A look-ahead algorithm for the solution of block Toeplitz systems

Marc Van Barel and Adhemar Bultheel

Report TW 224, March 1995
(Revised: May, September 1995)



Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

A look-ahead algorithm for the solution of block Toeplitz systems

Marc Van Barel and Adhemar Bultheel

Report TW 224, March 1995
(Revised: May, September 1995)

Department of Computer Science, K.U. Leuven

Abstract

In this paper, we give a weakly stable algorithm to solve a block Toeplitz system of linear equations. If the "look-ahead" steps taken to compute the parameters of the inversion formula for the block Toeplitz matrix are small compared to the order n of the matrix, the algorithm requires $O(n^2)$ floating point operations. The parameters of the inversion formula are interpreted and computed in a recursive way as solutions of certain interpolation problems given the formal Laurent series based on the data of the block Toeplitz matrix.

Keywords : (block) Toeplitz matrix, look-ahead algorithm, roundoff error analysis.
AMS(MOS) Classification : Primary : 65F05, Secondary : 41A05, 65G05.

A look-ahead algorithm for the solution of block Toeplitz systems*

Marc Van Barel and Adhemar Bultheel†

March 1995, Revised: May, September 1995

Abstract

In this paper, we give a weakly stable algorithm to solve a block Toeplitz system of linear equations. If the "look-ahead" steps taken to compute the parameters of the inversion formula for the block Toeplitz matrix are small compared to the order n of the matrix, the algorithm requires $O(n^2)$ floating point operations. The parameters of the inversion formula are interpreted and computed in a recursive way as solutions of certain interpolation problems given the formal Laurent series based on the data of the block Toeplitz matrix.

1 Introduction

In this paper, we look for the solution of a system of linear equations

$$T_{m,n} V_{m,n} = Y_{m,n} \tag{1}$$

where $T_{m,n}$ is a block Toeplitz coefficient matrix, i.e. the elements are equal along each block diagonal

$$\begin{aligned} T_{m,n} &= [t_{i-j}]_{i=0,1,\ldots,m-1}^{j=0,1,\ldots,n-1} \\ &= \begin{bmatrix} t_0 & t_{-1} & \cdots & t_{-n+2} & t_{-n+1} \\ t_1 & t_0 & \cdots & t_{-n+3} & t_{-n+2} \\ \vdots & \vdots & & \vdots & \vdots \\ t_{m-1} & t_{m-2} & \cdots & t_{m-n+1} & t_{m-n} \end{bmatrix} \end{aligned}$$

where the entries t_i are $p \times q$ real or complex matrices.

The classical algorithms to solve (1) exploit the Toeplitz structure and require only $O(m^2)$ operations, compared to $O(m^3)$ for general linear systems. The flow of these methods is determined by the exact singularity of the square principal block submatrices of $T_{m,n}$. They compute the solutions for successive nonsingular principal submatrices. However, in finite-precision arithmetic it turns out that also the principal submatrices which are nonsingular but ill-conditioned should be avoided. For scalar Toeplitz matrices, several of these look-ahead algorithms have been designed [5, 4, 8, 7, 10, 17]. Even superfast, i.e. requiring $O(m \log^2 m)$ operations, look-ahead algorithms where developed [13, 14, 16]. Also for Hankel matrices, i.e. matrices having the same entries on each antidiagonal, such look-ahead schemes have been constructed [3, 9, 1]. Note that the Hankel case is simpler than the Toeplitz case because Hankel matrices are symmetric and when translating these problems into series, the

*This research was supported by the National Fund for Scientific Research (NFWO), project Lanczos, grant #2.0042.93 and by the Human Capital and Mobility project ROLLS of the European Community under contract ERBCHRXCT930416.

†Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200 A, B-3001 Heverlee, Belgium. e-mail: Marc.VanBarel@cs.kuleuven.ac.be

Hankel problem can be translated using formal power series while for the Toeplitz case we need formal Laurent series. For the Hankel case, [3] gives the error analysis. As far as we know, such a detailed error analysis was not done for the look-ahead procedures to solve the (scalar) Toeplitz problem. In this paper, we want to design a look-ahead algorithm solving the block Toeplitz system (1) and motivate the look-ahead strategy by a detailed error analysis.

In [2], Cabay et al. give the error analysis for a "look-ahead" algorithm computing multi-dimensional Padé systems and inverting generalized Sylvester matrices. In a similar way, we want to give the error analysis for the inversion of block Toeplitz matrices and the solution of the related interpolation problems. I.e., we follow the same path of reasoning as in [2] but the content will be adapted to block Toeplitz instead of generalized Sylvester matrices. This change of content turns out to be nontrivial, especially for finding the corresponding interpolation problems and the updating formulas for the parameters of the inversion formula of the block Toeplitz matrix. To get a similar flow of the error analysis, we derive the "look-ahead" algorithm in a specific way. Other implementations of the updating formulas are possible leading to another error analysis.

Section 2 shows the inversion formula which we will use for a block Toeplitz matrix. Section 3 shows how some parameters of this inversion formula can be interpreted as solutions of an interpolation problem. A so-called ***basis matrix*** allows to represent all solutions of this interpolation problem in a compact way. In Section 4, we develop an updating procedure for this basis matrix. Section 5 describes the dual interpolation problem which has to be solved to get the remaining parameters of the inversion formula. In the error analysis, we need a connection between the basis matrices of the original and the dual interpolation problem. This duality is given for so-called *extended* basis matrices in Section 6. Hence, instead of working with the original basis matrices, we derive in Section 7 an updating algorithm for these extended basis matrices. Before we give the error analysis in Section 9, Section 8 defines the norms. Based on the computed inversion parameters, Section 10 gives formulas for the inverse of the block Toeplitz matrix. Section 11 shows that the "look-ahead" algorithm computes the basis matrix in a weakly stable way. In Section 12, we indicate that using the inversion formula to compute the solution of a block Toeplitz system of linear equations is weakly stable. Section 13 gives some experimental results. The look-ahead strategy requires the user to input a threshold value. In Section 14, we design another look-ahead strategy where the user has only to give the maximum size of the look-ahead step taken. The experimental results, however, are less accurate compared to the results given by Freund and Zha in [8] for scalar Toeplitz matrices. It turns out that using the inversion formula leads to this loss of accuracy. Therefore, we constructed the updating formula of Section 15 for the solution of the block Toeplitz system of linear equations. Note that the inversion formula can be used for iterative refinement as was done in [14]. In Section 16, we refine the look-ahead strategy to get experimental results which are now comparable to those of Freund and Zha for scalar matrices. We also apply the procedure to 100 block Toeplitz matrices generating acceptable results and indicating that this last look-ahead procedure together with the recursive updating of the solution is weakly stable. However, in some rare cases there is some greater loss of accuracy. For the moment, the mechanism behind this is not clear. In Section 17, we give the error analysis for the recursive updating formula. We tried to incorporate these results into another look-ahead criterion not only taking into account the error computing the basis matrices but also the error using the recursive updating. In some cases this gives a better result but generally the result is a little bit worse compared to the look-ahead criterion of Section 16.

2 Inversion formula

In this section, we give an inversion formula for a block Toeplitz matrix. The parameters of this inversion formula will be interpreted as solutions of certain interpolation problems in Section 3 and 5. We repeat the result of Gohberg and Shalom [11, Th. 6.1].

Theorem 2.1 *Let T be a square block Toeplitz matrix of the form $T_{m,n} = T = [t_{i-j}]_{i=0,1,\ldots,m-1}^{j=0,1,\ldots,n-1}$ with entries $t_k \in \mathbb{R}^{p\times q}$ and $mp = nq$.*[1]

The matrix T is invertible if (and only if) there exist solutions to the following systems of matrix equations:

$$\begin{bmatrix} t_0 & t_{-1} & \cdots & t_{-n+1} & t_{-n} \\ t_1 & t_0 & \cdots & t_{-n+2} & t_{-n+1} \\ \vdots & \vdots & & \vdots & \vdots \\ t_{m-1} & t_{m-2} & \cdots & t_{m-n} & t_{m-n-1} \end{bmatrix} \begin{bmatrix} x_0 & -z_0 \\ x_1 & -z_1 \\ \vdots & \vdots \\ x_{n-1} & -z_{n-1} \\ 0 & I_q \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{2}$$

and

$$\begin{bmatrix} 0 & u_{m-1} & \cdots & u_1 & u_0 \\ I_p & -w_{m-1} & \cdots & -w_1 & -w_0 \end{bmatrix} \begin{bmatrix} t_{-1} & t_{-2} & \cdots & t_{-n+1} & t_{-n} \\ t_0 & t_{-1} & \cdots & t_{-n+2} & t_{-n+1} \\ \vdots & \vdots & & \vdots & \vdots \\ t_{m-1} & t_{m-2} & \cdots & t_{m-n+1} & t_{m-n} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & I_q \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

with $x_j \in \mathbb{R}^{q\times p}$, $z_j \in \mathbb{R}^{q\times q}$, $u_j \in \mathbb{R}^{q\times p}$ and $w_j \in \mathbb{R}^{p\times p}$. Note that $z_0u_0 = x_0w_0$. Moreover, T^{-1} admits the following representations:

$$\begin{aligned} T^{-1} &= L_{(n,\eta)}(x_0, x_1, \ldots, x_{n-1})U_{(\eta,m)}(I_p, -w_{m-1}, -w_{m-2}, \ldots, -w_1) + \\ &\quad L_{(n,\eta)}(z_0, z_1, \ldots, z_{n-1})U_{(\eta,m)}(0, u_{m-1}, u_{m-2}, \ldots, u_1) \end{aligned}$$

and

$$\begin{aligned} T^{-1} &= U_{(n,\eta)}(I_q, -z_{n-1}, -z_{n-2}, \ldots, -z_1)L_{(\eta,m)}(u_0, u_1, \ldots, u_{m-1}) + \\ &\quad U_{(n,\eta)}(0, x_{n-1}, x_{n-2}, \ldots, x_1)L_{(\eta,m)}(w_0, w_1, \ldots, w_{m-1}) \end{aligned}$$

for $\eta = \min(m, n)$. For $k \geq l$,

$$L_{(k,l)}(g_0, g_1, \ldots, g_{k-1}) = \begin{bmatrix} g_0 & 0 & \cdots & 0 \\ g_1 & g_0 & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & & g_0 \\ \vdots & \vdots & & \vdots \\ g_{k-1} & g_{k-2} & \cdots & g_{k-l} \end{bmatrix}.$$

For $k \leq l$,

$$L_{(k,l)}(g_0, g_1, \ldots, g_{l-1}) = \begin{bmatrix} g_{l-k} & \cdots & g_0 & 0 & \cdots & 0 \\ g_{l-k+1} & \cdots & g_1 & g_0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ g_{l-1} & \cdots & g_{k-1} & g_{k-1} & \cdots & g_0 \end{bmatrix}.$$

Similarly, for $k \geq l$, $U_{(k,l)}(g_0, g_1, \ldots, g_{k-1}) = L_{(l,k)}(g_0^T, g_1^T, \ldots, g_{k-1}^T)^T$ and for $k \leq l$, $U_{(k,l)}(g_0, g_1, \ldots, g_{l-1}) = L_{(l,k)}(g_0^T, g_1^T, \ldots, g_{l-1}^T)^T$.

[1] The results developed in this report for real numbers can be easily adapted to complex numbers.

proof. This theorem can be obtained as a special result of Theorem 10.1 when the computations are considered to be done with infinite precision, i.e. $\theta_{VIII} = 0$. □

The couple (m, n) with m and n natural numbers is called a *normal point* if $mp = nq$ and if the corresponding block Toeplitz matrix $T_{m,n}$ (square) is nonsingular. In the next section, we show that the parameters x_i, z_i of the inversion formula are connected to a right interpolation problem in such a normal point (m, n).

3 Connection to an interpolation problem

Consider the formal Laurent series $T(z) = \sum_{k=-\infty}^{+\infty} t_k z^k$. In this section, we assume the values of m and n to be fixed.

Definition 3.1 (right interpolation problem) *For each natural number α, the set $\mathcal{S}_\alpha$ is defined as the set of all vector polynomials $q(z) \in \mathbb{R}[z]^{q\times 1}$ satisfying the interpolation condition*

$$T(z)q(z) = O_-(z^{-n+\alpha-1}) + O_+(z^m) \tag{3}$$

and the degree condition $\deg q \leq \alpha$.

Note that if $m + n - \alpha \leq 0$ the interpolation condition is trivially satisfied. In this section, we shall show that for a given α each element of $\mathcal{S}_\alpha$ can be written in a unique way as a linear polynomial combination of the columns of two matrix polynomials. Therefore, we need the following lemma.

Lemma 3.1 *The block Toeplitz matrix $T_{m,n}$ is nonsingular iff all nontrivial solutions of S_n have a nonzero highest degree coefficient.*

proof. The vector q of coefficients of a nontrivial solution $q(z)$ of S_n having degree $< n$ should satisfy the set of linear homogeneous equations: $T_{m,n}q = 0$. Hence, such a nontrivial $q(z)$ exists iff $T_{m,n}$ is singular. □

Theorem 3.2 *If the block Toeplitz matrix T is nonsingular, each element $q(z)$ of $\mathcal{S}_\alpha$ can be written in a unique way as*

$$q(z) = [X(z) \quad Z(z)] \begin{bmatrix} q_x(z) \\ q_z(z) \end{bmatrix}$$

with $\deg q_x \leq \min(\alpha - n - 1, m - 1)$ *and* $\deg q_z \leq \alpha - n$. $Z(z) = -z_0 - z_1 z - \cdots - z_{n-1}z^{n-1} + I_q z^n$ *and* $X(z) = x_0 + x_1 z + \cdots + x_{n-1}z^{n-1}$, x_k, z_k *as in Theorem 2.1.*

proof. The proof is by induction on α. It is clear that the only element of $\mathcal{S}_\alpha$ for $\alpha < n$ is the zero vector polynomial. For $\alpha = n$, the columns of $Z(z)$ are elements of $\mathcal{S}_\alpha$. Because its highest degree coefficients are the columns of I_q, hence linearly independent, any element $q(z)$ of $\mathcal{S}_\alpha$ can be written in a unique way as $q(z) = Z(z)q_z + q'(z)$ with $\deg q_z \leq 0$ (hence, q_z is constant) and $\deg q' < n$. However, because $\mathcal{S}_{n-1}$ only contains the zero vector polynomial and $q' \in \mathcal{S}_{n-1}$, q' is equal to zero. This proves the theorem for $\alpha = n$. Suppose the theorem is true for $\alpha < m + n$. Because $T(z)X(z) = O_-(z^{-1}) + I_p + O_+(z^m)$, we can write every element $q(z)$ of $\mathcal{S}_{m+n}$ in a unique way as $q(z) = z^m Z(z)q_{z,m} + z^{m-1}X(z)q_{x,m-1} + q'(z)$ by making the highest degree coefficient of q and the coefficient of z^{m-1} in $T(z)q(z)$ zero. In other words, $q'(z) \in \mathcal{S}_{\alpha-1}$. Hence, we can write $q(z)$ in a unique way as $q(z) = Z(z)(q_z(z) + z^m q_{z,m}) + X(z)(q_x(z) + z^{m-1}q_{x,m-1})$. This proves the theorem for $\alpha = m + n$.
For $\alpha \geq m + n$, the elements q of $\mathcal{S}_\alpha$ only have to satisfy the degree condition $\deg q \leq \alpha$. Hence, any element $q(z)$ of $\mathcal{S}_\alpha$ with $\alpha > m + n$ can be written in a unique way as $q(z) = z^{\alpha-n}Z(z)q_{z,\alpha-n} + q'(z)$ with $q'(z) \in \mathcal{S}_{\alpha-1}$. Applying the same reasoning as above this proves the theorem for $\alpha > m + n$. □

The polynomial matrix $B_{m,n}(z) = [X(z) \;\; Z(z)]$ is called the normalized basis matrix in the normal point (m, n). It is also possible to construct a basis matrix in points (m, n) which are not normal but we do not need this result in the sequel. It is clear that the polynomial matrix

$$B_{m,n}(z)\Gamma_R = [X(z) \;\; Z(z)] \begin{bmatrix} \Gamma_x & 0 \\ 0 & \Gamma_z \end{bmatrix}$$

with $\Gamma_R = \mathrm{diag}(\Gamma_x, \Gamma_z)$ and Γ_x and Γ_z nonsingular diagonal matrices, is also a basis matrix. In the sequel, we will denote this more general form of a basis matrix also by $B_{m,n}(z)$.

