SCIENCE CHINA
Mathematics

• ARTICLES •

January 2022 Vol. 65 No. 1: 203–220
https://doi.org/10.1007/s11425-018-1711-x

Optimal orthogonalization processes

Marko Huhtanen* & Pauliina Uusitalo

Department of Electrical and Information Engineering, University of Oulu, Oulu 90570, *Finland*

Email: Marko.Huhtanen@aalto.fi, Pauliina.Uusitalo@oulu.fi

Received November 13, 2018; accepted June 5, 2020; published online August 22, 2020

Abstract Two optimal orthogonalization processes are devised to orthogonalize, possibly approximately, the columns of a very large and possibly sparse matrix $A \in \mathbb{C}^{n\times k}$. Algorithmically the aim is, at each step, to optimally decrease nonorthogonality of all the columns of A. One process relies on using translated small rank corrections. Another is a polynomial orthogonalization process for performing the Löwdin orthogonalization. The steps rely on using iterative methods combined, preferably, with preconditioning which can have a dramatic effect on how fast the nonorthogonality decreases. The speed of orthogonalization depends on how bunched the singular values of A are, modulo the number of steps taken. These methods put the steps of the Gram-Schmidt orthogonalization process into perspective regarding their (lack of) optimality. The constructions are entirely operator theoretic and can be extended to infinite dimensional Hilbert spaces.

Keywords optimal orthogonalization, sparse matrix, Gram-Schmidt orthogonalization, Löwdin orthogonalization, polynomial orthogonalization, implicit orthogonalization, preconditioning, Gram matrix, frame inequality

MSC(2020) 65T99, 65F25, 65F50

Citation: Huhtanen M, Uusitalo P. Optimal orthogonalization processes. Sci China Math, 2022, 65: 203–220, https://doi.org/10.1007/s11425-018-1711-x

1 Introduction

This paper is concerned with optimal orthogonalization, possibly approximate, of the columns of a very large and possibly sparse matrix $A \in \mathbb{C}^{n\times k}$ with linearly independent columns. Preconditioning is investigated to speed up orthogonalization processes. Denote by $\mathcal{Q}$ the set of n-by-k matrices with orthonormal columns. By interpreting the steps of the Gram-Schmidt process as translated rank-one corrections, among such methods they are far from optimal. For optimality, consider

$$\min_{X\in\mathcal{X},\, Q\in\mathcal{Q}} \|AX - Q\|, \tag{1.1}$$

where the subset $\mathcal{X} \subset \mathbb{C}^{k\times k}$ is determined by the constraints that X should satisfy. It is assumed that $\mathcal{X}$ is set in such a way that the minimum exists. There should also exist a parameter j to increase so as to improve approximations. Moreover, for moderate values of j, matrix-vector products with X should be inexpensive. We are primarily interested in translated small rank corrections $\mathbb{C}I + \mathcal{F}_j$, where $\mathcal{F}_j \subset \mathbb{C}^{k\times k}$ denotes the set of matrices of rank j at most with $j \ll n$. Also a polynomial

* Corresponding author

orthogonalization scheme is devised to perform a sparse Löwdin orthogonalization of quantum chemistry [25,26]. In both cases the aim is, at each step, to optimally decrease nonorthogonality of all the columns of A. Based on iterative methods, preferably combined with preconditioning, the schemes are parallelizable. Orthogonalization may also take place implicitly, i.e., to retain sparsity, AX need not be explicitly formed. Aside from computational harmonic analysis problems involving families of almost orthogonal wavelets (Riesz sequences), the task arises naturally in preconditioning for the normal equations (see [20] for polynomial preconditioning, i.e., polynomial orthogonalization). The problem occurs also in finite frame theory in connection with the so-called frame inequality (see Example 3.3). In finite element method (FEM) discretizations there is a balance; it is not acceptable to use elements giving rise to a very ill-conditioned basis, although full orthogonality is not the goal either[1)]. For approximate orthogonalization deliberately avoiding the Gram-Schmidt process in signal processing applications, see [40].

Let $\mathcal{X} = \mathbb{C}I + \mathcal{F}_j$. Then the speed of orthogonalization, as a function of the number of steps j, is shown to depend on appropriately translated extreme singular values $\sigma_j(A)$ of A such that (1.1) equals

$$\omega_j(A) = \min_{r\in\mathbb{R}} \max\{|1-\sigma_{j+1}(rA)|,\, |1-\sigma_{k-j}(rA)|\}$$

for $j = 0, 1, \ldots, \lfloor \frac{k-1}{2} \rfloor$. This is a non-increasing sequence such that the associated orthogonalization process requires, unlike the Gram-Schmidt process, computational information about all the columns of A. Besides optimality, it also retains locality since $2j$ columns get orthonormalized after j steps. (In the Gram-Schmidt process j columns of A get orthonormalized after j steps.) In particular, zero is attained after at most $j = \lfloor \frac{k}{2} \rfloor$ steps with an explicit connection to the Löwdin orthogonalization. For the Löwdin orthogonalization and its applications, see, e.g., [9, 12, 27, 36, 40] and the references therein. By requiring computing right singular vectors, in large scale problems the steps should rely on iterative methods combined, preferably, with preconditioning. This is affordable if matrix-vector products with A and its Hermitian transpose are inexpensive. In the square matrix case the orthogonalization process is symmetric in the sense that $\omega_j(A^*) = \omega_j(A)$ holds, i.e., the speed is the same for the column and row orthogonalization.

Let $\mathcal{X} = \{p(A^*A) \mid \deg(p) \leqslant j-1\}$. This provides a natural option to polynomially perform an optimal incomplete orthogonalization (1.1). An approximate Löwdin orthogonalization can be performed by solving a positive polynomial approximation problem on the spectrum of the Gram matrix A^*A. In particular, this determines the speed of the process. This alternative is attractive also for preconditioning the case $\mathcal{X} = \mathbb{C}I + \mathcal{F}_j$. Both of these processes described are unitarily invariant, admitting an operator theoretic interpretation in a natural manner. This is of importance since in applications the matrix A typically results from a discretely sampled infinite dimensional system. Then, in large dimensions, the speed of orthogonalization of these schemes depend to a lesser extent on the dimension of the discretization. Moreover, $\mathcal{X} = \{p(A^*A) \mid \deg(p) \leqslant j-1\}$ can be regarded as providing a more general scheme, allowing orthogonalizing infinite dimensional systems without any immediate barriers. As opposed to this, $\mathcal{X} = \mathbb{C}I + \mathcal{F}_j$ requires certain compactness assumptions, natural though.

These orthogonalization processes can be accelerated by preconditioning. Preconditioning orthogonalization is needed, e.g., in [31]. Denote the preconditioner by $M \in \mathbb{C}^{k\times k}$. The aim is then to attain a faster decrease in nonorthogonality with the matrix AM. Since the proposed algorithms rely on performing matrix-vector products, explicitly forming AM is not necessary. For an obvious and very inexpensive choice, take M to be the diagonal scaling of the columns of A. (Scaling is a classical technique (see [13, 24, 38] and the references therein).) This can be improved by extending it, e.g., to sparse upper triangular matrices. We propose formulating the task of finding the QR factorization as an optimization problem. This can then be modified to compute an approximate upper triangular factor in an optimal way. There are also other alternatives. We show with an example how in FEM discretizations circulant-like preconditioning can result in tremendous speed-ups in orthogonalization.

The rest of the paper is organized as follows. Section 2 starts with a description of how constraints

[1)] Full orthogonality typically spoils sparsity of matrices. In addition, since the dimensions can be huge, full orthogonalization by using the Gram-Schmidt process is unrealistic in practice.

in orthogonalization processes arise. Thereafter, bearing in mind the Löwdin and Gram-Schmidt orthogonalizations, optimal orthogonalization processes are devised and algorithms are derived. In Section 3, applications and examples are presented. Associated operator theoretic remarks are made. In Section 4, options to preconditioning are addressed, primarily in terms of scaling and an optimal formulation of an approximate QR factorization. Numerical experiments are conducted in Section 5. Finally, Section 6 concludes the paper.

2 Optimal orthogonalization processes

Consider the problem (1.1) of approximate orthogonalization of the columns of a large and possibly sparse matrix $A \in \mathbb{C}^{n\times k}$ with linearly independent columns. (See [3] for the problem of having a sparse basis matrix.) Although we deal with matrices, the constructions that follow extend to infinite dimensions, i.e., let

$$\{h_1, \dots, h_k\} \subset \mathcal{H},$$

where $\mathcal{H}$ is a separable complex Hilbert space equipped with an inner product $(\cdot, \cdot)$. For orthogonalization, consider the "matrix"

$$A = [\, h_1 \ \cdots \ h_k \,], \tag{2.1}$$

which can be regarded as a linear map from $\mathbb{C}^k$ to $\mathcal{H}$. The adjoint operator is then

$$A^* = \begin{bmatrix} (\cdot, h_1) \\ \vdots \\ (\cdot, h_k) \end{bmatrix}.$$

These suffice to carry out the computations analogously. In practice k can be huge. For example, in FEM discretizations of 3D problems, the number of FEM basis elements can easily be hundreds of millions. Then methods such as Gram-Schmidt are not really applicable for full orthogonalization.

2.1 Constraints of orthogonalization

Regardless of the method to attain zero in (1.1), the solution X must satisfy the following constraints. Here, $\mathcal{Q}$ denotes the set of n-by-k matrices with orthonormal columns.

Proposition 2.1. *Suppose the columns of $A \in \mathbb{C}^{n\times k}$ are linearly independent. If $AX = Q$ with $Q \in \mathcal{Q}$, then*

$$\|X\| = \frac{1}{\sigma_k(A)} \quad \textit{and} \quad \|X\|_F = \sqrt{\sum_{l=1}^{k} \frac{1}{\sigma_l(A)^2}} \tag{2.2}$$

in the spectral and Frobenius norms. Moreover, $\{X \in \mathbb{C}^{k\times k} \mid AX = Q \text{ with } Q \in \mathcal{Q}\}$ is a connected compact set of the real dimension k^2.

