2011

TR-2011011: Randomized and Derandomized Matrix Computations II

Victor Y. Pan

Guoliang Qian

Ai-Long Zheng

Randomized and Derandomized Matrix Computations II *

Victor Y. Pan[1,2],[a], Guoliang Qian[2],[b], and Ai-Long Zheng[2],[c]

[1] Department of Mathematics and Computer Science
Lehman College of the City University of New York
Bronx, NY 10468 USA
[2] Ph.D. Programs in Mathematics and Computer Science
The Graduate Center of the City University of New York
New York, NY 10036 USA
[a] victor.pan@lehman.cuny.edu
http://comet.lehman.cuny.edu/vpan/
[b] gqian@gc.cuny.edu
[c] azheng-1999@yahoo.com

Abstract

We propose new techniques and algorithms for approximation by low-rank matrices and by structured matrices, numerical stabilization of Gaussian elimination with no pivoting, and preconditioning and block diagonalization of an ill conditioned matrix having a small positive numerical nullity or rank. Our technical advances include estimates for the condition numbers of random Toeplitz matrices, dual Sherman–Morrison–Woodbury formula, novel techniques of randomized preprocessing, a proof of their preconditioning power, and application to preconditioning general and structured matrices. Our extensive tests support the results of our analysis and show effectiveness of the proposed algorithms.

Key Words: Low-rank matrices, Randomized preconditioning, Derandomization, Numerical rank

1 Introduction

1.1 Derandomized approximation by low-rank matrices

Approximation of $m \times n$ matrices A having small numerical ranks $q \ll \min\{m, n\}$ by low-rank matrices and approximate matrix decompositions based on such approximation are a thriving research area with numerous important applications and extensions [HMT11]. To the wealth of such extensions listed in [HMT11] we can add approximation of matrices with small numerical displacement ranks by matrices with displacement structure (see Section 7.4) and tensor computations [T00], [MMD08], [OT09].

Low-rank approximation is actually needed just for $\mathcal{R}(A)$, that is the range of A. The most popular algorithms approximate it based on multiplication of an $m \times n$ input matrix A by $q + p$ random vectors; with a high probability it is sufficient to choose $p \ll \min\{m, n\}$ [HMT11].

*Supported by NSF Grant CCF-1116736 and PSC CUNY Awards 62230–0040 and 63153–0041. Some results of this paper have been presented at the ACM-SIGSAM International Symposium on Symbolic and Algebraic Computation (ISSAC '2011), San Jose, CA, 2011, the 3nd International Conference on Matrix Methods in Mathematics and Applications (MMMA 2011) in Moscow, Russia, June 22-25, 2011, and the 7th International Congress on Industrial and Applied Mathematics (ICIAM 2011), in Vancouver, British Columbia, Canada, July 18-22, 2011

Our distinct recipe is the approximation of the range $\mathcal{R}(A)$ by $\mathcal{R}(C_-U_-)$ where

$$C_- = A - AU_-H^{-1}V_-^TA,\ H = I_q + V_-^TAU_-, \tag{1.1}$$

U_- and V_- are scaled random matrices of sizes $m \times q$ and $n \times q$, respectively, and M^T denotes the transpose of a matrix M. Furthermore we derandomize the computations (see Section 7.5 and the very end of Section 6). Our techniques supporting these algorithms also have some other important applications, e.g., to approximation of the leading and trailing singular spaces of a matrix. Combination of our algorithms with the ones of [GTZ97], [GT01], [GOS08] and [HMT11] may lead to further progress.

1.2 Numerically safe Gaussian elimination with no pivoting

Hereafter $\sigma_j(A)$ denotes the jth largest singular value of a matrix A of a rank ρ for $j = 1, \dots, \rho$; $\kappa(A) = \sigma_1(A)/\sigma_\rho(A)$ denotes its condition number. Recall that if $\kappa(A)$ is large (in context), then the matrix A is ill conditioned, that is lies near a rank deficient matrix; its numerical inversion with rounding to the IEEE standard single or double precision as well as the solution of a linear system with such a matrix can be easily corrupted. Unless $\kappa(A)$ is large, the matrix A is well conditioned, and if it also has full rank, then it can be safely treated numerically with single or double precision as long as no ill conditioned auxiliary matrices are involved.

Pivoting, that is row or column interchange is applied to avoid dealing with such auxiliary matrices. *Gaussian elimination with no pivoting* (hereafter we refer to it as *GENP*) can easily fail in numerical computations with rounding errors, except for some special classes of input matrices such as diagonally dominant and positive definite matrices. For these matrix classes, GENP and its pivoting-free variations outperform Gaussian elimination with pivoting [GL96, page 119]. We dramatically expand these classes by proving in Corollary 4.3 that pre- as well as post-multiplication of a well conditioned coefficient matrix by a Gaussian random square matrix is expected to support safe numerical performance of GENP as well as block Gaussian elimination. Here and hereafter "expected" and "likely" mean "with a probability close to one", and quite typically "with probability one", due to our derandomization techniques.

Our formal study (cf. Section 3.3 and Remark 4.1) supports such results already in the case of Gaussian random circulant multipliers, and in our tests (cf. Table 9.6) we consistently observed the same results even where we filled these multipliers with ± 1 for random choice of the n signs $\pm$.

1.3 Randomized preconditioning and its applications

Can we extend the above advance by applying randomized multipliers M and N to yield a much better conditioned matrix product MAN? No, because random square matrices M and N are expected to be nonsingular and well conditioned [D88], [E88], [ES05], [CD05], [SST06] and to satisfy the bound $\kappa(MAN) \geq \kappa(A)/(\kappa(M)\kappa(N))$. Approximate inverses $X \approx A^{-1}$ are popular multipliers, but their computation is noncostly only for some special although important classes of matrices.

It is customary to call the map $A \Longrightarrow B$ *preconditioning* wherever it simplifies the solution of a nonsingular linear system $A\mathbf{y} = \mathbf{b}$ of n equations, e.g., where B is the product MA, AN or MAN and the matrix $I - B$ has a small rank or a small numerical rank for I denoting the identity matrix. Indeed in this case most popular iterative algorithms such as CG and GMRES converge to the solution of a linear system $B\mathbf{x} = \mathbf{f}$ very fast even if the condition number $\kappa(B)$ is not small. Here and hereafter we use the acronym "CG" for "Conjugate Gradient".

Additive preprocessing $A \Longrightarrow C = A + B$ produces $C = I$ for $B = I - A$, but this observation is not easy to utilize for solving a linear system $A\mathbf{y} = \mathbf{b}$. Assume, however, that an $n \times n$ nonsingular input matrix A has a numerical nullity at most r, that is has at most r singular values that are much smaller than the norm $||A||$. For a small r these matrices make up a large and important subclass in the class of ill conditioned matrices (cf. [CDG03] and our Remarks 2.1 and 5.3).

We scale such a matrix A to yield $||A||_2 \approx 1$, choose $n \times r$ standard Gaussian random matrices U and V, write $C = A + UV^T$, and prove (cf. Corollary 5.1 and Remark 5.1) that with a probability close to one $\kappa(C)$ has order $\sigma_1(A)/\sigma_{n-r}(A)$ and thus is not large, whereas $\kappa(A) = \sigma_1(A)/\sigma_n(A)$ is

large. Typically a well conditioned matrix C can be more readily inverted than an ill conditioned matrix A, whereas the matrices $I - AC^{-1} = UV^TC^{-1}$ and $I - C^{-1}A = C^{-1}UV^T$ have ranks at most r, and so C^{-1} can be used as a multiplicative preconditioner for A.

If the matrix A is singular and has a nullity at most r, then the same map $A \to C = A+UV^T$ for $n \times r$ random matrices U and V under both Gaussian and uniform probability distribution produces a nonsingular matrix C with probability one.

The matrix C_-^{-1} of (1.1) for scaled random $n \times q$ matrices U_- and V_- is also likely to be a preconditioner for a nonsingular matrix A having a small numerical rank q. This is because

$$C_-^{-1} = A^{-1} + U_-V_-^T \tag{1.2}$$

(and so the matrices $I - C_-^{-1}A = -U_-V_-^TA$ and $I - AC_-^{-1} = -AU_-V_-^T$ have ranks at most q) and because the matrix C_- is expected to be well conditioned and thus typically more readily invertible than the ill conditioned matrix A.

Given a nonsingular $n \times n$ matrix A that has numerical rank q and numerical nullity $r = n - q$, we applied our additive and dual additive preprocessing to compute 2×2 block diagonalization of this matrix with diagonal blocks of sizes $r \times r$ and $q \times q$, both expected to be better conditioned than the matrix A (see Sections 7.6 and 7.7).

In yet another application we compute the solution $\mathbf{y} = A^{-1}\mathbf{b}$ to a linear system $A\mathbf{y} = \mathbf{b}$ by expressing A^{-1} via C^{-1} by means of the Sherman–Morrison–Woodbury formula [GL96, page 50], [S98, Corollary 4.3.2], hereafter referred to as the *SMW formula*. In the case of an ill conditioned matrix A having small numerical nullity, small numerical rank or structure of Toeplitz type the resulting algorithms accelerate customary solutions of a linear system $A\mathbf{y} = \mathbf{b}$ by order of magnitude and nearly reach optimality (see Section 7.8).

We extend all our results to the case where we employ scaled Gaussian random Toeplitz matrices U, V, U_- and V_- (each defined by $O(n)$ random parameters). This enables us to keep Toeplitz structure of an input matrix, but we can preserve matrix structure and sparseness most perfectly by applying randomized augmentation, e.g., defined by the map $A \Longrightarrow K = \begin{pmatrix} A & -U \\ V^T & W \end{pmatrix}$, where we can choose random matrices U, V and W with fixed patterns of structure and sparseness. Augmentation is closely linked to additive preprocessing, has similar preconditioning power, and allows similar applications.

1.4 Our progress: a brief summary

Besides our novel approximation by low-rank matrices and by structured matrices and our randomized support of GENP and block Gaussian elimination, we extend the study of conditioning of random matrices in [D88], [E88], [ES05], [CD05], [SST06] to the case of random structured matrices (thus answering the challenge of [SST06]) and to randomized preconditioning of matrices having small numerical nullity, small numerical rank or structure of Toeplitz type. Further applications to various fundamental matrix computations can be developed based on the techniques in [PGMQ], [PIMR10], [PQ10], [PQa], [PQZa], [PQZC], and [PZ11].

1.5 The test results

The results of our extensive tests (the contribution of the second and the third authors) are in good accordance with our theoretical estimates. In particular we match output accuracy of the customary algorithms but outperform them in terms of the CPU time in the case of Toeplitz inputs (see Table 9.10). Some results of our experiments may be of independent interest, e.g., in our tests (see Tables 9.1 and 9.3 and Section 3.3) random Toeplitz matrices tended to be reasonably well conditioned, unlike some important classes of Toeplitz matrices in [BG05].

1.6 Organization of the paper and selective reading

We devote the next section to the definitions and basic results on matrix computations and Section 3 to estimates for the condition numbers of Gaussian random general, Toeplitz and circulant matrices.

In Section 4 we estimate the condition numbers of randomized matrix products; the results support GENP with randomized circulant multipliers. In Section 5 we prove that our randomized additive preprocessing is expected to transform an ill conditioned matrix into a well conditioned matrix provided the input matrix has a small positive numerical nullity. In Section 6 we recall the SMW formula and define its dual version and dual additive preprocessing, which enables randomized preconditioning of matrices that have a small numerical rank. In Section 7 we estimate numerical rank and nullity, apply our results to approximation by low-rank matrices and by structured matrices, approximate trailing and leading singular spaces, and cover randomized additive and multiplicative preconditioning for linear systems of equations and derandomization techniques. In Section 8 we first discuss randomized augmentation and randomized structured preprocessing and then solve ill conditioned Toeplitz linear systems based on these techniques. In Section 9 we cover numerical tests, which are the contribution of the second and the third authors. In Section 10 we briefly recall Newton's iteration for the inversion of structured matrices, refine it with our preprocessing and discuss its heuristic acceleration. In Section 11 we comment on the related works, our technical contributions and some directions for further study. In the Appendix we recall some auxiliary results on randomized regularization of matrices.

In Sections 6–10 we devise algorithms whose effectiveness is implied by Corollary 5.1 and Remark 5.1, but otherwise these sections can be read independently of the analytical Sections 3–5.

2 Some definitions and basic results on matrix computations

We use and extend the customary definitions of matrix computations (cf. [GL96], [S98]).

2.1 Some basic definitions

A flop stands for an arithmetic operation.

Throughout we assume computations in the field $\mathbb{R}$ of real numbers and in Section 3.5 comment on the extension to the field $\mathbb{C}$ of complex numbers.

A^T is the transpose of a matrix A.

We write $A^{-T} = (A^{-1})^T = (A^T)^{-1}$.

A real matrix A is Hermitian (or symmetric) if $A = A^T$ and is Hermitian (or symmetric) positive definite if $A = B^T B$ for a nonsingular matrix B.

$(B_1 \mid \ldots \mid B_k) = (B_j)_{j=1}^k$ is a $1 \times k$ block matrix with blocks $B_1, \ldots, B_k$.

$\mathrm{diag}(B_1, \ldots, B_k) = \mathrm{diag}(B_j)_{j=1}^k$ is a $k \times k$ block diagonal matrix with diagonal blocks $B_1, \ldots, B_k$.

I_n denotes the $n \times n$ identity matrix $(\mathbf{e}_j)_{j=1}^n = (\mathbf{e}_1 \mid \ldots \mid \mathbf{e}_n)$.

J_n denotes the $n \times n$ reflection matrix $(\mathbf{e}_j)_{j=n}^1 = (\mathbf{e}_n \mid \ldots \mid \mathbf{e}_1)$.

$O_{k,l}$ denotes the $k \times l$ matrix filled with zeros. $\mathbf{0}_k$ denotes the vector $O_{k,1}$.

We drop the subscripts and write I, J, O, and $\mathbf{0}$ where the size of a matrix or a vector is not important or is defined by context.

2.2 Range, null space, rank, nullity, and nmbs

$\mathcal{R}(A)$ denotes the range of an $m \times n$ matrix A, that is the linear space $\{\mathbf{z} : \mathbf{z} = A\mathbf{x}\}$ generated by its columns, $\mathcal{N}(A)$ its null space $\{\mathbf{v} : A\mathbf{v} = \mathbf{0}\}$, $\mathrm{rank}(A) = \dim \mathcal{R}(A)$ its rank, and $\mathrm{nul}\, A = \dim \mathcal{N}(A) = n - \rho$ its nullity. $\mathbf{v}$ is its null vector if $A\mathbf{v} = \mathbf{0}$.

Fact 2.1. *The set $\mathbb{M}$ of $m \times n$ matrices of rank ρ is an algebraic variety of dimension $(m+n-\rho)\rho$.*

Proof. Let M be an $m \times n$ matrix of a rank ρ with a nonsingular $\rho \times \rho$ leading block M_{00} and write $M = \begin{pmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{pmatrix}$. Then the $(m-\rho) \times (n-\rho)$ Schur complement $M_{11} - M_{10}M_{00}^{-1}M_{01}$ must vanish, which imposes $(m-\rho)(n-\rho)$ algebraic equations on the entries of M. Similar argument can be applied where any $\rho \times \rho$ submatrix of the matrix M is nonsingular. Therefore $\dim \mathbb{M} = mn - (m-\rho)(n-\rho) = (m+n-\rho)\rho$. □

Suppose a matrix B has full column rank and $\mathcal{R}(B) = \mathcal{N}(A)$. Then we call B a *null matrix basis* or a *nmb* for a matrix A and write $B = \mathrm{nmb}(A)$.

The nullity, the null space, null vectors, and nmbs of the transposed matrix A^T are said to be the left nullity, the left null space, left null vectors, and left nmbs of a matrix A, respectively.

$A^{(k)}$ denotes the $k \times k$ leading, that is northwestern block submatrix of a matrix A.

A matrix of a rank ρ has *generic rank profile* if all its $i \times i$ leading blocks are nonsingular for $i = 1, \dots, \rho$. If such a matrix is itself nonsingular, then it is called *strongly nonsingular*.

2.3 Norms, orthogonalization, SVD and inverses

$||A||_h$ is the h-norm of a matrix $A = (a_{i,j})_{i,j=1}^{m,n}$ for $h = 1, 2, \infty$. We write $||A|| = ||A||_2$ and recall from [GL96, Section 2.3.2 and Corollary 2.3.2] that

$$\max_{i,j=1}^{m,n} |a_{i,j}| \le ||A|| = ||A^T|| \le \sqrt{mn}\, \max_{i,j=1}^{m,n} |a_{i,j}|, \tag{2.1}$$

$$||A||^2 \le ||A||_1 ||A||_\infty. \tag{2.2}$$

$\mathbf{v}$ is a *unit* or *normalized* vector if $||\mathbf{v}|| = \sqrt{\mathbf{v}^T\mathbf{v}} = 1$.

An $m \times n$ matrix U (for $m \le n$) is *unitary* or *orthonormal* if $U^T U = I$.

Fact 2.2. *[GL96, Theorem 5.2.2]. QR factorization $A = QR$ of a matrix A having full column rank into the product of a unitary matrix $Q = Q(A)$ and an upper triangular matrix $R = R(A)$ is unique if the factor R is a square matrix with positive diagonal entries.*

$A = S_A \Sigma_A T_A^T$ is an *SVD* or *full SVD* of an $m \times n$ matrix A of a rank ρ provided $S_A S_A^T = S_A^T S_A = I_m$, $T_A T_A^T = T_A^T T_A = I_n$, $\Sigma_A = \mathrm{diag}(\widehat{\Sigma}_A, O_{m-\rho,n-\rho})$, $\widehat{\Sigma}_A = \mathrm{diag}(\sigma_j(A))_{j=1}^{\rho}$, $\sigma_j = \sigma_j(A) = \sigma_j(A^T)$ is the jth largest singular value of a matrix A. These values have the minimax characterization

$$\sigma_j = \max_{\dim(\mathbb{S})=j} \min_{\mathbf{x}\in\mathbb{S},\ ||\mathbf{x}||=1} ||A\mathbf{x}||,\ j = 1, \dots, \rho, \tag{2.3}$$

where $\mathbb{S}$ denotes linear spaces [GL96, Theorem 8.6.1], and so $\sigma_j = 0$ for $j > \rho$, $\sigma_j(A)$ is the distance from the matrix A to the nearest matrix of a rank $j-1$ for $j = 1, \dots, \rho+1$, $\sigma_1 = \max_{||\mathbf{x}||=1} ||A\mathbf{x}|| = ||A||$, and if $m \ge n$, then $\sigma_n = \min_{||\mathbf{x}||=1} ||A\mathbf{x}||$.

The minimax characterization (2.3) also implies

Fact 2.3. *If A_0 is a $p \times q$ submatrix of a matrix A, then $\sigma_j(A) \ge \sigma_j(A_0)$ for all j.*

If $\sigma_q > \sigma_{q+1}$, in which case $q \le \rho$, then the first q columns of the matrices S_A and T_A generate the leading left and right singular spaces $\mathbb{S}_A^{(q)} = \mathcal{R}(S_A(I_q \mid O_{q,m-q})^T)$ and $\mathbb{T}_A^{(q)} = \mathcal{R}(T_A(I_q \mid O_{q,n-q})^T)$, respectively, associated with the q largest singular values of the matrix A. The orthogonal complements $\mathbb{S}_{A,m-q}$ and $\mathbb{T}_{A,n-q}$ of these singular spaces are the left and right trailing singular spaces, respectively.

A matrix $X = A^{(I)}$ is a left inverse of a matrix A if $XA = I$ (in this case $m \ge n$); a matrix $Y = A^{(I)}$ is its right inverse if $AY = I$ (in this case $m \le n$); $A^{(I)} = A^{-1}$ for a nonsingular matrix A.

$\Sigma_A^+ = \mathrm{diag}((\widehat{\Sigma}_A)^{-1}, O_{n-\rho,m-\rho})$ and $A^+ = T_A \Sigma_A^+ S_A^T$ are the Moore–Penrose pseudo-inverses of the matrices Σ_A and A, respectively. A^+ is a left inverse of a matrix A of full rank for $m \ge n$ and its right inverse for $m \le n$; $A^+ = A^{-1}$ for a nonsingular matrix A. $||A^+|| = 1/\sigma_\rho(A)$ for a matrix A of a rank ρ. We write A^{+T} for $(A^+)^T = (A^T)^+$.

Theorem 2.1. *[GL96, Section 5.5.5]. Assume two matrices $C \in \mathbb{C}^{m\times n}$ and $\tilde{C} \in \mathbb{C}^{m\times n}$ having full rank and write $E = \tilde{C} - C$. Then $||\tilde{C}^+ - C^+|| \le 2\sqrt{n}||E|| \max\{||\tilde{C}^+||^2, ||C^+||^2\}$*

2.4 Condition number, numerical rank and nullity

Hereafter the concepts "large", "small", "near", "closely approximate", "ill conditioned" and "well conditioned" are quantified in the context.

For two positive parameters a and b we write $a \gg b$ and $b \ll a$ if the ratio a/b is large and write $a \approx b$ if the ratio is close to one or if $b = 0$ and $|a|$ is small. For two matrices A and B we write $A \approx B$ if $||A - B|| \ll ||A||$.

