2013

TR-2013011: Fast Approximation Algorithms for Cauchy Matrices, Polynomials and Rational Functions

Victor Y. Pan

Fast Approximation Algorithms for Cauchy Matrices, Polynomials and Rational Functions *

Victor Y. Pan

Department of Mathematics and Computer Science
Lehman College and the Graduate Center of the City University of New York
Bronx, NY 10468 USA
victor.pan@lehman.cuny.edu,
home page: http://comet.lehman.cuny.edu/vpan/

Abstract. The papers [MRT05], [CGS07], [XXG12], and [XXCBa] have combined the advanced FMM techniques with transformations of matrix structures (traced back to [P90]) in order to devise numerically stable algorithms that approximate the solutions of Toeplitz, Hankel, Toeplitz-like, and Hankel-like linear systems of equations in nearly linear arithmetic time, versus classical cubic time and quadratic time of the previous advanced algorithms. We show that the power of these approximation algorithms can be extended to yield similar results for computations with other matrices that have displacement structure, which includes Vandermonde and Cauchy matrices, as well as to polynomial and rational evaluation and interpolation. The resulting decrease of the running time of the known approximation algorithms is again by order of magnitude, from quadratic to nearly linear. We present detailed description and analysis of the algorithms and provide new insights into the subject, formal complexity estimates, and the support for further advances in [Pa]. The techniques of our study can be of independent interest.

Key words: Cauchy matrices, Fast Multipole Method, HSS matrices, Matrix compression, Polynomial evaluation, Rational interpolation

AMS Subject Classification: 12Y05, 15A04, 47A65, 65D05, 68Q25

1 Introduction

An important area of recent progress in Numerical Linear Algebra is the design and implementation of numerically stable algorithms for approximate solution of Toeplitx, Hankel, Toeplitz-like, and Hankel-like linear systems of equations in nearly linear arithmetic time, versus the classical cubic time and the previous

* Some results of this paper have been presented at the 18th Conference of the International Linear Algebra Society (ILAS'2013), Providence, RI, 2013 and at the 15th Annual Conference on Computer Algebra in Scientific Computing (CASC 2003), September 9–13, 2013, Berlin, Germany

record quadratic time (see [MRT05], [CGS07], [XXG12], and [XXCBa]). The algorithms transform the matrix structures of Toeplitz and Hankel types into the structure of Cauchy type (which is a special case of the general technique proposed in [P90]) and then approximate a basic Cauchy matrix by HSS matrices, by applying a variant of the fast numerically stable FMM. "HSS" and "FMM" are the acronyms for "Hierarchically Semiseparable" and "Fast Multipole Method", respectively. "Historically HSS representation is just a special case of the representations commonly exploited in the FMM literature" [CDG06].

Our present subjects are the analysis of the successful algorithms of [MRT05], [CGS07], [XXG12], and [XXCBa] and their extension to computations with Vandermonde and Cauchy matrices linked to multipoint polynomial and rational evaluation and interpolation. For the solution of these tasks the arithmetic time of the known numerical approximation algorithms was quadratic (cf. [BF00], [BEGO08], and [KZ08]), and we decrease it to nearly linear by extending the power of the cited algorithms of [MRT05], [CGS07], [XXG12], and [XXCBa] from the case of Toeplitz and Hankel inputs. We note, however, that in our case the solution of linear systems of equations as well as the interpolation problems are ill conditioned and thus unfavorable to approximate numerical solution, except for some important special cases.

Our extensions may be surprising because we cover a substantially wider class of Cauchy matrices, which are unitary equivalent to the class of our inputs. Technically, as in the papers [MRT05], [CGS07], [XXG12], and [XXCBa], we rely on the approximation of these Cauchy matrices by HSS matrices and exploit the HSS matrix structure. As in these papers our basic computational blocks are the well studied numerically stable FFT and FMM algorithms that have been highly efficiently implemented on both serial and parallel computers [GS66], [B99], [BY13].

Unlike the cited papers, however, we treat a large subclass of Cauchy matrices $C = (\frac{1}{s_i - t_j})_{i,j=0}^{n-1}$ (we call them CV matrices because they are linked to Vandermonde matrices via FFT-based unitary transformations) rather than just the single CV matrix involved in the fast Toeplitz solvers. For that matrix $\{s_0, \dots, s_{n-1}\}$ is the set of the nth roots of unity, and $\{t_0, \dots, t_{n-1}\}$ is the set of the other $(2n)$th roots of unity, whereas for a CV matrix C, $\{s_0, \dots, s_{n-1}\}$ is an unrestricted set of n knots, and only the knots $\{t_0, \dots, t_{n-1}\}$ are equally spaced on the unit circle. The latter property of the knots still enables us to simplify the HSS approximation of CV matrices (by exploiting a partition of the complex plane into sectors sharing the origin 0). We supply a detailed complexity analysis of the approximation by HSS matrices and their compression, covering the harder case where the diagonal blocks are rectangular and have row indices that pairwise overlap. Our study can be technically interesting, it provides new insights and background for further progress in [Pa] where our present algorithms have been extended to various other classes of structured matrices and have been accelerated a little further in the case of Toeplitz and Hankel inputs.

We refer the reader to the papers and books [GKK85],[DV98], [T00], [EGHa], [EGHb], [VVGM05], [MRT05], [CDG06], [CGS07], [VVM07], [VVM08], [X12],

[XXG12], [X13], [XXCBa], [B10], [BY13], [GR87], [DGR96], [CGR98], [LRT79], [P93], [PR93], and the bibliography therein on the study of the FMM, HSS matrices, and the Matrix Compression (e.g., Nested Dissection) Algorithms.

We organize our paper as follows. In the next section we recall some definitions and basic results on computations with general matrices. In Section 3 we state the evaluation and interpolation tasks for polynomials and rational functions and link to each other these tasks and the computations with Vandermonde and Cauchy matrices. In Section 4 we recall the definition and basic properties of HSS matrices and extend the known results on HSS matrix computations to cover subsequent application to CV matrices. In Section 5 we prove the claimed results of Section 4 by extending the algorithms of [MRT05], [CGS07], [XXG12] and [XXCBa] and elaborate upon the complexity analysis, not completed in these papers. In Section 6 we outline a further extension of the class of HSS matrices that we employ in Section 7, where we treat CV matrices efficiently based on computing their compressed approximations of HSS type. In Section 8 we extend our fast approximation algorithms to computations with Vandermonde matrices and their transposes as well as with polynomials and rational functions, and we recall further extensions from [Pa]. In Section 9 we summarize our study. The Appendix includes figures and legends.

2 Definitions and auxiliary results

We measure the arithmetic cost by the number of arithmetic operations involved in the computations and performed in the field $\mathbb{C}$ of complex numbers with no error. $|\mathcal{S}|$ denotes the cardinality of a set $\mathcal{S}$.

$M = (m_{i,j})_{i,j=0}^{m-1,n-1}$ is an $m \times n$ matrix. M^T is its transpose, M^H is its Hermitian transpose. $\mathcal{C}(B)$ and $\mathcal{R}(B)$ are the index sets of the rows and columns of its submatrix B, respectively. For two sets $\mathcal{I} \subseteq \{1, \dots, m\}$ and $\mathcal{J} \subseteq \{1, \dots, n\}$ define the submatrix $M(\mathcal{I}, \mathcal{J}) = (m_{i,j})_{i \in \mathcal{I}, j \in \mathcal{J}}$. $\mathcal{R}(B) = \mathcal{I}$ and $\mathcal{C}(B) = \mathcal{J}$ if and only if $B = M(\mathcal{I}, \mathcal{J})$. Write $M(\mathcal{I}, .) = M(\mathcal{I}, \mathcal{J})$ where $\mathcal{J} = \{1, \dots, n\}$. Write $M(., \mathcal{J}) = M(\mathcal{I}, \mathcal{J})$ where $\mathcal{I} = \{1, \dots, m\}$. $(B_0 \ \dots \ B_{k-1})$ and $(B_0 \mid \dots \mid B_{k-1})$ denote a $1 \times k$ block matrix with k blocks $B_0, \dots, B_{k-1}$, whereas $\mathrm{diag}(B_0, \dots, B_{k-1}) = \mathrm{diag}(B_j)_{j=0}^{k-1}$ is a $k \times k$ block diagonal matrix with k diagonal blocks $B_0, \dots, B_{k-1}$, possibly rectangular. $O = O_{m,n}$ is the $m \times n$ matrix filled with zeros. $I = I_n$ is the $n \times n$ identity matrix. M is a $k \times l$ *unitary* matrix if $M^H M = I_l$ or $MM^H = I_k$. An $m \times n$ matrix M has a nonunique *generating pair* (F, G^T) *of a length* ρ if $M = FG^T$ for two matrices $F \in \mathbb{C}^{m \times \rho}$ and $G \in \mathbb{C}^{n \times \rho}$. The rank of a matrix is the minimum length of its generating pairs. An $m \times n$ matrix is *regular* or nonsingular if it has full rank $\min\{m, n\}$.

Theorem 1. *A matrix M has a rank at least ρ if and only if it has a nonsingular $\rho \times \rho$ submatrix $M(\mathcal{I}, \mathcal{J})$, and if so, then $M = M(., \mathcal{J})M(\mathcal{I}, \mathcal{J})^{-1}M(\mathcal{I}, .)$.*

The theorem defines a *generating triple* $(M(., \mathcal{J}), M(\mathcal{I}, \mathcal{J})^{-1}, M(\mathcal{I}, .))$ and two generating pairs $(M(., \mathcal{J}), M(\mathcal{I}, \mathcal{J})^{-1}M(\mathcal{I}, .)$ and $(M(., \mathcal{J})M(\mathcal{I}, \mathcal{J})^{-1}, M(\mathcal{I}, .)$ for a matrix M of a length ρ. We call such pairs and triples *generators*. One

can obtain some generating triples of the minimum length for a given matrix by computing its SVD or its less costly rank revealing factorizations such as ULV and URV factorizations in [CGS07], [XXG12], and [XXCBa], where the factors are unitary, diagonal or triangular.

$\alpha(M)$ and $\beta(M)$ denote the arithmetic cost of computing the vectors $M\mathbf{u}$ and $M^{-1}\mathbf{u}$, respectively, maximized over all unit vectors $\mathbf{u}$ and minimized over all algorithms, and we write $\beta(M) = \infty$ where the matrix M is singular. The straightforward algorithm supports the following bound.

Theorem 2. *$\alpha(M) \le 2(m+n)\rho - \rho - m$ for an $m \times n$ matrix M given with its generating pair of a length ρ.*

$||M|| = ||M||_2$ denotes the spectral norm of an $m \times n$ matrix $M = (m_{i,j})_{i,j=0}^{m-1,n-1}$, and we also write $|M| = \max_{i,j} |m_{i,j}|$, $||M|| \le \sqrt{mn}|M|$. It holds that $||U|| = 1$ and $||MU|| = ||UM|| = ||M||$ for a unitary matrix U. A vector $\mathbf{u}$ is unitary if and only if $||\mathbf{u}|| = 1$, and if this holds we call it a *unit vector*. A matrix $\tilde{M}$ is an *ϵ-approximation* of a matrix M if $|\tilde{M} - M| \le \epsilon$. The *$\epsilon$-rank* of a matrix M denotes the integer $\min_{|\tilde{M}-M| \le \epsilon} \text{rank}(\tilde{M})$. An *$\epsilon$-basis* for a linear space $\mathbb{S}$ of dimension k is a set of vectors that ϵ-approximate the k vectors of a basis for this space. An ϵ-generator of a matrix is a generator of its ϵ-approximation. $\alpha_\epsilon(M)$ and $\beta_\epsilon(M)$ replace the bounds $\alpha(M)$ and $\beta(M)$ where we ϵ-approximate the vectors $M\mathbf{u}$ and $M^{-1}\mathbf{u}$ instead of evaluating them. The *numerical rank* of a matrix M, which we denote $\text{nrank}(M)$, is its ϵ-rank for a small ϵ. A matrix M is *ill conditioned* if its rank exceeds its numerical rank.

Theorem 3. *(See [S98, Corollary 1.4.19] for $P = -M^{-1}E$.) Suppose M and $M + E$ are two nonsingular matrices of the same size and $||M^{-1}E|| = \theta < 1$. Then $||I - (M+E)^{-1}M|| \le \frac{\theta}{1-\theta}$ and $||(M+E)^{-1} - M^{-1}|| \le \frac{\theta}{1-\theta}||M^{-1}||$. In particular $||(M+E)^{-1} - M^{-1}|| \le 1.5\ \theta\ ||M^{-1}||$ if $\theta \le 1/3$.*

3 Polynomial and rational evaluation and interpolation and structured matrices

Write $T = (t_{i-j})_{i,j=0}^{m-1,n-1}$, $H = (h_{i+j})_{i,j=0}^{m-1,n-1}$, $V = V_{\mathbf{s}} = (s_i^j)_{i,j=0}^{m-1,n-1}$, and $C = C_{\mathbf{s},\mathbf{t}} = \left(\frac{1}{s_i - t_j}\right)_{i,j=0}^{m-1,n-1}$ to denote *$m \times n$ Toeplitz, Hankel, Vandermonde, and Cauchy* matrices, respectively, which are four classes of highly popular structured matrices, each having mn entries defined by at most $m + n$ parameters.

Remark 1. The four matrix structures have quite distinct features. The matrix structure of Cauchy type is invariant in row and column interchange (in contrast to the structures of Toeplitz and Hankel types) and can be closely approximated by HSS matrices (unlike the structures of the three other types). The paper [P90], however, linked the four structures to each other by means of structured matrix multiplication and *proposed to exploit this link in order to extend any successful matrix inversion algorithm for the matrices of any of the four classes to the matrices of the three other classes.*

Problem 1. Multipoint polynomial evaluation or Vandermonde-by-vector multiplication.
INPUT: $m+n$ complex scalars $p_0,\dots,p_{n-1}; s_0,\dots,s_{m-1}$.
OUTPUT: n complex scalars $v_0,\dots,v_{m-1}$ satisfying

$$v_i = p(s_i) \text{ for } p(x) = p_0 + p_1 x + \cdots + p_{n-1}x^{n-1} \text{ and } i = 0,\dots,m-1 \tag{1}$$

or equivalently

$$V\mathbf{p} = \mathbf{v} \text{ for } V = V_{\mathbf{s}} = (s_i^j)_{i,j=0}^{m-1,n-1},\ \mathbf{p} = (p_j)_{j=0}^{n-1},\ \text{and } \mathbf{v} = (v_i)_{i=0}^{m-1}. \tag{2}$$

Problem 2. Polynomial interpolation or the solution of a Vandermonde linear system of equations.
INPUT: $2n$ complex scalars $v_0,\dots,v_{n-1}; s_0,\dots,s_{n-1}$, the last n of them distinct.
OUTPUT: n complex scalars $p_0,\dots,p_{n-1}$ satisfying equations (1) and (2) for $m=n$.