Proof. We have $AX = Q$ with orthonormal columns, i.e., if $A = Q_1 \Sigma Q_2^*$ is the singular value decomposition of A, then $\Sigma Y = Q_1^* Q$ with $Y = Q_2^* X$. This forces the last $n - k$ rows of $Q_1^* Q$ to be zero. Solving for Y and using the unitary invariance of the spectral and Frobenius norms yields the claim.

Regarding the dimension of the solution set, let q_j for $j = 1, \dots, n$ denote the columns of Q_1. Since the last $n - k$ rows of $Q_1^* Q$ are zero, $Q = [q_1 \cdots q_k]V$ with $V \in \mathbb{C}^{k\times k}$ unitary. Any unitary V yields a solution

$$X = Q_2 \tilde{\Sigma}^{-1} V, \tag{2.3}$$

where $\tilde{\Sigma} \in \mathbb{R}^{k\times k}$ consists of the first k rows of Σ. Hence the real dimension is k^2, the dimension of k-by-k unitary matrices. Consider the representation (2.3) of the orthogonalizations of A. Connectedness and compactness follow from the continuity of the associated map $V \mapsto Q_2 \tilde{\Sigma}^{-1} V$. □

In this sense we are dealing with a compact factorization problem. For non-compact factoring, see [18] where the problem formulation is conceptually similar except that the dimension of the solution set is strongly problem dependent.

Practical algorithms have led to some elegant additional constraints, i.e., the QR factorization[2)] and the polar decomposition[3)] are the best-known alternatives to solve the problem. (For the QR factorization and Gram-Schmidt orthogonalization process, see [2,22]. See also [29, Chapter 4] for a concise introduction to applications of orthogonality.) Then $\mathcal{X}$ is of real dimension k^2 consisting of upper triangular matrices with non-negative diagonal entries and positive definite matrices, respectively. The associated orthogonalizations are Gram-Schmidt and Löwdin.

In very large scale problems (typically discretizations of infinite dimensional problems) constraints arise in abundance because of computational complexity, i.e., then $\mathcal{X}$ can end up being chosen such that zero is not attained in (1.1), so that the columns of A can get only incompletely orthogonalized. There are typically two reasons for this. First, it may be too time-consuming to perform a complete orthogonalization, i.e., to complete the task within a certain time frame, one must accept a nonorthogonal set of vectors. Second, for the sake of storage limitations, X may be allowed to have only $O(k)$ nonzero entries. These both take place in preconditioning [33, Subsection 10.8.3]. For another example, retaining sparsity is the reason for not using orthonormal bases in discretizing PDE with FEM, imposing restrictions on how to compute X.

These remarks give rise to the following sparse orthogonalization problem (see also [3]).

Definition 2.2. Let $\mathcal{X} = \mathbb{C}^{k\times k}$. Then the sparsest X yielding zero in (1.1) is said to be a sparsest orthogonalization of the columns of $A \in \mathbb{C}^{n\times k}$.

Solving this problem appears difficult. Since there is a lot of freedom to define $\mathcal{X}$, more accessible alternatives exist requiring only $O(k)$ parameters to fully orthogonalize the columns of A. One such orthogonalization is the polynomial orthogonalization process for square matrices devised in [20]. It also admits performing an incomplete orthogonalization. Another polynomial process is described in the section that follows.

2.2 Orthogonalization with respect to $\mathcal{X} = \{p(A^*A) \mid \deg(p) \leqslant j-1\}$

Let us first focus on the Löwdin orthogonalization [25] since it admits a polynomial interpretation for computing approximations in a natural way. Related with Hermitian orthogonalizations, Löwdin [25] designed his method to solve the matrix nearness problem on the left-hand side in (2.4).

The following is well known where $\sigma_1(A) \geqslant \cdots \geqslant \sigma_k(A) \geqslant 0$ denote the singular values of $A \in \mathbb{C}^{n\times k}$ (see, e.g., [17]).

Proposition 2.3. *Let $A \in \mathbb{C}^{n\times k}$. Then in the spectral norm,*

$$\min_{Q\in\mathcal{Q}} \|A-Q\| = \max\{|1-\sigma_1(A)|, |1-\sigma_k(A)|\}. \tag{2.4}$$

Proof. Let $A = Q_1\Sigma Q_2^*$ be the singular value decomposition of A. By the unitary invariance of the spectral norm, $\min_{Q\in\mathcal{Q}} \|A-Q\| = \min_{\hat{Q}\in\mathcal{Q}} \|\Sigma-\hat{Q}\|$ holds. Choosing $\hat{Q}$ such that its first k-by-k block is the identity matrix yields $\|A-Q\| = \max\{|1-\sigma_1(A)|, |1-\sigma_k(A)|\}$. On the other hand, Q is any n-by-k matrix with orthonormal columns, then $\|\Sigma - I\| \leqslant \|A-Q\|$ by [17, Theorem 7.4.9.1]. □

There exists a polynomial orthogonalization process to perform the Löwdin orthogonalization in terms of the Gram matrix $A^*A \in \mathbb{C}^{k\times k}$. It also provides a rapid way to optimally perform an incomplete orthogonalization as follows, where $\Lambda(M)$ denotes the eigenvalues of a matrix $M \in \mathbb{C}^{n\times n}$.

Theorem 2.4. *Assume $A \in \mathbb{C}^{n\times k}$ has linearly independent columns and let*

$$\mathcal{X} = \operatorname{span}\{I, A^*A, (A^*A)^2, (A^*A)^3, \ldots, (A^*A)^{j-1}\}. \tag{2.5}$$

[2)] Algorithmically the Gram-Schmidt orthogonalization process (or applications of Householder transformations) yields $A = QR$. Because of our problem formulation, this means $X = R^{-1}$.

[3)] With P being the positive definite polar factor of A, take $X = P^{-1}$.

Then (1.1) *in the spectral norm equals*

$$\min_{\deg(p)\leqslant j-1}\max_{\lambda\in\Lambda(A^*A)}|\sqrt{\lambda}p(\lambda)-1|. \tag{2.6}$$

Proof. Let $A=Q_1\tilde{\Sigma}Q_2^*$ be the reduced singular value decomposition of A, where $Q_1\in\mathbb{C}^{n\times k}$ with orthonormal columns and $Q_2\in\mathbb{C}^{k\times k}$ is unitary while $\tilde{\Sigma}\in\mathbb{R}^{k\times k}$ is diagonal with non-increasing nonnegative diagonal entries. To have an element of $\mathcal{X}$, take a polynomial p of degree $j-1$ at most. Then we have $p(A^*A)=Q_2p(\tilde{\Sigma}^2)Q_2^*$, so that

$$Ap(A^*A)=Q_1\tilde{\Sigma}p(\tilde{\Sigma}^2)Q_2^*. \tag{2.7}$$

Then (1.1) reads

$$\min_{\deg(p)\leqslant j-1,\,Q\in\mathcal{Q}}\|Q_1\tilde{\Sigma}p(\tilde{\Sigma}^2)Q_2^*-Q\|=\min_{\deg(p)\leqslant j-1,\,Q\in\mathcal{Q}}\|\tilde{\Sigma}p(\tilde{\Sigma}^2)-Q\|$$

by the unitary invariance of the spectral norm. The claim then follows by using Proposition 2.3 and the fact that the singular values are non-negative. □

Recall that the degree of a matrix $M\in\mathbb{C}^{n\times n}$ is the degree of the monic polynomial of the least degree annihilating M. Because of (2.7) and the freedom of choosing p, the real dimension of orthogonalizations of A with $\mathcal{X}$ is $\deg(A^*A)$. It seems natural to restrict to using positive polynomials[4)] since then the corresponding elements of $\mathcal{X}$ are positive definite.

Corollary 2.5. *There exists a polynomial p of degree* $\deg(A^*A)-1$ *such that $Ap(A^*A)$ has orthonormal columns. Choosing, in addition, p to be positive on $\Lambda(A^*A)$ yields the Löwdin orthogonalization of A.*

Proof. Let D be any diagonal unitary matrix. By the Lagrange interpolation, there exists a unique polynomial p of degree $\deg(A^*A)-1$ yielding $\tilde{\Sigma}p(\tilde{\Sigma}^2)=D$. Imposing p to be positive forces D to be the identity matrix, so that

$$Ap(A^*A)=Q_1Q_2^*, \tag{2.8}$$

the Löwdin orthogonalization of A. □

Generically, Hermitian orthogonalizations are based on $\mathcal{X}$ as follows.

Corollary 2.6. *Suppose* $\deg(A^*A)=k$. *If $AX=Q$ with $Q\in\mathcal{Q}$ and a Hermitian matrix X, then $X=p(A^*A)$ for a polynomial p.*

Proof. If X is Hermitian, then, by (2.3), $VQ_2\tilde{\Sigma}^{-1}=\tilde{\Sigma}^{-1}Q_2^*V^*$, so that $\tilde{\Sigma}VQ_2\tilde{\Sigma}^{-1}$ is unitary. Thereby

$$\tilde{\Sigma}VQ_2\tilde{\Sigma}^{-2}Q_2^*V^*\tilde{\Sigma}=I$$

which gives $VQ_2\tilde{\Sigma}^{-2}Q_2^*V^*=\tilde{\Sigma}^{-2}$, i.e., $VQ_2\tilde{\Sigma}^{-2}=\tilde{\Sigma}^{-2}VQ_2$. Since $\deg(A^*A)=k$, it follows that $V=DQ_2^*$ with a diagonal matrix D. By using (2.3) and the Lagrange interpolation, the claim follows. □

This is clearly a unitarily invariant orthogonalization process, i.e., the value of (1.1) is independent on applying A with unitary matrices from the left and right. In particular, it seems natural to call $\deg(A^*A)$ the degree of the basis (2.1). Moreover, since $\mathcal{X}$ is polynomially generated, the process is particularly well suited for producing an incomplete Löwdin orthogonalization in terms of (2.6). Then only some information about the singular values is needed, i.e., there is a lot flexibility to use incomplete data.

