HIGH ORDER NUMERICAL METHODS TO APPROXIMATE THE SINGULAR VALUE DECOMPOSITION*

DIEGO ARMENTANO † AND JEAN-CLAUDE YAKOUBSOHN ‡

Abstract. In this paper, we present a class of high order methods to approximate the singular value decomposition of a given complex matrix (SVD). To the best of our knowledge, only methods up to order three appear in the the literature. A first part is dedicated to define and analyse this class of method in the regular case, i.e., when the singular values are pairwise distinct. The construction is based on a perturbation analysis of a suitable system of associated to the SVD (SVD system). More precisely, for an integer p be given, we define a sequence which converges with an order $p+1$ towards the left-right singular vectors and the singular values if the initial approximation of the SVD system satisfies a condition which depends on three quantities : the norm of initial approximation of the SVD system, the greatest singular value and the greatest inverse of the modulus of the difference between the singular values. From a numerical computational point of view, this furnishes a very efficient simple test to prove and certifiy the existence of a SVD in neighborhood of the initial approximation. We generalize these result in the case of clusters of singular values. We show also how to use the result of regular case to detect the clusters of singular values and to define a notion of deflation of the SVD. Moreover numerical experiments confirm the theoretical results.

Key words. singular value decomposition,

MSC codes. 65F99,68W25

1. Introduction.

1.1. Notations and main goal. Let us consider an $m \times n$ complex matrix $M \in \mathbb{C}^{m\times n}$ where we can assume $m \geqslant n$ without loss of generalty. The terminology "diagonal" for a matrix of $\mathbb{C}^{m\times n}$ is understood if it is of the form $\begin{pmatrix} \mathrm{diag}(\sigma_1, \dots, \sigma_n) \\ 0 \end{pmatrix}$ and design by $\mathbb{D}^{m\times n}$ the set of such type matrices and also $\mathbb{E}^{m\times \ell}_{n\times q} = \mathbb{C}^{m\times \ell} \times \mathbb{C}^{n\times q} \times \mathbb{D}^{\ell\times q}$. For $\ell \geqslant 1$, we denote the identity matrix in $\mathbb{C}^{\ell\times\ell}$ by I_ℓ and for $W \in \mathbb{C}^{m\times \ell}$ we define $E_\ell(W) = W^*W - I_\ell$. The variety of Stiefel matrices is $\mathrm{St}_{m,\ell} = \{W \in \mathbb{C}^{m\times\ell} : E_\ell(W) = 0\}$. For each ℓ, $1 \leqslant \ell \leqslant m$ and q, $1 \leqslant q \leqslant n$, we know that there exists two Stiefel matrices $U \in \mathrm{St}_{m,\ell}$, $V \in \mathrm{St}_{n,q}$, and a diagonal matrix $\Sigma \in \mathbb{D}^{\ell\times q}_{\geqslant 0}$ be such that

$$f(U, V, \Sigma) \quad = \begin{pmatrix} E_\ell(U) \\ E_q(V) \\ U^*MV - \Sigma \end{pmatrix} = \ 0. \tag{1.1}$$

When $\ell = m$ and $q = n$, the triplet (U, V, Σ) is the classical singular value decompsition (SVD) of the matrix M. If $\ell < m$ or $q < n$ this abbreviated version of the SVD is referred as the thin SVD. The problem of computing a numerical thin SVD of M is to approximate the triplet (U, V, Σ) by a sequence $(U_i, V_i, \Sigma_i,)_{i\geqslant 0}$ such that the quantities $f(U_i, V_i, \Sigma_i)_{i\geqslant 0}$ converge to 0. We name *SVD sequence* a such type sequence $(U_i, \Sigma_i, V_i)_{i\geqslant 0}$.

In the context of this paper we will say that a sequence $(T_i)_{i\geqslant 0}$ of a normed space with a norm $\|.\|$ converges to T_∞ with an order $p+1 \geqslant 2$ if there exists a positive constant c be such that $\|T_i - T_\infty\| \leqslant c2^{-(p+1)^i+1}$. We then say that the numerical

---

*Submitted to the editors DATE.

†Centro de Matemática, Universidad de la República, Montevideo, Uruguay (diegoax@gmail.com)

‡Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cédex, France (yak@mip.ups-tlse.fr)

method which defines the sequence $(T_i)_{i\geqslant 0}$ is of order $p+1$. If $p=1$ (respectively $p=2$) we say that the method is quadratic (respectively cubic). Finally we say that a method associated to a map H is of order p if there exists a sequence $x_{k+1}=H(x_k)$, $k\geqslant 0$, which converges at the order p. Moreover we shall consider the matrix norm $\|A\|=\max(\|A\|_1,\|A^*\|_1)$ where

$$\|A\|_1 := \max_{1\leqslant i\leqslant m}\sum_{j=1}^{n}|M_{i,j}|.$$

Fundamental quantities occur throughout this study. From a triplet $(U,V,\Sigma)\in\mathbb{E}_{n\times q}^{m\times \ell}$ we introduce :

1. $\Delta = U^*MV-\Sigma$.
2. $\kappa(\Sigma)=\max\left(1,\max_{1\leqslant i\leqslant q}\dfrac{1}{|\sigma_i|},\max_{i\neq j}\left(\dfrac{1}{|\sigma_i-\sigma_j|}+\dfrac{1}{|\sigma_i+\sigma_j|}\right)\right)$ where the σ_i's constitute the diagonal of Σ.
3. $K(\Sigma)=\max\left(1,\max_i\sigma_i\right)$.

Throughout the text p is a given integer greater or equal to one. The goal of this paper is the construction and the convergence analysis of a class of methods of order $p+1$. The classical methods to compute the SVD are linear or quadratic : to best of our knowledge, there is no mention of any study in the literature on this subject of a method of order greater than three. These methods only use matrix addition and multiplication : there is no linear system to solve nor matrix to invert.

1.2. Construction of a quadratic method. We begin by explain how to construct a quadratic method to approximate the SVD. Let us given U,V,Σ and denote $\Delta=U^*MV-\Sigma$. The first step is to consider multiplicative perturbations such type $U\Omega$, $V\Lambda$ and S of U, V, Σ respectively and also $U_2=U_1(I_\ell+X)$ and $V_2=V_1(I_q+Y)$ multiplicative perturbations of $U_1=U(I_\ell+\Omega)$ and $V_1=V(I_q+\Lambda)$ respectively. Expanding the quantities $E_\ell(U_1)$, $E_q(V_1)$ and $\Delta_2:=U_2^*MV_2-\Sigma-S$, we get

$$E_\ell(U_1)=E_\ell(U)+\Omega+\Omega^*+\Omega^*E_\ell(U)+E_\ell(U)\Omega+\Omega^*\Omega+\Omega^*E_\ell(U)\Omega, \quad \text{idem for } E_q(V_1) \tag{1.2}$$

$$\Delta_2=\Delta_1-S+X^*\Sigma+\Sigma Y+X^*\Delta_1+\Delta_1 Y+X^*(\Delta_1+\Sigma)Y. \tag{1.3}$$

where $\Delta_1=U_1^*MV_1-\Sigma$. Denoting $\varepsilon=\max(\|E_\ell(U)\|,||E_q(V)||,\|\Delta\|)$, the second step is to determine two Hermitian matrices Ω, Λ, a diagonal matrix S, and two skew Hermitian matrices X, Y in order to get

$$\max(||E_\ell(U_2)||,||E_q(V_2)||,||\Delta_2||)\leqslant O(\varepsilon^2). \tag{1.4}$$

This occurs with $\Omega=-E_\ell(U)/2$, $\Lambda=-E_q(V)/2$ and (X,Y,S) a solution of the equation $\Delta_1-S+X^*\Sigma+\Sigma Y=0$. We will give in section 4 explicit formulas to solve this the linear equation where a solution is given by $S=\mathrm{diag}(\Delta_1)$ and X, Y that are two skew Hermitian matrices. In fact a straighforward calculation shows that

$$E_\ell(U_1)=-(3I_\ell+2\Omega)\Omega^2, \quad \text{idem for } E_q(V_1) \tag{1.5}$$

$$\Delta_1=\Delta+\Omega(\Delta+\Sigma)+(\Delta+\Sigma)\Omega+\Omega(\Delta+\Sigma)\Omega \tag{1.6}$$

$$\Delta_2=-X\Delta_1+\Delta_1 Y-X(\Delta_1+\Sigma)Y \qquad \text{since } X^*=-X \tag{1.7}$$

$$E_\ell(U_2)=(I_\ell-X)E_\ell(U_1)(I_\ell+X)+(I_\ell-X)(I_\ell+X)-I_\ell, \quad \text{idem for } E_q(V_2). \tag{1.8}$$

The formula (1.5-1.6) imply $\|E_\ell(U_1)\| \leqslant O(\varepsilon^2)$ and $\|\Delta_1\| \leqslant O(\varepsilon)$. Similarly we have $\|E_q(V_1)\| \leqslant O(\varepsilon^2)$. Moreover we will prove that $\|X\|, \|Y\| \leqslant O(\varepsilon)$ in section 4. Plugging these estimates in the formulas (1.7-1.8) we find that the inequality (1.4) holds. From the point of view of the complexity this step is the key point of the methods presented here since this requires no matrix inversion. These ingredients pave the way for the construction of a quadradic method. The third step is to introduce the map

$$H_1(U,V,\Sigma) = \begin{pmatrix} U(I_\ell+\Omega)(I_\ell+X) \\ V(I_q+\Lambda)(I_q+Y) \\ \Sigma+S \end{pmatrix}$$

where $\Omega = -\dfrac{1}{2}E_\ell(U)$, $\Lambda = -\dfrac{1}{2}E_q(V)$, $S \in \mathbb{D}^{m\times n}$ is a diagonal matrix and X, $Y$are skew Hermitian matrices be such that $\Delta_1 - S - X\Sigma + \Sigma Y = 0$. The behaviour of the sequence $(U_i, V_i, \Sigma_i)_{i\geqslant 0}$ defined by $(U_{i+1}, V_{i+1}, \Sigma_{i+1}) = H_1(U_i, V_i, \Sigma_i)$, $i \geqslant 0$ is given by Theorem 1.2.

Remark 1.1. The Newton's method is based on the cancellation of the affine part of a Taylor expansion closed to a root of the function. Here we remark that only the cancellation of a part of the affine part is enough to build a numerical quadratic method. For instance in the expression (1.2), we cancel $E_\ell(U) + \Omega + \Omega^*$ rather than $E_\ell(U) + \Omega + \Omega^* + \Omega^* E_\ell(U) + E_\ell(U)\Omega$. In the same way $\Delta_1 - S + X^*\Sigma + \Sigma Y$ is cancelled rather than $\Delta_1 - S + X^*\Sigma + \Sigma Y + X^*\Delta_1 + \Delta_1 Y$ in the expression (1.3).

1.3. Construction of a method of order $p+1$. We explain the main ideas that allow to generalize the previous method with the care to improve the condition of convergence. Taking in account the formulas (1.5-1.8) we notice that to generalize the previous construction we need the following tools. We first require a method of order $p+1$ to approximate the variety of Stiefel matrices. This is realized in considering a multiplicative perturbation $Us_p(\Omega)$ of U where $s_p(u)$ is an univariate polynomial of degree p in order that $U_1 = U(I_\ell + s_p(\Omega))$ satisfies $E_\ell(U_1) = O(E_\ell(U)^{p+1})$. This is motivated by (1.5). Next we introduce a multiplicative perturbation $U_1 c_p(U_1)$ where $c_p(u)$ is an univariate polynomial of degree p such that $(1+c_p(-u))(1+c_p(u)) - 1 = O(u^{p+1})$. This is motivated by (1.8) where appears the expression $(I_\ell - X)(I_\ell + X) - I_\ell$. The polynomials $s_p(u)$ and $c_p(u)$ as well as the matrices Ω and X are defined respectively below and their properties will be precisely studied in sections 3 and 5. Under these previous conditions a we will prove in Section 3 that a perturbation such type $U_2 = U(I_\ell + s_p(\Omega))(I_\ell + c_p(X))$ satisfies $E_\ell(U_2) = O(E_\ell(U)^{p+1})$. Finally the third tool is to determine X, Y, and S in order to get the condition $||\Delta_{p+1}|| = O(||\Delta||^{p+1})$ where $\Delta_{p+1} = U_2^* M V_2 - \Sigma - S$.

To introduce the map on which is based the method of order $p+1$ we define the following quantities:

1. Let $s_p(u)$ the truncated polynomial of degree p of the series expansion of $-1 + (1+u^2)^{-1/2}$.
2. Let $c_p(u)$ the truncated polynomial of degree p of the series expansion of $(1+u^2)^{1/2} + u - 1$.

With these preliminaries we introduce the map H_p :

$$(1.9) \qquad (U,V,\Sigma) \in \mathbb{E}^{m\times n} \to \qquad H_p(U,V,\Sigma) = \begin{pmatrix} U(I_\ell+\Omega)(I_\ell+\Theta) \\ V(I_q+\Lambda)(I_q+\Psi) \\ \Sigma+S \end{pmatrix} \in \mathbb{E}^{m\times n}$$

where :

1. $\Omega = s_p(E_\ell(U))$ and $\Lambda = s_p(E_q(V))$.
2. $\Theta = c_p(X)$ and $\Psi = c_p(Y)$ where X and Y are defined below.
3. $S = S_1 + \cdots + S_p \in \mathbb{D}^{m\times n}$, $X = X_1 + \cdots + X_p$ and $Y = Y_1 + \cdots + Y_p$ with each X_k, Y_k are skew Hermitian matrices in $\mathbb{C}^{\ell\times\ell}$ and $\mathbb{C}^{q\times q}$ respectively. Moreover each triplet (S_k, X_k, Y_k) are solutions of the following linear systems :

$$\Delta_k - S_k - X_k\Sigma + \Sigma Y_k = 0, \qquad 1 \leqslant k \leqslant p \tag{1.10}$$

where the Δ_k's for $2 \leqslant k \leqslant p+1$, are defined as

$$\Delta_1 = (I_\ell + \Omega)(\Delta + \Sigma)(I_q + \Lambda) - \Sigma, \qquad S_1 = \mathrm{diag}(\Delta_1)$$
$$\Theta_k = c_p(X_1 + \cdots + X_k), \quad \Psi_k = c_p(Y_1 + \cdots + Y_k), \quad 1 \leqslant k \leqslant p,$$
$$\Delta_k = (I_\ell + \Theta_{k-1}^*)(\Delta_1 + \Sigma)(I_q + \Psi_{k-1}) - \Sigma - \sum_{j=1}^{k-1} S_j, \tag{1.11}$$
$$S_k = \mathrm{diag}(\Delta_k), \ \ 2 \leqslant k \leqslant p.$$

We will see in section 5 that the formulas (1.10) cancel respectively the linear parts of each Δ_k. We will show that $||\Delta_{p+1}|| = O(||\Delta_1||^{p+1})$.

1.4. Main result. Then we state the folowing result which precisely shows the method associated to the map H_p is of order $p+1$.

THEOREM 1.2. *Let $p \geqslant 1$. From (U_0, V_0, Σ_0), let us define the sequence*

$$(U_{i+1}, V_{i+1}, \Sigma_{i+1}) = H_p(U_i, V_i, \Sigma_i), \quad i \geqslant 0.$$

We denote $\Delta = U_0^* M V_0 - \Sigma_0$, $K = K(\Sigma_0)$ and $\kappa = \kappa(\Sigma_0)$. We consider the constants defined in Table 1 :

	$p=1$	$p=2$	$p\geqslant 3$
a	2	4/3	4/3
u_0	0.0289	0.046	0.0297
γ_1	6.1	9.41	10.2
σ	1.67	2.1	2.62

TABLE 1

If

$$\max((\kappa K)^a ||E_\ell(U_0)||, (\kappa K)^a, ||E_q(V_0)||, \kappa^a K^{a-1} ||\Delta_0||) = \varepsilon \leqslant u_0 \tag{1.12}$$

then the sequence $(U_i, V_i, \Sigma_i)_{i\geqslant 0}$ converges to a solution $(U_\infty, V_\infty, \Sigma_\infty)$ of system (1.1) with an order of convergence equal to $p+1$. More precisely we have for $i \geqslant 0$:

$$\|U_i - U_\infty\| \leqslant \gamma_1 \sqrt{\ell} 2^{-(p+1)^i+1} \varepsilon$$
$$\|V_i - F_\infty\| \leqslant \gamma_1 \sqrt{q} 2^{-(p+1)^i+1} \varepsilon$$
$$\|\Sigma_i - \Sigma_\infty\| \leqslant \sigma \times 2^{-(p+1)^i+1} \varepsilon.$$

1.5. Arithmetic Complexity. The computation of $H_p(U,V,\Sigma)$ only requires matrix additions and multiplications without resolution of linear systems. This is possible since there are explicit formulas for the equations (1.10). Table 2 gives the number of addition and multiplications to evaluate $H_p(U,V,\Sigma)$ where $L_k := \Delta_k - S_k - X_k\Sigma + \Sigma Y_k$.

	$E_\ell(U)$	$s_p(E_\ell(U))$	$c_p(X)$	L_k	S_k	Δ_k
matrix additions	1	p	p^2		p	
matrix multiplications	1	p	p^2			$2p+2$
additions				$10np$		$(m+4n)p$
multiplications				$(m-n+8)np$		$(m+n)mnp$

TABLE 2

This implies $2(p+1)(m^2+n^2)+(m+14n)p$ additions and $2(p+1)(m^3+n^3)+(m^2+mn+m-n+8)np$ multiplications.

1.6. Outline of this paper. In section 2 we give a short overview on the computational methods for the SVD and we discuss about the method of Davies-Smith to update the SVD. We exhibit the links with the method associated to the map H_2. We also state a result on Davies-Smith method which will be proved in section 10. In section 3 we study the approximation of the unitary group by high order methods. We will use the polynomial $s_p(u)$ to define the sequence $U_{i+1} = U_i(I_\ell + s_p(E_\ell(U_i)))$, $i \geqslant 0$, from a matrix U_0 closed to the unitary group. The result is that under condition $\|E_\ell(U_0)\| < 1/4$ the sequence $(U_i)_{i\geqslant 0}$ converges to the polar projection of U_0. In section 4 we show how to explicitely solve the equation $\Delta - S - X\Sigma + \Sigma Y = 0$. We also state a condition-like result that shows the quantity κ is the condition number of this resolution. In fact we will prove that : $\|X\|, \|Y\| \leqslant \kappa\|\Delta\|$. This bound plays a signifiant role in the convergence analysis. The section 5 is devoted to the convergence analysis. We introduce the notion of p-map for the SVD. This is convenient to states in Theorem 5.2 that the method associated to a p-map is of order $p+1$. Then Theorem 1.2 derives from Theorem 5.2. The proof is done in sections 6, 7 and 8 for $p=1$, $p=2$, and $p=3$ respectively. In section 11, we study the case of clusters of singular values and we show how to use the condition (1.12) to separate clusters of singular values. We introduce a notion of deflation for the SVD : the idea is to compute a thin SVD with one singular value per cluster. Finally we illustrate this by numerical experiments in section 12.

2. Related works and discussion.

2.1. Short overview on the SVD and the methods to compute it. "The practical and theoretical importance of the SVD is hard to overestimate".This sentence from Golub and Van Loan [27] perfectly sums up the role of SVD in science and more particularly in the world of computation. The SVD was discovered by Belrami in 1873 and Jordan in 1874, see the historical survey of Stewart [43] that traces the contributions of Sylvester, Schmidt and Weyl, the first precursors of the SVD. A recent overview of numerical methods for the SVD can be found in the Hanbook of Linear Algebra [32] mainly in chapters 58 and 59. On the aspects developments on modern computers, Dongarra and all [14] give a survey of algorithms and their implementations for dense and tall matrices with comparison of performances of most bidiagonalization and Jacobi type methods. From a numerical linear algebra point of view, the SVD is at the center of the significant problems. Let us mention a few : the generalized inverse of a matrix [6], the best subspace problem [28], the orthogonal Procrustes problem [20], the linear least square problem [27], the low rank approximation problem[27]. Finally, a very stimulating article of Martin and Porter [38] describes the vitality of SVD in all areas by showing surprising examples.

There are two classes of methods to compute the SVD : bidiagonalizations methods and Jacobi methods. Since the time of precursors, Golub and Kahan in 1965 [26] for bidiagonalization with QR iteration and Kogbeliantz in 1955 [35] for Jacobi two-sided method, many various evolutions and ameliorations have been proposed. In our context ($m \geqslant n$), the bidiagonalzation methods reduce first the complex matrix under the form $M = UM'V^*$ where U, $V$are unitary and M' real and upper bidiagonal [15]. Next the SVD is computed roughly by QR iteration with notable improvements as implicit zero-shift QR [12] and differential qd algorihms [23]. In this vein of bidiagonalization methods, other alternatives to QR iteration have been developped. Let us mention the divide and conquer methods [29], [25], [37], the bisection and inverse iteration methods [34], [32] in chapter 55 and methods based on multiple relatively robust representation [13], [46]. The Jacobi methods consist to successively apply rotations now called Givens rotations on the left and right of the original matrix in order to eliminate a pair of elements at each steps. Wilkinson [45] proves that the method is ultimately quadratic for the eigenvalue problem. After Kogbetliantz, the properties of two-sided Jacobi method applying two different rotations has been studied a lot : global convergence [22], [24], quadratic convergence at the end of the algorithm [42], [2], behaviour in presence of clusters [8], reliability and accuracy [17], [18], [30], [39], [40]. Let us also mention main improvements for the one-sided Jacobi method due to several forms of preconditionning [17], [18] and [16] which uses a preconditionner QR to get high accuracy for the SVD. Finally the simultaneous use of block Jacobi methods and preconditionning improve convergence [4], [41] and computing time [14].

Other ways have been investigated related to classical topics studied in the field of numerical analysis. For instance, Chatelin [9] studies the Newton method for the eigenproblem. This requires a resolution of a Sylvester equation. Since the resolution of Sylvester is expensive, several variants of Newton method are proposed but the quadratic convergence is lost. There is also the purpose of Edelman et al. [19] which explores the geometry of Grassmann and Stiefel manifolds in the context of numerical algorithms and propose Newton method in this context. It also requires to solve a Sylvester equation to get numerical results. These ideas also have been developped by Absil et al. [1] in the context of the optimization on manifolds. Finally let us mention differential point of view developped by Chu [10] where an O.D.E. is derived for the SVD in the context of bidiagonal matrices. The methods mentioned above have a most quadratic order of convergence.

2.2. The Davies-Smith method. The method of Davies and Smith [11] to update the singular decomposition of matrices in $\mathbb{R}^{m\times n}$ is probably the closest study to our. In our framework of notations, it consists to define the map

$$(U,V,\Sigma)\to \mathrm{DS}(U,V,\Sigma)=\begin{pmatrix} U\left(I_\ell+X+\frac{1}{2}X_1^2\right)=:U\Gamma_1\\ V\left(I_q+Y+\frac{1}{2}Y_1^2\right)=:V\mathrm{K}_1\\ \Sigma+S=:\Sigma_1\end{pmatrix} \tag{2.1}$$

with $S=S_1+S_2$, $X=X_1+X_2$, $Y=Y_1+Y_2$ where the S_i's, $i=1,2,$ are diagonal matrices, the X_i's and Y_i's are skew Hermitian matrices that verify

$$X_1\Sigma-\Sigma Y_1+S_1=\Delta_1:=\Delta=U^*MV-\Sigma \tag{2.2}$$

$$X_2\Sigma-\Sigma Y_2+S_2=\Delta_2:=-\frac{1}{2}X_1(\Delta+S_1)+\frac{1}{2}(\Delta+S_1)Y_1. \tag{2.3}$$

This gives an approximation at the order three of the SVD in the regular case under the condition that the quantity $\|\Delta+\Sigma\|$ is small enough. More precisely Davies and Smith states that if the condition $\kappa^3\varepsilon^3\leqslant$ tol where tol is a given tolerance then $U\Gamma_1\Sigma K_1^*V_1^*$is an approximation of the SVD of M, such that :

1. $\|E_\ell(U\Gamma_1)\|,\|E_q(VK_1)\leqslant 2(\kappa\varepsilon)^3+O(\kappa^4\varepsilon^4)$.
2. $\dfrac{1}{\|M\|}\|\Gamma_1^*U^*MVK_1-\Sigma_1\|\leqslant\dfrac{28}{3}(\kappa\varepsilon)^3+O(\kappa^4\varepsilon^4)$

where the considered norm is that of Frobenius. Thanks to the map H_p defined in the introduction with $p=2$, we improve the previous method and its analysis on several points.

1. The norm of $E_\ell(U(I_\ell+\Omega)(I_q+\Theta))$ is in $O(\varepsilon^3)$, see Theorem 2.1 below, while the norm of $E_\ell(U\Gamma_1)$ depends on the norm of $E_\ell(U)$. In fact
$$E_\ell(U\Gamma_1)=\Gamma_1^*E_\ell(U)\Gamma_1+E_\ell(\Gamma_1).$$
For this reason, Davies and Smith suggest to use a Givens type method after their update of the SVD to iterate the method.
2. Note that $\Theta_2=X_1+X_2+\dfrac{1}{2}(X_1+X_2)^2$ is computed with the same arithmetic complexity as Γ_1. There is a gain in the error analysis.
3. The analysis of the map H_2 takes in account all the terms of the series expansion of $H_2(U,V,\Sigma)$ with respect U,V,Σ. In this way, the Theorem 2.1 show that $\kappa^{5/4}K^{2/5}\varepsilon$ (and not $\kappa\varepsilon$) is the quantity on which the method Davies Smith rests. This shows that the quantity K is not negligible in the error analysis.
4. The tolerance tol in the method associated to the map H_p is determined by imposing a condition of contraction which is not the case in the Davies-Smith method, see the algorithm 2.3 of [11].

