

Perturbation Bounds for Hyperbolic Matrix Factorizations

Michael Berhanu *

June 14, 2005

Abstract

Several matrix factorizations depend on orthogonal factors, matrices that preserve the Euclidean scalar product. Some of these factorizations can be extended and generalized to (J, \tilde{J}) -orthogonal factors, that is, matrices that satisfy $H^T JH = \tilde{J}$, where J and \tilde{J} are diagonal with diagonal elements ± 1 . The purpose of this work is to analyze the perturbation of matrix factorizations that have a (J, \tilde{J}) -orthogonal or orthogonal factor and to give first order perturbation bounds. For each factorization analyzed, we give the sharpest possible first order bound, which yields a condition number. The cost of computing these condition numbers is high. It is usually equivalent to computing the 2-norm of an $n^2 \times n^2$ matrix for a problem of size n . Thus, we also propose less sharp bounds that are less expensive to compute.

Key words: Condition number, hyperbolic matrix factorization, manifold.

AMS subject classifications: 65F35, 15A23, 15A60.

1 Introduction

Matrix factorization is a common tool in different branches of mathematics. A general definition is given in [1]:

A matrix factorization theorem is an assertion that a matrix A can be factorized into a product $A = A_1 A_2$, of two special matrices A_1, A_2 . Some conditions may be necessary for such a decomposition to exist, and some further conditions may ensure the uniqueness of the factorization.

Throughout this paper, we encounter matrix factorizations in which more than two matrices are involved. The aim of this work is to analyze the sensitivity of some matrix factorizations that involve at least one hyperbolic matrix and to give a first order perturbation bound for the factors, where a hyperbolic matrix is a matrix that satisfy $H^T JH = \tilde{J}$, with J and \tilde{J} diagonal with diagonal elements ± 1 . The optimal first order perturbation bound yields a condition number of the relevant

*School of Mathematics, University of Manchester, Sackville Street Manchester, M60 1QD, England (mberhanu@ma.man.ac.uk)

matrix in the factorization, which measures its sensitivity to perturbations in the data. For $A, X, Y \in \mathbb{R}^{n \times n}$, let $\varphi(X, Y) = A$ be a factorization of A , where φ is a function describing the factorization. For instance for the QR factorization, $\varphi(X, Y) = XY$, where X is unitary and R upper triangular. The classical theory of condition numbers [17] employs the definitions

$$\begin{aligned}\kappa_X &= \limsup_{\epsilon \rightarrow 0} \{ \epsilon^{-1} \|\Delta X\|, \varphi(X + \Delta X, Y + \Delta Y) = A + \Delta A, \|\Delta A\| \leq \epsilon \}, \\ \kappa_Y &= \limsup_{\epsilon \rightarrow 0} \{ \epsilon^{-1} \|\Delta Y\|, \varphi(X + \Delta X, Y + \Delta Y) = A + \Delta A, \|\Delta A\| \leq \epsilon \}.\end{aligned}$$

This definition has the advantage of being simple to present, although in most cases the necessary computations to bound the condition number or to show it is attained are far from being trivial. The method used in this paper is quite different. Our aim is to define a function g in a neighborhood of A such that $\varphi(\tilde{X}, \tilde{Y}) = A + \Delta A$ with $(\tilde{X}, \tilde{Y}) = g(A + \Delta A)$. We define the condition number as the norm of $\|dg(A)\|$, the differential of g at A . The main tool for this analysis is the implicit function theorem. Our method is described in detail in Section 2.3.

Several results that we cite later on are available concerning orthogonal matrix factorizations. In most cases, these results are only bounds and not the condition number. In the literature, condition numbers for hyperbolic matrix factorization have not been reported. In this paper, we investigate some matrix factorizations and for each of them we compute the condition number. Our proof technique is not new. It was used in [1] and [7] to investigate perturbation bounds for several matrix factorizations.

The HR factorization is the generalization of the usual QR factorization when the orthogonal factor is allowed to be (J, \tilde{J}) -orthogonal. Perturbation bounds of the QR factorization are given for example in [1], [6], [7], [14] and [19]. In this paper, we compute the condition number of the HR factorization and show that the classical perturbation bounds for the QR factorization are very weak. Our analysis (of the HR factorization) is closer to the ones presented in [1], [7] and [14]. For the singular value decomposition (SVD), it is well known that the condition number for the singular values is 1 (see for example [20]). Perturbation bounds for the singular vectors are also available in [18], [21]. In our case, we compute the condition number of the hyperbolic SVD (see Section 4), which is the generalization of the usual SVD. In several papers, the polar factorization have been analyzed (see for example [5], [8], [10], [13], [15]). Once more, we compute the condition number of the indefinite polar factorization and apply our results to the usual polar factorization, which allows us to give a short and easy computation of its condition number. We also refer to two surveys, [1] where perturbation bounds for several matrix factorizations are given and [10] where various conditioning problems are treated.

This paper is organized as follows. In Section 2, the notations, definitions and all the necessary background material are presented. Then, from Section 3 to Section 5, the HR factorization, the hyperbolic singular value decomposition and the indefinite polar factorization are analyzed. Numerical experiments were carried out in each section. Test matrices were generated using MATLAB 6.5 with unit roundoff $u \approx 1e - 16$.

2 Background and Preliminary Results

2.1 Notations, Definitions and Matrix Operators Properties

For $X = (x_{ij}) \in \mathbb{K}^{n \times n}$, $\mathbb{K} = \mathbb{R}$ or \mathbb{C} , the 2-norm $\|\cdot\|_2$ and Frobenius norm $\|\cdot\|_F$ are defined by

$$\|X\|_2 = \sup_{\|y\|_2=1} \|Xy\|_2, \quad \|X\|_F = (\text{trace}(X^*X))^{\frac{1}{2}} = \left(\sum_{i,j=1}^n |x_{ij}|^2 \right)^{\frac{1}{2}}.$$

The condition number of $X \in \mathbb{K}^{n \times n}$ is $\kappa_\nu(X) = \|X\|_\nu \|X^{-1}\|_\nu$, $\nu = 2, F$. For an operator or a linear map \mathcal{T} defined on $\mathbb{K}^{n \times n}$, the 2-norm is defined by

$$\|\mathcal{T}\|_2 = \sup_{\|X\|_F=1} \|\mathcal{T}X\|_F.$$

The choice of this norm is justified by its differentiability properties and its computational simplicity. We now present some notations and we give some results that are needed throughout this work.

e_k denotes the k -th column of the identity and we define $e_{ij} = e_i e_j^T$. Let $x \in \mathbb{K}^n$. Throughout this work, \mathcal{V}_x denotes an open neighborhood of x . For a differentiable function f in \mathcal{V}_x , $df(x)$ denotes the differential at the point x . All the vector spaces are vector spaces on \mathbb{R} and thus all the functions are considered as functions of real variables and the differentiation is real.

$\Delta(\mathbb{K})$ denotes the set of upper triangular matrices in $\mathbb{K}^{n \times n}$ with a real diagonal. $\mathbf{Sym}(\mathbb{K})$ (respectively $\mathbf{Skew}(\mathbb{K})$) is the linear subspace of symmetric matrices (respectively skew-symmetric matrices) with coefficients in \mathbb{K} . \mathbf{Herm} (respectively \mathbf{SkewH}) is the linear subspace of Hermitian matrices (respectively skew-Hermitian matrices). \dim denotes the dimension of linear space in \mathbb{R} . We recall that

$$\dim \Delta(\mathbb{R}) = \dim \mathbf{Sym}(\mathbb{R}) = \frac{n^2 + n}{2}, \quad (1)$$

$$\dim \Delta(\mathbb{C}) = \dim \mathbf{Herm} = \dim \mathbf{SkewH} = n^2, \quad (2)$$

$$\dim \mathbf{Skew}(\mathbb{R}) = \frac{n^2 - n}{2}. \quad (3)$$

For $x \in \mathbb{K}^n$, $\text{diag}(x)$ denotes the $n \times n$ diagonal matrix whose diagonal is x . For $X \in \mathbb{K}^{n \times n}$, we denote $\Pi_d(X) = Y$, the diagonal part, $\Pi_u(X)$, the strictly upper part and $\Pi_l(X)$, the strictly lower part of X .

For $A, B \in \mathbb{K}^{n \times n}$ with $A = (a_{ij})$ and $B = (b_{ij})$, the *Schur product* is defined by $A \circ B = (a_{ij}b_{ij})$ and the *Kronecker product* is defined by $A \otimes B = (a_{ij}B)$. The matrix operator $\text{vec} : \mathbb{K}^{n \times n} \rightarrow \mathbb{K}^{n^2}$ stacks the columns of the matrix into one long vector. We define $\mathbf{Tvec}(A) = \text{vec}(A^T)$. \mathbf{T} is an $n^2 \times n^2$ permutation matrix.

Theorem 2.1 *Let $A, B, X \in \mathbb{K}^{n \times n}$. We define the operators $\mathcal{T}_1X = X \circ A$ and $\mathcal{T}_2X = AXB$. Then,*

$$\|\mathcal{T}_1\|_2 = \max_{ij} |a_{ij}|, \quad (4)$$

$$\|\mathcal{T}_2\|_2 = \|A \otimes B\|_2 = \|A\|_2 \|B\|_2, \quad (5)$$

$$\min_{\|X\|_F=1} \|\mathcal{T}_2(AXB)\|_F = \|A^{-1}\|_2^{-1} \|B^{-1}\|_2^{-1}. \quad (6)$$

Proof. For (4), it is straightforward to show that the right hand side of (4) is an upper bound. Let $|a_{pq}| = \max_{ij} |a_{ij}|$. Then, the bound is attained by $e_p e_q^T$.

Let $A = Q_1 S_1 Z_1^T$ and $B = Q_2 S_2 Z_2^T$ be the singular value decompositions of A and B . Then, $(A \otimes B) = (Q_1 \otimes Q_2)(S_1 \otimes S_2)(Z_1^T \otimes Z_2^T)$. Thus, we obtain the first part of (5) $\|(A \otimes B)\|_2 = \|(S_1 \otimes S_2)\|_2 = \|A\|_2 \|B\|_2$. For (5) and (6), the result is easily obtained by using the singular value decompositions of A and B , then the Lagrange multipliers theorem. \square

2.2 (J, \tilde{J}) -orthogonal and (J, \tilde{J}) -unitary matrices

We denote by $\text{diag}_n^k(\pm 1)$ the set of all $n \times n$ diagonal matrices with k diagonal elements equal to 1 and $n - k$ equal to -1 . A matrix $J \in \text{diag}_n^k(\pm 1)$ for some k is called a *signature* matrix. A matrix $H \in \mathbb{R}^{n \times n}$ is said to be (J, \tilde{J}) -orthogonal if $H^T J H = \tilde{J}$, where $J, \tilde{J} \in \text{diag}_n^k(\pm 1)$. We denote by $\mathcal{O}_n(J, \tilde{J})$ the set of $n \times n$ (J, \tilde{J}) -orthogonal matrices. If $J = \tilde{J}$ then we say that H is J -orthogonal or pseudo-orthogonal and the set of J -orthogonal matrices is denoted by $\mathcal{O}_n(J)$. We say that a matrix is *hyperbolic* if it is (J, \tilde{J}) -orthogonal or pseudo-orthogonal with $J \neq \pm I$. We recall that if $J = \pm I$, then $\mathcal{O}_n(\pm I) = \mathcal{O}_n$ is the usual set of orthogonal matrices.