Example 2.7. In large scale problems $\Lambda(A^*A)$ cannot be assumed to be available. Instead, suppose we have $E\subset\mathbb{R}$ such that $\Lambda(A^*A)\subset E$. (For example, it is very inexpensive to estimate the largest and the smallest eigenvalue of the Gram matrix A^*A with the Hermitian Lanczos method. This does not require forming A^*A and consumes a fixed amount of storage.) Find p solving the polynomial approximation problem

$$\min_{\deg(p)\leqslant j-1}\max_{\lambda\in E}|\sqrt{\lambda}p(\lambda)-1|.$$

[4)] A positive polynomial on a given set is a polynomial whose values are positive on that set.

This can be accomplished by executing the Remez algorithm (see, e.g., [6, p. 478]). Then $Ap(A^*A)$ yields an incomplete Löwdin orthogonalization. Observe that it need not be explicitly formed. However, if it is formed, then A^*A should be sparse and p should be of moderate degree. (For polynomial evaluation, see [15, Chapter 5].)

If the degree of A^*A is small, then Corollary 2.5 provides a remarkably swift way to perform the Löwdin orthogonalization, i.e., $\Lambda(A^*A)$ is quickly found by executing the Hermitian Lanczos method, requiring storing only three vectors. Then (2.8), when not formed explicitly (store only the coefficients of p), does not consume any more storage than storing A. Finding the Fourier coefficients of a vector $x \in \mathbb{C}^k$ consists of computing the matrix-vector product $(Ap(A^*A))^*x$. This is impressive since, by the implicit orthogonalization process described, one can then reap all the benefits of optimal sparse matrix computations. For non-optimal iterations, requiring some assumptions on the Gram matrix, see [30]. For the Taylor expansions of Löwdin, see [31, Equation (10)].

2.3 Orthogonalization with respect to $\mathcal{X} = \mathbb{C}I + \mathcal{F}_j$

Let us now focus on making the Gram-Schmidt orthogonalization process optimal. This interpretation relies on assessing how far the columns of A are from being orthogonal under translated small rank corrections[5)]. In assessing optimality it is useful to bear in mind, aside from the optimality condition (1.1), the constraints (2.2) that any solution must satisfy.

Denote by $\mathcal{F}_j \subset \mathbb{C}^{k\times k}$ the set of matrices of rank j at most with $j \geqslant 1$.

Theorem 2.8. *Suppose $A \in \mathbb{C}^{n\times k}$ has linearly independent columns and let $\mathcal{X} = \mathbb{C}I + \mathcal{F}_j$ for $j \leqslant \lfloor \frac{k-1}{2} \rfloor$. Then* (1.1) *in the spectral norm equals*

$$\omega_j(A) = \min_{r\in\mathbb{R}} \max\{|1 - \sigma_{j+1}(rA)|,\, |1 - \sigma_{k-j}(rA)|\} \tag{2.9}$$

with a solution $X = r_jI + F_j$ satisfying $\|X\| = \frac{1}{\sigma_k(A)}$ and

$$\|X\|_F = \sqrt{\sum_{l=1}^{j}\left(\frac{1}{\sigma_l(A)^2} + \frac{1}{\sigma_{k-l+1}(A)^2}\right) + \frac{4(k-2j)}{(\sigma_{j+1}(A) + \sigma_{k-1}(A))^2}}$$

such that AX has $2j$ orthonormal columns. Moreover, zero is attained for $j = \lfloor \frac{k}{2} \rfloor$.

Proof. We have

$$\min_{\alpha\in\mathbb{C},\,\mathrm{rank}(F_j)\leqslant j,\,Q\in\mathcal{Q}} \|A(\alpha I + F_j) - Q\| \geqslant \min_{\alpha\in\mathbb{C},\,\mathrm{rank}(G_j)\leqslant j,\,Q\in\mathcal{Q}} \|\alpha A + G_j - Q\|. \tag{2.10}$$

(Now $G_j \in \mathbb{C}^{n\times k}$.) To solve the minimization problem on the right, we use the singular value decomposition of A. It is clear that α can be real and non-negative. For any fixed $r \in \mathbb{R}^+$ and G_j of rank j at most, the value of

$$\min_{Q\in\mathcal{Q}} \|rA + G_j - Q\| \tag{2.11}$$

is determined by Proposition 2.3. For $j \leqslant \lfloor \frac{k-1}{2} \rfloor$, apply now the singular value inequalities [16, Theorem 3.3.16(a)] to $rA + G_j$ to have

$$\sigma_{j+1}(rA) \leqslant \sigma_1(rA - G_j) + \sigma_{j+1}(G_j) = \sigma_1(rA - G_j)$$

and

$$\sigma_k(rA - G_j) \leqslant \sigma_{k-j}(rA) + \sigma_{j+1}(-G_j) = \sigma_{k-j}(rA).$$

[5)] Only when the Gram-Schmidt orthogonalization process is completed, the set of upper triangular matrices with non-negative diagonal entries is reached and available. When the process is under execution, one exclusively deals with $\mathbb{C}I + \mathcal{F}_j$ for $1 < j < k$.

Consequently, no matter how r and G_j are chosen,

$$\begin{aligned}\max\{|1-\sigma_1(rA-G_j)|,|1-\sigma_k(rA-G_j)|\} &\geqslant \max\{|1-\sigma_{j+1}(rA)|,|1-\sigma_{k-j}(rA)|\}\\ &\geqslant \min_{r\in\mathbb{R}}\max\{|1-\sigma_{j+1}(rA)|,|1-\sigma_{k-j}(rA)|\}\end{aligned}$$

holds. Next, it is shown that there exist choices such that the equalities hold.

Choose first $r_j\in\mathbb{R}^+$ such that the right-hand side of (2.9) is attained, i.e., by imposing $1-r_j\sigma_{k-j}(A)=r_j\sigma_{j+1}(A)-1$ we obtain

$$r_j=\frac{2}{\sigma_{j+1}(A)+\sigma_{k-j}(A)}. \tag{2.12}$$

Next, let us construct G_j. To this end, let $A=Q_1\Sigma Q_2^*$ be the singular value decomposition of A with $\Sigma\in\mathbb{R}^{n\times k}$ having the (l,l) entries $\sigma_l(A)$, for $l=1,\ldots,k$. By the unitary invariance of the spectral norm, consider

$$\min_{Q\in\mathcal{Q}}\|r_jA+G_j-Q\|=\min_{Q\in\mathcal{Q}}\|r_j\Sigma+\hat{G}_j-Q\| \tag{2.13}$$

with $\hat{G}_j=Q_1^*G_jQ_2\in\mathbb{C}^{n\times k}$. To have $\hat{G}_j$, we resort to the construction devised in [19, Proposition 2.1] (see Appendix A). Choose $u_l\in\mathbb{C}^n$ all zeros except the l-th entry to be

$$-r_j\sqrt{1-\frac{1}{r_j^2\sigma_l^2}}(\sqrt{\sigma_l^2-\sigma_{k-l+1}^2}+\mathrm{i}\sigma_{k-l+1})$$

and the $(k-l+1)$-th entry to be

$$-\mathrm{i}\sqrt{1-r_j^2\sigma_{k-l+1}^2}.$$

Choose $v_l\in\mathbb{C}^k$ all zeros except the l-th entry to be

$$\frac{1}{r_j}\sqrt{\frac{r_j^2\sigma_l^2-1}{\sigma_l^2-\sigma_{k-l+1}^2}}$$

and the $(k-l+1)$-th entry to be

$$\frac{1}{r_j}\sqrt{\frac{1-r_j^2\sigma_{k-l+1}^2}{\sigma_l^2-\sigma_{k-l+1}^2}}.$$

Then put $\hat{G}_j=\sum_{l=1}^j u_lv_l^*$. With this choice, the $2j$ singular values of $r_j\Sigma+\hat{G}_j$ are equal to one and the remaining singular values are $r_j\sigma_{j+1}(A),r_j\sigma_{j+2}(A),\ldots,r_j\sigma_{k-j}(A)$. Consequently, choose $G_j=Q_1\hat{G}_jQ_2^*$. Now with these choices, we have shown that (2.11) equals the right-hand side of (2.9). For k odd, this is zero for $j=\lfloor\frac{k-1}{2}\rfloor=\lfloor\frac{k}{2}\rfloor$. For k even and $j=\lfloor\frac{k}{2}\rfloor=\frac{k}{2}$, take r_j as in (2.12). Then repeat the construction to have $\hat{G}_j$ giving zero in (2.13).

Next, we need to recover F_j leading to the equality in (2.10) while r_j is as before in (2.12). We get this by imposing $AF_j=G_j=Q_1\hat{G}_jQ_2^*$, so that $\Sigma Q_2^*F_j=\hat{G}_jQ_2^*$. The last $n-k$ rows of Σ and $\hat{G}_j$ are zeros. Denote by $\tilde{\Sigma}$ and $\tilde{G}_j$ the k-by-k matrices coinciding with the first k rows of Σ and $\hat{G}_j$. This allows us to determine F_j by setting

$$F_j=Q_2\tilde{\Sigma}^{-1}\tilde{G}_jQ_2^* \tag{2.14}$$

and thus establishing the equality in (2.10).

