Preconditioning and conditioning of systems arising from Boundary Value Methods. *

F. Iavernaro † , D. Trigiante ‡

Abstract

The application of Boundary Value Methods to several classes of Differential Equations requires the solution of large dimension and sparse linear systems having (block) quasi-Toeplitz coefficient matrices. This has naturally suggested the use of Krylov subspace methods in combination with well known preconditioners suitable for Toeplitz matrices. However, the behaviour of such methods is closely related to the continuous problem (in the simplest case the system to be solved depends on a complex parameter) and some aspects need to be carefully studied in order to determine the effectiveness of these preconditioners and even their compatibility with some basic concepts in this area. Considerations about the choice of an optimal preconditioner are also presented.

Key words. Circulant preconditioners, Toeplitz-like matrices, Initial value problems, Linear multistep formulae, Boundary value methods.

AMS(MOS) subject classifications. 65F10, 65L05, 65L20, 65F15, 15A18

1 Introduction

Boundary Value Methods (BVMs) are a relatively recent class of methods for the numerical treatment of a wide variety of differential equations (IVPs, BVPs, DAEs, PDEs) (see for example [2, 3, 7, 10, 11, 12, 13, 17]). Their application transforms a continuous differential problem of dimension m into a discrete one of dimension mn, represented by a system of the form

$$(A_n \otimes I_m)Y - h(B_n \otimes I_m)F(Y) = \boldsymbol{\delta}. \tag{1}$$

The matrices A_n, B_n are square of dimension n, $\boldsymbol{\delta}$ is a known vector of length mn, I_m is the identity matrix of dimension m, $Y = [\mathbf{y}_1, \dots, \mathbf{y}_n]^T$ is a block vector of length mn whose components $\mathbf{y}_i \in \mathbb{R}^m$ are approximations to the true solution at given mesh points. The vector $F(Y) = [\mathbf{f}(\mathbf{y}_1), \dots, \mathbf{f}(\mathbf{y}_n)]^T$ contains the evaluations over $\mathbf{y}_i$ of a function $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^m$ which is typically defined by the continuous problem, while the step-length h depends on n and the time integration interval. For instance, choosing $\boldsymbol{\delta} = -\mathbf{a}_0 \otimes I_m \mathbf{y}_0 + h\mathbf{b}_0 \otimes I_m \mathbf{f}(t_0, \mathbf{y}_0)$ ($\mathbf{a}_0$ and $\mathbf{b}_0$ are vectors of length n), the system (1) may be considered as the discrete counterpart of the IVP

$$\left\{ \begin{array}{rcl} \mathbf{y}'(t) & = & \mathbf{f}(\mathbf{y}), \qquad t \in [t_0, t_0 + T], \\ \mathbf{y}(t_0) & = & \mathbf{y}_0, \end{array} \right. , \tag{2}$$

*Work supported by MURST and GNIM.

†Dipartimento di Matematica, Università di Bari, Via Orabona 4, I-70125 Bari (Italy), felix@dm.uniba.it, fax:+39 080 5460612

‡Dip. di Energetica, Università di Firenze, via C. Lombroso 6/17, I-50134 Firenze (Italy), trigiant@cesit1.unifi.it.

where now $h = T/n$. By definition BVMs give to the matrices A_n and B_n a banded quasi Toeplitz structure with bandwidth k independent of n. Applied to the problem (2) they take the form

$$A_n = \begin{pmatrix} \alpha_1^{(1)} & \alpha_2^{(1)} & \dots & \alpha_k^{(1)} & & & & \\ \vdots & \vdots & \vdots & \vdots & & & & \\ \alpha_1^{(k_1-1)} & \alpha_2^{(k_1-1)} & \dots & \alpha_k^{(k_1-1)} & & & & \\ \alpha_1 & \alpha_2 & \dots & \alpha_k & & & & \\ & \alpha_0 & \alpha_1 & \dots & \alpha_k & & & \\ & & \ddots & \ddots & \ddots & \ddots & & \\ & & & \alpha_0 & \alpha_1 & \dots & \alpha_k & \\ & & & \alpha_0^{(k_1)} & \alpha_1^{(k_1)} & \dots & \alpha_k^{(k_1)} & \\ & & & \vdots & \vdots & \vdots & \vdots & \\ & & & \alpha_0^{(k-1)} & \alpha_1^{(k-1)} & \dots & \alpha_k^{(k-1)} & \end{pmatrix}_{n\times n},$$

$$B_n = \begin{pmatrix} \beta_1^{(1)} & \beta_2^{(1)} & \dots & \beta_k^{(1)} & & & & \\ \vdots & \vdots & \vdots & \vdots & & & & \\ \beta_1^{(k_1-1)} & \beta_2^{(k_1-1)} & \dots & \beta_k^{(k_1-1)} & & & & \\ \beta_1 & \beta_2 & \dots & \beta_k & & & & \\ & \beta_0 & \beta_1 & \dots & \beta_k & & & \\ & & \ddots & \ddots & \ddots & \ddots & & \\ & & & \beta_0 & \beta_1 & \dots & \beta_k & \\ & & & \beta_0^{(k_1)} & \beta_1^{(k_1)} & \dots & \beta_k^{(k_1)} & \\ & & & \vdots & \vdots & \vdots & \vdots & \\ & & & \beta_0^{(k-1)} & \beta_1^{(k-1)} & \dots & \beta_k^{(k-1)} & \end{pmatrix}_{n\times n},$$

while $\mathbf{a}_0 = [\alpha_0^{(1)}, \alpha_0^{(k_1-1)}, \alpha_0, 0, \dots, 0]$ and $\mathbf{b}_0 = [\beta_0^{(1)}, \beta_0^{(k_1-1)}, \beta_0, 0, \dots, 0]$. The ith component of (1) is actually a k-step linear formula with k_1 initial and $k_2 = k - k_1$ final conditions. Its coefficients $\alpha_i,\ \beta_i,\ i = 0, \dots, k$ are determined imposing that $\mathbf{y}_i$ is an approximation of order p to the true solution $\mathbf{y}(t_i)$. In such a case p is also the order of the BVM, and the local truncation error assumes the form

$$\boldsymbol{\tau}(h) \equiv A_n\hat{Y} - hB_nF(\hat{Y}) + \mathbf{a}_0 \otimes I_m\mathbf{y}_0 - h\mathbf{b}_0 \otimes I_m\mathbf{f}(t_0, \mathbf{y}_0) = h^{p+1}G(\boldsymbol{\xi}), \tag{3}$$

where $\hat{Y} = [\mathbf{y}(t_1), \dots, \mathbf{y}(t_n)]^T$ is the vector of evaluations of the true solution $\mathbf{y}(t)$ of (2) at the internal mesh times t_i and $G(\boldsymbol{\xi}) = [c_1\mathbf{y}^{(p+1)}(\boldsymbol{\xi}_1), \dots, c_n\mathbf{y}^{(p+1)}(\boldsymbol{\xi}_n)]^T$, with c_i the error constant of the ith formula. The first $k_1 - 1$ and the final k_2 components of (1) are called respectively initial and final methods, and they cause the loss of Toeplitz structure which is instead conferred by the main method in the remaining rows. We remark that similar arguments are also valid for the other class of evolutionary problems to which BVMs have been applied.

The system (1) is nonlinear if $\mathbf{f}$ is so and its solution Y is therefore obtained as the limit of a sequence of vectors Y^k computed as solution of suitable linear systems. Here we suppose to linearize (1) in a neighborhood of its solution according to a simplified Newton iteration that gives rise to the scheme

$$(A_n \otimes I_m - hB_n \otimes J_k)\left(Y^{k+1} - Y^k\right) = G(Y^k), \tag{4}$$

where $G(Y^k) = \boldsymbol{\delta} - (A_n \otimes I_m)Y^k + h(B_n \otimes I_m)F(Y^k)$ and J_k is the Jacobian of $\mathbf{f}(y)$ evaluated at a suitable component of the current vector Y^k (in the simplest case J_k is independent of k). We observe that a similar system as (4) is to be solved when the continuous problem is linear and autonomous, namely $\mathbf{f}(y) = J\mathbf{y} + \mathbf{b}$. In this paper we are interested in analysing the properties of some Krylov subspace methods (see [14]) such as GMRES or BICGSTAB as applied to such linear systems subject to preconditioning and hence, until the convergence of the procedure (4) will be considered, it is reasonable to confine our analysis to linear problems only. The block quasi-Toeplitz and banded structure of the matrix $M_n = (A_n \otimes I_m - hB_n \otimes J)$, has suggested the use of preconditioners that normally work well when applied to Toeplitz or block Toeplitz matrices. In [8] the authors compare the efficiency of some preconditioning techniques showing, on the basis of their experiments, that good results, in terms of computational complexity, is achieved considering the block circulant preconditioner $S_n = C_n^A \otimes I_m - hC_n^B \otimes J$, where C_n^A and C_n^B are the Strang circulant preconditioners generated by the main method [15]:

$$C_n^A = \begin{pmatrix}
\alpha_{k_1} & \cdots & \alpha_k & & & & \alpha_0 & \cdots & \alpha_{k_1-1} \\
\vdots & \ddots & & \ddots & & & & \ddots & \vdots \\
\alpha_1 & & \ddots & & \ddots & & & & \alpha_0 \\
& \ddots & & \ddots & & \ddots & & & \\
& & \ddots & & \ddots & & \ddots & & \\
& & & \ddots & & \ddots & & \ddots & \\
\alpha_k & & & & \ddots & & \ddots & & \alpha_{k-1} \\
\vdots & \ddots & & & & \ddots & & \ddots & \vdots \\
\alpha_{k_1+1} & \cdots & \alpha_k & & & & \alpha_0 & \cdots & \alpha_{k_1}
\end{pmatrix}_{n \times n},$$

and analogously for C_n^B, with β_i instead of α_i (for simplicity, S_n will also be referred to as the Strang preconditioner).

