Homotopic Residual Correction Algorithms for General and Structured Matrices

by

Hülya CEBECIOĞLU

A dissertation submitted to the Graduate faculty in Mathematics in

partial fulfillment for the degree of Doctor of Philosophy.

The City University of New York.

2001

This manuscript has been read and accepted for the Graduate Faculty in Mathematics in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Victor Pan

Date — Chair of Examining Committee

Józef Dodziuk e.z.

Date — Executive Officer

MICHAEL ANSHEL

ALEXEI MIASNIKOV

VICTOR Y. PAN Victor Pan

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Homotopic Residual Correction Algorithms for General and Structured Matrices

by

Hülya CEBECIOĞLU

Advisor: Professor Victor Y. PAN

Newton's Iteration is a fundamental algorithm of numerical and algebraic computing. We focus on its application to the approximation of the inverse or Moore–Penrose generalized inverse of a matrix. This application was studied by Schultz in 1933 and since then by many authors. The algorithm is strongly stable numerically (in fact it is self-correcting) and converges quadratically, provided that an initial approximation to the (generalized) inverse is available.

An initial approximation can be crude, but must have a resiual norm less than 1, and some known recipes (see in particular [PS91]) give approximations with the initial norms slightly below 1. Most effective, however, are applications of Newton's iteration where the input matrices are structured (celebrated examples are the Toeplitz, Henkel, Vandermonde, and Cauchy

matrices), and in this case approximations obtained based on the above recipes is generally too crude. The computation by N.I. is reduced essentially to repeatedly performing matrix multiplication, and this operation can be performed very effectively, in nearly linear time, when the matrices are structured (versus cubic time for general matrices). The structure, however, tends to deteriorate gradually as N.I. progresses. Special techniques for preserving the structure have been proposed in [P92], [PBRZ99], [PRWa].These techniques, however, require sufficiently close initial approximations, substantially closer than those supplied by the known recipes, including the recipes of [PS91]. The problem can be solved, however, based on a new homotopic approach which is our subject in the thesis.

We apply this approach to general input matrices. The resulting homotopic version of N.I. turns out to be competitive with non-homotopic version for Hermitian (or real symmetric) input matrices. For the homotopic version, the computational cost is roughly the same as for the non-homotopic one where the input matrix is positive definite and is substantially less where it is indefinite. We also study application of this approach to structured matrices, in which case each iteration step is dramatically accelerated and is performed in nearly linear time, because some special techniques preserve the structure during the entire iterative process. Numerical experiments reported in [PKRCa] confirm the effectiveness of the resulting algorithms in the case of Toeplitz input matrices.

Acknowledgements

I would like to express my deepest appreciation to my mentor, Prof. Pan, for his invaluable guidance, generous assistance and friendship. His seminar on Numerical and Algebraic computations introduced me to the subject and his interest in my mathematical development provided the additional stimulus to study it further. It has been an honor and a privilege to work with him, and I look forward with pleasure to our continued collaboration.

I also want to thank Prof. Anshel and Prof. Miasnikov for serving on my thesis committee. I wish to thank all the faculty members and staff of the Mathematics Program at the CUNY Graduate Center, especially to Prof. Józéf Dodziuk, Prof. Martin Moskowitz, Prof. Alvany Rocha and the Assistant Program Officer Mr. Robert Landsman.

Completion of my education would have been impossible without the support of my family members. I would like to thank my parents, Elif and Bekir, for their unstinting support for me all these years and taking care of my son, Berkay; Thank you, Berkay, your smiles and hugs have helped me to keep everything in its proper perspective. My very special thanks go to my husband and colleague, Oktay Cebecioglu, whose constant help, encouragement and support were indispensable. My deepest gratitute to my sister, Fatma Gulser, who was a great source of encouragement and moral support for years. They believed that I was capable of achieving anything I wanted to do in life and made me believe it too. I would like to extend thanks to my brother-in-law, Gokhan Gulser, for his technical support on computers, and my brother, Kerim Kodal, for his moral support.

Special thanks to friends and colleagues upon whose personal support and intellectual insights I have depended and have been enriched: Bilal Khan, Ann Marie Morhaime, Rhys E. Rosholt, Erez Schocrat, and Xinmao Wang.

I also want to thank the Higher Educational Council of Turkey, which through Kocaeli University, provided a full scholarship for my doctoral study.

I wish to dedicate this thesis to my parents.

Contents

1 Introduction

Summary: We present and analyze homotopic Newton's iteration algorithms for the computation of the inverses of general and structured matrices. For unstructured indefinite Hermitian input matrices, we substantially accelerate the known best non-homotopic algorithms, with no sacrifice in their numerical stability and self-correction property. For structured matrices, the homotopic algorithms, like the known residual correction methods, perform each iteration in nearly linear time. Unlike the non-homotopic algorithms, however, superlinear convergence to the inverse is guaranteed even where no initial approximation is available. Numerical tests with Toeplitz input matrices show greater power of both homotopic and non-homotopic approaches than the theoretical study predicts.

Key Words: Newton's iteration, residual correction, homotopic algorithms, structured matrices, rank of a matrix, generalized inverse

1.1 Newton's iteration and Residual correction (RC) processes

Residual correction processes (in particular *Newton's iteration*) compute the inverse or the Moore-Penrose generalized inverse of a general $n \times n$ matrix M [S33], [B-I66], [B-IC66], [IK66], [SS74], [PS91]. The processes involve p matrix multiplications in each step to achieve convergence of order p, for any $p \geq 2$. With appropriate scaling of the process, however, one may reach the order of $p > 2$ by using only two matrix multiplications per step [PS91]. Hereafter, we will write RC for "residual correction" and MM for "matrix multiplication". The RC processes can be directed to the numerical generalized inverse and are known for their strong numerical stability and self-correcting property [PS91].

Let us recall the two main problems with these processes. (For simplicity here and actually throughout until Section 11, we assume non-singularity of the input matrices M. In Section 11, we show extension to the singular case and to the numerical computation of the Moore–Penrose generalized inverses.)

a) The RC processes require additional techniques for the computation of an initial approximation to the inverse. The known techniques of [B-I66], [B-IC66], [SS74], and [PS91] produce a crude initial approximation. Then it takes the order of $\log_2 \kappa(M)$ RC steps ($\kappa(M) = \mathrm{cond}(M)$ denoting the condition number of the matrix M) to refine the approximation to the level from which the iteration very rapidly converges.

b) For general matrices, MM is an expensive operation, comparable to matrix inversion in its computational cost. Such an operation, however, is dramatically simplified in the highly important case of structured matrices, represented by their displacements in a compressed form. Namely, the displacement of an $n \times n$ matrix occupies memory space $O(n)$, and multiplication of $n \times n$ compressed structured matrices uses $O(n \log n)$ or $O(n \log^2 n)$ flops. Consequently, the RC processes can be also performed by using small memory space and little computer time as long as compression performed throughout the computation does not destroy rapid convergence of the process. Here some advanced compression techniques are applied, first proposed in [P92] and then elaborated in [PZHD97], [PBRZ99], [PR01], [PRWa].

For the sake of completeness of our study, we briefly review this development in Sections 2–4, and in the thesis we refer to the RC processes covering Newton's iteration as their special case for $p = 2$.

1.2 Homotopic RC processes

The solution techniques for problems a) and b) do not always match one another, however. That is, compression perturbs the computed approximation and may easily destroy convergence at the initial stages of the RC processes where the convergence is fragile. This implies additional requirements of achieving either much stronger initial approximations than can be yielded by the known techniques of [B-I66], [B-IC66], [SS74], and [PS91] or compression causing much smaller perturbations of the computed approximations to the inverse than the current study ensures. The original approach of [P92] allows a very natural heuristic modification towards the latter goal. Recent experiments show that such a heuristic is surprisingly effective in the important case of Toeplitz input matrices, but no convincing theoretical results support such a development. In this thesis, our main subjects are new techniques for computing the initial approximation. Their efficiency is confirmed by both experiments and proofs of the estimates for the computational work of the resulting algorithms. The methods are homotopic, based on the inversion of an auxiliary readily invertible matrix M_0 such as $M_0 = I$ and on the subsequent homotopic transition to the matrix M along the trajectories

$$M_h = (1 - t_h)M + t_h M_0, \ h = 0, 1, \ldots, \tag{1.1}$$

or

$$M_h = M + t_h M_0, \ h = 0, 1, \ldots, \tag{1.2}$$

where

$$t_0 > t_1 > \ldots > t_H = 0, \tag{1.3}$$

$t_0 = 1$ in (1.1) and is a sufficiently large value in (1.2). We arrange the homotopy to keep the trajectories $M(t)$ away from singularities for $t_0 \geq t \geq 0$; we prove that for $t \geq 0$ the condition numbers of the matrices $M(t)$ reach their maximums where $t = 0$. Then, by choosing the step sizes $t_h - t_{h+1}$ sufficiently small, we may always ensure that the matrix $M_h^{-1}M_{h+1}$ is close enough to the identity matrix; then the approximation to the inverse M_h^{-1} computed at the h-th homotopic step would serve as a good initial approximation at the next, $(h+1)$-st homotopic step.

1.3 Preceding and related works

Newton's iteration for the inverse of a matrix was covered in some detail in the papers [S33], [B-I66], [B-IC66], [SS74], [PS91]. Higher order RC processes were also well studied (see [IK66, pp. 88-89], [PS91]). The paper [PS91] accelerated Newton's iteration by using scaling, extended the iteration to the computation of the numerical generalized inverses of a matrix, and proved strong numerical stability of the original and modified iterations. The paper [P92] worked out Newton's iteration for Toeplitz-like matrices (with the compression of the displacement by the truncation of its singular values) as well as the homotopic process for the initialization. The paper also estimated the perturbation of the computed approximations to the inverse caused by the compression (the problem was further studied in [P93]) and proved that nearly linear overall number of flops is sufficient for Toeplitz-like inversion provided that $\log \kappa(M) = O(\log n)$. Parallel implementation of this approach was studied in the papers [P92] and [P93a] in the Toeplitz-like case. [PZHD97] studied extension to the Cauchy-like input (with a distinct policy of compression). The paper [PBRZ99] published in [KS99] presented some elaboration of Newton's iteration under both approaches to the compression in the Toeplitz-like case; [BM,a] did the same with the compression approach; technically, the study of Newton's iteration in both papers remained within the frameworks of [P92] and [PZHD97]. No further works on the homotopy approach followed since [P92], except for the proceedings paper [P01]. A unified method for the extension of Newton's iteration to various classes of structured matrices was proposed and analyzed in [PR01] and [PRWa]. On an alternative general approach to the unification, based on transformation of the associated displacement operators, see Remark 4.1.

1.4 Organization of the thesis

In Sections 2–4, we recall some known results on the RC processes for general and structured matrices. In Sections 5–8 and 10, we elaborate the choices of the initial approximations and the step sizes, which use fewer RC steps for positive definite and indefinite Hermitian input matrices; we prove substantial acceleration in the latter case versus the non-homotopic approach. We briefly cover the extension to structured input matrices in Section 9. In this case the homotopic approach supplies the only known proof of convergence of the RC processes in nearly linear time where no initial approximation is available from the outside sources and the input matrix is well-conditioned. In the cases where numerical generalized inverse is structured, the same approach can be extended to its effective numerical computation (Section 11).

