Helsinki University of Technology Institute of Mathematics Research Reports
Teknillisen korkeakoulun matematiikan laitoksen tutkimusraporttisarja
Espoo 1999 A404

TOEPLITZ PRECONDITIONING OF TOEPLITZ MATRICES — AN OPERATOR THEORETIC APPROACH

Jarmo Malinen

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Jarmo Malinen: Toeplitz Preconditioning of Toeplitz Matrices — an Operator Theoretic Approach; Helsinki University of Technology Institute of Mathematics Research Reports A404 (1999).

Abstract: *We study the preconditioning of Toeplitz matrices $T_n[f]$ by some other Toeplitz matrices. We divide the preconditioned matrix into two parts. One of these parts has a singular value decay that depends on the smoothness of the symbol f. The remaining part is regarded as preconditioning error, and it has the structure of Toeplitz matrix. We indicate that the preconditioners can be generated by a multitude of approximation theoretic methods applied upon the symbol of such system. Moreover, the Toeplitz preconditioning error can be made arbitrarily small if the degree of preconditioning is increased. Our results show that the Krylov subspace methods (such as GMRES or CG) will perform initially at superlinear speed when applied upon such preconditioned system. The connection between smoothness of the symbol f and the numerical properties of matrices $T_n[f]$ with increasing n is presented.*

AMS subject classifications: 65T10, 65T20, 47B35.

Keywords: Toeplitz operator, Krylov subspace methods, preconditioning, speed estimate, superlinear, linear.

ISBN 951-22-4354-7
ISSN 0784-3143
Edita, Espoo, 1999

Helsinki University of Technology
Department of Engineering Physics and Mathematics
Institute of Mathematics
P.O. Box 1100, 02015 HUT, Finland
email: *math@hut.fi*
downloadables: *http://www.math.hut.fi/*

author's email: *Jarmo.Malinen@hut.fi*

1 Introduction

A standard line of attack in the solution of linear systems is the following: First, the "big lines" of the problem are investigated, so that a rough approximation of the inverse operator, called preconditioner, can be found. The application of a preconditioner eliminates the "large amount of disorder" in the problem. A successfully preconditioned linear operator is, in some appropriate sense, "almost" an identity operator. The "remaining disorder" is eliminated by various iterative procedures (Krylov subspace methods, such as CG of GMRES), possibly using parallel computation techniques. In this paper, we apply these ideas to a special class of matrices, namely Toeplitz matrices. It seems that Toeplitz matrices appear practically throughout all the applied mathematics. For example, they are used in the numerical solution of convolution type equations when they approximate an infinite dimensional object (e.g. Toeplitz operator), see [6].

Let us consider some requirements that a good preconditioner should have. In practice, the preconditioning can be done either before or after the discretization of the problem. In the first way, the preconditioning is done to the original (infinite dimensional) operator abstractly "on a piece of paper", and the iterator software works with (a finite dimensional discretization of) the preconditioned operator. For the parallel implementation of the iterator to be effective, the preconditioned operator should, in some sense, be easily "decomposable" into smaller "blocks" that do not communicate too much with each other; the reader is instructed to think of the structures that resemble the decomposition of a compact operator into its generalized eigenspaces. It is also clear that the iteration of the "blocks" should converge fast for a good preconditioner.

The second way to precondition is the following: The infinite dimensional problem is discretized first, and then the finite dimensional matrix of the discretized problem is preconditioned. Now the preconditioned matrix would not necessarily exists in the memory of the computer, because this would require a numerical computation of a matrix-matrix product. Instead, the preconditioner is applied inside each iteration loop over and over again, so that only matrix-vector products are calculated by the computer software. It is now desirable that the preconditioner-vector product is numerically convenient.

In 1986, G. Strang [14] proposed that a circulant Toeplitz preconditioner could be applied on Toeplitz matrices. An iterative conjugate gradient method (CG) is then used to complete the inversion. Several other classes of Toeplitz preconditioners have been studied by many authors; band Toeplitz preconditioners for systems with positive symbols [1], circulant Toeplitz preconditioners associated to convolution kernels [3], circulant Toeplitz preconditioners with complex symbols [4], preconditioners arising from "inverse symbol" [2], to mention few.

The common foundation for all these approaches is the possibility of calculating the Toeplitz-vector product in $n \log n$ time by FFT, where n is the dimension of the matrix (regarded as the level of discretization if the original problem is infinite dimensional). It follows that the calculation of a single iteration step for a Toeplitz system is fast, but the convergence of the iteration is poor, without proper preconditioning. Because of the mentioned $n \log n$ complexity property, it seems a computationally attractive alternative to use a Toeplitz matrix as a preconditioner to a Toeplitz system. Of course, if the iteration of the preconditioned system would still converge poorly, or a good preconditioner could not easily be found, this attraction would be totally lost. Fortunately, this is not the case, as indicated by the results of this paper. We remark that a plenty of results about the clustering of the spectrum of such preconditioned systems have been presented by other authors, as discussed above. In these works, the superlinear convergence of CG is established for such Toeplitz systems. These results have been obtained by matrix algebra tools which is a quite different approach from ours.

We now discuss the outline of this paper. We apply approximation theoretic methods upon the symbol (also known as the generating function) $f \in C(\mathbf{T})$ of the $n \times n$ Toeplitz matrix $T_n[f]$. This gives us a $n \times n$ Toeplitz preconditioner $T_n[g]$ for the original Toeplitz matrix $T_n[f]$. We remark that, contrary to the case of Toeplitz operators $\mathcal{T}[f]$, $T_n[f]$ does not uniquely determine its symbol f. In our approach, much of this uniqueness problem resolves because we are more interested in the families of matrices $\{T_n[f]\}_{n \geq 1}$ for fixed $f \in C(\mathbf{T})$, rather than any of the matrices $T_n[f]$ alone. This setting is the one adopted in [6], where Toeplitz matrix equations of increasing dimension n serve as discretized Toeplitz operator equations.

In our approach, the Toeplitz preconditioned operator is divided into two parts. First of these parts — truncation effect $K_{f,g}^{(n)}$ — is what remains large in operator norm, even if the Toeplitz preconditioning is successfully chosen. This is the part of problem that has to be "iterated away". The other part — perturbation matrix $B_{f,g}^{(n)}$ — is a Toeplitz matrix of small norm, corresponding to the nonoptimality of the Toeplitz preconditioning. Note that after the preconditioning, one should aim to kill only the truncation effect part $K_{f,g}^{(n)}$, at least if n is large. The iteration of the Toeplitz part $B_{f,g}^{(n)}$ is increasingly expensive with increasing n, and thus should not be attempted.

As the dimension n grows, the effect of the smoothness of the symbol f will be seen in the limit process. Smoothness is measured by requiring the rth derivative of symbol to be Lipschitz continuous of index α, for $r + \alpha > 0$. Roughly, matrices $T_n[f]$ with smooth symbols f are computationally more simple and remain that way, for large n. The smoothness of f will get encoded into the decay of singular values of the truncation effect part $K_{f,g}^{(n)}$, in a manner that is essentially independent of n. The performance of the iterative solver depends upon this decay rate, as discussed in [8], [10] and [11]. In other words, it is not the cost of a single iteration step alone that gives the whole price of the computation. We also need to consider how many steps we have to calculate in order to get the required precision. Our conclusion is that the Krylov subspace method (such as GMRES) applied upon the preconditioned system initially converges at increasing speed (or "superlinearly"), until the truncation effect part has been "killed off" and the small Toeplitz matrix part begins to dominate.

We emphasize that our results do not require any normality of the matrices we study. The symbols of Toeplitz matrices can be complex valued continuous functions, and our convergence results are equally valid for the GMRES algorithm for nonsymmetric systems, as they are for the CG algorithm for the symmetric systems. In our approach, the decay of the singular values is the valuable information that we know about the linear operators of interest. The preconditioned system, being a product of two Toeplitz matrices, is not an object whose spectrum is easily available. In this work we try to say as much as possible about the properties of iteration for the preconditioned system, without saying much (nontrivial) about the spectrum.

In the strategy we have adopted, there is a quite unavoidable cost we have to pay. In the final speed estimate for the convergence, a generally unknown constant remains, measuring the ill-conditioning of the preconditioned system. To actually determine this constant, we would have to know the spectrum (with multiplicities) of the preconditioned system. This is the bad news. The good news is that the effect of the constant (or equivalently; the normalization of the polynomial sequence in equation (26)) is not significant in the asymptotics of the estimate, as the iteration number $k \to \infty$. This makes it possible to draw the conclusion about the superlinear convergence of iteration.

Our approach resorts to a multitude of operator theoretic arguments that are based upon the Hankel and Toeplitz structure of the problem. The tools of matrix algebra are not central for us. The abstract numerical analysis framework is mostly from [10]. The treatment here is analogous to that given in the companion paper [9] for infinite dimensional Toeplitz operators. In this sense, the present work differs from what is already done in the literature. We remark that, in comparison to [9], a lot of extra algebraic structure emerges, as we have to study two truncation effects and some detail about their interaction. We conclude that [9] can be seen as an instructive limit case of this work, when the dimension of the Toeplitz matrix becomes infinite.

2 Definitions and basic theory

We use the following notations throughout the paper: $\mathbf{Z}$ is the set of integers. $\mathbf{Z}_+ := \{j \in \mathbf{Z} \quad | \quad j \geq 0\}$. $\mathbf{N} := \{j \in \mathbf{Z} \quad | \quad j > 0\}$. $\mathbf{T}$ is the unit circle of the complex plane. $C(\mathbf{T})$ denotes the class of continuous functions on $\mathbf{T}$ equipped with sup-norm $||\cdot||_\infty$. Given $f \in C(\mathbf{T})$ and $\alpha > 0$, the number $||f||_{Lip_\alpha(\mathbf{T})}$ is defined by

$$(1) \qquad ||f||_{Lip_\alpha}(\mathbf{T}) = ||f||_\infty + \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^\alpha}$$

is called the Lipschitz norm of f. $Lip_\alpha(\mathbf{T}) \subset C(\mathbf{T})$ is the set of such f for which the expression (1) is finite. For $r \in \mathbf{Z}_+$, $C^{r,\alpha}(\mathbf{T})$ are those functions whose r.th derivative is in $Lip_\alpha(\mathbf{T})$. If $\alpha = 0$, then $C^{r,\alpha}(\mathbf{T}) := C^r(\mathbf{T})$. If H is a Hilbert space, then $\mathcal{L}(H)$ denotes the bounded and $\mathcal{LC}(H)$ the compact linear operators in H.