4 Updating procedure

In the previous section, we kept the indices m and n fixed. To get a *square* block Toeplitz matrix, m and n should satisfy $pm = qn$. In this section, we consider all possible normal points $(m^{(\sigma)}, n^{(\sigma)})$, i.e. such that the (nested) block Toeplitz matrices $T_{m^{(\sigma)},n^{(\sigma)}}$ based on the data $t_i \in \mathbb{R}^{p\times q}$ are square and nonsingular. If the basis matrix $B_{m^{(\sigma)},n^{(\sigma)}} = B_{m,n}$ in the normal point (m, n) is known, we construct a method to compute the basis matrix $B_{m^{(\sigma+k)},n^{(\sigma+k)}} = B_{M,N}(z)$ in one of the next normal points (M, N). We use the following notation:

$$T(z)[X(z) \;\; Z(z)] = [R_x^-(z) + R_x^+(z) \;\; R_z^-(z) + R_z^+(z)]$$

with $R_x^-(z) = R_{x,0}^- + R_{x,1}^- z^{-1} + R_{x,2}^- z^{-2} + \cdots$ with $R_{x,0}^- = \Gamma_x^{(m,n)}$, called the $(-)$-*residual series* for $X(z)$, and $R_x^+(z) = R_{x,1}^+ z^m + R_{x,2}^+ z^{m+1} + R_{x,3}^+ z^{m+2} + \cdots$, called the $(+)$-residual series for $X(z)$. Similarly for $R_z^-(z)$ and $R_z^+(z)$ with $R_{z,0}^- = 0$. For an $R_v^-(z) = O_-(z^0)$ and an $R_v^+(z) = O_+(z^m)$ series, we define the block Toeplitz matrix $S_{v,r,c}$ as

$$S_{v,r,c} = \begin{bmatrix} R_{v,0}^- & R_{v,1}^- & \cdots & R_{v,c-1}^- \\ 0 & R_{v,0}^- & \cdots & R_{v,c-2}^- \\ \vdots & & \ddots & \vdots \\ 0 & & & \vdots \\ R_{v,1}^+ & 0 & \cdots & \vdots \\ R_{v,2}^+ & R_{v,1}^+ & \ddots & \vdots \\ \vdots & \vdots & & \vdots \\ R_{v,r}^+ & R_{v,r-1}^+ & \cdots & R_{v,r-c+1}^+ \end{bmatrix}$$

where the number of block rows is $r + c$ if $c < m$ and $m + r$ if $c \geq m$. The matrix $S_{m,n}$ is defined as

$$S_{m,n} = [S_{x,\delta_m,\delta} \;\; S_{z,\delta_m,\delta_n}] \tag{4}$$

with $\delta_m = M - m$, $\delta_n = N - n$ and $\delta = \min(\delta_n, m)$. Note that $S_{m,n}$ is square iff $pM = qN$. We consider the following system of linear equations

$$S_{m,n} \begin{bmatrix} a_0 & c_0 \\ a_1 & c_1 \\ \vdots & \vdots \\ a_{\delta-1} & c_{\delta-1} \\ b_0 & d_0 \\ b_1 & d_1 \\ \vdots & \vdots \\ b_{\delta_n-1} & d_{\delta_n-1} \end{bmatrix} = \begin{bmatrix} I_p & -R_{z,\delta_n}^- \\ 0 & -R_{z,\delta_n-1}^- \\ \vdots & \vdots \\ 0 & -R_{z,\delta_m-\delta_n-1}^+ \\ 0 & -R_{z,\delta_m-\delta_n}^+ \end{bmatrix} \begin{bmatrix} \Gamma_x^{(M,N)} & 0 \\ 0 & \left(\Gamma_z^{(m,n)}\right)^{-1} \left(\Gamma_z^{(M,N)}\right) \end{bmatrix}. \tag{5}$$

We define the polynomial matrices A, B, C, D as

$$\begin{aligned} A(z) &= a_0 + a_1 z + \cdots a_{\delta-1} z^{\delta-1} \\ C(z) &= c_0 + c_1 z + \cdots c_{\delta-1} z^{\delta-1} \\ B(z) &= b_0 + b_1 z + \cdots b_{\delta_n-1} z^{\delta_n-1} \\ D(z) &= d_0 + d_1 z + \cdots d_{\delta_n-1} z^{\delta_n-1} + \left(\Gamma_z^{(m,n)}\right)^{-1} \left(\Gamma_z^{(M,N)}\right) z^{\delta_n}. \end{aligned}$$

Theorem 4.1 *If $B_{m,n}(z)$ is a basis matrix in the normal point (m,n) then (M,N) is also a normal point iff the matrix $S_{m,n}$ is nonsingular. Moreover, if $B_{m,n} = [X(z) \;\; Z(z)]$, then*

$$[X'(z) \;\; Z'(z)] = [X(z) \;\; Z(z)] \begin{bmatrix} A(z) & C(z) \\ B(z) & D(z) \end{bmatrix} \tag{6}$$

is a basis matrix $B_{M,N}$ for the normal point (M,N).

proof. All elements $q(z)$ of $\mathcal{S}_N^{(M,N)} \subset \mathcal{S}_N^{(m,n)}$ having degree $< N$ can be written in a unique way as $q(z) = Z(z)q_z(z) + X(z)q_x(z)$ with $\deg q_z < \delta_n$ (equality means that $\deg q = N$ because the highest degree coefficients of Z are linearly independent) and $\deg q_x \leq \delta - 1$ where the coefficients of $q_z(z) = q_{z,0} + \cdots + q_{z,\delta_n-1} z^{\delta_n-1}$ and $q_x(z) = q_{x,0} + \cdots + q_{x,\delta-1} z^{\delta-1}$ are solutions of the following system of linear equations

$$S_{m,n} \begin{bmatrix} q_{x,0} \\ q_{x,1} \\ \vdots \\ q_{x,\delta-1} \\ q_{z,0} \\ q_{z,1} \\ \vdots \\ q_{z,\delta_n-1} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{7}$$

If the matrix $S_{m,n}$ is nonsingular, only the zero vector is a solution of $S_{m,n}v = 0$. Hence, from Lemma 3.1 it follows that (M,N) is a normal point. On the other hand, if (M,N) is a normal point, the only solution of (7) should be the zero vector. Hence, $S_{m,n}$ should be nonsingular.

If $S_{m,n}$ is nonsingular, the reader can easily check that $[X'(z) \;\; Z'(z)]$ satisfies the degree and interpolation conditions to be the basis matrix for the normal point (M,N). □

The previous theorem leads to the updating procedure 4.1 to compute the basis matrices $B_{m^{(\sigma)},n^{(\sigma)}}$, $\sigma = 0, 1, \ldots$. To get a *square* block Toeplitz matrix $T_{m,n}$, m and n should satisfy $pm = qn$. The next points $(m + \delta m, n + \delta n)$ for which $T_{m+\delta m, n+\delta n}$ is square should satisfy $p(m + \delta m) = q(n + \delta n)$ or $p\delta m = q\delta n = \lambda$. Hence, the smallest value for δm and δn is obtained by making λ as small as possible, i.e. by taking λ equal to the least common multiple of p and q.

Algorithm 4.1 (updating procedure)

$(m^{(0)}, n^{(0)}) \leftarrow (0,0)$
$B_{m^{(0)},n^{(0)}} \leftarrow [0 \;\; \Gamma_z^{(0)}]$ *($\Gamma_z^{(0)}$ is arbitrary nonsingular diagonal)*
λ is the least common multiple of p and q
$\delta m \leftarrow \lambda/p; \;\; \delta n \leftarrow \lambda/q; \;\; \sigma \leftarrow 0$
repeat
    $M \leftarrow m^{(\sigma)} + \delta m$ *and* $N \leftarrow n^{(\sigma)} + \delta n$
    while $S_{m^{(\sigma)},n^{(\sigma)}}$ *is singular* **do**
        $M \leftarrow M + \delta m$ *and* $N \leftarrow N + \delta n$

endwhile
$m^{(\sigma+1)} \leftarrow M$ *and* $n^{(\sigma+1)} \leftarrow N$
compute the polynomial matrices $A(z), B(z), C(z), D(z)$
compute $B_{M,N}$
$\sigma \leftarrow \sigma + 1$
until $M = \overline{m}$

This algorithm allows to update $X(z)$, $Z(z)$, part of the inversion parameters, for a sequence of nested block Toeplitz matrices $T_{m^{(\sigma)},n^{(\sigma)}}$ until the desired matrix $T_{\overline{m},\overline{n}}$ is reached. In the next section, we see how the remaining inversion parameters can be interpreted as building blocks of a basis matrix for a left interpolation problem.

5 Left interpolation problem

Similar to Definition 3.1, we define the left interpolation problem as follows.

Definition 5.1 (left interpolation problem) *For each natural number* α*, the set* $\mathcal{S}_{L,\alpha}$ *is defined as the set of all vector polynomials* $q(z) \in \mathbb{R}[z]^{1\times p}$ *satisfying the interpolation condition*

$$q(z)T(z) = O_-(z^{-n+\alpha-1}) + O_+(z^m) \tag{8}$$

and the degree condition $\deg q \leq \alpha$.

The next theorem shows that all solutions of the left interpolation problem can be represented as a linear polynomial combination of the rows of a basis matrix

$$B_{L,m,n} = \begin{bmatrix} W(z) \\ U(z) \end{bmatrix}$$

with $U(z) = u_0 + u_1 z + \cdots + u_{m-1}z^{m-1}$ and $W(z) = w_0 + w_1 z + \cdots + w_{m-1}z^{m-1} - \Gamma_w z^m$, u_k, w_k as in Theorem 2.1.

Theorem 5.1 *If the block Toeplitz matrix* T *is nonsingular, each element* $q(z)$ *of* $\mathcal{S}_{L,\alpha}$ *can be written in a unique way as*

$$q(z) = [q_w(z) \;\; q_u(z)] \begin{bmatrix} W(z) \\ U(z) \end{bmatrix}$$

with $\deg q_u \leq \min(\alpha - m - 1, n - 1)$ *and* $\deg q_w \leq \alpha - m$.

Note that the normal points of the left and the right problem are just the same because they are defined by looking at the singularity of the block Toeplitz matrices $T_{m,n}$. If the basis matrix $B_{L,m^{(\sigma)},n^{(\sigma)}} = B_{L,m,n}$ in the normal point (m,n) is known, we construct a method to compute the basis matrix $B_{L,m^{(\sigma+k)},n^{(\sigma+k)}} = B_{L,M,N}$ in one of the next normal points (M,N). We use the following notation:

$$\begin{bmatrix} W(z) \\ U(z) \end{bmatrix} T(z) = \begin{bmatrix} R_w^-(z) + R_w^+(z) \\ R_u^-(z) + R_u^+(z) \end{bmatrix}$$

with $R_u^-(z) = z^{m-n}(R_{u,0}^- + R_{u,1}^- z^{-1} + R_{u,2}^- z^{-2} + \cdots)$ with $R_{u,0}^- = \Gamma_u^{(m,n)}$ and $R_u^+(z) = R_{u,1}^+ z^m + R_{u,2}^+ z^{m+1} + R_{u,3}^+ z^{m+2} + \cdots$. Similarly for $R_w^-(z)$ and $R_w^+(z)$ with $R_{w,0}^- = 0$. For $R_v^-(z) = z^{m-n}(R_{v,0}^- + R_{v,1}^- z^{-1} + R_{v,2}^- z^{-2} + \cdots)$ and $R_v^+(z) = R_{v,1}^+ z^m + R_{v,2}^+ z^{m+1} + R_{v,3}^+ z^{m+2} + \cdots$, we define the block Hankel matrix $H_{v,\delta_m,\delta}$ as

$$H_{v,\delta_m,\delta} = \begin{bmatrix} R_{v,\delta_m}^+ & R_{v,\delta_m-1}^+ & \cdots & \cdots & \\ R_{v,\delta_m-1}^+ & \cdots & & & \vdots \\ \vdots & & & \cdots & R_{v,\delta+\delta_n-\delta_m-2}^- \\ & \cdots & \cdots & R_{v,\delta+\delta_n-\delta_m-2}^- & R_{v,\delta+\delta_n-\delta_m-1}^- \end{bmatrix}$$

with $\delta_m = M - m$ and $\delta_n = N - n$, where the number of block columns is N if $\delta \geq n$ and $\delta_n + \delta$ if $\delta \leq n$. The matrix $H_{m,n}$ is defined as

$$H_{m,n} = \begin{bmatrix} H_{w,\delta_m,\delta_m} \\ H_{u,\delta_m,\delta} \end{bmatrix}$$

with $\delta = \min(\delta_m, n)$. Note that $H_{m,n}$ is square iff $pM = qN$. We consider the following system of linear equations

$$H_{m,n}^T \begin{bmatrix} a_0^T & c_0^T \\ a_1^T & c_1^T \\ \vdots & \vdots \\ a_{\delta_m-1}^T & c_{\delta_m-1}^T \\ b_0^T & d_0^T \\ b_1^T & d_1^T \\ \vdots & \vdots \\ b_{\delta-1}^T & d_{\delta-1}^T \end{bmatrix} = \begin{bmatrix} \vdots & 0 \\ \vdots & 0 \\ \vdots & \vdots \\ -R_{w,\delta_n-1}^{+T} & 0 \\ -R_{w,\delta_n}^{+T} & I_q \end{bmatrix} \begin{bmatrix} \left(\Gamma_w^{(m,n)}\right)^{-1} \left(\Gamma_w^{(M,N)}\right) & 0 \\ 0 & \Gamma_u^{(M,N)} \end{bmatrix}. \quad (9)$$

We define the polynomial matrices A_L, B_L, C_L, D_L as

$$\begin{aligned} A_L(z) &= a_0 + a_1 z + \cdots a_{\delta_m-1} z^{\delta_m-1} + \left(\Gamma_w^{(M,N)}\right) \left(\Gamma_w^{(m,n)}\right)^{-1} z^{\delta_m} \\ C_L(z) &= c_0 + c_1 z + \cdots c_{\delta_m-1} z^{\delta_m-1} \\ B_L(z) &= b_0 + b_1 z + \cdots b_{\delta-1} z^{\delta-1} \\ D_L(z) &= d_0 + d_1 z + \cdots d_{\delta-1} z^{\delta-1}. \end{aligned}$$

Theorem 5.2 *If $B_{L,m,n}(z)$ is a left basis matrix in the normal point (m,n) then (M,N) is also a normal point iff the matrix $H_{m,n}$ is nonsingular. Moreover, if $B_{L,m,n}(z) = [W(z)^T \;\; U(z)^T]^T$, then*

$$\begin{bmatrix} W'(z) \\ U'(z) \end{bmatrix} = \begin{bmatrix} A_L(z) & B_L(z) \\ C_L(z) & D_L(z) \end{bmatrix} \begin{bmatrix} W(z) \\ U(z) \end{bmatrix} \quad (10)$$

is a left basis matrix $B_{L,M,N}(z)$ for the normal point (M,N).

proof. The proof is similar to the proof of Theorem 4.1. □

The previous theorem leads to an updating procedure for left basis matrices similar to the one for right basis matrices.

6 Duality

The results of this section will be used in the error analysis of Section 9. We fix the values m and n. If the interpolation condition (3) is not satisfied exactly, but

$$T(z)[X(z) \;\; Z(z)] = [\delta R_x(z) + R_x^+(z) + R_x^-(z) \;\; \delta R_z(z) + R_z^+(z) + R_z^-(z)], \quad (11)$$

with $\delta R_x(z) = O_+(z^0)$ and $\delta R_z(z) = O_+(z^0)$ as small residual errors, we call $B_R = [X \;\; Z]$ a right *numerical* basis matrix (NBM). In the error analysis, we need a duality result, which can be written in terms of *extended* basis matrices.

Definition 6.1 (extended basis matrix) *A right extended basis matrix (EBM) is defined as*

$$\overline{B}_R = \begin{bmatrix} R_x^+ & R_z^+ \\ X & Z \end{bmatrix}.$$

Similarly an extended left basis matrix is defined as

$$\overline{B}_L = \begin{bmatrix} W & -R_w^+ \\ U & -R_u^+ \end{bmatrix}.$$

Similar definitions could be given for the $(-)$-series but we do not need them in the sequel. Note that in [2], the basis matrices are already "extended", i.e. are square in a natural way. Here, we have to introduce the $(+)$-series in a rather artificial way to make the resulting basis matrix square. Before we give the next lemma, we need the definition of the projection operator.

Definition 6.2 (projection operator) *For a formal Laurent series* $T(z) = \sum_{k=-\infty}^{+\infty} t_k z^k$, *we define the projection operator* $\Pi_{m:n}$ *as*

$$\Pi_{m:n} T(z) = \sum_{k=m}^{n} t_k z^k.$$

The next lemma is used in the proof of Theorem 11.1.

