Non-negative Matrix Factorization via Archetypal Analysis

Hamid Javadi* and Andrea Montanari†

May 8, 2017

Abstract

Given a collection of data points, non-negative matrix factorization (NMF) suggests to express them as convex combinations of a small set of 'archetypes' with non-negative entries. This decomposition is unique only if the true archetypes are non-negative and sufficiently sparse (or the weights are sufficiently sparse), a regime that is captured by the separability condition and its generalizations.

In this paper, we study an approach to NMF that can be traced back to the work of Cutler and Breiman [CB94] and does not require the data to be separable, while providing a generally unique decomposition. We optimize the trade-off between two objectives: we minimize the distance of the data points from the convex envelope of the archetypes (which can be interpreted as an empirical risk), while minimizing the distance of the archetypes from the convex envelope of the data (which can be interpreted as a data-dependent regularization). The archetypal analysis method of [CB94] is recovered as the limiting case in which the last term is given infinite weight.

We introduce a 'uniqueness condition' on the data which is necessary for exactly recovering the archetypes from noiseless data. We prove that, under uniqueness (plus additional regularity conditions on the geometry of the archetypes), our estimator is robust. While our approach requires solving a non-convex optimization problem, we find that standard optimization methods succeed in finding good solutions both for real and synthetic data.

1 Introduction

Given a set of data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \in \mathbb{R}^d$, it is often useful to represent them as convex combinations of a small set of vectors (the 'archetypes' $\boldsymbol{h}_1, \dots, \boldsymbol{h}_\ell$):

$$\boldsymbol{x}_i \approx \sum_{\ell=1}^{r} w_{i,\ell} \boldsymbol{h}_\ell \,, \quad w_{i,\ell} \geq 0, \ \sum_{\ell=1}^{r} w_{i,\ell} = 1 \,. \tag{1.1}$$

Decompositions of this type have wide ranging applications, from chemometrics [PT94] to image processing [LS99] and topic modeling [XLG03]. As an example, Figure 1 displays the infrared reflection spectra[1] of four molecules (caffeine, sucrose, lactose and trioctanoin) for wavenumber between 1186 cm^{-1} and 1530 cm^{-1}. Each spectrum is a vector $\boldsymbol{h}_{0,\ell} \in \mathbb{R}^d$, with $d = 87$ and $\ell \in \{1, \dots, 4\}$. If a mixture of these substances is analyzed, the resulting spectrum will be a convex combination of the spectra of the four analytes. This situation arises in hyperspectral imaging [MBDC+14], where a main focus is to estimate spatially varying proportions of a certain

*Department of Electrical Engineering, Stanford University

†Department of Electrical Engineering and Statistics, Stanford University

[1]Data were retrieved from the NIST Chemistry WebBook dataset [LM].

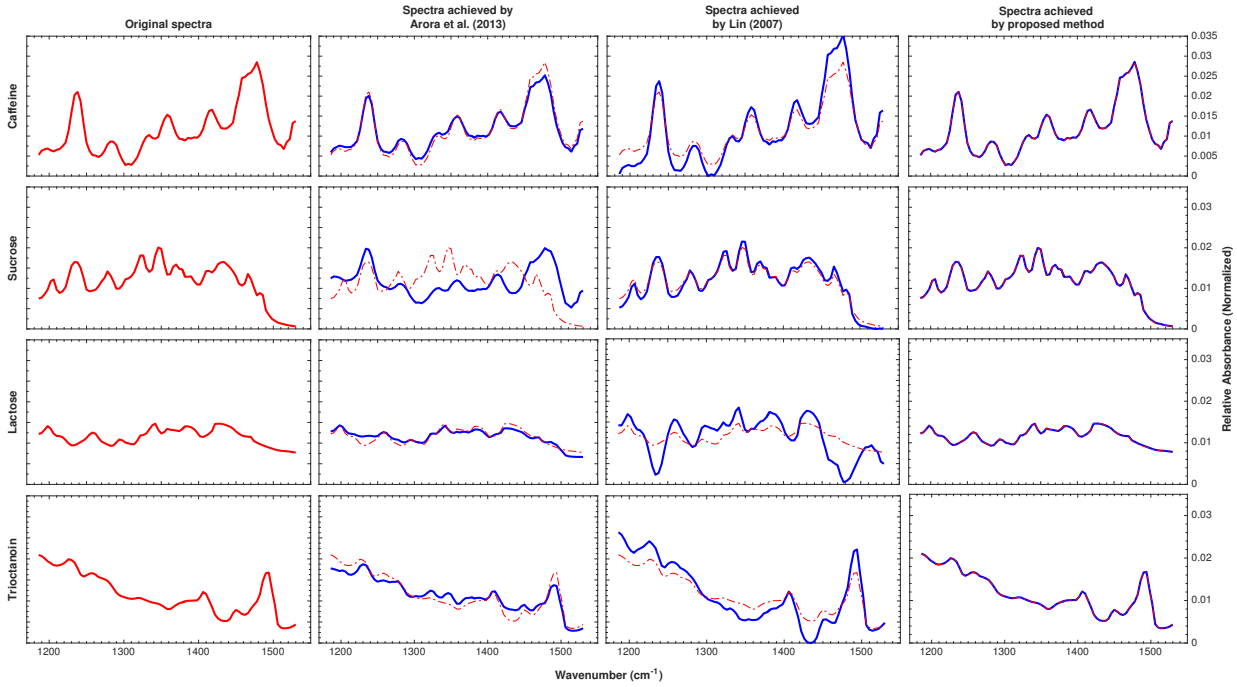

Figure 1: Left column: Infrared reflection spectra of four molecules. Subsequent columns: Spectra estimated from $n = 250$ spectra of mixtures of the four original substances (synthetic data generated by taking random convex combinations of the pure spectra, see Appendix A for details). Each column reports the results obtained with a different estimator: continuous blue lines correspond to the reconstructed spectra; dashed red lines correspond to the ground truth.

number of analytes. In order to mimic this setting, we generated $n = 250$ synthetic random convex combinations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ of the four spectra $\boldsymbol{h}_{0,1}$, $\dots$, $\boldsymbol{h}_{0,4}$, each containing two or more of these four analytes, and tried to reconstruct the pure spectra from the $\boldsymbol{x}_i$'s. Each column in Figure 1 displays the reconstruction obtained using a different procedure. We refer to Appendix A for further details.

Without further constraints, the decomposition (1.1) is dramatically underdetermined. Given a set of valid archetypes $\{\boldsymbol{h}_{0,\ell}\}_{\ell \le r}$, any set $\{\boldsymbol{h}_\ell\}_{\ell \le r}$ whose convex hull contains the $\{\boldsymbol{h}_{0,\ell}\}_{\ell \le r}$ also satisfies Eq. (1.1). For instance, we can set $\boldsymbol{h}_\ell = \boldsymbol{h}_{0,\ell}$ for $\ell \le r-1$, and $\boldsymbol{h}_r = (1+s)\boldsymbol{h}_{0,r} - s\boldsymbol{h}_{0,1}$ for any $s \ge 0$, and obtain an equally good representation of the data $\{\boldsymbol{x}_i\}_{i \le n}$.

How should we constrain the decomposition (1.1) in such a way that it is generally unique (up to permutations of the r archetypes)? Since the seminal work of Paatero and Tapper [PT94, Paa97], and of Lee and Seung [LS99, LS01], an overwhelming amount of work has addressed this question by making the assumptions that the archetypes are componentwise non-negative $\boldsymbol{h}_\ell \ge 0$. Among other applications the non-negativity constraint is justified for chemometrics (reflection or absorption spectra are non-negative), and topic modeling (in this case archetypes correspond to topics, which are represented as probability distributions over words). This formulation has become popular as non-negative matrix factorization (NMF).

Under the non-negativity constraint $\boldsymbol{h}_\ell \ge 0$ the role of weights and archetypes becomes symmetric, and the decomposition (1.1) is unique provided that the archetypes or the weights are sufficiently sparse (without loss of generality one can assume $\sum_{\ell=1}^r h_{\ell,i} = 1$). This point was clar-

ified by Donoho and Stodden [DS03], introduced a separability condition that ensure uniqueness. The non-negative archetypes $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r$ are separable if, for each $\ell \in [r]$ there exists an index $i(\ell) \in [d]$ such that $(\boldsymbol{h}_\ell)_{i(\ell)} = 1$, and $(\boldsymbol{h}_{\ell'})_{i(\ell)} = 0$ for all $\ell' \neq \ell$. If we exchange the roles of weights $\{w_{i,\ell}\}$ and archetypes $\{h_{\ell,i}\}$, separability requires that $\ell \in [r]$ there exists an index $i(\ell) \in [n]$ such that $w_{i(\ell),\ell} = 1$, and $w_{i(\ell),\ell'} = 0$ for all $\ell' \neq \ell$ This condition has a simple geometric interpretation: the data are separable if for each archetype $\boldsymbol{h}_\ell$ there is at least one data point $\boldsymbol{x}_i$ such that $\boldsymbol{x}_i = \boldsymbol{h}_\ell$. A copious literature has developed algorithms for non-negative matrix factorization under separability condition or its generalizations [DS03, AGKM12, RRTB12, AGH$^+$13, GZ15].

Of course this line of work has a drawback: in practice we do not know whether the data are separable. (We refer to the Section 5 for a comparison with [GZ15], which relaxes the separability assumption.) Further, there are many cases in which the archetypes $\boldsymbol{h}_1, \dots, \boldsymbol{h}_\ell$ are not necessarily non-negative. For instance, in spike sorting, the data are measurements of neural activity ad the archetypes correspond to waveforms associated to different neurons [RCP09]. In other applications the archetypes $\boldsymbol{h}_\ell$ are non-negative, but –in order to reduce complexity– the data $\{\boldsymbol{x}_i\}_{i\le n}$ are replaced by a random low-dimensional projection [KSD08, WL10]. The projected archetypes loose the non-negativity property. Finally, the decomposition (1.1) is generally non-unique, even under the constraint $\boldsymbol{h}_\ell \ge 0$. This is illustrated, again, in Figure 1: all the spectra are strictly positive, and hence we can find archetypes $\boldsymbol{h}_1, \dots, \boldsymbol{h}_4$ that are still non-negative and whose convex envelope contains $\boldsymbol{h}_{0,1}, \dots, \boldsymbol{h}_{0,4}$.

Since NMF is underdetermined, standard methods fail in such applications, as illustrated in Figure 1 . We represent the data as a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ whose i-th row is the vector $\boldsymbol{x}_i$, the weights by a matrix $\boldsymbol{W} = (w_{i,\ell})_{i\le n,\ell\le r} \in \mathbb{R}^{n\times d}$ and the prototypes by a matrix $\boldsymbol{H} = (h_{\ell,j})_{\ell\le r,j\le d} \in \mathbb{R}^{r\times d}$. The third column of Figure 1 uses a projected gradient algorithm from [Lin07] to solve the problem

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 , \\ \text{subject to} \quad & \boldsymbol{W} \ge 0, \ \ \boldsymbol{H} \ge 0 . \end{aligned} \tag{1.2}$$

Empirically, projected gradient converges to a point with very small fitting error $\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2$, but the reconstructed spectra (rows of $\boldsymbol{H}$) are inaccurate. The second column in the same figure shows the spectra reconstructed using an algorithm from [AGH$^+$13], that assumes separability: as expected, the reconstruction is not accurate.

In a less widely known paper, Cutler and Breiman [CB94] addressed the same problem using what they call 'archetypal analysis.' Archetypal analysis presents two important differences with respect to standard NMF: (1) The archetypes $\boldsymbol{h}_\ell$ are not necessarily required to be non-negative (although this constraint can be easily incorporated); (2) The under-determination of the decomposition (1.1) is addressed by requiring that the archetypes belong to the convex hull of the data points: $\boldsymbol{h}_\ell \in \mathrm{conv}(\{\boldsymbol{x}_i\}_{i\le n})$.

In applications the condition $\boldsymbol{h}_\ell \in \mathrm{conv}(\{\boldsymbol{x}_i\}_{i\le n})$ is too strict. This paper builds on the ideas of [CB94] to propose a formulation of NMF that is uniquely defined (barring degenerate cases) and provides a useful notion of optimality. In particular, we present the following contributions.

Archetypal reconstruction. We propose to reconstruct the archetypes $\boldsymbol{h}_1, \dots, \boldsymbol{h}_r$ by optimizing a combination of two objectives. On one hand, we minimize the error in the decomposition (1.1). This amounts to minimizing the distance between the data points and the convex hull of the archetypes. On the other hand, we minimize the distance of the archetypes from the convex hull of data points. This relaxes the original condition imposed in [CB94] which required the archetypes to lie in $\mathrm{conv}(\{\boldsymbol{x}_i\})$, and allows to treat non-separable data.

Robustness guarantee. We next assume that that the decomposition (1.1) approximately hold for some 'true' archetypes $\boldsymbol{h}_\ell^0$ and weights $w_{i,\ell}^0$, namely $\boldsymbol{x}_i = \boldsymbol{x}_i^0 + \boldsymbol{z}_i$, where $\boldsymbol{x}_i^0 = \sum_{\ell=1}^r w_{i,\ell}^0 \boldsymbol{h}_\ell^0$ and

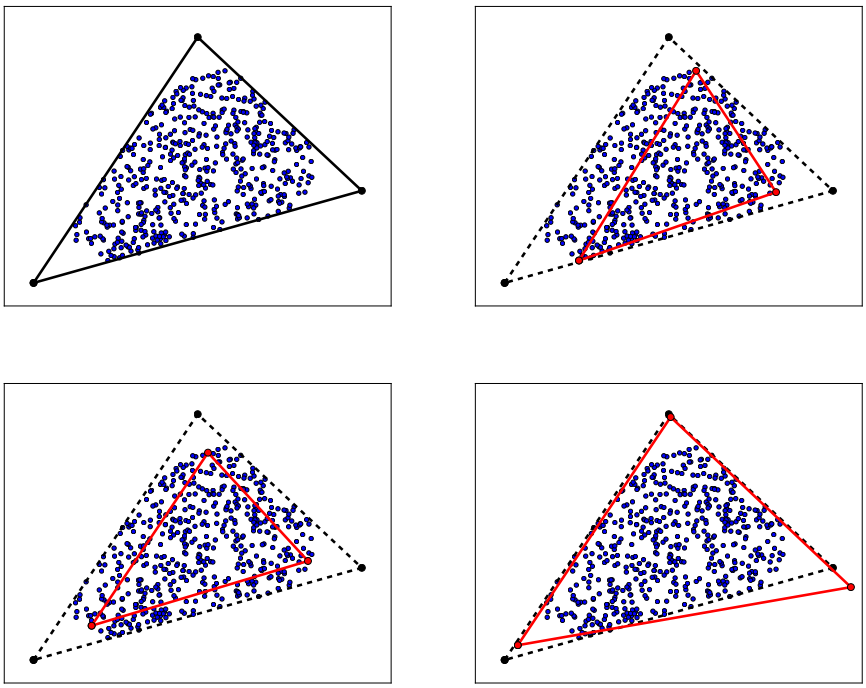

Figure 2: Toy example of archetype reconstruction. Top left: data points (blue) are generated as random linear combinations of $r = 3$ archetypes in $d = 2$ dimensions (red, see Appendix A for details). Top right: Initialization using the algorithm of [AGH$^+$13]. Bottom left: Output of the alternate minimization algorithm of [CB94] with initialization form the previous frame. Bottom right: Alternate minimization algorithm to compute the estimator (2.4), with $\lambda = 0.0166$.

$\boldsymbol{z}_i$ captures unexplained effects. We introduce a 'uniqueness condition' on the data $\{\boldsymbol{x}_i^0\}_{i\le n}$ which is necessary for exactly recovering the archetypes from the noiseless data. We prove that, under uniqueness (plus additional regularity conditions on the geometry of the archetypes), our estimator is robust. Namely it outputs archetypes $\{\hat{\boldsymbol{h}}_\ell\}_{\ell\le r}$ whose distance from the true ones $\{\boldsymbol{h}_\ell^0\}_{\ell\le r}$ (in a suitable metric) is controlled by $\sup_{i\le n}\|\boldsymbol{z}_i\|_2$.

Algorithms. Our approach reconstructs the archetypes $\boldsymbol{h}_1,\dots,\boldsymbol{h}_r$ by minimizing a non-convex risk function $\mathscr{R}_\lambda(\boldsymbol{H})$. We propose three descent algorithms that appear to perform well on realistic instances of the problem. In particular, Section 4 introduces a proximal alternating linearized minimization algorithm (PALM) that is guaranteed to converge to critical points of the risk function. Appendix E discusses two alternative approaches. One possible explanation for the success of such descent algorithms is that reasonably good initializations can be constructed using spectral methods, or approximating the data as separable, cf. Section 4.1. We defer a study of global convergence of this two-stages approach to future work.

2 An archetypal reconstruction approach

Let $\mathcal{Q} \subseteq \mathbb{R}^d$ be a convex set and $D : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$, $(\boldsymbol{x}, \boldsymbol{y}) \mapsto D(\boldsymbol{x}; \boldsymbol{y})$ a loss function on $\mathcal{Q}$. For a point $\boldsymbol{u} \in \mathcal{Q}$, and a matrix $\boldsymbol{V} \in \mathbb{R}^{m \times d}$, with rows $\boldsymbol{v}_1, \dots, \boldsymbol{v}_m \in \mathcal{Q}$, we let

$$\mathscr{D}(\boldsymbol{u}; \boldsymbol{V}) \equiv \min \left\{ D\big(\boldsymbol{u}; \boldsymbol{V}^{\mathsf{T}} \boldsymbol{\pi}\big) \ : \quad \boldsymbol{\pi} \in \Delta^m \right\}, \tag{2.1}$$

$$\Delta^m \equiv \left\{ \boldsymbol{x} \in \mathbb{R}^m_{\ge 0} : \ \ \langle \boldsymbol{x}, \boldsymbol{1} \rangle = 1 \right\}. \tag{2.2}$$

In other words, denoting by $\mathrm{conv}(\boldsymbol{V}) = \mathrm{conv}(\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_m\})$ the convex hull of the rows of matrix $\boldsymbol{V}$, $\mathscr{D}(\boldsymbol{u}; \boldsymbol{V})$ is the minimum loss between $\boldsymbol{x}$ and any point in $\mathrm{conv}(\boldsymbol{V})$. If $\boldsymbol{U} \in \mathbb{R}^{k \times d}$ is a matrix with rows $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k \in \mathcal{Q}$, we generalize this definition by letting

$$\mathscr{D}(\boldsymbol{U}; \boldsymbol{V}) \equiv \sum_{\ell=1}^{k} \mathscr{D}(\boldsymbol{u}_\ell; \boldsymbol{V}). \tag{2.3}$$

While this definition makes sense more generally, we have in mind two specific examples in which $D(\boldsymbol{x}; \boldsymbol{y})$ is actually separately convex in its arguments $\boldsymbol{x}$ and $\boldsymbol{y}$. (Most of our results will concern the first example.)

Example 2.1 (Square loss)**.** In this case $\mathcal{Q} = \mathbb{R}^d$, and $D(\boldsymbol{x}; \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. This is the case originally studied by Cutler and Breiman [CB94].

Example 2.2 (KL divergence)**.** We take $\mathcal{Q} = \Delta^d$, the d-dimensional simplex, and $D(\boldsymbol{x}; \boldsymbol{y})$ to be the Kullback-Leibler divergence between probability distributions $\boldsymbol{x}$ and $\boldsymbol{y}$, namely $D(\boldsymbol{x}; \boldsymbol{y}) \equiv \sum_{i=1}^d x_i \log(x_i/y_i)$.

Given data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ organized in the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, we estimate the archetypes by solving the problem[2]

$$\widehat{\boldsymbol{H}}_\lambda \in \arg\min \left\{ \mathscr{D}(\boldsymbol{X}; \boldsymbol{H}) + \lambda\, \mathscr{D}(\boldsymbol{H}; \boldsymbol{X}) : \ \ \boldsymbol{H} \in \mathcal{Q}^r \right\}, \tag{2.4}$$

where we denote by $\mathcal{Q}^r$ the set of matrices $\boldsymbol{H} \in \mathbb{R}^{r \times d}$ with rows $\boldsymbol{h}_1, \dots, \boldsymbol{h}_r \in \mathcal{Q}$. A few values of λ are of special significance. If we set $\lambda = 0$, and $\mathcal{Q} = \Delta^d$, we recover the standard NMF objective (1.2), with a more general distance function $D(\,\cdot\,, \cdot\,)$. As pointed out above, in general this objective has no unique minimum. If we let $\lambda \to 0+$ after the minimum is evaluated, $\widehat{\boldsymbol{H}}_\lambda$ converges to the minimizer of $\mathscr{D}(\boldsymbol{X}; \boldsymbol{H})$ which is the 'closest' to the convex envelope of the data $\mathrm{conv}(\boldsymbol{X})$ (in the sense of minimizing $\mathscr{D}(\boldsymbol{H}; \boldsymbol{X})$). Finally as $\lambda \to \infty$, the archetypes $\boldsymbol{h}_\ell$ are forced to lie in $\mathrm{conv}(\boldsymbol{X})$ and hence we recover the method of [CB94].