2 Residual Correction Processes (RC Processes)

Hereafter, M^T, $\mathbf{v}^T$, M^*, and $\mathbf{v}^*$ denote the transposes and Hermitian (conjugate) transposes of a matrix M and a vector $\mathbf{v}$, respectively. We write $\sigma_j = \sigma_j(M)$, $\kappa(M) = \sigma_1/\sigma_r$. σ_j denote the singular values of a matrix M where $r = \mathrm{rank}(M)$, $j = 1, \ldots, r$; $0 < \sigma_- \le \sigma_r \le \ldots \le \sigma_1 \le \sigma_+$; $\kappa(M)$ is the condition number of M. $\mathbf{e}_{i-1}$ denotes the i-th coordinate vector, $i = 1, \ldots, n$. $\lceil x \rceil$ is the smallest among the integers not exceeded by a real x.

A sufficiently close initial approximation X_0 to the inverse of a nonsingular matrix M can be rapidly improved by means of a scaled RC process [S33], [IK66], [PS91]:

$$X_{i+1} = c_{i+1} X_i \sum_{k=0}^{p-1} R_i^k, \quad i = 0, 1, \ldots, \tag{2.1}$$

where we write

$$R_i = R(M, X_i) = I - MX_i. \tag{2.2}$$

Already for the unscaled process, that is, under the simplest choice of

$$c_i = 1 \text{ for all } i, \tag{2.3}$$

(2.1) and (2.2) imply that

$$R_i = (R_0)^{p^i}, \quad \|R_i\| \le \|R_0\|^{p^i}, \quad i = 1, 2, \ldots. \tag{2.4}$$

This shows that the unscaled RC process (2.1), (2.3) converges with the order p to the matrix M^{-1} provided that

$$\|R_0\|_2 \le \theta < 1, \quad R_0 = R(M, X_0).$$

Suppose that the latter bound holds for a fixed θ. Then the computational work per step (2.1), (2.3) or, equivalently, the number of steps required to ensure the desired upper bound on the norm $\|R_i\|$ is minimized for $p = 3$ [IK66, pages 86–88].

Now suppose that no initial approximation X_0 to the matrix M^{-1} is available. Then one may choose

$$X_0 = c_0 M^*, \quad c_0 = \frac{2}{\sigma_+ + \sigma_-} \tag{2.5}$$

to yield that

$$||R_0||_2 \le 1 - \frac{2}{1 + \kappa_+^2}, \quad \kappa_+ = \kappa_+(M) = \sigma_+/\sigma_-. \tag{2.6}$$

Now, it is sufficient to apply at first

$$i = 2 \log_p \kappa_+ + O(1)$$

unscaled *critical RC steps* (2.1), (2.3), to decrease the residual norm $||R_i||_2$ below $1/2$, and then

$$j = \lceil \log_p \log_2(1/\epsilon) \rceil$$

additional *refinement RC steps* (2.1), (2.3), to decrease the norm below any fixed positive $\epsilon \leq 1/2$ [SS74]. In Section 7 we will use the threshold value $1/e = 0.367819\ldots$ instead of $1/2$; this may change i at most by 1.

The scaling policy of choosing c_{i+1} in (2.1) was optimized in [PS91] in the case of Newton's iteration,

$$X_{i+1} = c_{i+1}X_i(I + R_i), \tag{2.7}$$

that is, of RC process (2.1) for $p = 2$. Namely, by choosing $p = 2$,

$$c_0^- = \frac{2\sigma_-}{\sigma_+ + \sigma_-}, \quad c_{i+1} = \frac{2}{1 + (2 - c_i^-)c_i^-}, \quad c_{i+1}^- = (2 - c_i^-)c_i^- c_{i+1} \tag{2.8}$$

for $i = 0, 1, \ldots$, one obtains that

$$||R_l||_2 \leq \max_{\sigma_- \leq x \leq \sigma_+} |T_{2^l}(\gamma x + \delta)|/|T_{2^l}(\delta)| \leq \frac{1}{|T_{2^l}(\delta)|}$$

where $\gamma = 2/(\sigma_+ - \sigma_-)$, $\delta = -1 - \gamma\sigma$, and $T_j(x) = \cos(j \arccos x)$ is the j-th degree Chebyshev polynomial of the first kind on $[-1, 1]$. It follows [FF63, Chapter 9, Section 9] that

$$||R_l||_2 \leq \frac{2}{(\delta + \sqrt{\delta^2 - 1})^L + (\delta - \sqrt{\delta^2 - 1})^L}, \quad L = 2^l;$$

this bound is substantially smaller than δ^L. In particular, the number of critical steps decreases roughly by twice versus policy (2.3), namely to the level

$$i = \log_2 \kappa_+(M) + O(1/\kappa_+^2(M)). \tag{2.9}$$

In other words, the optimal scaling of (2.8) is equivalent to increasing the order of convergence of the critical steps from $q = 2$ to $q = 4$ for the same RC process (2.7) (that is, (2.1) for $p = 2$).

The asymptotic bound $i_- = \log_2 \kappa(M) + O(1)$ on the number of critical RC steps is achieved also under the simpler initial choice of

$$X_0 = M^*/(||M||_1||M||_\infty).$$

Furthermore, for a Hermitian (or real symmetric) and positive definite matrix M, *one may further decrease the number of critical RC steps* (2.7) *roughly by twice* [PS91] with an appropriate choice of the initial approximation X_0. In particular we have the desired decrease where

$$X_0 = I/||M||_F, \quad ||R_0||_2 \leq \frac{1}{\sqrt{n}\kappa(M)}, \tag{2.10}$$

$||M||_F = \text{trace}\,(M^+M)$ denotes the Frobenius norm of the matrix M, and M is a Hermitian and positive definite matrix.

The paper [PS91] has also shown some advantages of using a scaled cubic RC process (2.1) for $p = 3$ and a modification where Newton's RC processes for $p = 2$ converged to numerical generalized (Moore–Penrose) inverse M_ϵ^+, that is, the generalized inverse of the matrix M_ϵ formed via the truncation of the smallest singular values of M (up to a fixed tolerance ϵ). This is achieved by first applying iteration (2.7)–(2.8) but with

$$c_0 = \sigma_+ c, \; c_0^- = c\epsilon^2, \; c = \min(2/(\sigma_+ + \epsilon^2), \rho/\epsilon^2), \tag{2.11}$$

$\rho = (1 + \sqrt{3}/2 = 1.366\ldots$ (Under the scaling of (2.11) the value ρ partitions the range for the spectrum of the matrix X_0M; the partition is induced by the respective partition by ϵ of the singular values of the matrix M. Note that the bound σ_- is not needed in this variation of the iteration.) The iteration is performed until we arrive at $c_i^- \geq \rho$ for some integer i. Then the matrix X_i is scaled, that is, replaced by the matrix $(\rho/c_i^-)X_i$ and the iteration is continued based on the expressions

$$X_{i+1} = (-2X_iM + 3I)X_iMX_i, \quad i = 0, 1, \ldots. \tag{2.12}$$

The singular values $\sigma_j(M)$ are partitioned by ϵ into two groups: those exceeding ϵ correspond to the eigenvalues $\lambda^{(i)}$ of X_iM that lie in the interval $1/2 < \lambda^{(i)} \leq \rho$; iteration (2.12) sends them towards 1. The other eigenvalues of X_iM lie in the interval $[0, 1/2)$; they correspond to the singular values

$\sigma_j(M) < \epsilon$. Iteration (2.12) sends them towards 0. This is exactly the desired convergence to the matrix M_ϵ^+. Convergence is ultimately quadratic but is slow near $1/2$ and ρ. Iteration can be immediately extended to the computation of the matrices $M_\epsilon = MM_\epsilon^+M$ and $\widetilde{M}_\epsilon = M - M_\epsilon$ and the numerical rank trace$(M_\epsilon M_\epsilon^+)$.

It was proved in [PS91] that both original and modified Newton's (RC) processes are numerically stable.

3 Toeplitz Residual Correction Processes

If $M = T = (t_{i-j})_{j=0}^{n-1}$ is a non-singular Toeplitz matrix, then RC processes (2.1) can be accelerated dramatically, based on the known formulae for the inverse matrix $X = T^{-1}$ via a pair of its products by vectors [GS72], [HR84], [AG89], [BP94], [VHKa], [BM,a].

Let us recall two such formulas and describe respective accelerations of Newton's RC processes by following [PBRZ99]. The same techniques immediately produce an RC process for Toeplitz inputs wherever a basic RC process (2.1) and a Toeplitz inversion formula are specified. Write

$$J = \begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix} = (\mathbf{e}_{n-1}, \dots, \mathbf{e}_0)^T$$

for the $n \times n$ reflection matrix, and

$$Z_f = \begin{pmatrix} 0 & \dots & 0 & f \\ 1 & \ddots & & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & 0 \end{pmatrix} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{n-1}, f\mathbf{e}_0)$$

for a unit f-circulant matrix, where $\mathbf{e}_i$ is the $(1+i)$-th coordinate vector of dimension n. Then $Z_f(\mathbf{v}) = \sum_{i=0}^{n-1} v_i Z_f^i$ denotes the f-circulant matrix

of size $n \times n$ with the first column vector $\mathbf{v} = (v)_{i=0}^{n-1}$. $Z(\mathbf{v}) = Z_0(\mathbf{v})$ is a lower triangular Toeplitz matrix, $Z_1(\mathbf{v})$ is circulant. In the next sections, we denote a diagonal matrix by $D(\mathbf{v}) = \mathrm{diag}(v_i)_{i=0}^{n-1}$ for $\mathbf{v} = (v_i)_{i=0}^{n-1}$.

Write

$$T\mathbf{y} = \mathbf{e}_0, \quad T\mathbf{x} = \mathbf{t} \tag{3.1}$$

where

$$\mathbf{t} = (w, at_1 - bt_{1-n}, at_2 - bt_{2-n}, \ldots, at_{n-1} - bt_{-1})^T \tag{3.2}$$

for three fixed scalars w, a, and b.

In particular by choosing $a = 0$, $b = -1$, and any w, we obtain that

$$\mathbf{t} = (w, t_{1-n}, \ldots, t_{-1})^T, \tag{3.3}$$

and then we have the following expressions for $X = T^{-1}$ via the vectors $\mathbf{y} = X\mathbf{e}_0$ and $\mathbf{x} = X\mathbf{t}$:

$$X = Z(\mathbf{x})Z^T(ZJ\mathbf{y}) - Z(\mathbf{y})Z^T(ZJ\mathbf{x} - \mathbf{e}_0). \tag{3.4}$$

To yield an alternative expression via f-circulant matrices instead of triangular Toeplitz matrices, fix any pair of values $b \neq 0$ and w, write $a = 1$, $f \neq 1/b$, and obtain the vector

$$\mathbf{t} = (w, t_1 - bt_{1-n}, t_2 - bt_{2-n}, \ldots, t_{n-1} - bt_{-1})^T \tag{3.5}$$

and the equation

$$X = \frac{1}{1-bf}\left(Z_f(\mathbf{y})Z_{1/b}(\mathbf{x}) - Z_f(\mathbf{x} - (1-bf)\mathbf{e}_0)Z_{1/b}(\mathbf{y})\right), \tag{3.6}$$

which expresses the matrix X via the vectors $\mathbf{y} = X\mathbf{e}_0$ and $\mathbf{x} = X\mathbf{t}$.