Denote by $e_1,\ldots,e_k$ the standard basis vectors of $\mathbb{C}^k$. By regarding the claim concerning the norm of $X=r_jI+F_j$, the construction is done such that the matrix $A(r_jI+F_j)$ is isometry when restricted to

$$Q_2\mathrm{span}\{e_1,e_2,\ldots,e_j,e_{k-j+1},e_{k-j+2},\ldots,e_k\} \tag{2.15}$$

and unitarily equivalent to $r_j\Sigma$ when restricted to the orthogonal complement of (2.15). From this the claim follows. □

Since A is assumed to have linearly independent columns, we have

$$1 > \omega_j(A) \geqslant 0$$

for any $j = 1, 2, \ldots$ Moreover, in the square matrix case we have $\omega_j(A^*) = \omega_j(A)$, i.e., the speed is the same for the column and row orthogonalization. In practice, the speed of decay of $\omega_j(A)$ is, of course, the primary object of interest. A swift decay allows quickly attaining almost orthonormality.

The proof of this theorem is constructive such that the solution $X = r_j I + F_j$ can be found in a numerically stable manner, i.e., by (2.14), the rank-j matrix F_j is built from the extreme $2j$ right singular vectors of A corresponding to the j largest and the j smallest singular values. Due to the factorization (2.14), we only need to collect these vectors and $4j$ scalars to store $\tilde{\Sigma}^{-1}\tilde{G}_j$. This information is sufficient to recover F_j (see Algorithm 1).

Algorithm 1 Compute $X = r_j I + F_j$

1: Read n-by-k matrix A with linearly independent columns.
2: Compute the right singular pairs $(\sigma_1(A), q_1), \ldots, (\sigma_{j+1}(A), q_{j+1})$ and $(\sigma_{k-j}(A), q_{k-j}), \ldots, (\sigma_k(A), q_k)$.
3: Set $r_j = \frac{2}{\sigma_{j+1}(A)+\sigma_{k-j}(A)}$
4: **for** $l = 1, \ldots, j$ **do**
5: $\quad u_l = \frac{-r_j}{\sigma_l}\sqrt{1 - \frac{1}{r_j^2\sigma_l^2}}(\sqrt{\sigma_l^2 - \sigma_{k-l+1}^2} + \mathrm{i}\sigma_{k-l+1})e_l + \frac{-\mathrm{i}}{\sigma_{k-l+1}}\sqrt{1 - r_j^2\sigma_{k-l+1}^2}e_{2j-l+1}$
6: $\quad v_l = \frac{1}{r_j}\sqrt{\frac{r_j^2\sigma_l^2 - 1}{\sigma_l^2 - \sigma_{k-l+1}^2}}e_l + \frac{1}{r_j}\sqrt{\frac{1 - r_j^2\sigma_{k-l+1}^2}{\sigma_l^2 - \sigma_{k-l+1}^2}}e_{2j-l+1}$
7: **end for**
8: set $F_j = [q_1 \cdots q_j\, q_{k-j+1} \cdots q_k] \sum_{l=1}^{j} u_l v_l^* [q_1 \cdots q_j\, q_{k-j+1} \cdots q_k]^*$.

The associated factorization of A reads

$$A = \hat{Q}Y, \tag{2.16}$$

where $\hat{Q}$ denotes the approximate orthogonalization

$$\hat{Q} = AX = A(r_j I + F_j)$$

and $Y = X^{-1} \in \mathcal{X}$. (Observe that the inverse of X is readily computable, by applying the Sherman-Morrison-Woodbury formula, whenever j is moderate.) The first j and the last j columns of $\hat{Q}$ are orthonormal. Otherwise the columns of $\hat{Q}$ are almost orthonormal as soon as $\omega_j(A)$ reaches a given tolerance. Moreover, whenever A is assumed to be sparse, the approximate orthogonalization $\hat{Q}$ should not be explicitly formed, i.e., the resulting basis vectors consisting of the columns of the matrix $A(r_j I + F_j)$ can be expected to be dense. Basis vectors can always be recovered by performing matrix-vector products with standard basis vectors of $\mathbb{C}^k$.

Corollary 2.9. *There exist $r \in \mathbb{R}$ and $F_j \in \mathcal{F}_j$ with $j \leqslant \lfloor \frac{k}{2} \rfloor$ such that*

$$A(rI + F_j)$$

has orthonormal columns.

The equality $A(r_j I + F_j) = Q_1(r_j\tilde{\Sigma} + \tilde{G}_j)Q_2^*$ then provides an explicit connection of this orthogonalization process with the Löwdin orthogonalization (2.8) of A.

3 Applications and operator theoretic remarks

Let us consider examples and make some operator theoretic comments.

Example 3.1. The factorization (2.16) can actually be redundant. For example, if AX has (approximately) orthonormal columns, then the least squares problem

$$\min_{x \in \mathbb{C}^k} \|Ax - b\|^2$$

is (approximately) solved by $x = XX^*A^*b$. In particular, now inverting X is not needed. For example, suppose we have the orthogonalization (2.8) available. Then $x = p(A^*A)^2A^*b$.

In large scale problems it is not uncommon that A results from a discretization of an infinite dimensional system. For example, in certain applications one deals with the Gaussian basis functions. (For applications in quantum chemistry, see, e.g., [32, pp. 265–266].) If the centers of the Gaussian basis functions are well separated, one can expect, because of locality, to have a swift decay of $\omega_j(A)$.

Example 3.2. In applications such as computational quantum chemistry, there arise generalized eigenvalue problems

$$Mx = \lambda Nx, \tag{3.1}$$

where $M \in \mathbb{C}^{n\times n}$ is Hermitian and $N \in \mathbb{C}^{n\times n}$ is the so-called overlap matrix, i.e., a positive definite Gram matrix collecting the inner products of the basis used (see [35, pp. 137–138] or [37, Chapter 3.1]). For sparsity issues, see [32]. Since there are good reasons to aim at using orthonormal bases [31], or almost so, one can expect some near unitarity of N (see, e.g., [27, 28]). In fact, it may even be that $N = U + P$, where the unitary (the so-called zero-order part) matrix U is available. Consequently, suppose Algorithm 1 yields, for some moderate j, a nearly unitary NX. Then (3.1) can be converted into an equivalent standard eigenvalue problem $X^*N^*MXy = \lambda y$ on which iterative methods can be executed.

Example 3.3. In finite frame theory [5, 8] one deals with the matrices $B \in \mathbb{C}^{k\times n}$ satisfying $n \geqslant k$ such that the columns of B span $\mathbb{C}^k$. If we denote B^* by A, then the classical

$$\sigma_k(A)\|x\|^2 \leqslant \|Ax\|^2 \leqslant \sigma_1(A)\|x\|^2$$

is called the frame inequality. A problem is to make the frame inequality "tighter" which, in our context, means solving (1.1) and putting X^*B. Using the so-called frame operator to this end means taking X to be the Löwdin orthogonalization of A (see, e.g., [8, p. 8]). As we have illustrated, there are many other alternatives to make the frame B tighter.

Unlike the Gram-Schmidt process, Algorithm 1 yields a unitarily invariant orthogonalization process, i.e., it holds that $\omega_j(UAV) = \omega_j(A)$ for any unitary matrices $U \in \mathbb{C}^{n\times n}$ and $V \in \mathbb{C}^{k\times k}$. It is noteworthy that the steps of the Gram-Schmidt process can also be expressed in terms of translated rank-one corrections. These corrections are particularly local and hence non-optimal as follows.

Example 3.4. The first j steps of the Gram-Schmidt process can be expressed as a factorization

$$A = \hat{Q}_j\hat{R}_j$$

of A, where the first j columns of $\hat{Q}_j$ are orthonormal while the remaining columns coincide with the columns of A. Then $\hat{R}_j = I + T_j$, where the upper-triangular matrix T_j can be written as $T_j = \sum_{l=1}^{j} t_l e_l^*$, where e_l denotes the l-th standard basis vector. Consequently, $A(I + F_j) = \hat{Q}_j$ with $F_j \in \mathcal{F}_j$. It is clear that $\hat{Q}_j$ cannot be expected to be near $\mathcal{Q}$ for $j < k$.

Example 3.5. The difference between Algorithm 1 and the Gram-Schmidt process can be underscored by taking $A = \alpha Q + F$ with $\alpha \in \mathbb{C}$, where Q has orthonormal columns and F is of rank l with $l \ll k$. Then, generically, the Gram-Schmidt process requires k steps to reach $\mathcal{Q}$. The optimal orthogonalization requires just at most $2l$ steps by the fact that $A^*A = \alpha I + U^*F + F^*(U + F)$, so that $\omega_{2l}(A) = 0$. This is actually not an artificial construction. For small rank unitary perturbations and their appearance in physics, see [7, 23] and the references therein.

Lack of optimality of the Gram-Schmidt process can also be seen in the overall number of steps with Algorithm 1 (see Corollary 2.9). Regarding the number of parameters, recall that k-by-k matrices of rank r at most is an affine variety of dimension $r(2k - r)$ which for $r = \frac{k}{2}$ yields $\frac{3}{4}k^2$, i.e., of real dimension $\frac{3}{2}k^2$.