For a convergent BVM ($p \geq 1$), one has from (3), $\sum_{i=0}^{k} \alpha_i = 0$ and hence $C_n^A \mathbf{e} = \mathbf{0}$, with $\mathbf{e} = [1, \ldots, 1]^T$. It follows that C_n^A is singular for all values of n and this causes the singularity of the preconditioner S_n when $det(J) = 0$. Indeed, if $\mathbf{x} \in \mathbb{R}^m - \{\mathbf{0}\}$ is such that $Jx = \mathbf{0}$, one also has $S_n X = \mathbf{0}$, with $X = \mathbf{e} \otimes \mathbf{x}$. It is not difficult to realize that this fact produces undesirable effects also when $det(J) \simeq 0$ due to a bad conditioning of the matrix S_n. Despite the good behaviour presented in [8] (which has favourably impressed the present authors), other elements must be considered that show how the use of S_n as preconditioner of M_n could be unappropriate in several cases. A comparison of preconditioners in terms of their conditioning is in our case indispensable but not new (see for example [16]); in [5, 6] the present problem is outlined and solved by P-circulant preconditioners.

In sections 2 and 3 we weigh up in more details the pros and cons of this strategy and propose (section 4) a modification in S_n that prevents a number of drawbacks. Lately (section 5), we also introduce a modification in the method itself that allow to the Strang preconditioner to work well when $det(J) = 0$. The properties of all these preconditioners are analysed to show their effectiveness.

2 Circulant preconditioners for BVMs

As seen for the Strang preconditioner, in general a circulant matrix is a Toeplitz matrix (that is its entries are constant along diagonals) for which the last entry in each row is the first one in the subsequent row. Multiplication of a circulant matrix of dimension n by a vector requires only $O(n \log(n))$ arithmetic operations if the Fast Fourier Transform (FFT) is performed. A circulant matrix $\mathcal{C}$ is in fact similar to a diagonal matrix D via a Fourier transformation matrix V. More precisely we have $\mathcal{C} = VDV^H$, where the diagonal matrix $D = diag(d_1, \ldots, d_n)$ and the Fourier matrix $V = \{v_{jk}\}$ have elements (i is the imaginary unit):

$$d_j = e^{\frac{2\pi i}{n} j}, \qquad v_{jk} = \frac{1}{\sqrt{n}} e^{\frac{2\pi i}{n} jk}, \qquad j,k = 0,\ldots,n-1. \tag{5}$$

A consequence of (5) is that

$$||\mathcal{C}|| = \max_j |d_j|, \qquad ||\mathcal{C}^{-1}|| = \frac{1}{\min\limits_j |d_j|} \qquad \text{and} \qquad \mu(\mathcal{C}) = \frac{\max_j |d_j|}{\min_j |d_j|},$$

where here and in the rest of the paper $||\cdot||$ will denote the 2-norm and $\mu(\mathcal{C}) = ||\mathcal{C}||\,||\mathcal{C}^{-1}||$ is the conditioning number (in 2-norm) of $\mathcal{C}$. Concerning the basic properties of circulant matrices that we will exploit during our discussion, we refer to [9].

To account for the choice of S_n as preconditioner of the matrix M_n, it is sufficient to observe that the preconditioned matrix P_n may be recast as

$$P_n \equiv S_n^{-1} M_n = I_{nm} + S_n^{-1} E_n,$$

with $E_n = M_n - S_n$. Since the rank of E_n is at most km, it follows that, for n large, most of the eigenvalues of P_n coincide with 1, which allows fast convergence of iterative methods like GMRES or BICGSTAB. However the other eigenvalues of P_n also play a role that cannot be neglected. For example, it is not possible to bound them inside a finite region of the complex plane independently of the function $\mathbf{f}$, a circumstance that may be critical when dealing with some classes of problems. To go into the question we will consider, here and in the rest of the paper, a class of BVMs called Generalized Backward Differentiation Formulae (GBDFs) over which a test problem will be performed and mathematical results will be derived. In passing, we emphasize that similar considerations may be easily extended to other classes of methods. The k-step GBDF is defined by choosing B_n as the identity matrix I_n, $\mathbf{b}_0 \equiv \mathbf{e}_1 = (1,0,\ldots,0)^T$, the index $k_1 = \nu$ according to the formula

$$\nu = \begin{cases} (k+2)/2, & \text{for even } k, \\ (k+1)/2, & \text{for odd } k, \end{cases} \tag{6}$$

and all the coefficients α_i in order that the formula has the highest possible order $p = k$.
As test problem we consider the linear pendulum system

$$\mathbf{y}' = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} \mathbf{y} \tag{7}$$

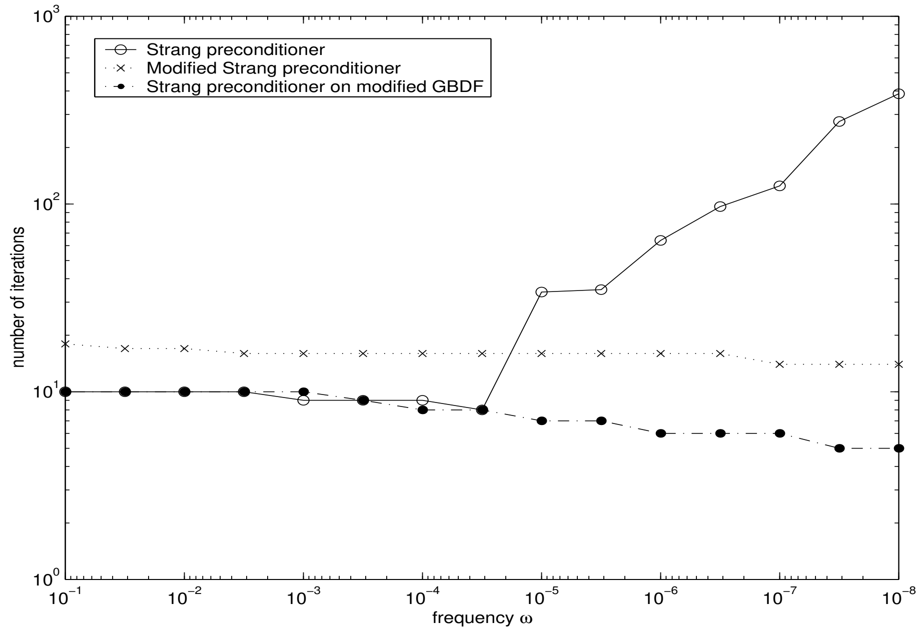
in the time interval $[0, 2\pi]$ and study the numerical solution obtained by the order 5 GBDF for different values of the frequency $\omega/(2\pi)$ and dimension $n = 100$ (the stepsize is therefore $h = 2\pi/100$).

The behaviour of this simple problem is also typical of more general dynamical systems in a neighborhood of marginally stable equilibrium points or even in a small time interval during which an equilibrium point loses or acquires stability due to the occurrence of a Hopf bifurcation.

The linear system originated by the BVM is solved by the GMRES routine of MATLAB using 10^{-12} as control of the relative residual and the Strang preconditioner S_n as input parameter. To state the inefficiency of S_n for small values of $|h \det(J)|$, we set $\omega = 10^{-m}$, $m = 1,2,\ldots,8$ and consider, for each value of the frequency, the number of iterations needed to get the numerical solution; this in fact is proportional to the overall cost of the algorithm (numbers of floating point operations).

Figure 1 shows an unexpected increase of the computational cost while ω decreases (the smaller the frequency the easier the numerical treatment of the problem should be). The reason of that may

Figure 1: Computational cost of GMRES applied to problem (7)



be understood looking at the three columns of Table 1 that report the conditioning numbers (in 2-norm) of the matrices M_n, S_n and P_n. It is seen that while the conditioning of the GBDF formula (the matrix M_n) stays constant independently of ω, the same is not true for the Strang preconditioner S_n and consequently for the preconditioned matrix P_n. They are indeed proportional to $1/\omega^2$ and as ω decreases, the use of finite precision arithmetic causes a drop in the convergence properties of GMRES and a loss of accuracy in the results. For instance the error is $1.5 \cdot 10^{-12}$ at $\omega = 10^{-1}$ and $5.7 \cdot 10^{-1}$ at $\omega = 10^{-8}$. Such problems also occur fixing a small value for ω and decreasing the stepsize $h = 2\pi/n$. In such a case the global error should decrease as $O(h^p)$ but once again, since $\mu(S_n)$ is proportional to n, loss of accuracy is experienced.