We defined a Davies-Smith revisited method introducing the map

$$(U,V,\Sigma)\to\overline{\mathrm{DS}}(U,V,\Sigma)=\begin{pmatrix} U(I_\ell+\Theta_2)\\ V(I_q+\Psi_2)\\ \Sigma+S=:\Sigma_1\end{pmatrix} \tag{2.4}$$

with $S=S_1+S_2$, $X=X_1+X_2$, $Y=Y_1+Y_2$ where the S_i's, $i=1,2,$ are diagonal matrices, the X_i's and Y_i's are skew Hermitian matrices defined by (2.2-2.3). The following result specifies the behaviour of $\mathrm{DS}(U,V,\Sigma)$ and $\overline{\mathrm{DS}}(U,V,\Sigma)$.

Theorem 2.1. *Let us consider M, U, V, Σ as in the introduction, $\Delta=U^*MV-\Sigma$ and $\varepsilon_1=\|\Delta\|$. Let $\kappa=\kappa(\Sigma)$ and $K=K(\Sigma)$.*

1. *Let us assume that* $\kappa^{5/4}K^{2/5}\varepsilon_1 \leqslant \varepsilon \leqslant 0.1$. *Then the triplet* $(U_1, V_1, \Sigma_1) = \mathrm{DS}(U, V, \Sigma)$ *defined by* (2.1) *satisfies*

$$\|\Delta_1\| := \|U_1^* M V_1 - \Sigma_1\| \leqslant (8 + 18\varepsilon + 33\varepsilon^2)\varepsilon^3. \tag{2.5}$$

2. *Let us assume that* $\kappa^{6/5}K^{3/10}\varepsilon_1 \leqslant \varepsilon \leqslant 0.1$. *Then the triplet* $(\bar{U}_1, \bar{V}_1, \bar{\Sigma}_1) = \overline{\mathrm{DS}}(U, V, \Sigma)$ *defined by* (2.4) *satisfies*

$$\|\bar{\Delta}_1\| := \|\bar{U}_1^* M \bar{V}_1 - \bar{\Sigma}_1\| \leqslant (6 + 21\varepsilon + 54\varepsilon^2)\varepsilon^3. \tag{2.6}$$

Since $\kappa^{6/5}K^{3/10} < \kappa^{5/4}K^{2/5}$, the condition to update the singular value decomposition is better with the Davies Smith method revisited than the Davies Smith method.

3. Approximation of Stiefel matrices. The Stieffel manifold $\mathrm{St}_{m,\ell}$ generalizes the Unitary group. An important tool is the polar decomposition $U_0 = \pi(U_0)H$ of rectangular matrix U_0 where the polar projection $\pi(U_0)$ is a Stiefel matrix and H is Hermitian positive semidefinite [33]. It is also well known that $\pi(U_0)$ is indeed the closest element in $\mathrm{St}_{m,l}$ to U_0 for every unitarily norm [21, Theorem 1]. Since we are doing approximate computations, the Stiefel matrices in an SVD are not given exactly, so we may wish to estimate the distance between an approximate Stiefel matrix and the closest actual Stiefel matrix. This is related to the following problem: given an approximately Stiefel $m \times \ell$ matrix U, find a good approximation $U + \dot{U}$ for its projection on the manifold $\mathrm{St}_{m,\ell}$. We define a class of high order iterative methods for this problem and provide a detailed analysis of its convergence, see also [36, 7, 31]. The theorem 3.3 establishes that our method converges towards the polar projection of the matrix $U_0 \in \mathbb{C}^{m\times\ell}$ if U_0 is sufficiently close to the Stiefel manifold. In this case the matrix H is positive definite and can uniquely be written as the exponential of another Hermitian matrix.

3.1. A class of high order iterative methods. We wish to compute $\dot{U}$ using an appropriate Newton iteration. Since the normal space in U of Stiefel manifol is composed of $U\Omega$'s where eΩ is an Hermitian matrix,it turns out that it is more convenient to write $U + \dot{U} = U(I_\ell + \Omega)$. The following lemma gives the expression Ω so that $U + \dot{U} \in \mathrm{St}_{m,\ell}$ it is the polar projection of U.

LEMMA 3.1. *Let* $U \in \mathbb{C}^{m\times\ell}$ *such that the spectral radius of* $E_\ell(U)$ *is strictly less than* 1. *Then*

$$\Omega = -I_\ell + (I_\ell + E_\ell(U))^{-1/2} \Rightarrow E_\ell(U + U\Omega) = 0. \tag{3.1}$$

Hence $U(I_\ell + E_\ell(U))^{-1/2} \in \mathrm{St}_{m,\ell}$ *is the polar projection of* U.

Proof. If the spectral radius of $E_\ell(U)$ is strictly less than 1 then the matrix $(I_\ell + E_\ell(U))^{1/2}$ exists and $\Omega = -I_\ell + (I_\ell + E_\ell(U))^{-1/2}$ is Hermitian positive definite matrix. With $E_\ell(U) = U^*U - I_\ell$ and $\dot{U} = U\Omega$, we have

$$\begin{aligned} E_\ell(U + U\Omega) &= (I_\ell + \Omega^*)(I_\ell + E_\ell(U))(I_\ell + \Omega) - I_\ell \\ &= E_\ell(U) + 2\Omega + \Omega E_\ell(U) + E_\ell(U)\Omega + \Omega^2 + \Omega E_\ell(U)\Omega. \end{aligned}$$

A straighforward calculation implies $E_\ell(U+U\Omega) = 0$. Then $U = U(I_\ell+\Omega)(I_\ell+\Omega)^{-1}$. Hence $U(I_\ell + E_\ell(U))^{-1/2} \in \mathrm{St}_{m,\ell}$ is the polar projection of U. ∎

Consequently an high order approximation of $\Omega = -I_\ell + (I_\ell + E_\ell(U))^{-1/2}$ will permit to define an high order method to numerically compute the polar projection. Evidently Ω commutes with U. The approximation of Ω can be obtained as follows. Let us consider the Taylor serie of $-1+(1+u)^{-1/2}$ at $u = 0$:

$$s(u) = \sum_{k\geqslant 1}(-1)^k \frac{1}{4^k}\begin{pmatrix} 2k \\ k \end{pmatrix} u^k = -\frac{1}{2}u + \frac{3}{8}u^2 - \frac{5}{16}u^3 + \cdots$$

For $p \geqslant 1$ we introduce $s_p(u) = \sum_{k=1}^{p}(-1)^k t_k u^k$ and $r_p(u) = s(u) - s_p(u)$. The quantities

$$\Omega_p = s_p(E_\ell(U)), \quad R_p = r_p(E_\ell(U)) \tag{3.2}$$

commute with U^*U. We have $\Omega_p = \Omega - R_p$ and $E_\ell(U + U\Omega) = 0$. A straightforward calculation shows that

$$\begin{aligned} E_\ell(U + U\Omega_p) &= (U^* + \Omega_p U^* - R_p U^*)(U + U\Omega_p - UR_p) - I_\ell \\ &= E(U + U\Omega) - 2(I_\ell + \Omega)U^*UR_p + R_p^2 U^*U \\ &= (I_\ell + E_\ell(U))R_p(-2I_\ell - 2\Omega + R_p) \qquad \text{since } U^*U = I_\ell + E_\ell(U) \end{aligned} \tag{3.3}$$

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

$$U_{i+1} = U_i(I_\ell + s_p(E_\ell(U_i))), \qquad i \geqslant 0. \tag{3.4}$$

Theorem 3.3 below shows the convergence of the sequence (3.4) towards the polar projection of U_0 with a p order of convergence under the universal condition $\|E(U_0)\| < 1/4$.

3.2. Error analysis.

Proposition 3.2. *Let $p \geqslant 1$. Let U be an $m \times \ell$ matrix with $\varepsilon := \|E_\ell(U)\| < 1$ and $\Omega_p = s_p(E_\ell(U))$. Let $U_1 = U(I_\ell + \Omega)$ and write $\varepsilon_1 := \|E_\ell(U_1)\|$. Then $\|\Omega_p\| \leqslant |s_p(\varepsilon)| \leqslant -1 + (1-\varepsilon)^{-1/2}$ and*

$$\varepsilon_1 \leqslant \varepsilon^{p+1}. \tag{3.5}$$

Proof. Let $\Omega_p = s_p(E_\ell(U))$. We have

$$\begin{aligned} \|\Omega_p\| &\leqslant |s_p(\varepsilon)| \\ &\leqslant -1 + (1-\varepsilon)^{-1/2}. \end{aligned}$$

Since Ω is Hermitian which commutes with U we have

$$\begin{aligned} E_\ell(U_1) &= (I_\ell + \Omega_p)U^*U(I_\ell + \Omega_p) - I_\ell \\ &= (I_\ell + \Omega_p)^2 E_\ell(U) + \Omega_p^2 + 2\Omega_p \\ &= (I_\ell + E_\ell(U))(\Omega_p^2 + 2\Omega_p) + E_\ell(U). \end{aligned}$$

Then using Lemma 3.4 below in sub-section, it follows easily that

$$E_\ell(U_1) = \left(\sum_{k=0}^{p} \alpha_k E_\ell(U)^k\right) E_\ell(U)^{p+1}$$

where $\sum_{k=0}^{p} |\alpha_k| \leqslant 1$. Hence $\varepsilon_1 \leqslant \varepsilon^{p+1}$. ∎

Proposition 3.2 permits to analyse the behaviour of the sequence $(U_i)_{i\geqslant 0}$ deftined by (3.4).

THEOREM 3.3. *let $p \geqslant 1$. Let $U_0 \in \mathbb{C}^{m\times\ell}$ be such that $\|E(U_0)\| \leqslant \varepsilon < 1/2$. Then the sequence defined by*

$$U_{i+1} = U_i(I_\ell + s_p(E(U_i)) \qquad i \geqslant 0, \tag{3.6}$$

converges to a Stiefel matrix $U_\infty \in \mathrm{St}_{m,\ell}$. More precisely, for all $i \geqslant 0$, we have

$$\|U_i - U_\infty\| \leqslant \sqrt{\ell}\frac{2^{-(p+1)^i+1}2\varepsilon}{1-2\varepsilon}. \tag{3.7}$$

Moreover if $\varepsilon < 1/4$ then this sequence converges to the polar projection $\pi(U_0) \in \mathrm{St}_{m,\ell}$ of U_0.

Proof. The Newton sequence (3.6) defined from $U_0 = U$ gives

$$U_{i+1} = U_0(I_\ell + \Omega_{0,p})\cdots(I_\ell + \Omega_{i,p})$$

with $\Omega_{i,p} = s_p(E_\ell(U_i))$. An obvious induction using Proposition 3.2 yields $\|E_\ell(U_i)\| \leqslant 2^{-(p+1)^i+1}\varepsilon$. In fact we have

$$\begin{aligned}
||E_\ell(U_{i+1})|| &\leqslant ||E_\ell(U_i)||^{p+1} && \text{from Proposition 3.2}\\
&\leqslant 2^{-(p+1)^{i+1}+p+1}\varepsilon^{p+1}\\
&\leqslant (2\varepsilon)^p 2^{-(p+1)^{i+1}+1}\varepsilon\\
&\leqslant 2^{-(p+1)^{i+1}+1}\varepsilon && \text{since} \quad \varepsilon < 1/2.
\end{aligned}$$

We are using Lemma 3.6 to conclude. We have $\|\Omega_{k,p}\| \leqslant -1 + (1-2^{-(p+1)^k+1}\varepsilon)^{-1/2}$. Since $\varepsilon \leqslant 1/2$ then $-1 + (1-2^{-(p+1)^k+1}\varepsilon)^{-1/2} \leqslant 2^{-(p+1)^k+1}\varepsilon$. Considering $u_0 = \varepsilon$, $\alpha_1 = 1$ and $\alpha_2 = 0$, the assumptions of Lemma 3.6 below are satisfied. Hence the sequence $(U_i)_{i\geqslant 0}$ converges to a matrix U_∞ which is an unitary matrix since the sequence $(E_\ell(U_i)_{i\geqslant 0}$ converges towards 0. We then have

$$\begin{aligned}
\|U_i - U_\infty\| &\leqslant \sqrt{\ell}\frac{2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)}{1-2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)u_0}2^{-(p+1)^i+1}\alpha_0\varepsilon\\
&\leqslant \sqrt{\ell}\frac{2^{-(p+1)^i+1}2\varepsilon}{1-2\varepsilon}.
\end{aligned}$$

We denote $Z_0 = \prod_{j\geqslant 0}(I_\ell + \Omega_{j,p})$. We have $U_\infty = U_0 Z_0$. From Lemma 3.6 Z_0 is invertible with $\|Z_0\| \leqslant 2\varepsilon$. By induction on i, it can also be checked that all the $\Omega_{i,p}$'s commute. Whence Z_0 and Z_0^{-1} are actually Hermitian matrices. If $\varepsilon < 1/4$ we have $\|Z_0^{-1} - I_\ell\| \leqslant \|Z_0^{-1}\|\|I_\ell - Z_0\| \leqslant 2\varepsilon/(1-2\varepsilon) < 1$. Then 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. Since $U_0 = U_\infty Z_0^{-1}$, we conclude that $U_\infty = \pi(U_0)$ the polar projection of U_0 from the polar decomposition theorem. ∎

3.3. Technical Lemmas. This following Lemma is used in the proof of Proposition 3.2.

LEMMA 3.4. *Let $p \geqslant 1$. We have*

$$(u+1)(s_p(u)^2 + 2s_p(u)) + u = \left(\sum_{k=0}^{p}\alpha_k u^k\right)u^{p+1}$$

where $\sum_{k=0}^{p} |\alpha_k| \leqslant 1$.

Proof. Let $t_i = (-1)^i \frac{1}{4^i} \binom{2i}{i}$ for $i \geqslant 0$. The convolution of sequence binomial t_i with itself is the sequence with general terms (-1)^i. In fact it is sufficient to square $(1+u)^{-1/2}$:

$$\frac{1}{1+u} = \sum_{k\geqslant 0} (-1)^k u^k = \sum_{k\geqslant 0} \left(\sum_{i+j=k} t_i t_j \right) u^k.$$

We proceed by induction. When $p = 1$ the lemma holds since

$$\begin{aligned} (u+1)(h_1(u)^2 + 2h_1(u)) + u &= (u+1)\left(\frac{u^2}{4} - u\right) + u \\ &= \left(-\frac{3}{4} + \frac{1}{4}u\right) u^2 \end{aligned}$$

and $\frac{1}{4} + \frac{3}{4} = 1$. Let us suppose that the lemma holds for an indice $p \geqslant 1$ be given. We first remark that $\alpha_0 = -2t_{p+1}$. In fact since α_0 is the coefficient of u^{p+1} in $(u+1)(s_p(u)^2 + 2s_p(u)) + u$. Then

$$\begin{aligned} \alpha_0 &= \sum_{\substack{i+j=p \\ 1\leqslant i,j\leqslant p}} t_i t_j + \sum_{\substack{i+j=p+1 \\ 1\leqslant i,j\leqslant p}} t_i t_j + 2t_p \\ &= (-1)^p - 2t_0 t_p + (-1)^{p+1} - 2t_0 t_{p+1} + 2t_p \\ &= -2t_{p+1}. \end{aligned}$$

Next, writing $h_{p+1}(u) = s_p(u) + t_{p+1}u^{p+1}$ we get by straightforward calculations :

$$\begin{aligned} &(u+1)(s_p(u)^2 + 2s_p(u)) + u \\ &= \left(\sum_{k=0}^{p} \alpha_k u^k\right) u^{p+1} + (u+1)(2t_{p+1}s_p(u)u^{p+1} + t_{p+1}^2 u^{2(p+1)} + 2t_{p+1}u^{p+1}) \\ &= (\alpha_1 + 2t_{p+1}(t_1+1))u^{p+2} + \sum_{k=2}^{p} (\alpha_k + 2t_{p+1}(t_k + t_{k-1}))u^{p+k+1} \\ &\quad + t_{p+1}(2t_p + t_{p+1})u^{2(p+1)} + t_{p+1}^2 u^{2p+3} \\ &:= \left(\sum_{k=0}^{p+1} \beta_k u^k\right) u^{p+2} \end{aligned}$$

Let us prove that $\sum_{k=0}^{p+1} |\beta_k| \leqslant 1$. In fact since $t_1 = -1/2$ and $\sum_{k=1}^{p} |\alpha_k| = 1 - 2|t_{p+1}|$ it follows:

$$\begin{aligned} \sum_{k=0}^{p+1} |\beta_k| &\leqslant \sum_{k=1}^{p} |\alpha_k| + |t_{p+1}| + 2|t_{p+1}| \sum_{k=2}^{p} (|t_{k-1}| - |t_k|) + |t_{p+1}|(2|t_p| - |t_{p+1}|) + t_{p+1}^2 \\ &\leqslant 1 - 2|t_{p+1}| + |t_{p+1}| + 2|t_{p+1}|(|t_1| - |t_p|) + |t_{p+1}|(2|t_p| - |t_{p+1}|) + t_{p+1}^2 \\ &\leqslant 1. \end{aligned}$$

The Lemma is proved. □

The following Lemma 3.5 is used in the proof of Lemma 3.6.

LEMMA 3.5. 1. *Let* $0 \leqslant u < 1$. *We have* $\prod_{j\geqslant 0}(1+u^{2^j}) = \dfrac{1}{1-u}$.

2. *Let* $p \geqslant 1$ *and* $0 \leqslant \varepsilon < 1$. *We have for* $i \geqslant 0$,

$$\prod_{j\geqslant 0}(1+2^{-(p+1)^{j+i}+1}\varepsilon) \leqslant 1+2^{-(p+1)^i+1}2\varepsilon \tag{3.8}$$

3. *Let* $p \geqslant 1$ *and* $0 \leqslant \varepsilon \leqslant 1/2$. *We have for* $i \geqslant 0$,

$$\prod_{j\geqslant 0}(1-2^{-(p+1)^{j+i}+1}\varepsilon)^{-1/2} \leqslant 1+2^{-(p+1)^i+1}2\varepsilon \tag{3.9}$$

Proof. For the item 1 we prove by induction that $\prod_{j=0}^{k}(1+u^{2^j}) = \dfrac{1-u^{2^{k+1}}}{1-u}$. This holds when $k=0$. Next, assuming the property for k be given we have

$$\begin{aligned}\prod_{j=0}^{k+1}(1+u^{2^j}) &= \frac{1-u^{2^{k+1}}}{1-u}(1+u^{2^{k+1}})\\ &= \frac{1-u^{2^{k+2}}}{1-u}.\end{aligned}$$

Item 1 is proved. The item 2 follows from

$$\begin{aligned}\prod_{j\geqslant 0}(1+2^{-(p+1)^{j+i}+1}\varepsilon) &\leqslant \prod_{j\geqslant 0}(1+(2^{-(p+1)^i})^{2^j}2\varepsilon)\\ &\leqslant 1+\left(\prod_{j\geqslant 0}(1+(2^{-(p+1)^i})^{2^j})-1\right)2\varepsilon\\ &\leqslant 1+\left(\frac{1}{1-2^{-(p+1)^i}}-1\right)2\varepsilon \qquad \text{from item 1.}\\ &\leqslant 1+2^{-(p+1)^i}4\varepsilon.\end{aligned}$$

Since $\varepsilon \leqslant 1/2$ we have $(1-u)^{-1/2} \leqslant 1+u$, item 3 follows from :

$$\begin{aligned}\prod_{j\geqslant 0}(1-2^{-(p+1)^{j+i}+1}\varepsilon)^{-1/2} &\leqslant \prod_{j\geqslant 0}(1+2^{-(p+1)^{i+j}+1}\varepsilon)\\ &\leqslant 1+2^{-(p+1)^i+1}2\varepsilon \quad \text{from item 2.}\end{aligned}$$
□

The Lemma 3.6 is used in Theorems 3.3 and 5.2.

LEMMA 3.6. *Let* ε, u_0, *and* α_i, $i=1,2$, *be real numbers such that* $\varepsilon \leqslant u_0$ *and* $2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)u_0 < 1$. *Let us consider a sequence of matrices defined by*

$$U_{i+1} = U_i(I_\ell+\Omega_i)(I_l+\Theta_i), \qquad i \geqslant 0,$$

where the norms of the Ω_i*'s and the* Θ_i*'s satisfy*

$$\|\Omega_i\| \leqslant \alpha_1 2^{-(p+1)^i+1}\varepsilon \qquad \text{and} \qquad \|\Theta_i\| \leqslant \alpha_2 2^{-(p+1)^i+1}\varepsilon.$$

Then the sequence $(U_i)_{i\geqslant 0}$ converges to a matrix U_∞. If U_∞ is an unitary matrix then each U_i is invertible and we have

$$\|U_i - U_\infty\| \leqslant \sqrt{\ell}\frac{2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)}{1-2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)u_0}2^{-(p+1)^i+1}\varepsilon.$$

Moreover each $N_i = \prod_{j\geqslant 0}(I_\ell+\Omega_{i+j})(I_\ell+\Theta_{i+j})$ is invertible and satisfies

$$\|N_i - I_\ell\| \leqslant 1 - 2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)u_0.$$

Proof. We remark that $U_i = U_0\prod_{j=0}^{i-1}(I_\ell+\Omega_j)(I_\ell+\Theta_j)$. Let $N_i = \prod_{j\geqslant 0}(I_\ell + \Omega_{i+j})(I_\ell+\Theta_{i+j})$. Let us consider $U_\infty = U_0N_0$. From assumption we know that $\|\Omega_j\| \leqslant \alpha_1 2^{-(p+1)^j+1}\varepsilon$ and $\|\Theta_k\| \leqslant \alpha_2 2^{-(p+1)^j+1}\varepsilon$. Taking in account that $\varepsilon \leqslant u_0$, it follows

$$(1+\|\Omega_{i+j}\|)(1+\|\Theta_{i+j}\|) \leqslant 1 + (\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)\times 2^{-(p+1)^{i+j}+1}\varepsilon.$$

The matrix $N_i - I_\ell$ is written an infinite sum of homogeneous polynomials of degree $k \geqslant 1$:

$$N_i - I_\ell = \sum_{k\geqslant 1} P_k(\Omega_i,\ldots,\Omega_{i+j},\ldots\Theta_i,\ldots,\Theta_{i+j},\ldots)$$

Consequently for $i \geqslant 0$ we have :

$$\begin{aligned}
\|N_i - I_\ell\| &\leqslant \sum_{k\geqslant 1} P_k(\|\Omega_i\|,\ldots\|\Omega_{i+j}\|,\ldots,\|\Theta_i\|,\ldots,\|\Theta_{i+J}\|,\ldots)\\
&\leqslant \prod_{j\geqslant 0}(1+\|\Omega_{i+j}\|)(1+\|\Theta_{i+j}\|) - 1\\
&\leqslant \prod_{j\geqslant 0}(1+(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)\times 2^{-(p+1)^{i+j}+1}\varepsilon) - 1\\
&\leqslant 2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)2^{-(p+)^i+1}\varepsilon \quad \text{from Lemma ??}\\
&\leqslant 2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)u_0 \quad \text{since} \quad \varepsilon \leqslant u_0
\end{aligned}$$

Since $2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)u_0 < 1$ it follows that each N_i is invertible. Since $U_\infty = U_0N_0$ it is easy to see

$$\|U_\infty\| \leqslant \|U_0\|(1+2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)\varepsilon).$$

We have $U_i = U_\infty N_i^{-1}$. We deduce that

$$\begin{aligned}
\|U_i - U_\infty\| &\leqslant \|U_\infty N_i^{-1}(I_\ell - N_i)\|\\
&\leqslant \|U_\infty\|\frac{1}{1-2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)u_0}2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)2^{-(p+1)^i+1}\varepsilon.
\end{aligned}$$

If U_∞ is an unitary matrix then each U_i is invertible and $\|U_\infty\| \leqslant \sqrt{\ell}$. The result is proved. ∎

LEMMA 3.7. *From $U_0 \in \mathbb{C}^{m\times\ell}$ be given, let us define the sequence for $i \geqslant 0$, $U_{i+1} = U_i(I_\ell + \Omega_{i,p})$ with $\Omega_{i,p} = s_p(E_\ell(U_i))$. Let $\varepsilon = \|E_\ell(U_0)\|$. Then we have*

$$\|\Omega_{i,p}\| \leqslant (-1 + (1-\varepsilon)^{-1/2})\varepsilon^{(p+1)^i-1}$$

Proof. From Proposition 3.2 we know that $\|E_\ell(U_i)\| \leqslant \varepsilon^{(p+1)^i}$. Since $s_p(u) \leqslant -1 + (1-u)^{-1/2}$ we can write $\|\Omega_{i,p}\| \leqslant -1 + (1-\varepsilon^{(p+1)^i})^{-1/2}$. The function $u \to \frac{1}{u}(-1 + (1-u)^{-1/2})$ is defined and is increasing on $[0,1]$. We then find that

$$\|\Omega_{i,p}\| \leqslant \frac{1}{\varepsilon}(-1 + (1-\varepsilon)^{-1/2})\varepsilon^{(p+1)^i}.$$

We are done. □

4. SVD for perturbed diagonal matrices.

4.1. Solving the equation $\Delta - S - X\Sigma + \Sigma Y = 0$. The following proposition shows how to explicitly solve this linear equation under these constraints without inverting a matrix.

