Generalized frame operator distance problems

Pedro G. Massey [*], Noelia B. Rios [*] and Demetrio Stojanoff [*] [†]
CMaLP-FCE-UNLP and IAM-CONICET, Argentina

Abstract

Let $S \in \mathcal{M}_d(\mathbb{C})^+$ be a positive semidefinite $d \times d$ complex matrix and let $\mathbf{a} = (a_i)_{i \in \mathbb{I}_k} \in \mathbb{R}_{>0}^k$, indexed by $\mathbb{I}_k = \{1, \ldots, k\}$, be a k-tuple of positive numbers. Let $\mathbb{T}_d(\mathbf{a})$ denote the set of families $\mathcal{G} = \{g_i\}_{i \in \mathbb{I}_k} \in (\mathbb{C}^d)^k$ such that $\|g_i\|^2 = a_i$, for $i \in \mathbb{I}_k$; thus, $\mathbb{T}_d(\mathbf{a})$ is the product of spheres in $\mathbb{C}^d$ endowed with the product metric. For a strictly convex unitarily invariant norm N in $\mathcal{M}_d(\mathbb{C})$, we consider the generalized frame operator distance function $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$ defined on $\mathbb{T}_d(\mathbf{a})$, given by

$$\Theta_{(N\,,\,S\,,\,\mathbf{a})}(\mathcal{G}) = N(S - S_{\mathcal{G}}) \quad \text{where} \quad S_{\mathcal{G}} = \sum_{i \in \mathbb{I}_k} g_i\, g_i^* \in \mathcal{M}_d(\mathbb{C})^+ \, .$$

In this paper we determine the geometrical and spectral structure of local minimizers $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ of $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$. In particular, we show that local minimizers are global minimizers, and that these families do not depend on the particular choice of N.

AMS subject classification: 42C15, 15A60.
Keywords: matrix approximation, unitarily invariant norms, majorization, frame operator distance.

Contents

[*]Partially supported by CONICET (PIP 0150/15), FONCyT (PICT 1506/15) and FCE-UNLP (11X829), Argentina.
[†]e-mail addresses: massey@mate.unlp.edu.ar, nbrios@mate.unlp.edu.ar, demetrio@mate.unlp.edu.ar

1 Introduction

Matrix approximation problems are ubiquitous in applications of matrix analysis. Following [13] these problems can be briefly described as follows: given $S \in \mathcal{M}_d(\mathbb{C})$, a complex matrix of size d, a matrix norm N in $\mathcal{M}_d(\mathbb{C})$, and a set $\mathcal{X} \subset \mathcal{M}_d(\mathbb{C})$ then we search for the minimal distance

$$d_N(S, \mathcal{X}) = \min\{N(S - A) : \ A \in \mathcal{X}\} \ ,$$

and for the best approximations of S from $\mathcal{X}$ (or nearest members in $\mathcal{X}$)

$$\mathcal{A}_N^{\rm op}(S\,,\,\mathcal{X}) = \{A \in \mathcal{X} : \ N(S - A) = d_N(S\,,\,\mathcal{X})\} \ .$$

Solving these problems, that are also known as matrix nearness or Procrustes problems in the literature (see for example the recent text [12], and the classic books of Bhatia [3] and Kato [14]) amounts to provide a characterization and, if possible, an explicit computation (in some cases sharp estimations) of $d_N(S\,,\,\mathcal{X})$ and of the set of best approximations $\mathcal{A}_N^{\rm op}(S\,,\,\mathcal{X})$. A typical choice for N is the Frobenius norm (also called 2-norm) since it is an euclidean norm (i.e. it is the norm associated with an inner product in $\mathcal{M}_d(\mathbb{C})$). Still, some other norms are also of interest such as weighted norms, the p-norms for $1 \leq p$ (that contain the Frobenius norm), or the more general class of unitarily invariant norms. Some of the most important choices for $\mathcal{X}$ are the set of: selfadjoint matrices, positive semidefinite matrices, correlation matrices, orthogonal projections, oblique projections, matrices with rank bounded by a fix number (see [10, 11, 13, 15, 24]).

Once the nearness problem above has been solved for some S, some set $\mathcal{X}$ and norm N in $\mathcal{M}_d(\mathbb{C})$ then, a natural proximity problem arises: for a fixed $A_0 \in \mathcal{X}$, we search for (some sharp upper bound of) the distance

$$d_{\mathcal{X}}(A_0\,,\,\mathcal{A}_N^{\rm op}(S\,,\,\mathcal{X})) = \min\{d_{\mathcal{X}}(A_0\,,\,A) : \ A \in \mathcal{A}_N^{\rm op}(S\,,\,\mathcal{X})\} \ ,$$

where $d_{\mathcal{X}}$ denotes a metric in $\mathcal{X}$. In case $\mathcal{X}$ can be endowed with a smooth structure that is compatible with $d_{\mathcal{X}}$ and such that $\Psi(A) = N(A_0 - A)$ is also a smooth function on $\mathcal{X}$, then estimations of $d_{\mathcal{X}}(A_0\,,\,\mathcal{A}_N^{\rm op}(S\,,\,\mathcal{X}))$ can be obtained by applying gradient descent algorithms for Ψ or by studying the evolution of the solutions of flows in $\mathcal{X}$ associated with the gradient of Ψ.

Motivated by some optimization problems in finite frame theory, in [17] we considered the following matrix nearness problem. Fix an arbitrary positive semidefinite $S \in \mathcal{M}_d(\mathbb{C})^+$ and a finite sequence of positive numbers $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}_{>0}^k)^\downarrow$, indexed by $\mathbb{I}_k = \{1\,,\,\dots\,,\,k\}$; we considered the sets

$$\mathbb{T}_d(\mathbf{a}) = \Big\{\{g_i\}_{i\in\mathbb{I}_k} \in (\mathbb{C}^d)^k : \ \|g_i\|^2 = a_i\,,\, i \in \mathbb{I}_k\Big\} \quad \text{and} \quad \mathcal{X}_{\mathbf{a}} = \left\{\sum_{i\in\mathbb{I}_k} g_i\,g_i^* : \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})\right\} \ .$$

With this notation we solved the matrix nearness problem corresponding to $\mathcal{X}_{\mathbf{a}} \subset \mathcal{M}_d(\mathbb{C})^+$, for an arbitrary strictly convex unitarily invariant norm N in $\mathcal{M}_d(\mathbb{C})$. That is, we obtained an explicit description of $d_N(S\,,\,\mathcal{X}_{\mathbf{a}}) = d_N(S\,,\,\mathbf{a})$ and $\mathcal{A}_N^{\rm op}(S\,,\,\mathcal{X}_{\mathbf{a}}) = \mathcal{A}_N^{\rm op}(S\,,\,\mathbf{a})$. We point out that the set $\mathcal{X}_{\mathbf{a}}$ above can also be described as the set of frame operators $S_{\mathcal{G}}$ of finite sequences $\mathcal{G} = \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ (see Section 2 for details).

It is then natural to consider the proximity problem associated to the matrix nearness problem that we just described. Indeed, because of our initial motivation on this problem, we further pose the following (stronger) version: for $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ search for a (sharp) upper bound of

$$d(\mathcal{G}_0\,,\,\mathcal{B}_N^{\rm op}(S\,,\,\mathbf{a})) = \min\left\{d_{\mathbb{T}_d(\mathbf{a})}(\mathcal{G}_0\,,\,\mathcal{G}) : \ \mathcal{G} \in \mathcal{B}_N^{\rm op}(S\,,\,\mathbf{a})\right\} \ ,$$

where

$$\mathcal{B}_N^{\rm op}(S\,,\,\mathbf{a}) = \left\{\mathcal{G} \in \mathbb{T}_d(\mathbf{a}) : \ S_{\mathcal{G}} \in \mathcal{A}_N^{\rm op}(S\,,\,\mathbf{a})\right\} \quad \text{and} \quad d^2_{\mathbb{T}_d(\mathbf{a})}(\mathcal{G}_0\,,\,\mathcal{G}) = \sum_{i\in\mathbb{I}_k} \|g_i^0 - g_i\|^2 \ ,$$

for $\mathcal{G}_0 = \{g_i^0\}_{i\in\mathbb{I}_k}$, $\mathcal{G} = \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$. That is, we shift our attention from frame operators $S_{\mathcal{G}}$ to finite sequences $\mathcal{G} \in \mathbb{T}_d(\mathbf{a})$. Notice that in the particular case $S = \frac{k}{d}\, I$, $a_i = 1$ for $i \in \mathbb{I}_k$ and N is the Frobenius norm, this problem is related with Paulsen's proximity problem [4, 5, 6], which is a central open problem in finite frame theory.

In case the norm N is sufficiently smooth, we could apply gradient descent algorithms to the function $\Theta = \Theta_{(N\,,\,S\,,\,\mathbf{a})}$ defined on $\mathbb{T}_d(\mathbf{a})$ - which is a smooth manifold in $(\mathbb{C}^d)^k$ - given by $\Theta(\mathcal{G}) = N(S-S_{\mathcal{G}})$, starting at $\mathcal{G}_0$. Such an approach was considered by N. Strawn [22, 23] for the Frobenius norm N. Also, we could study the evolution of solutions of gradient flows as considered in [16].

In the general case, the analysis of the behavior of gradient descent algorithms leads to the study the local behavior of the map

$$\mathbb{T}_d(\mathbf{a}) \ni \mathcal{G} = \{g_i\}_{i\in\mathbb{I}_k} \mapsto \Theta(\mathcal{G}) = N(S - S_{\mathcal{G}}) \quad \text{around} \quad \mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})\,.$$

One important issue is determining whether local minimizers of Θ (that are natural attractors of gradient descent algorithms) are actually global minimizers. In [17] we settled this question in the affirmative for the Frobenius norm (thus solving a conjecture in [22]), by relating frame operator distance problems in the Frobenius norm with frame completion problems for the Benedetto-Fickus frame potential introduced in [2]. Unfortunately, the techniques used in [17] do not apply for arbitrary N (not even for p-norms with $p > 1$, $p \neq 2$). In the present work we tackle this problem and show that, in case N is an arbitrary strictly convex u.i.n., local minimizers of Θ are characterized by a spectral condition that does not depend on N, but only on S and $\mathbf{a}$. In particular, we conclude that local minimizers are global minimizers and do not depend on the particular choice of N. Our techniques rely on majorization theory and Lidskii's local theorems for unitarily invariant norms obtained in [18]; indeed, in that paper we showed that in some particular cases, local minimizers of the generalized frame operator distance (GFOD) functions (i.e. $\Theta(\mathcal{G}) = N(S - S_{\mathcal{G}})$) are global minimizers. Based on the features of these particular cases, we introduce the notion of co-feasible GFOD problems. Although in general GFOD problems are not co-feasible, this notion plays a crucial role in the study of the spectral structure of local minimizers. Using that the map $\mathbb{T}_d(\mathbf{a}) \ni \mathcal{G} \mapsto S_{\mathcal{G}} \in \mathcal{X}_{\mathbf{a}}$ is continuous, as a byproduct we obtain that local minimizers $S_{\mathcal{G}_0} \in \mathcal{X}_{\mathbf{a}}$ of the function

$$\Psi : \mathcal{X}_{\mathbf{a}} \to \mathbb{R}_{\geq 0} \quad \text{given by} \quad \Psi(S_{\mathcal{G}}) = N(S - S_{\mathcal{G}})$$

are global minimizers and do not depend on the choice of strictly convex u.i.n. N. This last fact is weaker than the result for the functions Θ, since the continuous map $\mathbb{T}_d(\mathbf{a}) \ni \mathcal{G} \mapsto S_{\mathcal{G}} \in \mathcal{X}_{\mathbf{a}}$ does not have local cross sections around an arbitrary $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$.

The paper is organized as follows. In Section 2 we include some preliminary material on matrix analysis and finite frame theory that is used throughout the paper. In Section 3 we state our main problem namely, the study of the geometrical and spectral structure of local minimizers of the GFOD functions (i.e. Θ as above), associated to a strictly convex unitarily invariant norm. We begin by obtaining a series of results related with what we call the inner structure of such local minimizers. In section 4 we state our main results namely, that local minimizers of GFOD functions are global minimizers, and give an algorithmic construction of the eigenvalues of such families. Finally, in Section 5 we give detailed proofs of some results stated in Section 3.

2 Preliminaries

In this section we introduce the notation, terminology and results from matrix analysis (see the text [3]) and finite frame theory (see the texts [7, 8, 9]) that we will use throughout the paper.

2.1 Matrix Analysis

Notation and terminology. We let $\mathcal{M}_{k,d}(\mathbb{C})$ be the space of complex $k\times d$ matrices and write $\mathcal{M}_{d,d}(\mathbb{C})=\mathcal{M}_d(\mathbb{C})$ for the algebra of $d\times d$ complex matrices. We denote by $\mathcal{H}(d)\subset\mathcal{M}_d(\mathbb{C})$ the real subspace of selfadjoint matrices and by $\mathcal{M}_d(\mathbb{C})^+\subset\mathcal{H}(d)$ the cone of positive semi-definite matrices. We let $\mathcal{U}(d)\subset\mathcal{M}_d(\mathbb{C})$ denote the group of unitary matrices.

For $d\in\mathbb{N}$, let $\mathbb{I}_d=\{1,\ldots,d\}$. Given a vector $x\in\mathbb{C}^d$ we denote by D_x the diagonal matrix in $\mathcal{M}_d(\mathbb{C})$ whose main diagonal is x. Given $x=(x_i)_{i\in\mathbb{I}_d}\in\mathbb{R}^d$ we denote by $x^\downarrow=(x_i^\downarrow)_{i\in\mathbb{I}_d}$ the vector obtained by rearranging the entries of x in non-increasing order. We also use the notation $(\mathbb{R}^d)^\downarrow=\{x\in\mathbb{R}^d\ :\ x=x^\downarrow\}$ and $(\mathbb{R}^d_{\geq 0})^\downarrow=\{x\in\mathbb{R}^d_{\geq 0}\ :\ x=x^\downarrow\}$. For $r\in\mathbb{N}$, we let $\mathbb{1}_r=(1,\ldots,1)\in\mathbb{R}^r$.

Given a matrix $A\in\mathcal{H}(d)$ we denote by $\lambda(A)=\lambda^\downarrow(A)=(\lambda_i(A))_{i\in\mathbb{I}_d}\in(\mathbb{R}^d)^\downarrow$ the eigenvalues of A counting multiplicities and arranged in non-increasing order. For $B\in\mathcal{M}_d(\mathbb{C})$ we let $s(B)=\lambda(|B|)$ denote the singular values of B, i.e. the eigenvalues of $|B|=(B^*B)^{1/2}\in\mathcal{M}_d(\mathbb{C})^+$; we also let $\sigma(B)\subset\mathbb{C}$ denote the spectrum of B. If $x,\,y\in\mathbb{C}^d$ we denote by $x\otimes y=x\,y^*\in\mathcal{M}_d(\mathbb{C})$ the rank-one matrix given by $(x\otimes y)\,z=\langle z\,,\,y\rangle\,x$, for $z\in\mathbb{C}^d$.

Next we recall the notion of majorization between vectors, that will play a central role throughout our work.