A modified Strang preconditioner $\bar{S}_n$, to be defined in the sequel, has also been used with the same set of parameters. The fifth and sixth columns of Table 1 tell us that the conditioning of $\bar{S}_n$ and $\bar{P}_n \equiv \bar{S}_n^{-1} M_n$ is comparable to that of M_n and what's more, they are independent of ω which make $\bar{S}_n$ suitable for small values of the frequency. A different but more appealing approach consists in modifying the GBDF formula via a similarity transformation (see section 5). The new matrix $\widehat{M}_n$ generates a nonsingular circulant matrix $\widehat{S}_n$ even if $h\det(J) = 0$. Figure 1 and the conditioning of $\widehat{M}_n$, $\widehat{S}_n$ and $\widehat{P}_n \equiv \widehat{S}_n^{-1} \widehat{M}_n$ in Table 1 prove the good behaviour of this technique.

In the next section the dependence of the conditioning of P_n on both the problem and the dimension n is analysed. It is custom in numerical analysis, to carry out the study of the discrete problem as applied to the scalar test equation

$$y' = \lambda y, \;\; \lambda \in \mathbb{C}. \qquad (8)$$

This approach reduces the complexity of calculus and may be easily generalized to the vector case in many cases of interest (for example when the system has a complete set of eigenfunctions).

Table 1: Comparison of conditioning numbers of the matrices $M_n,\ S_n,\ P_n,\ \bar{S}_n,\ \bar{P}_n,\ \widehat{M}_n,\ \widehat{S}_n,\ \widehat{P}_n$

ω	$\mu(M_n)$	$\mu(S_n)$	$\mu(P_n)$	$\mu(\bar{S}_n)$	$\mu(\bar{P}_n)$	$\mu(\widehat{M}_n)$	$\mu(\widehat{S}_n)$	$\mu(\widehat{P}_n)$
10^{-1}	$3.3 \cdot 10^{3}$	$2.6 \cdot 10^{3}$	$4.3 \cdot 10^{5}$	$7.6 \cdot 10^{2}$	$1.2 \cdot 10^{5}$	$1.7 \cdot 10^{3}$	$7.6 \cdot 10^{2}$	$6.2 \cdot 10^{4}$
$5 \cdot 10^{-2}$	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{4}$	$4.5 \cdot 10^{6}$	$1.0 \cdot 10^{3}$	$1.7 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$8.9 \cdot 10^{4}$
10^{-2}	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{5}$	$4.5 \cdot 10^{7}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
$5 \cdot 10^{-3}$	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{6}$	$4.5 \cdot 10^{8}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
10^{-3}	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{7}$	$4.5 \cdot 10^{9}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
$5 \cdot 10^{-4}$	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{8}$	$4.5 \cdot 10^{10}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
10^{-4}	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{9}$	$4.5 \cdot 10^{11}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
$5 \cdot 10^{-5}$	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{10}$	$4.5 \cdot 10^{12}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
10^{-5}	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{11}$	$4.5 \cdot 10^{13}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
$5 \cdot 10^{-6}$	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{12}$	$4.5 \cdot 10^{14}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
10^{-6}	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{13}$	$4.5 \cdot 10^{15}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
$5 \cdot 10^{-7}$	$3.4 \cdot 10^{3}$	$2.6 \cdot 10^{14}$	$4.5 \cdot 10^{16}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
10^{-7}	$3.4 \cdot 10^{3}$	$2.5 \cdot 10^{15}$	$4.3 \cdot 10^{17}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
$5 \cdot 10^{-8}$	$3.4 \cdot 10^{3}$	$1.6 \cdot 10^{17}$	$1.5 \cdot 10^{19}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$
10^{-8}	$3.4 \cdot 10^{3}$	$1.8 \cdot 10^{16}$	$1.5 \cdot 10^{21}$	$1.0 \cdot 10^{3}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{3}$	$1.0 \cdot 10^{3}$	$9.3 \cdot 10^{4}$

3 Preconditioning and conditioning

The discrete problem corresponding to (8) has dimension n and is now defined for the GBDFs by the matrix $M_n = A_n - h\lambda I_n$, with $h = T/n$. From the arbitrariness of λ, it follows that it is not a restriction to consider $T = 1$.

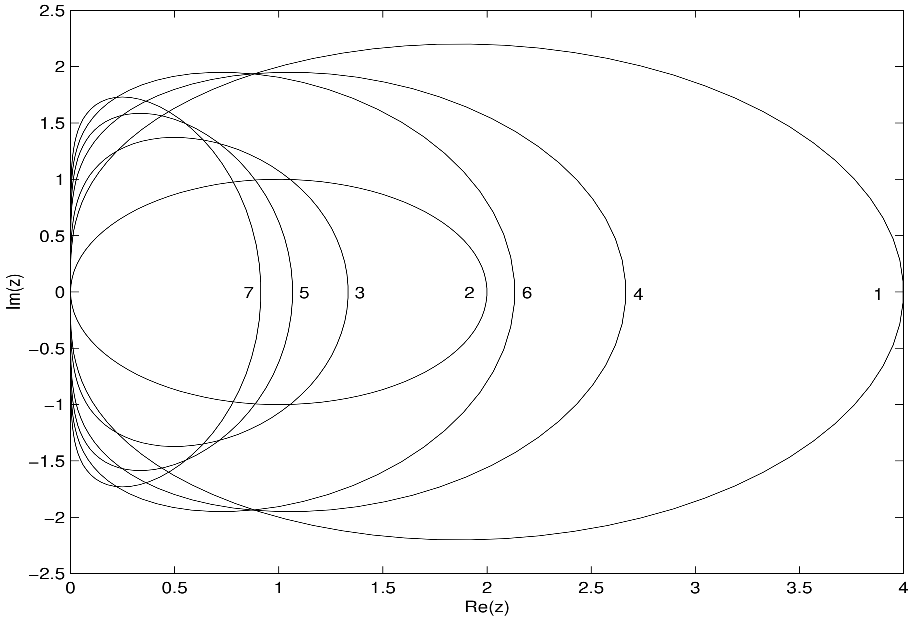
Of particular interest in the following is the main method of the GBDF, defined by the polynomial pair (ρ, σ):

$$\rho(z) = \sum_{j=0}^{k} \alpha_j z^j, \qquad \sigma(z) = z^{\nu}.$$

A link between the method and the algebraic properties of the preconditioner is in the function $g(z) = \rho(z)/\sigma(z)$ which generates the boundary locus of the former when evaluated at $z = e^{i\theta}$, $\theta \in [0, 2\pi]$ (i is the imaginary unit), and represents the symbol of the latter apart from a translation of size $-\lambda/n$ in the complex plane. Figure 2 reports the boundary loci of the main method of GBDFs up to the order 7. These curves also approximate the boundaries of the A-stability regions of the methods when n is large and state that GBDFs are indeed A-stable methods.

A necessary condition for A-stability is that all the eigenvalues of the matrix M_n have positive real part, when $\lambda \in \mathbb{C}^-$, where $\mathbb{C}^-$ is the left half of the complex plane. It follows that the solution of the equivalent method identified by the matrix $P_n = S_n^{-1} M_n$ will retain all the stability properties of the original one if none of the eigenvalues of S_n lies in $\mathbb{C}^-$ when $\lambda \in \mathbb{C}^-$. However the eigenvalues of the circulant matrix S_n are $g(e^{(2\pi i/n)j}) + \lambda/n$, $j = 0, \ldots, n-1$, and since $Re(g(e^{i\theta})) \geq 0$ they actually have nonnegative real part. Unfortunately the matrix S_n has $d_1 = \lambda/n$ as eigenvalue of minimum real part and consequently $d_1 = 0$ if $\lambda = 0$. Taking into account that the conditioning number of a circulant matrix is the ratio between the maximum and minimum modulus of its eigenvalues, it follows that $\mu(S_n)$ behaves at least as $O(n/\lambda)$. This means that, although both $\mu(M_n)$ and $\mu(S_n)$ are proportional to their dimension n, the latter cannot be bounded from below by a quantity independent of the problem: despite M_n, the preconditioner S_n may become ill conditioned if $\lambda \simeq 0$. In Figure 3, the location of the eigenvalues of M_n and S_n is displayed for $n = 80$, $\lambda = -1$ and order $p = 5$. We see that all the eigenvalues of M_n (except two) are inside the region delimited by

Figure 2: Boundary loci of the main formulae of GBDFs up to the order 7



the boundary locus and away from zero (see [4] for a characterization of the asymptotic spectra of banded quasi-Toeplitz matrices), whereas the eigenvalues of S_n place themselves on the boundary locus which in turns passes near zero for small values of λ/n.

