boundary conditions to the main difference equation. Some algebraic manipulations give the following general formulas:

\[
\gamma_{\text{in}}(z) = \pi_0(z) - \frac{\pi(z)}{\tilde{\pi}(z)} R[z^s \tilde{\pi}], \quad \gamma_{\text{fin}}(z) = z^{s-2} \pi_s(z) - \frac{\pi(z)}{\tilde{\pi}(z)} R[z^{s-1} \tilde{\pi}],
\]

where the notation \( R[P] \) means the remainder of \( P \) modulo \( z^{s+1} - 1 \), and

\[
\pi_0(z) := \frac{\pi(z) - \pi(0)}{z}, \quad \pi_s(z) := (1-q)z^2 - z
\]

are polynomials associated with the first and last rows of (4.8), respectively.

An useful simplification arises by observing that the fraction \( \pi/\tilde{\pi} \) always equals \( \lambda \) and therefore needs not to be evaluated in \( \zeta_2 \) when we form the second column of \( K \). The determinant is clearly given by

\[
\det(\zeta; q) = \gamma_{\text{in}}(\zeta)\gamma_{\text{fin}}(\zeta_2) - \gamma_{\text{in}}(\zeta_2)\gamma_{\text{fin}}(\zeta)
\]

and a direct look shows that it is a rational function whose denominator is \( \tilde{\pi}(\zeta)^2 D(\zeta)^* \).

It follows that the function

\[
d(\zeta; q) := \tilde{\pi}(\zeta)^2 D(\zeta)^* \det(\zeta; q)
\]
is a \((2s + 4)\)-degree polynomial in \(\zeta\), whose \(N(\zeta) - \zeta D(\zeta)\) is an exact divisor. Its significant roots occur in pairs \((\zeta(q), \zeta_2(q))\), each of them providing a unique value of \(\lambda(q)\).

The formulas derived so far simplify very much if we are concerned with the nonpreconditioned case: it suffices to put formally \(\tilde{\pi}(z) := z\), so that

\[
\zeta_2 = \frac{\gamma_0}{\gamma_2 \zeta} \quad \gamma_{\text{in}}(z) := \pi_0(z) - \frac{\pi(z)}{z} = -\frac{\gamma_0}{z} \quad \gamma_{\text{fin}}(z) := z^{s-2}(\pi_s(z) - z\pi(z))
\]

whence

\[
\det(\zeta; q) = -\gamma_2 \zeta_2^{s-1}((\pi_s(\zeta_2) - \zeta_2 \pi(\zeta_2)) + \gamma_2 \zeta^{s-1}((\pi_s(\zeta) - \zeta \pi(\zeta)) ;
\]

here the denominator is just \(\zeta^{s+2}\), and the spurious roots of \(d(\zeta; q) := \zeta^{s+2} \det(\zeta; q)\) are \(\pm \sqrt{\gamma_0/\gamma_2}\).

It is important to observe that, in every case, \(d(\zeta; q)\) is a sparse polynomial, which makes a first-order analysis quite feasible.

We sketch below the essential formulas arising for the specific examples under consideration, which represent the individual instances of equations (4.10), (4.11) and (4.6).

**Nonpreconditioned matrices** \(M(q)\)

**Midpoint (MP):**

\[
\zeta_2 = -\frac{1}{\zeta} \\
\lambda(q) = \zeta(q) - 2q - \frac{1}{\zeta(q)}
\]

**Simpson (S):**

\[
\zeta_2 = \frac{\gamma}{\zeta} \quad \gamma := \frac{q/3 + 1}{q/3 - 1} \\
\lambda(q) = \left(1 - \frac{q}{3}\right) \zeta(q) - \frac{4}{3}q - \left(1 + \frac{q}{3}\right) /\zeta(q)
\]
Adams-Moulton (AM):

\[ \zeta_2 = \frac{\gamma}{\zeta}, \quad \gamma := -5 - \frac{12}{q} \]

\[ d(\zeta; q) = \gamma^{s-1} \left( 1 + \frac{5}{12} q \right) \left( \gamma \left( 1 + \frac{5}{12} q \right) - \frac{1}{3} \gamma q \zeta + \frac{5}{12} q \zeta^2 \right) \]

\[ + \zeta^{2s+2} \frac{q^2}{3} \left( \frac{5}{4} - \zeta - \frac{1}{4} \zeta^2 \right) \]

\[ \lambda(q) = \frac{q}{12} \zeta(q) + 1 - \frac{2}{3} q - \left( 1 + \frac{5}{12} q \right) / \zeta(q) \]