Figure 2 illustrates the advantages of the estimator (2.4) on a small synthetic example, with $d = 2$, $r = 3$, $n = 500$: in this case the data are non separable. We first use the successive projections algorithm of [AGH+13] (that is designed to deal with separable data) in order to estimate the archetypes. As expected, the reconstruction is not accurate because this algorithm assumes separability and hence estimates the archetypes with a subset of the data points. We then use these estimates as initialization in the alternate minimization algorithm of [CB94], which optimizes the objective (2.4) with $\lambda = \infty$. The estimates improve but not substantially: they are still constrained to lie in $\mathrm{conv}(\boldsymbol{X})$. A significant improvement is obtained by setting λ to a small value. We

[2]This problem can have multiple global minima if $\lambda = 0$ or in degenerate settings. One minimizer is selected arbitrarily when this happens.

(approximately) minimize the cost function (2.4) by generalizing the alternate minimization algorithm, cf. Section 4. The optimal archetypes are no longer constrained to $\mathrm{conv}(\boldsymbol{X})$, and provide a better estimate of the true archetypes. The last column in Figure 1 uses the same estimator, and approximately solves problem (2.4) by gradient descent algorithm.

In our analysis we will consider a slightly different formulation in which the Lagrangian of Eq. (2.4) is replaced by a hard constraint:

$$\begin{aligned} \text{minimize} \quad & \mathscr{D}(\boldsymbol{H};\boldsymbol{X})\,, \\ \text{subject to} \quad & \mathscr{D}(\boldsymbol{x}_i;\boldsymbol{H}) \le \delta^2 \quad \text{for all } i \in \{1,\dots,n\}\,. \end{aligned} \tag{2.5}$$

We will use this version in the analysis presented in the next section, and denote the corresponding estimator by $\widehat{\boldsymbol{H}}$.

3 Robustness

In order to analyze the robustness properties of estimator $\widehat{\boldsymbol{H}}$, we assume that there exists an approximate factorization

$$\boldsymbol{X} = \boldsymbol{W}_0\boldsymbol{H}_0 + \boldsymbol{Z}\,, \tag{3.1}$$

where $\boldsymbol{W}_0 \in \mathbb{R}^{n\times r}$ is a matrix of weights (with rows $\boldsymbol{w}_{0,i} \in \Delta^r$), $\boldsymbol{H}_0 \in \mathbb{R}^{r\times d}$ is a matrix of archetypes (with rows $\boldsymbol{h}_{0,\ell}$), and we set $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$. The deviation $\boldsymbol{Z}$ is arbitrary, with rows $\boldsymbol{z}_i$ satisfying $\max_{i\le n}\|\boldsymbol{z}_i\|_2 \le \delta$. We will assume throughout r to be known.

We will quantify estimation error by the sum of distances between the true archetypes and the closest estimated archetypes

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}) \equiv \sum_{\ell=1}^{r} \min_{\ell'\le r} D(\boldsymbol{h}_{0,\ell}, \hat{\boldsymbol{h}}_{\ell'})\,. \tag{3.2}$$

In words, if $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})$ is small, then for each true archetype $\boldsymbol{h}_{0,\ell}$ there exists an estimated archetype $\hat{\boldsymbol{h}}_{\ell'}$ that is close to it in D-loss. Unless two or more of the true archetypes are close to each other, this means that there is a one-to-one correspondence between estimated archetypes and true archetypes, with small errors.

Assumption (Uniqueness)**.** *We say that the factorization $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ satisfies uniqueness with parameter $\alpha > 0$ (equivalently, is α-*unique*) if for all $\boldsymbol{H} \in \mathcal{Q}^r$ with $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\boldsymbol{H})$, we have*

$$\mathscr{D}(\boldsymbol{H},\boldsymbol{X}_0)^{1/2} \ge \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)^{1/2} + \alpha\left\{\mathscr{D}(\boldsymbol{H},\boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{H})^{1/2}\right\}. \tag{3.3}$$

The rationale for this assumption is quite clear. Assume that the data lie in the convex hull of the true archetypes $\boldsymbol{H}_0$, and hence Eq. (3.1) holds without error term $\boldsymbol{Z} = 0$, i.e. $\boldsymbol{X} = \boldsymbol{X}_0$. We reconstruct the archetypes by demanding $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\boldsymbol{H})$: any such $\boldsymbol{H}$ is a plausible explanation of the data. In order to make the problem well specified, we define $\boldsymbol{H}_0$ to be the matrix of archetypes that are the closest to $\boldsymbol{X}_0$, and hence $\mathscr{D}(\boldsymbol{H},\boldsymbol{X}_0) \ge \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)$ for all $\boldsymbol{H}$. In order for the reconstruction to be unique (and hence for the problem to be identifiable) we need to assume $\mathscr{D}(\boldsymbol{H},\boldsymbol{X}_0) > \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)$ strictly for $\boldsymbol{H} \neq \boldsymbol{H}_0$. The uniqueness assumption provides a quantitative version of this condition.

Remark 3.1. Given $\boldsymbol{X}_0$, $\boldsymbol{H}_0$, the best constant α such that Eq. (3.3) holds for all $\boldsymbol{H}$ is a geometric property that depend on $\boldsymbol{X}_0$ only through $\mathrm{conv}(\boldsymbol{X}_0)$. In particular, if $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ is a separable factorization, then it satisfies uniqueness with parameter $\alpha = 1$. Indeed in this case $\mathrm{conv}(\boldsymbol{H}_0) = \mathrm{conv}(\boldsymbol{X}_0)$, whence $\mathscr{D}(\boldsymbol{H}, \boldsymbol{X}_0) = \mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)$ and $\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0) = \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H}) = 0$.

It is further possible to show that $\alpha \in [0, 1]$ for all $\boldsymbol{H}_0, \boldsymbol{X}_0$. Indeed, we took $\boldsymbol{H}_0$ to be the matrix of archetypes that are closest to $\boldsymbol{X}_0$. In other words, $\mathscr{D}(\boldsymbol{H}, \boldsymbol{X}_0) \geq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)$ and hence, $\alpha \geq 0$. In addition, since $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\boldsymbol{H}_0)$, for $\boldsymbol{h}_i$ an arbitrary row of $\boldsymbol{H}$ we have

$$\mathscr{D}(\boldsymbol{h}_i, \boldsymbol{X}_0) \leq \mathscr{D}(\boldsymbol{h}_i, \boldsymbol{H}_0). \tag{3.4}$$

Hence, $\mathscr{D}(\boldsymbol{H}, \boldsymbol{X}_0) \leq \mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)$ and therefore

$$\mathscr{D}(\boldsymbol{H}, \boldsymbol{X}_0)^{1/2} \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + \left\{\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2}\right\}. \tag{3.5}$$

Thus, $\alpha \leq 1$.

We say that the convex hull $\mathrm{conv}(\boldsymbol{X}_0)$ has *internal radius* (at least) μ if it contains an $r-1$-dimensional ball of radius μ, i.e. if there exists $\boldsymbol{z}_0 \in \mathbb{R}^d$, $\boldsymbol{U} \in \mathbb{R}^{d\times(r-1)}$, with $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \mathbf{I}_d$, such that $\boldsymbol{z}_0 + \boldsymbol{U}\mathsf{B}_{r-1}(\mu) \subseteq \mathrm{conv}(\boldsymbol{X}_0)$. We further denote by $\kappa(\boldsymbol{M})$ the condition number of matrix $\boldsymbol{M}$.

Theorem 1. *Assume $\boldsymbol{X} = \boldsymbol{W}_0\boldsymbol{H}_0 + \boldsymbol{Z}$ where the factorization $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ satisfies the uniqueness assumption with parameter $\alpha > 0$, and that $\mathrm{conv}(\boldsymbol{X}_0)$ has internal radius $\mu > 0$. Consider the estimator $\widehat{\boldsymbol{H}}$ defined by Eq. (2.5), with $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ (square loss) and $\delta = \max_{i\leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2$. If*

$$\max_{i\leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \frac{\alpha\mu}{30r^{3/2}}, \tag{3.6}$$

then, we have

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}) \leq \frac{C_*^2\, r^5}{\alpha^2} \max_{i\leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2^2, \tag{3.7}$$

where C_* is a coefficient that depends uniquely on the geometry of $\boldsymbol{H}_0$, $\boldsymbol{X}_0$, namely $C_* = 120(\sigma_{\max}(\boldsymbol{H}_0)/\mu)\cdot \max(1, \kappa(\boldsymbol{H}_0)/\sqrt{r})$.

4 Algorithms

While our main focus is on structural properties of non-negative matrix factorization, we provide evidence that the optimization problem we defined can be solved in practical scenarios. A more detailed study is left to future work.

From a computational point of view, the Lagrangian formulation (2.4) is more appealing. For the sake of simplicity, we denote the regularized risk by

$$\mathscr{R}_\lambda(\boldsymbol{H}) \equiv \mathscr{D}(\boldsymbol{X}; \boldsymbol{H}) + \lambda\, \mathscr{D}(\boldsymbol{H}; \boldsymbol{X}), \tag{4.1}$$

and leave implicit the dependence on the data $\boldsymbol{X}$. Notice that this function is non-convex and indeed has multiple global minima: in particular, permuting the rows of a minimizer $\boldsymbol{H}$ yields other minimizers. We will describe two greedy optimization algorithms: one based on gradient descent, and one on alternating minimization, which generalizes the algorithm of [CB94]. In both cases it is helpful to use a good initialization: two initialization methods are introduced in the next section.

4.1 Initialization

We experimented with two initialization methods, described below.

(1) Spectral initialization. Under the assumption that the archetypes $\{\boldsymbol{h}_{0,\ell}\}_{\ell\le r}$ are linearly independent (and for non-degenerate weights $\boldsymbol{W}$), the 'noiseless' matrix $\boldsymbol{X}_0$ has rank exactly r. This motivates the following approach. We compute the singular value decomposition $\boldsymbol{X} = \sum_{i=1}^{n\wedge d} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}$, $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_{n\wedge d}$, and initialize $\widehat{\boldsymbol{H}}$ as the matrix $\widehat{\boldsymbol{H}}^{(0)}$ with rows $\hat{\boldsymbol{h}}_1^{(0)} = \boldsymbol{v}_1, \dots, \hat{\boldsymbol{h}}_r^{(0)} = \boldsymbol{v}_r$.

(2) Successive projections initialization. We initialize $\widehat{\boldsymbol{H}}^{(0)}$ by choosing archetypes $\{\hat{\boldsymbol{h}}_\ell^{(0)}\}_{1\le\ell\le r}$ that are a subset of the data $\{\boldsymbol{x}_i\}_{1\le i\le n}$, selected as follows. The first archetype $\hat{\boldsymbol{h}}_1^{(0)}$ is the data point which is farthest from the origin. For each subsequent archetype, we choose the point that is farthest from the affine subspace spanned by the previous ones.

ARCHETYPE INITIALIZATION ALGORITHM

Input : Data $\{\boldsymbol{x}_i\}_{i\le n}$, $\boldsymbol{x}_i \in \mathbb{R}^d$; integer r;

Output : Initial archetypes $\{\hat{\boldsymbol{h}}_\ell^{(0)}\}_{1\le\ell\le r}$;

1: Set $i(1) = \arg\max\{D(\boldsymbol{x}_i; \boldsymbol{0}) : i \le n\}$;
2: Set $\hat{\boldsymbol{h}}_1^{(0)} = \boldsymbol{x}_{i(1)}$;
3: For $\ell \in \{1, \dots, r\}$
4: Define $V_\ell \equiv \mathrm{aff}(\hat{\boldsymbol{h}}_1^{(0)}, \hat{\boldsymbol{h}}_2^{(0)}, \dots, \hat{\boldsymbol{h}}_\ell^{(0)})$;
5: Set $i(\ell+1) = \arg\max\{\mathscr{D}(\boldsymbol{x}_i; V_\ell) : i \le n\}$;
6: Set $\hat{\boldsymbol{h}}_{\ell+1}^{(0)} = \boldsymbol{x}_{i(\ell+1)}$;
7: End For;
8: Return $\{\hat{\boldsymbol{h}}_\ell^{(0)}\}_{1\le\ell\le r}$l

---

This coincides with the successive projections algorithm of [ASG+01], with the minor difference that V_ℓ is the affine subspace spanned by the first ℓ vectors, instead of the linear subspace[3] This method can be proved to return the exact archetypes if data are separable the archetypes are affine independent [AGH+13, GV14]. When data are not separable it provides nevertheless a good initial assignment.

4.2 Proximal alternating linearized minimization

The authors of [BST14] develop a proximal alternating linearized minimization algorithm (PALM) to solve the problems of the form

$$\text{minimize} \qquad \Psi(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + g(\boldsymbol{y}) + h(\boldsymbol{x}, \boldsymbol{y}) \tag{4.2}$$

where $f : \mathbb{R}^m \to (-\infty, +\infty]$ and $g : \mathbb{R}^n \to (-\infty, \infty]$ are lower semicontinuous and $h \in C^1(\mathbb{R}^m \times \mathbb{R}^n)$. PALM is guaranteed to converge to critical points of the function Ψ [BST14].

We apply this algorithm to minimize the cost function (4.1), with $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ which we write as

$$\mathscr{R}_\lambda(\boldsymbol{H}) = \min_{\boldsymbol{W}} \Psi(\boldsymbol{H}, \boldsymbol{W}) = f(\boldsymbol{H}) + g(\boldsymbol{W}) + h(\boldsymbol{H}, \boldsymbol{W})\,. \tag{4.3}$$

[3]The same modification is also used in [AGH+13], but we do not apply the full algorithm of this paper.

where,

$$f(\boldsymbol{H}) = \lambda\, \mathscr{D}(\boldsymbol{H}, \boldsymbol{X}), \tag{4.4}$$

$$g(\boldsymbol{W}) = \sum_{i=1}^{n} \mathbf{I}\left(\boldsymbol{w}_i \in \Delta^r\right), \tag{4.5}$$

$$h(\boldsymbol{H}, \boldsymbol{W}) = \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 . \tag{4.6}$$

In above equations $\boldsymbol{w}_i$ are the rows of $\boldsymbol{W}$ and the indicator function $\mathbf{I}(\boldsymbol{x} \in \Delta^r)$ is equal to zero if $\boldsymbol{x} \in \Delta^r$ and is equal to infinity otherwise.

By using this decomposition, the iterations of the PALM iteration reads

$$\widetilde{\boldsymbol{H}}^k = \boldsymbol{H}^k - \frac{1}{\gamma_1^k}(\boldsymbol{W}^k)^{\mathsf{T}}\left(\boldsymbol{W}^k\boldsymbol{H}^k - \boldsymbol{X}\right), \tag{4.7}$$

$$\boldsymbol{H}^{k+1} = \widetilde{\boldsymbol{H}}^k - \frac{\lambda}{\lambda + \gamma_1^k}\left(\widetilde{\boldsymbol{H}}^k - \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}\left(\widetilde{\boldsymbol{H}}^k\right)\right), \tag{4.8}$$

$$\boldsymbol{W}^{k+1} = \boldsymbol{\Pi}_{\Delta^r}\left(\boldsymbol{W}^k - \frac{1}{\gamma_2^k}\left(\boldsymbol{W}^k\boldsymbol{H}^{k+1} - \boldsymbol{X}\right)(\boldsymbol{H}^{k+1})^{\mathsf{T}}\right), \tag{4.9}$$

where γ_1^k, γ_2^k are step sizes and, for $\boldsymbol{M} \in \mathbb{R}^{m_1 \times m_2}$, and $\mathcal{S} \subseteq \mathbb{R}^{m_2}$ a closed convex set, $\boldsymbol{\Pi}_{\mathcal{S}}(\boldsymbol{M})$ is the matrix obtained by projecting the rows of $\boldsymbol{M}$ onto the simplex $\mathcal{S}$.

Proposition 4.1. *Consider the risk (4.1), with loss $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$, and the corresponding cost function $\Psi(\boldsymbol{H}, \boldsymbol{W})$. If the step sizes are chosen such that $\gamma_1^k > \left\|\boldsymbol{W}^{k\mathsf{T}}\boldsymbol{W}^k\right\|_F$, $\gamma_2^k > \max\left\{\left\|\boldsymbol{H}^{k+1}\boldsymbol{H}^{k+1\mathsf{T}}\right\|_F, \varepsilon\right\}$ for some constant $\varepsilon > 0$, then $(\boldsymbol{H}^k, \boldsymbol{W}^k)$ converges to a stationary point of the function $\Psi(\boldsymbol{H}, \boldsymbol{W})$.*

The proof of this statement is deferred to Appendix D.

It is also useful to notice that the gradient of $\mathscr{R}_\lambda(\boldsymbol{H})$ can be computed explicitly (this can be useful to devise a stopping criterion).

Proposition 4.2. *Consider the risk (4.1), with loss $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$, and assume that the rows of $\boldsymbol{H}$ are affine independent. Then, $\mathscr{R}_\lambda$ is differentiable at $\boldsymbol{H}$ with gradient*

$$\nabla\mathscr{R}_\lambda(\boldsymbol{H}) = 2\sum_{i=1}^{n} \boldsymbol{\alpha}_i^*\left(\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H})}(\boldsymbol{x}_i) - \boldsymbol{x}_i\right) + 2\lambda\big(\boldsymbol{H} - \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}(\boldsymbol{H})\big), \tag{4.10}$$

$$\boldsymbol{\alpha}_i^* = \arg\min_{\boldsymbol{\alpha} \in \Delta^r} \left\|\boldsymbol{H}^{\mathsf{T}}\boldsymbol{\alpha} - \boldsymbol{x}_i^{\mathsf{T}}\right\|_2 . \tag{4.11}$$

where we recall that $\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}(\boldsymbol{H})$ denotes the matrix with rows $\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}(\boldsymbol{H}_{1,\cdot}), \ldots, \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}(\boldsymbol{H}_{r,\cdot})$.

The proof of this proposition is given in Appendix C. Appendix E also discusses two alternative algorithms.

4.3 Numerical experiments

We implemented both the PALM algorithm described in the previous section, and the two algorithms described in Appendix E. The outcomes are generally similar.

Figures 3 and 4 repeat the experiment already described in the introduction. We generate $n = 250$ convex combinations of $r = 4$ spectra $\boldsymbol{h}_{0,1}, \ldots, \boldsymbol{h}_{0,4} \in \mathbb{R}^d$, $d = 87$, this time adding white

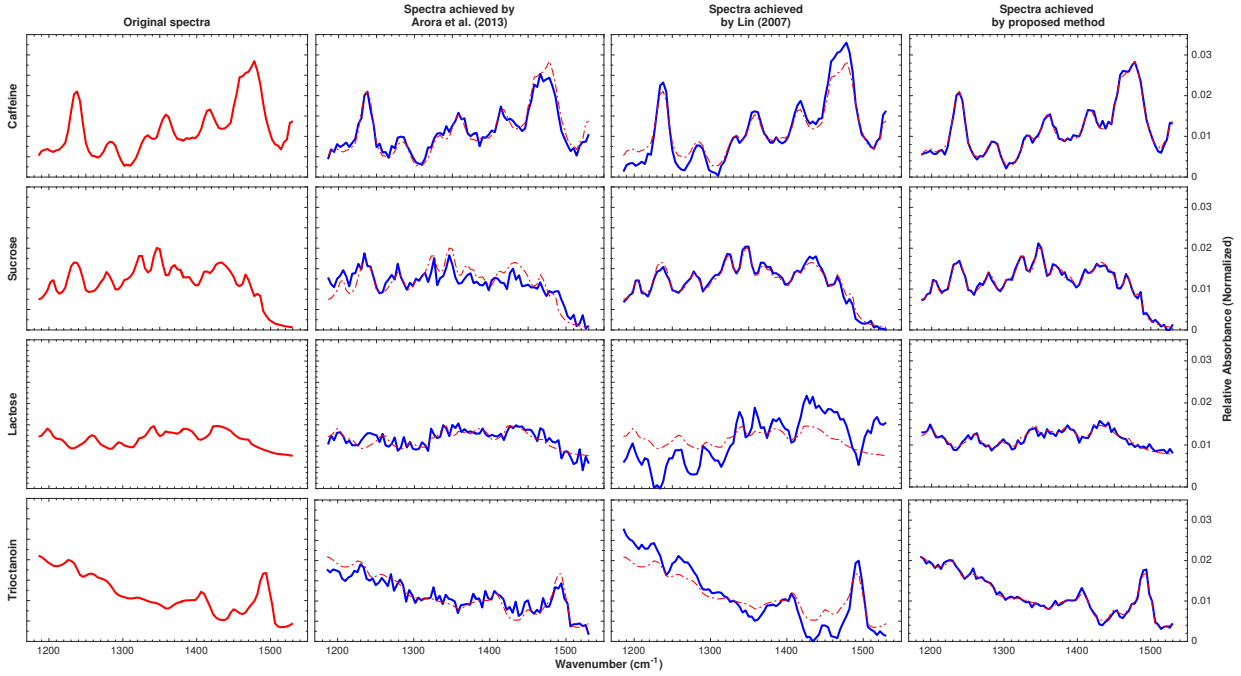

Figure 3: Reconstructing infrared spectra of four molecules, from noisy random convex combinations. Noise level $\sigma = 10^{-3}$. Left column: original spectra. The other columns correspond to different reconstruction methods.

