Certified Singular Value Decomposition

Joris van der Hoeven, Jean-Claude Yakoubsohn

Certified Singular Value Decomposition

JORIS VAN DER HOEVEN
CNRS, Laboratoire LIX
Campus de l'École Polytechnique
1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing CS35003
91120 Palaiseau
France
Email: vdhoeven@lix.polytechnique.fr

JEAN-CLAUDE YAKOUBSOHN
Institut de Mathématiques de Toulouse
Université Paul Sabatier
118 route de Narbonne
31062 Toulouse Cedex 9
France
Email: yak@mip.ups-tlse.fr

December 2, 2018

In this paper, we present an efficient algorithm for the certification of numeric singular value decompositions (SVDs) in the regular case, i.e., in the case when all the singular values are pairwise distinct. Our algorithm is based on a Newton-like iteration that can also be used for doubling the precision of an approximate numerical solution.

1. INTRODUCTION

Let $\mathbb{F}$ be the set of floating point numbers for a fixed precision and a fixed exponent range. We denote $\mathbb{F}^{\geqslant} = \{x \in \mathbb{F} : x \geqslant 0\}$. Consider an $m \times n$ matrix $M \in \mathbb{F}[\mathrm{i}]^{m \times n}$ with complex floating entries, where $m \geqslant n$. The problem of computing the numeric *singular value decomposition* of M is to compute unitary transformation matrices $U \in \mathbb{F}[\mathrm{i}]^{m \times m}$, $V \in \mathbb{F}[\mathrm{i}]^{n \times n}$, and a diagonal matrix $\Sigma \in (\mathbb{F}^{\geqslant})^{m \times n}$ such that

$$M \approx U \Sigma V^{*}. \tag{1}$$

If $m > n$, then $\Sigma \in \mathbb{F}[\mathrm{i}]^{m \times n}$ is understood to be "diagonal" if it is of the form

$$\Sigma = \begin{pmatrix} \mathrm{Diag}(\sigma_1, \ldots, \sigma_n) \\ 0 \end{pmatrix}, \qquad \mathrm{Diag}(\sigma_1, \ldots, \sigma_n) = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}.$$

The diagonal entries $\sigma_1, \ldots, \sigma_n$ of Σ are the approximate singular values of the matrix M and throughout this paper we will assume them to be pairwise distinct and ordered

$$\sigma_1 > \sigma_2 > \cdots > \sigma_n > 0.$$

There are several well-known algorithms for the computation of numeric singular value decompositions [8, 4].

Now (1) is only an approximate equality. It is sometimes important to have a rigorous bound for the distance between an approximate solution and some exact solution. More precisely, we may ask for a diagonal matrix $\Sigma_r \in (\mathbb{F}^{\geqslant})^{m \times n}$ and matrices $U_r \in \mathbb{F}[\mathrm{i}]^{m \times m}$, $V_r \in \mathbb{F}[\mathrm{i}]^{n \times n}$, such that there exist unitary matrices $\tilde{U} \in \mathbb{C}^{m \times m}$, $\tilde{V} \in \mathbb{C}^{n \times n}$, and a diagonal matrix $\tilde{\Sigma} \in \mathbb{C}^{m \times n}$ for which

$$M = \tilde{U} \tilde{\Sigma} \tilde{V}^{*}$$

and

$$\begin{aligned} |\tilde{\Sigma}_{i,i} - \Sigma_{i,i}| &\leqslant (\Sigma_r)_{i,i} \\ |\tilde{U}_{i,j} - U_{i,j}| &\leqslant (U_r)_{i,j} \\ |\tilde{V}_{i,j} - V_{i,j}| &\leqslant (V_r)_{i,j} \end{aligned}$$

for all i, j. This task will be called the *certification problem* for the given numeric singular value decomposition (1). The matrices Σ_r, U_r and V_r can be thought of as reliable error bounds for the matrices Σ, U and V of the numerical solution.

It will be convenient to rely on *ball arithmetic* [13, 19], which is a systematic technique for this kind of bound computations. When computing with complex numbers, ball arithmetic is more accurate than more classical interval arithmetic [22, 1, 23, 18, 21, 24], especially in multiple precision contexts. We will write $\mathbb{B} = \mathcal{B}(\mathbb{F}[\mathrm{i}], \mathbb{F}^{\geqslant})$ for the set of balls $\boldsymbol{z} = \mathcal{B}(z_c, z_r) = \{z \in \mathbb{C} : |z - z_c| \leqslant z_r\}$ with centers z_c in $\mathbb{F}[\mathrm{i}]$ and radii z_r in $\mathbb{F}^{\geqslant}$. In a similar way, we may consider matricial balls $\boldsymbol{M} = \mathcal{B}(M_c, M_r) \in \mathcal{B}(\mathbb{F}[\mathrm{i}]^{m \times n}, (\mathbb{F}^{\geqslant})^{m \times n})$: given a center matrix $M_c \in \mathbb{F}[\mathrm{i}]^{m \times n}$ and a radius matrix $M_r \in (\mathbb{F}^{\geqslant})^{m \times n}$, we have

$$\boldsymbol{M} \;=\; \mathcal{B}(M_c, M_r) \;=\; \{M \in \mathbb{C}^{m \times n} : \forall i, j, |(M_c)_{i,j} - M_{i,j}| \leqslant (M_r)_{i,j}\}.$$

Alternatively, we may regard $\mathcal{B}(M_c, M_r)$ as the set of matrices in $\mathbb{B}^{m \times n}$ with ball coefficients:

$$\mathcal{B}(M_c, M_r)_{i,j} \;=\; \mathcal{B}((M_c)_{i,j}, (M_r)_{i,j}).$$

Standard arithmetic operations on balls are carried out in a reliable way. For instance, if $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{B}$, then the computation of the product $\boldsymbol{w} = \boldsymbol{u}\,\boldsymbol{v}$ using ball arithmetic has the property that $u\,v \in \boldsymbol{w}$ for any $u \in \boldsymbol{u}$ and $v \in \boldsymbol{v}$.

In the language of ball arithmetic, it is natural to allow for small errors in the input and replace the numeric input $M \in \mathbb{F}[\mathrm{i}]^{m \times n}$ by a ball input $\mathcal{B}(M_c, M_r) \in \mathbb{B}^{m \times n}$. Then we may still compute a numeric singular value decomposition of the center matrix M_c:

$$D_c \;\approx\; U_c M_c V_c^*. \tag{2}$$

The generalized *certification problem* now consists of the computation of matrices $U_r \in (\mathbb{F}^{\geqslant})^{m \times m}$, $V_r \in (\mathbb{F}^{\geqslant})^{n \times n}$, and a diagonal matrix $\Sigma_r \in (\mathbb{F}^{\geqslant})^{m \times n}$ such that, for every $M \in \mathcal{B}(M_c, M_r)$, there exist unitary matrices $U, V \in \mathcal{B}(\Sigma_c, \Sigma_r)$, and a diagonal matrix $\Sigma \in \mathcal{B}(\Sigma_c, \Sigma_r)$ with

$$\Sigma \;=\; U M V^*.$$

In this paper we propose an efficient solution for this problem in the case when all singular values are simple. Our algorithm relies on an efficient Newton iteration that is also useful for doubling the precision of a given numeric singular value decomposition. The iteration admits a quadratic convergence and only requires matrix sums and products of size at most $m \times m$. In [13, 15], a similar approach was used for the certification of eigenvalues and eigenvectors.

We are not aware of similarly efficient and robust algorithms in the literature. Jacobi-like methods from Kogbeliantz' SVD algorithm admit quadratic convergence in the presence of cluster in the Hermitian case [3], but only linear convergence is achieved in general [5]. Gauss-Newton type methods have also been proposed for the approximation the regular real SVD in [17, 16].