Definition 2.1. Let $x\in\mathbb{R}^k$ and $y\in\mathbb{R}^d$. We say that x is *submajorized* by y, and write $x\prec_w y$, if

$$\sum_{i=1}^{j}x_i^\downarrow\leq\sum_{i=1}^{j}y_i^\downarrow\quad\text{for every}\quad 1\leq j\leq\min\{k\,,\,d\}\,.$$

If $x\prec_w y$ and $\operatorname{tr}x=\sum_{i=1}^k x_i=\sum_{i=1}^d y_i=\operatorname{tr}y$, then x is *majorized* by y, and write $x\prec y$. If $k=d$, we say that x is *strictly* majorized by y if $x\prec y$ and $x^\downarrow\neq y^\downarrow$. △

Remark 2.2. Given $x,\,y\in\mathbb{R}^d$ we write $x\leqslant y$ if $x_i\leq y_i$ for every $i\in\mathbb{I}_d$. It is a standard exercise to show that:

1. $x\leqslant y\implies x^\downarrow\leqslant y^\downarrow\implies x\prec_w y$.
2. $x\prec y\implies|x|\prec_w|y|$, where $|x|=(|x_i|)_{i\in\mathbb{I}_d}\in\mathbb{R}^d_{\geq 0}$.
3. $x\prec y$ and $|x|^\downarrow=|y|^\downarrow\implies x^\downarrow=y^\downarrow$.
4. $x\prec y$ and $z\prec w\in\mathbb{R}^e\implies(x,z)\prec(y,w)\in\mathbb{R}^{d+e}$. △

Although majorization is not a total order in $\mathbb{R}^d$, there are several fundamental inequalities in matrix theory that can be described in terms of this relation. As an example of this phenomenon we can consider Lidskii's (additive) inequality (see [3]). In the following result we also include the characterization of the case of equality obtained in [21].

Theorem 2.3 (Lidskii's inequality). Let $A,\,B\in\mathcal{H}(d)$. Then

1. $\lambda(A)-\lambda(B)\prec\lambda(A-B)$.
2. $\lambda(A-B)=\big(\lambda(A)-\lambda(B)\big)^\downarrow$ if and only if there exists $\{v_i\}_{i\in\mathbb{I}_d}$ an ONB of $\mathbb{C}^d$ such that

$$A=\sum_{i\in\mathbb{I}_d}\lambda_i(A)\ v_i\otimes v_i\quad\text{and}\quad B=\sum_{i\in\mathbb{I}_d}\lambda_i(B)\ v_i\otimes v_i\ .\tag{1}$$

Notice that in this case, A and B commute. □

Recall that a norm N in $\mathcal{M}_d(\mathbb{C})$ is **unitarily invariant** (briefly u.i.n.) if

$$N(UAV) = N(A) \quad \text{for every} \quad A \in \mathcal{M}_d(\mathbb{C}) \quad \text{and} \quad U, V \in \mathcal{U}(d) ,$$

and N is **strictly convex** if its restriction to diagonal matrices is a strictly convex norm in $\mathbb{C}^d$. Examples of u.i.n. are the spectral norm $\| \cdot \|$ and the p-norms $\| \cdot \|_p$, for $p \geq 1$ (strictly convex if $p > 1$). It is well known that (sub)majorization relations between singular values of matrices are intimately related with inequalities with respect to u.i.n's. The following result summarizes these relations (see for example [3]):

Theorem 2.4. Let $A, B \in \mathcal{M}_d(\mathbb{C})$ be such that $s(A) \prec_w s(B)$. Then:

1. For every u.i.n. N in $\mathcal{M}_d(\mathbb{C})$ we have that $N(A) \leq N(B)$.
2. If N is a strictly convex u.i.n. in $\mathcal{M}_d(\mathbb{C})$ and $N(A) = N(B)$, then $s(A) = s(B)$. □

2.2 Finite frames

We consider some notions and results from the theory of finite frames. In what follows we adopt:

Notation and terminology: let $\mathcal{F} = \{f_i\}_{i \in \mathbb{I}_k}$ be a finite sequence in $\mathbb{C}^d$. Then,

1. $T_\mathcal{F} \in \mathcal{M}_{d,k}(\mathbb{C})$ denotes the synthesis operator of $\mathcal{F}$ given by $T_\mathcal{F} \cdot (\alpha_i)_{i \in \mathbb{I}_k} = \sum_{i \in \mathbb{I}_k} \alpha_i f_i$.
2. $T^*_\mathcal{F} \in \mathcal{M}_{k,d}(\mathbb{C})$ denotes the analysis operator of $\mathcal{F}$ and it is given by $T^*_\mathcal{F} \cdot f = (\langle f, f_i \rangle)_{i \in \mathbb{I}_k}$.
3. $S_\mathcal{F} \in \mathcal{M}_d(\mathbb{C})^+$ denotes the frame operator of $\mathcal{F}$ and it is given by $S_\mathcal{F} = T_\mathcal{F} T^*_\mathcal{F}$. Hence,
$$S_\mathcal{F} = \sum_{i \in \mathbb{I}_k} f_i \otimes f_i \quad \text{and} \quad R(S_\mathcal{F}) = \operatorname{span}\{f_i : i \in \mathbb{I}_k\} . \tag{2}$$
4. We say that $\mathcal{F}$ is a frame for $\mathbb{C}^d$ if it spans $\mathbb{C}^d$; equivalently, $\mathcal{F}$ is a frame for $\mathbb{C}^d$ if $S_\mathcal{F}$ is a positive invertible operator acting on $\mathbb{C}^d$. △

Hence, in case $\mathcal{F} = \{f_i\}_{i \in \mathbb{I}_k}$ is a frame for $\mathbb{C}^d$ we get the so-called canonical reconstruction formulas: for $x \in \mathbb{C}^d$,

$$x = \sum_{i \in \mathbb{I}_k} \langle x \,,\, S_\mathcal{F}^{-1} f_i \rangle \, f_i = \sum_{i \in \mathbb{I}_k} \langle x \,,\, f_i \rangle \, S_\mathcal{F}^{-1} f_i \ .$$

In several applications of finite frame theory, it is important to construct families $\mathcal{F} = \{f_i\}_{i \in \mathbb{I}_k} \in (\mathbb{C}^d)^k$ in such a way that the frame operator $S_\mathcal{F}$ and the squared norms $(\|f_i\|^2)_{i \in \mathbb{I}_k}$ are prescribed in advance. This problem is known as the frame design problem, and its solution can be obtained in terms of the Schur-Horn theorem for majorization.

Theorem 2.5 (See [1])**.** *Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and let $\mathbf{a} = (a_i)_{i \in \mathbb{I}_k} \in (\mathbb{R}^k_{>0})$. Then, the following statements are equivalent:*

1. *There exists $\mathcal{F} = \{f_i\}_{i \in \mathbb{I}_k} \in (\mathbb{C}^d)^k$ such that $S_\mathcal{F} = S$ and $\|f_i\|^2 = a_i$, for $i \in \mathbb{I}_k$;*
2. *$\mathbf{a} \prec \lambda(S)$.* □

3 Generalized frame operator distance functions

In this section we state our main problem namely, the study of the geometrical and spectral structure of local minimizers of generalized frame operator distance (GFOD) functions. After recalling some preliminary results from [18], we obtain a description of what we call the inner structure of local minimizers of GFOD's functions. Since the proofs of some results in this section are quite technical, they are developed in Section 5.

3.1 Statement of the problem and related results

Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$. In this case we consider the torus

$$\mathbb{T}_d(\mathbf{a}) \stackrel{\text{def}}{=} \left\{ \mathcal{G} = \{g_i\}_{i\in\mathbb{I}_k} \in (\mathbb{C}^d)^k \ : \|g_i\|^2 = a_i \, , \text{ for every } i \in \mathbb{I}_k \right\} \ .$$

By definition, $\mathbb{T}_d(\mathbf{a})$ is the (Cartesian) product of spheres in $\mathbb{C}^d$; we endow $\mathbb{T}_d(\mathbf{a})$ with the product metric of the Euclidean metrics in each of these spheres, namely

$$d^2(\mathcal{G} \, , \, \mathcal{G}') = \sum_{i\in\mathbb{I}_k} \|g_i - g_i'\|^2 \quad \text{for} \quad \mathcal{G} = \{g_i\}_{i\in\mathbb{I}_k} , \, \mathcal{G}' = \{g_i'\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a}) \ .$$

Thus, $\mathbb{T}_d(\mathbf{a})$ is a compact smooth manifold. Given a strictly convex u.i.n $N : \mathcal{M}_d(\mathbb{C}) \to \mathbb{R}_{\geq 0}$, we can consider the generalized frame operator distance (G-FOD) in $\mathbb{T}_d(\mathbf{a})$ (see [17]) given by

$$\Theta_{(N \, , \, S \, , \, \mathbf{a})} = \Theta : \mathbb{T}_d(\mathbf{a}) \to \mathbb{R}_{\geq 0} \quad \text{given by} \quad \Theta(\mathcal{G}) = N(S - S_\mathcal{G}) \ ,$$

where $S_\mathcal{G} = \sum_{i\in\mathbb{I}_k} g_i \otimes g_i$ denotes the frame operator of a family $\mathcal{G} \in \mathbb{T}_d(\mathbf{a})$. This notion is based on the frame operator distance (FOD) $\Theta_{(\|\cdot\|_2 \, , \, S \, , \, \mathbf{a})}$ introduced by Strawn in [22], where $\|A\|_2^2 = \operatorname{tr}(A^*A)$ denotes the Frobenius norm, $A \in \mathcal{M}_d(\mathbb{C})$. Based on his work and on numerical evidence, Strawn conjectured that local minimizers of $\Theta_{(\|\cdot\|_2 \, , \, S \, , \, \mathbf{a})}$ were also global minimizers. In [17] we settled Strawn's conjecture in the affirmative, by relating FOD problems in the norm $\| \cdot \|_2$ with optimal frame completion problems for the Benedetto-Fickus frame potential. It is then natural to ask whether local minimizers of the G-FOD $\Theta_{(N \, , \, S \, , \, \mathbf{a})}$ are also global minimizers, where N denotes an arbitrary strictly convex u.i.n. on $\mathcal{M}_d(\mathbb{C})$ (e.g. p-norms, with $p \in (1, \infty)$). Unfortunately, the techniques used in [17] do not apply in this general case, leaving untouched the following

Problems 3.1. Let $S \in \mathcal{M}_d(\mathbb{C})^+$, $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ and fix a strictly convex u.i.n. N on $\mathcal{M}_d(\mathbb{C})$. Then

- P1. Compute the spectral and geometrical structure of local minimizers of $\Theta_{(N \, , \, S \, , \, \mathbf{a})}$ in $\mathbb{T}_d(\mathbf{a})$.
- P2. Determine whether local minimizers are global minimizers of $\Theta_{(N \, , \, S \, , \, \mathbf{a})}$ in $\mathbb{T}_d(\mathbf{a})$.
- P3. Determine whether these minimizers depend on the chosen u.i.n. □

In what follows we completely solve the three problems above in an algorithmic way, thus settling in the affirmative the questions in P2. and P3. (see Theorem 4.12 in Section 4.2).

Next, we recall some results from [18] that we use throughout our work.

Theorem 3.2 (See [18]). Fix $S \in \mathcal{M}_d(\mathbb{C})^+$, $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$, and a strictly convex u.i.n. N on $\mathcal{M}_d(\mathbb{C})$. Consider the map $\Theta_{(N \, , \, S \, , \, \mathbf{a})} = \Theta : \mathbb{T}_d(\mathbf{a}) \to \mathbb{R}_{\geq 0}$ given by $\Theta(\mathcal{G}) = N(S - S_\mathcal{G})$.

Fix a local minimizer $\mathcal{G}_0 = \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ of $\Theta_{(N \, , \, S \, , \, \mathbf{a})}$, with frame operator $S_0 = S_{\mathcal{G}_0}$. Denote by $W = R(S_0) = \operatorname{span}\{g_i : \ i \in \mathbb{I}_k\} \subseteq \mathbb{C}^d$. Then,

1. There exists $\mathcal{B} = \{v_i\}_{i\in\mathbb{I}_d}$ an ONB of $\mathbb{C}^d$ such that

$$S = \sum_{i\in\mathbb{I}_d} \lambda_i(S) \ v_i \otimes v_i \quad \text{and} \quad S_0 = \sum_{i\in\mathbb{I}_d} \lambda_i(S_0) \ v_i \otimes v_i \, .$$

In particular, we have that $\lambda(S - S_0) = \big(\lambda(S) - \lambda(S_0)\,\big)^\downarrow$.

2. The subspace W reduces $S - S_0 \in \mathcal{H}(d)$; hence, $D \stackrel{\text{def}}{=} (S - S_0)|_W \in L(W)$ satisfies $D^* = D$.

3. All vectors g_i ($i \in \mathbb{I}_k$) are eigenvectors of D and $S - S_0$.

4. Let $\sigma(D) = \{c_1, \dots, c_p\}$ be such that $c_1 < c_2 < \dots < c_p$. Denote by

$$J_j = \{\ell \in \mathbb{I}_k : \; D\, g_\ell = c_j\, g_\ell\} \quad \text{and} \quad W_j = \operatorname{span}\{g_\ell : \; \ell \in J_j\} \quad \text{for} \quad j \in \mathbb{I}_p \, .$$

Then the subspaces W_j reduces both S and S_0, for $j \in \mathbb{I}_p$. Moreover,

$$\mathbb{I}_k = \overset{\text{D}}{\bigcup_{j\in\mathbb{I}_p}} J_j \quad \text{(disjoint union)} \quad \text{and} \quad W = \bigoplus_{j\in\mathbb{I}_p} W_j \, . \tag{3}$$

5. If $j \in \mathbb{I}_p$ and $c_j \neq \max \sigma(S - S_0)$ (for example, when $1 \le j < p$), then the family $\{g_\ell\}_{\ell \in J_j}$ is linearly independent. □

Remark 3.3. With the notation of Theorem 3.2, if we assume that

$$k \ge d \quad \Longrightarrow \quad c_p = \max \sigma(S - S_0) \, . \tag{4}$$

Indeed, if $W = \mathbb{C}^d$ then $\sigma(S - S_0) = \{c_1, \dots, c_p\}$. Otherwise $\dim W < d \le k$ so, by items 4 and 5 of Theorem 3.2, the family $\{g_i\}_{i\in J_p}$ can not be linearly independent (because the families $\{g_i\}_{i\in J_j}$ are linearly independent for $1 \le j < p$, and all families are mutually orthogonal). By item 5 again, we deduce that $c_p = \max \sigma(S - S_0)$. △

3.2 Inner structure of local minimizers of GFOD's

In this section, based on Theorem 3.2 above, we obtain a detailed description of what we call the inner structure of local minimizers. In order to do this, we introduce the following

Notation 3.4. Fix $S \in \mathcal{M}_d(\mathbb{C})^+$, $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ and a strictly convex u.i.n. N on $\mathcal{M}_d(\mathbb{C})$. Also consider the notions introduced in Theorem 3.2. As before, consider

1. $\Theta_{(N, S, \mathbf{a})} = \Theta : \mathbb{T}_d(\mathbf{a}) \to \mathbb{R}_{\ge 0}$ given by $\Theta(\mathcal{G}) = N(S - S_{\mathcal{G}})$.

2. A local minimizer $\mathcal{G}_0 = \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ of $\Theta_{(N, S, \mathbf{a})}$, with frame operator $S_0 = S_{\mathcal{G}_0}$.

3. We denote by $\lambda = (\lambda_i)_{i\in\mathbb{I}_d} = \lambda(S) \in (\mathbb{R}^d_{\ge 0})^\downarrow$ and $\mu = (\mu_i)_{i\in\mathbb{I}_d} = \lambda(S_0) \in (\mathbb{R}^d_{\ge 0})^\downarrow$.

4. We fix $\mathcal{B} = \{v_i\}_{i\in\mathbb{I}_d}$ an ONB of $\mathbb{C}^d$ as in Theorem 3.2. Hence,

$$S = \sum_{i\in\mathbb{I}_d} \lambda_i \; v_i \otimes v_i \quad \text{and} \quad S_0 = \sum_{i\in\mathbb{I}_d} \mu_i \; v_i \otimes v_i \, . \tag{5}$$

5. We consider $W = R(S_0)$, $D = (S - S_0)|_W$ and $\sigma(D) = \{c_1, \dots, c_p\}$ where $c_1 < c_2 < \dots < c_p$.

6. Let $s_D = \max\{i \in \mathbb{I}_d : \mu_i \neq 0\} = \operatorname{rk} S_0$.

7. We denote by $\delta = \lambda - \mu \in \mathbb{R}^d$ so that, by Eq. (5),

$$S - S_0 = \sum_{i\in\mathbb{I}_d} \delta_i \; v_i \otimes v_i \quad \text{and} \quad D = \sum_{i=1}^{s_D} \delta_i \; v_i \otimes v_i \, .$$

Notice that δ is constructed by pairing the entries of ordered vectors (since $\lambda = \lambda(S)$ and $\mu = \lambda(S_0)$. Nevertheless, we have that $\lambda(S - S_0) = \delta^\downarrow$. In what follows we obtain some properties of (the unordered vector) δ.