Lemma 6.1 (see Lemma 1 in [2]) *If* $\overline{B}_R(z)$ *is a right extended numerical basis matrix (ENBM), then*

$$(-1)^p \det \overline{B}_R(z) = \gamma z^{pm} + \theta_I(z)$$

with

$$\begin{aligned} \theta_I(z) &= \Pi_{pm:+\infty} \left(\det \begin{bmatrix} \delta R_x(z) & \delta R_z(z) \\ X(z) & Z(z) \end{bmatrix} \right) \qquad \textit{and} \\ \gamma &= \prod_{i=1}^{p+q} \gamma_{R,i}, \qquad \gamma_{R,i} \textit{ is the ith diagonal element of } \Gamma_R. \end{aligned}$$

proof. We start from the equality

$$\begin{bmatrix} -I_p & T(z) \\ 0 & I_q \end{bmatrix} \begin{bmatrix} R_x^+(z) & R_z^+(z) \\ X(z) & Z(z) \end{bmatrix} = \begin{bmatrix} \delta R_x(z) + R_x^-(z) & \delta R_z(z) + R_z^-(z) \\ X(z) & Z(z) \end{bmatrix}.$$

Because the determinant of the left hand side is

$$(-1)^p \det \overline{B}_R(z) = O_+(z^{pm}),$$

we get the result. □

The next duality result is used in the proofs of Lemma 9.8 and Theorem 9.9.

Theorem 6.2 (duality (See Theorem 4 of [2])) *If* $\overline{B}_L(z)$ *and* $\overline{B}_R(z)$ *are left and right ENBM respectively, then*

$$\overline{B}_L(z)\overline{B}_R(z) = z^m \Gamma_L \Gamma_R + \theta_{II}(z)$$

where

$$\theta_{II}(z) = \Pi_{m:+\infty} \left(\begin{bmatrix} W(z) \\ U(z) \end{bmatrix} [-\delta R_x \quad -\delta R_z] + \begin{bmatrix} \delta R_w \\ \delta R_u \end{bmatrix} [X \quad Z] \right) \tag{12}$$

proof. The proof is very similar to the proof of Theorem 4 in [2]. We work out the product

$$\begin{aligned} \overline{B}_L(z)\overline{B}_R(z) &= \begin{bmatrix} W(z) \\ U(z) \end{bmatrix} [R_x^+(z) \quad R_z^+(z)] + \begin{bmatrix} -R_w^+(z) \\ -R_u^+(z) \end{bmatrix} [X(z) \quad Z(z)] \\ &= \begin{bmatrix} W(z) \\ U(z) \end{bmatrix} \left([R_x^+(z) \quad R_z^+(z)] - T(z)[X(z) \quad Z(z)] \right) \\ &\quad + \left(\begin{bmatrix} -R_w^+(z) \\ -R_u^+(z) \end{bmatrix} + \begin{bmatrix} W(z) \\ U(z) \end{bmatrix} T(z) \right) [X(z) \quad Z(z)] \\ &= \begin{bmatrix} W(z) \\ U(z) \end{bmatrix} [-\delta R_x(z) \quad -\delta R_z(z)] + \begin{bmatrix} \delta R_w(z) \\ \delta R_u(z) \end{bmatrix} [X(z) \quad Z(z)] \\ &\quad + \begin{bmatrix} W(z) \\ U(z) \end{bmatrix} [-R_x^-(z) \quad -R_z^-(z)] + \begin{bmatrix} R_w^-(z) \\ R_u^-(z) \end{bmatrix} [X(z) \quad Z(z)] \end{aligned}$$

Because W, U, X and Z are $O_+(z^0)$ and $R_w^+, R_u^+, R_x^+, R_z^+$ are $O_+(z^m)$, it is easy to see that the product $\overline{B}_L\overline{B}_R = O_+(z^m)$. It then follows that

$$\begin{aligned}\overline{B}_L(z)\overline{B}_R(z) &= z^m \begin{bmatrix} \Gamma_w\Gamma_x & 0 \\ 0 & \Gamma_u\Gamma_z \end{bmatrix} + \theta_{II}(z) \\ &= z^m\Gamma_L\Gamma_R + \theta_{II}(z)\end{aligned}$$

□

The next two corollaries will be used in the proof of Theorem 11.1.

Corollary 6.1 (see Corollary 5 of [2]) *If $\overline{B}_R(z)$ is a right normalized and $\overline{B}_L(z)$ a left normalized ENBM for the normal point (m,n) of $T(z)$, then*

$$\overline{B}_R(z)\overline{B}_L(z) = z^m I_{p+q} + \theta_{III}(z)$$

with

$$\theta_{III}(z) = \overline{B}_R(z)\Gamma_L^{-1}\theta_{II}(z)\Gamma_R^{-1}\overline{B}_R^{-1}(z).$$

proof. Recall if $\overline{B}_R$ is not normalized, then $\overline{B}_R(z)\Gamma_R^{-1}$ is normalized. Multiplying both sides of (12) on the left by $\overline{B}_R(z)\Gamma_L^{-1}$ and on the right by $\Gamma_R^{-1}\overline{B}_R^{-1}(z)$ gives the result. □

For the next corollary, we need the following notation

$$\begin{aligned}R_R^-(z) &= [R_x^-(z) \;\; R_z^-(z)] \quad \text{(similarly for } R_R^+(z)\text{)} \\ \delta R_R(z) &= [\delta R_x(z) \;\; \delta R_z(z)] \\ \overline{T}(z) &= [-I_p \;\; T(z)].\end{aligned}$$

Corollary 6.2 (see Corollary 6 of [2]) *The residual series $R_R^-(z)$ for a normalized ENBM $\overline{B}_R(z)$ satisfies*

$$R_R^-(z)\overline{B}_L(z) = \overline{T}(z)z^m + \theta_{IV}(z)$$

with

$$\theta_{IV}(z) = \overline{T}(z)\theta_{III}(z) - \delta R_R(z)\overline{B}_L(z).$$

proof. Multiplying both sides of

$$\overline{T}(z)\overline{B}_R(z) = \delta R_R(z) + R_R^-(z)$$

to the right by $\overline{B}_L(z)$ and applying the previous corollary, leads to

$$\begin{aligned}R_R^-(z)\overline{B}_L(z) &= \overline{T}(z)\overline{B}_R(z)\overline{B}_L(z) - \delta R_R(z)\overline{B}_L(z) \\ &= \overline{T}(z)\left(z^m I_{p+q} + \theta_{III}(z)\right) - \delta R_R(z)\overline{B}_L(z).\end{aligned}$$

□

7 Updating procedure for the extended basis matrix

We can only use the updating procedure for the right (but we drop the index R) extended basis matrix

$$\overline{B}^{(\sigma+1)}(z) = \overline{B}^{(\sigma)}(z)\hat{B}^{(\sigma)}(z),$$

if $\delta_n = n^{(\sigma+1)} - n^{(\sigma)} \leq m^{(\sigma)}$. If $\delta_n > m^{(\sigma)}$, also the $R_z^{(\sigma)-}(z)$ series is needed to update the $R_R^{(\sigma)+}(z)$ series. In this case the dimension of the matrix $S_{m^{(\sigma)},n^{(\sigma)}}$ is equal to the dimension of the matrix $T_{m^{(\sigma+1)},n^{(\sigma+1)}}$. Hence, we can just use $T_{m^{(\sigma+1)},n^{(\sigma+1)}}$ to compute an extended basis matrix and consider this as connected to $\sigma = 1$. A similar reasoning

is true for updating the left EBM. This leads to the updating procedure 7.1. The formal Laurent series $T(z)$ is scaled such that[2] $\max\{||T(z)||, ||T^T(z)||\} \leq 1$. This can be done by first scaling the columns of $T(z)$ such that they get a norm equal to one and then scaling the rows of $T(z)$ having a norm greater than one in the same way. The downscaling of the rows will not increase the norms of the columns. The diagonal matrices $\Gamma_L^{(\sigma)}$ and $\Gamma_R^{(\sigma)}$ are chosen such that $||\overline{B}_R^{(\sigma)}(z)|| = ||\left(\overline{B}_L^{(\sigma)}(z)\right)^T|| = 1$. The error analysis will show that the quantity

$$\kappa^{(\sigma)} = \sum_{i=1}^{p+q} (\gamma_{L,i}\gamma_{R,i})^{-1} \tag{13}$$

gives a measure for the stability of the point $(m^{(\sigma)}, n^{(\sigma)})$. This quantity is compared to a user-supplied tolerance value τ. The final value for M in the algorithm is $\overline{m}$. The variable *choice* indicates at each step whether the solution was accepted or not, based on the comparison between $\kappa^{(\sigma)}$ and τ. It also indicates wether the extended basis matrices are computed using (part of) the given (square) block Toeplitz matrix $T_{\overline{m},\overline{n}}$ ($separated = false$) or using the updating formulas ($separated = true$). The value of *choice* will be plotted in Section 13 when showing some experimental results.

Algorithm 7.1 (updating procedure for the extended basis matrix)

UPDATING$(T(z), \overline{m}, \tau)$
$\lambda \leftarrow$ *least common multiple* (p, q)
$\delta m \leftarrow \lambda/p$; $\delta n \leftarrow \lambda/q$
$(M, N) \leftarrow (0,0)$; $(m^{(0)}, n^{(0)}) \leftarrow (0,0)$; $\sigma \leftarrow 0$
repeat
  $M \leftarrow M + \delta m$; $N \leftarrow N + \delta n$
  separated $\leftarrow (N - n^{(\sigma)} \leq m^{(\sigma)})$ *and* $(M - m^{(\sigma)} \leq n^{(\sigma)})$
  if *not separated* **then**
    if $T_{M,N}$ *is numerically nonsingular* **then**
      compute $\overline{B}_R^{(1)}(z)$ *and* $\overline{B}_L^{(1)}(z)$ *(using Def. 6.1, (11) and Th. 2.1)*
      stable $\leftarrow \kappa^{(1)} \leq \tau$
      if *stable* **then**
        $(m^{(1)}, n^{(1)}) \leftarrow (M, N)$
        $\sigma \leftarrow 1$
        choice $\leftarrow 2$
      else
        choice $\leftarrow -1$
      endif
    endif
  else if $S_{m^{(\sigma)},n^{(\sigma)}}$ *is numerically nonsingular* **then**
      $(m^{(\sigma+1)}, n^{(\sigma+1)}) \leftarrow (M, N)$
      compute $A(z), B(z), C(z), D(z)$ *(Gaussian elimination on* $S^{(\sigma)T}$ *of (5))*
      $\overline{B}_R^{(\sigma+1)}(z) \leftarrow \overline{B}_R^{(\sigma)}\hat{B}_R^{(\sigma)}(z)$
      compute $A_L(z), B_L(z), C_L(z), D_L(z)$ *(Gaussian elimination on* $H^{(\sigma)}$ *of (9))*
      $\overline{B}_L^{(\sigma+1)}(z) \leftarrow \hat{B}_L^{(\sigma)}(z)\overline{B}_L^{(\sigma)}$
      stable $\leftarrow \kappa^{(\sigma+1)} \leq \tau$
      if *stable* **then**
        $\sigma \leftarrow \sigma + 1$
        choice $\leftarrow 1$
      else
        choice $\leftarrow 0$

[2] The norms used will be defined in Section 8.

```
        endif
    endif
until M = m̄
if stable then
    return (B̄_R^(σ)(z), B̄_L^(σ)(z), κ^(σ))
else
    return (B̄_R^(σ+1)(z), B̄_L^(σ+1)(z), κ^(σ+1))
endif
```

If the order of the block Toeplitz matrices $T_{M,N}$ and of the matrices $S^{(\sigma)}$ and $H^{(\sigma)}$ is small during the algorithm, this algorithm requires $O((p\overline{m})^2)$ floating point operations. Hence, this algorithm is an order of magnitude faster than using the classical algorithms to compute the final basis matrix from the block Toeplitz system of equations of order $p\overline{m}$ without taking into account the special structure. Hence, Algorithm 7.1 can be called fast. In the literature, there have appeared also so-called superfast algorithms. However, the stability of these algorithms is still an open question.

8 Norms

Before we give the error analysis in the next section, we define the norms here. We will assume that all the series considered have only a finite number of nonzero coefficients. This is no limitation for practical applications where we can assume that all coefficients of $T(z)$ not appearing in the block Toeplitz matrix $T_{\overline{m},\overline{n}}$ are zero. So, we can define a norm for polynomials, polynomial vectors and matrices as well as for series. The set of polynomials $\mathbb{R}[z]$ is considered as a subset of $\mathbb{R}(z)$, i.e. the set of all formal Laurent series (with only a finite number of nonzero coefficients). The norm of $a(z) = \sum_{i=-\infty}^{+\infty} a_i z^i \in \mathbb{R}(z)$ is defined as

$$||a(z)|| = \sum |a_i|$$

where the sum is taken over all nonzero coefficients. The norm of $A(z) \in \mathbb{R}(z)^{p\times q}$ is defined as

$$||A(z)|| = \max_{j=1,\ldots,q} \sum_{i=1}^{p} ||a_{i,j}(z)||.$$

It is easy to prove that for two elements A_1 and A_2 of $\mathbb{R}(z)^{p\times q}$: $||A_1A_2|| \leq ||A_1||\,||A_2||$, $||A_1|| \leq p||A_1^T||$ and $||A_1^T|| \leq q||A_1||$. Note that in [2] different norms are used for polynomials and series. This makes the proofs harder to read but the advantage is that no assumption has to be made concerning the number of nonzero coefficients of a series.

9 Error analysis

To make clear the similarities and the differences between the error analysis of [2] and this report, the same path of reasoning is followed with a different content. Let μ denote the unit roundoff or machine precision.

Lemma 9.1 *If $\mu n \leq 0.01$, then*

$$fl\left(\sum_{i=1}^{n} u_i v_i\right) = \sum_{i=1}^{n} u_i v_i (1+\delta_i),$$

with $|\delta_i| \leq 1.01 n\mu$.

proof. This is a direct consequence of Theorem (20.18) of [6]. □

Lemma 9.2 (see Lemma 10 in [2]) *Let $B(z)$ be a NBM of type (m,n) for $T(z)$, then*

$$fl(T(z)B(z)) = T(z)B(z) + \psi(z)$$

where

$$||\psi(z)|| \leq 1.01\mu(n+q)||T(z)||\,||B(z)||.$$

proof.

$$\begin{aligned} fl[\sum_{k=1}^{q} t_{i,k}(z)b_{k,j}(z)] &= \sum_{l=-\infty}^{+\infty} z^l fl(\sum_{k=1}^{q}\sum_{r=0}^{n} t_{i,k}^{(l-r)} b_{k,j}^{(r)}) \\ &= \sum_{l=-\infty}^{+\infty} z^l \sum_{k=1}^{q}\sum_{r=0}^{n} t_{i,k}^{(l-r)} b_{k,j}^{(r)}(1+\delta_{i,j,k,r,l}) \end{aligned}$$

where $|\delta_{i,j,k,r,l}| \leq 1.01\mu(n+q)$. So,

$$\psi_{i,j}(z) = \sum_{l=-\infty}^{+\infty} z^l \sum_{k=1}^{q}\sum_{r=0}^{n} t_{i,k}^{(l-r)} b_{k,j}^{(r)} \delta_{i,j,k,r,l}$$

and

$$\begin{aligned} ||\psi(z)|| &= \max_{j=1,2,\ldots,p+q} \sum_{i=1}^{p} ||\psi_{i,j}(z)|| \\ &\leq 1.01\mu(n+q)\max_j \sum_{i=1}^{p}\sum_{l=-\infty}^{+\infty}\left(\sum_{k=1}^{q}\sum_{r=0}^{n} |t_{i,k}^{(l-r)}|\,|b_{k,j}^{(r)}|\right) \\ &\leq 1.01\mu(n+q)\max_j \sum_{i=1}^{p}\sum_{k=1}^{q}\sum_{r=0}^{n} ||t_{i,k}(z)||\,|b_{k,j}^{(r)}| \\ &\leq 1.01\mu(n+q)\max_j \sum_{k=1}^{q}\sum_{r=0}^{n} |b_{k,j}^{(r)}| \sum_{i=1}^{p} ||t_{i,k}(z)|| \\ &\leq 1.01\mu(n+q)||T(z)||\,||B(z)||. \end{aligned}$$

□

To construct $S_{m^{(\sigma)},n^{(\sigma)}}$, we need the first coefficients of the $(+)$-residual series $R^{(\sigma)+}(z)$ which is part of the right ENBM $\overline{B}^{(\sigma)}(z)$. We also have to compute the first coefficients of the $(-)$-residual series $R^{(\sigma)-}(z)$. For the constant coefficient, we can take $[\Gamma_x^{(\sigma)} \;\; 0]$.

Lemma 9.3 (Lemma 12 of [2]) *The computed residual $R^{(\sigma)-}(z)$ satisfies*

$$R^{(\sigma)-}(z) = [\Gamma_x^{(\sigma)} \;\; 0] + \Pi_{-\infty:-1}\left(T(z)B^{(\sigma)}(z)\right) + \theta_V^{(\sigma)}(z),$$

where

$$||\theta_V^{(\sigma)}(z)|| \leq 1.01(n^{(\sigma)}+q)\mu.$$

proof. Follows immediately from the previous Lemma. □

$\hat{B}^{(\sigma)}(z)$ is obtained by solving (5) using Gaussian elimination on the transpose of $S^{(\sigma)}$. Taking the transpose leads to the 1-norm instead of the ∞-norm in the backward error analysis for Gaussian elimination.

