2002

# TR-2002015: Residual Correction Algorithms for General and Structured Matrices

V. Y. Pan

M. Kunin

R. E. Rosholt

H. Cebecioglu

# Residual Correction Algorithms for General and Structured Matrices *†

V. Y. Pan [‡,§,¶] M. Kunin [§] R. E. Rosholt [§] H. Cebecioğlu [¶]

November 7, 2001

**Summary:** We present and analyze residual correction algorithms for the computation of the inverses, generalized inverses, and numerical generalized inverses of general and structured matrices. For structured matrices, we extend the known algorithms by new policies of compression delay. For both structured and unstructured matrices, we propose and analyze the homotopic (continuation) methods, which supply close initial approximations. For unstructured indefinite Hermitian input matrices, the homotopy methods enable substantial acceleration of the known best non-homotopic algorithms. Furthermore, by using homotopic techniques, we guarantee superlinear convergence to the inverses of structured matrices even where no initial approximation is available and where compression of displacement generators is ensured in every residual correction step. Numerical tests with Toeplitz input matrices show a greater power of both homotopic and non-homotopic approaches than the theoretical study predicts.

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

**2000 AMS Math. Subject Classification: 65F10, 65F30**

*Supported by NSF Grant CCR9732206, PSC CUNY Awards 62435-0031 and 63383-0032, and a Grant from the CUNY Institute for Software Design and Development (CISDD).

†The results of this paper have been presented at the Second Conference on Numerical Analysis and Applications, Rousse, Bulgaria, in June 2000, and at the AMS/IMS/SIAM Summer Research Conference on Fast Algorithms in Mathematics, Computer Science, and Engineering in South Hadley, Massachusetts, in August 2001.

‡Mathematics and Computer Science Department, Lehman College, CUNY, Bronx, NY 10468, USA (vpan@lehman.cuny.edu)

§Ph. D. Program in Computer Science, Graduate Center, CUNY

¶Ph. D. Program in Mathematics, Graduate Center, CUNY

# 1 Introduction

## 1.1 Residual correction (RC) processes

We study *residual correction processes* converging to 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 basic processes perform matrix multiplication $p$ times in each step to achieve convergence of the order of $p$, for any $p \geq 2$. With appropriate scaling, however, one may reach the order of $p > 2$ by performing matrix multiplication only twice per step [PS91]. Hereafter, we write RC for "residual correction" and MM for "matrix multiplication". For $p = 2$ unscaled RC processes turn into *Newton's iteration*. The RC processes can be directed to converge to numerical generalized inverses and are known for their strong numerical stability and self-correcting property [PS91].

For simplicity until Section 11, we assume non-singularity of the input matrices $M$. Here are the two main problems with RC processes.

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 quite a few, the order of $\log_2 \kappa(M)$, RC steps ($\kappa(M) = \text{cond}(M)$ denoting the condition number of the matrix $M$) to refine the approximation to the level from which the iteration very rapidly converges. Our main topic is an alternative approach based on the homotopic (continuation) techniques.

b) For general matrices, MM is an expensive operation, comparable to matrix inversion in its computational cost, although substantially simpler than the computation of the generalized inverse and allowing effective parallel implementation.

   Furthermore, MM 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 matrix structure and compressed representation of matrices are preserved throughout the computation without destroying rapid convergence of the process. Here some advanced compression techniques are applied, first proposed in [P92] and then elaborated upon in [PZHD97], [PBRZ99], [PR01], [PRW01], [P01a].

For the sake of completeness of our study, we briefly review this development in Sections 2–4. Furthermore, we propose some new techniques to improve practical performance of the known algorithms in the case of structured input matrices. These techniques rely on the *delayed compression* and *modular arithmetic in the real field* (see Section 2.5, Remarks 3.2 and 4.2, and equations (4.8)–(4.11)).

## 1.2 Homotopic RC processes

The solution techniques for problems a) and b) do not always match one another. That is, compression perturbs computed approximations and may easily destroy convergence at the initial stages of the RC processes where the convergence is fragile. This implies some additional requirements. We must either yield much closer initial approximations versus the known techniques of [B-I66], [B-IC66], [SS74], and [PS91] or perform much more work per iteration step to yield compression with much smaller perturbation of the computed approximations to the inverse. The original approach in [P92] achieves the latter goal for the class of Toeplitz and Toeplitz-like matrices by developing the *homotopic* (or *continuation*) method but also allows a very natural heuristic modification towards this goal. The approach in [P92] relies on truncating the smallest singular values of the displacements. Tests show that the heuristic is surprisingly effective in the important case of Toeplitz input matrices, but no theoretical results support this development.

In the present paper, we recall the RC algorithms for general and structured matrices, show their improved variations for Toeplitz and other sturctured matrices (see (3.7), (3.8), (4.8))–(4.11)), and report the results of numerical experiments in Section 12; otherwise our main subjects are new techniques for computing the initial approximation. Their efficiency is confirmed by both experiments and the proved estimates for the computational work. The new methods extend the homotopic (continuation) techniques of [P92] relying on the inversion of a readily invertible matrix $M_0$ (e.g., $M_0 = I$) and the subsequent homotopic transition to the matrix $M$ along the trajectories

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

or

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

where

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

$t_0$ is 1 in (1.1) and a sufficiently large value in (1.2). We arrange the homotopy to keep the trajectories $M(t)$ away from singular matrices 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 $t = 0$ (cf. our earlier *techniques of variable diagonal* [P00b]). 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 $(h + 1)$-st homotopic step.

## 1.3 Related works

Newton's iteration for the inverse and generalized 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]). In [PS91] Newton's iteration was accelerated by using scaling and extended to the computation of the numerical generalized inverses of a matrix; furthermore strong numerical stability of the original and modified iterative processes was proved. In [P92] Newton's iteration was worked out for Toeplitz-like matrices (with the compression of the displacement by means of truncating its singular values), and the homotopic process for the initialization was proposed and analyzed. The paper also included estimates for the perturbation of the computed approximations to the inverse caused by the compression (the problem was further studied in [P93]) and the proof 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 described 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]) elaborated upon Newton's iteration under both approaches to the compression in the Toeplitz-like case; the subsequent paper [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]. Further work on the homotopy approach, extending [P92], was reported in the short proceedings paper [P01] and surveyed in the book [P01a]. A unified method for the extension of Newton's iteration to various classes of structured matrices was proposed and analyzed in [P01], [PR01], and [PRW01]. 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 paper

In Sections 2–4, we recall some known results on the RC processes for general input matrices and show some improvements in the case of structured matrices. In Sections 5–8 and 10, we elaborate upon the choice 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). We show the results of some numerical experiments in Section 12. Section 13 is left for a brief conclusion. Section 12 was co-authored by the first three authors, Section 9 by Pan and Cebecioğlu, other sections are due to Pan.