Remark 2.1. *Unlike the nullity and the rank, numerical nullity and numerical rank are not well defined for a large class of ill conditioned matrices, in particular for all matrices A having nested clusters of small singular values but also for the matrix class represented by a 1000×1000 matrix A with singular values $\sigma_j(A) = 2^{1000-j}$, $j = 1, 2, \dots, 1000$, e.g., by* $\mathrm{diag}(2^{1000-j})_{j=1}^{1000}$.

$\kappa(A) = \frac{\sigma_1(A)}{\sigma_\rho(A)} = ||A||\ ||A^+||$ is the condition number of a matrix A of a rank ρ. Such a matrix is *ill conditioned* if $\sigma_1(A) \gg \sigma_\rho(A)$ and is *well conditioned* otherwise. See [D83], [GL96, Sections 2.3.2, 2.3.3, 3.5.4, 12.5], [H02, Chapter 15], and [S98, Section 5.3] on the estimation of norms and condition numbers. $\kappa(A) = ||A||\ ||A^{-1}||$ for a nonsingular matrix A.

An $m \times n$ matrix A has *numerical rank* ρ and *numerical nullity* $n - \rho$ if the ratios $\sigma_j(A)/||A||$ are small for $j > \rho$ but are not small for $j \le \rho$.

If an $m \times n$ well conditioned matrix A has a rank $\rho < l = \min\{m, n\}$, then all its sufficiently close neighbours $\tilde{A}$ have numerical rank ρ, even though almost all of them have rank l. The minimax characterization implies that conversely, every $m \times n$ matrix $\tilde{A}$ having a positive numerical rank $\rho < l$ is close to the well conditioned matrix A of rank ρ obtained by setting to zero all but ρ largest singular values of $\tilde{A}$. The range $\mathcal{R}(A)$ of the matrix A approximates the leading singular space $\mathbb{T}_{\tilde{A}}^{(\rho)}$ associated with the ρ largest singular values of $\tilde{A}$; the null space $\mathcal{N}(A)$ approximates the trailing singular space $\mathbb{T}_{\tilde{A},n-\rho}$ (the orthogonal complement of $\mathbb{T}_{\tilde{A}}^{(\rho)}$).

The map $M \Longrightarrow M^T$ transforms singular spaces of M and its numerical nullity into the respective left singular spaces of M^T and its left numerical nullity.

A map $A \Longrightarrow B$ is called *preconditioning* if $\kappa(B) \ll \kappa(A)$ and if the solution of a linear system $A\mathbf{y} = \mathbf{b}$ is readily reduced to the solution of a linear system $B\mathbf{x} = \mathbf{f}$, e.g., if $B = AM$ for a readily computable matrix M. Preconditioning of the coefficient matrix accelerates convergence of CG, GMRES and other popular iterative algorithms to the solution of a linear system $B\mathbf{x} = \mathbf{f}$, but the convergence is particularly fast where the matrix $I - B$ has a small rank or a small numerical rank, even if the condition number $\kappa(B)$ is large [A94], [B02], [G97]. This prompts us to call a square matrix M a left (resp. right) *r-preconditioner* for a matrix A if the matrix $MA - I$ (resp. $AM - I$) has a rank or numerical rank at most r. A r-preconditioner is of interest if it is easier to compute it than the 0-preconditioner $M = A^{-1}$ and if r is small enough.

By extending the definitions in Section 2.2 we say that a matrix having a numerical rank ρ has *generic conditioning profile* if its $i \times i$ leading blocks are well conditioned for $i = 1, \dots, \rho$. If such a matrix is well conditioned itself, then we call it *strongly well conditioned*.

One can readily verify the following property (see [PQZa]).

Theorem 2.2. *Suppose Gaussian elimination with no pivoting has been applied to a matrix A of a rank (resp. numerical rank) ρ to compute LU factorizations of the leading block submatrices $A^{(j)}$ for $j = 1, \dots, \rho$. Then the computations involve no divisions by zeros (resp. by values that are absolutely small relative to the norm $||A||$) if and only if the matrix A has generic rank (resp. generic conditioning) profile.*

Similar property holds for block Gaussian elimination (see [PQZa]).

These results motivate the search for multipliers that map a well conditioned matrix of full rank into a matrix having generic conditioning profile. We call such a map *generic preconditioning* and obtain it in Corollary 4.3.

2.5 Toeplitz and circulant matrices

$m \times n$ Toeplitz matrix $T = (t_{i-j})_{i,j=1}^{m,n}$ is defined by its first row and column, that is by the vector $(t_h)_{h=1-n,2-n,\dots,m-1}$ of dimension $m+n-1$.

An $n \times n$ lower triangular Toeplitz matrix $T = (t_{i-j})_{i,j=1}^{n}$ where $t_k = 0$ for $k < 0$ is defined by its first column $T\mathbf{e}_1$; hereafter $Z(\mathbf{v})$ denotes such a matrix with the first column $\mathbf{v} = Z(\mathbf{v})\mathbf{e}_1$. $Z = Z(\mathbf{e}_2)$ is the $n \times n$ downshift matrix whose all entries are zeros except for the first subdiagonal filled with ones; $Z\mathbf{v} = (v_i)_{i=0}^{n-1}$ for a vector $\mathbf{v} = (v_i)_{i=1}^{n}$ and $v_0 = 0$; furthermore $Z^n = O$, $Z(\mathbf{v}) = \sum_{i=0}^{n-1} v_{i-1}Z^i$.

Observe that $||Z(\mathbf{v})||_1 = ||Z(\mathbf{v})||_\infty = ||\mathbf{v}||_1$. By combining these equations with bound (2.2) obtain that

$$||Z(\mathbf{v})|| \le ||\mathbf{v}||_1. \tag{2.4}$$

Suppose $X = (x_{ij})_{i,j=1}^{n}$ is the inverse of a nonsingular $n \times n$ Toeplitz matrix and $x_{11} \neq 0$. Then Gohberg and Sementsul's celebrated formula in [GS72] expresses the matrix X through its two columns $\mathbf{x}_1 = X\mathbf{e}_1$ and $\mathbf{x}_n = X\mathbf{e}_n$ as follows,

$$x_{11}X = Z(\mathbf{x}_1)Z^T(J\mathbf{x}_n) - Z(Z\mathbf{x}_n)Z^T(ZJ\mathbf{x}_1). \tag{2.5}$$

$Z_f = Z + f\mathbf{e}_n^T\mathbf{e}_1$ is the unit f-circulant matrix. An *f-circulant matrix* $Z_f(\mathbf{v}) = \sum_{i=0}^{n-1} v_i Z_f^i = (z_{i-j \mod n})_{i,j=1}^{n}$ is an $n \times n$ Toeplitz matrix defined by its first column vector $\mathbf{v} = (v_i)_{i=0}^{n-1}$ and a scalar $f \neq 0$.

f-circulant matrix is called *circulant* if $f = 1$ and *skew circulant* if $f = -1$. By replacing f by zero we arrive at a lower triangular Toeplitz matrix $Z(\mathbf{v})$.

Theorem 2.3. *(See [CPW74].) We have $Z_1(\mathbf{v}) = \Omega^{-1}D(\Omega\mathbf{v})\Omega$. More generally, for any $f \neq 0$, we have $Z_{f^n}(\mathbf{v}) = U_f^{-1}D(U_f\mathbf{v})U_f$ where $U_f = \Omega D(\mathbf{f})$, $\mathbf{f} = (f^i)_{i=0}^{n-1}$, $D(\mathbf{u}) = \mathrm{diag}(u_i)_{i=0}^{n-1}$ for a vector $\mathbf{u} = (u_i)_{i=0}^{n-1}$, and $\Omega = (\omega_n^{ij})_{i,j=0}^{n-1}$ is the $n \times n$ matrix of the discrete Fourier transform at n points, $\omega_n = \exp(2\pi\sqrt{-1}/n)$ being a primitive n-th root of one.*

The theorem implies that products and inverses of f-circulant matrices (wherever defined) are f-circulant and can be computed in $O(n \log n)$ flops for $n \times n$ inputs.

3 Conditioning of random general, Toeplitz and circulant matrices

3.1 Random variables and random matrices

Definition 3.1. *$F_X(y) = Probability\{X \le y\}$ for a real random variable X is the* cumulative distribution function (CDF) *of X evaluated at y. $F_{g(\mu,\sigma)}(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} \exp(-\frac{(x-\mu)^2}{2\sigma^2})dx$ for a Gaussian random variable $g(\mu,\sigma)$ with a mean μ and a variance σ^2, so that*

$$\mu - 4\sigma \le y \le \mu + 4\sigma \text{ with a probability near one.} \tag{3.1}$$

Definition 3.2. *$F_M(y) = F_{\sigma_l(M)}(y)$ for an $m \times n$ matrix M and an integer $l = \min\{m,n\}$.*

Definition 3.3. *A matrix or a vector is a* Gaussian random matrix or vector *with a mean μ and a variance σ^2 if it is filled with independent Gaussian random variables, all having the same mean μ and variance σ^2. $\mathcal{G}_{\mu,\sigma}^{m\times n}$ denotes the set of $m \times n$ Gaussian random matrices. For $\mu = 0$ and $\sigma^2 = 1$ they are* standard Gaussian random matrices.

Definition 3.4. *(Cf. Subsection 2.5). We write $T \in \mathcal{T}_{\mu,\sigma}^{m\times n}$ and call an $m \times n$ Toeplitz matrix $T = (t_{i-j})_{i,j=1}^{m,n}$ a* Gaussian random Toeplitz matrix *with a mean μ and a variance σ^2 if $\mathbf{t} = (t_h)_{h=1-n,2-n,\dots,m-1} \in \mathcal{G}_{\mu,\sigma}^{(m+n-1)\times 1}$. We write $Z_f(\mathbf{v}) \in \mathcal{Z}_{f,\mu,\sigma}^{n\times n}$ and call $Z_f(\mathbf{v})$ a* Gaussian random f-circulant matrix *with a mean μ and a variance σ^2 if $\mathbf{v} \in \mathcal{G}_{\mu,\sigma}^{n\times 1}$.*

Definition 3.5. *$\chi_{\mu,\sigma,n}(y)$ is the CDF of the random function $(\sum_{i=1}^{n} X_i^2)^{1/2} = ||(Y_i)_{i=1}^{n}||$ where $(X_i)_{i=1}^{n} \in \mathcal{G}_{\mu,\sigma}^{n\times 1}$. For $y \ge 0$ we have $\chi_{0,1,n}(y) = \frac{2}{2^{n/2}\Gamma(n/2)} \int_{-\infty}^{y} x^{n-1}\exp(-x^2/2)dx$ where $\Gamma(h) = \int_0^\infty x^{h-1}\exp(-x)dx$, $\Gamma(n+1) = n!$ for nonnegative integers n.*

3.2 Conditioning of Gaussian random matrices

Gaussian random matrices in Definition 3.3 tend to be well conditioned [D88], [E88], [ES05], [CD05]. Moreover the sum $M+W$ for $M \in \mathbb{R}^{m\times n}$ and $W \in \mathcal{G}_{\mu,\sigma}^{m\times n}$ is expected to be well conditioned unless the ratio $\sigma/||M||$ is small [SST06].

The following upper bound on the probability that for a Gaussian random matrix W the smallest singular value of a matrix $A = W + M$ is less than a scalar y can also be viewed as a probabilistic lower bound on the smallest singular value of the matrix A.

Theorem 3.1. *Suppose $M \in \mathbb{R}^{m\times n}$, $W \in \mathcal{G}_{\mu,\sigma}^{m\times n}$, $A = W + M$, $l = \min\{m,n\}$, and $y \geq 0$. Then $F_A(y) \leq 2.35\, y\sqrt{l}/\sigma$.*

Proof. Clearly the matrix $A = M + W$ has full rank with probability one because W is a Gaussian random matrix (see the Appendix). In view of Fact 2.3 it is sufficient to prove the claimed bound on $F_A(y)$ in the case where $m = n$, and in this case the theorem turns into [SST06, Theorem 3.3]. □

The two following theorems supply lower bounds on the probabilities that $||W|| \leq y$ and $\kappa(A) \leq y$ for a scalar y, $A = W + M$, and a Gaussian random matrix W. They can also be viewed as probabilistic upper bounds on the norm $||W||$ and the condition number $\kappa(A)$.

Theorem 3.2. *(See [DS01, Theorem II.7]). Suppose $W \in \mathcal{G}_{0,\sigma}^{m\times n}$, $l = \min\{m,n\}$ and $y \geq 2\sigma\sqrt{l}$. Then $F_{||W||}(y) \geq 1 - \exp(-(y - 2\sigma\sqrt{l})^2/(2\sigma^2))$.*

Theorem 3.3. *(See [SST06, Theorem 3.1]). Suppose $M \in \mathbb{R}^{m\times n}$, $W \in \mathcal{G}_{0,\sigma}^{m\times n}$, $A = W + M$, $l = \min\{m,n\}$, $||M|| \leq \sqrt{l}$, $\sigma \leq 1$, $y \geq 1$. Then $F_{\kappa(A)}(y) \geq 1 - (14.1 + 4.7\sqrt{(2\ln y)/n})n/(y\sigma)$.*

This bound has been improved by a factor $\sqrt{\log n}$ in [W04] and in the case of $M = O$ by a factor $y^{|m-n|}\sqrt{\ln y}$ in [ES05] and [CD05].

Theorem 3.3 is deduced in [SST06] based on combining Theorems 3.1 and 3.2. The proof of Theorem 3.1 relies on the two lemmas below that we use in the next subsections.

Lemma 3.1. *Suppose y is a positive number, $\mathbf{w} \in \mathbb{R}^{n\times 1}$ is any fixed real unit vector, $||\mathbf{w}|| = 1$, $M \in \mathbb{R}^{n\times n}$ is a fixed real matrix, $W \in \mathcal{G}_{\mu,\sigma}^{n\times n}$, and $A = W + M$. Let Q be a unitary matrix such that $Q\mathbf{w} = \mathbf{e}_1$ and let $B = QA = (\mathbf{b}_1 \mid \ldots \mid \mathbf{b}_n)$. Then*

$$\text{Probability}\{||A^{-1}\mathbf{w}|| > y\} \leq \max_{\mathbf{b}_1,\ldots,\mathbf{b}_n} \text{Probability}\{|\mathbf{t}^T\mathbf{b}_1| < 1/y\}$$

where $||\mathbf{t}|| = 1$ and $\mathbf{t}^T\mathbf{b}_i = 0$ for $i = 2,\ldots,n$.

Proof. See [SST06, the proof of Lemma 3.2]. □

Lemma 3.2. *[SST06, Lemma A.2]. For a nonnegative y, a fixed unit real vector $\mathbf{t}$ of a dimension n, and a vector $\mathbf{b} \in \mathcal{G}_{\mu,\sigma}^{n\times 1}$ we have $F_{|\mathbf{t}^T\mathbf{b}|}(y) = \text{Probability}\{|\mathbf{t}^T\mathbf{b}| \leq y\} \leq \sqrt{\frac{2}{\pi}}\frac{y}{\sigma}$.*

Remark 3.1. *The latter bound is independent of μ and n, and so it holds even if all coordinates of the vector $\mathbf{b}$ are fixed, except for a single coordinate in $\mathcal{G}_{\mu,\sigma}$.*

3.3 Conditioning of Gaussian random Toeplitz matrices

A matrix $T \in \mathcal{T}_{\mu,\sigma}^{n\times n}$ is the sum of two triangular Toeplitz matrices, and so (2.4) implies that

$$F_{||T||}(y) \geq \chi_{\mu,\sigma,n}(y/2). \tag{3.2}$$

Next we estimate the norm $||T^{-1}||$.

Theorem 3.4. *Let $T = (T_{i-j})_{i,j=1}^n \in \mathcal{T}_{\mu,\sigma}^{n\times n}$, $T_{11} = (T_{i-j})_{i,j=2}^n$ and $y \geq 0$. Then (a) with probability one the matrix T is nonsingular and the entry $x_{11} = \det T_{11}/\det T$ of its inverse $X = (x_{ij})_{i,j=1}^n$ does not vanish; (b) furthermore $||x_{11}X|| \leq 2RS$ for two random variables R and S such that $F_R(y) \leq \frac{y}{\sigma}\sqrt{\frac{2n}{\pi}}$ and $F_S(y) \leq \frac{y}{\sigma}\sqrt{\frac{2n}{\pi}}$.*

Proof. Part (a) is easy to deduce (see the Appendix). To prove part (b) we first readily extend Lemma 3.1 (cf. [SST06, Lemma 3.2]) as follows.

Lemma 3.3. *Suppose y is a positive number, $T \in \mathcal{T}_{\mu,\sigma}^{n\times n}$, j is an integer, $1 \le j \le n$, $\bar{\mathbf{x}}_j \in \mathbb{R}^{n\times 1}$ is the unit vector orthogonal to all vectors $T\mathbf{e}_i$ for $i \ne j$, $M \in \mathbb{R}^{n\times n}$ is a fixed real matrix, and $A = W + M$. Then*

$$\text{Probability}\{||T^{-1}\mathbf{e}_j|| > 1/y\} \le \text{Probability}\{|\bar{\mathbf{x}}_j^T T\mathbf{e}_j| < y\}.$$

Now observe that each of the column vectors $T\mathbf{e}_1$ and $T\mathbf{e}_n$ has an entry not shared with the other entries of T, recall Remark 3.1 and deduce that

$$\text{Probability}\{|\bar{\mathbf{x}}_j^T T\mathbf{e}_j| < y\} \le \frac{y}{\sigma}\sqrt{\frac{2}{\pi}}$$

for $j = 1$ and $j = n$. It remains to combine these estimates with (2.4) and (2.5). □

Let us extend Theorem 3.4 to estimate the norm $||X||$, equal to $||x_{11}X||/|x_{11}| = ||x_{11}X||\ |s|$. Here $s = t_0 - t_0^{-1}((t_k)_{k=1}^{n-1})^T(t_{-k})_{k=1}^{n-1}$ is the Schur complement of $T_{11} = (t_{i-j})_{i,j=2}^{n}$ in $T = (t_{i-j})_{i,j=1}^{n}$.

Since $T \in \mathcal{T}_{\mu,\sigma}^{n\times n}$ we obtain that $|s| \le (1 + (n-1)g)g/|t_0|$ where $g = |g(\mu,\sigma)|$ for $g(\mu,\sigma)$ in Definition 3.1 and that $\text{Probability}\{|t_0| < y\} \le \frac{y}{\sigma}\sqrt{\frac{2}{\pi}}$ by virtue of Lemma 3.2.

The equations $||X|| = ||x_{11}X||\ |s|$, the latter upper bounds on $|s|$ and $1/|t_0|$, and Theorem 3.4 together imply a probabilistic upper bound on the norm $||X||$ in terms of g. Combining it with (3.1) and (3.2) implies that for $T \in \mathcal{T}_{\mu,\sigma}^{n\times n}$ the condition number $\kappa(T)$ does not tend to grow very fast as $n \to \infty$ (cf. Table 9.3). We also recall that $\kappa(T)$ is invariant in scaling the matrix T, which enables some control over the values μ and σ.

Remark 3.2. *For another direction to estimating the factor $1/|x_{11}| = |\det T|/|\det T_{11}|$ from above, recall that $|\det T|$ is the volume $V_{n,\mu,\sigma}$ defined by the matrix $T = T_n$. For matrices $T_k \in \mathcal{T}_{\mu,\sigma}^{k\times k}$ we can expect that this volume grows more or less uniformly as n grows, and so it is informative to observe that Hadamard's upper bound $V_{n,\mu,\sigma} \le V_{n,\mu,\sigma}^{(+)} = n^{n/2}g^n$ increases by a factor $(n-1)^{1/2}g$ versus $V_{n-1,\mu,\sigma}^{(+)}$. Indeed $V_{n,\mu,\sigma}^{(+)}/V_{n-1,\mu,\sigma}^{(+)} \le n^{n/2}g^n/((n-1)^{(n-1)/2}g^{n-1}) \le g\exp(\frac{n}{2n-2})(n-1)^{1/2}$ because $n^{n/2} = (1+\frac{1}{n-1})^{n/2}(n-1)^{n/2} \le \exp(\frac{n}{2n-2})(n-1)^{n/2}$.*

3.4 Conditioning of Gaussian random circulant matrices

Next we readily estimate the norms of a random Gaussian f-circulant matrix and its inverse.

Theorem 3.5. *Assume an $n \times n$ circulant matrix $A = Z_1(\mathbf{v})$ for $\mathbf{v} \in \mathcal{G}_{\mu,\sigma}^{n\times 1}$. Then a) $F_{||A||}(y) \ge \chi_{\mu,\sigma,n}(y/2)$ and b) $F_{\sigma_n(A)}(y) \le \frac{yn}{\sigma}\sqrt{\frac{2}{\pi}}$ for all nonnegative y and $\chi_{\mu,\sigma,n}(y)$ in Definition 3.5.*

Proof. Part a) of the theorem follows from (3.2) because A is a Toeplitz matrix.

To prove part b), represent the matrix A as in Theorem 2.3 and write $B = \Omega A\Omega^{-1} = D(\Omega\mathbf{v})$, $\mathbf{u} = (u_i)_{i=0}^{n-1} = \Omega\mathbf{v}$. We have $\sigma_j(A) = \sigma_j(B)$ for all j because $\frac{1}{\sqrt{n}}\Omega$ and $\sqrt{n}\Omega^{-1}$ are unitary matrices.

