Finding k in Latent k-Polytope

Chiranjib Bhattacharyya [*1] **Ravi Kannan** [*2] **Amit Kumar** [*3]

Abstract

The recently introduced Latent $k-$ Polytope(LkP) encompasses several stochastic Mixed Membership models including Topic Models. The problem of finding k, the number of extreme points of LkP, is a fundamental challenge and includes several important open problems such as determination of number of components in Ad-mixtures. This paper addresses this challenge by introducing Interpolative Convex Rank(ICR) of a matrix defined as the minimum number of its columns whose convex hull is within Hausdorff distance ε of the convex hull of all columns. The first important contribution of this paper is to show that under *standard assumptions* k equals the ICR of a *subset smoothed data matrix* defined from Data generated from an LkP. The second important contribution of the paper is a polynomial time algorithm for finding k under standard assumptions. An immediate corollary is the first polynomial time algorithm for finding the *inner dimension* in Non-negative matrix factorisation(NMF) with assumptions which are qualitatively different than existing ones such as *Separability*.

1. Introduction

Latent variable models have found large number of applications in the real world. Such models specify a generative process between the observed variables and a set of underlying un-observed variables, often called Latent variables. Examples of latent variable models include Clustering like k-means, Finite Mixture models, Topic models (LDA), and Stochastic block models. In such models learning the parameters of the generative process is often intractable and remains an active area of research.

[1]Department of Computer Science and Automation, IISc Bangalore, India [2]Microsoft Research India Lab., Bangalore, India [3]Department of Computer Science and Engineering, IIT Delhi, India. Correspondence to: Chiranjib Bhattacharya <chiru@iisc.ac.in>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

(Bhattacharyya & Kannan, 2020) suggested Latent k polytope(LkP), an unifying model which includes all of the above mentioned instances of latent variable modelis as special cases. LkP defines a latent polytope K on k vertices and data is generated by first picking *latent points* from this polytope, denoted by columns of a matrix $\mathbf{P}$, and then perturbing them in an adversarial manner to produce data-points. The perturbations can be quite large in magnitude, and the data points can lie quite far from the actual polytope K. The problem of recovering (vertices of) K from data generated in this manner is in general hard, unless we make assumptions about K and the data matrix $\mathbf{A}$, each of whose columns represents one data point. They crystallized a set of assumptions, which will be referred in the sequel as *standard assumptions*, and showed that the vertices of the polytope K can be provably recovered to an ε approximation from a data Matrix $\mathbf{A}$ satisfying the assumptions.

The algorithm generally applies to all special cases and results in Model estimation with guaranteed approximation from a finite sample of data. However, as is usual practice in Latent Variable models the value of k is assumed known. For example, most algorithms for $k-$ Means or LDA with k topics would require the value of k.

In this paper we rely on the geometry of the problem to find k from the data Matrix $\mathbf{A}$. Straightforward attempts such as examining the rank etc of $\mathbf{A}$ are fruitless because most of the points in $\mathbf{A}$ can be very far from the actual polytope K. We use two key insights here: (i) Suppose first we have access to the latent points matrix $\mathbf{P}$ instead of $\mathbf{A}$. In this case, the convex hull of $\mathbf{P}$, $\mathsf{CH}(\mathbf{P})$, would be a close approximation of K. Still it is not clear how to identify k from the $\mathsf{CH}(\mathbf{P})$. (ii) Such a strategy is still not practical as $\mathbf{P}$ is not available. However the eigen structure of $\mathbf{A}$ can provide important clues.

Contributions: This paper addresses some of these challenges and a summary of contributions are listed below.

- The paper introduces the notion of Interpolative Convex Rank(ICR) of a matrix, and shows that $k =$ ICR of a *subset smoothed data matrix* where k is the number of vertices in LkP (see details in Theorem 1). The notion of ICR should be of independent interest.
- The paper introduces new techniques based on the hy-

perplane separator theorem for proving lower bounds on the ICR of a matrix (see details in Section 3.4).

- Under standard assumptions, the paper gives a polynomial time algorithm for finding the correct value of the number of vertices of the polytope K (see Theorem 18). We do this by using the familiar singular value thresholding method, but from a different perspective than commonly used in numerical analysis.
- In this paper we show that Non-negative Matrix factorization(NMF) is a special case of Latent-k-polytope. In Theorem 20 we show that the algorithm recovers the correct inner dimension for NMF under standard assumptions.

2. Preliminaries and Problem Definition

Notation: For any natural number n we will denote $[n] = \{1, \ldots, n\}$. The points will lie in the d-dimensional Euclidean space $\Re^d$. For a matrix $\mathbf{B}$ and index ℓ, we shall use $B_{\cdot,\ell}$ to denote the ℓ^{th} column of $\mathbf{B}$. Given a subset S of columns of $\mathbf{B}$, the notation $\mathbf{B}_{\cdot,S}$ refers to $\frac{1}{|S|}\sum_{\ell \in S} B_{\cdot,\ell}$. The spectral norm of $\mathbf{B}$ will be denoted by $||\mathbf{B}||$. The 2-norm (i.e., Euclidean norm) of a vector v is denoted by $|v|$. The r^{th} singular value of a matrix $\mathbf{B}$ is denoted by $s_r(\mathbf{B})$. The convex hull of a set of points $\mathbf{x}_i \in \Re^d, i \in [m]$, is denoted by $\mathsf{CH}(\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}) = \{\sum_{i=1}^m \alpha_i \mathbf{x}_i | 0 \leq \alpha_i, \sum_{i=1}^m \alpha_i = 1\}$. We abuse notation and use $\mathsf{CH}(\mathbf{B})$, where $\mathbf{B}$ is a matrix, to denote the convex hull of the columns of $\mathbf{B}$. An extreme point of a convex set, $\mathbf{C}$, is any point, $\mathbf{x} \in \mathbf{C}$ such that it cannot be expressed as $\mathbf{x} = \beta\mathbf{x}_1 + (1-\beta)\mathbf{x}_2, 0 \leq \beta \leq 1, \mathbf{x}_1 \neq \mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_2 \in \mathbf{C}$.

Definition 1. *Given two sets $S, T \subseteq \Re^d$, the Hausdorff distance* $\mathsf{dist}(S,T)$ *between them is defined as*

$$\mathsf{dist}(S,T) := \max\left(\max_{x \in S} d(x,T), \max_{y \in T} d(y,S)\right),$$

where $d(z,A)$, for a point z and a set $A \subseteq \Re^d$, is defined as $\min_{t \in A} |z - t|$.

2.1. Review of LkP and Problem statement

In this subsection we introduce relevant definitions and present the problem statement.

2.1.1. Review of Latent-k-Polytope

In this subsection we briefly review the Latent-k-polytope introduced in (Bhattacharyya & Kannan, 2020).