Now let us modify RC processes (2.1) by expressing similarly the approximation matrices X_i via the pair of vectors $X_i\mathbf{e}_0$ and $X_i\mathbf{t}$, for all i. Fix the

vector $\mathbf{t}$ of (3.2) and post-multiply (2.1) by the $n \times 2$ matrix $(\mathbf{e}_0, \mathbf{t})$, having the columns $\mathbf{e}_0$ and $\mathbf{t}$:

$$X_{i+1}(\mathbf{e}_0, \mathbf{t}) = c_{i+1} X_i \sum_{i=0}^{p-1} R_i^k(\mathbf{e}_0, \mathbf{t}). \tag{3.7}$$

In particular for $p = 2$ we obtain the following extension of process (2.7):

$$X_{i+1}(\mathbf{e}_0, \mathbf{t}) = c_{i+1} X_i (I + R_i)(\mathbf{e}_0, \mathbf{t}). \tag{3.8}$$

Now, instead of defining the matrix X_{i+1} via X_i based on (2.1), we define it via the vectors $\mathbf{y}_{i+1} = X_{i+1}\mathbf{e}_0$ and $\mathbf{x}_{i+1} = X_{i+1}\mathbf{t}$, by substituting X_{i+1} for X, $\mathbf{y}_{i+1}$ for $\mathbf{y}$, and $\mathbf{x}_{i+1}$ for $\mathbf{x}$ in (3.4) or (3.6), respectively. This completely defines a Toeplitz RC process. For each i, its i-th step is reduced to a few multiplications of Toeplitz matrices by vectors, which are performed fast based on FFT, that is, each step uses $O(n \log n)$ flops versus the order of n^2 flops required for multiplication of a general $n \times n$ matrix by a vector.

Remark 3.1. *For a Hermitian or real symmetric non-singular Toeplitz matrix T, one may represent the inverse matrix $X = T^{-1}$ via its first column only [GS72], [AG89]; this would save memory space but would involve divisions by the $(0,0)$-th entry of X, which may vanish or nearly vanish for indefinite matrices T, thus causing numerical stability problems.*

Let us recall the estimates of [PBRZ99] for the convergence rate of the *Newton-Toeplitz Iteration* defined by (3.8), in both cases for $c_{i+1} = 1$. Let us write $\rho(i) = \|I - X_i T\|_1$, $e(i) = \max(\|\mathbf{x}_i - \mathbf{x}\|_1/\|\mathbf{x}\|_1, \|\mathbf{y}_i - \mathbf{y}\|_1/\|\mathbf{y}\|_1)$. Furthermore, let us write either $\mu = \|\mathbf{y}_1\|_1(2(n-1)(2+\rho(0)e(0))\|\mathbf{x}\|_1+1)$ provided that the Toeplitz RC process relies on (3.1)-(3.4), or $\mu = \|\mathbf{y}\|_1(\|\mathbf{x}\|_1(1+\rho(0)e(0)) + 1)$ provided that the Toeplitz RC process relies on (3.1), (3.2), (3.5), and (3.6). Assume that

$$\rho(0) \le \theta, \quad e(0)\|T\|_1\mu \le \theta \tag{3.9}$$

for a fixed θ, $0 < \theta < 1$. Then it is proved in [PBRZ99] that $\rho(i) < \theta^{2^i}$, $e(i) < \theta^{2^i-1}e(0)$, $i = 1, 2, \ldots$, which shows quadratic convergence under assumptions (3.9). To satisfy (3.9), however, we must have a sufficiently close initial approximation to the inverse matrix T^{-1}.

4 Residual Correction Processes for Structured Matrices

Extensions of unscaled RC processes (2.7), (2.3) to Toeplitz-like matrices can be found in [P92], [PBRZ99, Section 7.4]. Let us next follow [PZHD97], [PBRZ99], [PR01], [PRWa], [Pa] to outline such extensions in a unified way - simultaneously to various classes of structured matrices, in particular, to Toeplitz, Hankel, Vandermonde, and Cauchy matrices (see Table 4.1) and the matrices with the structures of these four types. This covers the most popular classes of structured matrices.

4.1 Structured matrices and the displacement rank approach

With two *operator matrices* A and B we associate a linear displacement operator L, of Sylvester type $L = \nabla_{A,B}$,

$$\nabla_{A,B}(M) = AM - MB \tag{4.1}$$

or Stein type $L = \Delta_{A,B}$,

$$\Delta_{A,B}(M) = M - AMB \tag{4.2}$$

where M is an $n \times n$ matrix.

Typically, $A, B \in \{D(\mathbf{s}), D(\mathbf{t}), Z_e, Z_f^T\}$ for appropriate vectors $\mathbf{s}$ and $\mathbf{t}$ and scalars e and f, which covers the cited four most popular classes of structured matrices. We have the following properties:

Table 4.1: **Four classes of structured matrices**

Toeplitz matrices $(t_{i-j})_{i,j=0}^{n-1}$ $\begin{pmatrix} t_0 & t_{-1} & \cdots & t_{1-n} \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \cdots & t_1 & t_0 \end{pmatrix}$	Hankel matrices $(h_{i+j})_{i,j=0}^{n-1}$ $\begin{pmatrix} h_0 & h_1 & \cdots & h_{n-1} \\ h_1 & h_2 & \iddots & h_n \\ \vdots & \iddots & \iddots & \vdots \\ h_{n-1} & h_n & \cdots & h_{2n-2} \end{pmatrix}$
Vandermonde matrices $(t_i^j)_{i,j=0}^{n-1}$ $\begin{pmatrix} 1 & t_0 & \cdots & t_0^{n-1} \\ 1 & t_1 & \cdots & t_1^{n-1} \\ \vdots & & & \vdots \\ 1 & t_{n-1} & \cdots & t_{n-1}^{n-1} \end{pmatrix}$	Cauchy matrices $\left(\frac{1}{s_i-t_j}\right)_{i,j=0}^{n-1}$ $\begin{pmatrix} \frac{1}{s_0-t_0} & \cdots & \frac{1}{s_0-t_{n-1}} \\ \frac{1}{s_1-t_0} & \cdots & \frac{1}{s_1-t_{n-1}} \\ \vdots & & \vdots \\ \frac{1}{s_{n-1}-t_0} & \cdots & \frac{1}{s_{n-1}-t_{n-1}} \end{pmatrix}$

a) the matrix $L(M)$ has a small rank, r for a structured matrix M and an associated displacement operator L (r is called the *displacement rank* of the matrix M),

b) the operator L^{-1} is linear, furthermore there are simple expressions for the matrix $M = L^{-1}(L(M))$ through its displacement $L(M)$, and

c) an $n \times n$ structured matrix can be multiplied by a vector fast, in $O(nr\log^d n)$ flops for $d \leq 2$ (cf. Table 4.2).

The first and the most celebrated demonstration of these properties was given in the seminal paper [KKM79] for Toeplitz-like matrices M, associated with the operators $L_+ = \Delta_{Z,Z^T}$ and $L_- = \Delta_{Z^T,Z}$. In particular, it was proved that the matrix equations

$$L(M) = GH^T,\ G = (\mathbf{g}_1, \ldots, \mathbf{g}_r),\ H = (\mathbf{h}_1, \ldots, \mathbf{h}_r) \tag{4.3}$$

Table 4.2: **Parameter and flop count for matrix representation and multiplication by a vector**

Matrices M	Number of parameters per an $m \times n$ matrix M	Number of flops for computation of $M\mathbf{v}$
general	mn	$2mn - n$
Toeplitz	$m+n-1$	$O((m+n)\log(m+n))$
Hankel	$m+n-1$	$O((m+n)\log(m+n))$
Vandermonde	m	$O((m+n)\log^2(m+n))$
Cauchy	$m+n$	$O((m+n)\log^2(m+n))$

imply that

$$M = \sum_{j=1}^{r} Z(\mathbf{g}_j) Z^T(\mathbf{h}_j) \tag{4.4}$$

for $L = L_+$ and

$$M = \sum_{j=1}^{r} Z^T(J\mathbf{g}_j) Z(J\mathbf{h}_j) \tag{4.5}$$

for $L = L_-$. It is easy to observe that

$$|\operatorname{rank}(L_+(M)) - \operatorname{rank}(L_-(M))| \leq 2,$$

for any matrix M, and that

$$\operatorname{rank}(L_+(M)) \leq 2, \ \operatorname{rank}(L_-(M)) \leq 2$$

where M is a Toeplitz matrix. This motivated the definition of Toeplitz-like matrices M as the ones with displacements $L_+(M)$ and $L_-(M)$ having small

ranks. Expressions (4.4), (4.5) enable multiplications of such a matrix by a vector in $O(rn \log n)$ flops.

Similar simple expressions have been obtained for displacement operators associated with matrices of Hankel, Vandermonde, and Cauchy types [HR84], [BP94], [GO94], [PWa], [Pa], enabling *compressed representations* of an $n \times n$ structured matrix via $2nr$ entries of the matrices G and H. Note that orthogonal representations (4.3) for a given matrix $L(M)$ can be immediately obtained from its SVD [P92], [P93] (e.g., in the real case, $L(M) = U\Sigma^2 V^T$, $U^T U = V^T V = I_r$, $G = U\Sigma$, $H = V\Sigma$) and if $L(M)$ is a Hermitian matrix then from its eigendecomposition too.

Compressed representations can be also derived based on some singular displacement operators. For instance, in [PBRZ99] the following known representation of an $n \times n$ Toeplitz-like matrix has been exploited,

$$M = Z_{f,lc}(M\mathbf{e}_{n-1}) + \frac{e}{e-f}\sum_{j=1}^{r} Z_f(Z_f \mathbf{g}_j) Z_{1/e}^T(\mathbf{h}_j) \tag{4.6}$$

provided that (4.3) holds for $L = \nabla_{Z_f^{-1}, Z_f^{-1}}$, where e and f are two scalars, $e \neq f$, $ef \neq 0$, and $Z_{f,lc}(\mathbf{v})$ denotes the f-circulant matrix of size $n \times n$ with the last column $\mathbf{v}$. (Note that $Z_f^{-1} = Z_{1/f}^T$.) Table 4.3 shows some displacement operators associated with structured matrices.

According to *the displacement rank approach*, one should operate with structured matrices M represented in a compressed form such as (4.3)–(4.6) and when required, recover the output (such as the solution of a linear system of equations) based on their linear expressions via the displacement $L(M)$. The entire approach can be represented by the following flowchart:

COMPRESS, OPERATE, DECOMPRESS.

At the OPERATE stage, the following simple results can be used [KKM79], [CKL-A87], [P90], [P00a], [Pa].

Table 4.3: **Some pairs of operators $\nabla_{A,B}$ and structured matrices**

operator matrices		class of structured	rank of
A	B	matrices M	$\nabla_{A,B}(M)$
Z_1	Z_0	Toeplitz and its inverse	≤ 2
Z_1	Z_0^T	Hankel and its inverse	≤ 2
$Z_0 + Z_0^T$	$Z_0 + Z_0^T$	Toeplitz+Hankel	≤ 4
$D(\mathbf{t})$	Z_0	Vandermonde	≤ 1
Z_0	$D(\mathbf{t})$	inverse of Vandermonde	≤ 1
Z_0^T	$D(\mathbf{t})$	transposed Vandermonde	≤ 1
$D(\mathbf{s})$	$D(\mathbf{t})$	Cauchy	≤ 1
$D(\mathbf{t})$	$D(\mathbf{s})$	inverse of Cauchy	≤ 1

Theorem 4.1. *For any linear operator L (in particular, for $L = \nabla_{A,B}$ and $L = \Delta_{A,B}$, for any pair of matrices A and B) and any pair of scalars a and b, we have $L(aM + bN) = aL(M) + bL(N)$.*

Theorem 4.2. *For any 5-tuple $\{A, B, C, M, N\}$ of $n \times n$ matrices, we have*

$$\nabla_{A,C}(M, N) = \nabla_{A,B}(M)N + M\nabla_{B,C}(N),$$
$$\Delta_{A,C}(M, N) = \Delta_{A,B}(M)N + AM\nabla_{B,C}(N).$$

Furthermore,

$$\Delta_{A,C}(MN) = \Delta_{A,B}(M)N + AMB\Delta_{B^{-1},C}(N),$$

if B is a non-singular matrix, whereas

$$\Delta_{A,C}(MN) = \Delta_{A,B}(M)N - AM\Delta_{B,C^{-1}}(N)C,$$

if C is a non-singular matrix.