Combine the equations $u_i = \mathbf{e}_i^T\Omega\mathbf{v}$, the bounds $||\Re(\mathbf{e}_i^T\Omega)|| \ge 1$ for all i, and Lemma 3.2 and deduce that $F_{|\Re(u_i)|}(y) \le \frac{y}{\sigma}\sqrt{\frac{2}{\pi}}$ for any i, $i = 1,\dots,n$. We have $F_{\sigma_n(B)}(y) = F_{\min_i|u_i|}(y)$ because $B = \text{diag}(u_i)_{i=0}^{n-1}$. Clearly $|u_i| \ge |\Re(u_i)|$, and part b) of the theorem follows. □

Remark 3.3. *Our extensive experiments suggest that the estimates of Theorem 3.5 are overly pessimistic (cf. Table 9.4).*

Combining Theorem 2.3 with minimax characterization implies that

$$\frac{1}{g(f)}\sigma_j(Z_1(\mathbf{v})) \le \sigma_j(Z_f(\mathbf{v})) \le g(f)\sigma_j(Z_1(\mathbf{v}))$$

for all vectors $\mathbf{v}$, scalars $f \neq 0$, $g(f) = \max\{1, |f|\}\ \max\{1, \frac{1}{|f|}\}$, and $j = 1, 2, \dots, n$. This enables us to extend the estimates of Theorem 3.5 to f-circulant matrices for $f \neq 0$. In particular these estimates do not change in the case of skew circulant matrices (for which $f = -1$) and show that f-circulant matrices of size $n \times n$ tend to be well conditioned unless $f \approx 0$ or $1/f \approx 0$.

3.5 Extension to the case of complex inputs

For simplicity we assume real random matrices and vectors throughout this paper, but most of our study can be readily extended to the computations in the field $\mathbb{C}$ of complex numbers. To extend our study in Sections 6–10, we usually just need to replace transposes A^T by Hermitian transposes A^H. Below are some basic results for the extension of our probabilistic estimates for norms and singular values to the case of complex matrices.

Definition 3.6. *The set $\mathcal{G}^{m\times n}_{\mathbb{C},\mu,\sigma}$ of $m \times n$ complex Gaussian random matrices with a mean μ and a variance σ is the set $\{A + B\sqrt{-1}\}$ for $(A \mid B) \in \mathcal{G}^{m\times 2n}_{\mu,\sigma}$ (cf. Definition 3.3).*

We can immediately extend Theorem 3.2 to the latter matrices. Let us extend Lemma 3.2.

Lemma 3.4. *The bound of Lemma 3.2 also holds provided $\mathbf{t} = \mathbf{r} + \mathbf{q}\sqrt{-1}$ is a fixed complex unit vector and $\mathbf{b} = \mathbf{f} + \mathbf{g}\sqrt{-1} \in \mathcal{G}^{n\times 1}_{\mathbb{C},\mu,\sigma}$ is a complex vector such that $\mathbf{f}$, $\mathbf{g}$, $\mathbf{r}$ and $\mathbf{q}$ are real vectors, $||\mathbf{t}|| = 1$, and vectors $\mathbf{f}$ and $\mathbf{g}$ are in $\mathcal{G}^{n\times 1}_{\mu,\sigma}$.*

Proof. We have $\mathbf{t}^H\mathbf{b} = \mathbf{r}^T\mathbf{f} - \mathbf{q}^T\mathbf{g} + (\mathbf{r}^T\mathbf{g} + \mathbf{q}^T\mathbf{f})\sqrt{-1}$, and so $|\mathbf{t}^H\mathbf{b}|^2 = |\mathbf{r}^T\mathbf{f} - \mathbf{q}^T\mathbf{g}|^2 + |\mathbf{r}^T\mathbf{g} + \mathbf{q}^T\mathbf{f}|^2$. Hence $|\mathbf{t}^H\mathbf{b}| \geq |\mathbf{r}^T\mathbf{g} + \mathbf{q}^T\mathbf{f}| = |\mathbf{u}^T\mathbf{v}|$ where $\mathbf{u}^T = (\mathbf{r}^T \mid \mathbf{q}^T)$ and $\mathbf{v}^T = (\mathbf{g}^T \mid \mathbf{f}^T)$, and so $\mathbf{v} \in \mathcal{G}^{1\times 2n}_{\mu,\sigma}$ and $||\mathbf{u}|| = ||\mathbf{t}|| = 1$. It remains to apply Lemma 3.2 to real vectors $\mathbf{u}$ and $\mathbf{v}$ replacing $\mathbf{b}$ and $\mathbf{t}$. □

By combining Lemmas 3.1 and 3.4 we extend [SST06, Lemma 3.2] to the case of complex inputs.

Corollary 3.1. *Suppose y is a positive number, $\mathbf{w} \in \mathbb{C}^{n\times 1}$ is any fixed complex unit vector, $||\mathbf{w}|| = 1$, $M \in \mathbb{C}^{n\times n}$ is a fixed matrix, $W \in \mathcal{G}^{n\times n}_{\mathbb{C},\mu,\sigma}$, and $A = W + M$. Then*

$$\text{Probability}\{||A^{-1}\mathbf{w}|| > y\} \leq \frac{2}{y\sigma\sqrt{\pi}} .$$

Proof. Lemmas 3.1 and 3.4 together imply the corollary for a unit vector $\mathbf{w}$ in $\mathbb{R}^{n\times 1}$ and for the upper bound $\frac{\sqrt{2}}{y\sigma\sqrt{\pi}}$. By allowing its increase by a factor $\sqrt{2}$ we extend the bound to any unit complex vector $\mathbf{w}$ because $\max\{||\mathbf{u}||, ||\mathbf{v}||\} \geq 1/\sqrt{2}$ provided $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{n\times 1}$, $\mathbf{w} = \mathbf{u} + \mathbf{v}\sqrt{-1}$, $\mathbf{u} = \Re(\mathbf{w})$, $\mathbf{v} = \Im(\mathbf{w})$, and $||\mathbf{w}||^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 = 1$. □

Corollary 3.2. *Suppose y is a positive number, $M \in \mathbb{C}^{n\times n}$ is a fixed matrix, $W \in \mathcal{G}^{n\times n}_{\mathbb{C},\mu,\sigma}$, and $A = W + M$. Then*

$$\text{Probability}\{||A^{-1}|| > y\} \leq \frac{2n}{y\sigma\sqrt{\pi}} .$$

Proof. Recall that $||B|| = \max_{j=1}^{n} ||B\mathbf{e}_j||$ for any $n \times n$ matrix B, in particular for $B = A^{-1}$. It remains to apply Corollary 3.1 to the vectors $\mathbf{w} = \mathbf{e}_j$ for $j = 1, \dots, n$ and to deduce that

$$\text{Probability}\{\max_{j=1,\dots,n} ||A^{-1}\mathbf{e}_j|| > y\} \leq \frac{2n}{y\sigma\sqrt{\pi}} .$$

□

Based on the latter result, one can readily extend Theorems 3.1, 3.4 and 3.5 to the case of complex inputs.

Remark 3.4. *Corollary 3.2 extends [SST06, Theorem 3.3] to the case of complex matrices but increases the bound of the theorem by a factor $\sqrt{2n}$. This estimated increase must be overly pessimistic because random complex matrices tend to be a little better conditioned than random real matrices (see [E88], [ES05], [CD05] and our Table 9.2).*

4 Conditioning of randomized matrix products and generic preconditioning

Next we extend the estimates of Theorem 3.1 to yield probabilistic lower bounds on the smallest singular values of the products of fixed and random matrices. We begin with three lemmas. The first two of them easily follow from minimax characterization (2.3).

Lemma 4.1. *Suppose* $A = \mathrm{diag}(\sigma_i)_{i=1}^n$, $G \in \mathbb{R}^{r\times n}$, $H \in \mathbb{R}^{n\times r}$, $\mathrm{rank}(A) = n$, $\mathrm{rank}(G) = r(G)$, *and* $\mathrm{rank}(H) = r(H)$. *Then* $\mathrm{rank}(GA) = r(G)$, $\mathrm{rank}(AH) = r(H)$, $\sigma_j(GA) \ge \sigma_j(G)\sigma_n$, $\sigma_j(AH) \ge \sigma_j(H)\sigma_n$ *for all* j.

Lemma 4.2. $\sigma_j(GA) = \sigma_j(AH) = \sigma_j(A)$ *for all* j *if* G *and* H *are square unitary matrices.*

Lemma 4.3. *[SST06, Proposition 2.2]. Suppose* $W \in \mathcal{G}_{\mu,\sigma}^{m\times n}$, $SS^T = S^TS = I_m$, $TT^T = T^TT = I_n$. *Then* $SW \in \mathcal{G}_{\mu,\sigma}^{m\times n}$ *and* $WT \in \mathcal{G}_{\mu,\sigma}^{m\times n}$.

Theorem 4.1. *Suppose* $G \in \mathbb{R}^{r(G)\times m}$, $H \in \mathbb{R}^{n\times r(H)}$, $\mathrm{rank}(G) = r(G)$, $\mathrm{rank}(H) = r(H)$, $y \ge 0$, $M \in \mathbb{R}^{m\times n}$, $W \in \mathcal{G}_{\mu,\sigma}^{m\times n}$, $A = M + W$ *(cf. Theorem 3.1). Write* $l(G) = \min\{r(G), n\}$, $l(H) = \min\{m, r(H)\}$, $c(r) = 2.35$ *for* $r > 1$ *and* $c(1) = \sqrt{\frac{2}{\pi}}$. *Then*

(a) $F_{GA}(y) \le yc(r(G))\sqrt{l(G)}/(\sigma_{r(G)}(G)\sigma)$ *and*

(b) $F_{AH}(y) \le yc(r(H))\sqrt{l(H)}/(\sigma_{r(H)}(H)\sigma)$.

The theorem shows that $\sigma_{\mathrm{rank}(M)} \le y$ with a probablilty of at most the order y for $M = GA$ and $M = AH$, and so it is unlikely that multiplication by a square or rectangular Gaussian random matrix can dramatically decrease the smallest singular value of a matrix, even though we can have $UV = O$ for two rectangular unitary matrices U and V.

Proof. Lemma 3.2 and Fact 2.3 together imply part (a) for $r(G) = 1$. Now let $r(G) > 1$ and let $G = S_G\Sigma_G T_G^T$ be full SVD where $\Sigma_G = (\widehat{\Sigma}_G \mid O) = \widehat{\Sigma}_G(I_{r(G)} \mid O)$ and $\widehat{\Sigma}_G = \mathrm{diag}(\sigma_j(G))_{j=1}^{r(G)}$. Write $M_{r(G)} = (I_{r(G)} \mid O)T_G^TM$, $W_{r(G)} = (I_{r(G)} \mid O)T_G^TW$, $A_{r(G)} = (I_{r(G)} \mid O)T_G^TA = M_{r(G)} + W_{r(G)}$.

We have $GA = S_G\Sigma_G T_G^TA$, and so $\sigma_j(GA) = \sigma_j(\Sigma_G T_G^TA) = \sigma_j(\widehat{\Sigma}_G A_{r(G)})$ for all j by virtue of Lemma 4.2 (since S_G is a square unitary matrix) and because $\widehat{\Sigma}_G A_{r(G)} = \Sigma_G T_G^TA$. Furthermore $\sigma_j(GA) = \sigma_j(\widehat{\Sigma}_G A_{r(G)}) \ge \sigma_{r(G)}(G)\sigma_j(A_{r(G)})$ for all j by virtue of Lemma 4.1. For $j = r(G)$ obtain

$$\sigma_{r(G)}(GA) \ge \sigma_{r(G)}(G)\sigma_{r(G)}(A_{r(G)}). \tag{4.1}$$

We have $T_G^TW \in \mathcal{G}_{\mu,\sigma}^{m\times n}$ by virtue of Lemma 4.3, since T_G is a square unitary matrix; consequently $W_{r(G)} \in \mathcal{G}_{\mu,\sigma}^{r(G)\times n}$. Now estimate $F_{A_{r(G)}}(y)$ by applying Theorem 3.1 for $A = A_{r(G)}$, $M = M_{r(G)}$, and $W = W_{r(G)}$, combine this estimate with bound (4.1) and obtain part (a) of Theorem 4.1. Part (a) implies part (b) because $\sigma_j(AH) = \sigma_j((AH)^T) = \sigma_j(H^TA^T)$ for all j. □

Corollary 4.1. *Assume integers* m, n *and* l *and matrices* G, H, M, W *and* A *as in Theorem 4.1 and choose two scalars* y *and* z *such that* $y > 0$ *and* $z \ge 2\sigma\sqrt{l}$. *Then*

(a) $F_{\kappa(GA)}(||G||yz) \ge 2 - \exp(-\frac{(z-2\sigma\sqrt{l})^2}{2\sigma^2}) - yc(r(G))\sqrt{l(G)}/(\sigma_{r(G)}(G)\sigma)$ *and*

(b) $F_{\kappa(AH)}(||H||yz) \ge 2 - \exp(-\frac{(z-2\sigma\sqrt{l})^2}{2\sigma^2}) - yc(r(H))\sqrt{l(H)}/(\sigma_{r(H)}(H)\sigma)$.

Proof. Combine Theorems 3.2 for $y = z$ and 4.1. □

The following corollary extends the bound of Theorem 4.1 for a randomized matrix product to the respective bounds for its leading blocks; this implies that *randomized multiplication of a well conditioned matrix is expected to be generic preconditioning*, that is, to ensure (with probability one or near one) generic rank and conditioning profiles for the product.

Corollary 4.2. *Suppose $j, k, m,, n,$ q and s are integers, $1 \le j \le q$, $1 \le k \le s$, $G \in \mathbb{R}^{q\times m}$, $H \in \mathbb{R}^{n\times s}$, $M \in \mathbb{R}^{m\times n}$, $W \in \mathcal{G}_{\mu,\sigma}^{m\times n}$, $A = M + W$, and $y \ge 0$. Write $G_j = (I_j \mid O_{j,m-j})G$, $r(G_j) = \operatorname{rank}(G_j)$ and $l(G_j) = \min\{r(G_j), n\}$ for $j = 1, \dots, q$, $H_k = H \begin{pmatrix} I_k \\ O_{n-k,k} \end{pmatrix}$, $r(H_k) = \operatorname{rank}(H_k)$ and $l(H_k) = \min\{m, r(H_k)\}$ for $k = 1, \dots, s$. Then for all j and k in the above ranges we have*

(a) $F_{(GA)^{(j)}}(y) \le yc(r(G))\sqrt{l(G_j)}/(\sigma_{r(G_j)}\sigma)$,

(b) $F_{(AH)^{(k)}}(y) \le yc(r(H))\sqrt{l(H_k)}/(\sigma_{r(H_k)}\sigma)$, and consequently

(c) with probability one the matrices GA and AH have generic rank profile unless $\sigma = 0$.

Proof. For every j and every k apply Theorem 4.1 replacing G by G_j, H by H_k, and A by either $A \begin{pmatrix} I_j \\ O_{n-j,j} \end{pmatrix}$ or $(I_k \mid O_{k,m-k})A$. □

Combining the latter results with Theorem 2.2 implies that randomized multiplication is expected to make Gaussian elimination with no pivoting numerically safe, and similarly for block Gaussian elimination (cf. [PQZa]).

Corollary 4.3. *Suppose M is a normalized $m \times n$ well conditioned matrix of full rank, $||M|| = 1$, $B \in \mathcal{G}_{0,1}^{m\times m}$ and $C \in \mathcal{G}_{0,1}^{n\times n}$. Then Gaussian elimination with no pivoting applied to computing LU factorization of the matrices BM and MC is expected to involve no divisions by absolutely small values.*

Remark 4.1. *According to Section 3.3 the matrices in $\mathcal{T}_{\mu,\sigma}^{m\times n}$ are likely to be well conditioned; consequently Corollary 4.3 still holds where $B \in \mathcal{Z}_{f,0,\sigma}^{m\times m}$ and $C \in \mathcal{Z}_{f,0,\sigma}^{n\times n}$ unless $f \approx 0$ or $1/f \approx 0$.*

5 Randomized additive preconditioning

Suppose a matrix $\tilde{A} \in \mathbb{R}^{m\times n}$ has numerical nullity r, $0 < r < l = \min\{m, n\}$, $U \in \mathcal{G}_{0,\sigma}^{m\times r}$, $V \in \mathcal{G}_{0,\sigma}^{n\times r}$ for $\sigma = ||\tilde{A}||/(2\sqrt{r})$, and $\tilde{C} = \tilde{A} + UV^T$. (We reuse some notation of Section 2.4). Our goal is to employ the results of the previous section to prove that the additive preprocessing $\tilde{A} \Longrightarrow \tilde{C} = \tilde{A} + UV^T$ is expected to decrease the condition number of the matrix $\tilde{A}$ (cf. Remark 5.4). We first reduce the study of this nearly rank deficient matrix $\tilde{A}$ to the study of the nearby rank deficient matrix $A = \tilde{A} + E$ of rank $\rho = l - r$ obtained by setting to zero the singular values $\sigma_j(\tilde{A})$ for $j > l - r$ in the SVD $\tilde{A} = S\Sigma T^T$; consequently $||E|| = \sigma_{l-r+1}(\tilde{A})$.

Write $C = A + UV^T$ and $\tilde{C} = \tilde{A} + UV^T = C - E$, assume that the ratio $\frac{||E||}{||C||}$ is small, and in the rest of this section estimate the ratio $\frac{\kappa(C)}{\kappa(A)}$, which in our case closely approximates the ratio $\frac{\kappa(\tilde{C})}{\kappa(\tilde{A})}$.

The following results are readily verified.

Theorem 5.1. *Let $A = S\Sigma T^T$ be full SVD of an $m \times n$ matrix A of a rank ρ where $\rho < l = \min\{m, n\}$, S and T are square unitary matrices, $S \in \mathbb{R}^{m\times m}$, $T \in \mathbb{R}^{n\times n}$, $\Sigma = \operatorname{diag}(\Sigma_A, O_{m-\rho,n-\rho})$ is an $m \times n$ matrix, and $\Sigma_A = \operatorname{diag}(\sigma_j)_{j=1}^{\rho}$ is the $\rho \times \rho$ diagonal matrix of the positive singular values of the matrix A. Write $r = l - \rho$ and suppose $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, and the $m \times n$ matrix $C = A + UV^T$ has full rank l. Write*

$$S^T U = \begin{pmatrix} \bar{U} \\ U_r \end{pmatrix},\ T^T V = \begin{pmatrix} \bar{V} \\ V_r \end{pmatrix},\ R_U = \begin{pmatrix} I_{m-r} & \bar{U} \\ O_{r,m-r} & U_r \end{pmatrix},\ R_V = \begin{pmatrix} I_{n-r} & \bar{V} \\ O_{r,n-r} & V_r \end{pmatrix}.$$

Then $R_U \Sigma R_V^T = \Sigma$, whereas $R_U \operatorname{diag}(O_{m-r,n-r}, I_r) R_V^T = S^T U V^T T$, so that

$$C = S R_U D R_V^T T^T,\ D = \Sigma + \operatorname{diag}(O_{m-r,n-r}, I_r) = \operatorname{diag}(\Sigma_A, O_{m-\rho,n-\rho}, I_r)\ . \tag{5.1}$$

Theorem 5.2. *Under the assumptions of Theorem 5.1, write $p = ||R_U^{-1}||\ ||R_V^{-1}||$ and suppose $||A|| = 1$ and the $r \times r$ matrices U_r and V_r are nonsingular. Then (a) $p \ge \frac{||C^+||}{||A^+||} = \frac{\sigma_{l-r}(A)}{\sigma_l(C)}$ and (b) $1 \le p \le (1 + ||U||)(1 + ||V||)f_r$ where $f_r = \max\{1, ||U_r^{-1}||\}\ \max\{1, ||V_r^{-1}||\}$.*

Proof. (a) Combine the equations $S^{-1} = S^T$, $T^{-T} = T$ and (5.1) and obtain $C^+ = TR_V^{-T}D^+R_U^{-1}S^T$. Recall that S and T are unitary matrices, and so $||C^+|| = ||R_V^{-T}D^+R_U^{-1}|| \le ||R_V^{-T}||\ ||D^+||\ ||R_U^{-1}||$. Note that $D^+ = \Sigma^+ + \operatorname{diag}(O_{n-r,m-r}, I_r)$ (verify the Moore–Penrose conditions [GL96, Section 5.5.3]), recall that $\sigma_{l-r}(A) \le ||A|| = 1$ by assumption, deduce that $||D^+|| = ||\operatorname{diag}(\Sigma_A^{-1}, I_r)|| = 1/\sigma_{l-r}(A) = ||A^+||$, and obtain the claimed bound $p \ge \frac{||C^+||}{||A^+||}$.