Definition 2. *Let the columns of $\mathbf{M} \in \Re^{d \times k}$ be the extreme points of the Latent k Polytope, $\mathrm{K} = \mathsf{CH}(\mathbf{M})$. The data matrix $\mathbf{A} \in \Re^{d \times n}$ is said to be generated from* LkP *with parameters $\mathbf{M}, \sigma$, $\sigma > 0$, if there exists a $d \times n$ matrix $\mathbf{P}$, whose columns lie in* K*, such that $\sigma = \frac{\|\mathbf{A}-\mathbf{P}\|}{\sqrt{n}}$.*

For each $j \in [n]$, $A_{\cdot,j}$, a column of the data-matrix as the perturbation of the *latent* data point $P_{\cdot,j}$. The parameter σ, referred in the sequel as *perturbation parameter*, can be thought of as a measure of the maximum mean squared perturbation in any direction. In previous work (Bhattacharyya & Kannan, 2020) the problem of learning $\mathbf{M}$ from $\mathbf{A}$ was addressed. In general this is an intractable problem but under certain standard assumptions stated in the next subsection, a polynomial time algorithm was given for recovering $\mathbf{M}$ to an additive error. The dimension of $\mathbf{M}$, i.e. the value of k, was assumed to be known.

2.1.2. Standard Assumptions

(Bhattacharyya & Kannan, 2020) defined a set of assumptions on LkP which will be referred in the sequel as *Standard Assumptions*. A LkP model with parameters $\mathbf{M}, \sigma$, and the data natrix $\mathbf{A}$ is said to satisfy the standard assumptions if the following are true.

- **Separation**: Every vertex of K is far from the affine or the convex hull of the remaining vertices of K.
- **Proximity**: There is a parameter $\delta \in (0,1)$ such that for every vertex of K, there are at least δn points (columns of $\mathbf{A}$) whose corresponding latent points (columns in $\mathbf{P}$) are close to the vertex.
- **Spectrally Bounded Perturbations**: The parameter $\sigma\ (= ||A - P||/\sqrt{n})$ is much smaller than the size of K.

The first assumption is on the model,$\mathbf{M}$, the second is on observed data,$\mathbf{A}$, and last assumption is on the perturbation parameter, σ. In the sequel these assumptions will be stated more precisely. We prove that the set of standard assumptions (see Theorem 1) imply that there are k subsets of data each of cardinality δn whose averages nearly contain in their convex hull the averages of all δn sized subsets of data.This observation will be leveraged to give a data driven characterization of k. It is to be noted that the oft used notion of *separability* in Non-negative matrix Factorization(see (Gillis & Luce, 2018)) assumes that there are k actual data points whose hull nearly contains all data points. This is restrictive and does not hold even for simple cases of LkP.

2.1.3. Problem Definiton

The objective of the paper is to address the problem of finding k under the above mentioned assumptions.

Definition 3. *Given* $\mathbf{A}$ *generated from* LkP *and δ, find the number of extreme points of* K.

Note that $\sigma, k, \mathbf{P}$ and $\mathbf{M}$ are not given as input parameters. Only data, $\mathbf{A}$, is available. In (Bhattacharyya & Kannan,

2020) the emphasis was on computing $\mathbf{M}$ given k. In contrast this paper attempts to address the problem of finding k from $\mathbf{A}$ under standard assumptions.

2.2. Non-negative Matrix factorization(NMF)

LkP is an interesting geometric alternative to several Unsupervised learning paradigms such as Topic Models, Clustering and Community Detection (Bhattacharyya & Kannan, 2020). An important aim of this paper is to explore NMF as a special case of LkP.

Over the last two decades NMF have attracted significant research interest. In this section we do not review the vast literature on NMF but point the reader to many readable surveys(e.g. (Gillis, 2014)). However, to aid further discussion the definition of NMF is briefly reviewed.

Exact NMF seeks a factorization of the data matrix $\mathbf{A}$, $\mathbf{A}_{d\times n} = \mathbf{M}_{d\times k}\mathbf{W}_{k\times n}$ where n is the number of datapoints,d is the ambient data dimension and all entries of $\mathbf{M}, \mathbf{W}, \mathbf{A}$ are assumed to be positive. Each column of $\mathbf{M}$ represents a feature and each column of $\mathbf{W}$ denotes the corresponding weights on the individual features. The parameter k is often called *inner-dimension* (Vavasis, 2009). It is possible that Exact NMF may not exist and bulk of the research in NMF has been directed towards understanding the case of approximate NMF defined as follows

$$\mathbf{A}_{d\times n} = \mathbf{M}_{d\times k}\mathbf{W}_{k\times n} + \mathbf{N}. \qquad \texttt{(Approx-NMF)}$$

Here again $\mathbf{W}, \mathbf{M}$ are assumed to have *non-negative* entries. Each column of $\mathbf{N}$ can be thought of as perturbation or noise.

The Inexact NMF is a special case of Latent-k-Polytope with the requirement that each column of $\mathbf{W}$ should sum to 1. Each datapoint, $\mathbf{A}_{\cdot,j}$ is a perturbation of a point $\mathbf{P}_{\cdot,j} = \sum_{l=1}^{r} w_{jl}\mathbf{M}_{\cdot,l}$, lying on the polytope $\mathrm{K} = \mathsf{CH}(\mathbf{M})$. The paramater σ depends on the norm of the matrix $\mathbf{N}$. It is straightforward to see that if we are give data-matrix $\mathbf{A}$, which satisfies the standard assumptions, and the inner dimension r, the algorithm mentioned in (Bhattacharyya & Kannan, 2020) directly applies and one can recover the features.

Our goal in this paper is to recover the *inner-dimension* for NMF.

2.3. Related Literature: finding the number of vertices of a Latent Polytope

To the best of our knowledge the problem of finding the number of vertices of the latent polytope has not been studied. The problem of enumerating all the vertices of a polytope, described by linear inequalites, is known to be hard (Khachiyan et al., 2008). The problem described here is much more complicated as we have access only to some perturbed points. Since LkP is a geometric construct it maybe difficult to adopt probabilistic approaches such as Hierarchical Dirichlet Process (Teh et al., 2006) which have shown empirical success in many applications. However, to the best of our understanding it does not provide a guarantee for finding the correct k, the number of extreme points of K, from a finite number of samples.

3. Finding the number of vertices of LkP from data: An interpolative approach

Finding the number of vertices of K, the Latent k Polytope, defined in Definition 2 is in general a hard problem. Many open problems of active research such as finding the number of topics in a topic model are important special cases of this problem. The challenges in addressing the issue of determing the number of vertices can be distilled into two following issues.

Firstly, one needs to establish a suitable quantity which relates the observed data $\mathbf{A}$ to the columns of $\mathbf{M}$. Existing attempts in special cases (e.g.use of Hierarchical Dirichlet processes for Topic models(Teh et al., 2006)) do not translate into a finite-time algorithm which guarantees the recovery of the correct number of vertices. A well-formulated quantity which can yield tractable algorithmic insights under realistic assumptions is important aim to have. Secondly, even if such a quantity is devised, the overall goal of finding a tractable algorithm is a harder challenge.