From the theoretical bit complexity point of view, our algorithm essentially reduces the certification problem to a constant number of numeric matrix multiplications. When using a precision of p bits for numerical computations, it has recently been shown [10] that two $n \times n$ matrices can be multiplied in time

$$\mathsf{MM}(n,p) = O(n^2 \mathsf{I}(p) + n^\omega p\, 2^{O(\lg^* p - \lg^* n)}\, \mathsf{I}(\lg d)/\lg d).$$

Here $\mathsf{I}(p) = O(p \lg p\, K^{\lg^* p})$ with $K \leqslant 4$ is the cost of p-bit integer multiplication [11, 9] and $\omega < 2.3728639$ is the exponent of matrix multiplication [7]. If p is large enough with respect to the log of the condition number, then $O(\mathsf{MM}(n,p))$ yields an asymptotic bound for the bit complexity of our certification problem.

We have implemented unoptimized versions of the new algorithms in MATHEMAGIX [14] and MATLAB. These toy implementations indeed confirmed the quadratic convergence of our Newton iteration and the efficiency of the new algorithms. We intend to report more extensively on implementation issues in a forthcoming paper.

2. NOTATIONS

2.1. Matrix norms

Throughout this paper, we will use the max-norm for vectors and the corresponding matrix norm. More precisely, given positive integers m, n, a vector $v \in \mathbb{C}^n$, and an $m \times n$ matrix $M \in \mathbb{C}^{m \times n}$, we set

$$\begin{aligned} \|v\| &= \max\{|v_1|, \ldots, |v_n|\} \\ \|M\| &= \max_{\|v\|=1} \|Mv\|. \end{aligned}$$

For a second matrix $N \in \mathbb{C}^{m \times n}$, we clearly have

$$\begin{aligned} \|M+N\| &\leqslant \|M\| + \|N\| \\ \|MN\| &\leqslant \|M\|\,\|N\|. \end{aligned}$$

We also define

$$\|M\|_* = \max(\|M\|, \|M^*\|). \tag{3}$$

Explicit machine computation of the matrix norm is easy using the formula

$$\|M\| = \max\{|M_{i,1}| + \cdots + |M_{i,n}| : 1 \leqslant i \leqslant m\}. \tag{4}$$

Given a second matrix $N \in \mathbb{C}^{m \times n}$ it follows that the coefficientwise product

$$M \odot N = \begin{pmatrix} M_{1,1}N_{1,1} & \cdots & M_{1,n}N_{1,n} \\ \vdots & & \vdots \\ M_{m,1}N_{m,1} & \cdots & M_{m,n}N_{m,n} \end{pmatrix}$$

satisfies

$$\|M \odot N\| \leqslant \|M\| \max\{|N_{i,j}| : 1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n\}. \tag{5}$$

In particular, when changing certain entries of a matrix M to zero, its matrix norm $\|M\|$ can only decrease. We will write $\mathcal{B}(0,1)_{m \times n}$ for the $m \times n$ ball matrix whose entries are all unit balls $\mathcal{B}(0,1)$. This matrix has the property that $M \in \|M\|\, \mathcal{B}(0,1)_{m \times n}$ for all $m \times n$ matrices M.

2.2. Miscellaneous notations

In the sequel we consider two integers $m \geqslant n$ and we introduce the sets of matrices:

$$\mathbb{D}^{m\times n} := \left\{ \begin{pmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \\ 0 \end{pmatrix} \in \mathbb{R}_+^{m\times n} : \sigma_1 > \cdots > \sigma_p > 0 \right\} \tag{6}$$

and

$$\mathbb{E}^{m\times n} = \mathbb{C}^{m\times n} \times \mathbb{D}^{m\times n} \times \mathbb{C}^{m\times m} \times \mathbb{C}^{n\times n}.$$

We also write $\operatorname{diag}: \mathbb{C}^{m\times n} \to \mathbb{D}^{m\times n}$ for the natural projection that replaces all non-diagonal entries by zeros. For any integer k we finally define the map $E_k: \mathbb{C}^{k\times k} \to \mathbb{C}^{k\times k}$ by

$$E_k(U) = U^* U - I_k,$$

where I_k is the identity matrix of size $k \times k$.

3. OVERVIEW OF OUR METHOD

Given $M \in \mathbb{C}^{m\times n}$, the triple (Σ, U, V) with $(M, \Sigma, U, V) \in \mathbb{E}^{m\times n}$ forms an SVD for M if and only if it satisfies the following system of equations:

$$F(M, \Sigma, U, V) = \begin{pmatrix} E_m(U) \\ E_n(V) \\ \Sigma - U^* M V \end{pmatrix} = 0. \tag{7}$$

This is a system of $m^2 + n^2 + m\,n$ equations with $m^2 + n^2 + n$ unknowns. Our efficient numerical method for solving this system will rely on the following principles:

1. For a well-chosen *ansatz* U_0 close to the unitary group $\mathbf{U}_m$, we prove that
$$X_0 = U_0 (I_m - E(U_0)/2)$$
is even closer to the unitary group than U_0: see section 4. Similarly, for an *ansatz* V_0 close to $\mathbf{U}_n$, we take $Y_0 = V_0 (I_m - E(V_0)/2)$
2. From Σ_0, X_0 and Y_0, we prove that is possible to explicitly compute $\dot{\Sigma}_0$, and two skew Hermitian matrices $\dot{X}_0$ and $\dot{Y}_0$ such that
$$X_0^* M Y_0 - \Sigma_0 = \dot{X}_0 \Sigma_0 - \Sigma_0 \dot{Y}_0 + \dot{\Sigma}_0,$$
after which $(I_m + \dot{X}_0)(\Sigma_0 + \dot{\Sigma}_0)(I_n - \dot{Y}_0)$ is a first-order approximation of $\Sigma_0 - X_0^* M Y_0$: see section 5.
3. Let $\Sigma_1 := \Sigma_0 + \dot{\Sigma}_0$, $U_1 := X_1(1 + \dot{X}_0)$, and $V_1 := Y_1(1 + \dot{Y}_0)$. If $U_0 \Sigma_0 V_0^*$ is sufficiently close to M, then we will prove that $U_1 \Sigma_1 V_1^*$ is a better approximation of the matrix M than $U_0 \Sigma_0 V_0^*$:
$$\|U_1 \Sigma_1 V_1^* - M\| = O(\|U_0 \Sigma_0 V_0^* - M\|^2).$$

More precisely, given $\Sigma_0 \in \mathbb{D}^{m\times n}$, $U_0 \in \mathbb{C}^{m\times m}$, and $V_0 \in \mathbb{C}^{n\times n}$, we define the following sequence of matrices $(\Sigma_i, U_i, V_i)_{i\geqslant 0}$

$$X_i = U_i \left(I_m - \frac{E_m(U_i)}{2} \right) \tag{8}$$

$$Y_i = V_i \left(I_n - \frac{E_n(V_i)}{2} \right) \tag{9}$$

$$\Sigma_{i+1} = \Sigma_i + \dot{\Sigma}_i \tag{10}$$

$$U_{i+1} = X_i (I_m + \dot{X}_i) \tag{11}$$

$$V_{i+1} = Y_i (I_n + \dot{Y}_i), \tag{12}$$

where $\dot{\Sigma}_i$ is a diagonal matrix and $\dot{X}_i$, $\dot{Y}_i$ are two skew Hermitian matrices such that

$$X_i^* M Y_i - \Sigma_i = \dot{X}_i \Sigma_i - \Sigma_i \dot{Y}_i + \dot{\Sigma}_i. \tag{13}$$

In order to measure the quality of the *ansatz*, we define

$$\|F(M,\Sigma,U,V)\| = \max(\|\Sigma - U^* M V\|, \|E_m(U)\|, \|E_n(V)\|).$$

The main result of the paper is the following theorem that gives explicit conditions for the quadratic convergence of the sequence $(\Sigma_i, U_i, V_i)_{i\geqslant 0}$, together with explicit error bounds.