From the relation

$$\mu(S_n) = \mu(M_n M_n^{-1} S_n) = \mu(M_n P_n^{-1}) \leq \mu(M_n)\mu(P_n)$$

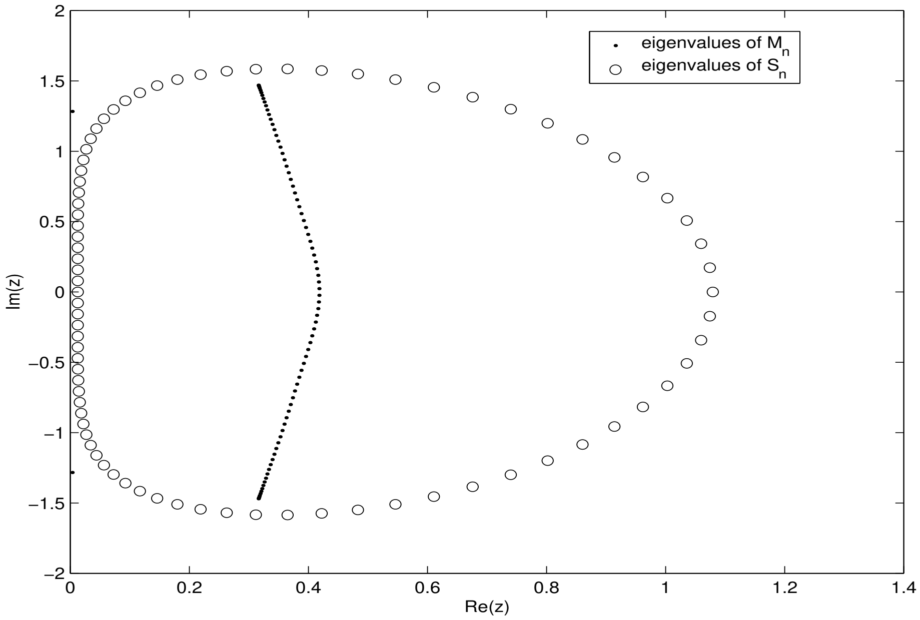
it follows that $\mu(P_n) \geq \mu(S_n)/\mu(M_n)$ and the same considerations hold true for the preconditioned matrix P_n. As seen above for the pendulum problem, the possible dangerous effects of that, are the loss of accuracy in the numerical computation and the weakening of the convergence properties of the iteration procedure used to determine the solution of the linear system.
As concerns nonlinear dynamical systems, the overall integration interval is usually partitioned into adjacent subintervals in each of which a scheme of the form (4), based on the Newton method, is performed to get the solution. It is then clear that analogous problems may be encountered in the convergence properties of (4) when $\det(J_k) \simeq 0$ and S_n is used as preconditioner.

For the sake of simplicity, we shall suppose in the following $\lambda \in]-\epsilon, 0]$, $\epsilon > 0$. The restriction to the real case makes the calculation easier and, using a continuity argument, it describes as well the behaviour of the complex problem in a neighborhood of zero, which is the primary objective of the present analysis. In the rest of the paper the matrix M_n will therefore assume the expression

$$M_n = A_n + \frac{|\lambda|}{n} I_n, \tag{9}$$

and since $C_n^B = I_n$, to simplify the notation C_n will stand for C_n^A.

Figure 3: Eigenvalues of M_n and S_n of the order 5 GBDF



4 A modified Strang preconditioner

We focus now our attention to the conditioning of the preconditioned matrix $P_n = S_n^{-1}M_n$. The final purpose is to introduce a family of preconditioners depending on a real parameter γ in order that for the new preconditoned matrix $\bar{P}_n(\gamma)$ the inequality

$$\mu(\bar{P}_n(\gamma)) \leq c\mu(M_n) \tag{10}$$

may hold true with the constant $c \geq 1$ independent of n and of moderate size. To begin, we introduce the family of preconditioners

$$\mathcal{S}_n(\gamma) = C_n + \gamma/n I_n,$$

and the associated preconditioned matrices

$$\mathcal{P}_n(\gamma) = (\mathcal{S}_n(\gamma)^{-1})M_n,$$

which will be related later on to the family $\bar{P}_n(\gamma)$.

Lemma 4.1 *For the main method (ρ, σ) of a GBDF of order $p \geq 1$, the functions $\varphi(\theta) = Re(g(e^{i\theta}))$ and $\xi(\theta) = Im(g(e^{i\theta}))$ satisfy:*

$$(a)\ \varphi(\theta) = \begin{cases} O(\theta^{p+2}), & \text{if } p \text{ is even,} \\ O(\theta^{p+1}), & \text{if } p \text{ is odd;} \end{cases}$$

$$(b)\ \xi(\theta) = \begin{cases} \theta + O(\theta^{p+1}), & \text{if } p \text{ is even,} \\ \theta + O(\theta^{p+2}), & \text{if } p \text{ is odd.} \end{cases}$$

Proof. The order conditions for the main method of a p order GBDF are:

$$\sum_{j=0}^{k} j^s \alpha_j = s\nu^{s-1}, \qquad s = 0, \ldots, p, \tag{11}$$

where ν is as in (6). For $s = 0, 1, \ldots,$ define the quantities

$$c_s = \sum_{j=0}^{k} (j-\nu)^s \alpha_j.$$

The $p+1$ independent conditions (11) are seen to be equivalent to the following ones:

$$c_0 = 0, \quad c_1 = 1, \quad c_s = 0, \quad s = 2, \ldots, p. \tag{12}$$

Indeed, by direct comparison, $c_0 = 0$ and $c_1 = 1$ are equivalent to (11) for $s = 0, 1$. Consider now $s \in \{2, \ldots, p\}$. We have

$$\begin{aligned} c_s = & \sum_{j=0}^{k} (j-\nu)^s \alpha_j = \sum_{j=0}^{k} \alpha_j \sum_{t=0}^{s} (-1)^{s-t} \binom{s}{t} j^t \nu^{s-t} = \sum_{t=0}^{s} (-1)^{s-t} \nu^{s-t} \binom{s}{t} \sum_{j=0}^{k} j^t \alpha_j \\ = & \sum_{t=0}^{s} (-1)^{s-t} \nu^{s-t} \binom{s}{t} t\nu^{t-1} = \sum_{t=0}^{s} (-1)^{s-t} \binom{s}{t} t\nu^{s-1} = \nu^{s-1} \sum_{t=1}^{s} (-1)^{s-t} \binom{s}{t} t. \end{aligned}$$

Exploiting the equality

$$\binom{s}{t} t = \binom{s-1}{t-1} s,$$

it follows that

$$c_s = s\nu^{s-1} \sum_{t=1}^{s} (-1)^{s-t} \binom{s-1}{t-1} = s\nu^{s-1} \sum_{t=0}^{s} (-1)^{s-t-1} \binom{s-1}{t} = s\nu^{s-1}(1-1)^{s-1} = 0.$$

The assertion follows considering that the Taylor expansion of $\varphi(\theta)$ and $\xi(\theta)$ in a neighborhood of zero are respectively

$$\varphi(\theta) = \sum_{j=0}^{k} \alpha_j cos(j-\nu)\theta = \sum_{j=0}^{k} \alpha_j \sum_{n=0}^{\infty} (-1)^n \frac{(j-\nu)^{2n}}{(2n)!} \theta^{2n} = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} c_{2n} \theta^{2n},$$

and

$$\xi(\theta) = \sum_{j=0}^{k} \alpha_j sin(j-\nu)\theta = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} c_{2n+1} \theta^{2n+1}.$$

□

LEMMA 4.2 *Consider the family of circulant matrices $\mathcal{S}_n(\gamma) = C_n + \gamma/nI_n$, with γ a positive parameter and denote by d_j the eigenvalues of $\mathcal{S}_n(\gamma)$ defined as*

$$d_j = \frac{\gamma}{n} + g(e^{\frac{2\pi i}{n} j}), \qquad j = 0, \ldots, n-1. \tag{13}$$

For n sufficiently large, the following estimation holds true:

$$\frac{1}{n^2} \sum_{j=0}^{n} \frac{1}{|d_j|^2} \simeq \frac{i}{2\pi\gamma}[\Psi(-\frac{i}{2\pi}\gamma) - \Psi(\frac{i}{2\pi}\gamma)] - \frac{1}{\gamma^2}, \tag{14}$$

where Ψ is the digamma function (i is the imaginary unit).