In this paper we address both these issues by appealing to convex geometry. It involves conceptualizing a new notion of Matrix rank (see Subsection 3.1), and following which we develop a polynomial time algorithm(see Subsection 3.2).

3.1. Interpolative Convex Rank (ICR) of a Matrix

In this section the concept of Interpolative Convex Rank (ICR) of a matrix is introduced.

Definition 4. *Given a parameter $\varepsilon > 0$, the Interpolative Convex Rank (ICR) of a matrix $\mathbf{C} \in \Re^{d\times m}$ denoted $\texttt{ICR}(\varepsilon, \mathbf{C}) = l$, where l is the minimum number such that there exits columns $\mathbf{C}_{\cdot,i_1}, \ldots, \mathbf{C}_{\cdot,i_l}$ of $\mathbf{C}$ such that*

$$\mathsf{dist}(\mathsf{CH}(\mathbf{C}), \mathsf{CH}(\mathbf{C}_{\cdot,i_1}, \ldots, \mathbf{C}_{\cdot,i_l})) \le \varepsilon$$

In words it is the minimum number of columns of $\mathbf{C}$ whose convex hull is ε close the convex hull of all columns of $\mathbf{C}$. A geometric insight behind the definition can be obtained by observing that if $\varepsilon = 0$ then ICR of $\mathbf{C}$ is equal to the number of *extreme points* of $\mathsf{CH}(\mathbf{C})$. For any $\varepsilon > 0$, $\texttt{ICR}(\varepsilon, \mathbf{C})$ is the minimum number of vertices drawn from the columns of $\mathbf{C}$ whose convex hull approximates the convex hull of all the columns of $\mathbf{C}$ with error at most ε. Such approximations have been studied with different notions of distance in Convex Geometry (Barvinok, 2014; Gruber, 1993).

For any non zero ε, the quantity ICR rank is the smallest number of extreme points of a polytope whose convex hull approximates the polytope.

One of the goals of this paper is to relate NMF to LkP. We argue that Non-Negative Rank (NNR), a often used measure for number of factors, may not be suited and ICR is a better alternative. To support this claim we first recall the definition of NNR((Vavasis, 2009)); we define it through the *inner dimension*.

Definition 5. *The* **non-negative rank** *(NNR) of a matrix is the minimum inner dimension of a non-negative factorization of the matrix.*

There is no known provable method for identifying the correct *inner dimension*; it is mostly determined by heuristics (see Section 3 in (Gillis, 2014)). It is folklore that (under some conditions) NNR reflects or is equal to the correct inner dimension. However, in many cases NNR can be very different than inner-dimension. In the Supplememt this is illustrated with several examples where ICR recovers the true inner-dimension while NNR could be much lower. The intuitive reason for the contrast between NNR and ICR in these examples is that the interpolative property forces ICR to restrict attention to the columns of the given data matrix, but NNR does not have such restriction and hence can choose suitable but arbitrary points not from the data-matrix to yield a low value of NNR.

The use of interpolative assumption for NMF is not new and is mentioned in (Recht et al., 2012). However the algorithms assume that the value of inner-dimension is known and using ICR to characterize the inner-dimension is new.

3.2. Recovering the number of vertices of LkP from ICR

In this section, we prove the result that the parameter k can be recovered from the data matrix $\mathbf{A}$ as ICR of a suitably smoothed data matrix $\widetilde{\mathbf{A}}$ obtained from $\mathbf{A}$. As pointed out earlier, the raw data matrix $\mathbf{A}$ does not yield k since $\mathsf{CH}(\mathbf{A})$ can be much larger than K.

We now define the subset smoothed version of $\mathbf{A}$.

Definition 6. (Subset smoothed data matrix) *Let $\mathcal{R}_\delta$ denote the collection of all subsets $R \subseteq [n], |R| = \delta n$. Then the subset smoothed data matrix $\widetilde{\mathbf{A}}$ has $|\mathcal{R}_\delta|$ columns indexed by the elements in $\mathcal{R}_\delta$. For each $R \in \mathcal{R}_\delta$, the column $\widetilde{\mathbf{A}}_{\cdot,R}$ of $\widetilde{\mathbf{A}}$ is same as $A_{\cdot,R}$, i.e., $\frac{1}{|R|}\sum_{\ell \in R} A_{\cdot,\ell}$.*

We now state the main result of this section.

Theorem 1. *Assume that the input data matrix $\mathbf{A}$ is generated by* LkP *(as in Definition 2). Let $\rho > 0, \delta \in (0,1)$ and suppose the following three assumptions are satisfied*

1. *Every vertex $M_{\cdot,\ell}$ of K satisfies the following condition:*
$$d(M_{\cdot,\ell}, \mathsf{CH}(\{M_{\cdot,\ell'} : \ell' \neq \ell\}) \geq 7\rho. \tag{1}$$
2. *For every $\ell \in [k]$, there exists a subset $S_\ell \subseteq [n]$ of size δn such that*
$$|M_{\cdot,\ell} - P_{\cdot,S_\ell}| \leq \frac{\rho}{30} \tag{2}$$
3. *The perturbation parameter σ is defined as follows*
$$\frac{\sigma}{\sqrt{\delta}} \leq \frac{\rho}{36}. \tag{3}$$

Then the $\mathrm{ICR}(\rho/3, \widetilde{\mathbf{A}}) = k$.

Proof. The proof consists of computing an upper-bound (see Lemma 3) and a lower bound (see Corollary 8). ☐

Remarks: Before proceeding to give formal details we wish to make some remarks to convey the main intuitions. The assumptions are re-statements of the three standard assumptions defined in Subsection 2.1.2. We would like to note that the first condition (1) above is much weaker than requiring that $M_{\cdot,\ell}$ is far from the affine hull of the other columns of $M_{\cdot,\ell}$, made in (Bhattacharyya & Kannan, 2020). We now give some intuitions behind the proof. It proceeds in two parts. In the upper bound proof, we show tha $\mathrm{ICR}(\rho/3, \widetilde{\mathbf{A}})$ is at most k (see Lemma 3). The idea of the proof is as follows. For $R \in \mathcal{R}_\delta$, $A_{\cdot,R}$ and $P_{\cdot,R}$ are close to each other (this is where the subset smoothening operation helps). Now, $P_{\cdot,R}$ lies inside K, and so can be expressed as a convex combination of the columns $M_{\cdot,\ell}$ of $\mathbf{M}$, each of which is close to the corresponding point $A_{\cdot,S_\ell}$, where S_ℓ is the subset defined in (2). It follows that $A_{\cdot,R}$ is close to the convex hull of the points $\{A_{\cdot,S_\ell}, \ell \in [k]\}$. This exhibits a convex polytope on k vertices which are close to $\mathsf{CH}(\widetilde{\mathbf{A}})$, proving the upper bound.