Bi-infinite sequences of complex numbers are denoted $\tilde{a} := \{a_j\}_{j=-\infty}^{\infty}$. The set of such square summable sequences are denoted by $\ell^2(\mathbf{Z})$. We define the following operators in sequence spaces

Definition 1. *Orthogonal projections in $\ell^2(\mathbf{Z})$ are*

(i) the interval projections for $j, k \in \mathbf{Z}$:

$$\pi_{[j,k]}\tilde{a} := \{w_j\}; \quad w_i = a_i \quad \textit{for} \quad j \leq i \leq k, \quad 0 \quad \textit{otherwise};$$
$$\pi_j := \pi_{[j,j]},$$

(ii) the future and past projections:

$$\pi_+ := \pi_{[1,\infty]}, \quad \pi_- := \pi_{[-\infty,-1]},$$

(iii) the composite projections:

$$\bar{\pi}_+ := \pi_0 + \pi_+, \quad \bar{\pi}_- := \pi_0 + \pi_-.$$

We define the spaces of semi-infinite, square summable sequences by $\ell^2(\mathbf{Z}_+) := \bar{\pi}_+\ell^2(\mathbf{Z})$ and $\ell^2(\mathbf{Z}_-) := \pi_-\ell^2(\mathbf{Z})$. The unitary bilateral shift U in $\ell^2(\mathbf{Z})$ is given by

$$(2) \qquad U\tilde{a} := \{w_j\} \quad \text{with} \quad w_j = a_{j-1}.$$

It is well known that the operator polynomials of U and $U^* = U^{-1}$ form a normed commutative subalgebra in $\mathcal{L}(\ell^2(\mathbf{Z}))$, whose norm closure has a particularly simple commutative C^*-algebra structure. This is the content of the following lemma.

Lemma 2. *Let A be the operator norm closure in $\mathcal{L}(\ell^2(\mathbf{Z}))$ of the set of operators $p(U, U^*)$, where p ranges over all polynomials $p(x, y)$ with complex coefficient. Then A is a commutative C^*-algebra that is isometrically isomorphic to $C(\mathbf{T})$. Moreover, there is an isometric isomorphism $\Psi : C(\mathbf{T}) \to A$ which satisfies:*

(i) $\Psi\bar{f} = (\Psi f)^*$

(ii) $(\forall \xi \in \mathbf{T}) : f(\xi) = \xi \implies \Psi f = U$

Proof. First note that $\sigma(U) = \mathbf{T}$, because U is a bilateral shift. Because U and U^* commute, the claim follows from [13, Theorem 11.19]. □

In the light of previous lemma, let us denote the elements of A by $\mathcal{C}[f] := \Psi f$ — the convolution operator with symbol $f \in C(\mathbf{T})$. Now we are in position to introduce the Toeplitz operators and matrices, together with the causal and anti-causal Hankel operators and matrices.

Definition 3. *Let $f \in C(\mathbf{T})$ be arbitrary and $\mathcal{C}[f]$ be the convolution operator with symbol f.*

(i) The Toeplitz operator $\mathcal{T}[f]$ with symbol f is the operator in $\mathcal{L}(\ell^2(\mathbf{Z}_+))$ defined by

$$\mathcal{T}[f] := \bar{\pi}_+ \mathcal{C}[f] \bar{\pi}_+.$$

(ii) The causal (anti-causal) Hankel operator $\mathcal{H}_+[f]$ ($\mathcal{H}_-[f]$) with symbol f is the operator in $\mathcal{L}(\ell^2(\mathbf{Z}_+), \ell^2(\mathbf{Z}_-))$ ($\mathcal{L}(\ell^2(\mathbf{Z}_-), \ell^2(\mathbf{Z}_+))$) defined by

$$\mathcal{H}_+[f] := \bar{\pi}_+ \mathcal{C}[f] \pi_- \quad (\mathcal{H}_-[f] := \pi_- \mathcal{C}[f] \bar{\pi}_+).$$

(iii) The $n \times n$ Toeplitz matrix $T_n[f]$ with symbol f is the operator in $\mathcal{L}(\ell^2(\mathbf{Z}_+))$ defined by

$$T_n[f] := \pi_{[0,n-1]} \mathcal{C}[f] \pi_{[0,n-1]}$$

(iv) The $n \times n$ causal (anti-causal) Hankel matrix $H_+[f]$ ($H_-[f]$) with symbol f is the operator in $\mathcal{L}(\ell^2(\mathbf{Z}))$ defined by

$$H_+[f] := \pi_{[0,n-1]} \mathcal{C}[f] \pi_{[-n,-1]} \quad (H_-[f] := \pi_{[-n,-1]} \mathcal{C}[f] \pi_{[0,n-1]}).$$

In this paper, all the $n \times n$ matrices are regarded as operators of forms $\pi_{[0,n-1]} T \pi_{[0,n-1]}$, $\pi_{[0,n-1]} T \pi_{[-n,-1]}$ or $\pi_{[-n,-1]} T \pi_{[0,n-1]}$, where $T \in \mathcal{L}(\ell^2(\mathbf{Z}))$. The identification of $\mathbf{C}^n$ (with Euclidean inner product) and range $\left(\pi_{[0,n-1]}\right)$ is obvious. With this identification, all the matrices are naturally interpreted as finite dimensional operators in space $\ell^2(\mathbf{Z})$. The matrices are regarded invertible (or nonsingular) in $\mathbf{C}^n$ when they are bijections in range $\left(\pi_{[0,n-1]}\right)$. Clearly this is not equivalent with the invertibility when the matrix is regarded as an operator in the whole space. Inverse of a nonsingular $n \times n$ matrix T_n is defined by the Moore-Penrose pseudoinverse

$$T_n^{-1} = \lim_{\nu \to 0+} (T_n^* T_n + \nu I_n)^{-1} T_n^* \in \mathcal{L}(\ell^2(\mathbf{Z}_+)),$$

where I_n denotes the identity matrix of $\mathbf{C}^n$, identified with the projection $\pi_{[0,n-1]}$. Clearly this definition of matrix invertibility is in harmony with the usual notion of nonsingularity of a matrix. The ordinary inverse matrix is just extended with zeroes, from $\mathbf{C}^n$ to the whole of $\ell^2(\mathbf{Z}_+)$.

The symbols of Hankel operators, Hankel matrices and Toeplitz matrix are not unique. However, as $n \to \infty$, all the Fourier coefficients of f will appear in infinitely many Toeplitz matrices $T_n[f]$. In this paper, we are more interested in the families $\{T_n[f]\}_{n\in\mathbf{N}}$ of Toeplitz matrices with a common symbol f, rather than any of $T_n[f]$ separately. A special attention is payed to the limit processes as $n \to \infty$, but f is kept fixed. In some sense, this resolves the nonuniqueness problem of the symbol

The following norm estimates are basic:

Lemma 4. *It is true that for any $f \in C(\mathbf{T})$ and $n \in \mathbf{N}$*

$$||\mathcal{T}[f]|| = ||f||_\infty, \tag{3}$$

$$||\mathcal{H}_\pm[f]|| \le ||f||_\infty, \tag{4}$$

$$||T_n[f]|| \le ||f||_\infty, \quad ||H^{(n)\pm}[f]|| \le ||\mathcal{H}_\pm[f]||. \tag{5}$$

Proof. For equation (3) see [6]. Estimate (4) is an easy consequence of Lemma 2 giving the norm of the convolution operator $||\mathcal{C}[f]|| = ||f||_\infty$ and the fact that the norm of the orthogonal projections equal 1. The same comment goes also for equations (5). $\square$

Note that the estimate (5) for Hankel norm is nothing but optimal. The Nehari extension theorem characterizes the norm of Hankel operator $\mathcal{H}_+[f]$ as the infimum $\inf ||f+g||$ over all g with vanishing positively indexed Fourier coefficients. We use the estimate (5) to connect the smoothness of the symbol (via trigonometric polynomials and Jackson's theorems) to the singular value decay of certain Hankel operator.

The invertibility condition for Toeplitz operators in terms of symbols is quite simple to state, see [9] or [6]:

Proposition 5. *Let $f \in C(\mathbf{T})$ be arbitrary. Then $\mathcal{T}[f]$ is invertible if and only if $0 \notin f(\mathbf{T})$ and $Ind\,(f) = 0$, where*

$$Ind\,(f) := \frac{1}{2\pi}\, arg\, f(e^{i\theta})|_0^{2\pi}$$

For Toeplitz matrices things are not so simple. A simple counter example shows that a given nonunique symbol does not determine the invertibility of $T_n[f]$ as in Proposition 5, but the dimension n is also important. The following lemma is a slight modification of [6, Theorem 2.1].

Lemma 6. *Let $f \in C(\mathbf{T})$ be arbitrary. Then the following are equivalent:*

(i) $\liminf_{n\to\infty} ||T_n[f]^{-1}|| < \infty$,

(ii) *f satisfies the invertibility condition of Proposition 5,*

(iii) *There is a $N(f) \in \mathbf{N}$ such that $T_n[f]$ is invertible in $\mathbf{C}^n$ for $n > N(f)$. Furthermore, if $\tilde{b} \in \ell^2(\mathbf{Z}_+)$ is arbitrary, then $\lim_{n\to\infty} \tilde{a}_n = \tilde{a}$, where $\tilde{b}_n = T_n[f]\tilde{a}_n$ and $\tilde{b} = \mathcal{T}[f]$.*

Proof. The implications (i) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (iii) follow from [6, Theorem 2.1]. Assume that claim (iii) holds. Let $n > N(f)$. Now

$$T_n[f]^{-1}\tilde{b} = T_n[f]^{-1}\tilde{b}_n = \tilde{a}_n \to \tilde{a} := \mathcal{T}[f]^{-1}\tilde{b}$$

for all $\tilde{b} \in \ell^2(\mathbf{Z}_+)$ by assumption, where $\tilde{b}_n = \pi_{[0,n-1]}\tilde{b}$. It follows from the Banach–Steinhaus Theorem that $||T_n[f]^{-1}|| \le C < \infty$ for $n \ge N(f)$. But then $\liminf_{n\to\infty} ||T_n[f]^{-1}|| \le \limsup_{n\to\infty} ||T_n[f]^{-1}|| \le C < \infty$, and claim (i) holds. $\square$

We complete this section with a word of explanation. In [6], the Toeplitz operators are defined as the closure of polynomials $p(S, S^*)$ for the unilateral shift S on $\ell^2(\mathbf{Z}_+)$. In our earlier work [9], S was unitarily dilated into a bilateral shift. By using functional calculus (i.e. Lemma 2), we obtain discrete convolution operators, associated to symbols $f \in \mathbf{C}(\mathbf{T})$. Finally, the Toeplitz operators were recovered as compressions of the convolution operators. In discrete time control theory, one could regard the bilateral (time) shift the starting object, and unilateral shift only something that we get when future and past projections (causality) enter the game.