# 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 $\sigma_j = \sigma_j(M)$, $\kappa(M) = \sigma_1/\sigma_r$. $\sigma_j$ denote the singular values of a matrix $M$ where $r = \text{rank}(M)$, $j = 1, \dots, r$; $0 < \sigma_- \le \sigma_r \le \dots \le \sigma_1 \le \sigma_+$; $\kappa(M)$ is the condition number of $M$. $\mathbf{e}_{i-1}$ denotes the $i$-th coordinate vector, $i = 1, \dots, n$. $\lceil x \rceil$ is the smallest among the integers not exceeded by a real $x$.

## 2.1 A Basic RC Process

A sufficiently close initial approximation $X_0$ to the inverse of a non-singular matrix $M$ can be rapidly improved by means of a scaled RC process [IK66]

$$\Delta_i = \Delta(X_i) = X_{i+1} - c_{i+1}X_i = c_{i+1}\sum_{k=1}^{p-1} R_i^k X_i, \quad i = 0, 1, \dots, \tag{2.1}$$

where we write

$$R_i = R(M, X_i) = I - X_i M = R_{i-1} - M(X_{i+1} - X_i). \tag{2.2}$$

For $p = 2$, $c_{i+1} = 1$ for all $i$, we arrive at Newton's iteration [S33]. For $p = 2^h$, we may compute $\sum_{k=0}^{p-1} R_i^k$ as $\prod_{j=0}^{h-1}(I + R_i^{2^j})$ using fewer additions. Already for the unscaled process, that is, under the simplest choice

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

(2.1) and (2.2) imply that

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

That is, 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 $\theta$. Then the computational work required to ensure the desired upper bound on the norm $\|R_i\|$ is minimized for $p = 3$ [IK66, pages 86-88].

## 2.2 An Initial Approximation

For an initial approximation $X_0$ to the matrix $M^{-1}$, one may choose [B-I66], [SS74]

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

to yield that

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

Now it follows that the first

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

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

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

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

The asymptotic bound $i_- = \log_2 \kappa(M) + O(1)$ on the number of critical RC steps is achieved in [B-I66] 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]: we have

$$||R_0||_2 \le 1 - \frac{1}{\sqrt{n}\kappa(M)} \quad \text{for} \quad X_0 = I/||M||_F \tag{2.7}$$

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

## 2.3 Scaled Newton's Iteration

The choice of $c_{i+1}$ in (2.1) was optimized in [PS91] in the case of RC process (2.1) for $p = 2$:

$$X_{i+1} = c_{i+1}(I + R_i)X_i = c_{i+1}(2X_i - X_iMX_i). \tag{2.8}$$

Namely, by choosing $p = 2$,

$$c_0^- = \frac{2\sigma_-}{\sigma_+ + \sigma_-}, \quad c_{i+1} = \frac{2}{1 + (2 - c_i^-)c_i^-}, \quad c_{i+1}^- = (2 - c_i^-)c_i^- c_{i+1} \tag{2.9}$$

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

$$||R_l||_2 \le \max_{\sigma_- \le x \le \sigma_+} |T_{2^l}(\gamma x + \delta)|/|T_{2^l}(\delta)| \le \frac{1}{|T_{2^l}(\delta)|} \tag{2.10}$$

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

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

This bound is substantially smaller than $\delta^L$. The number of critical steps decreases roughly by twice versus policy (2.3), reaching the level

$$i = \log_2 \kappa_+(M) + O(1/\kappa_+^2(M)). \tag{2.11}$$

In other words, the impact of the optimal scaling of (2.9) is equivalent to increasing the order of convergence of the critical steps from $q = 2$ to $q = 4$.

## 2.4 RC Processes for Numerical Generalized Inverse

The paper [PS91] also proposes a modification where Newton's RC processes for $p = 2$ converge to a numerical generalized (Moore–Penrose) inverse $M_\epsilon^+$, that is, the generalized inverse of the matrix $M_\epsilon$ formed via truncating the smallest singular values of $M$ (up to a fixed tolerance $\epsilon$). This is achieved by first applying iteration (2.8)–(2.9) with

$$c_0 = \sigma_+ c, \ c_0^- = c\epsilon^2, \ c = \min(2/(\sigma_+ + \epsilon^2), \rho/\epsilon^2), \tag{2.12}$$

$\rho = (1 + \sqrt{3})/2 = 1.366\ldots$ (Under the scaling of (2.12), the value $\rho$ partitions the range for the spectrum of the matrix $X_0 M$; the partition is induced by the respective partition by $\epsilon$ of the singular values of the matrix $M$. Note that the bound $\sigma_-$ is not needed in this variation of the iteration.) The iteration is performed until we arrive at $c_i^- \ge \rho$ 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

$$X_{i+1} = (-2X_i M + 3I)X_i M X_i, \quad i = 0, 1, \ldots . \tag{2.13}$$

(This is a generalizations of RC process (2.1) and a special case of a more general process

$$X_{i+1} = p_i(X_i M)X_i \tag{2.14}$$

where $p_i(y)$ are selected polynomials, $\quad i = 0, 1, \ldots.$) Based on (2.13), the singular values $\sigma_j(M)$ are partitioned by $\epsilon$ into two groups: those exceeding $\epsilon$ correspond to the eigenvalues $\lambda^{(i)}$ of $X_i M$ that lie in the interval $1/2 < \lambda^{(i)} \le \rho$; iteration (2.13) sends them towards 1. The other eigenvalues of $X_i M$ 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 $M_\epsilon^+$. Convergence is ultimately quadratic but is slow near $1/2$ and $\rho$. Iteration can be immediately extended to computing the matrices $M_\epsilon = MM_\epsilon^+ M$ and $\tilde{M}_\epsilon = M - M_\epsilon$ and the numerical rank trace$(M_\epsilon M_\epsilon^+)$.