Consider (2.1) with $k = \infty$, i.e., there are operator theoretic interpretations associated with the orthogonalizations schemes corresponding to the families

$$\mathcal{X} = \{p(A^*A) \mid \deg(p) \leqslant j - 1\} \tag{3.2}$$

and

$$\mathcal{X} = \mathbb{C}I + \mathcal{F}_j. \tag{3.3}$$

This is important when A is a discretization of an infinite dimensional system, i.e., these interpretations can be expected to set limits to what is attainable when dimensions grow. With (3.2) there are no immediate barriers to produce approximations, i.e., under some natural assumptions, the polar factor is used to solve the problem of orthogonalization (see [21]). As opposed to positive definite operators, the family (3.3) extends in infinite dimensions to scalar-plus-compact operators. Scalar-plus-compact operators are very important in operator theory, on Banach spaces in particular [1]. In orthogonalization it then seems that one should be dealing with bounded linear operators on a separable Hilbert space representable as $\alpha Q+K$, where $\alpha \in \mathbb{C}$ is nonzero, Q is a partial isometry[6)] and K is a compact operator. Only these can be regarded as being orthogonalizable under this process, with the speed depending on the structure of the compact part and, thereby, to a lesser extent on the dimension of the discretization. If a system is not initially representable as $\alpha Q+K$, it should be preconditioned so as to satisfy this condition. (It is noteworthy that in computational physics, orthogonalization problems are occasionally expressed as perturbations of unitary operators (see, e.g., [27, Chapter 4]).) This means that the Gram-Schmidt process is limited accordingly. However, since the Gram-Schmidt process operates entirely locally, it may not decrease the overall nonorthogonality of the system in a finite number of steps[7)]. This explains the popularity and importance of the polar factor in infinite dimensions where, moreover, the notion of "upper triangular matrix" makes no sense unless an orthonormal basis, or a basis in the first place, is available. In addition, that is very rarely the case in realistic problems. As opposed to this, (3.2) and (3.3) give rise to basis independent orthogonalization processes.

4 Preconditioning orthogonalization

Preconditioning is a central operation to speed up computational processes. As is well known, preconditioning is typically the most critical part of solving any realistic practical problem. Next, we consider ways of achieving this for orthogonalization. (This is needed, e.g., in [31].) This means replacing A with AM, where $M \in \mathbb{C}^{k\times k}$ is constructed with the aim at attaining a faster decrease in nonorthogonality with AM. The following example illustrates this well. It also shows in which way an FEM basis can occasionally be viewed as being "almost orthogonal", i.e., easily orthogonalizable with the techniques proposed.

Example 4.1. In the eigenvalue problem (3.1), the matrix N is typically a Gram matrix. Consider the case of an interval, let us say, $[0,1]$, so that the elements of N are the L^2 inner-products $(h_j,h_i) = \int_0^1 h_j(x)\overline{h_i}(x)dx$ of the FEM basis

$$\{h_1,\dots,h_k\} \tag{4.1}$$

used (see (2.1)). A natural question is, how far is this basis from being orthonormal? In particular, can this basis be orthonormalized by using $O(k)$ number of parameters? Occasionally there is some sort of circulant plus small rank structure involved. In the linear C^0 B-spline discretization we have

$$N = \frac{1}{6k}\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix}.$$

6) A bounded linear operator on a complex separable Hilbert space is a partial isometry if it preserves the norms of vectors orthogonal to its kernel. Such operators appear in the extension of polar decomposition to Hilbert space operators.

7) We measure orthogonalization using the norm topology among operators. It is, of course, possible to consider weaker topologies to get "local convergence" results [14].

If we put the $(1,k)$ and $(k,1)$ entries to be one, we have $N = C + G$ with a circulant matrix C and G of rank two. (If we use quadratic C^1 B-splines, then the rank of F is four at most[8].) Denoting by F the Fourier matrix, we have $C = F\Lambda^2 F^*$ with the diagonal matrix Λ having positive entries. Consequently, the basis (4.1) is almost orthogonalized by taking

$$AM = [\, h_1 \ \cdots \ h_k \,]M \quad \text{with} \quad M = F\Lambda^{-1}$$

by the fact that $(AM)^*AM = \Lambda^{-1}F^*NF\Lambda^{-1} = I + \Lambda^{-1}F^*GF\Lambda^{-1}$. This takes just k parameters. Thereby running Algorithm 1 with AM at most two steps completes the orthogonalization. Or, alternatively, polynomial orthogonalization requires just three steps of the Hermitian Lanczos method and thereafter interpolation of the function $\frac{1}{\sqrt{\lambda}}$ at the spectrum of $(AM)^*AM$.

Consequently, whenever some sort of (sparse) Toeplitzness appears in the Gram matrix, such as with B-splines [34, 39], it is an appealing alternative to consider circulant structures for preconditioning orthogonalization.

Let us next proceed more algebraically, without resorting to the Gram matrix. For computational reasons, in large scale problems it is often beneficial to assume $\mathcal{X}$ to be a subspace of $\mathbb{C}^{k\times k}$. Such an assumption is particularly appropriate in devising preconditioners for the orthogonalization processes just described. The goal of preconditioning is then simple to formulate. The aim is to favorably relocate the singular values of $A \in \mathbb{C}^{n\times k}$ in view of the optimization problems (2.6) and (2.9). In what follows we describe some options to this end.

Diagonal scaling of the columns fits into this since then $\mathcal{X}$ is the subspace of diagonal matrices. (Diagonal scaling is a standard operation in numerical linear algebra (see [38] for its optimality).) It can be interpreted as preconditioning, although it can be redundant since mere orthogonality of the columns may be sufficient for the construction of preconditioners. In this the following structure is helpful.

Definition 4.2. A subset $\mathcal{X}$ of $\mathbb{C}^{k\times k}$ is said to be diagonally right invariant if $XD \in \mathcal{X}$ for any $X \in \mathcal{X}$ and any diagonal matrix D.

Proposition 4.3. *Let $A \in \mathbb{C}^{n\times k}$ and assume $\mathcal{X} \subset \mathbb{C}^{k\times k}$ is diagonally right invariant. Let D be a nonsingular diagonal matrix. Then solving*

$$\min_{X\in\mathcal{X}, Q\in\mathcal{Q}} \|AX - QD\|_F$$

is equivalent to solving (1.1) *in the Frobenius norm.*

Proof. Denote by x_j the columns of X and q_j the columns of Q for $j = 1, \ldots, k$. Since the Frobenius norm is used, we may consider the problem column-wise. This means that we are concerned with minimizing $\|Ax_j - q_j d_j\|$. Since $\mathcal{X}$ is diagonally right invariant, this is equivalent to minimizing $\|Ax_j - q_j\|$. □

To extend diagonal scaling, recall that the Gram-Schmidt process is an algorithm to produce orthonormal vectors starting from an ordered set of vectors. The associated QR factorization consists of the product of an element of $\mathcal{Q}$ and an upper triangular k-by-k matrix with positive diagonal entries. Neither pivoting nor preprocessing is needed in computing the QR factorization. With parallel processing and sparse problems in mind, we proceed by formulating the notion of QR factorization as a Frobenius norm minimization problem

$$\min_{Y\in\mathcal{Y}} \|A(I - Y)\|_F, \tag{4.2}$$

where $\mathcal{Y} \subset \mathbb{C}^{k\times k}$ consists of strictly upper triangular matrices.

Proposition 4.4. *Assume $A \in \mathbb{C}^{n\times k}$ has linearly independent columns. Let*

$$A(I - Y) = E, \tag{4.3}$$

where $Q = ED$ and $R = D^{-1}(I - Y)^{-1}$ with a diagonal matrix D scaling the columns of E to be unit vectors.

[8] For B-spline discretizations leading to this type of Gram matrices, see, e.g., [10].

Proof. The l-th column of $A(I-Y)$ reads

$$a_l - \sum_{j=1}^{l-1} y_{jl} a_j, \tag{4.4}$$

where a_j denotes the j-th column of A. Because the Frobenius norm is used, this gets minimized in the norm by choosing the column entries y_{jl} such that a_l gets orthogonalized against the vectors $a_1, \ldots, a_{l-1}$ for $l = 2, \ldots, k$. Once complete, the columns of $A(I-Y)$ are orthogonal. □

Now (4.4) is intriguing since minimizing the norm to have the l-th column y_l of Y means solving a least squares problem, i.e., a central application of the QR factorization, originally suggested in [4], is solving least squares problems. Here, the roles are entirely interchanged since this construction of the QR factorization relies on solving least squares problems[9].

Based on Proposition 4.4, the minimization problem (4.2) provides a way to construct optimal incomplete orthogonalization schemes with respect to low dimensional matrix subspaces. Let us make the formulation rigorous as follows.

Definition 4.5. Assume $\mathcal{Y} \subset \mathbb{C}^{k\times k}$ is a subspace of strictly upper triangular matrices. Then solving

$$\min_{Y\in\mathcal{Y}} \|A(I-Y)\|_F \tag{4.5}$$

is said to be an incomplete upper triangular orthogonalization of the columns of $A \in \mathbb{C}^{n\times k}$ corresponding to $\mathcal{Y}$.

We say that the corresponding incomplete upper triangular orthogonalization is of dimension $\dim \mathcal{Y}$. Then the approximate orthonormal basis is given by the rows of

$$\hat{Q} = A(I-Y)D \tag{4.6}$$

with the diagonal matrix D scaling the columns of $A(I-Y)$ to be unit vectors. Clearly, if there is no $\mathcal{Y}$, then this is just standard scaling. In large scale problems it is not advisable to explicitly form $\hat{Q}$. Instead, store the matrices A, Y and D. Computing Y can be done in parallel for each column y_l. It consists of solving the least squares problems associated with (4.5) with the constraint that, for every l, only the prescribed $O(1)$ entries y_{jl} in (4.4) are allowed to be nonzero.

The Gram-Schmidt process is not invariant with respect to a permutation of the columns. Because of this, ordering the columns of A is customary in preprocessing large orthogonalization processes (see, e.g., [2, Chapter 6]). This corresponds to forming AP, where P is a permutation matrix. This converts the minimization problem into

$$\min_{Y\in\mathcal{Y}} \|AP(I-Y)\|_F = \min_{Y\in P\mathcal{Y}P^*} \|A(I-Y)\|_F.$$

This allows an operator theoretic, i.e., basis independent interpretation of the Gram-Schmidt process. The relevant structure to look at here is nilpotent, i.e., each element of the matrix subspace $P\mathcal{Y}P^*$ is a nilpotent matrix[10]. For nilpotent matrix subspaces, see [11].