8. For each $j \in \mathbb{I}_p$, we consider the following sets of indexes:

$$K_j = \{i \in \mathbb{I}_{s_D} : \delta_i = \lambda_i - \mu_i = c_j\} \quad \text{and} \quad J_j = \{i \in \mathbb{I}_k : D\, g_i = c_j\, g_i\} \ .$$

Theorem 3.2 assures that $\mathbb{I}_{s_D} = \overset{\text{D}}{\bigcup}_{j\in\mathbb{I}_p} K_j$ and $\mathbb{I}_k = \overset{\text{D}}{\bigcup}_{j\in\mathbb{I}_p} J_j$ (disjoint unions).

9. By Eq. (2), $R(S_0) = \operatorname{span}\{g_i : i \in \mathbb{I}_k\} = W = \bigoplus_{i\in\mathbb{I}_p} \ker(D - c_i\, I_W)$. Then, for every $j \in \mathbb{I}_p$,

$$W_j = \operatorname{span}\{g_i : i \in J_j\} = \ker(D - c_j\, I_W) = \operatorname{span}\{v_i : i \in K_j\} \ ,$$

because $g_i \in \ker(D - c_j\, I_W)$ for every $i \in J_j$. Note that, by Theorem 3.2, each W_j reduces both S and S_0. △

The next proposition describes the structure of the sets J_j and K_j for $j \in \mathbb{I}_p$, as defined in Notation 3.4. In turn, these sets play a central role in the proof of Theorem 3.8 below.

Proposition 3.5. *Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ be as in Notation 3.4. Then there exist indexes $0 = s_0 < s_1 < \ldots < s_{p-1} < s_p = \operatorname{rk} S_0 \le d$ such that*

$$\begin{aligned} K_j &= J_j = \{s_{j-1}+1, \ldots, s_j\} , \qquad \textit{for} \quad j \in \mathbb{I}_{p-1} \quad (\textit{if } p > 1), \\ K_p &= \{s_{p-1}+1, \ldots, s_p\} \ , \quad J_p = \{s_{p-1}+1, \ldots, k\} \ . \end{aligned} \tag{6}$$

Proof. See Section 5. □

Remark 3.6. Consider Notation 3.4 for $S \in \mathcal{M}_d(\mathbb{C})^+$ and a local minimizer $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ of the map $\Theta_{(N, S, \mathbf{a})}$. Let $s_0 = 0 < s_1 < \ldots < s_p \le d$, where $s_p = \operatorname{rk}(S_0)$, be as in Proposition 3.5. In terms of these indexes we also get that $\lambda(S - S_0) = \delta(S, \mathbf{a}, \mathcal{G}_0)^{\downarrow}$ for $\delta(S, \mathbf{a}, \mathcal{G}_0) = \lambda(S) - \lambda(S_0)$, and

$$\delta(S, \mathbf{a}, \mathcal{G}_0) = \big(c_1\, \mathbb{1}_{s_1}, c_2\, \mathbb{1}_{s_2 - s_1}, \ldots, c_p\, \mathbb{1}_{s_p - s_{p-1}}, \lambda_{s_p+1}, \ldots, \lambda_d\big) \quad \text{if} \quad s_p < d \tag{7}$$

or

$$\delta(S, \mathbf{a}, \mathcal{G}_0) = \big(c_1\, \mathbb{1}_{s_1}, c_2\, \mathbb{1}_{s_2 - s_1}, \ldots, c_p\, \mathbb{1}_{s_p - s_{p-1}}\big) \quad \text{if} \quad s_p = d\,. \tag{8}$$

In the next result, we obtain a characterization of the indexes $s_1 < \ldots < s_{p-2}$ and constants $c_1 < \ldots c_{p-1}$ in terms of the index s_{p-1} (when $p > 1$). In the next section we complement these results and show the key role played by the index s_{p-1} and give a characterization of c_p. We begin by fixing some notation, which is independent of the norm N and the local minimizer $\mathcal{G}_0$. △

Notation 3.7. Let $S \in \mathcal{M}_d(\mathbb{C})^+$, $\mathbf{a} \in (\mathbb{R}_{>0}^k)^{\downarrow}$, $\lambda(S) = (\lambda_i)_{i\in\mathbb{I}_d} \in (\mathbb{R}^d)^{\downarrow}$ and $m = \min\{k, d\}$.

1. We let $h_i \overset{\text{def}}{=} \lambda_i - a_i$, for every $i \in \mathbb{I}_m$.
2. Given $1 \le j \le r \le m$, let

$$P_{j,r} = \frac{1}{r-j+1} \sum_{i=j}^{r} h_i = \frac{1}{r-j+1} \sum_{i=j}^{r} \lambda_i - a_i \ . \tag{9}$$

We abbreviate $P_{1,r} = P_r$ for the initial averages. △

Theorem 3.8. *Consider Notation 3.4 for $S \in \mathcal{M}_d(\mathbb{C})^+$ and a local minimizer $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ of the map $\Theta_{(N, S, \mathbf{a})}$. Assume further that $p > 1$. Let $s_0 = 0 < s_1 < \ldots < s_p \le d$ be such that Eq. (6) holds. Then, we have the following relations:*

1. *The index $s_1 = \max\big\{1 \le r \le s_{p-1} : P_r = \min_{1\le i\le s_{p-1}} P_i\big\}$, and $c_1 = P_{s_1}$.*
2. *Recursively, if $s_j < s_{p-1}$, then*

$$s_{j+1} = \max\big\{s_j < r \le s_{p-1} : P_{s_j+1, r} = \min_{s_j < i \le s_{p-1}} P_{s_j+1, i}\big\} \quad \textit{and} \quad c_{j+1} = P_{s_j+1, s_{j+1}} \ .$$

Proof. See Section 5. □

3.3 The co-feasible case for $k \geq d$.

Throughout this section we assume that $k \geq d$. In [18] we showed that in some cases, local minimizers of G-FOD functions are also global minimizers. We recall this fact in the following

Theorem 3.9 (See [18])**.** Consider Notation 3.4 with $k \geq d$ for $S \in \mathcal{M}_d(\mathbb{C})^+$ and a local minimizer $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ of the map $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$. Assume further that $p = 1$ i.e., that there exists $c = c_1$ that satisfies $(S - S_0)g_i = c\,g_i$, for every $i \in \mathbb{I}_k$. Then there exists an ONB $\{v_i\}_{i\in\mathbb{I}_d}$ of $\mathbb{C}^d$ such that

$$S = \sum_{i\in\mathbb{I}_d} \lambda_i \; v_i \otimes v_i \quad \text{ and } \quad S_0 = \sum_{i\in\mathbb{I}_d} (\lambda_i - c)^+ \; v_i \otimes v_i \,,$$

where $(\lambda_i)_{i\in\mathbb{I}_d} = \lambda(S) \in (\mathbb{R}^d_{\geq 0})^\downarrow$. Moreover, $\mathcal{G}_0$ is a global minimizer of Θ in $\mathbb{T}_d(\mathbf{a})$. □

Corollary 3.10. *With the hypotheses and notation in Theorem 3.9 we have that:*

1. *The constant* $c = \max \sigma(S - S_0)$ *is the largest eigenvalue of* $S - S_0$.
2. *The eigenvalue* $\lambda_i(S_0) = (\lambda_i - c)^+$, *for every* $i \in \mathbb{I}_d$.
3. *The list of norms* $\mathbf{a} \prec \big(\,(\lambda_i - c)^+\big)_{i\in\mathbb{I}_d}$. *In particular*

$$\operatorname{tr}(\mathbf{a}) = \sum_{i\in\mathbb{I}_k} a_i = \sum_{i\in\mathbb{I}_d} (\lambda_i - c)^+ \,.$$

Proof. 1. We are assuming that $k \geq d$. Then Remark 3.3 assures that $c = c_p = \max \sigma(S - S_0)$.

2. This is a direct consequence of Theorem 3.9 above and the fact that $(\lambda_i)_{i\in\mathbb{I}_d} = \lambda(S) \in (\mathbb{R}^d)^\downarrow$, so that also $\big(\,(\lambda_i - c)^+\big)_{i\in\mathbb{I}_d} \in (\mathbb{R}^d)^\downarrow$.

3. Since $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ (it is a family of vectors with norms given by $\mathbf{a}$), then Theorem 2.5 assures that

$$\mathbf{a} \prec \lambda(S_{\mathcal{G}_0}) = \big(\,(\lambda_i - c)^+\big)_{i\in\mathbb{I}_d} \,.$$

The rest of the statement is a direct consequence of this majorization relation. □

The previous results motivate the following notion, which only depends on some $\lambda \in (\mathbb{R}^d_{\geq 0})^\downarrow$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$, with $k \geq d$ (and does not require any norm N nor a local minimizer $\mathcal{G}_0$).

Definition 3.11. Let $\lambda \in (\mathbb{R}^d_{\geq 0})^\downarrow$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$, with $k \geq d$. We say that the pair $(\lambda\,,\,\mathbf{a})$ is **co-feasible** if there exists a constant

$$c < \lambda_1 \quad \text{ such that } \quad \mathbf{a} \prec \big(\,(\lambda_i - c)^+\big)_{i\in\mathbb{I}_d} \,. \tag{10}$$

In this case, the co-feasibility constant c is uniquely determined by $\operatorname{tr}(\mathbf{a}) = \sum_{i\in\mathbb{I}_d} (\lambda_i - c)^+$. △

Proposition 3.12. *Let* $S \in \mathcal{M}_d(\mathbb{C})^+$ *and* $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ *with* $k \geq d$. *Then the pair* $(\lambda(S)\,,\,\mathbf{a})$ *is co-feasible if and only if the following conditions hold:*

1. *There exist* $\mathcal{G} = \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ *and* $c \in \mathbb{R}$ *such that* $(S - S_{\mathcal{G}})\,g_i = c\,g_i$, *for every* $i \in \mathbb{I}_k$.
2. *This constant* $c = \max \sigma(S - S_{\mathcal{G}})$.

Proof. Assume that there exist $c \in \mathbb{R}$ and $\mathcal{G} = \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ which satisfy items 1 and 2. By Eq. (2), $W = R(S_\mathcal{G}) = \text{span}\{g_i : i \in \mathbb{I}_k\}$. Since $(S - S_\mathcal{G})\big|_W = c\, I_W$, then $S(W) \subseteq W$. Let $r = \dim W$. Then, considering separately the eigenvalues of $S\big|_W$ and $S\big|_{W^\perp} = (S - S_\mathcal{G})\big|_{W^\perp}$, the fact that $c = \max \sigma(S - S_\mathcal{G})$ implies that

$$c < \lambda_i(S) \quad \text{for} \quad i \in \mathbb{I}_r \quad \text{and} \quad c \geq \lambda_i(S) \quad \text{for} \quad r < i \leq d\ .$$

Therefore $\lambda(S_\mathcal{G}) = \lambda(S-(S-S_\mathcal{G})\,) = \big(\,(\lambda_i(S)-c)^+\big)_{i\in\mathbb{I}_d}$. Hence, arguing as in the proof of Corollary 3.10, we conclude that this c satisfies Eq. (10). Note that $c < \lambda_1(S)$ because $\operatorname{tr}\mathbf{a} \neq 0$.

Conversely, if there exists c which satisfies Eq. (10), let $\mathcal{B} = \{v_i\}_{i\in\mathbb{I}_d}$ be an ONB for $\mathbb{C}^d$ such that

$$S = \sum_{i\in\mathbb{I}_d} \lambda_i(S)\, v_i \otimes v_i \quad , \quad \text{and set} \quad S_0 \overset{\text{def}}{=} \sum_{i\in\mathbb{I}_d} (\lambda_i(S) - c)^+\, v_i \otimes v_i \in \mathcal{M}_d(\mathbb{C})^+\ .$$

By Theorem 2.5, there exists $\mathcal{G} = \{g_j\}_{j\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ such that $S_0 = S_\mathcal{G}$. Note that

$$\lambda_i(S) - (\lambda_i(S) - c)^+ = \min\{\lambda_i(S)\,,\, c\} \quad \text{for every} \quad i \in \mathbb{I}_d\ . \tag{11}$$

Then $c \geq \max \sigma(S - S_\mathcal{G})$. If we let

$$r = \operatorname{rk} S_\mathcal{G} = \max\{i \in \mathbb{I}_d :\ (\lambda_i(S) - c)^+ > 0\} = \max\{i \in \mathbb{I}_d :\ \lambda_i(S) > c\} \geq 1\ ,$$

then $\{0\} \neq W \overset{\text{def}}{=} R(S_\mathcal{G}) = \text{span}\{v_i :\ i \in \mathbb{I}_r\}$, and it satisfies that $(S - S_\mathcal{G})\big|_W = c\, I_W$. The proof finishes by noticing that, by Eq. (2), $g_j \in W$ and hence $(S - S_\mathcal{G})g_j = c\, g_j$ for every $j \in \mathbb{I}_k$. □

Remark 3.13. Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ (with $k \geq d$) such that the pair $(\lambda(S)\,,\,\mathbf{a})$ is co-feasible. Let $\mathcal{G} = \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ and $c \in \mathbb{R}$ be as in the proof of the second part of Proposition 3.12. Then, by Theorem 3.9, $\mathcal{G}$ is a global (and local) minimizer of the map $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$, with $p = 1$. Nevertheless, a priori this fact does not imply that every local minimizers should have the same structure (namely, to have also $p = 1$). We shall prove soon that the spectral structure of local minimizers is indeed unique (in general, and then also in the co-feasible cases). △

It is worth pointing out that there are GFOD problems that are not co-feasible. In order to see this we include the following:

Example 3.14. Consider $S \in \mathcal{M}_4(\mathbb{C})^+$ be such that $\lambda := \lambda(S) = (2, 2, 1, 1) \in (\mathbb{R}^4_{>0})^\downarrow$ and let $\mathbf{a} = (3, 1, 1, 1) \in (\mathbb{R}^4_{>0})^\downarrow$. Then, the pair $(\lambda\,,\,\mathbf{a})$ is not co-feasible. Indeed, the unique solution $c < 2$ to the equation $6 = \operatorname{tr}(\mathbf{a}) = 2\,(2 - c)^+ + 2\,(1 - c)^+$ is $c = 0$. Thus $((\lambda_i - c)^+)_{i\in\mathbb{I}_4} = \lambda$. But it can be easily checked that $\mathbf{a} \not\prec \lambda$. △

Although in general, given $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$, the pair $(\lambda(S)\,,\,\mathbf{a})$ corresponding to this data is not co-feasible, the GFOD problems contain a co-feasible part. Indeed, if we further consider a strictly convex u.i.n. N in $\mathcal{M}_d(\mathbb{C})$, then local minimizers of $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$ allow us to locate such co-feasible parts. In order to describe this situation, we introduce the following

Definition 3.15. Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ with $k \geq d$. For $r \in \mathbb{I}_{d-1} \cup \{0\}$ we consider the truncated data

$$\lambda^{(r)}(S) = (\lambda_{r+1}(S)\,,\,\dots\,,\,\lambda_d(S)) \in (\mathbb{R}^{d-r}_{\geq 0})^\downarrow \quad \text{and} \quad \mathbf{a}^{(r)} = (a_{r+1}\,,\,\dots\,,\,a_k) \in (\mathbb{R}^{k-r}_{>0})^\downarrow\ .$$

We say that r is a **co-feasible index** for S and $\mathbf{a}$ if the pair $(\lambda^{(r)}(S)\,,\,\mathbf{a}^{(r)})$ is co-feasible (according to Definition 3.11 with dimensions $d - r \leq k - r$). △

Remark 3.16. Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ with $k \geq d$. Let $\mathcal{B} = \{v_i\}_{i\in\mathbb{I}_d}$ be an ONB for $\mathbb{C}^d$ such that $S\,v_i = \lambda_i(S)\,v_i$ for $i \in \mathbb{I}_d$. Then, by Proposition 3.12, an index $r \in \mathbb{I}_{d-1} \cup \{0\}$ is co-feasible if and only if the conditions 1 and 2 of Proposition 3.12 hold for the space $V_r = \operatorname{span}\{v_i : \ r+1 \leq i \leq d\}$, the positive operator $S_r = S|_{V_r} \in L(V_r)$ and the vector of norms $\mathbf{a}^{(r)} = (a_{r+1}(S), \ldots, a_k) \in (\mathbb{R}^{k-r}_{>0})^\downarrow$. This means that there exist $c \in \mathbb{R}$ and

$$\mathcal{G} = \{g_i\}_{i\in\mathbb{I}_{k-r}} \in \mathbb{T}_{V_r}(\mathbf{a}^{(r)}) \overset{\text{def}}{=} \mathbb{T}_{k-r}(\mathbf{a}^{(r)}) \cap V_r^{k-r}$$

such that $(S_r - S_\mathcal{G})\,g_i = c\,g_i$, for every $i \in \mathbb{I}_{k-r}$, and $c = \max\sigma(S_r - S_\mathcal{G})$. Note that this statement seems to depend on the basis $\mathcal{B}$. But actually, the list of eigenvalues $\lambda(S_r) = \lambda^{(r)}(S) \in (\mathbb{R}^{d-r}_{\geq 0})^\downarrow$, so it does not depend on $\mathcal{B}$. △

The next result complements Theorem 3.8.