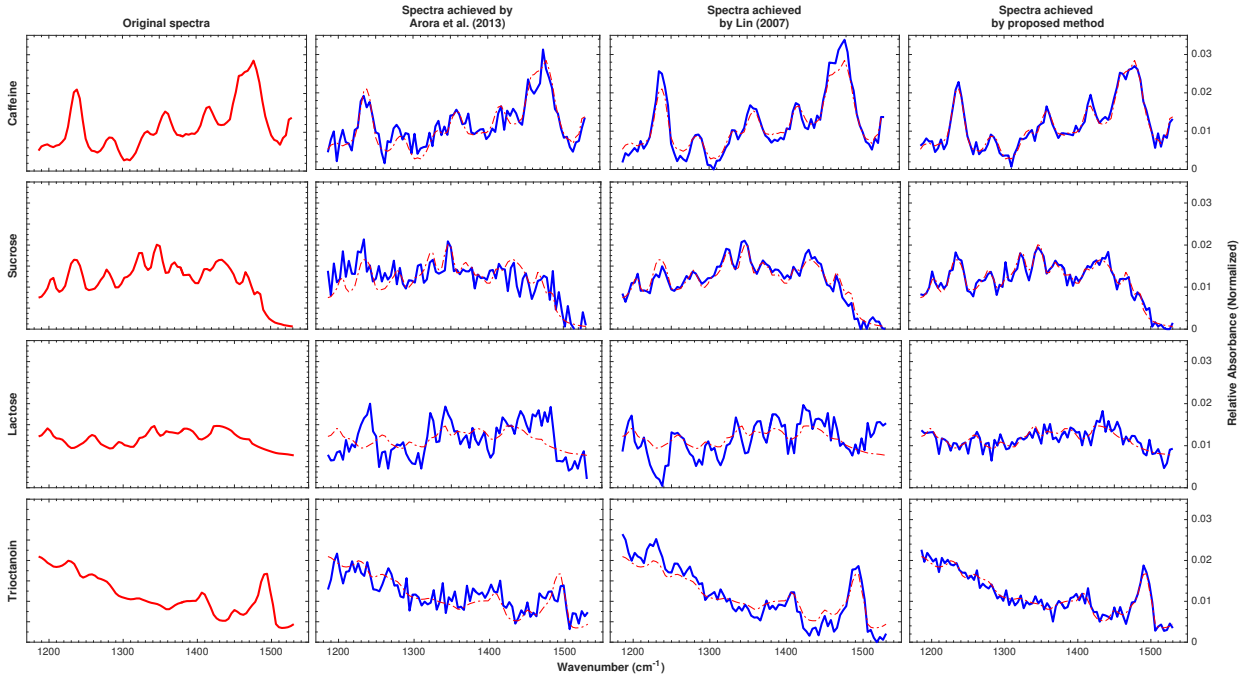

Figure 4: As in Figure 3, with $\sigma = 2 \cdot 10^{-3}$ (in blue).

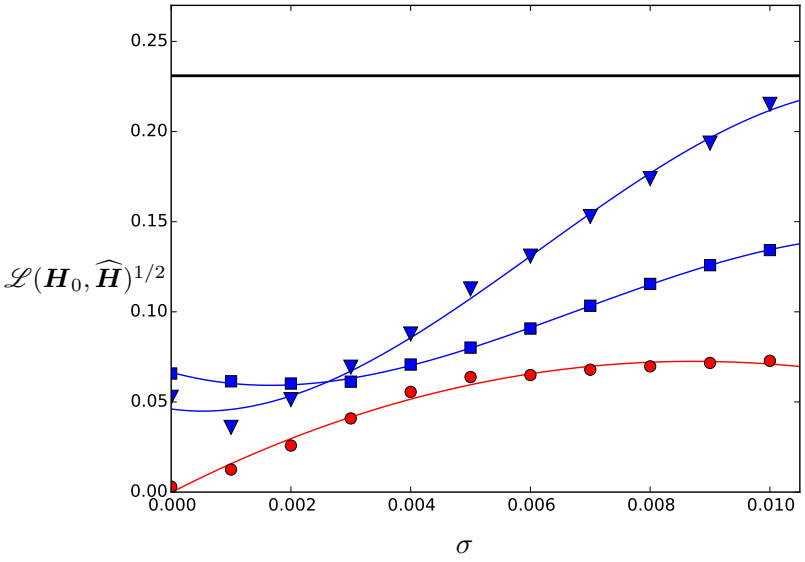

Figure 5: Risk $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}$ vs σ for different reconstruction methods. Triangles (blue): anchor words algorithm from [AGH$^+$13]. Squares (blue): minimizing the objective function (1.2) using the projected gradient algorithm of [Lin07]. Circles (red): archetypal reconstruction approach in this paper. Interpolating lines are just guides for the eye. The thick horizontal line corresponds to the trivial estimator $\widehat{\boldsymbol{H}} = 0$.

Gaussian noise with variance σ^2. We minimize the Lagrangian $\mathscr{R}_\lambda(\boldsymbol{H})$, with[4] $\lambda = 4$ (for Figure 3) and $\lambda = 0.8$ (for Figure 4). The reconstructed spectra of the pure analytes appear to be accurate and robust to noise.

In Figure 5 we repeated the same experiment systematically for 10 noise realizations for each noise level σ, and report the resulting average loss. Among various reconstruction methods, the approach described in this paper seem to have good robustness to noise and achieves exact reconstruction as $\sigma \to 0$.

5 Discussion

We introduced a new optimization formulation of the non-negative matrix factorization problem. In its Lagrangian formulation, our approach consists in minimizing the cost function $\mathscr{R}_\lambda(\boldsymbol{H})$ defined in Eq. (4.1). This encompasses applications in which only one of the factors is required to be non-negative. A special case of this formulation ($\lambda \to \infty$) corresponds to the 'archetypal analysis' of [CB94]. In this case, the archetype estimates coincide with a subset of the data points, which is appropriate only under the separability assumption of [DS03].

Our main technical result (Theorem 1) is a robustness guarantee for the reconstructed archetypes. This holds under the uniqueness assumption, which appears to hold for generic geometries of the dataset. In particular, while separability implies uniqueness (with optimal constant $\alpha = 1$), uniqueness holds for non-separable data as well. To the best of our knowledge, similar robustness results have been obtained under separability [RRTB12, AGH$^+$13, GV14, GV15] (albeit these works ob-

[4]These values were chosen as to approximately minimize the estimation error.

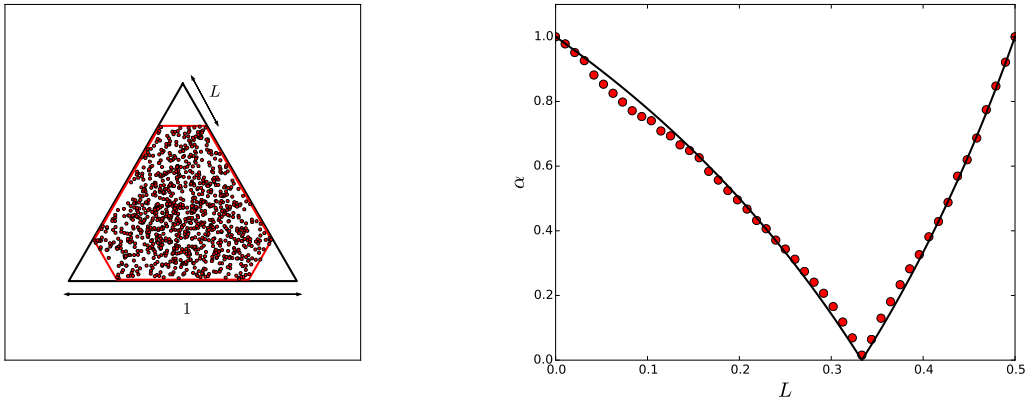

Figure 6: Numerical computation of the uniqueness parameter α. Left: data geometry. The red hexagon corresponds to $\mathrm{conv}(\boldsymbol{X})$, and the black (equilateral) triangle to the archetypes $\boldsymbol{H}_0$, for $L < 1/3$. For $L = 1/3$ the archetypes are not unique, and for $L \in (1/3, 1/2]$, they are given by an equilateral triangle rotated by $\pi/3$ (pointing down). Right: numerical evaluation of the uniqueness constant (red circles). The continuous line corresponds to an analytical upper bound (triangle rotated by $\pi/3$ with respect to $\boldsymbol{H}_0$).

tain a better dependence on r). The only exception is the recent work of Ge and Zou [GZ15] who prove robustness under a 'subset separability' condition, which provides a significant relaxation of separability. Under this condition, [GZ15] develops a polynomial-time algorithm to estimate the archetypes by identifying and intersecting the faces of $\mathrm{conv}(\boldsymbol{H}_0)$. However, the algorithm of [GZ15] exploits collinearities to identify the faces, and this requires additional 'genericity' assumptions.

Admittedly, the uniqueness constant α is difficult to evaluate analytically, even for simple geometries of the data. However, by definition it does not vanish except in the case of multiple minimizers, and we expect it typically to be of order one. Figure 6 illustrate this point by computing numerically α for a simple one-parameter family of geometries with $r = 3$, $d = 2$. The parameter α vanishes at a single point, corresponding to a degenerate problem with multiple solutions.

Finally, several earlier works addressed the non-uniqueness problem in classical non-negative matrix factorization. Among others, Miao and Qi [MQ07] penalize a matrix of archetypes $\boldsymbol{H}$ by the corresponding volume. Closely related to our work is the approach of Mørup and Hansen [MH12] tha also builds on archetypal analysis. To the best of our knowledge, none of these works establishes robustness of the proposed methods.

We conclude by mentioning three important problems that are not addressed by this paper: (1) Are there natural condition under which the risk function $\mathscr{R}_\lambda(\boldsymbol{H})$ of Eq. (4.1) can be optimized in polynomial time? We only provided an algorithm that is guaranteed to converge to a critical point. (2) We assumed the rank r to be known. In practice it will need to be estimated from the data. (3) Similarly, the regularization parameter λ should be chosen from data.

Acknowledgements

This work was partially supported by the NSF grant CCF-1319979 and a Stanford Graduate Fellowship.

References

[AGH+13] Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu, *A practical algorithm for topic modeling with provable guarantees.*, ICML (2), 2013, pp. 280–288.

[AGKM12] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra, *Computing a nonnegative matrix factorization–provably*, Proceedings of the forty-fourth annual ACM symposium on Theory of computing, ACM, 2012, pp. 145–162.

[ASG+01] Mário César Ugulino Araújo, Teresa Cristina Bezerra Saldanha, Roberto Kawakami Harrop Galvao, Takashi Yoneyama, Henrique Caldas Chame, and Valeria Visani, *The successive projections algorithm for variable selection in spectroscopic multicomponent analysis*, Chemometrics and Intelligent Laboratory Systems **57** (2001), no. 2, 65–73.

[BST14] Jérôme Bolte, Shoham Sabach, and Marc Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming **146** (2014), no. 1-2, 459–494.

[CB94] Adele Cutler and Leo Breiman, *Archetypal analysis*, Technometrics **36** (1994), no. 4, 338–347.

[DS03] David L Donoho and Victoria Stodden, *When does non-negative matrix factorization give a correct decomposition into parts?*

[GV14] Nicolas Gillis and Stephen A Vavasis, *Fast and robust recursive algorithmsfor separable nonnegative matrix factorization*, IEEE transactions on pattern analysis and machine intelligence **36** (2014), no. 4, 698–714.

[GV15] ———, *Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization*, SIAM Journal on Optimization **25** (2015), no. 1, 677–698.

[GZ15] Rong Ge and James Zou, *Intersecting faces: Non-negative matrix factorization with new guarantees*, Proceedings of the 32nd International Conference on Machine Learning (ICML-15), 2015, pp. 2295–2303.

[KSD08] Dongmin Kim, Suvrit Sra, and Inderjit S Dhillon, *Fast projection-based methods for the least squares nonnegative matrix approximation problem*, Statistical Analysis and Data Mining **1** (2008), no. 1, 38–51.

[Lin07] Chih-Jen Lin, *Projected gradient methods for nonnegative matrix factorization*, Neural computation **19** (2007), no. 10, 2756–2779.

[LM] P.J. Linstrom and W.G. Mallard (eds.), *Nist chemistry webbook, nist standard reference database number 69*, National Institute of Standards and Technology, Gaithersburg MD, 20899, `http://webbook.nist.gov`, (retrieved January 5, 2017).

[LS99] Daniel D Lee and H Sebastian Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature **401** (1999), no. 6755, 788–791.

[LS01] ———, *Algorithms for non-negative matrix factorization*, Advances in neural information processing systems, 2001, pp. 556–562.

[MBDC+14] Wing-Kin Ma, José M Bioucas-Dias, Tsung-Han Chan, Nicolas Gillis, Paul Gader, Antonio J Plaza, ArulMurugan Ambikapathi, and Chong-Yung Chi, *A signal processing perspective on hyperspectral unmixing: Insights from remote sensing*, IEEE Signal Processing Magazine **31** (2014), no. 1, 67–81.

[MH12] Morten Mørup and Lars Kai Hansen, *Archetypal analysis for machine learning and data mining*, Neurocomputing **80** (2012), 54–63.

[MN13] Boris S Mordukhovich and Nguyen Mau Nam, *An easy path to convex analysis and applications*, Synthesis Lectures on Mathematics and Statistics **6** (2013), no. 2, 1–218.

[MQ07] Lidan Miao and Hairong Qi, *Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization*, IEEE Transactions on Geoscience and Remote Sensing **45** (2007), no. 3, 765–777.

[Paa97] Pentti Paatero, *Least squares formulation of robust non-negative factor analysis*, Chemometrics and intelligent laboratory systems **37** (1997), no. 1, 23–35.

[PT94] Pentti Paatero and Unto Tapper, *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics **5** (1994), no. 2, 111–126.

[RCP09] Jonathan L Roux, Alain D Cheveigné, and Lucas C Parra, *Adaptive template matching with shift-invariant semi-nmf*, Advances in neural information processing systems, 2009, pp. 921–928.

[RRTB12] Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf, *Factoring nonnegative matrices with linear programs*, Advances in Neural Information Processing Systems, 2012, pp. 1214–1222.

[WL10] Fei Wang and Ping Li, *Efficient nonnegative matrix factorization with random projections*, Proceedings of the 2010 SIAM International Conference on Data Mining, SIAM, 2010, pp. 281–292.

[XLG03] Wei Xu, Xin Liu, and Yihong Gong, *Document clustering based on non-negative matrix factorization*, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, 2003, pp. 267–273.

[Zie12] Günter M Ziegler, *Lectures on polytopes*, vol. 152, Springer Science & Business Media, 2012.

A Further details on numerical experiments

The data in Figures 1, 3, and 4 were generated as follows. We retrieved infrared reflection spectra of caffeine, sucrose, lactose and trioctanoin from the NIST Chemistry WebBook dataset [LM]. We restricted these spectra to the wavenumbers between 1186 cm^{-1} and 1530 cm^{-1}, and denote by $\boldsymbol{h}_{0,1}, \dots, \boldsymbol{h}_{0,4} \in \mathbb{R}^d$, $d = 87$ the vector representations of these spectra. We then generates data $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \le n = 250$ by letting

$$\boldsymbol{x}_i = \sum_{\ell=1}^{4} w_{i,\ell} \boldsymbol{h}_\ell + \boldsymbol{z}_i \,, \tag{A.1}$$

where $\boldsymbol{z}_i \sim \mathsf{N}(0, \sigma^2 \mathrm{I}_d)$ are i.i.d. Gaussian noise vectors. The weights $\boldsymbol{w}_i = (w_{i,\ell})_{\ell \le 4}$ were generated as follows. The weight vectors $\{\boldsymbol{w}_i\}_{1 \le i \le 9}$ are generated such that they have 2 nonzero entries. In other words, 9 data points are on one dimensional facets of the polytope generated by $\boldsymbol{h}_{0,1}, \dots, \boldsymbol{h}_{0,4}$. In order to randomly generate these weight vectors, for each $1 \le i \le 9$, a pair of indices (ℓ_1, ℓ_2) between 1 and 4 is chosen uniformly at random. Then $\{\widetilde{\boldsymbol{w}}\}_{1 \le i \le 9}$, $\widetilde{\boldsymbol{w}} \in \mathbb{R}^2$ are generated as independent Dirichlet random vectors with parameter $(5, 5)$. Then we let $w_{i,\ell_1} = \tilde{w}_{i,1}$ and $w_{i,\ell_2} = \tilde{w}_{i,2}$ for $1 \le i \le 9$. The weight vectors $\{\boldsymbol{w}_i\}_{10 \le i \le 20}$ each have 3 nonzero entries. Similar to above, for each of these weight vectors a 3-tuple of indices (ℓ_1, ℓ_2, ℓ_3) between 1 and 4 is chosen uniformly at random. Then we let $w_{i,\ell_1} = \tilde{w}_{i,1}$, $w_{i,\ell_2} = \tilde{w}_{i,2}$, $w_{i,\ell_3} = \tilde{w}_{i,4}$ for $10 \le i \le 20$, where $\{\widetilde{\boldsymbol{w}}\}_{10 \le i \le 20}$, $\widetilde{\boldsymbol{w}} \in \mathbb{R}^3$ are i.i.d. Dirichlet random vectors with parameter $(5, 5, 5)$. The rest of the weight vectors have cardinality equal to 4. Hence, for $21 \le i \le 250$, $\boldsymbol{w}_i$ are generated as i.i.d. Dirichlet random vectors with parameter $(5, 5, 5, 5)$.

B Proof of Theorem 1

In this appendix we prove Theorem 1. We start by recalling some notations already defined in the main text, and introducing some new ones. We will then state a stronger form of the theorem (with better dependence on the problem geometry in some regimes). Finally, we will present the actual proof.

Throughout this appendix, we assume the square loss $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.

B.1 Notations and definitions

We use bold capital letters (e.g. $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$,...) for matrices, bold lower case for vectors (e.g. $\boldsymbol{x}$, $\boldsymbol{y}$, ...) and plain lower case for scalars (aa, b, c and so on). In particular, $\boldsymbol{e}_i \in \mathbb{R}^d$ denotes the i'th vector in the canonical basis, $E^{r,d} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_r\}$ and for $r \le d$, $\boldsymbol{E}_{r,d} \in \{0,1\}^{r \times d}$ is the matrix whose i'th column is $\boldsymbol{e}_i$, and whose columns after the r-th one are equal to $\boldsymbol{0}$. For a matrix $\boldsymbol{X}$, $\boldsymbol{X}_{i,\cdot}$ and $\boldsymbol{X}_{\cdot,i}$ are its i'th row and column, respectively.

As in the main text, we denote by Δ^m the m-dimensional standard simplex, i.e. $\Delta^m = \{\boldsymbol{x} \in \mathbb{R}^m_{\ge 0}, \langle \boldsymbol{x}, \boldsymbol{1} \rangle = 1\}$, where $\boldsymbol{1} \in \mathbb{R}^m$ is the all ones vector. For a matrix $\boldsymbol{H} \in \mathbb{R}^{r \times d}$, we use $\sigma_{\max}(\boldsymbol{H})$, $\sigma_{\min}(\boldsymbol{H})$ to denote its largest and smallest nonzero singular values and $\kappa(\boldsymbol{H}) = \sigma_{\max}(\boldsymbol{H})/\sigma_{\min}(\boldsymbol{H})$ to denote its condition number. We denote by $\mathrm{conv}(\boldsymbol{H}), \mathrm{aff}(\boldsymbol{H})$ the convex hull and the affine hull of the rows of $\boldsymbol{H}$, respectively. In other words,

$$\mathrm{conv}(\boldsymbol{H}) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} = \boldsymbol{H}^{\mathsf{T}} \boldsymbol{\pi}, \boldsymbol{\pi} \in \Delta^r\}, \tag{B.1}$$

$$\mathrm{aff}(\boldsymbol{H}) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} = \boldsymbol{H}^{\mathsf{T}} \boldsymbol{\alpha}, \langle \boldsymbol{1}, \boldsymbol{\alpha} \rangle = 1\}. \tag{B.2}$$

We denote by $Q_{r,n}$ is the set of r by n row stochastic matrices. Namely,

$$Q_{r,n} = \left\{ \boldsymbol{\Pi} \in \mathbb{R}_{\geq 0}^{r \times n} : \langle \boldsymbol{\Pi}_{i,\cdot}, \mathbf{1} \rangle = 1 \right\}. \tag{B.3}$$

with use $Q_r \equiv Q_{r,r}$. Further, S_r is defined as

$$S_r = \{ \boldsymbol{\Pi} \in Q_r : \Pi_{i,j} \in \{0,1\} \}. \tag{B.4}$$

As a consequence, given $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{H}_1, \boldsymbol{H}_2 \in \mathbb{R}^{r \times d}$, the loss functions $\mathscr{D}(\,\cdot\,,\,\cdot\,)$ and $\mathscr{L}(\,\cdot\,,\,\cdot\,)$ take the form

$$\mathscr{D}(\boldsymbol{H}_1, \boldsymbol{X}) = \min_{\boldsymbol{\Pi} \in Q_{r,n}} \|\boldsymbol{H}_1 - \boldsymbol{\Pi}\boldsymbol{X}\|_F^2, \tag{B.5}$$

$$\mathscr{L}(\boldsymbol{H}_1, \boldsymbol{H}_2) = \min_{\boldsymbol{\Pi} \in S_r} \|\boldsymbol{H}_1 - \boldsymbol{\Pi}\boldsymbol{H}_2\|_F^2. \tag{B.6}$$

We use $\mathsf{B}_m(\rho)$ to denote the closed ball with radius ρ in m dimensions, centered at 0. In addition, for $\boldsymbol{H} \in \mathbb{R}^{m \times d}$ we define the ρ-neighborhood of $\mathrm{conv}(\boldsymbol{H})$ as

$$\mathsf{B}_r(\rho; \boldsymbol{H}) := \{ \boldsymbol{x} \in \mathbb{R}^d : \mathscr{D}(\boldsymbol{x}, \boldsymbol{H}) \leq \rho^2 \}. \tag{B.7}$$

For a convex set $\mathcal{C}$ we denote the set of its extremal points by $\mathrm{ext}(\mathcal{C})$ and the projection of a point $\boldsymbol{x} \in \mathbb{R}^d$ onto $\mathcal{C}$ by $\boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})$. Namely,

$$\boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y} \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{B.8}$$

Also, for a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, and a mapping (not necessarily linear) $\boldsymbol{P} : \mathbb{R}^d \to \mathbb{R}^d$, $\boldsymbol{P}(\boldsymbol{X}) \in \mathbb{R}^{n \times d}$ is the matrix whose i'th row is $\boldsymbol{P}(\boldsymbol{X}_{i,\cdot})$.