Preconditioned matrices \( P(q)^{-1} M(q) \)

MP, P-circulant:

\[ N(\zeta) = q - \zeta, \quad D(\zeta) = 1 + q \zeta, \quad \zeta_2 = \frac{N(\zeta)}{D(\zeta)} \]

\[ d(\zeta; q) = \frac{2\pi(\zeta)}{s+1} (N - \zeta D) \left[ - \left( 1 - \frac{1}{s+1} \right) \zeta^s N^s + \left( 1 + \frac{1}{s+1} \right) D^s \right] \]

\[ + \left( 1 - \frac{1}{(s+1)^2} \right) \pi(\zeta)^2 [N^s - (\zeta D)^s] \]

\[ - \frac{2(1+q)\pi(\zeta)}{s+1} [N^{s+1} - (\zeta D)^{s+1}] + \frac{4}{(s+1)^2} [N^{s+2} - (\zeta D)^{s+2}] \]

\[ \pi(\zeta) = \zeta^2 - 2q\zeta - 1, \quad \tilde{\pi}(\zeta) = \pi(\zeta) + \frac{1}{s+1}(\zeta^2 + 1), \quad \lambda(q) = \frac{\pi(\zeta(q))}{\tilde{\pi}(\zeta(q))} \]

S, P-circulant:

\[ N(\zeta) = (3+q)(2q + (q-3)\zeta), \quad D(\zeta) = (3-q)(q+3+2q \zeta) \]

\[ d(\zeta; q) = \pi(\zeta)(N - \zeta D)(\tilde{g}_1 D^{s-1} \phi_1 + \tilde{g}_s (\zeta N)^{s-1} \psi_0) \]

\[ - \left( \tilde{g}_1 \tilde{g}_s \pi(\zeta)^2 + \frac{2D\psi_0}{9(s+1)} \right) [N^s - (\zeta D)^s] - \frac{2\psi_1}{9(s+1)} [N^{s+1} - (\zeta D)^{s+1}], \]

(\( \tilde{g}_i := \tilde{a}_i - q \tilde{b}_i \) where \( \tilde{a}_i \) and \( \tilde{b}_i \) are given by Table 4.2)

\[ \phi_1 := \frac{2D}{9(s+1)}, \quad \psi_0 := -\tilde{\pi} + \left( 1 - \frac{1}{s+1} \right) \left( 1 + \frac{q}{3} \right) \pi, \quad \psi_1 := (1-q)\tilde{\pi} + \frac{4q}{3} \pi \]

\[ \pi(\zeta) = \zeta^2 - 1 - \frac{q}{3}(\zeta^2 + 4\zeta + 1), \quad \tilde{\pi}(\zeta) = \pi(\zeta) + \frac{1}{s+1} \left( 1 + \zeta^2 + \frac{q}{3}(1 - \zeta^2) \right) \]
AM, T. Chan:

\[ N(\zeta) = 5q + 12 \quad D(\zeta) = -q\zeta \]

\[
d(\zeta; q) = \left(1 - \frac{1}{s+1}\right)\pi(\zeta)(N - \zeta D) \left[\frac{q}{12}D^{s-1}\phi_1 - \left(1 + \frac{5}{12}q\right)(\zeta N)^{s-1}\psi_0\right] \\
+ \zeta ND\phi_1\psi_0[N^{s-2} - (\zeta D)^{s-2}] + (\zeta N\phi_1\psi_1 + D\phi_0\psi_0)[N^{s-1} - (\zeta D)^{s-1}] \\
+ \left[\phi_0\psi_1 + \frac{q}{12} \left(1 - \frac{1}{s+1}\right)^2 \left(1 + \frac{5}{12}q\right)\pi^2\right][N^s - (\zeta D)^s],
\]