Proposition 3.17. *Consider Notation 3.4 with $k \geq d$ for $S \in \mathcal{M}_d(\mathbb{C})^+$ and a local minimizer $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ of the map $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$. Let $0 = s_0 < s_1 < \ldots < s_{p-1} < s_p \leq d$ be as in Proposition 3.5. Then $c_p = \max\sigma(S - S_{\mathcal{G}_0})$ and s_{p-1} is a co-feasible index for S and $\mathbf{a}$.*

In particular, the constant c_p and the index $s_p = \operatorname{rk} S_{\mathcal{G}_0}$ are uniquely determined by the equations

$$\begin{aligned} \sum_{i=s_{p-1}+1}^{k} a_i \ &= \sum_{i=s_{p-1}+1}^{d} (\lambda_i(S) - c_p)^+ \quad \textit{and} \\ s_p \ &= \max\big\{\, i \in \mathbb{N} : s_{p-1}+1 \leq i \leq d \quad \textit{and} \quad \lambda_i(S) - c_p > 0 \,\big\}\,. \end{aligned} \tag{12}$$

Proof. Let $S_0 = S_{\mathcal{G}_0}$. Note that $c_p = \max\sigma(S - S_0)$ by Remark 3.3, since we are assuming that $k \geq d$. In order to show that s_{p-1} is a co-feasible index we shall use Remark 3.16. Let $r = s_{p-1}$. Recall from Notation 3.4 and Proposition 3.5 that $J_p = \{i \in \mathbb{I}_k : (S - S_0)g_i = c_p\,g_i\} = \{r+1, \ldots, k\}$ and that $W_p = \operatorname{span}\{g_i : i \in J_p\} = \operatorname{span}\{v_j : r+1 \leq j \leq s_p\}$. Since

$$W = R(S_0) = \operatorname{span}\{v_i : j \in \mathbb{I}_{s_p}\} \quad \text{then} \quad V_r = \operatorname{span}\{v_i : \ r+1 \leq i \leq d\} = W_p \oplus W^\perp\,.$$

Then, $\mathcal{G}_r = \{g_i\}_{i=r+1}^k \in \mathbb{T}_{V_r}(\mathbf{a}^{(r)}) = \mathbb{T}_{k-r}(\mathbf{a}^{(r)}) \cap V_r^{k-r}$ is such that $S_{\mathcal{G}_r} = S_0|_{V_r}$ (here we use that, by Eq. (3), $g_j \in W_p^\perp$ for every $j \notin J_p$). So that, if $P_\mathcal{M}$ denotes the orthogonal projection onto a subspace $\mathcal{M} \subseteq \mathbb{C}^n$,

$$S|_{V_r} - S_{\mathcal{G}_r} = (S - S_0)|_{V_r} = c_p\,P_{W_p} + S\,P_{W^\perp} \implies (S|_{V_r} - S_{\mathcal{G}_r})\,g_i = c_p\,g_i \quad \text{for} \quad r+1 \leq i \leq k\,.$$

Hence $\max\sigma(S|_{V_r} - S_{\mathcal{G}_r}) \leq \max\sigma(S - S_0) = c_p$ and, by Remark 3.16, $s_{p-1} = r$ is a co-feasible index for S and $\mathbf{a}$. Then, by Definition 3.15, s_p and c_p are determined by Eq. (12). □

Remark 3.18. Consider Notation 3.4 with $k \geq d$ for $S \in \mathcal{M}_d(\mathbb{C})^+$ and a local minimizer $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ of the map $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$. Taking into account all objects and facts detailed in Notation 3.4, Remark 3.6, Theorem 3.8, Eq. (11) and Proposition 3.17, we conclude that $\lambda(S - S_{\mathcal{G}_0}) = \delta(S\,,\,\mathbf{a}\,,\,\mathcal{G}_0)^\downarrow$, with

$$\delta(S\,,\,\mathbf{a}\,,\,\mathcal{G}_0) \overset{\text{def}}{=} \Big(c_1\,\mathbb{1}_{s_1}\,,\,c_2\,\mathbb{1}_{s_2-s_1}\,,\,\ldots\,,\,c_{p-1}\,\mathbb{1}_{s_{p-1}-s_{p-2}}\,,\,\big(\min\{\lambda_i(S)\,,\,c_p\}\big)_{i=s_{p-1}+1}^{d}\Big)\,, \tag{13}$$

or $\delta(S\,,\,\mathbf{a}\,,\,\mathcal{G}_0) = \big(\min\{\lambda_i(S)\,,\,c_1\}\big)_{i\in\mathbb{I}_d}$ (if $p = 1$, the co-feasible case), where all data in this formula can be explicitly computed in terms of S, $\mathbf{a}$ and the index s_{p-1}. Indeed, this expression depends on $\mathcal{G}_0$ and N only through the index s_{p-1} which determines the previous indexes and constants by Theorem 3.8, and the co-feasible part which begins at s_{p-1}, so it determines s_p and c_p, by Proposition 3.17 via Eq. (12). Hence we shall denote $s_{p-1} = s_{p-1}(\mathcal{G}_0)$. △

We end this section with the following result, which compares the co-feasibility constants corresponding to different co-feasible indexes.

Corollary 3.19. *Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ with $k \geq d$ and assume that r, $s \in \mathbb{I}_{d-1}$ are co-feasible indexes for S and $\mathbf{a}$. Denote by $c(s)$ and $c(r)$ their co-feasibility constants. Then,*

$$s < r \implies c(s) \geq c(r) \ .$$

Proof. By Proposition 3.12, $\mathbf{a}^{(s)} \prec \big(\,(\lambda_i(S) - c(s)\,)^+\big)_{i=s+1}^d$ and $\mathbf{a}^{(r)} \prec \big(\,(\lambda_i(S) - c(r)\,)^+\big)_{i=r+1}^d$. Then

$$\sum_{i=r+1}^{k} a_i = \sum_{i=r+1}^{d} (\lambda_i(S) - c(r)\,)^+ \quad \text{and} \quad \sum_{i=s+1}^{r} a_i \leq \sum_{i=s+1}^{r} (\lambda_i(S) - c(s)\,)^+ \ . \tag{14}$$

Therefore,

$$\sum_{i=s+1}^{d} (\lambda_i(S) - c(s)\,)^+ = \sum_{i=s+1}^{k} a_i \leq \sum_{i=s+1}^{r} (\lambda_i(S) - c(s)\,)^+ + \sum_{i=r+1}^{d} (\lambda_i(S) - c(r)\,)^+ \ .$$

But if $c(s) < c(r)$ then $(\lambda_i(S) - c(s)\,)^+ \geq (\lambda_i(S) - c(r)\,)^+$ for every $i \in \mathbb{I}_d$, and moreover, we have that $(\lambda_{r+1}(S) - c(s)\,)^+ > (\lambda_{r+1}(S) - c(r)\,)^+$ because $\sum\limits_{i=r+1}^{k} a_i > 0 \overset{(14)}{\Longrightarrow} c(r) < \lambda_{r+1}(S)$. □

4 Main results

In this section we state and prove our main result namely, that local minimizers of GFOD's are actually global minimizers. This is achieved by considering in detail the results obtained in Section 3 related with the spectral structure of local minimizers of GFOD's functions, and the notion of co-feasible index. We first consider the case when $k \geq d$.

4.1 When $k \geq d$

Throughout this subsection we assume that $k \geq d$. Notice that Eqs. (7) and (8) together with Theorem 3.8 and Proposition 3.17 give a detailed description of the spectral structure of local minimizers of GFOD problems. With the notation of these results, it is worth pointing out the key role played by the (co-feasible) index s_{p-1} in the determination of the complete spectral structure of $S - S_0$ and S_0 (see Definition 3.11).

The basic idea for what follows is to replace s_{p-1} by an arbitrary co-feasible index r, to reproduce the algorithm given in Theorem 3.8 and get indexes and constants in terms of r (which a priori are not associated to any minimizer $\mathcal{G}_0$). Then, we shall show that there exists a unique "correct" index r (i.e. co-feasible and admissible, see Definition 4.1 below) which only depends on $\lambda(S)$ and $\mathbf{a}$, so that it must coincide with $s_{p-1}(\mathcal{G}_0)$.

Definition 4.1. Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$. For a co-feasible index $r \in \mathbb{I}_{d-1} \cup \{0\}$ let $q = q(r) \in \mathbb{I}_d$, $0 = s_0(r) < s_1(r) < \ldots < s_{q-1}(r) = r < s_q \leq d \leq k$ and $c_1(r), \ldots, c_q(r)$ be computed according to the following recursive algorithm (which only depends on r, $\lambda(S)$ and $\mathbf{a}$):

1. If $r = 0$, set $q = q(r) = 1$ and $s_0(r) = s_{q-1}(r) = r = 0$ (and go to item 4.).
2. If $r > 0$, using the numbers $P_{i\,,\,j}$ defined in Notation 3.7, the index

$$s_1(r) = \max\left\{1 \leq j \leq r \,:\, P_{1\,,\,j} = \min_{i\leq r} P_{1\,,\,i}\right\} \ , \qquad \text{and} \qquad c_1(r) = P_{1\,,\,s_1(r)} \ .$$

3. If the index $s_j(r)$ is already computed and $s_j(r) < r$, then

$$s_{j+1}(r) = \max \left\{ s_j(r) < j \le r \,:\, P_{s_j(r)+1\,,\,j} = \min_{s_j(r)<i\le r} P_{s_j(r)+1\,,\,i} \right\} , \tag{15}$$

and $c_{j+1}(r) = P_{s_j(r)+1\,,\,s_{j+1}(r)}$.

4. If $s_j(r) = r$, we set $q = q(r) = j+1$ (so that $s_{q-1}(r) = r$), and we define $c_q(r)$ and $s_q(r)$ (with $c_q(r) < \lambda_{r+1}$ and $r = s_{q-1}(r) < s_q(r) \le d$) that are uniquely determined by

$$\sum_{i=r+1}^{k} a_i = \sum_{i=r+1}^{d} \big(\lambda_i(S) - c_q(r)\,\big)^+ \quad \text{and} \tag{16}$$

$$s_q(r) = \max\{r+1 \le i \le d : \lambda_i(S) - c_q(r) > 0\} . \tag{17}$$

In particular, $s_q(r) = \max\{i \in \mathbb{I}_d : \lambda_i(S) - c_q(r) > 0\}$ since $\lambda(S) = \lambda(S)^\downarrow$.

5. If $r > 0$ we denote by $\delta(\lambda(S)\,,\,\mathbf{a}\,,\,r) \in \mathbb{R}^d$ the vector given by

$$\delta(\lambda(S)\,,\,\mathbf{a}\,,\,r) = \Big(c_1(r)\,\mathbb{1}_{s_1(r)}\,,\,\ldots\,,\,c_{q-1}(r)\,\mathbb{1}_{s_{q-1}(r)-s_{q-2}(r)}\,,\,\big(\,\min\{\lambda_i(S)\,,\,c_q(r)\}\,\big)_{i=r+1}^{d}\Big) , \tag{18}$$

and $\delta(\lambda(S)\,,\,\mathbf{a}\,,\,0) = \big(\,\min\{\lambda_i(S)\,,\,c_1(0)\}\,\big)_{i\in\mathbb{I}_d}$. It is easy to see (by construction) that

$$\operatorname{tr}\delta(\lambda(S)\,,\,\mathbf{a}\,,\,r) = \operatorname{tr}(S) - \operatorname{tr}(\mathbf{a}) . \tag{19}$$

Finally, we shall say that the index r is **admissible** if $r = 0$ or $r > 0$ and $c_{q-1}(r) < c_q(r)$. △

Remark 4.2. Consider a fixed strictly convex u.i.n. N in $\mathcal{M}_d(\mathbb{C})$. Let $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ be a local (or global) minimizer of $\Theta_{(N\,,\,S\,,\,\mathbf{a})} = \Theta : \mathbb{T}_d(\mathbf{a}) \to \mathbb{R}_{\ge 0}$. Assume that $k \ge d$.

We can apply the previous results to $\mathcal{G}_0$; thus, we consider $p \ge 1$ and constants $c_1 < \ldots < c_p$ and indexes $s_0 = 0 < s_1 < \ldots < s_p \le d$ as in Theorem 3.8 and Proposition 3.17. In particular, we get that s_{p-1} is a co-feasible index which is also admissible since, if $s_{p-1} > 0$, then $c_{p-1} < c_p$ by definition (see Theorem 3.2). The idea of what follows is to show that s_{p-1} (denoted $s_{p-1}(\mathcal{G}_0)$ in Remark 3.18) is the **unique** index which has both properties (for any norm N). First, we need to verify some properties of the vector $\delta(\lambda(S)\,,\,\mathbf{a}\,,\,r)$ for a co-feasible and admissible index. △

Proposition 4.3. *Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}_{>0}^k)^\downarrow$ (with $k \ge d$). Let $r \in \mathbb{I}_{d-1} \cup \{0\}$ be a co-feasible index. Then, with $p = q(r)$, $s_j = s_j(r)$, $c_j = c_j(r)$ for $j \in \mathbb{I}_p$, and $\delta = \delta(\lambda(S)\,,\,\mathbf{a}\,,\,r)$ as in Definition 4.1, we have that:*

1. *If $p > 1$ then $c_1 < \ldots < c_{p-1}$.*

If we also assume that r is admissible, then $c_{p-1} < c_p = \max_{i\in\mathbb{I}_d} \delta_i$ and:

2. *$\lambda_{s_{p-1}+1} \ge \lambda_{s_p} > c_p$ and $\lambda_i(S) > c_j$, for every $s_{j-1}+1 \le i \le s_j$ and $j \in \mathbb{I}_{p-1}$. Then*

$$\delta_i \le \min\{c_p\,,\,\lambda_i\} \quad \textit{for every} \quad i \in \mathbb{I}_d . \tag{20}$$

3. *If $p > 1$ then $(a_i)_{i=s_{j-1}+1}^{s_j} \prec \big(\lambda_i(S) - c_j\big)_{i=s_{j-1}+1}^{s_j} \in \mathbb{R}_{>0}^{s_j - s_{j-1}}$, for every $j \in \mathbb{I}_{p-1}$.*

4. *$(a_i)_{i=s_{p-1}+1}^{k} \prec \big((\lambda_i(S) - c_p)^+\big)_{i=s_{p-1}+1}^{d} \in \mathbb{R}_{\ge 0}^{d-s_{p-1}}$.*

Proof. 1. The case $p = 2$ is trivial. If $p > 2$, assume that there exists $j \in \mathbb{I}_{p-2}$ such that $c_j \geq c_{j+1}$. Then, notice that

$$P_{s_{j-1}+1\,,\,s_{j+1}} = \frac{s_j - s_{j-1}}{s_{j+1} - s_{j-1}}\; c_j + \frac{s_{j+1} - s_j}{s_{j+1} - s_{j-1}}\; c_{j+1} \leq c_j \ ,$$

which contradicts the definition of s_j in Eq. (15), since $s_{j+1} \leq s_{p-1} = r$. Thus, $c_1 < \ldots < c_{p-1}$. If $r = s_{p-1}$ is an admissible index, then $c_{p-1} < c_p = \max\limits_{i \in \mathbb{I}_d} \delta_i$ by definition and Eq. (18).