The lower bound part of the proof (see Corollary 8) is harder and involves new tools from Convex Geometry. Suppose there is a subset of points W whose convex hull is close to $\mathsf{CH}(\widetilde{\mathbf{A}})$. It is not hard to show that the convex hull of W is also close to $\mathsf{CH}(\mathbf{M})$ (again using subset smoothening). Now we show that we can partition (a slightly larger set than) $\mathsf{CH}(\mathbf{M})$ into k parts by cutting it with suitable hyperplanes obtained from the Separating Hyperplane Theorem, and W must have a non-empty intersection with each of these parts. This shows that $|W|$ must be at least k.

3.3. Upper Bound Proof

Let H denote $\mathsf{CH}(\widetilde{\mathbf{A}})$. Let H' denote the convex hull of $\{A_{.,S_\ell}, \ell \in [k]\}$, where $S_\ell \in \mathcal{R}_\delta$ are as defined in (2). We show that H is within Hausdorf distance $\rho/3$ of H'. Since $H' \subseteq H$, it suffices to show that for every point in H, there is a point in H' within distance at most $\rho/3$. H is a convex set and $d(x, H')$ is a convex function of x, so the maximum of $d(x, H')$ over H is attained at a vertex of H. Hence it suffices to show that for each of the points $A_{.,R}, R \in \mathcal{R}_\delta$, we have $d(A_{.,R}, H') \le \rho/3$.

The proof proceeds by first showing that for any $R \in \mathcal{R}_\delta$, the vector $A_{.,R}$ is close to $P_{.,R}$, in fact within distance at most $\sigma/\sqrt{\delta}$. Now, any vector of the form $P_{.,R}$ can be written as a convex combination of the vertices of K, i.e., the vectors $M_{.,\ell}, \ell \in [k]$. The same argument also shows that each of the vectors $M_{.,\ell}$ is close to the corresponding vector $A_{.,S_\ell}$. Thus $A_{.,R}$ is close to a convex combination of the vectors $A_{.,S_\ell}, \ell \in [k]$. We now prove these claims formally.

Claim 2. *For all $S \subseteq [n], |S| \ge \delta n,\ |A_{.,S} - P_{.,S}| \le \frac{\sigma}{\sqrt{\delta}}$.*

Proof. We first note that $|A_{.,S} - P_{.,S}| \le ||\mathbf{A} - \mathbf{P}||/\sqrt{|S|}$. Indeed, let u be the unit vector which is $1/\sqrt{|S|}$ on coordinates corresponding to S, 0 otherwise. Then

$$|A_{.,S} - P_{.,S}| = \frac{1}{\sqrt{|S|}} \cdot |(\mathbf{A} - \mathbf{P})u| \le \frac{||\mathbf{A} - \mathbf{P}||}{\sqrt{|S|}}.$$

The claim now follows by using the definition of σ, i.e., $||\mathbf{A} - \mathbf{P}|| = \sigma\sqrt{n}$. □

We are now ready to show the upper bound result.

Lemma 3. *Under the conditions of Theorem 1,* $\mathtt{ICR}(\rho/3, \widetilde{\mathbf{A}})$ *is at most k.*

Proof. Fix an $R \in \mathcal{R}_\delta$. Since $P_{.,R} \in K$, it can be written as a convex combination $\sum_{\ell \in [k]} \lambda_\ell M_{.,\ell}$ of the vertices of K. Then, we have

$$\left| A_{.,R} - \sum_{\ell \in [k]} \lambda_\ell A_{.,S_\ell} \right| \le |A_{.,R} - P_{.,R}| + \sum_\ell \lambda_\ell \Big(|M_{.,\ell} - P_{.,S_\ell}| + |A_{.,S_\ell} - P_{.,S_\ell}| \Big) \le \frac{2\sigma}{\sqrt{\delta}} + \frac{\rho}{30} \le \rho/3,$$

since the first and third term (in the RHS of the first inequality) are each at most $\sigma/\sqrt{\delta}$ by the Claim 2 and the second term is at most $\rho/30$ by (2). This shows that that the $\mathtt{ICR}(\rho/3, \widetilde{\mathbf{A}})$ is at most k. □

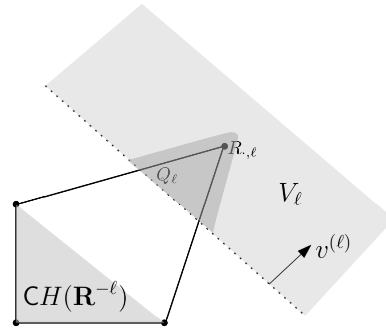

Figure 1. Illustration of the proof of Theorem 5: the matrix $\mathbf{R}$ has four points. The lightly shaded half-space around $R_{.,\ell}$ denotes V_ℓ, whereas the darker shaded region around the same vertex denotes Q_ℓ.

3.4. Lower Bound Proof

We now prove the more difficult part of Theorem 1, i.e., $\mathtt{ICR}(\rho/3, \widetilde{\mathbf{A}})$ cannot be strictly less than k. We will prove a stronger assertion, namely, that there is no set of $k-1$ points in $\Re^d$ whose convex hull is within distance $\rho/3$ of H, where $H = \mathsf{CH}(\widetilde{\mathbf{A}})$.

Assume, for sake of contradiction, that there exist points $w_1, \ldots, w_{k-1} \in \Re^d$ such that

$$\mathsf{dist}(H, \mathsf{CH}(\{w_1, \ldots, w_{k-1}\})) \le \rho/3. \tag{4}$$

Recall that $\mathrm{K} = \mathsf{CH}(\mathbf{M})$. Let W_0 denote $\mathsf{CH}(\{w_1, \ldots, w_{k-1}\})$. We first begin by showing that $\mathsf{dist}(\mathrm{K}, W_0)$ is at most ρ. The intuition is that the Hausdorff distance between H and K is small – the argument follows along the same lines as in the upper bound proof. It follows from (4) that $\mathsf{dist}(\mathrm{K}, W_0)$ is also small. We give the complete proof of the following result in the supplementary material.

Claim 4. *Let W_0 denote the convex hill of the points $w_1, \ldots, w_{k-1}$. Then,* $\mathsf{dist}(\mathrm{K}, W_0) \le \rho$.