For maximal parallelizability, we are primarily interested in standard matrix subspaces, i.e., when the dimension of $\mathcal{Y}$ is determined by its sparsity. This allows solving the problem column-wise. In practice $\dim \mathcal{Y} \ll n^2/2$, so that the columns of (4.6) can be expected to be only very approximately orthogonal. This deviation can be measured in many ways. For example, inspect some appropriately defined distance of $\hat{Q}^*\hat{Q}$ from the set of diagonal matrices. For preconditioning purposes even rough approximations and estimates may well suffice. Still, successfully choosing a sparse upper triangular matrix subspace $\mathcal{Y}$ is a nontrivial task.

[9] The l-th column y_l of Y can be found (independently of the other columns) by applying the conjugate gradient (CG) method on $A_l^*A_l y_l = A_l^* a_l$, where $A_l \in \mathbb{C}^{n\times l}$ consists of the first l columns of A. Since the CG method is well understood, it is easy to control the accuracy of the numerical solution.

[10] A bounded linear operator T is nilpotent if $T^j = 0$ for some $j \in \mathbb{N}$.

Let us describe one option, assuming the permutation of the columns has already been performed. Because of (4.4), it must be assessed how linearly dependent the l-th column of A is on the columns preceding it. For commensurability, first scale the columns of A to be unit vectors and denote this matrix by $\tilde{A}$. Then, for a sparsity structure determined row-wise, allow the (j,l) entry of $\mathcal{Y}$ to be nonzero if

$$|(\tilde{m}_l, \tilde{m}_j)| > \mathrm{tol} \tag{4.7}$$

for $j < l$ for some tolerance $0 < \mathrm{tol} < 1$, i.e., $\mathcal{Y}$ is determined by a sparsification of the scaled Gram matrix $\tilde{A}^*\tilde{A}$. For feasibility, such computations should be performed parallel in a sparse mode.

5 Numerical experiments

Numerical experiments are conducted with MATLAB to demonstrate properties of the orthogonalization methods introduced. The matrices considered are benchmarks arising in various disciplines and they are all sparse. The object of interest is the decay of $1 > \omega_j(A) \geqslant 0$ for $j = 1, 2, \ldots$ Effects of preconditioning are illustrated.

Example 5.1. This example is concerned with the **illc1033**-matrix from the Harwell-Boeing sparse matrix collection, i.e., $A \in \mathbb{R}^{1033\times 320}$ arises in the least squares problems related with surveying. Full orthogonalization with Algorithm 1 is available after at most $\frac{320}{2} = 160$ steps. However, near orthogonality is reached much earlier (see Figure 1(a) for the decay of $\omega_j(A)$). In Figure 1(b), we have a plot of the singular values of A. In this plot three plateaus can be identified. This renders polynomial preconditioning attractive, assuming the "heights" of the plateaus are known. We take p of degree two such that these plateaus are mapped to 1 (see Figure 2). Regarding near orthogonality, we now have

$$\#\{\omega_j(A) \leqslant 10^{-3}\} = 34, \quad \#\{\omega_j(A) \leqslant 10^{-2}\} = 34, \quad \#\{\omega_j(A) \leqslant 10^{-1}\} = 35,$$

and for $B = Ap(A^*A)$,

$$\#\{\omega_j(B) \leqslant 10^{-3}\} = 47, \quad \#\{\omega_j(B) \leqslant 10^{-2}\} = 57, \quad \#\{\omega_j(B) \leqslant 10^{-1}\} = 68,$$

i.e., a considerable improvement.

Next, we look at square matrices. Observe that if one is preconditioning corresponding linear systems, then attaining zero in (1.1) means that $A^{-1} = XX^*A^*$. Hence the preconditioner resulting from approximate orthogonalization means taking XX^*A^*.

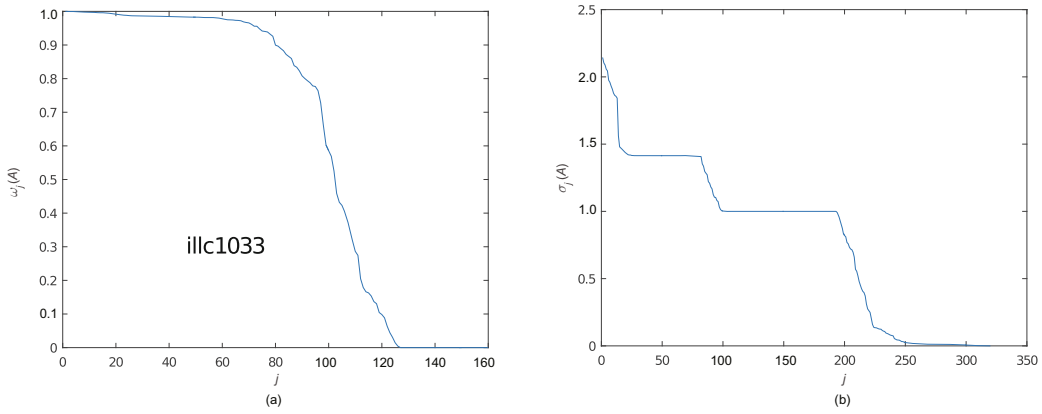

Figure 1 For the **illc1033**-matrix of Example 5.1, we have $\omega_j(A)$ on (a) and the singular values $\sigma_j(A)$ on (b)

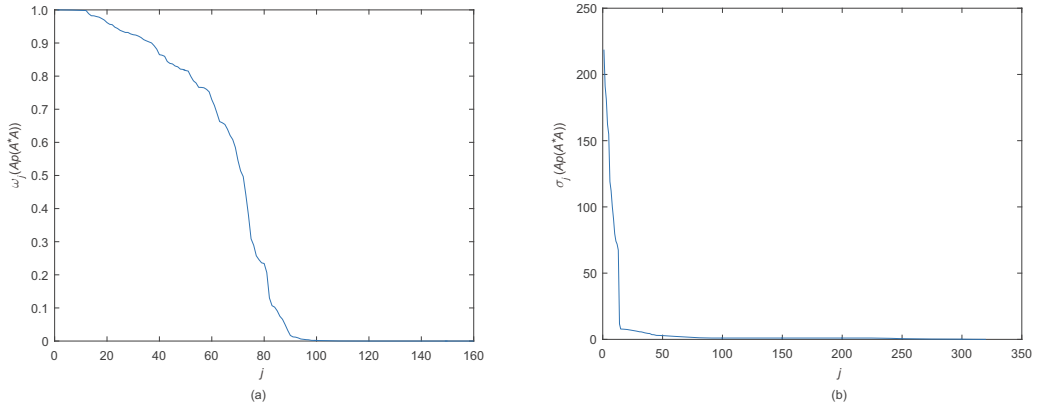

Figure 2 For the **illc1033**-matrix of Example 5.1, we have $\omega_j(Ap(A^*A))$ on (a) and the singular values $\sigma_j(Ap(A^*A))$ on (b)

Example 5.2. This example is concerned with the **tols1090**-matrix from the Harwell-Boeing sparse matrix collection, i.e.,

$$A \in \mathbb{R}^{1090\times 1090}$$

arises in stability analysis. Full orthogonalization with Algorithm 1 is available after at most $\frac{1090}{2} = 545$ steps (see Figure 3(a) for the decay of $\omega_j(A)$). In Figure 4 we have AD, i.e., preconditioned A by scaling its columns to be unit vectors. This simple preconditioning has a dramatic effect on the speed of orthogonalization.

Regarding near orthogonality, we now have

$$\#\{\omega_j(A) \leqslant 10^{-3}\} = 103, \quad \#\{\omega_j(A) \leqslant 10^{-2}\} = 106, \quad \#\{\omega_j(A) \leqslant 10^{-1}\} = 108,$$

and

$$\#\{\omega_j(AD) \leqslant 10^{-3}\} = 273, \quad \#\{\omega_j(AD) \leqslant 10^{-2}\} = 273, \quad \#\{\omega_j(AD) \leqslant 10^{-1}\} = 274.$$

In Figure 5 we have used (4.7) with a tolerance 0.9 to choose 24 columns to perform the partial orthogonalization (4.4) of each column. Denoting the resulting preconditioner by P we have

$$\#\{\omega_j(ADP) \leqslant 10^{-3}\} = 281, \quad \#\{\omega_j(ADP) \leqslant 10^{-2}\} = 281, \quad \#\{\omega_j(ADP) \leqslant 10^{-1}\} = 282.$$

For comparison, let us mention that randomly choosing 24 columns to perform a partial orthogonalization did not lead to any improvement.