Theorem 4.3. *Let M be a non-singular matrix. Then*

$$\nabla_{B,A}(M^{-1}) = -M^{-1}\nabla_{A,B}(M)M^{-1}.$$

Furthermore,

$$\Delta_{B,A}(M^{-1}) = BM^{-1}\Delta_{A,B}(M)B^{-1}M^{-1},$$

if B is a non-singular matrix, whereas

$$\Delta_{B,A}(M^{-1}) = M^{-1}A^{-1}\Delta_{A,B}(M)M^{-1}A,$$

if A is a non-singular matrix.

4.2 Structured RC processes

Based on the latter results and properties a)-c) of structured matrices listed in the previous subsection, one may perform structured matrix multiplications fast. So $O(qnr^2 \log^d n)$ flops are sufficient per an RC step (2.1), which outputs a short displacement generator of the matrix X_{i+1}, provided that the matrices M and X_i are given in compressed form (4.3) and q is the order of convergence of a process (2.1). Special care is required, however, to contain the growth of $\mathrm{rank}(L(X_{i+1}))$. With no care the rank rapidly increases; it may be tripled already in each Newton step (2.7). Thus processes (2.1) should be modified as follows where the input matrix M is structured:

$$X_{i+1} = X(Y_{i+1}), \quad Y_{i+1} = c_{i+1} X_i \sum_{k=0}^{p-1} R_i^k. \tag{4.7}$$

Here, the matrix $X_{i+1} = X(Y_{i+1})$ approximates the matrices Y_{i+1} and M^{-1}, and $r_{i+1} = \mathrm{rank}(L(X_{i+1}))$ either equals or only slightly exceeds r. To complete the definition of the structured RC process (4.7) for fixed parameters p, c_{i+1}, let us specify the transition from the matrix Y_{i+1} to the matrix X_{i+1}, where both structured matrices Y_{i+1} and X_{i+1} are represented by their displacements [P92], [P92a], [BP93], [PZHD97], [PBRZ99], [PR01], [PRWa].

Approach I. Truncation of the smallest singular values of the displacement. Compute the SVD of $L(Y_{i+1})$ and truncate the smallest singular values to obtain a displacement matrix $L(X_{i+1})$ having r_{i+1} (non-zero) singular values, for $r_{i+1} = r$ or $r_{i+1} \approx r$. (In the case where $L(X_i)$ is a Hermitian matrix, one may rely on its eigendecomposition instead of its SVD.)

Approach II. Substitution of a computed approximation for the inverse in the inversion formulae. Compute the displacement $L(X_{i+1})$

based on Theorem 4.3, where M^{-1} is replaced by X_i. That is, write

$$\nabla_{B,A}(X_{i+1}) = -X_i \nabla_{A,B}(M) X_i, \tag{4.8}$$

$$\Delta_{B,A}(X_{i+1}) = X_i A^{-1} \Delta_{A,B}(M) X_i A, \tag{4.9}$$

where the operator matrix A is non-singular, or

$$\Delta_{B,A}(X_{i+1}) = B X_i \Delta_{A,B}(M) B^{-1} X_i, \tag{4.10}$$

where the operator matrix B is non-singular. The previous section actually covered Approach II specified to Toeplitz input matrices M and based on two known explicit formulae for the Toeplitz inverse M^{-1}.

Approach I relies on the observation that

$$||L(X_{i+1}) - L(Y_{i+1})|| \leq ||L(X_{i+1}) - L(M^{-1})||$$

under the 2-norm and the Frobenius norm. This observation is due to Theorem 4.3 and to the well known results on the lower rank approximation based on the truncation of the singular values [GL96]. Thus we bound the norms $||L(X_{i+1}) - L(M^{-1})||$ and $||X_{i+1} - M^{-1}|| \leq ||L^{-1}||\ ||L(X_{i+1}) - L(M^{-1})||$ in terms of the norm $||L(Y_{i+1}) - L(M^{-1})||$.

In Approach II, we bound the same norms by combining (4.8)-(4.10) with Theorem 4.3.

Specific estimates for the approximation errors, the convergence rate, and the initial residual or error norms which ensure rapid convergence for both approaches can be found in [P92], [PZHD97], [PBRZ99], [PRWa], and [Pa].

Algorithm 7.4.1 of [PBRZ99] computes the displacements $L_-(X_{i+1}) = L_-(X(Y_{i+1}))$ by applying Approach I to Toeplitz-like matrices M and by using the displacements $L_+(M)$ and $L_-(X_i)$ and expressions (4.4), (4.5).

It is proved that in this case

$$||X_{i+1} - M^{-1}||_2 \leq (1 + 2(r_i - r)n)||X_i - M^{-1}||_2 \tag{4.11}$$

where $r_i = \text{rank}(L_-(Y_i))$.

Algorithm 7.4.2 of [PBRZ99] implements Approach II and relies on (2.3), (2.8), and (4.6). In this case the matrix X_{i+1} is defined by its displacement

$$\nabla_{Z_f^{-1},Z_f^{-1}}(X_{i+1}) = G_{i+1}H_{i+1}^T,$$

$$G_{i+1} = Y_{i+1}(2I - MY_i)\widetilde{G}_{i+1}, \quad H_{i+1}^T = \widetilde{H}_{i+1}^T Y_{i+1}(2I - MY_{i+1})$$

and its last column

$$X_{i+1}\mathbf{e}_{n-1} = Y_{i+1}(2I - MY_{i+1})\mathbf{e}_{n-1},$$

provided that $\nabla_{Z_f^{-1},Z_f^{-1}}(Y_{i+1}) = \widetilde{G}_{i+1}\widetilde{H}_{i+1}^T$.

In [PRWa] both Approaches I and II have been elaborated and analyzed in a unified way for various classes of structured matrices (based on the displacement rank approach). The results of [SS74] and [PS91] on the convergence of Newton's and other RC processes cited in Section 2 do not apply to processes (4.7) because of the compression of the displacements $L(Y_i)$. The following theorems from [PRWa] (extending their preliminary versions of [P92], [PZHD97], [PBRZ99], and [PR01]) state the estimates for the error norms of the computed approximations. The statements of the theorems involve the norm $||L^{-1}||_l$ of the inverse of the displacement operator L,

$$||L^{-1}||_l = \sup_M(||M||_l/||L(M)||_l), \quad l = 1, 2, \infty.$$

Upper estimates for this norm, $||L^{-1}||_l$ for various customary operators L associated with the most popular classes of structured matrices have been deduced in [PRWa] and [PWa].

Theorem 4.4. *[PRWa]. Let unscaled Newton's process (2.7), (2.3) be applied to a non-singular structured matrix M. Let all its steps be performed with compression according to (4.7) and Approach I such that all the singular*

values of the displacements $L(Y_i)$, except for the r largest ones were truncated where $r = \mathrm{rank}(L(M^{-1}))$. Then we have

$$||X_i - M^{-1}||_2 \le ||I - X_iM||_2 \; ||M^{-1}||_2 \le \theta^{2^i}||M^{-1}||_2/\eta,$$

$i = 1, 2, \ldots,$ *provided that*

$$\theta = ||I - X_0M||_2\eta,$$

$$\eta = (1 + (||A||_2 + ||B||_2)||L^{-1}||_2)\sigma_1(M)/\sigma_n(M) \quad \text{for } L = \nabla_{A,B},$$

$$\eta = (1 + (1 + ||A||_2||B||_2)||L^{-1}||_2)\sigma_1(M)/\sigma_n(M) \quad \text{for } L = \Delta_{A,B}.$$

Theorem 4.5. *[PRWa]. Let structured unscaled Newton's process (2.7), (2.3) be applied to invert a non-singular structured matrix M. Let (4.7) and Approach II be used for the compression of the displacements $L(Y_i)$, $i = 1, 2, \ldots$. Write*

$$r_{i,l} = ||I - X_iM||_l,$$

$$e_{i,l} = ||Y_i - M^{-1}||_l,$$

$$\hat{e}_{i,l} = ||X_i - M^{-1}||_l,$$

$$l = 1, 2, \infty; \; i = 0, 1, 2, \ldots$$

Let $r_0 \le 1, \quad e_{i,l} \le ||M^{-1}||_l, \quad l = 1, 2, \infty; \quad i = 0, 1, 2, \ldots,$

$$C_l = 3||L^{-1}||_l \; ||L(M)||_l \; ||X_0||_l/(1 - r_{0,l}) \text{ for } L = \nabla_{A,B},$$

$$C_l = 3||L^{-1}||_l \; ||L(M)||_l \; ||M||_l \; ||M^+||_l \; ||X_0||_l/(1 - r_{0,l}) \text{ for } L = \Delta_{A,B}.$$

Then

$$\hat{e}_{i,l} \le C_l e_{i,l}, \quad e_{i+1,l} \le (C_l e_{i,l})^2||M||_l,$$

and therefore,

$$\gamma_l e_{i+1,l} \le (\gamma_l e_{1,l})^{2^i}, \quad i = 1, 2, \ldots; \; l = 1, 2, \infty,$$

where $\gamma_l = C_l^2||M||_l$.

Remark 4.1. *Newton-Structured Iteration with compression was first studied for Toeplitz-like matrices (see [P92]). In the papers [PR01], [PRWa] the algorithms were extended to various other classes of structured matrices in a unified way, adapted in this section. In an alternative general approach of [P90], it was proposed to extend successful algorithms available for one class of structured matrices to various other classes by means of the transformation of the associated displacement operators, and sample transformation techniques were shown for the transformation in all directions among the operators associated with the matrices having structures of Toeplitz, Hankel, Vandermonde, and Cauchy types. In particular, these techniques apply to matrix inversion and thus enable immediate extension of our RC and HRC processes. For input matrices with the structures of Cauchy and Vandermonde types, the transformation approach may lead to some addtional advantages. Namely, multiplication by a vector requires fewer flops for a Toeplitz or Hankel matrix than for a Cauchy or Vandermonde matrix (see Table 4.2) and is more stable numerically. Since the algorithms of this section are ultimately reduced to multiplication by a vector of structured matrices of a given type, we may decrease the overall computational cost by reducing the problem to the Toeplitz-like or Hankel-like case. So far, the most acclaimed application of the transformation approach has been the reduction of the practical solution of Toeplitz and Toeplitz-like linear systems of equations to the Cauchy-like case via the transformation of the associated displacement operators [H95], [GKO95].*

5 A Homotopic Residual Correction (HRC) Algorithm for a Positive Definite Matrix

A reliable solution of the initialization problem for the RC processes is given by *homotopic RC processes*, to be referred to as *HRC processes* and studied next. RC processes (2.1) (both scaled and unscaled and with any selected levels of compression in the case of structured input) may serve as a black box subroutine in each homotopic step.