We extend the definition of (J, \tilde{J}) -orthogonal matrices to rectangular matrices in $\mathbb{R}^{m \times n}$, with $m \geq n$. $H \in \mathbb{R}^{m \times n}$ is (J, \tilde{J}) -orthogonal if $H^T J H = \tilde{J}$ with $J \in \text{diag}_m^k(\pm 1)$ and $\tilde{J} \in \text{diag}_n^q(\pm 1)$. We denote by $\mathcal{O}_{mn}(J, \tilde{J})$ the set of (J, \tilde{J}) -orthogonal in $\mathbb{R}^{m \times n}$. (J, \tilde{J}) -unitary matrices are the complex counterpart of (J, \tilde{J}) -orthogonal matrices and we say that a matrix $H \in \mathbb{C}^{n \times n}$ is (J, \tilde{J}) -unitary matrix if $H^* J H = \tilde{J}$ where J and \tilde{J} are signature matrices. We denote by $\mathcal{U}_n(J, \tilde{J})$ the set of $n \times n$ (J, \tilde{J}) -orthogonal matrices and by $\mathcal{U}_n(J, \tilde{J})$ we denote the set of $n \times n$ (J, \tilde{J}) -unitary matrices.

We show that $\mathcal{O}_{mn}(J, \tilde{J})$ and $\mathcal{U}_{mn}(J, \tilde{J})$ can respectively be identified with \mathbb{R}^d and \mathbb{R}^{n^2} with $d = \frac{n^2 - n}{2}$. We show that each of these sets are manifolds and we compute their dimension. Then, the introduction of local coordinate systems enables us to make the identification mentioned above.

Lemma 2.2 $\mathcal{O}_{mn}(J, \tilde{J})$ and $\mathcal{U}_{mn}(J, \tilde{J})$ are manifolds with respective dimension d and n^2 , where $d = \frac{n^2 - n}{2}$.

Proof. Let $q_1 : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{n \times n}$ and $q_2 : \mathbb{C}^{m \times n} \rightarrow \mathbb{C}^{n \times n}$ be defined by

$$q_1(X) = X^T J X - \tilde{J} \quad \text{and} \quad q_2(X) = X^* J X - \tilde{J}.$$

We recall that $\mathcal{O}_n(J, \tilde{J}) = q_1^{-1}(\{0\})$. q_1 is clearly differentiable. We have that $dq_1(H)\Delta H = H^T J \Delta H + \Delta H^T J H$. To compute the dimension of the manifold we need to determine the tangent space, that is, the null space of $dq_1(H)$, with $H \in \mathcal{O}_n(J, \tilde{J})$. We have that $\text{null}(dq_1(H)) = JH^{-T} \mathbf{Skew}(\mathbb{R})$. Thus, following (2)-(3), $\mathcal{O}_n(J, \tilde{J})$ is an $\frac{n^2 - n}{2}$ dimensional manifold.

Similarly, $\text{null}(dq_2(H)) = JH^{-*} \mathbf{Skew}(\mathbb{H})$ and $\mathcal{U}_n(J, \tilde{J})$ is a n^2 dimensional manifold. \square

When, $m \neq n$, since H has full rank, then the proof is identical to the one above. Let $X \in \mathcal{O}_{mn}(J, \tilde{J})$ and $Y \in \mathcal{U}_{mn}(J, \tilde{J})$. There exists differentiable one-to-one functions ϕ_k , $1 \leq k \leq 2$, open sets $\mathcal{V}_1 \subset \mathbb{R}^d$ and $\mathcal{V}_2 \subset \mathbb{R}^{n^2}$, $\mathcal{V}_X \subset \mathbb{R}^{m \times n}$ and $\mathcal{V}_Y \subset \mathbb{C}^{m \times n}$ such that

$$\phi_1(\mathcal{V}_1) = \mathcal{V}_X \cap \mathcal{O}_{mn}(J, \tilde{J}) \quad \text{and} \quad \phi_2(\mathcal{V}_2) = \mathcal{V}_Y \cap \mathcal{U}_{mn}(J, \tilde{J}). \quad (7)$$

Moreover, the differential of these maps ϕ_k have full rank over the entire space where they are defined.

2.3 General Method for Computing the Condition Number

Let \mathbb{S} be a linear subspace of $\mathbb{K}^{n \times n}$, $X \in \mathbb{S}$ and let $\mathcal{V}_X \subset \mathbb{S}$ be an open neighborhood of X . \mathbb{S} can be regarded as a set of matrices that have a particular structure such as symmetry, Hermitian or a sparse structure such as upper triangular. In this section, let \mathcal{H} denote either $\mathcal{O}_{mn}(J, \tilde{J})$ or $\mathcal{U}_{mn}(J, \tilde{J})$. Let \mathcal{E} be a linear subspace of $\mathbb{K}^{n \times n}$. A (J, \tilde{J}) -orthogonal or (J, \tilde{J}) -unitary factorization of a matrix $A \in \mathcal{E}$ can be described by a function

$$\varphi(X, Y) = A, \quad X \in \mathcal{V}_X \quad \text{and} \quad Y \in \mathcal{H}$$

Our aim is to derive perturbation bounds for the X and Y factors when A is subject to some perturbation ΔA . The main tool for this analysis is the implicit function theorem. This technique was also used by Bhatia in [1]. This method is divided into three steps.

Step 1 Using (7), we define

$$\begin{aligned} f : \mathcal{V}_A \times \mathcal{V}_X \times \mathbb{K}^p &\rightarrow \mathcal{E}, \\ (\tilde{A}, \tilde{X}, \tilde{y}) &\mapsto \varphi(\tilde{X}, \tilde{Y}) - \tilde{A}, \end{aligned}$$

where $\tilde{Y} = \phi(\tilde{y})$ is defined according to \mathcal{H} in Lemma 2.2 and p is the dimension of \mathcal{H} . Note that $f(A, X, y) = 0$, with $Y = \phi(y)$. Assume that f is differentiable. We denote the differential of f in the X and y direction by

$$df_2(A, X, y) = \frac{\partial f}{\partial X} + \frac{\partial f}{\partial y}.$$

For all the factorizations, $df_2(A, X, y)$ can be easily computed because φ is linear in X and at most quadratic in Y .

Step 2 In order to apply the implicit function theorem to f at (A, X, y) , $df_2(A, X, y)$ has to be nonsingular. Thus, $\text{null}(df_2(A, X, y))$ needs to be computed, that is we need to solve the equation

$$df_2(A, X, y)(\Delta X, \Delta Y) = 0, \quad \Delta X \in \mathcal{E}, \quad \Delta Y = d\phi(y)\Delta y,$$

with $\Delta y \in \mathbb{K}^p$. Using Section 2.2, ΔY is in the tangent space of \mathcal{H} . Assume that $\text{null}(df_2(A, X, y)) = 0$. Then, by computing $d\varphi(X, Y)$, we have that

$$\{0\} = \text{range} \left(\frac{\partial \varphi}{\partial X} \right)_{|\mathbb{S}} \cap \text{range} \left(\frac{\partial \varphi}{\partial y} \right)_{|T(\mathcal{H})}.$$

Additionally, using (1)-(3) if $\dim \mathcal{E} = \dim \mathbb{S} + p$ then we have that $df_2(A, X, y)$ is invertible and the following splitting of \mathcal{E} into a direct sum decomposition of the type

$$\mathcal{E} = \text{range} \left(\frac{\partial \varphi}{\partial X} \right)_{|\mathbb{S}} \oplus \text{range} \left(\frac{\partial \varphi}{\partial y} \right)_{|T(\mathcal{H})}, \quad (8)$$

holds, where $T(\mathcal{H}_n)$ is the tangent space of \mathcal{H}_n at Y . The advantage of (8) is that it enable us to invert $df_2(A, X, y)$ by using the corresponding projector to the direct sum. Then, by the implicit function theorem, there exists a differentiable function $g = (g_X, g_Y)$ and an open neighborhood \mathcal{V}_A of A satisfying

$$\begin{aligned} g : \mathcal{V}_A &\rightarrow \mathcal{V}_X \times \mathcal{V}_Y, \\ \tilde{A} &\mapsto (g_X(\tilde{A}), g_Y(\tilde{A})), \end{aligned} \quad (9)$$

where $\mathcal{V}_X \times \mathcal{V}_Y$ is an open neighborhood of (X, Y) . Moreover, g satisfies $g_X(A) = X$, $g_Y(A) = Y$ and

$$dg(A) = -(d_2 f(A, X, y))^{-1} \frac{\partial f}{\partial A}, \quad (10)$$

$$f(\tilde{A}, g_X(\tilde{A}), g_Y(\tilde{A})) = 0, \quad \text{for all } \tilde{A} \in \mathcal{V}_A, \quad (11)$$

that is $\tilde{A} = \varphi(g_1(\tilde{A}), g_1(\tilde{A}))$ is the factorization of \tilde{A} . Let $\Pi_{\mathbb{S}}$ and $\Pi_{T(\mathcal{H}_n)}$ denote the projectors corresponding to (8). We have that $\frac{\partial f}{\partial A} = -I$. Thus, (10) becomes

$$dg_X(A)\Delta A = \left(\frac{\partial \varphi}{\partial X}\right)^{-1} \Pi_{\mathbb{S}}\Delta A, \quad (12)$$

$$dg_Y(A)\Delta A = \left(\frac{\partial \varphi}{\partial Y}\right)^{-1} \Pi_{T(\mathcal{H})}\Delta A. \quad (13)$$

Step 3 The condition number of the factorization is given by the norm of the linear map $dg(A)$. In some cases, only a bound for the norm of $dg(A)$ will be given. Finally, for $\tilde{A} \in \mathcal{V}_A$ and $\varphi(\tilde{X}, \tilde{Y}) = \tilde{A}$, the first order perturbation bounds and expansion are obtained using Taylor's theorem

$$\|\tilde{X} - X\|_F \leq \|dg_X(A)\|_2 \epsilon + O(\epsilon^2), \quad (14)$$

$$\|\tilde{Y} - Y\|_F \leq \|dg_Y(A)\|_2 \epsilon + O(\epsilon^2), \quad (15)$$

where $\epsilon = \|\tilde{A} - A\|_F$.

3 The HR Factorization

We say that $A \in \mathbb{C}^{n \times n}$ admits an HR factorization with respect to a signature matrix $J \in \text{diag}_n^k(\pm 1)$ if

$$A = HR, \quad R \in \Delta(\mathbb{C}), \quad H \in \mathcal{U}_n(J, \tilde{J}),$$

where $\tilde{J} \in \text{diag}_n^k(\pm 1)$. The next theorem from [4] shows that almost every matrix has an HR factorization with respect to J , where J is a signature matrix.