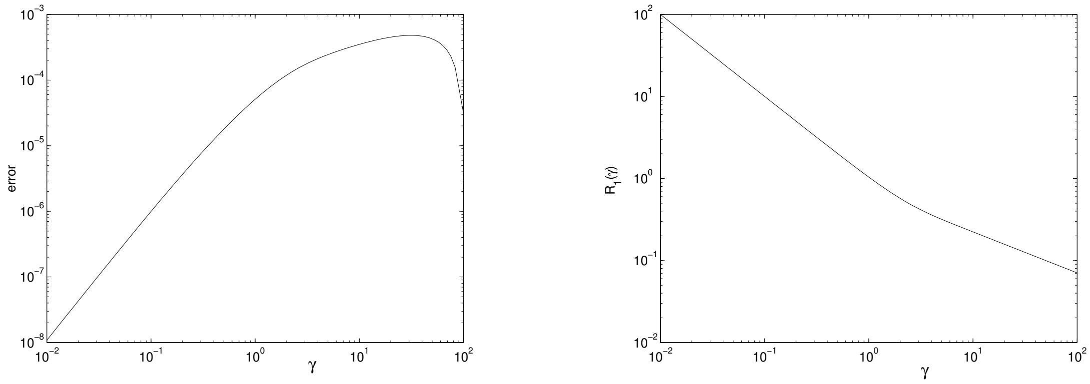
Proof. The increase of the dimension n reduces the shift of the boundary locus of the method (ρ, σ) of a term γ/n and, as Figures 2, 3 and formula (13) suggest, gives rise to the accumulation of a number of eigenvalues, proportional to n, into a neighborhood of the origin. All of these eigenvalues (say $\pm d_i$, $i = 1, \ldots, c(n)$, by the symmetry of the distribution), will provide the significative contribution to the sum in the left hand side of (14) and therefore neglecting the remaining terms will not produce a consistent error. The neighborhood may be chosen so that in the expressions (a) and (b) of Lemma 4.1 we can also neglect the higher order terms. Under these assumptions we have

$$\begin{aligned} \sum_{j=0}^{n} \frac{1}{|d_j|^2} &= \sum_{j=0}^{n} \frac{1}{(\frac{\gamma}{n} + \varphi(\frac{2\pi i}{n}j))^2 + \xi^2(\frac{2\pi i}{n}j)} \\ &\simeq 2\sum_{j=0}^{c(n)} \frac{1}{(\frac{\gamma}{n})^2 + (\frac{2\pi}{n})^2 j^2} - \frac{n^2}{\gamma^2} \leq 2n^2 \sum_{j=0}^{\infty} \frac{1}{\gamma^2 + (2\pi)^2 j^2} - \frac{n^2}{\gamma^2}. \end{aligned} \tag{15}$$

The assertion follows noting that the last series converges to half the first term in the right hand side of (14). □

To check for the reliability of the estimation (14) we report in Figure 4 the relative error of the computed values that the expressions in its left and right hand side assume in a wide range of values of γ of interest.

Figure 4: Error in the estimation (14) (left), and a plot of the function $R(\gamma)$ (right)



A plot of the function

$$R(\gamma) \equiv \left(\frac{i}{2\pi\gamma}[\Psi(-\frac{i}{2\pi}\gamma) - \Psi(\frac{i}{2\pi}\gamma)] - \frac{1}{\gamma^2} \right)^{1/2}$$

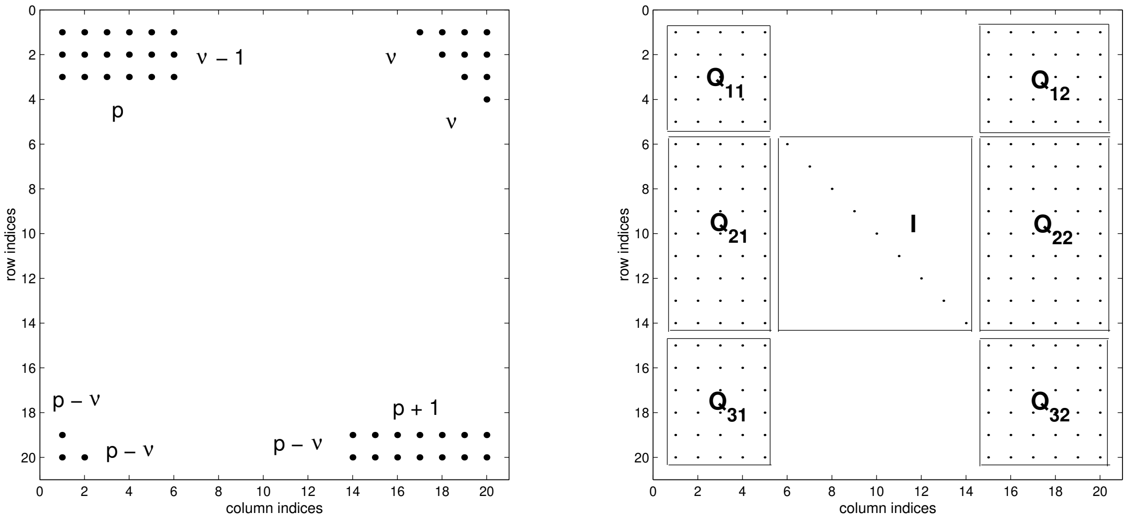
is also reported. This function is strictly decreasing for $\gamma \geq 0$ and its range is $(0, \infty)$; furthermore the principal part in its Laurent expansion is $1/\gamma$. It will be used in the sequel to obtain the main result of this approach.

In the proof of the main result, expressed by Theorem 4.5, the structure of the error matrix $E_n = A_n - C_n$ plays an important role. For a GBDF of order p, E_n has rank p and its nonzero elements are located in the four corners as sketched in Figure 5 for the case $p = 6$ and $n = 20$. It is easy to realize that the 2-norm of E_n remains constant for $n \geq 2p + 1$; such constant has been computed and reported in Table 2 for the GBDFs up to the order 9. Multiplication of a square matrix W_n of dimension n by E_n satisfies the property $W_n E_n = W_n^* E_n$, where W_n^* has all zero columns apart from the first ν and the last $p - \nu$ ones that agree with those of W_n. The asterisk upon a generic square matrix will assume hereafter the same meaning as for W_n^*.

Table 2: Norm of the matrix E_n for $p = 1, \ldots, 9$ and $n \geq 2p + 1$

p	1	2	3	4	5	6	7	8	9
$\|\|E_n\|\|$	1	2.62	2.90	4.17	7.59	11.69	19.76	32.51	55.55

Figure 5: Structures of the error matrix E_n (left) and of a matrix $Q \in \mathcal{H}$ (right)



For reasons that will be clear in the sequel, we are interested in investigating some properties of the set of square matrices of dimension n

$$\mathcal{H} = \{Q | Q = I + W,\ W \text{ has zero columns except the first } r \text{ and the last } s \text{ ones}\},$$

defined by the integers r and s, $r + s \leq n$. A picture of how an element of $\mathcal{H}$ looks like is in Figure 5 in the case $n = 20,\ r = 5,\ s = 6$. It is easy to verify that the product of two matrices in $\mathcal{H}$ belongs to $\mathcal{H}$. This also holds true for the inverse, as the following lemma states.

LEMMA 4.3 *The inverse of a nonsingular matrix $Q \in \mathcal{H}$ belongs to $\mathcal{H}$.*

Proof. We refer to the partition of Q by means of the blocks Q_{ij} given in Figure 5 (to simplify the notation subscripts describing the dimension of the blocks have been omitted), and consider a matrix $H \in \mathcal{H}$:

$$Q = \begin{pmatrix} Q_{11} & 0 & Q_{12} \\ Q_{21} & I & Q_{22} \\ Q_{31} & 0 & Q_{32} \end{pmatrix}_{n \times n}, \qquad H = \begin{pmatrix} H_{11} & 0 & H_{12} \\ H_{21} & I & H_{22} \\ H_{31} & 0 & H_{32} \end{pmatrix}_{n \times n}.$$

The condition $QH = I$ is expressed in terms of the blocks Q_{ij} and H_{ij} by means of the following six equations:

$$\left\{\begin{array}{l} Q_{11}H_{11}+Q_{12}H_{31}=I, \\ Q_{11}H_{12}+Q_{12}H_{32}=0, \\ Q_{21}H_{11}+H_{21}+Q_{22}H_{31}=0, \\ Q_{21}H_{12}+H_{22}+Q_{22}H_{32}=0, \\ Q_{31}H_{11}+Q_{32}H_{31}=0, \\ Q_{31}H_{12}+Q_{32}H_{32}=I. \end{array}\right. \tag{16}$$

The first two and the last two equations may be recast as $\tilde{Q}\tilde{H}=I$, where

$$\tilde{Q}=\left(\begin{array}{cc} Q_{11} & Q_{12} \\ Q_{31} & Q_{32} \end{array}\right), \qquad \text{and} \qquad \tilde{H}=\left(\begin{array}{cc} H_{11} & H_{12} \\ H_{31} & H_{32} \end{array}\right).$$

Since $\det(\tilde{Q})=\det(Q)$, from the invertibility of Q we conclude that the blocks H_{11}, H_{12}, H_{31} and H_{32} are uniquely determined from the relation $\tilde{H}=\tilde{Q}^{-1}$. The remaining blocks H_{21} and H_{22}, comes from the third and the fourth equations in (16). □

We recall (see for example [7]) that a polynomial $p(z)=\sum_{j=0}^{k} z^k$ is of type (s,u,l) (s, u and l are integers such that $k=s+u+l$), if it has s zeros with modulus smaller than 1, u zeros with unit modulus and l zeros with modulus larger than 1.