The desired contradiction follows easily from the above claim and the following result:

Theorem 5. *Let W be a set of points such that the Hausdorff distance between the the convex hull of W and that of the columns of $\mathbf{M}$ is at most ρ, i.e.,* $\mathsf{dist}(\mathsf{CH}(W), \mathsf{CH}(\mathbf{M})) \le \rho$. *Then $|W| \ge k$.*

The intuition behind the proof is as follows: each vertex $M_{.,\ell}$ of $\mathsf{CH}(\mathbf{M})$ is far from the convex hull formed by rest of the columns of $\mathbf{M}$, and hence we can find a separating hyperplane which separates $M_{.,\ell}$ from $\mathsf{CH}(\mathbf{M}^{-\ell})$ by some large enough margin (see the region V_ℓ in Figure 1). We define a further refinement of this region, shown as Q_ℓ in Figure 1. The proof consists of two conceptual steps: (i) we show that the regions Q_ℓ are disjoint for distinct columns $M_{\cdot,\ell}$ of $\mathbf{M}$, and (ii) we argue that each of the regions Q_ℓ must contain at least one vertex of W. Since there are k columns in $\mathbf{M}$, this would implies that $|W|$ must be at least k.

Proof. For two subsets A and B of points in $\Re^d$, define their Minkowski sum, $A+B$, as $\{a+b : a \in A, b \in B\}$. For an index $\ell \in [k]$, let $\mathbf{M}^{-\ell}$ denote the matrix $\mathbf{M}$ with column ℓ removed. Since the columns of $\mathbf{M}$ satisfy condition (1), this implies that (here $\mathbf{B}$ denotes the unit ball)

$$[M_{\cdot,\ell} + 7\rho\mathbf{B}] \cap \mathsf{CH}(\mathbf{M}^{-\ell}) = \emptyset.$$

Therefore, the Separating Hyperplane Theorem from Convex Geometry implies that there is a unit vector $v^{(\ell)}$ such that for all $\ell' \neq \ell$,

$$v^{(\ell)} \cdot M_{\cdot,\ell} > v^{(\ell)} \cdot M_{\cdot,\ell'} + 7\rho \tag{5}$$

For an index $\ell \in [k]$, let V_ℓ denote the set

$$V_\ell := \{x : v^{(\ell)} \cdot x > v^{(\ell)} \cdot M_{\cdot,\ell} - 2\rho\},$$

and define $Q_\ell := (\mathsf{CH}(\mathbf{M}) + \rho\mathbf{B}) \cap V_\ell$.

Claim 6. *For every $\ell \in [k]$, there is a point $w \in W$ such that $w \in Q_\ell$.*

Proof. Assume for the sake of contradiction that there is an index $\ell \in [k]$ such that $Q_\ell \cap W = \emptyset$. Since $\mathsf{dist}(\mathsf{CH}(W), \mathsf{CH}(\mathbf{M})) \leq \rho$, $W \subseteq \mathsf{CH}(\mathbf{M}) + \rho\mathbf{B}$, and so it must be the case that $V_\ell \cap W = \emptyset$. The definition of V_ℓ implies that for every $w \in W$,

$$w \cdot v^{(\ell)} \leq M_{\cdot,\ell} \cdot v^{(\ell)} - 2\rho.$$

Consequently, for every point $y \in \mathsf{CH}(W)$,

$$v^{(\ell)} \cdot y \leq v^{(\ell)} \cdot M_{\cdot,\ell} - 2\rho. \tag{6}$$

Since $\mathsf{dist}(\mathsf{CH}(\mathbf{M}), \mathsf{CH}(W)) \leq \rho$, there is a point $y \in \mathsf{CH}(W)$ with $|y - M_{\cdot,\ell}| \leq \rho$. But this contradicts (6). This proves the claim. □

So, we must have that for each $\ell \in [k]$, $Q_\ell \cap W$ is non-empty. We now show that the sets Q_ℓ are disjoint, which will yield the desired result.

Lemma 7. *The sets $Q_\ell, \ell \in [k]$ are mutually disjoint.*

Proof. Suppose, for the sake of contradiction, that there is a point $z \in Q_\ell \cap Q_{\ell'}$ for distinct indices ℓ, ℓ'. Since $z \in \mathsf{CH}(\mathbf{M}) + \rho\mathbf{B}$, there is a point $y \in \mathsf{CH}(\mathbf{M})$, with $|y - z| \leq \rho$.

So y can be written as a convex combination $\sum_{\ell'' \in [k]} \alpha_{\ell''} M_{\cdot,\ell''}$. We now also write y as

$$y = \alpha_\ell M_{\cdot,\ell} + (1-\alpha_\ell)x \ , \ x \in \mathsf{CH}(M_{\cdot,\ell''} : \ell'' \neq \ell) \tag{7}$$

Inequality (5) implies that $v^{(\ell)} \cdot x < v^{(\ell)} \cdot M_{\cdot,\ell} - 7\rho$. So,

$$v^{(\ell)} \cdot y < v^{(\ell)} \cdot M_{\cdot,\ell} - (1-\alpha_\ell)7\rho.$$

Since $v^{(\ell)}$ is unit vector and $|y - z| \leq \rho$,

$$v^{(\ell)} \cdot z \leq v^{(\ell)} \cdot M_{\cdot,\ell} - (1-\alpha_\ell)7\rho + \rho.$$

But since $z \in V_\ell$, we have by the definition of V_ℓ

$$v^{(\ell)} \cdot z > v^{(\ell)} \cdot M_{\cdot,\ell} - 2\rho.$$

Thus by the last two inequalities, $\alpha_\ell > 4/7 > 0.5$. A similar argument shows that $\alpha_{\ell'} > 0.5$. But this is not possible because the quantities $\alpha_{\ell''}$ are non-negative and add to 1. This proves the result. □

Combining Lemma 7 and Claim 6, we see that $|W| \geq k$. This proves the desired theorem. □

As a corollary, we get the lower bound result:

Corollary 8. *Under the assumptions of Theorem 1,* $\texttt{ICR}(\rho/3, \widetilde{\mathbf{A}}) \geq k$.

Proof. Combining Theorem 5 and Claim 4, we see that if W is any set of points for $\mathsf{dist}(\mathsf{CH}(\widetilde{\mathbf{A}}), \mathsf{CH}(W)) \leq \rho/3$, then $|W| \geq k$. By definition, if $\texttt{ICR}(\rho/3, \widetilde{\mathbf{A}}) = k'$, then there is a set of k' points whose convex hull is within Hausdorff distance $\rho/3$ of $\mathsf{CH}(\widetilde{\mathbf{A}})$. Therefore, $k' \geq k$. □

It is worth noting that the matrix $\widetilde{\mathbf{A}}$ has exponential (in n) number of columns, and so cannot be used directly by an efficient algorithm. We have given an algebraic characterization of the number of vertices of K in terms of ICR, but no polynomial time algorithm is known to find ICR. In the next section, we give a polynomial time algorithm to find the number of vertices of K.

4. Algorithm for finding the number of vertices of the latent polytope

In this section, we consider a slightly stronger set of standard assumptions than those in Theorem 1. We prove that if the input points satisfy these conditions, then we can recover the correct value of k in polynomial time. In fact, we will show that there is a polynomial time computable threshold τ such that

$$k = \text{Max}_r \ : \ s_r(\mathbf{A}) \geq \tau.$$

Singular value thresholding is a well known procedure which yields provable guarantees in many applications (Cai et al., 2010; Chatterjee, 2015; Donoho & Gavish, 2014). However the techniques considered in the literature do not apply to the problem at hand and hence new methods are warranted.