Now, when treating the Toeplitz matrices, the unilateral shift has lost much of its special appeal over the bilateral shift. For this reason, we have here only introduced the bilateral shift and its polynomials as convolution operators, and Toeplitz matrices as their compressions. Note that, in a sense, (inverse of the) Toeplitz matrix is simultaneously an "approximation" to both the (inverse of the) convolution operator as well as to the (inverse of the) Toeplitz operator, all sharing the same symbol. Because the convolution and Toeplitz operators are two fundamentally different kinds of objects, the "approximation process" is rather interesting. This work, together with [9], sheds some indirect light upon this process through the study of the truncation effect matrices, as defined later. Trivially, the Toeplitz matrices could be described by purely matrix algebraic notions, but this would be inconvenient in the present approach; the description of the truncation effect requires Hankel operators, and thus either unilateral of bilateral shifts.

The reader is instructed to see that all the different (but equivalent) ways of introducing the Toeplitz operators and matrices are matters of technical convenience, preferences and traditions, rather than reflections of such deep mathematical structure that would require some particular type of formalism.

3 Preconditioning of Toeplitz matrices

In this section, we develop a Toeplitz preconditioning theory for Toeplitz matrices, in the spirit of our earlier work [9] for Toeplitz operators. The dimension of the Toeplitz matrix is denoted by $n \in \mathbf{N}$. By $\tilde{a} = \{a_j\}_{j=0}^{\infty}$ denote a sequence in $\ell^2(\mathbf{Z}_+)$, and define $\tilde{a}_n := \pi_{[0,n-1]}\tilde{a}$. Now $\tilde{a}_n$ can be regarded as an arbitrary element of $\mathbf{C}^n$, as considered above. Our problem is to find $\tilde{b}_n \in \mathbf{C}^n$ satisfying

$$\tilde{a}_n = T_n[f]\tilde{b}_n \tag{6}$$

for all n large enough. For this to be possible, we assume that the conditions of Lemma 6 hold. Let us multiply the both sides of equation (6) by an invertible Toeplitz matrix $T_n[g]$ for a suitably chosen symbol g. This gives:

$$T_n[g]\tilde{a}_n = T_n[g]T_n[f]\tilde{b}_n,$$

or equivalently, in the form of a fixed point problem:

$$\tilde{b}_n = (I_n - T_n[g]T_n[f])\tilde{b}_n + T_n[g]\tilde{a}_n, \tag{7}$$

where I_n is the identity matrix of $\mathbf{C}^n$. We say that the equation (7) has been Toeplitz preconditioned, at least if $T_n[g]$ is in some sense close to $T_n[f]^{-1}$, see [1] and [2]. The matrix $I_n - T_n[g]T_n[f]$ is almost Toeplitz, but not quite. The following decomposition theorem makes this point precise:

Theorem 7. *Assume that $f, g \in C(\mathbf{T})$ and $n \in \mathbf{N}$. Then $I_n - T_n[g]T_n[f]$ can be decomposed as:*

$$\begin{aligned} I_n - T_n[g]T_n[f] &= (T_n[gf] - T_n[g]T_n[f]) + T_n[1-gf]) \\ &=: K_{f,g}^{(n)} + B_{f,g}^{(n)}, \end{aligned} \tag{8}$$

where $K_{f,g}^{(n)\pm}$ is the $n \times n$ matrix given by

$$K_{f,g}^{(n)} := K_{f,g}^{(n)+} + K_{f,g}^{(n)-}, \tag{9}$$

with the $n \times n$ matrices

$$\begin{aligned} K_{f,g}^{(n)+} &:= \pi_{[0,n-1]}\mathcal{H}_+[g]\mathcal{H}_-[f]\pi_{[0,n-1]} \\ K_{f,g}^{(n)-} &:= \pi_{[0,n-1]}U^n\mathcal{H}_-[g]\mathcal{H}_+[f]U^{*n}\pi_{[0,n-1]}. \end{aligned} \tag{10}$$

The matrix $B_{f,g}^{(n)}$ is a $n \times n$ Toeplitz matrix.

Proof. The claim of equation (8) if trivial, because the mapping $f \mapsto T_n[f]$ is linear. In order to prove equations (9) and (10), we write

$$\begin{aligned} K_{f,g}^{(n)} &:= \pi_{[0,n-1]}\left(\mathcal{C}[gf] - \mathcal{C}[g]\pi_{[0,n-1]}\mathcal{C}[f]\right)\pi_{[0,n-1]} \\ &= \pi_{[0,n-1]}\left(\mathcal{C}[gf] - \mathcal{C}[g](\mathcal{I} - \pi_- - \pi_{[n,\infty]})\mathcal{C}[f]\right)\pi_{[0,n-1]} \\ &= \pi_{[0,n-1]}\left(\mathcal{C}[g]\pi_-\mathcal{C}[f]\right)\pi_{[0,n-1]} + \pi_{[0,n-1]}\left(\mathcal{C}[g]\pi_{[n,\infty]}\mathcal{C}[f]\right)\pi_{[0,n-1]}, \end{aligned} \tag{11}$$

where the last equality holds because $\mathcal{C}[gf]-\mathcal{C}[g]\mathcal{C}[f]=0$ by Lemma 2. The first term in the left of (11) is equal to $\pi_{[0,n-1]}\mathcal{H}_+[g]\mathcal{H}_-[f]\pi_{[0,n-1]}=K_{f,g}^{(n)+}$. The equations (9),(10) are proved if we show that the latter term in the left of (11) is equal to $\pi_{[0,n-1]}U^n\mathcal{H}_-[g]\mathcal{H}_+[f]U^{*n}\pi_{[0,n-1]}=K_{f,g}^{(n)-}$.

It is a matter of an easy computation that $\pi_{[0,n-1]}\mathcal{C}[g]\pi_{[n,\infty]}$ $=\pi_{[0,n-1]}U^n\mathcal{H}_-[g]U^{*n}$ and $\pi_{[n,\infty]}\mathcal{C}[f]\pi_{[0,n-1]}=U^n\mathcal{H}_+[f]U^{*n}\pi_{[0,n-1]}$. Combination of these gives

$$\begin{aligned}&\pi_{[0,n-1]}\left(\mathcal{C}[g]\pi_{[n,\infty]}\mathcal{C}[f]\right)\pi_{[0,n-1]}=\pi_{[0,n-1]}\mathcal{H}_-[f]U^nU^{*n}\mathcal{H}_+[f]\pi_{[0,n-1]}\\&=\pi_{[0,n-1]}\mathcal{H}_-[f]\mathcal{H}_+[f]\pi_{[0,n-1]}=K_{f,g}^{(n)-}.\end{aligned}$$

This completes the proof of equations (9) and (10). To conclude the proof, we note that $B_{f,g}^{(n)}$ is Toeplitz by definition. ☐

Now that we have our basic objects in hands, it it time to name them. We propose the following:

Definition 8. *Let $K_{f,g}^{(n)\pm}$, $B_{f,g}^{(n)}$ be as in Theorem 7. The matrix $K_{f,g}^{(n)}$ is the truncation effect matrix of order n. The matrices $K_{f,g}^{(n)+}$, $K_{f,g}^{(n)-}$ are the upper and lower truncation effect matrix of order n, respectively. The matrix $B_{f,g}^{(n)}$ is called the perturbation matrix of order n.*

We call matrix $B_{f,g}^{(n)}$ perturbation matrix because it is regarded an an unstructured perturbation to the (compact) truncation effect $K_{f,g}^{(n)}$ in the frame of reference of [8], as studied in Section 5. An analogous theorem to Theorem 7 for Toeplitz operators is [9, Theorem 3.1]. There we introduced the truncation effect operator $K_{f,g}:=\mathcal{H}_+[g]\mathcal{H}_-[f]$. Clearly $K_{f,g}^{(n)+}=\pi_{[0,n-1]}K_{f,g}\pi_{[0,n-1]}$. We proceed to discuss the implications of Theorem 7, especially from the numerical analysis point of view. We also compare the analogous Toeplitz matrix and operator results.

We first remark the that left hand side of (8) does not depend upon the Fourier coefficients of f, g with index j satisfying $|j|>n$. However, both the operators $K_{f,g}^{(n)}$, $B_{f,g}^{(n)}$ depend on all the Fourier coefficients of f and g. Different choices of the symbols f, g give different decompositions in the right of (8) for the same preconditioned Toeplitz systems given in the left of (8). This is in contrast to the case of Toeplitz operators where $\mathcal{T}[f]$ and its symbol $f\in C(\mathbf{T})$ are in bijective correspondence. The Toeplitz operator truncation effect is always compact for $f\in C(\mathbf{T})$, see [8, Theorem 3.3]. A Toeplitz operator is compact if and only if it vanishes, by a spectral argument. It follows that the decomposition of $\mathcal{I}-\mathcal{T}[g]\mathcal{T}[f]$ into truncation effect $K_{f,g}$ and perturbation operator $B_{f,g}$ is unique in [9, Theorem 3.1]. For the Toeplitz matrices, $K_{f,g}^{(n)}$ can be even Toeplitz; consider the circulant example $f(e^{i\theta}):=e^{-i(n-1)\theta}+e^{in\theta}$ and $g(e^{i\theta}):=f(e^{-i\theta})$. Then

$$K_{f,g}^{(n)}=T_n[fg]-T_n[g]T_n[f]=2I_n-I_n=I_n.$$

However, this example works only because the symbols f and g depend on the dimension n of the Toeplitz matrix. For the same f and g as above, $K_{f,g}^{(n+1)}$ is no longer Toeplitz. Because our results are stated for "large n", these "accidents" play no significant role.