Problem 3. Multipoint rational evaluation or Cauchy-by-vector multiplication.
INPUT: $2m+n$ complex scalars $s_0,\dots,s_{m-1}; t_0,\dots,t_{n-1}; v_0,\dots,v_{m-1}$.
OUTPUT: m complex scalars $v_0,\dots,v_{m-1}$ satisfying

$$v_i = \sum_{j=0}^{n-1} \frac{u_j}{s_i - t_j} \text{ for } i = 0,\dots,m-1 \tag{3}$$

or equivalently

$$C\mathbf{u} = \mathbf{v} \text{ for } C = C_{\mathbf{s},\mathbf{t}} = \Big(\frac{1}{s_i - t_j}\Big)_{i,j=0}^{m-1,n-1},\ \mathbf{u} = (u_j)_{j=0}^{n-1},\ \text{and } \mathbf{v} = (v_i)_{i=0}^{m-1}. \tag{4}$$

Problem 4. Rational interpolation or the solution of a Cauchy linear system of equations.
INPUT: $3n$ complex scalars $s_0,\dots,s_{n-1}; t_0,\dots,t_{n-1}; v_0,\dots,v_{n-1}$, the first $2n$ of them distinct.
OUTPUT: n complex scalars $u_0,\dots,u_{n-1}$ satisfying equations (3) and (4) for $m=n$.

The $m+n$ scalars $s_0,\dots,s_{n-1},t_0,\dots,t_{n-1}$ define the Vandermonde and Cauchy matrices $V_{\mathbf{s}}$ and $C_{\mathbf{s},\mathbf{t}}$ and are basic for Problems 1–4. Hereafter we call these scalars *knots*.

Theorem 4. *(i) An $m\times n$ Vandermonde matrix $V_{\mathbf{s}} = (s_i^j)_{i,j=0}^{m-1,n-1}$ has full rank if and only if all m knots $s_0,\dots,s_{m-1}$ are distinct. (ii) An $m\times n$ Cauchy matrix $C_{\mathbf{s},\mathbf{t}} = \Big(\frac{1}{s_i-t_j}\Big)_{i,j=0}^{m-1,n-1}$ is well defined if and only if its two knot sets $s_0,\dots,s_{m-1}$ and $t_0,\dots,t_{n-1}$ share no elements. (iii) If this matrix is well defined, then it has full rank if and only if all its $m+n$ knots $s_0,\dots,s_{m-1},t_0,\dots,t_{n-1}$ are distinct and also (iv) if and only if all its submatrices have full rank.*

Proof. Parts (i)–(iii) are implied by the following equations of independent interest (see, e.g., [P01, Section 3.6]),

$$\det V_{\mathbf{s}} = \prod_{i>j}(s_i - s_j), \quad \det C_{\mathbf{s},\mathbf{t}} = \prod_{i<j}(s_j - s_i)(t_i - t_j)/\prod_{i,j}(s_i - t_j). \tag{5}$$

Part (iv) follows from part part (iii) and the observation that every submatrix of a Cauchy matrix is a Cauchy matrix itself.

How many arithmetic operations do we need for solving Problems 1–4? The algorithms of [F72], [GGS87], and [MB72] solve Problems 1–3 by using $O((m+n)\log^2(n)\log(\log(n)))$ arithmetic operations over any field of constants. For $m \geq n$ this is within a factor of $\log(n)\log(\log(n))$ from the optimum [S73], [B-O83]. Equation (6) of this subsection extends the latter upper bound to Problem 4. For numerical solution of Problems 1–4, however, the users employ quadratic time algorithms to avoid error propagation (cf. [KZ08], [BF00], [P64], [BP70], [BEGO08]), in spite of substantial research progress reported in [PRT92], [PSLT93], [P95], [PZHY97], and particularly [DGR96].

We can solve Problems 1–4 numerically by using $O(n\log(n))$ arithmetic operations in the important special case where the knots $s_i = \omega^i$ are the nth roots of 1, $\omega = \omega_n = \exp(2\pi\sqrt{-1}/n)$, $i = 0, \dots, n-1$, and $V_{\mathbf{s}} = (\omega^{ij})_{i,j=0}^{n-1}$, and hereafter we write $\Omega = \frac{1}{\sqrt{n}}(\omega^{ij})_{i,j=0}^{n-1}$. In this case Problems 1 and 2 turn into the computational problems of the forward and inverse discrete Fourier transforms (hereafter *DFT* and *IDFT*). The *FFT* (Fast Fourier transform) and Inverse FFT are two numerically stable algorithms that perform DFT and IDFT at the arithmetic cost $1.5n\log_2(n)$ and $1.5n\log_2(n)+n$, respectively, if $m = n$ is a power of 2 (cf. [BP94, Sections 1.2 and 3.4]), whereas Generalized FFT and Generalized Inverse FFT use $O(n\log(n))$ arithmetic operations to perform DFT and IDFT for any n [P01, Problem 2.4.2]. Note that $\Omega^H\Omega = I_n$, that is $\Omega = \Omega^T$ and $\Omega^H = \Omega^{-1} = \frac{1}{\sqrt{n}}(\omega^{-ij})_{i,j=0}^{n-1}$ are unitary matrices. The following equation links Problems 1 and 2 to Cauchy matrix computations (cf. [P01, Section 3.6]),

$$C_{\mathbf{s},\mathbf{t}} = \operatorname{diag}(t(s_i)^{-1})_{i=0}^{m-1} V_{\mathbf{s}} V_{\mathbf{t}}^{-1} \operatorname{diag}(t'(t_j))_{j=0}^{n-1}. \tag{6}$$

Remark 2. One can compute the values $v(t_0) = -t_0^n \dots, v(t_{n-1}) = -t_{n-1}^n$ of the polynomial $v(x) = t(x) - x^m$ by using $O(n\log n)$ arithmetic operations, and then the problem of computing the coefficients of this polynomial when we are given its values $v(t_i) = -t_i^n$ for $i = 0, \dots, n-1$ turns into a special case of Problem 2 of polynomial interpolation.

For $\mathbf{t} = (f\omega^j)_{j=0}^{n-1}$, the knots t_j are the nth roots of 1 scaled by f, $t(x) = x^n - f^n$, $t'(x) = nx^{n-1}$, $V_{\mathbf{t}} = \sqrt{n}\Omega\operatorname{diag}(f^j)_{j=0}^{n-1}$, $V_{\mathbf{t}}^{-1} = \frac{1}{\sqrt{n}}\operatorname{diag}(f^{-j})_{j=0}^{n-1}\Omega^H$, and we deduce from equation (6) that

$$C_{\mathbf{s},f} = \sqrt{n}\operatorname{diag}\Big(\frac{f^{n-1}}{s_i^n - f^n}\Big)_{i=0}^{m-1} V_{\mathbf{s}} \operatorname{diag}(f^{-j})_{j=0}^{n-1}\Omega^H \operatorname{diag}(\omega^{-j})_{j=0}^{n-1}, \tag{7}$$

$$V_{\mathbf{s}} = \frac{f^{1-n}}{\sqrt{n}} \operatorname{diag}\left(s_i^n - f^n\right)_{i=0}^{m-1} C_{\mathbf{s},f} \operatorname{diag}(\omega^j)_{j=0}^{n-1} \Omega \operatorname{diag}(f^j)_{j=0}^{n-1}, \tag{8}$$

$$V_{\mathbf{s}}^T = -\frac{f^{1-n}}{\sqrt{n}} \operatorname{diag}(f^j)_{j=0}^{n-1} \Omega \operatorname{diag}(\omega^j)_{j=0}^{n-1} C_{f,\mathbf{s}} \operatorname{diag}(s_i^n - f^n)_{i=0}^{m-1}, \tag{9}$$

and for $m = n$ also

$$V_{\mathbf{s}}^{-1} = \sqrt{n} \operatorname{diag}(f^{-j})_{j=0}^{n-1} \Omega^H \operatorname{diag}(\omega^{-j})_{j=0}^{n-1} C_{\mathbf{s},f}^{-1} \operatorname{diag}\left(\frac{f^{n-1}}{s_i^n - f^n}\right)_{i=0}^{n-1}, \tag{10}$$

$$V_{\mathbf{s}}^{-T} = -\sqrt{n} \operatorname{diag}\left(\frac{f^{n-1}}{s_i^n - f^n}\right)_{i=0}^{n-1} C_{f,\mathbf{s}}^{-1} \operatorname{diag}(\omega^{-j})_{j=0}^{n-1} \Omega^H \operatorname{diag}(f^{-j})_{j=0}^{n-1}. \tag{11}$$

The latter equations link Vandermonde matrices and their inverses to the $m \times n$ Cauchy matrices with the knot set $\mathcal{T} = \{t_j = f\omega^j,\ j = 0, \dots, n-1\}$ (for $f \neq 0$), which we call *CV matrices* and denote $C_{\mathbf{s},f}$. The $n \times m$ matrices $C_{e,\mathbf{t}} = -C_{\mathbf{t},e}^T = \left(\frac{1}{\omega^i - t_j}\right)_{i,j=0}^{n-1,m-1}$ for $e \neq 0$ have the knot set $\mathcal{S} = \{s_i = e\omega^i,\ i = 0, \dots, n-1\}$, are linked to transposed Vandermonde matrices, and are said to be *CV^T matrices*. Note that

$$C_{a\mathbf{s},a\mathbf{t}} = \frac{1}{a} C_{\mathbf{s},\mathbf{t}} \text{ and } C_{\mathbf{s}+a\mathbf{e},\mathbf{t}+a\mathbf{e}} = C_{\mathbf{s},\mathbf{t}} \text{ for } a \neq 0 \text{ and } \mathbf{e} = (1, \dots, 1)^T, \tag{12}$$

and so we can define a Cauchy matrix up to shift and scaling by constants and then choose $e = f = 1$ in the above expressions.

Remark 3. By linking together the Vandermonde and Cauchy matrix structures, equations (6)–(11) also link Problems 1 and 2 to Problems 3 and 4. To establish the latter links directly, assume $p(x)$ of equation (1), $t(x) = \prod_{j=0}^{n-1}(x - t_j)$, and n distinct knots $t_0, \dots, t_{n-1}$ and then represent the rational function $v(x) = \frac{p(x)}{t(x)}$ as $v(x) = \sum_{j=0}^{n-1} \frac{u_j}{x - t_j}$. We obtain equations (3) by writing $v_i = v(s_i)$ for $i = 0, \dots, m-1$.

Theorem 5 below as well as [P01, Equation (3.4.1)] link together a Vandermonde matrix, its transpose, inverse and the inverse of the transpose. [P01, Sections 4.7 and 4.8] and [Pa] cover more comprehensively such links as well as the links between the computations with structured matrices and polynomials, exemplified by the links between Problems 1–4 and the problems of multiplication of Vandermonde and Cauchy matrices and their inverses by a vector.

Theorem 5. *(i) JH and HJ are Toeplitz matrices if H is a Hankel matrix, and vice versa. (ii) $H = V^T V = (\sum_{k=0}^{m-1} s_k^{i+j})_{i,j=0}^{n-1}$ is a Hankel matrix for any $m \times n$ Vandermonde matrix $V = (s_i^j)_{i,j=0}^{m-1,n-1}$.*

4 Quasiseparable and HSS matrices

Next we define HSS matrices and study their multiplication by vectors and the solution of nonsingular HSS linear systems of equations.

4.1 Quasiseparable matrices and generators

Definition 1. *A matrix given with its block diagonal is* (l, u)-quasiseparable *if* l *and* u *are the maximum ranks of its sub- and superdiagonal blocks, respectively. By replacing ranks with* ϵ*-ranks we define* (ϵ, l, u)-quasiseparable *matrices.*

In particular the matrices with a lower bandwidth l and an upper bandwidth u as well as their inverses (if defined) are (l, u)-quasiseparable.

Theorem 6. *[DV98], [EG02]. Suppose that an* (l, u)*-quasiseparable matrix* M *is given with* $m_q \times n_q$ *diagonal blocks* Σ_q, $q = 0, \dots, k-1$, *such that* $\sum_{q=0}^{k-1} m_q = m$, $\sum_{q=0}^{k-1} n_q = n$, *and* $s = \sum_{q=0}^{k-1} m_q n_q = O((l+u)(m+n))$. *Then*

$$\alpha(M) \le 2\sum_{q=0}^{k-1}((m_q + n_q)(l+u) + s) + 2l^2k + 2u^2k = O((l+u)(m+n)).$$

Furthermore if $m_q = n_q$ *for all* q *and if the matrix* M *is nonsingular, then*

$$\beta(M) = O(\sum_{q=0}^{k-1}((l+u)^2(l+u+n_q)n_q + n_q^3)).$$

The algorithms of [DV98], [EG02] supporting the theorem as well as the study in [CGS07], [VVM07], [VVM08], [XXG12], [EGHa] and [EGHb] rely on the representation of (l, u)-quasiseparable matrices with *quasiseparable generators*, demonstrated by the following 4×4 example and defined in Theorem 7,

$$M = \begin{pmatrix} \Sigma_0 & S_0T_1 & S_0B_1T_2 & S_0B_1B_2T_3 \\ P_1Q_0 & \Sigma_1 & S_1T_2 & S_1B_2T_3 \\ P_2A_1Q_0 & P_2Q_1 & \Sigma_2 & S_2T_3 \\ P_3A_2A_1Q_0 & P_3A_2Q_1 & P_3Q_2 & \Sigma_3 \end{pmatrix}. \tag{13}$$

Note that M is a block tridiagonal matrix where $A_p = B_q = O$ for all p and q.