Lemma 9.4 (Lemma 13 of [2]) *If $S^{(\sigma)}$ is nonsingular, $\hat{B}^{(\sigma)}(z)$ is obtained by solving (5) using Gaussian elimination on $S^{(\sigma)T}$, and $\|\delta R^{(\sigma)}\| = O(\mu)$, then*

$$R^{(\sigma)}(z)\hat{B}^{(\sigma)}(z) = [\Gamma_x^{(\sigma+1)} \;\; 0] + \theta_{VI}^{(\sigma)}(z) + O_-(z^{-1}) + O_+(z^{m^{(\sigma+1)}}),$$

where

$$\begin{aligned} R^{(\sigma)}(z) &= R^{(\sigma)-}(z) + R^{(\sigma)+}(z), \\ \|\theta_{VI}^{(\sigma)}(z)\| &\leq (8p^3(\delta_m^{(\sigma)} + \delta_n^{(\sigma)})^3\rho_\sigma\mu + f_\sigma(\mu)\|\hat{B}^{(\sigma)}\| \\ f(\mu) &= O(\mu^2) \end{aligned}$$

with ρ_σ the growth factor of the Gaussian elimination and $f_\sigma(\mu)$ the $O(\mu^2)$ term appearing in the error analysis of the Gaussian elimination (see (14)).

proof. Because we use Gaussian elimination on the transpose of $S^{(\sigma)}$, the coefficients X of each column of $\hat{B}^{(\sigma)}$ satisfy

$$(S^{(\sigma)} + \epsilon)X = Y$$

where Y is the corresponding column of the right-hand side of (5) and

$$\|\epsilon\|_1 \leq 8p^3(\delta_m^{(\sigma)} + \delta_n^{(\sigma)})^3\rho_\sigma\|S^{(\sigma)}\|_1\mu + f_\sigma(\mu) \tag{14}$$

with $f_\sigma(\mu) = O(\mu^2)$. The number ρ_σ is the growth factor of the Gaussian elimination. Note that we have an upper bound for the 1-norm of the error instead of the ∞-norm because we take the *transpose* of $S^{(\sigma)}$. For the ∞-norm result, we refer the reader to (4.4-13) of [12]. Because

$$T(z)B^{(\sigma)}(z) = R^{(\sigma)}(z) + \delta R^{(\sigma)} - \theta_V^{(\sigma)}(z),$$

we get

$$\begin{aligned} \|S^{(\sigma)}\|_1 &\leq \|R^{(\sigma)}(z)\| \\ &\leq 1 + 1.01(n^{(\sigma)} + q)\mu + \|\delta R^{(\sigma)}(z)\|. \end{aligned}$$

So,

$$S^{(\sigma)}X - Y = -\epsilon X$$

where

$$\|\epsilon X\|_1 \leq \|\epsilon\|_1\|X\|_1 \leq (8p^3(\delta_m^{(\sigma)} + \delta_n^{(\sigma)})^3\rho_\sigma\mu + O(\mu^2))\|X\|_1.$$

Hence,

$$\|\theta_{VI}^{(\sigma)}(z)\| \leq (8p^3(\delta_m^{(\sigma)} + \delta_n^{(\sigma)})^3\rho_\sigma\mu + O(\mu^2))\|\hat{B}^{(\sigma)}(z)\|.$$

□

Lemma 9.5 (Lemma 14 of [2]) *If $\overline{B}^{(\sigma+1)}(z) = \text{proj } fl(\overline{B}^{(\sigma)}(z)\hat{B}^{(\sigma)}(z))$, where* proj *means the projection $\Pi_{m^{(\sigma+1)}:+\infty}$ of the upper block, i.e. the first p rows, then*

$$\overline{B}^{(\sigma+1)}(z) = \overline{B}^{(\sigma)}(z)\hat{B}^{(\sigma)}(z) + \theta_{VII}^{(\sigma)'}(z)$$

where

$$\begin{aligned} \theta_{VII}^{(\sigma)'}(z) &= \theta_{VII}^{(\sigma)}(z) - \begin{bmatrix} \theta_{VI}^{(\sigma)+}(z) \\ 0 \end{bmatrix}, \\ \|\theta_{VII}^{(\sigma)}(z)\| &\leq 1.01(\delta_n^{(\sigma)} + p + q)\|\overline{B}^{(\sigma)}(z)\|\,\|\hat{B}^{(\sigma)}(z)\|\mu, \\ \theta_{VI}^{(\sigma)+}(z) &= \Pi_{m^{(\sigma)}:m^{(\sigma+1)}-1}\theta_{VI}^{(\sigma)}(z). \end{aligned}$$

proof. The upper bound for $\|\theta_{VII}^{(\sigma)}(z)\|$ is found similarly as in Lemma 9.2. By the projection, we just set the $\Pi_{m^{(\sigma)}:m^{(\sigma+1)}-1}$ part of the upper block of $\overline{B}^{(\sigma)}(z)\hat{B}^{(\sigma)}(z)$ equal to zero. But this part is just equal to $\theta_{VI}^{(\sigma)+}(z)$. □

Note that compared to Lemma 14 of [2] we also apply a projection operator introducing the error $\theta_{VI}^{(\sigma)+}(z)$.

The next lemma gives the error $\delta R^{(\sigma+1)}(z)$ at the $(\sigma+1)$th iteration in terms of the residual error $\delta R^{(\sigma)}(z)$ at the σth iteration and the floating point errors introduced locally by the σth iteration.

Lemma 9.6 (Lemma 15 of [2])

$$\delta R^{(\sigma+1)}(z) = \delta R^{(\sigma)}(z)\hat{B}^{(\sigma)}(z) + \mathcal{L}^{(\sigma)}(z), \tag{15}$$

where

$$\mathcal{L}^{(\sigma)}(z) = \Pi_{0:+\infty}\left(T(z)\theta_{VII}^{(\sigma)}(z) + \theta_{VI}^{(\sigma)}(z) - \theta_{V}^{(\sigma)}(z)\hat{B}^{(\sigma)}(z)\right). \tag{16}$$

proof. We start from the definition of $\delta R^{(\sigma+1)}$

$$\left[\Gamma_x^{(\sigma+1)} \;\; 0\right] + \delta R^{(\sigma+1)} + O(z^{-1}) = \overline{T}(z)\overline{B}^{(\sigma+1)}(z) \tag{17}$$

where

$$\overline{T}(z) = [-I \;\; T(z)].$$

Using Lemmas 9.3, 9.4 and 9.5, the right-hand side of (17) can be written as

$$\begin{aligned}
\overline{T}(z)\overline{B}^{(\sigma+1)}(z) &= \overline{T}(z)\left(\overline{B}^{(\sigma)}(z)\hat{B}^{(\sigma)}(z) + \theta_{VII}^{(\sigma)'}(z)\right) \\
&= \left(\delta R^{(\sigma)}(z) + R^{(\sigma)-}(z) - \theta_{V}^{(\sigma)}(z)\right)\hat{B}^{(\sigma)}(z) + \overline{T}(z)\theta_{VII}^{(\sigma)'}(z) \\
&= \delta R^{(\sigma)}(z)\hat{B}^{(\sigma)}(z) + \left[\Gamma_x^{(\sigma+1)} \;\; 0\right] + \theta_{VI}^{(\sigma)}(z) + \overline{T}(z)\theta_{VII}^{(\sigma)}(z) - \theta_{V}^{(\sigma)}(z)\hat{B}^{(\sigma)}(z)
\end{aligned}$$

The lemma follows from (17). □

By recursively applying (15), we can write the residual error in function of the local errors at each iteration as follows.

Theorem 9.7 (Theorem 16 of [2]) *The residual error satisfies*

$$\delta R^{(\sigma+1)}(z) = \sum_{j=0}^{\sigma} \mathcal{L}^{(j)}(z)\mathcal{G}_j^{(\sigma)}(z), \tag{18}$$

where

$$\mathcal{G}_j^{(\sigma)}(z) = \begin{cases} \hat{B}^{(j+1)}(z)\hat{B}^{(j+2)}(z)\cdots\hat{B}^{(\sigma)}(z), & 0 \le j < \sigma \\ I_{p+q}, & j = \sigma. \end{cases} \tag{19}$$

proof. Follows immediately from Lemma 9.6. □

The local error $\mathcal{L}^{(j)}(z)$ at iteration j is multiplied by $\mathcal{G}_j^{(\sigma)}(z)$. Hence, it is important to have a sharp upper bound for $\|\mathcal{G}_j^{(\sigma)}(z)\|$. We start by giving an upper bound for $\|\mathcal{G}_{\sigma-1}^{(\sigma)}(z)\| = \|\hat{B}^{(\sigma)}(z)\|$.

Lemma 9.8 (Lemma 17 of [2]) *If μ, $\|\delta R_R^{(\sigma)}(z)\|$ and $\|\delta R_L^{(\sigma)}(z)\|$ are so small that*

$$\begin{aligned}\kappa^{(\sigma)} \Big\{ \Big(&(p+q)\|\delta R_R^{(\sigma)}(z)\| + \|\delta R_L^{(\sigma)}(z)\| \Big) \Big) \\ &+ \Big((p+q)1.01(\delta_n^{(\sigma)} + p + q)\mu \\ &+ (p+q)8p^3(\delta_m^{(\sigma)} + \delta_n^{(\sigma)})^3 \rho_\sigma \mu + (p+q) f_\sigma(\mu) \Big) \Big\} \leq \frac{1}{2},\end{aligned}$$

then

$$\|\hat{B}^{(\sigma)}(z)\| \leq 2(p+q)\kappa^{(\sigma)}.$$

proof. By definition of $\kappa^{(\sigma)}$ in (13), we get

$$\begin{aligned}\|(\Gamma_L^{(\sigma)}\Gamma_R^{(\sigma)})^{-1}\overline{B}_L^{(\sigma)}(z)\overline{B}_R^{(\sigma+1)}(z)\| &\leq \|(\Gamma_L^{(\sigma)}\Gamma_R^{(\sigma)})^{-1}\|\|\overline{B}_L^{(\sigma)}(z)\| \\ &\leq \kappa^{(\sigma)}(p+q).\end{aligned}$$

But using Theorem 6.2 and Lemma 9.5

$$\begin{aligned}\|(\Gamma_L^{(\sigma)}\Gamma_R^{(\sigma)})^{-1}\overline{B}_L^{(\sigma)}(z)\overline{B}_R^{(\sigma+1)}(z)\| &= \|(\Gamma_L^{(\sigma)}\Gamma_R^{(\sigma)})^{-1}\overline{B}_L^{(\sigma)}(z)\left\{\overline{B}_R^{(\sigma)}(z)\hat{B}_R^{(\sigma)}(z) + \theta_{VII}^{(\sigma)'}(z)\right\} \\ &= \|(\Gamma_L^{(\sigma)}\Gamma_R^{(\sigma)})^{-1}\left\{z^{m^{(\sigma)}}\Gamma_L^{(\sigma)}\Gamma_R^{(\sigma)} + \theta_{II}^{(\sigma)}(z)\right\}\hat{B}_R^{(\sigma)}(z) \\ &\quad + (\Gamma_L^{(\sigma)}\Gamma_R^{(\sigma)})^{-1}\overline{B}_L^{(\sigma)}(z)\theta_{VII}^{(\sigma)'}(z)\| \\ &\geq \|\hat{B}_R^{(\sigma)}(z)\| \\ &\quad - \|(\Gamma_L^{(\sigma)}\Gamma_R^{(\sigma)})^{-1}\|\,\|\theta_{II}^{(\sigma)}(z)\|\,\|\hat{B}_R^{(\sigma)}(z)\| \\ &\quad - \|(\Gamma_L^{(\sigma)}\Gamma_R^{(\sigma)})^{-1}\|\,\|\overline{B}_L^{(\sigma)}(z)\|\,\|\theta_{VII}^{(\sigma)'}(z)\| \\ &\geq \|\hat{B}_R^{(\sigma)}(z)\|/2.\end{aligned}$$

This proves the lemma. □

Theorem 9.9 (Theorem 18 of [2]) *If μ, $\|\delta R_R^{(j)}(z)\|$ and $\|\delta R_L^{(j)}(z)\|$ are so small that*

$$\begin{aligned}\kappa^{(j)} \Big\{ \Big(&(p+q)\|\delta R_R^{(j)}(z)\| + \|\delta R_L^{(j)}(z)\| \Big) \\ &+ \Big((p+q)1.01(\delta_n^{(j)} + p + q)\mu \\ &+ (p+q)8p^3(\delta_m^{(\sigma)} + \delta_n^{(\sigma)})^3 \rho_j \mu + (p+q) f_j(\mu) \Big) \Big\} \leq \frac{1}{2}, \quad j \leq \sigma,\end{aligned}$$

then

$$\|\mathcal{G}_{j-1}^{(\sigma)}(z)\| \leq 2(p+q)\kappa^{(j)} + O(\mu), \qquad j \leq \sigma.$$

proof. The proof is by induction on j. From the previous lemma, we get that the theorem is true for $j = \sigma$. Assume the theorem is true for $\mathcal{G}_{\sigma-1}^{(\sigma)}(z), \mathcal{G}_{\sigma-2}^{(\sigma)}(z), \ldots, \mathcal{G}_j^{(\sigma)}(z)$. From the definition (13) of $\kappa^{(j)}$,

$$\begin{aligned}\|(\Gamma_L^{(j)}\Gamma_R^{(j)})^{-1}\overline{B}_L^{(j)}(z)\overline{B}_R^{(\sigma+1)}(z)\| &\leq \|(\Gamma_L^{(j)}\Gamma_R^{(j)})^{-1}\|\|\overline{B}_L^{(j)}(z)\| \\ &\leq \kappa^{(\sigma)}(p+q).\end{aligned}$$

From (19) and Lemma 9.5,

$$\overline{B}_R^{(\sigma+1)}(z) = \overline{B}_R^{(j)}(z)\mathcal{G}_{j-1}^{(\sigma)}(z) + \sum_{l=j}^{\sigma} \theta_{VII}^{(l)'}(z)\mathcal{G}_l^{(\sigma)}(z).$$

Using Lemma 9.5, Theorem 6.2 and the inductive hypothesis,

$$\begin{aligned}
&\|(\Gamma_L^{(j)}\Gamma_R^{(j)})^{-1}\overline{B}_L^{(j)}(z)\overline{B}_R^{(\sigma+1)}(z)\| \\
&= \|(\Gamma_L^{(j)}\Gamma_R^{(j)})^{-1}\overline{B}_L^{(j)}(z)\overline{B}_R^{(j)}(z)\mathcal{G}_{j-1}^{(\sigma)}(z) \\
&\quad + \sum_{l=j}^{\sigma}(\Gamma_L^{(j)}\Gamma_R^{(j)})^{-1}\overline{B}_L^{(j)}(z)\theta_{VII}^{(l)'}(z)\mathcal{G}_l^{(\sigma)}(z)\| \\
&\geq \|(\Gamma_L^{(j)}\Gamma_R^{(j)})^{-1}\left\{z^{m^{(j)}}\Gamma_L^{(j)}\Gamma_R^{(j)} + \theta_{II}^{(j)}(z)\right\}\mathcal{G}_{j-1}^{(\sigma)}(z)\| \\
&\quad -\kappa^{(j)}\sum_{l=j}^{\sigma}(p+q)\|\theta_{VII}^{(l)'}(z)\|\left\{2\kappa^{(l+1)}(p+q)+O(\mu)\right\} \\
&\geq \|\mathcal{G}_{j-1}^{(\sigma)}(z)\|\left\{1-\kappa^{(j)}\|\theta_{II}^{(j)}(z)\|\right\} - O(\mu) \\
&\geq \|\mathcal{G}_{j-1}^{(\sigma)}(z)\|/2 - O(\mu).
\end{aligned}$$

□

To analyse the residual error $\delta R^{(\sigma+1)}(z)$ given by (18), we split up the local error $\mathcal{L}^{(\sigma)}(z)$ given by (16) as

$$\mathcal{L}^{(\sigma)}(z) = \mathcal{L}_1^{(\sigma)}(z) + \mathcal{L}_2^{(\sigma)}(z) + \mathcal{L}_3^{(\sigma)}(z)$$

where

$$\begin{aligned}
\mathcal{L}_1^{(\sigma)}(z) &= \Pi_{0:+\infty}\left(-\theta_V^{(\sigma)}(z)\hat{B}^{(\sigma)}(z)\right) \\
\mathcal{L}_2^{(\sigma)}(z) &= \theta_{VI}^{(\sigma)}(z) \\
\mathcal{L}_3^{(\sigma)}(z) &= \Pi_{0:+\infty}\left(T(z)\theta_{VII}^{(\sigma)}(z)\right).
\end{aligned}$$

For each of these local errors, we investigate the influence in (18), i.e. we look at

$$\mathcal{E}_i^{(\sigma+1)}(z) = \sum_{j=0}^{\sigma}\mathcal{L}_i^{(j)}(z)\mathcal{G}_j^{(\sigma)}(z), \qquad i = 1,2,3.$$

The sum of these errors gives us the residual error

$$\delta R^{(\sigma+1)}(z) = \sum_{i=1}^{3}\mathcal{E}_i^{(\sigma+1)}(z).$$

In the following lemmas, we give upper bounds for each of the norms $\|\mathcal{E}_i^{(\sigma+1)}(z)\|$.