\[
\phi_0 := \left(1 - \frac{2}{3}q\right)(\bar{\pi} - \pi), \quad \phi_1 := \frac{q}{12} \left[\bar{\pi} - \left(1 - \frac{1}{s+1}\right)\pi\right],
\]

\[
\psi_0 := -\bar{\pi} + \left(1 - \frac{1}{s+1}\right)\left(1 + \frac{5}{12}q\right)\pi, \quad \psi_1 := (1 - q)\bar{\pi} - \left(1 - \frac{2}{3}q\right)\pi
\]

\[
\pi(\zeta) = \zeta - 1 - \frac{q}{12}(\zeta^2 - 8\zeta - 5), \quad \bar{\pi}(\zeta) = \pi(\zeta) + \frac{1}{s+1}\left(-\frac{q}{12}\zeta^2 + 1 + \frac{5}{12}q\right)
\]

AM, P-circulant:

\[ N(\zeta) = \left(1 + \frac{5}{12}q\right)\left(1 - \frac{2}{3}q + \frac{1}{6}q^2\zeta\right), \quad D(\zeta) = \frac{q}{12} \left[\left(1 - \frac{2}{3}q\right)\zeta - 2 + \frac{5}{6}q\right] \]

\[
d(\zeta; q) = \pi(\zeta)(N - \zeta D)(\bar{\gamma}_1 D^{s-1}\phi_1 + \bar{\gamma}_s(\zeta N)^{s-1}\psi_0) + \zeta ND\phi_1\psi_0[N^{s-2} - (\zeta D)^{s-2}] \\
+ (\zeta N\phi_1\psi_1 + D\phi_0\psi_0)[N^{s-1} - (\zeta D)^{s-1}] + (\phi_0\psi_1 - \bar{\gamma}_1\bar{\gamma}_s\pi(\zeta^2))[N^s - (\zeta D)^s],
\]

\[(\bar{\gamma}_i := \bar{a}_i - \bar{\theta}_i) \text{ where } \bar{a}_i \text{ and } \bar{\theta}_i \text{ are given by Table 4.2} \]

\[
\phi_0 := \left(1 - \frac{2}{3}q\right)(\bar{\pi} - \pi), \quad \phi_1 := \frac{q}{12} \left[\bar{\pi} - \left(1 + \frac{1}{s+1}\right)\pi\right],
\]

\[
\psi_0 := -\bar{\pi} + \left(1 - \frac{1}{s+1}\right)\left(1 + \frac{5}{12}q\right)\pi, \quad \psi_1 := (1 - q)\bar{\pi} - \left(1 - \frac{2}{3}q\right)\pi
\]

\[
\pi(\zeta) = \zeta - 1 - \frac{q}{12}(\zeta^2 - 8\zeta - 5), \quad \bar{\pi}(\zeta) = \pi(\zeta) + \frac{1}{s+1}\left(-\frac{q}{12}\zeta^2 + 1 + \frac{5}{12}q\right) \quad .
\]

4.2. A first-order analysis. The parameterization obtained so far of \(\lambda\) as a function of \(q\) allows us to investigate the behavior of the eigenvalues for \(q\) small.

We recall that \(q = h\mu\), where \(h\) is the time discretization step and \(\mu\) represents any eigenvalue of the Jacobian matrix \(J\) in (2.4) related to the space discretization. Thus, a small value of \(q\) is a physically meaningful situation, occurring whenever, e.g., the Jacobian matrix has eigenvalues with small modulus (as in the examples sketched...
in §5) and/or a small time step is used. A particular care is required in the latter instance: we stress that $s \to \infty$ as $h \to 0$, so that the polynomial $d(\zeta; q)$ raises its degree increasing the number of the roots $\zeta(q)$. Anyway, the insight given by the first-order analysis is generally in good agreement with the real localization of $\lambda$, as we will see in the numerical experiments of §6.

In the sequel, we present the first-order results for all the three nonpreconditioned methods (MP, S, AM) and for two examples of preconditioned matrices (P-circulant approximations for MP and AM).

The starting point is the continuity of polynomial roots with respect of coefficients (provided that the degree remains constant). Hence $\zeta(q)$ is very close to $\zeta(0)$ for small $q$, and its first-order dependence on $q$ can be made explicit.