Theorem 7. *(Cf. [EGHa], [VVM07], [X12], the bibliography therein, and our Table 1.) Assume a* $k \times k$ *matrix* M *with a block diagonal* $\widehat{\Sigma} = (\Sigma_0, \dots, \Sigma_{k-1})$, *where* $\Sigma_q = M(I_q, J_q)$, $q = 0, \dots, k-1$. *Then* M *is an* (l, u)*-quasiseparable matrix if and only if there exists a nonunique family of* quasiseparable generators $\{P_i, Q_h, S_h, T_i, A_g, B_g\}$ *such that*

$$M(\mathcal{I}_i, \mathcal{J}_h) = P_i A_{i-1} \cdots A_{h+1} Q_h \text{ and } M(\mathcal{I}_h, \mathcal{J}_i) = S_h B_{h+1} \cdots B_{i-1} T_i$$

for $0 \le h < i < k$. *Here* P_i, Q_h *and* A_g *are* $|\mathcal{I}_i| \times l_i$, $l_{h+1} \times |\mathcal{J}_h|$, *and* $l_{g+1} \times l_g$ *matrices, respectively, whereas* S_h, T_i *and* B_g *are* $|\mathcal{I}_h| \times u_{h+1}$, $u_i \times |J_i|$, *and* $u_g \times u_{g+1}$ *matrices, respectively,* $g = 1, \dots, k-2$, $h = 0, \dots, k-2$, $i = 1, \dots, k-1$, *and the integers* $l = \max_g\{l_g\}$ *and* $u = \max_h\{u_h\}$ *are called the* lower and upper lengths *or* orders *of the quasiseparable generators.*

Table 1. The sizes of quasiseparable generators of Theorem 7

P_i	Q_h	A_g	S_h	T_i	B_g
$\|\mathcal{I}_i\| \times l_i$	$l_{h+1} \times \|\mathcal{J}_h\|$	$l_{g+1} \times l_g$	$\|\mathcal{I}_h\| \times u_{h+1}$	$u_i \times \|J_i\|$	$u_g \times u_{g+1}$

By virtue of this theorem one can redefine the (l, u)-quasiseparable matrices as the ones allowing representation with the families $\{P_h,\ Q_i,\ A_g\}$ and $\{S_h,\ T_i,\ B_g\}$ of quasiseparable generators having lower and upper orders l and u, respectively. Definition 1 and Theorem 7 provide two useful insights into the properties of (l, u)-quasiseparable matrices. In the next subsections we recall and employ the third equivalent definition, providing yet another insight and applied to the study of the $n \times n$ Cauchy matrix $C_{1,\omega_{2n}}$ in the papers [CGS07], [XXG12], [XXCBa].

4.2 Recursive merging of diagonal blocks of a matrix

Definition 2. *Assume a* $1 \times k$ *block matrix* $M = (M_0\ \ldots\ M_{k-1})$ *with* k *block columns* M_q*, each partitioned into a diagonal block* Σ_q *and a* basic neutered block column N_q*,* $q = 0, \ldots, k-1$ *(cf. our Figures 2–4 and [MRT05, Section 1]).* Merge *the* l *basic block columns* $M_{q_0}, \ldots, M_{q_{l-1}}$*, the* l *diagonal blocks* $\Sigma_{q_0}, \ldots, \Sigma_{q_{l-1}}$*, and the* l *basic neutered block columns* $N_{q_0}, \ldots, N_{q_{l-1}}$ *into their* union $M_{q_0,\ldots,q_{l-1}} = M(., \cup_{j=0}^{l-1} \mathcal{C}(\Sigma_{q_j}))$*, their* diagonal union $\Sigma_{q_0,\ldots,q_{l-1}}$*, and their* neutered union $N_{q_0,\ldots,q_{l-1}}$*, respectively, such that* $\mathcal{R}(\Sigma_{q_0,\ldots,q_{l-1}}) = \cup_{j=0}^{l-1} \mathcal{R}(\Sigma_{q_j})$ *and the block column* $M_{q_0,\ldots,q_{l-1}}$ *is partitioned into the diagonal union* $\Sigma_{q_0,\ldots,q_{l-1}}$ *and the neutered union* $N_{q_0,\ldots,q_{l-1}}$*.*

The complete binary tree of Figure 1 represents *recursive merging* of eight diagonal blocks $\Sigma_0, \Sigma_1, \ldots, \Sigma_7$ at first into the four diagonal unions of the four pairs $\Sigma_{0,1} = \Sigma(\Sigma_0, \Sigma_1), \ldots,\ \Sigma_{6,7} = \Sigma(\Sigma_6, \Sigma_7)$, then into the two diagonal unions of two quadruples

$$\Sigma_{0,1,2,3} = \Sigma(\Sigma_{0,1}, \Sigma_{2,3}) = \Sigma(\Sigma_0, \Sigma_1, \Sigma_2, \Sigma_3),$$

$$\Sigma_{4,5,6,7} = \Sigma(\Sigma_{4,5}, \Sigma_{6,7}) = \Sigma(\Sigma_4, \Sigma_5, \Sigma_6, \Sigma_7),$$

and finally into the diagonal union of the single 8-tuple,

$$\Sigma_{0,1,2,3,4,5,6,7} = \Sigma(\Sigma_{0,1,2,3}, \Sigma_{4,5,6,7}) = \Sigma(\Sigma_0, \Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4, \Sigma_5, \Sigma_6, \Sigma_7).$$

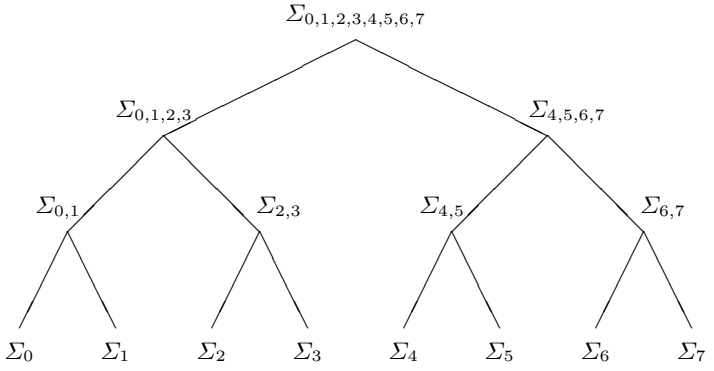
Appropriate processes of recursive merging can produce the diagonal union of any fixed set of diagonal blocks.

Every merging is represented by a binary tree. Let $L(v)$ and $R(v)$ denote the two sets of leaves that are the left and right descendants of a vertex v of the tree, respectively. Then for any leaf set define a unique *balanced binary tree* with the minimum number of edges such that $0 \le |L(v)| - |R(v)| \le 1$ for all its nodes v.

Such a tree with n leaves can be uniquely defined by removing $2^{l(n)}-n$ properly chosen leaves from the leaf level of the complete binary tree with $2^{\widehat{l}}$ leaves for $l(n)=\lceil \log_2(n) \rceil$. Alternatively we can remove the $2^{l(n)}-n$ rightmost leaves and arrive at the so called heap structure with n leaves. In both ways we define a unique tree whose all leaves lie in its two lowest levels.

For every set of diagonal blocks, either of the trees identifies a unique process of their merging (such as the one of Figure 1). Hereafter we refer to such a process identified by the balanced binary tree as *balanced merging*, although in our application to computations with Cauchy matrices we only use processes identified with complete binary trees.

Fig. 1. Balanced merging of diagonal blocks.



4.3 HSS and balanced HSS matrices, their link to quasiseparable matrices, and the cost of basic operations with them

Definition 3. *A matrix given with its block diagonal is* basically ρ-neutered *if all its basic neutered block columns have ranks at most ρ. By replacing ranks with ϵ-ranks we define basically (ϵ,ρ)-neutered matrices.*

Theorem 8. *Given the block diagonal matrix $\Sigma=\mathrm{diag}(\Sigma_q)_{q=0}^{k-1}$ and k generators $(F_0,G_0),\dots,(F_{k-1},G_{k-1})$ of lengths at most ρ for the k basic neutered block columns of an $m\times n$ matrix M, it holds that*

$$\alpha(M)\le\alpha(\Sigma)+(2m+2n-1)k\rho.$$

Proof. Write $M=M'+\mathrm{diag}(\Sigma_q)_{q=0}^{k-1}$. Note that $\alpha(M)\le\alpha(\Sigma)+\alpha(M')+m$. The basic neutered block columns of the matrix M share their entries with the matrix M', whose other entries are zeros. So the k pairs $(F_0,G_0),\dots,(F_{k-1},G_{k-1})$

combined form a single generating pair of a length at most $k\rho$ for the matrix M'. Therefore $\alpha(M') \le (2m+2n-1)k\rho - m$ by virtue of Theorem 2.

Basically ρ-neutered matrices are precisely the input class of Theorem 8, whose cost estimates are weaker than in Theorem 6. By adding row-wise links among the basic neutered block columns of a basically ρ-neutered matrix we can turn it into (ρ,ρ)-quasiseparable, as we show next (see Theorem 9).

Definition 4. *(i) A matrix given with its block diagonal is a* balanced ρ-HSS *matrix if it is basically ρ-neutered throughout the process of balanced merging of its diagonal blocks, that is if all neutered unions of its basic neutered block columns involved into this process have ranks at most ρ. (ii) This is a* ρ-HSS *matrix if it is basically ρ-neutered throughout any process of recursive merging of its diagonal blocks. (iii) By replacing ranks with ϵ-ranks we define* balanced (ϵ,ρ)-HSS matrices *and* (ϵ,ρ)-HSS matrices.

Theorem 9. *(i) Every (l,u)-quasiseparable matrix M is an $(l+u)$-HSS matrix. (ii) Every ρ-HSS matrix is (ρ,ρ)-quasiseparable.*

Proof. A basic block neutered column N_q of a matrix can be partitioned into its basic block sub- and superdiagonal parts L_q and U_q, respectively, and so $\operatorname{rank}(N_q) \le \operatorname{rank}(L_q) + \operatorname{rank}(U_q)$, which implies that $\operatorname{rank}(N_q) \le l+u$ for $q = 0,\dots,k-1$ if the matrix M is (l,u)-quasiseparable. This proves part (i). Next note that the union N of any set of basic neutered block columns of a matrix M can be turned into a basic neutered block column at some stage of an appropriate process of recursive merging. Therefore $\operatorname{rank}(N) \le \rho$ where M is a ρ-HSS matrix. Now for every off-diagonal block B of a matrix M define the set of its basic neutered block columns that share some column indices with the block B and then note that the block B is a submatrix of the neutered union of this set. Therefore $\operatorname{rank}(B) \le \operatorname{rank}(N) \le \rho$, and we obtain part (ii).

By combining Theorems 6 and 9 we obtain the following results.

Corollary 1. *Assume a ρ-HSS matrix M given with $m_q \times n_q$ diagonal blocks Σ_q, $q = 0,\dots,k-1$, and write $m = \sum_{q=0}^{k-1} m_q$, $n = \sum_{q=0}^{k-1} n_q$, and $s = \sum_{q=0}^{k-1} m_q n_q$. Then $\alpha(M) < 2s + 4\rho^2 k + 4\sum_{q=0}^{k-1}(m_q+n_q)\rho = O((m+n)\rho+s)$. Furthermore if $m_q = n_q$ for all q and if the matrix M is nonsingular, then $\beta(M) = O(\sum_{q=0}^{k-1}((\rho+n_q)\rho^2 n_q + n_q^3))$.*

For a balanced ρ-HSS matrices M we only have a little weaker representation than in Theorem 7, and so the proof of the estimates of Corollary 1 for $\alpha(M)$ and $\beta(M)$ does not apply. Nevertheless our next theorem matches these two estimates up to logarithmic and constant factors, respectively. We obtain our bound on $\alpha(M)$ by means of a simple recursion and obtain our bound on $\beta(M)$ by analyzing the algorithms of [CGS07, Sections 3 and 4], [XXG12], and [XXCBa]. Unlike Theorem 6 and Corollary 1, we allow $m_q \ne n_q$ for all q.

Theorem 10. *Given a balanced ρ-HSS matrix M and its block diagonal matrix $\Sigma = \mathrm{diag}(\Sigma_q)_{q=0}^{k-1}$ with $m_q \times n_q$ blocks Σ_q, $q = 0, \dots, k-1$, and $s = \sum_{q=0}^{k-1} m_q n_q$ entries overall, write $l = \lceil \log_2(k) \rceil$, $m = \sum_{q=0}^{k-1} m_q$, $n = \sum_{q=0}^{k-1} n_q$, $m_+ = \max_{q=0}^{k-1} m_q$, $n_+ = \max_{q=0}^{k-1} n_q$, and $s = \sum_{q=0}^{k-1} m_q n_q$, $s \le \min\{m_+ n, m n_+\}$. Then it holds that*

$$\alpha(M) < 2s + (m + 2(m+n)\rho)l. \tag{14}$$

Moreover if $m = n$ and if the matrix M is nonsingular, then

$$\beta(M) = O(n_+ s + (n_+^2 + \rho n_+ + l\rho^2)n + (k\rho + n)\rho^2). \tag{15}$$

Furthermore the same bounds (14) and (15) hold where the matrix M is the transpose of a balanced ρ-HSS matrix with $n_q \times m_q$ diagonal blocks Σ_q for $q = 0, \dots, k-1$.

Corollary 2. *Suppose that under the assumptions of Theorem 10 it holds that $k\rho = O(n)$, and $n_+ + \rho = O(\log(n))$. Then $\alpha(M) = O((m+n)\log^2(n))$ and $\beta(M) = O(n \log^3(n))$.*

5 Proof of Theorem 10

5.1 A proof of bound (14)

With no loss of generality assume that the $(l-1)$st, final stage of a balanced merging process has produced a 2×2 block representation

$$M = \begin{pmatrix} \bar{\Sigma}_0^{(l)} & \bar{S}_{01}^{(l)} \bar{T}_1^{(l)} \\ \bar{S}_{10}^{(l)} \bar{T}_0^{(l)} & \bar{\Sigma}_1^{(l)} \end{pmatrix}$$

where $\bar{\Sigma}_j^{(l)}$ is an $\bar{m}_j^{(l)} \times \bar{n}_j^{(l)}$ matrix, $\bar{T}_j^{(l)}$ is an $\bar{n}_j^{(l)} \times \bar{\rho}_j^{(l)}$ matrix, $\bar{\rho}_j^{(l)} \le \rho$, $j = 0, 1$, $\bar{m}_1^{(l)} + \bar{m}_2^{(l)} = m$, and $\bar{n}_1^{(l)} + \bar{n}_2^{(l)} = n$.

Clearly $\alpha(M) \le m + \sum_{j=0}^{1} \alpha(\bar{\Sigma}_j^{(l)}) + \sum_{j=0}^{1} \alpha(\bar{T}_j^{(l)}) + \alpha(\bar{S}_{01}^{(l)}) + \alpha(\bar{S}_{10}^{(l)})$. Apply Theorem 2 and obtain that $\sum_{j=0}^{1} \alpha(\bar{T}_j^{(l)}) + \alpha(\bar{S}_{01}^{(l)}) + \alpha(\bar{S}_{10}^{(l)}) < 2(m+n)\rho$.

The second last stage of the balanced merging process produces a similar 2×2 block representation for each of the diagonal blocks $\bar{\Sigma}_j^{(l)}$, $j = 0, 1$, and therefore $\sum_{j=0}^{1} \alpha(\bar{\Sigma}_j^{(l)}) < m + 2(m+n)\rho + \sum_{j=0}^{k(1)} \alpha(\bar{\Sigma}_j^{(1-1)})$ where $\bar{\Sigma}_0^{(1-1)}, \dots, \bar{\Sigma}_{k(1)-1}^{(1-1)}$ are the diagonal blocks output at the second last merging stage. Recursively going back through the merging process we obtain that $\alpha(M) < (m + 2(m+n)\rho)l + \sum_{j=0}^{k-1} \alpha(\Sigma_j)$ where $\Sigma_q = \bar{\Sigma}_q^{(0)}$ is an $m_q \times n_q$ matrix for $m_q = \bar{m}_q^{(0)}$, $n_q = \bar{n}_q^{(0)}$, and $q = 0, \dots, k-1$. Consequently $\sum_{q=0}^{k-1} \alpha(\Sigma_q) < 2 \sum_{q=0}^{k-1} m_q n_q = 2s$, and we arrive at bound (14).