Fix $f, g \in \mathbf{C}(\mathbf{T})$. The truncation effect can further be decomposed as

$$K_{f,g}^{(n)} = K_{f,\frac{1}{f}}^{(n)} + K_{f,(g-\frac{1}{f})}^{(n)}. \tag{12}$$

The first part $K_{f,\frac{1}{f}}^{(n)}$ is a measure how far from Toeplitz the matrix $T_n[f]^{-1}$ is. Namely,

$$T_n[f]^{-1} - T_n[\frac{1}{f}] = K_{f,\frac{1}{f}}^{(n)} T_n[f]^{-1}.$$

This part is independent of the preconditioners symbol g. The singular value decay of the right hand side can be bounded above by an estimate not depending on n, if some smoothness of f is assumed. The latter part $K_{f,(g-\frac{1}{f})}^{(n)}$ in (12) is due to the nonoptimality of the preconditioning symbol g. As n increases, the major part of the computational cost of nonoptimal preconditioning is in the Toeplitz perturbation matrix $B_{f,g}^{(n)}$, not in $K_{f,(g-\frac{1}{f})}^{(n)}$, as will be implied by this work.

The rest of this section is dedicated to a more detailed study of the matrices $K_{f,g}^{(n)\pm}$. We present some connections to the Toeplitz operator case as studied in [9]. Also some results are established that lead to the proof of Theorem 20. It is interesting to see in what sense the Toeplitz matrix case given in Theorem 7 relates to the Toeplitz operator case given in [9] as $n \to \infty$. Lemma 10 gives us the result, but first we need a functional analytic proposition.

Proposition 9. *Let H be a Hilbert space and $B, B_n \in \mathcal{L}(H)$ for $n \in \mathbf{N}$. Let $K \in \mathcal{LC}(H)$. Then the following holds:*

(i) If $B_n x \to Bx$ for all $x \in H$ (i. e. $B_n \to B$ strongly), then $||B_n K - BK|| \to 0$.

(ii) If $B_n^* x \to B^* x$ for all $x \in H$, then $||KB_n - KB|| \to 0$.

Lemma 10. *Let the operators $K_{f,g}$, $K_{f,g}^{(n)\pm}$ be defined in Theorem 7. Then*

$$\lim_{n\to\infty} ||K_{f,g} - K_{f,g}^{(n)+}|| = 0 \tag{13}$$

and

$$\lim_{n\to\infty} ||K_{f,g}^{(n)-} \tilde{a}|| = 0 \quad \textit{for all} \quad \tilde{a} \in \ell^2(\mathbf{Z}_+), \tag{14}$$

i. e. $K_{f,g}^{(n)-} \to 0$ strongly. Moreover, $K_{f,g}^{(n)} \to K_{f,g}$ strongly.

Proof. To prove (13), write

$$K_{f,g} - K_{f,g}^{(n)+} = \pi_{[n,\infty]} K_{f,g} + \pi_{[0,n-1]} K_{f,g} \pi_{[n,\infty]}.$$

Now use Proposition 9 with $B_n = B_n^* = \pi_{[n,\infty]}$ and $B = 0$. Because $K_{f,g}$ is compact (see [9]), equation (13) follows.

The proof of (14) is somewhat more technical. Because

$$K_{f,g}^{(n)-} = \pi_{[0,n-1]} U^n \mathcal{H}_-[g] \mathcal{H}_+[f] U^{*n} \pi_{[0,n-1]}$$

by Theorem 7 and $\pi_{[0,n-1]} U^n \mathcal{H}_-[g]$ is bounded, it suffices to show that $\mathcal{H}_+[f] U^{*n} \pi_{[0,n-1]} \to 0$ strongly. To this end, let $\epsilon > 0$ and $\tilde{a} \in \ell^2(\mathbf{Z}_+)$ be arbitrary. Choose $m \in \mathbf{N}$ so large that $||\pi_{[m,\infty]} \tilde{a}|| < \epsilon/(2||f||_\infty)$. It is a matter of easy manipulation to show that $\mathcal{H}_+[f] U^{*n} \pi_{[0,n-1]} = (\bar{\pi}_+ U^{*n}) \mathcal{C}[f] \pi_{[0,n-1]}$. Using this we may estimate for $n \leq m$:

(15)
$$||\mathcal{H}_+[f] U^{*n} \pi_{[0,n-1]} \tilde{a}|| \leq ||(\bar{\pi}_+ U^{*n}) \mathcal{C}[f] \pi_{[0,m-1]} \tilde{a}|| + ||(\bar{\pi}_+ U^{*n}) \mathcal{C}[f] \pi_{[m,\infty]} \tilde{a}||$$

The second part of (14) is less than $\epsilon/2$, because $||(\bar{\pi}_+ U^{*n})|| = 1$, $||\mathcal{C}[f]|| = ||f||_\infty$ and $||\pi_{[m,\infty]} \tilde{a}|| < \epsilon/(2||f||_\infty)$. The first part is under $\epsilon/2$ if n is large enough, because the unilateral backward shift $\bar{\pi}_+ U^{*n} \to 0$ strongly. This completes the proof. ☐

The operator sequence $K_{f,g}^{(n)-}$ does not generally converge in norm. When it does, the operator family itself is trivial:

Corollary 11. *Let the operators $K_{f,g}$, $K_{f,g}^{(n)\pm}$ be defined in Theorem 7. Then the following are equivalent:*

(i) $K_{f,g}^{(n)} \to K_{f,g}$ *in norm,*

(ii) $K_{f,g}^{(n)-} \to 0$ *in norm,*

(iii) $\mathcal{H}_-[g] \mathcal{H}_+[f] = 0$ *for all* $n \in \mathbf{N}$,

(iv) $K_{f,g}^{(n)} = 0$.

Proof. The only nontrivial part is to check that $K_{f,g}^{(n)-} \to 0$ in norm implies $\mathcal{H}_-[g] \mathcal{H}_+[f] = 0$. So assume that $K_{f,g}^{(n)-} \to 0$ in norm. By $\{e_i\}_{i \in \mathbf{Z}}$ denote the natural basis of $\ell^2(\mathbf{Z})$. Let $i, j \in \mathbf{Z}_-$, and $n < \max(-i, -j)$. An elementary calculation, based upon formula (10), gives the following:

$$\langle e_i, \mathcal{H}_-[g] \mathcal{H}_+[f] e_j \rangle = \left\langle e_{i+n}, K_{f,g}^{(n)-} e_{j+n} \right\rangle \tag{16}$$

The left hand side of (16) does not depend upon n. The right hand side does and approaches zero, because $K_{f,g}^{(n)-} \to 0$ in norm. It follows that $\langle e_i, \mathcal{H}_-[g] \mathcal{H}_+[f] e_j \rangle = 0$ for all $i, j \in \mathbf{Z}_-$, and the proof is complete. ☐

The development of Lemma 10 and Corollary 11 has an important implication from the numerical analysis point of view. If one is inverting an infinite dimensional Toeplitz operator by preconditioning and iterating a finite dimensional section of it (i.e. a Toeplitz matrix), the iterative solver will attack both $K_{f,g}^{(n)+}$ and $K_{f,g}^{(n)-}$. Only the data in $K_{f,g}^{(n)+}$ is present in the limit case of the infinite dimensional Toeplitz operator. However, $K_{f,g}^{(n)-}$ is always present for all large n, if for any n at all, by Corollary 11. It is by formula (10), modulo truncation and unitary shift, a product of two Hankel operators — a structure numerically quite comparable to $K_{f,g}^{(n)+}$.

The Proposition 12 and Lemmas 13, 14 are results that we need in our main result, Corollary 22. We start with a fundamental symmetry between $K_{f,g}^{(n)+}$ and $K_{f,g}^{(n)-}$.

Proposition 12. *Let $f, g \in C(\mathbf{T})$. Then we have the unitary equivalence*

$$K_{f,g}^{(n)-} = Flip_n^* \cdot K_{\tilde{f},\tilde{g}}^{(n)+} \cdot Flip_n, \tag{17}$$

where $\tilde{f}(e^{i\theta}) = f(e^{-i\theta})$, $\tilde{g}(e^{i\theta}) = g(e^{-i\theta})$, and the operator $Flip_n : \operatorname{range}\left(\pi_{[0,n-1]}\right) \to \operatorname{range}\left(\pi_{[0,n-1]}\right)$ is the permutation

$$Flip_n \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} u_{n-1} \\ \vdots \\ u_1 \\ u_0 \end{pmatrix} \tag{18}$$

Proof. Note first that the Fourier coefficients of $\tilde{f} \in C(\mathbf{T})$ ($\tilde{g} \in C(\mathbf{T})$) are related to the Fourier coefficients of f (g) by

$$\tilde{f}_j = f_{-j} \quad (\tilde{g}_j = g_{-j}) \quad \text{for all} \quad j \in \mathbf{Z}. \tag{19}$$

Define the operator $flip : \ell^2(\mathbf{Z}; U) \to \ell^2(\mathbf{Z}; U)$ by $(flip\, \tilde{u})_j = u_{-j}$ for all $j \in \mathbf{Z}$ and $\tilde{u} \in \ell^2(\mathbf{Z}; U)$. With this notation, we can show that

$$\mathcal{H}_-[g]\mathcal{H}_+[f] = U^* \, flip \cdot \mathcal{H}_+[\tilde{g}]\mathcal{H}_-[\tilde{f}] \cdot flip \; U,$$

where U is the bilateral shift given in equation (2), and $\tilde{f}$, $\tilde{g}$ are given by (19). Now

$$\begin{aligned} K_{f,g}^{(n)-} &:= \pi_{[0,n-1]} U^n \mathcal{H}_-[g]\mathcal{H}_+[f] U^{*n} \pi_{[0,n-1]} \\ &= \left(\pi_{[0,n-1]} U^n\right) \left(U^* \cdot flip \cdot \mathcal{H}_+[\tilde{g}]\mathcal{H}_-[\tilde{f}] \cdot flip \cdot U\right) \left(U^{*n} \pi_{[0,n-1]}\right) \\ &= \left(\pi_{[0,n-1]} U^{n-1}\, flip\right) \cdot \left(\mathcal{H}_+[\tilde{g}]\mathcal{H}_-[\tilde{f}]\right) \left(flip\, U^{*(n-1)} \pi_{[0,n-1]}\right). \end{aligned}$$

By looking at (18), $Flip_n = \pi_{[0,n-1]} U^{n-1}\, flip = \pi_{[0,n-1]} U^{n-1}\, flip\, \pi_{[0,n-1]}$. This, together with the above calculation implies equation (17). ☐

Note that the matrix $Flip_n$ gives a unitary equivalence of a Toeplitz matrix $T_n[f]$ to its transpose $T_n[\tilde{f}]$.