2. By Eq. (17), we have that $c_p < \lambda_i(S)$ for $s_{p-1} + 1 \leq i \leq s_p$. Therefore, if

$$j \in \mathbb{I}_{p-1} \quad \text{ and } \quad s_{j-1} + 1 \leq i \leq s_j \implies c_j < c_p < \lambda_{s_{p-1}+1}(S) \leq \lambda_i(S) \ ,$$

since $i \leq s_j \leq s_{p-1} < s_{p-1} + 1$ and $\lambda(S) \in (\mathbb{R}^d)^{\downarrow}$.

3. For $j \in \mathbb{I}_{p-1}$ and $s_{j-1} + 1 \leq m \leq s_j$, we have that

$$\sum_{i=s_{j-1}+1}^{m} a_i \leq \sum_{i=s_{j-1}+1}^{m} (\lambda_i(S) - c_j) \iff c_j \leq \frac{1}{m - s_{j-1}} \sum_{i=s_{j-1}+1}^{m} \lambda_i(S) - a_i \overset{(9)}{=} P_{s_{j-1}+1\,,\,m} \tag{21}$$

(the equivalence also holds for equalities). Using the definition of c_j (item 2. of Definition 4.1), we see that the inequalities to the right in Eq. (21) hold for every such index m, with equality for $m = s_j$ (by definition of c_j and s_j). We have proved that $(a_i)_{i=s_{j-1}+1}^{s_j} \prec \big(\lambda_i(S) - c_j\big)_{i=s_{j-1}+1}^{s_j}$.

Item 4 follows immediately from the fact that $r = s_{p-1}$ is a co-feasible index (see Definition 3.15). □

Corollary 4.4. *Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i \in \mathbb{I}_k} \in (\mathbb{R}^k_{>0})^{\downarrow}$ (with $k \geq d$). Let $r \in \mathbb{I}_{d-1} \cup \{0\}$ be a co-feasible index which is also admissible. Then $\mathbf{a} \prec \lambda(S) - \delta(\lambda(S)\,,\,\mathbf{a}\,,\,r)$.*

Proof. The relation $\mathbf{a} \prec \lambda(S) - \delta(\lambda(S)\,,\,\mathbf{a}\,,\,r)$ follows from items 3 and 4 of Proposition 4.3, since $x \prec y$ and $z \prec w \implies (x\,,\,z) \prec (y\,,\,w)$ (Remark 2.2). □

Theorem 4.5. *Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and $\mathbf{a} = (a_i)_{i \in \mathbb{I}_k} \in (\mathbb{R}^k_{>0})^{\downarrow}$ (with $k \geq d$). Then there is a unique co-feasible and admissible index $s \in \mathbb{I}_{d-1} \cup \{0\}$, and this s is the minimal co-feasible index.*

Proof. Assume that there exist two co-feasible indexes $0 \leq s < r \leq d - 1$ such that r is admissible. We show that this leads to a contradiction. Indeed, let $s_0 = 0 < s_1 < \ldots < s_{p-1} = r < s_p \leq d$ and $c_1 < \ldots < c_p$ be the indexes and constants corresponding to Definition 4.1, for the index r (i.e., we rename $p = q(r)$, $s_j = s_j(r)$ and $c_j = c_j(r)$ for $j \in \mathbb{I}_p$). Let $\lambda \overset{\text{def}}{=} \lambda(S) \in (\mathbb{R}^d_{\geq 0})^{\downarrow}$ and consider

$$\delta = \delta(\lambda\,,\,\mathbf{a}\,,\,r) = \Big(c_1\,\mathbb{1}_{s_1 - s_0}\,,\,\ldots\,,\,c_{p-1}\,\mathbb{1}_{s_{p-1} - s_{p-2}}\,,\,\big(\min\{\lambda_i\,,\,c_p\}\big)_{i=s_{p-1}+1}^{d}\Big)\ . \tag{22}$$

Similarly, consider $q = q(s)$ and $s^*_0 = 0 < s^*_1 < \ldots < s^*_{q-1} = s < s^*_q \leq d$ and $c^*_1 < \ldots < c^*_{q-1}$ and c^*_q be the indexes and constants corresponding to Definition 4.1, for the index s. We also consider

$$\delta^* = \delta(\lambda\,,\,\mathbf{a}\,,\,s) = \Big(c^*_1\,\mathbb{1}_{s^*_1 - s^*_0}\,,\,\ldots\,,\,c^*_{q-1}\,\mathbb{1}_{s^*_{q-1} - s^*_{q-2}}\,,\,\big(\min\{\lambda_i\,,\,c^*_q\}\big)_{i=s^*_{q-1}+1}^{d}\Big)\ . \tag{23}$$

If $\delta^* = \delta$ then by Eqs. (17), (22) and (23), $s^*_q = s_p = \max\{i \in \mathbb{I}_d : \delta_i < \lambda_i(S)\}$ and $c^*_q = c_p = \delta_{s_p}$. But in this case $r = s_{p-1} = \min\{i \in \mathbb{I}_{d-1} : \delta_{i+1} = c_p\} \leq s^*_{q-1} = s$, a contradiction. Hence $\delta^* \neq \delta$.

Case 1. Assume that there exists $1 \leq j \leq \min\{p-1\,,\,q-1\}$ such that

$$s_i = s^*_i \quad \text{(and then also } c_i = c^*_i\text{)} \quad \text{for } \ 0 \leq i \leq j-1 \quad , \quad \text{but } \ s_j \neq s^*_j\ .$$

Next we show that this leads to a contradiction ($\delta^* = \delta$). Indeed, since $s_{j-1} = s^*_{j-1}$, by construction

$$s_j = \max\{s_{j-1} < i \le r : \ P_{s_{j-1}+1\,,\,i} = \min_{s_{j-1}+1\le \ell\le r} P_{s_{j-1}+1\,,\,\ell}\} \quad \text{with} \quad c_j = P_{s_{j-1}+1\,,\,s_j}$$

and

$$s^*_j = \max\{s_{j-1} < i \le s : \ P_{s_{j-1}+1\,,\,i} = \min_{s_{j-1}+1\le \ell\le s} P_{s_{j-1}+1\,,\,\ell}\} \quad \text{with} \quad c^*_j = P_{s_{j-1}+1\,,\,s^*_j} \ .$$

Using that the limits $s < r$, then $\min\limits_{s_{j-1}+1\le \ell\le s} P_{s_{j-1}+1\,,\,\ell} \ge \min\limits_{s_{j-1}+1\le \ell\le r} P_{s_{j-1}+1\,,\,\ell}$. Since $s^*_j \neq s_j$, this fact easily shows that

$$c_j \le c^*_j \quad \text{and} \quad s^*_{q-1} = s < s_j \le r \ . \tag{24}$$

On the other hand, by Corollary 3.19 we have that $c^*_q = c_q(s) \ge c_p(r) = c_p$, since they are the co-feasible constants corresponding to the co-feasible indexes $s^*_{q-1} = s < r = s_{p-1}$.

With these facts we can compare δ and δ^*:

- We have that $\delta_i = \delta^*_i$ for $1 \le i \le s_{j-1} = s^*_{j-1}$ by hypothesis.
- By Eq. (22), (23), and item 1 of Proposition 4.3 ($c^*_j < \cdots < c^*_{q-1}$),

$$\delta^*_i \ge c^*_j \overset{(24)}{\ge} c_j = \delta_i \quad \text{for} \quad s_{j-1} = s^*_{j-1} < i \le s^*_{q-1} \overset{(24)}{<} s_j \ .$$

- Since $c_p = \max\{\delta_j : j \in \mathbb{I}_d\}$ by Proposition 4.3 (r is admissible), then

$$\delta^*_i = c^*_q \ge c_p \ge \delta_i \quad \text{for} \quad s^*_{q-1} < i \le s^*_q \ .$$

- Finally, $\delta^*_i = \lambda_i \ge \delta_i$, for $s^*_q < i \le d$ (item 2 in Proposition 4.3).

Therefore $\delta \leqslant \delta^*$. Since $\operatorname{tr}(\delta) = \operatorname{tr}(S) - \operatorname{tr}(\mathbf{a}) = \operatorname{tr}(\delta^*)$ by Eq. (19), we get that $\delta = \delta^*$, a contradiction.

Case 2. If we assume that $p \le q$ and $s_j = s^*_j$ (and hence $c_j = c^*_j$) for $0 \le j \le p-1$, then

$$s_{p-1} = s^*_{p-1} \le s^*_{q-1} = s < r = s_{p-1} \ .$$

Case 3. Finally, if $q < p$ and $s_j = s^*_j$ (and hence $c_j = c^*_j$) for $0 \le j \le q-1$, then we have that $\delta_i = \delta^*_i$ for $1 \le i \le s^*_{q-1} = s_{q-1}$. Then, by Proposition 4.3, we have that

$$c_p \le c^*_q \implies \delta_i \overset{(20)}{\le} \min\{c_p\,,\,\lambda_i\} \le \min\{c^*_q\,,\,\lambda_i\} \overset{(23)}{=} \delta^*_i \quad \text{for} \quad s^*_{q-1} < i \le d \ .$$

Hence, $\delta \le \delta^*$. Using that $\operatorname{tr}(\delta) = \operatorname{tr}(\delta^*)$, also in this case we conclude that $\delta = \delta^*$. The proof finishes once we notice that one of these three cases should occur. □

Definition 4.6. Let $S \in \mathcal{M}_d(\mathbb{C})^+$ with $\lambda = \lambda(S) \in (\mathbb{R}^d_{\ge 0})^\downarrow$ and $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ (with $k \ge d$). If $s \in \mathbb{I}_{d-1} \cup \{0\}$ is the unique co-feasible and admissible index for S and $\mathbf{a}$ (which exists by Remark 4.2), then we denote by $\delta(\lambda\,,\,\mathbf{a}) \overset{\text{def}}{=} \delta(\lambda\,,\,\mathbf{a}\,,\,s)$ as in Eq. (18) of Definition 4.1. △

Remark 4.7. With the notation of Definition 4.6 above, notice that the vector $\delta(\lambda\,,\,\mathbf{a})$ can be computed using a fast algorithm. Indeed, the notion of co-feasible and admissible index is algorithmic and can be checked using a fast routine; once the unique co-feasible and admissible index is computed, the vector $\delta(\lambda\,,\,\mathbf{a})$ can also be computed using a fast algorithm (Definition 4.1). △

Theorem 4.8. *Let $S \in \mathcal{M}_d(\mathbb{C})^+$ with $\lambda = \lambda(S) \in (\mathbb{R}^d_{\ge 0})^\downarrow$, $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ (with $k \ge d$) and $\delta(\lambda\,,\,\mathbf{a})$ as in Definition 4.6. If N is a strictly convex u.i.n. in $\mathcal{M}_d(\mathbb{C})$ and $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ then, the following statements are equivalent:*

1. $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ *is a global minimizer of* $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$;

2. $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ *is a local minimizer of* $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$;

3. $\lambda(S - S_{\mathcal{G}_0}) = \delta(\lambda\,,\,\mathbf{a})^{\downarrow}$.

Hence, the global (and local) minimizers are the same for every strictly convex u.i.n. N.

Proof. Clearly, 1. ⇒ 2. In order to see 2. ⇒ 3., we recall Remarks 3.6, 3.18 and 4.2, where we have seen that $\lambda(S - S_{\mathcal{G}_0}) = \delta(S\,,\,\mathbf{a}\,,\,\mathcal{G}_0)^{\downarrow}$, for the vector $\delta(\lambda\,,\,\mathbf{a}\,,\,\mathcal{G}_0)$ given in Eq. (13) and completely determined by the index called $s_{p-1}(\mathcal{G}_0)$. By Remark 4.2 and Theorem 4.5, this $s_{p-1}(\mathcal{G}_0)$ is the unique co-feasible and admissible index of Theorem 4.5. Therefore, by Equations (13) and (18),

$$\delta(\lambda(S)\,,\,\mathbf{a}) = \delta(S\,,\,\mathbf{a}\,,\,\mathcal{G}_0) \implies \lambda(S - S_{\mathcal{G}_0}) = \delta(\lambda\,,\,\mathbf{a})^{\downarrow} \ .$$

3. ⇒ 1. Notice that Θ is a continuous function defined on a compact metric space, so then there exists $\mathcal{G}_1 \in \mathbb{T}_d(\mathbf{a})$ that is a global minimizer of Θ and, in particular, a local minimizer. By the already proved 2. ⇒ 3., we must have that $\lambda(S - S_{\mathcal{G}_1}) = \delta(\lambda\,,\,\mathbf{a})^{\downarrow} = \lambda(S - S_{\mathcal{G}_0})$.

In particular, since N is unitarily invariant

$$\Theta_{(N\,,\,S\,,\,\mathbf{a})}(\mathcal{G}_0) = N(S - S_{\mathcal{G}_0}) = N(D_{\delta(\lambda\,,\,\mathbf{a})}) = N(S - S_{\mathcal{G}_1}) = \Theta_{(N\,,\,S\,,\,\mathbf{a})}(\mathcal{G}_1) \ ,$$

where $D_{\delta(\lambda\,,\,\mathbf{a})} \in \mathcal{M}_d(\mathbb{C})$ denotes the diagonal matrix with main diagonal $\delta(\lambda\,,\,\mathbf{a})$. □

We end this section with the following examples.