B.2 Theorem statement

The statement below provides more detailed result with respect to the one in Theorem 1.

Theorem 2. *Assume $\boldsymbol{X} = \boldsymbol{W}_0\boldsymbol{H}_0 + \boldsymbol{Z}$ where the factorization $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ satisfies the uniqueness assumption with parameter $\alpha > 0$, and that $\mathrm{conv}(\boldsymbol{X}_0)$ has internal radius $\mu > 0$. Consider the estimator $\widehat{\boldsymbol{H}}$ defined by Eq. (2.5), with $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ (square loss) and $\delta = \max_{i \leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2$. If*

$$\max_{i \leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \frac{\alpha\mu}{30 r^{3/2}}, \tag{B.9}$$

then, setting $\delta = \max_{i \leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2$ in the problem (2.5) we get

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}) \leq \frac{C_*^2 r^5}{\alpha^2} \max_{i \leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2^2, \tag{B.10}$$

where C_* is a coefficient that depends uniquely on the geometry of $\boldsymbol{H}_0$, $\boldsymbol{X}_0$, namely $C_* = 120(\sigma_{\max}(\boldsymbol{H}_0)/\mu) \cdot \max(1, \kappa(\boldsymbol{H}_0)/\sqrt{r})$.

Further, if

$$\max_{i \leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \frac{\alpha\mu}{330\kappa(\boldsymbol{H}_0) r^{5/2}}, \tag{B.11}$$

then, setting $\delta = \max_{i \leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2$ in the problem (2.5) we get

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}) \leq \frac{C_{**}^2 r^4}{\alpha^2} \max_{i \leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2^2, \tag{B.12}$$

where $C_{**} = 120 \max(\kappa(\boldsymbol{H}_0), (\sigma_{\max}(\boldsymbol{H}_0)/r + \|\boldsymbol{z}_0\|_2)/(\mu r^{1/2})) \cdot \max(1, \kappa(\boldsymbol{H}_0)/\sqrt{r})$.

B.3 Proof

B.3.1 Lemmas

Lemma B.1. *Let $\mathcal{R}$ be a convex set and $\mathcal{C}$ be a convex cone. Define*

$$\gamma_{\mathcal{C}} = \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{v}\in\mathcal{C},\|\boldsymbol{v}\|_2=1} \langle \boldsymbol{u}, \boldsymbol{v}\rangle. \tag{B.13}$$

We have

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 + (1+\gamma_{\mathcal{C}}) \max_{x\in\mathrm{ext}(\mathcal{R})} \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2 \geq \gamma_{\mathcal{C}} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \|\boldsymbol{x}\|_2. \tag{B.14}$$

An illustration of this lemma in the case of $\mathcal{R} \subset \mathcal{C}$ is given in Figure 7. Note that, $\gamma_{\mathcal{C}}$ measures the pointedness of the cone $\mathcal{C}$. Geometrically (for $\mathcal{R} \subseteq \mathcal{C}$) the lemma states that the cosine of the angle between $\arg\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2$ and $\arg\min_{\boldsymbol{x}\in\mathrm{ext}(R)} \|\boldsymbol{x}\|_2$ is smaller than $\gamma_{\mathcal{C}}$.

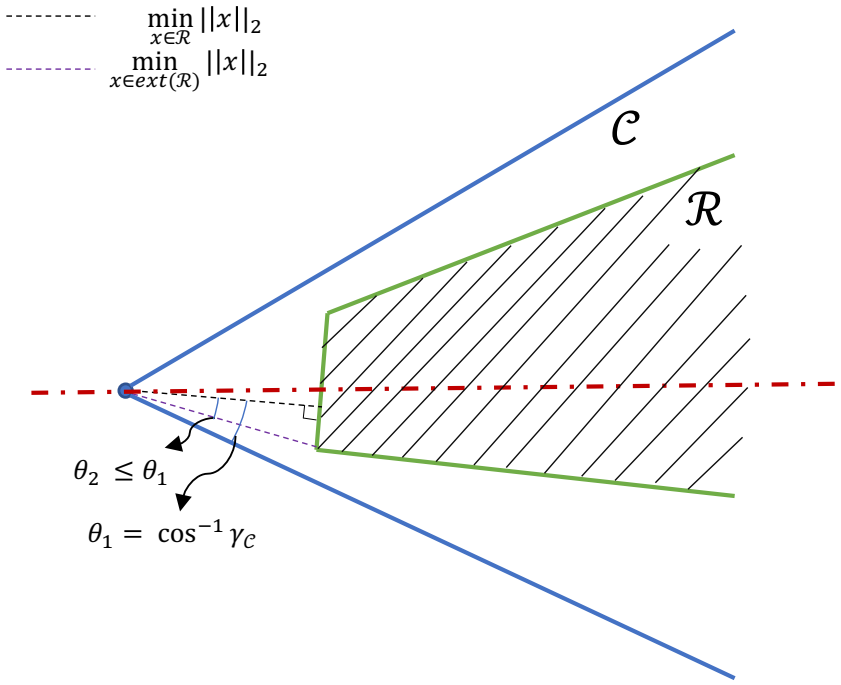

Figure 7: Picture of Lemma B.1, in the case, $\mathcal{R} \subset \mathcal{C}$.

Proof. We write

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 = \min_{\boldsymbol{x}\in\mathcal{R}} \max_{\|\boldsymbol{u}\|_2=1} \langle \boldsymbol{u}, \boldsymbol{x}\rangle \geq \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathcal{R}} \langle \boldsymbol{u}, \boldsymbol{x}\rangle = \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \langle \boldsymbol{u}, \boldsymbol{x}\rangle. \tag{B.15}$$

Replacing

$$\boldsymbol{x} = \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x}))\,, \tag{B.16}$$

we get

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 \geq \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \langle \boldsymbol{u}, \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x}))\rangle \tag{B.17}$$

$$\geq \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \langle \boldsymbol{u}, \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\rangle - \max_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2. \tag{B.18}$$

Hence, using the definition of $\gamma_{\mathcal{C}}$, we have

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 \geq \gamma_{\mathcal{C}} \min_{\boldsymbol{x}\in \mathrm{ext}(\mathcal{R})} \|\boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2 - \max_{\boldsymbol{x}\in \mathrm{ext}(\mathcal{R})} \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2. \tag{B.19}$$

Note that

$$\|\boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2 \geq \|\boldsymbol{x}\|_2 - \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2. \tag{B.20}$$

Therefore,

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 \geq \gamma_{\mathcal{C}} \min_{\boldsymbol{x}\in \mathrm{ext}(\mathcal{R})} \|\boldsymbol{x}\|_2 - (1+\gamma_{\mathcal{C}}) \max_{\boldsymbol{x}\in \mathrm{ext}(\mathcal{R})} \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2, \tag{B.21}$$

and this completes the proof. □

The next lemma is a consequence of Lemma B.1.

Lemma B.2. *Let $\boldsymbol{H}, \boldsymbol{H}_0 \in \mathbb{R}^{r\times d}$, $r \leq d$, be matrices with linearly independent rows. We have*

$$\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \leq \sqrt{2}\kappa(\boldsymbol{H}_0)\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} + (1+\sqrt{2})\sqrt{r}\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}. \tag{B.22}$$

Proof. Consider the cone $\mathcal{C}_1 \subset \mathbb{R}^d$, generated by vectors $\boldsymbol{e}_2 - \boldsymbol{e}_1, \ldots, \boldsymbol{e}_r - \boldsymbol{e}_1 \in \mathbb{R}^d$, i.e,

$$\mathcal{C}_1 = \left\{ \boldsymbol{v} \in \mathbb{R}^d; \boldsymbol{v} = \sum_{i=2}^{r} v_i(\boldsymbol{e}_i - \boldsymbol{e}_1), v_i \geq 0 \right\}. \tag{B.23}$$

For $\boldsymbol{v} \in \mathcal{C}_1, \|\boldsymbol{v}\|_2 = 1$ we have

$$\boldsymbol{v} = (-\langle \mathbf{1}, \boldsymbol{x}\rangle, \boldsymbol{x}, 0, 0, \ldots, 0), \tag{B.24}$$

where $\boldsymbol{x} \in \mathbb{R}^{r-1}_{\geq 0}$ and

$$\|\boldsymbol{x}\|_2^2 + \langle \mathbf{1}, \boldsymbol{x}\rangle^2 = 1. \tag{B.25}$$

Since, $\langle \mathbf{1}, \boldsymbol{x}\rangle = \|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x}\|_2$, we get $\langle \mathbf{1}, \boldsymbol{x}\rangle \geq 1/\sqrt{2}$. Thus, for $\boldsymbol{u} = -\boldsymbol{e}_1$, we have $\langle \boldsymbol{u}, \boldsymbol{v}\rangle \geq 1/\sqrt{2}$. Therefore, for $\gamma_{\mathcal{C}_1}$ defined as in Lemma B.1, we have $\gamma_{\mathcal{C}_1} \geq 1/\sqrt{2}$. In addition, by symmetry, for $i \in \{1, 2, \ldots, r\}$, for the cone $\mathcal{C}_i \subset \mathbb{R}^d$, generated by vectors $\boldsymbol{e}_1 - \boldsymbol{e}_i, \boldsymbol{e}_2 - \boldsymbol{e}_i, \ldots, \boldsymbol{e}_r - \boldsymbol{e}_i \in \mathbb{R}^d$ we have $\gamma_{\mathcal{C}_i} = \gamma \geq 1/\sqrt{2}$. Hence, using Lemma B.1 for $\boldsymbol{H} \in \mathbb{R}^{r\times d}$, $\mathcal{R} = \mathrm{conv}(\boldsymbol{H}) - \boldsymbol{e}_j$ (the set obtained by translating $\mathrm{conv}(\boldsymbol{H})$ by $-\boldsymbol{e}_j$), $\mathcal{C} = \mathcal{C}_j$ we get for $j = 1, 2, \ldots, r$

$$\min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 \geq \gamma \min_{\boldsymbol{q}\in E^{r,r}} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 - (1+\gamma) \max_{i\in[r]} \min_{\boldsymbol{q}\in\mathbb{R}^r_{\geq 0}} \|\boldsymbol{H}^\mathsf{T}_{i,\cdot} - \boldsymbol{e}_j - \boldsymbol{E}^\mathsf{T}_{r,d}\boldsymbol{q} + \boldsymbol{e}_j\langle \mathbf{1}, \boldsymbol{q}\rangle\|_2 \tag{B.26}$$

$$\geq \gamma \min_{\boldsymbol{q}\in E^{r,r}} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 - (1+\gamma) \max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{H}^\mathsf{T}_{i,\cdot} - \boldsymbol{E}^\mathsf{T}_{r,d}\boldsymbol{q}\|_2. \tag{B.27}$$

Hence,

$$\sum_{j=1}^{r} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2^2 \geq \gamma^2 \sum_{j=1}^{r} \min_{\boldsymbol{q}\in E^{r,r}} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2^2 + (1+\gamma)^2 r \max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{H}^\mathsf{T}_{i,\cdot} - \boldsymbol{E}^\mathsf{T}_{r,d}\boldsymbol{q}\|_2^2$$
$$- 2\gamma(1+\gamma)\left(\max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{H}^\mathsf{T}_{i,\cdot} - \boldsymbol{E}^\mathsf{T}_{r,d}\boldsymbol{q}\|_2\right) \sum_{j=1}^{r} \min_{\boldsymbol{q}\in E^{r,r}} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 \tag{B.28}$$
$$\geq \left[\gamma\left(\sum_{j=1}^{r} \min_{\boldsymbol{q}\in E^{r,r}} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2^2\right)^{1/2}\right.$$
$$\left. - (1+\gamma)\sqrt{r}\left(\max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{H}^\mathsf{T}_{i,\cdot} - \boldsymbol{E}^\mathsf{T}_{r,d}\boldsymbol{q}\|_2\right)\right]^2. \tag{B.29}$$

Therefore,

$$\min_{\boldsymbol{Q}\in Q_r} \|\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{H}\|_F \geq \gamma \min_{\boldsymbol{Q}\in S_r} \|\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{H}\|_F - (1+\gamma)\sqrt{r} \max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{H}_{i,\cdot}^\mathsf{T} - \boldsymbol{E}_{r,d}^\mathsf{T}\boldsymbol{q}\|_2. \tag{B.30}$$

Now consider $\boldsymbol{H}_0 \in \mathbb{R}^{r\times d}$ where $\boldsymbol{H}_0 = \boldsymbol{E}_{r,d}\boldsymbol{M}$, $\boldsymbol{H} = \boldsymbol{Y}\boldsymbol{M}$, where $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ is invertible. We have

$$\begin{aligned}
\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} &= \min_{\boldsymbol{Q}\in Q_r} \|\boldsymbol{H}_0 - \boldsymbol{Q}\boldsymbol{H}\|_F = \min_{\boldsymbol{Q}\in Q_r} \|(\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{Y})\boldsymbol{M}\|_F && \text{(B.31)}\\
&\geq \sigma_{\min}(\boldsymbol{M}) \min_{\boldsymbol{Q}\in Q_r} \|\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{Y}\|_F && \text{(B.32)}\\
&\geq \gamma\sigma_{\min}(\boldsymbol{M}) \min_{\boldsymbol{Q}\in S_r} \|\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{Y}\|_F - \sigma_{\min}(\boldsymbol{M})\sqrt{r}(1+\gamma) \max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{Y}_{i,\cdot}^\mathsf{T} - \boldsymbol{E}_{r,d}^\mathsf{T}\boldsymbol{q}\|_2 && \text{(B.33)}\\
&= \gamma\sigma_{\min}(\boldsymbol{M}) \min_{\boldsymbol{Q}\in S_r} \|(\boldsymbol{H}_0 - \boldsymbol{Q}\boldsymbol{H})\boldsymbol{M}^{-1}\|_F &&\\
&\quad - \sigma_{\min}(\boldsymbol{M})\sqrt{r}(1+\gamma) \max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|(\boldsymbol{M}^{-1})^\mathsf{T}(\boldsymbol{H}_{i,\cdot}^\mathsf{T} - \boldsymbol{H}_0^\mathsf{T}\boldsymbol{q})\|_2. && \text{(B.34)}
\end{aligned}$$

Thus, using the fact that $\sigma_{\max}(\boldsymbol{M})/\sigma_{\min}(\boldsymbol{M}) = \kappa(\boldsymbol{M}) = \kappa(\boldsymbol{H}_0)$,

$$\begin{aligned}
\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} &\geq \frac{\gamma}{\kappa(\boldsymbol{H}_0)}\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} - \frac{(1+\gamma)\sqrt{r}}{\kappa(\boldsymbol{H}_0)} \max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{H}_{i,\cdot}^\mathsf{T} - \boldsymbol{H}_0^\mathsf{T}\boldsymbol{q}\|_2 && \text{(B.35)}\\
&\geq \frac{\gamma}{\kappa(\boldsymbol{H}_0)}\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} - \frac{(1+\gamma)\sqrt{r}}{\kappa(\boldsymbol{H}_0)}\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}. && \text{(B.36)}
\end{aligned}$$

Therefore,

$$\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \leq \frac{\kappa(\boldsymbol{H}_0)}{\gamma}\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} + \frac{(1+\gamma)\sqrt{r}}{\gamma}\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}. \tag{B.37}$$

Finally, note that the function $f(x) = (1+x)/x$ is monotone decreasing over $\mathbb{R}_{>0}$. Hence, for $\gamma \geq 1/\sqrt{2}$, $(1+\gamma)/\gamma \leq 1+\sqrt{2}$. Therefore, we get

$$\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \leq \sqrt{2}\kappa(\boldsymbol{H}_0)\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} + (1+\sqrt{2})\sqrt{r}\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2} \tag{B.38}$$

and this completes the proof. □

We continue with the following lemmas on the condition number of the matrix $\boldsymbol{H}$.

Lemma B.3. *Let $\boldsymbol{H}_0, \boldsymbol{H} \in \mathbb{R}^{r\times d}$, $r \leq d$,with $\boldsymbol{H}$ having full row rank. We have*

$$\sigma_{\max}(\boldsymbol{H}) \leq \mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2} + \sqrt{r}\sigma_{\max}(\boldsymbol{H}_0),. \tag{B.39}$$

In addition, if

$$\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \leq \frac{\sigma_{\min}(\boldsymbol{H}_0)}{2}, \tag{B.40}$$

then

$$\kappa(\boldsymbol{H}) \leq \frac{2r\sigma_{\max}(\boldsymbol{H}_0) + 2\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}\sqrt{r}}{\sigma_{\min}(\boldsymbol{H}_0)}. \tag{B.41}$$

Further, if

$$\mathscr{D}(\boldsymbol{H},\boldsymbol{H}_0)^{1/2}+\mathscr{D}(\boldsymbol{H}_0,\boldsymbol{H})^{1/2}\le\frac{\sigma_{\min}(\boldsymbol{H}_0)}{6\sqrt{r}}, \tag{B.42}$$

then

$$\sigma_{\max}(\boldsymbol{H})\le 2\sigma_{\max}(\boldsymbol{H}_0), \tag{B.43}$$
$$\kappa(\boldsymbol{H})\le(7/2)\kappa(\boldsymbol{H}_0). \tag{B.44}$$

Proof. For the sake of simplicity, we will write $\mathscr{D}_1=\mathscr{D}(\boldsymbol{H},\boldsymbol{H}_0)^{1/2}$, $\mathscr{D}_2=\mathscr{D}(\boldsymbol{H}_0,\boldsymbol{H})^{1/2}$ Note that using the assumptions of Lemma B.3 we have

$$\begin{aligned}\boldsymbol{H}_0&=\boldsymbol{P}\boldsymbol{H}+\boldsymbol{A}_2;\quad \|\boldsymbol{A}_2\|_F=\mathscr{D}_2,\\ \boldsymbol{H}&=\boldsymbol{R}\boldsymbol{H}_0+\boldsymbol{A}_1;\quad \|\boldsymbol{A}_1\|_F=\mathscr{D}_1,\end{aligned} \tag{B.45}$$

where $\boldsymbol{P},\boldsymbol{R}\in\mathbb{R}^{r\times r}_{\ge 0}$ are row-stochastic matrices and $\boldsymbol{A}_1,\boldsymbol{A}_2\in\mathbb{R}^{r\times d}$. Also, $\sigma_{\max}(\boldsymbol{A}_1)\le\|\boldsymbol{A}_1\|_F=\mathscr{D}_1$, $\sigma_{\max}(\boldsymbol{A}_2)\le\|\boldsymbol{A}_2\|_F=\mathscr{D}_2$. Therefore,

$$\sigma_{\max}(\boldsymbol{P})\sigma_{\min}(\boldsymbol{H})\ge\sigma_{\min}(\boldsymbol{P}\boldsymbol{H})\ge\sigma_{\min}(\boldsymbol{H}_0)-\sigma_{\max}(\boldsymbol{A}_2)\ge\sigma_{\min}(\boldsymbol{H}_0)-\mathscr{D}_2. \tag{B.46}$$

In addition, note that for a row stochastic matrix $\boldsymbol{P}\in Q_r$, we have

$$\sigma_{\max}(\boldsymbol{P})\le\|\boldsymbol{P}\|_F=\left(\sum_{i=1}^{r}\|\boldsymbol{P}_{i,\cdot}\|_2^2\right)^{1/2}\le\left(\sum_{i=1}^{r}\|\boldsymbol{P}_{i,\cdot}\|_1^2\right)^{1/2}\le\sqrt{r}. \tag{B.47}$$

Hence, for $\mathscr{D}_2\le\sigma_{\min}(\boldsymbol{H}_0)$ we get

$$\sigma_{\min}(\boldsymbol{H})\ge\frac{\sigma_{\min}(\boldsymbol{H}_0)-\mathscr{D}_2}{\sqrt{r}}. \tag{B.48}$$

In addition,

$$\sigma_{\max}(\boldsymbol{H})\le\sigma_{\max}(\boldsymbol{R}\boldsymbol{H}_0)+\sigma_{\max}(\boldsymbol{A}_1)\le\sigma_{\max}(\boldsymbol{R})\sigma_{\max}(\boldsymbol{H}_0)+\mathscr{D}_1\le\sqrt{r}\sigma_{\max}(\boldsymbol{H}_0)+\mathscr{D}_1. \tag{B.49}$$

Hence, using (B.48), (B.49), for $\mathscr{D}_2\le\sigma_{\min}(\boldsymbol{H}_0)$ we have

$$\kappa(\boldsymbol{H})\le\frac{r\sigma_{\max}(\boldsymbol{H}_0)+\mathscr{D}_1\sqrt{r}}{\sigma_{\min}(\boldsymbol{H}_0)-\mathscr{D}_2}. \tag{B.50}$$

Thus, for $\mathscr{D}_2\le\sigma_{\min}(\boldsymbol{H}_0)/2$, we get Eqs. (B.39), (B.41).