Let M be a Hermitian positive definite matrix, and let spectrum$(M) = \{\lambda_1, \dots, \lambda_n\}$, where

$$\lambda_1^+ \geq \lambda_1 = ||M||_2 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq \lambda_n^- > 0 \tag{5.1}$$

and where λ_1^+ is a known precomputed value. Fix some values θ_h, $0 < \theta_h < 1$, $h = 0, 1, \dots$, and write (cf. (1.2), (1.3))

$$M_0 = M + t_0 I, \ t_0 = \lambda_1^+/\theta_0, \ X_0 = t_0^{-1} I, \tag{5.2}$$

$$M_{h+1} = t_{h+1} I + M = M_h - \Delta_h I, \ \ \Delta_h = t_h - t_{h+1} > 0, \ \ h = 0, 1, \dots. \tag{5.3}$$

Then, for the residual R_0 of (2.2), we have

$$R_0 = R(M_0, t_0^{-1} I) = I - t_0^{-1} M_0 = -t_0^{-1} M, \ r_0 = ||R(M_0, t_0^{-1} I)||_2 \leq \theta_0. \tag{5.4}$$

Therefore, the matrix M_0 can be inverted rapidly by process (2.1) unless the bound θ_0 is close to 1.

Further, deduce from (5.3) that

$$R(M_{h+1}, M_h^{-1}) = \Delta_h M_h^{-1},$$
$$r_{h+1} = ||R(M_{h+1}, M_h^{-1})||_2 = \Delta_h ||M_h^{-1}||_2 = \Delta_h/(t_h + \lambda_n). \tag{5.5}$$

Compute an upper bound η_h on the norm

$$||M_h^{-1}||_2 = \frac{1}{t_h + \lambda_n}. \tag{5.6}$$

Write

$$\lambda_{n,h} = 1/\eta_h - t_h \le 1/||M_h^{-1}||_2 - t_h = \lambda_n \tag{5.7}$$

and observe that the value $\lambda_n^- = \lambda_{n,h}$ satisfies bound (5.1). Choose

$$\Delta_h = \theta_h/\eta_h, \quad h = 1, 2, \ldots, H-1, \tag{5.8}$$

which implies that

$$r_{h+1} \le \theta_h \text{ for all } h. \tag{5.9}$$

Recursively invert the matrices M_{h+1} by applying a selected RC process (2.1) as long as t_{h+1} remains positive. As soon as we arrive at $t_H \le 0$, we invert the input matrix M instead of M_H.

The algorithm is completely defined as soon as we fix the parameters θ_h and RC processes (2.1) (including their stopping criteria and, for structured matrices M, their policies of the compression of the displacements) applied at the h-th homotopic steps for $h = 0, 1, \ldots, H$.

In the next two sections, we estimate the overall numbers of the RC steps required for the inversion of a general unstructured Hermitian positive definite matrix M and optimize this number by choosing appropriate bounds θ_h for a fixed order of convergence q of the basic RC process. In Section 8, we extend the algorithm to the case of a general indefinite matrix M. In Section 9–11, we cover extensions to the cases where the matrix M is structured and compression of the displacements is applied, where the matrix M is singular, and/or where a numerical inverse of M is computed.

Remark 5.1. *We have* $\|M_h^{-1}\|_1/\sqrt{n} \le \|M_h^{-1}\|_2 \le \|M_h^{-1}\|_1$ *for an* $n \times n$ *Hermitian matrix* M_h^{-1}. *Sharper upper bound* η_h *on the matrix norm can be obtained by applying the power or Lanczos methods [GL96]. If an estimate* η_h *is sufficiently sharp for a fixed* $h = k$ *(say for* $h = 1$*), close upper bounds* η_{k+i} *can be computed based on the following simple expression:*

$$\eta_{k+i} = \frac{1}{t_{k+i} + \lambda_{n,k}}, \quad \lambda_{n,k} = 1/\eta_k - t_k, \quad i = 1, 2, \ldots$$

(see (5.3)-(5.8)).

Remark 5.2. *The homotopic process of (5.2), (5.3) has trajectory* $M(t) = M + tI$ *which for* $t > 0$ *is better conditioned than the input matrix* M. *That is, one may easily verify that*

$$\kappa(M(t)) < \kappa(M) \text{ for } t > 0. \tag{5.10}$$

The same inequality can be easily verified for the modification of the homotopic process of Section 8 proposed in the indefinite Hermitian case.

Remark 5.3. *The approach allows variations. For instance, instead of process (5.2), (5.3), we may apply homotopic process (1.1) or the dual process*

$$M_{h+1} = I + t_{h+1}M = M_h + (t_{h+1} - t_h)M, \quad h = 0, 1, \ldots,$$

followed at the end by a single step (5.3) or a few steps (5.3). The resulting computations can be analyzed similarly to process (5.3).

Remark 5.4. *Homotopic processes (5.3) and (8.5) of Section 8 exploit the* techniques of variable diagonal *(cf. [P00b] and references therein).*

6 The Number of Homotopic Steps

To simplify our subsequent analysis, we next assume that the values $\lambda_{n,h}$ are invariant in h, that is, $\lambda_{n,h} = \lambda_n^-$ for all $h \geq 1$ (cf. (5.6) and Remark 5.1). Then by virtue of (5.3), (5.6), (5.7), and (5.8), we have $t_{h+1} + \lambda_n^- = (1 - \theta_h)(t_h + \lambda_n^-)$, $h = 0, 1, \ldots, H-1$. Therefore,

$$t_{h+1} + \lambda_n^- = (t_0 + \lambda_n^-)\prod_{i=0}^{h}(1 - \theta_i), \quad h = 0, 1, \ldots, H-1,$$

$$t_H \leq 0 \text{ if } \lambda_n^- \geq (t_0 + \lambda_n^-)\prod_{h=0}^{H-1}(1 - \theta_h).$$

Let us next estimate the number H of homotopic steps. For simplicity, assume that the parameter θ_h is invariant in h, that is, let $\theta_h = \theta$ for all h. Substitute $t_0 = \lambda_1^+/\theta$ of (5.2) and rewrite the latter inequality as follows:

$$\frac{1}{(1-\theta)^H} \geq \lambda_1^+/(\theta\lambda_n^-) + 1,$$

$$H \geq -\log(1 + \lambda_1^+/(\theta\lambda_n^-))/\log(1-\theta).$$

Choose the minimum integer H satisfying this bound, that is,

$$H = \left\lceil \frac{\log(1 + \lambda_1^+/(\theta\lambda_n^-))}{\log(1/(1-\theta))} \right\rceil \tag{6.1}$$

homotopic steps are sufficient. Substitute

$$\theta = K/(1+K) \tag{6.2}$$

and rewrite (6.1) as follows:

$$H = \left\lceil \frac{\log(1 + (K+1)\lambda_1^+/(K\lambda_n^-))}{\log(1+K)} \right\rceil. \tag{6.3}$$

7 The Overall Number of the Residual Correction (RC) Steps

Let us next complement estimates (6.1)–(6.3) by counting the RC steps. At each homotopic step, their number depends on the bound θ on the initial residual norm (to be assumed invariant at all homotopic steps), the order q of convergence of the selected RC process, and the stopping criterion for this process. We assume some fixed order q for each process (2.1) given a general unstructured matrix M and scalars p and c_{i+1}, $i = 0, 1, \ldots$. In particular, $q = p$ for unscaled processes (2.1), (2.3).

7.1 Critical and refinement stages of an RC process

Estimating the number of RC steps at the i-th homotopic step, we will treat separately its initial *critical stage*, where the residual norm decreases below $1/e = 1/2.718281\ldots = 0.367819\ldots$, and the subsequent *refinement stage*, where the residual norm decreases below a fixed target bound ν_i for the output approximation X_j to M_i^{-1} (compare a similar partition of a non-homotopic process in Section 2). We write $\nu_H = \epsilon$ and $\nu_i = \nu$ for all $i < H$, and choose the scalar $\nu = \nu(\theta)$ sufficiently small to ensure that the computed approximations are close enough to the matrices M_i^{-1} to serve as initial approximations at the next homotopic steps.

7.2 The number of RC steps at the refinement stages

Processes (2.1) with the order of convergence q decrease the residual norm from $1/e$ to e^{-q^g} in g RC steps (cf. (2.4)). Therefore, at the H-th homotopic step, the refinement requires

$$\gamma = \lceil (\log \ln(1/\epsilon))/\log q \rceil \tag{7.1}$$

RC steps, whereas

$$\beta = \lceil (\log \ln(1/\nu))/\log q \rceil \tag{7.2}$$

refinement steps suffice for the transition from $1/e$ to ν for each $i < H$

Summarizing, we have a total of at most

$$P = \gamma + (H-1)\beta \tag{7.3}$$

RC steps at the refinement stages of all homotopic steps of the HRC algorithm. Bound (7.1) applies to the number of all refinement RC steps of the non-homotopic processes of Section 2 (for the same q and ϵ). Bound (7.2) covers the $(H-1)\beta$ refinement RC steps particular to the HRC processes. Practically, β is quite small. For instance, for $q = 4$, the bound e^{-16} is

achieved in two steps. The specific choice of the bound ν can be guided by the following simple estimate.

Proposition 7.1. *Let*

$$\|I - XM_{h-1}\| \leq \nu, \tag{7.4}$$

$$\|I - M_{h-1}^{-1}M_h\| \leq \theta_h \tag{7.5}$$

for any fixed matrix norm. Then

$$\|I - XM_h\| \leq (1+\nu)\theta_h + \nu.$$

Proof. $\|I-XM_h\| \leq \nu+\|XM_{h-1}+XM_h\| \leq \nu+\|XM_{h-1}\|\,\|I-M_{h-1}^{-1}M_h\| \leq \nu + (1+\nu)\theta_h$. □

7.3 The number of RC steps at the critical stages

Let α denote the number of RC steps used at the critical stage of a homotopic step. Then we have

$$1/\theta_h^{q^\alpha} = (1+1/K)^{q^\alpha} \approx e,\ q^\alpha \approx 1/\ln(1+1/K) \approx K, \alpha \approx (\log K)/\log q \tag{7.6}$$

provided that θ is close to 1, that is, that K is large.

By combining (6.3) and (7.6) for $\theta_h = \theta$ for all h, we estimate the overall number of RC steps at all critical stages of the entire HRC process:

$$\begin{aligned} N = \alpha H &\approx ((\log(\lambda_1^+/\lambda_n^-))\log K)/((\log(K+1))\log q) \\ &\leq N^+ = (\log(\lambda_1^+/\lambda_n^-))/\log q. \end{aligned} \tag{7.7}$$

7.4 The overall number of RC steps in homotopic and non-homotopic processes

Based on (6.3), (7.2)–(7.5), and (7.7), one may immediately estimate the overall number,

$$N + P = \alpha H + \gamma + (H-1)\beta$$

of the RC steps of the entire HRC algorithm. This is the same bound as in Section 2 for non-homotopic RC processes both with scaling (for q=4) and without it (for q=2).