(b) We have $R_U^{-1} = \begin{pmatrix} I_{m-r} & -\bar{U} \\ O & I_r \end{pmatrix} \begin{pmatrix} I_{m-r} & O \\ O & U_r^{-1} \end{pmatrix}$, $R_V^{-1} = \begin{pmatrix} I_{n-r} & -\bar{V} \\ O & I_r \end{pmatrix} \begin{pmatrix} I_{n-r} & O \\ O & V_r^{-1} \end{pmatrix}$, $||\bar{U}|| \le ||U||$ and $||\bar{V}|| \le ||V||$. Combine these relationships. □

Theorem 5.3. *Assume $A \in \mathbb{R}^{m\times n}$, $l = \min\{m,n\}$, $U \in \mathcal{G}_{0,\sigma}^{m\times r}$, $V \in \mathcal{G}_{0,\sigma}^{n\times r}$, U_r, V_r, and f_r in Theorem 5.2, and $z = (y - 2\sigma\sqrt{l})/(\sigma\sqrt{2})$. Then*

(a) $F_{||W||}(y) \ge 1 - \exp(-z^2)$ provided $z \ge 0$ and either $W = U$ or $W = V$ and

(b) the probability that $||C|| \ge ||A|| + y$, $p \ge f_r + y$, or $\frac{\kappa(C)}{\kappa(A)} \ge f_r + y$ decreases to zero exponentially in z^2 as $z \to \infty$.

Proof. Part (a) follows from Theorem 3.2. Part (b) follows from part (a), the bounds of Theorem 5.2 and the inequalities $||C|| \le ||A|| + ||UV^T||$ and $||UV^T|| \le ||U||\ ||V^T||$. □

Next we probabilistically bound the norms $||U_r^{-1}|| = \frac{1}{\sigma_r(U_r)}$ and $||V_r^{-1}|| = \frac{1}{\sigma_r(V_r)}$ based on Theorem 4.1 (see Section 7.5 on derandomization of our computations).

Theorem 5.4. *Suppose $A \in \mathbb{R}^{m\times n}$, $U \in \mathcal{G}_{\mu,\sigma}^{m\times r}$, $V \in \mathcal{G}_{\mu,\sigma}^{n\times r}$, U_r and V_r denote the five matrices in Theorem 5.1, $0 < r < l = \min\{m,n\}$, $f_r = \max\{||U_r^{-1}||^2, ||V_r^{-1}||^2\}$, and Theorem 4.1 holds*

(i) for $r(G) = r$, $G = (O \mid I_r)S^T$, and A replaced by U (in this case GA is replaced by U_r) as well as

(ii) for $r(G) = r$, $G = (O \mid I_r)T^T$, A replaced by V (in this case GA is replaced by V_r).

Also recall Definition 3.2 and assume that $y \ge 0$, $c(r) = 2.35$ for $r > 1$ and $c(1) = \sqrt{\frac{2}{\pi}}$ (cf. Theorem 4.1). Then

(a) $F_{U_r}(y) \le c(r)\ y\sqrt{r}/\sigma$, $F_{V_r}(y) \le c(r)\ y\sqrt{r}/\sigma$ and

(b) the matrix C is rank deficient with probability zero.

Proof. Apply part (a) of Theorem 4.1 for $r(G) = r$, $G = (O, I_r)S^T$ and $A = U$ and obtain that $F_{U_r}(y) \le c(r)y\sqrt{r}/(\sigma_r((O \mid I_r)S^T)\sigma)$. Then apply part (a) of Theorem 4.1 for $r(G) = r$, $G = (O \mid I_r)T^T$ and $A = V$ and obtain that $F_{V_r}(y) \le c(r)y\sqrt{r}/(\sigma_r((O, I_r)T^T)\sigma)$. Observe that $\sigma_r((O \mid I_r)S^T) = \sigma_r((O \mid I_r)T^T) = 1$ because S and T are unitary matrices. Substitute these equations into the above bounds on $F_{U_r}(y)$ and $F_{V_r}(y)$ and obtain part (a) of Theorem 5.4, which implies that the matrices U_r and V_r are singular with probability zero. Therefore part (b) follows from equation (5.1). □

Corollary 5.1. *Under the assumptions of Theorem 5.4 let $z = (y - 2\sigma\sqrt{l})/(\sigma\sqrt{2})$ and $\mu = 0$. Then Probability$(\kappa(C) \ge (f_r + y)\frac{\sigma_1(A)}{\sigma_{l-r}(A)})$ decreases to zero exponentially in z^2 as $y \to \infty$.*

Remark 5.1. *Under the assumptions of Corollary 5.1 suppose $\tilde{A} = A + E \approx A$ and $\tilde{C} = \tilde{A} + UV^T$. Then $\tilde{C} = C + E$, $||\tilde{C} - C|| \le ||E||$ and Theorem 2.1 bounds the norm $||\tilde{C}^+ - C^+||$. Therefore we can extend the estimates of Corollary 5.1 to $\kappa(\tilde{C})$ if the norm $||E||$ is sufficiently small.*

Remark 5.2. *The estimates of Section 3.3 imply that matrices in $\mathcal{T}_{\mu,\sigma}^{m\times n}$ are likely to have full rank and to be well conditioned; consequently the claims of Corollary 5.1 and Remark 5.1 also hold where $U \in \mathcal{T}_{0,\sigma}^{m\times r}$, $V \in \mathcal{T}_{0,\sigma}^{n\times r}$.*

Remark 5.3. *How large is our class of $m\times n$ matrices $\tilde{A}$ having a numerical rank $\rho = \min\{m,n\} - r$? We characterize it indirectly, by noting that by virtue of Fact 2.1 the nearby matrices A of rank ρ form a variety of dimension $(m + n - \rho)\rho$, which increases as ρ decreases.*

Remark 5.4. *One can relax the assumption that $\mu = 0$ in Theorem 3.2 (at the price of some complication of the probability estimates) and then extend Theorem 5.4, Corollary 5.1 and Remark 5.1 by inreasing the expected value of $||C||$ to $||A|| + \mu^2$ and the expected upper bound on $\kappa(C)$ to $2f_r||A^+||(||A|| + \mu^2)$ provided that $|\mu|$ has at most order $||A||$ (orthewise we would have expected that $C \approx UV^T$ and consequently that $\kappa(C) \gg \sigma_1(A)/\sigma_{l-r}(A)$). Note similar results for randomized augmentation in [PQa], linked to additive preprocessing in Section 8.1. In actual implementation of randomized additive preprocessing one can scale the matrix A, so that, say $||A|| \approx 1$, and then choose standard Gaussian random matrices U and V.*

6 The SMW and dual SMW formulae and dual additive preprocessing

Theorem 6.1. *Suppose $A \in \mathbb{R}^{m\times n}$, $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, $C = A + UV^T$, $0 < r < l = \min\{m, n\}$. Then*

$$\begin{pmatrix} C & U \\ V^T & I_r \end{pmatrix} = \begin{pmatrix} I_m & U \\ O_{r,m} & I_r \end{pmatrix} \begin{pmatrix} A & O_{m,r} \\ O_{r,n} & I_r \end{pmatrix} \begin{pmatrix} I_n & O_{n,r} \\ V^T & I_r \end{pmatrix}. \tag{6.1}$$

Furthermore suppose the matrix C has full rank l and define the $r \times r$ matrix $G = I_r - V^T C^+ U$, which is the Schur complement of the block C in $\begin{pmatrix} C & U \\ V^T & I_r \end{pmatrix}$. Then this matrix is nonsingular if and only if the matrix A has full rank l. Furthermore if $\mathrm{rank}(A) = l$, *then*

$$\begin{pmatrix} C & U \\ V^T & I_r \end{pmatrix} = \begin{pmatrix} I_m & O_{m,r} \\ V^T C^+ & I_r \end{pmatrix} \begin{pmatrix} C & O_{m,r} \\ O_{r,n} & G \end{pmatrix} \begin{pmatrix} I_n & C^+ U \\ O_{r,n} & I_r \end{pmatrix} \tag{6.2}$$

and

$$A^+ = (C - UV^T)^+ = C^+ + C^+ U G^{-1} V^T C^+. \tag{6.3}$$

We call (6.3) *generalized SMW formula*; it turns into the SMW classical formula of [GL96, page 50], [S98, Corollary 4.3.2] where $m = n$, $A^+ = A^{-1}$, and $C^+ = C^{-1}$.

Proof. Equations (6.1) and (6.2) are readily verified. The four unit block triangular matrices in these equations are nonsingular, and so the matrices $\mathrm{diag}(A, I_r)$ and $\mathrm{diag}(C, G)$ have full rank if and only if so does the matrix $\begin{pmatrix} C & U \\ V^T & I_r \end{pmatrix}$. Therefore $\mathrm{rank}(A) = l$ if and only if $\mathrm{rank}(G) = r$ provided $\mathrm{rank}(C) = l$ (as claimed).

Factorization (6.1) implies that

$$\mathrm{diag}(A, I_r) = \begin{pmatrix} A & O_{m,r} \\ O_{r,n} & I_r \end{pmatrix} = \begin{pmatrix} I_m & -U \\ O_{r,m} & I_r \end{pmatrix} \begin{pmatrix} C & U \\ V^T & I_r \end{pmatrix} \begin{pmatrix} I_n & O_{n,r} \\ -V^T & I_r \end{pmatrix}.$$

Substitute (6.2) and obtain

$$\mathrm{diag}(A, I_r) = \begin{pmatrix} I_m & -U \\ O_{r,m} & I_r \end{pmatrix} \begin{pmatrix} I_m & O_{m,r} \\ V^T C^+ & I_r \end{pmatrix} \mathrm{diag}(C, G) \begin{pmatrix} I_n & C^+ U \\ O_{r,n} & I_r \end{pmatrix} \begin{pmatrix} I_n & O_{n,r} \\ -V^T & I_r \end{pmatrix}.$$

Consequently

$$\mathrm{diag}(A^+, I_r) = \begin{pmatrix} I_n & O_{n,r} \\ V^T & I_r \end{pmatrix} \begin{pmatrix} I_n & -C^+ U \\ O_{r,n} & I_r \end{pmatrix} \mathrm{diag}(C^+, G^{-1}) \begin{pmatrix} I_m & O_{m,r} \\ -V^T C^+ & I_r \end{pmatrix} \begin{pmatrix} I_m & U \\ O_{r,m} & I_r \end{pmatrix} =$$

$$\begin{pmatrix} I_n & -C^+ U \\ V^T & G \end{pmatrix} \mathrm{diag}(C^+, G^{-1}) \begin{pmatrix} I_m & U \\ -V^T C^+ & G \end{pmatrix} = \begin{pmatrix} C^+ & -C^+ U G^{-1} \\ V^T C^+ & I_r \end{pmatrix} \begin{pmatrix} I_m & U \\ -V^T C^+ & G \end{pmatrix} =$$

$$\mathrm{diag}(C^+ + C^+ U G^{-1} V^T C^+, I_r).$$

Restrict the matrix equation $\mathrm{diag}(A, I_r) = \mathrm{diag}(C^+ + C^+ U G^{-1} V^T C^+, I_r)$ to its $n \times m$ leading blocks and obtain (6.3). □

Assume as before that $A \in \mathbb{R}^{m\times n}$, $U_- \in \mathbb{R}^{n\times q}$, $V_- \in \mathbb{R}^{m\times q}$, $0 < q < l = \min\{m, n\}$ and let the matrices C_- of equation (1.1) and

$$C_-^+ = A^+ + U_- V_-^T \tag{6.4}$$

have full rank l. Apply the generalized SMW formula (6.3) to the matrices A^+, U_- and V_- replacing C, $-U$ and V, respectively, obtain that $H = I_q + V_-^T A U_-$ is a nonsingular matrix and arrive at the *dual SMW formula* (1.1), (6.4). Alternatively define the matrices $H = I_q + V_-^T A U_-$ and $C_- = A - AU_- H^{-1} V_-^T A$ of (1.1), assume that they have full ranks, deduce that the matrix A has full rank as well, and obtain (6.4).

We naturally extend our additive preprocessing $A \Longrightarrow C = A + UV^T$ to the *dual additive preprocessing* $A^+ \Longrightarrow (C_-)^+$ for C_- of (1.1) and (6.4). Our analysis implies that $\kappa(C)_-$ is expected to have order $\sigma_{q+1}(A)/\sigma_l(A)$ provided $U_- \in \mathcal{G}_{0,1}^{n\times q}$, $V_- \in \mathcal{G}_{0,1}^{n\times q}$, and the norm $||A^+||$ is neither large nor small (cf. Section 7.5).

We can control the norm $||A^+||$ by scaling the matrix A. The randomized algorithm of [D83] produces a crude estimate for the norm $||A^+||$ at a low computational cost; in most of applications (e.g., to approximation of nmbs and matrix bases for singular spaces and to the solution of linear systems of equations) we can work with a matrix $\widehat{A} = \mathrm{diag}(A, \epsilon)$ instead of a matrix A and then can choose a positive ϵ sufficiently small to ensure that $||\widehat{A}^+|| = 1/\epsilon$.

7 Applications, derandomization and extensions of randomized additive preprocessing

7.1 Application to multiplicative preprocessing

Assume that $A \in \mathbb{R}^{m\times n}$, $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, and the matrix $C = A + UV^T$ has full rank. Then $C^+A = I_n - C^+UV^T$ where $m \geq n$, whereas $AC^+ = I_m - UV^TC^+$ where $m \leq n$. In both cases C is a r-preconditioner for the matrix A.

Furthermore suppose matrix A has numerical nullity r, the norm $||A||$ is neither large nor small, $U \in \mathcal{G}_{0,1}^{m\times r}$ and $V \in \mathcal{G}_{0,1}^{n\times r}$. Then according to Remarks 5.4 and 5.1, the matrix C is expected to be well conditioned and therefore more readily invertible than the ill conditioned matrix A.

Likewise assume matrices $A \in \mathbb{R}^{m\times n}$, H and C_- of (1.1) where C_- has full rank. Then $(C_-)^+ = A^+ + U_-V_-^T$ (cf. (6.4)), and so $(C_-)^+A = I_n + U_-V_-^TA$ where $m \geq n$, $A(C_-)^+ = I_m + AU_-V_-^T$ where $m \leq n$. In both cases C_- is a q-preconditioner for the matrix A.

Furthermore suppose matrix A has numerical rank q, the norm $||A^+||$ is neither large nor small, $U \in \mathcal{G}_{0,1}^{n\times q}$ and $V \in \mathcal{G}_{0,1}^{m\times q}$. Then the matrix C_- is expected to be well conditioned and therefore more readily invertible than the ill conditioned matrix A.

7.2 Estimation of numerical rank and numerical nullity and compression of preprocessors

Given an $m \times n$ matrix A having a numerical nullity r and numerical rank $q = n - r$, one can compute both integers r and q by means of at most $2\lceil \log_2 r \rceil$ steps of binary search whose every search step tests whether the matrix $C = A + UV^T$ has full rank and is well conditioned for a pair of $n \times s$ random and properly scaled matrices U and V and a candidate integer s, $s = 0, 1, 2, 4, \ldots$. Instead one can begin binary search with an upper bound $r_+ \geq r$, e.g., with $r_+ = n - 1$, and in at most $2\lceil \log_2(n - r) \rceil$ steps compute r as the minimum integer for which the matrix C has full rank and is well conditioned and the ratio $\frac{||AC^{-1}U||}{||A||\ ||C^{-1}U||}$ is small [PQ10, Algorithm 6.7]. This variant of the binary search is most attractive where $q = n - r \ll r$.

We can begin with scaled random additive preprocessors U and V of larger sizes expecting to obtain a better conditioned matrix $C = A + UV^T$, and then we can try to decrease the size of the preprocessor UV^T as much as we can still keeping the matrix C well conditioned.

To facilitate the binary search, one can apply the power transforms $A \Longrightarrow B = (AA^T)^h A$ for positive integers h. They increase the gaps between consecutive singular values of A because $\sigma_j(B) = (\sigma_j(A))^{2h+1}$.

7.3 Computation of nmbs and approximation of leading and trailing singular spaces

Theorem 7.1. *[PQ10, Theorem 3.1]. Suppose a matrix $A \in \mathbb{R}^{m\times n}$ has rank q, $0 < q < l = \min\{m, n\}$, $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, $r = l - q$, and the matrix $C = A + UV^T$ has full rank l. Then the matrix C^+U is a* nmb(A) *if $m \geq n$, whereas the matrix $C^{+T}V$ is a left* nmb(A) *if $n \geq m$.*

The following theorem extends these results to the matrices $C = A + UV^T$ and $(C_-)^+ = A^+ + U_-V_-^T$ of (1.1) and (6.4) where the matrix A has numerical rank q.

Theorem 7.2. *Assume a matrix $A \in \mathbb{R}^{m\times n}$ having numerical rank q where $0 < q < l = \min\{m, n\}$.*

(a) Write $r = l - q$ and suppose $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, and the matrix $C = A + UV^T$ has full rank and is well conditioned. Then there is a scalar c independent of A, U, V, m, n and q such that $\mathcal{R}(B_U) = \mathbb{T}_{A,r}$ and $||C^+U - B_U|| \leq c\sigma_{q+1}(A)||B_U||$ for some matrix $B_U \in \mathbb{R}^{m\times r}$ if $m \geq n$, whereas $\mathcal{R}(B_V) = \mathbb{S}_{A,r}$ and $||C^{+T}V - B_V|| \leq c\sigma_{q+1}(A)||B_V||$ for some matrix $B_V \in \mathbb{R}^{n\times r}$ if $n \geq m$.

(b) Assume four matrices of full ranks, $U_- \in \mathbb{R}^{n\times q}$, $V_- \in \mathbb{R}^{m\times q}$, $H = I_q + V_-AU_-^T$, and $C_- = A - AU_-H^{-1}V_-^TA$ (cf. (1.1)) where the matrix C_- is well conditioned. Then there exists a scalar c_- independent of A, U_-, V_-, m, n and q and such that $\mathcal{R}(B_{U_-}) = \mathbb{T}_A^{(q)}$ and $||C_-U_- - B_{U_-}|| \leq c_-\sigma_{r+1}(A)||B_{U_-}||$ for some matrix $B_{U_-} \in \mathbb{R}^{n\times q}$ if $m \geq n$, whereas $\mathcal{R}(B_{V_-}) = \mathbb{S}_A^{(q)}$ and $||C_-^TV_- - B_{V_-}|| \leq c_-\sigma_{r+1}(A)||B_{V_-}||$ for some matrix $B_{V_-} \in \mathbb{R}^{m\times q}$ if $n \geq m$.

Proof. See [PQ10, Section 7.1]. □

Part (a) of Theorem 7.2 states that $\mathcal{R}(C^+U) \approx \mathbb{T}_{A,r}$ if $n \leq m$ and $\mathcal{R}(C^{+T}V) \approx \mathbb{S}_{A,r}$ if $n \geq m$, that is, the linear spaces $\mathcal{R}(C^+U)$ for $n \leq m$ and $\mathcal{R}(C^{+T}V)$ for $n \geq m$ approximate the right and left trailing singular spaces associated with the r smallest singular values of the matrix A, respectively. (Some of these values can be zero). Likewise part (b) states that $\mathcal{R}(C_-U_-) \approx \mathbb{T}_A^{(q)}$ if $n \geq m$ and $\mathcal{R}(C_-^TV_-) \approx \mathbb{S}_A^{(q)}$ if $n \leq m$, that is, the linear spaces $\mathcal{R}(C_-U_-)$ for $n \geq m$ and $\mathcal{R}(C_-^TV_-)$ for $n \leq m$ approximate the right and left leading singular spaces associated with the q largest singular values of the matrix A, respectively.

Remark 7.1. *In the case where $m = n$ Theorems 7.1 and 7.2 define both left and right nmbs or approximations to both left and right trailing and leading singular spaces. This case is actually general. Indeed (a) $\mathcal{N}(A) = \mathcal{N}(A^TA)$, (b) $\mathcal{N}(A) = \mathcal{N}(B^TA)$ if $A, B \in \mathbb{R}^{m\times n}$ and B has full rank $m \leq n$, and (c) $(A \mid O_{m,n-m})\mathbf{u} = \mathbf{0}_m$ if and only if $A\widehat{\mathbf{u}} = \mathbf{0}_m$ provided $m > n$ and $\widehat{\mathbf{u}} = (I_n \mid O_{n,m-n})\mathbf{u}$, whereas $\mathbf{v}^T \begin{pmatrix} A \\ O_{n-m,n} \end{pmatrix} = \mathbf{0}_n^T$ if and only if $\widehat{\mathbf{v}}^TA = \mathbf{0}_n^T$ provided $n > m$ and $\widehat{\mathbf{v}} = \mathbf{v}^T \begin{pmatrix} I_m \\ O_{n-m,m} \end{pmatrix}$. Furthermore given an $m \times n$ matrix A for $m > n$, we can represent it as the sum $A = \sum_{i=1}^h A_i$ where $A_i = (0, B_i^T, 0)^T$ and B_i are $k_i \times n$ matrices for $i = 1, \ldots, h$, $\sum_{i=1}^h k_i \geq m$. Then $N(A) = \cap_{i=1}^h N(B_i)$, and [GL96, Theorem 12.4.1] simplifies the computation of the intersection of nmbs. Alternatively we can compute both left and right nmbs and approximate bases for both left and right trailing singular spaces of rectangular matrices by applying Theorem 8.2.*

7.4 Low-rank matrix approximation

If a matrix $A = S_A^T\Sigma_AT_A$ has a numerical rank q and if Q is a unitary matrix of rank q such that $\mathcal{R}(Q) \approx \mathbb{T}_A^{(q)}$, then the matrix AQQ^T has rank q and $AQQ^T \approx A$. Part (b) of Theorem 7.2 defines a randomized algorithm for computing such a matrix $Q = Q(C_-U_-)$. The computation is division-free except for the orthogonalization of the $n \times q$ matrix C_-U_- and the inversion of the $q \times q$ matrix H.