In the MP method, $\zeta = \zeta(0)$ is a root of $d(\zeta; 0) = (-1)^{s}(1 + \zeta) + \zeta^{2s+3}(1 - \zeta)$, and therefore

$$|\zeta|^{2s+3} = \left|\frac{1+\zeta}{1-\zeta}\right|.$$

Squaring both sides of the previous equation and letting $\zeta = \rho e^{i\theta}$, after some algebraic manipulations, we get

$$\cos \theta = \frac{1 + \rho^2}{2\rho} \frac{\rho^{4s+6} - 1}{\rho^{4s+6} + 1}.$$  \hspace{1cm} (4.12)

(4.12) is the equation, in polar coordinates, of a curve containing all the roots $\zeta(0)$ and lying in the following region of the complex plane:

$$\Omega = \left\{ \theta \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right), \rho < 1 \right\} \cup \left\{ |\theta| < \frac{\pi}{2}, \rho > 1 \right\} \cup \{ \pm i \},$$

where $\pm i$ are exactly the spurious roots for which $\zeta = \zeta_2$.

Therefore $\lambda(0)$ can be localized through the transformation $\lambda = \zeta - 1/\zeta$ of the previous curve. In particular, since $\text{Re}\lambda = (\rho - 1/\rho)\cos \theta$, it is straightforward to observe that $\text{Re}\lambda > 0$ whenever $\zeta \in \Omega$ (except for the spurious roots). By continuity, we have the useful result that the eigenvalues of $M$ lie on $C^+$ for $q$ small enough. We recall that projection methods like GMRES or BiCGstab show often a faster convergence behavior whenever the matrix of the linear systems we have to solve has all eigenvalues in one half plane; see [19].

If we are interested in a deeper analysis, we can check that the roots $\zeta(0)$ are distinct and therefore

$$\zeta(q) \doteq \zeta(0) + \zeta'(0)q,$$

where $\doteq$ denotes a first-order approximation of the function in the left hand side. Therefore,

$$\lambda(q) \doteq \zeta(0) - \frac{1}{\zeta(0)} + \left[ \zeta'(0) + \frac{\zeta'(0)}{\zeta(0)^2} - 2 \right] q;$$

the explicit expression of $\zeta'(0)$, if desired, can be retrieved from the classical theory on the conditioning of zeros of polynomials (see, e.g., [27, §5.8]).
The Simpson method has a quite similar analysis. In addition, since the matrix $A$ is the same as the previous case, the zero-order terms of $\zeta(q)$ and $\lambda(q)$ are exactly equal to the corresponding ones for MP. However, new computations should be done for the terms of order 1, obtaining in particular

$$\lambda(q) \doteq \zeta(0) - \frac{1}{\zeta(0)} + \left[ \zeta'(0) + \frac{\zeta'(0)}{\zeta(0)^2} - \frac{4}{3} - \frac{\zeta(0)}{3} \right] q$$

(with a value for $\zeta'(0)$ different from the MP method).

The analysis of the AM method shows a further complication with respect to the previous cases. It is evident that $d(\zeta; q)$ loses several degrees when $q$ goes to zero, so that many roots $\zeta(q)$ become infinite. Hence we are not able to predict the behavior of $\lambda(q)$, unless we apply an appropriate change of variable. For this purpose, let

$$\xi := q^{1/2} \zeta, \quad \beta := \gamma q$$

and rewrite the polynomial $d$ in terms of the new variable $\xi$. We obtain

$$q^s d(\xi; q) = \beta^{s-1} \left( 1 + \frac{5}{12} q \right) \left( \beta \left( 1 + \frac{5}{12} q \right) - \frac{1}{3} \beta q^{1/2} \xi + \frac{5}{12} q \xi^2 \right) + \xi^{2s+2}/36 \left( \frac{5}{4} q - q^{1/2} \xi - \frac{1}{4} \xi^2 \right),$$

whence $\xi(q) \doteq \xi(0) + \xi'(0) q^{1/2}$ where $\xi(0)$ solves the equation

$$(-12)^s - \xi^{2s+4}/144 = 0,$$

that is $\xi(0) = 2 \sqrt{3} \exp \left( i \left( \frac{l \pi}{s+2} + \frac{\pi}{2} \right) \right)$, $l = 1, \ldots, s+1$ (other values of the index $l$ would give a spurious root or the same values of $\lambda$).