Now assume that $\mathscr{D}_1+\mathscr{D}_2\le\sigma_{\min}(\boldsymbol{H}_0)/(6\sqrt{r})$. In this case, using (B.45) we have

$$\boldsymbol{H}_0=\boldsymbol{P}(\boldsymbol{R}\boldsymbol{H}_0+\boldsymbol{A}_1)+\boldsymbol{A}_2. \tag{B.51}$$

Therefore,

$$(\mathbf{I}-\boldsymbol{P}\boldsymbol{R})\boldsymbol{H}_0=\boldsymbol{P}\boldsymbol{A}_1+\boldsymbol{A}_2, \tag{B.52}$$

hence,

$$\mathbf{I}-\boldsymbol{P}\boldsymbol{R}=(\boldsymbol{P}\boldsymbol{A}_1+\boldsymbol{A}_2)\boldsymbol{H}_0^{\dagger} \tag{B.53}$$

and

$$\boldsymbol{P}\boldsymbol{R} = \mathbf{I} - \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{H}_0^\dagger - \boldsymbol{A}_2\boldsymbol{H}_0^\dagger. \tag{B.54}$$

where $\boldsymbol{H}_0^\dagger$ is the right inverse of matrix $\boldsymbol{H}_0$. Note that

$$\sigma_{\max}(\boldsymbol{H}_0^\dagger) = \sigma_{\min}(\boldsymbol{H}_0)^{-1}. \tag{B.55}$$

By permuting the rows and columns of $\boldsymbol{H}_0$, without loss of generality, we can assume that $R_{ii} = \|\boldsymbol{R}_{\cdot,i}\|_\infty$. We can write

$$\begin{aligned}
R_{ii} \geq \langle \boldsymbol{P}_{i,\cdot}, \boldsymbol{R}_{\cdot,i} \rangle &= 1 - (\boldsymbol{P}\boldsymbol{A}_1\boldsymbol{H}_0^\dagger)_{ii} - (\boldsymbol{A}_2\boldsymbol{H}_0^\dagger)_{ii} && \text{(B.56)}\\
&\geq 1 - \|(\boldsymbol{P}\boldsymbol{A}_1\boldsymbol{H}_0^\dagger)_{i,\cdot}\|_2 - \|(\boldsymbol{A}_2\boldsymbol{H}_0^\dagger)_{i,\cdot}\|_2 && \text{(B.57)}\\
&\geq 1 - \max_{\boldsymbol{u}\in\Delta^r} \|\boldsymbol{A}_1^\mathsf{T}\boldsymbol{u}\|_2 \sigma_{\max}(\boldsymbol{H}_0^\dagger) - \|(\boldsymbol{A}_2)_{i,\cdot}\|_2 \sigma_{\max}(\boldsymbol{H}_0^\dagger) && \text{(B.58)}\\
&\geq 1 - \max_{\boldsymbol{u}\in\Delta^r} \|\boldsymbol{u}\|_2 \sigma_{\max}(\boldsymbol{A}_1)\sigma_{\max}(\boldsymbol{H}_0^\dagger) - \|\boldsymbol{A}_2\|_F \sigma_{\max}(\boldsymbol{H}_0^\dagger) && \text{(B.59)}\\
&\geq 1 - \frac{\mathscr{D}_1 + \mathscr{D}_2}{\sigma_{\min}(\boldsymbol{H}_0)}. && \text{(B.60)}
\end{aligned}$$

Hence, for all $i, j \in [r], i \neq j$, since $\boldsymbol{R}$ is row-stochastic,

$$R_{ji} \leq \frac{\mathscr{D}_1 + \mathscr{D}_2}{\sigma_{\min}(\boldsymbol{H}_0)}. \tag{B.61}$$

Thus,

$$\begin{aligned}
\langle \boldsymbol{P}_{i,\cdot}, \boldsymbol{R}_{\cdot,i} \rangle = R_{ii}P_{ii} + \sum_{j\neq i} P_{ij}R_{ji} &\leq R_{ii}P_{ii} + \left(\max_{j\neq i} R_{ji}\right) \sum_{j\neq i} P_{ij} && \text{(B.62)}\\
&\leq P_{ii} + \frac{\mathscr{D}_1 + \mathscr{D}_2}{\sigma_{\min}(\boldsymbol{H}_0)}(1 - P_{ii}). && \text{(B.63)}
\end{aligned}$$

Therefore, using (B.60),

$$P_{ii} \geq \frac{\sigma_{\min}(\boldsymbol{H}_0) - 2(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.64}$$

Thus, we can write

$$\boldsymbol{P} = \mathbf{I} + \Delta; \quad \|\Delta_{i,\cdot}\|_1 \leq \frac{2(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.65}$$

Therefore,

$$\sigma_{\max}(\Delta) \leq \|\Delta\|_F = \left(\sum_{i=1}^r \|\Delta_{i,\cdot}\|_2^2\right)^{1/2} \leq \left(\sum_{i=1}^r \|\Delta_{i,\cdot}\|_1^2\right)^{1/2} \leq \frac{2(\mathscr{D}_1 + \mathscr{D}_2)\sqrt{r}}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.66}$$

Hence,

$$\sigma_{\max}(\boldsymbol{P}) \leq 1 + \frac{2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}, \qquad \sigma_{\min}(\boldsymbol{P}) \geq 1 - \frac{2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.67}$$

From (B.45) we have $\sigma_{\min}(\boldsymbol{PH}) \geq \sigma_{\min}(\boldsymbol{H}_0) - \mathscr{D}_2$. Using $\sigma_{\min}(\boldsymbol{PH}) \leq \sigma_{\max}(\boldsymbol{P})\sigma_{\min}(\boldsymbol{H})$, we get

$$\sigma_{\min}(\boldsymbol{H}) \geq \frac{(\sigma_{\min}(\boldsymbol{H}_0) - \mathscr{D}_2)(\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2))}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2) + 2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.68}$$

Further, from (B.45) we have $\sigma_{\max}(\boldsymbol{PH}) \leq \sigma_{\max}(\boldsymbol{H}_0) + \mathscr{D}_2$. Using $\sigma_{\max}(\boldsymbol{PH}) \geq \sigma_{\min}(\boldsymbol{P})\sigma_{\max}(\boldsymbol{H})$, we get

$$\sigma_{\max}(\boldsymbol{H}) \leq \frac{(\sigma_{\max}(\boldsymbol{H}_0) + \mathscr{D}_2)(\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2))}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2) - 2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.69}$$

Hence, for $\mathscr{D}_1 + \mathscr{D}_2 \leq \sigma_{\min}(\boldsymbol{H}_0)/(6\sqrt{r})$, we have $\sigma_{\max}(\boldsymbol{H}) \leq 35\sigma_{\max}(\boldsymbol{H}_0)/18 < 2\sigma_{\max}(\boldsymbol{H}_0)$. In addition,

$$\kappa(\boldsymbol{H}) \leq \left(\frac{\sigma_{\max}(\boldsymbol{H}_0) + \mathscr{D}_2}{\sigma_{\min}(\boldsymbol{H}_0) - \mathscr{D}_2}\right)\left(1 + \frac{4\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2) - 2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}\right) \tag{B.70}$$

$$\leq \frac{6\kappa(\boldsymbol{H}_0) + 1}{5}\left(1 + \frac{4}{3}\right) \leq \frac{42\kappa(\boldsymbol{H}_0) + 7}{15} < \frac{7\kappa(\boldsymbol{H}_0)}{2}, \tag{B.71}$$

and this completes the proof. □

Lemma B.4. *Let $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0 \in \mathbb{R}^{n\times d}$ be such that $\operatorname{conv}(\boldsymbol{X}_0)$ has internal radius at least $\mu > 0$, and $\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{Z}$ with $\max_{i\leq n}\|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \delta$. If $\boldsymbol{H} \in \mathbb{R}^{r\times d}, \boldsymbol{H}_{i,\cdot} \in \operatorname{aff}(\boldsymbol{H}_0)$ is feasible for problem (2.5) and has linearly independent rows, then we have*

$$\sigma_{\min}(\boldsymbol{H}) \geq \sqrt{2}(\mu - 2\delta)\,. \tag{B.72}$$

Proof. Let

$$\boldsymbol{X}'_{i,\cdot} = \boldsymbol{\Pi}_{\operatorname{conv}(\boldsymbol{H})}(\boldsymbol{X}_{i,\cdot}) \equiv \arg\min_{\boldsymbol{x}\in\operatorname{conv}(\boldsymbol{H})}\|\boldsymbol{X}_{i,\cdot} - \boldsymbol{x}\|_2\,. \tag{B.73}$$

Note that since $\boldsymbol{H}$ is feasible for problem (2.5) and $\max_{i\leq n}\|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \delta$

$$\|(\boldsymbol{X}_0)_{i,\cdot} - \boldsymbol{X}'_{i,\cdot}\|_2 \leq \|(\boldsymbol{X}_0)_{i,\cdot} - \boldsymbol{X}_{i,\cdot}\|_2 + \|\boldsymbol{X}_{i,\cdot} - \boldsymbol{X}'_{i,\cdot}\|_2 \leq 2\delta. \tag{B.74}$$

Therefore, for any $\boldsymbol{x}_0 \in \operatorname{conv}(\boldsymbol{X}_0)$, writing $\boldsymbol{x}_0 = \boldsymbol{X}_0^{\mathsf{T}}\boldsymbol{a}_0$, $\boldsymbol{a}_0 \in \Delta^n$, we have

$$\mathscr{D}(\boldsymbol{x}_0, \boldsymbol{X}')^{1/2} = \min_{\boldsymbol{a}\in\Delta^n}\left\|\boldsymbol{X}_0^{\mathsf{T}}\boldsymbol{a}_0 - \boldsymbol{X}'^{\mathsf{T}}\boldsymbol{a}\right\|_2 \leq \left\|\boldsymbol{X}_0^{\mathsf{T}}\boldsymbol{a}_0 - \boldsymbol{X}'^{\mathsf{T}}\boldsymbol{a}_0\right\|_2 \tag{B.75}$$

$$\leq \left(\sum_{i=1}^{n}(a_0)_i\right)\|(\boldsymbol{X}_0)_{i,\cdot} - \boldsymbol{X}'_{i,\cdot}\|_2 \leq 2\delta. \tag{B.76}$$

Since $\operatorname{conv}(\boldsymbol{X}_0)$ has internal radius at least μ, there exists $\boldsymbol{z}_0 \in \mathbb{R}^d$, and an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{d\times r'}$, $r' = r - 1$, such that $\boldsymbol{z}_0 + \boldsymbol{U}\mathsf{B}_{r'}(\mu) \subseteq \operatorname{conv}(\boldsymbol{X}_0)$. Hence, for every $\boldsymbol{z} \in \mathbb{R}^{r'}$, $\|\boldsymbol{z}\|_2 = 1$ there exists $\boldsymbol{a} \in \Delta^n$ such that

$$\mu\boldsymbol{U}\boldsymbol{z} + \boldsymbol{z}_0 = \boldsymbol{X}_0^{\mathsf{T}}\boldsymbol{a}. \tag{B.77}$$

Therefore, for any unit vector $\boldsymbol{u}$ in column space of $\boldsymbol{U}$, for the line segment

$$l_{\boldsymbol{u},\mu} = \{\boldsymbol{z} : \boldsymbol{z} = \boldsymbol{z}_0 + \alpha\boldsymbol{u}, |\alpha| \leq \mu\} \subseteq \operatorname{conv}(\boldsymbol{X}_0)\,. \tag{B.78}$$

Thus,

$$l_{\boldsymbol{u},\mu} \subseteq \boldsymbol{P}_{\boldsymbol{u}}(\mathrm{conv}(\boldsymbol{X}_0)) \tag{B.79}$$

where $\boldsymbol{P}_{\boldsymbol{u}}$ is the orthogonal projection onto the line containing $l_{\boldsymbol{u},\mu}$. Note that using (B.76), for any $\boldsymbol{x}_0 \in \mathrm{conv}(\boldsymbol{X}_0)$ we have

$$\mathscr{D}(\boldsymbol{P}_{\boldsymbol{u}}(\boldsymbol{x}_0), \boldsymbol{P}_{\boldsymbol{u}}(\mathrm{conv}(\boldsymbol{X}')))^{1/2} \leq \mathscr{D}(\boldsymbol{x}_0, \boldsymbol{X}')^{1/2} \leq 2\delta. \tag{B.80}$$

In other words, for any $\boldsymbol{x}_0 \in \boldsymbol{P}_{\boldsymbol{u}}(\mathrm{conv}(\boldsymbol{X}_0))$, $D(\boldsymbol{x}_0, \boldsymbol{P}_{\boldsymbol{u}}(\mathrm{conv}(\boldsymbol{X}'))) \leq 2\delta$. Therefore, using (B.78) for any $\boldsymbol{u}$ in column space of $\boldsymbol{U}$, we have

$$l_{\boldsymbol{u},\mu-2\delta} \subseteq \boldsymbol{P}_u(\mathrm{conv}(\boldsymbol{X}')). \tag{B.81}$$

This implies that

$$\boldsymbol{z}_0 + \boldsymbol{U}\mathsf{B}_{r'}(\mu - 2\delta) \subseteq \mathrm{conv}(\boldsymbol{X}') \subseteq \mathrm{conv}(\boldsymbol{H}). \tag{B.82}$$

Hence, for every $\boldsymbol{z} \in \mathbb{R}^{r'}$, $\|\boldsymbol{z}\|_2 = 1$ there exists $\boldsymbol{a} \in \Delta^r$ such that

$$(\mu - 2\delta)\boldsymbol{U}\boldsymbol{z} + \boldsymbol{z}_0 = \boldsymbol{H}^{\mathsf{T}}\boldsymbol{a}. \tag{B.83}$$

Note that $\boldsymbol{H}^{\mathsf{T}}$ has linearly independent columns. Multiplying the previous equation by $(\boldsymbol{H}^{\mathsf{T}})^{\dagger}$ the left inverse of $\boldsymbol{H}^{\mathsf{T}}$, we get

$$(\mu - 2\delta)(\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}\boldsymbol{z} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0 = \boldsymbol{a}. \tag{B.84}$$

Let

$$\boldsymbol{a}_1 = (\mu - 2\delta)(\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}\boldsymbol{v} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0\,, \tag{B.85}$$

$$\boldsymbol{a}_2 = -(\mu - 2\delta)(\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}\boldsymbol{v} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0\,, \tag{B.86}$$

where $\boldsymbol{v}$ is the right singular vector corresponding to the largest singular value of $(\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}$. Therefore, we have

$$\boldsymbol{a}_1 = (\mu - 2\delta)\sigma_{\max}((\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U})\boldsymbol{v} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0, \tag{B.87}$$

$$\boldsymbol{a}_2 = -(\mu - 2\delta)\sigma_{\max}((\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U})\boldsymbol{v} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0. \tag{B.88}$$

Thus, for $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \Delta^r$

$$\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|_2 = 2(\mu - 2\delta)\sigma_{\max}((\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}). \tag{B.89}$$

Note that

$$\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|_2 \leq \sqrt{2}. \tag{B.90}$$

Thus,

$$2(\mu - 2\delta)\sigma_{\max}((\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}) = \frac{2(\mu - 2\delta)}{\sigma_{\min}(\boldsymbol{H})} \leq \sqrt{2}. \tag{B.91}$$

Hence,

$$\sigma_{\min}(\boldsymbol{H}) \geq \sqrt{2}(\mu - 2\delta). \tag{B.92}$$

□

The following lemma states an important property of $\widehat{\boldsymbol{H}}$ the optimal solution of problem (2.5).

Lemma B.5. *If* $\max_i \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \delta$ *and* $\widehat{\boldsymbol{H}}$ *is the optimal solution of problem* (2.5)*, then we have*

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + 3\delta\sqrt{r}. \tag{B.93}$$

Proof. First note that since $\delta \geq \max_i \|\boldsymbol{Z}_{i,\cdot}\|_2$, we have

$$\max_{i\leq n} \mathscr{D}(\boldsymbol{X}_{i,\cdot}, \mathrm{conv}(\boldsymbol{H}_0))^{1/2} \leq \max_{i\leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \delta. \tag{B.94}$$

Hence, $\boldsymbol{H}_0$ is a feasible solution for the problem (2.5). Therefore, we have

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}). \tag{B.95}$$

Letting $\tilde{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha}\in\Delta^n} \|\widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}^\mathsf{T}\boldsymbol{\alpha}\|_2$, we have

$$\begin{aligned}
\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) &= \sum_{i=1}^r \min_{\boldsymbol{\alpha}_i\in\Delta^r} \|\widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}_0^\mathsf{T}\boldsymbol{\alpha}_i - \boldsymbol{Z}^\mathsf{T}\boldsymbol{\alpha}_i\|_2^2 && \text{(B.96)}\\
&= \sum_{i=1}^r \min_{\boldsymbol{\alpha}_i\in\Delta^r} \left(\|\widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}_0^\mathsf{T}\boldsymbol{\alpha}_i\|_2^2 - 2\left\langle \boldsymbol{Z}^\mathsf{T}\boldsymbol{\alpha}_i, \widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}_0^\mathsf{T}\boldsymbol{\alpha}_i\right\rangle + \|\boldsymbol{Z}^\mathsf{T}\boldsymbol{\alpha}_i\|_2^2\right) && \text{(B.97)}\\
&= \sum_{i=1}^r \left(\|\widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}_0^\mathsf{T}\tilde{\boldsymbol{\alpha}}_i\|_2^2 - 2\left\langle \boldsymbol{Z}^\mathsf{T}\tilde{\boldsymbol{\alpha}}_i, \widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}_0^\mathsf{T}\tilde{\boldsymbol{\alpha}}_i\right\rangle + \|\boldsymbol{Z}^\mathsf{T}\tilde{\boldsymbol{\alpha}}_i\|_2^2\right). && \text{(B.98)}
\end{aligned}$$

Using the fact that (by triangle inequality) $\|\boldsymbol{Z}^\mathsf{T}\tilde{\boldsymbol{\alpha}}_i\|_2 \leq \delta$, we have

$$\begin{aligned}
\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) &\geq \sum_{i=1}^r \left(\|\widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}_0^\mathsf{T}\tilde{\boldsymbol{\alpha}}_i\|_2^2 - 2\delta\|\widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}_0^\mathsf{T}\tilde{\boldsymbol{\alpha}}_i\|_2\right) && \text{(B.99)}\\
&\geq U^2 - 2\delta\sqrt{r}U && \text{(B.100)}
\end{aligned}$$

where $U^2 = \sum_{i=1}^r \|\widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}_0^\mathsf{T}\tilde{\boldsymbol{\alpha}}_i\|_2^2$. Note that $\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) \geq 0$ and for $U \geq 2\delta\sqrt{r}$, the function $U^2 - 2\delta\sqrt{r}U$ is increasing. Hence, since

$$U \geq \left(\sum_{i=1}^r \min_{\boldsymbol{\alpha}_i} \|\widehat{\boldsymbol{H}}_{i,\cdot}^\mathsf{T} - \boldsymbol{X}_0^\mathsf{T}\boldsymbol{\alpha}_i\|_2^2\right)^{1/2} = \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2}, \tag{B.101}$$

we have

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) \geq (U^2 - 2\delta\sqrt{r}U)\mathbb{I}_{U\geq 2\delta\sqrt{r}} \geq \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0) - 2\delta\sqrt{r}\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2}. \tag{B.102}$$

Therefore,

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X})^{1/2} \geq \left(\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0) - 2\delta\sqrt{r}\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2}\right)_+^{1/2} \geq \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} - 2\delta\sqrt{r}. \tag{B.103}$$

In addition,

$$\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}) = \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^n} \|(\boldsymbol{H}_0)_{i,\cdot} - \boldsymbol{X}_0^\mathsf{T} \boldsymbol{\alpha}_i - \boldsymbol{Z}^\mathsf{T} \boldsymbol{\alpha}_i\|_2^2 \tag{B.104}$$

$$\leq \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^n} \left\{ \|(\boldsymbol{H}_0)_{i,\cdot} - \boldsymbol{X}_0^\mathsf{T} \boldsymbol{\alpha}_i\|_2 + \|\boldsymbol{Z}^\mathsf{T} \boldsymbol{\alpha}_i\|_2 \right\}^2 \tag{B.105}$$

$$\leq \sum_{i=1}^{r} \left\{ \min_{\boldsymbol{\alpha}_i \in \Delta^n} \|(\boldsymbol{H}_0)_{i,\cdot} - \boldsymbol{X}_0^\mathsf{T} \boldsymbol{\alpha}_i\|_2 + \max_{\boldsymbol{\alpha}_i \in \Delta^n} \|\boldsymbol{Z}^\mathsf{T} \boldsymbol{\alpha}_i\|_2 \right\}^2 \tag{B.106}$$

$$\leq \left\{ \left(\sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^n} \|(\boldsymbol{H}_0)_{i,\cdot} - \boldsymbol{X}_0^\mathsf{T} \boldsymbol{\alpha}_i\|_2^2 \right)^{1/2} + \delta\sqrt{r} \right\}^2 \tag{B.107}$$

$$\leq \left(\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + \delta\sqrt{r} \right)^2. \tag{B.108}$$

Hence,

$$\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X})^{1/2} \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + \delta\sqrt{r}. \tag{B.109}$$

Combining equations (B.103), (B.109), and (B.95), we get

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + 3\delta\sqrt{r}. \tag{B.110}$$

This completes the proof of lemma. □

Lemma B.6. *Let $\boldsymbol{X}_0$ be such that the uniqueness assumption holds with parameter $\alpha > 0$, and $\mathrm{conv}(\boldsymbol{X}_0)$ has internal radius at least $\mu > 0$. In particular, we have $\boldsymbol{z}_0 + \boldsymbol{U}\mathsf{B}_{r-1}(\mu) \subseteq \mathrm{conv}(\boldsymbol{X}_0)$ for $\boldsymbol{z}_0 \in \mathbb{R}^d$, and an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{d \times (r-1)}$. Finally assume $\max_{i \leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \delta$. Then for $\widehat{\boldsymbol{H}}$ the optimal solution of problem (2.5), we have*

$$\alpha(\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}) \leq 2(1+2\alpha) \left[r^{3/2} \delta \kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) + \frac{\delta\sqrt{r}}{\mu} \sigma_{\max}(\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T}) \right] + 3\delta\sqrt{r} \tag{B.111}$$

where $\boldsymbol{P}_0 : \mathbb{R}^d \to \mathbb{R}^d$ is the orthogonal projector onto $\mathrm{aff}(\boldsymbol{H}_0)$ (in particular, $\boldsymbol{P}_0$ is an affine map).