Theorem 3.1 *Let $A \in \mathbb{C}^{n \times n}$ be nonsingular and $J \in \text{diag}_n^k(\pm 1)$. There exist $H, R \in \mathbb{C}^{n \times n}$ such that $H^* J H \in \text{diag}_n^k(\pm 1)$, R is upper triangular and $A = HR$ if and only if all principal minors of $A^* J A$ are nonzero.*

For rectangular matrix $A \in \mathbb{C}^{m \times n}$, the HR factorization with respect to a signature matrix $J \in \text{diag}_m^k(\pm 1)$ is defined by

$$A = HR, \quad R \in \Delta(\mathbb{C}), \quad H \in \mathcal{U}_{mn}(J, \tilde{J}),$$

where $\tilde{J} \in \text{diag}_n^q(\pm 1)$. The following theorem extends Theorem 3.1 to rectangular matrices that have an HR factorization.

Theorem 3.2 *Let $A \in \mathbb{C}^{m \times n}$ with $m \geq n$ having full rank and $J \in \text{diag}_m^k(\pm 1)$. A has an HR factorization with respect to J if and only if all the principal minors of $A^* J A$ are nonzero.*

Proof. Assume that all the principal minors of A^*JA are nonzero. Then, like in Theorem 3.4, A^*JA can be factorized as

$$A^*JA = L|D|^{\frac{1}{2}}\tilde{J}_1|D|^{\frac{1}{2}}L^*,$$

where $L \in \mathbb{C}^{n \times n}$ is unit lower triangular, $D \in \mathbb{R}^{n \times n}$ is nonsingular diagonal and $\tilde{J} \in \text{diag}_n^q(\pm 1)$ for some integer q . Let $R = |D|^{\frac{1}{2}}L^*$ and define $H = AR^{-1} \in \mathbb{C}^{m \times n}$. We have that $H^*JH = \tilde{J}$.

We suppose now that $A = HR$, where H is a (J, \tilde{J}) -orthogonal and R upper triangular. Then

$$A^TJA = R^T H^T JHR = R^T \tilde{J}R.$$

Since A has full rank, R is nonsingular. Moreover,

$$A^TJA(1:k, 1:k) = R^T(1:k, 1:k)\tilde{J}(1:k, 1:k)R(1:k, 1:k), \quad k = 1:n,$$

which shows that all the leading principal submatrices of A are nonsingular. \square

3.1 Perturbation of the HR Factorization

Now that the definition of the HR factorization is given, our aim is to derive perturbation bounds for the H factor and the R factor when A is subject to some perturbation ΔA . In this section, we first generalize the results on the perturbation bounds for the HR factorization of square matrices in [1] to complex rectangular matrices and then we extend the results concerning the QR factorizations in [6, 19, 14] to the HR factorization of complex rectangular matrices. We also compute the condition number of the HR factorization.

Let $\mathcal{V}_h \subset \mathbb{R}^p$ with $p = \frac{n^2-n}{2}$ and according to Section 2.3 $H = \phi(h)$. Following the general method developed in Section 2.3, we define

$$\begin{aligned} f : \mathbb{C}^{m \times n} \times \Delta(\mathbb{C}) \times \mathcal{V}_h &\rightarrow \mathbb{C}^{m \times n}, \\ (\tilde{A}, \tilde{R}, \tilde{h}) &\mapsto \tilde{H}\tilde{R} - \tilde{A}, \end{aligned}$$

where from (7) $\tilde{H} = \phi(\tilde{h})$. We have that $\varphi(H, R) = HR$. We get

$$d_2f(A, h, R)(\Delta h, \Delta R) = \Delta HR + H\Delta R, \quad (16)$$

$$H^*Jd_2f(A, h, R)(\Delta h, \Delta R)R^{-1} = H^*J\Delta H + \tilde{J}\Delta RR^{-1}, \quad (17)$$

where $\Delta H = d\phi(h)\Delta h$. Note that $H^*J\Delta H \in \mathbf{SkewH}$ and $\tilde{J}\Delta RR^{-1} \in \Delta(\mathbb{C})$. We define the two projectors Π_1 and Π_2 by

$$\begin{aligned} \Pi_1 : \mathbb{C}^{n \times n} &\rightarrow \Delta(\mathbb{C}), & \Pi_1 &= \Pi_d + \Pi_u + \Pi_l^*, \\ \Pi_2 : \mathbb{C}^{n \times n} &\rightarrow \mathbf{SkewH}, & \Pi_2 &= \Pi_l - \Pi_l^*. \end{aligned}$$

Note that $X = (\Pi_1 + \Pi_2)X$ and $\text{range}(\Pi_1) \cap \text{range}(\Pi_2) = \emptyset$. Hence

$$\mathbb{C}^{n \times n} = \Delta(\mathbb{C}) \oplus \mathbf{SkewH}.$$

We have $\|\Pi_2(X)\|_F^2 = 2\|\Pi_l(X)\|_F^2$, thus, since Π_l is an orthogonal projection $\|\Pi_2\|_2 = \sqrt{2}$. It is straightforward to show that $\|\Pi_1\|_2 \leq \sqrt{2}$. This bound is attained by $X = \frac{\sqrt{2}}{2}(e_i e_j^T + e_j e_i^T)$. Thus, from (16) and (16) and using (12) we get

$$dg_R(A)\Delta A = \Pi_1(H^*J\Delta AR^{-1})R.$$

If $m = n$ then

$$dg_H(A)\Delta A = H\tilde{J}\Pi_2(H^*J\Delta AR^{-1}).$$

If $m > n$, then there exists $G = [H \ H_0] \in \mathbb{C}^{m \times m}$ such that $G^* J G \in \text{diag}_m^k(\pm 1)$ for some integer k . Note that G and k are obtained by a Gram-Schmidt type process. Thus,

$$\begin{aligned} dg_H(A)\Delta A &= JG^{-*} \begin{bmatrix} \Pi_2(H^* J \Delta A R^{-1}) \\ H_0^* J \Delta A R^{-1} \end{bmatrix}, \\ \|dg_H(A)\Delta A\|_F &= \|G\|_2 (\|\Pi_2\|_2^2 \|H\|_2^2 + \|H_0\|_2^2) \|R^{-1}\|_2^2 \|\Delta A\|_F^2)^{\frac{1}{2}}, \\ \|dg_H(A)\|_2 &\leq \sqrt{3}\kappa_2(G) \|R^{-1}\|_2. \end{aligned}$$

Finally, we obtain the bounds

$$\|dg_R(A)\|_2 \leq \sqrt{2}\kappa_2(R) \|H\|_2, \quad (18)$$

$$\|dg_H(A)\|_2 \leq \sqrt{2}\kappa_2(G) \|R^{-1}\|_2. \quad (19)$$

Following (14) and (15), we obtain the following theorem.

Theorem 3.3 *Let $A = HR$, $H \in \mathcal{U}_{mn}(J, \tilde{J})$ be the HR factorization of A and for $\Delta A \in \mathbb{C}^{n \times n}$ such that $\epsilon = \|\Delta A\|_F$ is small enough, let $A + \Delta A = \tilde{H}\tilde{R}$ be the HR factorization of $A + \Delta A$. Then*

$$\|\tilde{R} - R\|_F \leq \sqrt{2}\kappa_2(R) \|H\|_2 \epsilon + O(\epsilon^2), \quad (20)$$

$$\|\tilde{H} - H\|_F \leq \sqrt{2}\kappa_2(H) \|R^{-1}\|_2 \epsilon + O(\epsilon^2). \quad (21)$$

Theorem 3.3 generalizes the result in [1] to complex rectangular matrices and also extends the results concerning the QR factorizations in [6, 19, 14] to the HR factorization of complex rectangular matrices. The bounds are similar to those obtained in [1, 14].

If we apply our result to the particular case of the QR factorization, then we get the well-known bounds

$$\|\tilde{R} - R\|_2 \leq \sqrt{2}\kappa_2(A) \epsilon + O(\epsilon^2), \quad (22)$$

$$\|\tilde{H} - H\|_2 \leq \sqrt{2} \|A^{-1}\|_2 \epsilon + O(\epsilon^2). \quad (23)$$

(18) and (19) give a bound on the condition number of the HR factorization. The exact condition number can be obtained by using a Kronecker product approach. Let M_1 , M_2 , C and \hat{r} be defined by

$$\begin{aligned} M_1 &= (I \otimes R^* \tilde{J}) + (R^* \tilde{J} \otimes I) C \mathbf{T}, \\ M_2 &= (I \otimes A^* J) + (A^* J \otimes I) C \mathbf{T}, \\ \text{vec}(A^*) &= C \text{vec}(A). \end{aligned}$$

From (17) or by differentiating $(A, R) \mapsto R^* \tilde{J} R - A^* J A$, we get

$$R^* \tilde{J} \tilde{R} + \hat{R}^* \tilde{J} R = A^* J \Delta A + \Delta A^* J A. \quad (24)$$

Applying the vec operator to (24), we obtain

$$\begin{aligned} M_1 \hat{r} &= M_2 \text{vec}(\Delta A), \\ \hat{r} &= (M_1)_{|\Delta(\mathbb{C})}^{-1} M_2 \text{vec}(\Delta), \\ \|dg_R(A)\|_2 &= \|(M_1)_{|\Delta(\mathbb{C})}^{-1} M_2\|_2, \end{aligned} \quad (25)$$

where $(M_1)_{|\Delta(\mathbb{C})}$ is the restriction of M_1 to $\text{vec}(\Delta(\mathbb{C}))$. Combining (25) with the direct sum decomposition, we obtain

$$\begin{aligned} dg_H(A)\Delta A &= \Delta A R^{-1} - H(dg_R(A)\Delta A)R^{-1}, \\ \|dg_H(A)\|_2 &= \|R^{-T} \otimes I - (M_1)_{|\Delta(\mathbb{C})}^{-1} M_2\|_2, \end{aligned} \quad (26)$$

Using (25) and (26), we have the following theorem.

Theorem 3.4 *Let $A = HR$, $H \in \mathcal{U}_{mn}(J, \tilde{J})$ be the HR factorization of A and for $\Delta A \in \mathbb{C}^{n \times n}$ such that $\epsilon = \|\Delta A\|_F$ is small enough, let $A + \Delta A = \tilde{H}\tilde{R}$ be the HR factorization of $A + \Delta A$. Then, the sharpest perturbation bounds to first order are given by*

$$\|\tilde{R} - R\|_F \leq \|(M_1)_{|\Delta(C)}^{-1} M_2\|_2 \epsilon + O(\epsilon^2), \quad (27)$$

$$\|\tilde{H} - H\|_F \leq \|R^{-T} \otimes I - (M_1)_{|\Delta(C)}^{-1} M_2\|_2 \epsilon + O(\epsilon^2). \quad (28)$$

Theorem 3.3 is a generalization of the HR and QR factorization perturbation bounds that can be found in the literature. Although Theorem 3.3 and 3.4 are similar, the bounds in Theorem 3.4 are the best possible. In Table 2, we compare the bounds that are stated in these two theorems.