Taking into account the change of variable, the behaviors of $\zeta$ and $\lambda$ are respectively

$$\zeta(q) \doteq \xi(0) q^{-1/2} + \xi'(0),$$

$$\lambda(q) \doteq 1 + \frac{i}{\sqrt{3}} \cos \frac{l \pi}{s+2} q^{1/2} + \left[ \frac{\xi'(0)}{12} + \frac{\xi'(0)}{\xi(0)^2} - \frac{2}{3} \right] q.$$

Hence, for small values of $q$, the eigenvalues of $M$ are close to a segment on $\mathbb{C}^+$ with the midpoint placed at 1.

The main complication arising in the preconditioned case is the presence of the spurious divisor $N - \zeta D$ in all instances of $d(\zeta; q)$. In order to perform the analysis, it is worth to consider the quotient $\hat{d}(\zeta; q) := d(\zeta; q) / (N - \zeta D)$, which must be studied for $q \approx 0$. Let

$$F_m(\zeta; q) := \frac{N^m - (\zeta D)^m}{N - \zeta D} = \sum_{j=0}^{m-1} N^j (\zeta D)^{m-j-1};$$

this expression appears in almost all the terms of $\hat{d}(\zeta; q)$, and will determine the first-order behavior of the significant roots.
Concerning the P-circulant preconditioner for the MP method, for \( q = 0 \) we have \( N(\zeta) = -\zeta, \ D(\zeta) = 1 \). Therefore
\[
F_m(\zeta; 0) = \zeta^{m-1} \sum_{j=0}^{m-1} (-1)^j = \begin{cases} \zeta^{m-1} & \text{if } m \text{ is odd,} \\ 0 & \text{otherwise.} \end{cases}
\]

Thus, the zero-order localization of the roots \( \zeta(q) \) strictly depends on the parity of \( s \): more specifically, when \( s \) is odd they solve the equation
\[
(\zeta^2 - 1)\zeta^{s+1} + \frac{2}{s+1}(\zeta^2 - 1)(\zeta^{2s} + 1) + \frac{1}{(s+1)^2}[4\zeta^{s+1} - (\zeta^2 - 1)\zeta^{s-1} + 2(\zeta^2 - 1)(1 - \zeta^{2s})] = 0 ,
\]
when \( s \) is even the equation becomes
\[
(\zeta^2 - 1)(1 - \zeta^s - \zeta^{2s}) + \frac{1}{s+1}[(\zeta^2 - 1)(\zeta^{2s} + 1) - (\zeta^2 + 1)\zeta^s] = 0 .
\]

For \( q \) very small, the eigenvalues of \( P^{-1}M \) can be estimated from the roots \( \zeta = \zeta(0) \) through the relation
\[
\lambda(0) = \frac{\zeta^2 - 1}{\zeta^2 - 1 + \frac{1}{s+1}(\zeta^2 + 1)}.
\]

After more heavy computations we are able to obtain the first-order terms in the expansions of \( \zeta(q) \) and \( \lambda(q) \). In §6 we will present explicit estimates based on the formulas derived so far, comparing them with numerical experiments.

The complication found in the analysis of the AM method naturally arises in the P-circulant preconditioned case too. Many ingredients of \( d(\zeta; q) \) degenerate for \( q = 0 \): among others, \( N(\zeta), \ D(\zeta), \pi(\zeta) \) and \( \tilde{\pi}(\zeta) \) drop their degree. This causes several roots \( \zeta(q) \) going to infinity: also here we need a suitable change of variable.