Unlike the customary low-rank matrix approximation in [HMT11] we use no auxiliary matrices of sizes exceeding $n \times q$.

By applying our algorithm to the displacement of a matrix A having a small numerical displacement rank [BM01], that is lying near a matrix with displacement structure, we obtain an approximation of A by a matrix having a small displacement rank. By using such approximations we can simplify Newton's structured matrix inversion (see Section 10).

7.5 Derandomization of additive and dual additive preprocessing

Let us derandomize our additive and dual additive preprocessing at the computational cost of the same order as the cost of randomized additive preprocessing.

Theorem 7.3. *Suppose $A \in \mathbb{R}^{m\times n}$, $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, $0 < q = n - r < n \le m$, and $\mathrm{rank}(UV^T) = r$. Write $C = A + UV^T$, $\sigma_j = \sigma_j(A)$ for all j and assume that $\sigma_1 \ge 1 \ge \sigma_q$ and $\mathrm{rank}(C) = n$ (in this case $\mathrm{rank}(U) = \mathrm{rank}(V) = r$). Write $U_1 = Q(C^+U)$, $V_1 = Q(C^{+T}V)$, and $C_1 = A + U_1V_1^T$.*

(a) If the matrix A has rank q, then the matrix C_1 has full rank and $\kappa(C_1) = \kappa(A) = \sigma_1(A)/\sigma_q(A)$.

(b) If the matrix A has numerical rank q, then the matrix C_1 has full rank and $\kappa(C_1) \approx \sigma_1(A)/\sigma_q(A)$.

Proof. Due to Theorem 7.1, the updated matrices U_1 and V_1 are the right and left nmbs for the matrix A, respectively. Let $A = \sum_{j=1}^{q} \sigma_j \mathbf{s}_j \mathbf{t}_j^T$ be SVD of the matrix A. Write $U_1 = (\mathbf{u}_j)_{j=1}^r$ and $V_1 = (\mathbf{v}_j)_{j=1}^r$ and obtain the SVD $C_1 = A + U_1V_1^T = \sum_{j=1}^{r} \mathbf{u}_j\mathbf{v}_j^T + \sum_{j=1}^{q} \sigma_j \mathbf{s}_j\mathbf{t}_j^T$. This implies part (a) of the theorem because $r = n - q$ and $\sigma_1(A) \ge 1 \ge \sigma_q(A)$.

Now set to zero the $n-q$ smallest singular values of the matrix A, apply part (a) to the resulting matrix, and go back to the input matrix to obtain part (b) by the continuity argument. □

Theorem 7.4. *Suppose a matrix $A \in R^{m\times n}$ has numerical rank $q < m \le n$, $U_- \in R^{n\times q}$, $V_- \in R^{m\times q}$ and $\sigma_{m-q}(A) \ge 1 \ge \sigma_m(A) > 0$. Define the matrices H and C_- of (1.1). Suppose they have full rank, in which case $\mathrm{rank}(U_-) = \mathrm{rank}(V_-) = q$. Write $U_-^{(1)} = Q(C_-U_-)$, $V_-^{(1)} = Q(V_-^TC_-)$, $H_1 = I_q + V_-^{(1)}A(U_-^{(1)})^T$, assume that the latter matrix is nonsingular, and define the matrix $C_-^{(1)} = A - AU_-^{(1)}H_1^{-1}(V_-^{(1)})^TA$ (cf. (1.1)). Then this matrix has full rank m, $(C_-^{(1)})^+ = A^+ + U_-^{(1)}(U_-^{(1)})^T$, and $\kappa(C_-^{(1)}) \approx \sigma_{q+1}(A)/\sigma_n(A)$.*

Proof. Apply part (b) of Theorem 7.3 to matrices A^+, U_- and V_- replacing A, U and V, respectively. □

With these theorems and the trick at the very end of Section 6 for computing $\sigma_n(A) = 1/||A^+||$ we can derandomize the algorithms supporting Theorem 7.2 and their application in the previous section; the singularity of the matrices H and H_1, rank deficiency of the matrices $A + UV^T$ in part (a) of Theorem 7.2 or $A^+ + U_-V_-^T$ in its part (b), and ill conditioning of these matrices are the only remaining potential sources of troubles with our preconditioning.

7.6 Block diagonalization with approximate trailing singular spaces

Theorem 7.5. *Let a matrix $A \in \mathbb{R}^{m\times n}$ have numerical rank $q < l = \min\{n, m\}$.*

(a) For $r = n - q$ assume matrices $L_0 \in \mathcal{G}_{0,1}^{n\times q}$ and $L_1 \in \mathbb{R}^{n\times r}$ such that $\mathcal{R}(L_1) \approx \mathbb{T}_{A,r}$ and $||L_1|| = 1$. Then the expected order is $\sigma_1(A)/\sigma_q(A)$ for $\kappa(AL_0)$ and at most $\sigma_{q+1}(A)$ for $||AL_1||$.

(b) For $\bar{r} = m - q$ assume two matrices $K_0 \in \mathcal{G}_{0,1}^{q\times m}$ and $K_1 \in \mathbb{R}^{\bar{r}\times m}$ such that $\mathcal{R}(K_1) \approx \mathbb{S}_A^{(\bar{r})}$ and $||K_1|| = 1$. Then the expected order is $\sigma_1(A)/\sigma_q(A)$ for $\kappa(K_0A)$ and at most $\sigma_{q+1}(A)$ for $||K_1A||$.

Proof. Estimate $\kappa(AL_0)$ and $\kappa(K_0A)$ by combining Theorems 3.2 and 4.1; estimate $||AL_1||$ and $||K_1A||$ by applying Theorems 7.2. □

Let the assumptions of both parts (a) and (b) hold, that is suppose that $r = n-q$, $\bar{r} = m-q$, $K_0 \in \mathcal{G}_{0,1}^{q\times m}$, $L_0 \in \mathcal{G}_{0,1}^{n\times q}$, $K_1 \in \mathbb{R}^{\bar{r}\times m}$, $L_1 \in \mathbb{R}^{n\times r}$, $\mathcal{R}(K_1) \approx \mathbb{S}_A^{(\bar{r})}$ and $\mathcal{R}(L_1) \approx \mathbb{T}_{A,r}$. Then Theorem 7.5 implies that the $q\times q$ leading block W_{00} of the $m\times n$ matrix $W = \begin{pmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{pmatrix} = \begin{pmatrix} K_0 \\ K_1 \end{pmatrix} A(L_0 \mid L_1)$ is expected to be well conditioned and to dominate the three other blocks; namely the theorem implies the expected order $\sigma_1(A)/\sigma_q(A)$ for $\kappa(W_{00})$ and at most $\sigma_{q+1}(A)$ for $||W_{01}|| + ||W_{10}|| + ||W_{11}||$.

Next assume for simplicity that $m = n$, $r = \bar{r}$ and compute the matrices K_1 and L_1 based on our randomized additive preconditioning. Alternatively one can rely on other algorithms that approximate trailing singular spaces, e.g., see Remark 8.1.

Algorithm 7.1. Block diagonalization with approximate trailing singular spaces.

INPUT: *Three integers n, r and q, $0 < q = n - r < n$, a matrix $A \in \mathbb{R}^{n\times n}$ having numerical rank q and scaled so that the norm $||A||$ is neither large nor small, and a Subroutine* LIN·SOLVE *that either solves a linear system of equations if it is nonsingular and well condtioned or outputs FAILURE otherwise.*

OUTPUT: *FAILURE or four random matrices K_0 and L_0 in $\mathbb{R}^{n\times q}$ and K_1 and L_1 in $\mathbb{R}^{n\times r}$ such that $W = (K_0 \mid K_1)^T A(L_0 \mid L_1) = \begin{pmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{pmatrix}$ and with a probability near one the $q \times q$ block submatrix $W_{00} = K_0^T A L_0$ is nonsingular, well conditioned, and strongly dominant, such that $\sigma_q(W_{00}) \gg \max\{||W_{01}||,\ ||W_{10}||,\ ||W_{11}||\}$.*

COMPUTATIONS:

1. *Generate four matrices $K_0, L_0 \in \mathcal{G}_{0,1}^{n\times q}$; $U, V \in \mathbb{G}_{0,1}^{n\times r}$. Output the matrices K_0 and L_0.*
2. *Compute the matrix $C = A + UV^T$ (expected to be nonsingular and well conditioned).*
3. *Apply the Subroutine* LIN·SOLVE *to compute and to output the matrices $K_1 = C^{-T}V$ and $L_1 = C^{-1}U$. Stop and output FAILURE if so does the subroutine.*

Correctness of the algorithm follows from Theorem 7.5.

Block Gauss–Jordan factorization

$$W = \begin{pmatrix} I & O \\ W_{10}W_{00}^{-1} & I \end{pmatrix} \begin{pmatrix} W_{00} & O \\ O & G \end{pmatrix} \begin{pmatrix} I & W_{00}^{-1}W_{01} \\ O & I \end{pmatrix}$$

for $G = W_{11} - W_{10}W_{00}^{-1}W_{01}$ reduces the inversion of the matrices W and A and the solution of a linear system $A\mathbf{y} = \mathbf{b}$ to the similar operations with the matrices W_{00} and G of smaller sizes, where the matrix W_{00} is expected to be nonsingular and well conditioned.

Remark 7.2. *Under the adopted assumptions on the output of Algorithm 7.1 we expect that the norms $||W_{00}^{-1}W_{01}||$ and $||W_{10}W_{00}^{-1}||$ are small and consequently $W \approx \mathrm{diag}(W_{00}, G)$.*

Remark 7.3. *Computation of the Schur complement G involves $O(n^2 r)$ flops. We generally need its highly accurate approximation to compute an uncorrupted solution $\mathbf{y}$ to a linear system $A\mathbf{y} = \mathbf{b}$ because we must counter the expected cancellation of the leading digits of some computed values. This computation, however, only involves an r/n fraction of the overall computational time required for the solution of this linear system by the customary algorithms. One can decrease the required computational precision by replacing the matrices K_1, L_1, K_0 and L_0 with the orthogonal matrices $Q(K_1)$, $Q(L_1)$, $I_n - Q(K_1)Q(K_1)^T$ and $I_n - Q(L_1)Q(L_1)^T$, respectively, and by applying derandomization in part (b) of Theorem 7.3 and iterative refinement; the number of flops involved in the refinement is proportional to its output precision.*

Remark 7.4. *Our proofs can be extended to the case where U, V, K_0, and L_0 are standard Gaussian random Toeplitz matrices with the respective decrease of the number of random parameters.*

Tables 9.8 and 9.9 demonstrate the power of this approach versus standard algorithms.

7.7 Block diagonalization with approximate leading singular spaces

Let us keep assuming a square matrix A. If it has a small positive numerical rank q, one can arrive at a dual variation of Algorithm 7.1 based on part (b) of Theorem 7.2 for simplicity. In this variation matrix inversions are limited to the $q \times q$ matrices H, $K_0^T K_0$ and $L_0^T L_0$. Alternatively one can employ other algorithms that approximate leading singular spaces, e.g., ones in [HMT11].

Algorithm 7.2. Block diagonalization with approximate leading singular spaces.

INPUT: *A Subroutine* LIN·SOLVE *that either solves a linear system of equations if it is nonsingular and well conditioned or outputs FAILURE otherwise, integers n, q and r, $0 < q = n - r < n$, and a nonsingular ill conditioned matrix $A \in \mathbb{R}^{n \times n}$ having numerical rank q and scaled so that the norm $||A^{-1}||$ is neither large nor small, in which case the norm $||A||$ is small (see the end of Section 6 on the approximation of this norm).*

OUTPUT: *FAILURE or four matrices $K_0,\ L_0 \in \mathbb{R}^{n \times q}$ and $K_1,\ L_1 \in \mathbb{R}^{n \times r}$ such that*

$$W = \begin{pmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{pmatrix} = (K_0 \mid K_1)^T A (L_0 \mid L_1)$$

and with a high probability the block submatrix $W_{00} = K_0^T A L_0$ is nonsingular, well conditioned, and strongly dominant, so that $\sigma_q(W_{00}) \gg \max\{||W_{01}||,\ ||W_{10}||,\ ||W_{11}||\}$.

COMPUTATIONS *(cf. (1.1)):*

1. *Generate four matrices $F, G \in \mathcal{G}_{0,1}^{n \times r}$; $U_-, V_- \in \mathcal{G}_{0,1}^{n \times q}$.*
2. *Compute the matrix $H = I_q + V_- A U_-^T$.*
3. *Apply the Subroutine* LIN·SOLVE *to compute the matrix H^{-1}. Stop and output FAILURE if so does the subroutine.*
4. *Compute the matrix $C_- = A - A U_- H^{-1} V_-^T A$ of (1.1).*
5. *Compute and output the matrices $K_0 = C_-^T V_-$ and $L_0 = C_- U_-$.*
6. *Compute and output the matrices*

$$K_1 = (I_n - K_0 (K_0^T K_0)^{-1} K_0^T) F \text{ and } L_1 = (I_n - L_0 (L_0^T L_0)^{-1} L_0^T) G\ .$$

Remarks 7.2–7.4 can be readily extended. We only specify an extension of Remark 7.3.

Remark 7.5. *The computation of the $q \times q$ auxiliary matrix H takes q/n fraction of the overall time involved in the customary algorithms for the solution of a linear system $A\mathbf{y} = \mathbf{b}$. This matrix should be computed with high accuracy to counter the expected cancellation of the leading digits of some computed values. One can rely on iterative refinement and can facilitate the task by means of orthogonalization of the matrices K_0, L_0, U_- and V_- and derandomization in Theorem 7.4 (toward numerical stabilization of the computations). As by-product of the orthogonalization of the matrices K_0 and L_0, we would have $K_0^T K_0 = L_0^T L_0 = I_q$, which would simplify Stage 6.*

7.8 Randomized additive preconditioning with the SMW recovery and the optimality of the computations

Suppose we seek the solution $\mathbf{y} = A^{-1}\mathbf{b}$ of a nonsingular linear system $A\mathbf{y} = \mathbf{b}$ of n equations where the real matrix A has a small positive numerical nullity r. Then randomized additive preprocessing $A \Longrightarrow C = A + UV^T$ for $U, V \in \mathcal{G}_{0,1}^{n \times r}$ and a matrix A having a norm bounded from above and below is expected to produce a well conditioned matrix C (cf. Remark 5.1). We can strengthen this expectation with derandomization of Section 7.5 and of the end of Section 6. The generalized SMW formula (6.3) implies that $\mathbf{y} = C^{-1}\mathbf{b} + C^{-1}UG^{-1}V^T C^{-1}\mathbf{b}$ for $G = I_r - V^T C^{-1} U$. Substitute $W_U = C^{-1}U$ and $\mathbf{w_b} = C^{-1}\mathbf{b}$ and obtain $\mathbf{y} = \mathbf{w_b} + W_U G^{-1} V^T \mathbf{w_b}$ for $G = I_r - V^T W_U$. This reduces the computation of $\mathbf{y}$ essentially to the solution of the matrix equation $CW = (U \mid \mathbf{b})$ for $W = (W_U \mid w_{\mathbf{b}})$, computing the matrix G, and its inversion. Here is a flowchart of this solution where we incorporate iterative refinement.

Flowchart 7.1. Randomized Solution of a Linear System with Iterative Refinement

INPUT: $\mathbf{b} \in \mathbb{R}^{n\times 1}$, *a nonsingular matrix* $A \in \mathbb{R}^{n\times n}$ *having a small positive numerical nullity* r.

OUTPUT: $\tilde{\mathbf{y}} \approx A^{-1}\mathbf{b}$.

COMPUTATIONS*:*

1. *Generate two matrices* $U, V \in \mathcal{G}_{0,\sigma}^{n\times r}$.
2. *Compute the matrix* $C = A + UV^T$ *(expected to be nonsingular and well conditioned).*
3. *Apply Gaussian elimination (or another direct algorithm) involving order* n^3 *flops to compute an approximate inverse* $X \approx C^{-1}$. *(Perform the computations by using single or double precision. Application of the same algorithm to the original ill conditioned linear system* $A\mathbf{y} = \mathbf{b}$ *would require about as many flops but in extended precision).*
4. *Employ this inverse as the basis for iterative refinement to compute sufficiently accurate solution* W *of the matrix equation* $CW = (U \mid \mathbf{b})$ *and then recover a close approximation to the vector* $\mathbf{y} = A^{-1}\mathbf{b}$ *via the generalized SMW formula (6.3).*

Handling an ill conditioned input A, we must perform the computations with extended precision to counter magnification of rounding errors, but we confine this mostly to computing the Schur complement $G = I_r - V^T C^{-1} U$, which takes the fraction r/n of the computational time of the customary algorithms for a linear system $A\mathbf{y} = \mathbf{b}$.

More precisely every loop of iterative refinement produces order $p - \log_2 \kappa(C)$ new correct bits per output value and is essentially reduced to multiplication of the matrices C and X by two vectors, that is to $4n^2 - 2n$ flops in a low (e.g., single or double) precision p. The refinement algorithm outputs order rn values; they can be accumulated with high accuracy as the sums of sufficiently many low precision summands, similarly to symbolic lifting [GG03], [P09/11].

Gaussian elimination uses $\frac{2}{3}n^3 + O(n^2)$ flops in extended precision of $p_+ \approx p_{\text{out}} + \log_2 \kappa(A)$ bits to output the solution to the ill conditioned linear system $A\mathbf{y} = \mathbf{b}$ with a prescribed precision p_{out}. We compute an approximate inverse X of the well conditioned matrix C by using also $O(n^3)$ flops, but in the low precision p.

Let $\mu(q)$ denote the number of bitwise operations involved in a flop performed with a precision q; $\mu(q)$ has order ranging from $(q \log q) \log\log q$ to q^2 depending on the magnitude of q and computer environment [GG03], [F07]. Since C is well conditioned and A is not, we can assume that $2\log_2 \kappa(A) \le p \ll p_+$. Then we immediately obtain that Flowchart 7.1 involves $O(n^3\mu(p)) + rn^2\mu(p)p_+/p$ bitwise operations overall. For large n and p_+ this is dramatically less than the order $n^3\mu(p_+)$ in Gaussian elimination; furthermore we yield optimality up to polylogarithmic factors in n and $\log_2 \kappa(A)$ provided $\mu(q) = O((q \log q) \log\log q)$. Indeed we have the information lower bound $\frac{1}{2}(n+1)np_+$ on the overall number of bitwise operations involved. This folows because we must process the $(n+1)n$ entries of A and $\mathbf{b}$, each represented with the precision of p_+ bits, to obtain the output with precision p_{out}, whereas every bitwise operation can process at most two bits.

To represent a Toeplitz-like input matrix A with a displacement generator of a length $d \ll n$, we process $2dnp_+$ input bits, which imples the information lower bound of dnp_+ bitwise operations. In this case the customary algorithms solve a linear system $A\mathbf{y} = \mathbf{b}$ and invert C by using $O(dn^2)$ flops (cf. [GKO95], [P10], [R06]), whereas iterative refinement takes $O(dn \log n)$ flops per iteration. The overall bitwise operation count decreases to $O(dn^2\mu(p) + (dn \log n)\mu(p)p_+/p)$ in Flowchart 7.1, which is dramatically less than the order $dn^2\mu(p_+)$ in the customary solutions and is within a polylogarithmic factor in n and $\log_2 \kappa(A)$ from the information lower bound dnp_+.

Remark 7.6. *One can replace iterative refinement with the CG or GMRES algorithms. They use no approximate inverse but are more sensitive to the success of preconditioning. In particular every CG loop (essentially multiplication of the matrices* C *and* C^T *by two vectors) produces order of* $1/\kappa(C)$ *new correct bits per an output value versus* $p - \log_2 \kappa(C)$ *in iterative refinement. Thus we need stronger upper bounds on* $\kappa(C)$ *to ensure progress in the presence of rounding errors.*

8 Randomized augmentation and structured preprocessing

8.1 Randomized augmentation

The solution of a nonsingular linear system of n equations $A\mathbf{y} = \mathbf{b}$ can be readily recovered from a null vector $\begin{pmatrix} \mathbf{y} \\ -1/\beta \end{pmatrix}$ of the matrix $K = (A \mid \beta\mathbf{b})$ for a nonzero scalar β. If the matrix A has numerical nullity one and if the ratio $||A||/||\beta\mathbf{b}||$ is neither large nor small, then on average vector $\mathbf{b}$ the matrix K is well conditioned [PQa]. Our next theorem links additive preprocessing $A \Longrightarrow C = A + UV^T$ to an extension of such augmentation techniques.

Theorem 8.1. *Suppose* $K = \begin{pmatrix} A & -U \\ WV^T & W \end{pmatrix} \in \mathbb{R}^{(m+r)\times(n+r)}$, $W \in \mathbb{R}^{r\times r}$ *is a nonsingular matrix,* $C = A + UV^T$. *Then* $K = \operatorname{diag}(I_m, W)\widehat{U}\operatorname{diag}(C, I_r)\widehat{V}$ *for* $\widehat{U} = \begin{pmatrix} I_m & -U \\ O_{r,m} & I_r \end{pmatrix}$, $\bar{U} = \widehat{U}^{-1} = \begin{pmatrix} I_m & U \\ O_{r,m} & I_r \end{pmatrix}$, $\widehat{V} = \begin{pmatrix} I_n & O_{n,r} \\ V^T & I_r \end{pmatrix}$, $\bar{V} = \widehat{V}^{-1} = \begin{pmatrix} I_n & O_{n,r} \\ -V^T & I_r \end{pmatrix}$. *Moreover if the matrices* C *and* W *or the matrix* K *have full rank, then all the three matrices* C, W, *and* K *have full rank,* $K^+ = \bar{V}\operatorname{diag}(C^+, I_r)\bar{U}\operatorname{diag}(I_m, W^{-1})$, *and* $C^+ = (I_n \mid O_{n,r})K^+(I_m \mid O_{m,r})^T$.