3.2 Numerical Experiments

The sensitivity of the HR factorization of A with respect to a signature matrix $J \neq \pm I$ is closely related to the minors of A^*JA . If one of the minors of A^*JA vanishes or is close to zero, then R is ill conditioned which implies that H is also ill conditioned or does not exist (if R is singular). To illustrate this fact numerically, we construct a sequence of matrices A_ϵ such that their first column $a_\epsilon = A_\epsilon(:, 1)$ is almost isotropic, that is, $a_\epsilon^T J a_\epsilon \rightarrow 0$ as $\epsilon \rightarrow 0$. We denote $\delta_\epsilon = \|A_{\epsilon_0} - A_\epsilon\|_F$ and $A_{\epsilon_0} = H_{\epsilon_0} R_{\epsilon_0}$. The results are in Table 1. In the second column of Table 1, the values of δ_ϵ are relatively small. We see that the values of $\|R_\epsilon - R_{\epsilon_0}\|_F$ in the third column and the values of $\|H_\epsilon - H_{\epsilon_0}\|_F$ in the fourth column do not depend on δ_ϵ . They depend instead on the values of $a_\epsilon^T J a_\epsilon$, in the sense that the bounds in the third and the sixth column get more accurate when $a_\epsilon^T J a_\epsilon$ increases and in the meantime the value δ_ϵ increases slowly. It confirms the fact that the sensitivity of the HR factorization depends on the minors of A^*JA . Note that the errors in R , in the third column (respectively H in the fifth column) are very close to the expected value in the fourth column (respectively the sixth column). This is due to the fact that we use the condition number. In the next numerical experiment, with the QR factorization, we see that if the bound is not sharp, then the expected values do not reflect the errors that are obtained.

Table 1: Perturbation bounds of the HR factorization.

$a_\epsilon^T J a_\epsilon$	δ_ϵ	$\ R_\epsilon - R\ _F$	$\ dg_R(A)\ _2 \delta_\epsilon$	$\ H_\epsilon - H\ _F$	$\ dg_H(A)\ _2 \delta_\epsilon$
$-7e - 8$	$2e - 15$	$1.77e - 4$	$5.98e - 4$	$6.3e - 4$	$2.37e - 4$
$-5e - 6$	$2e - 14$	$2.14e - 6$	$2.97e - 6$	$5.21e - 7$	$2.02e - 6$
$-2e - 4$	$2e - 13$	$1.34e - 7$	$2e - 7$	$3.78e - 8$	$1.51e - 7$
$-7e - 3$	$2e - 12$	$4.67e - 8$	$6.61e - 8$	$1.88e - 8$	$1.36e - 7$

For the QR factorization, we compare numerically (18) and (25). We consider the following 2×2 example

$$A_\epsilon = \begin{bmatrix} 1 - \epsilon & 1 \\ 1 & 1 + \epsilon \end{bmatrix}, \quad \kappa_2(A_\epsilon) = \epsilon^{-2}(1 + \sqrt{1 + \epsilon^2})^2.$$

Let $Q_\epsilon R_\epsilon = A_\epsilon$ be the QR factorization of A_ϵ and let $\Delta A_\epsilon = A_0 - A_\epsilon$. We have that $\|\Delta A_\epsilon\|_F = |\epsilon|\sqrt{2}$. The numerical results are in Table 2. Note that the expected values in the second column, computed with our condition number, are just twice the error on the R factor. Note that $\|A_0\|_F = 2$. Thus, if we use relative errors our bounds are the same as the computed values. The expected values obtained with

the usual bound are quite poor since the bound given by (22) is very poor. These results suggest that in this example the QR factorization of A_ϵ is a well conditioned problem independent of the condition number of the matrix that is factorized.

Table 2: Bounds of $\|dg_R(A)\|_2\|\Delta A_\epsilon\|_F$ and $\sqrt{2}\kappa_2(A_\epsilon)\|\Delta A_\epsilon\|_F$ as $\epsilon \rightarrow 0$.

ϵ	$\ R_\epsilon - R\ _F$	$\ dg_R(A)\ _2\epsilon$	$\sqrt{2}\kappa_2(A_\epsilon)\epsilon$
10^{-1}	$1.001e - 1$	$2.107e - 1$	$8e1$
10^{-2}	$1e - 2$	$2.01e - 2$	$8e2$
10^{-3}	$1e - 3$	$2e - 3$	$8e3$
10^{-4}	$1e - 4$	$2e - 4$	$8e4$
10^{-5}	$1e - 5$	$2e - 5$	$8e5$
10^{-6}	$1e - 6$	$2e - 6$	$8e6$

4 The Hyperbolic Singular Value Decomposition

Let $A \in \mathbb{R}^{m \times n}$ with $m \geq n$. We say that A admits a hyperbolic singular value decomposition (HSVD) if

$$A = QDH^T$$

with D diagonal, Q orthogonal and $H \in \mathcal{O}_n(J, \tilde{J})$. The hyperbolic singular value decomposition (HSVD) and the indefinite least square problem were analyzed in [2], [3] and [16]. \mathcal{E}_D denotes the set of real diagonal matrices. The following theorem establishes the existence of the HSVD. The theorem and the proof are similar to those in [3, Sec. 2].

Theorem 4.1 *Let $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ be a full rank matrix, $J \in \text{diag}_n^k(\pm 1)$ and assume that $\text{rank}(AJA^T) = n$. Then, there exists a positive nonsingular diagonal matrix $D \in \mathbb{R}^{m \times n}$, Q orthogonal, $\tilde{J} \in \text{diag}_n^k(\pm 1)$ and $H \in \mathcal{U}_n(J, \tilde{J})$ such that*

$$A = QDH^T.$$

Proof. Let $AJA^T = QSQ^T$ be an eigendecomposition. Assume that AJA^T is nonsingular. We define $D = |S|^{\frac{1}{2}}$ and $\tilde{J} = \text{sign}(S)$. Let

$$H = A^T Q \begin{bmatrix} D^{-1} \\ 0 \end{bmatrix}.$$

We have

$$H^T J H = \begin{bmatrix} D^{-1} \\ 0 \end{bmatrix}^T Q^T A J A^T Q \begin{bmatrix} D^{-1} \\ 0 \end{bmatrix} = \tilde{J}. \quad \square$$

In the definition of the HSVD, we see that only the n first columns of Q are necessary to define the decomposition. Thus, in the rest of this section, we consider that the HSVD of $A \in \mathbb{R}^{m \times n}$ is given by

$$A = QDH^T, \quad Q \in \mathcal{O}_{mn}(I), \quad H \in \mathcal{O}_n(J, \tilde{J}), \quad J, \tilde{J} \in \text{diag}_n^k(\pm 1).$$

4.1 Perturbation of the HSVD

The linear subspace of $n \times n$ real diagonal matrices is identified with \mathbb{R}^n and it is denoted by \mathcal{E}_D . Let

$$\begin{aligned} f : \mathbb{R}^{m \times n} \times \mathcal{E}_D \times \mathcal{V}_q \times \mathcal{V}_h &\rightarrow \mathbb{R}^{n \times n}, \\ (\tilde{A}, \tilde{D}, \tilde{q}, \tilde{h}) &\mapsto \tilde{Q}\tilde{D}\tilde{H}^T - \tilde{A}, \end{aligned}$$

with $\tilde{Q} = \phi_1(\tilde{q})$ and $\tilde{H} = \phi_2(\tilde{h})$, where ϕ_1 and ϕ_2 are defined by (7). Note that $f(A, D, q, h) = 0$. We define $d_2f = \frac{\partial f}{\partial D} + \frac{\partial f}{\partial q} + \frac{\partial f}{\partial h}$, $\Delta Q = d\phi_1(q)\Delta q$ and $\Delta H = d\phi_2(h)\Delta h$. We have

$$\begin{aligned} d_2f(A, D, q, h)(\Delta A, \Delta Q, \Delta H) &= \Delta Q D H^T + Q D \Delta H^T + Q \Delta D H^T, \\ Q^T d_2f(A, D, q, h)(\Delta A, \Delta Q, \Delta H) H^{-T} &= Q^T \Delta Q D + D \Delta H^T H^{-T} + \Delta D, \end{aligned}$$

with $Q^T \Delta Q$ and $\Delta H^T H^{-T} \tilde{J}$ skew-symmetric. The following lemma establishes the direct sum decomposition (8).

Lemma 4.2 *Let $D = \text{diag}(\lambda_k)$. If the diagonal elements of $\tilde{J}D^2$ are distinct, then we have the following direct sum decomposition*

$$\mathbb{R}^{n \times n} = \mathcal{E}_D \oplus \mathbf{Skew}(\mathbb{R})D \oplus D\mathbf{Skew}(\mathbb{R})\tilde{J}.$$

The corresponding projector Π_1 on \mathcal{E}_D is just Π_d whereas for all $Z \in \mathbb{R}^{n \times n}$ the projector on $\mathbf{Skew}\mathbf{H}D$ is $\Pi_2 D$ and the projector on $D\mathbf{Skew}\mathbf{H}\tilde{J}$ is $D\Pi_3\tilde{J}$ where

$$\Pi_2(Z) = (\tilde{J}ZD + DZ^T\tilde{J}) \circ \Lambda, \quad \Pi_3(Z) = (DZ + Z^T D) \circ \Lambda, \quad (29)$$

$$\Lambda = (\mu_{ij}), \quad \mu_{ij} = \begin{cases} 0 & \text{if } i = j, \\ \frac{1}{\tilde{\sigma}_j \lambda_i^2 - \tilde{\sigma}_i \lambda_j^2} & \text{otherwise,} \end{cases} \quad (30)$$

where $D = \text{diag}(\lambda_i)$ and $\tilde{J} = \text{diag}(\tilde{\sigma}_i)$. Moreover, the norms of the operators, $\|\Pi_2\|_2$ and $\|\Pi_3\|_2$ are given by

$$\|\Pi_2\|_2 = \|\Pi_3\|_2 = \sqrt{2} \max_{i \neq j} \frac{\sqrt{\lambda_i^2 + \lambda_j^2}}{|\tilde{\sigma}_i \lambda_j^2 - \tilde{\sigma}_j \lambda_i^2|}. \quad (31)$$

Proof. Let $Z \in \mathbb{R}^{m \times n}$, $Z = (z_{ij})$ and assume that $\Delta D + XD + DY\tilde{J} = Z$ where $\Delta D \in \mathcal{E}_D$ and $X, Y \in \mathbf{Skew}(\mathbb{R})$. Since X and Y are skew symmetric, we have $\Pi_d(XD + DY) = 0$. Thus,

$$\Delta D = \Pi_d(Z).$$

By computing the elements of $(\Pi_l + \Pi_u)(XD - DY\tilde{J})$, we get $\frac{n^2-n}{2}$ 2×2 linear systems

$$E_{ij} \begin{bmatrix} x_{ij} & y_{ij} \end{bmatrix}^T = \begin{bmatrix} z_{ij} & z_{ji} \end{bmatrix}^T,$$

where E_{ij} is defined by

$$E_{ij} = \begin{bmatrix} \lambda_j & \tilde{\sigma}_i \lambda_i \\ -\lambda_i & -\tilde{\sigma}_j \lambda_j \end{bmatrix}.$$

We have $\det E_{ij} = \tilde{\sigma}_i \tilde{\sigma}_j (\tilde{\sigma}_j \lambda_j^2 - \tilde{\sigma}_i \lambda_i^2) \neq 0$. Hence, for $i \neq j$, we obtain the $\frac{n^2-n}{2}$ solutions

$$x_{ij} = -\frac{\sigma_i \lambda_j z_{ij} + \tilde{\sigma}_j \lambda_i z_{ji}}{\tilde{\sigma}_j \lambda_i^2 - \tilde{\sigma}_i \lambda_j^2}, \quad (32)$$

$$y_{ij} = \frac{\lambda_i z_{ij} + \lambda_j z_{ji}}{\tilde{\sigma}_j \lambda_i^2 - \tilde{\sigma}_i \lambda_j^2}. \quad (33)$$

With (32) and (33), we obtain

$$X = -(\tilde{J}ZD + DZ^T\tilde{J}) \circ \Lambda, \quad (34)$$

$$Y = (DZ + Z^TD) \circ \Lambda, \quad (35)$$

where $\Lambda = (\mu_{ij})$ is given by (30). Finally, we obtain $\Pi_1 = \Pi_d$, $\Pi_2(Z) = X$ and $\Pi_3(Z) = Y$.