The main difficulty in choosing such a threshold τ is that it cannot be kept at some suitable multiple of the largest singular value of the data matrix $\mathbf{A}$: the matrix $\mathbf{A}$ may be ill-conditioned, and so the largest singular value may not give us any information about the parameter k. Instead, we write down a convex program whose output is used for deciding this threshold τ. Once we have the correct value of k, we can use ideas based on ICR to recover the actual vertices of $\mathbf{M}$ (see for example Algortihm 2 and Theorem 20).

Algorithm 1 Algorithm for finding the number of extreme points of K.

Input: $d \times n$ data matrix $\mathbf{A}$ and a parameter $\delta \in (0, 1)$.
Let opt be the optimal value of the convex program (12).

Let $k^\star$ be the maximum value of $k' \in [d]$ such that the singular value $s_{k'}(\mathbf{A}) \geq \delta^2 \mathsf{opt}/8$.
Output $k^\star$.

We now state the main result of this section. We give some definitions. For a subspace V and a point x, let $\mathbf{proj}(x, V)$ denote the orthogonal projection of x on V. For a subset S of points, let $\mathbf{Null}(S)$ denote the subspace orthogonal to the linear span of S.

Theorem 9. *Suppose that the input data satisfies the following assumptions. There exists a $\delta \in (0, 1)$ such that*

Every vertex $M_{.,\ell}$ of K satisfies the following condition:

$$|\mathbf{proj}\left(M_{.,\ell}, \mathbf{Null}\left(\mathbf{M} \setminus M_{.,\ell}\right)\right)| \geq \delta |M_{.,\ell}|. \tag{8}$$

The LHS above is same as the distance of $M_{.,\ell}$ from the affine hull of the other columns of $\mathbf{M}$.

$$\textit{For } S_\ell := \{j : |M_{.,\ell} - P_{.,j}| \leq \frac{4\sigma}{\sqrt{\delta}}\}, \quad |S_\ell| \geq \delta n \tag{9}$$

$$\sigma \leq \delta^3 \min_\ell |M_{.,\ell}|/20. \tag{10}$$

Further, assume that all the entries of $\mathbf{M}$ are non-negative. Then, given δ and the data matrix $\mathbf{A}$, Algorithm 1 outputs the correct value of k.

Note that the LHS of condition (8) is smaller than the LHS of the corresponding constraint (1).

The following result follows as corollary of the above result. It shows that the ICR of the subset smoothed data matrix $\widetilde{\mathbf{A}}$ is equal to the number of vertices k of K. Details are given in the supplementary material.

Lemma 10. *Suppose the conditions (8), (9) and (10) stated in Theorem 9 are satisfied by the input data. Then* $\mathtt{ICR}(6\sigma/\sqrt{\delta}, \widetilde{\mathbf{A}}) = k$.

4.1. Proof of Theorem 9

We now proceed to prove Theorem 9. There are two phases in Algorithm 1. In the first phase, we estimate $\min_\ell |M_{.,\ell}|$ by writing a suitable convex program. We use this estimate as a threshold for the singular values of $\mathbf{A}$ to determine the value of k.

The convex program is as follows (recall that $\mathbf{A}$ is the data matrix generated from LkP defined in Definition 2):

$$\min \ \mathsf{opt} := \quad |\mathbf{A} \cdot x| \tag{11}$$

$$\sum_{j \in [n]} x_j = 1, \quad 0 \leq x_j \leq 1/(\delta n), \forall j \in [n]. \tag{12}$$

We now show that the optimal value of the convex program can be used to approximate $\min_\ell |M_{.,\ell}|$ within a constant factor.

Lemma 11. *Let $\rho_0 := \min_\ell |M_{.,\ell}|$, and* opt *be defined in (11). Then*

$$\rho_0/(2k) \leq \mathsf{opt} \leq 2\rho_0.$$

Proof. The proof is done by two subclaims, each of which prove one of the above inequalities. Let $x^\star$ denote the vector which satisfies $\mathsf{opt} = |A \cdot x^\star|$.

Claim 12. $|\mathbf{A} \cdot x^\star| \geq \rho_0/(2k)$.

Proof. The vector $\mathbf{P} \cdot x^\star$ is a convex combination of the columns of $\mathbf{P}$. Each column of $\mathbf{P}$ can be expressed as a convex combination of the columns of $\mathbf{M}$. It follows that $\mathbf{P} \cdot x^\star$ is a convex combination of the columns of $\mathbf{M}$, i.e., it is of the form $\sum_{j \in [k]} \alpha_j M_{.,j}$, where α_j's are non-negative and sum to 1. It follows that there is an index j with $\alpha_j \geq 1/k$. Since all the columns of $\mathbf{M}$ are non-negative, $\mathbf{P} \cdot x$ is component-wise at least $M_{.,j}/k$, and so, has length at least ρ_0/k. Finally,

$$|\mathbf{A} \cdot x| \geq |\mathbf{P} \cdot x| - ||P - A|| \cdot |x| \geq \rho_0/k - \sigma/\delta \overset{(10)}{\geq} \rho_0/(2k),$$

where the second last inequality follows from the fact that $x_j \leq 1/(\delta n)$ and (10). ☐

Claim 13. $|\mathbf{A} \cdot x^\star| \leq 2\rho_0$

Proof. Let ρ_0 be $|M_{.,\ell}|$. Define a feasible solution x to the convex program $\mathcal{P}$ as follows: $x_j = 1/(\delta n)$ if $j \in S_\ell$, 0 otherwise. Observe that $|x| = \frac{1}{\sqrt{\delta}\sqrt{n}}$. Now,

$$|\mathbf{A} \cdot x| = |A_{.,S_\ell}| \leq |P_{.,S_\ell}| + \frac{\sigma}{\sqrt{\delta}} \leq |M_{.,\ell}| + \frac{4\sigma}{\sqrt{\delta}} + \frac{\sigma}{\sqrt{\delta}} \leq 2\rho_0,$$

where the first inequality is by Claim 4 and the second by (9). □

The above two results show that $\rho_0/(2k) \leq \mathsf{opt} \leq 2\rho_0$. □

We now use opt as a threshold for the singular values of $\mathbf{A}$, and show that k can be recovered from this threshold. This will also prove correctness of Algorithm 1, and hence yield a provable data determined threshold for finding the ICR of $\mathbf{A}$.

Theorem 14. *Define* $k^\star = \max_{k' \in [d]} : s_{k'}(\mathbf{A})/\sqrt{n} \geq \delta^2 \mathsf{opt}/8$. *Then* $k = k^\star$.