Lemma 9.10 (Lemma 19 of [2])

$$\begin{aligned}
\|\mathcal{E}_1^{(\sigma+1)}(z)\| \leq\ & 4\kappa^{(\sigma)}(p+q)(n^{(\sigma)}+q)\mu \\
& +8(p+q)^2\mu\sum_{j=0}^{\sigma-1}\kappa^{(j)}\kappa^{(j+1)}(n^{(j)}+q) \\
& +O(\mu^2)
\end{aligned}$$

proof.

$$\|\mathcal{E}_1^{(\sigma+1)}(z)\| \leq \|\sum_{j=0}^{\sigma}\mathcal{L}_1^{(j)}(z)\mathcal{G}_j^{(\sigma)}(z)\|$$

$$
\begin{aligned}
&\leq \|\theta_V^{(\sigma)}(z)\| \, \|\hat{B}_R^{(\sigma)}(z)\| + \sum_{j=0}^{\sigma-1} \|\theta_V^{(j)}(z)\| \, \|\hat{B}_R^{(j)}(z)\| \, \|\mathcal{G}_j^{(\sigma)}(z)\| \\
&\leq \left\{1.01(n^{(\sigma)}+q)\mu\right\}\left\{2(p+q)\kappa^{(\sigma)}\right\} \\
&\quad + \sum_{j=0}^{\sigma-1}\left\{1.01(n^{(j)}+q)\mu\right\}\left\{2(p+q)\kappa^{(j+1)}+O(\mu)\right\} \\
&\leq 4\kappa^{(\sigma)}(p+q)(n^{(\sigma)}+q)\mu \\
&\quad +8(p+q)^2\mu\sum_{j=0}^{\sigma-1}\kappa^{(j)}\kappa^{(j+1)}(n^{(j)}+q) \\
&\quad +O(\mu^2).
\end{aligned}
$$

□

Lemma 9.11 (Lemma 20 of [2])

$$
\begin{aligned}
\|\mathcal{E}_2^{(\sigma+1)}(z)\| \leq\ & 16\kappa^{(\sigma)}(p+q)p^3(\delta_n^{(\sigma)}+\delta_m^{(\sigma)})^3\rho_\sigma\mu \\
& +32(p+q)^2p^3\mu\sum_{j=0}^{\sigma-1}\kappa^{(j)}\kappa^{(j+1)}\rho_j(\delta_n^{(j)}+\delta_m^{(j)})^3 \\
& +O(\mu^2)
\end{aligned}
$$

proof.

$$
\begin{aligned}
\|\mathcal{E}_2^{(\sigma+1)}(z)\| &\leq \|\sum_{j=0}^{\sigma}\mathcal{L}_2^{(j)}(z)\mathcal{G}_j^{(\sigma)}(z)\| \\
&\leq \|\theta_{VI}^{(\sigma)}(z)\| + \sum_{j=0}^{\sigma-1}\|\theta_{VI}^{(j)}(z)\| \, \|\mathcal{G}_j^{(\sigma)}(z)\| \\
&\leq (8p^3(\delta_n^{(\sigma)}+\delta_m^{(\sigma)})^3\rho_\sigma\mu+O(\mu^2))\|\hat{B}_R^{(\sigma)}\| \\
&\quad +\sum_{j=0}^{\sigma-1}(8p^3(\delta_n^{(j)}+\delta_m^{(j)})^3\rho_j\mu+O(\mu^2))\|\hat{B}_R^{(j)}\| \, \|\mathcal{G}_j^{(\sigma)}(z)\| \\
&\leq (8p^3(\delta_n^{(\sigma)}+\delta_m^{(\sigma)})^3\rho_\sigma\mu+O(\mu^2))(2\kappa^{(\sigma)}(p+q)) \\
&\quad +\sum_{j=0}^{\sigma-1}(8p^3((\delta_n^{(j)}+\delta_m^{(j)})^3\rho_j\mu+O(\mu^2))(2\kappa^{(j)}(p+q))(2\kappa^{(j+1)}(p+q)+O(\mu)).
\end{aligned}
$$

□

Lemma 9.12 (Lemma 21 of [2])

$$
\begin{aligned}
\|\mathcal{E}_3^{(\sigma+1)}(z)\| \leq\ & 4\kappa^{(\sigma)}(p+q)(\delta_n^{(\sigma)}+p+q)\mu \\
& +8(p+q)^2\mu\sum_{j=0}^{\sigma-1}\kappa^{(j)}\kappa^{(j+1)}(\delta_n^{(j)}+p+q) \\
& +O(\mu^2)
\end{aligned}
$$

proof.

$$\begin{aligned}
\|\mathcal{E}_3^{(\sigma+1)}(z)\| &\leq \|\sum_{j=0}^{\sigma} \mathcal{L}_3^{(j)}(z)\mathcal{G}_j^{(\sigma)}(z)\| \\
&\leq \|\theta_{VII}^{(\sigma)}(z)\| + \sum_{j=0}^{\sigma-1} \|\theta_{VII}^{(j)}(z)\| \, \|\mathcal{G}_j^{(\sigma)}(z)\| \\
&\leq 1.01(\delta_n^{(\sigma)} + p + q)\|\hat{B}_R^{(\sigma)}\|\mu \\
&\quad + \sum_{j=0}^{\sigma-1} (1.01(\delta_n^{(j)} + p + q)\mu)\|\hat{B}_R^{(j)}\| \, \|\mathcal{G}_j^{(j)}(z)\| \\
&\leq 1.01(\delta_n^{(\sigma)} + p + q)(2(p+q)\kappa^{(\sigma)})\mu \\
&\quad + \sum_{j=0}^{\sigma-1} (1.01(\delta_n^{(j)} + p + q)\mu)(2(p+q)\kappa^{(j)})(2(p+q)\kappa^{(j+1)} + O(\mu)).
\end{aligned}$$

□

Finally, this gives us an upper bound for the norm of the residual error.

Theorem 9.13 (Theorem 22 of [2]) *If μ, $\|\delta R_R^{(j)}(z)\|$ and $\|\delta R_L^{(j)}(z)\|$ are so small that*

$$(\overline{n} + p + q)\mu \leq 0.01$$

and

$$\begin{aligned}
&\kappa^{(j)} \Big\{ \Big((p+q)\|\delta R_R^{(j)}(z)\| + \|\delta R_L^{(j)}(z)\| \Big) \\
&\quad + \Big((p+q)1.01(\delta_n^{(j)} + p + q)\mu \\
&\quad + (p+q)8p^3(\delta_m^{(\sigma)} + \delta_n^{(\sigma)})^3\rho_j\mu + (p+q)f_j(\mu) \Big) \Big\} \leq \frac{1}{2}, \quad j \leq \sigma,
\end{aligned}$$

then

$$\|\delta R^{(\sigma+1)}(z)\| \leq F_\sigma + 2(p+q)\sum_{j=0}^{\sigma-1} \kappa^{(j+1)} F_j$$

with

$$\begin{aligned}
F_j &= 4\kappa^{(j)}(p+q)\mu \\
&\quad \Big\{ (n^{(j)} + q) + 4p^3(\delta_n^{(j)} + \delta_m^{(j)})^3\rho_\sigma + (\delta_n^{(j)} + p + q) \Big\}. \qquad (20)
\end{aligned}$$

proof. Summation of the error bounds given in Lemmas 9.10, 9.11 and 9.12 gives the result. □

This theorem shows that if $\|\delta R_R^{(j)}(z)\|$ and $\|\delta R_L^{(j)}(z)\|$ are small for $j \leq \sigma$ also $\|\delta R^{(\sigma+1)}(z)\|$ will be small if $\kappa^{(\sigma)}$ is not too large. An analogous error analysis can be made for $\delta R_L^{(\sigma+1)}(z)$. The norm $\|.\|_L$ is defined as $\|.\|_L = \|.^T\|$.

Theorem 9.14 (Theorem 33 of [2]) *If μ, $\|\delta R_R^{(j)}(z)\|_L$ and $\|\delta R_L^{(j)}(z)\|_L$ are so small that*

$$(\overline{m} + p + q)\mu \leq 0.01$$

and

$$\begin{aligned}
&\kappa^{(j)} \Big\{ \Big(\|\delta R_R^{(j)}(z)\|_L + (p+q)\|\delta R_L^{(j)}(z)\|_L \Big) \\
&\quad + \Big((p+q)1.01(\delta_m^{(j)} + p + q)\mu \\
&\quad + (p+q)8q^3(\delta_m^{(\sigma)} + \delta_n^{(\sigma)})^3\rho_{j,L}\mu + (p+q)f_{j,L}(\mu) \Big) \Big\} \leq \frac{1}{2}, \quad j \leq \sigma,
\end{aligned}$$

then

$$\|\delta R_L^{(\sigma+1)}(z)\|_L \leq F_{\sigma,L} + 2(p+q)\sum_{j=0}^{\sigma-1} \kappa^{(j+1)} F_{j,L}$$

with

$$\begin{aligned} F_{j,L} &= 4\kappa^{(j)}(p+q)\mu \\ &\quad \left\{(m^{(j)}+p) + 4q^3(\delta_n^{(j)} + \delta_m^{(j)})^3 \rho_{\sigma,L} + (\delta_m^{(j)} + p + q)\right\}. \end{aligned} \tag{21}$$

10 The inverse of a block Toeplitz matrix

In this section, we give formulas for the inverse of a block Toeplitz matrix based on *numerical* basis matrices. These results will be used in the next section to get an upper bound for the norm of the error between the exact and the computed right basis matrix. Let $[x^{f(i,j)}]$ denote a matrix whose element at position (i,j) is $x^{f(i,j)}$.

Theorem 10.1 (see Theorem 34 of [2]) *If $B_R(z)$ is a right and $B_L(z)$ a left* normalized *NBM for the normal point (m,n) of $T(z)$, the inverse of the block Toeplitz matrix $T_{m,n}$ satisfies*

$$\begin{aligned} T_{m,n}^{-1}(I + \theta_{VIII}) &= L_{(n,\eta)}(x_0, x_1, \ldots, x_{n-1}) U_{(\eta,m)}(I_p, -w_{m-1}, \ldots, -w_1) \qquad (22) \\ &\quad + L_{(n,\eta)}(z_0, z_1, \ldots, z_{n-1}) U_{(\eta,m)}(0, u_{m-1}, \ldots, u_1) \\ &= M_{m,n} \qquad (23) \end{aligned}$$

with

$$\begin{aligned} \theta_{VIII} &= \left[\theta_{IV,L}^{(m+i-j)}\right] - \left[t^{(-n+i-j)}\right]\left[\theta_{III,L,L}^{(m+n+i-j)}\right] \\ &\quad + \left[\delta R_x^{(i-j)}\right]\left[w^{(m+i-j)}\right] + \left[\delta R_z^{(i-j)}\right]\left[u^{(m+i-j)}\right]. \end{aligned}$$

Here, $\theta_{IV,L}(z)$ denotes the left block of $\theta_{IV}(z)$, i.e. the first p columns. Similarly, $\theta_{III,L,L}(z)$ denotes the lower, left block of $\theta_{III}(z)$, i.e. the last q rows of the first p columns of $\theta_{III}(z)$.

proof.

$$T_{m,n} L_{(n,\eta)}(X) = \left[\delta R_x^{(i-j)} + R_x^{-(i-j)}\right] - \left[t^{(-n+i-j)}\right]\left[x^{(n+i-j)}\right].$$

Similarly for $T_{m,n} L_{(n,\eta)}(Z)$. Hence,

$$\begin{aligned} T_{m,n} M_{m,n} &= \left(\left[\delta R_x^{(i-j)} + R_x^{-(i-j)}\right] - \left[t^{(-n+i-j)}\right]\left[x^{(n+i-j)}\right]\right)\left[w^{(m+i-j)}\right] \\ &\quad + \left(\left[\delta R_z^{(i-j)} + R_z^{-(i-j)}\right] - \left[t^{(-n+i-j)}\right]\left[z^{(n+i-j)}\right]\right)\left[u^{(m+i-j)}\right]. \end{aligned}$$

Using Corollary 6.1 and 6.2,

$$\begin{aligned} T_{m,n} M_{m,n} &= I + \left[\theta_{IV,L}^{(m+i-j)}\right] - \left[t^{(-n+i-j)}\right]\left[\theta_{III,L,L}^{(m+n+i-j)}\right] \\ &\quad + \left[\delta R_x^{(i-j)}\right]\left[w^{(m+i-j)}\right] + \left[\delta R_z^{(i-j)}\right]\left[u^{(m+i-j)}\right]. \end{aligned}$$

This proves the theorem. □

Note that this proof shows that $M_{m,n}$ is the inverse of $T_{m,n}$ if all errors are zero, i.e. if $B_R(z)$ and $B_L(z)$ are *exact* basis matrices.

Corollary 10.1 (see Corollary 35 of [2]) *In terms of the scaled NBM $B_R(z)$ and $B_L(z)$, the inverse of $T_{m,n}$ is given by*

$$\begin{aligned} T_{m,n}^{-1}(I+\theta_{IX}) &= L_{(n,\eta)}(X)\,\mathrm{diag}(\Gamma_x^{-1}\Gamma_w^{-1})U_{(\eta,m)}(W) \\ &\quad + L_{(n,\eta)}(Z)\,\mathrm{diag}(\Gamma_z^{-1}\Gamma_u^{-1})U_{(\eta,m)}(U) \end{aligned}$$

with

$$\begin{aligned} \theta_{IX} &= \left[\theta_{IV,L}^{(m+i-j)}\right] - \left[t^{(-n+i-j)}\right]\left[\theta_{III,L,L}^{(m+n+i-j)}\right] \\ &\quad + \left[\delta R_x^{(i-j)}\right]\mathrm{diag}(\Gamma_x^{-1}\Gamma_w^{-1})\left[w^{(m+i-j)}\right] \\ &\quad + \left[\delta R_z^{(i-j)}\right]\mathrm{diag}(\Gamma_z^{-1}\Gamma_u^{-1})\left[u^{(m+i-j)}\right]. \end{aligned}$$

proof. Follows directly from the previous theorem by taking scaled basis matrices instead of normalized ones. □

11 Stability

In this section, we investigate how accurately the scaled right basis matrix is computed. In Section 9, we have given an upper bound for $\|\delta R_R(z)\|$. A similar one holds for $\|\delta R_L(z)\|$. We will use this result to give an upper bound for $\|\delta B_R(z)\|$ with $\delta B_R(z) = B_R(z) - B_{R,E}(z)$ where $B_{R,E}(z)$ is the scaled *exact* BM with the same scaling as $B_R(z)$. Because

$$T(z)\delta B_R(z) = \Pi_{0:m-1}\left(\delta R_R(z)\right) + O_-(z^{-1}) + O_+(z^m),$$

we get that

$$\|\delta B_R(z)\| \leq \|T_{m,n}^{-1}\|_1 \|\delta R_R(z)\|.$$

Theorem 11.1 (see Theorem 40 of [2]) *If μ is so small and $\|\delta R_R(z)\|$ and $\|\delta R_L(z)\|$ are not too large such that the conditions of Theorem 9.13 are satisfied and*

$$(p+q)!\|\delta R_R(z)\|^p \leq \gamma/2 \tag{24}$$

and

$$2\kappa(p+q)\|\delta R_R(z)\| + \frac{4\kappa(p+q)!}{\gamma}\left\{(p+q)\|\delta R_R(z)\| + \|\delta R_L(z)\|\right\} \leq 1/2, \tag{25}$$

then

$$\|\delta B_R(z)\| \leq 2\kappa(p+q)\left\{F_\sigma + 2(p+q)\sum_{j=0}^{\sigma-1}\kappa^{(j+1)}F_j\right\}$$

where F_j is defined by (20).