Let \( \xi := q\zeta \) and rewrite all the polynomials \( N, D, \pi, \tilde{\pi}, \phi, \psi \ (i = 0, 1) \) in terms of the new variable, in particular
\[
N(\xi) = \left(1 + \frac{5}{12}q\right)\left(1 - \frac{2}{3}q + \frac{1}{6}\xi\right) , \quad D(\xi) = \frac{1}{12}\left[\left(1 - \frac{2}{3}q\right)\xi - q\left(2 + \frac{5}{6}q\right)\right] .
\]

The “clean” polynomial \( \hat{d} \) takes the following expression after some algebra:
\[
d(\zeta; q) = q^{-s-1}d(\xi; q) ,
\]
where \( \hat{d} \) has constant degree \( 2s + 2 \) (independently on \( q \)) and its zero-order form is
\[
\hat{d}(\xi; 0) = \left(-\frac{1}{s+1}\right)^s \xi^{s+1} \left\{ -\frac{1}{12(s+1)} \left[\xi^{s-3}(12 + 2\xi)^2 + \xi^{s+1}\right] - \left(1 - \frac{1}{s+1}\right) \left(1 + \frac{1}{12}\xi\right) \left(1 + \frac{1}{6}\xi\right)^s \right\} .
\]

Notice that \( s + 1 \) roots of \( \hat{d} \) are distinct and behave as \( \zeta(q) \approx \xi(0) + \xi'(0)q \), whence
\[
\zeta(q) \approx \xi(0)q^{-1} + \xi'(0) .
\]
These are the roots going to infinity, associated with the values of \( \lambda(q) \) with
\[
\lambda(0) = \frac{12 + \xi(0)}{12 + \xi(0) \left( 1 - \frac{1}{s+1} \right)};
\]
the same eigenvalues are associated with the corresponding “dual” roots given by \( \zeta_2 = N(\xi)/D(\xi) \) which are finite for \( q = 0 \), as a direct look at \( N \) and \( D \) shows. Through the transformation \( \xi = q\zeta_2 \) we find the remaining \( s + 1 \) roots of \( \tilde{d} \), which collapse at the origin.

Explicit estimates for this case too will be presented in §6.

5. Model problems. As a first benchmark of our analysis, we consider two simple model problems which encompass two important types of spectra for their Jacobian matrices: real and negative and pure imaginary eigenvalues, respectively. Only 1D problems are considered but extensions to 2D and 3D cases are straightforward and not necessary in our setting.

Diffusion equation. As a typical example of a problem whose Jacobian matrix has negative (real) eigenvalues we report the variable coefficient 1D diffusion equation with homogeneous Dirichlet boundary conditions at both ends. Let \( a = a(x) \geq 0 \) be a suitably smooth function.

\[
\begin{align*}
\frac{\partial u}{\partial t} - c(a(x) \frac{\partial u}{\partial x}) &= 0, \quad x \in [0, x_{\text{max}}], \ t \in (0, T] \\
u(0, t) &= u(x_{\text{max}}, t) = 0 \quad t \in (0, T] \\
u(x, 0) &= g(x), \quad x \in [0, x_{\text{max}}]
\end{align*}
\]

Discretizing the operator \( \partial/\partial x \) in (5.1) with centered differences and stepsize \( \Delta x = x_{\text{max}}/(m + 1) \) gives the set of \( m \) equations

\[
\begin{align*}
y'(t) &= T_m y(t), \quad t \in [0, T] \\
y(0) &= \eta, \quad \eta = (g(x_1) \cdots g(x_m))^T
\end{align*}
\]

where \( x_j = j\Delta x \) and

\[
T_m = \frac{c}{(\Delta x)^2} \begin{pmatrix}
a_1 & b_1 & & \\
& b_1 & \ddots & & \\
& & \ddots & \ddots & \\
& & & b_{m-1} & b_m \\
& & & & a_m
\end{pmatrix},
\]

where

\[
a_j = -(a(x_{j-1/2}) + a(x_{j+1/2})), \quad b_j = a(x_{j+1/2}).
\]

The Jacobian matrix \( T_m \) is \( m \times m \) symmetric, tridiagonal, weakly diagonally dominant with negative diagonal entries. From an extensive study performed in [12] and [25] we get for each eigenvalue \( \mu_j \) of \( T_m \) the bounds

\[
-\frac{4c}{(\Delta x)^2} \max_x \{a(x)\} \leq \mu_j \leq -\frac{c\pi^2}{(x_{\text{max}})^2} \min_x \{a(x)\}.
\]