Together with Remark 5.1 the theorem implies that $\kappa(K)$ is expected to have order $\sigma_1(A)/\sigma_q(A)$ for $q = l - r$, $l = \min\{m, n\}$, so that the matrix K is expected to be well conditioned provided $U \in \mathcal{G}_{0,1}^{m\times r}$, $V \in \mathcal{G}_{0,1}^{n\times r}$, $W \in \mathcal{G}_{0,1}^{r\times r}$, the matrix A has numerical nullity at most r and the norm $||A||$ is neither large nor small.

[PQa] employs Theorem 4.1 to prove similar preconditioning property for the more general classes of augmentations, e.g., the northwestern augmentation

$$K = \begin{pmatrix} W & V^T \\ -U & A \end{pmatrix} \tag{8.1}$$

provided that $U \in \mathcal{G}_{0,\sigma}^{m\times r}$, $V \in \mathcal{G}_{0,\sigma}^{n\times r}$, $W \in \mathcal{G}_{0,\sigma}^{r\times r}$ or $W = I_r$ and the ratio $\sigma/||A||$ is neither large nor small. Together with Theorems 8.1 above and 8.2 below this leads to alternative derivation of the bounds on $\kappa(C)$ for $C = A + UV^T$ and Gaussian random U and V (cf. Remarks 5.1 and 7.1). Indeed the augmentation matrix K in Theorem 8.1 turns into the one of (8.1) for $W = I_r$ (up to block row and column interchange).

Next under (8.1) let the matrices A, W and K have full rank and write $S = A + UW^{-1}V^T$ and $R = I - V^TUW^{-1}$. Then the matrix S has full rank, S^+ is the $m \times n$ trailing (southwestern) block of K^+, and (6.3) for C replaced by S and U by UW^{-1} implies that

$$A^+ = S^+ + S^+UW^{-1}R^{-1}V^TS^+. \tag{8.2}$$

[PQa, Section 3.1] extends Theorem 7.1 as follows for both $m \geq n$ and $m < n$.

Theorem 8.2. *Assume two matrices* $A \in \mathbb{R}^{m\times n}$ *of a rank* $\rho < n$ *and* $V \in \mathbb{R}^{r\times n}$ *for* $r = n - \rho$. *Suppose the matrix* $K = \begin{pmatrix} V \\ A \end{pmatrix}$ *has full column rank* n. *Then* $B = K^{(I)}\begin{pmatrix} I_r \\ O \end{pmatrix}$ *is a* $nmb(A)$.

For $K^{(I)} = K^+$ the theorem supports approximation of trailing singular spaces of A.

Remark 8.1. *By applying Theorem 8.2 to both matrices* A *and* A^T *we can compute bases for both left and right singular spaces of the matrix* A *associated with its smallest singular values. We can employ these bases in Theorem 7.5 to extend Algorithm 7.1 to rectangular matrices* A.

Remark 8.2. *In the next subsections our augmentations use fewer random parameters by exploiting matrix structure, but saving random parameters by means of symmetization can lead to a pitfall: the map* $A \Longrightarrow K = \begin{pmatrix} W & V^T \\ V & A \end{pmatrix}$ *cannot decrease the condition number* $\kappa(A)$ *if* K *is a symmetric positive definite matrix because of the Interlacing Property of its eigenvalues [GL96, Theorem 8.1.7]. In contrast scaled randomized symmetric additive preprocessing* $A \Longrightarrow C = A + VV^T$ *is expected to work as preconditioning for an ill conditioned matrix* A *having small numericall nullity [W07].*

Remark 8.3. *One can embed an $m \times n$ matrix A into an $(m+r) \times (n+q)$ matrix banded with zeros and then view augmentation as its $2r$-rank modification. Alternatively one can apply such an augmentation to an $(m-r) \times (n-q)$ block of the matrix A and arrive at an $m \times n$ matrix K with r modified rows and q modified columns. We refer to this special case of randomized additive preprocessing, closely linked to augmentation, as* randomized (r, q) updating. *It can be analyzed similarly to augmentation.*

8.2 Randomized structured preprocessing

Would the $n \times n$ preprocessed matrices $C = A + UV^T$ inherit the structure of an $n \times n$ matrix A where $U, V \in \mathbb{R}^{n \times r}$? For a small value r the adverse impact of involving the $2nr$ entries of the matrices U and V on the structure is small, but it could be even smaller if we choose such matrices having structure consistent with the one of the matrix A. In the case of Gaussian random circulant multipliers B and C and scaled Gaussian random Toeplitz matrices U and V we have proved the preconditioning power of multiplicative and additive preprocessing, and we also confirmed such power empirically (see Remarks 3.3 and 4.1 and Section 3.3), but in our tests we consistently observed it even where we used other highly structured and sparse preconditioners B, C, U and V (see [PIMR10] and our Table 9.6).

Furthermore in the case of a nonsingular matrix M given with a displacement generator of a small length d we ontain the same results where such a matrix M has any numerical rank ρ because in this case the inverse M^{-1} can be readily expressed via the solution of $2d$ linear systems of equations with the matrices M and M^T.

Similar comments apply to dual additive preprocessing in Section 6 and randomized augmentation of (8.1).

8.3 A randomized Toeplitz solver

Gohberg and Sementsul in [GS72] express the inverse of a nonsingular Toeplitz matrix $T = (t_{i-j})_{i,j=1}^{n}$ via the columns $T^{-1}\mathbf{e}_1$ and $T^{-1}\mathbf{e}_n$ (see some extensions in [H79], [HR84], [T90]). The following theorem generates T^{-1} by pairs of columns $K^{-1}\mathbf{e}_1$ and $K^{-1}\mathbf{e}_{n+1}$ of the inverses K^{-1} of $(n+1) \times (n+1)$ Toeplitz matrices K that embed T as a block submatrix.

Theorem 8.3. *Suppose $K = (t_{i,j})_{i,j=0}^{n}$ is a nonsingular $(n+1) \times (n+1)$ Toeplitz matrix, write $T = (t_{i,j})_{i,j=0}^{n-1}$, $\widehat{\mathbf{v}} = (v_i)_{i=0}^{n} = K^{-1}\mathbf{e}_1$, $\mathbf{v} = (v_i)_{i=0}^{n-1}$, $\mathbf{v}' = (v_i)_{i=1}^{n}$, $\widehat{\mathbf{w}} = (w_i)_{i=0}^{n} = K^{-1}\mathbf{e}_{n+1}$, $\mathbf{w} = (w_i)_{i=0}^{n-1}$, and $\mathbf{w}' = (w_i)_{i=1}^{n}$. (a) If $v_0 \neq 0$, then the matrix $T = (t_{i,j})_{i,j=0}^{n-1}$ is nonsingular and $v_0 T^{-1} = Z(\mathbf{v})Z^T(J\mathbf{w}') - Z(\mathbf{w})Z^T(J\mathbf{v}')$. (b) If $v_n \neq 0$, then the matrix $T_{10} = (t_{i,j})_{i=1,j=0}^{n,n-1}$ is nonsingular and $v_n T^{-1} = Z(\mathbf{w})Z^T(J\mathbf{v}') - Z(\mathbf{v})Z^T(J\mathbf{w}')$.*

Proof. See [GS72] on part (a), [GK72] on part (b), reproduced in [BGY80, Theorem 7]. □

In the case of a nonsingular real symmetric matrix K the first and the last columns of the matrix K^{-1} turn into one another up to reflection, that is $K^{-1}\mathbf{e}_1 = J_{n+1}K^{-1}\mathbf{e}_{n+1}$, because in this case the inverse K^{-1} is both symmetric and persymmetric. Then part (a) of Theorem 8.3 expresses the matrix T^{-1} via the first column of the matrix K^{-1} alone.

Let us apply Theorem 8.3 to support our randomized augmentation techniques for solving a nonsingular Toeplitz linear system $T\mathbf{y} = \mathbf{b}$ of n equations provided the matrix T has numerical nullity one.

To compute the solution vector $\mathbf{y} = T^{-1}\mathbf{b}$, we first embed the matrix T into an $(n+1) \times (n+1)$ Toeplitz matrix $K = \begin{pmatrix} w & \mathbf{v}^T \\ \mathbf{f} & T \end{pmatrix}$ (cf. [GS72]). Here $w = \mathbf{e}_1^T T \mathbf{e}_1$ and the vectors $\mathbf{f} = (f_i)_{i=1}^{n}$ and $\mathbf{v} = (v_i)_{i=1}^{n}$ are filled with the respective entries of the matrix T except for the two coordinates f_n and v_n, which we choose at random and then scale to have the ratios $\frac{|f_n|}{||K||}$ and $\frac{|v_n|}{||K||}$ neither large nor small.

By virtue of Corollary A.2 this policy is likely to produce a nonsingular matrix K whose inverse is likely to have a nonzero entry $\mathbf{e}_1^T K^{-1}\mathbf{e}_1$. Our tests were in very good accordance with these two implications of Corollary A.2 and moreover consistently produced well conditioned matrices K.

Part (a) of Theorem 8.3 expresses the inverse T^{-1} via the first column $\mathbf{v} = K^{-1}\mathbf{e}_1$ and the last column $\mathbf{w} = K^{-1}\mathbf{e}_{n+1}$ of the inverse matrix K^{-1}.

To summarize, we reduce the solution of the original ill conditioned Toeplitz linear system $T\mathbf{y} = \mathbf{b}$ to computing highly accurate solutions of two linear systems $K\mathbf{x} = \mathbf{e}_1$ and $K\mathbf{z} = \mathbf{e}_{n+1}$, both expected to be well conditioned. High accuracy is needed to counter magnification of the input and rounding errors, expected in the case of ill conditioned input.

To solve the two latter systems, we first employ the Toeplitz linear solver of [KV99], [V99], [VBHK01], and [VK98], and then apply iterative refinement with double precision. We refer to the resulting algorithm as **Algorithm 8.1**.

In the important special case where the Toeplitz matrix T is real symmetric, we can choose a real scalar w and a real vector $\mathbf{f} = \mathbf{v}$ to yield a real symmetric matrix $K = \begin{pmatrix} w & \mathbf{v}^T \\ \mathbf{v} & T \end{pmatrix}$. Then Algorithm 8.1 is simplified because $K^{-1}\mathbf{e}_{n+1} = J_{n+1}\mathbf{v} = J_{n+1}K^{-1}\mathbf{e}_1$, and we only need to solve a single linear system with the matrix K. In Section 9.6 we test the resulting algorithm for solving an ill conditioned real symmetric Toeplitz linear system.

One can readily extend the approach of this section to the case of Toeplitz-like, Hankel and Hankel-like inputs as well as to augmenting the input matrix with $r > 1$ rows and columns and to randomized (r, r) updating in Remark 8.3. The transition to the solution of the original problem can employ expression (8.2) and either recursive application of Theorem 8.3 in the case of augmentation or the generalized SMW formula (6.3) in the case of randomized (r, r) updating.

9 Numerical Experiments

Our numerical experiments with random general, Hankel, Toeplitz and circulant matrices have been performed in the Graduate Center of the City University of New York on a Dell server with a dual core 1.86 GHz Xeon processor and 2G memory running Windows Server 2003 R2. The test Fortran code was compiled with the GNU gfortran compiler within the Cygwin environment. Random numbers were generated with the random_number intrinsic Fortran function, assuming the uniform probability distribution over the range $\{x : -1 \le x < 1\}$.

9.1 Conditioning tests

We have computed the condition numbers of $n \times n$ random general matrices for $n = 2^k$, $k = 5, 6, \ldots$, with the entries sampled in the range $[-1, 1)$ as well as complex general, Toeplitz, and circulant matrices whose entries had real and imaginary parts sampled at random in the same range $[-1, 1)$. We have performed 100 tests for each class of inputs, each dimension n, and each nullity r. Tables 9.2–9.4 represent the test results. The last four columns of each table display the average (mean), minimum, maximum, and standard deviation of the computed condition numbers of the input matrices, respectively. Namely we have computed the values $\kappa(A) = ||A||\ ||A^{-1}||$ for general and circulant matrices A and the values $\kappa_1(A) = ||A||_1\ ||A^{-1}||_1$ for Toeplitz matrices A. We have computed and displayed in Table 9.3 the 1-norms of Toeplitz matrices and their inverses rather than their 2-norms to facilitate the computations in the case of inputs of large sizes. Table 9.1 shows that the 1-norms and 2-norms are quite close to each other. It displays the data on $n \times n$ general, Toeplitz, and circulant matrices A for $n = 32, 64, \ldots, 1024$.

9.2 Preconditioning tests

Table 9.5 reproduces some results of testing the preconditioning power of additive preprocessing in [PIMR10]. We have tested the input matrices of the following classes.

1n. *Nonsymmetric matrices of type I with numerical nullity r.* $A = S\Sigma_r T^T$ are $n \times n$ matrices where G and H are $n \times n$ random orthogonal matrices, that is, the factors Q in the QR factorizations

of random real matrices; $\Sigma_r = \mathrm{diag}(\sigma_j)_{j=1}^n$ is the diagonal matrix such that $\sigma_{j+1} \le \sigma_j$ for $j = 1, \dots, n-1$, $\sigma_1 = 1$, the values $\sigma_2, \dots, \sigma_{n-r-1}$ are randomly sampled in the semi-open interval $[0.1, 1)$, $\sigma_{n-r} = 0.1$, $\sigma_j = 10^{-16}$ for $j = n-r+1, \dots, n$, and therefore $\kappa(A) = 10^{16}$ [H02, Section 28.3].

1s. *Symmetric matrices of type I with numerical nullity r.* The same as in part 1n, but for $G = H$.

The matrices of six other classes were constructed in the form of $\frac{A}{||A||} + \beta I$ where the recipes for defining the matrices A and scalars β are specified below.

2n. *Nonsymmetric matrices of type II with numerical nullity r.* $A = (W \mid WZ)$ where W and Z are random orthogonal matrices of sizes $n \times (n-r)$ and $(n-r) \times r$, respectively.

2s. *Symmetric matrices of type II with numerical nullity r.* $A = WW^T$ where W are random orthogonal matrices of size $n \times (n-r)$.

3n. *Nonsymmetric Toeplitz-like matrices with numerical nullity r.* $A = c(T \mid TS)$ for random Toeplitz matrices T of size $n \times (n-r)$ and S of size $(n-r) \times r$ and for a positive scalar c such that $||A|| \approx 1$.

3s. *Symmetric Toeplitz-like matrices with numerical nullity r.* $A = cTT^T$ for random Toeplitz matrices T of size $n \times (n-r)$ and a positive scalar c such that $||A|| \approx 1$.

4n. *Nonsymmetric Toeplitz matrices with numerical nullity one.* $A = (a_{i,j})_{i,j=1}^n$ is an $n \times n$ Toeplitz matrix. Its entries $a_{i,j} = a_{i-j}$ are random for $i - j < n - 1$. The entry $a_{n,1}$ is selected to ensure that the last row is linearly expressed through the other rows.

4s. *Symmetric Toeplitz matrices with numerical nullity one.* $A = (a_{i,j})_{i,j=1}^n$ is an $n \times n$ Toeplitz matrix. Its entries $a_{i,j} = a_{i-j}$ are random for $|i - j| < n - 1$, whereas the entry $a_{1,n} = a_{n,1}$ is a root of the quadratic equation $\det A = 0$. We have repeatedly generated the matrices A until we arrived at the quadratic equation having real roots.

We have set $\beta = 10^{-16}$ for the symmetric matrices A in the classes 2s, 3n, and 4s, so that $\kappa(A) = 10^{16} + 1$ in these cases. For the nonsymmetric matrices A we have defined the scalar β by an iterative process such that $||A|| \approx 1$ and $10^{-18}||A|| \le \kappa(A) \le 10^{-16}||A||$ [PIMR10, Section 8.2].

Table 9.5 displays the average values of the condition numbers $\kappa(C)$ of the matrices $C = A + UU^T$ over 100,000 tests for the inputs in the above classes, $r = 1, 2, 4, 8$ and $n = 100$. We have defined the additive preprocessor UU^T by a normalized $n \times r$ matrix $U = U/||U||$ where $U^T = (\pm I \mid O_{r,r} \mid \pm I \mid O_{r,r} \mid \dots \mid O_{r,r} \mid \pm I \mid O_{r,s})$, we have chosen the integer s to obtain $n \times r$ matrices U and have chosen the signs for the matrices $\pm I$ at random.

In our further tests the condition numbers of the matrices $C = A + 10^p UV^T$ for $p = -10, -5, 5, 10$ were steadily growing within a factor $10^{|p|}$ as the value $|p|$ was growing. This have showed the importance of proper scaling of the additive preprocessor UV^T.

9.3 Solution of general linear systems of equations with random circulant multipliers

Table 9.6 (cf. [PQZa, Table 2]) displays the results of our tests of the solution of a nonsingular well conditioned linear system $A\mathbf{y} = \mathbf{b}$ of n equations whose coefficient matrix had an ill conditioned $n/2 \times n/2$ submatrix for n ranging from 64 to 1024. We have performed 100 numerical tests for each dimension n and computed the maximum, minimum and average relative residual norms $||A\mathbf{y} - \mathbf{b}||/||\mathbf{b}||$ as well as standard deviation. GENP applied to these systems has output corrupted solutions with the residual norms ranging from 10 to 10^8. When we preprocessed the systems with circulant multipliers filled with ± 1 (with the n signs $\pm$ chosen at random), the norms decreased to at worst 10^{-7} for all inputs. Table 9.6 also shows further decrease of the norm in a single step of iterative refinement.

9.4 Approximation of the tails of SVDs and low-rank approximation of a matrix

Table 9.7 (cf. [PQ10, Section 10.6]) displays the data from our tests on the approximation of trailing singular spaces of the SVD of an $n \times n$ matrix A having numerical nullity $r = n - q$ and on the

approximation of this matrix with a matrix of rank $q = n - r$.

For $n = 64, 128, 256$ we have generated pairs of $n \times n$ random unitary matrices S and T and diagonal matrices $\Sigma = \mathrm{diag}(\sigma_j)_{j=1}^{n}$ such that $\sigma_j = 1/j,\ j = 1, \dots, q,\ \sigma_j = 10^{-10},\ j = q+1, \dots, n$. Then we computed the input matrices $A = S\Sigma T^T$ (with $\kappa(A) = 10^{10}$) as well as the matrix bases $T_r = T \begin{pmatrix} O_{q,r} \\ I_r \end{pmatrix}$ for the trailing singular spaces $\mathbb{T}_r$ of these matrices. Namely we have generated pairs of $n \times r$ random matrices U and V for $r = 1, 8, 32$, scaled them to have $||UV^T|| \approx ||A|| = 1$, and computed the matrices $C = A + UV^T$, $B_r = C^{-1}U$, AB_r, Y_r which minimized the norms $||B_rY_r - T_r||$; B_rY, $B_rY_r - T_r$, $Q = Q(B_r)$, and $AQQ^T = A - A(I_n - QQ^T)$.

Table 9.7 displays the data on the values $\kappa(A)$ and the relative residual norms $\mathrm{rrn}_1 = \frac{||B_rY_r - T_r||}{||B_rY_r||}$, $\mathrm{rrn}_2 = \frac{||AB_r||}{||A||\ ||B_r||}$, and $\mathrm{rrn}_3 = \frac{||AQQ^T||}{||A||}$ obtained in our tests.

9.5 Solution of linear systems of equations based on approximation of trailing singular spaces of the SVDs

At first we have chosen $n = 32, 64$ and $r = 1, 2, 4$ and for every pair (n, r) generated 100 instances of vectors $\mathbf{b}$ and matrices A, U, and V as follows.

We have generated (a) random vectors $\mathbf{b}$ of dimension n, (b) the matrices A as the error-free products $S\Sigma T^T$ where S and T were $n \times n$ random real orthonormal matrices (generated with double precision), $\Sigma = \mathrm{diag}(\sigma_j)_{j=1}^{n}$, $\sigma_{n-j} = 10^{j-17}$ for $j = 0, 1, \dots, r-1$, and $\sigma_{n-j} = 1/(n-j)$ for $j = r, \dots, n-1$ [H02, Section 28.3], and (c) $n \times r$ random matrices U and V such that $||A| = ||U|| = ||V|| = 1$.

For every choice of these matrices we have solved the linear systems $A\mathbf{y} = \mathbf{b}$ based on Algorithm 7.1. We first generated $n \times (n-r)$ random matrices K_0 and L_0 and then computed the matrices $C = A + UV^T$ (which were always nonsingular and well conditioned in our tests), $K_1 = C^{-T}V$, $L_1 = C^{-1}U$, and $W = (K_0 \mid K_1)^T A (L_0 \mid L_1) = \begin{pmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{pmatrix}$. In all our tests the $(n-r) \times (n-r)$ leading principal $(n-r) \times (n-r)$ block $W_{00} = K_0^T A L_0$ was well conditioned and strongly dominated the three other blocks W_{01}, W_{10}, and W_{11} in the 2×2 block matrix W, as we expected to see based on our analysis in Section 7.6. We computed the dominated blocks W_{01}, W_{10}, and W_{11} with extended precision. Then we solved the linear system $W\mathbf{x} = (K_0 \mid K_1)^T \mathbf{b}$. We first applied Gaussian elimination to eliminate the subdiagonal block. Then we readily computed the solution of the resulting block triangular linear system, whose both diagonal blocks were expected and consistently turned out to be much better conditioned than the original matrix A. Finally we computed and output the vector $\mathbf{y} = (L_0 \mid L_1)\mathbf{x}$.