Proof. Let $\widetilde{\boldsymbol{H}}$ be such that $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\widetilde{\boldsymbol{H}})$. The uniqueness assumption implies

$$\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \geq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + \alpha\big(\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}})^{1/2}\big). \tag{B.112}$$

Note that Lemma B.5 implies

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + 3\delta\sqrt{r}. \tag{B.113}$$

Therefore,

$$\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \geq \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} - 3\delta\sqrt{r} + \alpha\big(\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}})^{1/2}\big). \tag{B.114}$$

Hence,

$$\alpha\big(\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}})^{1/2}\big) \leq \mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} - \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} + 3\delta\sqrt{r}. \tag{B.115}$$

In addition, for a convex set S, by triangle inequality we have

$$\left[\sum_{i=1}^{n} \mathscr{D}(\widehat{\boldsymbol{H}}_{i,.}, S)\right]^{1/2} - \left[\sum_{i=1}^{n} \mathscr{D}(\widetilde{\boldsymbol{H}}_{i,.}, S)\right]^{1/2} \leq \left[\sum_{i=1}^{n} \left(\mathscr{D}(\widehat{\boldsymbol{H}}_{i,.}, S)^{1/2} - \mathscr{D}(\widetilde{\boldsymbol{H}}_{i,.}, S)^{1/2}\right)^2\right]^{1/2} \tag{B.116}$$

Therefore, using

$$|\mathscr{D}(\widetilde{\boldsymbol{H}}_{i,.}, S)^{1/2} - \mathscr{D}(\widehat{\boldsymbol{H}}_{i,.}, S)^{1/2}| \leq \|\widetilde{\boldsymbol{H}}_{i,.} - \widehat{\boldsymbol{H}}_{i,.}\|_2 \tag{B.117}$$

we have

$$\left[\sum_{i=1}^{n} \mathscr{D}(\widehat{\boldsymbol{H}}_{i,.}, S)\right]^{1/2} - \left[\sum_{i=1}^{n} \mathscr{D}(\widetilde{\boldsymbol{H}}_{i,.}, S)\right]^{1/2} \leq \left[\sum_{i=1}^{n} \|\widehat{\boldsymbol{H}}_{i,.} - \widetilde{\boldsymbol{H}}_{i,.}\|_2^2\right]^{1/2} = \|\widehat{\boldsymbol{H}} - \widetilde{\boldsymbol{H}}\|_F. \tag{B.118}$$

Hence,

$$|\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} - \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2}| \leq \|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F \tag{B.119}$$

and

$$|\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} - \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2}| \leq \|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F. \tag{B.120}$$

In addition, similarly to the proof of Lemma B.5, we can write

$$\begin{aligned}
\mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}}) &= \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \|(\boldsymbol{H}_0)_{i,.} - \widetilde{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2^2 && \text{(B.121)}\\
&= \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \|(\boldsymbol{H}_0)_{i,.} - \widehat{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}_i - (\widehat{\boldsymbol{H}} - \widetilde{\boldsymbol{H}})^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2^2 && \text{(B.122)}\\
&\leq \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \left\{\|(\boldsymbol{H}_0)_{i,.} - \widehat{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2 + \|(\widehat{\boldsymbol{H}} - \widetilde{\boldsymbol{H}})^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2\right\}^2 && \text{(B.123)}\\
&\leq \sum_{i=1}^{r} \left\{\min_{\boldsymbol{\alpha} \in \Delta^r} \|(\boldsymbol{H}_0)_{i,.} - \widehat{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}\|_2 + \max_{\boldsymbol{\alpha} \in \Delta^r} \|(\widehat{\boldsymbol{H}} - \widetilde{\boldsymbol{H}})^{\mathsf{T}} \boldsymbol{\alpha}\|_2\right\}^2 && \text{(B.124)}\\
&\leq \left\{\left(\sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \|(\boldsymbol{H}_0)_{i,.} - \widehat{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2^2\right)^{1/2} + \sqrt{r} \max_{i \in [r]} \|\widehat{\boldsymbol{H}}_{i,.} - \widetilde{\boldsymbol{H}}_{i,.}\|_2\right\}^2 && \text{(B.125)}\\
&\leq \left(\mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} + \sqrt{r} \max_{i \in [r]} \|\widehat{\boldsymbol{H}}_{i,.} - \widetilde{\boldsymbol{H}}_{i,.}\|_2\right)^2. && \text{(B.126)}
\end{aligned}$$

Thus,

$$|\mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}})^{1/2} - \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}| \leq \sqrt{r} \max_{i \in [r]} \|\widetilde{\boldsymbol{H}}_{i,.} - \widehat{\boldsymbol{H}}_{i,.}\|_2 \tag{B.127}$$

Therefore, combining (B.115), (B.119), (B.120), (B.127), we get

$$\alpha\big(\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}\big) \leq (1+\alpha)\|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F + \alpha\sqrt{r} \max_{i \in [r]} \|\widetilde{\boldsymbol{H}}_{i,.} - \widehat{\boldsymbol{H}}_{i,.}\|_2 + 3\delta\sqrt{r}. \tag{B.128}$$

Now, we would like to bound the terms $\|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F$, $\max_{i\in[r]} \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2$. Note that using the fact that $\widehat{\boldsymbol{H}}$ is feasible for Problem (2.5), we have

$$\mathscr{D}(\boldsymbol{X}_{i,\cdot}, \widehat{\boldsymbol{H}}) \leq \delta^2 \,. \tag{B.129}$$

Thus,

$$\mathscr{D}((\boldsymbol{X}_0)_{i,\cdot}, \widehat{\boldsymbol{H}})^{1/2} \leq \|\boldsymbol{X}_{i,\cdot} - (\boldsymbol{X}_0)_{i,\cdot}\|_2 + \mathscr{D}(\boldsymbol{X}_{i,\cdot}, \widehat{\boldsymbol{H}})^{1/2} \leq 2\delta. \tag{B.130}$$

In addition, we know that $(\boldsymbol{X}_0)_{i,\cdot} \in \mathrm{aff}(\boldsymbol{H}_0)$, where $\mathrm{aff}(\boldsymbol{H}_0)$ is a $r-1$ dimensional affine subspace. Therefore, $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{aff}(\boldsymbol{H}_0)$ and, by convexity of $\mathsf{B}_d(2\delta, \widehat{\boldsymbol{H}})$, we get

$$\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathsf{B}_d(2\delta, \widehat{\boldsymbol{H}}) \cap \mathrm{aff}(\boldsymbol{H}_0). \tag{B.131}$$

First consider the case in which $\widehat{\boldsymbol{H}}_{i,\cdot} \in \mathrm{aff}(\boldsymbol{H}_0)$. for all $i \in \{1, 2, \ldots, r\}$. By a perturbation argument, we can assume that the rows of $\widehat{\boldsymbol{H}}$ are linearly independent, and hence $\mathrm{aff}(\widehat{\boldsymbol{H}}) = \mathrm{aff}(\boldsymbol{H}_0)$. Consider $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{r\times d}$ defined by

$$\widetilde{Q}_{ii} = 1 + \xi, \qquad \text{if } \; i = j \in \{1, 2, \ldots, r\}, \tag{B.132}$$

$$\widetilde{Q}_{ij} = -\frac{\xi}{r-1}, \quad \text{if } \; i \neq j \in \{1, 2, \ldots, r\}, \tag{B.133}$$

$$\widetilde{Q}_{ij} = 0, \qquad \text{if } \; j \in \{r+1, r+2, \ldots, d\} \tag{B.134}$$

where $\xi = 2r\delta_0$. Note that for every $\boldsymbol{y} \in \mathsf{B}_d(2\delta_0; \boldsymbol{E}_{r,d}) \cap \mathrm{aff}(\boldsymbol{E}_{r,d})$, we have $\mathscr{D}(\boldsymbol{y}, \boldsymbol{E}_{r,d})^{1/2} \leq 2\delta_0$. In addition, since $\boldsymbol{y} \in \mathrm{aff}(\boldsymbol{E}_{r,d})$, $\langle \boldsymbol{y}, \boldsymbol{1}\rangle = 1$. Hence, for $\boldsymbol{y} \in \mathsf{B}_d(2\delta_0; \boldsymbol{E}_{r,d}) \cap \mathrm{aff}(\boldsymbol{E}_{r,d})$, we can write

$$\boldsymbol{y} = \boldsymbol{\pi} + \boldsymbol{x} \tag{B.135}$$

where $\boldsymbol{\pi} \in \mathrm{conv}(\boldsymbol{E}_{r,d})$, $\boldsymbol{x} \in \mathbb{R}^d$, $\langle \boldsymbol{1}, \boldsymbol{x}\rangle = 0$, $\|\boldsymbol{x}\|_2 \leq 2\delta_0$. It is easy to check that for this $\boldsymbol{y}$ we have

$$\boldsymbol{y} = \sum_{i=1}^{r} \beta_i \widetilde{\boldsymbol{Q}}_{i,\cdot} \tag{B.136}$$

where $\boldsymbol{\beta} \in \mathbb{R}^r$ is such that for $i = 1, 2, \ldots, r$,

$$\beta_i = \frac{r-1}{r-1+\xi r}(\pi_i + x_i) + \frac{\xi}{r-1+\xi r}. \tag{B.137}$$

Further, note that since $\boldsymbol{\pi} \in \mathrm{conv}(\boldsymbol{E}_{r,d})$, $\pi_i \geq 0$ and $x_i \geq -\|\boldsymbol{x}\|_2 \geq -2\delta_0$, we have $\pi_i + x_i \geq -2\delta_0$. Hence, for $i \in \{1, 2, \ldots, r\}$,

$$\beta_i \geq \frac{-2\delta_0(r-1) + \xi}{r-1+\xi r} = \frac{2\delta_0}{r-1+\xi r} \geq 0. \tag{B.138}$$

In addition,

$$\sum_{i=1}^{r} \beta_i = \frac{r\xi}{r-1+\xi r} + \frac{r-1}{r-1+\xi r}\left(\sum_{i=1}^{r}(\pi_i + x_i)\right) = 1. \tag{B.139}$$

Therefore, every $\boldsymbol{y} \in \mathsf{B}_d(2\delta_0; \boldsymbol{E}_{r,d}) \cap \operatorname{aff}(\boldsymbol{E}_{r,d})$ can be written as a convex combination of the rows of $\widetilde{\boldsymbol{Q}}$. Hence,

$$\mathsf{B}_d(2\delta_0; \boldsymbol{E}_{r,d}) \cap \operatorname{aff}(\boldsymbol{E}_{r,d}) \subseteq \operatorname{conv}(\widetilde{\boldsymbol{Q}}). \tag{B.140}$$

Let $\widehat{\boldsymbol{H}} = \boldsymbol{E}_{r,d}\boldsymbol{M}$, $\boldsymbol{M} \in \mathbb{R}^{d\times d}$. Since $\operatorname{aff}(\widehat{\boldsymbol{H}}) = \operatorname{aff}(\boldsymbol{H}_0)$, by taking $\widetilde{\boldsymbol{H}} = \widetilde{\boldsymbol{Q}}\boldsymbol{M}$, we have

$$\begin{aligned}
\operatorname{conv}(\widetilde{\boldsymbol{H}}) &\supseteq \left[\cup_{\boldsymbol{x}\in\operatorname{conv}(\boldsymbol{E}_{r,d})} \boldsymbol{M}^{\mathsf{T}}\mathsf{B}_d(2\delta_0;\boldsymbol{x})\right] \cap \operatorname{aff}(\widehat{\boldsymbol{H}}) && \text{(B.141)}\\
&\supseteq \left[\cup_{\boldsymbol{x}\in\operatorname{conv}(\widehat{\boldsymbol{H}})} \mathsf{B}_d(2\delta_0\sigma_{\min}(\boldsymbol{M});\boldsymbol{x})\right] \cap \operatorname{aff}(\widehat{\boldsymbol{H}}) && \text{(B.142)}\\
&\supseteq \mathsf{B}_d(2\delta;\widehat{\boldsymbol{H}}) \cap \operatorname{aff}(\boldsymbol{H}_0), && \text{(B.143)}
\end{aligned}$$

provided that $\delta_0 = \delta/\sigma_{\min}(\boldsymbol{M}) = \delta/\sigma_{\min}(\widehat{\boldsymbol{H}})$. Hence, using (B.131) for this δ_0, $\operatorname{conv}(\boldsymbol{X}_0) \subseteq \operatorname{conv}(\widetilde{\boldsymbol{H}})$. Note that for $\widetilde{\boldsymbol{Q}}$, we have $\|\widetilde{\boldsymbol{Q}}_{i,\cdot} - \boldsymbol{e}_i\|_2 \le 2r\delta_0$. Thus,

$$\|\widetilde{\boldsymbol{Q}} - \boldsymbol{E}_{r,d}\|_F \le 2r^{3/2}\delta_0. \tag{B.144}$$

Therefore, there exists $\widetilde{\boldsymbol{H}} \in \mathbb{R}^{r\times d}$ such that $\operatorname{conv}(\boldsymbol{X}_0) \subseteq \operatorname{conv}(\widetilde{\boldsymbol{H}})$ and

$$\begin{aligned}
&\|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F = \|(\widetilde{\boldsymbol{Q}} - \boldsymbol{E}_{r,d})\boldsymbol{M}\|_F \le 2r^{3/2}\delta_0\sigma_{\max}(\boldsymbol{M}) = 2r^{3/2}\delta_0\sigma_{\max}(\widehat{\boldsymbol{H}}) = 2r^{3/2}\delta\kappa(\widehat{\boldsymbol{H}}),\\
&\max_{i\in[r]} \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2 = \max_{i\in[r]} \|(\widetilde{\boldsymbol{Q}}_{i,\cdot} - \boldsymbol{e}_i)\boldsymbol{M}\|_2 \le 2r\delta_0\sigma_{\max}(\boldsymbol{M}) = 2r\delta_0\sigma_{\max}(\widehat{\boldsymbol{H}}) = 2r\delta\kappa(\widehat{\boldsymbol{H}}).
\end{aligned}$$

Now consider the general case in which $\operatorname{aff}(\widehat{\boldsymbol{H}}) \neq \operatorname{aff}(\boldsymbol{H}_0)$. Let $\boldsymbol{H}' \in \mathbb{R}^{r\times d}$ be such that $\boldsymbol{H}'_{i,\cdot}$ is the projection of $\widehat{\boldsymbol{H}}_{i,\cdot}$ onto $\operatorname{aff}(\boldsymbol{H}_0)$. Assuming that the rows of $\boldsymbol{H}'$ are linearly independent, $\operatorname{aff}(\boldsymbol{H}') = \operatorname{aff}(\boldsymbol{H}_0)$. Note that since $\operatorname{conv}(\boldsymbol{X}_0) \in \operatorname{aff}(\boldsymbol{H}_0)$, for every point $\boldsymbol{x} \in \operatorname{conv}(\boldsymbol{X}_0)$, $\mathscr{D}(\boldsymbol{x}, \boldsymbol{H}')^{1/2} \le \mathscr{D}(\boldsymbol{x}, \widehat{\boldsymbol{H}})^{1/2} \le 2\delta$. Thus,

$$(\boldsymbol{X}_0)_{i,\cdot} \in \mathsf{B}_d(2\delta, \boldsymbol{H}') \cap \operatorname{aff}(\boldsymbol{H}'). \tag{B.145}$$

Therefore, using the above argument for the case where $\operatorname{aff}(\widehat{\boldsymbol{H}}) = \operatorname{aff}(\boldsymbol{H}_0)$, we can find $\widetilde{\boldsymbol{H}}$ such that $\operatorname{conv}(\boldsymbol{X}_0) \subseteq \operatorname{conv}(\widetilde{\boldsymbol{H}})$ and

$$\begin{aligned}
&\|\widetilde{\boldsymbol{H}} - \boldsymbol{H}'\|_F \le 2r^{3/2}\delta\kappa(\boldsymbol{H}'), && \text{(B.146)}\\
&\max_{i\in[r]} \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \boldsymbol{H}'_{i,\cdot}\|_2 \le 2r\delta\kappa(\boldsymbol{H}'). && \text{(B.147)}
\end{aligned}$$

Hence, for every $i = 1, 2, \ldots, r$,

$$\begin{aligned}
\|\widetilde{\boldsymbol{H}}_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2 &\le \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \boldsymbol{H}'_{i,\cdot}\|_2 + \|\boldsymbol{H}'_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2\\
&\le 2r\delta\kappa(\boldsymbol{H}') + \|\boldsymbol{P}_0(\widehat{\boldsymbol{H}}_{i,\cdot}) - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2
\end{aligned} \tag{B.148}$$

where $\boldsymbol{P}_0$ orthogonal projection onto $\operatorname{aff}(\boldsymbol{H}_0)$. We next use the assumption on the internal radius of $\operatorname{conv}(\boldsymbol{X}_0)$ to upper bound the term $\|\boldsymbol{P}_0(\widehat{\boldsymbol{H}}_{i,\cdot}) - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2$. Note that since $\operatorname{conv}(\boldsymbol{X}_0) \subseteq \mathsf{B}_d(2\delta, \widehat{\boldsymbol{H}})$, letting $\bar{\boldsymbol{H}} = \widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^{\mathsf{T}}$, for some orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{d\times r'}$, $r' = r - 1$, we have

$$\begin{aligned}
\max_{\|\boldsymbol{z}\|_2\le\mu} \min_{\langle\boldsymbol{a},\mathbf{1}\rangle=1,\boldsymbol{a}\ge 0} \|\boldsymbol{U}\boldsymbol{z} - \bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2^2 &= \max_{\|\boldsymbol{z}\|_2\le\mu} \min_{\langle\boldsymbol{a},\mathbf{1}\rangle=1,\boldsymbol{a}\ge 0} \|\boldsymbol{U}\boldsymbol{z} - (\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^{\mathsf{T}})^{\mathsf{T}}\boldsymbol{a}\|_2^2 && \text{(B.149)}\\
&\le \max_{\|\boldsymbol{z}\|_2\le\mu} \min_{\langle\boldsymbol{a},\mathbf{1}\rangle=1,\boldsymbol{a}\ge 0} \|\boldsymbol{U}\boldsymbol{z} + \boldsymbol{z}_0 - \widehat{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2^2 \le 4\delta^2. && \text{(B.150)}
\end{aligned}$$

Now, using Cauchy-Schwarz inequality we can write

$$\max_{\|\boldsymbol{z}\|_2\leq\mu}\min_{\|\boldsymbol{a}\|_2\leq 1}\|\boldsymbol{U}\boldsymbol{z}-\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2^2\leq\max_{\|\boldsymbol{z}\|_2\leq\mu}\min_{\langle\boldsymbol{a},\mathbf{1}\rangle=1,\boldsymbol{a}\geq 0}\|\boldsymbol{U}\boldsymbol{z}-\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2^2\leq 4\delta^2. \tag{B.151}$$

Note that,

$$\min_{\|\boldsymbol{a}\|_2\leq 1}\|\boldsymbol{U}\boldsymbol{z}-\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2^2=\max_{\rho\geq 0}\min_{\boldsymbol{a}}\left\{\|\boldsymbol{z}\|_2^2-2\left\langle\boldsymbol{z},\boldsymbol{U}^{\mathsf{T}}\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\right\rangle+\left\langle\boldsymbol{a},(\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^{\mathsf{T}}+\rho\mathbf{I})\boldsymbol{a}\right\rangle-\rho\right\} \tag{B.152}$$

$$=\max_{\rho\geq 0}\left\{\|\boldsymbol{z}\|_2^2-\left\langle\bar{\boldsymbol{H}}\boldsymbol{U}\boldsymbol{z},(\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^{\mathsf{T}}+\rho\mathbf{I})^{-1}\bar{\boldsymbol{H}}\boldsymbol{U}\boldsymbol{z}\right\rangle-\rho\right\} \tag{B.153}$$

Hence, using (B.151)

$$\mu^2\max_{\rho\geq 0}\left\{\lambda_{\max}(\mathbf{I}-\boldsymbol{U}^{\mathsf{T}}\bar{\boldsymbol{H}}^{\mathsf{T}}(\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^{\mathsf{T}}+\rho\mathbf{I})^{-1}\bar{\boldsymbol{H}}\boldsymbol{U})-\rho\leq 4\delta^2\right\}. \tag{B.154}$$

In particular, for $\rho=0$ we get

$$\mu^2\lambda_{\max}(\mathbf{I}-\boldsymbol{U}^{\mathsf{T}}\bar{\boldsymbol{H}}^{\mathsf{T}}(\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^{\mathsf{T}})^{-1}\bar{\boldsymbol{H}}\boldsymbol{U})\leq 4\delta^2. \tag{B.155}$$