PROPOSITION 4.1. *Let $\Sigma = \mathrm{diag}(\sigma_1, \ldots \sigma_q) \in \mathbb{D}^{\ell\times q}$ and $\Delta = (\delta_{i,j}) \in \mathbb{C}^{\ell\times q}$. Consider the diagonal matrix $S \in \mathbb{D}^{\ell\times q}$ and the two skew Hermitian matrices $X = (x_{i,j}) \in \mathbb{C}^{\ell\times\ell}$ and $Y = (y_{i,j}) \in \mathbb{C}^{q\times q}$ that are dend the tfined by the following formulas :*

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

$$S_{i,i} = \mathrm{Re}\,\delta_{i,i} \tag{4.1}$$

$$x_{i,i} = -y_{i,i} = \frac{\mathrm{Im}\,\delta_{i,i}}{2\sigma_i}\mathrm{i} \tag{4.2}$$

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

$$x_{i,j} = \frac{1}{2}\left(\frac{\delta_{i,j} + \overline{\delta_{j,i}}}{\sigma_j - \sigma_i} + \frac{\delta_{i,j} - \overline{\delta_{j,i}}}{\sigma_j + \sigma_i}\right) \tag{4.3}$$

$$y_{i,j} = \frac{1}{2}\left(\frac{\delta_{i,j} + \overline{\delta_{j,i}}}{\sigma_j - \sigma_i} - \frac{\delta_{i,j} - \overline{\delta_{j,i}}}{\sigma_j + \sigma_i}\right) \tag{4.4}$$

- *For $q+1 \leqslant i \leqslant \ell$ and $1 \leqslant j \leqslant q$, we take*

$$x_{i,j} = \frac{1}{\sigma_j}\delta_{i,j}. \tag{4.5}$$

- *For $q+1 \leqslant i \leqslant \ell$ and $q+1 \leqslant j \leqslant \ell$, we take*

$$x_{i,j} = 0. \tag{4.6}$$

Then we have

$$\Delta - S - X\Sigma + \Sigma Y = 0 \tag{4.7}$$

Proof. Since X and Y are skew Hermitian matrices, we have $\mathrm{diag}(\mathrm{Re}(X\Sigma - \Sigma Y)) = 0$. In view of (4.1), we thus get

$$\mathrm{diag}(\mathrm{Re}\,\Delta) = \mathrm{diag}\,\mathrm{Re}(X\Sigma - \Sigma Y + S).$$

By skew symmetry, for the equation

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

holds, it is sufficient to have

$$\sigma_i x_{i,i} - \sigma_i y_{i,i} = \mathrm{i}\operatorname{Im}\delta_{i,i}, \qquad 1 \leqslant i \leqslant q. \tag{4.8}$$

$$\begin{pmatrix} \sigma_i x_{i,i} & \sigma_j x_{i,j} \\ -\sigma_i \overline{x_{i,j}} & \sigma_j x_{j,j} \end{pmatrix} - \begin{pmatrix} \sigma_i y_{i,i} & \sigma_i y_{i,j} \\ -\sigma_j \overline{y_{i,j}} & \sigma_j y_{j,j} \end{pmatrix} = \begin{pmatrix} \mathrm{i}\operatorname{Im}\delta_{i,i} & \delta_{i,j} \\ \delta_{j,i} & \mathrm{i}\operatorname{Im}\delta_{j,j} \end{pmatrix}, \quad 1 \leqslant i < j \leqslant q \tag{4.9}$$

$$\sigma_j x_{i,j} = \delta_{i,j}, \qquad q+1 \leqslant i \leqslant \ell, \quad 1 \leqslant j \leqslant q. \tag{4.10}$$

The formulas (4.2) clearly imply (4.8). The $x_{i,j}$ from (4.3) clearly satisfy (4.10) as well. For $1 \leqslant i < j \leqslant q$, the formulas (4.9) 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 (4.3–4.4) indeed provide us with a solution. The entries $x_{i,j}$ with $q+1 \leqslant i, j \leqslant \ell$ do not affect the product $X\Sigma$, so they can be chosen as in (4.6). 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 X and Y are completely defined. ∎

DEFINITION 4.2. *Let $\Sigma = \operatorname{diag}(\sigma_1, \dots \sigma_q) \in \mathbb{D}^{\ell\times q}$ and $\Delta \in \mathbb{C}^{\ell\times q}$. We name condition number of equation $X\Sigma - \Sigma Y = \Delta - S$ the quantity*

$$\kappa = \kappa(\Sigma) = \max\left(1, \max_{1\leqslant i\leqslant q} \frac{1}{\sigma_i}, \max_{1\leqslant i<j\leqslant q} \frac{1}{\sigma_i - \sigma_j} + \frac{1}{\sigma_i + \sigma_j}\right) \tag{4.11}$$

4.2. Error analysis.

PROPOSITION 4.3. *Under the notations and assumptions of Proposition 4.1, assume that X, Y and S are computed using* (4.1–4.4)*. Given ε with $\|\Delta\| \leqslant \varepsilon$, the matrices X, Y and S solutions of $\Delta - S - X\Sigma + \Sigma Y = 0$ satisfy*

$$\|S\| \leqslant \varepsilon \tag{4.12}$$

$$\|X\|, \|Y\| \leqslant \kappa\varepsilon \tag{4.13}$$

Proof. From the formula (4.1) we clearly have $\|S\| \leqslant \|\Delta\| \leqslant \varepsilon$.

Since $\Sigma \in \mathbb{D}^{\ell\times q}$ we know that $\sigma_i > \sigma_j$ for $i < j$. It follows

$$\begin{aligned} |x_{i,j}| &\leqslant \frac{|\delta_{i,j}|}{2}\left(\frac{1}{\sigma_i - \sigma_j} + \frac{1}{\sigma_i + \sigma j}\right) + \frac{|\overline{\delta_{i,j}}|}{2}\left(\frac{1}{\sigma_i - \sigma_j} + \frac{1}{\sigma_i + \sigma_j}\right) \\ &\leqslant \kappa|\delta_{i,j}| \qquad \text{since} \quad |\delta_{i,j}| = |\overline{\delta_{i,j}}|. \end{aligned}$$

We also have $|x_{i,i}| \leqslant \dfrac{|\delta_{i,i}|}{\sigma_i}$ and for $q+1 \leqslant i \leqslant \ell$ and $1 \leqslant j \leqslant q$, $|x_{i,i}| \leqslant \dfrac{|\delta_{i,i}|}{\sigma_j}$. Combined with the fact that $\|\Delta\| \leqslant \varepsilon$, we get $\|X\| \leqslant \kappa\varepsilon$. In the same way we also have $\|Y\| \leqslant \kappa\varepsilon$. ∎

5. Convergence analysis : a general result.

DEFINITION 5.1. *Let an integer $p \geqslant 1$. Let $\delta = 1$ if p is odd and $\delta = 2$ if p is even. Let us consider the map*

$$(5.1) \qquad (U, V, \Sigma) \in \mathbb{E}_{n\times q}^{m\times \ell} \to \qquad H(U, V, \Sigma) = \begin{pmatrix} U(I_\ell + \Omega)(I_\ell + \Theta) \\ V(I_q + \Lambda)(I_q + \Psi) \\ \Sigma + S \end{pmatrix} \in \mathbb{E}_{n\times q}^{m\times \ell}$$

where Ω, Λ are Hermitian matrices, S a diagonal matrix and Θ, Ψ are skew Hermitian matrices. Let $\Delta = U^*MV - \Sigma$ and $\Delta_1 = (I_\ell + \Theta^*)(I_\ell + \Omega)U^*MV(I_q + \Lambda)(I_q + \Psi) - \Sigma - S$. We said that H is a p-map if there exists quantities $a \geqslant 1$, $b \geqslant 0$, τ, ζ_1, ζ_2, α_1, α_2, α_0, α, ε be such that for all (U, V, Σ) satisfying $\max\left(\kappa^a K^b\|\Delta\|, \kappa^a K^{b+1}\|E_\ell(U)\|, \kappa^a K^{b+1}\|E_q(V)\|\right) \leqslant \varepsilon$ we have :

$$(5.2) \qquad ||E_\ell(U(I_\ell + \Omega))|| \leqslant \|E_\ell(U)\|^{p+1} \text{ and } \|E_q(V(I_q + \Lambda))\| \leqslant \|E_q(V)\|^{p+1}$$

$$(5.3) \qquad \kappa^a K^b\|\Delta_1\| \leqslant \tau\|\Delta\|^{p+1} \text{ and } \kappa^a K^b\|S\| \leqslant \alpha\|\Delta\|$$

$$(5.4) \qquad \begin{array}{l} \|I_\ell + \Theta\|^2, \quad \|I_q + \Psi\|^2 \leqslant \zeta_1 \\ \|(I_\ell + \Theta^*)(I_\ell + \Theta) - I_\ell\|, \quad \|(I_q + \Psi^*)(I_q + \Psi) - I_q\| \leqslant \dfrac{1}{\kappa^a K^{b+1}}\zeta_2\varepsilon^{p+\delta} \end{array}$$

$$(5.5) \qquad \|\Omega\|, \|\Lambda\| \leqslant \alpha_1\|\Delta\| \text{ and } \|\Theta\|, \|\Psi\| \leqslant \alpha_2\alpha_0\varepsilon.$$

We are proving that the theorems cited in the introduction result from the following
satement.

THEOREM 5.2. *Let an integer $p \geqslant 1$ and three reals $a \geqslant 1$, $b, \varepsilon \geqslant 0$. Let $\delta = 1$ if p is odd and $\delta = 2$ if p is even. Let us consider a p-map H as in (5.1). Let us consider a triplet (U_0, V_0, Σ_0) and define the sequence for $i \geqslant 0$, $(U_{i+1}, V_{i+1}, \Sigma_{i+1}) = H(U_i, V_i, \Sigma_i)$. Let $\Delta_i = U_i^*MV_i - \Sigma$, $K_i := K(\Sigma_i)$ and $\kappa_i = \kappa(\Sigma_i)$ with $K = K_0$ and $\kappa = \kappa_0$. Let us suppose*

$$(5.6) \qquad \max\left(\kappa^a K^b\|\Delta_0\|, \kappa^a K^{b+1}\|E_\ell(U_0)\|, \kappa^a K^{b+1}\|E_q(V_0)\|\right) \leqslant \varepsilon$$

$$(5.7) \qquad \frac{(1 + \alpha\varepsilon)^b}{(1 - 2\alpha\varepsilon)^a}(2\varepsilon)^p\tau \leqslant 1.$$

$$(5.8) \qquad (2\varepsilon)^p\frac{(1 + \alpha\varepsilon)^{b+1}}{(1 - 2\alpha\varepsilon)^a}(\zeta_1 + \zeta_2\varepsilon^{\delta-1}) \leqslant 1.$$

$$(5.9) \qquad 1 - 8\alpha\varepsilon > 0$$

where the quantities α, τ, ζ_1 and ζ_2 are as in Definition 5.1. Then the sequence $(U_i, V_i, \Sigma_i)_{i\geqslant 0}$ converge to an SVD of M and we have

$$(5.10) \qquad \max\left(\kappa_i^a K_i^b\|\Delta_i\|, \kappa_i^a K_i^{b+1}\|E_\ell(U_i)\|, \kappa_i^a K_i^{b+1}\|E_q(V_i)\|\right) \leqslant \varepsilon_i \leqslant 2^{-(p+1)^i+1}\varepsilon$$

$$(5.11) \qquad \|\Sigma_i - \Sigma_0\| \leqslant (2 - 2^{2-(p+1)^i})\frac{\alpha c}{\kappa}\varepsilon$$

where $c(1-4\alpha\varepsilon)=1$. *The inequality (5.11) implies* $K-2\alpha c\varepsilon \leqslant K_i \leqslant K+2\alpha c\varepsilon$ *and* $\frac{\kappa}{c} \leqslant \kappa_i \leqslant \frac{\kappa}{1-4\alpha c\varepsilon}$. *Morever if there exist positive constant* u_0 *such that* $\varepsilon \leqslant u_0$ *and* $2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)u_0 < 1$, *then by denoting* $\gamma = 2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)$ *and* $\sigma = 0.82 \times \alpha$ *we have*

$$\|U_i - U_\infty\| \leqslant 2^{-(p+1)^i+1}\sqrt{m}\frac{\gamma}{1-\gamma u_0}\varepsilon \tag{5.12}$$

$$\|V_i - V_\infty\| \leqslant 2^{-(p+1)^i+1}\sqrt{n}\frac{\gamma}{1-\gamma u_0}\varepsilon \tag{5.13}$$

$$\|\Sigma_i - \Sigma_\infty\| \leqslant 2^{-(p+1)^i+1}\sigma\varepsilon \tag{5.14}$$

Proof. Let us denote for each $i \geqslant 0$, $U_{i,1} = U_i(I_\ell + \Omega_i)$ and $U_{i+1} = U_{i,1}(I_\ell + \Theta_i)$ with similar notations for $V_{i,1}$ and V_{i+1}. Let $\Delta_i + \Sigma_i = U_i^* M V_i$, $\Sigma_{i+1} = \Sigma_i + S_i$ and also

$$\begin{array}{rclrcl} \varepsilon_0 & = & \varepsilon & \varepsilon_i & = & \max(\kappa_i^a K_i^b\|\Delta_i\|, \kappa_i^a K_i^{b+1}\|E_\ell(U_i)\|, \kappa_i^a K_i^{b+1}\|E_q(V_i)\|) \\ \kappa_0 & = & \kappa & \kappa_i & = & \kappa(\Sigma_i) \\ K_0 & = & K & K_i & = & K(\Sigma_i) \end{array}$$

We proceed by induction to prove (5.10-5.11). The property evidently hold for $i=0$. By assuming this for a given i, let us prove it for $i+1$. We first prove that $\|\Sigma_{i+1} - \Sigma_0\| \leqslant (2-2^{2-(p+1)^{i+1}})\frac{\alpha c}{\kappa}\varepsilon$ under the assumption $\|\Sigma_i - \Sigma_0\| \leqslant (2-2^{2-(p+1)^i})\frac{\alpha c}{\kappa}\varepsilon$ with $c = 1+4\alpha c\varepsilon$. From Lemma 5.3 we have $K-2\alpha c\varepsilon \leqslant K_i \leqslant K+2\alpha c\varepsilon$ and $\frac{\kappa}{c} \leqslant \kappa_i \leqslant \frac{\kappa}{1-4\alpha c\varepsilon} = \frac{1-4\alpha\varepsilon}{1-8\alpha\varepsilon}\kappa$. Using these bounds and assumption (5.3)it follows that

$$\begin{aligned} \|\Sigma_{i+1} - \Sigma_i\| = \|S_i\| &\leqslant \frac{1}{\kappa_i^a K_i^b}\alpha\varepsilon_i \\ &\leqslant \frac{c}{\kappa}2^{-(p+1)^i+1}\alpha\varepsilon \quad \text{since } a \geqslant 1 \quad K \geqslant 1 \text{ and } \kappa_i \geqslant \frac{\kappa}{c}. \end{aligned} \tag{5.15}$$

By applying the bound (5.15) we get

$$\begin{aligned} \|\Sigma_{i+1} - \Sigma_0\| &\leqslant \|S_i\| + \|\Sigma_i - \Sigma_0\| \\ &\leqslant 2^{1-(p+1)^i}\frac{1}{\kappa}\alpha c\varepsilon + (2-2^{2-(p+1)^i})\frac{1}{\kappa}\alpha c\varepsilon \\ &\leqslant (2-2^{1-(p+1)^i}(2-1))\frac{\alpha c}{\kappa}\varepsilon \\ &\leqslant (2-2^{-(p+1)^i})\frac{\alpha c}{\kappa}\varepsilon. \end{aligned}$$

But it is easy to see that $p \geqslant 1$ implies $2^{1-(p+1)^i} \geqslant 2^{2-(p+1)^{i+1}}$. Hence

$$\|\Sigma_{i+1} - \Sigma_0\| \leqslant (2-2^{2-(p+1)^{i+1}})\frac{\alpha c}{\kappa}\varepsilon.$$

Then inequality (5.11) holds for all i. From (5.3) we have $\|\Sigma_{i+1} - \Sigma_i\| = \|S_i\| \leqslant \frac{\alpha}{\kappa_i}\varepsilon_i$. We then deduce

$$K_i - \frac{\alpha}{\kappa_i}\varepsilon_i \leqslant K_{i+1} \leqslant \|\Sigma_i\| + \|\Sigma_{i+1} - \Sigma_i\| \leqslant K_i + \frac{\alpha}{\kappa_i}\varepsilon_i. \tag{5.16}$$

As in the proof of Lemma 5.3 we can obtain

$$\frac{\kappa_i}{1+2\alpha\varepsilon} \leqslant \kappa_{i+1} \leqslant \frac{\kappa_i}{1-2\alpha\varepsilon} \tag{5.17}$$

We now prove that $\kappa_{i+1}^a K_{i+1}^b \|\Delta_{i+1}\| \leqslant 2^{-2^{i+1}+1}\varepsilon$. Using both the assumption (5.3) and (5.16-5.17) it follows

$$\begin{aligned}
\kappa_{i+1}^a K_{i+1}^b \|\Delta_{i+1}\| &\leqslant \frac{(1+\alpha\varepsilon)^b}{(1-2\alpha\varepsilon)^a}\kappa_i^a K_i^b \tau \|\Delta_i\|^{p+1} \\
&\leqslant \frac{(1+\alpha\varepsilon)^b}{(1-2\alpha\varepsilon)^a}\tau\varepsilon_i^{p+1} \\
&\leqslant \frac{(1+\alpha\varepsilon)^b}{(1-2\alpha\varepsilon)^a}(2\varepsilon)^p\tau 2^{-(p+1)^{i+1}+1}\varepsilon \\
&\leqslant 2^{-(p+1)^{i+1}+1}\varepsilon \quad \text{since} \quad \frac{(1+\alpha\varepsilon)^b}{(1-2\alpha\varepsilon)^a}(2\varepsilon)^p\tau \leqslant 1 \quad \text{from} \quad (5.7)\,.
\end{aligned}$$

We now can bound $\|E_\ell(U_{i+1})\|$. We have

$$\begin{aligned}
\|E_\ell(U_{i+1})\| &\leqslant \|(I_\ell+\Theta_i^*)U_{i,1}{}^*U_{i,1}(I_\ell+\Theta_i)\| \\
&\leqslant \|(I_\ell+\Theta_i^*)E_\ell(U_{i,1})(I_\ell+\Theta_i)+(I_\ell+\Theta_i^*)(I_\ell+\Theta_i)-I_\ell\| \\
&\leqslant (1+\|\Theta_i\|)^2\|E_\ell(U_{i,1})\|+\|(I_\ell+\Theta_i^*)(I_\ell+\Theta_i)-I_\ell\|. \qquad (5.18)
\end{aligned}$$

From assumption (5.2)we know $\|E_\ell(U_{i,1})\| \leqslant \|E_\ell(U_i)\|^{p+1} \leqslant \dfrac{1}{\kappa_i^a K_i^{b+1}}\varepsilon_i^{p+1}$. It follows using both assumption (5.4), (5.22-5.16) that

$$\begin{aligned}
\kappa_{i+1}^a K_{i+1}^{b+1}\|E_\ell(U_{i+1})\| &\leqslant \frac{(1+\alpha\varepsilon)^{b+1}}{(1-2\alpha\varepsilon)^a}(\zeta_1\varepsilon_i^{p+1}+\zeta_2\varepsilon_i^{p+\delta}) \\
&\leqslant \frac{(1+\alpha\varepsilon)^{b+1}}{(1-2\alpha\varepsilon)^a}(2\varepsilon)^p(\zeta_1+\zeta_2\varepsilon^{\delta-1})2^{-(p+1)^{i+1}+1}\varepsilon \\
&\leqslant 2^{-(p+1)^{i+1}+1}\varepsilon \\
&\text{since} \quad \frac{(1+\alpha\varepsilon)^{b+1}}{(1-2\alpha\varepsilon)^a}(2\varepsilon)^p(\zeta_1+\zeta_2\varepsilon^{\delta-1}) \leqslant 1 \quad \text{from}\,(5.8)\,.
\end{aligned}$$

Hence $\kappa_{i+1}^a K_{i+1}^{b+1}\|E_\ell(U_{i+1})\| \leqslant 2^{-(p+1)^{i+1}+1}\varepsilon$. In the same way $\kappa_{i+1}^a K_{i+1}^{b+1}\|E_q(V_{i+1})\| \leq 2^{-2^{i+1}+1}\varepsilon$. Hence we have shown that $\varepsilon_{i+1} \leqslant 2^{-2^{i+1}+1}\varepsilon$. This completes the proof of (5.10–5.11).

By applying Lemma 3.6 we conclude that the sequences $(U_i)_{i\geqslant 0}$ and $(V_i)_{i\geqslant 0}$ converges respectively towards U_∞ and V_∞ which are two unitary matrices since $\|E_\ell(U_i)\|, \|E_q(V_i) \leqslant 2^{-2^i+1}\varepsilon$. Hence the bounds (5.12-5.13) hold. Finally the bound

(5.14) follows from

$$\begin{aligned}
\|\Sigma_{i+j} - \Sigma_i\| &\leqslant \sum_{k=i}^{i+j-1} \|\Sigma_{k+1} - \Sigma_k\| \\
&\leqslant \sum_{k\geqslant i} 2^{-(p+1)^k+1}\alpha\varepsilon \\
&\leqslant \left(\sum_{k\geqslant 0} 2^{-(p+1)^k}\right) 2^{-(p+1)^i+1}\alpha\varepsilon \\
&\leqslant 2^{-(p+1)^i+1} \times 0.82\alpha\varepsilon \quad \text{since} \quad \sum_{k\geqslant 0} 2^{-(p+1)^k} \leqslant \sum_{k\geqslant 3} 2^{-2^k} \leqslant 0.82.
\end{aligned}$$

Hence the sequence $(\Sigma_i)_{i\geqslant 0}$ admits a limit Σ_∞. The triplet $(U_\infty, V_\infty, \Sigma_\infty)$ is a solution of SVD system (1.1). The theorem is proved. ∎

Lemma 5.3. *Using the notations and asumptions of the proof of Theorem 5.2 we have with $c = 1 + 4\alpha c\varepsilon$:*

$$\begin{aligned}
K - 2\alpha c\varepsilon \leqslant K_i \leqslant K + 2\alpha c\varepsilon \\
\frac{\kappa}{c} \leqslant \kappa_i \leqslant \frac{\kappa}{1 - 4\alpha c\varepsilon}
\end{aligned}$$

Proof. Let us prove that $K_i \leqslant K + 2\alpha\varepsilon$. We have

$$\begin{aligned}
K_i := \|\Sigma_i\| &\leqslant \|\Sigma_0\| + \|\Sigma_i - \Sigma_0\| \\
&\leqslant K + (2 - 2^{-(p+1)^i+1})\frac{\alpha c}{\kappa}\varepsilon \\
&\leqslant K + 2\alpha c\varepsilon \quad \text{since} \quad \kappa \geqslant 1.
\end{aligned}$$

In the same way $K_i \geqslant K - 2\alpha c\varepsilon$. We have also $\kappa_i \leqslant \dfrac{\kappa}{1 - 4\alpha c\varepsilon}$. In fact, if $\sigma_{i,j}$'s be the diagonal values of Σ_i, the Weyl's bound [44] implies that

$$|\sigma_{i,j} - \sigma_{0,j}| \leqslant \|\Sigma_i - \Sigma_0\| \leqslant 2\frac{\alpha c}{\kappa}\varepsilon \qquad 1 \leqslant j \leqslant n, \tag{5.19}$$

and

$$K - 2\frac{\alpha c}{\kappa}\varepsilon \leqslant \sigma_{i,j} \leqslant K + 2\frac{\alpha c}{\kappa}\varepsilon \qquad 1 \leqslant j \leqslant n.$$

Hence, since $\kappa, K \geqslant 1$ we get

$$\frac{\kappa}{1 + 2\alpha c\varepsilon} \leqslant \sigma_{i,j}^{-1} \leqslant \frac{\kappa}{1 - 2\alpha c\varepsilon} \tag{5.20}$$

Moreover for $1 \leqslant j < k \leqslant n$, we have :

$$\begin{aligned}
|\sigma_{i,k} \pm \sigma_{i,j}| &\geqslant |\sigma_{0,k} \pm \sigma_{0,j}| - |\sigma_{i,k} - \sigma_{0,k}| - |\sigma_{i,j} - \sigma_{0,j}| \\
&\geqslant |\sigma_{0,k} \pm \sigma_{0,j}| \left(1 - \frac{1}{\kappa|\sigma_{0,k} \pm \sigma_{0,j}|} 4\alpha c\varepsilon\right) \quad \text{from} \quad (5.19) \\
&\geqslant |\sigma_{0,k} \pm \sigma_{0,j}|(1 - 4\alpha c\varepsilon) = |\sigma_{0,k} \pm \sigma_{0,j}| \frac{1 - 8\alpha\varepsilon}{1 - 4\alpha\varepsilon} > 0 \\
&\qquad \text{since}\, \kappa|\sigma_{0,k} \pm \sigma_{0,j}| \geqslant 1 \text{ and } (5.9)
\end{aligned} \tag{5.21}$$

Taking in account the definition of κ and the inequalities (5.20), (5.21), we then get

$$\begin{aligned}\kappa_i &= \max\left(1, \max_j \frac{1}{\sigma_{i,j}}, \max_{k\neq j}\left(\frac{1}{|\sigma_{i,k}-\sigma_{i,j}|}+\frac{1}{|\sigma_{i,k}+\sigma_{i,j}|}\right)\right)\\ &\leqslant \kappa \max\left(\frac{1}{1-2\alpha c\varepsilon}, \frac{1}{1-4\alpha c\varepsilon}\right)\\ &\leqslant \frac{\kappa}{1-4\alpha c\varepsilon} = \frac{1-4\alpha\varepsilon}{1-8\alpha\varepsilon}.\end{aligned}$$

In the same way we have

$$\begin{aligned}|\sigma_{i,k}\pm\sigma_{i,j}| &\leqslant |\sigma_{0,k}\pm\sigma_{0,j}| + |\sigma_{i,k}-\sigma_{0,k}| + |\sigma_{i,j}-\sigma_{0,j}|\\ &\leqslant |\sigma_{0,k}\pm\sigma_{0,j}|(1+4\alpha c\varepsilon) = |\sigma_{0,k}\pm\sigma_{0,j}|c.\end{aligned}$$

We deduce that

$$\kappa_i \geqslant \frac{\kappa}{c} = (1-4\alpha\varepsilon)\kappa. \tag{5.22}$$