proof. Because $\|T_{m,n}^{-1}\|_1$ is connected to $\|\theta_{IX}\|_1$ by Corollary 10.1, we have to look at $\|\theta_I(z)\|$, $\|\theta_{II}(z)\|$, $\|\theta_{III}(z)\|$ and $\|\theta_{IV}(z)\|$. From Lemma 6.1 and (24), we get

$$\|\theta_I(z)\| \leq (p+q)!\|\delta R_R(z)\|^p \leq \gamma/2.$$

From Lemma 6.1,

$$\|\det \overline{B}_R(z)\| \geq \gamma/2.$$

From Theorem 6.2,

$$\|\theta_{II}(z)\| \leq (p+q)\|\delta R_R(z)\| + \|\delta R_L(z)\|.$$

From Corollary 6.1,

$$\begin{aligned}
\|\theta_{III}(z)\| &= \|\overline{B}_R(z)(\Gamma_l\Gamma_R)^{-1}\theta_{II}(z)\overline{B}_R^{-1}(z)\| \\
&= \|\overline{B}_R(z)(\Gamma_l\Gamma_R)^{-1}\theta_{II}(z)\overline{B}_R^{\mathrm{adj}}(z)/\det\overline{B}_R(z)\| \\
&\leq 2\kappa(p+q)!\|\theta_{II}(z)\|/\gamma \\
&\leq \frac{2\kappa(p+q)!}{\gamma}\{(p+q)\|\delta R_R(z)\| + \|\delta R_L(z)\|\}.
\end{aligned}$$

From Corollary 6.2,

$$\begin{aligned}
\|\theta_{IV}(z)\| &= \|\overline{T}(z)\theta_{III}(z) - \delta R_R(z)(\Gamma_l\Gamma_R)^{-1}\overline{B}_L(z)\| \\
&\leq \kappa(p+q)\|\delta R_R(z)\| + \frac{2\kappa(p+q)!}{\gamma}\{(p+q)\|\delta R_R(z)\| + \|\delta R_L(z)\|\}.
\end{aligned}$$

From Corollary 10.1 and (25),

$$\begin{aligned}
\|\theta_{IX}\|_1 &= \| \left[\theta_{IV,L}^{(m+i-j)}\right] - \left[t^{(-n+i-j)}\right]\left[\theta_{III,L,L}^{(m+n+i-j)}\right] \\
&\quad + \left[\delta R_x^{(i-j)}\right]\mathrm{diag}(\Gamma_x^{-1}\Gamma_w^{-1})\left[w^{(m+i-j)}\right] \\
&\quad + \left[\delta R_z^{(i-j)}\right]\mathrm{diag}(\Gamma_z^{-1}\Gamma_u^{-1})\left[u^{(m+i-j)}\right]\|_1 \\
&\leq \|\theta_{IV}(z)\| + \|\theta_{III}(z)\| + (p+q)\kappa\|\delta R_R(z)\| \\
&\leq 2\kappa(p+q)\|\delta R_R(z)\| + \frac{4\kappa(p+q)!}{\gamma}\{(p+q)\|\delta R_R(z)\| + \|\delta R_L(z)\|\} \qquad (26) \\
&\leq 1/2. \qquad (27)
\end{aligned}$$

Hence, $I + \theta_{IX}$ is invertible and

$$\|(I+\theta_{IX})^{-1}\|_1 \leq \frac{1}{1-\|\theta_{IX}\|_1} \leq 2.$$

From Corollary 10.1,

$$\begin{aligned}
\|T_{m,n}^{-1}\|_1 &\leq \|(I+\theta_{IX})^{-1}\|_1\|L_{(n,\eta)}(X)\,\mathrm{diag}(\Gamma_x^{-1}\Gamma_w^{-1})U_{(\eta,m)}(W) \\
&\quad + L_{(n,\eta)}(Z)\,\mathrm{diag}(\Gamma_z^{-1}\Gamma_u^{-1})U_{(\eta,m)}(U)\|_1 \qquad (28) \\
&\leq 2\kappa(p+q). \qquad (29)
\end{aligned}$$

Together with Theorem 9.13, this proves the theorem. □

Hence, if we allow in Algorithm 7.1 only those points $(m^{(\sigma)}, n^{(\sigma)})$ where the $\kappa^{(\sigma)}$ value is smaller than a tolerance value τ (stable points), the conditions of Theorem 11.1 will be satisfied. Hence, the right basis matrices will be accurately computed. Note that from (29) in these points, the 1-norm condition number of $T_{m^{(\sigma)},n^{(\sigma)}}$ is bounded by $2\kappa^{\sigma}(p+q)$. So, algorithm 7.1 is a weakly stable algorithm to compute the right basis matrices in the stable points.

12 The solution of a block Toeplitz system

In this section, we investigate the final error by using the approximation $M = M_{m,n}$ given by (23) of the inverse of the block Toeplitz matrix $T = T_{m,n}$ to solve a block Toeplitz system of linear equations

$$TV = Y, \qquad T \in \mathbb{R}^{pm\times qn}, V \in \mathbb{R}^{qn\times 1}, Y \in \mathbb{R}^{pm\times 1}$$

where T and Y are given and V is the unknown vector.

Lemma 12.1 *The computed value V satisfies*

$$V = MY + \theta_X$$

with

$$\|\theta_X\|_1 \leq \left(1.01\mu(2m + 2n + 3p + q + 2) + O(\mu^2)\right)(p + q)\|Y\|_1.$$

proof. The floating point computation of the γth element in the ith block of

$$L_{(n,\eta)}(X)U_{(\eta,m)}(W)Y = \left[x^{(i-j)}\right]\left[w^{(m+i-j)}\right]\left[b^{(i)}\right]$$

equals

$$\sum_{k=0}^{n-1}\sum_{\alpha=1}^{p} x_{\gamma,\alpha}^{(i-k)}\left(\sum_{j=0}^{m-1}\sum_{\beta=1}^{p} w_{\alpha,\beta}^{(m+k-j)} b_\beta^j (1 + \delta_{\alpha,\beta,k,j})\right)(1 + \epsilon_{\gamma,\alpha,i,k}),$$

with

$$|\delta_{\alpha,\beta,k,j}| \leq 1.01(m + p)\mu$$
$$|\epsilon_{\gamma,\alpha,i,k}| \leq 1.01(n + p)\mu.$$

Hence,

$$\begin{aligned}
\|\theta_X^{(1)}\|_1 &\leq \sum_{\gamma=1}^{q}\sum_{i=0}^{n-1}\sum_{k=0}^{n-1}\sum_{\alpha=1}^{p}\sum_{j=0}^{m-1}\sum_{\beta=1}^{p} |x_{\gamma,\alpha}^{(i-k)}||w_{\alpha,\beta}^{(m+k-j)}||b_\beta^{(j)}| \\
&\quad (|\delta_{\alpha,\beta,k,j}| + |\epsilon_{\gamma,\alpha,i,k}| + |\delta_{\alpha,\beta,k,j}||\epsilon_{\gamma,\alpha,i,k}|) \\
&\leq \left(1.01\mu(m + n + 2p) + O(\mu^2)\right)\|L_{(n,\eta)}(x)\|_1\|U_{(\eta,m)}(W)\|_1\|Y\|_1 \\
&\leq \left(1.01\mu(m + n + 2p) + O(\mu^2)\right)(p + q)\|Y\|_1.
\end{aligned}$$

Similarly, we derive

$$\|\theta_X^{(2)}\|_1 \leq \left(1.01\mu(m + n + p + q) + O(\mu^2)\right)(p + q)\|Y\|_1.$$

Finally adding the two parts together leads to

$$\begin{aligned}
V &= (MY + \theta_X^{(1)} + \theta_X^{(2)})(1 + \delta) \\
&= MY + \theta_X
\end{aligned}$$

with

$$|\delta| \leq \mu.$$

This proves the lemma. □

Theorem 12.2 *If the conditions of Theorem 11.1 are satisfied, then the computed solution V of the equation*

$$TV = Y$$

satisfies

$$V = V_E + \theta_{XI}$$

with

$$\|\theta_{XI}\|_1 \leq (p + q)\left[\kappa + 1.01\mu(2m + 2n + 3p + q + 2) + O(\mu^2)\right]\|Y\|_1$$

and

$$\|Y\|_1 \leq 2\kappa(p + q)\|V_E\|_1,$$

with

$$V_E = T^{-1}Y.$$

proof. Using the results of Theorem 11.1, i.e. (27) and (29),

$$\begin{aligned} \|\theta_{XI}\|_1 &\leq \|T^{-1}\|_1\|\theta_{IX}\|_1\|Y\|_1 + \|\theta_X\|_1 \\ &\leq [2\kappa(p+q)]\left[\frac{1}{2}\right]\|Y\|_1 + \|\theta_X\|_1 \\ &\leq (p+q)\left\{\kappa + 1.01\mu(2m+2n+3p+q+2) + O(\mu^2)\right\}\|Y\|_1. \end{aligned}$$

On the other hand,

$$\begin{aligned} \|Y\|_1 &\leq \|T^{-1}V_E\|_1 \\ &\leq \|T^{-1}\|_1\|V_E\|_1 \\ &\leq 2\kappa(p+q)\|V_E\|_1. \end{aligned}$$

This proves the theorem. □

This shows that executing Algorithm 7.1 and multiplying the approximating inverse M by the right-hand side Y gives a weakly stable algorithm to solve the block Toeplitz system of linear equations.

13 Experimental results

In this section, we show how accurately Algorithm 7.1 together with the inversion formula computes the solution of each block-Toeplitz set of linear equations

$$T_{m,n}V_{m,n} = Y_{m,n}$$

for each $T_{m,n}$ appearing explicitly or implicitly in Algorithm 7.1. The elements of the block Toeplitz matrix $T_{\overline{m},\overline{n}}$ were taken as random numbers uniformly distributed in the interval $(0,1)$ except for one element in the first row making at least one leading principal submatrix ill-conditioned. The corresponding formal Laurent series $T(z)$ was scaled as indicated in Section 7 such that $\max\{\|T(z)\|, \|T^T(z)\|\} \leq 1$ leading to the final elements of $T_{\overline{m},\overline{n}}$. The right-hand side $Y_{m,n}$ of the system of linear equations $T_{m,n}V_{m,n} = Y_{m,n}$ is computed by setting $V_{m,n} = [1,1,\ldots,1]^T$. All the computations were done in MATLAB Version 4.2a and 4.2c on a DEC 5000/240 workstation. The machine precision μ is of the order 10^{-16}.

Example 13.1 In Figure 1, we show the influence of the threshold parameter τ. We take $\overline{m} = 60$, $\overline{n} = 90$, $p = 3$ and $q = 2$. If we set $V_{E,m,n} = [1,1,\ldots,1]^T$, the plot *error* gives for each step $-\log_{10}(\|V_{m,n} - V_{E,m,n}\|_1/\|V_{m,n}\|_1)$ where $V_{m,n}$ is computed by using the inversion formula. Also $\log_{10}(\kappa)$ and *choice* appearing in Algorithm 7.1 are plotted as well as the $\log_{10}$ of the (2-norm) condition number of $T_{m,n}$. Note the big loss of accuracy corresponding to the 15th block leading principal submatrix which is ill-conditioned. When the threshold parameter τ is lowered, we avoid these bad intermediate results. In this case, the plots of κ and the condition number have a similar shape showing that κ gives a good estimate of the relative increase or decrease of the condition number of $T_{m,n}$. ◇

Example 13.2 Figure 2 shows the influence of the threshold parameter τ for an example with $\overline{m} = \overline{n} = 50$, $p = q = 1$ where the 25th leading principal submatrix is ill-conditioned. ◇

Note that in the scalar case, i.e., $p = q = 1$, $X(z) = U(z)$ and $Z(z) = W(z)$.

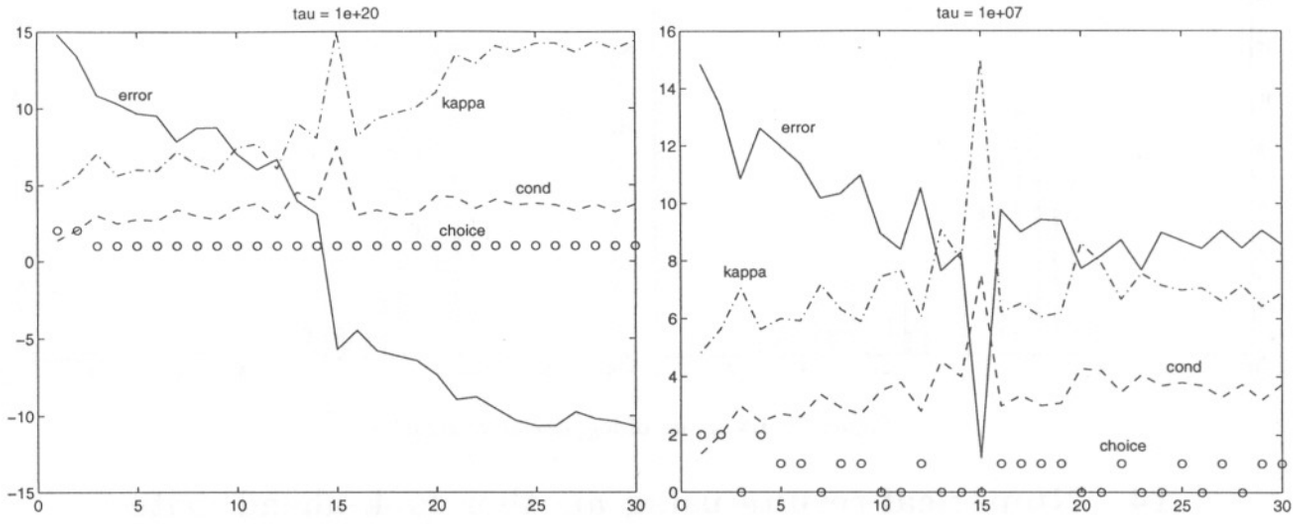


Figure 1: *error*, κ, *choice* and $\text{cond}(T_{m,n})$ for $\tau = 10^{20}$ and $\tau = 10^{7}$

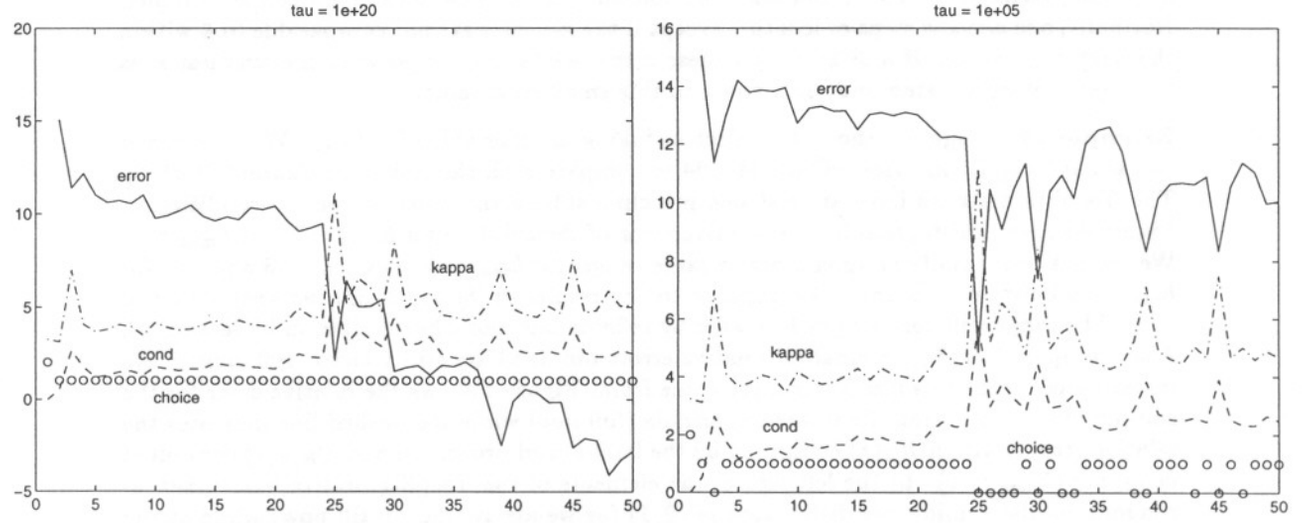


Figure 2: *error*, κ, *choice* and $\text{cond}(T_{m,n})$ for $\tau = 10^{20}$ and $\tau = 10^{5}$

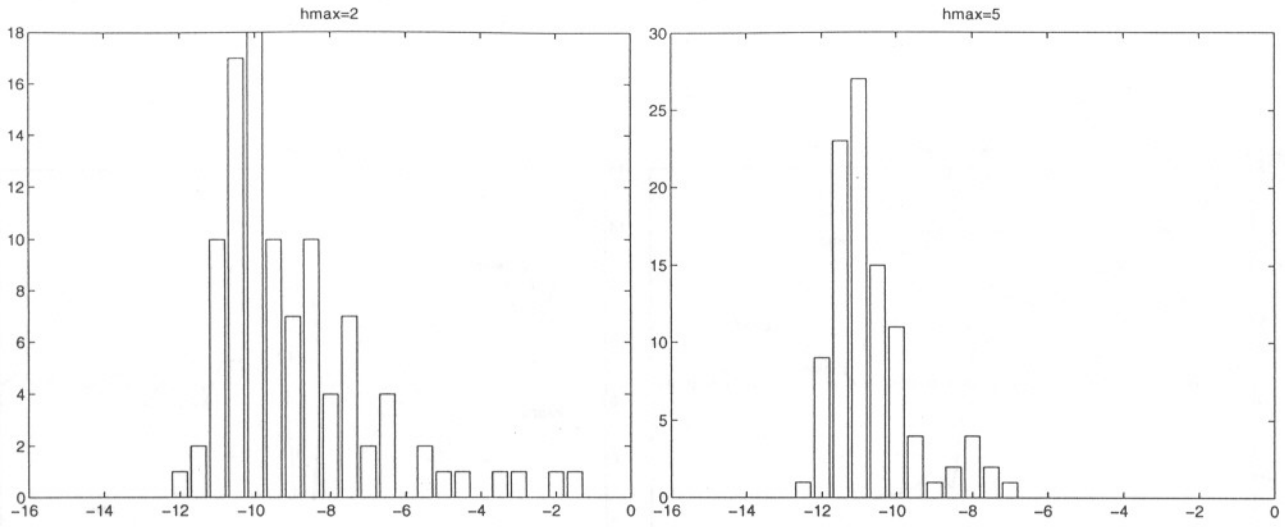


Figure 3: Histogram of $\log_{10}$(relative error)

14 Numerical results using another look-ahead criterion

The problem with the look-ahead scheme used in Algorithm 7.1 is that the user has to choose the threshold value τ. The choice of this τ determines where and how big the look-ahead steps are taken. There is no limit on the size of these look-ahead steps. In this section, we want to develop a more sofisticated look-ahead criterion inspired by the one given by Chan and Hansen in [4, 5]. The user gives $h_{\max}$ which is the maximal length of a look-ahead step. During the algorithm κ is compared to a reference value $\bar{\kappa}$. This reference value is initialized as the minimal κ-value of the first $h_{\max}$ steps. The point where this minimum occurs is the first look-ahead step. From then on, we look for the first step with $\kappa < 2\bar{\kappa}$ within the next $h_{\max}$ steps. The algorithm runs more efficiently as long as the steps taken are smaller. Normally, one expects steps of length 1 except if the κ-values are not comparable to $\bar{\kappa}$ within the next $h_{\max}$ steps. If $\kappa \geq 2\bar{\kappa}$ for all these steps, we take the step with the smallest κ as the next look-ahead step and set $\bar{\kappa}$ equal to this smallest κ value.