5.2 Some introductory comments for proving bound (15)

The algorithm of [CGS07, Section 3] factors an (l, u)-quasiseparable matrix M into the product of unitary and block triangular matrices. This enables unitary reduction of a nonsingular linear systems of equations $M\mathbf{y} = \mathbf{b}$ to triangular linear systems, and then one can compute the solution vector $\mathbf{y}$ in nearly linear arithmetic time. We apply the algorithm to a balanced ρ-HSS matrix M and obtain similar factorization and unitary reduction to deduce the cost bounds of Theorem 10. We rearrange the computations to facilitate the proof of the arithmetic cost estimates (not presented in [CGS07]). As in [CGS07, Section 3] we demonstrate the algorithm for a 4×4 block matrix, although instead of (l, u)-HSS matrix of (13) we work with a basically ρ-HSS matrix

$$M = \begin{pmatrix} \Sigma_0 & S_{01}T_1 & S_{02}B_{12}T_2 & S_{03}B_{13}B_{23}T_3 \\ S_{10}T_0 & \Sigma_1 & S_{12}T_2 & S_{13}B_{23}T_3 \\ S_{20}B_{20}T_0 & S_{21}T_1 & \Sigma_2 & S_{23}T_3 \\ S_{30}B_{20}B_{10}T_0 & S_{31}B_{32}T_1 & S_{23}T_2 & \Sigma_3 \end{pmatrix} \tag{16}$$

having $m_q \times n_q$ diagonal blocks Σ_q for any pairs $m_q \times n_q$ and $q = 0, 1, 2, 3$. For balanced ρ-HSS matrices M we could have written $B_{p,q} = I$ for all pairs of p and q, but we use expression (16) to simplify comparison with [CGS07]). As soon as we complete the description of the construction for $k = 4$, we outline its generalization to the case of any positive integer k.

5.3 Compression and merging stages

At first, for $k = 4$ and $q = 0, 1, 2, 3$, we compute the QR factors of the matrices T_q^H, that is compute square unitary matrices U_q (in factored form) and $\rho_q \times \widehat{n}_q$ matrices $\widehat{T}_q$ of full column ranks $\widehat{n}_q$ such that $T_q U_q^H = (O \mid \widehat{T}_q)$ and $\widehat{n}_q \le \rho_q \le \rho$. Write $\widehat{U} = \mathrm{diag}(U_q)_{q=0}^3$, $\widehat{M} = M\widehat{U}^H$, $M = \widehat{M}\widehat{U}$, and obtain

$$\widehat{M} = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} & O & S_{01}\widehat{T}_1 & O & S_{02}B_{12}\widehat{T}_2 & O & S_{03}B_{13}B_{23}\widehat{T}_3 \\ O & S_{10}\widehat{T}_0 & \Sigma_{10} & \Sigma_{11} & O & S_{12}\widehat{T}_2 & O & S_{13}B_{2,3}\widehat{T}_3 \\ O & S_{20}B_{20}\widehat{T}_0 & O & S_{21}\widehat{T}_1 & \Sigma_{20} & \Sigma_{21} & O & S_{23}\widehat{T}_3 \\ O & S_{30}B_{20}B_{10}\widehat{T}_0 & O & S_{31}B_{32}\widehat{T}_1 & O & S_{32}\widehat{T}_2 & \Sigma_{30} & \Sigma_{31} \end{pmatrix}.$$

Choose a permutation matrix P_0 such that $\widehat{M}P_0 = (\mathrm{diag}(\Sigma_{q0})_{q=0}^3 \mid M_1)$,

$$M_1 = \begin{pmatrix} \Sigma_{01} & S_{01}\widehat{T}_1 & S_{02}B_{12}\widehat{T}_2 & S_{03}B_{13}B_{23}\widehat{T}_3 \\ S_{10}\widehat{T}_0 & \Sigma_{11} & S_{12}\widehat{T}_2 & S_{13}B_{2,3}\widehat{T}_3 \\ S_{20}B_{20}\widehat{T}_0 & S_{21}\widehat{T}_1 & \Sigma_{21} & S_{23}\widehat{T}_3 \\ S_{30}B_{20}B_{10}\widehat{T}_0 & S_{31}B_{32}\widehat{T}_1 & S_{32}\widehat{T}_2 & \Sigma_{31} \end{pmatrix},$$

and the four diagonal blocks Σ_{q0} have sizes $m_q \times (n_q - \widehat{n}_q)$ for $q = 0, 1, 2, 3$. Note that M_1 is a basically ρ-HSS matrix. Write $M = M^{(0)}$, $\Sigma^{(0)} = \mathrm{diag}(\Sigma_{q0})_{q=0}^3$, and $U^{(0)} = \widehat{U}P_0$, and obtain that

$$M^{(0)} = (\Sigma^{(0)} \mid M_1)U^{(0)}. \tag{17}$$

By following [CGS07] we call the above computation of the matrices $U^{(0)}$, $\Sigma^{(0)}$ and M_1 the *compression* of the matrix M. For a fixed ρ we cannot compress the matrix M_1 any further because its every diagonal block Σ_{q0} has at most ρ columns. At this point (cf. [CGS07]) we *merge* pairwise the diagonal blocks Σ_{01}, Σ_{11}, Σ_{21} and Σ_{31} of the matrix M_1 into the diagonal unions of the two pairs, $\Sigma_0^{(1)} = \begin{pmatrix} \Sigma_0^{(1)} & \widehat{S}_{01}^{(1)} T_1^{(1)} \\ \widehat{S}_{10}^{(1)} T_0^{(1)} & \Sigma_1^{(1)} \end{pmatrix}$ and $\Sigma_1^{(1)} = \begin{pmatrix} \Sigma_2^{(1)} & \widehat{S}_{23}^{(1)} T_3^{(1)} \\ \widehat{S}_{32}^{(1)} T_2^{(1)} & \Sigma_3^{(1)} \end{pmatrix}$. By definition, merging preserves the property of being a basically ρ-HSS matrix, and we redefine M_1 as a 2×2 block matrix $M^{(1)} = \begin{pmatrix} \Sigma_0^{(1)} & \widehat{S}_{01}^{(1)} T_1^{(1)} \\ \widehat{S}_{10}^{(1)} T_0^{(1)} & \Sigma_1^{(1)} \end{pmatrix}$ where $\Sigma_q^{(1)}$ are $m_q^{(1)} \times n_q^{(1)}$ matrices, $\widehat{S}_{pq}^{(1)}$ are $m_p^{(1)} \times \rho_q^{(1)}$ matrices, $T_q^{(1)}$ are $\rho_q^{(1)} \times \bar{n}_q$ matrices, $m_q^{(1)} = m_{2q} + m_{2q+1}$, $\bar{n}_q = \widehat{n}_{2q} + \widehat{n}_{2q+1} \le 2\rho$, and $\rho_q^{(1)} \le \rho$ for $p, q \in \{0, 1\}$.

5.4 Recursive alternation of compression and merging

By following [CGS07, Section 3] we recursively alternate compression and merging, and next we compress the 2×2 block matrix $M^{(1)}$. We compute unitary matrices $U_0^{(1)}$ and $U_1^{(1)}$ (the Q factors) such that $T_q^{(1)}(U_q^{(1)})^H = (O \mid \widehat{T}_q^{(1)})$ and $\widehat{T}_q^{(1)}$ is an $n_q^{(1)} \times \rho_q^{(1)}$ matrix of full rank $n_q^{(1)}$ for $n_q^{(1)} \le \rho_q^{(1)} \le \rho$ and $q = 0, 1$. Then we write $\widehat{U}^{(1)} = \mathrm{diag}(U_0^{(1)}, U_1^{(1)})$ and obtain $M^{(1)} = \widehat{M}^{(1)} \widehat{U}^{(1)}$,

$$\widehat{M}^{(1)} = M^{(1)}(\widehat{U}^{(1)})^H = \begin{pmatrix} \Sigma_{00}^{(1)} & \Sigma_{01}^{(1)} & O & S_{01}^{(1)} \widehat{T}_1^{(1)} \\ O & S_{10}^{(1)} \widehat{T}_0^{(1)} & \Sigma_{10}^{(1)} & \Sigma_{11}^{(1)} \end{pmatrix}$$

and $\widehat{M}^{(1)} P_1 = \begin{pmatrix} \Sigma_{00}^{(1)} & O & \Sigma_{01}^{(1)} & S_{01}^{(1)} \widehat{T}_1^{(1)} \\ O & \Sigma_{10}^{(1)} & S_{10}^{(1)} \widehat{T}_0^{(1)} & \Sigma_{11}^{(1)} \end{pmatrix}$ for a permutation matrix P_1. Now write $\Sigma_q^{(1)} = \Sigma_{q0}^{(1)}$ for $q = 0, 1$, $\Sigma^{(1)} = \mathrm{diag}(\Sigma_q^{(1)})_{q=0}^{1}$, $U^{(1)} = \widehat{U}^{(1)} P_1$, and $M_2 = \begin{pmatrix} \Sigma_{01}^{(1)} & S_{01}^{(1)} \widehat{T}_1^{(1)} \\ S_{10}^{(1)} \widehat{T}_0^{(1)} & \Sigma_{11}^{(1)} \end{pmatrix}$ and obtain

$$M^{(1)} = (\Sigma^{(1)} \mid M_2) U^{(1)}. \tag{18}$$

We cannot compress the 2×2 block matrix M_2 any further because each of its diagonal blocks $\Sigma_{q1}^{(1)}$, $q = 0, 1$, has at most ρ columns. We merge these two blocks to rewrite M_2 as a 1×1 block matrix, to which we refer hereafter as $\Sigma^{(2)}$. Now we combine equations (17) and (18) and write $U = U^{(0)} \mathrm{diag}(I, U^{(1)})$ to obtain $M = M^{(0)} = DU$ for $D = (\Sigma^{(0)} \mid \Sigma^{(1)} \mid \Sigma^{(2)})$, $\Sigma^{(0)} = \mathrm{diag}(\Sigma_q^{(0)})_{q=0}^{3}$, $\Sigma^{(1)} = \mathrm{diag}(\Sigma_0^{(1)}, \Sigma_1^{(1)})$, and so

$$D = \left(\begin{array}{cccc|cc|c} \Sigma_0^{(0)} & & & & \Sigma_0^{(1)} & & \\ & \Sigma_1^{(0)} & & & & & \Sigma_0^{(2)} \\ & & \Sigma_2^{(0)} & & & \Sigma_1^{(1)} & \\ & & & \Sigma_3^{(0)} & & & \end{array}\right),$$

where (cf. (16)) $\Sigma_q^{(0)} = \Sigma_{q0}$ for $q = 0, 1, 2, 3$, $\Sigma_0^{(1)} = \begin{pmatrix} \Sigma_{01} & S_{01}\widehat{T}_1 \\ S_{10}\widehat{T}_0 & \Sigma_{11} \end{pmatrix} U_1^{(0)}$, and $\Sigma_1^{(1)} = \begin{pmatrix} \Sigma_{21} & S_{23}\widehat{T}_3 \\ S_{32}\widehat{T}_2 & \Sigma_{31} \end{pmatrix} U_1^{(1)}$. This completes the recursive process of compression and merging of the 4×4 block matrix M.

Given an $m \times n$ basically ρ-HSS matrix M with k diagonal blocks Σ_q of sizes $m_q \times n_q$ for $q = 0, \dots, k-1$, we generalize this recursive process and successively obtain matrices $U^{(j)}$ (unitary), $\Sigma^{(j)}$ (block diagonal), and $M_{j+1} = M^{(j+1)}$ (basically ρ-HSS) for $j = 0, \dots, l-1$ and $l = \lceil \log_2(k) \rceil$. At the end we arrive at the factorization $M = DU$. Here $U = U^{(0)} \prod_{j=0}^{l-1} \mathrm{diag}(I, U^{(j)})$ is a unitary matrix, $D = (\Sigma^{(0)} \mid \Sigma^{(1)} \mid \dots \mid \Sigma^{(l-1)})$, $\Sigma^{(j)} = \mathrm{diag}(\Sigma_q^{(j)})_{q=0}^{k(j)-1}$, $\Sigma_q^{(0)} = \Sigma_{q0}$ is an $m_q \times (n_q - \rho_q^{(0)})$ matrix for $\rho_q^{(0)} \le \rho$ and $q = 0, \dots, k-1$, whereas $\Sigma_q^{(j)}$ is an $m_q^{(j)} \times \rho_q^{(j)}$ matrix for $q = 0, \dots, k(j)-1$ and $k(j) \le \lceil k/2^j \rceil$, $m_q^{(j)} = m_{2q-1}^{(j-1)} + m_{2q}^{(j-1)}$, $m_q^{(0)} = m_q$ for $q < k$, $m_q^{(0)} = 0$ for $q \ge k$, $\rho_q^{(j)} \le \rho$ for $q = 0, \dots, k(j)-1$ and $j = 1, \dots, l-1$.

5.5 Reduction to an auxiliary linear system

Observe that

$$\beta(M) \le \beta(D) + \beta(U) + \sum_{j=0}^{l-1} (a(U^{(j)}) + a(\Sigma^{(j)})) \tag{19}$$

where $a(W)$ denotes the arithmetic cost of computing a matrix W. For the solution of a linear system $D\mathbf{y} = \mathbf{b}$ we need the entries of the matrices $\Sigma^{(j)}$, and so bound (19) includes the terms $a(\Sigma^{(j)})$.

The value $a(U^{(0)})$ is equal to the arithmetic cost of computing the QR factorization of the $n_q \times \rho_q$ matrices $T_1^H, \dots, T_{k-1}^H$ where $\rho_q \le \rho$ for all q, and so $a(U^{(0)}) = O(\sum_{q=0}^{k-1} \rho_q^2 n_q) = O(\rho^2 n)$. The values $a(U^{(j)})$ for $j > 0$ are bounded similarly, except that we compute the QR factors of $k(j) \le \lceil k/2^j \rceil$ matrices of sizes at most $\rho \times \rho$ for every $j > 0$, and so $\sum_{j=1}^{l-1} a(U^{(j)}) = O(k\rho^3)$ and

$$\sum_{j=0}^{l-1} a(U^{(j)}) = O((n + k\rho)\rho^2). \tag{20}$$

Next estimate $\beta(U) = \alpha(U^H)$. At the jth merging the block diagonal matrix $U^{(j)}$ has $k(j) \le \lceil k/2^j \rceil$ diagonal blocks, which are the Q factors of the QR factorization for the matrices of sizes of at most $\rho \times n_q$ for $q = 0, \dots, k-1$ and $j = 0$ and of at most $(2\rho) \times \rho$ for all positive j and all q. Therefore $\alpha(U^{(0)}) = O(\rho \sum_{q=0}^{k-1} n_q) = O(n\rho)$, whereas $\alpha(U^{(j)}) \le ck\rho^2/2^j$ for a constant c and all $j > 0$, and so

$$\beta(U) = O((n + k\rho)\rho), \tag{21}$$

dominated by bound (20). It remains to estimate $\beta(D)$ and $a(\Sigma^{(j)})$ for $j = 0, \dots, l-1$.