Note that, as \( \Delta x \) tends to zero, the differential equations (5.2) become increasingly stiff, spreading the eigenvalues of the Jacobian matrix \( T_m \) along an interval in \((-4c\max_x \{a\}/(\Delta x)^2, 0)\) whose left boundary tends to \(-\infty\) with \( O((\Delta x)^{-2}) \).
More precisely, the spectrum is equally distributed [28] as the values of the bivariate function $a(x)f(\theta)$, being

$$f(\theta) = \frac{2c}{(\Delta x)^2} (\cos(\theta) - 1), \quad \theta \in (-\pi, \pi)$$

the so called "generating function" related to the constant-coefficient version of the problem. As stated in [24, 22], if $a(x)$ has a zero at the origin of order $\alpha$, the smallest eigenvalue shows an asymptotic behaviour like $(\Delta x)^2/m^{\max(2, \alpha)}$.

**Transport equation.** The linear 1D transport equation with periodic boundary conditions and constant coefficient $c > 0$ in its simplest form reads:

$$\begin{aligned}
& \left\{ 
\begin{array}{l}
\frac{du}{dt} + cu_x = 0, \\
u(x, 0) = g(x), \quad x \in [0, \pi] \\
u(\pi, t) = u(0, t), \quad t \in [0, 2\pi].
\end{array}
\right. 
\end{aligned} \tag{5.4}$$

Discretizing the partial derivative $\partial/\partial x$ with central differences and stepsize $\Delta x = \pi/m$, $x_j = j\Delta x$ gives the set of equations

$$\begin{aligned}
& \left\{ 
\begin{array}{l}
y'(t) = C_my(t), \quad t \in [0, 2\pi] \\
y(0) = \eta, \quad \eta = (g(x_0) \cdots g(x_{m-1}))^T,
\end{array}
\right. 
\end{aligned} \tag{5.5}$$

with

$$C_m = \frac{c}{2\Delta x} \begin{pmatrix}
0 & -1 & 1 \\
1 & \ddots & \ddots \\
& \ddots & \ddots & -1 \\
-1 & 1 & 0
\end{pmatrix}. \tag{5.6}$$

The matrix $C_m$ is circulant $m \times m$ with generating function

$$\hat{f}(\theta) = \frac{c}{2\Delta x} \left( e^{-i\theta} - e^{i\theta} - e^{-i(m-1)\theta} + e^{i(m-1)\theta} \right) = \frac{-ic}{\Delta x} \left( \sin \theta - \sin(m-1)\theta \right), \tag{5.7}$$

where $\theta \in (-\pi, \pi)$. Therefore, the eigenvalues of $C_m$ are distributed as $\hat{f}(\theta)$ in (5.7) and lie in the purely imaginary (closed) interval

$$[-2ic/\Delta x, 2ic/\Delta x]$$

which becomes wider as $O(1/\Delta x)$ as we refine the discretization. This implies that a finer mesh for the time-step integrator is required to resolve the (oscillatory) solution as $\Delta x$ (the step for the discretization in space) decreases to zero.

An explicit expression of the spectrum of $C_m$ can be obtained by observing that

$$C_m = FAF^*, \tag{5.8}$$

where $\Lambda$ is a diagonal matrix containing the eigenvalues $\mu_j$ of $C_m$ and $F$ is the Fourier matrix; see, e.g., [16]. Thus, from the expression of the eigenvalues of a circulant matrix, we have

$$\mu_j = -\frac{2ic}{\Delta x} \left( \sin \frac{2\pi j}{m} \right), \quad j = 0, \ldots, m-1.$$
i.e., the generating function computed in the points \( \theta_j = 2\pi j/m, j = 0, \ldots, m - 1 \), as usual.

It is worth to note that the Jacobian matrices for both the proposed model problems are normal and therefore can be diagonalized by unitary matrices. This feature is useful in order to use the bounds for the convergence of a Krylov accelerator which uses the preconditioners analyzed here; see [7, Theorems 3.1 and 3.2]. In particular, by applying the cited results, for the underlying problems we can immediately predict convergence in at most \( O(\log n) \) (preconditioned) iterations.