Example 4.9. Consider $\mathcal{B} = \{e_1\,,\,e_2\}$ the canonical basis of $\mathbb{C}^2$. Let $S = 3\,e_1 \otimes e_1 + e_2 \otimes e_2 \in \mathcal{M}_2(\mathbb{C})^+$ and $\mathbf{a} = (1\,,\,1)$ (i.e. $k = d = 2$). Then S is an invertible operator. Consider the vectors $g_1 = g_2 = e_1$, and $\mathcal{G}_0 = \{g_1\,,\,g_2\} \in \mathbb{T}_2(\mathbf{a})$. Then $\lambda(S - S_{\mathcal{G}_0}) = \lambda(e_1 \otimes e_1 + e_2 \otimes e_2) = (1\,,\,1)$. If $\mathcal{G} \in \mathbb{T}_2(\mathbf{a})$ is arbitrary, then $\operatorname{tr} \lambda(S - S_{\mathcal{G}}) = \operatorname{tr}\, S - \operatorname{tr}\, S_{\mathcal{G}} = 2$. Hence

$$\lambda(S - S_{\mathcal{G}_0}) = (1\,,\,1) \prec \lambda(S - S_{\mathcal{G}}) \implies s(S - S_{\mathcal{G}_0}) = (1\,,\,1) \prec_w s(S - S_{\mathcal{G}}) \ ,$$

by Remark 2.2 and Theorem 2.4. Then $\Theta_{(N\,,\,S\,,\,\mathbf{a})}(\mathcal{G}_0) \leq \Theta_{(N\,,\,S\,,\,\mathbf{a})}(\mathcal{G})$, for every u.i.n. N. Thus, $\mathcal{G}_0 = \{e_1\,,\,e_1\}$ is a global minimizer of $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$ in $\mathbb{T}_2(\mathbf{a})$. Therefore this problem is co-feasible, so that $p = 1$, $s_1 = \operatorname{rk} S_{\mathcal{G}_0} = 1$ and $c_1 = \lambda_2(S) = 1$. Notice that in this case $\mathcal{G}_0$ is not a frame for $\mathbb{C}^2$ (even when $S \in \mathcal{M}_2(\mathbb{C})^+$ is invertible and $k \geq d$). △

Example 4.10. Consider $\mathcal{B} = \{e_1\,,\,e_2\}$ the canonical basis of $\mathbb{C}^2$. Let $S = e_1 \otimes e_1 \in \mathcal{M}_2(\mathbb{C})^+$ and $\mathbf{a} = (2, 1)$ (with $k = d = 2$ again). Then S is a non-invertible operator. We shall see that $\mathcal{G}_0 = \{\sqrt{2}\,e_1\,,\,e_2\} \in \mathbb{T}_2(\mathbf{a})$ is a global minimizer of $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$, for every u.i.n. N. Indeed,

$$\lambda(S - S_{\mathcal{G}_0}) = \lambda(-e_1 \otimes e_1 - e_2 \otimes e_2) = (-1\,,\,-1) \implies s(S - S_{\mathcal{G}_0}) = |\lambda(S - S_{\mathcal{G}_0})| = (1, 1)$$

and, if $\mathcal{G} \in \mathbb{T}_2(\mathbf{a})$ is arbitrary, then $\operatorname{tr}\, \lambda(S - S_{\mathcal{G}}) = 1 - 3 = -2$, so that $\operatorname{tr}\, s(S - S_{\mathcal{G}}) \geq 2$. This last fact implies that $s(S - S_{\mathcal{G}_0}) \prec_w s(S - S_{\mathcal{G}})$ and therefore $\Theta_{(N\,,\,S\,,\,\mathbf{a})}(\mathcal{G}_0) \leq \Theta_{(N\,,\,S\,,\,\mathbf{a})}(\mathcal{G})$. Also this problem is co-feasible, with $p = 1$, $s_1 = \operatorname{rk} S_{\mathcal{G}_0} = 2$ and $c_1 = -1$. Notice that in this case $\mathcal{G}_0$ is a frame for $\mathbb{C}^2$ (even when $S \in \mathcal{M}_2(\mathbb{C})^+$ is not an invertible operator). △

4.2 The general case

So far, we have considered the case of local minimizers of GFOD functions when the number of vectors k is greater than or equal to the dimension of the space d. This was essentially needed in Section 3.3. In this section we add the case when $k < d$, thus covering all possible cases. Our approach is based on a reduction to the case considered in Section 4.1.

Definition 4.11. Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and let $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ with $k < d$. Let $\mathcal{B} = \{v_i\}_{i\in\mathbb{I}_d}$ be an ONB of $\mathbb{C}^d$ such that $S = \sum_{i\in\mathbb{I}_d} \lambda_i(S)\ v_i \otimes v_i$. Let

$$V_k = \operatorname{span}\{v_i : i \in \mathbb{I}_k\} \quad \text{and} \quad S_k \overset{\text{def}}{=} S|_{V_k} = \sum_{i\in\mathbb{I}_k} \lambda_i(S)\ v_i \otimes v_i \in L(V_k)^+ \ .$$

Since $k = \dim V_k$ (the "new d") we can take $\delta(\lambda(S_k)\,,\,\mathbf{a}) \in \mathbb{R}^k$ using Definition 4.6, for the data $\lambda(S_k) = (\lambda_1(S)\,,\,\dots\,,\,\lambda_k(S)\,) \in (\mathbb{R}^k_{\geq 0})^\downarrow$ and $\mathbf{a} \in (\mathbb{R}^k_{>0})^\downarrow$. We define the vector

$$\delta(\lambda(S)\,,\,\mathbf{a}) \overset{\text{def}}{=} \big(\,\delta(\lambda(S_k)\,,\,\mathbf{a})\,,\,\lambda_{k+1}(S)\,,\,\dots\,,\,\lambda_d(S)\,\big)\ ,$$

which does not really depends on S_k and $\mathcal{B}$, but only on $\lambda(S)$ and $\mathbf{a}$. △

Theorem 4.12. *Let $S \in \mathcal{M}_d(\mathbb{C})^+$, let $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$ and let N be a strictly convex u.i.n. in $\mathcal{M}_d(\mathbb{C})$. Given $\mathcal{G}_0 = \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ the following are equivalent:*

1. *$\mathcal{G}_0$ is a global minimizer of $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$;*
2. *$\mathcal{G}_0$ is a local minimizer of minimizer of $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$;*
3. *$\lambda(S - S_{\mathcal{G}_0}) = \delta(\lambda(S)\,,\,\mathbf{a})^\downarrow$ (see Definition 4.6 if $k \geq d$, and Definition 4.11 if $k < d$).*

Proof. If $k \geq d$ this is Theorem 4.8. Let us assume that $k < d$.

Clearly 1. ⇒ 2. If we assume 2 we can apply Theorem 3.2, Proposition 3.5 and Theorem 3.8 (these statements do not assume that $k \geq d$). With the notation of these results (i.e., with Notation 3.4), there exists $\mathcal{B} = \{v_i\}_{i\in\mathbb{I}_d}$ an ONB of $\mathbb{C}^d$ such that

$$S = \sum_{i\in\mathbb{I}_d} \lambda_i(S)\ v_i \otimes v_i \quad \text{and} \quad S_0 = S_{\mathcal{G}_0} = \sum_{i\in\mathbb{I}_d} \lambda_i(S_0)\ v_i \otimes v_i\,.$$

We have that $r \overset{\text{def}}{=} \operatorname{rk} S_0 \leq k$, and $W = R(S_{\mathcal{G}_0}) = \operatorname{span}\{v_i\}_{i\in\mathbb{I}_r} \subseteq \operatorname{span}\{v_i\}_{i\in\mathbb{I}_k} = V_k$, as in Definition 4.11. Since $\lambda_i(S_0) = 0$ for $i > k$, the vector $\delta \overset{\text{def}}{=} \big(\,\lambda_i(S) - \lambda_i(S_0)\,\big)_{i\in\mathbb{I}_k} \in \mathbb{R}^k$ satisfies that

$$\lambda(S - S_0) = \big(\,\lambda(S) - \lambda(S_0)\,\big)^\downarrow \quad \text{and} \quad \lambda(S) - \lambda(S_0) = \big(\,\delta\,,\,\lambda_{k+1}(S)\,,\,\dots\,,\,\lambda_d(S)\,\big)\ . \tag{25}$$

With the notation of Definition 4.11, we have to prove that $\delta = \delta(\lambda(S_k)\,,\,\mathbf{a})$. Since $r = s_p \leq k < d$, we can apply Remark 3.6 (to $\lambda(S) - \lambda(S_0) \in \mathbb{R}^d$), so that

$$\delta = \big(\,c_1\,\mathbb{1}_{s_1}\,,\,c_2\,\mathbb{1}_{s_2-s_1}\,,\,\dots\,,\,c_p\,\mathbb{1}_{s_p-s_{p-1}}\,,\,\lambda_{s_p+1}(S)\,,\,\dots\,,\,\lambda_k(S)\,\big) \quad \text{if} \quad s_p < k \tag{26}$$

or

$$\delta = \big(\,c_1\,\mathbb{1}_{s_1}\,,\,c_2\,\mathbb{1}_{s_2-s_1}\,,\,\dots\,,\,c_p\,\mathbb{1}_{s_p-s_{p-1}}\,\big) \quad \text{if} \quad s_p = k\ , \tag{27}$$

where the indexes $s_1 < \ldots < s_{p-2}$ and constants $c_1 < \ldots c_{p-1}$ are constructed (for $\lambda(S - S_{\mathcal{G}_0})$ and therefore also for δ) in terms of the index s_{p-1} (when $p > 1$) using the algorithm given in Theorem 3.8 (and also in Definition 4.1, with respect to $\lambda(S_k)$, $\mathbf{a}$ and s_{p-1}). Also $c_{p-1} < c_p$ by Theorem 3.2.

Therefore, in order to show that $\delta = \delta(\lambda(S_k)\,,\,\mathbf{a})$, by Theorem 4.5 we just need to prove that the index $s_{p-1} \in \mathbb{I}_{k-1} \cup \{0\}$ is co-feasible (and admissible) with respect to S_k and $\mathbf{a}$. By Theorems 3.2 and 3.5 we know that $(S - S_0)g_i = c_p\, g_i \iff s_{p-1} + 1 \le i \le k$, and

$$W_p = \operatorname{span}\{g_i :\ s_{p-1} + 1 \le i \le k\ \} = \operatorname{span}\{v_i :\ s_{p-1} + 1 \le i \le s_p\ \}\ .$$

Hence, if we let $X = \operatorname{span}\{v_i :\ s_{p-1} + 1 \le i \le k\ \}$ and $\mathcal{G}_p = \{g_i\}_{i=s_{p-1}+1}^{k} \in \mathbb{T}_X(\mathbf{a}^{(s_{p-1})})$ then

$$(S_k|_X - S_{\mathcal{G}_p})\, g_i = (S - S_0)\, g_i = c_p\, g_i\ , \qquad \text{for} \qquad s_{p-1} + 1 \le i \le k\ .$$

By Remark 3.16 (for S_k and $\mathbf{a}$), we only need to show that $c_p = \max\limits_{s_{p-1}+1 \le i \le k} \delta_i\ (= \max\limits_{i \in \mathbb{I}_k} \delta_i\ )$.

Suppose that $c_p < \max \sigma(S - S_0)$. Then, by item 5 of Theorem 3.2, the set $\mathcal{G}_0$ is linearly independent (since each set $\{g_j\}_{j \in J_j}$ is linearly independent, and they are sets of eigenvectors of the different eigenvalues c_j). Then $s_p = \operatorname{rk} S_0 = k$, so we can apply Eq. (27), and automatically $c_p = \max\limits_{i \in \mathbb{I}_k} \delta_i$.

Otherwise we have that $c_p = \max \sigma(S - S_0) \ge \max\limits_{i \in \mathbb{I}_k} \delta_i$. Then, in any case $c_p = \max\limits_{i \in \mathbb{I}_k} \delta_i$. We have proved that the index s_{p-1} is co-feasible (and also admissible, because $c_{p-1} < c_p$) with respect to S_k and $\mathbf{a}$. Then $\delta = \delta(\lambda(S_k)\,,\,\mathbf{a})$ by Theorem 4.5 and $\lambda(S - S_{\mathcal{G}_0}) = \delta(\lambda(S)\,,\,\mathbf{a})^{\downarrow}$ by Eq. (25).

3. $\Rightarrow$ 1. An argument analogous to that in the proof of Theorem 4.8 (3. $\Rightarrow$ 1.) proves this implication. □

Remark 4.13. The proof of 2. $\Rightarrow$ 3. of Theorem 4.12 becomes trivial if we assume that (the vectorial version of) the norm N satisfies that, for $x\,,\,y \in \mathbb{R}^k$ and $z \in \mathbb{R}^{d-k}$,

$$N(x\,,\,z) \le N(y\,,\,z) \implies N(x\,,\,0) \le N(y\,,\,0)\ , \tag{28}$$

since in this case $\mathcal{G}_0$ is still a local minimizer for S_k and $\mathbf{a}$ in V_k. The most usual strictly convex norms (for example p-norms, for $p \in (1, \infty)$) satisfy Eq. (28), but this property fails in general. Take $N = \|\cdot\|_\infty + \|\cdot\|_2$ which is a strictly convex u.i.n. In this case, if $r = \frac{\sqrt{2}}{2}$, then ($d = 3$, $k = 2$)

$$N\big((0\,,\,1)\,,\,1\big) = 1 + \sqrt{2}\ = N\big((r\,,\,r)\,,\,1\big) \ \text{ but } \ N\big((0\,,\,1)\,,\,0\big) = 2 > r + 1 = N\big((r\,,\,r)\,,\,0\big)\ . \quad \triangle$$

Corollary 4.14. *With the notation of Theorem 4.12, we have that*

$$|\delta(\lambda\,,\,\mathbf{a})| \prec_w |\lambda(S - S_{\mathcal{G}})| \quad , \qquad \textit{for every} \qquad \mathcal{G} \in \mathbb{T}_d(\mathbf{a})\ .$$

Proof. For $h \in \mathbb{I}_d$ and $\varepsilon > 0$ let

$$N_{(h\,,\,\varepsilon)}(A) = N_{(h)}(A) + \varepsilon\,\|A\|_2 = \sum_{i \in \mathbb{I}_h} s_i(A) + \varepsilon\,\|A\|_2\ , \qquad \text{for} \qquad A \in \mathcal{M}_d(\mathbb{C})\ .$$

Then, $N_{(h\,,\,\varepsilon)}$ is a strictly convex u.i.n. in $\mathcal{M}_d(\mathbb{C})$ such that $\lim\limits_{\varepsilon \to 0^+} N_{(h\,,\,\varepsilon)}(A) = N_{(h)}(A)$, for $A \in \mathcal{M}_d(\mathbb{C})$. If we let $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ be such that $\lambda(S - S_{\mathcal{G}_0}) = \delta(\lambda\,,\,\mathbf{a})^{\downarrow}$ then, by Theorem 4.12,

$$\begin{aligned}\textstyle\sum_{i \in \mathbb{I}_h} |\delta(\lambda\,,\,\mathbf{a})|_i^{\downarrow} &= N_{(h)}(S - S_{\mathcal{G}_0}) = \lim_{\varepsilon \to 0^+} N_{(h\,,\,\varepsilon)}(S - S_{\mathcal{G}_0}) \\ &\le \lim_{\varepsilon \to 0^+} N_{(h\,,\,\varepsilon)}(S - S_{\mathcal{G}}) = \sum_{i \in \mathbb{I}_h} |\lambda(S - S_{\mathcal{G}})|_i^{\downarrow}\ .\end{aligned}$$

Since this occurs for every $h \in \mathbb{I}_d$, then $|\delta(\lambda\,,\,\mathbf{a})| \prec_w |\lambda(S - S_{\mathcal{G}})|$. □

5 Proof of some technical results

In this section we prove some results stated in Section 3.2. We begin by re-stating Notation 3.4, that we will use again throughout this section.

Notation 3.4 (repeated). Fix $S \in \mathcal{M}_d(\mathbb{C})^+$, $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}^k_{>0})^\downarrow$, and a strictly convex u.i.n. N on $\mathcal{M}_d(\mathbb{C})$. Also consider the notions introduced in Theorem 3.2. As before, let

1. $\Theta_{(N\,,\,S\,,\,\mathbf{a})} = \Theta : \mathbb{T}_d(\mathbf{a}) \to \mathbb{R}_{\geq 0}$ given by $\Theta(\mathcal{G}) = N(S - S_\mathcal{G})$.
2. A local minimizer $\mathcal{G}_0 = \{g_i\}_{i\in\mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ of $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$, with frame operator $S_0 = S_{\mathcal{G}_0}$.
3. We denote by $\lambda = (\lambda_i)_{i\in\mathbb{I}_d} = \lambda(S) \in (\mathbb{R}^d_{\geq 0})^\downarrow$ and $\mu = (\mu_i)_{i\in\mathbb{I}_d} = \lambda(S_0) \in (\mathbb{R}^d_{\geq 0})^\downarrow$.
4. We fix $\mathcal{B} = \{v_i\}_{i\in\mathbb{I}_d}$ an ONB of $\mathbb{C}^d$ as in Theorem 3.2. Hence,
$$S = \sum_{i\in\mathbb{I}_d} \lambda_i \ v_i \otimes v_i \quad \text{and} \quad S_0 = \sum_{i\in\mathbb{I}_d} \mu_i \ v_i \otimes v_i \ ,$$
5. We consider $W = R(S_0)$, $D = (S - S_0)|_W$ and $\sigma(D) = \{c_1\,,\,\dots\,,\,c_p\}$ where $c_1 < c_2 < \dots < c_p$.
6. Let $s_D = \max\{i \in \mathbb{I}_d : \mu_i \neq 0\} = \operatorname{rk} S_0$.
7. We denote by $\delta = \lambda - \mu \in \mathbb{R}^d$ so that
$$S - S_0 = \sum_{i\in\mathbb{I}_d} \delta_i \ v_i \otimes v_i \quad \text{and} \quad D = \sum_{i=1}^{s_D} \delta_i \ v_i \otimes v_i \ .$$
Notice that δ is constructed by pairing the entries of ordered vectors (since $\lambda = \lambda(S)$ and $\mu = \lambda(S_0)$ Nevertheless, we have that $\lambda(S - S_0) = \delta^\downarrow$. In what follows we obtain some properties of (the unordered vector) δ.
8. For each $j \in \mathbb{I}_p$, we consider the following sets of indexes:
$$K_j = \{i \in \mathbb{I}_{s_D} : \delta_i = \lambda_i - \mu_i = c_j\} \quad \text{and} \quad J_j = \{i \in \mathbb{I}_k : D\, g_i = c_j\, g_i\} \ .$$
Theorem 3.2 assures that $\mathbb{I}_{s_D} = \overset{\text{D}}{\bigcup}_{j\in\mathbb{I}_p} \ K_j$ and $\mathbb{I}_k = \overset{\text{D}}{\bigcup}_{j\in\mathbb{I}_p} \ J_j$ (disjoint unions).
9. By Eq. (2), $R(S_0) = \operatorname{span}\{g_i : i \in \mathbb{I}_k\} = W = \bigoplus_{i\in\mathbb{I}_p} \ker(D - c_i\, I_W\,)$ then, for every $j \in \mathbb{I}_p$,
$$W_j = \operatorname{span}\{g_i : i \in J_j\} = \ker(D - c_j\, I_W\,) = \operatorname{span}\{v_i : i \in K_j\} \ , \tag{29}$$
because $g_i \in \ker(D - c_j\, I_W\,)$ for every $i \in J_j$. Note that, by Theorem 3.2, each W_j reduces both S and S_0 . △

In order to prove Proposition 3.5 we first present the following two results. Recall that, given $x\,,\,y \in (\mathbb{R}^d)^\downarrow$, we say that x is **strictly** majorized by y if $x \prec y$ and $x \neq y$.