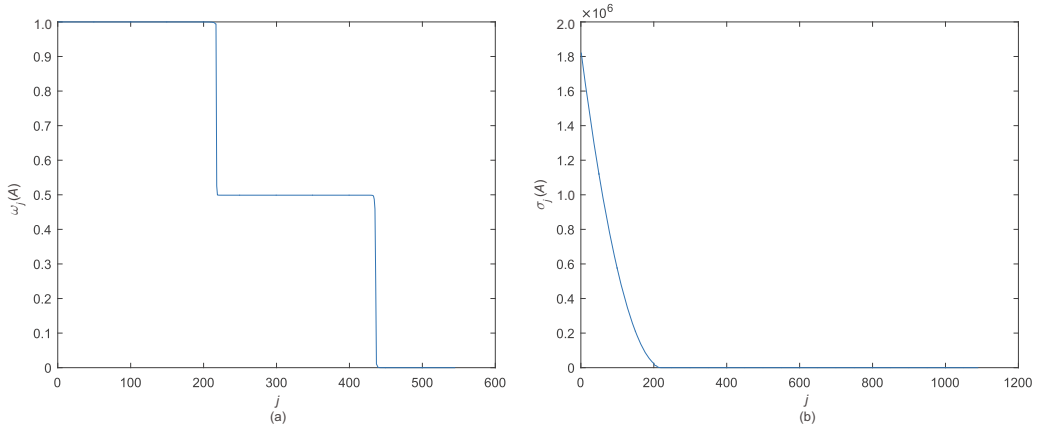

Figure 3 For the **tols1090**-matrix of Example 5.2, we have $\omega_j(A)$ on (a) and the singular values $\sigma_j(A)$ on (b)

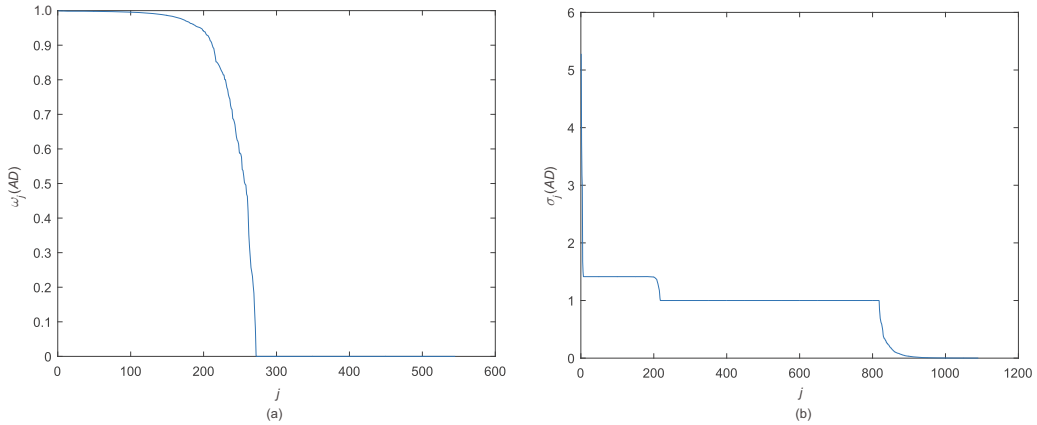

Figure 4 For the **tols1090**-matrix of Example 5.2, we have $\omega_j(AD)$ on (a) and the singular values $\sigma_j(AD)$ on (b)

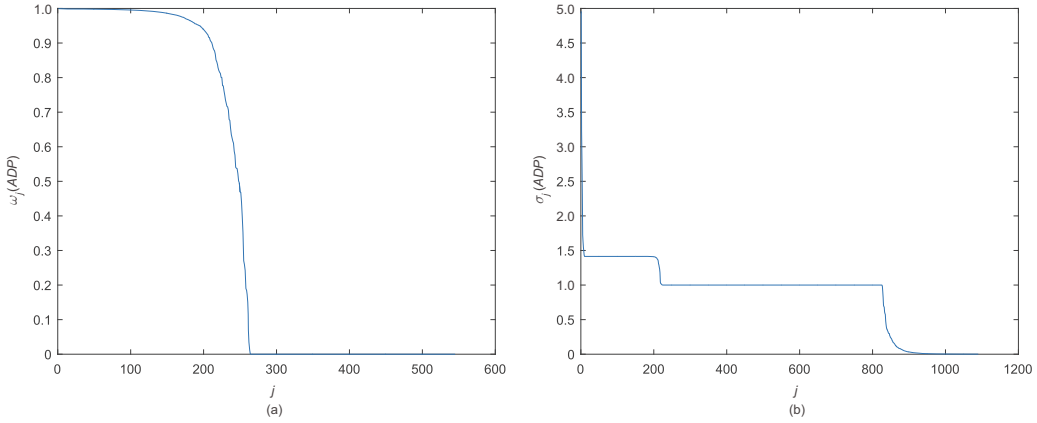

Figure 5 For the **tols1090**-matrix of Example 5.2, we have $\omega_j(ADP)$ on (a) and the singular values $\sigma_j(ADP)$ on (b)

The following example is of different nature in the sense that Algorithm 1 is used as a corrector to the Gram-Schmidt process (or, stated in other words, the modified Gram-Schmidt process is now used as a preconditioner), i.e., there is some loss of orthogonality when the modified Gram-Schmidt process is executed. It seems, however, that this loss of orthogonality is very structured in favor of using optimal orthogonalization processes.

Example 5.3. This example is concerned with the **fpga**-matrix from Sandia National Laboratory[11], i.e.,

$$A \in \mathbb{C}^{1200\times 1200}$$

arises in circuit simulations. When the modified Gram-Schmidt method is executed to have a numerically computed QR factorization $A = QR$ of A, some loss of orthogonality of the columns of Q takes place, as is well known. This appears to happen in such a way that only a small number of vectors suffer from it. Therefore, the problem can be very efficiently fixed by executing Algorithm 1 (see Figure 6 for the swift decay of $\omega_j(Q)$).

Example 5.4. This example is also concerned with the **fpga**-matrix. We precondition A by using the incomplete modified Gram-Schmidt of Saad [33, Subsection 10.8.3], i.e., we have $A = QR$ such that R is sparse and Q is not unitary. Only those elements of R are kept nonzero which are larger than the drop tolerance $\tau = 0.05$ used. Then we have only 3,519 nonzero elements in R (in the modified Gram-Schmidt 527,246). We take R^{-1} to be the preconditioner (see Figure 7 for the decay of $\omega_j(AR^{-1})$).

[11] Https://www.cise.ufl.edu/research/sparse/matrices/Sandia/

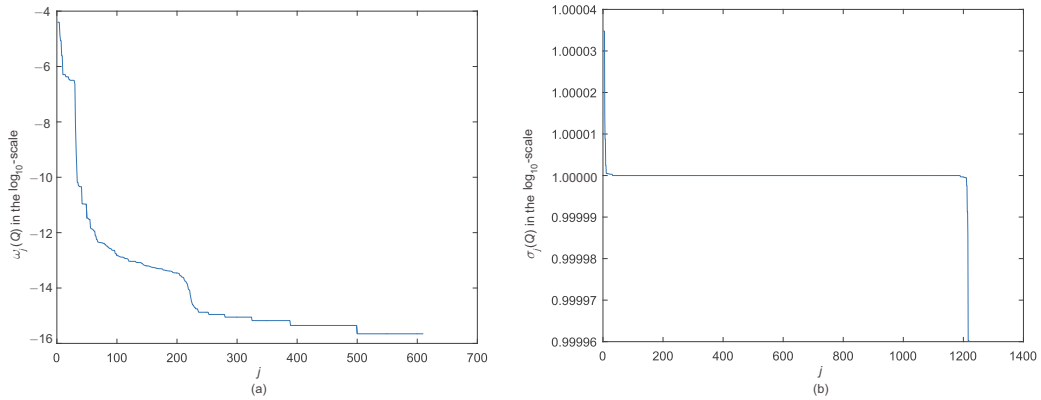

Figure 6 For the **fpga**-matrix of Example 5.3, we have computed Q with the modified Gram-Schmidt method. The columns are not fully orthogonal as some 50 extreme singular values deviate from 1 by at most of order 10^{-5} (see the singular values $\sigma_j(Q)$ on (b)). We then have $\omega_j(Q)$ in the $\log_{10}$-scale on (a)

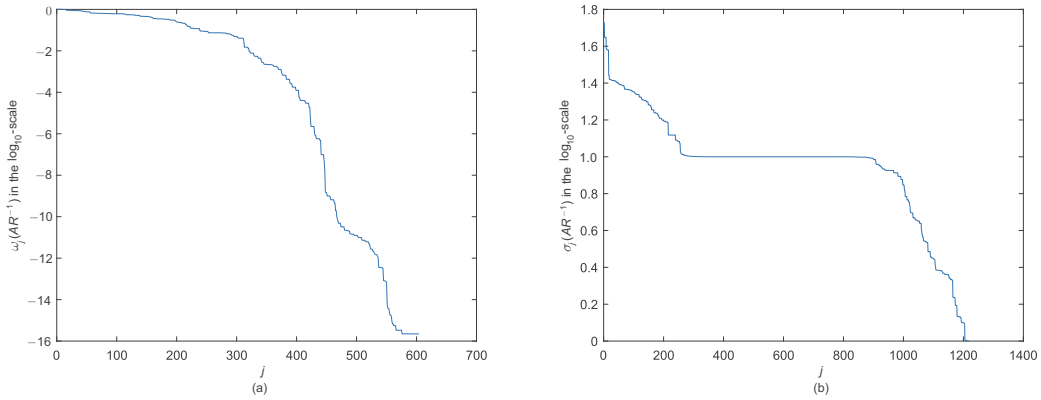

Figure 7 For the **fpga**-matrix of Example 5.4, we have computed $A = QR$ with the incomplete modified Gram-Schmidt method. The columns of $AR^{-1} = Q$ are not orthonormal (see the singular values $\sigma_j(AR^{-1})$ on (b)). We then have $\omega_j(AR^{-1})$ in the $\log_{10}$-scale on (a). Observe that initially we are far from orthogonality. Nonorthogonality starts decreasing at a moderate pace and gains speed after around 300 steps

6 Conclusions

The problem of optimally decreasing nonorthogonality of a system of vectors with respect to a limited number of free parameters has been addressed. Being a fundamental problem in a Hilbert space, the task arises in several computational disciplines. In this paper the Löwdin and the Gram-Schmidt orthogonalization processes were interpreted polynomially and in terms of translated small-rank corrections, respectively. Optimality was attained such that there exists a natural way of increasing the number of parameters to improve approximations in both cases. The latter alternative gets the best of both worlds since it simultaneously retains locality by yielding $2j$ exactly orthonormalized columns after j steps, for $j = 1, 2, \ldots$ Preconditioning was considered and shown to be of significance in speeding up orthogonalization processes. Numerical experiments were conducted.