Claim (i) of the following Lemma has been used in the construction of the numerical example in [8, Section 5]. Claim (ii) has a direct application in Corollary 22, one of the main results of this paper.

Lemma 13. *Assume that $f, g \in C(\mathbf{T})$. Define*

$$\begin{aligned}\tilde{K}_{f,g}^{(n)+} &:= \pi_{[0,\lfloor n/2 \rfloor]} K_{f,g}^{(n)+} \pi_{[0,\lfloor n/2 \rfloor]} = K_{f,g}^{(\lfloor n/2 \rfloor)+}, \\ \tilde{K}_{f,g}^{(n)-} &:= \pi_{[\lceil n/2 \rceil, n]} K_{f,g}^{(n)-} \pi_{[\lceil n/2 \rceil, n]}, \\ \tilde{K}_{f,g}^{(n)} &:= \tilde{K}_{f,g}^{(n)+} + \tilde{K}_{f,g}^{(n)-},\end{aligned}$$

where $\lfloor j \rfloor$ is the integral part of $j \in \mathbf{Z}$, and $\lceil j \rceil = \lfloor j \rfloor + 1$. Then

(i) $\lim_{n\to\infty} ||K_{f,g}^{(n)} - \tilde{K}_{f,g}^{(n)}|| = 0$, *and*

(ii) $\lim_{n\to\infty} ||K_{f,g}^{(n)}|| = \lim_{n\to\infty} ||\tilde{K}_{f,g}^{(n)}|| = \max(||K_{f,g}||, ||K_{\tilde{f},\tilde{g}}||)$.

Proof. In order to prove claim (i), we estimate:

$$(20) \qquad ||K_{f,g}^{(n)} - \tilde{K}_{f,g}^{(n)}|| \le ||K_{f,g}^{(n)+} - \tilde{K}_{f,g}^{(n)+}|| + ||K_{f,g}^{(n)-} - \tilde{K}_{f,g}^{(n)-}||.$$

The estimation of the first part in the right hand side of (20) goes as follows:

$$||K_{f,g}^{(n)+} - \tilde{K}_{f,g}^{(n)+}|| \le ||K_{f,g}^{(n)+} - K_{f,g}|| + ||K_{f,g} - \tilde{K}_{f,g}^{(n)+}||.$$

Because, in fact $\tilde{K}_{f,g}^{(n)+} = K_{f,g}^{(\lfloor n/2 \rfloor)+}$, the both terms in the right hand side converge to zero, by equation (13) of Lemma 10. The second part of equation (20) is similar, when we use the identity

$$K_{f,g}^{(n)-} - \tilde{K}_{f,g}^{(n)-} = Flip_n^* (K_{\tilde{f},\tilde{g}}^{(n)+} - \tilde{K}_{\tilde{f},\tilde{g}}^{(n)+}) Flip_n,$$

implied by Proposition 12. Claim (ii) is a consequence of the first claim. We obtain

$$(21) \qquad ||\tilde{K}_{f,g}^{(n)}|| - ||K_{f,g}^{(n)} - \tilde{K}_{f,g}^{(n)}|| \le ||K_{f,g}^{(n)}|| \le ||\tilde{K}_{f,g}^{(n)}|| + ||K_{f,g}^{(n)} - \tilde{K}_{f,g}^{(n)}||,$$

which implies $\lim_{n\to\infty} ||K_{f,g}^{(n)}|| = \lim_{n\to\infty} ||\tilde{K}_{f,g}^{(n)}||$. But $||\tilde{K}_{f,g}^{(n)}|| = \max(||\tilde{K}_{f,g}^{(n)+}||, ||\tilde{K}_{f,g}^{(n)-}||)$, because $\tilde{K}_{f,g}^{(n)+}$ and $\tilde{K}_{f,g}^{(n)-}$ operate in their disjoint reducing subspaces; this was the reason why we defined them in the first place. By equation (13) of Lemma 10, and Proposition 12, we have

$$\lim_{n\to\infty} ||\tilde{K}_{f,g}^{(n)+}|| = ||K_{f,g}||, \quad \lim_{n\to\infty} ||\tilde{K}_{f,g}^{(n)-}|| = ||K_{\tilde{f},\tilde{g}}||.$$

It now follows that $\max(||\tilde{K}_{f,g}^{(n)+}||, ||\tilde{K}_{f,g}^{(n)-}||) \to \max(||K_{f,g}||, ||K_{\tilde{f},\tilde{g}}||)$, and the proof is complete. □

In Corollary 22, we need to exclude the situations when $K_{f,g}^{(n)}$ vanishes. This is because the convergence speed estimate (25) we use, has a norm of the truncation effect in the denominator. Given a fixed n, there are two totally different situations when this happens. The situation of the first kind appears when $K_{f,g}^{(n)} = 0$ for all n large enough. This can happen only when $K_{f,g}^{(n)+} = K_{f,g}^{(n)-} = 0$ for all n, as can be shown by a similar calculation as in the proof of Corollary 11. This first case is explicitly excluded in the assumptions of Corollary 22, by using the following Lemma 14. The situation of the second kind appears when generally $K^{(n)}_{f,g} \neq 0$ except for some finite number of particular $n \in \mathbf{N}$. Two Toeplitz matrices $T_n[f]$, $T_n[g]$ with analytic symbols have vanishing truncation effects $K^{(n)}_{f,g} = 0$ for all n. A small co-analytic perturbation to such f and g will cause that $K^{(n)}_{f,g} = 0$ only for n small. This second case is excluded in the assumptions of Corollary 22, by saying that the result only holds for large n.

So, we need only be able to identify the first kind of situation, as described above. The following lemma is the result we need, stated only for $K^{(n)+}_{f,g}$. A similar result for the lower truncation effect $K^{(n)-}_{f,g}$ can be obtained by an application of Proposition 12.

Lemma 14. *Let $f, g \in C(\mathbf{T})$. Then the following are equivalent:*

(i) $K^{(n)+}_{f,g} = 0$ *for all* n,

(ii) $K_{f,g} = 0$,

(iii) $\mathcal{H}_+[g] = 0$ *or* $\mathcal{H}_-[f] = 0$.

Proof. (i) $\Rightarrow$ (ii) follows because $0 \equiv K^{(n)+}_{f,g} \to K_{f,g}$ in operator norm as $n \to \infty$, by equation (13) of Lemma 10. The implication (ii) $\Rightarrow$ (iii) is somewhat nontrivial. Because $K_{f,g} = \mathcal{H}_+[g]\mathcal{H}_-[f]$ it suffices to show that if a product of two arbitrary Hankel operators (with complex valued symbols) vanishes, the at least one of the Hankel operators vanishes. By multiplying two semi-infinite Hankel matrices, we obtain for the matrix element

$$\pi_i \mathcal{H}_+[g]\mathcal{H}_-[f]\pi_j = \sum_{k,l \geq 1} g_{k+i} f_{-k-j} \quad \text{for all} \quad i, j \geq 0, \tag{22}$$

where f_j, g_j are the Fourier coefficients of f, g. Note that the sum in (22) must converge, because the Hankel operators are bounded. Now $\pi_i \mathcal{H}_+[g]\mathcal{H}_-[f]\pi_j = 0$ for all $i, j \geq 0$ by assumption. Furthermore, $\pi_{i+1}\mathcal{H}_+[g]\mathcal{H}_-[f]\pi_{j+1} - \pi_i\mathcal{H}_+[g]\mathcal{H}_-[f]\pi_j = g_i f_{-j} = 0$, for all $i, j \geq 0$. For definiteness, assume that $\mathcal{H}_+[g] \neq 0$. Then $g_{i'} \neq 0$ for some $i' \geq 0$. Because $g_{i'} f_{-j} = 0$ for all $j \geq 0$, it follows that $f_{-j} = 0$ for all $j \geq 0$, which is equivalent to saying that $\mathcal{H}_-[f] = 0$. Thus at least one of $\mathcal{H}_+[g], \mathcal{H}_-[f]$ vanishes. The remaining part (iii) $\Rightarrow$ (i) is trivial. This completes the proof. $\square$

We conclude this section with the following dimension lemma which has a direct application in Theorem 19 that has been used in [8, Section 5]. It gives a basic approximation property of the truncation effect operators. By trigonometric polynomial in $\mathbf{T}$, we mean the finite sums of form

$$h(e^{i\theta}) = \sum_{j=-\nu}^{\nu} h_j e^{ij\theta}, \quad \nu \in \mathbf{Z}_+.$$

The least ν such that $h_j = 0$ for all j such that $|j| \geq \nu + 1$ is the degree of the polynomial, and denoted by $\deg h$. In many practical applications, the symbol g of the preconditioner matrix $T_n[g]$ would be a trigonometric polynomial.

Lemma 15. *Assume that $f, g \in C(\mathbf{T})$, $n \in \mathbf{N}$.*

(i) If f is trigonometric polynomial with $\nu_1 := \deg f$ and $n > \nu_1$, then

$$K_{f,g}^{(n)+} = K_{f,g}^{(n)+} \pi_{[0,\nu_1-1]}, \quad K_{f,g}^{(n)-} = K_{f,g}^{(n)-} \pi_{[n-\nu_1,n-1]}.$$

(ii) If g is trigonometric polynomial $\nu_2 := \deg f$ and $n > \nu_2$, then

$$K_{f,g}^{(n)+} = \pi_{[0,\nu_2-1]} K_{f,g}^{(n)+}, \quad K_{f,g}^{(n)-} = \pi_{[n-\nu_2,n-1]} K_{f,g}^{(n)-}.$$

(iii) If both f, g are trigonometric polynomials, and $n \geq 2\nu$, where $\nu := \max(\deg f, \deg g)$, then the upper and lower truncation effects $K_{f,g}^{(n)+}$, $K_{f,g}^{(n)-}$ operate in their reducing subspaces $\operatorname{range}\left(\pi_{[0,\nu-1]}\right)$, $\operatorname{range}\left(\pi_{[n-\nu+1,\nu-1]}\right)$, respectively.