## 2.5 Bounding the Precision of Computing

It was proved in [PS91] that both original and modified Newton's processes are numerically stable. Process (2.1), however, involves the expression

$$c_{i+1}(I + \sum_{k=0}^{p-1} R_i^k)X_i,$$

whose representation for a smaller $\|R_i\|$ requires the $p$-fold precision versus the single precision for representation of $M$ and $X_i$. For $p = 2$, $c_{i+1} = 1$, the precision growth in the RC process (2.1) can be avoided based on using *modular arithmetic in the real field* [P92b], [EPY98].

For the task of solving a linear system $M\mathbf{x} = \mathbf{b}$:

$$\mathbf{x}_1 = X_0\mathbf{b}, \quad \mathbf{r_1} = \mathbf{b} - M\mathbf{x_1}, \tag{2.15}$$

the precision can be controlled if we apply iterative improvement process

$$\Delta_i = \mathbf{x}_{i+1} - \mathbf{x}_i = X_0\mathbf{r}_i, \quad \mathbf{r}_{i+1} = \mathbf{r}_i - M(\mathbf{x}_{i+1} - \mathbf{x}_i), \quad i = 1, \dots, s. \tag{2.16}$$

This process invloves neither residual matrices $R_i$ nor higher precision approximations $X_i$ to $M^{-1}$, but the approximation error norm $\|\mathbf{x} - \mathbf{x}_{i-1}\|$ decreases by the factor of $\|R_0\| = \|I - X_0M\|$ in each iteration step: $\mathbf{x} - \mathbf{x}_i = (I - X_0M)(\mathbf{x} - \mathbf{x}_{i-1}) = (I - X_0M)^i(\mathbf{x} - \mathbf{x}_0)$. That is, the process converges linearly. The computations can be performed with a single/double precision, where the output vector $\mathbf{x}_s = \mathbf{x}_1 + \sum_{i=1}^{s-1} \Delta_i$ is represented by the sequence $\mathbf{x}, \Delta_1, \dots, \Delta_{s-1}$. This is an advantage of process (2.15)–(2.16) versus processes (2.1).

# 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].

Let us recall two such formulas and describe respective accelerations of Newton's RC processes (2.1) for $p = 2$ by following [PBRZ99] (similarly for $p > 2$). 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

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

for three fixed scalars $w$, $a$, and $b$.

By choosing $a = 0$, $b = -1$, and any $w$, we obtain

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

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

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

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 $\mathbf{x} = X\mathbf{t}$.

**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 the 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.*

Now let us modify RC processes (2.1) by expressing 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 $P = (\mathbf{e}_0, \mathbf{t})$:

$$X_{i+1}P = S_iX_iP, \quad S_i = c_{i+1}(I + \sum_{k=0}^{p-1} R_i^k). \tag{3.7}$$

Each step (3.7) requires multiplication of the matrix $R_i$ by the $2p-2$ vectors $R_i^k X_i \mathbf{e}_0$, $R^k X_i \mathbf{t}$ for $k = 0, 1, \dots, p-1$, versus $p$ matrix multiplications per step (2.1). For $p = 2$ we obtain the following extension of process (2.8):

$$X_{i+1}P = c_{i+1}(I + R_i)X_iP, \tag{3.8}$$

with only four matrix-by-vector multiplications per step versus two matrix multiplications per step (2.8).

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 (3.7) for fixed $c_{i+1}$, $i = 0, 1, \dots$. For each $i$, its $i$-th step is reduced to multiplication of five Toeplitz matrices, that is, $M$ and either $Z(\mathbf{x}_i)$, $Z^T(ZJ\mathbf{y}_i)$, $Z(\mathbf{y}_i)$, and $Z^T(ZJ\mathbf{x}_i - \mathbf{e}_0)$ or $Z_f(\mathbf{y}_i)$, $Z_{1/b}(\mathbf{x}_i)$, $Z_f(\mathbf{x}_i - (1-bf)\mathbf{e}_0)$, and $Z_{1/b}(\mathbf{y}_i)$, by a few vectors. The multiplication is performed fast based on FFT, that is, uses $\mathrm{O}(n \log n)$ flops.

**Remark 3.2.** *Expression (3.4) and (3.6) hold for $X = M^{-1}$. Extending them to $X_{i+1} \approx M^{-1}$ is justified less as $X_{i+1}$ deviates from $M^{-1}$. This has two negative impacts:*

*a) the residual norm of the computed approximation to $M^{-1}$ may increase versus $\|R_{i+1}\|$ for $R_{i+1} = I - X_{i+1}M$ defined by $X_{i+1}$ of (2.1),*

*b) the policy (2.9) of choosing the scalars $c_{i+1}$ for $i = 0, 1, \dots$ is not supported by the estimates of [PS91] based on (2.14) anymore.*

*The negative impact a) is most serious at the initial iteration steps where the residual norm can be close to 1 (see (2.6), (2.7)), so the convergence could be easily destroyed. To counter the problem, one may apply RC processes (2.1) with larger $p$ at the initial steps. The negative impact b) cannot occur at the first iteration step, that is, before the first compression step. Thus to avoid the impact b), we replace the $i$-th step for $i > 0$ by the initial step for the inversion of $MX_i$ where $X_i$ is the current approximation to $M^{-1}$. That is, we compute an initial approximation $Y_i$ to $(MX_i)^{-1}$ according to the recipes of Section 2.5 where $M$ is replaced by $MX_i$ and $X_0$ by $Y_i$. Then we improve the approximation $Y_i$ by computing $Y_{i+1} = p_i(Y_iX_iM)Y_i$ for a selected polynomial $p_i(t)$ (see (2.14)), and finally compute the vectors $X_{i+1}\mathbf{e}_0$ and $X_{i+1}\mathbf{t}$ defining $X_{i+1} \approx M^{-1} = (X_iM)^{-1}X_i$ as follows:*

$$X_{i+1}P = p_i(Y_iX_iM)Y_iX_iP, \quad i = 0, 1, \dots, \tag{3.9}$$

*where $P = (\mathbf{e}_0, \mathbf{t})$, $Y_0 = I$, and $X_0$ is selected as in Section 2.5. The coefficients of the polynomials $p_i(t)$ can be chosen to decrease the residual norm $\|I - p_i(Y_iX_iM)Y_iX_iM\|$. We may write $\tilde{X}_0 = Y_iX_i$ and $X_{i+1} = \tilde{X}_k$ and adopt policy (2.9) as follows:*

$$\tilde{X}_1P = c_1(2I - \tilde{X}_0M)\tilde{X}_0P,$$

$$\tilde{X}_2P = c_2(2I - c_1(2I - \tilde{X}_0M)\tilde{X}_0M)\tilde{X}_1P,$$

$$\tilde{X}_3P = c_3(2I - c_2(2I - c_1(2I - \tilde{X}_0M)\tilde{X}_0M)c_1(2I - \tilde{X}_0M)\tilde{X}_0M)\tilde{X}_2P, \quad (3.10)$$