Example 14.1 Figure 3 shows the effect of taking another value for $h_{\max}$. We have taken 100 scalar Toeplitz matrices of size 64×64 to compare with the results of Example 1 of [8]. The Toeplitz matrices have at least one principal submatrix which is very ill-conditioned. Figure 3 shows a histogram for the relative error of the solution for $h_{\max} = 2$ and $h_{\max} = 5$. We see that our results ar spread over a large range, for $h_{\max} = 2$ between -13 and -1, for $h_{\max} = 5$ between -13 and -6 compared to the results of [8] which are between -14 and -10. The range will become much smaller by refining our look-ahead criterion in Section 16. Also, compared to [8], the smallest relative errors obtained are still a factor 100 bigger! We investigated where this loss of accuracy came from. Figure 4 shows the relative error for the computed solution using the inversion formula (full line) while the dashed line indicates the relative error between $B_R(z)$ computed by the look-ahead procedure and $B_{R,E}(z)$ computed using (2) ($h_{\max} = 2$). In the left figure, the elements of the Toeplitz matrix were taken as random numbers uniformly distributed in $(0, 1)$ (as we always did up till now) while at the right-hand side, from $(-1, 1)$. Note that the difference between the relative errors for $B_R(z)$ and $V(z)$ is smaller for the $(-1, 1)$ case than for the $(0, 1)$ case. Hence, we concentrated on the $(0, 1)$ case. It turned out that applying the inversion formula led to the loss of two decimal digits. If we use the notation

$$V_1 \quad = \quad L_{(n,\eta)}(X)U_{(\eta,m)}(W)Y$$

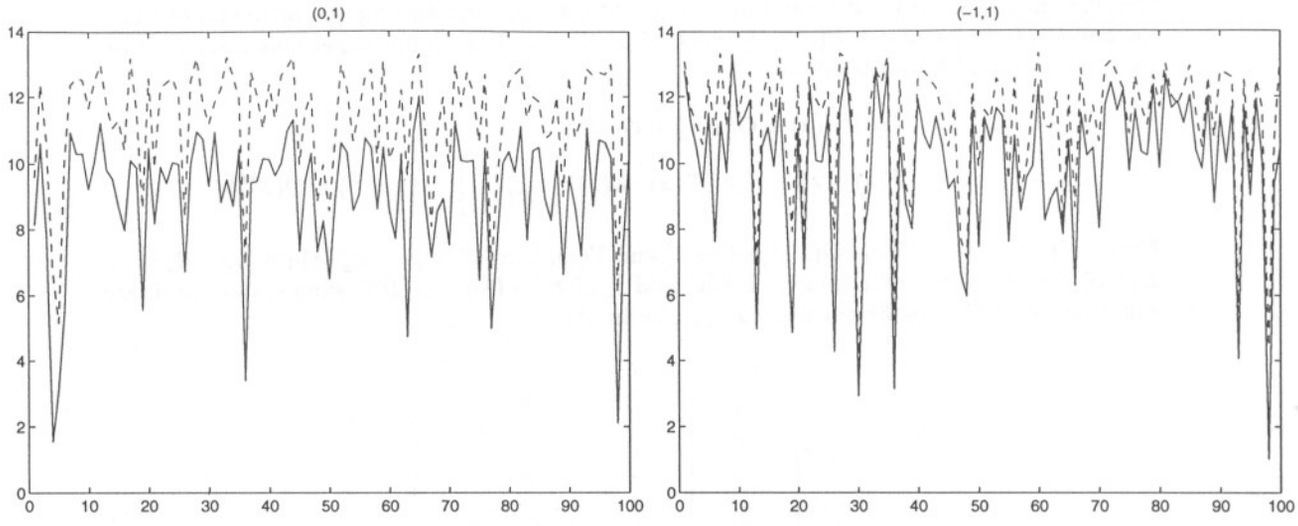


Figure 4: $-\log_{10}$(relative error) for $B_R(z)$ (dashed line) and $V(z)$ (full line)

$$V_2 \quad = \quad L_{(n,\eta)}(Z)U_{(\eta,m)}(U)Y,$$

we get for one of the 100 test cases (with similar results for the others) $\|V_1\|_2 = 497.6939$ and $\|V_2\|_2 = 498.3299$ while V is computed as the sum $V_1 + V_2 = V \approx [1, 1, \ldots, 1]^T$, so $\|V\|_2 \approx 8$. This explains the loss of about 2 decimal places. $\diamond$

In the next section, another method is given to compute the solution of a block Toeplitz system without using the inversion formula while in Section 16 the look-ahead criterion is refined. These two enhancements will lead to results comparable to those of Example 1 in [8]. The stability of inversion formulas for scalar Toeplitz matrices is studied in [15].

15 Updating of the solution of a block Toeplitz system

The block Toeplitz system of linear equations

$$TV = Y$$

at point (m, n) can be written as a nonhomogeneous (right) interpolation problem.

Definition 15.1 (nonhomogeneous interpolation problem) *Let $Y(z) \in \mathbb{R}[z]^{p\times 1}$ and $T(z) \in \mathbb{R}[z]^{p\times q}$ be given. For each normal point (m, n) of $T(z)$, we call $V(z) \in \mathbb{R}[z]^{q\times 1}$ the solution of the nonhomogeneous interpolation problem iff*

$$T(z)V(z) = Y(z) + O(z^{-1}) + O(z^m)$$

and $\deg V(z) < n$.

Note that for a normal point, the solution always exists and is unique.

We construct now the updating formulas to go from the solution $V(z)$ at the normal point (m, n) to the solution $V'(z)$ at the normal point (M, N). Let us try to find the polynomial vectors $E(z) \in \mathbb{R}[z]^{p\times 1}$ and $F(z) \in \mathbb{R}[z]^{q\times 1}$ such that

$$V'(z) = V(z) - [X(z) \;\; Z(z)] \begin{bmatrix} E(z) \\ F(z) \end{bmatrix}. \tag{30}$$

Because the degree of $V'(z)$ should be smaller than N, we limit the degrees of $E(z)$ and $F(z)$, i.e. $\deg E(z) < \delta_n$ and $\deg F(z) < \delta_n$ with $\delta_n = N - n$. The nonhomogeneous interpolation conditions should be satisfied, i.e.

$$\begin{aligned} Y(z) &= T(z)V'(z) + O(z^{-1}) + O(z^M) \\ &= T(z)V(z) - T(z)[X(z) \;\; Z(z)] \begin{bmatrix} E(z) \\ F(z) \end{bmatrix} + O(z^{-1}) + O(z^M) \end{aligned}$$

Because $T(z)V(z) = Y(z) + O(z^{-1}) + O(z^m)$ and $T(z)[X(z) \;\; Z(z)] = [R_x^-(z) + R_x^+(z) \;\; R_z^-(z) + R_z^+(z)]$, we find the coefficients e_i of $E(z)$ and f_i of $F(z)$ from the following system of linear equations with the coefficient matrix $S_{m,n}$ as in (4)

$$S_{m,n} \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_{\delta-1} \\ f_0 \\ f_1 \\ \vdots \\ f_{\delta_n-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ R_{v,m}^+ - Y_m \\ R_{v,m+1}^+ - Y_{m+1} \\ \vdots \\ R_{v,M}^+ - Y_M \end{bmatrix} \tag{31}$$

with $T(z)V(z) = R_v^+(z) + O(z^{-1}) = R_{v,0}^+ + R_{v,1}^+ z^1 + \cdots + O(z^{-1})$ and $Y(z) = Y_0 + Y_1 z^1 + \cdots$. In Section 17, we give the error analysis of the updating formula (30). From now on, we will update the solution of the block Toeplitz system instead of using the inversion formula.

16 Refinement of the look-ahead criterion

Instead of comparing κ with a fixed threshold τ, like we did in Algorithm 7.1, we compared κ with a value $\bar{\kappa}$ which was adapted during the execution of the algorithm. However, the value $\bar{\kappa}$ could only increase. This led in some cases to bad results. Therefore, if we find within the next $h_{\max}$ steps a κ which satisfies $\kappa < 2\bar{\kappa}$, we set the new value of $\bar{\kappa}$ equal to this κ if $\kappa < \bar{\kappa}$.

Example 16.1 If we take hundred 64×64 scalar Toeplitz matrices with elements randomly distributed within $(-1, +1)$ and at least one of the principal submatrices is ill-conditioned, Figure 5 shows the effect of the new look-ahead criterion on the histogram of the relative errors for two different values of $h_{\max}$, 2 and 5. Figure 6 compares the old look-ahead criterion (dashed line) with the new one (full line) plotting the $-\log_{10}$ of the relative errors for the two values of $h_{\max}$, 2 and 5. $\diamond$

Example 16.2 If we take hundred 210×210 block Toeplitz matrices with $p = 2$, $q = 3$, $M = 105$ and $N = 70$ with elements randomly distributed within $(-1, +1)$ and at least one of the principal block-submatrices is ill-conditioned, Figure 7 shows the effect on the histogram of the relative errors when using the new look-ahead criterion. Figure 8 compares the old and the new look-ahead criterion. Note that if a look-ahead step of size $h_{\max} = 5$ is taken, the $S_{m,n}$ matrix is 30×30. For one of the matrices the $\log_{10}$(relative error) is between -6 and -4. Figure 9 shows $-\log_{10}$(relative error) of $B_R(z)$ and $V(z)$ and the $\log_{10}$ of the condition number of the corresponding principal submatrix in each of the normal points taken as look-ahead steps. This example shows that the look-ahead criterion does not always give satisfying results. If we compare the basis matrix $B_R(z)$ computed by the look-ahead procedure in each step with the basis matrix computed using the system of linear equations (2), we get the same plot for the relative error. In the right part of Figure 10, however, we show that the relative error for the basis matrix (dash-dotted line) and the

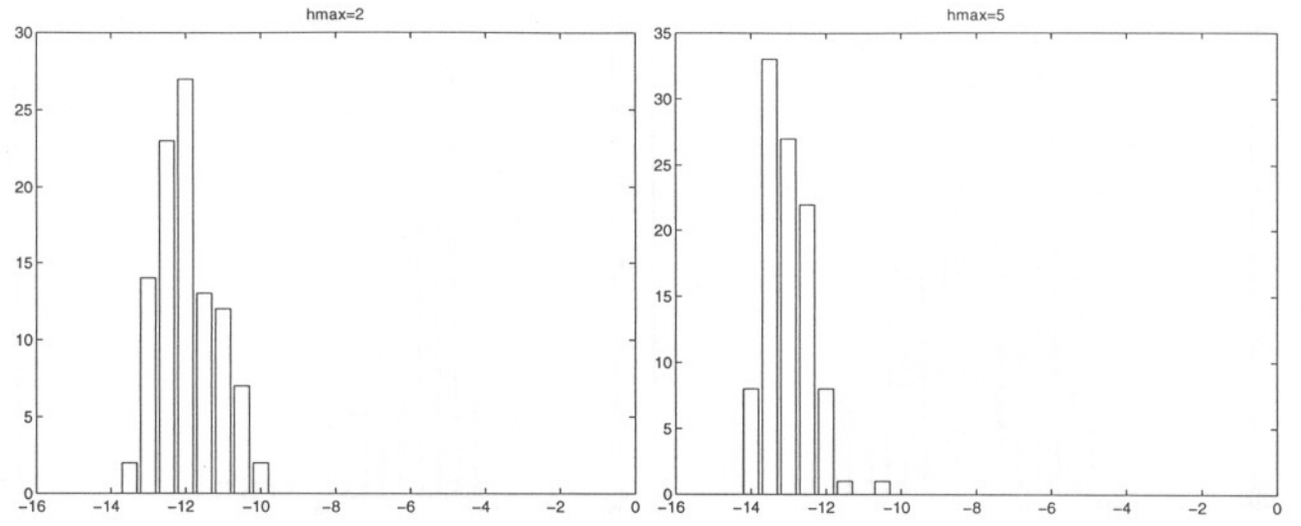


Figure 5: Histogram of $\log_{10}$(relative error)

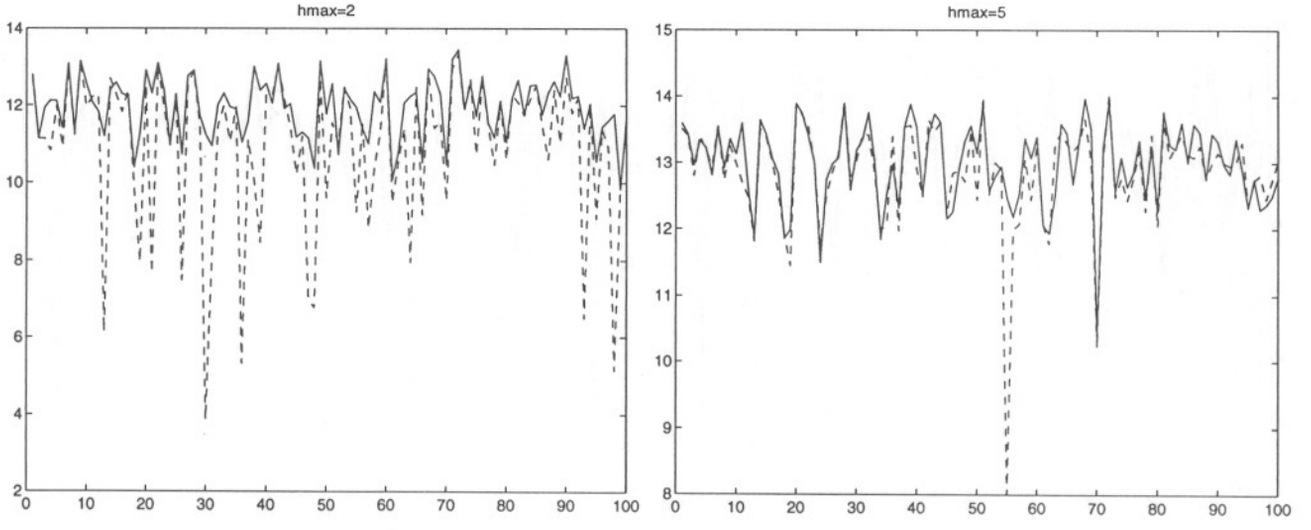


Figure 6: $-\log_{10}$ (rel. err.) for the old (dashed) and the new look-ahead criterion(full line)