The Lemma is proved. ∎

6. Proof of Theorem 1.2 : case $p = 1$.

Let

$$s = \left(1+\frac{1}{2}\varepsilon\right)^2 + 1 + \frac{1}{4}\varepsilon, \quad \tau = (3+s\varepsilon)s^2, \qquad a = 2, \quad b = 1, \quad u_0 = 0.0289.$$

It consists to verify the assumptions of Theorem 5.2. Remember that (5.6) is satisfied from assumption since

$$\max\left(\kappa^a K^{b+1}\|E_\ell(U)\|, \kappa^a K^{b+1}\left\|E_q(V)\right\|, \kappa^a K^b\|\Delta\|\right) \leqslant \varepsilon$$

where U, V, Δ stand for U_0, V_0, Δ_0 respectively. The item (5.2) follows of Proposition 3.2 since $\Omega = -\frac{1}{2}E_\ell(U)$ and $\Lambda = -\frac{1}{2}E_q(V)$. Let us prove the item (5.3). To do that we denote $\Delta_{0,1} = (I_\ell+\Omega)(\Delta+\Sigma)(I_q+\Lambda) - \Sigma$ and $\varepsilon_{0,1} = \|\Delta_{0,1}\|$. From Proposition 3.2 and $\|E_\ell(U)\|, \|E_q(V)\| \leqslant \frac{\varepsilon}{\kappa^a K^{b+1}}$ we know that $\|\Omega\|, \|\Lambda\| \leqslant \frac{1}{2\kappa^a K^{b+1}}\varepsilon$. We then apply Proposition 6.1 with $w = \frac{1}{2}$ to get

$$\begin{aligned}\varepsilon_{0,1} &\leqslant \left(\left(1+\frac{1}{2}\varepsilon\right)^2 + 1 + \frac{1}{4}\varepsilon\right)\frac{\varepsilon}{\kappa^a K^b}\\ &\leqslant \frac{s\varepsilon}{\kappa^a K^b}. \qquad (6.1)\end{aligned}$$

From Lemma 4.3 we have $\|X\|, \|Y\| \leqslant \kappa\varepsilon_{0,1}$. We deduce that the quantity

$$\begin{aligned}\Delta_1 &= (I_\ell - X)(\Delta_{0,1}+\Sigma)(I_q+Y) - \Sigma - S\\ &= -X\Delta_{0,1} + \Delta_{0,1}Y - X\Delta_{0,1}Y - X\Sigma Y \quad \text{since} \quad \Delta_{0,1} - S - X\Sigma + \Sigma Y = 0,\end{aligned}$$

can be bounded by

$$\begin{aligned}\|\Delta_1\| &\leqslant 2\kappa\varepsilon_{0,1}^2 + \kappa^2\varepsilon_{0,1}^3 + \kappa^2 K\varepsilon_{0,1}^2\\ &\leqslant \left(\frac{2}{\kappa^3K^2} + \frac{s\varepsilon}{\kappa^4K^3} + \frac{1}{\kappa^2K}\right)s^2\varepsilon^2 \text{ since } \kappa, K \geqslant 1 \text{ and } \varepsilon_{0,1} \leqslant \frac{s\varepsilon}{\kappa^2 K} \text{ from (6.1)}.\\ &\leqslant \frac{1}{\kappa^2K}(3+s\varepsilon)s^2\varepsilon^2 = \frac{1}{\kappa^2K}\tau\varepsilon^2\end{aligned}$$

On the other hand $S = \mathrm{diag}(\Delta_{0,1})$. It follows $\|S\| \leqslant \varepsilon_{0,1} \leqslant \frac{s\varepsilon}{\kappa^2 K}$. The quantity α of Definition 5.1 is equal to s. This allows to prove the assumption (5.7) that is

$$\begin{aligned} 2\varepsilon \frac{1+s\varepsilon}{(1-2s\varepsilon)^2}\tau &\leqslant 2\frac{1+s\varepsilon}{(1-2s\varepsilon)^2}(3+s\varepsilon)s^2\varepsilon \\ &\leqslant 1 \quad \text{since} \quad \varepsilon \leqslant u_0 = 0.0289. \end{aligned}$$

We now prove the item (5.4). We have

$$\begin{aligned} \|I_\ell + \Theta\|^2 &\leqslant (1+\|X\|)^2 \\ \|(I_\ell - X)(I_\ell + X) - I_\ell\| &= \|X\|^2. \end{aligned}$$

Using Lemma 9.4 we know that $\|X\| \leqslant \kappa\varepsilon_{0,1} \leqslant \frac{s\varepsilon}{\kappa^{a-1}K^b}$. We deduce that

$$\begin{aligned} (1+\|X\|)^2 &\leqslant (1+s\varepsilon)^2 = \zeta_1 \\ \|(I_\ell - X)(I_\ell + X) - I_\ell\| &\leqslant \frac{\zeta_2\varepsilon^2}{\kappa^{2a-2}K^{2b}} \quad \text{where}\, \zeta_2 = s^2. \\ &\leqslant \frac{1}{\kappa^a K^{b+1}}\zeta_2\varepsilon^2 \quad \text{since}\, a = 2 \text{ and } b = 1. \end{aligned}$$

This allows to prove the assumption (5.8) that is

$$\begin{aligned} &(2\varepsilon)\frac{(1+s\varepsilon)^2}{(1-2s\varepsilon)^2}(\zeta_1 + \zeta_2\varepsilon^{\delta-1}) \\ &\qquad \leqslant 2\frac{(1+s\varepsilon)^2}{(1-2s\varepsilon)^2}((1+s\varepsilon)^2 + s^2)\varepsilon \quad \text{since} \quad p = 1\,\text{implies} \quad \delta = 1 \\ &\qquad \leqslant 0.443 \quad \leqslant 1 \quad \text{since} \quad u \leqslant u_0. \end{aligned}$$

Finally $1 - 8s\varepsilon \geqslant 0.46 > 0$. This proves the item (5.9).

We now verify the assumption (5.5). We have seen that $\|\Omega\|, \|\Lambda\| \leqslant \frac{1}{2}\varepsilon$. Hence $\alpha_1 = \frac{1}{2}$. On the other hand one has $\Theta = X$ and $\Psi_i = Y$. From $\|X\|, \|Y\| \leqslant s\varepsilon \leqslant 2.042\varepsilon$ since $u \leqslant u_0$, we can take $\alpha_2 = 2.042$. Since $\gamma u_0 = 2(\alpha_1 + \alpha_2 + \alpha_1\alpha_2 u_0)u_0 < 0.15$ then the bounds (5.12-5.14) of Theorem 5.2 hold with

$$\begin{aligned} \gamma &= 5.14 \\ \frac{\gamma}{1-\gamma u_0} &\leqslant 6.1 \\ \sigma = 0.82s &\leqslant 1.67. \end{aligned}$$

The Theorem 1.2 is proved in the case $p = 1$.□

Proposition 6.1. *Let $\varepsilon \geqslant 0$ and $a, b > 0$. Let $\Delta_1 = (I_\ell + \Omega)(\Delta + \Sigma)(I_q + \Lambda) - \Sigma$ with $\Omega^* = \Omega$. Let us suppose $\|\Delta\| \leqslant \frac{\varepsilon}{\kappa^a K^b}$ and $\|\Omega\|, \|\Lambda\| \leqslant \frac{w\varepsilon}{\kappa^a K^{b+1}}$ with $\kappa = \kappa(\Sigma)$ and $K = K(\Sigma)$. We have*

$$\|\Delta_1\| \leqslant \left((1+w\varepsilon)^2 + 2w + w^2\varepsilon\right)\frac{\varepsilon}{\kappa^a K^b}.$$

Proof. We have $\Omega^* = \Omega$. A straightforward calculation shows that

$$\begin{aligned}\Delta_1 &= (I_\ell + \Omega)\Delta(I_q + \Lambda) + (I_\ell + \Omega)\Sigma(I_q + \Lambda) - \Sigma \\ &= (I_\ell + \Omega)\Delta(I_q + \Lambda) + \Omega\Sigma + \Sigma\Lambda + \Omega\Sigma\Lambda.\end{aligned}$$

Bounding $\|\Delta_1\|$ we get

$$\begin{aligned}\|\Delta_1\| &\leqslant \left(1 + \frac{w\varepsilon}{\kappa^a K^{b+1}}\right)^2 \frac{\varepsilon}{\kappa^a K^b} + 2\frac{w\varepsilon}{\kappa^a K^b} + \left(\frac{w\varepsilon}{\kappa^a K^{b+1}}\right)^2 K \\ &\leqslant \left((1 + w\varepsilon)^2 + 2w + w^2\varepsilon\right) \frac{\varepsilon}{\kappa^a K^b} \quad \text{since} \quad \kappa, K \geqslant 1.\end{aligned}$$

The proposition is proved. ∎

7. Proof of Theorem 1.2 : case $p = 2$. Let us introduce some constants and quantities.

$$\begin{aligned} & w = \frac{1}{2}\left(1 + \frac{3}{4}\varepsilon\right), \quad s = (1 + w\varepsilon)^2 + 2w + w^2\varepsilon, \\ & a = \frac{4}{3}, \qquad b = \frac{1}{3}, \quad u_0 = 0.046. \end{aligned} \tag{7.1}$$

We also introduce

$$\tau_1 = 2 + 2\varepsilon + \frac{5}{4}\varepsilon^2 + \frac{1}{4}\varepsilon^3$$

$$\begin{aligned}\tau_2 = 3 &+ \frac{1}{2}(11 + 2\tau_1)\varepsilon + \frac{1}{2}(8 + 7\tau_1)\varepsilon^2 + \frac{1}{2}(2 + 7\tau_1 + \tau_1^2)\varepsilon^3 \\ &+ \frac{1}{2}(3 + 2\tau_1)\tau_1\varepsilon^4 + \tau_1^2\varepsilon^5 + \frac{1}{4}\tau_1^3\varepsilon^6\end{aligned} \tag{7.2}$$

$$\tau = \tau_1\tau_2 \tag{7.3}$$

$$\alpha = (1 + \tau_1(s\varepsilon)s\varepsilon)s$$

Let us verify the assumptions of Theorem 5.2. The item (5.2) follows of Proposition 3.2 since $\Omega = s_2(E_\ell(U))$ and $\Lambda = s_2(E_q(V))$. Let us prove the item (5.3). We first bound $\|\Delta_1\|$ where $\Delta_1 = U_1^* M V - \Sigma_1$. We use the $\Delta_{0,i}$, $1 \leqslant i \leqslant 3$, the quantities defined by the formulas (1.10-1.11). By definition of the map H_2, we have $\Delta_1 = \Delta_{0,3}$. We introduce the quantities $\varepsilon_{0,i} = \|\Delta_{0,i}\|$. From Proposition 3.2 in the case $p = 2$ and assumption $||E_\ell(U)||, \|E_q(V)\| \leqslant \dfrac{\varepsilon}{\kappa^a K^{b+1}}$ we know that $\|\Omega\|, \|\Lambda\| \leqslant \dfrac{w}{\kappa^a K^{b+1}}\varepsilon$ with $w = \dfrac{1}{2}\left(1 + \dfrac{3}{4}\varepsilon\right)$. We then apply Proposition 6.1 to get

$$\begin{aligned}\varepsilon_{0,1} &\leqslant ((1 + w\varepsilon)^2 + 2w + w^2\varepsilon)\frac{\varepsilon}{\kappa^a K^b} \\ &\leqslant \frac{s\varepsilon}{\kappa^a K^b} \qquad \text{from} \quad (7.1)\,.\end{aligned} \tag{7.4}$$

From Proposition 7.1 we can write

$$\|\Delta_1\| = \|\Delta_{0,3}\| \leqslant \frac{1}{\kappa^{4/3}K^{1/3}}\tau(s\varepsilon)s^3\varepsilon^3.$$

We now bound the norm of $S = S_1 + S_2$. We have always from Proposition 7.1

$$\|S\| \leqslant \|\Delta_{0,1}\| + \|\Delta_{0,2}\| \leqslant \frac{1}{\kappa^{4/3}K^{1/3}}(1+\tau_1(s\varepsilon)s\varepsilon)s\varepsilon = \frac{1}{\kappa^{4/3}K^{1/3}}\alpha\varepsilon. \tag{7.5}$$

A numerical computation shows that the inequality $(2\varepsilon)^2\dfrac{(1+\alpha\varepsilon)^{1/3}}{(1-2\alpha\varepsilon)^{4/3}}\tau(s\varepsilon)s^3 \leqslant 1$ is verified for all $u \leqslant u_0$. Then the assumption (5.7) holds.

We now prove the item (5.4). We have

$$\|I_\ell + \Theta\|^2 \leqslant (1+||c_2(X)||)^2$$
$$\|(I_\ell+\Theta^*)(I_\ell+\Theta) - I_\ell\| \leqslant (1+c_2(-\|X\|))(1+c_2(\|X\|)) - 1$$

From the bound (7.5) we deduce that $\|X\| \leqslant \|X_1\| + \|X_2\| \leqslant \dfrac{\kappa x}{\kappa^{4/3}K^{1/3}} = \dfrac{x}{\kappa^{1/3}K^{1/3}}$ with $x = \alpha\varepsilon$. On the other hand $c_2(u) = u + \dfrac{1}{2}u^2$ and $(1+c_2(-u))(1+c_2(u)) - 1 = \dfrac{u^4}{4}$. It follows :

$$\|I_\ell+\Theta\|^2 \leqslant \left(1+x+\frac{1}{2}x^2\right)^2 = \zeta_1$$
$$\|(I_\ell+\Theta^*)(I_\ell+\Theta) - I_\ell\| \leqslant \frac{1}{4\kappa^{4/3}K^{4/3}}(\alpha\varepsilon)^4 = \frac{1}{\kappa^{4/3}K^{4/3}}\zeta_2\varepsilon^4 \quad \text{where} \quad \zeta_2 = \frac{1}{4}\alpha^4\varepsilon^4.$$

We now prove a part of assumption (5.8) that is $(2\varepsilon)^2\dfrac{(1+\alpha\varepsilon)^{4/3}}{(1-2\alpha\varepsilon)^{4/3}}(\zeta_1+\zeta_2\varepsilon) \leqslant 1$. We have

$$(2\varepsilon)^2\frac{(1+\alpha\varepsilon)^{4/3}}{(1-2\alpha\varepsilon)^{4/3}}(\zeta_1+\zeta_2\varepsilon) \leqslant 0.025 \qquad \text{since} \quad u \leqslant u_0.$$

This proves the item (5.8). The item 5.9 holds since $1 - 8\alpha\varepsilon \geqslant 0.05 > 0$ when $\varepsilon \leqslant u_0$.

Let us prove the assumption (5.5). Using $\varepsilon \leqslant u_0$ we have $\|\Omega\|, \|\Lambda\| \leqslant w\varepsilon \leqslant \alpha_1\varepsilon$ with $\alpha_1 = 0.52$ and $\|\Theta\|, \|\Psi\| \leqslant (1+x/2)\alpha\varepsilon \leqslant \alpha_2\varepsilon$ with $\alpha_2 = 2.7$ Moreover

$$2(\alpha_1+\alpha_2+\alpha_1\alpha_2u_0)u_0 \leqslant 0.304 < 1$$

Then the bounds (5.12-5.14) of Theorem 5.2 hold with

$$\gamma = 6.56$$
$$\frac{\gamma}{1-\gamma u_0} \leqslant 9.41$$
$$\sigma = 0.82\alpha \leqslant 2.1.$$

The Theorem 1.2 is proved for $p = 2$. □

Proposition 7.1. *Let $p = 2$, $\varepsilon \geqslant 0$. Let us consider $\Delta_1 = U_1^*MV_1 - \Sigma$ such that $\|\Delta_1\| = \varepsilon_1 \leqslant \dfrac{\varepsilon}{\kappa^{4/3}K^{1/3}}$ where $\kappa = \kappa(\Sigma)$ and $K = K(\Sigma)$. Let us consider $\tau_1 := \tau_1(\varepsilon)$ and $\tau := \tau(\varepsilon)$ as in (7.3) Then we have*

$$\|\Delta_2\| \leqslant \frac{1}{\kappa^{4/3}K^{1/3}}\tau_1\varepsilon^2,$$
$$\tau_3 := \|\Delta_3\| \leqslant \frac{1}{\kappa^{4/3}K^{1/3}}\tau\varepsilon^3,$$

where $\Delta_2 = (I_\ell+\Theta_1^*)(\Delta_1+\Sigma)(I_q+\Psi_1) - \Sigma - S_1$ and $\Delta_3 = (I_\ell+\Theta_2^*)(\Delta_1+\Sigma)(I_q+\Psi_2) - \Sigma - S_1 - S_2$ with Θ_2 and Ψ_2 are defined by the formulas (1.11) for $p = 2$.

Proof. We denote $e_2(X) = X^2/2$, $\Theta_1 = X_1 + e_2(X_1)$ and $\Psi_1 = Y_1 + e_2(Y_1)$. Remember $\Delta_1+\Sigma = U^*\Sigma V$ and $\Delta_2 = (I_\ell+\Theta_1^*)(\Delta_1+\Sigma)(I_q+\Psi_1)-\Sigma-S_1$. Expanding Δ_2 we find

$$\begin{aligned}
\Delta_2 &= \Delta_1 - S_1 - X_1\Sigma + \Sigma Y_1 - X_1\Sigma Y_1 + \frac{1}{2}X_1^2\Sigma + \Sigma\frac{1}{2}Y_1^2 + \frac{1}{4}X_1^2\Sigma Y_1^2 \\
&\quad + \frac{1}{2}X_1^2\Sigma Y_1 - \frac{1}{2}X_1\Sigma Y_1^2 - X_1\Delta_1 + \Delta_1 Y_1 - X_1\Delta_1 Y_1 + \frac{1}{2}X_1^2\Delta_1 + \frac{1}{2}\Delta_1 Y_1^2 \\
&\quad + \frac{1}{4}X_1^2\Delta_1 Y_1^2 + \frac{1}{2}X_1^2\Delta_1 Y_1 - \frac{1}{2}X_1\Delta_1 Y_1^2 \\
&= \frac{1}{2}(X_1(-\Sigma Y_1 + X_1\Sigma) + (-X_1\Sigma + \Sigma Y_1)Y_1) + \frac{1}{4}X_1^2\Sigma Y_1^2 \\
&\quad + \frac{1}{2}X_1(X_1\Sigma - \Sigma Y_1)Y_1 - X_1\Delta_1 + \Delta_1 Y_1 - X_1\Delta_1 Y_1 + \frac{1}{2}X_1^2\Delta_1 + \frac{1}{2}\Delta_1 Y_1^2 \\
&\quad e + \frac{1}{4}X_1^2\Delta_1 Y_1^2 + \frac{1}{2}X_1^2\Delta_1 Y_1 + \frac{1}{2}X_1\Delta_1 Y_1^2 \\
&= \frac{1}{2}(X_1(-\Delta_1 - S_1) + (S_1 + \Delta_1)Y_1) + \frac{1}{4}X_1^2\Sigma Y_1^2 + \frac{1}{2}X_1(-\Delta_1 - S1)Y_1 \qquad (7.6)\\
&\quad + \frac{1}{2}X_1^2\Delta_1 + \frac{1}{2}\Delta_1 Y_1^2 + \frac{1}{4}X_1^2\Delta_1 Y_1^2 + \frac{1}{2}X_1^2\Delta_1 Y_1 - \frac{1}{2}X_1\Delta_1 Y_1^2.
\end{aligned}$$

We know that $\|\Delta_1\| \leqslant \varepsilon_1$. From the formula (7.6) we deduce

$$\begin{aligned}
\|\Delta_2\| &\leqslant 2\kappa\varepsilon_1^2 + \frac{1}{4}\kappa^4 K\varepsilon_1^4 + 2\kappa^2\varepsilon_1^3 + \frac{1}{4}\kappa^4\varepsilon_1^5 + \kappa^3\varepsilon_1^4 \\
&\leqslant q_1\varepsilon_1^2 \quad \text{with} \quad q_1 = 2\kappa + 2\kappa^2\varepsilon_1 + \frac{5}{4}\kappa^4 K\varepsilon_1^2 + \frac{1}{4}\kappa^4\varepsilon_1^3 \qquad (7.7)
\end{aligned}$$

Since $\varepsilon_1 \leqslant \dfrac{\varepsilon}{\kappa^{4/3}K^{1/3}}$ it follows $q_1\varepsilon_1 \leqslant \tau_1\varepsilon$ with $\tau_1 = 2 + 2\varepsilon + \frac{5}{4}\varepsilon^2 + \frac{1}{4}\varepsilon^3$. Hence we have obtained $\|\Delta_2\| \leqslant \tau_1\dfrac{\varepsilon^2}{\kappa^{4/3}K^{1/3}}$.

From definition $\Theta_2 = c_2(X_1 + X_2)$. Hence we can write $\Theta_2 = \Theta_1 + X_2 + A_2$ with

$$\begin{aligned}
A_2 := A_2(X_1, X_2) &= c_2\,(X_1{+}X_2) - c_2(X_1) - X_2 \\
&= \frac{1}{2}((X_1 + X_2)^2 - X_1^2) \\
&= \frac{1}{2}(X_2^2 + X_1X_2 + X_2X_1)
\end{aligned}$$

In the same way $\Psi_2 = \Psi_1{+}Y_2{+}B_2$ where $B_2 = A_2(Y_1, Y_2)$. Expanding $(I_\ell{+}\Theta_2^*)(\Delta_1{+}\Sigma)(I_q + \Psi_2)$ we get

$$\begin{aligned}
\Delta_3 &= (I_\ell + \Theta_2^*)(\Delta_1 + \Sigma)(I_q + \Psi_2) - \Sigma - S_1 - S_2 \\
&= (I_\ell + \Theta_1^* - X_2 + A_2)(\Delta_1 + \Sigma)(I_q + \Psi_1 + Y_2 + B_2) - \Sigma - S_1 - S_2 \\
&= (I_\ell + \Theta_1^*)(\Delta_1 + \Sigma)(I_q + \Psi_1) - \Sigma - S_1 - S_2 + (I_\ell + \Theta_1^*)(\Delta_1 + \Sigma)(Y_2 + B_2) \\
&\quad + (-X_2 + A_2)(\Delta_1 + \Sigma)(I_q + \Psi_1) + (-X_2 + A_2)(\Delta_1 + \Sigma)(Y_2 + B_2)
\end{aligned}$$

We know that

$$(I_\ell + \Theta_1^*)(\Delta_1 + \Sigma)(I_q + \Psi_1) - \Sigma - S_1 - S_2 = \Delta_2 - S_2 - X_2\Sigma + \Sigma Y_2 = 0.$$

Expanding more Δ_3, we then can write by grouping the terms appropriately :

$$\Delta_3 = -X_2\Delta_1 Y_2 + \Delta_1 B_2 + A_2\Delta_1 - X_2\Delta_1 B_2 + A_2\Delta_1 Y_2 + A_2\Delta_1 B_2 \tag{7.8}$$
$$+ \Theta_1^*\Delta_1 Y_2 - X_2\Delta_1\Psi_1 + \Theta_1^*\Delta_1 B_2 + A_2\Delta_1\Psi_1 \tag{7.9}$$
$$+ G,$$

where $G = -X_2\Delta_1 + \Delta_1 Y_2 - X_2\Sigma Y_2 + \Sigma B_2 + A_2\Sigma + \Theta_1^*\Sigma Y_2 - X_2\Sigma\Psi_1 + \Theta_1^*\Sigma B_2 + A_2\Sigma\Psi_1 - X_2\Sigma B_2 + A_2\Sigma Y_2 + A_2\Sigma B_2$. The Lemma 7.2 modifies the quantity as sum of the following G_i's :

$$G_1 = \frac{1}{2}X_2(\Delta_2 - S_2) + \frac{1}{2}(S_2 - \Delta_2)Y_2 \tag{7.10}$$
$$G_2 = \frac{1}{2}(X_1(\Delta_2 - S_2) + (S_2 - \Delta_2)Y_1) + \frac{1}{2}(X_2(-\Delta_1 - S_1) + (S_1 + \Delta_1)Y_2) \tag{7.11}$$
$$G_3 = \frac{1}{2}\left(X_1(\Delta_2 - S_2)Y_1 + X_2(\Delta_1 - S_1)Y_2 + X_1\left(\Delta_2 - S_2\right)Y_2\right) \tag{7.12}$$
$$+ \frac{1}{2}(X_2(\Delta_1 - S_1)Y_1 + X_1(\Delta_1 - S_1)Y_2 + X_2(\Delta_2 - S_2)Y_1)$$
$$G_4 = \frac{1}{2}X_2(S_2 - \Delta_2)Y_2 \tag{7.13}$$
$$G_5 = e_2(X_1)\Sigma R_{2,1} + Q_{2,1}\Sigma e_2(Y_1) + e_2(X_1)\Sigma e_2(Y_2) + e_2(X_2)\Sigma e_2(Y_1) \tag{7.14}$$

where $Q_{2,1} = \frac{1}{2}(X_1X_2 + X_2X_1)$ and $R_{2,1} = \frac{1}{2}(Y_1Y_2 + Y_2Y_1)$. We are going to prove $\|\Delta_3\| \leqslant q_1q_2\varepsilon_1^3$ where q_2 is defined below in (7.16). To do that we will use the bounds

1. $\|X_1\|, \|Y_1\| \leqslant \kappa\varepsilon_1$, $\|\Delta_2\| \leqslant q_1\varepsilon_1^2$ and
$$\|X_2\|, \|Y_2\| \leqslant \kappa q_1\varepsilon_1^2. \tag{7.15}$$
2. $\|\Theta_1\|, \|\Psi_1\| \leqslant \left(1 + \frac{1}{2}\kappa\varepsilon_1\right)\kappa\varepsilon_1$.
3. $\|Q_{2,1}\|, \|R_{2,1}\| \leqslant q_1\kappa^2\varepsilon_1^3$.
4. $\|A_2\|, \|B_2\| \leqslant \frac{1}{2}(q_1^2\kappa^2\varepsilon_1^4 + 2q_1\kappa^2\varepsilon_1^3) = \frac{1}{2}(q_1\varepsilon_1 + 2)q_1\kappa^2\varepsilon_1^3$.