Using the Cauchy-Schwarz inequality, we have that

$$x_{ij}^2 \leq \frac{\lambda_i^2 + \lambda_j^2}{\tilde{\sigma}_i \lambda_j^2 - \tilde{\sigma}_j \lambda_i^2} (z_{ij}^2 + z_{ji}^2),$$

$$\|X\|_F^2 \leq 2 \max_{ij, i \neq j} \frac{\lambda_i^2 + \lambda_j^2}{|\tilde{\sigma}_i \lambda_j^2 - \tilde{\sigma}_j \lambda_i^2|} \|Z\|_F^2.$$

The bound (61) is attained by

$$E = \frac{\sigma_p \lambda_q e_{pq} + \sigma_q \lambda_p e_{qp}}{\sqrt{\lambda_p^2 + \lambda_q^2}},$$

with $\|E\|_F = 1$ and where (p, q) are the indices where

$$\max_{ij, i \neq j} \frac{\sqrt{\lambda_i^2 + \lambda_j^2}}{|\tilde{\sigma}_i \lambda_j^2 - \tilde{\sigma}_j \lambda_i^2|}$$

is attained. Similarly, using the same method, we show the second part of (61) and that the bound is attained by

$$\tilde{E} = \frac{\lambda_p e_{pq} + \lambda_q e_{qp}}{\sqrt{\lambda_p^2 + \lambda_q^2}},$$

with $\|\tilde{E}\|_F = 1$. \square

To characterize g , we proceed as in (10) and (12-13). We have $\frac{\partial f}{\partial A} \Delta A = -\Delta A$. We set $(\hat{D}, \hat{Q}, \hat{H}) = (dg_D(A)\Delta A, dg_Q(A)\Delta A, dg_H(A)\Delta A)$. We obtain the linear system

$$\begin{aligned} \hat{Q}DH^T + QD\hat{H}^T + Q\hat{D}H &= \Delta A, \\ Q^T\hat{Q}D + D\hat{H}^T JH\tilde{J} + \hat{D} &= Q^T \Delta A JH\tilde{J}, \\ Q^T\hat{Q} + \hat{Q}^T Q &= 0, \\ \hat{H}^T JH + H^T J\hat{H} &= 0. \end{aligned} \quad (36)$$

Thus, by Lemma 4.2,

$$dg_D(A)\Delta A = \Pi_d(Q^T \Delta A JH\tilde{J}), \quad (37)$$

$$dg_H(A)\Delta A = \tilde{J}H^T J \Pi_3(Q^T \Delta A JH\tilde{J}). \quad (38)$$

If $m = n$, then

$$dg_Q(A)\Delta A = Q \Pi_2(Q^T \Delta A JH\tilde{J}). \quad (39)$$

If $m > n$, then we know that there exist $G = [Q, Q_0] \in \mathbb{R}^{n \times n}$ such that $G^T G = I$. G is obtained as in Section 3 by the classical Gram-Schmidt process. Using (36), we have

$$dg_Q(A)\Delta A = G \begin{bmatrix} \Pi_2(Q^T \Delta A JH\tilde{J}) \\ Q_0^T \Delta A H^{-T} D^{-1} \end{bmatrix}. \quad (40)$$

Let \tilde{h}_k denote the k -th column $H\tilde{J}$. We have

$$\begin{aligned}\|dg_D(A)\Delta A\|_2 &= \sup_{\|\Delta A\|_F=1} \|\Pi_d(Q^T \Delta A J H \tilde{J})\|_F, \\ &= \sup_{\|\Delta A\|_F=1} \|\Pi_d(\Delta A H \tilde{J})\|_F, \\ &= \|W\|_2,\end{aligned}$$

where $W \in \mathbb{R}^{n \times n^2}$ has its k -th row defined by $\tilde{h}_k^T \otimes e_k^T$. Thus,

$$\|dg_D(A)\Delta A\|_2 = \|W\|_2 = \max_k \|\tilde{h}_k\|_2 = \max_k \|H(k, :)\|_2. \quad (41)$$

We define

$$M_1 = \tilde{J}H^T J \otimes Q^T, \quad (42)$$

$$M_2 = D \otimes \tilde{J} + (\tilde{J} \otimes D)\mathbf{T}, \quad \tilde{M}_2 = I \otimes D + (D \otimes I)\mathbf{T}. \quad (43)$$

Applying the vec operator to (38) and taking norms, we obtain

$$\|dg_H(A)\|_2 = \|(I \otimes H^T J) \text{diag}(\text{vec}(\Lambda)) \tilde{M}_2 M_1\|_2. \quad (44)$$

Similarly, for Q factor, we obtain from (40)

$$\|dg_Q(A)\|_2 = \begin{cases} \|\text{diag}(\text{vec}(\Lambda))M_2 M_1\|_2, & \text{if } m = n, \\ \|\text{diag}(\text{vec}(\Lambda))M_2 M_1 + (D^{-1}H^{-1}) \otimes I_n\|_2, & \text{if } m > n. \end{cases} \quad (45)$$

We are now able to give the first order expansion of the three factors of the HSVD.

Theorem 4.3 *Let $A = QDH^T$, $H \in \mathcal{O}_n(J, \tilde{J})$ be the HSVD of A and for $\Delta A \in \mathbb{R}^{n \times n}$ such that $\epsilon = \|\Delta A\|_F$ is small, let $(A + \Delta A) = \tilde{Q}\tilde{D}\tilde{H}^T$ be the HSVD of $A + \Delta A$. Then, using (45-44) and (41-50)*

$$\|D - \tilde{D}\|_F \leq \max_k \|H(k, :)\|_2 \epsilon + O(\epsilon^2), \quad (46)$$

$$\|Q - \tilde{Q}\|_F \leq \|dg_Q(A)\|_2 \epsilon + O(\epsilon^2), \quad (47)$$

$$\|H - \tilde{H}\|_F \leq \|dg_H(A)\|_2 \epsilon + O(\epsilon^2), \quad (48)$$

where $\|dg_Q(A)$ and $\|dg_H(A)\|_2$ are given by (44) and (45). These bounds are the sharpest possible to first order.

Using (39) and (38), note that the condition number of the HSVD can be bounded by

$$\|dg_Q(A)\|_2 \leq \begin{cases} \frac{2}{m} \|D\|_2 \|H\|_2, & \text{if } m = n, \\ \left(\frac{4}{m^2} \|D\|_2^2 \|H\|_2^2 + \|H^{-T} D^{-1}\|_2^2 \right)^{\frac{1}{2}}, & \text{if } m > n, \end{cases} \quad (49)$$

$$\|dg_H(A)\|_2 \leq \frac{2}{m} \|D\|_2 \kappa_2(H), \quad (50)$$

where $m = \min_{i,j} |\tilde{\sigma}_i \lambda_i^2 - \tilde{\sigma}_j \lambda_j^2| = \|\text{diag}(\text{vec}(\Lambda))\|_2$. These bounds are less sharp than (45) and (44) but they are easily computable. We also can give better bounds than

(49)-(50), using (31),

$$\|dg_Q(A)\|_2 \leq \begin{cases} \alpha\|H\|_2, & \text{if } m = n, \\ \left(\alpha^2\|H\|_2^2 + \|H\tilde{J}D^{-1}\|_2^2\right)^{\frac{1}{2}}, & \text{if } m > n, \end{cases} \quad (51)$$

$$\|dg_H(A)\|_2 \leq \alpha\kappa_2(H), \quad (52)$$

where $\alpha = \|\Pi_2\|_2 = \|\Pi_3\|_2$ is defined in (31).

For the usual SVD, H is orthogonal. We get the well-known result (see for example [20]) for the singular values

$$\|dg_D(A)\|_2 = 1, \quad \|D - \tilde{D}\|_F \leq \epsilon + O(\epsilon^2).$$

The condition numbers of Q and H can be easily computed since H is also orthogonal

$$\|dg_Q(A)\|_2 = \begin{cases} \alpha, & \text{if } m = n, \\ \left(\alpha^2 + \frac{1}{\lambda_n^2}\right)^{\frac{1}{2}}, & \text{if } m > n, \end{cases} \quad (53)$$

$$\|dg_H(A)\|_2 = \alpha, \quad (54)$$

where $\alpha = \|\Pi_2\|_2 = \|\Pi_3\|_2$ is defined in (31) and λ_n is the smallest singular value of A . In [12] and [18], a bound for the singular vectors is proposed. This bound is obtained by applying the fact that H is orthogonal in (49) and (50).

4.2 Numerical Experiments

We consider a 3×3 example with

$$D_0 = \text{diag}(10, 9.9, 1) \quad \text{and} \quad A_0 = U_0 D_0 V_0^T,$$

where (U_0, V_0) is a randomly generated orthogonal-hyperbolic matrix pair and the signature matrices (J, \tilde{J}) such that $V_0^T J V = \tilde{J}$ are defined by

$$\begin{aligned} J &= \text{diag}(-1, -1, 1), \\ \tilde{J} &= \text{diag}(-1, 1, -1). \end{aligned}$$

We construct a sequence of matrices $A_\epsilon = U_0 D_\epsilon V_0^T$ with $D_\epsilon = D_0 + \epsilon e_2 e_2^T$ where e_2 denotes the second column of the identity. The results are in Table 3 for the singular values and in Table 4 for the orthogonal and hyperbolic factor, with $A_\epsilon = U_\epsilon D_\epsilon V_\epsilon^T$ be the HSVD of A_ϵ , $\delta_\epsilon = \|A_0 - A_\epsilon\|_F$. We see that the expected bound for the hyperbolic singular values are very close to the computed values. It is due to the fact that the bound on the hyperbolic singular value depends only on the norm of the hyperbolic factor

$$V_\epsilon = A_\epsilon^T U_\epsilon D_\epsilon^{-1},$$

with $\kappa_2(D_\epsilon) = 10$. The orthogonal and hyperbolic factors are more sensitive to the fact that one of the hyperbolic singular values is becoming double which does not appear easily in Theorem 4.3. But, we see in the expressions of dg_Q and dg_H in (38) and (39)–(40) that the sensitivity of the orthogonal and hyperbolic factors depend on Π_2 and Π_3 in (29). Moreover, the norms of these projectors (31) vary proportionally to the inverse of $\min_{i,j} |\lambda_j| - |\lambda_i|$ which explains the numerical test

in Table 4. The bounds in the last row of Table 4 (column 3 and 5) are quite poor. The first explanation is the fact that the value of $\delta_\epsilon = 10^{-5}$ is big, the corresponding perturbation is not in the required neighborhood \mathcal{V}_A (see Section 2.3) in order to apply the implicit function theorem. Consequently, the result on the condition number and the perturbation expansion in Theorem 4.3 are not valid. Another fact that we need to keep in mind is that the perturbation expansion given in Theorem 4.3 gives a bound for the predicted result but it does not guaranty any accuracy of these bounds.