8 Extensions to the Inversion of Indefinite Non-singular Input Matrices

We may extend our HRC algorithm of Section 5 to compute numerically the inverse M^{-1} of any non-singular matrix based on the equations

$$M^{-1} = M^*(MM^*)^{-1} = (M^*M)^{-1}M^* \tag{8.1}$$

because the matrices MM^* and M^*M are Hermitian (or real symmetric) and positive definite. Such a standard symmetrization, however, has the well-known price of squaring the condition number and, consequently, of a substantial slowdown of the HRC algorithm (cf. (7.7)). Let us next show a simple remedy in the case where M is a non-singular Hermitian (or a real symmetric) but indefinite matrix M. Recall that the inversion of any non-singular input matrix M reduces to the inversion of the Hermitian or real symmetric matrix

$$N = \begin{pmatrix} 0 & M \\ M^* & 0 \end{pmatrix}, \tag{8.2}$$

$$N^{-1} = \begin{pmatrix} 0 & (M^*)^{-1} \\ M^{-1} & 0 \end{pmatrix},$$

$\kappa(N) = \kappa(M)$.

Let λ^- and λ^+ be two fixed positive values such that

$$\lambda^- \le |\lambda| \le \lambda^+$$

for every eigenvalue λ of M. Then for any fixed sequence of real θ_h, $0 < \theta_h < 1$, $h = 0, 1, \ldots$, we define an HRC process by (5.2), (5.3), and (5.8), for η_h

still denoting an upper bound on the norm $||M_h^{-1}||_2$ but with the matrix I replaced by the matrix $I\sqrt{-1}$. That is, our new HRC algorithm (which can be applied to any Hermitian input matrix M) is defined by the equations

$$M_0 = M + t_0 I\sqrt{-1}, \quad t_0 = \lambda^+/\theta_0, \tag{8.3}$$

$$X_0 = -t_0^{-1} I\sqrt{-1} \tag{8.4}$$

(replacing (5.2)), and

$$M_{h+1} = t_{h+1} I\sqrt{-1} + M = M_h - \Delta_h I\sqrt{-1}, \quad \Delta_h = t_h - t_{h+1} > 0, \quad h = 0, 1, \ldots \tag{8.5}$$

(replacing (5.3)). (8.3)–(8.5) immediately imply bounds (5.4) and (5.9) for $\eta_h \geq ||M_h^{-1}||_2$ and Δ_h of (5.8).

Let us extend our analysis presented in Sections 6 and 7. First note that the equation

$$||M_h^{-1}||_2 = ((t_h^2 + (\lambda^-)^2)^{-1/2} \text{ for all } h \tag{8.6}$$

replaces (5.6). Then again let us simplify the analysis, similarly to Sections 6 and 7. Assume that $\eta_h = (t_h^2 + (\lambda^-)^2)^{-1/2}$ (cf. Remark 5.1) and $\theta_h = \theta$ for all h. It follows that

$$t_{h+1} = t_h - \Delta_h = t_h - (t_h^2 + (\lambda^-)^2)^{1/2}\theta \ < t_h - \theta \max\{t_h, \lambda^-\}, \quad h = 0, 1, \ldots.$$

Therefore, $t_{h+1} < 0$ where $(1-\theta)^h t_0 \leq \theta\lambda^-$. Substitute $t_0 = \lambda^+/\theta$ and obtain that $t_H \leq 0$ where

$$H - 1 = \lceil \log(\lambda^+/(\theta^2\lambda^-))/\log(1/(1-\theta)) \rceil.$$

The latter bound is within the term $\eta = 1 + \lceil (\log(1/\theta))/\log(1/(1-\theta)) \rceil$ from bound (6.1) for $\lambda_1^+ = \lambda^+$ and $\lambda_n^- = \lambda^-$. This term is at most 2 for $\theta \geq 1/2$. On the other hand, our estimates of Section 7 for the numbers of critical and refinement steps performed in each homotopic step remain unchanged (these

estimates are completely defined by the parameters ϵ, ν, and θ). Therefore, up to the replacements of λ_n^- by λ^- and λ_1^+ by λ^+ and performing at most $a = \eta\lceil \log_q((\log \nu)/\log \theta)\rceil$ additional RC steps, the estimates of Sections 6-7 apply to the Hermitian indefinite case as well. We view the latter bound a as relatively small and ignore it in Table 8.1, which summarizes our estimates for the overall numbers of RC steps in the HRC processes and non-homotopic RC processes applied to the same general Hermitian matrix M. (Table 8.1 uses γ of (7.1), H of (6.1), (6.3), and $\kappa_+(M)$ equal to either λ_1^+/λ_n^- or λ^+/λ^-.) According to these estimates, the HRC processes use roughly as many RC steps as non-homotopic RC processes for the inversion of a Hermitian positive definite input matrix M where M is positive definite and roughly by twice fewer critical RC steps and as many refinement RC steps where M is indefinite.

Table 8.1: **Numbers of RC steps required for numerical inversion of Hermitian matrices M.**

	RC Processes	HRC Processes
indefinite M	$\log_2 \kappa_+(M) + \gamma + O(1)$	$0.5 \log_2 \kappa_+(M) + \gamma + O(H)$
positive definite M	$0.5 \log_2 \kappa_+(M) + \gamma + O(1)$	$0.5 \log_2 \kappa_+(M) + \gamma + O(H)$

9 RC and HRC Processes with Compression for Structured Matrices

Suppose an RC process with compression has been applied to a structured input matrix M. Then compression of the displacements perturbs the computed approximations to the inverse, and this may destroy convergence, particularly at the critical RC steps, at which the convergence is more fragile. A natural recipe is to use no compression or limited compression until close approximations X_i to M^{-1} are computed. (Recall that Approach I of Section 4 allows us to vary the level of compression by truncating more or fewer singular values.) How close should these approximations be?

(3.9) and Theorems 4.4 and 4.5 show the level of approximation starting at which rapid convergence is guaranteed even under the *maximal compression*, such that the number of the untruncated singular values of the displacements of the computed approximations is set to be equal to the displacement rank of M. On the other hand, the techniques of Section 2 (cf. (2.7) and (2.10)) fall short of even approaching this level. If we start with an initial approximation supplied by the recipes of Section 2, then to reach the desired levels of (3.9) or Theorems 4.4 and 4.5 with using no compression, we should allow an increase of the displacement rank to n, which means complete loss of the matrix structure. In this case already a single RC step would become too expensive in terms of the number of flops involved.

Practically, the non-homotopic structured RC processes may still be effective, however. That is, according to the experiments reported in [PKRCa], the initial approximation policies of Section 2 under the maximal compression or under compression close to the maximal frequently enable sufficiently rapid convergence in the Toeplitz case, suggesting that the estimates of (3.9) and Theorems 4.4 and 4.5 are overly pessimistic.

HRC processes with compression is an alternative approach supported

both experimentally (see [PKRCa]) and theoretically [P92]. It is proved in [P92] that $O((n\log^3 n)\log\kappa_+(M) + (n\log n)\log\log(1/\epsilon))$ flops are sufficient to approximate M^{-1} for an $n\times n$ Toeplitz-like matrix M. The latter bound is supported in [P92] by an HRC algorithm with the maximal compression (to the level of the displacement rank of M) throughout the computations, and the convergence is controlled via the choice of the sizes of the homotopic steps. Further progress could be achieved based on simultaneous optimization of two groups of parameters, that is, the tolerance values θ_h defining the step sizes Δ_h and the levels of compression based on experimental computations.

HRC processes could be further improved for specific structures of the input matrices. For instance, for real non-singular Toeplitz matrices T, one may achieve symmetrization without doubling the matrix size, simply in the transition to the Hankel matrices JT or TJ, which are real symmetric and satisfy the equations $T^{-1} = (JT)^{-1}J = J(TJ)^{-1}$.

On the other hand, the structure of Cauchy or Vandermonde types is not generally preserved in the transition from a matrix M to the matrices M_0 of (5.2) and (8.4). The problem is solved in the next section where we extend the HRC processes to the case where M is a Hermitian matrix and $M_0 = \widehat{M}$ or $M_0 = \widehat{M}\sqrt{-1}$ for any Hermitian and positive definite matrix $\widehat{M}$.

Example 9.1. *Pick matrices* $M = P = \left(\frac{\mathbf{u}_i^*\mathbf{v}_k}{z_i + z_k^*}\right)_{i,k=1}^{n}$ *where* $\mathbf{u}_i$ *and* $\mathbf{v}_k$ *are vectors of a fixed dimension* d; z_i *are scalars,* $Im\, z_i > 0$, *and* z_k^* *are the complex conjugates of* z_k *for all* i *and* k. *Pick matrices define the Nevanlinna–Pick celebrated problem of rational interpolation [BGR90] and the matrix Nehari problem of rational approximation [BGR90a], [GO94b], [OP98]. The problem is solvable if and only if the Pick matrix is positive definite. One may apply our HRC processes, but the Cauchy structure of the Pick matrices* $M = P$ *is not preserved in the transition to the matrices* M_0 *of (5.2) and (8.4). The structure is much better preserved, however, if we choose* $M_h =$

$M + t_h M_0$, $M_0 = \left(\frac{1}{z_i + z_k^*}\right)_{i,k=1}^{n}$ *or, more generally,* $M_0 = \left(\frac{\mathbf{x}_i^* \mathbf{y}_k}{z_i + z_k^*}\right)_{i,k=1}^{n}$ *where* $\mathbf{x}_i$ *and* $\mathbf{y}_k$ *are* l*-dimensional column vectors for a fixed small non-negative integer* l*. Our extension of the HRC processes in the next section covers the above initialization proposed in the case of Pick matrices.*

10 A Homotopic RC Process with a Generalized Initialization Rule

Motivated by the applications to the inversion of structured matrices, let us extend homotopic processes and their analysis by allowing more general choice of the initial matrix M_0.

First assume that M and M_0 is any fixed pair of positive definite matrices, where M_0 is readily invertible, spectrum$(M_0) = \{\mu_1, \dots, \mu_n\}$,

$$\mu_1^+ \geq \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq \mu_n^- > 0, \tag{10.1}$$

and the values μ_1^+ and μ_n^- are available. Now recursively define scalars $t_1, \dots, t_{H-1}$ and matrices

$$M_{h+1} = t_{h+1} M_0 + M = M_h + (t_{h+1} - t_h) M_0, \quad h = 0, 1, \dots, H-1, \tag{10.2}$$

where $t_1 > t_2 > \cdots > t_{H-1} > t_H = 0$.

One may rewrite (10.2) as $M_{h+1} = M_0(t_h I + M_0^{-1} M)$ and apply our previous study to the inversion of the matrix $M_0^{-1} M$, but we avoid shifting to this matrix directly. We deduce that

$$||I - (t_1 M_0)^{-1} M_1||_2 \leq ||M_0^{-1} M / t_1||_2 \leq ||M_0^{-1}||_2 ||M||_2 / t_1 \leq \lambda_1^+ / (t_1 \mu_n^-),$$

for λ_1^+ of (5.1) and choose

$$t_1 = \lambda_1^+ / (\theta_0 \mu_n^-) \tag{10.3}$$

so that $||I - (t_1 M_0)^{-1} M_1||_2 \leq \theta_0$. Invert M_1 by applying processes (2.1) for $X_0 = t_1 M_0$.