Considering the bounds of the norms of matrices given in (7.8-7.14), we get

$$\begin{aligned}
&\frac{1}{q_1\varepsilon_1^3}\|\Delta_3\| \\
&\leqslant \frac{1}{4}q_1^3\kappa^4\varepsilon_1^6 + q_1^2\kappa^4\varepsilon_1^5 + (\kappa + q_1)q_1\kappa^3\varepsilon_1^4 + 2\kappa^3q_1\varepsilon_1^3 + 2\kappa^2q_1\varepsilon_1^2 + 2\kappa^2\varepsilon_1 \quad \text{from (7.8)} \\
&\quad + \frac{1}{2}\kappa^4q_1\varepsilon_1^4 + \kappa^3(\kappa + q_1)\varepsilon_1^3 + 3\kappa^3\varepsilon_1^2 + 2\kappa^2\varepsilon_1 \quad \text{from (7.9)} \\
&\quad + \kappa q_1\varepsilon_1 + 3\kappa + \frac{3}{2}\kappa^2q_1\varepsilon_1^2 + \frac{3}{2}\kappa^2\varepsilon_1 + \frac{1}{2}\kappa^2q_1^2\varepsilon_1^3 \quad \text{from (7.10-7.13)} \\
&\quad + \frac{1}{2}\kappa^4Kq_1\varepsilon_1^3 + \kappa^4K\varepsilon_1^2. \quad \text{from (7.14)}
\end{aligned}$$

Collecting the previous bound we get $||\Delta_3|| \leqslant q_2q_1\varepsilon_1^3$ where

$$\begin{aligned}
q_2 = 3\kappa &+ \frac{1}{2}(11\kappa + 2q_1)\kappa\varepsilon_1 + \frac{1}{2}(2\kappa^2K + 6\kappa + 7q_1)\kappa^2\varepsilon_1^2 \\
&+ \frac{1}{2}(q_1\kappa^2K + 2\kappa^2 + 6\kappa q_1 + q_1^2)\kappa^2\varepsilon_1^3 + \frac{1}{2}(3\kappa + 2q_1)q_1\kappa^3\varepsilon_1^4 \\
&+ q_1^2\kappa^4\varepsilon_1^5 + \frac{1}{4}q_1^3\kappa^4\varepsilon_1^6.
\end{aligned} \tag{7.16}$$

Now we are bounding $q_2\varepsilon_1$. We remark that the monomials which appears in $q_2\varepsilon_1$ are of the form $q_1^i\kappa^jK^k\varepsilon_1^{i+l}$ for some $(i,j,k,l)\in\mathbb{N}^4$ such that $i\geqslant 0$, $3j\leqslant 4l$ and $3k\leqslant l$. Since $\varepsilon_1\leqslant\dfrac{\varepsilon}{\kappa^{4/3}K^{1/3}}$ and $q_1\varepsilon_1\leqslant\tau_1\varepsilon$ the we have :

$$\begin{aligned} q_1^i\kappa^jK^k\varepsilon_1^{i+l} &\leqslant (\tau_1\varepsilon)^i\kappa^{j-4l/3}K^{k-l/3}\varepsilon^l\\ &\leqslant \tau_1^i\varepsilon^{i+l} \qquad \text{since}\, \kappa, K\geqslant 1. \end{aligned}$$

From the expression of q_2 it follows after straightforward calculation that $q_2\varepsilon_1\leqslant\tau_2\varepsilon$ where

$$\begin{aligned} \tau_2 = 3 &+ \frac{1}{2}(11+2\tau_1)\varepsilon+\frac{1}{2}(8+7\tau_1)\varepsilon^2+\frac{1}{2}(\tau_1^2+7\tau_1+2)\varepsilon^3\\ &+\frac{1}{2}(3+2\tau_1)\tau_1\varepsilon^4+\tau_1^2\varepsilon^5+\frac{1}{4}\tau_1^3\varepsilon^6. \end{aligned}$$

Since we also have $q_1\varepsilon_1\leqslant\tau_1\varepsilon$ it follows

$$||\Delta_3||\leqslant\tau_1\tau_2\varepsilon^2\varepsilon_1\leqslant\frac{1}{\kappa^{4/3}K^{1/3}}\tau_2\tau_1\varepsilon^3. \tag{7.17}$$

The Proposition is proved. □

Lemma 7.2. *Let us consider*

$$\begin{aligned} G = &-X_2\Delta_1+\Delta_1Y_2-X_2\Sigma Y_2+A_2\Sigma+\Sigma B_2+\Theta_1^*\Sigma Y_2-X_2\Sigma\Psi_1\\ &+\Theta_1^*\Sigma B_2+A_2\Sigma\Psi_1-X_2\Sigma B_2+A_2\Sigma Y_2. \end{aligned}$$

Then $G=G_1+\cdots+G_5$ *with*

$$\begin{aligned} G_1 &= \frac{1}{2}X_2(\Delta_2-S_2)+\frac{1}{2}(S_2-\Delta_2)Y_2\\ G_2 &= \frac{1}{2}(X_1(\Delta_2-S_2)+(S_2-\Delta_2)Y_1)+\frac{1}{2}(X_2(-\Delta_1-S_1)+(S_1+\Delta_1)Y_2)\\ G_3 &= \frac{1}{2}\left(X_1(\Delta_2-S_2)Y_1+X_2(\Delta_1-S_1)Y_2+X_1\left(\Delta_2-S_2\right)Y_2\right)\\ &\quad+\frac{1}{2}(X_2(\Delta_1-S_1)Y_1+X_1(\Delta_1-S_1)Y_2+X_2(\Delta_2-S_2)Y_1)\\ G_4 &= \frac{1}{2}X_2(S_2-\Delta_2)Y_2\\ G_5 &= e_2(X_1)\Sigma R_{2,1}+Q_{2,1}\Sigma e_2(Y_1)+e_2(X_1)\Sigma e_2(Y_2)+e_2(X_2)\Sigma e_2(Y_1) \end{aligned}$$

where $Q_{2,1}=\frac{1}{2}(X_1X_2+X_2X_1)$ *and* $R_{2,1}=\frac{1}{2}(Y_1Y_2+Y_2Y_1)$.

Proof. Let $e_2(X)=X^2/2$. We have $A_2=e_2(X_2)+Q_{2,1}$ with

$$Q_{2,1}=\frac{1}{2}(X_1X_2+X_2X_1).$$

Moreover $\Theta_1=X_1+e_2(X_1)$. In the same way $B_2=e_2(Y_2)+R_{2,1}$ with $R_{2,1}=\frac{1}{2}(Y_1Y_2+Y_2Y_1)$and $\Psi_1=Y_1+e_2(Y_1)$. We also remark $e_2(X_2)=\dfrac{1}{2}X_2^2$. Expanding G

we can write G as the sum of the following quantities :

$$
\begin{aligned}
G_1 &= -X_2\Sigma Y_2 + \frac{1}{2}X_2^2\Sigma + \frac{1}{2}\Sigma Y_2^2\\
G_2 &= -X_2\Delta_1 + \Delta_1 Y_2 + Q_{2,1}\Sigma + \Sigma R_{2,1} - X_1\Sigma Y_2 - X_2\Sigma Y_1\\
G_3 &= -X_1\Sigma R_{2,1} + Q_{2,1}\Sigma Y_1 - X_2\Sigma R_{2,1} + Q_{2,1}\Sigma Y_2\\
&\qquad - X_1\Sigma e_2(Y_2) + e_2(X_2)\Sigma Y_1 + e_2(X_1)\Sigma Y_2 - X_2\Sigma e_2(Y_1)\\
G_4 &= -X_2\Sigma e_2(Y_2) + e_2(X_2)\Sigma Y_2\\
G_5 &= e_2(X_1)\Sigma R_{2,1} + Q_{2,1}\Sigma e_2(Y_1) + e_2(X_1)\Sigma e_2(Y_2) + e_2(X_2)\Sigma e_2(Y_1)
\end{aligned}
$$

We are going to transform the quantities G_i's. We first remark using $\Delta_2 - S_2 - X_2\Sigma + \Sigma Y_2 = 0$ that

$$
\begin{aligned}
-X_2\Sigma Y_2 + \frac{1}{2}X_2^2\Sigma + \frac{1}{2}\Sigma Y_2^2 &= \frac{1}{2}X_2(-\Sigma Y_2 + X_2\Sigma) + \frac{1}{2}(-X_2\Sigma + \Sigma Y_2)Y_2\\
&= \frac{1}{2}X_2(\Delta_2 - S_2) + \frac{1}{2}(S_2 - \Delta_2)Y_2.
\end{aligned}
$$

Hence

$$
G_1 = \frac{1}{2}X_2(\Delta_2 - S_2) + \frac{1}{2}(S_2 - \Delta_2)Y_2.
$$

Next we remember that $Q_{2,1} = \frac{1}{2}(X_1X_2 + X_2X_1)$ and $R_{2,1} = \frac{1}{2}(Y_1Y_2 + Y_2Y_1)$. On the other hand we have : $\Delta_i - S_i - X_i\Sigma + \Sigma Y_i = 0$ for $i = 1, 2$. Hence we can write G_2 as

$$
\begin{aligned}
G_2 &= -X_2\Delta_1 + \Delta_1 Y_2 + Q_{2,1}\Sigma + \Sigma R_{2,1} - X_1\Sigma Y_2 - X_2\Sigma Y_1\\
&= -X_2\Delta_1 + \Delta_1 Y_2 + \frac{1}{2}(X_1(X_2\Sigma - \Sigma Y_2) + (-X_2\Sigma + \Sigma Y_2)Y_1)\\
&\qquad + \frac{1}{2}(X_2(-\Sigma Y_1 + X_1\Sigma) + (-X_1\Sigma + \Sigma Y_1)Y_2)\\
&= -X_2\Delta_1 + \Delta_1 Y_2 + \frac{1}{2}(X_1(\Delta_2 - S_2) + (S_2 - \Delta_2)Y_1)\\
&\qquad + \frac{1}{2}(X_2(\Delta_1 - S_1) + (S_1 - \Delta_1)Y_2)\\
&= \frac{1}{2}(X_1(\Delta_2 - S_2) + (S_2 - \Delta_2)Y_1) + \frac{1}{2}(X_2(-\Delta_1 - S_1) + (S_1 + \Delta_1)Y_2)
\end{aligned}
$$

Next, by proceeding as above we see that

$$
\begin{aligned}
G_3 &= -X_1\Sigma R_{2,1} + Q_{2,1}\Sigma Y_1 - X_2\Sigma R_{2,1} + Q_{2,1}\Sigma Y_2\\
&\qquad - X_1\Sigma e_2(Y_2) + e_2(X_2)\Sigma Y_1 + e_2(X_1)\Sigma Y_2 - X_2\Sigma e_2(Y_1)\\
&= \frac{1}{2}(-X_1\Sigma Y_2Y_1 + X_1X_2\Sigma Y_1 - X_2\Sigma Y_1Y_2 + X_2X_1\Sigma Y_2)\\
&\qquad + \frac{1}{2}(X_1X_2\Sigma Y_2 + X_2X_1\Sigma Y_1 - X_1\Sigma Y_1Y_2 - X_2\Sigma Y_2Y_1)\\
&\qquad + \frac{1}{2}(-X_1\Sigma Y_2^2 - X_2\Sigma Y_1^2 + X_1^2\Sigma Y_2 + X_2^2\Sigma Y_1)\\
&= \frac{1}{2}\left(X_1(\Delta_2 - S_2)Y_1 + X_2(\Delta_1 - S_1)Y_2 + X_1\left(\Delta_2 - S_2\right)Y_2\right)\\
&\qquad + \frac{1}{2}(X_2(\Delta_1 - S_1)Y_1 + X_1(\Delta_1 - S_1)Y_2 + X_2(\Delta_2 - S_2)Y_1)
\end{aligned}
$$

We now see that

$$\begin{aligned} G_4 &= -X_2\Sigma e_2(Y_2) + e_2(X_2)\Sigma Y_2 \\ &= \frac{1}{2}(-X_2\Sigma Y_2^2 + X_2^2\Sigma Y_2) \\ &= \frac{1}{2}X_2(S_2 - \Delta_2)Y_2. \end{aligned}$$

Finally

$$G_5 = e_2(X_1)\Sigma R_{2,1} + Q_{2,1}\Sigma e_2(Y_1) + e_2(X_1)\Sigma e_2(Y_2) + e_2(X_2)\Sigma e_2(Y_1). \qquad \square$$

8. Proof of Theorem 1.2 : case $p \geqslant 3$.

8.1. Notations. Let us introduce some quantities to simplify the reading of expressions. We introduce the constants

$$\theta = 0.354, \qquad \eta = \frac{1}{1-\theta}, \; a = \frac{4}{3}, \quad b = \frac{1}{3}, \quad u_0 = 0.0297. \tag{8.1}$$

and the quantities :

$$\begin{aligned} &w = \tfrac{1}{\varepsilon}(-1 + (1-\varepsilon)^{-1/2}), \quad s = (1 + w\varepsilon)^2 + 2w + w^2\varepsilon = 2(1-\varepsilon)^{-1}, \\ &a_1(\varepsilon) = (1 + \sqrt{1-\varepsilon^2})^{-1}, \quad a_2(\varepsilon) = \frac{1}{\varepsilon^2}(a_1(\varepsilon) - 1/2) \\ &b_1(\varepsilon) = \frac{\varepsilon^2 a_1(\varepsilon)^2}{\sqrt{1-\varepsilon^2}} + 2a_1(\varepsilon), \quad b_2(\varepsilon) = \frac{a_1(\varepsilon)^2}{\sqrt{1-\varepsilon^2}} + 2a_2(\varepsilon) \\ &\alpha = \eta s, \end{aligned} \tag{8.2}$$

For $i = 1, 2$ we introduce

$$x_i = a_i(\eta\varepsilon), \quad y_i = b_i(\eta\varepsilon), \quad z_i = a_i(\theta\varepsilon), \quad r_1 = \theta^2 z_1 + \eta y_1, \quad t_1 = 1 + \eta x_1\varepsilon.$$

and

$$\begin{aligned} \tau(\varepsilon) = 2(1+\eta) &+ \left(2r_1 + \theta^2 + 2t_1\eta + \frac{3}{2}\eta^2 + \frac{1}{2}\eta\theta^2 + \frac{1}{2}\theta^4\right)\varepsilon_1 \\ &+ \left((z_1^2 + 2z_2)\theta^6 + 2y_1z_1\theta^4 + \left(2r_1 + 2x_1z_1\eta^2 + \eta^2y_1^2\right)\theta^2\right)\varepsilon_1^2 \\ &+ \left(2\left(y_2 + x_1y_1\right)\eta^3 + 2\eta r_1t_1\right)\varepsilon_1^2 \\ &+ (2z_2\theta^8 + 2z_2\eta\theta^6 + (2y_2\eta^3 + r_1^2)\theta^2 + 2(x_2 + y_2)\eta^4)\varepsilon_1^3. \end{aligned} \tag{8.3}$$

The following lemma justifies these notations and will be use in the sequel.

Lemma 8.1. *We have $\tau(s\varepsilon)s\varepsilon - \theta \leqslant 0$ and $2\dfrac{(1+\alpha\varepsilon)^{b/3}}{(1-2\alpha\varepsilon)^{a/3}}s^{4/3}\tau(s\varepsilon) \leqslant 1$ and for all $\varepsilon \in [0, u_0]$.*

Proof. From straighforward computations. $\square$

8.2. Proof. It consists to verify the assumptions of Theorem 5.2. Remember that

$$\max(\kappa^a K^{b+1}\|E_\ell(U)\|, \kappa^a K^{b+1}\|E_q(V)||, \kappa^a K^b\|\Delta||) \leqslant \varepsilon$$

where U, V, Δ stand for U_0, V_0, Δ_0 respectively. The item (5.2) follows of Proposition 3.2 since $\Omega = s_p(E_\ell(U))$ and $\Lambda = s_p(E_q(V))$. Let us prove the item (5.3). To do that we denote $\Delta_{0,1} = (I_\ell+\Omega)(\Delta+\Sigma)(I_q+\Lambda)-\Sigma$ and $\varepsilon_{0,1} = \|\Delta_{0,1}\|$. From Proposition 3.2 and assumption $||E_\ell(U)||, \|E_q(V)\| \leqslant \dfrac{\varepsilon}{\kappa^a K^{b+1}}$ we know that $\|\Omega\|, \|\Lambda\| \leqslant \dfrac{w}{\kappa^a K^{b+1}}\varepsilon$. We then apply Proposition 6.1 to get

$$
\begin{aligned}
\varepsilon_{0,1} &\leqslant ((1+w\varepsilon)^2 + 2w + w^2\varepsilon)\frac{\varepsilon}{\kappa^a K^b} \\
&\leqslant \frac{s\varepsilon}{\kappa^a K^b} \qquad \text{from (8.2)}\,.
\end{aligned} \tag{8.4}
$$

In view to use the Propositon 8.2, let us prove that $\tau(\varepsilon_{0,1})\varepsilon_{0,1} \leq \theta$. Using Lemma 8.1 we have

$$
\begin{aligned}
\tau(\varepsilon_{0,1})\varepsilon_{0,1} &\leqslant \tau(s\varepsilon)s\varepsilon \qquad \text{since } \varepsilon_{0,1} \leqslant s\varepsilon \\
&\leqslant \theta \qquad \text{from Lemma 8.1} \quad \text{since}\, \varepsilon \leqslant u_0.
\end{aligned}
$$

From formulas (1.11) we have

$$
\Delta_1 = \Delta_{0,p+1} = (I_\ell + \Theta_p^*)(\Delta_{0,1} + \Sigma)(I_q + \Psi_p) - \Sigma - \sum_{k=1}^{p} S_k.
$$

The quantity τ which appears in (5.7) is equal to $\tau(s\varepsilon)^p s^{p+1}$. Using Propositon 8.2 with $\tau := \tau(s\varepsilon)^p s^{p+1}$, we then get

$$
\begin{aligned}
\|\Delta_1\| &= \|\Delta_{0,p+1}\| \\
&\leqslant \frac{1}{\kappa^a K^b}(\tau(s\varepsilon)s^{\frac{p+1}{p}})^p \varepsilon^{p+1} \qquad \text{since} \quad \varepsilon_{0,1} \leqslant s\varepsilon.
\end{aligned}
$$

On the other hand from definition $S = S_1 + \cdots + S_p$ where $S_k = \operatorname{diag}(\Delta_{0,k})$. It follows $\|S_i\| \leqslant \varepsilon_{0,k} = \|\Delta_{0,k}\|$. From Proposition 8.2 one has

$$
\begin{aligned}
\varepsilon_{0,k} &\leqslant \tau(s\varepsilon)^{k-1}\varepsilon_{0,1}^k \\
&\leqslant \theta^{k-1}\varepsilon_{0,1} \qquad \text{since} \quad \tau(s\varepsilon)s\varepsilon \leqslant \theta \text{ and } \varepsilon_{0,1} \leqslant \frac{s\varepsilon}{\kappa^a K^b}
\end{aligned}
$$

We deduce

$$
\|S\| \leqslant \sum_{k=1}^{p} \varepsilon_{0,k} \leqslant \frac{1}{1-\theta}\varepsilon_{0,1} \leqslant \frac{\alpha\varepsilon}{\kappa^a K^b}. \tag{8.5}
$$

The assumption (5.7) is satisfied. In fact we have

$$
\begin{aligned}
(2\varepsilon)^p \frac{(1+\alpha\varepsilon)^b}{(1-2\alpha\varepsilon)^a}\tau(s\varepsilon)^p s^{p+1} &\leqslant \left(2\frac{(1+\alpha\varepsilon)^{b/3}}{(1-2\alpha\varepsilon)^{a/3}}\tau(s\varepsilon)s^{4/3}\varepsilon\right)^p \qquad \text{since}\, p \geqslant 3 \text{ and } s \geqslant 1 \\
&\leqslant 1 \qquad \text{from Lemma 8.1} \quad \text{since}\, \varepsilon \leqslant u_0.
\end{aligned} \tag{8.6}
$$

We now prove the item (5.4). We have

$$
\begin{aligned}
\|I_\ell + \Theta\|^2 &\leqslant (1 + ||c_p(X)||)^2 \\
\|(I_\ell + \Theta^*)(I_\ell + \Theta) - I_\ell\| &\leqslant (1 + c_p(-\|X\|))(1 + c_p(\|X\|)) - 1
\end{aligned}
$$

Using Lemma 9.4 and $\varepsilon_{0,1} \leqslant s\dfrac{\varepsilon}{\kappa^a K^b}$ we know that $\|X\| \leqslant \eta\kappa\varepsilon_{0,1} \leqslant \dfrac{x}{\kappa^{a-1}K^b} = \dfrac{x}{\kappa^b K^b}$ with $x = \alpha\varepsilon$. We deduce both from Lemma 9.4 that

$$(8.7) \qquad (1 + ||c_p(X)|)^2 \leqslant (1 + x + x^2 a_1(x))^2 = \zeta_1$$

and from Lemma 9.9 that

$$\begin{aligned}
(8.8) \qquad & (1 + c_p(-\|X\|))(1 + c_p(\|X\|)) - 1 \\
& \leqslant \left(2\sqrt{1-x^2} + a_1(x)x^{p+1}\right) a_1(x) \left(\frac{1}{\kappa^{a-1}K^b}\alpha\varepsilon\right)^{p+\delta} \\
& \leqslant \left(2\sqrt{1-x^2} + a_1(x)x^3\right) a_1(x)\alpha^{p+\delta} \left(\frac{1}{\kappa^b K^b}\right)^{p+\delta} \varepsilon^{p+1} \\
& \leqslant \frac{\zeta_2}{\kappa^a K^{b+1}}\varepsilon^{p+1} \text{ since } \ p \geqslant 3 \text{ implies } \ (p+\delta)b \geqslant b+1
\end{aligned}$$

where $\delta = 1$ if p is odd and $\delta = 2$ if p is even from Lemma 9.9. We then remark that

$$(8.9) \qquad (2\epsilon)^p \alpha^{p+\delta}\varepsilon^{\delta-1} \leqslant (2\alpha^{5/3}\varepsilon)^p \quad \text{since} \quad \frac{p+\delta}{p} \leqslant \frac{5}{3}$$

This allows to prove the assumption (5.8) that is $(2\varepsilon)^p \dfrac{(1+\alpha\varepsilon)^{b+1}}{(1-2\alpha\varepsilon)^a}(\zeta_1 + \zeta_2\varepsilon^{\delta-1}) \leqslant 1$. We first have since $b+1 = a$

$$\begin{aligned}
(2\varepsilon)^p \left(\frac{1+\alpha\varepsilon}{1-2\alpha\varepsilon}\right)^a \zeta_1 & \leqslant \left(2\left(\frac{1+\alpha\varepsilon}{1-2\alpha\varepsilon}\right)^{a/3} (1 + x + x^2 a_1(x))^{2/3}\varepsilon\right)^p \\
& \leqslant (0.037)^p \leqslant 0.00005 \quad \text{since} \quad \varepsilon \leqslant u_0 \text{ and } p \geqslant 3.
\end{aligned}$$

We now remark that

$$\zeta_2 = \left(2\sqrt{1-x^2} + a_1(x)x^3\right) a_1(x) \leqslant \quad 0.998 \quad \text{since } \varepsilon \leqslant u_0 \text{ implies } x \leqslant 0.098.$$

Taking in account (8.8-8.9) we get :

$$\begin{aligned}
(2\varepsilon)^p \left(\frac{1+\alpha\varepsilon}{1-2\alpha\varepsilon}\right)^a \zeta_2\varepsilon^{\delta-1} & \leqslant \left(2\left(\frac{1+\alpha\varepsilon}{1-2\alpha\varepsilon}\right)^{a/3} \alpha^{5/3}\varepsilon\right)^p \\
& \leqslant (0.24)^p \leqslant 0.013 \qquad \text{since} \quad \varepsilon \leqslant u_0 \text{ and } p \geqslant 3.
\end{aligned}$$

Consequently $(2\varepsilon)^p \dfrac{(1+\alpha\varepsilon)^a}{(1-2\alpha\varepsilon)^a}(\zeta_1 + \zeta_2\varepsilon^{\delta-1}) \leqslant 0.015 < 1$. This proves the item (5.8). The assumption (5.9) holds since $1 - 8\alpha\varepsilon \geqslant 0.25 > 0$ when $\varepsilon < u_0$.

We now verify the assumption (5.5). From above we know that $\|\Omega\|, \|\Lambda\| \leqslant \dfrac{w}{\kappa^a K^{b+1}}\varepsilon$ with $w = \dfrac{1}{\varepsilon}(-1 + (1-\varepsilon)^{-1/2})$. We can take $w \leqslant \alpha_1 = 0.52$ since $\varepsilon \leqslant u_0$.

On the other hand one has $\Theta = c_p(X)$ and $\Psi = c_p(Y)$. From above we know that

$$\begin{aligned}
\|c_p(X)\|, \|c_p(Y)\| & \leqslant (1 + x a_1(x))\, x \quad \text{with} \quad x = \alpha\varepsilon \\
& \leqslant \alpha_2\varepsilon \qquad \text{with} \quad \alpha_2 = 3.35 \quad \text{since } \varepsilon \leqslant u_0.
\end{aligned}$$

Since $\gamma u_0 = 2(\alpha_1+\alpha_2+\alpha_1\alpha_2 u_0)u_0 < 0.233 < 1$ then the bounds (5.12-5.14) of Theorem 5.2 hold with

$$\gamma = 7.82$$
$$\frac{\gamma}{1-\gamma u_0} \leqslant 10.2$$
$$\sigma = 0.82\alpha \leqslant 2.62$$

The Theorem 1.2 is proved for $p \geqslant 3$. □

PROPOSITION 8.2. *Let $p > 2$, $\varepsilon \geqslant 0$. Let us consider $\Delta_1 = U_1^* M V_1 - \Sigma$ such that $\|\Delta_1\| = \varepsilon_1 \leqslant \frac{\varepsilon}{\kappa^{4/3}K^{1/3}}$ where $\kappa = \kappa(\Sigma)$ and $K = K(\Sigma)$. Let us consider $\tau := \tau(\varepsilon)$ as in* (8.3) *and suppose $\tau\varepsilon \leq \theta$. Then we have*

$$\tau_{p+1} := \|\Delta_{p+1}\| \leqslant \frac{1}{\kappa^{4/3}K^{1/3}}\tau(\varepsilon)^p \varepsilon^{p+1}$$

where $\Delta_{p+1} = (I_\ell + \Theta_p^*)(\Delta_1+\Sigma)(I_q+\Psi_p) - \Sigma - \sum_{l=1}^{p} S_l$, with Θ_p and Ψ_p are defined by the formulas (1.11).