Lemma 4.4 *Consider the matrix $M_n=A_n+|\lambda|/nI_n$. Constants $\eta>0$ and $0<\zeta<1$ independent of n and λ exist such that the following two statements hold true:*

(a) The matrix $|M_n^{-1}|$, whose entries are the absolute values of the corresponding ones in M_n^{-1} satisfies the componentwise bound

$$|M_n^{-1}|\leq \eta(I_n+\Omega_n+\Delta_n^T), \tag{17}$$

where

$$\Omega_n=\left(\begin{array}{cccc} 0 & 0 & \dots & 0 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 1 & \ddots & 1 & 0 \end{array}\right)_{n\times n}, \qquad \Delta_n=\left(\begin{array}{cccc} 0 & 0 & \dots & 0 \\ \zeta & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \zeta^{n-1} & \ddots & \zeta & 0 \end{array}\right)_{n\times n};$$

(b) $||M_n^{-1}||_\infty$, $||M_n^{-1}||_1$, $||M_n^{-1}|| \leq \eta n$.

Proof. (a). The starting point is Theorem 4 in [1] which states the analogous result for Toeplitz matrices. More explicitly, let T_n be the Toeplitz matrix with associated symbol $g(z)+|\lambda|/n$. From A-stability of GBDFs, its characteristic polynomial $p(z)=z^\nu g(z)$ (of degree k), turns out to be of type $(\nu,0,k-\nu)$ for $|\lambda|\neq 0$ and of type $(\nu-1,1,k-\nu)$ if $|\lambda|=0$. As a result of the above mentioned theorem, a bound, uniform with respect to λ, of the form (17) holds true for T_n, and is attained at $\lambda=0$. The residual matrix $R_n=M_n-T_n$ differs from the zero matrix in the $(\nu-1)\times p$ upper left block and in the $(p-\nu)\times(p+1)$ lower right one. In terms of T_n and R_n, the matrix M_n^{-1} is recast as

$$M_n^{-1}=(I_n+T_n^{-1}R_n)^{-1}T_n^{-1}. \tag{18}$$

The matrix $Q_n\equiv(I_n+T_n^{-1}R_n)$ belongs to $\mathcal{H}$; using Lemma 4.3, we have also that $H_n\equiv(I_n+T_n^{-1}R_n)^{-1}\in\mathcal{H}$. Hence, considering (18), the assertion will follow if we prove that the entries of H_n are bounded with respect to n and λ. This is true for the matrix Q_n, as a direct consequence of its definition. As concerns H_n, we will see this in the simpler case $Q_{12} = 0$; this does not represent a loss of generality since actually $Q_{12} = O(\sigma^n)$ for some $\sigma \in]0, 1[$ (the elements of Q_{12} are in fact combinations of entries of the upper right corresponding block in T_n^{-1}) and a continuity argument may be considered. Exploiting the results presented in [1], it is possible to deduce that the blocks Q_{11}, Q_{31} and Q_{32} essentially remains the same independently of the dimension n (actually they are exponentially convergent as n tends to infinity). Consequently the blocks $H_{11} = Q_{11}^{-1}$, $H_{12} = 0$, $H_{32} = Q_{32}^{-1}$ and $H_{31} = -Q_{32}^{-1}Q_{31}Q_{11}^{-1}$ also have bounded coefficients. Finally, equations three and four in (16) lead to the same conclusions for the coefficients in the blocks H_{21} and H_{22}.

(b). The bound of the norm of the inverse of M_n is a consequence of (17):

$$
\begin{aligned}
&||M_n^{-1}||_\infty = ||\ |M_n^{-1}|\ ||_\infty \le \eta n, \\
&||M_n^{-1}||_1 = ||\ |M_n^{-1}|\ ||_1 \le \eta n, \\
&||M_n^{-1}|| \le \sqrt{||M_n^{-1}||_\infty\ ||M_n^{-1}||_1} \le \eta n.
\end{aligned}
$$

□

A consequence of this lemma is that, as mentioned in section 3, the matrix M_n is weakly well conditioned, uniformly with respect to λ, that is $\mu(M_n) \le cn$, with $c > 0$ independent of n and λ. This is in general not the case when the Strang preconditioner is used, unless some adjustment is introduced. The following theorem, that reports the main result, is in this direction.

THEOREM 4.5 *The conditioning of the preconditioned matrix $\mathcal{P}_n(\gamma) = \mathcal{S}_n(\gamma)^{-1}M_n$, with $\mathcal{S}_n(\gamma) = C_n + \gamma/nI_n$, satisfies:*

$$
\mu(\mathcal{P}_n(\gamma)) \le p^2\eta||E_n||^2\ R(\gamma)n \tag{19}
$$

where η is a positive constant independent of n and γ.

Proof.

An upper bound for the quantities $||\mathcal{S}_n(\gamma)^{-1}M_n||$ and $||M_n^{-1}\mathcal{S}_n(\gamma)||$ is derived in the two following steps.

Step 1. From

$$
\mathcal{S}_n(\gamma)^{-1}M_n = \mathcal{S}_n(\gamma)^{-1}\left(\mathcal{S}_n(\gamma) + E_n + \frac{|\lambda| - \gamma}{n}I_n\right) = \mathcal{S}_n(\gamma)^{-1}E_n + I_n + \frac{|\lambda| - \gamma}{n}\mathcal{S}_n(\gamma)^{-1},
$$

we deduce

$$
||\mathcal{S}_n(\gamma)^{-1}M_n|| \le ||\mathcal{S}_n(\gamma)^{-1}E_n|| + ||I_n + \frac{|\lambda| - \gamma}{n}\mathcal{S}_n(\gamma)^{-1}||.
$$

We analyse separately the two terms in the right hand side. Introducing the decomposition $C_n = V_nD_nV_n^H$ in the first one yields

$$
\begin{aligned}
||\mathcal{S}_n(\gamma)^{-1}E_n|| \le\ & ||(C_n + \tfrac{\gamma}{n}I_n)^{-1}E_n|| = ||V_n(D_n + \tfrac{\gamma}{n}I_n)^{-1}V_n^HE_n|| = ||(D_n + \tfrac{\gamma}{n}I_n)^{-1}(V_n^H)^*E_n|| \\
\le\ & ||E_n||\ ||(D_n + \tfrac{\gamma}{n}I_n)^{-1}(V_n^H)^*|| = ||E_n||\ \max_{||\mathbf{y}||=1} ||(D_n + \frac{\gamma}{n}I_n)^{-1}(V_n^H)^*\mathbf{y}|| \\
=\ & ||E_n||\ \max_{||\mathbf{y}||=1} ||y_1\mathbf{z}_1 + \ldots + y_\nu\mathbf{z}_\nu + y_{n-p+\nu+1}\mathbf{z}_{n-p+\nu+1} + \ldots + y_n\mathbf{z}_n|| \\
\le\ & p||E_n||\ \max\{||\mathbf{z}_1||, \ldots, ||\mathbf{z}_\nu||, ||\mathbf{z}_{n-p+\nu+1}||, \ldots, ||\mathbf{z}_n||\},
\end{aligned}
$$

where, $\mathbf{z}_i,\ i \in \{1,\ldots,\nu,n-p+\nu+1,\ldots,n\}$ are the nonzero columns of $(D_n + \frac{\gamma}{n} I_n)^{-1}(V_n^H)^*$. From (5) we deduce that these columns have constant norm

$$||\mathbf{z}_i|| = \frac{1}{\sqrt{n}} \left(\sum_{j=0}^{n} \frac{1}{|d_j|^2} \right)^{1/2}.$$

Hence Lemma 4.2 leads to

$$||\mathcal{S}_n(\gamma)^{-1} E_n|| \leq \quad p||E_n||\ R(\gamma)\sqrt{n}.$$

For the second term we have

$$||I_n + \frac{|\lambda| - \gamma}{n} \mathcal{S}_n(\gamma)^{-1}|| \leq 1 + \frac{||\lambda| - \gamma|}{n} ||\mathcal{S}_n(\gamma)^{-1}|| = 1 + \frac{||\lambda| - \gamma|}{\gamma}. \tag{20}$$

Since this term is bounded with respect to n, and γ is a fixed positive constant, for large n we can assume

$$||\mathcal{S}_n(\gamma)^{-1} M_n|| \leq p||E_n||\ R(\gamma)\sqrt{n}. \tag{21}$$

Step 2. Observe that

$$M_n^{-1}\mathcal{S}_n(\gamma) = -M_n^{-1}E_n + I_n + \frac{\gamma - |\lambda|}{n} M_n^{-1},$$

and consequently

$$||M_n^{-1}\mathcal{S}_n(\gamma)|| \leq ||M_n^{-1}E_n|| + ||I_n + \frac{\gamma - |\lambda|}{n} M_n^{-1}||. \tag{22}$$