*... and operate with the matrices* $\tilde{X}_0$, $M$, *and* $\tilde{X}_iP$, $P = (\mathbf{e}_0, \mathbf{t})$, $i = 0, 1, \dots, k$ *but not with* $\tilde{X}_1$, $\tilde{X}_2$, $\tilde{X}_3$, .... *We call these iterative schemes* Toeplitz RC processes with delayed compression. *The actual impact of compression on the convergence is hard to estimate theoretically; this impact is frequently positive according to our experiments, so the straightforward RC processes based on (2.1), (2.3), (3.4), (3.6) can be valuable.*

We conclude this section by recalling the estimates of [PBRZ99] for the convergence rate of the *Newton–Toeplitz Iteration* defined by (3.8) for $c_{i+1} = 1$. Let us write $\rho(i) = \|I - X_iT\|_1$, $e(i) = \max(\|\mathbf{x}_i - \mathbf{x}\|_1/\|\mathbf{x}\|_1, \|\mathbf{y}_i - \mathbf{y}\|_1/\|\mathbf{y}\|_1)$. 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

$$\rho(0) \le \theta, \quad e(0)\|T\|_1\mu \le \theta \quad (3.11)$$

for a fixed $\theta$, $0 < \theta < 1$. Then it is proved in [PBRZ99] that $\rho(i) < \theta^{2^i}$, $e(i) < \theta^{2^i-1}e(0)$, $i = 1, 2, \dots$, which shows quadratic convergence under assumptions (3.11). To satisfy (3.11), however, we must have a sufficiently close initial approximation to the inverse matrix $T^{-1}$. The critical parameter $\|T\|_1$ grows roughly proportionally to the product $\|T\|_1\|\mathbf{x}\|_1\|\mathbf{y}\|_1 = \|T\|_1\|T^{-1}\mathbf{t}\|_1\|T^{-1}\mathbf{e}_0\|_1$.

# 4 Residual Correction Processes for Structured Matrices

Extensions of unscaled RC processes (2.8), (2.3) to Toeplitz-like matrices can be found in [P92], [PBRZ99, Section 7.4]. Let us next follow [P01a] to outline these extensions in a unified way–simultaneously to various classes of structured matrices, including 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 a pair of $n \times n$ *operator matrices* $A$ and $B$ we associate linear *displacement operators* $L$, of Sylvester type $L = \nabla_{A,B}$,

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

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

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

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

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

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

The operator matrix pairs $A, B \in \{D(\mathbf{s}), D(\mathbf{t}), Z_e, Z_f^T\}$, for appropriate vectors $\mathbf{s}$ and $\mathbf{t}$ and scalars $e$ and $f$, cover the four cited most popular classes of structured matrices. The most used displacement operators satisfy the following properties:

a) the displacement $L(M)$ is a matrix having 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).

In particular, for the operators $L_+ = \Delta_{Z,Z^T}$ and $L_- = \Delta_{Z^T,Z}$, it was proved in the seminal paper [KKM79] that the matrix equations

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

imply that

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

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 | $m + n$ | $O((m+n)\log^2(m+n))$ |

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

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

for any matrix $M$, and that

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

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 a matrix $M$ by a vector in $O(rn \log n)$ flops.

Similar simple expressions have been obtained in the case of displacement operators $L$ associated with matrices of Hankel, Vandermonde, and Cauchy types [HR84], [BP94], [GO94], [PWa], [P01a], enabling *compressed representations* of an $n \times n$ structured matrix $M$ via $2nr$ entries of the matrices $G$ and $H$ where $r = \operatorname{rank}(L(M))$. 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 as well.

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,

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

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 | $\leq 2$ |
| $Z_1$ | $Z_0^T$ | Hankel and its inverse | $\leq 2$ |
| $Z_0 + Z_0^T$ | $Z_0 + Z_0^T$ | Toeplitz+Hankel | $\leq 4$ |
| $D(\mathbf{t})$ | $Z_0$ | Vandermonde | $\leq 1$ |
| $Z_0$ | $D(\mathbf{t})$ | inverse of Vandermonde | $\leq 1$ |
| $Z_0^T$ | $D(\mathbf{t})$ | transposed Vandermonde | $\leq 1$ |
| $D(\mathbf{s})$ | $D(\mathbf{t})$ | Cauchy | $\leq 1$ |
| $D(\mathbf{t})$ | $D(\mathbf{s})$ | inverse of Cauchy | $\leq 1$ |

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:

$$\text{COMPRESS} \longrightarrow \text{OPERATE} \longrightarrow \text{DECOMPRESS}.$$

At the OPERATE stage, the following simple results can be used [P01a].

**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*

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

*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 multiplication fast. $O(qnr^2 \log^d n)$ flops are sufficient per an RC step (2.1). This step 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's step (2.8). Thus processes (2.1) should be modified as follows where the input matrix $M$ is structured:

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

for $R_i$ of (2.2). 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], [PRW01].

**Approach I. Truncation of the smallest singular values of the displacement.** Compute the SVD of $L(Y_{i+1}) = G_{i+1}H_{i+1}^T$ (cf. [HLPW86]) and truncate the smallest singular values to obtain a displacement matrix $L(X_{i+1})$ having $r_{i+1}$ (non-zero) singular values, where $r_{i+1}$ is fixed according to a selected policy, say $r_{i+1} \le r$ or $r_{i+1} \le cr$ for a fixed constant $c$. (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.** This compression policy extends the policy of Section 3. Compute the displacement $L(X_{i+1})$ based on Theorem 4.3, where $M^{-1}$ is replaced by $X_i$. That is, for $S_{p,i} = S_i$ of (3.7), $\tilde{S}_{p,i} = c_{i+1}\sum_{k=o}^{p-1}\tilde{R}_i^k$, $\tilde{R}_i = I - MX_i$, $Y_{i+1}$ of (4.7), $i = 0, 1, \ldots$, write $L(M) = GH^T$,