(iv) If at least one of f, g is a trigonometric polynomial, then $\operatorname{rank} K_{f,g}^{(n)\pm} \leq \min(n, \deg f, \deg g)$ and $\operatorname{rank} K_{f,g}^{(n)\pm} \leq \min(n, 2\deg f, 2\deg g)$

Proof. All the formulae in claims (i) and (ii) are quite similar consequences of the easily proved equations

$$\mathcal{H}_+[f] = \pi_{[0,\nu_1-1]} \mathcal{H}_+[f] \pi_{[-\nu_1,-1]}, \quad \mathcal{H}_-[f] = \pi_{[-\nu_1,-1]} \mathcal{H}_-[f] \pi_{[0,\nu_1-1]}$$

where f trigonometric polynomial and $\nu_1 = \deg f$. Claims (iii) and (iv) are follow immediately from claims (i) and (ii). $\square$

4 Smoothness of symbols

In this section we study how the smoothness of the symbol $f \in C(\mathbf{T})$ affect the properties of iteration of the preconditioned system (7). We start with recalling the definitions of approximation numbers and Schatten classes of compact operators. A good reference for these is [5, p. 1089 - 1119].

Definition 16. *Let $T \in \mathcal{L}(\ell^2(\mathbf{Z}_+))$ and $k \in \mathbf{N}$. The approximation numbers by finite dimensional operators are defined by:*

$$\sigma_k(T) := \inf_{\operatorname{rank} F \leq k-1} ||T - F||$$

In a Hilbert space the approximation numbers $\sigma_k(T)$ equal the singular values of T. The closed ideal of compact operators $\mathcal{LC}(\ell^2(\mathbf{Z}_+))$ can now be divided into smaller spaces, if we look at the decay of the singular values. Consider the following definition:

Definition 17. *Let $p \in (0, \infty)$.*

(i) By $||.||_{S_p}$ denote the number in $[0, \infty]$ given by:

$$||T||_{S_p} := (\sum_{k=1}^{\infty} |\sigma_k(T)|^p)^{\frac{1}{p}}$$

for each $T \in \mathcal{LC}(\ell^2(\mathbf{Z}_+))$.

(ii) By S_p denote the set of such $T \in \mathcal{LC}(\ell^2(\mathbf{Z}_+))$ that $||T||_{S_p} < \infty$. The set S_p is the Schatten p-class.

The set S_p is always a vector space. Note that $||.||_{S_p}$ is not actually a norm if $p \in (0, 1)$ because the triangle inequality fails. However, for $p \in [1, \infty)$ the space S_p is Banach. One more detail is needed for the proof of Theorem 19, namely the result [9, Lemma 3.2], which is a combination of two Jackson's theorems.

Lemma 18. *Let $r \in \mathbf{Z}_+$, $\alpha \geq 0$ such that $r + \alpha > 0$. $f \in C^{r,\alpha}(\mathbf{T})$. For all $k \in \mathbf{Z}_+$, set $E_k(f) := \inf_{deg\, p_k \leq k} ||p_k - f||_\infty$, where the infimum is taken over all trigonometric polynomials p_k, $\deg p_k \leq k$. Then*

$$E_k(f) \leq \frac{\pi^{r+\alpha}}{2^r} ||f^{(r)}||_{Lip_\alpha(\mathbf{T})} (k+1)^{-(r+\alpha)}.$$

We are ready to present a result about the relation between the smoothness of the symbol f, and the decay of the singular values of $K^{(n)}_{f,g}$.

Theorem 19. *Let $f \in C(\mathbf{T})$ and $n \in N$, $r \in \mathbf{Z}_+$ and $\alpha \geq 0$ such that $r + \alpha > 0$.*

(i) If $f \in C^{r,\alpha}(\mathbf{T})$, then the approximation numbers of $K_{f,g}^{(n)}$ satisfy

$$\sigma_{2k+1}(K_{f,g}^{(n)}) \leq \frac{\pi^{r+\alpha}}{2^{r-1}} \, ||f^{(r)}||_{Lip_\alpha(\mathbf{T})} \, ||g||_\infty \, (k+1)^{-(r+\alpha)}$$

for k such that $0 \leq k \leq \lfloor (n-1)/2 \rfloor$.

(ii) Let $p \in (0, \infty)$. The Schatten information about $K_{f,g}^{(n)}$ is given by

$$||K_{f,g}^{(n)}||_{S_p} \leq \frac{\pi^{r+\alpha}}{2^{r-2}} \, ||f^{(r)}||_{Lip_\alpha(\mathbf{T})} \, ||g||_\infty \left(\sum_{k=1}^{\lfloor (n+1)/2) \rfloor} k^{-p(r+\alpha)} \right)^{1/p} .$$

In particular, if $p > 1/r + \alpha$, then $\{K_{f,g}^{(n)}\}_{n \in \mathbf{N}}$ is an uniformly bounded family in the norm of S_p.

Proof. Claim (i) is proved by the following calculation. For any $k \geq 0$, we get from Definition 17

$$\sigma_{2k+1}(K_{f,g}^{(n)}) = \inf_{\operatorname{rank} F \leq 2k} ||K_{f,g}^{(n)} - F|| \leq \inf_{\deg p_k \leq k} ||K_{f,g}^{(n)} - K_{p_k,g}^{(n)}||,$$

where the last estimate holds by claim (iv) of Lemma 15. Here p_k is a trigonometric polynomial, $\deg p_k \leq k$. By using formulae (10), we estimate

$$\begin{aligned}
&||K_{f,g}^{(n)} - K_{p_k,g}^{(n)}|| \leq ||K_{f,g}^{(n)+} - K_{p_k,g}^{(n)+}|| + ||K_{f,g}^{(n)-} - K_{p_k,g}^{(n)-}|| \\
&= ||\pi_{[0,n-1]} \mathcal{H}_+[g] \mathcal{H}_-[f] \pi_{[0,n-1]} - \pi_{[0,n-1]} \mathcal{H}_+[g] \mathcal{H}_-[p_k] \pi_{[0,n-1]}|| \\
&+ ||\pi_{[0,n-1]} U^n \mathcal{H}_-[g] \mathcal{H}_+[f] U^{*n} \pi_{[0,n-1]} - \pi_{[0,n-1]} U^n \mathcal{H}_-[g] \mathcal{H}_+[p_k] U^{*n} \pi_{[0,n-1]}|| \\
&= ||\pi_{[0,n-1]} \mathcal{H}_+[g] \mathcal{H}_-[f - p_k] \pi_{[0,n-1]}|| \\
&+ ||\pi_{[0,n-1]} U^n \mathcal{H}_-[g] \mathcal{H}_+[f - p_k] U^{*n} \pi_{[0,n-1]}|| \\
&\leq ||\mathcal{H}_+[g]|| \cdot ||\mathcal{H}_-[f - p_k]|| + ||\mathcal{H}_-[g]|| \cdot ||\mathcal{H}_+[f - p_k]|| \\
&\leq 2||g||_\infty \cdot ||f - p_k||_\infty.
\end{aligned}$$

where we have used Lemma 4. Now

$$\sigma_{2k+1}(K_{f,g}^{(n)}) \leq 2||g||_\infty \cdot \inf_{\deg p_k \leq k} ||f - p_k||_\infty,$$

and an application of Lemma 18 proves now (i).

In order to prove claim (ii), we first note that the singular values satisfy $\sigma_k(K_{f,g}^{(n)} = 0$ for $k \geq 0$. Furthermore, for $p > 0$

$$\begin{aligned}
&\sum_{j=1}^{n} \sigma_j(K_{f,g}^{(n)})^p = \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} \sigma_{2k+1}(K_{f,g}^{(n)})^p + \sum_{k=1}^{\lfloor n/2 \rfloor} \sigma_{2k}(K_{f,g}^{(n)})^p \\
&\leq 2 \cdot \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} \sigma_{2k+1}(K_{f,g}^{(n)})^p,
\end{aligned}$$

where the inequality holds because $\sigma_{2k+2}(K_{f,g}^{(n)}) \leq \sigma_{2k+1}(K_{f,g}^{(n)})$ for all $k \geq 0$. Summing things up, together with the first claim of this theorem, completes the proof. $\square$

The upper singular value estimate for the Hankel operators used in the proof of Theorem 19 is not optimal. It can be shown that there is an infimal symbol $h \in C(\mathbf{T})$ satisfying $\inf_{\operatorname{rank} F \leq n} ||\mathcal{H}_+[f] - F|| = ||\mathcal{H}_+[f] - \mathcal{H}_+[h]||$ where $\operatorname{rank} \mathcal{H}_+[h] \leq n$. This is a consequence of the Kronecker theorem for the finite dimensional Hankel operators and the AAK-theorem stating that the above infimum is actually attained by a Hankel operator, rather than just some unstructured operator. For details, see [12].

5 On the convergence of iterations

In this section, we study the Toeplitz preconditioners for Toeplitz matrices, such that the symbol of the preconditioner is a trigonometric polynomial. Consider the following theorem about the Toeplitz preconditioning:

Theorem 20. *Let $f \in C(\mathbf{T})$ satisfy the invertibility condition of claim (ii) of Lemma 6. Let $n > \max(N(f), N(\frac{1}{f}))$, where $N(f)$ and $N(\frac{1}{f})$ are the constants given by claim (iii) of Lemma 6. Then there is a preconditioning sequence of trigonometric polynomials $\{g_k\}_{k=0}^{\infty}$ ($\deg(g_k) \leq k$) satisfying the following conditions.*

(i) The perturbation matrices satisfy

$$||B^{(n)}_{f,g_k}|| \leq ||1 - fg_k||_\infty \leq ||f||_\infty||(\frac{1}{f} - g_k)||_\infty,$$

and $\lim_{k\to\infty} ||B^{(n)}_{f,g_k}|| = 0$. There is a constant $M < \infty$ such that g_k satisfies the invertibility condition of claim (ii) of Lemma 6, for all $k \geq M$. There is a constant $N < \infty$, such that for all $k \geq M$ and $n \geq N$, $T_n[g_k]$ is a nonsingular Toeplitz preconditioner for nonsingular $T_n[f]$, as written in equation (7).