Proof. Since the X_k's and Y_k's are skew Hermitian matrices, we have $\Theta_p = \Theta_{p-1} + X_p + A_p$ with

$$A_p := A_p(X_1+\ldots+X_{p-1}, X_p) = c_p(X_1+\cdots+X_p) - c_p(X_1+\cdots+X_{p-1}) - X_p$$

In the same way $\Psi_p = \Psi_{p-1} + Y_p + B_p$ where $B_p = A_p(Y_1+\cdots+Y_{p-1}, Y_p)$. We remark that A_p and B_p are Hermitian matrices. Expanding $(I_\ell+\Theta_p^*)(\Delta_1+\Sigma)(I_q+\Psi_p)$ we get

$$\begin{aligned}
\Delta_{p+1} &= (I_\ell+\Theta_p^*)(\Delta_1+\Sigma)(I_q+\Psi_p) - \Sigma - \sum_{l=1}^{p} S_l \\
&= (I_\ell+\Theta_{p-1}^* - X_p + A_p)(\Delta_1+\Sigma)(I_q+\Psi_{p-1}+Y_p+B_p) - \Sigma - \sum_{l=1}^{p} S_l \\
&= (I_\ell+\Theta_{p-1}^*)(\Delta_1+\Sigma)(I_q+\Psi_{p-1}) - \Sigma - \sum_{l=1}^{p-1} S_l - S_p - X_p\Sigma + \Sigma Y_p \\
&\qquad + (I_\ell+\Theta_{p-1}^*)(\Delta_1+\Sigma)(Y_p+B_p) + (-X_p+A_p)(\Delta_1+\Sigma)(I_q+\Psi_{p-1}) \\
&\qquad + (-X_p+A_p)(\Delta_1+\Sigma)(Y_p+B_p) + X_p\Sigma - \Sigma Y_p.
\end{aligned}$$

From definition we know that

$$(I_\ell+\Theta_{p-1}^*)(\Delta_1+\Sigma)(I_q+\Psi_{p-1}) - \Sigma - \sum_{l=1}^{p-1} S_l - S_p - X_p\Sigma + \Sigma Y_p = \Delta_p - S_p - X_p\Sigma + \Sigma Y_p = 0.$$

Expanding more Δ_{p+1}, we then can write by grouping the terms appropriately :

$$\Delta_{p+1} = -X_p\Delta_1 + \Delta_1 Y_p - X_p\Delta_1 Y_p + \Delta_1 B_p + A_p\Delta_1 - X_p\Delta_1 B_p + A_p\Delta_1 Y_p \tag{8.10}$$
$$\qquad + A_p\Delta_1 B_p + \Theta_{p-1}^*\Delta_1 Y_p - X_p\Delta_1\Psi_{p-1} + \Theta_{p-1}^*\Delta_1 B_p + A_p\Delta_1\Psi_{p-1} \tag{8.11}$$
$$\qquad + G,$$

where $G = -X_p\Sigma Y_p + \Sigma B_p + A_p\Sigma + \Theta_{p-1}^*\Sigma Y_p - X_p\Sigma\Psi_{p-1} + \Theta_{p-1}^*\Sigma B_p + A_p\Sigma\Psi_{p-1} - X_p\Sigma B_p + A_p\Sigma Y_p + A_p\Sigma B_p$. From the Lemma 8.3 the quantity G is sum of the following G_i's :

$$G_1 = d_p(X_p)\Sigma + \Sigma d_p(Y_p) \tag{8.12}$$

$$G_2 = Q_{p,2}\Sigma + \Sigma R_{p,2} + \frac{1}{2}C_{p-1}(\Delta_p - S_p) - \frac{1}{2}(\Delta_p - S_p)D_{p-1} + \frac{1}{2}X_p\sum_{k=1}^{p}(\Delta_k - S_k) + \frac{1}{2}\sum_{k=1}^{p}(S_k - \Delta_k)Y_p \tag{8.13}$$

$$G_3 = \frac{1}{2}C_{p-1}(\Delta_p - S_p)D_{p-1} - \frac{1}{2}X_p\sum_{k=1}^{p-1}(\Delta_k - S_k)Y_p + \frac{1}{2}X_p\sum_{k=1}^{p}(\Delta_k - S_k)D_{p-1} + \frac{1}{2}C_{p-1}\sum_{k=1}^{p}(S_k - \Delta_k)Y_p \tag{8.14}$$

$$G_4 = \frac{1}{2}X_p(S_p - \Delta_p)Y_p - X_p\Sigma d_p(Y_p) + d_p(X_p)\Sigma Y_p. \tag{8.15}$$

$$G_5 = e_p(C_{p-1})\Sigma R_{p,1} + Q_{p,1}\Sigma e_p(D_{p-1}) + e_p(C_{p-1})\Sigma e_p(Y_p) \tag{8.16}$$

$$+ e_p(X_p)\Sigma e_p(D_{p-1}) + Q_{p,1}\Sigma R_{p,1} + Q_{p,1}\Sigma e_p(Y_p) \tag{8.17}$$

$$+ e_p(X_p)\Sigma R_{p,1} + e_p(X_p)\Sigma e_p(Y_p). \tag{8.18}$$

$$G_6 = -C_{p-1}\Sigma R_{p,2} + Q_{p,2}\Sigma D_{p-1} - X_p\Sigma R_{p,2} + Q_{p,2}\Sigma Y_p - C_{p-1}\Sigma d_p(Y_p) + d_p(X_p)\Sigma D_{p-1} + d_p(C_{p-1})\Sigma Y_p - X_p\Sigma d_p(D_{p-1}). \tag{8.19}$$

where the quantities $Q_{p,i}$ and $R_{p,i}$ are defined at Lemma **??**. We now can bound $\|\Delta_{p+1}\|$. To do that introduce the quantities where $i = 1, 2$:

$$x_i = a_i(\eta\varepsilon), \quad y_i = b_i(\eta\varepsilon), \quad z_i = a_i(\theta\varepsilon), \quad r_1 = \theta^2 z_1 + \eta y_1, \quad t_1 = 1 + x_1\eta\varepsilon$$

and the polynomial $q := q(\kappa, K, \varepsilon_1)$

$$\begin{aligned} q = {} & 2(1+\eta)\kappa + \left(2r_1 + \theta^2 + 2t_1\eta + \frac{3}{2}\eta^2 + \frac{1}{2}\eta\theta^2 + \frac{1}{2}\theta^4\right)\kappa^2\varepsilon_1 \\ & + \left((z_1^2 + 2z_2)\theta^6 + 2\eta x_1 z_1\theta^4\right)K\kappa^4\varepsilon_1^2 \\ & + \left(\left(2r_1 + 2x_1z_1\eta^2 + \eta^2y_1^2\right)\theta^2 + 2\left(y_2 + x_1y_1\right)\eta^3 + 2\eta r_1 t_1\right)K\kappa^4\varepsilon_1^2 \\ & + (2z_2\theta^8 + 2z_2\eta\theta^6 + (2y_2\eta^3 + r_1^2)\theta^2 + 2(x_2+y_2)\eta^4)K\kappa^5\varepsilon_1^3. \end{aligned}$$

The inequality $\tau(\varepsilon)\varepsilon \leqslant \theta$ implies $q\varepsilon_1 \leqslant \theta$. In fact it is easy to see that the assumption $\varepsilon_1 \leqslant \dfrac{\varepsilon}{\kappa^{4/3}K^{1/3}}$ implies $q\varepsilon_1 \leqslant \tau(\varepsilon)\varepsilon$ since we simultaneously have $\kappa\varepsilon_1 \leqslant \varepsilon$, $\kappa^2\varepsilon_1^2 \leqslant \varepsilon^2$, $K\kappa^4\varepsilon_1^3 \leqslant \varepsilon^3$ and $K\kappa^5\varepsilon_1^4 \leqslant \varepsilon^4$. We know that $\|\Delta_1\| \leqslant \varepsilon_1$. Let us suppose $\|\Delta_k\| \leqslant q^{k-1}\varepsilon_1^k$ for $1 \leqslant k \leqslant p$ and, prove that $\|\Delta_{p+1}\| \leqslant q^p\varepsilon_1^{p+1}$. We remark $q \geqslant 2(\theta+\eta)$ in order that the Lemmas 9.4-9.8 apply. To bound $\|\Delta_{p+1}\|$ we use the following bounds :

1. We have for $i = 1, 2$, $a_i(\theta\kappa\varepsilon_1) \leqslant x_i$ $b_i(\eta\kappa\varepsilon_1) \leqslant y_i$.

2. For $1 \leqslant k \leqslant p$, we know that $\|X_k\|$, $\|Y_k\| \leqslant \kappa q^{k-1}\varepsilon_1^k$ from Proposition 4.3.

3. $\|C_k\|, \|D_k\| \leqslant \eta\kappa\varepsilon_1$ from Lemma 9.4 and also $\left\| \sum_{k=1}^{p-1} \Delta_k - S_k \right\| \leqslant \eta\varepsilon_1$ from Lemma 9.1.

4. $\|Q_{p,i}\|, \|R_{p,i}\| \leqslant \eta^{2i-1} y_i \kappa^{2i} q^{p-1} \varepsilon_1^{p+2i-1}$ from Lemma 9.7.

5. $\|e_p(X_p)\|, \|e_p(Y_p)\| \leqslant z_1\kappa^2 q^{2(p-1)} \varepsilon_1^{2p} \leqslant \theta^2 z_1 \kappa^2 q^{p-1} \varepsilon_1^{p+1}$ and $\|e_p(C_{p-1})\|, \|e_p(D_{p-1})\| \leqslant x_1\eta^2\kappa^2\varepsilon_1^2$ from Lemma 9.4, $q\varepsilon_1 \leqslant \theta$ and $p \geqslant 3$.

6. $\|d_p(X_p)\|, \|d_p(Y_p)\| \leqslant z_2\kappa^4 q^{4(p-1)}\varepsilon_1^{4p}$ and $\|d_p(C_{p-1})\|, \|d_p(D_{p-1})\| \leqslant x_2\eta^4\kappa^4\varepsilon_1^4$ from Lemma 9.5 and $q\varepsilon \leqslant \theta$.

7. $\|A_p\|, \|B_p\| \leq r_1\kappa^2\, q^{p-1}\varepsilon_1^{p+1}$ since $A_p = e_p(X_p) + Q_{p,1}$ and $B_p = e_p(Y_p) + R_{p,1}$.

8. $\|\Theta_{p-1}\|, \|\Psi_{p-1}\| \leq t_1\eta\kappa\varepsilon_1$ from Lemma 9.6.

9. $\| - X_p\Sigma d_p(Y_p) + d_p(X_p)\Sigma Y_p\| \leqslant 2Kz_2\kappa^5 q^{5(p-1)}\varepsilon_1^{5p}$ from Lemma 9.8.

Using the bounds above we then get $\|\Delta_{p+1}\| \leqslant \alpha_{p+1} q^{p-1}\varepsilon_1^{p+1}$ where

$$\begin{aligned}
&\alpha_{p+1} = \\
&2\kappa + \kappa^2 q^{p-1}\varepsilon_1^{p} + 2r_1\kappa^3 q^{p-1}\varepsilon_1^{p+1} + 2r_1\kappa^2\varepsilon_1 && \text{from (8.10)}\\
&+r_1^2\kappa^4 q^{p-1}\,\varepsilon_1^{p+2} + 2t_1\eta\kappa^2\,\varepsilon_1 + 2r_1t_1\eta\kappa^3\,\varepsilon_1^2 && \text{from (8.11)}\\
&+2z_2K\kappa^4 q^{3(p-1)}\varepsilon_1^{3p-1} + 2\eta^3 y_2 K\kappa^4\varepsilon_1^2 + 2\eta\kappa && \text{from (8.12 + 8.13)}\\
&+\tfrac{3}{2}\eta^2\kappa^2\varepsilon_1 + \tfrac{1}{2}\eta\kappa^2 q^{p-1}\varepsilon_1^{p} && \text{from (8.14)}\\
&+\tfrac{1}{2}\kappa^2 q^{2(p-1)}\varepsilon_1^{2p-1} + 2z_2K\kappa^5 q^{4(p-1)}\varepsilon_1^{4p-1} && \text{from (8.15)}\\
&+K\kappa^4(2x_1y_1\eta^3\varepsilon_1^2 + 2z_1x_1\eta^2 q^{p-1}\varepsilon_1^{p+1}) && \text{from (8.16)}\\
&+K\kappa^4(y_1^2\eta^2 q^{p-1}\varepsilon_1^{p+1} + 2z_1y_1\eta q^{2(p-1)}\varepsilon_1^{2p} + z_1^2 q^{3(p-1)}\varepsilon^{3p-1}) && \text{from (8.17 - 8.18)}\\
&+K\kappa^5(2\eta^4(x_2+y_2)\varepsilon_1^3 + 2z_2\eta q^{3(p-1)}\varepsilon_1^{3p} + 2y_2\eta^3 q^{p-1}\varepsilon_1^{p+2}) && \text{from (8.19)}
\end{aligned}$$

Since $p \geqslant 3$ and $\theta < 1$ it follows $(q\varepsilon_1)^{k(p-1)} \leqslant (q\varepsilon_1)^{2k} \leqslant (\tau\varepsilon)^{2k} \leqslant \theta^{2k}$. Plugging this in α_{p+1}, we then get

$$\begin{aligned}
\alpha_{p+1} \leqslant\ & 2\kappa + \kappa^2\theta^2\varepsilon_1 + 2r_1\kappa^3\theta^2\varepsilon_1^2 + 2r_1\kappa^2\,\varepsilon_1\\
&+r_1^2\kappa^4\theta^2\,\varepsilon_1^3 + 2t_1\eta\kappa^2\varepsilon_1 + 2r_1t_1\eta\kappa^3\,\varepsilon_1^2\\
&+2z_2K\kappa^4\theta^6\varepsilon_1^2 + 2\eta^3 y_2K\kappa^4\varepsilon_1^2 + 2\eta\kappa\\
&+\frac{3}{2}\eta^2\kappa^2\varepsilon_1 + \frac{1}{2}\eta\kappa^2\theta^2\varepsilon_1\\
&+\frac{1}{2}\kappa^2\theta^4\varepsilon_1 + 2z_2K\kappa^5\theta^8\varepsilon_1^3\\
&+K\kappa^4(2x_1y_1\eta^3\varepsilon_1^2 + 2z_1x_1\eta^2\theta^2\varepsilon_1^2)\\
&+K\kappa^4(y_1^2\eta^2\theta^2\varepsilon_1^2 + 2z_1y_1\eta\theta^4\varepsilon_1^2 + z_1^2\theta^6\varepsilon_1^2)\\
&+K\kappa^5(2\eta^4(x_2+y_2)\varepsilon_1^3 + 2z_2\eta\theta^6\varepsilon_1^3 + 2y_2\eta^3\theta^2\varepsilon_1^3).
\end{aligned}$$

Collecting the expression above following ε_1 and using that $\kappa, K \geqslant 1$, we finally find that $\alpha_{p+1} \leqslant q$. We then have proved that $\|\Delta_{p+1}\| \leqslant q^p \varepsilon_1^{p+1}$. We finally get

$$\begin{aligned}\|\Delta_{p+1}\| &\leqslant \tau(\varepsilon)^p \varepsilon^p \varepsilon_1 \\ &\leqslant \frac{1}{\kappa^{4/3} K^{1/3}} \tau(\varepsilon)^p \varepsilon^{p+1}.\end{aligned}$$

The theorem is proved. □

LEMMA 8.3. *Let us consider*

$$\begin{aligned}G = &-X_p \Sigma Y_p + A_p \Sigma + \Sigma B_p + \Theta_{p-1}^* \Sigma Y_p - X_p \Sigma \Psi_{p-1} \\ &+ \Theta_{p-1}^* \Sigma B_p + A_p \Sigma \Psi_{p-1} - X_p \Sigma B_p + A_p \Sigma Y_p + A_p \Sigma B_p.\end{aligned}$$

Let $C_{p-1} = X_1 + \cdots + X_{p-1}$ *and* $D_{p-1} = Y_1 + \cdots + Y_{p-1}$. *Then* $G = G_1 + \cdots + G_6$ *with*

$$\begin{aligned}
G_1 &= d_p(X_p)\Sigma + \Sigma d_p(Y_p) \\
G_2 &= Q_{p,2}\Sigma + \Sigma R_{p,2} + \frac{1}{2} C_{p-1}(\Delta_p - S_p) - \frac{1}{2}(\Delta_p - S_p) D_{p-1} \\
&\quad + \frac{1}{2} X_p \sum_{k=1}^{p} (\Delta_k - S_k) + \frac{1}{2} \sum_{k=1}^{p} (S_k - \Delta_k) Y_p. \\
G_3 &= \frac{1}{2} C_{p-1}(\Delta_p - S_p) D_{p-1} - \frac{1}{2} X_p \sum_{k=1}^{p-1} (\Delta_k - S_k) Y_p \\
&\quad + \frac{1}{2} X_p \sum_{k=1}^{p} (\Delta_k - S_k) D_{p-1} + \frac{1}{2} C_{p-1} \sum_{k=1}^{p} (S_k - \Delta_k) Y_p \\
G_4 &= \frac{1}{2} X_p (S_p - \Delta_p) Y_p - X_p \Sigma d_p(Y_p) + d_p(X_p) \Sigma Y_p. \\
G_5 &= e_p(C_{p-1}) \Sigma R_{p,1} + Q_{p,1} \Sigma e_p(D_{p-1}) + e_p(C_{p-1}) \Sigma e_p(Y_p) + e_p(X_p) \Sigma e_p(D_{p-1}) \\
&\quad + Q_{p,1} \Sigma R_{p,1} + Q_{p,1} \Sigma e_p(Y_p) + e_p(X_p) \Sigma R_{p,1} + e_p(X_p) \Sigma e_p(Y_p). \\
G_6 &= -C_{p-1} \Sigma R_{p,2} + Q_{p,2} \Sigma D_{p-1} - X_p \Sigma R_{p,2} + Q_{p,2} \Sigma Y_p \\
&\quad - C_{p-1} \Sigma d_p(Y_p) + d_p(X_p) \Sigma D_{p-1} + d_p(C_{p-1}) \Sigma Y_p - X_p \Sigma d_p(D_{p-1}).s
\end{aligned}$$

Proof. We have $A_p = e_p(X_p) + Q_{p,1} = \frac{1}{2} X_p^2 + d_p(X_p) + Q_{p,1}$ with

$$Q_{p,i} = \sum_{k=i}^{\max(k:2k \leqslant p)} c_k \sum_{\substack{i_1 + i_2 = 2k \\ i_1,\, i > 0}} L_{i_1,i_2}(C_{p-1}, X_p).$$

where the coefficients c_k and the polynomials L_{i_1,i_2} are defined at the beginning of the section 9. Moreover $\Theta_{p-1} = C_{p-1} + e_p(C_{p-1})$. In the same way $B_p = e_p(Y_p) + R_{p,1} = \frac{1}{2} Y_p^2 + d_p(Y_p) + R_{p,1}$ and $\Psi_{p-1} = D_{p-1} + e_p(D_{p-1})$. We also know that $\Theta_{p-1}^* = -C_{p-1} + e_p(C_{p-1})$ since C_{p-1} is a skew Hermitian matrix. Expanding

$$\begin{aligned}G = &-X_p \Sigma Y_p + A_p \Sigma + \Sigma B_p + \Theta_{p-1}^* \Sigma Y_p - X_p \Sigma \Psi_{p-1} \\ &+ \Theta_{p-1}^* \Sigma B_p + A_p \Sigma \Psi_{p-1} - X_p \Sigma B_p + A_p \Sigma Y_p + A_p \Sigma B_p,\end{aligned}$$

a straightforward calculation shows that we can write G as the sum of the following quantities :

$$
\begin{aligned}
G_1 &= d_p(X_p)\Sigma + \Sigma d_p(Y_p) \\
G_2 &= Q_{p,1}\Sigma + \Sigma R_{p,1} - C_{p-1}\Sigma Y_p - X_p\Sigma D_{p-1} - X_p\Sigma Y_p + \frac{1}{2}X_p^2\Sigma + \frac{1}{2}\Sigma Y_p^2 \\
G_3 + G_6 &= -C_{p-1}\Sigma R_{p,1} + Q_{p,1}\Sigma D_{p-1} - X_p\Sigma R_{p,1} + Q_{p,1}\Sigma Y_p \\
&\quad - C_{p-1}\Sigma e_p(Y_p) + e_p(X_p)\Sigma D_{p-1} + e_p(C_{p-1})\Sigma Y_p - X_p\Sigma e_p(D_{p-1}) \\
G_4 &= -X_p\Sigma e_p(Y_p) + e_p(X_p)\Sigma Y_p \\
G_5 &= e_p(C_{p-1})\Sigma R_{p,1} + Q_{p,1}\Sigma e_p(D_{p-1}) + e_p(C_{p-1})\Sigma e_p(Y_p) + e_p(X_p)\Sigma e_p(D_{p-1}) \\
&\quad + Q_{p,1}\Sigma R_{p,1} + Q_{p,1}\Sigma e_p(Y_p) + e_p(X_p)\Sigma R_{p,1} + e_p(X_p)\Sigma e_p(Y_p).
\end{aligned}
$$

We are going to transform some quantities G_i's. We first remark using $\Delta_p - S_p - X_p\Sigma + \Sigma Y_p = 0$ that

$$
\begin{aligned}
-X_p\Sigma Y_p + \frac{1}{2}X_p^2\Sigma + \frac{1}{2}\Sigma Y_p^2 &= \frac{1}{2}X_p(-\Sigma Y_p + X_p\Sigma) + \frac{1}{2}(-X_p\Sigma + \Sigma Y_p)Y_p \\
&= \frac{1}{2}X_p(\Delta_p - S_p) - \frac{1}{2}(\Delta_p - S_p)Y_p.
\end{aligned}
$$

Next we remark that $Q_{p,1} = \frac{1}{2}(C_{p-1}X_p + X_pC_{p-1}) + Q_{p,2}$ and $R_{p,1} = \frac{1}{2}(D_{p-1}Y_p + Y_pD_{p-1}) + R_{p,2}$. On the other hand we have : $\sum_{k=1}^{p-1}(\Delta_k - S_k) - C_{p-1}\Sigma + \Sigma D_{p-1} = 0$. Hence we can write G_2 as

$$
\begin{aligned}
G_2 &= Q_{p,1}\Sigma + \Sigma R_{p,1} - C_{p-1}\Sigma Y_p - X_p\Sigma D_{p-1} - X_p\Sigma Y_p + \frac{1}{2}X_p^2\Sigma + \frac{1}{2}\Sigma Y_p^2 \\
&= Q_{p,2}\Sigma + \Sigma R_{p,2} + \frac{1}{2}C_{p-1}(X_p\Sigma - \Sigma Y_p) + \frac{1}{2}(-X_p\Sigma + \Sigma Y_p)D_{p-1} \\
&\quad + \frac{1}{2}X_p(-\Sigma D_{p-1} + C_{p-1}\Sigma) + \frac{1}{2}(-C_{p-1}\Sigma + \Sigma D_{p-1})Y_p \\
&\quad + \frac{1}{2}X_p(\Delta_p - S_p) - \frac{1}{2}(\Delta_p - S_p)Y_p \\
&= Q_{p,2}\Sigma + \Sigma R_{p,2} + \frac{1}{2}C_{p-1}(\Delta_p - S_p) - \frac{1}{2}(\Delta_p - S_p)D_{p-1} \\
&\quad + \frac{1}{2}X_p\sum_{k=1}^{p}(\Delta_k - S_k) + \frac{1}{2}\sum_{k=1}^{p}(S_k - \Delta_k)Y_p.
\end{aligned}
$$

Next, by proceeding as above and using $e_p = \frac{1}{2}u^2 + d_p(u)$, we see that

$$\begin{aligned}
G_3 + G_6 &= -C_{p-1}\Sigma R_{p,1} + Q_{p,1}\Sigma D_{p-1} - X_p\Sigma R_{p,1} + Q_{p,1}\Sigma Y_p \\
&\quad - C_{p-1}\Sigma e_p(Y_p) + e_p(X_p)\Sigma D_{p-1} + e_p(C_{p-1})\Sigma Y_p - X_p\Sigma e_p(D_{p-1}) \\
&= \frac{1}{2}(-C_{p-1}\Sigma Y_p D_{p-1} + C_{p-1}X_p\Sigma D_{p-1} - X_p\Sigma D_{p-1}Y_p + X_pC_{p-1}\Sigma Y_p) \\
&\quad + \frac{1}{2}(C_{p-1}X_p\Sigma Y_p + X_pC_{p-1}\Sigma D_{p-1} - C_{p-1}\Sigma D_{p-1}Y_p - X_p\Sigma Y_p D_{p-1}) \\
&\quad + \frac{1}{2}(-C_{p-1}\Sigma Y_p^2 - X_p\Sigma D_{p-1}^2 + C_{p-1}^2\Sigma Y_p + X_p^2\Sigma D_{p-1}) \\
&\quad - C_{p-1}\Sigma R_{p,2} + Q_{p,2}\Sigma D_{p-1} - X_p\Sigma R_{p,2} + Q_{p,2}\Sigma Y_p \\
&\quad - C_{p-1}\Sigma d_p(Y_p) + d_p(X_p)\Sigma D_{p-1} + d_p(C_{p-1})\Sigma Y_p - X_p\Sigma d_p(D_{p-1}).
\end{aligned}$$