$$\begin{aligned} L(X_{i+1}) &= \nabla_{B,A}(X_{i+1}) = G_{i+1}H_{i+1}^T \\ &= -Y_{i+1}\nabla_{A,B}(M)Y_{i+1} = (-Y_{i+1}G)(H^TY_{i+1}) \\ &= -S_{p,i}(X_iG)(H^TX_i)\tilde{S}_{p,i}; \end{aligned} \tag{4.8}$$

also write either

$$\begin{aligned} A^{-1}L(X_{i+1}) &= A^{-1}\Delta_{B,A}(X_{i+1}) = A^{-1}G_{i+1}H_{i+1}A \\ &= A^{-1}Y_{i+1}A^{-1}\Delta_{A,B}(M)Y_{i+1}A = A^{-1}(Y_{i+1}A^{-1}G)(H^TY_{i+1}A) \\ &= A^{-1}S_{p,i}(X_iA^{-1}G)(H^TX_i)\tilde{S}_{p,i}A \end{aligned} \tag{4.9}$$

where the operator matrix $A$ is non-singular or

$$\begin{aligned} L(X_{i+1})B^{-1} &= \Delta_{B,A}(X_{i+1})B^{-1} = BG_{i+1}H_{i+1}^TB^{-1} \\ &= BY_{i+1}\Delta_{A,B}(M)B^{-1}Y_{i+1}B^{-1} = (BY_{i+1}G)(H^TB^{-1}Y_{i+1})B^{-1} \\ &= BS_{p,i}(X_iG)(H^TB^{-1}X_i)\tilde{S}_{p,i}B^{-1} \end{aligned} \tag{4.10}$$

where the operator matrix $B$ is non-singular. The computation of the displacement of $X_{i+1}$ in (4.8)–(4.10) essentially amounts to post-multiplying $S_{p,i}$ by $X_iG$ or $X_iA^{-1}G$ and either post-multiplying $H^TX_i$ by $\tilde{S}_{p,i}$ or $\tilde{S}_{p,i}A$ or pre-multiplying $\tilde{S}_{p,i}B^{-1}$ by $X_i$ and the product by $H^TB^{-1}$. In each case, we multiply each of the matrices $R_i$ and $R_i^T$ by $O(pr)$ vectors. RC process (4.8)–(4.10) can be applied to a Toeplitz matrix $M$. Then we would multiply each of the matrices $X_i$, $M$, $M^T$, and $X_i^T$ by two vectors for every $i$, whereas the RC process (3.7), (3.8) only requires multiplication of each of $M$ and $X_i$ by a pair of vectors for every $i$. Furthermore, since we assume that $X_i \approx Y_i$, we may replace $-X_iG$ by $G_i = -Y_iG$ and $H^TX_i$ by $H_i^T = H^TY_i$ in (4.8) and thus to replace (4.8) by a simpler expression

$$L(X_{i+1}) = G_{i+1}H_{i+1}^T = S_{p,i}G_iH_i^T\tilde{S}_{p,i} \tag{4.11}$$

Similarly, we may simplify (4.9) and (4.10) to write

$$A^{-1}L(X_{i+1}) = A^{-1}G_{i+1}H_{i+1}^T,$$

$$A^{-1}G_{i+1} = A^{-1}Y_{i+1}A^{-1}G = A^{-1}S_{p,i}G_i,$$

$$H_{i+1}^T = H^TY_{i+1} = H_i^T\tilde{S}_{p,i}$$

where $A$ in nonsingular and to write

$$L(X_{i+1})B^{-1} = G_{i+1}H_{i+1}^TB^{-1},$$

$$G_{i+1} = Y_{i+1}G = S_{p,i}G_i,$$
$$H_{i+1}^T B^{-1} = H^T B^{-1} Y_{i+1} B^{-1} = H_i^T \tilde{S}_{p,i} B^{-1}$$

where $B$ is nonsingular.

Approach I relies on the observation that

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

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 matrix approximation based on the truncation of the singular values [GL96]. Thus we bound the norms $||L(X_{i+1}) - L(M^{-1})||$ and $||X_{i+1} - M^{-1}|| \le ||L^{-1}||\ ||L(X_{i+1}) - L(M^{-1})||$ 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], [PRW01], and [P01a].

Algorithm 7.4.1 of [PBRZ99] applies Approach I to Toeplitz-like matrices $M$ and uses the displacements $L_+(M)$ and $L_-(X_i)$ and expressions (4.4), (4.5) to compute the displacements $L_-(X_{i+1}) = L_-(X(Y_{i+1}))$. It is proved in [PBRZ99] that in this case

$$||X_{i+1} - M^{-1}||_2 \le (1 + 2(r_i - r)n)||X_i - M^{-1}||_2 \tag{4.12}$$

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

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

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

$$G_{i+1} = X_i(2I - MX_i)G, \quad H_{i+1}^T = H^T X_i(2I - MX_i)$$

and by its last column

$$X_{i+1}\mathbf{e}_{n-1} = (2I - MX_i)X_i\mathbf{e}_{n-1}.$$

In [PRW01] both Approaches I and II have been elaborated upon 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 in Section 2 do not apply to processes (4.7) because of the compression of the displacements $L(Y_i)$, but the comments and recipes of Remark 3.2 can be extended to both Approaches I and II (see Remark 4.2).

The following theorems from [PRW01] (extending their preliminary versions of [P92], [PZHD97], [PBRZ99], and [PR01]) state the estimates for the error norms of the computed approximations under assumption (2.3). The error

norms grow proportionally to 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 [PRW01] and [PWb] (see also [P01]).

**Theorem 4.4.** *[PRW01]. Let the unscaled Newton-Structured process (2.8), (2.3) be applied to a non-singular matrix $M$. Let all its steps be performed with compression according to (4.7) and Approach I such that all 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*

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

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

$$\theta = ||I - X_0 M||_2 \eta,$$

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

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

**Theorem 4.5.** *[PRW01]. Let the unscaled Newton-Structured process (2.3), (2.8) be applied to invert a non-singular matrix $M$. Let (4.7) and Approach II be used for the compression of the displacements $L(Y_i)$, $i = 1, 2, \dots$. Write*

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

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

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

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

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

*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,*

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

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

The cited Algorithm 7.4.2 of [PBRZ99] can be modified according to (4.11) as follows:

$$\mathbf{x}_j = X_j \mathbf{e}_j,$$
$$\nabla_{Z_f^{-1}, Z_f^{-1}}(X_j) = G_j H_j^T, \quad j = 0, 1, \dots ;$$
$$G_{i+1} = (2I - MX_i)G_i,$$
$$H_{i+1}^T = H_i^T(2I - MX_i),$$
$$\mathbf{x}_{i+1} = (2I - MX_i)\mathbf{x}_i, \quad i = 0, 1, \dots,$$

saving multiplication of $r$ vectors by $X_i$ and $X_i$ by $r$ vectors in each iteration step $i$.