(ii) Assume, in addition, that $f \in C^{r,\alpha}(\mathbf{T})$ for $r + \alpha > 0$. Then the preconditioning sequence $\{g_k\}$ can be chosen so that the following speed estimate holds:

$$||B^{(n)}_{f,g_k}|| \leq \frac{\pi^{r+\alpha}}{2^r} ||f||_\infty||\left(\frac{1}{f}\right)^{(r)}||_{Lip_\alpha(\mathbf{T})} (k+1)^{-(r+\alpha)}. \tag{23}$$

(iii) If, in addition, $p > 1/(r+\alpha)$, then the family $\{K^{(n)}_{f,g_k}\}_{k\geq 0, n\in\mathbf{N}}$ is a bounded set in S_p.

Proof. By the definition of $B^{(n)}_{f,g_k}$ and Lemma 4 we have

$$||B^{(n)}_{f,g_k}|| = ||T_n[f(\frac{1}{f} - g_k)]|| \leq ||f(\frac{1}{f} - g_k)||_\infty \leq ||f||_\infty||(\frac{1}{f} - g_k)||_\infty.$$

By the Stone-Weierstrass approximation theorem, we can choose a sequence $\{g_k\}$ of trigonometric polynomials ($\deg(g_k) \leq k$) so that $||g_k - \frac{1}{f}||_\infty \to 0$. It now follows that $\lim_{k\to\infty} ||B^{(n)}_{f,g_k}|| = 0$ for this special sequence.

To continue the proof, note that if f satisfies the invertibility condition of Lemma 6, so does $\frac{1}{f}$, by a simple geometric argument. Let $\{g_k\}$ the sequence of polynomials as above. For the sequence $\{g_k\}$, there exists a $M < \infty$ such that for all $k \geq M$, we have

$$||g_k - \frac{1}{f}||_\infty < \frac{1}{2}\,||\mathcal{T}[\frac{1}{f}]^{-1}||^{-1}. \tag{24}$$

Note that the right hand side is finite and nonzero, because $\frac{1}{f}$ satisfies the invertibility condition of Lemma 6.

We now show that for all $k \geq M$, the Toeplitz operator $\mathcal{T}[g_k]$ is invertible. It is well known and easy to see that for any two operator S, T in a Hilbert space, T boundedly invertible, we have $S - T$ invertible if $||S|| < ||T^{-1}||^{-1}$. Now, define $S := \mathcal{T}[g_k] + \mathcal{T}[\frac{1}{f}]$ and $T := \mathcal{T}[\frac{1}{f}]$. Then $||\mathcal{T}[g_k] - \mathcal{T}[\frac{1}{f}]|| < ||\mathcal{T}[\frac{1}{f}]^{-1}||^{-1}$ implies that $\mathcal{T}[g_k]$ is boundedly invertible.

It follows that for any $k \geq M$, $T_n[g_k]$ is a nonsingular, if $n \geq N_k$ for some $N_k < \infty$, by Lemma 6. By the same lemma, the Toeplitz matrix $T_n[f]$ itself is nonsingular, if $n \geq N(f)$. It follows that $T_n[g_k]$ is a nonsingular preconditioner for nonsingular matrix $T_n[f]$, if $k \geq M$ and $n \geq \max(N_k, N(f),)$. However, we want to have the constants N_k independent of k.

By claim (iii) of Lemma 6, $\limsup_{n\to\infty} ||T_n[\frac{1}{f}]^{-1}|| := C < \infty$ because $T_n[\frac{1}{f}] \to \mathcal{T}[\frac{1}{f}]$ strongly. Let $N \geq N(\frac{1}{f})$ be so large that $||T_n[\frac{1}{f}]^{-1}|| < 2C$ for all $n \geq N$. Choose $M_2 \in \mathbf{N}$ so large that

$$||g_k - \frac{1}{f}||_\infty \leq \frac{1}{2C}$$

for all $k \geq M_2$. Then we can estimate for $k \geq M_2$ and $n \geq N$

$$\begin{aligned} ||T_n[g_k] - T_n[\frac{1}{f}]|| &= ||T_n[g_k - \frac{1}{f}]|| \leq ||g_k - \frac{1}{f}||_\infty \\ &\leq \frac{||T_n[\frac{1}{f}]^{-1}||}{2C} \cdot ||T_n[\frac{1}{f}]^{-1}||^{-1} < ||T_n[\frac{1}{f}]^{-1}||^{-1}. \end{aligned}$$

But this implies that $T_n[g_k]$ is invertible, as already considered above. Claim (i) is now proved.

Now claim (ii). If $f \in C^{r,\alpha}(\mathbf{T})$, so does $\frac{1}{f}$, by a routine argument. By Lemma 18, we can choose the sequence $\{g_k\}$ to satisfy equation (23). To prove the remaining claim, note that $\{g_k\}$ is an uniformly bounded family in $C(\mathbf{T})$ because it converges uniformly to a limit in $C(\mathbf{T})$. It follows that $\{K_{f,g_k}^{(n)}\}_{k\geq 0, n\in\mathbf{N}}$ is a uniformly bounded family in S_p, by part (ii) of Theorem 19. ☐

The companion paper [9] about Toeplitz operators contains a numerical example, which is equally valid for the present case of Toeplitz matrices, too. The rest of this chapter is dedicated to the interpretation of the results of Theorem 20 from the numerical analysis point of view.

Not all preconditioning sequences of trigonometric polynomials satisfy the speed estimate of equation (23). The preconditioning sequence $\{g_k\}$ can be constructed in great many ways, still preserving the speed estimate (23) in an asymptotic sense; see for example [7, pp.21, Ex. 2] for Lipschitz continuous symbols. So as to the numerical construction schemes for $\{g_k\}$, we refer to the ideas presented in [2].

Theorems 7 and 19 show that after Toeplitz preconditioning, the system consists of a truncation effect $K_{f,g}^{(n)}$ perturbed by a small $B_{f,g}^{(n)}$. Smoothness of the symbol f has thus a two-fold effect on the properties for iteration of $T_n[f]$: For smooth f it is easier to control the "preconditioning error" $B_{f,g_k}^{(n)}$ by increasing the degree of the trigonometric polynomial g_k. And then, by claim (ii) of Theorem 19, smoothness of the symbol $f \in C(\mathbf{T})$ specifies the approximation properties of $K_{f,g}^{(n)}$ by lower rank matrices. This gives us speed estimates for the Krylov subspace methods, as will be discussed below.

In [9], we treated the analogous preconditioning problem of Toeplitz operators by appealing to a solution of a more general problem: How does a Krylov subspace method perform if it is applied upon an operator consisting of a compact $K \in S_p$ perturbed by an unstructured small B. This problem is studied in [8] and [11]. There K corresponds to the truncation effect, and B to the perturbation operator of Toeplitz operators. In this paper we proceed along the same lines. Now both the operators $K_{f,g}^{(n)}$, $B_{f,g}^{(n)}$ are finite dimensional (thus compact) for each $n \in \mathbf{N}$. This does not stop us from treating $K_{f,g}^{(n)}$ as a compact operator and $B_{f,g}^{(n)}$ as an (unstructured) perturbation.

Let us briefly reiterate some of the terminology given in [8], [10] and [11]. In the study of Krylov subspace methods applied upon $K+B$, it is customary to look at how the sequence $||p_k(K+B)||^{\frac{1}{k}}$ behaves as $k \to \infty$, where $\{p_k\}$ is a sequence of normalized ($p_k(1) = 1$) degree k polynomials associated to the Krylov subspace method in question. Degree k of the polynomial corresponds to the number of iteration steps computed. For a brief reminder why this is done, look at the discussion at the end of this section.

A function theoretic argument proves the following theorem where the normalization of the polynomials $\{\tilde{p}_k\}$ is slightly different, but without effect on the asymptotics as $k \to \infty$:

Theorem 21. *Let $p \geq 1$. Let H be Hilbert space and $S_p(H)$ be the Schatten p-class. Take $K \in S_p(H)$, $K \neq 0$, and let $B \in \mathcal{L}(H)$ be a small perturbation such that $1 \notin \sigma(K+B)$. Then there exists an essentially monic sequence of polynomials $\{\tilde{p}_k\}_{k=1}^{\infty}$, $\deg p_k \leq k$, such that for all parameter values $\beta \in (0,1]$:*

$$
\begin{aligned}
(25) \qquad & ||\tilde{p}_k(K+B)||^{1/k} \\
& \leq p^{1/k} \left(||B|| + ||K||_{S_p} k^{-\beta/p}\right) \left(\frac{||B||\, k^{\beta/p}}{||K||_{S_p}} + 1\right)^{1/k} e^{1/(k^{1-\beta})}.
\end{aligned}
$$

Furthermore, $\lim_{k\to\infty} |\tilde{p}_k(1)| > 0$.

Proof. See [8, Theorem 6.7]. □

The expression "essentially monic" means that the leading term of all $\tilde{p}_k$ is a same nonzero complex number. The fact that $\lim_{k\to\infty} |\tilde{p}_k(1)| > 0$ makes it possible to normalize $\tilde{p}_k$ for large k, and define

$$
p_k(\lambda) := \frac{\tilde{p}_k(\lambda)}{\tilde{p}_k(1)}.
$$

Now the sequence $\{p_k\}$ has the correct normalization $p_k(1) = 1$, and the speed estimate like (25) holds also for p_k with an additional multiplicative constant sequence $|\tilde{p}_k(1)|^{-1/k}$, for all k large enough. Note that because $\lim_{k\to\infty} |\tilde{p}_k(1)|^{-1/k} = 1$, the effect of the incorrect normalization of $\tilde{p}_k$ does not change the nature of speed estimate (25) in an asymptotic sense.