Now deduce from (10.2) that

$$I - M_h^{-1} M_{h+1} = (t_h - t_{h+1}) M_h^{-1} M_0,$$
$$||I - M_h^{-1} M_{h+1}||_2 \leq (t_h - t_{h+1}) ||M_h^{-1}||_2 ||M_0||_2. \tag{10.4}$$

Substitute the bound

$$||M_0||_2 \leq \mu_1^+$$

and obtain that $||I - M_h^{-1} M_{h+1}||_2 \leq \theta_h$ if $(t_h - t_{h+1}) \mu_1^+ ||M_h^{-1}||_2 \leq \theta_h$ or, equivalently, if $t_{h+1} \geq t_h - \theta_h / (\mu_1^+ ||M_h^{-1}||_2)$. Recall that, clearly,

$$||M_h^{-1}||_2 \leq 1/(t_h \mu_n^- + \lambda_n^-)$$

for all h and for λ_n^- of (5.1) [Par80, p.191], write

$$t_{h+1} = t_h - (t_h \mu_n^- + \lambda_n^-) \theta_h / \mu_1^+, \tag{10.5}$$

and deduce (5.9). Now, invert the matrices M_{h+1} by applying processes (2.1) for $X_0 = M_h^{-1}$ and for $h = 1, 2, \ldots, H-2$, until the value t_{h+1} of (10.5) becomes non-positive for $h = H-1$. Then at the last homotopic step, invert M instead of M_H.

Clearly, the estimates of Section 7 for the number of RC steps at each homotopic step apply to the above generalized HRC process, too.

Let us next estimate the number of homotopic steps H, in terms of the parameters t_1, θ_h, $\kappa^+ = \mu_1^+ / \mu_n^-$, the lower bounds λ_n^- and μ_n^- on the eigenvalues of the matrices M and M_0. Substitute the expression $\kappa^+ = \mu_1^+ / \mu_n^-$ into (10.5) for $h = 0, 1, \ldots, H-1$ and obtain that

$$t_{h+1} = t_h (1 - \theta_h / \kappa^+) - \theta_h \lambda_n^- / \mu_1^+,$$

$$\begin{aligned} t_{h+1} + \kappa^+ \lambda_n^- / \mu_n^- &= (t_h + \kappa^+ \lambda_n^- / \mu_1^+)(1 - \theta_h / \kappa^+) \\ &= (t_1 + \kappa^+ \lambda_n^- / \mu_n^-) \prod_{i=0}^{h} (1 - \theta_i / \kappa^+). \end{aligned} \tag{10.6}$$

Therefore, we have $t_{h+1} \le 0$ if

$$(t_1 + \kappa^+\lambda_n^-/\mu_n^-)\prod_{i=0}^{h}(1-\theta_i/\kappa^+) \ge \kappa^+\lambda_n^-/\mu_n^-,$$

that is, if

$$1 + t_1\mu_n^-/(\lambda_n^-\kappa^+) \ge 1/\prod_{i=0}^{h}(1-\theta_i/\kappa^+).$$

Assuming that $\theta_h = \theta$ is invariant in h, we arrive at $t_H \le 0$ for

$$H = 1 + \lceil(\log(1 + t_1\mu_n^-/(\lambda_n^-\kappa^+)))/(\log(1-\theta/\kappa^+)^{-1})\rceil \qquad (10.7)$$

and t_1 of (10.3).

Finally, if M is any non-singular matrix, we may apply symmetrization recipes (8.1) or (8.2) to extend our algorithm of this section. In particular, recipe (8.2) reduces the problem to the case where M is a Hermitian (or real symmetric) but not necessarily positive definite matrix. Then we may extend HRC process (10.2)–(10.5) where we keep equations (10.2)–(10.3), choose the matrix M_0 equal to $\widehat{M}\sqrt{-1}$ for a fixed positive definite matrix $\widehat{M}$, and modify (10.4)–(10.5) to ensure that $||I - M_h^{-1}M_{h+1}||_2 \le \theta_h$ for all h.

Let us complete the description of such an extended homotopic process. Assume that bounds (10.1) still hold where $\{\mu_1 \ldots, \mu_n\} = \text{spectrum}(M)$ and each eigenvalue λ of the input matrix M satisfies the bounds

$$0 < \lambda^- \le |\lambda| \le \lambda^+ \qquad (10.8)$$

for two fixed positive values λ^- and λ^+. Now write

$$t_{h+1} = t_h - (\theta_h/\mu_1^+)((\lambda^-/\kappa^+)^2 + (t_h\mu_n^-)^2)^{1/2}, \quad \kappa^+ = \mu_1^+/\mu_n^-, \qquad (10.9)$$

$h = 0, 1, \ldots, H-1$.

Let us deduce bounds (5.9). Recall the following well-known theorem [Par80, proof of Theorem 15-3-3].

Theorem 10.1. *Let M and $\widehat{M}$ be two Hermitian matrices. Let the matrix $\widehat{M}$ be positive definite, such that*

$$\widehat{M} = U\Sigma^2 U^* \tag{10.10}$$

for a unitary matrix U, $U^*U = UU^* = I_n$, and a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_i)_{i=1}^n$, $\mu_1^+ \geq \sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_n^2 \geq \mu_n^- > 0$. Then there exists a unitary matrix V, $V^*V = VV^* = I_n$, such that

$$D = V^*\Sigma^{-1}U^*MU\Sigma^{-1}V \tag{10.11}$$

is a real diagonal matrix.

Corollary 10.1. *Under the notation of (10.1), (10.8), and Theorem 10.1, we have*

$$\|M_h^{-1}\|_2^2 \leq (\mu_n^-)^{-2}((\lambda^-/\mu_n^+)^2+t_h^2)^{-1} = ((\lambda^-/\kappa^+)^2+(t_h\mu_n^-)^2)^{-1} \text{ for } h = 1, 2, \ldots$$

where $\kappa^+ = \mu_1^+/\mu_n^-$.

Proof. By combining (10.10) and (10.11), obtain that

$$M_h = M + t_h\sqrt{-1}\widehat{M} = U\Sigma V(D + t_h I\sqrt{-1})V^*\Sigma U^*,$$

$$M_h^{-1} = U\Sigma^{-1}V(D + t_h I\sqrt{-1})^{-1}V^*\Sigma^{-1}U^*.$$

Therefore,

$$\|M_h^{-1}\|_2 \leq \|\Sigma^{-2}\|_2 \, \|(D + t_h I\sqrt{-1})^{-1}\|_2 \leq \left(\frac{1}{\mu_n^-}\right)^2 \left(\frac{1}{\|D^{-1}\|_2^2} + t_h^2\right)^{-0.5}$$

On the other hand, we deduce from (10.1), (10.8), and (10.11) that

$$\|D^{-1}\|_2 \leq \|\Sigma^2\|_2 \, \|M^{-1}\|_2 \leq \mu_1^+/\lambda^-.$$

Substitute the latter bound into our estimate for the norm $\|M_h^{-1}\|_2$ and obtain that

$$\|M_h^{-1}\|_2^2 \leq (\mu_n^-)^{-2}((\lambda^-/\mu_1^+)^2 + t_h^2)^{-1} = ((\lambda^-/\kappa^+)^2 + (t_h\mu_n^-)^2)^{-1}.$$

□

Relations (10.1), (10.2), (10.4), (10.9), and Corollary 10.1 together immediately imply (5.9). Let us compare the estimate of Corollary 10.1 and the bound $\|M_h^{-1}\|_2 \le 1/(t_h\mu_n^- + \lambda_n^-)$. The two estimates are close to one another provided that the terms λ_n^- and λ^-/κ^+ are dominated by the term $t_h\mu_n^-$. If the term λ^-/κ^+ dominates, the bound of Corollary 10.1 may be larger by roughly the factor of $\kappa^+\lambda_n^-/\lambda^-$.

(10.9) implies the crude bounds

$$t_{h+1} \le t_h - (\theta_h/\mu_1^+)(\lambda^-/\kappa^+ + t_h\mu_n^-),\ h = 1, 2, \ldots.$$

Consequently,

$$t_{h+1} + \lambda^-/\mu_1^+ \le (1 - \theta_h/\kappa^+)(t_h + \lambda^-/\mu_1^+) \le \ldots \le (t_1 + \lambda^-/\mu_1^+)\prod_{i=1}^{h}(1 - \theta_i/\kappa^+).$$

The latter inequality implies that the value t_H is non-positive for

$$H \le 1 + \lceil(\log(1 + t_1\mu_1^+/\lambda^-))/\log(1 - \theta/\kappa^+)^{-1}\rceil$$

provided that $\theta_h = \theta$ for all h.

11 Extensions and Generalizations

It is well known and easily verified that unscaled RC processes (2.1), (2.3) and scaled processes (2.7), (2.8) converge to the Moore–Penrose generalized inverse M^+ where the input matrix M is singular. Now recall that scaled RC processes (2.7), (2.8), (2.11), (2.12) converge to the numerical generalized inverse matrix M_ϵ^+. The analysis and the estimates of our thesis (including the ones for the HRC processes) can be extended provided that the 2-norms $\sigma_r^{-2}(W) = \|W^{-1}\|_2$ are replaced throughout by $\sigma_{r(\epsilon)}^{-2}(W)$, where $\sigma_{r(\epsilon)}^2(W)$ is the smallest singular value of the matrix W not exceeded by ϵ. This enables

various refinements from noisy perturbations of the input. Furthermore, the computation of M_ϵ^+ does not depend on whether the matrix M is singular or not. In particular, we may apply HRC processes to compute M_ϵ^+ for a positive ϵ where M is singular. If ϵ is small enough, the HRC processes output $M^+ = M_\epsilon^+$, even though the same processes may diverge if we apply them directly to M and use the iteration (2.1), (2.3) or (2.7), (2.8) as a Basic Subroutine.

For the extension of the RC and HRC methods to the computation of the numerical generalized inverse M_ϵ^+ (and in particular $M^+ = M_0^+$) for a structured matrix M, an additional problem is the compression because the displacement $L(M)$ does not completely define the matrix M_ϵ^+ even for $\epsilon = 0$. For Toeplitz and Hankel matrices and for $\epsilon = 0$, the problem can be circumvent [HH93], [HH94]. The following simple results solve the problem also for other classes of structured matrices wherever $\mathrm{rank}(M_\epsilon^+ M - I) = n - r_\epsilon$ is small.

Theorem 11.1. *For any positive ϵ and any triple of $n \times n$ matrices A, B, and M we have*

$$\nabla_{B,A}(M_\epsilon^+) = M_\epsilon^+ A(MM_\epsilon^+ - I) - (MM_\epsilon^+ - I)BM_\epsilon^+ - M_\epsilon^+ \nabla_{A,B}(M) M_\epsilon^+.$$

Corollary 11.1. *Under the assumptions of Theorem 11.1, we have*

$$\mathrm{rank}(\nabla_{B,A}(M_\epsilon^+)) \le \mathrm{rank}(\nabla_{A,B}(M)) + 2n - 2r_\epsilon$$

where $r_\epsilon = \mathrm{rank}(M_\epsilon^+) = \mathrm{rank}(M_\epsilon)$.

Note that the level of the truncation of the singular values in Approach I can be defined by Corollary 11.1.

Finally, here is a sample generalization of the HRC process (5.3):

$$M_{h+1} = M_0 F_{h+1} + M = M_h + M_0(F_{h+1} - F_h) \qquad (11.1)$$

where $F_0, F_1, F_2, \ldots$ is a sequence of matrices converging to 0 and satisfying $\|M_h^{-1} M_0(F_{h+1} - F_h)\| \le \theta_h$ for fixed scalars $\theta_h < 1$, $h = 0, 1, \ldots$.