**Remark 4.1.** *Newton-Structured Iteration with compression was first studied for Toeplitz-like matrices in [P92]. In [PR01], [PRW01] the algorithms were extended to various other classes of structured matrices in a unified way, adopted in this section. In an alternative* displacement transformation approach *due to [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; furthermore, sample displacement 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. So far, the most acclaimed application of the displacement 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].*

**Remark 4.2.** *Applying both Approaches I and II we may try to improve convergence by allowing more work per iteration step and using processes with the matrices $S_{p,i}$ and $\tilde{S}_{p,i}$ of (4.8)–(4.12) for larger $p_i$, replacing matrices $S_{p,i}$ and $\tilde{S}_{p,i}$ by $p_i(MX_i)$ for selected polynomials $p_i(y)$ (compare(2.14)) or generalizing the approaches of (3.9) and (3.10). Approach I also allows us to vary the level of compression by truncating more or fewer singular values of $L(X_i)$. Generalizing the nomenclature in Remark 3.2, we call the respective modifications the* Structured RC processes with delayed compression.

# 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.

**Algorithm 5.1. A homotopic RC process for a positive definite matrix.**

**Input:** *an $n \times n$ Hermitian positive definite matrix, a non-negative $\varepsilon$, a positive $\lambda_1^+$ such that* $\text{spectrum}(M) = \{\lambda_1, \dots, \lambda_n\}$, *where*

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

*and a black box RC process (2.1) (scaled or unscaled and with any selected levels of compression in the case of a structured input).*

**Initialization:** *Fix some values $\theta_h$, $0 < \theta_h < 1$, $h = 0, 1, \dots$, and write (cf. (1.2), (1.3))*

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

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

*Apply the selected black box RC process (2.1) for $M = M_0$ and $X_0$ of (5.2) to approximate $M^{-1}$ by $\tilde{X}_0$ such that*

$$\|R(\tilde{X}_0, M_0^{-1})\| \leq \varepsilon. \tag{5.4}$$

**Computations:** *Stage $h$, $h = 0, 1, \dots$. Compute an upper bound $\eta_h$ on the norm*

$$||M_h^{-1}||_2 = \frac{1}{t_h + \lambda_n} \tag{5.5}$$

*(see Remark 5.1). Compute*

$$\Delta_h = \theta_h/\eta_h. \tag{5.6}$$

*Apply the black box RC process with $X_0 = \tilde{X}_h$, replacing $M$ by $M_{h+1}$ if $t_{h+1} > 0$. If $t_{h+1} \geq 0$, write $H = h + 1$, use $M$ instead of $M_{h+1}$, and compute a matrix $\tilde{X}_{h+1}$ such that*

$$\|R(M_{h+1}, \tilde{X}_{h+1})\| \leq \varepsilon. \tag{5.7}$$

**Output:** *$\tilde{X}_H$ approximating $M^{-1}$ and stop.*

The algorithm is completely defined as soon as we fix an RC process (2.1) (including its stopping criterion and, for structured matrices $M$, the policy of the compression of the displacements), although we may modify the algorithm to allow this policy and the residual norm bound $\varepsilon$ vary with $h$.

Let us show correctness of the algorithm for $\varepsilon = 0$. First observe that for the residual $R_0$ of (2.2), we have

$$R_0 = R(M_0, t_0^{-1} I) = I - t_0^{-1} M_0 = -t_0^{-1} M,\ r_0 = \|R(M_0, t_0^{-1} I)\|_2 \leq \theta_0. \tag{5.8}$$

Further, deduce from (5.3) that

$$R(M_{h+1}, M_h^{-1}) = \Delta_h M_h^{-1},$$
$$r_{h+1} = \|R(M_{h+1}, M_h^{-1})\|_2 = \Delta_h \|M_h^{-1}\|_2 = \Delta_h/(t_h + \lambda_n). \tag{5.9}$$

Write

$$\lambda_{n,h} = 1/\eta_h - t_h \le 1/||M_h^{-1}||_2 - t_h = \lambda_n \tag{5.10}$$

and observe that the value $\lambda_n^- = \lambda_{n,h}$ satisfies bound (5.1). Finally, (5.6) implies that

$$r_{h+1} \le \theta_h \text{ for all } h. \tag{5.11}$$

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 $\theta_h$ 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 $\eta_h$ on the matrix norm (5.5) can be obtained by applying the power or Lanczos methods [GL96]. If an estimate $\eta_h$ is sufficiently sharp for a fixed $h = k$ (say for $h = 1$), a close upper bounds $\eta_{k+i}$ can be computed based on the following simple expression:*

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

*(see (5.3)–(5.11)).*

**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*

$$\kappa(M(t)) \le \kappa(M) \text{ for } t \le 0. \tag{5.12}$$

*The same inequality can be easily verified for the modification of the homotopic process in Section 8 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*

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

*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).*

# 6 The Number of Homotopic Steps

To simplify our subsequent analysis, we next assume that the values $\lambda_{n,h}$ and $\theta_h$ are invariant in $h$, that is, $\lambda_{n,h} = \lambda_n^-$ for all $h \geq 1$ (cf. (5.10) and Remark 5.1). Then by virtue of (5.3), (5.5), (5.6), and (5.10), we have

$$t_{h+1} + \lambda_n^- = (1 - \theta_h)(t_h + \lambda_n^-), \quad h = 0, 1, \dots, H-1.$$

Therefore,

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

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

Furthermore, let $\theta_h$ be 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

At each homotopic step, the number of RC steps depends on the bound $\theta$ 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 $c_{i+1}$, $i = 0, 1, \dots$. 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 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 $\nu_i$ 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 are sufficient for the transition from $1/e$ to $\nu$ 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 $\epsilon$). Bound (7.2) covers the $(H-1)\beta$ refinement RC steps particular to the HRC processes. Practically, $\beta$ is quite small. For instance, for $q = 4$, the bound $e^{-16}$ is achieved in two steps. The specific choice of the bound $\nu$ can be guided by the following simple estimate.