THEOREM 1. *Let $\varepsilon \geqslant 0$ and $(M_0, \Sigma_0, U_0, V_0) \in \mathbb{E}^{m\times n}$ be such that $\|F(M_0,\Sigma_0,U_0,V_0)\| \leqslant \varepsilon$. Denote*

$$\begin{aligned} \kappa &:= \kappa_0 = \max\left(1, \frac{1}{\sigma_{0,n}}, \max_{i<j} \frac{1}{|\sigma_{0,i} - \sigma_{0,j}|}\right) \\ K &:= K_0 = \max\left(1, \max_i \sigma_{0,i}\right), \end{aligned}$$

where $\sigma_{0,1}, \ldots, \sigma_{0,n}$ stand for the diagonal entries of Σ_0. If

$$K^3 \kappa^2 \varepsilon \leqslant 0.005,$$

then the sequence $(\Sigma_i, U_i, V_i)_{i\geqslant 0}$ defined by (10–12) converges quadratically towards an SVD (Σ, U, V) of the matrix M, i.e. $M = U\Sigma V^*$. More precisely, for each $i \geqslant 0$, we have

$$\begin{aligned} \|U_i - U\| &\leqslant 13.5\sqrt{m}\,\kappa K \varepsilon\, 2^{1-2^i} \\ \|V_i - V\| &\leqslant 13.5\sqrt{n}\,\kappa K \varepsilon\, 2^{1-2^i} \\ \|\Sigma_i - \Sigma\| &\leqslant 0.82\,\varepsilon 2^{1-2^i}. \end{aligned}$$

The proof of this theorem will be postponed to section 6. Assuming that the theorem holds, it naturally gives rise to the following algorithm for certifying an approximate SVD:

Algorithm 1
Input: an approximate SVD (Σ_0, U_0, V_0) for the center of a ball matrix $\boldsymbol{M} \in \mathbb{B}^{m\times n}$
Output: ball enclosures $\boldsymbol{\Sigma} \in \mathbb{B}^{m\times n}$, $\boldsymbol{U} \in \mathbb{B}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{B}^{n\times n}$ of Σ_0, U_0 and V_0 such that for any $M \in \boldsymbol{M}$, there exist $\Sigma \in \boldsymbol{\Sigma}$, $U \in \boldsymbol{U}$ and $V \in \boldsymbol{V}$ such that $M = U\Sigma V^*$ is an exact singular value decomposition of M

1. Compute $\boldsymbol{F} := F(\boldsymbol{M}, \Sigma_0, U_0, V_0)$ using ball arithmetic
2. Let $\bar{\varepsilon}$ be an upper bound for $\|\boldsymbol{F}\|$
3. Let $\bar{\kappa}$ and $\bar{K}$ be upper bounds for κ and K (with κ and K as in Theorem 1)
4. If $\bar{K}\bar{\kappa}^2\bar{\varepsilon} \geqslant 0.005$, then set $\varrho_\Sigma := \infty$, $\varrho_U := \infty$, $\varrho_V := \infty$
5. Else set $\varrho_U := 13.5\sqrt{m}\,\bar{\kappa}\bar{K}\bar{\varepsilon}$, $\varrho_V := 13.5\sqrt{n}\,\bar{\kappa}\bar{K}\bar{\varepsilon}$, $\varrho_\Sigma := 0.82\,\bar{\varepsilon}$ (using upward rounding)
6. Set $\boldsymbol{\Sigma} := \Sigma_0 + \varrho_\Sigma \operatorname{diag}(\mathcal{B}(0,1)_{m\times n})$
7. Set $\boldsymbol{U} := U_0 + \varrho_U \mathcal{B}(0,1)_{m\times m}$
8. Set $\boldsymbol{V} := V_0 + \varrho_V \mathcal{B}(0,1)_{n\times n}$
9. Return $(\boldsymbol{\Sigma}, \boldsymbol{U}, \boldsymbol{V})$

THEOREM 2. *Algorithm 1 is correct.*

Proof. If $\bar{K}\bar{\kappa}^2\bar{\varepsilon} \geqslant 0.005$, then we return matrix balls with infinite radii for which the result is trivially correct. If $\bar{K}\bar{\kappa}^2\bar{\varepsilon} \leqslant 0.005$, then for any $M_0 \in \boldsymbol{M}$, the actual values of ε, κ and K are bounded by $\bar{\varepsilon}$, $\bar{\kappa}$ and $\bar{K}$, so Theorem 1 applies for the *ansatz* $(M_0, \Sigma_0, U_0, V_0) \in \mathbb{E}^{m\times n}$. As a consequence, we obtain an SVD (Σ, U, V) for M_0 with the property that $\|U - U_0\| \leqslant 13.5\sqrt{m}\,\kappa K\varepsilon \leqslant \varrho_U$, $\|V - V_0\| \leqslant 13.5\sqrt{n}\,\kappa K\varepsilon \leqslant \rho_V$, and $\|\Sigma - \Sigma_0\| \leqslant 0.82\,\varepsilon \leqslant \rho_\Sigma$. We conclude that $U \in \boldsymbol{U}$, $V \in \boldsymbol{V}$, $\Sigma \in \boldsymbol{\Sigma}$, as desired. □

Remark 3. Notice that the algorithm does not use our Newton iteration in order to improve the quality of the approximate input SVD (in particular, the output is worthless whenever $\bar{K}\,\bar{\kappa}^2\,\bar{\varepsilon} \geqslant 0.005$). The idea is that Algorithm 1 is only used for the certification, and not for numerical approximation. The user is free to use any preferred algorithm for computing the initial approximate SVD. Of course, our Newton iteration can be of great use to increase the precision of a rough approximate SVD that was computed by other means.

4. POLAR PROJECTION

Since we are doing approximate computations, the unitary matrices in an SVD are not given exactly, so we may wish to estimate the distance between an approximate unitary matrix and the closest actual unitary matrix. This is related to the following problem: given an approximately unitary $n \times n$ matrix U, find a good approximation $U + \dot{U}$ for its projection on the group $\mathbf{U}(m)$ of unitary $m \times m$ matrices. We recall a Newton iteration for this problem [20, 2, 12] and provide a detailed analysis of its (quadratic) convergence.