Table 9.8 shows the average (mean), minimum and maximum values of the relative residual norms $||A\mathbf{y} - \mathbf{b}||/||\mathbf{b}||$ of the output vectors $\mathbf{y}$ as well as the standard deviations in these tests.

For the same ill conditioned inputs the Subroutine MLDIVIDE(A,B) for Gaussian elimination from MATLAB has produced corrupted outputs, as can be seen from Table 9.9.

9.6 Solution of a real symmetric Toeplitz linear system of equations with randomized augmentation

We have solved 100 real symmetric linear systems of equations $T\mathbf{y} = \mathbf{b}$ for each n where we used vectors $\mathbf{b}$ with random coordinates from the range $[-1, 1)$ and Toeplitz matrices $T = S + 10^{-9} I_n$ for an $n \times n$ singular symmetric Toeplitz matrices S having rank $n-1$ and nullity one and generated according to the recipe in [PQ10, Section 10.1b].

Table 9.10 shows the average CPU time of the solution by our Algorithm 8.1 and, for comparison, based on the QR factorization and SVD, which we computed by applying the LAPACK procedures DGEQRF and DGESVD, respectively.

The abbreviations "Alg. 8.1", "QR", and "SVD" indicate the respective algorithms. The last two columns of the table display the ratios of these data in the first and the two other columns.

We measured the CPU time with the mclock function by counting cycles. One can convert them into seconds by dividing their number by a constant CLOCKS_PER_SEC, which is 1000 on our platform. We marked the table entries by a "-" where the tests have run too long and were not completed.

We have obtained the solutions $\mathbf{y}$ with the relative residual norms of about 10^{-15} in all three algorithms, which showed that Algorithm 8.1 employing iterative refinement was as reliable as the QR and SVD based solutions but ran much faster.

We refer the reader to [PQZC, Table 3] on similar test results on the solution of ill conditioned homogeneous Toeplitz linear systems.

Table 9.1: Norms of random general, Toeplitz and circulant matrices and of their inverses

matrix A	n	$\|\|A\|\|_1$	$\|\|A\|\|_2$	$\frac{\|\|A\|\|_1}{\|\|A\|\|_2}$	$\|\|A^{-1}\|\|_1$	$\|\|A^{-1}\|\|_2$	$\frac{\|\|A^{-1}\|\|_1}{\|\|A^{-1}\|\|_2}$
General	32	1.9×10^1	1.8×10^1	1.0×10^0	4.0×10^2	2.1×10^2	1.9×10^0
General	64	3.7×10^1	3.7×10^1	1.0×10^0	1.2×10^2	6.2×10^1	2.0×10^0
General	128	7.2×10^1	7.4×10^1	9.8×10^{-1}	3.7×10^2	1.8×10^2	2.1×10^0
General	256	1.4×10^2	1.5×10^2	9.5×10^{-1}	5.4×10^2	2.5×10^2	2.2×10^0
General	512	2.8×10^2	3.0×10^2	9.3×10^{-1}	1.0×10^3	4.1×10^2	2.5×10^0
General	1024	5.4×10^2	5.9×10^2	9.2×10^{-1}	1.1×10^3	4.0×10^2	2.7×10^0
Toeplitz	32	1.8×10^1	1.9×10^1	9.5×10^{-1}	2.2×10^1	1.3×10^1	1.7×10^0
Toeplitz	64	3.4×10^1	3.7×10^1	9.3×10^{-1}	4.6×10^1	2.4×10^1	2.0×10^0
Toeplitz	128	6.8×10^1	7.4×10^1	9.1×10^{-1}	1.0×10^2	4.6×10^1	2.2×10^0
Toeplitz	256	1.3×10^2	1.5×10^2	9.0×10^{-1}	5.7×10^2	2.5×10^2	2.3×10^0
Toeplitz	512	2.6×10^2	3.0×10^2	8.9×10^{-1}	6.9×10^2	2.6×10^2	2.6×10^0
Toeplitz	1024	5.2×10^2	5.9×10^2	8.8×10^{-1}	3.4×10^2	1.4×10^2	2.4×10^0
Circulant	32	1.6×10^1	1.8×10^1	8.7×10^{-1}	9.3×10^0	1.0×10^1	9.2×10^{-1}
Circulant	64	3.2×10^1	3.7×10^1	8.7×10^{-1}	5.8×10^0	6.8×10^0	8.6×10^{-1}
Circulant	128	6.4×10^1	7.4×10^1	8.6×10^{-1}	4.9×10^0	5.7×10^0	8.5×10^{-1}
Circulant	256	1.3×10^2	1.5×10^2	8.7×10^{-1}	4.7×10^0	5.6×10^0	8.4×10^{-1}
Circulant	512	2.6×10^2	3.0×10^2	8.7×10^{-1}	4.5×10^0	5.4×10^0	8.3×10^{-1}
Circulant	1024	5.1×10^2	5.9×10^2	8.7×10^{-1}	5.5×10^0	6.6×10^0	8.3×10^{-1}

Table 9.2: Condition numbers $\kappa(A)$ of random matrices A

n	**input**	**min**	**max**	**mean**	**std**
32	real	2.4×10^1	1.8×10^3	2.4×10^2	3.3×10^2
32	complex	2.7×10^1	8.7×10^2	1.1×10^2	1.1×10^2
64	real	4.6×10^1	1.1×10^4	5.0×10^2	1.1×10^3
64	complex	5.2×10^1	4.2×10^3	2.7×10^2	4.6×10^2
128	real	1.0×10^2	2.7×10^4	1.1×10^3	3.0×10^3
128	complex	1.3×10^2	2.5×10^3	3.9×10^2	3.3×10^2
256	real	2.4×10^2	8.4×10^4	3.7×10^3	9.7×10^3
256	complex	2.5×10^2	1.4×10^4	1.0×10^3	1.5×10^3
512	real	3.9×10^2	7.4×10^5	1.8×10^4	8.5×10^4
512	complex	5.7×10^2	3.2×10^4	2.3×10^3	3.5×10^3
1024	real	8.8×10^2	2.3×10^5	8.8×10^3	2.4×10^4
1024	complex	7.2×10^2	1.3×10^5	5.4×10^3	1.4×10^4
2048	real	2.1×10^3	2.0×10^5	1.8×10^4	3.2×10^4
2048	complex	2.3×10^3	5.7×10^4	6.7×10^3	7.2×10^3

Table 9.3: Condition numbers $\kappa_1(A) = \frac{||A||_1}{||A^{-1}||_1}$ of random Toeplitz matrices A

n	**min**	**mean**	**max**	**std**
256	9.1×10^2	9.2×10^3	1.3×10^5	1.8×10^4
512	2.3×10^3	3.0×10^4	2.4×10^5	4.9×10^4
1024	5.6×10^3	7.0×10^4	1.8×10^6	2.0×10^5
2048	1.7×10^4	1.8×10^5	4.2×10^6	5.4×10^5
4096	4.3×10^4	2.7×10^5	1.9×10^6	3.4×10^5
8192	8.8×10^4	1.2×10^6	1.3×10^7	2.2×10^6

Table 9.4: Condition numbers $\kappa(A)$ of random circulant matrices A

n	**min**	**mean**	**max**	**std**
256	9.6×10^0	1.1×10^2	3.5×10^3	4.0×10^2
512	1.4×10^1	8.5×10^1	1.1×10^3	1.3×10^2
1024	1.9×10^1	1.0×10^2	5.9×10^2	8.6×10^1
2048	4.2×10^1	1.4×10^2	5.7×10^2	1.0×10^2
4096	6.0×10^1	2.6×10^2	3.5×10^3	4.2×10^2
8192	9.5×10^1	3.0×10^2	1.5×10^3	2.5×10^2
16384	1.2×10^2	4.2×10^2	3.6×10^3	4.5×10^2
32768	2.3×10^2	7.5×10^2	5.6×10^3	7.1×10^2
65536	2.4×10^2	1.0×10^3	1.2×10^4	1.3×10^3
131072	3.9×10^2	1.4×10^3	5.5×10^3	9.0×10^2
262144	6.3×10^2	3.7×10^3	1.1×10^5	1.1×10^4
524288	8.0×10^2	3.2×10^3	3.1×10^4	3.7×10^3
1048576	1.2×10^3	4.8×10^3	3.1×10^4	5.1×10^3

Table 9.5: Preconditioning tests

Type	r	Cond (C)
1n	1	3.21E+2
1n	2	4.52E+3
1n	4	2.09E+5
1n	8	6.40E+2
1s	1	5.86E+2
1s	2	1.06E+4
1s	4	1.72E+3
1s	8	5.60E+3
2n	1	8.05E+1
2n	2	6.82E+3
2n	4	2.78E+4
2n	8	3.59E+3
2s	1	1.19E+3
2s	2	1.96E+3
2s	4	1.09E+4
2s	8	9.71E+3
3n	1	2.02E+4
3n	2	1.53E+3
3n	4	6.06E+2
3n	8	5.67E+2
3s	1	2.39E+4
3s	2	2.38E+3
3s	4	1.69E+3
3s	8	6.74E+3
4n	1	4.93E+2
4n	2	4.48E+2
4n	4	2.65E+2
4n	8	1.64E+2
4s	1	1.45E+3
4s	2	5.11E+2
4s	4	7.21E+2
4s	8	2.99E+2

Table 9.6: Relative residual norms of the solutions by GENP with randomized circulant multiplicative preprocessing

dimension	**iterations**	**min**	**max**	**mean**	**std**
64	0	4.7×10^{-14}	8.0×10^{-11}	4.0×10^{-12}	1.1×10^{-11}
64	1	1.9×10^{-15}	5.3×10^{-13}	2.3×10^{-14}	5.4×10^{-14}
256	0	1.7×10^{-12}	1.4×10^{-7}	2.0×10^{-9}	1.5×10^{-8}
256	1	8.3×10^{-15}	4.3×10^{-10}	4.5×10^{-12}	4.3×10^{-11}
1024	0	1.7×10^{-10}	4.4×10^{-9}	1.4×10^{-9}	2.1×10^{-9}
1024	1	3.4×10^{-14}	9.9×10^{-14}	6.8×10^{-14}	2.7×10^{-14}

Table 9.7: Approximation of tails of the SVDs and low-rank approximation of a matrix (cf. [PQ10])

r	$\kappa(A)$ or rrn_i	n	**min**	**max**	**mean**	**std**
1	cond(A)	64	$2.38 \times 10^{+02}$	$1.10 \times 10^{+05}$	$6.25 \times 10^{+03}$	$1.68 \times 10^{+04}$
1	cond(A)	128	$8.61 \times 10^{+02}$	$7.48 \times 10^{+06}$	$1.32 \times 10^{+05}$	$7.98 \times 10^{+05}$
1	cond(A)	256	$9.70 \times 10^{+02}$	$3.21 \times 10^{+07}$	$3.58 \times 10^{+05}$	$3.21 \times 10^{+06}$
1	rrn_1	64	4.01×10^{-10}	1.50×10^{-07}	5.30×10^{-09}	1.59×10^{-08}
1	rrn_1	128	7.71×10^{-10}	5.73×10^{-07}	1.58×10^{-08}	6.18×10^{-08}
1	rrn_1	256	7.57×10^{-10}	3.2×10^{-07}	1.69×10^{-08}	5.02×10^{-08}
1	rrn_2	64	1.07×10^{-08}	4.71×10^{-06}	1.46×10^{-07}	4.90×10^{-07}
1	rrn_2	128	3.64×10^{-08}	3.05×10^{-05}	8.35×10^{-06}	3.29×10^{-06}
1	rrn_2	256	8.25×10^{-08}	3.30×10^{-05}	1.72×10^{-06}	5.03×10^{-06}
1	rrn_3	64	4.01×10^{-10}	1.50×10^{-07}	5.30×10^{-09}	1.59×10^{-08}
1	rrn_3	128	7.71×10^{-10}	5.73×10^{-07}	1.58×10^{-08}	6.18×10^{-08}
1	rrn_3	256	7.57×10^{-10}	3.22×10^{-07}	1.69×10^{-08}	5.02×10^{-08}
8	cond(A)	64	$1.26 \times 10^{+03}$	$1.61 \times 10^{+07}$	$2.68 \times 10^{+05}$	$1.71 \times 10^{+06}$
8	cond(A)	128	$2.92 \times 10^{+03}$	$3.42 \times 10^{+06}$	$1.58 \times 10^{+05}$	$4.12 \times 10^{+05}$
8	cond(A)	256	$1.39 \times 10^{+04}$	$8.75 \times 10^{+07}$	$1.12 \times 10^{+06}$	$8.74 \times 10^{+06}$
8	rrn_1	64	3.39×10^{-10}	2.27×10^{-06}	2.74×10^{-08}	2.27×10^{-07}
8	rrn_1	128	4.53×10^{-10}	1.91×10^{-07}	1.03×10^{-08}	2.79×10^{-08}
8	rrn_1	256	8.74×10^{-10}	1.73×10^{-07}	7.86×10^{-09}	1.90×10^{-08}
8	rrn_2	64	3.90×10^{-08}	1.47×10^{-04}	1.79×10^{-06}	1.47×10^{-05}
8	rrn_2	128	9.56×10^{-08}	2.97×10^{-05}	1.50×10^{-06}	4.12×10^{-06}
8	rrn_2	256	2.99×10^{-07}	3.91×10^{-05}	2.56×10^{-06}	5.70×10^{-06}
8	rrn_3	64	1.54×10^{-09}	7.59×10^{-06}	8.87×10^{-08}	7.58×10^{-07}
8	rrn_3	128	1.82×10^{-09}	7.27×10^{-07}	2.95×10^{-08}	8.57×10^{-08}
8	rrn_3	256	2.62×10^{-09}	3.89×10^{-07}	2.27×10^{-08}	5.01×10^{-08}
32	cond(A)	64	$1.77 \times 10^{+03}$	$9.68 \times 10^{+06}$	$1.58 \times 10^{+05}$	$9.70 \times 10^{+05}$
32	cond(A)	128	$1.65 \times 10^{+04}$	$6.12 \times 10^{+07}$	$1.02 \times 10^{+06}$	$6.19 \times 10^{+06}$
32	cond(A)	256	$3.57 \times 10^{+04}$	$2.98 \times 10^{+08}$	$4.12 \times 10^{+06}$	$2.98 \times 10^{+07}$
32	rrn_1	64	2.73×10^{-10}	3.29×10^{-08}	2.95×10^{-09}	4.93×10^{-09}
32	rrn_1	128	3.94×10^{-10}	1.29×10^{-07}	7.18×10^{-09}	1.64×10^{-08}
32	rrn_1	256	6.80×10^{-10}	4.00×10^{-07}	1.16×10^{-08}	4.27×10^{-08}
32	rrn_2	64	2.59×10^{-08}	2.11×10^{-06}	2.07×10^{-07}	3.29×10^{-07}
32	rrn_2	128	1.45×10^{-07}	1.82×10^{-05}	1.50×10^{-06}	2.76×10^{-06}
32	rrn_2	256	3.84×10^{-07}	7.06×10^{-05}	5.27×10^{-06}	1.14×10^{-05}
32	rrn_3	64	2.10×10^{-09}	1.49×10^{-07}	1.55×10^{-08}	2.18×10^{-08}
32	rrn_3	128	2.79×10^{-09}	3.80×10^{-07}	3.81×10^{-08}	6.57×10^{-08}
32	rrn_3	256	5.35×10^{-09}	1.05×10^{-06}	5.70×10^{-08}	1.35×10^{-07}

Table 9.8: Relative residual norms for a linear system of equations via nmb approximation

n	**r**	**min**	**max**	**mean**	**std**
32	1	1.49×10^{-13}	1.36×10^{-9}	4.25×10^{-11}	1.56×10^{-10}
32	2	3.70×10^{-13}	2.13×10^{-8}	3.83×10^{-10}	2.35×10^{-9}
32	4	9.33×10^{-13}	1.08×10^{-8}	3.37×10^{-10}	1.26×10^{-9}
64	1	1.11×10^{-12}	6.87×10^{-9}	2.03×10^{-10}	7.49×10^{-10}
64	2	1.53×10^{-12}	1.21×10^{-8}	5.86×10^{-10}	1.77×10^{-9}
64	4	2.21×10^{-12}	1.27×10^{-7}	1.69×10^{-9}	1.28×10^{-8}

Table 9.9: Relative residual norms for a linear system of equations with MLDIVIDE(A,B)

n	**r**	**min**	**max**	**mean**	**std**
32	1	6.34×10^{-3}	7.44×10^{1}	1.74×10^{0}	7.53×10^{0}
32	2	2.03×10^{-2}	1.32×10^{1}	9.19×10^{-1}	1.62×10^{0}
32	4	4.57×10^{-2}	1.36×10^{1}	1.14×10^{0}	1.93×10^{0}
64	1	3.82×10^{-3}	9.93×10^{0}	1.03×10^{0}	1.66×10^{0}
64	2	1.96×10^{-2}	1.27×10^{2}	3.09×10^{0}	1.40×10^{1}
64	4	7.13×10^{-3}	6.63×10^{0}	8.23×10^{-1}	1.20×10^{0}

Table 9.10: CPU time (in cycles) for solving an ill conditioned real symmetric Toeplitz linear system

n	**Alg. 8.1**	**QR**	**SVD**	**QR/Alg. 8.1**	**SVD/Alg. 8.1**
512	56.3	148.4	4134.8	2.6	73.5
1024	120.6	1533.5	70293.1	12.7	582.7
2048	265.0	11728.1	–	44.3	–
4096	589.4	–	–	–	–
8192	1304.8	–	–	–	–

10 Preprocessing for Newton–Toeplitz iteration

Recall Newton's iteration for matrix inversion

$$X_{i+1} = X_i(2I - CX_i),\ i = 0, 1, \dots. \tag{10.1}$$

Its ith loop squares the residual $I - CX_i$, that is,

$$I - CX_{i+1} = (I - CX_i)^2 = (I - CX_0)^{2^{i+1}}. \tag{10.2}$$

Therefore

$$||I - CX_{i+1}|| \le ||I - CX_i||^2 = ||I - CX_0||^{2^{i+1}},\ i = 0, 1, \dots, \tag{10.3}$$

so that the approximations X_i quadratically converge to the inverse C^{-1} right from the start provided that $||I - CX_0|| < 1$.

We can ensure that $||I - CX_0|| \le 1 - \frac{2n}{(\kappa(C))^2(1+n)}$ by choosing $X_0 = \frac{2nC^T}{(1+n)||C||_1||C||_\infty}$ [PS91].

Such a map $C \Longrightarrow X_0$ preserves the matrix structure of Toeplitz or Hankel type, but is the structure maintained throughout the iteration? Not automatically. In fact Newton's loop can triple the displacement rank of a matrix X_k. The structure can be maintained, however, via recursive compression of the displacement (also called recompression), in which case we arrive at *Newton's structured* (e.g., Newton–Toeplitz) iteration. In particular we can periodically set to zero the smallest singular values of the displacements of the matrices X_i to keep the length of the displacements within a fixed tolerance t, equal to or a little exceeding the displacement rank of the input matrix C. At this stage we can also apply the techniques for approximating the displacements of the matrices X_i by low-rank matrices (cf. Section 7.4 and [HMT11]).

We refer the reader to [P01, Chapter 6] on the history, variations, and analysis of this approach, proposed in [P92], [P93], and [P93a] for Toeplitz-like matrices. In [PBRZ99, Section 7.5.4] this iteration has been linked to iterative refinement that updated the input matrix. In [BM01] the extension of this study has naturally led to an important concept of approximate displacement rank of matrix. According to the estimates in [P01], the Newton–Toeplitz iteration converges quadratically right from the start provided $||I - CX_0|| < \frac{1}{(1+||Z_e||+||Z_f||)\kappa(C)}||L^{-1}||$, $||L^{-1}|| \le c_{e,f}n$, L denotes the associated displacement operator $L:\ C \to Z_eC - CZ_f$ for $e \ne f$ or $L:\ C \to C - Z_eCZ_f^T$ for $ef \ne 1$, and $c_{e,f}$ is a constant defined by e and f. Similar bounds can be deduced for other classes of matrices with displacement structure [P01, Section 6.6], [PRW02].

Newton's iteration can be incorporated into our randomized algorithms. E.g., it can be used instead of Gaussian elimination in Flowchart 7.1. Conversely one can apply preconditioning to decrease the initial residual norm $||I - CX_0||$ where it is close to one. The experiments reported in [P01, Table 6.21] suggest another combination of Newton's iteration with our preprocessing in the case of Toeplitz matrices C. Namely in this case the experiments show global convergence of Newton's structured iteration with compression (right from the start) in about 25% of tests, including the cases where the initial residual norm $||I - CX_0||$ was very close to one.