Proof. In order to prove this result, we need to estimate $s_k(\mathbf{A})$ and $s_{k+1}(\mathbf{A})$. We shall do this by estimating the singular values of some related matrices. We begin by bounding the singular values of $\mathbf{M}$.

Claim 15. $s_k(\mathbf{M}) \geq \frac{\delta \rho_0}{\sqrt{k}}$, *whereas* $s_{k+1}(\mathbf{M}) = 0$.

Proof. The second statement follows trivially since $\mathbf{M}$ has only k columns, and so is a rank k matrix. For the first statement, observe that

$$s_k(\mathbf{M}) = \min_{x:|x|=1} |\mathbf{M}x|.$$

Let x be any unit vector in $\Re^k$. There must be a coordinate $\ell \in [k]$ such that $|x_\ell| \geq \frac{1}{\sqrt{k}}$. Now, $|\mathbf{M}x| = \left|\sum_{j\in[k]} x_j M_{.,j}\right| \geq |x_\ell| \cdot |\mathbf{proj}(M_{.,\ell}, \mathbf{Null}(\mathbf{M} \setminus M_{.,\ell})| \geq \frac{\delta\rho_0}{\sqrt{k}}$, where the last inequality follows from condition (8). □

Let $S_1, \ldots, S_k$ be subsets of $[n]$ guaranteed by condition (9). We assume that each of these subsets have size exactly δn (otherwise we remove some elements from them). Note that these sets are mutually disjoint, since (8) and (10) imply that for any two distinct $\ell, \ell' \in [k]$, $|M_{.,\ell} - M_{.,\ell'}| > \frac{8\sigma}{\sqrt{\delta}}$.

Let $\mathbf{B} = (P_{.,S_1} \mid P_{.,S_2} \mid \ldots | P_{.,S_k})$. We relate $s_k(\mathbf{B})$ and $s_k(\mathbf{M})$.

Claim 16. $s_k(\mathbf{B}) \geq s_k(\mathbf{M}) - \frac{4\sigma\sqrt{k}}{\sqrt{\delta}} \geq \frac{\delta\rho_0}{2\sqrt{k}}$.

Proof. Condition (9) implies that $||\mathbf{B} - \mathbf{M}|| \leq \frac{4\sigma\sqrt{k}}{\sqrt{\delta}}$. This implies the first inequality in the claim, since we have (from Linear Algebra), $s_k(\mathbf{B}) \geq s_k(\mathbf{M}) - \cdot||\mathbf{M}-\mathbf{B}||$. The second inequality follows from Claim 15 and (10). □

We now relate $s_k(\mathbf{B})$ and $s_k(\mathbf{P})$.

Claim 17. $s_k(\mathbf{P}) \geq \sqrt{\delta n} \cdot s_k(\mathbf{B}) \geq \frac{\delta^{1.5}\rho_0}{2\sqrt{k}} \cdot \sqrt{n}$.

Proof. We shall exhibit a subspace of dimension k such that for any unit vector x in this subspace, $|\mathbf{P}\cdot x| \geq \sqrt{\delta n}\cdot s_k(\mathbf{P})$. This will prove the desired result by the Max-Min theorem of Linear Algebra which states:

$$s_k(\mathbf{B}) = \max_{V:\text{dim}(V)=k} \min_{x\in V, |x|=1} |\mathbf{P}x|.$$

For each index $\ell \in [k]$ defined a unit vector $e_\ell \in \Re^n$ as follows:

$$(e_\ell)_j = \begin{cases} \frac{1}{\sqrt{\delta n}} & \text{if } j \in S_\ell \\ 0 & \text{otherwise} \end{cases}$$

Since the sets $S_1, \ldots, S_k$ are pair-wise disjoint, the vectors $e_\ell, \ell \in [k]$, span a subspace U of dimension k (in fact these vectors are form an orthonormal family of k vectors).

Now consider any unit vector $v = \sum_{j\in[k]} \beta_j e_j$. Since e_j's are orthonormal, $\sum_j \beta_j^2 = 1$. Now,

$$\mathbf{P} \cdot v = \sum_{j\in[k]} \beta_j \sqrt{\delta n} \cdot P_{.,S_j} = \sqrt{\delta n} \cdot \mathbf{P} \cdot y,$$

where y is the vector whose j^{th} coordinate is β_j. Since y is a unit vector, $|\mathbf{P}\cdot y| \geq s_k(\mathbf{B})$. This implies the desired result (the last inequality in the claim follows from Claim 16). □

Finally, we give bounds on the singular values of the data matrix $\mathbf{A}$.

Lemma 18. $s_k(\mathbf{A}) \geq \delta^2 \mathsf{opt}/8 > s_{k+1}(\mathbf{A})$.

Proof. Since $||\mathbf{P} - \mathbf{A}|| \leq \sigma\sqrt{n}$, $|s_r(\mathbf{P})/\sqrt{n} - s_r(\mathbf{A})/\sqrt{n}|$ is at most σ for any $r \in [d]$. Using this fact and Claim 17, we see that

$$s_k(\mathbf{A})/\sqrt{n} \geq \frac{\delta^{1.5}\rho_0}{4\sqrt{k}} \geq \frac{\delta^{1.5}\mathsf{opt}}{8\sqrt{k}} \geq \frac{\delta^2\mathsf{opt}}{8},$$

where the second inequality follows from Claim 13, and the last inequality follows from the fact that $\delta \leq 1/k$.

Similarly, (since $s_{k+1}(\mathbf{P}) = 0$)

$$\begin{aligned} s_{k+1}(\mathbf{A})/\sqrt{n} \leq \sigma &\overset{(10)}{\leq} \delta^3\rho_0/20 \leq k\delta^3\mathsf{opt}/10 \\ &\leq \delta^2\mathsf{opt}/10 < \delta^2\mathsf{opt}/8, \end{aligned}$$

where the third inequality follows from Claim 12. □

This proves that $k^\star = k$. □

Thus we have shown that $k^\star = k$. This proves that we can recover the number of vertices in the latent polytope if the conditions (8), (9), (10) are satisfied, hence proving Theorem 9.

4.2. Finding a non-negative matrix factorization without the knowledge of inner-dimension

Given $\mathbf{A}$ finding an approximate non-negative factorization is intractable. In a long line of work starting from (Arora et al., 2012) several ingenious algorithms have emerged (Recht et al., 2012; Gillis & Vavasis, 2014; Rong & Zou, 2015) which using *Separability*, first introduced by (Donoho & Stodden, 2003a), provably recover an approximate factorization. These procedures assume that the inner-dimension k is known. By imposing the assumption that *Every data point is approximately a non-negative combination of k data points*, several algorithms (e.g. (Ambikapathi et al., 2013; Fu et al., 2015; Gillis & Luce, 2018)) can find minimum such k by solving Linear or Convex programs. However the assumption is too strong to hold for LkP and can be shown to fail in many simple instances. To the best of our knowledge there are no known algorithms which can provably recover the inner-dimension and the factorization under the conditions of LkP.