4.1. The Newton iteration

The tangent space to $\mathbf{U}(m)$ at U is

$$T_U\mathbf{U}(m) \;=\; \{UX : X^* = -X\}. \tag{14}$$

Consider the Riemannian metric inherited from the embedding space $\mathbb{C}^{m\times m}$

$$\langle X, Y\rangle_U \;:=\; \mathrm{Tr}(X^*Y).$$

Then the normal space is

$$T_U^{\perp}\mathbf{U}(m) \;=\; \{U\Delta : \Delta^* = \Delta\}.$$

We wish to compute $\dot{U}$ using an appropriate Newton iteration. From the characterization of the normal space, it turns out that it is more convenient to write $U + \dot{U} = U(1+\Delta)$, where Δ is Hermitian. With $E_m(U) = U^*U - I_m$ and $\dot{U} = U\Delta$, we have

$$\begin{aligned} E_m(U+\dot{U}) \;&=\; (I_m+\Delta^*)\,(I_m+E_m(U))\,(I_m+\Delta) - I_m \\ &=\; E_m(U) + 2\Delta + \Delta E_m(U) + E_m(U)\Delta + \Delta^2 + \Delta E_m(U)\Delta. \end{aligned}$$

Taking

$$\Delta \;=\; -\frac{E_m(U)}{2}, \tag{15}$$

it follows that

$$E_m(U+\dot{U}) \;=\; \left(-\frac{3}{4}I_m + \frac{1}{4}E_m(U)\right)E_m(U)^2. \tag{16}$$

We are thus lead to the following Newton iteration that we will further study below:

$$U_{i+1} \;=\; U_i\left(I_m - \frac{E_m(U_i)}{2}\right), \qquad i \geqslant 0. \tag{17}$$

Remark 4. Another way to construct the previous iteration is to remark that the derivative $DE_m(U)$ is onto from $\mathbb{C}^{m\times m}$ on the subset $\mathbb{H}^{m\times m} \subseteq \mathbb{C}^{m\times m}$ of Hermitian matrices. Then it is easy to see that for given $H \in \mathbb{H}^{m\times m}$ and $U \in \mathbf{U}(m)$, the matrix $\frac{1}{2}UH$ satisfies the equation $DE_m(U)\,X = H$, i.e,

$$X^*U + U^*X = H.$$

Consequently $DE_m(U)^{-1} E_m(U) = \frac{1}{2} U E_m(U)$. In this context the classical Newton operator thus becomes

$$N_{E_m}(U) = U - \frac{1}{2} U E_m(U).$$

4.2. Error analysis

PROPOSITION 5. *Let U be an $m \times m$ matrix with $\varepsilon := \|E_m(U)\| < 1$. Let $U_1 = U(1+\Delta)$, where $\Delta = -E_m(U)/2$ and write $\varepsilon_1 := \|E_m(U_1)\|$. Then $\|\Delta\| \leqslant \frac{\varepsilon}{2}$ and*

$$\varepsilon_1 \leqslant \varepsilon^2. \tag{18}$$

Proof. The conclusion follows from (16), since $\|E(U_1)\| \leqslant \frac{3}{4}\varepsilon^2 + \frac{1}{4}\varepsilon^3 \leqslant \varepsilon^2$. □

LEMMA 6. *Given $\varepsilon \leqslant 1/2$, $u \leqslant 1$, and $i \geqslant 0$, we have*

$$\prod_{j \geqslant 0} \left(1 + \frac{u}{2} \varepsilon^{2^{j+i}}\right) \leqslant 1 + 0.91\, u\, \varepsilon^{2^i}. \tag{19}$$

Proof. Modulo taking ε^{2^i} instead of ε, it suffices to consider the case when $i = 0$. Now

$$\varphi(\varepsilon, u) := \frac{\prod_{j \geqslant 0} \left(1 + \frac{u}{2} \varepsilon^{2^{j+i}}\right) - 1}{\varepsilon\, u}$$

is an increasing function in ε and u, since its power series expansion in ε and u admits only positive coefficients. Consequently, $\varphi(\varepsilon, u) \leqslant \varphi(1/2, 1) \approx 0.90607762222 < 0.91$. □

We recall that any invertible matrix $U \in \mathbb{C}^{m \times m}$ admits a unique polar decomposition

$$U = \pi(U)\, P,$$

where $\pi(U) \in \mathbf{U}(m)$ and $P \in \mathbb{C}^{m \times m}$ is a positive-definite Hermitian matrix. We call $\pi(U)$ the *polar projection* of U on $\mathbf{U}(m)$. The matrix P can uniquely be written as the exponential of another Hermitian matrix. It is also well known that $\pi(U)$ is indeed the closest element in $\mathbf{U}(m)$ to U for the Riemannian metric [6, Theorem 1].

THEOREM 7. *Let U be such that $\|E(U)\| \leqslant \varepsilon \leqslant 1/2$. Then the Newton sequence (17) defined from $U_0 = U$ converges quadratically to the polar projection $\pi(U) \in \mathbf{U}(m)$ of U. More precisely, for all $i \geqslant 0$, we have*

$$\|U_i - \pi(U)\|_* \leqslant 1.67 \sqrt{m}\, \varepsilon\, 2^{1-2^i}.$$

Proof. The Newton sequence (17) defined from $U_0 = U$ gives

$$U_{i+1} = U_0 (I_m + \Delta_0) \cdots (I_m + \Delta_i)$$

with $\Delta_i = -E_m(U_i)/2$. An obvious induction using Proposition 5 yields $\|\Delta_i\| \leqslant 1/2\, \varepsilon^{2^i}$ and $\|E_m(U_i)\| \leqslant \varepsilon^{2^i}$. Therefore this sequence converges to a limit $U_\infty \in \mathbf{U}(m)$ that is given by

$$U_\infty = U_0 Z_0, \qquad Z_0 = \prod_{j \geqslant 0} (I_m + \Delta_j).$$

Lemma 6 implies

$$\|Z_0 - I_m\| \leqslant \prod_{j \geqslant 0} \left(1 + \frac{\varepsilon^{2^j}}{2}\right) - 1 \leqslant 0.91\, \varepsilon. \tag{20}$$

More generally, we have

$$U_\infty \;=\; U_i Z_i, \qquad Z_i \;=\; \prod_{j \geqslant i} (I_m + \Delta_j), \qquad \|Z_i - I_m\| \;\leqslant\; 0.91\, \varepsilon^{2^i}.$$

Since U_∞ is unitary, we have $\|U_\infty\| \leqslant \sqrt{m}$. Neumann's lemma also implies that Z_i is invertible with

$$\|U_i - U_\infty\| \;=\; \|U_\infty\,(Z_i^{-1} - I_m)\| \;\leqslant\; \sqrt{m}\,\frac{0.91\,\varepsilon^{2^i}}{1 - 0.91\,\varepsilon^{2^i}} \;\leqslant\; 1.67\,\sqrt{m}\,\varepsilon^{2^i} \;\leqslant\; 1.67\,\sqrt{m}\,\varepsilon\, 2^{1-2^i}.$$

By induction on i, it can also be checked that $\Delta_i \in \mathbb{Q}[U_0^*\, U_0]$ for all i. This means that the Δ_i all commute, whence Z_0 and Z_0^{-1} are actually Hermitian matrices. Since $\|Z_0^{-1} - I_m\| \leqslant 0.91\,\varepsilon\,/\,(1 - 0.91\,\varepsilon) < 1$, the logarithm $\log Z_0^{-1}$ is well defined. We conclude that Z_0^{-1} is the exponential of a Hermitian matrix, whence it is positive-definite. □

5. SVDs FOR PERTURBED DIAGONAL MATRICES

5.1. Approximate solutions at order one

Let $\Sigma \in \mathbb{D}^{m \times n}$ be a matrix with diagonal entries $\sigma_1 > \cdots > \sigma_n$. Consider a perturbation

$$\Delta \;=\; \Sigma + \dot{\Delta}.$$

We wish to compute an approximate SVD

$$\Sigma + \dot{\Delta} \;\approx\; (I_m + \dot{X})\,(\Sigma + \dot{\Sigma})\,(I_n + \dot{Y})^*,$$

where $\dot{\Sigma} \in \mathbb{D}^{m \times n}$, $\dot{X} \in \mathbb{C}^{m \times m}$, and $\dot{Y} \in \mathbb{C}^{n \times n}$. Discarding higher order terms, this leads to the linear equation

$$\dot{\Delta} \;=\; \dot{X}\,\Sigma - \Sigma\,\dot{Y} + \dot{\Sigma},$$

with $\dot{X} \in T_{I_m}(\mathbf{U}(m))$ and $\dot{Y} \in T_{I_n}(\mathbf{U}(n))$. In view of (14), this means that $\dot{X}$ and $\dot{Y}$ are skew Hermitian. The following proposition shows how to solve the linear equation explicitly under these constraints.