Write $a(\Sigma^{(j)}) = a_0(\Sigma^{(j)}) + a_1(\Sigma^{(j)})$ where $a_0(\Sigma^{(j)})$ and $a_1(\Sigma^{(j)})$ denote the arithmetic cost of computing the block products $\Sigma^{(j)}U^{(j)} = \mathrm{diag}(\Sigma_p^{(j)}U_p^{(j)})_{p=0}^{k(j)}$ and the blocks appended to the diagonal blocks at the jth merging, respectively.

Compute the block product $\Sigma^{(j)}U^{(j)}$ by using less than $2\sum_{q=0}^{k-1} m_q n_q \rho \le 2mn_+\rho$ arithmetic operations for $j = 0$ and less than $2\sum_{q=0}^{k(j)-1} m_q^{(j)} n_q^{(j)} \rho \le 2m\rho^2$ for every positive $j = 0$. Hence $\sum_{j=0}^{l=1} a_0(\Sigma^{(j)}) \le 2(n_+ + l\rho)m\rho$.

Next observe that $a_1(\Sigma^{(0)})$ amounts to the cost of computing the products $S_{10}\widehat{T}_0$, $S_{01}\widehat{T}_1$, $S_{32}\widehat{T}_2$, and $S_{23}\widehat{T}_3$ in the displayed case of (16), where $k = 4$. In the general case the two factors of such a product in a block row q have sizes of at most $m_q \times \rho$ and $\rho \times \rho$, respectively, for $q = 0, \dots, k-1$. Therefore $a_1(\Sigma^{(0)}) < 2\rho^2 \sum_{q=0}^{k-1} m_q = 2\rho^2 m$. Likewise $a_1(\Sigma^{(j)}) < 2\rho^2 m$ for every j because the overall number of rows of the factors $S_{pq}^{(j)}$ is equal to m, whereas the factors $\widehat{T}_q^{(j)}$ have sizes at most $\rho \times \rho$. Consequently $\sum_{j=0}^{l-1} a_1(\Sigma^{(j)}) < 2l\rho^2 m$ and

$$\sum_{j=0}^{l-1} a(\Sigma^{(j)}) < 2(n_+ + 2l\rho)m\rho. \tag{22}$$

To estimate $\beta(M)$ it remains to bound $\beta(D)$.

5.6 The second recursive factorization

By following [CGS07, Section 3] compute the QR factors of the matrices $\Sigma_q^{(0)}$ for $q = 0, \dots, k-1$, that is compute some unitary matrices $V_q^{(0)}$ (in factored form) and $\widehat{\rho}_q^{(0)} \times \widehat{\rho}_q^{(0)}$ nonsingular upper triangular matrices $\widehat{\Sigma}_q^{(0)}$ such that $\Sigma_q^{(0)} = V_q^{(0)} \begin{pmatrix} O \\ \widehat{\Sigma}_q^{(0)} \end{pmatrix}$ and $\widehat{\rho}_q^{(0)} \le \min\{m_q, n_q\}$ for all q. Write $V^{(0)} = \mathrm{diag}(V_q^{(0)})_{q=0}^{k-1}$, $D_1^{(j)} = (V^{(0)})^H \Sigma^{(j)}$ for $j = 0, \dots, l-1$, and $\widehat{D} = (D_1^{(1)} \mid \dots \mid D_1^{(l-1)})$. Note that all nonzero blocks of the matrices $\Sigma^{(j)}$ for all positive j keep their sizes and positions and do not increase their ranks in the transition to the matrices $D_1^{(j)}$ and that the matrix $D_1^{(0)}$ has exactly $\sum_{q=0}^{k-1} \widehat{\rho}_q^{(0)}$ nonzero rows. Remove all rows of the matrix $\widehat{D}$ sharing indices with these rows and let D_1 denote the resulting matrix. Substitution reduces the solution of a linear system $D\mathbf{y} = \mathbf{b}$ to computing the matrices $D_1^{(j)}$, $j = 0, \dots, l-1$, and to solving two linear systems of equations with the matrices D_1 and $D_1^{(0)}$. Recursively apply this process to the matrix D_1 until, in l recursive steps, substitution reduces the original linear system $D\mathbf{y} = \mathbf{b}$ to block diagonal systems with the triangular diagonal blocks.

5.7 Completion of the proof of the theorem

We have showed that $\beta(D) \le \sigma + \sum_{j=0}^{l-1}(\beta(V^{(j)}) + a(V^{(j)}) + a(D_j))$. Here σ denotes the cost of the substitution and the solution of all triangular linear

systems involved, $V^{(j)}$ denotes the unitary multiplier computed at the jth stage of the above process for $j = 0, \dots, l-1$, and $a(D_j)$ denotes the arithmetic cost of the multiplication of the matrix $V^{(j)}$ by the submatrix (denote it $\widehat{D}_j$) obtained by removing the entries of the block column $D_j^{(j)}$ from the matrix D_j. Hereafter let $\nu(W)$ denote the overall number of the nonzero entries of a matrix W, observe that $\sigma < 2\nu(D)$, and obtain

$$\sigma < 2\nu(D) \le 2s + 2(l-1)m\rho. \tag{23}$$

The arithmetic cost of the computation of the unitary multipliers $V^{(j)}$ is $O(\sum_{q=0}^{k-1} m_q n_q^2) = O(mn_+^2)$ at Stage 0 of the process and $O(\sum_{p=0}^{k(j)-1} m_p^{(j)} (n_p^{(j)})^2)$ at its jth stage for every positive j. Here $n_p^{(j)} \le \rho$, and the sum $\sum_{p=0}^{k(j)-1} m_p^{(j)}$ is monotone decreasing from m as j increases from 0. Therefore $\sum_{j=1}^{l-1} a(V^{(j)}) = O(lm\rho^2)$, and so

$$\sum_{j=0}^{l-1} a(V^{(j)}) = O(n_+^2 + l\rho^2)m. \tag{24}$$

This bound strongly dominates the sum $\sum_{j=0}^{l-1}(\beta(V^{(j)}) = \sum_{j=0}^{l-1} \alpha((V^{(j)})^H)$.

To compute the product $V^{(j)}\widehat{D}_j$ we need $O(\rho m n_+)$ arithmetic operations for $j = 0$ and $O(m\rho^2)$ for any $j > 0$. Consequently we perform this computation for $j = 0, \dots, l-1$ by using $O((n_+ + l\rho)m\rho)$ arithmetic operations, which matches (22). By combining estimates (19)–(24) we deduce bound (15).

To complete the proof of the theorem, apply bounds (14) and (15) to the transposed matrix M^T, thus extending them to the case where the matrix M is the transpose of a balanced ρ-HSS matrix with $n_q \times m_q$ diagonal blocks Σ_q for $q = 0, \dots, k-1$.

Remark 4. At the jth stage of the merging process we deal with matrix D_j, which has at most $l_j = \lceil n_+(j)/n \rceil \le 2^{l-j}$ block columns, each of at most ρ columns, that is at most $\rho 2^{l-j}$ columns overall. Suppose we stop merging process at this stage and compute the QR factors of the matrix at the arithmetic cost $O(m\rho^2 2^{l-j+1})$. For $\rho = O(\log(n))$ this modification does not affect the cost bound $\beta(M) = O(n\log^3(n))$ of Corollary 2 if the integer $l - j$ is bounded from above by a constant and even if it just holds that $l - j = O(\log(\log(n)))$.

6 Extension of the block diagonal

In the next section we employ an extension of our study where we allow overlaps of the row sets of distinct diagonal blocks. For demonstration consider the following 8×8 block matrix, which turns into a $(1,2)$-block banded matrix if we glue together its lower and upper boundaries,

$$M = \begin{pmatrix} \Sigma_0 & B_0 & C_0 & O & O & O & O & A_0 \\ A_1 & \Sigma_1 & B_1 & C_1 & O & O & O & O \\ O & A_2 & \Sigma_2 & B_2 & C_2 & O & O & O \\ O & O & A_3 & \Sigma_3 & B_3 & C_3 & O & O \\ O & O & O & A_4 & \Sigma_4 & B_4 & C_4 & O \\ O & O & O & O & A_5 & \Sigma_5 & B_5 & C_5 \\ C_6 & O & O & O & O & A_6 & \Sigma_6 & B_6 \\ B_7 & C_7 & O & O & O & O & A_7 & \Sigma_7 \end{pmatrix}. \tag{25}$$

Define the eight *extended diagonal blocks*,

$$\Sigma_0^{(c)} = \begin{pmatrix} C_6 \\ B_7 \\ \Sigma_0 \\ A_1 \end{pmatrix}, \ \Sigma_1^{(c)} = \begin{pmatrix} C_7 \\ B_0 \\ \Sigma_1 \\ A_2 \end{pmatrix}, \ \Sigma_2^{(c)} = \begin{pmatrix} C_0 \\ B_1 \\ \Sigma_2 \\ A_3 \end{pmatrix}, \ \Sigma_3^{(c)} = \begin{pmatrix} C_1 \\ B_2 \\ \Sigma_3 \\ A_4 \end{pmatrix},$$

$$\Sigma_4^{(c)} = \begin{pmatrix} C_2 \\ B_3 \\ \Sigma_4 \\ A_5 \end{pmatrix}, \ \Sigma_5^{(c)} = \begin{pmatrix} C_3 \\ B_4 \\ \Sigma_5 \\ A_6 \end{pmatrix}, \ \Sigma_6^{(c)} = \begin{pmatrix} C_4 \\ B_5 \\ \Sigma_6 \\ A_7 \end{pmatrix}, \text{ and } \Sigma_7^{(c)} = \begin{pmatrix} C_5 \\ B_6 \\ \Sigma_7 \\ A_0 \end{pmatrix}.$$

Here $\Sigma_2^{(c)}$, $\Sigma_3^{(c)}$, $\Sigma_4^{(c)}$, $\Sigma_5^{(c)}$, and $\Sigma_6^{(c)}$ are five blocks of the matrix M of equation (25), whereas each of the submatrices $\Sigma_0^{(c)}$, $\Sigma_1^{(c)}$, and $\Sigma_7^{(c)}$ has been made up of a pair of blocks of this matrix. Each pair, however, turns into a single block if we glue together the lower and upper boundaries of the matrix M. We call a block column *basic* and denote it M_q if $\mathcal{C}(M_q) = \mathcal{C}(\Sigma_q^{(c)})$, that is if it is made up of the columns that overlap an extended diagonal block $\Sigma_q^{(c)}$. We partition such a block column into the block $\Sigma_q^{(c)}$ and its complement $N_q^{(c)}$, which we call a *contracted basic neutered block column* (cf. Definition 2) and which is filled with zeros in the case of the matrix M of (25) for every q, $q = 0, \ldots, 7$, but can be filled with any parameters for general 8×8 block matrix embedding the matrix M of (25).

Generalizing our 8×8 example to $k \times k$ block banded matrix M with four block diagonals we define the extended diagonal blocks $\Sigma_q^{(c)} = \begin{pmatrix} C_{q-2 \mod k} \\ B_{q-1 \mod k} \\ \Sigma_q \\ A_{q+1 \mod k} \end{pmatrix}$, for $q = 0, \ldots, k-1$, each of them made up of a quadruple of the blocks of the matrix M, and similarly we can proceed in the case of a block banded matrix M with any number of block diagonals. In a further natural generalization we narrow the extended diagonal blocks by removing some block rows adjacent to their lower and upper boundaries and assigning the entries of these block rows to the contracted neutered block columns, that is we diminish the extension of the diagonal blocks and the contraction of the basic neutered block columns (see Figures 5–8). In this case we assume no block partitioning of these matrices M as arrays anymore, but still extend our definitions of the diagonal blocks $\Sigma_q^{(c)}$,

basic block columns M_q, contracted basic neutered block columns $N_q^{(c)}$, recursive and balanced merging, diagonal and neutered unions, the basically ρ-neutered, balanced ρ-HSS, and ρ-HSS matrices M (cf. Definitions 3 and 4), as well as basically (ϵ,ρ)-neutered, balanced (ϵ,ρ)-HSS, and (ϵ,ρ)-HSS matrices M for a positive ϵ. Here are some sample diagonal unions of the extended diagonal blocks of the matrix M of (25), $\Sigma^{(c)}_{0,1,\dots,7}=M$,

$$\Sigma^{(c)}_{0,1,2,3}=\begin{pmatrix} C_6 & O & O & O\\ B_7 & C_7 & O & O\\ \Sigma_0 & B_0 & O & O\\ A_1 & \Sigma_1 & B_1 & O\\ O & A_2 & \Sigma_2 & B_2\\ O & O & A_3 & \Sigma_3\\ O & O & O & A_4 \end{pmatrix},\ \Sigma^{(c)}_{0,1}=\begin{pmatrix} C_6 & O\\ B_7 & C_7\\ \Sigma_0 & B_0\\ A_1 & \Sigma_1\\ O & A_2 \end{pmatrix},\ \text{and } \Sigma^{(c)}_{2,3}=\begin{pmatrix} C_0 & O\\ B_1 & C_1\\ \Sigma_2 & B_2\\ A_3 & \Sigma_3\\ O & A_4 \end{pmatrix}.$$

The following definitions and theorem specify an extension of Corollary 2 to a large class of extended balanced ρ-HSS matrices M.

Definition 5. *A balanced or an extended balanced ρ-HSS matrix is* hierarchically regular *if all its diagonal blocks at the second factorization stage of the associated balanced merging process have full rank. This matrix is* hierarchically well conditioned *if these blocks are also well conditioned.*

Remark 5. The second equation of (5) implies that a Cauchy matrix is regular if and only if it is hierarchically regular. If the knots of some pair in one of the sets $\mathcal{S}$ and $\mathcal{T}$ defining a Cauchy matrix lie close to one another, then the matrix is both ill conditioned and hierarchically ill conditioned. The condition numbers of a Cauchy matrix or its block can be large even where all knots in each of the sets $\mathcal{S}$ and $\mathcal{T}$ lie far from each other, but unlike Vandermonde matrices, Cauchy matrices are quite stable when we shift or scale their knots by a complex value $a\neq 0$ (cf. (12)).