Motivated by these tests we can concurrently apply Newton–Toeplitz iteration to a number of scaled randomized small rank modifications and (r, r)-updatings of the input matrix. As soon as one of these applications produces the inverse, we can readily recover the inverse of the original matrix via the SMW formula (6.3) or in case of augmentation via (8.2) or Theorem 8.1.

Of course it is interesting whether this approach can also work for other classes of structured matrices and under variations of the compression policy of Newton's structured iteration in [PS91], [P01, Chapter 6], and [PVWC04]. One can replace Newton's iteration with iterative refinement of Gohberg–Sementsul's pairs or displacement generators for the inverse, which enables iterative updating of the inverse and has local quadratic convergence (cf. [PBRZ99]). On further study of this approach see [PZa].

11 Related work, our technical contributions and further study

Early work on approximation by low-rank matrices with applications to matrix and tensor decomposition can be traced through [HMT11], [GTZ97], [GT01], [GOS08], [T00], [MMD08], [OT09], and the bibliography therein, but even much earlier advances in this subject appeared in [BCLR79], [B80], [B85], [B86], [BC87] in the study of the border rank for matrices and tensors, first directed to acceleration of matrix multiplication. Likewise, exploiting the link between tensor and matrix computations for their acceleration is a fashionable subject with applications to many important areas of modern computing (see, e.g., [T00], [MMD08], [OT09]), but then again its earliest examples appeared in the latter papers and in [P72], which introduced the technique of trilinear aggregation as a basic ingredient of fast algorithms for matrix multiplication [P84], [CW90], [LPS92], [K04], but perhaps more importantly this technique was the first example of the acceleration of fundamental matrix computations by means of tensor decomposition.

Preconditioned iterative algorithms for linear systems of equations is a classical subject [A94], [B02], [G97]. The open problem of creating inexpensive preconditioners for general use has been around for a long while as well. For earlier study of conditioning of random matrices see [D88], [E88], [ES05], [CD05], [SST06]; estimation of the condition numbers of random structured matrices was stated as a challenge in [SST06]; we provide such estimates for random Toeplitz and circulant matrices in Sections 3.3 and 3.4. In particular the estimates show that the expected condition numbers do not grow fast as the size of a Toeplitz matrix grows; this should be surprising in view of [BG05].

Our present study of randomized preconditioning formally supports and substantially advances the techniques proposed and developed by the first author in [PGMQ], [PIMR10], [PQa], [PQZa], [PQZC], and [PZ11]. Our technical novelties include extension from the study of conditioning of random matrices to the proof of preconditioning power of additive preprocessing with random general and Toeplitz preconditioners, randomized and derandomized low-rank matrix approximation, randomized structured multiplicative preconditioning, and block factorization in Sections 7.6 and 7.7 based on randomized approximation of trailing and leading singular spaces.

We plan to study our approximation by low-rank matrices further (as we stated in Section 1.1) and to extend them to tensor computations. Other natural directions for our further study include extensions to augmentation and randomized (r, q)-updating in Remark 8.3 as well as specification of our techniques to structured matrices, particularly to matrices with dispalcement structure [KKM79], treated in a unified way.

Unification of the computations with structured matrices of Toeplitz, Hankel, Vandermonde and Cauchy type based on operating with them in terms of their displacements and the method of displacement transformation can be traced to [P90] (cf. [P01]). Treatment of ill conditioned structured matrices is a well known challenge (cf. [VBHK01]); the best customary recipes employ displacement transformation and involve quadratic arithmetic time or large overhead constants [GKO95], [CGLX], [CGSXZ], [G98], [P10], [R06]; our present advance relies on randomized additive preconditioning and in [PQa] on augmentation.

Appendix

A Uniform random sampling and nonsingularity of random matrices

Let $|\Delta|$ denote the cardinality of a set Δ in any fixed ring. *Uniform random sampling* of elements from a set Δ is their selection from this set at random, independently of each other and under the uniform probability distribution on the set Δ.

The total degree of a multivariate monomial is the sum of its degrees in all its variables. The total degree of a polynomial is the maximal total degree of its monomials.

Lemma A.1. *[DL78], [S80], [Z79]. For a set Δ of cardinality $|\Delta|$ in any fixed ring let a polynomial in m variables have a total degree d and let it not vanish identically on this set. Then the polynomial vanishes in at most $d|\Delta|^{m-1}$ points.*

Lemma A.1 implies that a fixed nonvanishing polynomial vanishes with probablity zero or converging to zero if the values of its variables are sampled under variuos reasonable probability distributions (e.g., uniform and Gaussian) on the set Δ whose cardinality is infinite or grows to infinity. Under the uniform probability distribution the probability is readily estimated even for a fixed finite set S.

Corollary A.1. *Under the assumptions of Lemma A.1 let the values of the variables of the polynomial be randomly and uniformly sampled from the set Δ. Then the polynomial vanishes with a probability at most $\frac{d}{|\Delta|}$.*

Corollary A.2. *Let the entries of an $m \times n$ matrix have been randomly and uniformly sampled from a finite set Δ of cardinality $|\Delta|$ (in any fixed ring). Let $l = \min\{m, n\}$. Then (a) every $k \times k$ submatrix M for $k \le l$ is nonsingular with a probability at least $1 - \frac{k}{|\Delta|}$ and (b) is strongly nonsingular with a probability at least $1 - \sum_{i=1}^{k} \frac{i}{|\Delta|} = 1 - \frac{(k+1)k}{2|\Delta|}$. Furthermore (c) if the submatrix M is indeed nonsingular, then any entry of its inverse is nonzero with a probability at least $1 - \frac{k-1}{|\Delta|}$.*

Proof. The claimed properties of nonsingularity and nonvanishing hold for generic matrices. The singularity of a $k \times k$ matrix means that its determinant vanishes, but the determinant is a polynomial of total degree k in the entries. Therefore Corollary A.1 implies parts (a) and consequently (b). Part (c) follows because a fixed entry of the inverse vanishes if and only if the respective entry of the adjoint vanishes, but up to the sign the latter entry is the determinant of a $(k-1) \times (k-1)$ submatrix of the input matrix M, and so it is a polynomial of degree $k-1$ in its entries. □

References

[A94] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, England, 1994.

[B80] D. Bini, Border Rank of $p \times q \times 2$ Tensors and the Optimal Approximation of a Pair of Bilinear Forms, in *Lecture notes in Computer Science*, **85**, 98–108, Springer, 1980.

[B85] D. Bini, Tensor and Border Rank of Certain Classes of Matrices and the Fast Evaluation of Determinant, Inverse Matrix and Eigenvalues, *Calcolo*, **22**, 209–228, 1985.

[B86] D. Bini, Border Rank of $m \times n \times (mn-q)$ Tensors, *Linear Algebra and Its Applications*, **79**, 45–51, 1986.

[B02] M. Benzi, Preconditioning Techniques for Large Linear Systems: a Survey, *J. of Computational Physics*, **182**, 418–477, 2002.

[BA80] R. R. Bitmead, B. D. O. Anderson, Asymptotically Fast Solution of Toeplitz and Related Systems of Linear Equations, *Linear Algebra and Its Applications*, **34**, 103–116, 1980.

[BC87] D. Bini, M. Capovani, Tensor Rank and Border Rank of Band Toeplitz Matrices, *SIAM J. on Computing*, **2**, 252–258, 1987.

[BCLR79] D. Bini, M. Capovani, G. Lotti, F. Romani $O(n^{2.7799})$ Complexity for $n \times n$ Approximate Matrix Multiplication, *Information Processing Letters*, **8**, 234–235, 1979.

[BG05] A. Böttcher, S. M. Grudsky, *Spectral Properties of Banded Toeplitz Matrices*, SIAM Publications, Philadelphia, 2005.

[BGY80] R. P. Brent, F. G. Gustavson, D. Y. Y. Yun, Fast Solution of Toeplitz Systems of Equations and Computation of Padé Approximations, *J. Algorithms*, **1**, 259–295, 1980.

[BM01] D. A. Bini, B. Meini, Approximate Displacement Rank and Applications, in *AMS Conference "Structured Matrices in Operator Theory, Control, Signal and Image Processing"*, Boulder, 1999 (edited by V. Olshevsky), *American Math. Society*, 215–232, Providence, RI, 2001.

[CD05] Z. Chen, J. J. Dongarra, Condition Numbers of Gaussian Random Matrices, *SIAM. J. on Matrix Analysis and Applications*, **27**, 603–620, 2005.

[CDG03] B. Carpentieri, I.S. Duff, L. Giraud, A class of spectral two-level preconditioners, *SIAM J. Scientific Computing*, **25, 2**, 749–765, 2003.

[CGLX] S. Chandrasekaran, M. Gu, X. S. Li, J. Xia, Superfast Multifrontal Method for Large Structured Linear Systems of Equations, *SIAM. J. on Matrix Analysis and Applications*, **31**, 1382–1411, 2009.

[CGSXZ] S. Chandrasekaran, M. Gu, X. Sun, J. Xia, J. Zhu, A Superfast Algorithm for Toeplitz Systems of linear Equations, *SIAM. J. on Matrix Analysis and Applications*, **29**, **4**, 1247–1266, 2007.

[CPW74] R.E. Cline, R.J. Plemmons, and G. Worm, Generalized Inverses of Certain Toeplitz Matrices, *Linear Algebra and Its Applications,* **8**, 25–33, 1974.

[CW90] Coppersmith, S. Winograd, Matrix Multiplicaton via Arithmetic Progressions. *J. Symbolic Comput.*, **9, 3**, 251–280, 1990.

[D83] J. D. Dixon, Estimating Extremal Eigenvalues and Condition Numbers of Matrices, *SIAM J. on Numerical Analysis*, **20, 4**, 812–814, 1983.

[D88] J. Demmel, The Probability That a Numerical Analysis Problem Is Difficult, *Math. of Computation*, **50**, 449–480, 1988.

[DL78] R. A. Demillo, R. J. Lipton, A Probabilistic Remark on Algebraic Program Testing, *Information Processing Letters*, **7**, **4**, 193–195, 1978.

[DS01] K. R. Davidson, S. J. Szarek, Local Operator Theory, Random Matrices, and Banach Spaces, in *Handbook on the Geometry of Banach Spaces* (W. B. Johnson and J. Lindenstrauss editors), pages 317–368, North Holland, Amsterdam, 2001.

[E88] A. Edelman, Eigenvalues and Condition Numbers of Random Matrices, *SIAM J. on Matrix Analysis and Applications*, **9**, **4**, 543–560, 1988.

[ES05] A. Edelman, B. D. Sutton, Tails of Condition Number Distributions, *SIAM J. on Matrix Analysis and Applications*, **27**, **2**, 547–560, 1988.

[F07] M. Fürer, Faster Integer Multiplication, *Proceedings of 39th Annual Symposium on Theory of Computing (STOC 2007)*, 57–66, ACM Press, New York, 2007.

[G97] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[G98] M. Gu, Stable and Efficient Algorithms for Structured Systems of Linear Equations, *SIAM J. on Matrix Analysis and Applications*, **19**, 279–306, 1998.

[GG03] J. von zur Gathen, J. Gerhard, *Modern Computer Algebra*, Cambridge University Press, Cambridge, UK, 2003 (second edition).

[GK72] I. Gohberg, N. Y. Krupnick, A Formula for the Inversion of Finite Toeplitz Matrices, *Matematicheskiie Issledovaniia* (in Russian), **7**, **2**, 272–283, 1972.

[GKO95] I. Gohberg, T. Kailath, V. Olshevsky, Fast Gaussian Elimination with Partial Pivoting for Matrices with Displacement Structure, *Math. of Comp.,* **64(212)**, 1557–1576, 1995.

[GL96] G. H. Golub, C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, 1996 (third addition).

[GOS08] S. A. Goreinov, I. V. Oseledets, D. V. Savostyanov, E. E. Tyrtyshnikov, N. L. Zamarashkin, How to Find a Good Submatrix, Research Report 08-10, ICM HKBU, Kowloon Tong, Hong Kong, 2008.

[GS72] I. Gohberg, A. Sementsul, On the Inversion of Finite Toeplitz Matrices and Their Continuous Analogs, *Matematicheskiie Issledovaniia* (in Russian), **7**, **2**, 187–224, 1972.

[GT01] S. A. Goreinov, E. E. Tyrtyshnikov, The Maximal-volume Concept in Approximation by Low-rank Matrices, *Contemporary Mathematics*, **208**, 47–51, 2001

[GTZ97] S. A. Goreinov, E. E. Tyrtyshnikov, N. L. Zamarashkin, A Theory of Pseudo-skeleton Approximations, *Linear Algebra and Its Applications*, **261**, 1–22, 1997.

[H79] G. Heinig, Beitrage zur spektraltheorie von Operatorbuschen und zur algebraischen Theorie von Toeplitzmatrizen, Dissertation **B**, *TH Karl-Marx-Stadt*, 1979.

[H02] N. J. Higham, *Accuracy and Stability in Numerical Analysis*, SIAM, Philadelphia, 2002 (second edition).

[HMT11] N. Halko, P. G. Matrinsson, J. A. Tropp, Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions, *SIAM Reviews*, **53, 2**, 217–288, 2011.

[HR84] G. Heinig, K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators, Operator Theory*, **13**, Birkhäuser, 1984.

[K04] I. Kaporin, The Aggregation and Cancellation Techniques As a Practical Tool for Faster Matrix Multiplication, *Theoretical Computer Science*, **315**, **2–3**, 469–510, 2004.

[KKM79] T. Kailath, S. Y. Kung, M. Morf, Displacement Ranks of Matrices and Linear Equations, *Journal Math. Analysis and Appls*, **68(2)**, 395–407, 1979.

[KV99] P. Kravanja, M. Van Barel, Algorithms for Solving Rational Interpolation Problems Related to Fast and Superfast Solvers for Toeplitz Systems, *SPIE*, 359–370, 1999.

[LPS92] J. Laderman, V. Y. Pan, H. X. Sha, On Practical Algorithms for Accelerated Matrix Multiplication, *Linear Algebra and Its Applications*, **162–164**, 557–588, 1992.

[M80] M. Morf, Doubling Algorithms for Toeplitz and Related Equations, *Proceedings of IEEE International Conference on ASSP*, 954–959, IEEE Press, Piscataway, New Jersey, 1980.

[MMD08] M. W. Mahoney, M. Maggioni, P. Drineas, Tensor-CUR decompositions for tensor-based data, *SIAM Journal on Matrix Analysis and Applications*, **30, 2**, 957–987, 2008.

[OT09] I. V. Oseledets, E. E. Tyrtyshnikov, Breaking the Curse of Dimensionality, or How to Use SVD in Many Dimensions, *SIAM J. on Scientific Computing*, **31**, **5**, 3744–3759, 2009.

[P72] V. Y. Pan, On Schemes for the Evaluation of Products and Inverses of Matrices (in Russian), *Uspekhi Matematicheskikh Nauk*, **27**, **5 (167)**, 249–250, 1972.

[P84] V. Y. Pan, How Can We Speed up Matrix Multiplication? *SIAM Review,* **26**, **3**, 393–415, 1984.

[P90] V. Y. Pan, On Computations with Dense Structured Matrices, *Math. of Computation*, **55, 191**, 179–190, 1990. Proceedings version in *Proceedings of International Symposium on Symbolic and Algebraic Computation (ISSAC'89)*, 34–42, ACM Press, NY, 1989.

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

[P93] V. Y. Pan, Concurrent Iterative Algorithm for Toepliz-like Linear Systems, *IEEE Transactions on Parallel and Distributed Systems*, **4, 5**, 592–600, 1993.

[P93a] V. Y. Pan, Decreasing the Displacement Rank of a Matrix, *SIAM Journal on Matrix Analysis and Applications*, **14**, **1**, 118–121, 1993.

[P01] V. Y. Pan, *Structured Matrices and Polynomials: Unified Superfast Algorithms*, Birkhäuser/Springer, Boston/New York, 2001.

[P09/11] V. Y. Pan, Nearly Optimal Solution of Rational Linear Systems of Equations with Symbolic Lifting and Numerical Initialization, *Computers and Mathematics (with Applications)*, **62**, 1685–1706, 2011. Proceedings version in International Symposium on Symbolic-Numerical Computations (Kyoto, Japan, August 2009), (edited by Hiroshi Kai and Hiroshi Sekigawa), pp.105–113, ACM Press, New York (2009).

[P10] F. Poloni, A Note on the O(n)-Storage Implementation of the GKO Algorithm, *Numerical Algorithms*, **55**, 115–139, 2010.

[PBRZ99] V. Y. Pan, Brahnam, B. Murphy, R. E. Rosholt, Newton's Iteration for Structured Matrices and Linear Systems of Equations, *SIAM Volume on Fast Reliable Algorithms for Matrices with Structure* (T. Kailath and A. H. Sayed, editors), 189–210, SIAM Publications, Philadelphia, 1999.

[PGMQ] V. Y. Pan, D. Grady, B. Murphy, G. Qian, R. E. Rosholt, A. Ruslanov, Schur Aggregation for Linear Systems and Determinants, *Theoretical Computer Science*, *Special Issue on Symbolic–Numerical Algorithms* (D. A. Bini, V. Y. Pan, and J. Verschelde editors), **409**, **2**, 255–268, 2008.

[PIMR10] V. Y. Pan, D. Ivolgin, B. Murphy, R. E. Rosholt, Y. Tang, X. Yan, Additive Preconditioning for Matrix Computations, *Linear Algebra and Its Applications*, **432**, 1070–1089, 2010.

[PQ10] V. Y. Pan, G. Qian, Randomized Preprocessing of Homogeneous Linear Systems of Equations, *Linear Algebra and Its Applications*, **432**, 3272–3318, 2010.

[PQa] V. Y. Pan, G. Qian, Solving Linear System with Randomized Augmentation and Aggregation, Tech. Report TR 20110009, *Ph.D. Program in Computer Science*, *Graduate Center, the City University of New York*, 2011.
Available at http://www.cs.gc.cuny.edu/tr/techreport.php?id=407

[PQZa] V. Y. Pan, G. Qian, A. Zheng, Randomized Preprocessing versus Pivoting, *Linear Algebra and Its Applications*, in print.

[PQZC] V. Y. Pan, G. Qian, A. Zheng, Z. Chen, Matrix Computations and Polynomial Root-finding with Preprocessing, *Linear Algebra and Its Applications*, **434**, 854–879, 2011.

[PRW02] V. Y. Pan, Y. Rami, X. Wang, Structured Matrices and Newtons Iteration: Unified Approach, *Linear Algebra and Its Applications*, **343/344**, 233-265, 2002.

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

[PVWC04] V. Y. Pan, M. Van Barel, X. Wang, G. Codevico, Iterative Inversion of Structured Matrices, *Theoretical Computer Science*, 315, **2–3** (Special Issue on Algebraic and Numerical Computing, edited by I. Z. Emiris, B. Mourrain, and V. Y. Pan), 581–592, 2004.

[PZ11] V. Y. Pan, A. Zheng, New Progress in Real and Complex Polynomial Root-Finding, *Computers and Math. (with Applications)* **61**, 1305–1334, 2011.

[PZa] V. Y. Pan, P. Zlobich, Can We Employ Autocorrection of Newton's Iteration with Recompression for Structured Matrix Inversion?, preprint, 2011.

[R06] G. Rodriguez, Fast Solution of Toeplitz- and Cauchy-like Least Squares Problems, *SIAM J. Matrix Analysis and Applications*, **28, 3**, 724–748, 2006.

[S80] J. T. Schwartz, Fast Probabilistic Algorithms for Verification of Polynomial Identities, *Journal of ACM*, **27**, **4**, 701–717, 1980.

[S98] G. W. Stewart, *Matrix Algorithms, Vol I: Basic Decompositions*, SIAM, Philadelphia, 1998.

[SST06] A. Sankar, D. Spielman, S.-H. Teng, Smoothed Analysis of the Condition Numbers and Growth Factors of Matrices, *SIAM Journal on Matrix Analysis*, **28**, **2**, 446–476, 2006.

[T90] W. F. Trench, A Note on a Toeplitz Inversion Formula, *Linear Algebra and Its Applications*, **29**, 55–61, 1990.

[T00] E. E. Tyrtyshnikov, Incomplete Cross Approximation in Mosaic Skeleton Method. *Computing*, **64**, 367–380, 2000.

[V99] M. Van Barel, A Superfast Toeplitz Solver, 1999.
Available at http://www.cs.kuleuven.be/~ marc/software/index.html

[VBHK01] M. Van Barel, G. Heinig, P. Kravanja, A Stabilized Superfast Solver for Nonsymmetric Toeplitz Systems, *SIAM Journal on Matrix Analysis and Applications*, **23**, **2**, 494–510, 2001.

[VK98] M. Van Barel, P. Kravanja, A Stabilized Superfast Solver for Indefinite Hankel Systems, *Linear Algebra and its Applications*, **284**, **1–3**, 335–355, 1998.

[W04] M. Wschebor, Smoothed Analysis of $\kappa(a)$, *J. of Complexity*, **20**, 97–107, 2004.

[W07] X. Wang, Affect of Small Rank Modification on the Condition Number of a Matrix, *Computer and Math. (with Applications)*, **54**, 819–825, 2007.

[Z79] R. E. Zippel, Probabilistic Algorithms for Sparse Polynomials, *Proceedings of EUROSAM'79, Lecture Notes in Computer Science*, **72**, 216–226, Springer, Berlin, 1979.