Proposition 5.1. *Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and let $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ be as in Notation 3.4 and assume that $p > 1$. Assume that there exist*
$$i < r \leq p \quad , \quad h \in J_i \quad , \quad l \in J_r \quad \textit{with} \quad l < h \quad (\Rightarrow a_l \geq a_h)\,. \tag{30}$$
Then, there exists a continuous curve $\mathcal{G}(t) : [0,1) \to \mathbb{T}_d(\mathbf{a})$ such that $\mathcal{G}(0) = \mathcal{G}_0$ and $\lambda(S - S_{\mathcal{G}(t)}) \prec \lambda(S - S_0)$ with strict majorization for $t \in (0,\varepsilon)$, for some $\varepsilon > 0$.

Proof. Consider

$$w_h = g_h / \|g_h\| = a_h^{-1/2} g_h \quad \text{and} \quad w_l = g_l / \|g_l\| = a_l^{-1/2} g_l ,$$

(note that $\langle w_h, w_l\rangle = 0$ because $\langle g_h, g_l\rangle = 0$). Now define, for $t \in \mathbb{R}$ and for some convenient $\gamma \in \mathbb{R}\setminus\{0\}$ (which will be explicitly calculated later),

$$g_h(t) = \cos(t)\, g_h + \sin(t)\ \|g_h\|\ w_l \quad \text{and} \quad g_l(t) = \cos(\gamma t)\, g_l + \sin(\gamma t)\ \|g_l\|\ w_h .$$

Then consider the family $\mathcal{G}_\gamma(t)$, which is obtained from $\mathcal{G}_0$ by replacing the vectors g_h and g_l by $g_h(t)$ and $g_l(t)$ respectively, and denote by $S_\gamma(t)$ its frame operator. Note that $\mathcal{G}_\gamma(t) \in \mathbb{T}_d(\mathbf{a})$ for every $t \in \mathbb{R}$ and $\mathcal{G}_\gamma(0) = \mathcal{G}_0$.

Let $W_{h,l} = \operatorname{span}\{w_h, w_l\}$. This subspace reduces both $S - S_0$ and $S - S_\gamma(t)$. Since $g_h(t), g_l(t) \in W_{h,l}$, we can represent the following matrices with respect to the basis $\{w_h\,,\ w_l\}$ of $W_{h,l}$,

$$g_h \otimes g_h = \begin{pmatrix} a_h & 0 \\ 0 & 0 \end{pmatrix}, \quad g_h(t)\otimes g_h(t) = a_h \begin{pmatrix} \cos^2(t) & \cos(t)\sin(t) \\ \cos(t)\sin(t) & \sin^2(t) \end{pmatrix},$$

$$g_l \otimes g_l = \begin{pmatrix} 0 & 0 \\ 0 & a_l \end{pmatrix}, \quad g_l(t)\otimes g_l(t) = a_l \begin{pmatrix} \sin^2(\gamma t) & \cos(\gamma t)\sin(\gamma t) \\ \cos(\gamma t)\sin(\gamma t) & \cos^2(\gamma t) \end{pmatrix}.$$

Then,

$$S - S_\gamma(t) = S - S_0 - g_h(t)\otimes g_h(t) - g_l(t)\otimes g_l(t) + g_h\otimes g_h + g_l \otimes g_l .$$

Hence $(S - S_0)|_{W_{h,l}^\perp} = (S - S_\gamma(t))|_{W_{h,l}^\perp}$. On the other hand $(S - S_0)|_{W_{h,l}} = \begin{pmatrix} c_i & 0 \\ 0 & c_r \end{pmatrix}$ and

$$(S - S_\gamma(t))|_{W_{h,l}} = \begin{pmatrix} c_i + a_h \sin^2(t) - a_l \sin^2(\gamma t) & -a_h\cos(t)\sin(t) - a_l\cos(\gamma t)\sin(\gamma t) \\ -a_h\cos(t)\sin(t) - a_l\cos(\gamma t)\sin(\gamma t) & c_r + a_l\sin^2(\gamma t) - a_h\sin^2(t) \end{pmatrix}$$

Denote this matrix by $A_\gamma(t)$. Since $\operatorname{tr}(A_\gamma(t)) = c_i + c_r$ for every $t \in \mathbb{R}$, then we have the strict majorization $\lambda(A_\gamma(t)) \prec (c_r, c_i)$ if and only if $\|A_\gamma(t)\|_2^2 < c_r^2 + c_i^2$. So consider the map

$$m_\gamma : \mathbb{R} \to \mathbb{R} \quad \text{given by} \quad m_\gamma(t) = \|A_\gamma(t)\|_2^2 = \operatorname{tr}(A_\gamma(t)^2) \quad \text{for} \quad t \in \mathbb{R} .$$

Notice that $A_\gamma(0) = (S - S_0)|_{W_{h,l}}$, then $m_\gamma(0) = \operatorname{tr}((S - S_0)|_{W_{h,l}}^2) = c_r^2 + c_i^2$. By Eq. (31) below, $m_\gamma'(0) = 0$ for $\gamma \in \mathbb{R}\setminus\{0\}$. Then, the goal is to find some $\gamma \in \mathbb{R}\setminus\{0\}$ such that $m_\gamma''(0) < 0$. In this case we obtain the strict majorization $\lambda(A_\gamma(t)) \prec (c_r, c_i)$ for $t \in (0, \varepsilon)$, for some $\varepsilon > 0$. This last fact implies that $\lambda(S - S_\gamma(t)) \prec \lambda(S - S_0)$ strictly, for $t \in (0, \varepsilon)$, as desired.

Start computing the derivatives of the entries $a_{ij}(t)$ of $A_\gamma(t)$, for $1 \le i, j \le 2$:

$$\begin{aligned}
a_{11}'(t) &= a_h \sin(2\,t) - a_l\gamma \sin(2\,\gamma t) && \Rightarrow a_{11}'(0) = 0, \\
a_{12}'(t) &= -a_h\cos(2\,t) - a_l\gamma\cos(2\,\gamma t) = a_{21}'(0) && \Rightarrow a_{12}'(0) = -a_h - a_l\gamma, \\
a_{22}'(t) &= a_l\gamma\sin(2\,\gamma t) - a_h\sin(2\,t) && \Rightarrow a_{22}'(0) = 0, \\
a_{11}''(t) &= 2\,a_h\cos(2\,t) - 2\,a_l\gamma^2\cos(2\,\gamma t)) && \Rightarrow a_{11}''(0) = 2\,(a_h - a_l\gamma^2), \\
a_{12}''(t) &= 2\,a_h\sin(2\,t) + 2\,a_l\,\gamma^2\sin(2\,\gamma t) && \Rightarrow a_{12}''(0) = 0, \\
a_{22}''(t) &= 2\,a_l\gamma^2\cos(2\,\gamma t) - 2a_h\cos(2\,t) && \Rightarrow a_{22}''(0) = 2\,(a_l\gamma^2 - a_h).
\end{aligned}$$

Then

$$\begin{aligned}
m_\gamma'(0) &= 2\,a_{11}(0)\,a_{11}'(0) + 4\,a_{12}(0)\,a_{12}'(0) + 2\,a_{22}(0)\,a_{22}'(0) = 0\,, \qquad (31) \\
m_\gamma''(0) &= 2\,a_{11}(0)\,a_{11}''(0) + 4\,(a_{12}'(0))^2 + 2\,a_{22}(0)\,a_{22}''(0) \\
&= 4\,c_i(a_h - a_l\,\gamma^2) + 4(a_h + a_l\gamma)^2 + 4\,c_r(a_l\,\gamma^2 - a_h).
\end{aligned}$$

Note that $m''_\gamma(0)$ is a quadratic function depending on γ whose discriminant is

$$a_h^2\, a_l^2\, [a_h\, a_l - (a_l + c_r - c_i)(a_h + c_i - c_r)] > 0,$$

because we assume that $a_h \le a_l$ (and we have that $c_r > c_i$),

$$(a_l + (c_r - c_i))(a_h - (c_r - c_i)) = a_l\, a_h + (c_r - c_i)(a_h - a_l) - (c_r - c_i)^2 < a_l\, a_h.$$

Then, there exists $\gamma \in \mathbb{R} \setminus \{0\}$ such that $m''_\gamma(0) < 0$. □

The following result together with Proposition 5.1 will allow us to obtain a proof of Proposition 3.5 (see below).

Proposition 5.2. *Let $S \in \mathcal{M}_d(\mathbb{C})^+$ and let $\mathcal{G}_0 \in \mathbb{T}_d(\mathbf{a})$ be as in Notation 3.4 and assume that $p > 1$. Assume that there exist*

$$i \in K_e \quad \text{and} \quad j \in K_r \quad \text{with} \quad e < r \quad \text{such that} \quad j < i\,. \tag{32}$$

In this case, there exists a continuous curve $\mathcal{G}(t) : [0,1) \to \mathbb{T}_d(\mathbf{a})$ such that $\mathcal{G}(0) = \mathcal{G}_0$ and such that $\lambda(S - S_{\mathcal{G}(t)}) \prec \lambda(S - S_0)$ with strict majorization for $t \in (0, \varepsilon)$, for some $\varepsilon > 0$.

Proof. With the notation of the statement and Notation 3.4, notice that

$$\mu_i \le \mu_j \quad \text{and} \quad c_e = \lambda_i - \mu_i < c_r = \lambda_j - \mu_j\ .$$

As in Notation 3.4, consider $\mathcal{B} = \{v_l\}_{l \in \mathbb{I}_d}$ an ONB of $\mathbb{C}^d$ such that

$$S = \sum_{\ell \in \mathbb{I}_d} \lambda_\ell\ v_\ell \otimes v_\ell \quad \text{and} \quad S_0 = \sum_{\ell \in \mathbb{I}_d} \mu_\ell\ v_\ell \otimes v_\ell\ . \tag{33}$$

For $t \in [0,1)$ we let

$$g_l(t) = g_l + \big((1-t^2)^{1/2} - 1 \big)\ \langle g_l, v_i \rangle\ v_i + t\ \langle g_l, v_i \rangle\ v_j \quad \text{for} \quad l \in \mathbb{I}_k\,. \tag{34}$$

Notice that, if $l \in J_e$, then $(S - S_0)\, g_l = c_e\, g_l \implies \langle g_l\,,\, v_j \rangle = 0$. Similarly, if $l \in \mathbb{I}_k \setminus J_e$ then $\langle g_l\,,\, v_i \rangle = 0$ (so that $g_l(t) = g_l$). Therefore the sequence $\mathcal{G}(t) = \{g_l(t)\}_{l \in \mathbb{I}_k} \in \mathbb{T}_d(\mathbf{a})$ for $t \in [0,1)$. Let $P_i = v_i \otimes v_i$ and $P_{ji} = v_j \otimes v_i$ (so that $P_{ji}\, x = \langle x\,,\, v_i \rangle\ v_j$). Then, for every $t \in [0\,,\,1)$,

$$g_l(t) = \big(I + ((1-t^2)^{1/2} - 1)\ P_i + t\ P_{ji} \big)\ g_l \quad \text{for every} \quad l \in \mathbb{I}_k\ .$$

That is, if $V(t) = I + ((1-t^2)^{1/2} - 1)\ P_i + t\ P_{ji} \in \mathcal{M}_d(\mathbb{C})$ then $g_l(t) = V(t)\ g_l$ for every $l \in \mathbb{I}_k$ and $t \in [0,1)$. Therefore, we get that

$$\mathcal{G}(t) = V(t)\, \mathcal{G} = \{V(t)\, g_l\}_{l \in \mathbb{I}_k} \implies S_{\mathcal{G}(t)} = V(t)\, S_{\mathcal{G}}\, V(t)^* \quad \text{for} \quad t \in [0,1)\ .$$

Hence, we obtain the representation

$$S_{\mathcal{G}(t)} = \sum_{\ell \in \mathbb{I}_d \setminus \{i,\,j\}} \mu_\ell\ v_\ell \otimes v_\ell + \gamma_{11}(t)\ v_j \otimes v_j + \gamma_{12}(t)\ v_j \otimes v_i + \gamma_{21}(t)\ v_i \otimes v_j + \gamma_{22}(t)\ v_i \otimes v_i\ ,$$

where the functions $\gamma_{rs}(t)$ are the entries of $A(t) = \big(\gamma_{rs}(t) \big)_{r\,,\,s=1}^2 \in \mathcal{H}(2)$ defined by

$$A(t) = \begin{pmatrix} 1 & t \\ 0 & (1-t^2)^{1/2} \end{pmatrix} \begin{pmatrix} \mu_j & 0 \\ 0 & \mu_i \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t & (1-t^2)^{1/2} \end{pmatrix} \quad \text{for every} \quad t \in [0\,,\,1)\ .$$

It is straightforward to check that $\operatorname{tr}(A(t)) = \mu_j + \mu_i$ and that $\det(A(t)) = (1-t^2)\,\mu_j\,\mu_i$. These facts imply that if we consider the continuous function $L(t) = \lambda_{\max}(A(t))$ then $L(0) = \mu_j$ and $L(t)$ is strictly increasing in $[0,1)$. More straightforward computations show that we can consider continuous curves $x_i(t) : [0,1) \to \mathbb{C}^2$ which satisfy that $\{x_1(t),\, x_2(t)\}$ is ONB of $\mathbb{C}^2$ such that

$$A(t)\,x_1(t) = L(t)\,x_1(t) \quad \text{for} \quad t\in[0,1) \quad \text{and} \quad x_1(0) = e_1\ ,\ \ x_2(0) = e_2\,.$$

For $t \in [0,1)$ we let $X(t) = (u_{r,s}(t))_{r,s=1}^2 \in \mathcal{U}(2)$ with columns $x_1(t)$ and $x_2(t)$. By construction, $X(t) = [0,1) \to \mathcal{U}(2)$ is a continuous curve such that $X(0) = I_2$ and such that

$$X(t)^*\,A(t)\,X(t) = \begin{pmatrix} L(t) & 0 \\ 0 & \mu_i + \mu_j - L(t) \end{pmatrix}.$$