12 Conclusion

We first recalled RC (residual correction) algorithms and then specified and analyzed homotopic RC algorithms for matrix inversion and generalized inversion. The homotopic algorithms require as many matrix multiplications as the best non-homotopic RC algorithms for the inversion of unstructured positive definite matrices and substantially fewer matrix multiplications for Hermitian indefinite matrices. The homotopic RC processes (unlike their non-homotopic counterparts) generate initial approximation to the inverse where it is not available from outside. The latter feature supports the reliable application of the homotopic RC algorithms to the inversion of structured matrices where recursive compression of the displacements of computed approximate inverses enables all the matrix multiplications in nearly linear time. Non-homotopic RC algorithms have no proved rapid convergence results where compression is maintained throughout but in fact the results of experimental tests with Toeplitz matrices are more optimistic than the theoretical estimates.

We conclude with listing some natural directions for further theoretical and experimental study (in this listing we use the notation of our thesis).

1. Theoretical and experimental estimation of the order of convergence q for Structured RC Algorithms.

2. The choice of the parameters p and c_{i+1} for RC process (2.1) for structured matrices (cf. [FF63, Chapter 9] and [PS91] on the similar problem for general unstructured matrices) and for the homotopic versions of this process.

3. Generalization of the HRC processes such as (11.1).

4. Experimental tests of the presented results for general matrices and for structured matrices of various classes; experimental specification and

optimization of the parameters t_0 of (5.2), (8.4), t_1 of (10.3), θ_h and ν (ν defining the stopping criteria at the intermediate homotopic steps) provided that variation of the parameters θ_h with h is allowed in the case of RC processes with the compression of the displacements $L(X_i)$.

5. Specification and analysis of homotopic and non-homotopic processes for structured matrices where the compression is weakened (by the truncation of fewer singular values); the choice of the best balance in combining weakened compression with the homotopy, in particular simultaneous optimization of the compression level in Approach I of Section 4 and homotopic step sizes.

References

[AG89] G. S. Ammar, P. Gader, New Decompositions of the Inverse of a Toeplitz Matrix, *Proc. 1989 Int. Symp. on Math Theory of Networks and Systems (MTNS'89)*, Amsterdam, 1989.

[BGR90] J. A. Ball, I. Gohberg, L. Rodman, Interpolation of Rational Matrix Functions, *Operator Theory: Advances and Applications*, **45**, Birkhäuser, Basel, 1990.

[BGR90a] J. A. Ball, I. Gohberg, L. Rodman, Nehari Interpolation Problem for Rational Matrix Functions: the Generic Case, H_∞*-control Theory* (E. Mosca, L. Pandolfi, editors), 277–308, Springer, Berlin, 1990.

[B-I66] A. Ben-Israel, A Note on Iterative Method for Generalized Inversion of Matrices, *Math. Comp.*, **20**, 439–440, 1966.

[B-IC66] A. Ben-Israel, D. Cohen, On Iterative Computation of Generalized Inverses and Associated Projections, *SIAM J. Numer. Anal.*, **3**, 410–419, 1966.

[BM,a] D. A. Bini, B. Meini, Approximate Displacement Rank and Applications, preprint.

[BP93] D. A. Bini, V. Y. Pan, Improved Parallel Computations with Toeplitz-like and Hankel-like Matrices, *Linear Algebra and Its Applications*, **188/189**, 3–29, 1993.

[BP94] D. Bini, V. Y. Pan, *Polynomial and Matrix Computations, Volume 1: Fundamental Algorithms*, Birkhäuser, Boston, 1994.

[BP98] D. Bini, V. Y. Pan, Computing Matrix Eigenvalues and Polynomial Zeros Where the Output is Real, *SIAM J. on Computing*, **27, 4**, 1099–1115, 1998.

[CKL-A87] J. Chun, T. Kailath, H. Lev-Ari, Fast Parallel Algorithm for QR-factorization of Structured Matrices, *SIAM Journal on Scientific and Statistical Computing*, **8, 6**, 899–913, 1987.

[CN96] R. H. Chan, M. K. Ng, Conjugate Gradient Methods for Toeplitz Systems, *SIAM Review*, **38**, 427–482, 1996.

[CN99] R. H. Chan, M. K. Ng, Iterative Methods for Linear Systems with Matrix Structure, *SIAM Volume on Fast Reliable Algorithms for Matrices with Structure*, (T. Kailath, A.H. Sayed, Editors) 117–152, SIAM Publications, Philadelphia, 1999.

[FF63] D. K. Faddeev, V. N. Faddeeva, *Computational Methods of Linear Algebra*, W. H. Freeman, San Francisco, 1963.

[GE95] M. Gu, S. C. Eisenstat, A Divide-and-Conquer Algoritm for the Symmetric Tridiagonal Eigenproblem, *SIMAX*, **16, 1**, 172–191, 1995.

[GKO95] I. Gohberg, T. Kailath, V. Olshevsky, Fast Gaussian Elimination With Partial Pivoting for Matrices with Displacement Structure, *Math. of Computation*, **64**, 1557–1576, 1995.

[GL96] G. H. Golub, C. F. Van Loan, *Matrix Computations*, Johns Hopkins Univeristy Press, Baltimore, Maryland, 1989 (2nd edition), 1996 (3rd edition).

[GO94] I. Gohberg, V. Olshevsky, Complexity of Multiplication with Vectors for Structured Matrices, *Linear Algebra Appl.*, **202**, 163–192, 1994.

[GO94b] I. Gohberg, V. Olshevsky, Fast State Space Algorithms for Matrix Nehari and Nehari–Takagi Interpolation Problems, *Integral Equations and Operator Theory*, **20, 1**, 44–83, 1994.

[GS72] I. Gohberg, A. Semencul, On the Inversion of Finite Toeplitz Matrices and Their Continous Analogs, *Mat. Issled.*, **2**, 187–224, 1972.

[H95] G. Heinig, Inversion of Generalized Cauchy Matrices and the Other Classes of Structured Matrices, *Linear Algebra for Signal Processing, IMA Volume in Math. and its Applications*, **69**, 95–114, Springer, 1995.

[HH93] G. Heinig, F. Hellinger, On the Bezoutian Structure of the Moore–Penrose Inverses of Hankel Matrices, *SIAM J. on Matrix Analysis and Applications*, **14, 3**, 629–645, 1993.

[HH94] G. Heinig, F. Hellinger, Moore–Penrose Generalized Inverse of Square Toeplitz Matrices, *SIAM J. on Matrix Analysis and Applications*, **15, 2**, 418–450, 1994.

[HR84] G. Heinig, K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators, in Operator Theory: Advances and Applications*, (I. Gohberg editor), **13**, Birkhäuser, Basel, 1984.

[IK66] E. Issacson, H. B. Keller, *Analysis of Numerical Methods*, Wiley, New York, 1966.

[KKM79] T. Kailath, S. Y. Kung, M. Morf, Displacement Ranks of Matrices and Linear Equations, *J. Math. Anal. Appl.*, **68, 2**, 395–407, 1979.

[KS99] T. Kailath, A. H. Sayed (Editors), *Fast Reliable Algorithms for Matrices with Structure*, SIAM Publications, Philadelphia, 1999.

[OP98] V. Olshevsky, V. Y. Pan, A Unified Superfast Algorithm for Boundary Rational Tangential Interpolation Problem, *Proc. 39th Ann. IEEE Symp. Foundations of Comp. Sci.*, 192–201, IEEE Comp. Soc. Press, 1998.

[P90] V. Y. Pan, On Computations with Dense Structured Matrices, *Math. Comp.*, **55, 191**, 179–190, 1990. Proceeding version: *Proc. Intern. Symp. on Symbolic and Algebraic Comp. (ISSAC'89)*, 34–42, ACM Press, New York, 1989.

[P92] V. Y. Pan, Parallel Solution of Toeplitz-like Linear Systems, *J. of Complexity*, **8**, 1–21, 1992.

[P92a] V. Y. Pan, Parametrization of Newton's Iteration for Computations with Structured Matrices and Applications, *Computers & Mathematics (with Applications)*, **24, 3**, 61–75, 1992.

[P93] V. Y. Pan, Decreasing the Displacement Rank of a Matrix, *SIAM J. Matrix Anal. Appl.*, **14, 1**, 118–121, 1993.

[P93a] V. Y. Pan, Concurrent Iterative Algorithm for Toepliz-like Linear Systems, *IEEE Trans. on Parallel and Distributed Systems*, **4, 5**, 592–600, 1993.

[P00] V. Y. Pan, A Homotopic Residual Correction Process, *Proc. of the Second Conference on Numerical Analysis and Applications*, Rousse, Bulgaria, 2000 (P. Yalamov, Editor), *Lecture Notes in Computer Science*, **1196**, Springer, 2000.

[P00a] V. Y. Pan, New Techniques for the Computation of Linear Recurrence Coefficients, *Finite Fields and Their Applications*, **6**, 43–118, 2000.

[P00b] V. Y. Pan, Parallel Complexity of Computations with General and Toeplitz-like Matrices Filled with Integers and Extensions, *SIAM Journal on Computing*, **30**, 1080–1125, 2000.

[Pa] V. Y. Pan, New Effective Algorithms for Structured Matrices, to appear.

[P01] V. Y. Pan, A Homotopic Residual Correction Process, *Proceedings of the Second Conference on Numerical Analysis and Applications* (P. Yalamov, editor), *Lecture Notes in Computer Science*, **1988**, Springer, Berlin, 2001.

[Par80] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[PBRZ99] V. Y. Pan, S. Branham, R. Rosholt, A. Zheng, Newton's Iteration for Structured Matrices and Linear Systems of Equations, *SIAM volume on Fast Reliable Algorithms for Matrices with Structure*, (T. Kailath, A.H. Sayed, Editors) 189–210, SIAM Publications, Philadelphia, 1999.

[PKRCa] V. Y. Pan, M. Kunin, R. E. Rosholt, H. Cebecioğlu, Homotopic Residual Correction Algorithms for General and Structured Matrices, to appear.

[PR01] V. Y. Pan, Y. Rami, Newton's iteration for the Inversion of Structured Matrices, *Structured Matrices: Recent Developments in Theory and Computation*, edited by D. A. Bini, E. Tyrtyshnikov and P. Yalamov, Nova Science Publishers, USA, 2001.

[PRWa] V. Y. Pan, Y. Rami, X. Wang, Structured Matrices and Newton's Iteration: Unified Approach, preprint. (Proc. version in *Proc. 14th Intern. Symposium on Math. Theory of Network and Systems*

(MTNS'2000), University of Perpignan, Perpignan, France, June 2000.)

[PS91] V. Y. Pan, R. Schreiber, An Improved Newton Iteration for the Generalized Inverse of a Matrix, with Applications, *SIAM J. on Scientific and Statistical Computing*, **12, 5**, 1109–1131, 1991.

[PWa] V. Y. Pan, X. Wang, Inversion of Displacement Operators, to appear.

[PZHD97] V. Y. Pan, A. Zheng, X. Huang, O. Dias, Newton's Iteration for Inversion of Cauchy-like and Other Structured Matrices, *J. of Complexity*, **13**, 108–124, 1997.

[S33] G. Schultz, Iterative Berechnung der Reciproken Matrix, *Z. Angew. Meth. Mech.*, **13**, 57–59, 1933.

[SS74] T. Söderström, W. Stewart, On the Numerical Properties of an Iterative Method for Computing the Moore–Penrose Generalized Inverse, *SIAM J. Numer. Anal.*, **11**, 61–74, 1974.

[VHKa] M. Van Barel, G. Heinig, P. Kravanja, A Stabilized Superfast Solver for Nonsymmetric Toeplitz Systems, Report TW293