6. Numerical estimates and comparisons. We compare the results of zero- and of some first-order approximations presented in §4.2 with the eigenvalues computed by Matlab’s QR method for the model problems in section 5. We do not report plots generated by Simpson’s formula because very similar to those related to midpoint formula.

In all tests, unless specified otherwise, we consider \( s = m = 100, c = 1, T = 2\pi, x_{\text{max}} = \pi, t_0 = 0 \). The Jacobian matrix \( J \) is taken, in the light of Dahlquist’s hypothesis, as the smallest eigenvalue (in modulus) for each one of the model problems considered in the previous section. In the variable diffusion model problem, the diffusion function is of the form \( a(x) = x^k \), \( k > 0 \) integer, i.e., it has a zero in the origin of multiplicity \( k \). However, a similar eigenvalue distribution of the preconditioned and non preconditioned problems has been observed even in absence of zeros on the real axis for various functions like \( a(x) = x^k + \epsilon \), where \( \epsilon > 0 \) is a small constant, varying with \( O(m^{-1}) \). We stress that in both cases, eigenvalues of the Jacobian matrix (5.3) are negative but some of them go to zero as the space discretization gets refined. On the other hand, the same asymptotic behaviour holds for some nonzero eigenvalues of the Jacobian matrix (5.6) although the transport equation has constant coefficients.

Note that in all tests we get that even just zero-order approximations can give reasonable information on the qualitative behavior of the eigenvalues related to the smallest eigenvalues (in modulus) of the Jacobian matrix of the differential problem both in the nonpreconditioned and in the preconditioned cases, for variable and constant coefficients equations, provided that the mesh for the discretization in space is fine enough.

In order to emphasize the effect of the first-order approximations with respect to zero-order, just in Figure 6.1 (left) we use a more rough mesh with \( s = m = 20 \) for midpoint formula without using preconditioning.

It is surprising that for the transport equation (upper right plot in Figure 6.1), order 1 gives for some eigenvalues worse approximations than order 0: the “wrong” values come from roots \( \zeta(0) \) very close to the real axis (the same occurs for the derivatives \( \zeta'(0), \zeta''(0) \), etc.). This phenomenon is probably explained by observing that \( q \) is pure imaginary in this setting, so that in the power series \( \sum_{j=0}^{+\infty} \zeta^{(j)}(0)q^j \) just the even terms contribute to refine the real part, as well as the odd terms are only related to the imaginary part; this way the convergence radius of the series could be reduced, and the actual value of \( q \) could fall outside the region of analiticity. On the other hand, continuity still holds so that order 0 is always meaningful.

Our conjecture is confirmed by the lower right plot in Figure 6.1, where we have simply set \( c = 0.1 \): \( q \) has been divided by a factor of 10, and order 1 estimates become again more accurate than order 0.

For these moderate dimensions, every \( \zeta(0) \) has been computed through the Matlab function \texttt{roots}. If one is interested in locating the spectrum of much larger matrix...
Fig. 6.1. Midpoint (MP) method, smallest eigenvalue (in modulus) for (a) the diffusion equation with $a(x) = x^4$, $s = 20$, $m = 20$, (b) the transport equation, $s = 100$, $m = 100$, $c = 1$ and (c) the same equation with $c = 0.1$; $x=\text{order } 0$, $o=\text{eig}(P^{-1}M)$, $\square=\text{order } 1$ approximations.

Forces, we suggest the use of more efficient rootfinders specifically designed for sparse polynomials, like MPSolve proposed in [9].

REFERENCES

Fig. 6.2. Midpoint (MP) method with P-circulant preconditioning, smallest eigenvalue (in modulus) for (a) the diffusion equation with \( a(x) = x^4 \), \( s = 100 \), \( m = 100 \), and (b) the transport equation, \( s = 100 \), \( m = 100 \); \(+ = \) order 0, \( o = \text{eig}(P^{-1}M) \) approximations.

Fig. 6.3. Adams-Moulton with P-circulant preconditioning, smallest eigenvalue (in modulus) for (a) the diffusion equation with \( a(x) = x^4 \), \( s = 100 \), \( m = 100 \), and (b) the transport equation, \( s = 100 \), \( m = 100 \); \(+ = \) order 0, \( o = \text{eig}(P^{-1}M) \) approximations.