PROPOSITION 8. *Let $\Sigma \in \mathbb{D}^{m \times n}$ and $\dot{\Delta} = (\delta_{i,j}) \in \mathbb{C}^{m \times n}$. Consider the diagonal matrix $\dot{\Sigma} \in \mathbb{R}^{m \times n}$ and the two skew Hermitian matrices $\dot{X} = (x_{i,j}) \in \mathbb{C}^{m \times m}$ and $\dot{Y} = (y_{i,j}) \in \mathbb{C}^{n \times n}$ that are defined by the following formulas:*

- *For $1 \leqslant i \leqslant n$, we take*

$$\dot{\Sigma}_{i,i} \;=\; \operatorname{Re} \delta_{i,i} \tag{21}$$

$$x_{i,i} \;=\; -y_{i,i} \;=\; \frac{\operatorname{Im} \delta_{i,i}}{2\,\sigma_i}\,\mathrm{i}. \tag{22}$$

- *For $1 \leqslant i < j \leqslant n$, we take*

$$\operatorname{Re} x_{i,j} \;=\; \frac{1}{2}\left(\frac{\operatorname{Re} \delta_{i,j} + \operatorname{Re} \delta_{j,i}}{\sigma_j - \sigma_i} + \frac{\operatorname{Re} \delta_{i,j} - \operatorname{Re} \delta_{j,i}}{\sigma_j + \sigma_i}\right) \tag{23}$$

$$\operatorname{Re} y_{i,j} \;=\; \frac{1}{2}\left(\frac{\operatorname{Re} \delta_{i,j} + \operatorname{Re} \delta_{j,i}}{\sigma_j - \sigma_i} - \frac{\operatorname{Re} \delta_{i,j} - \operatorname{Re} \delta_{j,i}}{\sigma_j + \sigma_i}\right) \tag{24}$$

$$\operatorname{Im} x_{i,j} \;=\; \frac{1}{2}\left(\frac{\operatorname{Im} \delta_{i,j} - \operatorname{Im} \delta_{j,i}}{\sigma_j - \sigma_i} + \frac{\operatorname{Im} \delta_{i,j} + \operatorname{Im} \delta_{j,i}}{\sigma_j + \sigma_i}\right) \tag{25}$$

$$\operatorname{Im} y_{i,j} \;=\; \frac{1}{2}\left(\frac{\operatorname{Im} \delta_{i,j} - \operatorname{Im} \delta_{j,i}}{\sigma_j - \sigma_i} - \frac{\operatorname{Im} \delta_{i,j} + \operatorname{Im} \delta_{j,i}}{\sigma_j + \sigma_i}\right). \tag{26}$$

- *For $n+1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$, we take*

$$x_{i,j} = \frac{1}{\sigma_j} \delta_{i,j}. \tag{27}$$

- *For $n+1 \leqslant i \leqslant m$ and $n+1 \leqslant j \leqslant m$, we take*

$$x_{i,j} = 0. \tag{28}$$

Then we have

$$\dot{\Delta} = \dot{X}\Sigma - \Sigma\dot{Y} + \dot{\Sigma}. \tag{29}$$

Proof. Since $\dot{X}$ and $\dot{Y}$ are skew Hermitian, we have $\operatorname{diag}(\operatorname{Re}(\dot{X}\Sigma - \Sigma\dot{Y})) = 0$. In view of (21), we thus get

$$\operatorname{diag}(\operatorname{Re}\dot{\Delta}) = \operatorname{diag}\operatorname{Re}(\dot{X}\Sigma - \Sigma\dot{Y} + \dot{\Sigma}) = \dot{\Sigma}.$$

By skew symmetry, for the equation

$$\dot{X}\Sigma - \Sigma\dot{Y} = \dot{\Delta} - \operatorname{diag}(\operatorname{Re}\dot{\Delta}) = \dot{\Delta} - \dot{\Sigma}$$

to hold, it is sufficient to have

$$\sigma_i x_{i,i} - \sigma_i y_{i,i} = \mathrm{i}\operatorname{Im}\delta_{i,i} \qquad 1 \leqslant i \leqslant n \tag{30}$$

$$\begin{pmatrix} \sigma_i x_{i,i} & \sigma_j x_{i,j} \\ -\sigma_i \overline{x_{i,j}} & \sigma_j x_{j,j} \end{pmatrix} - \begin{pmatrix} \sigma_i y_{i,i} & \sigma_i y_{i,j} \\ -\sigma_j \overline{y_{i,j}} & \sigma_j y_{j,j} \end{pmatrix} = \begin{pmatrix} \mathrm{i}\operatorname{Im}\delta_{i,i} & \delta_{i,j} \\ \delta_{j,i} & \mathrm{i}\operatorname{Im}\delta_{j,j} \end{pmatrix} \qquad 1 \leqslant i < j \leqslant n \tag{31}$$

$$\sigma_j x_{i,j} = \delta_{i,j} \qquad n+1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n. \tag{32}$$

The formulas (22) clearly imply (30). The $x_{i,j}$ from (27) clearly satisfy (32) as well. For $1 \leqslant i < j \leqslant n$, the formulas (31) can be rewritten as

$$\begin{pmatrix} \sigma_j & -\sigma_i \\ -\sigma_i & \sigma_j \end{pmatrix} \begin{pmatrix} \operatorname{Re} x_{i,j} \\ \operatorname{Re} y_{i,j} \end{pmatrix} = \begin{pmatrix} \operatorname{Re}\delta_{i,j} \\ \operatorname{Re}\delta_{j,i} \end{pmatrix}$$

$$\begin{pmatrix} \sigma_j & -\sigma_i \\ \sigma_i & -\sigma_j \end{pmatrix} \begin{pmatrix} \operatorname{Im} x_{i,j} \\ \operatorname{Im} y_{i,j} \end{pmatrix} = \begin{pmatrix} \operatorname{Im}\delta_{i,j} \\ \operatorname{Im}\delta_{j,i} \end{pmatrix}.$$

Since $\sigma_i > \sigma_j$, the formulas (23–26) indeed provide us with a solution. The entries $x_{i,j}$ with $n+1 \leqslant i, j \leqslant m$ do not affect the product $\dot{X}\Sigma$, so they can be chosen as in (28). In view of the skew symmetry constraints $x_{j,i} = -\overline{x_{i,j}}$ and $y_{j,i} = -\overline{y_{i,j}}$, we notice that the matrices $\dot{X}$ and $\dot{Y}$ are completely defined. □

5.2. Error analysis

PROPOSITION 9. *Let $\Sigma \in \mathbb{D}^{m\times n}$. Assume that $\dot{\Sigma}$, $\dot{X}$ and $\dot{Y}$ are computed using (21–28). Denote*

$$\kappa = \max\left(1, \frac{1}{\sigma_n}, \max_{i<j} \frac{1}{|\sigma_i - \sigma_j|}\right)$$
$$K = \sigma_1.$$

Given ε with $\|\dot{\Delta}\|_* \leqslant \varepsilon$, we have

$$\|\dot{\Sigma}\| \leqslant \varepsilon \tag{33}$$
$$\|\dot{X}\|_* \leqslant 2\sqrt{2}\,\kappa\,\varepsilon \tag{34}$$
$$\|\dot{Y}\|_* \leqslant 2\sqrt{2}\,\kappa\,\varepsilon. \tag{35}$$