Proceeding analogously as in step 1, we obtain

$$\begin{aligned} ||M_n^{-1}E_n|| = ||(M_n^{-1})^* E_n|| \leq &\quad ||(M_n^{-1})^*||\ ||E_n|| \\ \leq &\quad ||E_n||\ \max\{||\mathbf{w}_1||,\ldots,||\mathbf{w}_\nu||,||\mathbf{w}_{n-p+\nu+1}||,\ldots,||\mathbf{w}_n||\}. \end{aligned}$$

where now $\mathbf{w}_i$ are the non-null columns of $(M_n^{-1})^*$. Exploiting point (a) of Lemma 4.4, we deduce that

$$||M_n^{-1}E_n|| \leq p\eta||E_n||\ \sqrt{n}.$$

Point (b) of Lemma 4.4 is invoked to state that

$$||I_n + \frac{\gamma - |\lambda|}{n} M_n^{-1}|| \leq 1 + |\gamma - |\lambda||\eta, \tag{23}$$

thus for large n we can write

$$||M_n^{-1}\mathcal{S}_n(\gamma)|| \leq p\eta||E_n||\sqrt{n}. \tag{24}$$

The bound (19) is finally derived by combining formulae (21) and (24). □

Two applications of this result are now discussed in points A_1 and A_2; they may be considered as corollaries of Theorem 4.5 corresponding to two different choices of the parameter γ.

A_1. The Strang preconditioner is obtained choosing $\gamma = |\lambda|$ and the associated preconditioned matrix P_n, defined in section 2, is consequently $P_n = \mathcal{P}_n(|\lambda|)$ (observe that in this case the bounds (20) and (23) become independent of γ and η). Considering (19) and the behaviour of the function $R(|\lambda|)$ we conclude that this preconditioner performs well when $|\lambda|$ is far off zero. On the contrary, for small values of $|\lambda|$ we have $R(|\lambda|) \simeq 1/|\lambda|$ and the right hand side of (19) reduces to $O(1/|\lambda|)$ showing that, as actually happens in the applications, the conditioning of P_n may arbitrarily increase.

A_2. Consider now the family of preconditioners $\bar{S}_n(\gamma) = \mathcal{S}_n(|\lambda| + \gamma)$, $\gamma \geq 0$; the corresponding preconditioned matrices are $\bar{P}_n(\gamma) = \mathcal{P}_n(|\lambda| + \gamma)$. The choice $\gamma = 0$ leads back to the case reported in A_1. We are rather interested in comparing the conditioning numbers of the two matrices $\bar{P}_n(\gamma)$ and M_n and in particular to solve the inequality (10) which, considering (19) is certainly fulfilled for all values of γ that satisfy

$$R(\gamma + |\lambda|) \leq \frac{c}{p^2\eta||E_n||^2}\frac{\mu(M_n)}{n}. \tag{25}$$

Taking into account that $\mu(M_n) = O(n)$ and that $R(\gamma)$ strictly decreases to zero as γ tends to infinity, we see that (25) has solutions $\gamma \in [\bar{\gamma}(c, |\lambda|), \infty[$ for some $\bar{\gamma}(c, |\lambda|) > 0$.

The approach presented in A_2 proves that a control of the conditioning during the preconditioning procedure is in principle possible, but two question about the setting up of the technique must be addressed:

(i) the dependence of $\bar{\gamma}(c, |\lambda|)$ on $|\lambda|$ should be removed because for more general problems of the form (4), it requires information about the location of the eigenvalues of J;

(ii) the determination of $\bar{\gamma}(c, |\lambda|)$ is impracticable, unless the quantities η and $\mu(M_n)/n$ are estimated in some way.

The problem pointed out in (i) is easily overcome restricting the analysis to the case $\lambda = 0$ and extending, by continuity, the results to the interval $|\lambda| \in]-\epsilon, 0[$ for some positive ϵ sufficiently small. As seen above, the choice of the Strang preconditioner represents the worst possible case when the conditioning of the problem is considered. Of particular interest is therefore the number $\gamma^*(c) = \bar{\gamma}(c, 0)$, which is the solution of the equation

$$R(\gamma) = \frac{c}{p^2\eta||E_n||^2}\frac{\mu(A_n)}{n}. \tag{26}$$

The question raised in point (ii) is conveniently solved as follows. Instead of searching approximations of η and $\mu(M_n)/n$, we go back to step2 of Theorem (4.5). From (22), where now $M_n = A_n$, we can assume

$$||A_n^{-1}\mathcal{S}_n(\gamma)|| \leq ||A_n^{-1}E_n||,$$

and therefore, without going into the inspection of this last term, we can simply conclude that

$$\mu(\mathcal{P}_n(\gamma)) \leq p||E_n||\ R(\gamma)||A_n^{-1}E_n||n.$$

Now observe that the quantity

$$\chi(p) = \frac{\sqrt{n}||A_n^{-1}E_n||}{\mu(A_n)}, \tag{27}$$

only depends on the particular GBDF used (namely on p) and therefore may be estimated and tabulated (see Table 3).

Table 3: Values of $\chi(p)$ for $p = 1, \ldots, 9$

p	1	2	3	4	5	6	7	8	9
$\chi(p) \simeq$	0.79	2.3	1.7	2.4	1.8	2.8	1.3	1.4	0.50

On the basis of these considerations, equation (26) may be replaced by the following one

$$R(\gamma) = \frac{c}{p\chi(p)||E_n||}, \tag{28}$$

where now all the quantities displayed in the right hand side are available. It is useful to notice that $\gamma^*(c)$, as solution of (28), turns out to be decreasing with respect to c (see also Figure 4).

The numbers c and $\gamma^*(c)$, which are in a one to one correspondence, are related respectively to conditioning and preconditioning properties of our problem and a pertinent question is what the best choice of c (or equivalently $\gamma^*(c)$) should be, namely how to define our *optimal* preconditioner.

If the preconditioner $\bar{S}_n(\gamma)$ is demanded to optimize the clustering rate of the eigenvalues of the preconditioned matrix $\bar{P}_n(\gamma)$, then $\gamma^*(c) = 0$ (the Strang preconditioner) would be the best choice because $\bar{P}_n(0)$ has almost all of its eigenvalues exactly centered in 1. However $\gamma^*(c) = 0$, gives also $c = \infty$ and the control of the conditioning over the preconditioned matrix is completely loss.

On the other hand, from the point of view of the conditioning, the best choice would be $c = 1$, because if so, the conditioning of the preconditioned system would not become worse with respect to that of the original one. In such a case however the cluster around 1 becomes wide and a slowing down in the convergence speed is noticed.

However if, as it should be, the optimality is linked to the property of the preconditioner of minimizing the algorithm cost, we deduce that we cannot define optimal neither the Strang preconditioner nor the preconditioner $\bar{S}(\gamma^*(1))$; the best choice is rather obtained for an intermediate value of $\gamma \in]0, \gamma^*(1)[$. We experienced that choosing $c = 10^j$, with small value of the integer j, provides the right compromise. In the example of section 2 for example we chose $\gamma^*(c) = 1$ and, considering Tables 2 and 3, the corresponding value of c is

$$c = 5 \cdot \chi(5) \cdot R(1) \cdot 7.59 \simeq 71,$$

which is the maximum amplification factor of $\mu(M_n)$ for small values of the eigenvalues. Indeed from Table 1, we get

$$\mu(M_n) \simeq 3.4 \cdot 10^3, \qquad c \cdot \mu(M_n) \simeq 2.4 \cdot 10^5, \qquad \mu(\bar{P}_n) = 1.8 \cdot 10^5,$$

in agreement with the obtained results of the test problem in section 2.