Acknowledgements This work was supported by the Academy of Finland (Grant No. 288641). The authors are grateful to the anonymous referees for their comments and suggestions which led to corrections of some mistakes and improved the paper notably.

References

1 Argyros S, Haydon R. A hereditarily indecomposable $\mathcal{L}_\infty$-space that solves the scalar-plus-compact problem. Acta Math, 2011, 206: 1–54

2 Björck A. Numerical Methods for Least Squares Problems. Philadelphia: SIAM, 1996
3 Brualdi R, Friedland S, Pothen A. The sparse basis problem and multilinear algebra. SIAM J Matrix Anal Appl, 1995, 16: 1–20
4 Businger P, Golub G. Linear least squares solutions by Householder transformations. Numer Math, 1965, 7: 269–276
5 Christensen O. Frames, Riesz bases, and discrete Gabor/wavelet expansions. Bull Amer Math Soc (NS), 2001, 38: 273–291
6 Dahlquist G, Björck A. Numerical Methods in Scientific Computing. Vol. I. Philadelphia: SIAM, 2008
7 Douglas R G, Liaw C. A geometric approach to finite rank unitary perturbations. Indiana Univ Math J, 2013, 62: 333–354
8 Feichtinger H, Luef F, Werther T. A guided tour from linear algebra to the foundations of Gabor analysis. In: Gabor and Wavelet Frames. IMS Lecture Notes Series, vol. 10. Hackensack: World Scientific, 2007, 1–49
9 Frank M, Paulsen V, Tiballi T. Symmetric approximation of frames and bases in Hilbert spaces. Trans Amer Math Soc, 2002, 354: 777–793
10 Garoni C, Speleers H, Ekström S-E, et al. Symbol-based analysis of finite element and isogeometric B-spline discretizations of eigenvalue problems: Exposition and review. Arch Comput Meth Eng, 2018, 26: 1639–1690
11 Gerstenhaber M. On nilalgebras and linear varieties of nilpotent matrices. I. Amer J Math, 1958, 80: 614–622
12 Goldstein J A, Levy M. Linear algebra and quantum chemistry. Amer Math Monthly, 1991, 98: 710–718
13 Golub G, Varah J. On the characterization of the best ℓ_2-scaling of a matrix. SIAM J Numer Anal, 1974, 11: 472–479
14 Hansen A C. On the approximation of spectra of linear Hilbert space operators. PhD Thesis. Cambridge: University of Cambridge, 2008
15 Higham N. Accuracy and Stability of Numerical Algorithms, 2nd ed. Philadelphia: SIAM, 2002
16 Horn R A, Johnson C R. Topics in Matrix Analysis. Cambridge: Cambridge University Press, 1991
17 Horn R A, Johnson C R. Matrix Analysis, 2nd ed. Cambridge: Cambridge University Press, 2013
18 Huhtanen M. Factoring matrices into the product of two matrices. BIT, 2007, 47: 793–808
19 Huhtanen M. Energy conservation and unitary approximation numbers. J Math Phys, 2007, 48: 073512
20 Huhtanen M, Nevanlinna O. Polynomials and lemniscates of indefiniteness. Numer Math, 2016, 133: 233–253
21 Jaffard S, Young R M. A representation theorem for Schauder bases in Hilbert space. Proc Amer Math Soc, 1998, 126: 553–560
22 Leon S J, Björck A, Gander W. Gram-Schmidt orthogonalization: 100 years and more. Numer Linear Algebra Appl, 2013, 20: 492–532
23 Liaw C. Rank one and finite rank perturbations—survey and open problems. arXiv:1205.4376v1, 2012
24 Livne O E, Golub G. Scaling by binormalization. Numer Algorithms, 2004, 35: 97–120
25 Löwdin P-O. On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. J Chem Phys, 1950, 18: 365–375
26 Löwdin P-O. On the nonorthogonality problem. Adv Quantum Chem, 1970, 5: 185–199
27 Mayer I. Simple Theorems, Proofs, and Derivations in Quantum Chemistry. New York: Kluwer, 2003
28 Mayer I, Surjan P R. Handling overlap as a perturbation. Croatica Chemica Acta, 1993, 66: 161–165
29 Olver P J, Shakiban C. Applied Linear Algebra, 2nd ed. Undergraduate Texts in Mathematics. New York: Springer, 2018
30 Philippe B. An algorithm to improve nearly orthonormal sets of vectors on a vector processor. SIAM J Algebraic Discrete Methods, 1987, 3: 393–403
31 Rubensson E H, Bock N, Holmstöm E, et al. Recursive inverse factorization. J Comput Chem, 2008, 128: 104105
32 Rubensson E H, Rudberg E, Salek P. Sparse matrix algebra for quantum modeling of large systems. In: Applied Parallel Computing, State of the Art in Scientific Computing. Lecture Notes in Computer Science, vol. 4699. Berlin-Heidelberg: Springer, 2007
33 Saad Y. Iterative Methods for Sparse Linear Systems, 2nd ed. Philadelphia: SIAM, 2003
34 Strang G. Piecewise polynomials and the finite element method. Bull Amer Math Soc, 1973, 79: 1128–1137
35 Szabo A, Ostlund N S. Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory. Mineola: Dover, 1996
36 Szczepanik D, Mrozek J. On several alternatives for Löwdin orthogonalization. Comput Theor Chem, 2013, 1008: 103–109
37 Thijssen J. Computational Physics, 2nd ed. Cambridge: Cambridge University Press, 2007
38 van der Sluis A. Condition numbers and equilibration of matrices. Numer Math, 1975, 14: 14–23
39 Varah J. On the condition number of local bases for piecewise cubic polynomials. Math Comp, 1977, 137: 37–44
40 Walk P, Jung P. Approximation of Löwdin orthogonalization to a spectrally efficient orthogonal overlapping PPM design for UWB impulse radio. Signal Processing, 2012, 92: 649–666

Appendix A Rank-one perturbations of positive semidefinite matrices

The following yields conditions on the existence of a unitary rank-one up-date of a positive semidefinite 2-by-2 matrix.

Proposition A.1 (See [19]). *Let $P \in \mathbb{C}^{2\times 2}$ be positive semidefinite with the eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant 0$. Then there are two vectors $u, v \in \mathbb{C}^2$ such that $P - uv^*$ is unitary if and only if $\lambda_1 \geqslant 1 \geqslant \lambda_2$.*

Proof. Let $V^*PV = \mathrm{diag}(\lambda_1, \lambda_2)$ be a diagonalization of P by a unitary matrix $V \in \mathbb{C}^{2\times 2}$. We may consider

$$V^*(V - uv^*)V = \mathrm{diag}(\lambda_1, \lambda_2) - kl^*$$

with $k = V^*u$ and $l = V^*v$. Now $\mathrm{diag}(\lambda_1, \lambda_2) - kl^*$ is unitary if and only if

$$\begin{aligned} &-\lambda_1\overline{k_2}l_1 - \lambda_2 k_1\overline{l_2} + k_1\overline{k_2}\|l\|^2 = 0,\\ &\lambda_1^2 - \lambda_1(\overline{k_1}l_1 + k_1\overline{l_1}) + |k_1|^2\|l\|^2 = 1,\\ &\lambda_2^2 - \lambda_2(\overline{k_2}l_2 + k_2\overline{l_2}) + |k_2|^2\|l\|^2 = 1. \end{aligned}$$

Without loss of generality, we may impose $\|l\| = 1$. Then the first equation gives $k_1 = \frac{\lambda_1\overline{k_2}l_1}{\overline{k_2} - \lambda_2\overline{l_2}}$. Inserting this into the second equation yields

$$\lambda_1^2|l_1|^2(\lambda_2(k_2\overline{l_2} + \overline{k_2}l_2) - |k_2|^2) = (1-\lambda_1^2)(\lambda_2^2|l_2|^2 + |k_2|^2 - \lambda_2(k_2\overline{l_2} + \overline{k_2}l_2)). \tag{A.1}$$

The third equation gives

$$\lambda_2(k_2\overline{l_2} + \overline{k_2}l_2) = \lambda_2^2 + |k_2|^2 - 1. \tag{A.2}$$

Inserting this into (A.1) with $|l_1|^2 = 1 - |l_2|^2$ yields $|l_1|^2 = (1-\lambda_1^2)/(\lambda_2^2 - \lambda_1^2)$. Hence $|\lambda_1| \geqslant 1$ for a solution to exist. Then we may take

$$l_1 = \sqrt{\frac{\lambda_1^2 - 1}{\lambda_1^2 - \lambda_2^2}} \quad \text{and} \quad l_2 = \sqrt{\frac{1-\lambda_2^2}{\lambda_1^2 - \lambda_2^2}}. \tag{A.3}$$

We may also impose $l_2 \geqslant 0$, so that (A.2) converts into

$$|k_2|^2 - 2\lambda_2 l_2 \mathrm{Re}k_2 = 1 - \lambda_2^2.$$

This is solvable by, for example, choosing k_2 to be pure imaginary. So we may take

$$k_2 = \mathrm{i}\sqrt{1-\lambda_2^2} \tag{A.4}$$

giving finally

$$k_1 = \lambda_1\overline{k_2}l_1/(\overline{k_2} - \lambda_2\overline{l_2}) = \sqrt{1 - \frac{1}{\lambda_1^2}}(\sqrt{\lambda_1^2 - \lambda_2^2} + \mathrm{i}\lambda_2) \tag{A.5}$$

after simplifications. □

The scalars (A.3)–(A.5) are used in the proof of Theorem 2.8.