**Proposition 7.1.** *Let*

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

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

*for any fixed matrix norm. Then*

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

**Proof.**

$$\begin{aligned} \|I - XM_h\| &\leq \nu + \|XM_{h-1} + XM_h\| \\ &\leq \nu + \|XM_{h-1}\| \, \|I - M_{h-1}^{-1} M_h\| \\ &\leq \nu + (1+\nu)\theta_h. \end{aligned}$$

□

### 7.3 The number of RC steps at the critical stages

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

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

provided that $\theta$ 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:

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

### 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 Inversion of Indefinite 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. This standard symmetrization, however, has the well-known price of squaring the condition number and, consequently, of a substantial slow-down of the HRC algorithm (cf. (7.7)). Let us show a simple remedy in the case where $M$ is a non-singular Hermitian (or a real symmetric) 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

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

where

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

Let $\lambda^-$ and $\lambda^+$ be two fixed positive values such that

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

for every eigenvalue $\lambda$ of $M$. Then for any fixed sequence of real $\theta_h,\ 0 < \theta_h < 1,\ h = 0, 1, \dots$, we define an HRC process by (5.2)–(5.7), for $\eta_h$ 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 HRC algorithm (which can be applied to any Hermitian input matrix $M$) is now 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

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

(replacing (5.3)). (8.3)–(8.5) immediately imply bounds (5.8) and (5.11) for $\eta_h \ge ||M_h^{-1}||_2$ and $\Delta_h$ of (5.6).

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

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

replaces (5.5). 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

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

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

$$H - 1 = \left\lceil \frac{\log(\lambda^+/(\theta^2\lambda^-))}{\log(1/(1-\theta))} \right\rceil.$$

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 $\theta \ge 1/2$. 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 $\epsilon, \nu$, and $\theta$). Therefore, up to replacing $\lambda_n^-$ by $\lambda^-$ and $\lambda_1^+$ by $\lambda^+$ 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. The latter bound $a$ is relatively small, and we 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 $\gamma$ of (7.1), $H$ of (6.1), (6.3), and $\kappa_+(M)$ equal to either $\lambda_1^+/\lambda_n^-$ or $\lambda^+/\lambda^-$.) 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 Remarks 3.2 and 4.2.) How close should these approximations be?

(3.11) and Theorems 4.4 and 4.5 show the level of approximation starting at which rapid convergence is guaranteed for the Newton–Toeplitz Iteration (3.8) and for the unscaled Newton-Structured RC process (2.8), (2.3), 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. (2.6) and (2.7)) fall short of even approaching this level. If we start with an initial approximation obtained according to the recipes of Section 2, then to ensure 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 are frequently effective, however. That is, according to the experiments reported in our Section 12, [P01a], and [BM,a], 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. Furthermore, in a large portion of test runs of non-homotopic processes (2.3), (4.7) for $p = 2$ and Toeplitz input matrices under Approach I, the residual 2-norm grew above 1 and sometimes well above 1 in the initial RC steps, and then iteration still converged. Since every RC step (2.3), (4.3) for $p = 2$ squares the 2-norm of the residual, the only explanation of this phenomenon is that the compression frequently decreases the residual norm, thus bringing the approximation *closer* to the inverse, so that the estimates of (3.11) and Theorems 4.4 and 4.5 are overly pessimistic.

HRC processes with compression is an alternative approach supported both experimentally (see Section 12) 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 $\theta_h$ defining the step sizes $\Delta_h$ 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 $M_0 = \hat{M}$ or $M_0 = \hat{M}\sqrt{-1}$ for any Hermitian and positive definite matrix $\hat{M}$.

**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*

$$M_h = M + t_h M_0, \quad M_0 = \left(\frac{1}{z_i + z_k^*}\right)_{i,k=1}^{n}$$

*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 $\mu_1^+$ and $\mu_n^-$ 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 $\lambda_1^+$ of (5.1) and choose

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

so that $||I-(t_1M_0)^{-1}M_1||_2 \le \theta_0$. Invert $M_1$ by applying processes (2.1) for $X_0 = t_1M_0$.

Now deduce from (10.2) that

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

Substitute the bound

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

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

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

for all $h$ and for $\lambda_n^-$ 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 (5.11). Now, invert the matrices $M_{h+1}$ by applying processes (2.1) for $X_0 = M_h^{-1}$ and for $h = 1, 2, \dots, H-2$, 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 as well.

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

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

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

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

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 $\hat{M}\sqrt{-1}$ for a fixed positive definite matrix $\hat{M}$, 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 this extended homotopic process. Assume that bounds (10.1) still hold where $\{\mu_1 \dots, \mu_n\} = \text{spectrum}(M)$ and each eigenvalue $\lambda$ of the input matrix $M$ satisfies the bounds

$$0 < \lambda^- \le |\lambda| \le \lambda^+ \tag{10.8}$$

for two fixed positive values $\lambda^-$ and $\lambda^+$. Now write

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

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

Let us deduce bounds (5.11). Recall the following well-known theorem [Par80, proof of Theorem 15-3-3].

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

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

*for a unitary matrix $U$, $U^*U = UU^* = I_n$, and a diagonal matrix*

$$\Sigma = \text{diag}(\sigma_i)_{i=1}^n, \ \mu_1^+ \ge \sigma_1^2 \ge \sigma_2^2 \ge \dots \ge \sigma_n^2 \ge \mu_n^- > 0.$$

*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*

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

*for $h = 1, 2, \dots$ where $\kappa^+ = \mu_1^+/\mu_n^-$.*

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

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

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

Therefore,

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

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

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

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

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

□

Relations (10.1), (10.2), (10.4), (10.9), and Corollary 10.1 together immediately imply (5.11). 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 $\lambda_n^-$ and $\lambda^-/\kappa^+$ are dominated by the term $t_h\mu_n^-$. If the term $\lambda^-/\kappa^+$ 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 the unscaled RC processes (2.1), (2.3) and the scaled processes (2.8), (2.9) converge to the Moore–Penrose generalized inverse $M^+$ where the input matrix $M$ is singular. Now recall that the scaled RC process (2.8), (2.9), (2.12)–(2.13) converges to the numerical generalized inverse matrix $M_\epsilon^+$. The analysis and the estimates of our paper (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 $\epsilon$. This enables various refinements from noisy perturbations of the input. Furthermore, the

computation of $M_\epsilon^+$ does not depend on whether the matrix $M$ is singular or not. In particular, we may apply HRC processes to compute $M_\epsilon^+$ for a positive $\epsilon$ where $M$ is singular. If $\epsilon$ 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 iteration (2.1), (2.3) or (2.8), (2.9) as a Basic Subroutine.