We group some terms of the previous expression :

$$\begin{aligned}
-C_{p-1}\Sigma Y_p D_{p-1} + C_{p-1}X_p\Sigma D_{p-1} &= C_{p-1}(\Delta_p - S_p)D_{p-1} \\
-X_p\Sigma D_{p-1}Y_p + X_pC_{p-1}\Sigma Y_p &= -X_p\sum_{k=1}^{p-1}(\Delta_k - S_k)Y_p \\
C_{p-1}X_p\Sigma Y_p - C_{p-1}\Sigma Y_p^2 &= C_{p-1}(\Delta_p - S_p)Y_p \\
X_pC_{p-1}\Sigma D_{p-1} - X_p\Sigma D_{p-1}^2 &= X_p\sum_{k=1}^{p-1}(\Delta_k - S_k)D_{p-1} \\
-C_{p-1}\Sigma D_{p-1}Y_p + C_{p-1}^2\Sigma Y_p &= C_{p-1}\sum_{k=1}^{p-1}(\Delta_k - S_k)Y_p \\
-X_p\Sigma Y_p D_{p-1} + X_p^2\Sigma D_{p-1} &= X_p(\Delta_p - S_p)D_{p-1}
\end{aligned}$$

In this way we get

$$\begin{aligned}
G_3 + G_6 &= \frac{1}{2}C_{p-1}(\Delta_p - S_p)D_{p-1} - \frac{1}{2}X_p\sum_{k=1}^{p-1}(\Delta_k - S_k)Y_p + \frac{1}{2}C_{p-1}(\Delta_p - S_p)Y_p \\
&\quad + \frac{1}{2}X_p\sum_{k=1}^{p-1}(\Delta_k - S_k)D_{p-1} + \frac{1}{2}C_{p-1}\sum_{k=1}^{p-1}(\Delta_k - S_k)Y_p \\
&\quad + \frac{1}{2}X_p(\Delta_p - S_p)D_{p-1} + G_6 \\
&= \frac{1}{2}C_{p-1}(\Delta_p - S_p)D_{p-1} - \frac{1}{2}X_p\sum_{k=1}^{p-1}(\Delta_k - S_k)Y_p \\
&\quad + \frac{1}{2}X_p\sum_{k=1}^{p}(\Delta_k - S_k)D_{p-1} + \frac{1}{2}C_{p-1}\sum_{k=1}^{p}(S_k - \Delta_k)Y_p + G_6
\end{aligned}$$

with

$$\begin{aligned}
G_6 = &-C_{p-1}\Sigma R_{p,2} + Q_{p,2}\Sigma D_{p-1} - X_p\Sigma R_{p,2} + Q_{p,2}\Sigma Y_p \\
&- C_{p-1}\Sigma d_p(Y_p) + d_p(X_p)\Sigma D_{p-1} + d_p(C_{p-1})\Sigma Y_p - X_p\Sigma d_p(D_{p-1}).
\end{aligned}$$

We now see that

$$
\begin{aligned}
G_4 &= -X_p \Sigma e_p(Y_p) + e_p(X_p)\Sigma Y_p \\
&= \frac{1}{2}(-X_p \Sigma Y_p^2 + X_p^2 \Sigma Y_p) - X_p \Sigma d_p(Y_p) + d_p(X_p)\Sigma Y_p \\
&= \frac{1}{2} X_p (S_p - \Delta_p) Y_p - X_p \Sigma d_p(Y_p) + d_p(X_p)\Sigma Y_p.
\end{aligned}
$$

Finally G_5 remains unchanged. □

9. Useful Lemmas and Propositions. The notations are those of the introduction and sections 6, 7 and 8. We also denote :

1. $e_p(u) = \displaystyle\sum_{k=1}^{max\{k\,:\,2k\leqslant p\}} c_k u^{2k}$ where $c_k = (-1)^{k+1} \dfrac{(2k)!}{4^k (k!)^2 (2k-1)}$.
2. $c_p(u) = u + e_p(u) = u + \dfrac{1}{2}u^2 + d_p(u)$ with $d_p(u) = \displaystyle\sum_{k=2}^{max\{k\,:\,2k\leqslant p\}} c_k u^{2k}$.
3. $L_{i_1,i_2}(X,Y)$ is the sum of monomials which the degree of each monomial with respect X is i_1 (respectively with respect Y is i_2).

LEMMA 9.1. *Let for* $1 \leqslant k \leqslant i$, $\|\Delta_k\| \leqslant q^{k-1}\varepsilon_1^k$ *with* $q\varepsilon_1 \leqslant \theta < 1$. *Then* $||\displaystyle\sum_{k=1}^{i} \Delta_i|| \leqslant \eta\varepsilon_1$ *with* $\eta = \dfrac{1}{1-\theta}$.

Proof. The proof is obvious. □

LEMMA 9.2. *Let us denote* $a_1(u) = \dfrac{1}{1+\sqrt{1-u^2}}$ *and* $a_2(u) = \dfrac{a_1(u) - 1/2}{u^2}$. *We have*

1. $|e_p(u)| = \displaystyle\sum_{k=1}^{max\{k\,:\,2k\leqslant p\}} |c_k| u^{2k} \leqslant u^2 a_1(u)$.
2. $|d_p(u)| = \displaystyle\sum_{k=2}^{max\{k\,:\,2k\leqslant p\}} |c_k| u^{2k} \leqslant u^4 a_2(u) = u^2 \left(a_1(u) - \dfrac{1}{2}\right)$.

Proof. It follows from classical Taylor series expansion. □

LEMMA 9.3. *Let* $b_1(u) = \dfrac{u^2 a_1(u)^2}{\sqrt{1-u^2}} + 2a_1(u)$ *and* $b_2(u) = \dfrac{a_1(u)^2}{\sqrt{1-u^2}} + 2a_2(u)$. *We have*

$$
(x+y)^{2i} a_i(x+y) - x^{2i} a_i(x) - y^{2i} a_i(y) \leqslant \qquad b_i(x+y) xy (x+y)^{2i-2}.
$$

Proof. To prove the case $i = 1$ we write

$$
\begin{aligned}
&(x+y)^2 a_1(x+y) - x^2 a_1(x) - y^2 a_1(y) \\
&= x^2(a_1(x+y) - a_1(x)) + y^2(a_1(x+y) - a_1(y)) + 2xy a_1(x+y) \\
&= \left(\frac{(2x+y)x a_1(x)}{\sqrt{1-x^2} + \sqrt{1-(x+y)^2}} + \frac{(2y+x)y a_1(y)}{\sqrt{1-y^2} + \sqrt{1-(x+y)^2}} + 2 \right) xy a_1(x+y)
\end{aligned}
$$

Using $y \leqslant x$, $a_1(y) \leqslant a_1(x)$ and $\sqrt{1-x^2}, \sqrt{1-y^2} \leqslant \sqrt{1-(x+y)^2}$ we get

$$\begin{aligned}(x+y)^2 a_1(x+y) - x^2 a_1(x) - y^2 a_1(y) &\leqslant \left(\frac{(x+y)^2 a_1(x+y)}{\sqrt{1-(x+y)^2}} + 2\right) xy a_1(x+y)\\ &= b_1(x+y)xy.\end{aligned}$$

To prove the case $i = 2$ we write from definition of $a_2(u)$:

$$\begin{aligned}(x+y)^4 a_2(x+y) - x^4 a_2(x) - y^4 a_2(y) &= (x+y)^2 a_1(x+y) - x^2 a_1(x) - y^2 a_1(y) - xy\\ &\leqslant \left(\frac{(x+y)^2 a_1(x+y)^2}{\sqrt{1-(x+y)^2}} + 2a_1(x+y) - 1\right) xy\\ &\leqslant \left(\frac{a_1(x+y)^2}{\sqrt{1-(x+y)^2}} + 2a_2(x+y)\right) xy(x+y)^2\\ &\leqslant b_2(x+y)xy(x+y)^2.\end{aligned}$$

We are done. □

LEMMA 9.4. *Let* $C_{p-1} = X_1 + \cdots + X_{p-1}$. *Let us suppose* $q \geqslant 2(\theta+\eta)\kappa$, $v = q\varepsilon_1 \leq \theta < 1$, $\eta = \dfrac{1}{1-\theta}$ *and* $\|X_k\| \leqslant \dfrac{\kappa}{q} v^k$, $1 \leq k \leq p-1$. *Then we have*

1. $\|C_{p-1}\| \leqslant \eta\kappa\varepsilon_1$.
€
2. $\|e_p(C_{p-1})\| \leqslant a_1(\eta\kappa\varepsilon_1)\eta^2\kappa^2\varepsilon_1^2$.

3. $||e_p(X_p)|| \leqslant a_1(\theta\kappa\varepsilon_1)\kappa^2 q^{2(p-1)}\varepsilon_1^{2p}$.

Proof. We have

$$\|C_{p-1}\| \leqslant \sum_{k=1}^{p-1} \|X_k\| \leqslant \sum_{k=1}^{p-1} \kappa q^{k-1}\varepsilon_1^k \leqslant \frac{1}{1-v}\kappa\varepsilon_1 \leqslant \eta\kappa\varepsilon_1.$$

From Lemma 9.2 we know that $|e_p(u)| \leqslant u^2 a_1(u)$. Since $q \geqslant 2(\theta+\eta)\kappa$ and $\varepsilon_1 \leqslant \dfrac{\theta}{q}$ it follows that $\eta\kappa\varepsilon_1 \leqslant \dfrac{\eta\theta}{2(\eta+\theta)} = \dfrac{\theta}{2(1+\theta-\theta^2)}$, we can see the quantity $a_1(\eta\kappa\varepsilon_1)$ is well defined when $\eta\kappa\varepsilon_1 \leqslant 1$. That is to say $\dfrac{\theta}{2(1+\theta-\theta^2)} \leqslant 1$. This is the case since $\theta < 1$. It follows

$$\|e_p(C_{p-1})\| \leqslant a_1(\eta\kappa\varepsilon_1)\,(\eta\kappa\varepsilon_1)^2.$$

We now bound $\|e_p(X_p)\|$. Always from Lemma 9.2 we have

$$\begin{aligned}\|e_p(X_p)\| &\leqslant a_1(\kappa q^{p-1}\varepsilon_1^p)(\kappa q^{p-1}\varepsilon_1^p)^2\\ &\leqslant a_1(\theta\kappa\varepsilon_1)\kappa^2 q^{2(p-1)}\varepsilon_1^{2p} \quad \text{since} \quad q\varepsilon_1 \leqslant \theta < 1.\end{aligned}$$

We are done. □

LEMMA 9.5. *Let us suppose* $2(\theta+\eta)\kappa \leq q$, $v = q\varepsilon_1 \leq \theta$ *and* $\|X_k\| \leqslant \frac{\kappa}{q}v^k$, $1 \leq k \leq p-1$. *Then we have*

$$\|d_p(C_{p-1})\| \leqslant a_2(\eta\kappa\varepsilon_1)\eta^4\kappa^4\varepsilon_1^4$$

and

$$||d_p(X_p)|| \leqslant a_2(\theta\kappa\varepsilon_1)\kappa^4 q^{4(p-1)}\varepsilon_1^{4p}.$$

Proof. The proof is like to that of Lemma 9.4. ∎

LEMMA 9.6. *Let us suppose* $2(\theta+\eta)\kappa \leq q$, $v = q\varepsilon_1 \leq \theta$ *and* $\|X_k\|, \|Y_k\| \leqslant \frac{\kappa}{q}v^k$, $1 \leq k \leq p$. *Then we have*

$$\|\Theta_{p-1}\| \leqslant (1+\eta\kappa\varepsilon_1 a_1(\eta\kappa\varepsilon_1))\eta\kappa\varepsilon_1.$$

Proof. We have $\|\Theta_{p-1}\| \leqslant \|C_{p-1}\| + \|e_p(C_{p-1})\|$. Using $\|C_{p-1}\| \leqslant\leq \eta\kappa\varepsilon_1$ and Lemma 9.4 the conclusion follows. ∎

LEMMA 9.7. *Let us suppose* $2(\theta+\eta)\kappa \leq q$, $v = q\varepsilon_1 \leq \theta$ *and* $\|X_k\| \leqslant \frac{\kappa}{q}v^k$, $1 \leq k \leq p$. *Let*

$$Q_{p,i} = \sum_{k=i}^{\max(k:2k\leqslant p)} c_k \sum_{\substack{i_1+i_2=2k \\ i_1,\, i>0}} L_{i_1,i_2}(C_{p-1}, X_p), \qquad i = 1, 2.$$

We have

$$\|Q_{p,i}\| \leqslant b_i(\eta\kappa\varepsilon_1)\eta^{2i-1}\kappa^{2i}q^{p-1}\varepsilon_1^{p+2i-1} \quad i = 1, 2.$$

Proof. Let $\|C_{p-1}\| \leqslant x$ and $\|X_p\| \leqslant y$. We have using Lemma 9.2 :

$$\begin{aligned}
\|Q_{p,i}\| &\leqslant \sum_{k=i}^{\max(k:2k\leqslant p)} |c_k| \sum_{\substack{i_1+i_2=2k \\ i_1>0, i_2>0}} \frac{(2k)!}{i_1!i_2!}x^{i_1}y^{i_2} \\
&\leqslant \sum_{k\geqslant i} |c_k|((x+y)^{2k} - x^{2k} - y^{2k}) \\
&\leqslant (x+y)^{2i}a_i(x+y) - x^{2i}a_i(x) - y^{2i}a_i(y).
\end{aligned}$$

We apply the Lemma 9.3 with the bounds $y \leqslant \frac{\kappa}{q}v^p \leqslant \kappa q^{p-1}\varepsilon_1^p$ and $x \leqslant x+y \leqslant \frac{\kappa}{q}\frac{v}{1-v} \leqslant \eta\kappa\varepsilon_1$. We then get :

$$\|Q_{p,1}\| \leqslant b_i(\eta\kappa\varepsilon_1)\eta^{2i-1}\kappa^{2i}q^{p-1}\varepsilon_1^{p+2i-1}.$$

The result follows. ∎

LEMMA 9.8. *Let* $\|X_p\|, \|Y_p\| \leqslant \kappa q^{p-1}\varepsilon_1^p$, $2(\theta+\eta)\kappa \leq q$ *and* $q\varepsilon_1 \leqslant \theta < 1$. *Then*

$$|| - X_p\Sigma d_p(Y_p) + d_p(X_p)\Sigma Y_p|| \leqslant 2Ka_2(\theta\kappa\varepsilon_1))\kappa^5 q^{5(p-1)}\varepsilon_1^{5p}.$$

Proof. Let $Z_p := -X_p\Sigma d_p(Y_p) + d_p(X_p)\Sigma Y_p$. Then from Lemma 9.5 we deduce

$$\|Z_p\| \leqslant 2Ka_2(\theta\kappa\varepsilon_1)\kappa^4 q^{5(p-1)}\varepsilon_1^{5p}.$$

We are done. □

LEMMA 9.9. *For $|u| < 1$ we have*

$$|(1+c_p(-u))(1+c_p(u)) - 1| \leqslant \left(2\sqrt{1+u^2} + a_1(u)u^{p+1}\right) a_1(u)u^{p+\delta}$$

where $\delta = 1$ if p is odd and $\delta = 2$ if p is even.

Proof. Remember that $e(u) = \sqrt{1+u^2} + u - 1$ and $e(u) = c_p(u) + r_p(u)$. Since $(1+e(u))(1+e(-u)) = 1$ and $r_p(u) = r_p(-u)$ it follows

$$\begin{aligned}
(1+c_p(-u))(1+c_p(u)) - 1 &= (1+e(-u)-r_p(-u))(1+e(u)-r_p(u)) - 1\\
&= (1+e(-u))(1+e(u)) - 1\\
&\quad - (1+e(-u))r_p(u) - (1+e(u))r_p(u) + r_p(u)^2\\
&= -(2+e(u)+e(-u)-r_p(u))\,r_p(u)\\
&= -\left(2\sqrt{1+u^2} - r_p(u)\right) r_p(u)
\end{aligned}$$

We have

$$\begin{aligned}
|r_p(u)| &\leqslant \sum_{i>\max\{k:2k\leqslant p\}} |c_{p,i}|u^{2i} =\\
&\leqslant \frac{1}{1+\sqrt{1-u^2}}u^{p+\delta} = a_1(u)u^{p+\delta}
\end{aligned}$$

where $\delta = 1$ if p is odd and $\delta = 2$ if p is even. We deduce that

$$|(1+c_p(-u))(1+c_p(u)) - 1| \leqslant \left(2\sqrt{1+u^2} + a_1(u)u^{p+\delta}\right) a_1(u)u^{p+\delta}.$$

We are done. □

LEMMA 9.10. *For $i \geqslant 0$, we have*

$$s_i := \sum_{k=0}^{i-1} 2^{-(p+1)^k+1} \leqslant 2 - 2^{2-(p+1)^i}.$$

Proof. We proceed by induction. The assertion holds for $i = 0$. By assuming for i let us prove it for $i+1$. We have

$$\begin{aligned}
s_{i+1} &\leqslant 2 - 2^{2-(p+1)^i} + 2^{-(p+1)^i+1} \leqslant 2 - 2^{2-(p+1)^i}(1-2^{-1}) = 2 - 2^{2-(p+1)^i-1}\\
&\leqslant 2 - 2^{2-(p+1)^{i+1}} \text{ since } \quad (p+1)^i + 1 \leqslant 2(p+1)^i \leqslant (p+1)^{i+1}.
\end{aligned}$$

We are done. □

10. Proof of Davies-Smith Theorem 2.1. Let us denote $\Delta_1 = U^*\Sigma V - \Sigma$ and $\Delta_2 = (I_\ell + \Theta_1^*)(\Delta_1 + \Sigma)(I_q + \Psi_1) - \Sigma - S_1$ with $\Theta_1 = X_1 + X_1^2/2$ and $\Psi_1 = Y_1 + Y_1^2/2$. From the definition of the map DS we have $U_1 = U(I_\ell + X_1 + X_2 + X_1^2/2)$, $V_1 = V(I_q + Y_1 + Y_2 + Y_1^2/2)$, $\Sigma_1 = \Sigma + S_1 + S_2$ where for $i = 1, 2$, one has $S_i = \mathrm{diag}(\Delta_i)$ and the X_i's are skew Hermitian matrices be such that $\Delta_i - S_i - X_i\Sigma + \Sigma Y_i = 0$. The goal is to bound the norm of $\Delta_3 := U_1^* M V_1 - \Sigma_1 = (I_\ell + \Theta_1^* - X_2)(\Delta_1 + \Sigma)(I_q + \Psi_1 + Y_2) - \Sigma - S_1 - S_2$. We first expand Δ_2 and as in the proof of Proposition 7.1 we have $\|\Delta_2\| \leqslant q_1\varepsilon_1^2$ where

$$q_1 = 2\kappa + 2\kappa^2\varepsilon_1 + \frac{5}{4}\kappa^4 K\varepsilon_1^2 + \frac{1}{4}\kappa^4\varepsilon_1^3, \tag{10.1}$$

and $q_1\varepsilon_1 \leqslant \tau_1\varepsilon$ with $\tau_1 = 2 + 2\varepsilon + \frac{5}{4}\varepsilon^2 + \frac{1}{4}\varepsilon^3$. We now expand Δ_3 to get :

$$\begin{aligned} \Delta_3 &= (I_\ell + \Theta_1^* - X_2)(\Delta_1 + \Sigma)(I_n + \Psi_1 + Y_2) - \Sigma - S_1 - S_2 \\ &= (I_\ell + \Theta_1^*)(\Delta_1 + \Sigma)(I_n + \Psi_1) - \Sigma - S_1 - S_2 \\ &\quad + (I_\ell + \Theta_1^*)(\Delta_1 + \Sigma)Y_2 - X_2(\Delta_1 + \Sigma)(I_n + \Psi_1) - X_2(\Delta_1 + \Sigma)Y_2 \end{aligned} \tag{10.2}$$

We know that

$$(I_\ell + \Theta_1^*)(\Delta_1 + \Sigma)(I_n + \Psi_1) - \Sigma - S_1 - S_2 = \Delta_2 - S_2 = X_2\Sigma - \Sigma Y_2.$$

Plugging the previous relation in (10.2) we find

$$\Delta_3 = -X_2\Delta_1 + \Delta_1 Y_2 - X_2\Delta_1 Y_2 + \Theta_1^*(\Delta_1 + \Sigma)Y_2 - X_2(\Delta_1 + \Sigma)\Psi_1 - X_2\Sigma Y_2 \tag{10.3}$$

We are going to prove $\|\Delta_3\| \leqslant q_1 q_2 \varepsilon_1^3$ where q_2 is defined below in (7.16). To do that we will use the bounds

1. $\|\Delta_2\| \leqslant q_1\varepsilon_1^2$ and $\|X_2\|, \|Y_2\| \leqslant \kappa q_1\varepsilon_1^2$.
2. $\|\Theta_1\|, \|\Psi_1\| \leqslant \left(1 + \frac{1}{2}\kappa\varepsilon_1\right)\kappa\varepsilon_1$.

Considering the bounds of the norms of matrices given in (10.3), we get $||\Delta_3|| \leqslant q_3 q_1 \varepsilon_1^3$ where

$$q_3 = 2\kappa(K\kappa + 1) + (K\kappa + 2 + Kq_1)\kappa^2\varepsilon_1 + (\kappa + q_1)\kappa^2\varepsilon_1^2.$$

A straighforward calculation shows that if $\varepsilon_1 \leqslant \dfrac{\varepsilon}{\kappa^{5/4}K^{2/5}}$ then

$$||\Delta_3|| \leqslant q_3 q_1 \varepsilon_1^3 \leqslant \tau_3\tau_1\varepsilon^3 \tag{10.4}$$

where

$$\tau_3 = 4 + (3 + \tau_1)\varepsilon + (1 + \tau_1)\varepsilon^2.$$

A straightforward computation shows that for all $\varepsilon \leqslant 0.1$ we have

$$\tau_3\tau_1 \leqslant 8 + 18\varepsilon + 28\varepsilon^2.$$

We finally get

$$\kappa^{5/4}K^{2/5}\|\Delta_3\| \leqslant (8 + 18\varepsilon + 33\varepsilon^2)\varepsilon^3.$$

Then the part 1 of Theorem 2.1 is proved.

We use the proof of Proposition 7.1 to proof the part 2 of Theorem. We have

$$\|\bar{U}_1^* M \bar{V}_1 - \bar{\Sigma}_1\| \leqslant q_2 q_1 \varepsilon_1^3$$

where q_1 is defined in (7.7) and q_2 in (7.16). A straightforward calculation shows that if $\varepsilon_1 \leqslant \dfrac{\varepsilon}{\kappa^{6/5} K^{3/10}}$ then

$$\|\bar{U}_1^* M \bar{V}_1 - \bar{\Sigma}_1\| \leqslant q_2 q_1 \varepsilon_1^3 \leqslant \tau_2 \tau_1 \varepsilon^3 \tag{10.5}$$

where $\tau = \tau_1 \tau_2$ given in (7.3). Moreover $\tau_2 \tau_1 \leqslant 6 + 21\varepsilon + 54\varepsilon^2$ for $\varepsilon \leqslant 0.1$. This proves te part 2. The Theorem holds. □

11. Application in the clusters case.

11.1. Definiton of Clusters and first properies. We use the Fortran or Matlab notation for submatrices, i.e., $A_{i:j,k:l}$ is the submatrix of A with lines and columns between the subscripts i, j and k, lrespectively. We consider e integers q_i's such that $\displaystyle\sum_{i=1}^{e} q_i = q$. We also associate the integers ℓ_i, $1 \leqslant i \leqslant e$, defined by $\ell_i = 1 + \displaystyle\sum_{j=1}^{i-1} q_j$ The first goal is to precise the notion of cluster of singular values.

DEFINITION 11.1. *Let e integers q_i's such that $\displaystyle\sum_{i=1}^{e} q_i = q$. We associate the integers ℓ_i, $1 \leqslant i \leqslant e$, defined by $\ell_i = 1 + \displaystyle\sum_{j=1}^{i-1} q_j$. From $\Delta \in \mathbb{C}^{\ell \times q}$ with $\ell \geqslant q$, we consider its sub-matrices $\Delta_i := \Delta_{\ell_i:\ell_{i+1}-1, \ell_i:\ell_{i+1}-1} \in \mathbb{C}^{q_i \times q_i}$, $1 \leqslant i \leqslant e$. We define the matrix*

$$\mathrm{Diag}_{q_1 \cdots q_e}(\Delta) = \begin{pmatrix} \Delta_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Delta_e \\ & 0 & \end{pmatrix}$$

We name by $\mathbb{D}^{\ell \times q}_{q_1, \ldots, q_e}$ the set of these matrices.

DEFINITION 11.2. *Let integers q_i's and ℓ_i's be as in Definition 11.1. Let $\delta \geqslant 0$ and define the set $\mathbb{D}^{\ell \times q}_{q_1 \cdots q_e}(\delta)$ of the matrices whose diagonal $\Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_q) \in \mathbb{D}^{\ell \times q}$ satisfies*

$$|\sigma_k - \sigma_j| \leqslant \delta \qquad \ell_i \leqslant j, k \leqslant \ell_{i+1} - 1, \quad 1 \leqslant i \leqslant e \tag{11.1}$$

$$|\sigma_j - \sigma_l| > \delta, \qquad \ell_i \leqslant j \leqslant \ell_{i+1} - 1, \quad \ell_k \leqslant l \leqslant \ell_{k+1} - 1, \quad 1 \leqslant i < k \leqslant e \tag{11.2}$$

We name $\mathbb{D}^{\ell \times q}_{q_1 \ldots q_e}(\delta)$ the set of clusters of size δ relatively to integers $q_1, \cdots, q_e$. We also name by $\mu = (q_1, \ldots, q_e)$ the multiplicity of cluster associated to Σ.