5 The Strang preconditioner on a modified GBDF

The ill conditioning of the Strang preconditioner for $\lambda = 0$, is generated by the consistency condition $\sum_{i=0}^{k} \alpha_i = 0$. In the previous section we overcame the problem introducing a modification in the preconditioner. An alternative is to modify the coefficients that define the method. In details we consider hereafter the approach used in [10] to deduce the global contractivity of GBDFs. In that paper, the authors proved that

$$\min_{-\pi \leq \theta < \pi} Re\left(\frac{\rho(e^{\frac{1}{n}+i\theta})}{\sigma(e^{\frac{1}{n}+i\theta})}\right) \geq \frac{s}{n}, \tag{29}$$

with s a positive constant independent of n (see Lemma 1.2 and Theorem 5.1). The modified symbol $\hat{g}(z) = g(e^{1/n}z)$ is generated by the matrix $\widehat{A}_n$ defined by a similarity transformation of A_n:

$$\widehat{A}_n = L_n A_n (L_n)^{-1}, \qquad L_n = \begin{pmatrix} e^{-\frac{1}{n}} & & & \\ & e^{-\frac{2}{n}} & & \\ & & \ddots & \\ & & & e^{-1} \end{pmatrix}_{n\times n}.$$

In details, the matrix L_n operates as follows: the original linear system (4) for a GBDF ($B_n = I_n$), is equivalent to

$$(L_n \otimes I_m)(A_n \otimes I_m - hI_n \otimes J)(L_n \otimes I_m)^{-1}(L_n \otimes I_m)(Y^{k+1} - Y^k) = (L_n \otimes I_m)G(Y^k). \tag{30}$$

Introducing the change of variables $Z^k = (L_n \otimes I_m)Y^k$, and considering that

$$(L_n \otimes I_m)(A_n \otimes I_m)(L_n \otimes I_m)^{-1} = (L_n A_n (L_n)^{-1}) \otimes I_m = \widehat{A} \otimes I_m$$

and

$$(L_n \otimes I_m)(I_n \otimes J)(L_n \otimes I_m)^{-1} = (L_n I_n (L_n)^{-1}) \otimes J = I_n \otimes J,$$

the equation (30) becomes

$$(\widehat{A}_n \otimes I_m - hI_n \otimes I_m)(Z^{k+1} - Z^k) = (L_n \otimes I_m)G((L_n \otimes I_m)^{-1}Z^k).$$

The matrix $\widehat{M}_n \equiv (\widehat{A}_n \otimes I_m - hI_n \otimes I_m)$ is therefore similar to M_n via the similarity transformation $(L_n \otimes I_m)$. Taking into account that $\mu(L_n) < e$, the conditioning of $\widehat{M}_n$ is close to that of M_n. However, contrary to what happens for M_n, the Strang circulant preconditioner $\widehat{S}_n$ associated to $\widehat{M}_n$ preserves a good conditioning. The lower bound (29) states in fact that even if $h \det(J) = 0$, $\widehat{S}_n$ is nonsingular and weakly well conditioned. As for the modified Strang preconditioner, we are now interested in studying the conditioning of the preconditioned matrix $\widehat{P}_n = \widehat{S}_n^{-1}\widehat{M}_n$ in the scalar case ($J = -|\lambda|$). The novelty that will make $\mu(\widehat{P}_n)$ independent of $|\lambda|$ is expressed in the following lemma that is the analogue of Lemma 4.1 for the function $\hat{g}(z)$.

Lemma 5.1 *For the modified GBDF of order $p \geq 1$, the functions $\hat{\varphi}(\rho, \theta) = Re(g(e^{\rho+i\theta}))$ and $\hat{\xi}(\rho, \theta) = Im(g(\rho + e^{i\theta}))$, $\rho, \theta \in I\!R$, satisfy in a neighborhood of the origin:*

(a) $\hat{\varphi}(\rho, \theta) = \rho +$ *higher order terms;*

(b) $\hat{\xi}(\rho, \theta) = \theta +$ *higher order terms.*

Proof. The Taylor expansion of $\hat{\varphi}(\rho,\theta)$ and $\hat{\xi}(\rho,\theta)$ about $(0,0)$ are respectively

$$\begin{aligned}\hat{\varphi}(\rho,\theta) = & \sum_{j=0}^{k} \alpha_j e^{(j-\nu)\rho} cos(j-\nu)\theta = \sum_{j=0}^{k}\alpha_j \sum_{s=0}^{\infty}\frac{(j-\nu)^s}{s!}\rho^s \sum_{n=0}^{\infty}(-1)^n\frac{(j-\nu)^{2n}}{(2n)!}\theta^{2n} \\ = & \sum_{n=0}^{\infty}\frac{(-1)^n}{(2n)!}\sum_{s=0}^{\infty}\frac{1}{s!}c_{2n+s}\theta^{2n}\rho^s,\end{aligned}$$

and

$$\hat{\xi}(\rho,\theta) = \sum_{j=0}^{k}\alpha_j e^{(j-\nu)\rho} sin(j-\nu)\theta = \sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)!}\sum_{s=0}^{\infty}\frac{1}{s!}c_{2n+s+1}\theta^{2n+1}\rho^s,$$

where the coefficients c_j were defined in Lemma 4.1. From this expressions and (12) we get the assertion. □

Define now $\widehat{C}_n$ as the Strang preconditioner of $\widehat{A}_n$. The proof of Lemma 4.2 remains the same for the circulant matrix $\widehat{S}_n = \widehat{C}_n + |\lambda|/nI_n$ except that, due to (a) of the previous lemma, the new term $\hat{\varphi}(1/n, 2\pi ij/n)$ can not longer be neglected as $\varphi(2\pi ij/n)$ in (15), rather it must be replaced by $1/n$. As a consequence, the result expressed by Lemma 4.2 holds as well in this case provided that γ is replaced by $1+|\lambda|$. To conclude, observing that $||\widehat{M}_n^{-1}|| \le e||M_n^{-1}||$ and denoting by $\widehat{E}_n = \widehat{M}_n - \widehat{S}_n$, we can reformulate Theorem 4.5 as follows.

THEOREM 5.2 *The conditioning of the preconditioned matrix* $\widehat{P}_n = \widehat{S}_n^{-1}\widehat{M}_n$ *satisfies:*

$$\mu(\widehat{P}_n) \le ep^2\eta||\widehat{E}_n||^2\ R(1+|\lambda|)n \tag{31}$$

where η is a positive constant independent of n and $|\lambda|$.

The bound (31) proves that $\widehat{P}_n$ is weakly well conditioned and non increasing for $\lambda \simeq 0$ (in particular for $|\lambda| = 0$ we get $R(1) \simeq 1.08$).

Acknowledgements

The authors wish to acknowledge P. Amodio and F. Mazzia for their very helpful suggestions that improved the paper.

References

[1] Amodio, P. and Brugnano, L. The conditioning of Toeplitz band matrices. *Math. Comput. Modelling*, **23 (10)**, (1996), 29–42.

[2] Amodio P. and Brugnano L. ParalleloGAM: a parallel code for ODEs. *Appl. Numer. Math.*, **28**, (1998), 95–106.

[3] Amodio, P. and Mazzia, F. Numerical solution of differential algebraic equations and computation of consistent initial/boundary conditions. *J. Comput. Appl. Math.*, **87 (1)**, (1997), 135–146.

[4] Beam, R.M. and Warming, R.F. The asymptotic spectra of Banded Toeplitz and quasi-Toeplitz matrices. *SIAM J. Sci. Comput.*, **14** (1993), 971–1006.

[5] Bertaccini, D. P-circulant preconditioners and the systems of ODE codes. In: Kincaid, D.R. and Elster, A.C. (Eds.), *Iterative Methods in Scientific computation II, IMACS Series in Computational and Applied Mathematics.* IMACS,New Brunswick, NJ, 1999, 179–193.

[6] Bertaccini, D. A circulant preconditioner for the systems of LMF-based ODE codes. *submitted.*

[7] Brugnano, L. and Trigiante, D. *Solving ODEs by Linear Multistep Initial and Boundary Value Methods.* Gordon & Breach, Amsterdam, 1998.

[8] Chan, R.H., Ng, M.K. and Jin, X.Q. Circulant Preconditioners for Solving Ordinary Differential Equations. In: D. Bini, D., Tyrtyshnikov, E. and Yalamov, P. (Eds.), *Structured Matrices: Recent Developments in Theory and Computation*, Nova Science Pub. Inc., 2000.

[9] Davis, P.J. *Circulant matrices*, John Wiley & Sons, New York, 1979.

[10] Iavernaro, F. and Mazzia, F. Convergence and stability of multistep methods solving nonlinear initial value problems. *SIAM J. Sci. Comput.*, **18**, (1997), 270–285.

[11] Iavernaro, F. and Mazzia, F. Solving ordinary differential equations by block Boundary Value Methods: properties and implementation techniques. *Appl. Num. Math.*, **28**, Issue 2-4,(1998) 107–126.

[12] Iavernaro, F. and Mazzia, F. Block-Boundary Value Methods for the solution of Ordinary Differential Equations. *Siam J. Sci. Comput.*, **21** (1999), 323–339.

[13] Mazzia, A., Mazzia, F. and Trigiante, D. Boundary value Methods for PDEs. In: Sivasundaram, S. (Eds.), *proceedings of the First International Conference on Nonlinear Problems in Aviation and Aerospace (Daytona Beach, FL, 1996).* Embry-Riddle Aeronaut. Univ. Press, Daytona Beach, FL, 1996, 421–436.

[14] Saad, Y. *Iterative Methods for Sparse Linear Systems.* PWS publishing company, Boston, MA, 1995.

[15] Strang, G. A proposal for Toeplitz matrix calculations. *Stud. Appl. Math.*, **74** (1986), 171–176.

[16] Strela, V.V. and Tyrtyshnikov, E.E. Which circulant preconditioner is better? *Math. Comp. Vol.*, **65** (1996), 137–150.

[17] Trigiante, D. Multipoint methods for linear Hamiltonian systems. Advances in Nonlinear Dynamics, series “Stability and Control: Theory Methods and Applications”, Gordon and Breach, Reading, UK, (1997), 335–348.