Setting

$$\dot{\Delta}_1 \;:=\; (I_m+\dot{X})\,(\Sigma+\dot{\Sigma})\,(I_n+\dot{Y})^* - (\Sigma+\dot{\Delta})$$

and $\varepsilon_1 = \|\dot{\Delta}_1\|_*$, *we also have*

$$\varepsilon_1 \;\leqslant\; (\sqrt{2}+2\,\kappa\,(K+\varepsilon))\,4\,\kappa\,\varepsilon^2.$$

Proof. From the formula (21) we clearly have $\|\dot{\Sigma}\| \leqslant \|\dot{\Delta}\| \leqslant \varepsilon$. The formula (22) implies $|x_{i,i}| = |y_{i,i}| \leqslant \kappa\,|\delta_{i,i}|/2$ for all $i \leqslant n$. For $1 \leqslant i < j \leqslant n$, the formulas (23–26) imply

$$\begin{aligned} |\mathrm{Re}\, x_{i,j} + \mathrm{Re}\, y_{i,j}| &\;\leqslant\; \kappa\,(|\mathrm{Re}\,\delta_{i,j}| + |\mathrm{Re}\,\delta_{j,i}|) \\ |\mathrm{Re}\, x_{i,j} - \mathrm{Re}\, y_{i,j}| &\;\leqslant\; \kappa\,(|\mathrm{Re}\,\delta_{i,j}| + |\mathrm{Re}\,\delta_{j,i}|), \end{aligned}$$

whence

$$\begin{aligned} |\mathrm{Re}\, x_{i,j}| &\;\leqslant\; \kappa\,(|\mathrm{Re}\,\delta_{i,j}| + |\mathrm{Re}\,\delta_{j,i}|) \\ |\mathrm{Re}\, y_{i,j}| &\;\leqslant\; \kappa\,(|\mathrm{Re}\,\delta_{i,j}| + |\mathrm{Re}\,\delta_{j,i}|). \end{aligned}$$

Similarly,

$$\begin{aligned} |\mathrm{Im}\, x_{i,j}| &\;\leqslant\; \kappa\,(|\mathrm{Im}\,\delta_{i,j}| + |\mathrm{Im}\,\delta_{j,i}|) \\ |\mathrm{Im}\, y_{i,j}| &\;\leqslant\; \kappa\,(|\mathrm{Im}\,\delta_{i,j}| + |\mathrm{Im}\,\delta_{j,i}|). \end{aligned}$$

It follows that

$$\begin{aligned} |x_{i,j}| &\;\leqslant\; \sqrt{2}\,\kappa\,(|\delta_{i,j}| + |\delta_{j,i}|) \\ |y_{i,j}| &\;\leqslant\; \sqrt{2}\,\kappa\,(|\delta_{i,j}| + |\delta_{j,i}|). \end{aligned}$$

From (27), and using (4), we also deduce that $|x_{i,j}| \leqslant \kappa\,|\delta_{i,j}|$, for $n+1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. Combined with the fact that $\|\dot{\Delta}\|_* \leqslant \varepsilon$, we get

$$\begin{aligned} \|\dot{X}\|_* &\;\leqslant\; 2\sqrt{2}\,\kappa\,\varepsilon \\ \|\dot{Y}\|_* &\;\leqslant\; 2\sqrt{2}\,\kappa\,\varepsilon. \end{aligned}$$

Since $\dot{\Delta}_1 := (I_m+\dot{X})\,(\Sigma+\dot{\Sigma})\,(I_n+\dot{Y})^* - (\Sigma+\dot{\Delta})$ and $\dot{\Delta} = \dot{X}\,\Sigma + \Sigma\,\dot{Y}^* + \dot{\Sigma}$, we now observe that

$$\|\dot{\Delta}_1\| \;\leqslant\; (\|\dot{X}\| + \|\dot{Y}\|)\,\|\dot{\Sigma}\| + \|\dot{X}\|\,\|\dot{Y}\|\,(\|\Sigma\| + \|\dot{\Sigma}\|).$$

Plugging in the above norm bounds, we deduce that

$$\|\dot{\Delta}_1\| \;\leqslant\; 4\sqrt{2}\,\kappa\,\varepsilon^2 + 8\,\kappa^2\,\varepsilon^2\,(K+\varepsilon) \;=\; (\sqrt{2}+2\,\kappa\,(K+\varepsilon))\,4\,\kappa\,\varepsilon^2.$$

In a similar way, one proves that $\|\dot{\Delta}_1^*\| \leqslant (\sqrt{2}+2\,\kappa\,(K+\varepsilon))\,4\,\kappa\,\varepsilon^2$. □

6. PROOF OF THEOREM 1

Let us denote

$$u \;=\; K^3\,\kappa^2\,\varepsilon \;\leqslant\; 0.005$$

and, for each $i \geqslant 0$,

$$\begin{aligned} \varepsilon_0 &= \varepsilon & \varepsilon_i &= \|F(M_0, \Sigma_i, U_i, V_i)\| \\ \kappa_0 &= \kappa & \kappa_i &= \max\left(1, \frac{1}{\sigma_{i,n}}, \max_{j<k} \frac{1}{|\sigma_{i,j} - \sigma_{i,k}|}\right) \\ K_0 &= K & K_i &= \max\left(1, \max_j \sigma_{i,j}\right), \end{aligned}$$

where $\sigma_{i,1}, \ldots, \sigma_{1,n}$ denote the diagonal entries of Σ_i. Let us show by induction on i that

$$\begin{aligned}
\varepsilon_i &\leqslant 2^{1-2^i}\,\varepsilon && (36)\\
\|\Sigma_i - \Sigma_0\| &\leqslant (2-2^{2-2^i})\,\varepsilon && (37)\\
\kappa_i &\leqslant \frac{\kappa}{1-4\,\kappa\,\varepsilon} \leqslant \frac{\kappa}{1-4u} && (38)\\
K_i &\leqslant K+2\,\varepsilon \leqslant (1+2u)\,K. && (39)
\end{aligned}$$

These inequalities clearly hold for $i=0$. Assuming that the induction hypothesis holds for a given i and let us prove it for $i+1$.

By the definition of ε_i, we have $\|E_m(U_i)\|_* \leqslant \varepsilon_i$ and $\|E_n(V_i)\|_* \leqslant \varepsilon_i$. Setting

$$\begin{aligned}
\dot\Delta_i &:= X_i^* M Y_i - \Sigma_i\\
W_i &:= U_i^* M V_i,
\end{aligned}$$

we have

$$\dot\Delta_i = W_i - \Sigma_i - \frac{1}{2}E_m(U_i)\,W_i - \frac{1}{2}W_i E_n(V_i) + \frac{1}{4}E_m(U_i)\,W_i E_n(V_i).$$

It follows that

$$\begin{aligned}
\|W_i\| &\leqslant \|W_i-\Sigma_i\| + \|\Sigma_i\| \leqslant K_i + \varepsilon_i\\
\|\dot\Delta_i\| &\leqslant \varepsilon_i + \left(\varepsilon_i + \tfrac{\varepsilon_i^2}{4}\right)(K_i+\varepsilon_i)\\
&\leqslant \left(2+\tfrac{1}{4}\varepsilon_i\right)(K_i+\varepsilon_i)\,\varepsilon_i\\
&\leqslant \left(2+\tfrac{1}{4}u\right)(1+3u)\,K\,\varepsilon_i\\
&\leqslant 2.04\,K\,\varepsilon_i.
\end{aligned}$$

Let $e_i = 2.04\,K\,\varepsilon_i$. Applying Proposition 9 to $\dot\Delta_i := X_i^* M Y_i - \Sigma_i$, we get

$$\begin{aligned}
\|\dot\Sigma_i\| &\leqslant e_i\\
\|\dot X_i\| &\leqslant 2\sqrt{2}\,\kappa_i e_i\\
\|\dot Y_i\| &\leqslant 2\sqrt{2}\,\kappa_i e_i.
\end{aligned}$$