Taking $\bar{\boldsymbol{H}}=\tilde{\boldsymbol{U}}\boldsymbol{\Sigma}\tilde{\boldsymbol{V}}^{\mathsf{T}}$, the singular value decomposition of $\bar{\boldsymbol{H}}$, we have $\sigma_{\max}(\bar{\boldsymbol{H}})=\sigma_{\max}(\widehat{\boldsymbol{H}}-\mathbf{1}\boldsymbol{z}_0^{\mathsf{T}})=\max_i\Sigma_{ii}$. Letting $\boldsymbol{U}^{\mathsf{T}}\tilde{\boldsymbol{V}}=\boldsymbol{Q}$, we get

$$\max_{\rho\geq 0}\lambda_{\max}\left(\mathbf{I}-\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}}\right)\leq\frac{4\delta^2}{\mu^2}. \tag{B.156}$$

Letting $q=\sigma_{\min}(\boldsymbol{Q})$, this results in

$$1-q^2\leq\frac{4\delta^2}{\mu^2}. \tag{B.157}$$

In addition, note that, by the internal radius assumption, for any $\boldsymbol{z}\in\mathbb{R}^{r'}$, $\boldsymbol{z}_0+\boldsymbol{U}\boldsymbol{z}\in\mathrm{aff}(\boldsymbol{H}_0)$. Further, since $\boldsymbol{z}_0\in\mathrm{aff}(\boldsymbol{H}_0)$,

$$\max_{i\in[r]}\|\boldsymbol{P}_0(\widehat{\boldsymbol{H}}_{i,\cdot})-\widehat{\boldsymbol{H}}_{i,\cdot}\|_2=\max_{i\in[r]}\|\boldsymbol{P}_{\boldsymbol{U}}(\bar{\boldsymbol{H}}_{i,\cdot})-\bar{\boldsymbol{H}}_{i,\cdot}\|_2 \tag{B.158}$$

$$\leq\max_{\|\boldsymbol{a}\|_2\leq 1}\|\boldsymbol{P}_{\boldsymbol{U}}(\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a})-\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2 \tag{B.159}$$

$$\leq\max_{\|\boldsymbol{a}\|_2\leq 1}\|\boldsymbol{P}_{\boldsymbol{U}}(\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a})-\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2 \tag{B.160}$$

$$\leq\max_{\|\boldsymbol{a}\|_2\leq 1}\min_{\boldsymbol{z}}\|\boldsymbol{U}\boldsymbol{z}-\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2^2 \tag{B.161}$$

where $\boldsymbol{P}_{\boldsymbol{U}}$ is the projector onto the column space of $\boldsymbol{U}$. Note that,

$$\max_{\|\boldsymbol{a}\|_2\leq 1}\min_{\boldsymbol{z}}\|\boldsymbol{U}\boldsymbol{z}-\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2^2=\max_{\|\boldsymbol{a}\|_2\leq 1}\left\{-\left\langle\boldsymbol{a},\bar{\boldsymbol{H}}\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\right\rangle+\left\langle\boldsymbol{a},\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\right\rangle\right\} \tag{B.162}$$

$$=\lambda_{\max}(\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^{\mathsf{T}}-\bar{\boldsymbol{H}}\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\bar{\boldsymbol{H}}^{\mathsf{T}}) \tag{B.163}$$

$$=\lambda_{\max}(\boldsymbol{\Sigma}(\mathbf{I}-\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q})\boldsymbol{\Sigma}) \tag{B.164}$$

$$\leq\sigma_{\max}(\bar{\boldsymbol{H}})^2\lambda_{\max}(\mathbf{I}-\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q}) \tag{B.165}$$

$$\leq\sigma_{\max}(\bar{\boldsymbol{H}})^2(1-q^2)\leq\frac{4\sigma_{\max}(\bar{\boldsymbol{H}})^2\delta^2}{\mu^2} \tag{B.166}$$

where the last inequality follows from (B.157). This results in

$$\max_{i\in[r]} \|\boldsymbol{P}_0(\widehat{\boldsymbol{H}}_{i,.}) - \widehat{\boldsymbol{H}}_{i,.}\|_2 \leq \frac{2\sigma_{\max}(\bar{\boldsymbol{H}})\delta}{\mu} = \frac{2\sigma_{\max}(\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T})\delta}{\mu} \tag{B.167}$$

Therefore, $\|\boldsymbol{P}_0(\widehat{\boldsymbol{H}}) - \widehat{\boldsymbol{H}}\|_F \leq 2\sigma_{\max}(\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T})\delta\sqrt{r}/\mu$. Hence, using (B.148) we get

$$\max_{i\in[r]} \|\widehat{\boldsymbol{H}}_{i,.} - \widetilde{\boldsymbol{H}}_{i,.}\|_2 \leq 2r\delta\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) + \frac{2\sigma_{\max}(\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T})\delta}{\mu}, \tag{B.168}$$

$$\|\widehat{\boldsymbol{H}} - \widetilde{\boldsymbol{H}}\|_F \leq 2r^{3/2}\delta\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) + \frac{2\sigma_{\max}(\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T})\delta\sqrt{r}}{\mu}. \tag{B.169}$$

Replacing this in (B.128) completes the proof. □

B.3.2 Proof of Theorem 2

For simplicity, let $\mathscr{D} = \alpha(\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2})$. First note that under the assumption of Theorem 2 we have

$$\boldsymbol{z}_0 + \boldsymbol{U}\mathsf{B}_{r'}(\mu) \subseteq \operatorname{conv}(\boldsymbol{X}_0) \subseteq \operatorname{conv}(\boldsymbol{H}_0). \tag{B.170}$$

Therefore, using Lemma B.4 with $\boldsymbol{H} = \boldsymbol{H}_0$ and $\delta = 0$, we have

$$\mu\sqrt{2} \leq \sigma_{\min}(\boldsymbol{H}_0) \leq \sigma_{\max}(\boldsymbol{H}_0). \tag{B.171}$$

In addition, since $\boldsymbol{z}_0 \in \operatorname{conv}(\boldsymbol{H}_0)$ we have $\boldsymbol{z}_0 = \boldsymbol{H}_0^\mathsf{T}\boldsymbol{\alpha}_0$ for some $\boldsymbol{\alpha}_0 \in \Delta^r$. Therefore,

$$\|\boldsymbol{z}_0\|_2 \leq \sigma_{\max}(\boldsymbol{H}_0)\|\boldsymbol{\alpha}_0\|_2 \leq \sigma_{\max}(\boldsymbol{H}_0). \tag{B.172}$$

Note that

$$\sigma_{\max}(\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T}) \leq \sigma_{\max}(\widehat{\boldsymbol{H}}) + \sigma_{\max}(\mathbf{1}\boldsymbol{z}_0^\mathsf{T}) = \sigma_{\max}(\widehat{\boldsymbol{H}}) + \sqrt{r}\|\boldsymbol{z}_0\|_2. \tag{B.173}$$

Therefore, using Lemma B.6 we have

$$\mathscr{D} \leq 2(1+2\alpha)\left(r^{3/2}\delta\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) + \frac{\sigma_{\max}(\widehat{\boldsymbol{H}})\delta r^{1/2}}{\mu} + \frac{r\delta\|\boldsymbol{z}_0\|_2}{\mu}\right) + 3\delta r^{1/2}. \tag{B.174}$$

In addition, Lemma B.2 implies that

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \leq \frac{1}{\alpha}\max\left\{(1+\sqrt{2})\sqrt{r}, \sqrt{2}\kappa(\boldsymbol{H}_0)\right\}\mathscr{D}\,. \tag{B.175}$$

Further, let $\boldsymbol{P}_0$ denote the orthogonal projector on $\operatorname{aff}(\boldsymbol{H}_0)$. Hence, $\boldsymbol{P}_0$ is a non-expansive mapping: for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $D(\boldsymbol{P}_0(\boldsymbol{x}), \boldsymbol{P}_0(\boldsymbol{y})) \leq D(\boldsymbol{x}, \boldsymbol{y})$. Therefore, since $\operatorname{conv}(\boldsymbol{H}_0) \subset \operatorname{aff}(\boldsymbol{H}_0)$, for any $\boldsymbol{h} \in \mathbb{R}^d$

$$\mathscr{D}(\boldsymbol{P}_0(\boldsymbol{h}), \boldsymbol{H}_0) \leq D(\boldsymbol{P}_0(\boldsymbol{h}), \boldsymbol{P}_0(\boldsymbol{\Pi}_{\operatorname{conv}(\boldsymbol{H}_0)}(\boldsymbol{h}))) \leq D(\boldsymbol{h}, \boldsymbol{\Pi}_{\operatorname{conv}(\boldsymbol{H}_0)}(\boldsymbol{h})) = \mathscr{D}(\boldsymbol{h}, \boldsymbol{H}_0). \tag{B.176}$$

Therefore,

$$\mathscr{D}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}), \boldsymbol{H}_0) \leq \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0). \tag{B.177}$$

First consider the case in which

$$\delta \leq \frac{\alpha\mu}{30\, r^{3/2}}. \tag{B.178}$$

Note that in this case $\delta \leq \mu/2$. Hence, using Lemma B.3 to upper bound $\sigma_{\max}(\widehat{\boldsymbol{H}})$, $\sigma_{\max}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))$ and Lemma B.4 to lower bound $\sigma_{\min}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))$, by (B.177), we get

$$\sigma_{\max}(\widehat{\boldsymbol{H}}) \leq \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + r^{1/2}\sigma_{\max}(\boldsymbol{H}_0) \leq \frac{\mathscr{D}}{\alpha} + r^{1/2}\sigma_{\max}(\boldsymbol{H}_0), \tag{B.179}$$

$$\begin{aligned}\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) &= \frac{\sigma_{\max}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))}{\sigma_{\min}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))} \leq \frac{\mathscr{D}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}), \boldsymbol{H}_0)^{1/2} + r^{1/2}\sigma_{\max}(\boldsymbol{H}_0)}{\sqrt{2}(\mu - 2\delta)} \\ &\leq \frac{\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + r^{1/2}\sigma_{\max}(\boldsymbol{H}_0)}{\sqrt{2}(\mu - 2\delta)} \leq \frac{\mathscr{D}}{\alpha(\mu - 2\delta)\sqrt{2}} + \frac{r^{1/2}\sigma_{\max}(\boldsymbol{H}_0)}{(\mu - 2\delta)\sqrt{2}}.\end{aligned} \tag{B.180}$$

Replacing these in (B.174) we have

$$\mathscr{D} \leq 2(1+2\alpha)\left[\frac{r^{3/2}\mathscr{D}\delta}{\alpha(\mu-2\delta)\sqrt{2}} + \frac{r^2\sigma_{\max}(\boldsymbol{H}_0)\delta}{(\mu-2\delta)\sqrt{2}} + \frac{\mathscr{D}r^{1/2}\delta}{\alpha\mu} + \frac{r\sigma_{\max}(\boldsymbol{H}_0)\delta}{\mu} + \frac{r\|\boldsymbol{z}_0\|_2\delta}{\mu}\right] + 3\delta\sqrt{r}. \tag{B.181}$$

Therefore,

$$\begin{aligned}&\mathscr{D}\left[1 - \frac{\sqrt{2}(1+2\alpha)r^{3/2}\delta}{\alpha(\mu-2\delta)} - \frac{2(1+2\alpha)r^{1/2}\delta}{\alpha\mu}\right] \\ &\qquad \leq 2(1+2\alpha)\left[\frac{r^2\sigma_{\max}(\boldsymbol{H}_0)\delta}{(\mu-2\delta)\sqrt{2}} + \frac{r\sigma_{\max}(\boldsymbol{H}_0)\delta}{\mu} + \frac{r\|\boldsymbol{z}_0\|_2\delta}{\mu}\right] + 3\delta\sqrt{r}\end{aligned} \tag{B.182}$$

Notice that condition (B.178) implies that $\mu - 2\delta \geq \mu/2$ and

$$\frac{\sqrt{2}(1+2\alpha)r^{3/2}\delta}{\alpha(\mu-2\delta)} + \frac{2(1+2\alpha)r^{1/2}\delta}{\alpha\mu} \leq \frac{1}{2}. \tag{B.183}$$

Using the previous two equations, under condition (B.178) we have

$$\begin{aligned}\mathscr{D} &\leq \frac{4(1+2\alpha)r\delta}{\mu}\left[\frac{5r\sigma_{\max}(\boldsymbol{H}_0)}{2} + \|\boldsymbol{z}_0\|_2\right] + 3\delta\sqrt{r} \\ &\leq \frac{4(1+2\alpha)r^2}{\mu}\left[\frac{5\sigma_{\max}(\boldsymbol{H}_0)}{2} + \frac{\|\boldsymbol{z}_0\|_2}{r} + \frac{3\mu}{4(1+2\alpha)r^{3/2}}\right]\delta.\end{aligned} \tag{B.184}$$

Combining this with (B.175), and using the fact that $1 + 2\alpha \leq 3$, we have under condition (B.178)

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \leq \frac{12r^2}{\mu\alpha}\left(\frac{5\sigma_{\max}(\boldsymbol{H}_0)}{2} + \frac{\|\boldsymbol{z}_0\|_2}{r} + \frac{3\mu}{4(1+2\alpha)r^{3/2}}\right)\max\left\{(1+\sqrt{2})\sqrt{r}, \sqrt{2}\kappa(\boldsymbol{H}_0)\right\}\delta \tag{B.185}$$

$$\leq \frac{29\sigma_{\max}(\boldsymbol{H}_0)r^{5/2}}{\alpha\mu}\max\left\{1, \frac{\kappa(\boldsymbol{H}_0)}{\sqrt{r}}\right\}\left(\frac{5}{2} + \frac{\|\boldsymbol{z}_0\|_2}{r\sigma_{\max}(\boldsymbol{H}_0)} + \frac{3\mu}{4(1+2\alpha)r^{3/2}\sigma_{\max}(\boldsymbol{H}_0)}\right)\delta. \tag{B.186}$$

Note that using (B.171), (B.172) and since $\alpha \geq 0$

$$\frac{\|\boldsymbol{z}_0\|_2}{r\sigma_{\max}(\boldsymbol{H}_0)} \leq 1, \qquad \frac{3\mu}{4(1+2\alpha)r^{3/2}\sigma_{\max}(\boldsymbol{H}_0)} \leq \frac{3}{4\sqrt{2}}. \tag{B.187}$$

Therefore,

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \leq \frac{120\sigma_{\max}(\boldsymbol{H}_0)r^{5/2}}{\alpha\mu} \max\left\{1, \frac{\kappa(\boldsymbol{H}_0)}{\sqrt{r}}\right\}\delta. \tag{B.188}$$

Thus,

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}) \leq \frac{C_*^2\, r^5}{\alpha^2} \max_{i\leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2^2\,, \tag{B.189}$$

where C_* is defined in Theorem 2.

Next, consider the case in which

$$\delta = \max_{i\leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \frac{\alpha\mu}{330\kappa(\boldsymbol{H}_0)r^{5/2}}, \tag{B.190}$$

Note that using (B.171),(B.172) and since $1+2\alpha \leq 3$, this condition on δ implies that

$$\delta \leq \frac{\alpha\mu\sigma_{\min}(\boldsymbol{H}_0)}{12r(1+2\alpha)(5r^{3/2}\sigma_{\max}(\boldsymbol{H}_0) + 2\|\boldsymbol{z}_0\|_2 r^{1/2} + 3\mu)}. \tag{B.191}$$

In particular, condition (B.178) holds. Hence, using equation (B.184) we get

$$\mathscr{D} \leq \frac{4(1+2\alpha)r^2}{\mu}\left[\frac{5\sigma_{\max}(\boldsymbol{H}_0)}{2} + \frac{\|\boldsymbol{z}_0\|_2}{r} + \frac{3\mu}{4(1+2\alpha)r^{3/2}}\right]\delta \leq \frac{\alpha\sigma_{\min}(\boldsymbol{H}_0)}{6\sqrt{r}}. \tag{B.192}$$

Further, note that since $\boldsymbol{P}_0$ is a projection onto an affine subspace, for $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{P}_0(\boldsymbol{x}) = \widetilde{\boldsymbol{P}_0}\boldsymbol{x} + \boldsymbol{x}_0$ for some $\widetilde{\boldsymbol{P}_0} \in \mathbb{R}^{d\times d}, \boldsymbol{x}_0 \in \mathbb{R}^d$. Hence, for any $\boldsymbol{\pi} \in \Delta^r$, $\boldsymbol{h} = \widehat{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{\pi} \in \mathrm{conv}(\widehat{\boldsymbol{H}})$, we have

$$\boldsymbol{P}_0(\boldsymbol{h}) = \widetilde{\boldsymbol{P}_0}\boldsymbol{h} + \boldsymbol{x}_0 = \widetilde{\boldsymbol{P}_0}\widehat{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{\pi} + \boldsymbol{x}_0 = \sum_{i=1}^r \pi_i\left(\widetilde{\boldsymbol{P}_0}\widehat{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{e}_i + \boldsymbol{x}_0\right) = \sum_{i=1}^r \pi_i \boldsymbol{P}_0(\widehat{\boldsymbol{h}}_i) \in \mathrm{conv}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) \tag{B.193}$$

where $\boldsymbol{e}_i$ is the i'th standard unit vector. Hence,

$$\boldsymbol{P}_0(\mathrm{conv}(\widehat{\boldsymbol{H}})) \subseteq \mathrm{conv}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})). \tag{B.194}$$

Thus, for $\boldsymbol{h}_0 \in \mathbb{R}^d$ an arbitrary row of $\boldsymbol{H}_0$, we have

$$\mathscr{D}(\boldsymbol{h}_0, \boldsymbol{P}_0(\widehat{\boldsymbol{H}})) = D(\boldsymbol{h}_0, \mathrm{conv}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))) \leq D(\boldsymbol{h}_0, \boldsymbol{P}_0(\mathrm{conv}(\widehat{\boldsymbol{H}}))) \leq D(\boldsymbol{h}_0, \boldsymbol{P}_0(\boldsymbol{\Pi}_{\mathrm{conv}(\widehat{\boldsymbol{H}})}(\boldsymbol{h}_0))). \tag{B.195}$$

In addition, using non-expansivity of $\boldsymbol{P}_0$, we have

$$D(\boldsymbol{h}_0, \boldsymbol{P}_0(\boldsymbol{\Pi}_{\mathrm{conv}(\widehat{\boldsymbol{H}})}(\boldsymbol{h}_0))) \leq D(\boldsymbol{h}_0, \boldsymbol{\Pi}_{\mathrm{conv}(\widehat{\boldsymbol{H}})}(\boldsymbol{h}_0)) = D(\boldsymbol{h}_0, \mathrm{conv}(\widehat{\boldsymbol{H}})) = \mathscr{D}(\boldsymbol{h}_0, \widehat{\boldsymbol{H}}). \tag{B.196}$$

This implies that

$$\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{P}_0(\widehat{\boldsymbol{H}})) \leq \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}). \tag{B.197}$$

Therefore, using (B.177), (B.197) and (B.192) we get

$$\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{P}_0(\widehat{\boldsymbol{H}}))^{1/2} + \mathscr{D}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}), \boldsymbol{H}_0)^{1/2} \le \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} + \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} \le \frac{\mathscr{D}}{\alpha} \le \frac{\sigma_{\min}(\boldsymbol{H}_0)}{6\sqrt{r}}. \tag{B.198}$$

Hence, in this case Lemma B.3 implies that

$$\sigma_{\max}(\widehat{\boldsymbol{H}}) \le 2\sigma_{\max}(\boldsymbol{H}_0), \tag{B.199}$$

$$\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) \le \frac{7\kappa(\boldsymbol{H}_0)}{2}. \tag{B.200}$$

Replacing this in (B.174), we have

$$\mathscr{D} \le (1+2\alpha) r^{1/2} \left(7r\delta\kappa(\boldsymbol{H}_0) + \frac{4\sigma_{\max}(\boldsymbol{H}_0)\delta + 2\sqrt{r}\|\boldsymbol{z}_0\|_2\delta}{\mu} \right) + 3\delta r^{1/2} \tag{B.201}$$

$$\le 3\delta\sqrt{r} \left(8r\kappa(\boldsymbol{H}_0) + \frac{4\sigma_{\max}(\boldsymbol{H}_0) + 2\sqrt{r}\|\boldsymbol{z}_0\|_2}{\mu} \right) \tag{B.202}$$

Hence, using (B.175) under assumption (B.190), we have

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \le 3\sqrt{r} \max\left\{ (1+\sqrt{2})\sqrt{r}, \sqrt{2}\kappa(\boldsymbol{H}_0) \right\} \left(8r\kappa(\boldsymbol{H}_0) + \frac{4\sigma_{\max}(\boldsymbol{H}_0) + 2\sqrt{r}\|\boldsymbol{z}_0\|_2}{\mu} \right) \frac{\delta}{\alpha} \tag{B.203}$$

$$\le 120 \max\left\{ 1, \frac{\kappa(\boldsymbol{H}_0)}{\sqrt{r}} \right\} \max\left\{ r\kappa(\boldsymbol{H}_0), \frac{\sigma_{\max}(\boldsymbol{H}_0) + \sqrt{r}\|\boldsymbol{z}_0\|_2}{\mu} \right\} \frac{r\delta}{\alpha}. \tag{B.204}$$

Hence, for C_*'' as defined in the statement of the theorem, we get

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \le \frac{C_*'' r}{\alpha} \max_{i \le n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \tag{B.205}$$

This completes the proof.