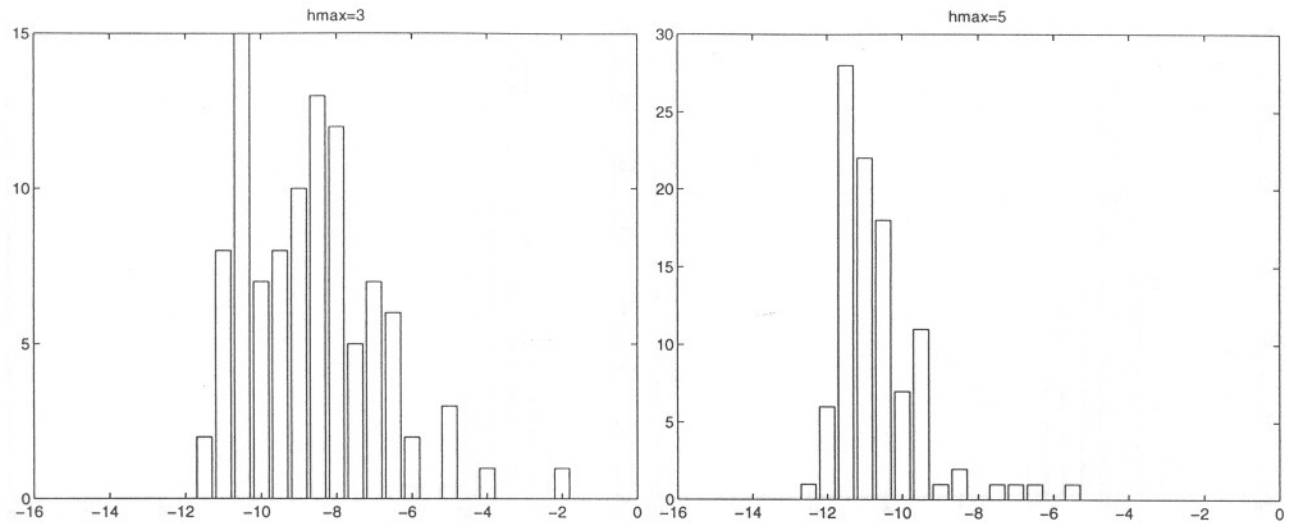


Figure 7: Histogram of $\log_{10}$(relative error)

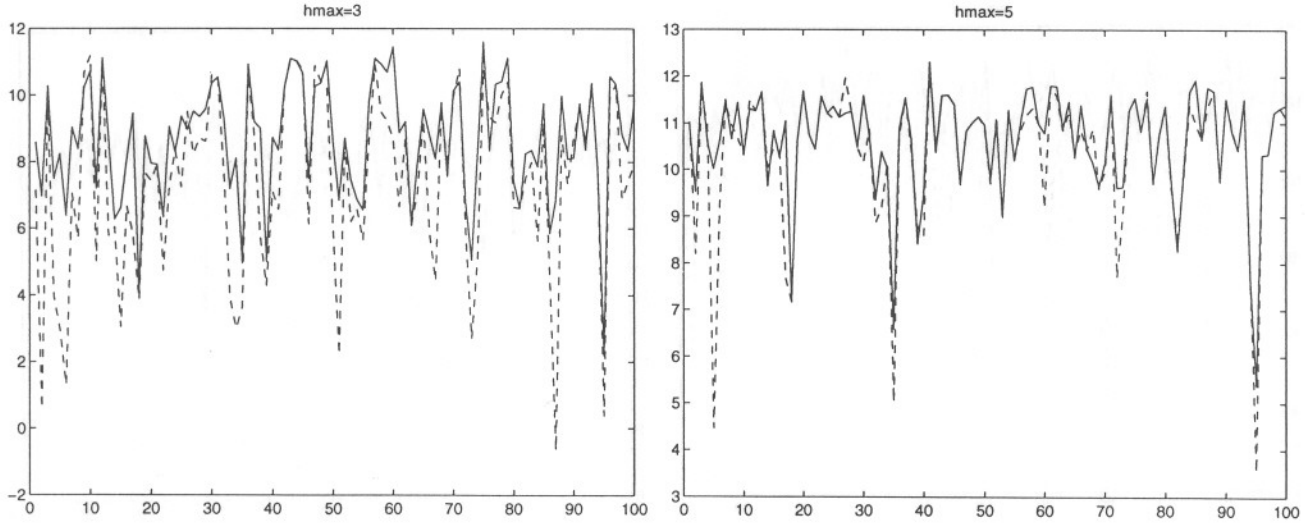


Figure 8: $-\log_{10}$ (rel. err.) for the old (dashed) and the new look-ahead criterion(full line)

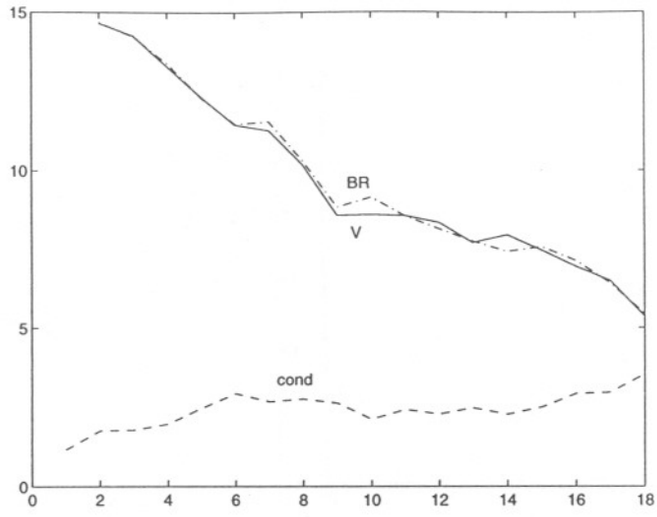


Figure 9: $-\log_{10}$(relative error) and $\log_{10}$(condition number)

relative error for the computed solution (full line) of the block Toeplitz system of equations can have a different behaviour. These cases are rare. In the left part of Figure 10 we show the typical behaviour. For both examples $p = q = 1$, $M = N = 64$, $h_{\max} = 5$ and the elements are uniformly random in $(0, 1)$. Note that till now we have not taken into account the error analysis for the updating of the solution $V(z)$ in the look-ahead strategy. $\diamond$

17 Error analysis

The computed solution $V(z)$ does not satisfy the nonhomogeneous interpolation conditions exactly, i.e.

$$T(z)V^{(\sigma)}(z) = Y^{(\sigma)}(z) + \delta Y^{(\sigma)}(z) + O_-(z^{-1}) + O_+(z^m).$$

We show how $\delta Y^{(\sigma)}(z)$ changes in each step σ. If it is clear from the context, we drop the index (σ). We have the following Lemma about the accuracy of the computed residual $R_v^+(z)$.

Lemma 17.1 *The computed residual $R_v^+(z)$ satisfies*

$$R_v^+(z) = \Pi_{0:+\infty}(T(z)V(z)) + \theta_{XII}(z),$$

with

$$\|\theta_{XII}(z)\| \leq 1.01\mu(q + n - 1)\|V(z)\|.$$

proof. Is similar to the proof of Lemma 9.2. $\square$

Lemma 17.2 *The computed difference of $R_v^+(z)$ and $Y(z)$ satisfies*

$$dRY(z) = (R_v^+(z) - Y(z)) + \theta_{XIII}(z)$$

with

$$\|\theta_{XIII}(z)\| \leq \mu(\|R_v^+(z)\| + \|Y(z)\|).$$

proof. Straightforward. $\square$

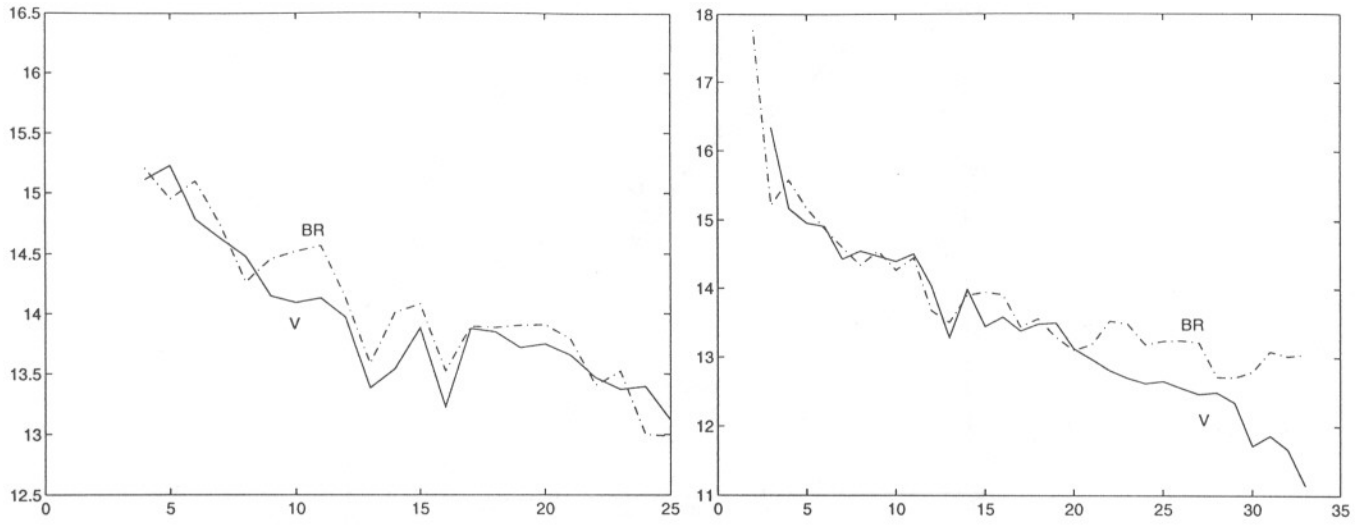


Figure 10: $-\log_{10}$(relative error) of $B_R(z)$ and of $V(z)$

Lemma 17.3 (see Lemma 9.4) *If $S^{(\sigma)}$ is nonsingular, $EF(z) = [E(z)^T \quad F(z)^T]^T$ is obtained by solving (31), then*

$$R^{(\sigma)}(z) \begin{bmatrix} E(z) \\ F(z) \end{bmatrix} = \Pi_{m:M-1}(dRY(z)) + \theta_{XIV}(z) + O_-(z^{-1}) + O_+(z^M)$$

with

$$\|\theta_{XIV}(z)\| \leq (8p^3(\delta_m + \delta_n)^3 \rho_\sigma \mu + O(\mu^2))\|EF(z)\|.$$

proof. The proof is similar to the proof of Lemma 9.4. □

Lemma 17.4 (see Lemma 9.5) *The computed $V^{(\sigma+1)}(z)$ satisfies*

$$V^{(\sigma+1)}(z) = V^{(\sigma)}(z) - [X(z) \quad Z(z)]EF(z) + \theta_{XV}(z)$$

with

$$\|\theta_{XV}(z)\| \leq \mu\|V^{(\sigma)}(z)\| + 1.01\mu(\delta_n^{(\sigma)} + p + q)\|EF(z)\|.$$

proof. Similar to the proof of Lemma 9.5. □

Theorem 17.5 (see Lemma 9.6) *The error $\delta Y^{(\sigma+1)}(z)$ satisfies*

$$\begin{aligned} \delta Y^{(\sigma+1)}(z) \quad = \quad & \delta Y^{(\sigma)}(z) \\ & + \Pi_{0:M-1}\left(-\theta_{XIII}(z) - \theta_{XIV}(z) + T(z)\theta_{XV}(z) + (\delta R^{(\sigma)}(z) - \theta_V(z))EF(z)\right). \end{aligned}$$

proof. Follows from combining Lemma's 17.1-17.4 and 9.3. □

Looking at the upper bound for the error $\delta Y^{(\sigma+1)}(z)$ we see that this error is additive over the different steps. The norm $\|EF(z)\|$ appears in several of the terms of the upper bound. We tried to design a look-ahead strategy in which the magnitude of $\|EF^{(\sigma)}(z)\|$ is compared to the magnitude of $\|Y^{(\sigma+1)}(z)\|$. For some single cases this gives an improvement but generally speaking the results are a little bit worse than using just the look-ahead criterion based on the κ's.

Example 17.1 We take the 100 matrices of Example 16.1 and $h_{\max} = 5$ together with the new look-ahead criterion. Figure 11 should be compared to the right part of Figure 5. ◇

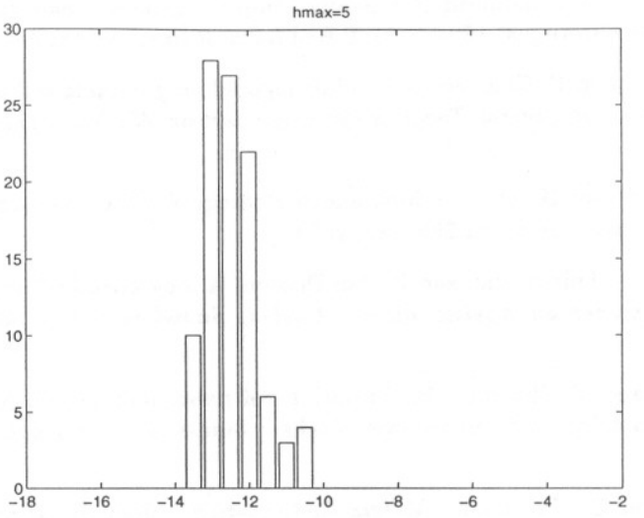


Figure 11: Histogram of $\log_{10}$(relative error)

18 Conclusion

In this report, we have designed an algorithm to compute basis matrices for certain left and right interpolation problems where the data come from a block Toeplitz matrix. This algorithm is weakly stable and efficient. We have shown that the use of the inversion formula gives results which in almost all cases are not as good as using an updating formula for the solution of the block Toeplitz system of linear equations. This leads to an efficient and weakly stable method to solve a block-Toeplitz system of linear equations. However, we have given examples for which the method does not work that well. It is not clear for the moment where the look-ahead strategy fails in these rare cases.

We could have given another implementation by updating in each step besides the whole $(+)$-residual series also the whole $(-)$-residual series. In the implementation we developed in this report, we only computed that part of the $(-)$-series needed to construct the $S_{m,n}$ matrix.

References

[1] A.W. Bojanczyk and G. Heinig. A multi-step algorithm for Hankel matrices. *Journal of Complexity*, 10(1):142–164, March 1994.

[2] S. Cabay, A.R. Jones, and G. Labahn. A stable algorithm for multi-dimensional Padé systems and the inversion of generalized Sylvester matrices. Technical Report TR 94-07, Department of Computing Science, The University of Alberta, Edmonton, Alberta, Canada, March 1994. To appear in SIMAX.

[3] S. Cabay and R. Meleshko. A weakly stable algorithm for Padé approximants and the inversion of Hankel matrices. *SIAM J. Matrix Anal. Appl.*, 14:735–765, 1993.

[4] T. F. Chan and Per C. Hansen. A look-ahead Levinson algorithm for general Toeplitz systems. *IEEE Trans. Sig. Proc.*, 40(5):1079–1090, 1992.

[5] T. F. Chan and Per C. Hansen. A look-ahead Levinson algorithm for indefinite Toeplitz systems. *SIAM J. Matrix Anal. Appl.*, 13(2):490–506, 1992.

[6] G.E. Forsythe and C.B. Moler. *Computer solution of linear algebraic systems.* Prentice-Hall, Englewood Cliffs, N. J., 1967.

[7] R.W. Freund. A look-ahead Bareis algorithm for general Toeplitz matrices. *Numer. Math.*, 68:35–69, August 1994. AT&T Bell Labs. Numerical Analysis Manuscript 93-11.

[8] R.W. Freund and H. Zha. Formally biorthogonal polynomials and a look-ahead Levinson algorithm for general Toeplitz systems. *Linear Algebra Appl.*, 188/189:255–303, 1993.

[9] R.W. Freund and H. Zha. A look-ahead strategy for the solution of general Hankel systems. *Numer. Math.*, 64:295–322, 1993.

[10] K. Gallivan, S. Thirumalai, and P. Van Dooren. A look-ahead schur algorithm. In *Fifth SIAM Conference on Applied Linear Algebra, Snowbird, Utah*, pages 450–454, June 1994.

[11] I. Gohberg and T. Shalom. On Bezoutians of nonsquare matrix polynomials and inversion of matrices with nonsquare blocks. *Linear Algebra Appl.*, 137/138:249–323, 1990.

[12] G. Golub and C. Van Loan. *Matrix Computations.* John Hopkins University Press, North Oxford Academic, 1983.

[13] M.H. Gutknecht. Stable row recurrences for the Padé table and generically superfast lookahead solvers for non-Hermitian Toeplitz systems. *Linear Algebra Appl.*, 188/189:351–422, 1993.

[14] M.H. Gutknecht and M. Hochbruck. Look-ahead Levinson and Schur algorithms for non-Hermitian Toeplitz systems. Technical Report 93-11, IPS, ETH Zürich, 1993.

[15] M.H. Gutknecht and M. Hochbruck. The stability of inversion formulas for Toeplitz matrices. Technical Report 93-13, IPS, ETH Zürich, 1993.

[16] M.H. Gutknecht and M. Hochbruck. Look-ahead Levinson- and Schur-type recurrences in the Padé table. *Electron. Trans. Numer. Anal.*, 2:104–129, 1994.

[17] T. Huckle. A look-ahead algorithm for solving nonsymmetric linear Toeplitz equation. In *Fifth SIAM Conference on Applied Linear Algebra, Snowbird, Utah*, pages 455–459, June 1994.