Since $\dot X_i^* = -\dot X_i$, we have

$$\begin{aligned}
U_{i+1}^* U_{i+1} - I_m &= (I_m - \dot X_i)\,X_i^* X_i\,(I_m + \dot X_i) - I_m\\
&= (I_m - \dot X_i)\,E_m(X_i)\,(I_m + \dot X_i) + (I_m - \dot X_i)\,(I_m + \dot X_i) - I_m\\
&= (I_m - \dot X_i)\,E_m(X_i)\,(I_m + \dot X_i) - \dot X_i^2
\end{aligned}$$

Using (18), we obtain

$$\begin{aligned}
\|U_{i+1}^* U_{i+1} - I_m\| &\leqslant (1+\|\dot X_i\|)^2 \|E_m(X_i)\| + \|\dot X_i\|^2\\
&\leqslant (1+\|\dot X_i\|)^2 \|E_m(U_i)\|^2 + \|\dot X_i\|^2\\
&\leqslant (1+2\sqrt{2}\,\kappa_i e_i)^2 \varepsilon_i^2 + (2\sqrt{2}\,\kappa_i)^2 e_i^2\\
&\leqslant (1+5.77\,\kappa_i K u)^2 \varepsilon_i^2 + (5.77\,\kappa_i K)^2 \varepsilon_i^2\\
&\leqslant 35\,\kappa_i^2 K^2 \varepsilon_i^2\\
&\leqslant \tfrac{35}{(1-4u)^2} 2^{2-2^{i+1}} \varepsilon^2 \leqslant \tfrac{70u}{(1-4u)^2} 2^{1-2^{i+1}} \varepsilon\\
&\leqslant 2^{1-2^{i+1}} \varepsilon.
\end{aligned}$$

Similarly,

$$\|V_{i+1}^* V_{i+1} - I_m\| \leqslant 2^{1-2^{i+1}}\,\varepsilon.$$

Using $\dot{X}_i^{\,*} = -\dot{X}_i$ and (13), we next have

$$\begin{aligned}
\Sigma_{i+1} - U_{i+1}^* M V_{i+1} &= \Sigma_i + \dot{\Sigma}_i - (I_m - \dot{X}_i)\, X_i^* M Y_i\, (I_n + \dot{Y}_i) \\
&= X_i^* M Y_i - \dot{X}_i \Sigma_i + \Sigma_i \dot{Y}_i - (I_m - \dot{X}_i)\, X_i^* M Y_i\, (I_n + \dot{Y}_i) \\
&= \Sigma_i + \dot{\Delta}_i - \dot{X}_i \Sigma_i + \Sigma_i \dot{Y}_i - (I_m - \dot{X}_i)\, (\Sigma_i + \dot{\Delta}_i)\, (I_n + \dot{Y}_i) \\
&= \dot{X}_i \dot{\Delta}_i - \dot{\Delta}_i \dot{Y}_i + \dot{X}_i\, (\Sigma_i + \dot{\Delta}_i)\, \dot{Y}_i
\end{aligned}$$

It follows that

$$\begin{aligned}
\|\Sigma_{i+1} - U_{i+1}^* M V_{i+1}\| &\leqslant 4\sqrt{2}\,\kappa_i\, e_i^2 + 8\,\kappa_i^2\,(K + 2\,\varepsilon + e_i)\, e_i^2 \\
&\leqslant (23.6\,\kappa_i + 33.3\,\kappa_i^2\,K\,(1+3\,u))\, K^2\,\varepsilon_i^2 \\
&\leqslant 57.4\,\kappa_i^2\,K^3\,\varepsilon_i^2 \leqslant \tfrac{57.4}{(1-4\,u)^2}\,\kappa^2 K^3\,\varepsilon_i^2 \\
&\leqslant 59.8\,\kappa^2 K^3\,\varepsilon_i^2 \\
&\leqslant 120\,\kappa^2 K^3\, 2^{1-2^{i+1}}\,\varepsilon^2 \leqslant 120\,u\, 2^{1-2^{i+1}}\,\varepsilon \\
&\leqslant 2^{1-2^{i+1}}\,\varepsilon.
\end{aligned}$$

This completes the proof that $\varepsilon_{i+1} \leqslant 2^{1-2^{i+1}}\,\varepsilon$. We also have

$$\|\Sigma_{i+1} - \Sigma_0\| \leqslant \|\dot{\Sigma}_i\| + \|\Sigma_i - \Sigma_0\| \leqslant 2^{1-2^i}\,\varepsilon + (2 - 2^{2-2^i})\,\varepsilon \leqslant (2 - 2^{2-2^{i+1}})\,\varepsilon.$$

We deduce that $\|\Sigma_{i+1}\| \leqslant \|\Sigma_0\| + 2\,\varepsilon$ and $K_{i+1} \leqslant K + 2\,\varepsilon$. Let us finally prove that $\kappa_{i+1} \leqslant \frac{\kappa}{1-4\,\kappa\,\varepsilon}$. From $\kappa\,\varepsilon \leqslant u \leqslant 0.005$, we get

$$\sigma_{i+1,j} \geqslant \sigma_{0,j} - 2\,\varepsilon \geqslant \sigma_{0,j}\,(1 - 2\,\kappa\,\varepsilon) > 0,$$

so that

$$\sigma_{i,j}^{-1} \leqslant \frac{\sigma_{0,j}^{-1}}{1 - 2\,\kappa\,\varepsilon}.$$

Similarly, using

$$\begin{aligned}
|\sigma_{i+1,j} - \sigma_{i+1,k}| &\geqslant |\sigma_{0,j} - \sigma_{0,k}| - |\sigma_{i+1,j} - \sigma_{0,j}| - |\sigma_{i+1,k} - \sigma_{0,k}| \\
&\geqslant |\sigma_{0,j} - \sigma_{0,k}|\,(1 - \kappa\,|\sigma_{i+1,j} - \sigma_{0,j}| - \kappa\,|\sigma_{i+1,k} - \sigma_{0,k}|) \\
&\geqslant |\sigma_{0,j} - \sigma_{0,k}|\,(1 - 4\,\kappa\,\varepsilon) > 0,
\end{aligned}$$

we get

$$|\sigma_{i+1,j} - \sigma_{i+1,k}|^{-1} \leqslant \frac{|\sigma_{0,j} - \sigma_{0,k}|^{-1}}{1 - 4\,\kappa\,\varepsilon}.$$

Hence $\kappa_{i+1} \leqslant \frac{\kappa}{1-4\,\kappa\,\varepsilon}$, which completes the proof of the four induction hypotheses (36–39) at order $i+1$.