Definition 6. *Assume positive integers m, n, $k=k_0<n$, l, and $k(j)=\lceil k/2^j\rceil$ where $j=0,\dots,l-1$. Suppose that the j-th stage of the recursive merging process for an $m\times n$ extended balanced ρ-HSS matrix M outputs diagonal blocks $\bar\Sigma_q^{(j)}$ of sizes $\bar m_q^{(j)}\times\bar n_q^{(j)}$ and extended diagonal blocks $\Sigma_{q,j}^{(c)}$ of sizes $m_{q,j}^{(c)}\times\bar n_q^{(j)}$ for $q=0,\dots,k(j)-1$ and $j=0,\dots,l-1$. Write $m_j^{(c)}=\sum_{q=0}^{k(j)-1}m_{q,j}^{(c)}$, $\bar s^{(j)}=\sum_{q=0}^{k(j)-1}\bar m_q^{(j)}\bar n_q^{(j)}$, and $s_j^{(c)}=\sum_{q=0}^{k(j)-1}m_{q,j}^{(c)}\bar n_q^{(j)}$. (Note that $m=\sum_{q=0}^{k-1}\bar m_q^{(j)}$ for every j and that our extensions of the diagonal blocks do not change the number of their columns.) Suppose that*

$$m_j^{(c)}\le\eta m \text{ and } s_j^{(c)}\le\eta\bar s^{(j)} \tag{26}$$

for a constant η and all j, $j=0,\dots,l-1$. Then the matrix M is called an η-dilated *balanced ρ-HSS matrix.*

Theorem 11. *For a constant $\eta \geq 1$ and an η-dilated balanced ρ-HSS matrix M of a size $m \times n$ it holds that $\alpha(M) = O((m+n)\log^2(m+n))$ where the parameters ϵ and ρ have order $\log(m+n)$. Furthermore $\beta(M) = O(n\log^3 n)$ provided that $m = n$ and the matrix M is hierarchically regular.*

Proof. Revisit the proof of the cost bound on $\alpha(M)$ of Theorem 10 replacing m and $\bar{s}^{(j)}$ by $m_j^{(c)}$ and $s_j^{(c)}$, respectively, and then, under this change, observe that with the only exception, all auxiliary and final bounds remain valid up to a factor of η, by virtue of (26). The exception is the impact of the QR factorizations at the second factorization stage in our proof of bound (15). Because of the extension of the diagonal blocks, the sizes and ranks of the nonzero blocks of the matrices $\Sigma^{(j)}$ for positive j can increase in the transition to the matrices $D_1^{(j)}$. To avoid this increase, we restrict the QR factorizations at that stage to the diagonal blocks and use the computed triangular factors as the pivot blocks to eliminate the other entries of the extended diagonal blocks in these columns by means of substitution. We readily verify that the recipe works (that is we avoid divisions by 0) and still supports bound (15) where the matrix M is hierarchically regular.

Remark 6. We conjecture that the associated balanced HSS process supporting Theorem 11 for a hierarchically regular input matrix M is numerically stable if the extended balanced ρ-HSS matrix M is hierarchically well conditioned, and clearly this process can be unstable otherwise.

7 Approximation of CV and CV^T matrices by HSS matrices and algorithmic implications

7.1 Small-rank approximation of certain Cauchy matrices

Definition 7. *(See [CGS07, page 1254].) For a* separation bound $\theta < 1$ *and a complex* separation center c, *two complex points s and t are (θ, c)-*separated *from one another if $|\frac{t-c}{s-c}| \leq \theta$. Two sets of complex numbers $\mathcal{S}$ and $\mathcal{T}$ are (θ, c)-*separated *from one another if every two points $s \in \mathcal{S}$ and $t \in \mathcal{T}$ are (θ, c)-separated from one another. $\delta_{c,\mathcal{S}} = \min_{s \in \mathcal{S}} |s - c|$ denotes the distance from the center c to the set $\mathcal{S}$.*

Lemma 1. *(See [R85] and [CGS07, equation (2.8)].) Suppose two complex values s and t are (θ, c)-separated from one another for $0 \leq \theta < 1$ and a complex* center *c and write $q = \frac{t-c}{s-c}$, $|q| \leq \theta$. Then for every positive integer ρ it holds that*

$$\frac{1}{s-t} = \frac{1}{s-c}\sum_{h=0}^{\rho-1}\frac{(t-c)^h}{(s-c)^h} + \frac{q_\rho}{s-c} \text{ where } |q_\rho| = \frac{|q|^\rho}{1-|q|} \leq \frac{\theta^\rho}{1-\theta}. \tag{27}$$

Proof. $\frac{1}{s-t} = \frac{1}{s-c}\frac{1}{1-q}$, $\frac{1}{1-q} = \sum_{h=0}^{\infty} q^h = (\sum_{h=0}^{\rho-1} q^h + \sum_{h=\rho}^{\infty} q^h) = (\sum_{h=0}^{\rho-1} q^h + \frac{q^\rho}{1-q})$.

Corollary 3. *(Cf. [CGS07, Section 2.2] and [B10].) Suppose two sets of* $2n$ *distinct complex numbers* $\mathcal{S} = \{s_0, \dots, s_{m-1}\}$ *and* $\mathcal{T} = \{t_0, \dots, t_{n-1}\}$ *are* (θ, c)*-separated from one another for* $0 < \theta < 1$ *and a* global complex center c*. Define the Cauchy matrix* $C = (\frac{1}{s_i - t_j})_{i,j=0}^{m-1,n-1}$ *and write* $\delta = \delta_{c,\mathcal{S}} = \min_{i=0}^{m-1} |s_i - c|$ *(cf. Definition 7). Then for every positive integer* ρ *it is sufficient to apply* $(m+n)\rho + m$ *arithmetic operations to compute the* $m \times \rho$ *matrix* $F = (1/(s_i - c)^{\nu+1})_{i,\nu=0}^{m-1,\rho-1}$ *and the* $n \times \rho$ *matrix* $G = ((t_j - c)^{\nu})_{j,\nu=0}^{n-1,\rho-1}$*, and it holds that*

$$C = FG^T + E, \ |E| \le \frac{\theta^\rho}{(1-\theta)\delta}. \tag{28}$$

Proof. Apply (27) for $s = s_i$, $t = t_j$, and all pairs (i, j) to deduce (28).

Remark 7. Assume an $m \times n$ Cauchy matrix $C = (\frac{1}{s_i - t_j})_{i,j=0}^{m-1,n-1}$ with $m+n$ distinct knots $s_0, \dots, s_{m-1}, t_0, \dots, t_{n-1}$. Then $\mathrm{rank}(C) = \min\{m, n\}$ (cf. Theorem 4). Further assume that the sets $\mathcal{S} = \{s_0, \dots, s_{m-1}\}$ and $\mathcal{T} = \{t_0, \dots, t_{n-1}\}$ are (θ, c)-separated from one another for a global complex center c and $0 < \theta < 1$ such that the value $(1-\theta)\delta/\sqrt{mn}$ is not small. Then by virtue of the corollary the matrix C can be closely approximated by a matrix FG^T of a smaller rank $\rho < \min\{m, n\}$, and therefore is ill conditioned. Furthermore if we have such (θ, c)-separation just for a $k \times l$ submatrix $C_{k,l}$ of the matrix C, implying that $\mathrm{nrank}(C_{k,l}) \le \rho$, then it follows that $\mathrm{nrank}(C) \le m - k + n - l + \rho$. Consequently if $m - k + n - l + \rho < \min\{m, n\}$, then again we conclude that the matrix C is ill conditioned. These classes of ill conditioned Cauchy matrices contain a large class of CV and CV^T matrices. In particular a CV matrix is ill conditioned if all its knots s_i or all knots s_i of its submatrix of a large size lie far enough from the unit circle $\{z : \ |z| = 1\}$, because in this case the origin serves as a global center for the matrix or submatrix.

Generally neither CV matrix nor its submatrices of a large size have global separation centers, but next we compute a *set of local centers* to approximate any CV matrix closely by an extended ρ-HSS matrix for a small ρ. We first (i) determine the required properties of the centers, then (ii) compute proper diagonal and *extended diagonal blocks*, and finally (iii) *merge* diagonal blocks recursively. We devote the next three subsections to these three stages.

7.2 Local separation centers

Definition 8. $\mathcal{A}(\phi, \phi') = \{z = \exp(\psi\sqrt{-1}) : \ 0 \le \phi \le \psi < \phi' \le 2\pi\}$ *is the* semi-open arc *of the unit circle* $\{z : \ |z| = 1|\}$ *having* length $\phi' - \phi$ *and the endpoints* $\tau = \exp(\phi\sqrt{-1})$ *and* $\tau' = \exp(\phi'\sqrt{-1})$*.* $\Gamma(\phi, \phi') = \{z = r\exp(\psi\sqrt{-1}) : \ r \ge 0, \ 0 \le \phi \le \psi < \phi' \le 2\pi\}$ *is the* semi-open sector *bounded by the two rays from the origin to the two endpoints of the arc.* $\bar{\Gamma}(\phi, \phi')$ *denotes the exterior (that is the complement) of this sector.* $D(c, r) = \{z : \ |z - c| \le r\}$ *is the* disc *on the complex plane with a center* c *and a radius* r*, having the exterior* $\bar{D}(c, r) = \{z : \ |z - c| > r\}$.

We readily verify the following result (cf. Figures 9 and 10).

Lemma 2. *Suppose* $0 \le \phi < \phi' < \psi \le 2\pi$, $\phi - \psi \mod (2\pi) \ge \phi' - \phi$, $\tau = \exp(\phi\sqrt{-1})$, *and* $\tau' = \exp(\phi'\sqrt{-1})$. *Then (i)* $|\tau' - \tau| = 2\sin((\phi' - \phi)/2)$ *and (ii) the distance from the point* τ *to the sector* $\Gamma(\phi', \psi)$ *is equal to* $\sin\mu$, *for* $\mu = \min\{(\phi - \psi) \mod (2\pi), \phi' - \phi\}$.

Apply the lemma to obtain the following theorem.

Theorem 12. *Assume the numbers* θ, ϕ, ϕ', *and* c *such that* $0 < \theta < 1$, $0 \le \phi < \phi \le 2\pi$, *and* $c = \exp(0.5(\phi' + \phi)\sqrt{-1})$ *is the midpoint of the arc* $\mathcal{A}(\phi, \phi')$. *Let* $\bar{D} = \bar{D}(c, (2/\theta)\sin((\phi' - \phi)/4))$ *denote the exterior of the disc* $D(c, (2/\theta)\sin((\phi' - \phi)/4))$. *Then the two sets* $\mathcal{A}(\phi, \phi')$ *and* $\mathcal{D}(\phi, \phi', \theta)$ *are* (θ, c)*-separated.*

Combine this theorem with Corollary 3 to obtain the following corollary.

Corollary 4. *Under the assumptions of Theorem 12 the* $\bar{m} \times \bar{n}$ *Cauchy matrix*

$$C(\phi, \phi') = \Big(\frac{1}{s_i - t_j}\Big)_{s_i \in \mathcal{D}(\phi, \phi', \theta),\ t_j \in \mathcal{A}(\phi, \phi')}$$

has $|E|$*-rank at most* ρ *provided that* $|E|$ *satisfies bound (28) for* $m = \bar{m}$, $n = \bar{n}$, *and* $\delta = \delta_{c, \mathcal{D}(\phi, \phi', \theta)} = \min_i |s_i - c|$.

We will apply Theorem 12 and Corollary 4 where the value $1 - \theta$ is not small and the knots s_i lie in the exterior $\bar{\Gamma}(\psi, \psi')$ of the smallest sector $\Gamma(\psi, \psi')$ that covers the disc $D(c, (2/\theta)\sin((\phi' - \phi)/4))$. Clearly $\bar{\Gamma}(\psi, \psi') \subset \bar{D}$, whereas the arcs $\mathcal{A}(\phi, \phi')$ and $\mathcal{A}(\psi, \psi')$ share their midpoint c. Next we estimate the ratio of the lengths of these arcs, $r(\theta) = (\phi' - \phi)/(\psi' - \psi)$. Lemma 2 implies that

$$\text{Distance}(c, \bar{\Gamma}(\psi, \psi')) = \sin((\psi' - \psi)/2),$$

but this distance is also equal to $(c - \tau)/\theta = (2/\theta)\sin((\phi' - \phi)/4)$, and so

$$\bar{r}(\theta) = (2\sin((\phi' - \phi)/4)/\sin((\psi' - \psi)/2) = 1/\theta. \tag{29}$$

Now recall that $y - \sin(y) = y^3/3! - y^5/5! + \dots$ and so $0 \le y - \sin(y) < y^3/3! = y^3/6$ where $y^2 < 7!/5! = 42$. Therefore the difference $y - \sin(y)$ is nonnegative and converges to 0 cubically in y as $y \to 0$. (In particular $y - \sin(y) < 1/48$ for $y \le 1/2$, $y - \sin(y) < 1/384$ for $y \le 1/4$, and $y - \sin(y) < 1/6000$ for $y \le 1/10$.) We can choose $\theta > 2/3$, say, and then $y = (\psi' - \psi)/2 \le 2x$ for $x = (\phi' - \phi)/2$ (see Figure 10). Furthermore, in our applications we can assume $\theta \ge 1/2$ (say), enforce a small upper bound on x, and therefore ensure close approximation

$$r(\theta) \approx \bar{r}(\theta) = 1/\theta. \tag{30}$$

7.3 Defining the extended diagonal blocks of a CV matrix

Next we partition the complex plane into k sectors, each defined by a pair of rays from the origin, then partition the sets of knots $\mathcal{S}$ and $\mathcal{T}$ accordingly, and finally define diagonal and extended diagonal blocks of CV matrices. Let us formalize this idea. Fix a positive integer l_+, write $k = 2^{l_+}$, $\phi_q = 2q\pi/k$, and $\phi'_q = \phi_{q+1 \mod k}$, partition the unit circle $\{z : |z = 1|\}$ by k equally spaced points $\phi_0, \dots \phi_{k-1}$ into k semi-open arcs $\mathcal{A}_q = \mathcal{A}(\phi_q, \phi'_q)$, each of the length $2\pi/k$, and define the semi-open sectors $\mathcal{S}_q = \mathcal{S}(\phi_q, \phi'_q)$ for $q = 0, \dots, k-1$.