Finally, consider the continuous curve $U(t) : [0,1) \to \mathcal{U}(d)$ given by

$$U(t) = u_{11}(t)\,v_j \otimes v_j + u_{12}(t)\,v_j \otimes v_i + u_{21}(t)\,v_i \otimes v_j + u_{22}(t)\,v_i \otimes v_i + \sum_{\ell\in\mathbb{I}_d\setminus\{i,j\}} v_l \otimes v_l\ .$$

Notice that $U(0) = I$; also, let $\tilde{\mathcal{G}}(t) = U(t)^*\,\mathcal{G}(t) \in \mathbb{T}_d(\mathbf{a})$ for $t \in [0,1)$, which is a continuous curve such that $\tilde{\mathcal{G}}(0) = \mathcal{G}_0$. In this case, for $t\in[0,1)$ we have that

$$S_{\tilde{\mathcal{G}}(t)} = U(t)^*\,S_{\mathcal{G}(t)}\,U(t) = L(t)\,v_j\otimes v_j + (\mu_i + \mu_j - L(t))\,v_i \otimes v_i + \sum_{\ell\in\mathbb{I}_d\setminus\{i,j\}} \mu_\ell\ v_\ell \otimes v_\ell\ .$$

In other words, $U(t)$ is constructed in such a way that $\mathcal{B} = \{v_l\}_{i\in\mathbb{I}_d}$ consists of eigenvectors of $S_{\tilde{\mathcal{G}}(t)}$ for every $t\in[0,1)$. Hence, if $E(t) = L(t) - \mu_j \geq 0$ for $t \in [0\,,\,1)$, we get that

$$S - S_{\tilde{\mathcal{G}}(t)} = (c_r - E(t)\,)\,v_j\otimes v_j + (c_e + E(t)\,)\,v_i\otimes v_i + \sum_{\ell\in\mathbb{I}_d\setminus\{i,j\}} (\lambda_\ell - \mu_\ell)\ v_\ell\otimes v_\ell\ .$$

Let $\varepsilon > 0$ be such that $E(t) = L(t) - \mu_j \leq \frac{c_r - c_e}{2}$ for $t \in [0,\varepsilon]$. (recall that $L(0) = \mu_j$ and that $c_e < c_r$). Since $L(t)$ (and hence $E(t)$) is strictly increasing in $[0,1)$, we see that

$$(c_r - E(t)\,,\, c_e + E(t)) \prec (c_r\,,\, c_e) \implies \lambda(S - S_{\tilde{\mathcal{G}}(t)}) \prec \lambda(S - S_0) \quad \text{for} \quad t\in(0\,,\,\varepsilon]\ ,$$

where the majorization relations above are strict. □

Proof of Proposition 3.5. Fix $S \in \mathcal{M}_d(\mathbb{C})^+$, $\mathbf{a} = (a_i)_{i\in\mathbb{I}_k} \in (\mathbb{R}_{>0}^k)^\downarrow$ and a strictly convex u.i.n. N on $\mathcal{M}_d(\mathbb{C})$. Consider $\mathcal{G}_0$ a local minimizer of $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$ in $\mathbb{T}_d(\mathbf{a})$. Then, $\mathcal{G}_0$ satisfies the assumptions in Notation 3.4; with this notation, assume that $p > 1$. Then, we show that there exist $0 = s_0 < s_1 < \ldots < s_{p-1} < s_p = \operatorname{rk} S_0 \leq d$ such that

$$\begin{aligned} K_j &= J_j = \{s_{j-1}+1\,,\,\ldots\,,\,s_j\}\ , \qquad \text{for} \quad j\in\mathbb{I}_{p-1}\ , \\ K_p &= \{s_{p-1}+1\,,\,\ldots\,,\,s_p\}\ \ ,\ \ J_p = \{s_{p-1}+1\,,\,\ldots\,,\,k\}\ . \end{aligned} \tag{35}$$

Indeed, in case the sets J_j for $j\in\mathbb{I}_p$ do not have the structure described above (i.e. increasing sets formed by consecutive indexes) then, we get that there exist indexes $i,\, r\in\mathbb{I}_p$ and $h,\, l\in\mathbb{I}_k$ for which Eq. (30) holds. In this case, Proposition 5.1 shows that there exists a continuous curve $\mathcal{G}(t) : [0,1) \to \mathbb{T}_d(\mathbf{a})$ such that $\mathcal{G}(0) = \mathcal{G}_0$ and such that $\lambda(S - S_{\mathcal{G}(t)}) \prec \lambda(S - S_0)$ with strict majorization for $t\in(0,\varepsilon)$ for some $\varepsilon > 0$. Since N is a strictly convex u.i.n. we conclude that

$$\Theta_{(N\,,\,S\,,\,\mathbf{a})}(\tilde{\mathcal{G}}(t)) = N(S - S_{\tilde{\mathcal{G}}(t)}) < N(S - S_0) = \Theta_{(N\,,\,S\,,\,\mathbf{a})}(\mathcal{G}_0) \quad \text{for} \quad t\in(0\,,\,\varepsilon]\ . \tag{36}$$

This last fact contradicts the local minimality of $\mathcal{G}_0$. Hence, there exist indexes $s_0 = 0 < s_1 < \ldots < s_{p-1} < s_p \leq d$ for which the representation of the sets J_j for $j \in \mathbb{I}_p$ as in Eq. (35) holds.

Similarly, in case K_j for $j \in \mathbb{I}_p$ are not increasing sets formed by consecutive indexes then, using Proposition 5.2, we also get that $\mathcal{G}_0$ is not a local minimizer; this last fact contradicts the hypothesis on $\mathcal{G}_0$. Finally, notice that by Theorem 3.2 we have that the family $\{g_i\}_{i\in J_j}$ is linearly independent for every $j \in \mathbb{I}_{p-1}$. In particular, by Eq. (29), we get that $\dim(W_j) = |K_j| = |J_j| \quad \text{for} \quad j \in \mathbb{I}_{p-1}$. Hence, we get that $J_j = K_j$ for $j \in \mathbb{I}_{p-1}$ and that $K_p = \{s_{p-1}+1\,,\,\ldots\,,\,s_p\}$ and the result follows. □

In what follows, we show Theorem 3.8. First, we consider a preliminary result.

Proposition 5.3. *Consider Notation 3.7 and 3.4, and assume that $p > 1$. Assume further that the sets J_j and K_j, for $j \in \mathbb{I}_p$, satisfy Eq. (35) above. Then,*

1. *We have that $(a_i)_{i\in J_j} \prec (\lambda_i - c_j)_{i\in K_j}$, for $j \in \mathbb{I}_p$.*
2. *If $0 \leq r < s \leq d$ then, $(a_j)_{j=r+1}^s \prec (\lambda_j - P_{r+1,s})_{j=r+1}^s$ if and only if*

$$P_{r+1,\,s} \leq P_{r,\,i}\ ,\quad r+1 \leq i \leq s \iff P_{r+1,\,s} = \min\{P_{r+1,\,i}:\ r+1 \leq i \leq s\}\,.$$

Proof. For each $j \in \mathbb{I}_p$, consider $W_j = \operatorname{span}\{g_i :\ i \in J_j\} = R(S_{\mathcal{G}_j})$, so that $\dim W_j = |K_j|$ and let Q_j be the orthogonal projection onto W_j; then, W_j reduces both S, S_0 and notice that $(S - S_0)\,Q_j = c_j\,Q_j$ and $S_0\,Q_j = S_{\mathcal{G}_j}$. Then,

$$S\,Q_j = (S - S_0)Q_j + S_0\,Q_j = c_j\,Q_j + S_{\mathcal{G}_j} \implies \lambda(S_{\mathcal{G}_j}) = ((\lambda_i - c_j)_{i\in K_j},\, 0_{d-|K_j|}) \in (\mathbb{R}^d)^\downarrow\,.$$

Hence, by the Schur-Horn theorem we get that $(a_i)_{i\in J_j} \prec \lambda(S_{\mathcal{G}_j})$ which is equivalent to the majorization relation $(a_i)_{i\in\mathbb{I}_j} \prec (\lambda_i - c_j)_{i\in K_j}$, and item 1 follows.

Let $0 \leq r < s \leq d$ and notice that by construction $(a_j)_{j=r+1}^s$, $(\lambda_j - P_{r+1\,,\,s})_{j=r+1}^s \in (\mathbb{R}^{s-r})^\downarrow$. On the other hand, if $r+1 \leq i \leq s$ then

$$\sum_{j=r+1}^{i} a_j \leq \sum_{j=r+1}^{i} \lambda_j - P_{r+1\,,\,s} \iff (i-r)P_{r+1\,,\,s} \leq \sum_{j=r+1}^{i} h_i \iff P_{r+1\,,\,s} \leq P_{r+1\,,\,i}\,.$$

This last fact shows item 2. □

Proof of Theorem 3.8. In case $\mathcal{G}_0$ is a local minimizer of $\Theta_{(N\,,\,S\,,\,\mathbf{a})}$ on $\mathbb{T}_d(\mathbf{a})$ for a strictly convex u.i.n., then the previous results imply that the sets J_j and K_j associated with $\mathcal{G}_0$ satisfy Eq. (35). Hence, we show that the following relations hold:

1. The index $s_1 = \max\big\{j \leq s_{p-1}\,:\, P_{1\,,\,j} = \min_{i\leq s_{p-1}} P_{1\,,\,i}\big\}$, and $c_1 = P_{1\,,\,s_1}$.
2. Recursively, if $s_j < s_{p-1}$, then

$$s_{j+1} = \max\big\{s_j < r \leq s_{p-1}\,:\, P_{s_j+1\,,\,r} = \min_{s_j<i\leq s_{p-1}} P_{s_j+1\,,\,i}\big\} \quad \text{and} \quad c_{j+1} = P_{s_j+1\,,\,s_{j+1}}\,.$$

Indeed, consider an arbitrary $0 \leq j \leq p-2$. By item 1. in Proposition 5.3 and the fact that $J_{j+1} = K_{j+1} = \{s_j + 1, \ldots, s_{j+1}\}$ then we see that

$$(a_i)_{i\in J_{j+1}} \prec (\lambda_i - c_{j+1})_{i\in K_{j+1}} \implies c_{j+1} = P_{s_j+1\,,\,s_{j+1}}\,. \tag{37}$$

Now, using the majorization relation in Eq. (37) and item 2 in Proposition 5.3 we also get that

$$P_{s_j+1\,,\,s_{j+1}} = \min\{P_{s_j+1\,,\,i} : s_j < i \leq s_j\}\,.$$

Therefore, in case the relations between the indexes $s_0 = 0 < \ldots < s_{p-1}$ and the constants $c_1 < \ldots < c_{p-1}$ in the statement do not hold, we get that there exists $0 \le j \le p-2$ such that

$$s_{j+1} < \max\left\{ s_j < r \le s_{p-1} \,:\, P_{s_j+1\,,\,r} = \min_{s_j < i \le s_{p-1}} P_{s_j+1\,,\,i} \right\} = t \le s_{p-1}\,.$$

By definition of t we get that

$$c_{j+1} = P_{s_j+1\,,\,s_{j+1}} \ge P_{s_j+1\,,\,t}\,. \tag{38}$$

Also, there exists $j+1 \le \ell \le p-2$ such that $s_\ell < t \le s_{\ell+1}$. Using the majorization relation in Eq. (37) we see that for $j \le r \le \ell - 1$:

$$(s_{r+1} - s_r)\, c_{r+1} = \sum_{i=s_r+1}^{s_{r+1}} h_i \quad \text{and} \quad (t - s_\ell)\, c_{\ell+1} \le \sum_{i=s_\ell+1}^{t} h_i\,.$$

Then, the previous inequalities allow us to bound

$$P_{s_j+1\,,\,t} = \frac{1}{t-s_j} \sum_{i=s_j+1}^{t} h_i \ge \sum_{r=j}^{\ell-1} \frac{s_{r+1}-s_r}{t}\, c_{r+1} + \frac{t-s_\ell}{t}\, c_{\ell+1} =: \beta$$

that represents the lower bound β as a convex combination of the constants $c_{j+1} < \ldots < c_{\ell+1}$. This last fact clearly implies that $P_{s_j+1\,,\,t} \ge \beta > c_{j+1}$, that contradicts Eq. (38). □

References

[1] J. Antezana, P. Massey, M. Ruiz and D. Stojanoff, The Schur-Horn theorem for operators and frames with prescribed norms and frame operator, Illinois J. Math., 51 (2007), 537-560.

[2] J.J. Benedetto, M. Fickus, Finite normalized tight frames, Adv. Comput. Math. 18, No. 2-4 (2003), 357-385 .

[3] R. Bhatia, Matrix Analysis, Berlin-Heildelberg-New York, Springer 1997.

[4] B.G. Bodmann, P.G: Casazza, The road to equal-norm Parseval frames. J. Funct. Anal. 258 (2010), no. 2, 397-420.

[5] J. Cahill, P.G. Casazza, The Paulsen problem in operator theory. Oper. Matrices 7 (2013), no. 1, 117-130.

[6] J. Cahill, M. Fickus, D.G. Mixon, Auto-tuning unit norm frames. Appl. Comput. Harmon. Anal. 32 (2012), no. 1, 1-15.

[7] P.G. Casazza, The art of frame theory, Taiwanese J. Math. 4 (2000), no. 2, 129-201.

[8] P. G. Casazza and G. Kutyniok eds., Finite Frames: Theory and Applications. Birkhauser, 2012. xii + 483 pp.

[9] O. Christensen, An introduction to frames and Riesz bases. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 2003. xxii+440 pp.

[10] G. Corach, P. Massey and M. Ruiz, Procrustes problems and Parseval quasi-dual frames, Acta Applicandae Mathematicae 131 (2014), 179–195

[11] L. Eldén, H. Park, A Procrustes problem on the Stiefel manifold.Numer. Math. 82, no. 4, 599-619 (1999)

[12] J. C. Gower, G. B. Dijksterhuis, Procrustes problems. Oxford Statistical Science Series, 30. Oxford University Press, Oxford, (2004)

[13] N.J. Higham, Matrix nearness problems and applications. Applications of matrix theory (Bradford, 1988), 1-27, Inst. Math. Appl. Conf. Ser. New Ser., 22, Oxford Univ. Press, New York, (1989)

[14] T. Kato, Perturbation Theory for Linear Operators, Springer-Verlag, Berlin Heidelberg New York 1980.

[15] U. Kintzel, Procrustes problems in finite dimensional indefinite scalar product spaces. Linear Algebra Appl. 402, 1-28 (2005)

[16] C.K. Li, Y.T. Poon, T. Schulte-Herbrüggen, Least-squares approximation by elements from matrix orbits achieved by gradient flows on compact Lie groups. Math. Comp. 80 (2011), no. 275, 1601-1621.

[17] P. Massey, N. B. Rios, D. Stojanoff; Frame completions with prescribed norms: local minimizers and applications. Adv. in Comput. Math. 44 (2018), no. 1, 54-86.

[18] P. Massey, N. B. Rios, D. Stojanoff; Local Liskii's theorems for unitarily invariant norms. Linear Algebra Appl. 557 (2018), 34-61.

[19] P. Massey, M.A. Ruiz, Tight frame completions with prescribed norms. Sampl. Theory Signal Image Process. 7 (2008), no. 1, 1-13.

[20] P. Massey, M. Ruiz; Minimization of convex functionals over frame operators. Adv. Comput. Math. 32 (2010), no. 2, 131-153.

[21] P.G. Massey, M.A. Ruiz, D. Stojanoff; Optimal frame completions. Advances in Computational Mathematics 40 (2014), 1011-1042.

[22] N. Strawn; Optimization over finite frame varieties and structured dictionary design, Appl. Comput. Harmon. Anal. 32 (2012) 413-434.

[23] N. Strawn; Finite frame varieties: nonsingular points, tangent spaces, and explicit local parameterizations. J. Fourier Anal. Appl. 17 (2011), no. 5, 821-853.

[24] G. A. Watson, The solution of orthogonal Procrustes problems for a family of orthogonally invariant norms. Adv. Comput. Math. 2 , no. 4, 393-405 (1994)