We have

PROPOSITION 11.3. *Let $\delta \geqslant 0$ and $\Delta \in \mathbb{D}^{\ell \times q}_{q_1 \cdots q_e}(\delta)$. The tuple $(q_1, \cdots, q_e)$ where each integer $q_i \geqslant 1$ is the only one such that the inequalities (11.1-11.2) hold.*

Proof. Let us suppose there exists two tuples $(m_1, \cdots, m_d)$ and $(q_1, \cdots, q_e)$ such that the inequalities (11.1-11.2) hold for the diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_q)$. Let us suppose for instance $m_1 < q_1$. Then we first have from the inequality (11.2) : $|\sigma_{m_1} - \sigma_{m_1+1}| > \delta$. In the other hand, since $m_1 < q_1$ we can write from the inequality (11.1) $|\sigma_{m_1} - \sigma_{m_1+1}| \leqslant \delta$. This is not possible and the proposition holds. □

11.2. Solving $\Delta - S - X\Sigma + \Sigma Y = 0$ in the clusters case. We state without proof the result that is generalizes the Proposition 4.1.

PROPOSITION 11.4. *Let $\Sigma \in \mathbb{D}^{\ell\times q}_{q_1\cdots q_e}(\delta)$ and $\Delta = (\delta_{i,j}) \in \mathbb{C}^{\ell\times q}$. Consider the matrix $S \in \mathbb{D}^{\ell\times q}_{q_1\cdots q_e}$ and the two skew Hermitian matrices $X = (x_{i,j}) \in \mathbb{C}^{\ell\times\ell}$ and $Y = (y_{i,j}) \in \mathbb{C}^{q\times q}$ that are defined by the following formulas:*

1. *The matrix S is defined by*

$$S = \mathrm{Diag}_{q_1\cdots q_e}(\Delta) \in \mathbb{D}^{\ell\times q}_{q_1\cdots q_e} \tag{11.3}$$

2.

$$\mathrm{Diag}_{q_1\cdots q_e}(X) = 0 \tag{11.4}$$

$$\mathrm{Diag}_{q_1\cdots q_e}(Y) = 0 \tag{11.5}$$

3. *For $1 \leqslant i < k \leqslant e$, $1 \leqslant j \leqslant q_i - 1$, and $1 \leqslant l \leqslant q_k - 1$ we take*

$$x_{\ell_i+j,\ell_k+l} = \frac{1}{2}\left(\frac{\delta_{\ell_i+j,\ell_k+l} + \overline{\delta_{\ell_k+l,\ell_i+j}}}{\sigma_{\ell_k+l} - \sigma_{\ell_i+j}} + \frac{\delta_{\ell_i+j,\ell_k+l} - \overline{\delta_{\ell_k+l,\ell_i+j}}}{\sigma_{\ell_k+l} + \sigma_{\ell_i+j}}\right) \tag{11.6}$$

$$y_{\ell_i+j,\ell_k+l} = \frac{1}{2}\left(\frac{\delta_{\ell_i+j,\ell_k+l} + \overline{\delta_{\ell_k+l,\ell_i+j}}}{\sigma_{\ell_k+l} - \sigma_{\ell_i+j}} - \frac{\delta_{\ell_i+j,\ell_k+l} - \overline{\delta_{\ell_k+l,\ell_i+j}}}{\sigma_{\ell_k+l} + \sigma_{\ell_i+j}}\right) \tag{11.7}$$

4. *For $q+1 \leqslant i \leqslant \ell$ and $1 \leqslant j \leqslant q$, we take*

$$x_{i,j} = \frac{1}{\sigma_j}\delta_{i,j}. \tag{11.8}$$

5. *For $q+1 \leqslant i \leqslant \ell$ and $q+1 \leqslant j \leqslant \ell$, we take*

$$x_{i,j} = 0. \tag{11.9}$$

Then we have

$$\Delta - S - X\Sigma + \Sigma Y = 0. \tag{11.10}$$

DEFINITION 11.5. *Under the previous framework, we name condition number of equation $\Delta - S - X\Sigma + \Sigma Y = 0$ the quantity*

$$\kappa(\Sigma) = \max\left(1, \max_{1\leqslant i\leqslant e}\frac{1}{|\sigma_i|}, \max_{\substack{1\leqslant i<k\leqslant e\\ |\sigma_k-\sigma_i|>\delta}} \frac{1}{|\sigma_k-\sigma_i|} + \frac{1}{|\sigma_k+\sigma_i|}\right|\right) \tag{11.11}$$

The analysis of error is given by the following result.

PROPOSITION 11.6. *Under the notations and assumptions of Proposition 11.4, assume that S, X and Y are computed using (11.3–11.9). Given ε with $\|\Delta\| \leqslant \varepsilon$, the matrices X, Y and S solutions of $\Delta - S - X\Sigma + \Sigma Y = 0$ satisfy*

$$\|S\| \leqslant \varepsilon \tag{11.12}$$

$$\|X\|, \|Y\| \leqslant \kappa\varepsilon \tag{11.13}$$

11.3. Method of order p+1 in the clusters case. Let $p \geqslant 2$ and $\mathbb{E}_{q_1,\dots,q_e}^{m\times\ell,n\times q} = \mathbb{C}^{m\times\ell} \times \mathbb{C}^{n\times q} \times \mathbb{D}_{q_1,\dots,q_e}^{m\times n}$.We denote $E_\ell(U) = U^*U - I_\ell$, $E_q(V) = V^*V - I_q$, $\Delta = U^*MV - \Sigma$ and we define the map H_p by

(11.14)

$$(U, V, \Sigma) \in \mathbb{E}_{q_1,\dots,q_e}^{m\times\ell,n\times q} \to \quad H_p(U, V, \Sigma) = \begin{pmatrix} U(I_\ell + \Omega)(I_\ell + \Theta) \\ V(I_q + \Lambda)(I_q + \Psi) \\ \Sigma + S \end{pmatrix} \in \mathbb{E}_{q_1,\dots,q_e}^{m\times\ell,n\times q}$$

where :

1. $\Omega = s_p(E_\ell(U))$ and $\Lambda = s_p(E_q(V))$.
2. $S = S_1 + \cdots + S_p \in \mathbb{D}_{q_1\dots q_l}^{m\times n}$, $X = X_1 + \cdots + X_p$ and $Y = Y_1 + \cdots + Y_p$ with each X_k, Y_k are skew Hermitian matrices. Moreover each triplet (S_k, X_k, Y_k) are solutions of the following linear systems :

$$\Delta_k - S_k - X_k\Sigma + \Sigma Y_k = 0, \qquad 1 \leqslant k \leqslant p$$

where the Δ_k's for $2 \leqslant k \leqslant p+1$, are defined as

$$(11.15)\quad \begin{aligned} &\Delta_1 = (I_\ell + \Omega)(\Delta + \Sigma)(I_q + \Lambda) - \Sigma, e \quad S_1 = \mathrm{Diag}_{q_1,\dots,q_e}(\Delta_1) \\ &\Theta_k = c_p(X_1 + \cdots + X_k), \quad \Psi_k = c_p(Y_1 + \cdots + Y_k), \quad 1 \leqslant k \leqslant p, \\ &\Delta_k = (I_\ell + \Theta_{k-1}^*)(\Delta_1 + \Sigma)(I_q + \Psi_{k-1}) - \Sigma - \sum_{l=1}^{k-1} S_l, \\ &S_k = \mathrm{Diag}_{q_1,\dots,q_e}(\Delta_k), \quad 2 \leqslant k \leqslant p. \end{aligned}$$

11.4. Result of convergence in the clusters case.

THEOREM 11.7. *If the sequence define by*

$$(U_{i+1}, V_{i+1}, \Sigma_{i+1}) = H_p(U_i, V_i, \Sigma_i), \quad i \geqslant 0$$

from $(U_0, V_0, \Sigma_0) \in \mathbb{E}_{q_1,\dots,q_e}^{m\times\ell,n\times q}$ *verifies the asumptions of Theorem 1.2 then it converges at the order* $p+1$ *to* $(U_\infty, V_\infty, \Sigma_\infty) \in \mathrm{St}_{m,\ell} \times \mathrm{St}_{n,q} \times \mathbb{D}_{q_1,\dots,q_e}^{m\times n}$ *such that* $U_\infty^* M V_\infty - \Sigma_\infty = 0$.

Proof. The proof is similar to that of Theorem 1.2. ∎

11.5. Deflation method for the SVD. The sequence $(U_i, V_i, \Sigma_i)_{i\geqslant 0}$ of Theorem 11.7 is not a SVD sequence since the Σ_i's belong to $\mathbb{D}_{q_1,\dots,q_e}^{m\times n}$. We can use the Theorem 1.2 to detect the presence of clusters of singular values.

To simplify the presentation we suppose $m = n$ in order that

$$\kappa(\Sigma) = \max\left(1, \max_{1\leqslant i<j\leqslant n} \frac{1}{|\sigma_i - \sigma_j|} + \frac{1}{|\sigma_i + \sigma_j|}\right).$$

To do that we introduce an index of deflation whose the existence is given by the following proposition.

PROPOSITION 11.8. *Let us consider* $(U_0, V_0, \Sigma_0) \in \mathbb{E}_{m\times m}^{m\times m}$ *and* $\Delta_0 = U_0^* M V_0 - \Sigma_0$. *Let*

$$e = \max\left(\frac{K^{a-1}\|\Delta_0\|}{u_0}, \frac{K^a}{u_0}\|E_m(U)\|, \frac{K^a}{u_0}\|E_m(V)\|\right)^{1/a}$$

Let us suppose $e \leqslant 1$. Then there exists an index $q \leqslant m$ be such that we can rewrite the diagonal matrix Σ_0 under the form $\begin{pmatrix} \Sigma_{0,q} & \\ & \Sigma_{0,n-q} \end{pmatrix}$ where $\kappa(\Sigma_{0,q})e \leqslant 1$. Let us consider $U_{0,q}$ and $V_{0,q}$ the sub matrices of U_0 and V_0 respectively corresponding to $\Sigma_{0,q}$. Then Theorem 1.2 applies for the sequence define from $(U_{0,q}, V_{0,q}, \Sigma_{0,q}) \in \mathbb{E}_{m\times q}^{m\times q}$ by $(U_{i+1,q}, V_{i+1,q}, \Sigma_{i+1,q}) = H_p(U_{i,q}, V_{i,q}, \Sigma_{i,q})$, $i \geqslant 0$.

Proof. The existence of the index q is obvious since q is at least equal at 1. In this case $\kappa(\Sigma_{0,1}) = 1$. ∎

DEFINITION 11.9. *Let us consider the notations and the assumption of Proposition 11.8.We name indice of deflation of (U_0, V_0, Σ_0) the maximum of indices q such that $\kappa(\Sigma_{0,q})e \leqslant 1$. If q is the index of deflation we name $(U_{0,q}, V_{0,q}, \Sigma_{0,q})$ a deflation of (U_0, V_0, Σ_0)*

To determine the index of deflation and a deflation of (U_0, V_0, Σ_0), we propose the following algorithm. We denote $\kappa_{i,j} = \max\left(1, \frac{1}{|\sigma_i - \sigma_j|} + \frac{1}{|\sigma_i + \sigma_j|}\right)$. Following the matlab notation if A is a matrix and k a vector of indices $A(:,k)$ means the matrix composed by the columns indexed by the vector k. Moreover $\#k$ is the size of k.

(11.16) **Algorithm to determine the index of deflation**

Input (U_0, V_0, Σ_0) such that $e \leqslant 1$
Ouput $(U_{0,q}, V_{0,q}, \Sigma_{0,q})$ a deflation of (U_0, V_0, Σ_0)
1. Let $\Sigma_0 = \mathrm{diag}(\sigma_{0,1}, \ldots, \sigma_{0,n})$ where $\sigma_{0,1} \geqslant \cdots \geqslant \sigma_{0,n}$
2. $k = 1$ $\quad i = 1$
3. while $i \leqslant m$ do
4. $\quad j = 1$
5. $\quad$ while $i + j \leqslant n$ and $\kappa_{i,i+j}e > 1$ do $\quad j = j + 1$ $\quad$ end while
6. $\quad$ if $i + j \leqslant n$ and $\kappa_{i,i+j} \leqslant 1$ then $\quad k = [k, i + j]$ $\quad$ end if
7. $\quad i = i + j$
8. end while
9. $q = \#k$
10. $\Sigma_{0,q} = \Sigma_0(k) \qquad U_{0,q} = U_0(k) \qquad V_{0,q} = V_0(k)$

THEOREM 11.10. *Let (U_0, V_0, Σ_0) that satisfies the Proposition 11.8. The algorithm 11.16 computes a deflation of (U_0, V_0, Σ_0).*

Proof. When $k = 1$ we have $\kappa(\Sigma_0(:,1)) = 1$ and $\kappa(\Sigma_0(:,1))e \leqslant 1$ from assumption. The loop 3-8 of the algorithm consists to determine an ordered list of indices k such that for all $i \in k$ such that $i + 1 \in k$ we have $\kappa_{i,i+1}e \leqslant 1$. Hence $\kappa(\Sigma_{0,q})e \leqslant 1$ and the Theorem follows. ∎

12. Numerical Experiments. Our numerical experiments are done with the `Julia Programming Language` [3] coupled with the library `ArbNumerics` of Jeffrey Sarnoff. To intialize our method we proceed in two steps

1. The triplet (U_0, V_0, Σ_0) is given by the function `svd` of `Julia` with 64-bit of precision unless otherwise stated.
2. From this (U_0, V_0, Σ_0) we determine $(U_{0,q}, V_{0,q}, \Sigma_{0,q})$ by the Algorithm 11.16.

We consider for $i \geqslant 0$ the quantities

$$\varepsilon_i = \max((\kappa_i K_i)^a \|E_\ell(U_i)\|, (\kappa_i K_i)^a \|E_q(V_i)\|, \kappa_i^a K_i^{a-1} \|\Delta_i\|)$$

where a, u_0 are defined in Theorem 1.2. All the Tables below show the behaviour of $e_i = -\lfloor \log_2(\varepsilon_i/u_0) \rfloor$.

The strategy of practical computations is to initialize the method with q bits of precision. Next the iteration i is done with $q(p+1)^i$ bits of precision. This setting of precision is done efficiently thanks to the library `ArbNumerics` at each iteration.

12.1. Random matrices. Table 3 confirms the behaviour of iterates expected by the convergence analysis.

Iterations / Order	2	3	4	5	6	7
0	7	8	9	8	8	8
1	18	35	47	59	69	85
2	44	112	194	311	427	604
3	92	346	787	1571	2580	4353

Table 3

12.2. Cauchy matrices. The classical Cauchy matrix is defined by

$$M = \left(\frac{1}{i+j}\right)_{1\leqslant i,j\leqslant n}.$$

Its singular values satisfy the inequalities $\sigma_{1+k} \geqslant 4\left(\exp\left(\frac{\pi^2}{2\,\mathrm{Log}(4n)}\right)\right)^{-2k}\sigma_1$ where σ_1 is the greatest singular values [5]. There is a strong decrease of singular values to 0. The computation of a deflation by the Algorithm 11.16 gives different values of q for $\Sigma_{0,q}$ following the value of p . For instance with 64-bit of precision and $n=200$, if $p=1$ then $q=11$: $\Sigma_{0,q}$ is constituted of the first ten singular values and one among the other 190's. If $p\geqslant 2$ then $q=15$: $\Sigma_{0,q}$ is constituted of the first fourteen singular values and one among the other 185's. Table 4 gives the behaviour of iterates from a computation of a deflation.

Iterations / Order	2	3	4	5	6	7
0	1	1	1	1	1	1
1	9	19	19	35	36	51
2	31	67	116	196	277	389
3	74	214	503	1003	1724	2757

Table 4

Table 5 gives the necessary precision that we need to get the size of Cauchy matrices as index of deflation.

n	$n\leqslant 7$	$8\leqslant n\leqslant 14$	$15\leqslant n$
bits precision	64	128	$\geqslant 256$

Table 5

12.3. Matrices with prescribed singular values. Let us define $M = U\Sigma V$ where U and V are two unitary matrices of size $4n \times 4n$ and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{4n})$ where

$$\sigma_{3(i-1)+j} = 2^i \qquad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant 3,$$
$$\sigma_{3n+i} = 2^{-i} \qquad 1 \leqslant i \leqslant n.$$

The condition $e \leqslant 1$ of the Proposition 11.8 holds if $\left(\frac{4 \times 2^n}{3}\right)^a \varepsilon_0 \leqslant u_0$ where $\varepsilon_0 = \max(\|\Delta_0\|, \|E_m(U_0)\|, \|E_m(V_0)\|)$. Table 6 gives the quantity $-\left\lfloor \log_2 \frac{3^a u_0}{4^a 2^{na}} \right\rfloor$ with respect n. For instance a C matrix of size 100×100, Proposition 11.8 applies if $\varepsilon_0 \leqslant 2^{-139}$ for $p \geqslant 2$ and for $p = 1$, it is necessary to have $\varepsilon_0 \leqslant 2^{-206}$. Hence the precision required on ε_0 to get

$p/4n$	4	20	40	60	80	100	120	140	160	180
$p = 1$	14	46	86	126	166	206	246	286	326	366
$p \geqslant 2$	11	33	59	86	113	139	166	193	219	246

Table 6

a deflation is greater in the case $p = 1$ than for $p \geqslant 2$. This is confirmed by numerical experimentation. If $p = 1$ then $n \leqslant 26$ (respectively if $p \geqslant 2$ then $n \leqslant 41$) a 64-bits precision is enough so that Proposition 11.8 holds. Table 7 shows for $p = 1$ (respectively $p \geqslant 2$) the quantities $q_+ = \#\{\sigma > 1\}$ and $q_- \#\{\sigma > 1\}$ from a $\Sigma_{0,q}$ given by the initialization. In each case of Table 7 the first number matches for q_+ and the second for q_-. The 64-bit precision used for $p = 1$ (respectively $p \geqslant 2$) until the size 100 (respectively 140). For larger sizes, 128-bits precision are used. The quantity q_+ is always equal to n which is the number of multiple singular values.

$q_+, q_-/4n$	4	20	40	60	80	100	120	140	160
$p = 1$	$1, 1$	$5, 5$	$10, 10$	$15, 10$	$20, 5$	$25, 1$	$30, 26$	$35, 21$	$40, 16$
$p \geqslant 2$	$1, 1$	$5, 5$	$10, 10$	$15, 15$	$20, 18$	$25, 13$	$30, 8$	$35, 3$	$40, 40$

Table 7

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2009.

[2] Z. Bai, *Note on the quadratic convergence of kogbetliantz's algorithm for computing the singular value decomposition*, Linear Algebra and its Applications, 104 (1988), pp. 131–140.

[3] I. Balbaert and A. Salceanu, *Julia 1.0 programming complete reference guide: discover Julia, a high-performance language for technical computing*, Packt Publishing Ltd, 2019.

[4] M. Bečka, G. Okša, and M. Vajteršic, *New dynamic orderings for the parallel one–sided block-jacobi svd algorithm*, Parallel Processing Letters, 25 (2015), p. 1550003.

[5] B. Beckermann and A. Townsend, *On the singular values of matrices with displacement structure*, SIAM Journal on Matrix Analysis and Applications, 38 (2017), pp. 1227–1248.

[6] A. Ben-Israel and T. N. Greville, *Generalized inverses: theory and applications*, vol. 15, Springer Science & Business Media, 2003.

[7] Å. Björck and C. Bowie, *An iterative algorithm for computing the best estimate of an orthogonal matrix*, SIAM J. on Num. Analysis, 8 (1971), pp. 358–364.

[8] J.-P. Charlier and P. Van Dooren, *On kogbetliantz's svd algorithm in the presence of clusters*, Linear Algebra and its Applications, 95 (1987), pp. 135–160.

[9] F. Chatelin, *Simultaneous newton's iteration for the eigenproblem*, in Defect correction methods, Springer, 1984, pp. 67–74.

[10] M. T. Chu, *A differential equation approach to the singular value decomposition of bidiagonal matrices*, Linear algebra and its applications, 80 (1986), pp. 71–79.

[11] P. I. Davies and M. I. Smith, *Updating the singular value decomposition*, Journal of computational and applied mathematics, 170 (2004), pp. 145–167.

[12] J. Demmel and W. Kahan, *Accurate singular values of bidiagonal matrices*, SIAM Journal on Scientific and Statistical Computing, 11 (1990), pp. 873–912.

[13] I. S. Dhillon and B. N. Parlett, *Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices*, Linear Algebra and its Applications, 387 (2004), pp. 1–28.

[14] J. Dongarra, M. Gates, A. Haidar, J. Kurzak, P. Luszczek, S. Tomov, and I. Yamazaki, *The singular value decomposition: Anatomy of optimizing an algorithm for extreme scale*, SIAM Review, 60 (2018), pp. 808–865.

[15] J. J. Dongarra, D. C. Sorensen, and S. J. Hammarling, *Block reduction of matrices to condensed forms for eigenvalue computations*, Journal of Computational and Applied Mathematics, 27 (1989), pp. 215–227.

[16] Z. Drmač, *Algorithm 977: A qr–preconditioned qr svd method for computing the svd with high accuracy*, ACM Transactions on Mathematical Software (TOMS), 44 (2017), p. 11.

[17] Z. Drmač and K. Veselić, *New fast and accurate jacobi svd algorithm. i*, SIAM Journal on matrix analysis and applications, 29 (2008), pp. 1322–1342.

[18] Z. Drmač and K. Veselić, *New fast and accurate jacobi svd algorithm. ii*, SIAM Journal on matrix analysis and applications, 29 (2008), pp. 1343–1362.

[19] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM journal on Matrix Analysis and Applications, 20 (1998), pp. 303–353.

[20] L. Eldén and H. Park, *A procrustes problem on the stiefel manifold*, Numerische Mathematik, 82 (1999), pp. 599–619.

[21] K. Fan and A. J. Hoffman, *Some metric inequalities in the space of matrices*, Proc. of the AMS, 6 (1955), pp. 111–116.

[22] K. V. Fernando, *Linear convergence of the row cyclic jacobi and kogbetliantz methods*, Numerische Mathematik, 56 (1989), pp. 73–91.

[23] K. V. Fernando and B. N. Parlett, *Accurate singular values and differential qd algorithms*, Numerische Mathematik, 67 (1994), pp. 191–229.

[24] G. E. Forsythe and P. Henrici, *The cyclic jacobi method for computing the principal values of a complex matrix*, Transactions of the American Mathematical Society, 94 (1960), pp. 1–23.

[25] M. Gates, S. Tomov, and J. Dongarra, *Accelerating the svd two stage bidiagonal reduction and divide and conquer using gpus*, Parallel Computing, 74 (2018), pp. 3–18.

[26] G. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis, 2 (1965), pp. 205–224.

[27] G. Golub and C. F. Van Loan, *Matrix computations*, Baltimoere, The Johns Hopkins University Press, fourth ed., 2013.

[28] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM journal on numerical analysis, 17 (1980), pp. 883–893.

[29] M. Gu and S. C. Eisenstat, *A divide-and-conquer algorithm for the bidiagonal svd*, SIAM Journal on Matrix Analysis and Applications, 16 (1995), pp. 79–92.

[30] V. Hari and J. Matejaš, *Accuracy of two svd algorithms for 2× 2 triangular matrices*, Applied Mathematics and Computation, 210 (2009), pp. 232–257.

[31] N. J. Higham, *Matrix nearness problems and applications*, in Applications of Matrix Theory, M. J. C. Gover and S. Barnett, eds., Oxford University Press, 1989, pp. 1–27.

[32] L. Hogben, *Handbook of linear algebra*, Chapman and Hall/CRC, 2013.

[33] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge university press, 2012.

[34] I. C. Ipsen, *Computing an eigenvector with inverse iteration*, SIAM review, 39 (1997), pp. 254–291.

[35] E. Kogbetliantz, *Solution of linear equations by diagonalization of coefficients matrix*, Quar-

terly of Applied Mathematics, 13 (1955), pp. 123–132.

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

[37] S. Li, M. Gu, L. Cheng, X. Chi, and M. Sun, *An accelerated divide-and-conquer algorithm for the bidiagonal svd problem*, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 1038–1057.

[38] C. D. Martin and M. A. Porter, *The extraordinary svd*, The American Mathematical Monthly, 119 (2012), pp. 838–851.

[39] J. Matejaš and V. Hari, *Accuracy of the kogbetliantz method for scaled diagonally dominant triangular matrices*, Applied mathematics and computation, 217 (2010), pp. 3726–3746.

[40] J. Matejaš and V. Hari, *On high relative accuracy of the kogbetliantz method*, Linear Algebra and its Applications, 464 (2015), pp. 100–129.

[41] G. Okša, Y. Yamamoto, M. Becka, and M. Vajteršic, *Asymptotic quadratic convergence of the two-sided serial and parallel block-jacobi svd algorithm*, SIAM Journal on Matrix Analysis and Applications, 40 (2019), pp. 639–671.

[42] C. Paige and P. Van Dooren, *On the quadratic convergence of kogbetliantz's algorithm for computing the singular value decomposition*, Linear algebra and its applications, 77 (1986), pp. 301–313.

[43] G. W. Stewart, *On the early history of the singular value decomposition*, SIAM review, 35 (1993), pp. 551–566.

[44] H. Weyl, *Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)*, Mathematische Annalen, 71 (1912), pp. 441–479, http://eudml.org/doc/158545.

[45] J. H. Wilkinson, *Note on the quadratic convergence of the cyclic jacobi process*, Numerische Mathematik, 4 (1962), pp. 296–300.

[46] P. R. Willems, B. Lang, and C. Vömel, *Computing the bidiagonal svd using multiple relatively robust representations*, SIAM Journal on Matrix Analysis and Applications, 28 (2006), pp. 907–926.