Let $C = C_{\mathbf{s},f} = \left(\frac{1}{s_i - t_j}\right)_{i,j=0}^{m-1,n-1}$ denote an $m \times n$ Cauchy matrix. Assume the polar representation $s_i = |s_i| \exp(\mu_i \sqrt{-1})$ and $t_j = |t_j| \exp(\nu_j \sqrt{-1})$, and reenumerate the knots in the counter-clockwise order of the angles μ_i and ν_j breaking ties arbitrarily and assigning the smallest subscripts i and j to all knots in the sector $\mathit{\Gamma}(\phi_0, \phi'_0)$, then to all knots in the sector $\mathit{\Gamma}(\phi_1, \phi'_1)$, and so on. The partition of the complex plane into the sectors induces a partition of the angles μ_i and ν_j in the polar representation of the knots s_i and t_j as well as the block partition of the matrix $C = (C_{p,q})_{p,q=0}^{k-1}$ such that $C_{p,q} = \left(\frac{1}{s_i - t_j}\right)_{s_i \in \mathit{\Gamma}_p, t_j \in \mathit{\Gamma}_q}$. We further partition the matrix C into its basic block columns $C = (C_0 \mid \dots \mid C_{k-1})$ and partition every basic block column C_q into the diagonal block $\mathit{\Sigma}_q = C_{q,q}$ and the basic neutered block column N_q for $q = 0, \dots, k-1$.

Now suppose that we are given the values θ and f such that $0 < \theta < 1$ and $|f| = 1$, and write $t_j = f\omega^j$ for $j = 0, \dots, n-1$. Then for $q = 0, \dots, k-1$, write $\phi = \phi_q$, $\phi' = \phi'_q$, define the real values $\psi = \psi_q$ and $\psi' = \psi'_q$ and the sectors $\mathit{\Gamma}_q^{(c)} = \mathit{\Gamma}(\psi_q, \psi'_q)$ by following the definitions of Section 7.1 (cf. (29)), and define the second partition of every basic block column C_q, this time into an extended diagonal block $\mathit{\Sigma}_q^{(c)} = \left(\frac{1}{s_i - t_j}\right)_{s_i \in \mathit{\Gamma}_q^{(c)}, t_j \in \mathit{\Gamma}_q}$ and the contracted basic neutered block column $N_q^{(c)}$, for $q = 0, \dots, k-1$. (Unlike the first partition this one depends on the parameter θ.) By following [B10], we refer to the contracted basic neutered block columns also as *admissible blocks* and define the partition of every basic neutered block column C_q into the triple of a diagonal block $\mathit{\Sigma}_q$, a pair of *neighboring blocks*, and an admissible block $N_q^{(c)}$ (cf. Figures 5–8).

By applying Corollary 4 obtain the following result.

Corollary 5. *For a fixed θ, $0 \le \theta < 1$, and an $m \times n$ CV matrix C, define the above admissible blocks $N_0^{(c)}, \dots, N_{k-1}^{(c)}$. Then all of them have the $|E|$-ranks at most ρ, that is C is an extended basically $(|E|, \rho)$-neutered matrix, provided $|E|$ and ρ satisfy bound (28) for $n = k$, $\delta = \min_{i,q=0}^{m-1,k-1} |s_i - c_q|$, and the fixed values of θ and m.*

Corollary 5 and Theorem 8 together imply that

$$\alpha_{|E|}(C) \le \alpha(\mathrm{diag}(\mathit{\Sigma}_q^{(c)})_{q=0}^{k-1}) + (2m + 2n - 1)k\rho. \tag{31}$$

One can deduce from this bound that $\alpha_{|E|}(C) = O(n\sqrt{n \log(n)})$ for an $n \times n$ CV matrix C (see [P13]), but we will strengthen these estimates by showing that C

is an extended balanced $(|E|, \rho)$-HSS matrix where $\log(1/|E|) = O(\log(m+n))$ and $\rho = O(\log(m+n))$.

7.4 Extended HSS approximation of a CV matrix

For a fixed value θ, $0 < \theta < 1$, and a positive integer l_+ we recursively merge the arcs $\mathcal{A}_0, \dots, \mathcal{A}_{k-1}$ pairwise. By applying Corollary 5 at every level of this merging process, we conclude that the CV matrix C of the previous subsection is an extended basically (ϵ_j, ρ_j)-neutered matrix at the jth level of the process for all j, $j = 0, \dots, l_+ - 1$. Actually we stop the merging process at the level $l < l_+$ to have reasonably many arcs of a sufficiently small length at the end of the process (cf. Remark 4).

Next we are going to specify a common pair (ϵ, ρ) for all levels, thus implying that C is an extended balanced (ϵ, ρ)-HSS matrix. We already have pairs (ϵ_j, ρ_j), for $j = 0, \dots, l-1$, $l = \log_2(k)$, that satisfy bound (28) for $|E| = \epsilon_j$ and $\rho = \rho_j$. Note that $m_{q,j}^{(c)} \le m$ and $\bar{n}_q^{(j)} \le n$ for admissible blocks of the sizes $m_{q,j}^{(c)} \times \bar{n}_q^{(j)}$. Therefore to unify the choice of the pair (ϵ, ρ) for all j, we just need to substitute the upper bounds m and n on $m_{q,j}^{(c)}$ and $\bar{n}_q^{(j)}$, respectively, and to fix a value $\delta \le \min_{i,j,q=0}^{m-1,l-1,k/2^j} |s_i - c_{j,q}|$. So far our construction still leaves us with the freedom of rotating the arcs $\mathcal{A}_q^{(j)}$ for all j and q by a fixed angle ψ. There are m knots s_i and $\sum_{j=0}^{l-1} k/2^j \le 2k$ centers $c_{j,q}$ overall, each separated from all other centers by arcs of lengths at least π/k. Therefore by choosing a proper rotation angle ψ we can ensure the angles of at least $\pi/(2km)$ between the two rays from the origin passing through the pair of a knot s_i and a center $c_{j,q}$, for every triple (i, j, q). Apply Lemma 2 and obtain

$$\delta \ge \delta_- = 2\sin(\pi/(4km)). \tag{32}$$

For larger integers km it holds that

$$\delta \ge \delta_- = 2\sin(\pi/(4km)) \approx \pi/(2km). \tag{33}$$

Theorem 13. *(i) An $m \times n$ CV matrix C is an extended balanced (ϵ, ρ)-HSS matrix where*

$$\epsilon \approx \frac{2mk\theta^\rho}{(1-\theta)\pi}, \tag{34}$$

θ and $1-\theta$ are positive constants, and $1 < k < n$. (ii) It holds that

$$\rho = O(\log(\frac{m+n}{\epsilon})) \tag{35}$$

and consequently $\rho = O(\log(m+n))$ provided that $\log(1/\epsilon) = O(\log(m+n))$.

Proof. Combine bounds (28) and (32) to obtain part (i), which immediately implies part (ii).

Remark 8. The lower bound $\delta_- = \pi/(2km)$ is overly pessimistic for many dispositions of the knots s_i on the complex plane. For example, δ is a constant for CV matrices where all these knots are separated from the unit circle, whereas $\delta \geq \pi/m$ where $s_i = \omega_m^i$, $i = 0, \dots, m-1$, and in many cases only a small number of the differences $c_{j,q} - s_i$ has absolute values close to the minimum δ.

7.5 Complexity of approximate computations with CV and CV^T matrices

Next we ensure that a CV matrix C is a 3.1-dilated balanced (ϵ, ρ)-HSS matrix and then apply Theorem 11. Suppose that our fan-in merging process applied to a balanced HSS approximation of an $m \times n$ CV matrix C begins with k diagonal blocks Σ_q of sizes $m_q \times n_q$ and k extended diagonal blocks $\Sigma_d^{(c)}$ of sizes $m_q^{(c)} \times n_q$, for $n_q \approx n/k$ of order $\log(m+n)$ and $q = 0, \dots, k-1$ and deduce the following result from equations (29) and (30).

Theorem 14. *An extended balanced (ϵ, ρ)-HSS CV matrix C of Theorem 11 is 3.1-dilated provided that the ratio $k/2^l$ is reasonably large (cf. Definition 6 and Remark 4) and that we have chosen a positive parameter θ such that $r(\theta) \leq 3$ (where $r(\theta)$ is the scalar parameter bounded according to (29) and (30)).*

Proof. Recall that at the initial level of our balanced recursive merging the unit circle is partitioned into the arcs $\mathcal{A}_q = \mathcal{A}(\phi_q, \phi'_q)$ that have the same length for all q. Therefore the $r(\theta)$-dilation $\mathcal{A}(\psi_q, \psi'_q)$, sharing the center c_q with the arc $\mathcal{A}_q$, has a length invariant in q as well. For $r(\theta) \leq 3$ such a dilation keeps the arc $\mathcal{A}(\psi_q, \psi'_q)$ in the union of the arc $\mathcal{A}_q$ and its two extensions on its both sides by two arcs, each of the length $(r(\theta)-1)(\phi'_q - \phi_q)/2$. This length is at most $\phi'_q - \phi_q$ for $r(\theta) \leq 3$. Therefore all arcs $\mathcal{A}(\psi_q, \psi'_q)$ together for all q cover the union $\cup_{q=0}^{k-1} \mathcal{A}(\phi_q, \phi'_q) = \{z : |z| = 1\}$ at most three times. Consequently $m_0^{(c)} = \sum_{q=0}^{k-1} m_{q,0}^{(c)} \leq 3m = \sum_{q=0}^{k-1} m_{q,0}$.

Next write $n_{+,0} = \max_{q=0}^{k-1} n_{q,0}$ and $n_{-,0} = \min_{q=0}^{k-1} n_{q,0}$ and observe that $0 \leq n_{+,0} - n_{-,0} \leq 1$ because the arcs $\mathcal{A}(\psi_q, \psi'_q)$ have the same length for all q and because $t_j = f\omega^j$ for $|f| = 1$. Furthermore note that $s_0^{(c)} = \sum_{q=0}^{k-1} m_{q,0}^{(c)} n_{q,0} \leq n_{+,0} \sum_{q=0}^{k-1} m_{q,0}^{(c)} = n_{+,0} m_0^{(c)}$. Similarly $s_0^{(c)} \geq n_{-,0} m_0^{(c)}$ and $n_{-,0} m \leq s_0 \leq n_{+,0} m$. Therefore $s_0^{(c)} \leq n_{+,0} m_0^{(c)} \leq (n_{-,0}+1) m_0^{(c)} \leq 3(n_{-,0}+1)m \leq (3 + 3/n_{-,0})s$. We can assume that $n_{-,0} \geq (n/k) - 1 \gg 1$, and thus $3/n_{-,0} < 0.1$.

Remark 9. Recall that the scalar parameter $r(\theta)$ cubically converges to $1/\theta$ as $k \to \infty$. Therefore for large integers k and for $\theta \approx 1$ a "typical" CV matrix is η-dilated with $\eta \approx 1$.

Combine Theorems 11 and 14 with Corollary 3 and obtain the following results (cf. Remark 5).

Theorem 15. *For an $m \times n$ CV matrix C it holds that $\alpha_\epsilon(C) = O((m+n)\rho\log(n))$ provided that $\rho = O(\log(1/(\delta\epsilon)))$ for δ of Corollary 3, which satisfies the bound $\log(1/\delta) = O(\log(m+n))$. If in addition $m = n$ and the matrix C is nonsingular and hierarchically well conditioned under a certain associated balanced (ϵ,ρ)-HSS merging process, then $\beta_\epsilon(C) = O(n\rho^2\log(n))$. In particular if $\log(1/\epsilon) = O(\log(m+n))$, then the above estimates imply that $\alpha_\epsilon(C) = O((m+n)\log(m+n)\log(n))$ and $\beta_\epsilon(C) = O(n\log^3(n))$.*

Because of the dual role of the rows and columns in our constructions we can readily extend all our results from CV matrices C to CV^T matrices C^T. In particular we can extend Theorem 15 as follows.

Corollary 6. *The estimates of Theorem 15 also hold for a CV^T matrix C.*

8 Extensions and modifications

8.1 Computations with Vandermonde matrices and their transposes

Next we employ equations (8)–(11) to extend Theorem 15 to computations with Vandermonde matrices, their transposes, and with polynomials.

Theorem 16. *Suppose that we are given two positive integers m and n and a vector $\mathbf{s} = (s_i)_{i=0}^{m-1}$ defining an $m \times n$ Vandermonde matrix $V = V_{\mathbf{s}}$. Write $s_+ = \max_{i=0}^{m-1}|s_i|$ and let $\log(1/\epsilon) = O(\log(m+n) + n\log(s_+))$. (i) Then*

$$\alpha_\epsilon(V) + \alpha_\epsilon(V^T) = O((m+n)(\rho\log(m+n) + n\log(s_+))). \tag{36}$$

(ii) Suppose that in addition $m = n$ and for some complex f, $|f| = 1$, the matrix $C_{\mathbf{s},f}$ of equation (7) is nosingular and hierarchically well conditioned under a certain associated balanced (ϵ,ρ)-HSS merging process. Then

$$\beta_\epsilon(V) + \beta_\epsilon(V^T) = O(n\rho^2\log(n)). \tag{37}$$

(iii) Bounds (36) and (37) on $\alpha_\epsilon(V)$ and $\beta_\epsilon(V)$ can be applied also to the solution of Problems 1 and 2 of Section 3, respectively.

Proof. Combine Theorem 15 and Corollary 6 with equations (8)–(11). The matrices $\mathrm{diag}(\omega^{-j})_{j=0}^{n-1}$, $\mathrm{diag}(f^{-j})_{j=0}^{n-1}$, and $\Omega/\sqrt{n} = (\sqrt{n}\Omega^H)^{-1}$ and their inverses are unitary, and so multiplication by them makes no impact on the output error norms. Multiplication by the matrix $\mathrm{diag}(s_i^n - f^n)_{i=0}^{m-1}$ can increase the value $\log_2(1/\epsilon)$ by at most $\log_2(s_+^n+1)$, whereas multiplication by its inverse for $m = n$ can increase this value by at most $\log_2(\Delta)$ for $\Delta = 1/\max_{f:\ |f|=1}\min_{i=0}^{m-1}|s_i^n - f^n|$. We can ensure that $\Delta \le 2m$ by choosing a proper value f, and so $\log_2(\Delta) \le 1+\log_2(m)$. Such an increase makes no impact on the asymptotic bounds of Theorem 16, and so we complete the proof of parts (i) and (ii). Equations (1) and (2) extend the proof to part (iii).

Note that the term $n\log(s_+)$ is dominated and can be removed from the bound on $\log(1/\epsilon)$ and (36) provided that $s_+ = 1 + O(\frac{\log^2(m+n)}{n})$.

8.2 Computations with other structured matrices, polynomials, and rational functions

The FMM/HSS techniques of [GR87], [DGR96], [CGR98], and [B10] combined with the algebraic techniques of [P90] and [Pa] work efficiently for other classes of structured matrices, and our complexity estimates can be extended and in some cases strengthened. Next we recall some relevant results from [Pa].