In Figure 1, we plot the logarithms of the exact condition number of the orthogonal and hyperbolic factor, the bounds given by (49), (50), (51) and (52), against the value of ϵ . The exact condition number for the orthogonal factor $\|dg_Q(A)\|_2$ and the condition number for hyperbolic factor $\|dg_H(A)\|_2$ are represented by \circ and \square . We denote by $c_{Q,1}$ and $c_{H,1}$ the bounds of the condition numbers given by (49), (50) and we denote by $c_{Q,2}$ and $c_{H,2}$ the bounds defined by (51) and (52). In Figure 1, the symbols $+$ and \triangleleft represent the logarithm of the bounds given by $c_{Q,1}$ and $c_{H,1}$ whereas the symbols \star and \triangleright are for the logarithms of $c_{Q,2}$ and $c_{H,2}$. These values are labeled by $\log_{10}(c)$ on the y -axes. We see that the condition number and the bounds are of the same order and seem to be the same asymptotically.

Table 3: Perturbation bounds for the hyperbolic singular values.

δ_ϵ	$\ D_0 - D_\epsilon\ _F$	$\ dg_D(A_0)\ _2 \delta_\epsilon$
10^{-13}	9.10^{-14}	10^{-13}
10^{-10}	2.10^{-10}	3.10^{-10}
10^{-6}	9.10^{-7}	1.10^{-6}
3.10^{-5}	3.10^{-5}	3.10^{-5}

Table 4: Perturbation bounds for the orthogonal and hyperbolic factor.

δ_ϵ	$\ Q_\epsilon - Q_0\ _F$	$\ dg_Q(A_0)\ _2 \delta_\epsilon$	$\ H_\epsilon - H_0\ _F$	$\ dg_H(A_0)\ _2 \delta_\epsilon$
10^{-13}	10^{-11}	10^{-9}	10^{-11}	$1.4.10^{-9}$
2.10^{-10}	10^{-12}	10^{-6}	10^{-12}	10^{-6}
10^{-6}	4.10^{-12}	10^{-1}	4.10^{-12}	1.10^{-1}
10^{-5}	10^{-12}	4	10^{-12}	4

The behaviour of the usual SVD and HSVD can be quite different and unexpected. For $n = 2$, if the two singular values are close then the condition number of the singular vector is large since the condition number for the orthogonal factors, (53)–(54) is unbounded. In the hyperbolic case, with $J = \text{diag}(1, -1)$, the condition number for the orthogonal factor and the hyperbolic factor is uniformly bounded on any subset of $\mathbb{R}^{2 \times 2} \setminus \{0\}$ at a positive distance of the zero matrix.

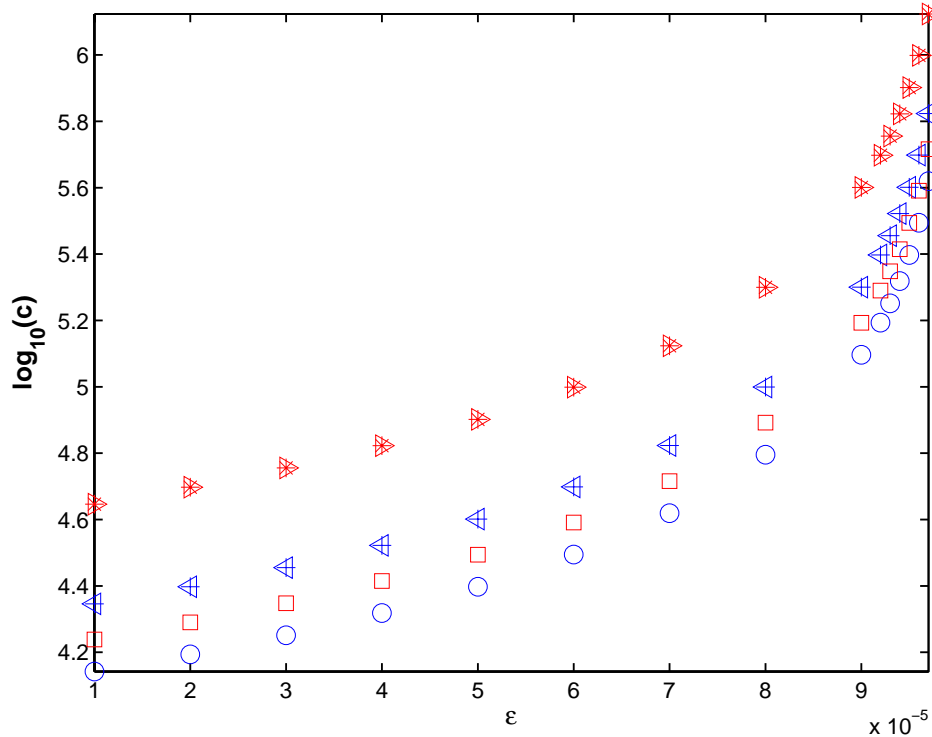


Figure 1: Comparison between the condition number and its bounds with $\log_{10}(\|dg_Q(A)\|_2)$ (\circ), $\log_{10}(\|dg_H(A)\|_2)$ (\square), $\log_{10}(c_{Q,1})$ (\triangleleft), $\log_{10}(c_{H,1})$ (\triangleright), $\log_{10}(c_{Q,2})$ ($*$) and $\log_{10}(c_{H,2})$ (\triangleright).

5 The Indefinite Polar Factorization

We say that $A \in \mathbb{R}^{n \times n}$ admits a polar factorization if $A = HS$ with H orthogonal and S symmetric definite positive. The indefinite polar factorization (IPF) is a generalization of the usual polar factorization, that is, we want to generalize the polar decomposition with H (J, \tilde{J})-orthogonal.

$$A = HS, \quad H^T JH = J,$$

The following theorem from [11] allows us to define this decomposition and it gives necessary conditions for the existence and uniqueness of the IPF.

Theorem 5.1 *If $A \in \mathbb{R}^{n \times n}$ and $JA^T J A$ has no eigenvalues on the nonpositive real axis, then A has a unique IPF $A = HS$, where H is (J, J)-orthogonal and S is J -symmetric with eigenvalues in the open right half-plane.*

In this paper, we define the IPF as in Theorem 5.1. Throughout this section, we assume that S is diagonalizable.

5.1 Perturbation of the IPF

We start by a preliminary result that will enable us to give the direct sum decomposition like in (8). We focus on some matrix equations that arise in the next sections. Let $A \in \mathbb{R}^{n \times n}$ be diagonalizable with the eigendecomposition $A = VDV^{-1}$, with

$D = \text{diag}(\lambda_k)$. For $X \in \mathbf{Sym}(\mathbb{R})$, we consider the equation

$$AX + XA^T = Z,$$

where $Z \in \mathbf{Sym}(\mathbb{R})$. The matrix operator that arise naturally is then defined on $\mathbf{Sym}(\mathbb{R})$ by

$$\mathcal{T}_A X = AX + XA^T. \quad (55)$$

We define $M \in \mathbb{C}^{n \times n}$

$$M = \left(\frac{1}{\lambda_i + \lambda_j} \right)_{ij}. \quad (56)$$

Lemma 5.2 \mathcal{T}_A is invertible if for all k_1, k_2 , such that $1 \leq k_1, k_2 \leq n$ and $k_1 \neq k_2$, $\lambda_{k_1} + \lambda_{k_2} \neq 0$. Then,

$$\mathcal{T}_A^{-1} Z = V((V^{-1} Z V^{-T}) \circ M) V^T,$$

where V is the matrix containing the eigenvectors of A .

Proof. We consider the equation $\mathcal{T}_A X = Z$. We have

$$\begin{aligned} V D V^{-1} X + X V^{-T} D V^T &= Z, \\ D \tilde{X} + \tilde{X} D &= \tilde{Z}, \end{aligned}$$

where $\tilde{X} = V^{-1} X V^{-T}$ and $\tilde{Z} = V^{-1} Z V^{-T}$. Note that \tilde{X} and \tilde{Z} are complex symmetric. If the eigenvalues have the properties required in each case then the solution exists and is unique. It is given by

$$X = V(\tilde{Z} \circ M) V^T.$$

We need to show now that X is real. Without loss of generality, assume that

$$\begin{aligned} V &= [V_1 \ V_2 \ \overline{V_2}], \quad V^{-T} = [U_1^T \ U_2^T \ \overline{U_2^T}] \quad \text{and} \\ D &= \text{diag}(D_1, D_2, \overline{D_2}), \end{aligned}$$

where V_1, U_1 and D_1 are real and V_2, U_2 and D_2 are complex with a nontrivial imaginary part. Then, $V = \overline{V} P$ and $V^T = P V^*$, where

$$P = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix}.$$

For $Y \in \mathbb{C}^{n \times n}$, $P \overline{Y} P = Y$ if and only if

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & \overline{Y_{12}} \\ Y_{21} & Y_{22} & Y_{23} \\ \overline{Y_{21}} & \overline{Y_{23}} & \overline{Y_{22}} \end{bmatrix},$$

with Y_{11} real. Note that

$$\begin{aligned} P \overline{\tilde{Z}} P &= \tilde{Z} \quad \text{and} \quad P \overline{M_{\pm}} P = M, \\ P \overline{\tilde{Z} \circ M_{\pm}} P &= \tilde{Z} \circ M. \end{aligned}$$

Hence, $\overline{X} = X$, X is real. \square

Similarly, we define on $\mathbf{Skew}(\mathbb{R})$, the matrix operator \tilde{T}_A by $\tilde{T}_A X = AX + XA^T$. Under the same assumption on the eigenvalues of A in Lemma 5.2, we can show that \tilde{T}_A is invertible and

$$\tilde{T}_A^{-1} X = V((V^{-1}ZV^{-T}) \circ M)V^T, \quad (57)$$

where M is defined by (56).

We now focus on the IPF. We assume that A is nonsingular and that it admits the IPF $A = HS$, $H \in \mathcal{O}_n(J, J)$. Our aim is to derive perturbation bounds for the H factor and the factor S when A is subject to some perturbation ΔA . Using (7), we define

$$\begin{aligned} f : \mathbb{R}^{n \times n} \times \mathcal{V}_h \times \mathbf{JSym}(\mathbb{R}) &\rightarrow \mathbb{R}^{n \times n}, \\ (\tilde{A}, \tilde{S}, \tilde{h}) &\mapsto \tilde{H}\tilde{S} - \tilde{A}, \end{aligned}$$

where $\tilde{H} = \phi(\tilde{h})$ and $H = \phi(h)$ and ϕ is defined by (7). Note that $f(A, h, S) = 0$. We define $d_2 f = \frac{\partial f}{\partial h} + \frac{\partial f}{\partial S}$. We have

$$\begin{aligned} d_2 f(A, h, S)(\Delta h, \Delta S) &= \Delta HS + H\Delta S, \\ H^T J d_2 f(A, h, S)(\Delta h, \Delta S)S^{-1} &= H^T J\Delta H + J\Delta SS^{-1}, \end{aligned}$$

where $\Delta H = d\phi(h)\Delta h$. Note that $H^T J\Delta H \in \mathbf{Skew}(\mathbb{R})$. In the following lemma, we establish the direct sum decomposition as in (8).