From the continuity of the maps E_m, E_n, and $(\Sigma, U, V) \mapsto \Sigma - U^* M V$, we deduce that the sequence $(\Sigma_i, U_i, V_i)_{i \geqslant 0}$ converges. Let (Σ, U, V) be the limit. By continuity, we have $E(U) = E(V) = \Sigma - U^* M V = 0$. The unitary matrix U is of the form $U = U_0 Z$ with

$$Z = \prod_{j \geqslant 0} \left(I_m - \frac{E(U_j)}{2} \right) (I_m + \dot{X}_j). \tag{40}$$

From above we know that

$$\|\dot{X}_j\| \leqslant 4.08\,\sqrt{2}\,\kappa_i K\,\varepsilon_i \leqslant \frac{4.08\,\sqrt{2}\,\kappa K}{1 - 4\,u}\, 2^{1-2^j}\,\varepsilon \leqslant 11.8\,\kappa K\,\varepsilon\, 2^{-2^j},$$

whence

$$\left\|\left(I_m-\frac{E(U_j)}{2}\right)(I_m+\dot{X}_j)-I_m\right\| \leqslant \left(\left(1+\varepsilon 2^{-2^j}\right)11.8\,\kappa K\varepsilon+\varepsilon\right)2^{-2^j}$$
$$\leqslant 12.9\,\kappa K\varepsilon 2^{-2^j} \leqslant \frac{26\,u}{2}2^{-2^j} \leqslant \frac{1}{2}2^{-2^j}.$$

Lemma 6 now implies

$$\|Z-I_m\| \leqslant \prod_{j\geqslant 0}\left(1+\frac{26\,u}{2}2^{-2^j}\right)-1 \leqslant \frac{0.91\cdot 26\,u}{2} \leqslant 0.06.$$

This shows that Z is invertible, with

$$\|Z^{-1}\| \leqslant \frac{1}{1-\|Z-I_m\|} \leqslant 1.07.$$

Hence

$$\|U_0\| \leqslant \|U\|\|Z^{-1}\| \leqslant 1.07\sqrt{m}.$$

From the definition of U_i we also have

$$U-U_i = U_0\prod_{j=0}^{i-1}\left(I_m-\frac{E(U_j)}{2}\right)(I_m+\dot{X}_j)\left(\prod_{j\geqslant i}\left(I_m-\frac{E(U_j)}{2}\right)(I_m+\dot{X}_j)-I_m\right),$$

Using Lemma 6, this yields

$$\|U_i-U\| \leqslant \|U_0\|\prod_{j=0}^{i-1}\left(1+\frac{26\,u}{2}2^{-2^j}\right)\left(\prod_{j\geqslant i}\left(1+\frac{26\,\kappa K\varepsilon}{2}2^{-2^j}\right)-1\right)$$
$$\leqslant \|U_0\|1.06\cdot 0.91\cdot 26\,\kappa K\varepsilon 2^{-2^i} \leqslant 26.9\sqrt{m}\,\kappa K\varepsilon 2^{-2^i}.$$

Similar bounds can be computed for $\|U_i^*-U^*\|$, $\|V_i-V\|$, and $\|V_i^*-V^*\|$. Altogether, this leads to

$$\|U_0\|_* \leqslant 1.07\sqrt{m} \tag{41}$$
$$\|V_0\|_* \leqslant 1.07\sqrt{n} \tag{42}$$
$$\|U_i-U\|_* \leqslant 13.5\sqrt{m}\ \kappa\ K\varepsilon\ 2^{1-2^i} \tag{43}$$
$$\|V_i-V\|_* \leqslant 13.5\sqrt{n}\ \kappa\ K\varepsilon\ 2^{1-2^i}. \tag{44}$$

We finally have

$$\|\Sigma_i-\Sigma\| \leqslant \sum_{k\geqslant i}\|\Sigma_{k+1}-\Sigma_k\| \leqslant \sum_{k\geqslant 0}2^{1-2^{k+i}}\varepsilon \leqslant \sum_{k\geqslant 0}2^{-2^k}2^{1-2^i}\varepsilon \leqslant 0.82\cdot 2^{1-2^i}\varepsilon,$$

since $\sum_{k\geqslant 0}2^{-2^k}\leqslant 0.82$. This completes the proof.

BIBLIOGRAPHY

[1] ALEFELD, G., AND HERZBERGER, J. *Introduction to interval analysis*. Academic Press, New York, 1983.

[2] BJÖRCK, Å., AND BOWIE, C. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM J. on Num. Analysis 8*, 2 (1971), 358–364.

[3] CHARLIER, J.-P., AND VAN DOOREN, P. On Kogbetliantz's SVD algorithm in the presence of clusters. *Linear Algebra and its Applications 95* (1987), 135–160.

[4] DEMMEL, J. W. *Applied Numerical Linear Algebra*, vol. 56. Siam, 1997.

[5] DRMAC, Z. A global convergence proof for cyclic Jacobi methods with block rotations. *SIAM journal on matrix analysis and applications 31*, 3 (2009), 1329–1350.

[6] FAN, K., AND HOFFMAN, A. J. Some metric inequalities in the space of matrices. *Proc. of the AMS 6*, 1 (1955), 111–116.

[7] GALL, F. L. Powers of tensors and fast matrix multiplication. In *Proc. ISSAC 2014* (Kobe, Japan, July 2014), pp. 296–303.

[8] GOLUB, G. H., AND LOAN, F. V. *Matrix Computations*. JHU Press, 1996.

[9] HARVEY, D., AND HOEVEN, J. V. D. Faster integer multiplication using short lattice vectors. Tech. rep., ArXiv, 2018. http://arxiv.org/abs/1802.07932, to appear in the Proceedings of ANTS 2018.

[10] HARVEY, D., AND HOEVEN, J. V. D. On the complexity of integer matrix multiplication. *JSC 89* (2018), 1–8.

[11] HARVEY, D., HOEVEN, J. V. D., AND LECERF, G. Even faster integer multiplication. *Journal of Complexity 36* (2016), 1–30.

[12] HIGHAM, N. J. Matrix nearness problems and applications. In *Applications of Matrix Theory* (1989), M. J. C. Gover and S. Barnett, Eds., Oxford University Press, pp. 1–27.

[13] HOEVEN, J. V. D. Ball arithmetic. In *Logical approaches to Barriers in Computing and Complexity* (February 2010), A. Beckmann, C. Gaßner, and B. Löwe, Eds., no. 6 in Preprint-Reihe Mathematik, Ernst-Moritz-Arndt-Universität Greifswald, pp. 179–208. International Workshop.

[14] HOEVEN, J. V. D., LECERF, G., MOURRAIN, B., ET AL. Mathemagix, 2002. http://www.mathemagix.org.

[15] HOEVEN, J. V. D., AND MOURRAIN, B. Efficient certification of numeric solutions to eigenproblems. In *Proc. MACIS 2017, Vienna, Austria* (Cham, 2017), J. Blömer, I. S. Kotsireas, T. Kutsia, and D. E. Simos, Eds., Lect. Notes in Computer Science, Springer International Publishing, pp. 81–94.

[16] JANOVSKÁ, D., JANOVSKÝ, V., AND TANABE, K. An algorithm for computing the analytic singular value decomposition. *World Academy of Science, Engineering and Technology 47* (2008), 135–140.

[17] JANOVSKÝ, V., JANOVSKÁ, D., AND TANABE, K. Computing the analytic singular value decomposition via a pathfollowing. In *Numerical Mathematics and Advanced Applications*. Springer, 2006, pp. 954–962.

[18] JAULIN, L., KIEFFER, M., DIDRIT, O., AND WALTER, E. *Applied interval analysis*. Springer, London, 2001.

[19] JOHANSSON, F. Arb: a C library for ball arithmetic. *ACM Commun. Comput. Algebra 47*, 3/4 (2014), 166–169.

[20] KOVARIK, Z. Some iterative methods for improving orthonormality. *SIAM J. on Num. Analysis 7*, 3 (1970), 386–389.

[21] KULISCH, U. W. *Computer Arithmetic and Validity. Theory, Implementation, and Applications*. No. 33 in Studies in Mathematics. de Gruyter, 2008.

[22] MOORE, R. E. *Interval Analysis*. Prentice Hall, Englewood Cliffs, N.J., 1966.

[23] NEUMAIER, A. *Interval methods for systems of equations*. Cambridge University Press, Cambridge, 1990.

[24] RUMP, S. M. INTLAB - INTerval LABoratory. In *Developments in Reliable Computing*, T. Csendes, Ed. Kluwer Academic Publishers, Dordrecht, 1999, pp. 77–104. http://www.ti3.tu-harburg.de/rump/