For the extension of the RC and HRC methods to the computation of the numerical generalized inverse $M_\epsilon^+$ (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 $M_\epsilon^+$ even for $\epsilon = 0$. For Toeplitz and Hankel matrices and for $\epsilon = 0$, the problem can be avoided [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, $r_\epsilon = \mathrm{rank}(M_\epsilon^+)$.

**Theorem 11.1.** *For any positive $\epsilon$ 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* $r_\epsilon = \mathrm{rank}(M_\epsilon^+)$.

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

# 12 Numerical experiments with Toeplitz matrices

The presented Newton-Structured Iteration algorithms (that is, structured RC processes for $p = 2$) were tested numerically for $n \times n$ Toeplitz input matrices $M$. The compression was achieved by means of the truncation of the singular values (according to Approach I) and was implemented as Algorithm 7.5.1 from [PBRZ99]. The tests were performed by M. Kunin at the Graduate Center of CUNY, in cooperation with R. Rosholt of the Lehman College of CUNY.

The tests used the following computational facilities:

- OS – Red Hat Linux 7.0
- compiler – GCC 2.96 (also using bundled random number generator)
- library – CLAPACK 3.0 (routines for computing SVD of real and complex matrices and eigenvalues of symmetric matrices)

Both non-homotopic and stiff homotopic versions of Newton-Structured Iteration were applied to the same input matrices $M$. In non-homotopic processes the compression level, that is, the number $l$ of untruncated singular values of the displacements, was chosen adaptively to minimize $l$ as long as convergence was achieved. More precisely, for a fixed threshold value $\epsilon$, the candidate compression level $l$ was calculated as follows. Let $\sigma_1^{(i)}, \dots, \sigma_n^{(i)}$ denote the singular

values of $X_i$ written in the non-increasing order. Then we chose $l = l(i)$ satisfying $\sigma_{l+1}^{(i)} < \epsilon\sigma_1^{(i)}$, $\sigma_l^{(i)} \geq \epsilon\sigma_1^{(i)}$. If Newton's process diverged for these $\epsilon$ and $l$, then all results of the computation (obtained by this moment) were discarded, except that the contribution to the overall work of the iteration was counted. Then the iteration process was repeated with the compression level $l/2$. In the case of divergence, this value was recursively halved further until convergence but at most 10 times. The experiments for homotopic processes were limited to recursive optimization of the tolerance values $\theta_h$ and consequently the homotopic step sizes, under a stiff 0.7-down policy. According to this policy, the initial value of $\theta$ was always set to 0.7 and never increased. In the case of divergence, $\theta$ was recursively halved, and the process was repeated until convergence. For a fixed $\theta$, the step sizes were calculated using LAPACK for computing the eigenvalues of the matrix $M$. The value $l$ of the compression level was fixed and remained invariant in all Newton's steps throughout the entire homotopic process. This value was chosen experimentally when convergence with this value was observed for one or two test runs. This was our simplified preliminary policy, subject to improvement in our future experiments.

The computations stopped at the final homotopic and non-homotopic steps where the residual norm decreased to the single precision 0; at all other homotopic steps, the computations stopped where the residual norm decreased below $10^{-6}$. (The latter bound was a little smaller than was necessary for convergence.)

The algorithm was tested for $n \times n$ Toeplitz matrices $M$ of the following classes (see details below).

1. Real symmetric tridiagonal Toeplitz matrices $(t_{i,j})_{i,j=0}^{n-1}$, $t_{i,j} = 0$ where $|i-j| > 1$, $t_{i+1,i} = t_{i,i+1} = 1$, $t_{i,i}$ equals 4 or $-2$.
2. The matrices $\left(\frac{1}{1+|i-j|}\right)_{i,j=0}^{n-1}$.
3. Randomly generated Toeplitz matrices.
4. Randomly generated real symmetric positive definite matrices with a specified condition number.
5. Randomly generated real symmetric indefinite Toeplitz matrices.

The tests results were differentiated further according to the condition number of the matrix $M$ and the size $n \times n$, for $n$ ranging from 50 to 350.

An $n \times n$ random real symmetric Toeplitz matrix of class 5 was defined by generating the $n$ random entries of its first row; for an unsymmetric Toeplitz matrix of class 3, also the $n-1$ remaining entries of its first column were generated. The random entries were generated as the random normally distributed variables with the mean 0 and the standard deviation 1. At this stage, a random number generator was applied with the **rand()** function from the standard C library that comes with the GCC compiler for Cygwin on Windows 2000. The condition number was most frequently quite small for random matrices, and

then the algorithms converged very rapidly. To make the results more meaningful, part of the experiments was restricted to the matrices with larger condition numbers. To form a matrix of class 4, that is, to achieve positive definiteness and a desired condition number, we computed the two extremal eigenvalues of a random real symmetric Toeplitz matrix and then added the matrix $aI$ for an appropriate positive $a$.

For unsymmetric and symmetric indefinite Toeplitz matrices $M$, the non-homotopic Newton's process was initialized with the matrices

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

For a symmetric positive definite matrix $M$, the same process was applied with $X_0 = I/\|M\|_F$. For the homotopic processes with the same symmetric input matrices $M$, the same Newton's processes were applied with the invariant truncation level $l = 2$, except that the initial matrices $X_0$ were determined by the homotopic rules and the choice of the matrix $M_0$. For unsymmetric input matrices $M$, both homotopic and non-homotopic processes also were applied to two symmetrized matrices $M^TM$ and $\left(\begin{smallmatrix} 0 & M \\ M^T & 0 \end{smallmatrix}\right)$ (cf. (8.1) and (8.2)). For homotopic processes, in these two cases, the invariant truncation levels $l = 12$ and $l = 6$ were selected, respectively. In the symmetric indefinite case, the initial choice (8.3) was used for the matrix $M_0$. In the positive definite case, the matrix $M_0$ was selected according to (5.2). The initial threshold bound $\epsilon$ in non-homotopic Newton's processes was selected at the levels 0.025 for the unsymmetric and symmetric indefinite matrices, 0.00005 for the symmetrized matrices $\left(\begin{smallmatrix} 0 & M \\ M^T & 0 \end{smallmatrix}\right)$