Lemma 5.3 *Let $J \in \text{diag}_n^q(\pm 1)$ and let S be nonsingular, J -symmetric such that the eigenvalues of JS are positive. Then,*

$$\mathbb{R}^{n \times n} = \mathbf{Skew}(\mathbb{R}) \oplus \mathbf{Sym}(\mathbb{R})S^{-1}.$$

Furthermore, let Π_1 be the projector on $\mathbf{Skew}(\mathbb{R})$ and Π_2 be the projector on $\mathbf{Sym}(\mathbb{R})S^{-1}$. Then,

$$\Pi_1(Z) = \tilde{T}_{S^T}^{-1}(ZS - S^T Z), \quad (58)$$

$$\Pi_2(Z) = \mathcal{T}_{S^T}^{-1}(S^T(Z + Z^T)S)S^{-1}, \quad (59)$$

where \mathcal{T}_{S^T} is defined in Lemma 5.2 and \tilde{T}_{S^T} is given by (57).

Proof. Let $Z \in \mathbb{R}^{n \times n}$ and consider the equation $X + YS^{-1} = Z$ with $X \in \mathbf{Skew}(\mathbb{R})$ and $Y \in \mathbf{Sym}(\mathbb{R})$. We have that $-X + S^{-T}Y = Z^T$. Thus,

$$\begin{aligned} S^T X + X S &= Z X - S^T Z, \\ S^T Y + Y S &= S^T(Z^T + Z)S. \end{aligned}$$

We see then the solutions are given by

$$X = \tilde{T}_{S^T}^{-1}(ZS - S^T Z) \quad \text{and} \quad Y = \mathcal{T}_{S^T}^{-1}(S^T(Z + Z^T)S). \quad \square$$

To characterize g , we proceed as follows. We have $\frac{\partial f}{\partial A}(A, S, h) = -\Delta A$. We set $(\hat{H}, \hat{S}) = (dg_H(A)\Delta A, dg_S(A)\Delta A)$. Thus,

$$\hat{H}S + H\hat{S} = \Delta A \quad \text{and} \quad \hat{H}^T JH + H^T J\hat{H} = 0.$$

Let $X = H^T J\hat{H} \in \mathbf{Skew}(\mathbb{R})$ and $\widetilde{\Delta A} = H^T J\Delta A$. Thus,

$$S^T X + X S = \widetilde{\Delta A} - \widetilde{\Delta A}^T, \quad (60)$$

$$S^T J\hat{S} + \hat{S}^T J S = S^T \widetilde{\Delta A} + \widetilde{\Delta A}^T S. \quad (61)$$

Thus, we obtain

$$dg_H(A)\Delta A = HJ\tilde{\mathcal{T}}_{S^T}^{-1}(H^T J\Delta A - \Delta A^T JH), \quad (62)$$

$$dg_S(A)\Delta A = J\mathcal{T}_{S^T}^{-1}(S^T)(A^T J\Delta A + \Delta A^T JA). \quad (63)$$

We define

$$\begin{aligned} M_1 &= (V \otimes V^T)\text{diag}(\text{vec}(M))(V^{-1} \otimes V^{-T}), \\ M_2 &= -(H^T J \otimes I)\mathbf{T} + I \otimes H^T J, \\ \tilde{M}_2 &= (A^T J \otimes I)\mathbf{T} + I \otimes A^T J. \end{aligned}$$

Then, applying the vec operator, we obtain

$$\|dg_H(A)\|_2 = \|(I \otimes HJ)M_1 M_2\|_2, \quad (64)$$

$$\|dg_S(A)\|_2 = \|M_1 \tilde{M}_2\|_2. \quad (65)$$

Using (64)-(65), we have the following theorem.

Theorem 5.4 *Let $A = HS$, $H \in \mathcal{O}_n(J)$ be the IPF of A and for $\Delta A \in \mathbb{R}^{n \times n}$ such that $\epsilon = \|\Delta A\|_F$ is small enough, let $A + \Delta A = \tilde{H}\tilde{S}$ be the IPF of $A + \Delta A$. Then*

$$\begin{aligned} \|\tilde{S} - S\|_F &\leq \|M_1 \tilde{M}_2\|_2 \epsilon + O(\epsilon^2), \\ \|\tilde{H} - H\|_F &\leq \|M_1 M_2\|_2 \epsilon + O(\epsilon^2), \end{aligned}$$

where M_1 , M_2 and \tilde{M}_2 are the matrices involved in the differential of the implicit function in (64) and (65)

The above theorem gives the perturbation expansion of the IPF for a nonsingular A . If A is singular and 0 is at most a simple eigenvalue of A then it is possible to give the perturbation bounds of the factor S . We just need to apply the implicit function theorem to $(A, S) \mapsto S^T JS - A^T JA$. Also, from (64)–(65), we can give bounds of the condition number that are less expensive to compute than the exact condition numbers:

$$\|dg_H(A)\|_2 \leq 2m\kappa_2(V)^2\kappa_2(H), \quad (66)$$

$$\|dg_S(A)\|_2 \leq 2m\kappa_2(V)^2\|A\|_2, \quad (67)$$

where $m = \max_{ij} |m_{ij}|$ and $M = (m_{ij})$ is defined by (56).

5.2 The Polar Factorization

The polar factorization is the particular case that is obtained when $J = \pm I$. Thus, $A = QS$ is the polar factorization of A , with Q orthogonal and S symmetric. Note that if A is complex then the perturbation bounds remain the same for the unitary Q factor and the Hermitian factor S . In [1], a perturbation bound for the Hermitian factor that involves the 2-norm of A is given but in [8] and [9], the author found a constant bound $\sqrt{2}$. With our method, we obtain the condition number for the Hermitian factor and for the unitary factor in a simpler way than [5]. We proceed as follows.

Lemma 5.5 *Let the two matrix operators \mathcal{T}_1 and \mathcal{T}_2 be defined by $\mathcal{T}_1 X = (X - X^T) \circ M$ and $\mathcal{T}_2 X = (DX + X^T D) \circ M$ where M is defined in (56) and D real diagonal matrix with positive entries. Then*

$$\|\mathcal{T}_1\|_2 = \frac{2}{\lambda_{n-1} + \lambda_n}, \quad (68)$$

$$\|\mathcal{T}_2\|_2 = \sqrt{2} \frac{\sqrt{\lambda_n^2 + \lambda_1^2}}{\lambda_n + \lambda_1}, \quad (69)$$

where λ_{n-1} and λ_n are the two smallest diagonal entries of D and λ_1 the largest diagonal entry of D .

Proof. Let $X = (x_{ij}) \in \mathbb{R}^{n \times n}$ and assume that $Y = \mathcal{T}_1(X)$ with $Y = (y_{ij})$. We have

$$\begin{aligned} \|Y\|_F^2 &= \sum_{i,j=1}^n \frac{(x_{ij} - x_{ji})^2}{(\lambda_i + \lambda_j)^2} \leq 4 \sum_{i,j=1}^n \frac{x_{ij}^2 + x_{ji}^2}{(\lambda_i + \lambda_j)^2}, \\ \|Y\|_F &\leq \frac{2}{\lambda_{n-1} + \lambda_n} \|X\|_F. \end{aligned}$$

The bound in (68) is attained by $E = \frac{1}{\sqrt{2}}(e_n e_{n-1}^T - e_{n-1} e_n^T)$ where e_k is the k -th column of the identity matrix.

We now focus on (69). Assume that $Y = \mathcal{T}_1(X)$ with $Y = (y_{ij})$. We have that $y_{ij} = \frac{1}{\lambda_i + \lambda_j}(\lambda_i x_{ij} + \lambda_j x_{ji})$, $y_{ii} = x_{ii}$. We define

$$\mu = \max_{i,j} \left(\frac{\lambda_i^2 + \lambda_j^2}{(\lambda_i + \lambda_j)^2} \right).$$

and we have that $y_{ij}^2 \leq \mu(x_{ij}^2 + x_{ji}^2)$. Thus,

$$\begin{aligned} \|Y\|_F^2 &= \sum_{i=1}^n x_{ii}^2 + \sum_{i=2}^n \sum_{j=1}^{i-1} 2y_{ij}^2 \leq 2 \max_{i,j} \left(\frac{\lambda_i^2 + \lambda_j^2}{(\lambda_i + \lambda_j)^2} \right) \|X\|_F^2, \\ \|\mathcal{T}_2\|_2 &\leq \sqrt{2\mu}. \end{aligned}$$

Let

$$E = \frac{1}{\sqrt{\lambda_p^2 + \lambda_q^2}} (\lambda_p e_p e_q^T + \lambda_q e_q e_p^T)$$

with (p, q) the indices where μ is attained. Note that $\|E\|_F = 1$ and $\|\mathcal{T}_2(E)\|_F = 1$. Without loss of generality, assume that $\lambda_p \leq \lambda_q$ and define $t = \frac{\lambda_p}{\lambda_q}$, with $0 \leq t \leq 1$.

We have that $\mu = \frac{1+t^2}{(1+t)^2}$. It is straightforward to see that $\tilde{\mu} : t \mapsto \frac{1+t^2}{(1+t)^2}$ is monotone and decreasing for $0 \leq t \leq 1$. Thus, $\tilde{\mu}$ attains its maximum for $t = 0$. Thus, $(p, q) = (n, 1)$. \square

Note that if A is nonsingular $\lambda_1 = \|A\|_2$ and $\lambda_n = 0$, thus $\|\mathcal{T}_2\|_2 = \sqrt{2}$. Otherwise if A is nonsingular $\lambda_n = \|A^{-1}\|_2^{-1}$ and we obtain

$$\|\mathcal{T}_2\|_2 = \sqrt{2} \frac{\sqrt{\|A^{-1}\|_2^{-2} + \|A\|_2^2}}{\|A^{-1}\|_2^{-1} + \|A\|_2} = \sqrt{2} \frac{\sqrt{1 + \kappa_2(A)^2}}{1 + \kappa_2(A)}, \quad (70)$$

We consider (60)-(61), with H orthogonal, S symmetric and $S = V^T D V$ the eigen-decomposition of S . Let $Z_1 = V \widehat{\Delta} A V^T$ and $Z_2 = V^T \widehat{\Delta} A V$. Then, (60)-(61) become

$$D \widetilde{X} + \widetilde{X} D = Z_1 - Z_1^T \quad \text{and} \quad D Y + Y D = D Z_2 + Z_2^T D,$$

where $\widetilde{X} = V X V^T$ and $Y = V \widehat{S} V^T$. Since $\|Z_1\|_F = \|Z_2\|_F = \|\Delta A\|_F$, applying Lemma 5.5 and using (70), we obtain

$$\|dg_H(A)\|_2 = \frac{2}{\lambda_{n-1} + \lambda_n} \quad \text{and} \quad \|dg_S(A)\|_2 = \sqrt{2} \frac{\sqrt{1 + \kappa_2(A)^2}}{1 + \kappa_2(A)}. \quad (71)$$

Note that $1 \leq \|dg_S(A)\|_2 \leq \sqrt{2}$. Both of these bounds are attained. If S is of the type $S = \lambda I$ or S is orthogonal, then $\|dg_S(A)\|_2 = 1$ and if A is singular then $\|dg_S(A)\|_2 = \sqrt{2}$. We have the following theorem.