For $m \times n$ Toeplitz and Hankel matrices W one yields the bound $\beta_\epsilon(W) = O((n)\log^2(1/\epsilon)\log(n))$ where $m = n$ (see [Pa]). Our estimates for CV matrices can be extended to general Cauchy matrices $C_{\mathbf{s},\mathbf{t}}$ with arbitrary sets of knots s_i and t_j provided that we allow to increase the approximation errors by factors $||C||\ ||C^{-1}||$ for $C = C_{\mathbf{s},f}$ or/and $C = C_{e,\mathbf{t}}$ for constants e and f of our choice such that $|e| = |f| = 1$. These estimates and the ones of the previous subsection are immediately extended to approximate solution of Problems 3 and 4 of rational interpolation and multipoint evaluation. Furthermore all algorithms and estimates can be extended from Cauchy to generalized Cauchy matrices $f(s_i - t_j)_{i,j=0}^{m-1,n-1}$ for various functions $f(z)$ such as x^{-p} for a positive integer p, $\ln z$, and $\tan z$.

Finally the classes of Toeplitz, Hankel, Vandermonde and Cauchy matrices W have been extended to larger classes of $m \times n$ matrices M that have structures of Toeplitz, Hankel, Vandermonde and Cauchy types. They allow compressed expressions through their displacements $AM - MB$ of small ranks d for operator matrices A and B fixed for each of the four structures, that is through at most $(m+n)d$ parameters per matrix. The known fast algorithms for computations with Toeplitz, Hankel, Vandermonde and Cauchy matrices are extended to these classes, and this includes fast approximation algorithms, with the estimates of this papers changed into $\alpha_{\epsilon'}(M) = O(d\alpha_\epsilon(W))$ and $\beta_{\epsilon''}(M) = O(d\beta_\epsilon(W))$ for $\epsilon' = O(d|F|\epsilon)$ and $\epsilon'' = O(d|F|\ ||M^{-1}||\epsilon)$ (cf. [Pa]).

8.3 Numerical rank of the admissible blocks and the impact on the implementation

To implement our algorithms one can compute the centers c_q and the admissible blocks $\widehat{N}_q$ of bounded ranks throughout the merging process, but one can avoid a large part of these computations by following the papers [CGS07], [X12], [XXG12], and [XXCBa]. They bypass the computation of the centers c_q and immediately compute the HSS generators for the admissible blocks $\widehat{N}_q$, defined by HSS trees. The length of the generators can be chosen equal to the available upper bound ρ on the numerical ranks of these blocks or can be adapted empirically. Our computational cost bounds $\alpha_\epsilon(M)$ and $\beta_\epsilon(M)$ are proportional to ρ and ρ^2, respectively, and so they decrease as the numerical rank ρ decreases. In particular our complexity bounds decrease to the level $\alpha_\epsilon(C) = O(n\log(1/\epsilon)\log(n))$ and $\beta_\epsilon(C) = O(n\log^2(1/\epsilon)\log(n))$ where the ϵ-rank decreases to the level $O(\log(1/\epsilon))$ (cf. our Remark 8), thus extending the latter bound to the case of Toeplitz and Hankel inputs.

9 Conclusions

The papers [MRT05], [CGS07], [XXG12], and [XXCBa] combined the advanced FMM/HSS techniques with a transformation of matrix structures (traced back to [P90]), in order to devise algorithms that compute approximate solution of Toeplitx, Hankel, Toeplitz-like, and Hankel-like linear systems of equations in nearly linear arithmetic time (versus cubic time of the classical algorithms). We analyzed these algorithms and showed that their power can be extended to yield similar results for computations with other structured matrices, in particular Vandermonde and Cauchy matrices (with the extensions to polynomial and rational evaluation and interpolation). The resulting decrease of the running time of the known approximation algorithms is by order of magnitude, from quadratic to nearly linear. We elaborated upon detailed description and analysis of the algorithms, providing new insights into the subject, formal complexity estimates, and background support for further advances in [Pa], which include the extension of our results to the case of other matrices having displacement structure and further acceleration of the known approximation algorithms in the case of Toeplitz and Hankel inputs.

Appendix

A Legends to Figures 2–10

In Figures 2–8 diagonal blocks are marked by green color. In Figures 2–4 basic neutered block columns are marked by blue color. In Figures 3 and 4 the pairs of smaller diagonal blocks (marked by light green color) are merged into their diagonal unions, each made up of four smaller blocks, marked by light and dark green colors. In Figures 5–8 the contracted basic neutered block columns, also called admissible blocks, and shown by blue, each (green) diagonal block has two red neighboring blocks, and their triples combined form the extended diagonal blocks. Figures 2, 5 and 6 share their diagonal blocks and are associated with the values θ equal to 1, 1.25, and 2, respectively. Accordingly, the neighboring blocks are absent from Figure 2 and are larger in Figure 6 than in Figure 5.

Figures 9 and 10 mark by black color an arc of the unit circle $\{z:\ |z=1|\}$. In Figure 9 this arc is intersected by a blue (internal) circle. The two intersection points τ and τ' are the endpoints of the arc $\mathcal{A}(\phi,\phi')$, having the center c. The red (external) circle bounds the disc $D(c,(2/\theta)\sin((\phi'-\phi)/4))$. In Figure 10 we mark by blue the five line intervals $[0,\tau]$, $[0,c]$, $[0,\tau']$, $[\tau,c]$, and $[c,\tau]$. We mark by red the two line intervals bounding the intersection of the sector $\mathit{\Gamma}(\psi,\psi')$ and the unit disc $D(0,1)$ as well as the two perpendiculars from the point c onto these two bounding lines.

Acknowledgements: Our research has been supported by the NSF Grant CC 1116736 and the PSC CUNY Awards 64512–0042 and 65792–0043.

References

[B99] R. Bracewell, *The Fourier Transform and Its Applications*. McGraw-Hill, New York, 1999 (3rd edition).

[B10] S. Börm, *Efficient Numerical Methods for Non-local Operators: $\mathcal{H}^2$-Matrix Compression, Algorithms and Analysis*, European Math. Society, 2010.

[BEGO08] T. Bella, Y. Eidelman, I. Gohberg, V. Olshevsky, Computations with Quasiseparable Polynomials and Matrices, *Theoretical Computer Science, Special Issue on Symbolic–Numerical Algorithms* (D. A. Bini, V. Y. Pan, and J. Verschelde editors), **409**, **2**, 158–179, 2008.

[BF00] D. A. Bini, G. Fiorentino, Design, Analysis, and Implementation of a Multiprecision Polynomial Rootfinder, *Numer. Algs.*, **23**, 127–173, 2000.

[B-O83] M. Ben-Or, Lower Bounds for Algebraic Computation Trees, *Proceedings of 15th Annual ACM Symposium on Theory of Computing (STOC'83)*, 80–86, ACM Press, New York, 1983.

[BP70] A. Björck, V. Pereyra, Solution of Vandermonde Systems of Equations, *Math. of Computation*, **24**, 893–903, 1970.

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

[BY13] L. A. Barba, R. Yokota, How Will the Fast Multipole Method Fare in Exascale Era? *SIAM News*, **46**, **6**, 1–3, July/August 2013.

[CDG06] S. Chandrasekaran, P. Dewilde, M. Gu, W. Lyons, T. Pals, A Fast Solver for HSS Representations via Sparse Matrices, *SIAM J. Matrix Anal. Appl.*, **29, 1**, 67–81, 2006.

[CGR98] J. Carrier, L. Greengard, V. Rokhlin, A Fast Adaptive Algorithm for Particle Simulation, *SIAM J. Scientific Computing*, **9**, 669–686, 1998.

[CGS07] S. Chandrasekaran, M. Gu, X. Sun, J. Xia, J. Zhu, A superfast algorithm for Toeplitz systems of linear equations, *SIAM J. Matrix Anal. Appl.*, **29**, 1247–1266, 2007.

[DGR96] A. Dutt, M. Gu, V. Rokhlin, Fast algorithms for polynomial interpolation, integration, and differentiation, *SIAM Journal on Numerical Analysis*, **33, 5**, 1689–1711, 1996.

[DV98] P. Dewilde and A. van der Veen, *Time-Varying Systems and Computations*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[EG02] Y. Eidelman, I. Gohberg, A Modification of the Dewilde–van der Veen Method for Inversion of Finite Structured Matrices, *Linear Algebra and Its Applications*, **343**, 419–450, 2002.

[EGHa] Y. Eidelman, I. Gohberg, I. Haimovici, *Separable type representations of matrices and fast algorithms. Volume 1. Basics. Completion problems. Multiplication and inversion algorithms*, Birkhauser, 2013.

[EGHb] Y. Eidelman, I. Gohberg, I. Haimovici, *Separable type representations of matrices and fast algorithms. Volume 2. Eigenvalue method*, Birkhauser, 2013.

[F72] C. M. Fiduccia, Polynomial Evaluation via the Division Algorithm: The Fast Fourier Transform Revisited, *Proc. 4th Annual ACM Symp. on Theory of Computing (STOC'72)*, 88–93, 1972.

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

[GI88] W. Gautschi, G. Inglese, Lower Bounds for the Condition Number of Vandermonde Matrices, *Numerische Mathematik*, **52**, 241–250, 1988.

[GKK85] I. Gohberg, T. Kailath, I. Kailath, Linear Complexity Algorithms for Semiseparable Matrices, *Integral Equations and Operator Theory*, **8, 6**, 780–804, 1985.

[GGS87] A. Gerasoulis, M. D. Grigoriadis, L. Sun, A Fast Algorithm for Trummer's Problem, *SIAM Journal on Scientific and Statistical Computing*, **8, 1**, 135–138, 1987.

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

[GL96] G. H. Golub, C. F. Van Loan, *Matrix Computations*, 3rd edition, The Johns Hopkins University Press, Baltimore, Maryland, 1996.

[GR87] L. Greengard, V. Rokhlin, A Fast Algorithm for Particle Simulation, *Journal of Computational Physics*, **73**, 325–348, 1987.

[GS66] W. Gentelman, G. Sande, Fast Fourier Transform for Fun and Profit, *Full Joint Comput. Conference*, **29**, 563–578, 1966.

[H72] E. Horowitz, A Fast Method for Interpolation Using Preconditioning, *Information Processing Letters*, **1, 4**, 157–163, 1972.

[KZ08] S. Köhler, M. Ziegler, On the Stability of Fast Polynomial Arithmetic, *Proc. 8-th Conference on Real Numbers and Computers* (J.D. Bruguera, M. Daumas, eds.), 147–156, 2008.

[LRT79] R. J. Lipton, D. Rose, R. E. Tarjan, Generalized Nested Dissection, *SIAM J. on Numerical Analysis*, **16**, **2**, 346–358, 1979.

[MB72] R. Moenck, A. Borodin, Fast Modular Transform via Division, *Proceedings of 13th Annual Symposium on Switching and Automata Theory*, 90–96, IEEE Computer Society Press, Washington, DC, 1972.

[MRT05] P. G. Martinsson, V. Rokhlin, M. Tygert, A fast algorithm for the inversion of Toeplitz matrices, *Comput. Math. Appl.*, **50**, 741–752, 2005.

[P64] F. Parker, Inverses of Vandermonde matrices, *Amer. Math. Monthly*, **71**, 410–411, 1964.

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

[P93] V. Y. Pan, Parallel Solution of Sparse Linear and Path Systems, in *Synthesis of Parallel Algorithms* (J.H. Reif, editor), Chapter 14, pp. 621–678, Morgan Kaufmann publishers, San Mateo, California (1993).

[P95] V. Y. Pan, An Algebraic Approach to Approximate Evaluation of a Polynomial on a Set of Real Points, *Advances in Computational Mathematics*, **3**, 41–58, 1995.

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

[P11] 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.

[P13] V. Y. Pan, Polynomial Evaluation and Interpolation and Transformations of Matrix Structures, *Proceedings of CASC 2013* (V.P. Gerdt et al. editors), *Lecture Notes in Computer Science*, **8136**, Springer, Heidelberg (2013, in print).

[Pa] V. Y. Pan, Transformations of Matrix Structures Work Again, Tech. Report TR 2013004, *PhD Program in Comp. Sci.*, *Graduate Center, CUNY*, 2013

Available at http://www.cs.gc.cuny.edu/tr/techreport.php?id=446

[PR93] V. Y. Pan, J. Reif, Fast and Efficient Parallel Solution of Sparse Linear Systems, *SIAM J. on Computing*, **22**, **6**, 1227–1250, 1993.

[PRT92] V. Y. Pan, J. H. Reif, S. R. Tate, The Power of Combining the Techniques of Algebraic and Numerical Computing: Improved Approximate Multipoint Polynomial Evaluation and Improved Multipole Algorithms, *33th Annual IEEE Symposium on Foundations of Computer Science (FOCS'92)*, 703–713, IEEE Computer Societty Press, 1992.

[PSLT93] V. Y. Pan, A. Sadikou, E. Landowne, O. Tiga, A New Approach to Fast Polynomial Interpolation and Multipoint Evaluation, *Computers and Math. (with Applications)*, **25, 9**, 25–30, 1993.

[PZHY97] V. Y. Pan, A. Zheng, X. Huang, Y.Yu, Fast Multipoint Polynomial Evaluation and Interpolation via Computation with Structured Matrices, *Annals of Numerical Math.*, **4**, 483–510, 1997.

[R85] V. Rokhlin, Rapid Solution of Integral Equations of Classical Potential Theory, *Journal of Computational Physics*, **60**, 187–207, 1985.

[S73] V. Strassen, Die Berechnungskomplexetät von elementarysymmetrischen Funktionen und von Interpolationskoeffizienten, *Numerische Mathematik*, **20, 3**, 238–251, 1973.

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

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

[VVGM05] R. Vandebril, M. Van Barel, G. Golub, N. Mastronardi, A Bibliography on Semiseparable Matrices, *Calcolo*, **42, 3–4,** 249–270, 2005.

[VVM07] R. Vandebril, M. Van Barel, N. Mastronardi, *Matrix Computations and Semiseparable Matrices: Linear Systems* (Volume 1), The Johns Hopkins University Press, Baltimore, Maryland, 2007.

[VVM08] R. Vandebril, M. Van Barel, N. Mastronardi, *Matrix Computations and Semiseparable Matrices: Eigenvalue and Singular Value Methods* (Volume 2), The Johns Hopkins University Press, Baltimore, Maryland, 2008.

[X12] J. Xia, On the Complexity of Some Hierarchical Structured Matrix Algorithms, *SIAM J. Matrix Anal. Appl.*, **33**, 388–410, 2012.

[X13] J. Xia, Randomized sparse direct solvers, *SIAM J. Matrix Anal. Appl.*, **34**, 197–227, 2013.

[XXCBa] J. Xia, Y. Xi, S. Cauley, and V. Balakrishnan, Superfast and stable structured solvers for Toeplitz least squares via randomized sampling, *SIAM J. Matrix Anal. and Applications*, in press.

[XXG12] J. Xia, Y. Xi, M. Gu, A superfast structured solver for Toeplitz linear systems via randomized sampling, *SIAM J. Matrix Anal. Appl.*, **33**, 837–858, 2012.