C Proof of Proposition 4.2

The proof follows immediately from the following two propositions.

Proposition C.1. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. Then the gradient of the function $\boldsymbol{u} \mapsto \mathscr{D}(\boldsymbol{u}, \boldsymbol{X})$ is given by*

$$\nabla_{\boldsymbol{u}} \mathscr{D}(\boldsymbol{u}, \boldsymbol{X}) = 2(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}(\boldsymbol{u}))\,. \tag{C.1}$$

Proof. Note that $\mathscr{D}(\boldsymbol{u}, \boldsymbol{X})$ is the solution of the following convex optimization problem.

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{u} - \boldsymbol{y}\|_2^2\,, \\ \text{subject to} \quad & \boldsymbol{y} = \boldsymbol{X}^\mathsf{T}\boldsymbol{\pi}, \\ & \boldsymbol{\pi} \ge 0, \\ & \langle \boldsymbol{\pi}, \mathbf{1} \rangle = 1. \end{aligned} \tag{C.2}$$

The Lagrangian for this problem is

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{\pi}, \boldsymbol{\rho}, \tilde{\rho}, \boldsymbol{\lambda}) = \|\boldsymbol{u} - \boldsymbol{y}\|_2^2 + \Big\langle \boldsymbol{\rho}, (\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{\pi}) \Big\rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\pi} \rangle + \tilde{\rho}(1 - \langle \boldsymbol{\pi}, \mathbf{1} \rangle). \tag{C.3}$$

The KKT condition implies that at the minimizer $(\boldsymbol{y}^*, \boldsymbol{\pi}^*, \boldsymbol{\rho}^*, \tilde{\rho}^*, \boldsymbol{\lambda}^*)$, we have

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{y}} = 0\,, \tag{C.4}$$

and therefore

$$\boldsymbol{\rho}^* = 2(\boldsymbol{u} - \boldsymbol{y}^*) \tag{C.5}$$

and the dual of the above optimization problem is

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{4}\|\boldsymbol{\rho}\|_2^2 + \langle \boldsymbol{\rho}, \boldsymbol{u} \rangle + \tilde{\rho}, \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq 0, \\ & \boldsymbol{X}\boldsymbol{\rho} + \tilde{\rho}\mathbf{1} + \boldsymbol{\lambda} = 0. \end{aligned} \tag{C.6}$$

Note that since (C.2) is strictly feasible, Slater condition holds and by strong duality the optimal value of (C.6) is equal to $f(\boldsymbol{u})$. Hence, we have written $f(\boldsymbol{u})$ as pointwise supremum of functions. Therefore, subgradient of $f(\boldsymbol{u})$ can be achieved by taking the derivative of the objective function in (C.6) at the optimal solution (see Section 2.10 in [MN13]). Note that the derivative of this objective function at the optimal solution is equal to $\boldsymbol{\rho}^* = 2(\boldsymbol{u} - \boldsymbol{y}^*) = 2(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}(\boldsymbol{u}))$ (where we used Eq. (C.5)). Since the dual optimum is unique (by strong convexity in $\boldsymbol{\rho}$), the function $\boldsymbol{u} \mapsto \mathscr{D}(\boldsymbol{u}, \boldsymbol{X})$ is differentiable with gradient given by Eq. (C.1). □

Proposition C.2. *Let $\boldsymbol{u} \in \mathbb{R}^d$ and $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$, and assume that the rows of $\boldsymbol{H}_0 \in \mathbb{R}^{r \times d}$ are affine independent. Then the function $\boldsymbol{H} \mapsto \mathscr{D}(\boldsymbol{u}, \boldsymbol{H})$ is differentiable at $\boldsymbol{H}_0$ with gradient*

$$\nabla_{\boldsymbol{H}} \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0) = 2\boldsymbol{\pi}_0(\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{u}) - \boldsymbol{u})^\mathsf{T}, \qquad \boldsymbol{\pi}_0 = \arg\min_{\boldsymbol{\pi} \in \Delta^r} \left\| \boldsymbol{H}_0^\mathsf{T}\boldsymbol{\pi} - \boldsymbol{u} \right\|_2^2 . \tag{C.7}$$

Proof. We will denote by $\boldsymbol{G}$ the right hand side of Eq. (C.7). For $\boldsymbol{V} \in \mathbb{R}^{r \times d}$, we have

$$\mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0 + \boldsymbol{V}) = \min_{\boldsymbol{\pi} \in \Delta^r} \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^\mathsf{T}\boldsymbol{\pi} - \boldsymbol{u} \right\|_2^2 . \tag{C.8}$$

Note that $(\boldsymbol{H}_0 + \boldsymbol{V})$ has affinely independent rows for $\boldsymbol{V}$ in a neighborhood of $\mathbf{0}$, and hence has a unique minimizer there, that we will denote by $\boldsymbol{\pi}_{\boldsymbol{V}}$. By optimality of $\boldsymbol{\pi}_{\boldsymbol{V}}$, we have

$$\begin{aligned} \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0 + \boldsymbol{V}) - \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0) &= \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^\mathsf{T}\boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{u} \right\|_2^2 - \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^\mathsf{T}\boldsymbol{\pi}_0 - \boldsymbol{u} \right\|_2^2 && \text{(C.9)} \\ &\leq \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^\mathsf{T}\boldsymbol{\pi}_0 - \boldsymbol{u} \right\|_2^2 - \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^\mathsf{T}\boldsymbol{\pi}_0 - \boldsymbol{u} \right\|_2^2 && \text{(C.10)} \\ &= \langle \boldsymbol{G}, \boldsymbol{V} \rangle + \|\boldsymbol{V}\boldsymbol{\pi}_0\|_2^2. && \text{(C.11)} \end{aligned}$$

On the other hand, by optimality of $\boldsymbol{\pi}_0$,

$$\begin{aligned} \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0 + \boldsymbol{V}) - \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0) &\geq \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^\mathsf{T}\boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{u} \right\|_2^2 - \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^\mathsf{T}\boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{u} \right\|_2^2 && \text{(C.12)} \\ &= \langle 2\boldsymbol{\pi}_{\boldsymbol{V}}(\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{u}) - \boldsymbol{u})^\mathsf{T}, \boldsymbol{V} \rangle + + \|\boldsymbol{V}\boldsymbol{\pi}_{\boldsymbol{V}}\|_2^2 && \text{(C.13)} \\ &= \langle \boldsymbol{G}, \boldsymbol{V} \rangle + 2\langle (\boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{\pi}_0)(\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{u}) - \boldsymbol{u})^\mathsf{T}, \boldsymbol{V} \rangle + \|\boldsymbol{V}\boldsymbol{\pi}_{\boldsymbol{V}}\|_2^2 . && \text{(C.14)} \end{aligned}$$

Letting $R(\boldsymbol{V}) = |\mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0 + \boldsymbol{V}) - \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0) - \langle \boldsymbol{G}, \boldsymbol{V} \rangle|$ denote the residual, we get

$$\frac{R(\boldsymbol{V})}{\|\boldsymbol{V}\|_F} \leq \|\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{u}) - \boldsymbol{u}\|_2 \|\boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{\pi}_0\|_2 + \|\boldsymbol{V}\|_F (\|\boldsymbol{\pi}_{\boldsymbol{V}}\|_2 + \|\boldsymbol{\pi}_0\|_2)\,. \tag{C.15}$$

Note that $\boldsymbol{\pi}_{\boldsymbol{V}}$ must converge to $\boldsymbol{\pi}_0$ as $\boldsymbol{V} \to 0$ because $\boldsymbol{\pi}_0$ is the unique minimizer for $\boldsymbol{V} = \boldsymbol{0}$. Hence we get $R(\boldsymbol{V})/\|\boldsymbol{V}\|_F \to 0$ as $\|\boldsymbol{V}\|_F \to 0$, which proves our claim. □

D Proof of Proposition 4.1

We use the results of [BST14] to prove Proposition 4.1. We refer the reader to [BST14] for the definitions of the technical terms in this section. First, consider the function

$$f(\boldsymbol{H}) = \lambda \mathscr{D}(\boldsymbol{H}, \boldsymbol{X}). \tag{D.1}$$

Note that using the main theorem of polytope theory (Theorem 1.1 in [Zie12]), we can write

$$\mathrm{conv}(\boldsymbol{X}) = \left\{ \boldsymbol{x} \in \mathbb{R}^d \,|\, \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \leq b_i \;\text{ for }\; 1 \leq i \leq m \right\} \tag{D.2}$$

for some $\boldsymbol{a}_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ and a finite m. Hence, using the definition of the semi-algebraic sets (see Definition 5 in [BST14]), the set $\mathrm{conv}(\boldsymbol{X})$ is semi-algebraic. Therefore, the function $f(\boldsymbol{H})$ which is proportional to the sum of squared ℓ_2 distances of the rows of $\boldsymbol{H}$ from a semi-algebraic set, is a semi-algebraic function (See Appendix in [BST14]). Further, the function

$$g(\boldsymbol{W}) = \sum_{i=1}^{n} \mathbf{I}\left(\boldsymbol{w}_i \in \Delta^r\right) \tag{D.3}$$

is the sum of indicator functions of semi-algebraic sets (Note that using the same argument used for $\mathrm{conv}(\boldsymbol{X})$, Δ^r is semi-algebraic). Therefore, the function g is semi-algebraic (See Appendix in [BST14]). In addition, the function

$$h(\boldsymbol{H}, \boldsymbol{W}) = \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 \tag{D.4}$$

is a polynomial. Hence, it is semi-algebraic. Therefore, we deduce that the function

$$\Psi(\boldsymbol{H}, \boldsymbol{W}) = f(\boldsymbol{H}) + g(\boldsymbol{W}) + h(\boldsymbol{H}, \boldsymbol{W}) \tag{D.5}$$

is semi-algebraic. In addition, since Δ^r is closed, Ψ is proper and lower semi-continuous. Therefore, $\Psi(\boldsymbol{H}, \boldsymbol{W})$ is a KL function (See Theorem 3 in [BST14]).

Now, we will show that the Assumptions 1,2 in [BST14] hold for our algorithm. First, note that since Δ^r is closed, the functions $f(\boldsymbol{H})$ and $g(\boldsymbol{W})$ are proper and lower semi-continuous. Further, $f(\boldsymbol{H}) \geq 0$, $g(\boldsymbol{W}) \geq 0$, $h(\boldsymbol{H}, \boldsymbol{W}) \geq 0$ for all $\boldsymbol{H} \in \mathbb{R}^{r \times d}$, $\boldsymbol{W} \in \mathbb{R}^{n \times r}$. In addition, the function $h(\boldsymbol{H}, \boldsymbol{W})$ is C^2. Therefore, it is Lipschitz continuous over the bounded subsets of $\mathbb{R}^{r \times d} \times \mathbb{R}^{n \times r}$. Also, the partial derivatives of $h(\boldsymbol{H}, \boldsymbol{W})$ are

$$\nabla_{\boldsymbol{H}} h(\boldsymbol{H}, \boldsymbol{W}) = 2\boldsymbol{W}^{\mathsf{T}}(\boldsymbol{W}\boldsymbol{H} - \boldsymbol{X}), \tag{D.6}$$

$$\nabla_{\boldsymbol{W}} h(\boldsymbol{H}, \boldsymbol{W}) = 2(\boldsymbol{W}\boldsymbol{H} - \boldsymbol{X})\boldsymbol{H}^{\mathsf{T}}. \tag{D.7}$$

It can be seen that for any fixed $\boldsymbol{W}$, the function $\boldsymbol{H} \mapsto \nabla_{\boldsymbol{H}} h(\boldsymbol{H}, \boldsymbol{W})$ is Lipschitz continuous with moduli $L_1(\boldsymbol{W}) = 2\|\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W}\|_F$. Similarly, for any fixed $\boldsymbol{H}$, the function $\boldsymbol{W} \mapsto \nabla_{\boldsymbol{W}} h(\boldsymbol{H}, \boldsymbol{W})$ is

Lipschitz continuous with moduli $L_2(\boldsymbol{H}) = 2\|\boldsymbol{H}\boldsymbol{H}^\mathsf{T}\|_F$. Note that since in each iteration of the algorithm the rows of $\boldsymbol{W}^k$ are in Δ^r. Hence,

$$\inf\left\{L_1(\boldsymbol{W}^k) : k \in \mathbb{N}\right\} \geq \lambda_1^- \qquad \sup\left\{L_1(\boldsymbol{W}^k) : k \in \mathbb{N}\right\} \leq \lambda_1^+ \tag{D.8}$$

for some some positive constants λ_1^-, λ_1^+. In addition, note that because the PALM algorithm is a descent algorithm, i.e., $\Psi(\boldsymbol{H}^k, \boldsymbol{W}^k) \leq \Psi(\boldsymbol{H}^{k-1}, \boldsymbol{W}^{k-1})$ for $k \in \mathbb{N}$, and since $f(\boldsymbol{H}) \to \infty$ as $\|\boldsymbol{H}\|_F \to \infty$, the value of $L_2(\boldsymbol{H}^k) = \|\boldsymbol{H}^k\boldsymbol{H}^{k^\mathsf{T}}\|_F$ remains bounded in every iteration. Finally, note that by taking $\gamma_2^k > \max\left\{\left\|\boldsymbol{H}^{k+1}\boldsymbol{H}^{k+1^\mathsf{T}}\right\|_F, \varepsilon\right\}$ for some constant $\varepsilon > 0$, we make sure that the steps in the PALM algorithm remain well defined (See Remark 3(iii) in [BST14]). Hence, we have shown that the assumptions of Theorem 1 in [BST14] hold. Therefore, using this theorem, the sequence $\left\{\boldsymbol{H}^k, \boldsymbol{W}^k\right\}_{k\in\mathbb{N}}$ generated by the iterations in (4.7) - (4.9) has a finite length and it converges to a stationary point $(\boldsymbol{H}^*, \boldsymbol{W}^*)$ of Ψ.

E Other optimization algorithms

Apart from the proximal alternating linearized minimization discussed in Section 4.2, we experimented with two other algorithms, obtaining comparable results. For the sake of completeness, we describe these algorithms here.

E.1 Stochastic gradient descent

Using any of the initializations discussed in Section 4.1 we iterate

$$\boldsymbol{H}^{(t+1)} = \boldsymbol{H}^{(0)} - \gamma_t \boldsymbol{G}^{(t)}\,. \tag{E.1}$$

The step size γ_t is selected by backtracking line search. Ideally, the direction $\boldsymbol{G}^{(t)}$ can be taken to be equal to $\nabla\mathscr{R}_\lambda(\boldsymbol{H}^{(t)})$. However, for large datasets this is computationally impractical, since it requires to compute the projection of each data point onto the set $\operatorname{conv}(\boldsymbol{H}^{(t)})$. In order to reduce the complexity of the direction calculation, we estimate this sum by subsampling. Namely, we draw a uniformly random set $S_t \subseteq [n]$ of fixed size $|S_t| = s \leq n$, and compute

$$\boldsymbol{G}^{(t)} = \frac{2n}{|S_t|}\sum_{i\in S_t} \boldsymbol{\alpha}_i^* \left(\boldsymbol{\Pi}_{\operatorname{conv}(\boldsymbol{H})}(\boldsymbol{x}_i) - \boldsymbol{x}_i\right) + 2\lambda\left(\boldsymbol{H} - \boldsymbol{\Pi}_{\operatorname{conv}(\boldsymbol{X})}(\boldsymbol{H})\right)\,, \tag{E.2}$$

$$\boldsymbol{\alpha}_i^* = \arg\min_{\boldsymbol{\alpha}\in\Delta^r} \left\|\boldsymbol{H}^\mathsf{T}\boldsymbol{\alpha} - \boldsymbol{x}_i^\mathsf{T}\right\|_2\,. \tag{E.3}$$

E.2 Alternating minimization

This approach generalizes the original algorithm of [CB94]. We rewrite the objective as a function of $\boldsymbol{W} = (bw_i)_{i\leq n}$, $\boldsymbol{w}_i \in \Delta^r$,$\boldsymbol{H} = (\boldsymbol{h}_i)_{i\leq r}$, $\boldsymbol{h}_i \in \mathbb{R}^d$ and $\boldsymbol{A} = (\boldsymbol{\alpha}_\ell)_{\ell\leq r}$, $\boldsymbol{\alpha}_\ell \in \Delta^n$

$$\mathscr{R}_\lambda(\boldsymbol{H}) = \min_{\boldsymbol{W},\boldsymbol{A}} F(\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{A})\,, \tag{E.4}$$

$$F(\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{A}) = \sum_{i=1}^n \left\|\boldsymbol{x}_i - \sum_{\ell=1}^r w_{i\ell}\boldsymbol{h}_\ell\right\|_2^3 + \lambda\sum_{\ell=1}^r \left\|\boldsymbol{h}_\ell - \sum_{i=1}^n \alpha_{\ell,i}\boldsymbol{x}_i\right\|_2^2. \tag{E.5}$$

The algorithm alternates between minimizing with respect to the weights $(\boldsymbol{w}_i)_{i\leq n}$ (this can be done independently across $i \in \{1, \ldots, n\}$) and minimizing over $(\boldsymbol{h}_\ell, \boldsymbol{\alpha}_\ell)$, which is done sequentially by

cycling over $\ell \in \{1, \dots, r\}$. Minimization over $\boldsymbol{w}_i$ can be performed by solving a non-negative least squares problem. As shown in [CB94], minimization over $(\boldsymbol{h}_\ell, \boldsymbol{\alpha}_\ell)$ is also equivalent to non-negative least squares. Indeed, by a simple calculation

$$F(\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{A}) = w_\ell^{\rm tot} \big\| \boldsymbol{h}_\ell - \boldsymbol{v}_\ell \big\|_2^2 + \lambda \Big\| \boldsymbol{h}_\ell - \sum_{i=1}^n \alpha_{\ell,i} \boldsymbol{x}_i \Big\|_2^2 + \widetilde{F}(\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{A}) \tag{E.6}$$

$$= f_\ell(\boldsymbol{h}_\ell, \boldsymbol{\alpha}_\ell; \boldsymbol{H}_{\neq \ell}, \boldsymbol{W}, \boldsymbol{A}) + \widetilde{F}(\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{A}) \, . \tag{E.7}$$

where $\boldsymbol{H}_{\neq \ell} = (\boldsymbol{h}_i)_{i \neq \ell, i \leq r}$, $\widetilde{F}(\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{A})$ does not depend on $(\boldsymbol{h}_\ell, \boldsymbol{\alpha}_\ell)$, and we defined

$$w_\ell^{\rm tot} \equiv \sum_{i=1}^n w_{i\ell}^2 \, , \tag{E.8}$$

$$\boldsymbol{v}_\ell \equiv \frac{1}{w_\ell^{\rm tot}} \sum_{i=1}^n w_{i,\ell} \left\{ \boldsymbol{x}_i - \sum_{j \neq \ell, j \leq r} w_{ij} \boldsymbol{h}_j \right\} . \tag{E.9}$$

It is therefore sufficient to minimize $f_\ell(\boldsymbol{h}_\ell, \boldsymbol{\alpha}_\ell; \boldsymbol{H}_{\neq \ell}, \boldsymbol{W}, \boldsymbol{A})$ with respect to its first two arguments, which is equivalent to a non-negarive least squares problem. This can be seen by minimizing $f_\ell(\cdots)$ explicitly with respect to $\boldsymbol{h}_\ell$ and writing the resulting objective function.

The pseudocode for this algorithm is given below.

ALTERNATING MINIMIZATION

Input : Data $\{\boldsymbol{x}_i\}_{i \leq n}$, $\boldsymbol{x}_i \in \mathbb{R}^d$; integer r; initial archetypes $\{\boldsymbol{h}_\ell^{(0)}\}_{1 \leq \ell \leq r}$; number of iterations T;
Output : Archetype estimates $\{\boldsymbol{h}_\ell^{(T)}\}_{1 \leq \ell \leq r}$;

For $\ell \in \{1, \dots, r\}$:
    Set $\boldsymbol{\alpha}_\ell^{(0)} = \arg\min_{\boldsymbol{\alpha} \in \Delta^n} \|\boldsymbol{h}_\ell^{(0)} - \boldsymbol{X}\boldsymbol{\alpha}_\ell\|_2$;
For $t \in \{1, \dots, T\}$:
    Set $\boldsymbol{W}^t = \arg\min_{\boldsymbol{W}} F(\boldsymbol{H}^{t-1}, \boldsymbol{W}, \boldsymbol{A}^{t-1})$
    For $\ell \in \{1, \dots, r\}$:
    Set $\boldsymbol{h}_\ell^{(t)}, \boldsymbol{\alpha}_\ell^{(t)} = \arg\min_{\boldsymbol{h}_\ell, \boldsymbol{\alpha}_\ell} f_\ell(\boldsymbol{h}_\ell, \boldsymbol{\alpha}_\ell; \boldsymbol{H}_{<\ell}^t, \boldsymbol{H}_{>\ell}^{t-1}, \boldsymbol{W}^t, \boldsymbol{A}_{<\ell}^t, \boldsymbol{A}_{>\ell}^{t-1})$;
End For;
Return $\{\hat{\boldsymbol{h}}_\ell^{(T)}\}_{1 \leq \ell \leq r}$

Here $\boldsymbol{H}_{<\ell} = (\boldsymbol{h}_i)_{i<\ell}$, $\boldsymbol{H}_{>\ell} = (\boldsymbol{h}_i)_{\ell<i\leq r}$, and similarly for $\boldsymbol{A}$.