Theorem 5.6 *Let $A = HS$, $H \in \mathcal{O}_n(I)$ be the polar factorization of A and for $\Delta A \in \mathbb{R}^{n \times n}$ such that $\epsilon = \|\Delta A\|_F$ is small enough, let $(A + \Delta A) = \tilde{H}\tilde{S}$ be the polar factorization of $A + \Delta A$. Then,*

$$\begin{aligned}\|\tilde{H} - H\|_F &\leq \frac{2}{\lambda_{n-1} + \lambda_n} \epsilon + O(\epsilon^2), \\ \|\tilde{S} - S\|_F &\leq \alpha \epsilon + O(\epsilon^2),\end{aligned}$$

where $\alpha = \sqrt{2}$ if A is singular or $\alpha = \sqrt{2} \frac{\sqrt{1 + \kappa_2(A)^2}}{1 + \kappa_2(A)}$ otherwise.

The bounds given in the above theorem are the sharpest possible to first order. Using the classical definition of condition number for the Hermitian factor, the same condition number as in (71) is obtained in [5]. Our method has the advantage of giving a shorter proof than [5] of several pages. Our method allows us also to compute explicitly the Fréchet derivative of the factors. In [15], the condition number in (71) for the orthogonal factor is given.

5.3 Numerical Experiments

To compute the indefinite polar factorization and the usual polar factorization, we used the iteration described in [11, Thm 5.2]. We recall that the iteration for the J -orthogonal factor is given by

$$H_0 = A, \quad H_{n+1} = \frac{1}{2}(H_n + JH_n^{-T}J).$$

This iteration is guaranteed to converge if JA^TJA has no eigenvalue with a negative real part. We present two series of numerical tests. The first ones are quite standard, their purpose being to illustrate the perturbation bounds given in Theorem 5.4. We generated a matrix A_0 using the function `randn` of MATLAB. Then, we build a sequence of matrices A_ϵ that converges to A_0 as ϵ tends to zero. We denote $\delta_\epsilon = \|A_0 - A_\epsilon\|_F$ and $A_0 = H_0S_0$, $A_\epsilon = H_\epsilon S_\epsilon$ the indefinite polar factorization of A_0 and A_ϵ . J was obtained by

```
J = (-1) .* randperm(n)
```

using MATLAB. We shifted all these matrices so that $JA_\epsilon^TJA_\epsilon$ has all its eigenvalues in the open right half-plane. The results are displayed in Table 5. We see that our perturbation bounds follow closely the computed values which confirms that in this case the bounds obtained by Theorem 5.4 are sharp.

We denote by c_H and c_S the bounds of the condition number of the hyperbolic and symmetric factors given by (66)-(67). Table 6 shows the first order perturbation bounds obtained by using c_H and c_S . The bounds obtained by using c_S and c_H in the first 4 rows in Table in 6 are accurate. In the last row, we see that the bound for the J -symmetric matrix is weak whereas the bound for the hyperbolic factor is more reliable. We conclude that the bounds c_S and c_H given by (66)-(67) should be used carefully when the norm of the perturbation is small.

The aim of the second numerical experiment series is to give an example where the bounds given by (66)-(67) are very poor approximations of the exact condition

Table 5: Perturbation bounds of the indefinite polar factorization.

δ_ϵ	$\ S_\epsilon - S_0\ _F$	$\ dg_{S_0}(A_0)\ _2\delta_\epsilon$	$\ H_\epsilon - H_0\ _F$	$\ dg_{H_0}(A_0)\ _2\delta_\epsilon$
$1e - 15$	$1e - 15$	$1e - 14$	$1e - 15$	$2e - 15$
$1e - 9$	$1e - 9$	$2e - 9$	$2e - 8$	$5e - 8$
$1e - 5$	$3e - 5$	$9e - 5$	$2e - 5$	$6e - 5$
$1e - 3$	$7e - 3$	$1.6e - 2$	$5e - 3$	$2e - 2$
$1e - 2$	$1e - 2$	$2.3 - 2$	$2e - 2$	$3.4e - 2$

Table 6: Perturbation bounds of the IPF using bounds of the condition number c_H and c_S .

δ_ϵ	$c_S\delta_\epsilon$	$c_H\delta A_\epsilon$
$1e - 15$	$3.7e - 13$	$7.5e - 14$
$1e - 9$	$3.7e - 7$	$7.5e - 8$
$1e - 5$	$3.7e - 3$	$7.5e - 4$
$1e - 3$	$3.7e - 1$	$7.5e - 2$
$1e - 2$	3.7	$7.5e - 1$

numbers. The test matrices are Hilbert matrices, built in MATLAB and they can be called by the function `hilb`. The (i, j) element of a Hilbert matrix is given by $1/(i + j - 1)$. These Hilbert matrices are symmetric and very ill conditioned. The signature matrix $J \in \text{diag}_n^k(\pm 1)$ is given by

$$J = \text{diag}(-I_{\lfloor n/2 \rfloor}, I_{\lfloor n/2 \rfloor}).$$

The logarithm of the condition number $\log_{10}(\|dg_S(A)\|_2)$ for the J -symmetric factor is represented by \star and by $+$ for $\log_{10}(\|dg_H(A)\|_2)$, the logarithm of the condition number of the hyperbolic factor. The logarithm of the bound denoted by c_S in (66) is represented by \square and by \circ the bound c_H in (67). We see in Figure 2, in all the test matrices the exact condition number is very small compare to c_S , the biggest ratio being of order 10^{18} . For the hyperbolic factor, the difference is less, the biggest ratio being of order 10^4 .

6 Conclusion

In this paper, we have analyzed the HR factorization, the hyperbolic SVD and the indefinite polar factorization. For each factorization, we gave a computable condition number for each factor. The condition number being quite expensive to compute, we also gave bounds that can be less expensive to compute but that are also less accurate. We also analyzed, for each factorization, the orthogonal case, where all the factors involved are J -orthogonal, with $J = \pm I$.

Our method is based on the implicit function theorem and the definition of local coordinates on manifolds which makes it quite different to the usual approach of perturbation bounds. Although, the definition of these notions might be long and complicated, this method has the advantage to give explicitly the condition operator and it is easily adaptable from one factorization to another.

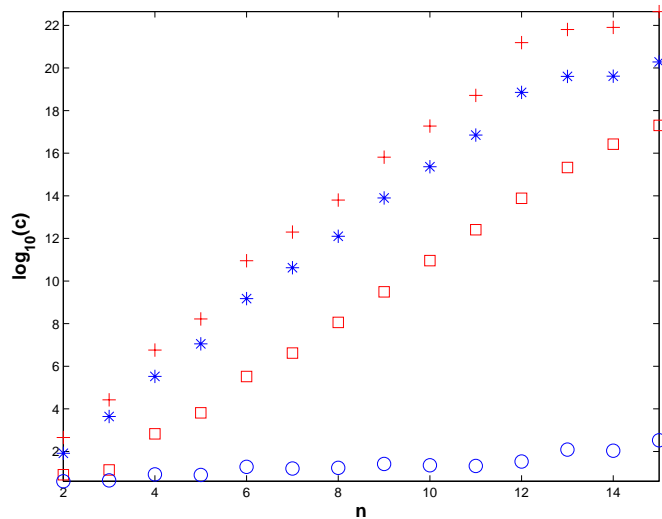


Figure 2: Condition number and perturbation bounds of the IPF of Hilbert matrices with $\log_{10}(\|d_{GS}(A)\|_2)$ (○), $\log_{10}(\|d_{GH}(A)\|_2)$ (□), $\log_{10}(c_S)$ (*) and $\log_{10}(c_H)$ (+).

References

- [1] Rajendra Bhatia. Matrix factorizations and their perturbations. *Linear Algebra Appl.*, 197/198:245–276, 1994.
- [2] Adam Bojanczyk, Nicholas J. Higham, and Harikrishna Patel. The equality constrained indefinite least squares problem: theory and algorithms. *BIT*, 43(3):505–517, 2003.
- [3] Adam W. Bojańczyk, Ruth Onn, and Allan O. Steinhardt. Existence of the hyperbolic singular value decomposition. *Linear Algebra Appl.*, 185:21–30, 1993.
- [4] A. Bunse-Gerstner. An analysis of the HR algorithm for computing the eigenvalues of a matrix. *Linear Algebra and Appl.*, 35:155–173, 1981.
- [5] F. Chaitin-Chatelin and S. Gratton. On the condition numbers associated with the polar factorization of a matrix. *Numer. Linear Algebra Appl.*, 7(5):337–354, 2000.
- [6] Xiao-Wen Chang, Christopher C. Paige, and G. W. Stewart. Perturbation analyses for the QR factorization. *SIAM J. Matrix Anal. Appl.*, 18(3):775–791, 1997.
- [7] Jean-Pierre Dedieu. Approximate solutions of numerical problems, condition number analysis and condition number theorem. In *The mathematics of numerical analysis (Park City, UT, 1995)*, volume 32 of *Lectures in Appl. Math.*, pages 263–283. Amer. Math. Soc., Providence, RI, 1996.
- [8] Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM J. Sci. Statist. Comput.*, 7(4):1160–1174, 1986.
- [9] Nicholas J. Higham. The matrix sign decomposition and its relation to the polar decomposition. In *Proceedings of the 3rd ILAS Conference (Pensacola, FL, 1993)*, volume 212/213, pages 3–20, 1994.

- [10] Nicholas J. Higham. A survey of componentwise perturbation theory in numerical linear algebra. In *Mathematics of Computation 1943–1993: a half-century of computational mathematics (Vancouver, BC, 1993)*, volume 48 of *Proc. Sympos. Appl. Math.*, pages 49–77. Amer. Math. Soc., Providence, RI, 1994.
- [11] Nicholas J. Higham. J -orthogonal matrices: properties and generation. *SIAM Rev.*, 45(3):504–519, 2003.
- [12] David W. Kammler. A perturbation analysis of the intrinsic conditioning of an approximate null vector computed with a SVD. *J. Comput. Appl. Math.*, 9(3):201–204, 1983.
- [13] Charles Kenney and Alan J. Laub. Polar decomposition and matrix sign function condition estimates. *SIAM J. Sci. Statist. Comput.*, 12(3):488–504, 1991.
- [14] A. Largillier. Bounds for relative errors of complex matrix factorizations. *Appl. Math. Lett.*, 9(6):79–84, 1996.
- [15] Roy Mathias. Perturbation bounds for the polar decomposition. *SIAM J. Matrix Anal. Appl.*, 14(2):588–597, 1993.
- [16] Ruth Onn and Adam Steinhardt, Allan O. and Bojanczyk. The hyperbolic singular value decomposition and applications. *Applied mathematics and computing, Trans. 8th Army Conf., Ithaca/NY (USA) 1990, ARO Rep. 91-1, 93-108*, 1991.
- [17] John R. Rice. A theory of condition. *SIAM J. Numer. Anal.*, 3(2):287–310, 1966.
- [18] G. W. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Rev.*, 15:727–764, 1973.
- [19] G. W. Stewart. Perturbation bounds for the QR factorization of a matrix. *SIAM J. Numer. Anal.*, 14(3):509–518, 1977.
- [20] G. W. Stewart. A note on the perturbation of singular values. *Linear Algebra Appl.*, 28:213–216, 1979.
- [21] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12(1):99–111, 1972.