Theorem 21 tells us that in the first stages the iteration the convergence factor $||p_k(K+B)||^{\frac{1}{k}}$ of order $||B|| + ||K||_{s_p} k^{-\frac{\beta}{p}}$ decreases (the "superlinear" stage) and is asymptotically only of order $||B||$ (the "linear" stage). Moreover, the rate of decrease of the convergence factor is dictated by the Schatten class of K. The concept "superlinear" is usually used to describe something that happens in the asymptotics of the speed estimates. Here we are a bit unorthodox (as we have been in [9]) and regard "superlinear" stage of an iteration as those iteration steps when "speed is being gained". By the "linear" stage we of course refer to the analogous phenomenon.

The following corollary of Theorem 21 is our convergence estimate for the iteration of $B_{f,g}^{(n)} + K_{f,g}^{(n)}$, the matrix of the preconditioned system (7). We invite the reader to regard g as an element of the preconditioning sequence $\{g_k\}$ of Theorem 20, with increasing degree k of the preconditioner.

To say that $f \in C(\mathbf{T})$ is strictly analytic (strictly coanalytic) means that the negatively (positively) indexed Fourier coefficients of f vanish. A strictly analytic symbol $f \in C(\mathbf{T})$ has an analytic continuation $f(z)$ inside the unit disk of the complex plane, and $f(0) = 0$.

Corollary 22. *Let $r \in \mathbf{Z}_+$, $r \geq 0$ be such that $r + \alpha > 0$. Assume that the nonconstant $f, g \in C^{r,\alpha}(\mathbf{T})$ satisfy the invertibility condition of Lemma 6. Furthermore, assume that not both f, g are simultaneously strictly analytic or coanalytic. Let N_1 be so large that both $T_n[f]$ and $T_n[g]$ are invertible for all $n > N_1$. Then the following holds:*

(i) For $p > 1/(r+\alpha)$, there are constants C_1, C_2, and N_2, such that

$$0 < C_1 \leq ||K_{f,g}^{(n)}||_{S_p} \leq C_2 < \infty.$$

for all $n > N_2$.

(ii) Assume, in addition, that $r + \alpha \geq 1$. For each fixed $n > N := \max(N_1, N_2)$, there exists an essentially monic sequence polynomials $\{\tilde{p}_k^{(n)}\}_{k=1}^{\infty}$ such that for all parameter values $\beta \in (0,1]$:

$$\begin{aligned} &||\tilde{p}_k^{(n)}(K_{f,g}^{(n)} + B_{f,g}^{(n)})||^{1/k} \\ &\leq p^{1/k} \left(||1 - fg||_\infty + C_2 \, k^{-\beta/p} \right) \cdot \\ &\cdot \left(\frac{||1 - fg||_\infty \, k^{\beta/p}}{C_1} + 1 \right)^{1/k} e^{3/(k^{1-\beta})}. \end{aligned} \tag{26}$$

Furthermore, $\lim_{k\to\infty} \tilde{p}_k^{(n)}(1)$ exists is bounded away from the origin.

Proof. We prove claim (i) about the constants C_1 and C_2. By the definition of the Schatten norm, always

$$||K_{f,g}^{(n)}|| = \sigma_1(K_{f,g}^{(n)}) \leq ||K_{f,g}^{(n)}||_{S_p}.$$

By Lemma 13, $\lim_{n\to\infty} ||K_{f,g}^{(n)}|| = \max(||K_{f,g}||, ||K_{\tilde{f},\tilde{g}}||)$. To show that the lower bound $C_1 > 0$ exists for n large enough, we have to show that it is not possible to have $K_{f,g} = K_{\tilde{f},\tilde{g}} = 0$ under the assumptions of this Corollary. By Lemma 14, $K_{f,g} = K_{\tilde{f},\tilde{g}} = 0$ if and only if one of the following conditions holds: (1) f is constant, (2) g is constant, (3) both f and g are strictly analytic, (4) both f and g are strictly coanalytic. However, all these possibilities are ruled out in the assumptions. So $C_1 > 0$ exists. The upper bound C_2 exists, by claim (iii) of Theorem 20. The latter claim (ii) is a straightforward application of Theorem 21. □

An analogous theorem to Theorem 21 can be proved for the Schatten classes $p \in (0, 1]$, see [8, Theorem 6.9]. An analogous corollary to Corollary 22 for $p \in (0, 1]$ is then a triviality. This establishes a convergence speed estimate of type (26) for all Toeplitz systems with symbol $f \in C^{r,\alpha}(\mathbf{T})$ for $r + \alpha > 0$, without the extra smoothness assumption $r + \alpha \geq 1$.

What is the meaning of the requirement in Corollary 22 that not both f, g are allowed to be, say, strictly analytic? For technical reasons only, the convergence estimate (25) is written so that the Schatten norm of the truncation effect is in the denominator. Suppose we could precondition optimally so that $g = \frac{1}{f}$ and the perturbation part $B_{f,g}^{(n)} = 0$ for all n. Then if both f were g are strictly analytic, then f is, by definition, would be an outer analytic function. But this is impossible, because $f(0) = 0$ by strict analyticity. We conclude that if $g \approx \frac{1}{f}$, not both f and g can be strictly analytic.

As we have seen, the upper bounds for both $||B_{f,g}^{(n)}||$ and $||K_{f,g}^{(n)}||_{S_p}$, given in Theorems 19 and 20, are not dependent of n, the dimension of the problem. It follows that the right hand side of the convergence estimate (26) is independent of n for n large. In order to obtain a similar speed estimate for the corresponding correctly normalized polynomial sequence satisfying $p_k^{(n)}(1) = 1$, with the right hand side independent of n, we would have to show at least that

$$\inf_{n>N} \lim_{k\to\infty} \tilde{p}_k^{(n)}(1) > 0.$$

Even to look at this infimum superficially, it requires long and complicated calculations about the limit process of the spectrum of $B_{f,g}^{(n)} + K_{f,g}^{(n)}$, as $n \to \infty$. This is no longer a subject of this paper, because our aim was to go as far as we can, without explicitly looking at the (difficult) spectral properties of the preconditioned operator.

How does this all relate to a particular Krylov subspace algorithm, namely GMRES? The GMRES method for the inversion of nonsymmetric problems can be regarded as a minimization algorithm that (at least implicitly) generates polynomial sequences to approximate the value of resolvent in certain points; this is the minimization of residuals. If the GMRES generates the polynomial sequence s_k with $\deg(s_k) = k$ and $s_k(1) = 1$, corresponding to the normalized sequence p_k given after Theorem 21. Then the residual d_k

after k steps is of size $||s_k(K+B)d_0||$, and we have

$$(27) \qquad ||s_k(K+B)d_0|| \leq ||p_k(K+B)d_0|| \leq ||p_k(K+B)|| \, ||d_0||,$$

see [8, Proposition 2.2] or [10, Chapter 1]. The former inequality is true because s_k is optimal polynomial of degree k for the initial residual d_0, and $\tilde{p}_k$ is possibly worse than optimal for the same initial residual d_0. This is to say that the upper estimates we have for $\tilde{p}_k$ are as well upper estimates for the GMRES residuals. The same kind of result is true so as to the error sequences with quite obvious modifications for the reasoning — we again refer to [8] or [10, Chapter 1].

References

[1] R. Chan. Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions. *IMA J. Numer. Anal.*, 11:333–345, 1991.

[2] R. Chan and Kwok-Po Ng. Toeplitz preconditioners for hermitian Toeplitz systems. *Linear Algebra and its Applications*, 190:181–205, 1993.

[3] R. Chan and Man-Chung Yeung. Circulant preconditioners constructed from kernels. *SIAM J. Numer. Anal.*, 29(4):1093–1103, 1992.

[4] R. Chan and Man-Chung Yeung. Circulant preconditioners for complex Toeplitz matrices. *SIAM J. Numer. Anal.*, 30(4):1193–1207, 1993.

[5] N. Dunford and J. Schwarz. *Linear Operators; Part II: Spectral Theory.* Interscience Publishers, Inc. (J. Wiley & Sons), New York, London, 1963.

[6] I.C. Gohberg and I.A. Feldman. *Convolution Equations and Projection Methods for Their Solution*, volume 41 of *AMS Translations of Mathematical Monographs.* American Mathematical Society, 1974.

[7] Y. Katznelson. *Introduction to Harmonic Analysis.* Dover Publications, Inc., New York, 2 edition, 1976.

[8] J. Malinen. On the properties for iteration of a compact operator with unstructured perturbation. *Helsinki University of Technology, Institute of mathematics, Research Report*, A360, 1996.

[9] J. Malinen. Properties of iteration of Toeplitz operators with Toeplitz preconditioners. *BIT Numerical Mathematics*, 38(2), June 1998.

[10] O. Nevanlinna. *Convergence of Iterations for Linear Equations.* Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, Boston, Berlin, 1993.

[11] O. Nevanlinna. Convergence of Krylov-methods for sums of two operators. *BIT Numerical Mathematics*, 36(4):775–785, 1996.

[12] J. R. Partington. *An introduction to Hankel operators*, volume 13 of *Student texts.* London Mathematical Society, London, 1988.

[13] W. Rudin. *Functional Analysis.* McGraw-Hill Book Company, New York, TMH edition, 1990.

[14] G. Strang. A proposal for Toeplitz matrix calculations. *Studies in Applied Mathematics*, 74:171–176, 1986.

(continued from the back cover)

HELSINKI UNIVERSITY OF TECHNOLOGY INSTITUTE OF MATHEMATICS
RESEARCH REPORTS

The list of reports is continued inside. Electronical versions of the reports are available at *http://www.math.hut.fi/reports/* .

A412 Marko Huhtanen
Ideal GMRES can be bounded from below by three factors, Jan 1999

A411 Juhani Pitkäranta
The first locking-free plane-elastic finite element: historia mathematica, Jan 1999

A410 Kari Eloranta
Bounded Triangular and Kagomé Ice, Jan 1999

A408 Ville Turunen
Commutator Characterization of Periodic Pseudodifferential Operators, Dec 1998

A407 Jarmo Malinen
Discrete time Riccati equations and invariant subspaces of linear operators, Feb 1999

ISBN 951-22-4354-7
ISSN 0784-3143
Edita, Espoo, 1999