In this section we state polynomial time procedures which provably recovers the factorization(the matrix $\mathbf{M}$) along with k, the inner-dimension (see equation (Approx-NMF)). Noting that Approx-NMF is a special case of LkP, the application of Algorithm 1 recovers k.

It is important to note that the *standard assumptions* stated earlier are different than separability. In fact conditions (9) and (10) of Theorem 9, are analogous to assumptions which have been used for NMF, namely, *nearly pure records* and *noise* assumptions of of (Bhattacharyya et al., 2016). Instead of the *Well-Separatedness* condition (8) of Theorem 9, we make the following "Dominant Features" assumptions which is similar to, but, weaker than their assumption with the same name.

Dominant Features This stipulates that for $\ell \in [k]$, there is a set T_ℓ of "dominant features" of column ℓ of $\mathbf{M}$. More precisely,

$$\exists T_1, T_2, \ldots, T_k \text{ disjoint } : \forall \ell, \sum_{i \in T_\ell} M_{i,\ell} \geq 4 \sum_{i \in T_\ell} \sum_{\ell' \neq \ell} M_{i,\ell'}.$$

Lemma 19. *The Dominant Features assumption above implies condition* (8) *of Theorem 9.*

The proof is based on diagonal dominance and is given in the Supplementary Material. Finally we state the main theorem, which gives a procedure, described in Algorithm 2, for finding close approximations to the vertices of $\mathbf{M}$.

The algorithm is a simplified version of (Bhattacharyya & Kannan, 2020), requiring new proofs. At the outset, the algorithm finds the left $k-$ singular subspace V of $\mathbf{A}$ and projects data points onto V. It runs in k iterations and in each iteration, finds an approximation to a new vertex of K. This approximation is of the form $A_{\cdot,R}$ for some $R \subseteq [n], |R| = \delta n$. R is obtained as follows: We pick a random vector u_r in $V\cap$ the null space of the vectors found so far and set: $R \leftarrow \arg\max_R |u \cdot A_{\cdot,R}|$, noting that the optimization can be carried out in polynomial time. Indeed, the optimization to find R is a 1-dimension problem. We project all the columns $A_{\cdot,j}$ along u_r (note that this is the matrix $\mathbf{A}$ which has only n columns). Now the optimal R would be either the δn columns of $\mathbf{A}$ with the highest dot product with u_r or the smallest, and so, can be done in polynomial time.

Theorem 20. *Assuming conditions* (9) *and* (10) *and also the Dominant Features assumption, Algorithm 1 finds k in polynomial time. Using this value of k, Algorithm 2 finds k points, each of which is within* $O(k^4\sigma/\delta^{1.5})$ *distance of a unique vertex of* $\mathbf{M}$.

The proof is given in the supplementary material. The above theorem yields an NMF without knowing the value of *inner-dimension*. A detailed discussion of the relative strengths and weakness of various assumptions is beyond the scope of this paper and will be discussed elsewhere. It suffices to stay that the assumptions, techniques and the results presented in this section are novel contributions to NMF.

Algorithm 2 Algorithm for finding the approximations to the vertices of K.

Input: $d \times n$ data matrix $\mathbf{A}$ and a parameter $\delta \in (0, 1), k$.

Let V be the subspace of $\Re^d$ spanned by the top k left singular vectors of $\mathbf{A}$.
Let $\hat{A}_{\cdot,j}$ be the projection of $A_{\cdot,j}$ on V, for $j = 1, \ldots, n$.
for $r = 0, \ldots, k-1$ **do**
  $u_r \leftarrow$ unit vector chosen uniformly at random from the subspace $U_r := V \cap \mathbf{Null}(\hat{A}_{\cdot,R_1}, \ldots, \hat{A}_{\cdot,R_r})$.
  $R_{r+1} \leftarrow \arg\max_{R \in \mathcal{R}_\delta} |u \cdot \hat{A}_{\cdot,R}|$.
end for
Output $\hat{A}_{\cdot,R_1}, \ldots, \hat{A}_{\cdot,R_k}$.

5. Conclusions

This paper uses Convex geometry in finding the number of extreme points of the Latent-k-polytope. This settles several interesting problems related to finding the number of components in Mixed-membership models.

Acknowledgement The authors thank the reviewers for their insightful comments which greatly helped in preparing the camera ready version. The author CB gratefully acknowledges the support of Microsoft Research.

References

Ambikapathi, A., Chan, T., Chi, C., and Keizer, K. Hyperspectral data geometry-based estimation of number of endmembers using p -norm-based pure pixel identification algorithm. *IEEE Trans. Geosci. Remote. Sens.*, 51(5-1):2753–2769, 2013. URL https://doi.org/10.1109/TGRS.2012.2213261.

Arora, S., Ge, R., Kannan, R., and Moitra, A. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 145–162. ACM, 2012.

Barvinok, A. Thrifty approximations of convex bodies by polytopes. *International Mathematics Research Notices*, 2014(16):4341–4356, 2014. doi: 10.1093/imrn/rnt078.

Bhattacharyya, C. and Kannan, R. Near-optimal sample complexity bounds for learning latent k-polytopes and applications to ad-mixtures. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 854–863. PMLR, 2020.

Bhattacharyya, C., Goyal, N., Kannan, R., and Pani, J. Non-negative matrix factorization under heavy noise. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1426–1434. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/bhattacharya16.html.

Cai, J., Candès, E. J., and Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, 20(4):1956–1982, 2010. doi: 10.1137/080738970. URL https://doi.org/10.1137/080738970.

Chatterjee, S. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43(1):177–214, 2015.

Donoho, D. and Gavish, M. Minimax risk of matrix denoising by singular value thresholding. *Annals of Statistics*, 42(6):2413–2440, 2014.

Donoho, D. and Stodden, V. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, pp. 1141–1148, 2003a.

Donoho, D. L. and Stodden, V. When does non-negative matrix factorization give a correct decomposition into parts? In Thrun, S., Saul, L. K., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pp. 1141–1148. MIT Press, 2003b. URL https://proceedings.neurips.cc/paper/2003/hash/1843e35d41ccf6e63273495ba42df3c1-Abstract.html.

Fu, X., Ma, W., Chan, T., and Bioucas-Dias, J. M. Self-dictionary sparse regression for hyperspectral unmixing: Greedy pursuit and pure pixel search are related. *IEEE J. Sel. Top. Signal Process.*, 9(6):1128–1141, 2015.

Gillis, N. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12, 2014.

Gillis, N. and Luce, R. A fast gradient method for nonnegative sparse regression with self-dictionary. *IEEE Trans. Image Process.*, 27(1):24–37, 2018. URL https://doi.org/10.1109/TIP.2017.2753400.

Gillis, N. and Vavasis, S. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(4):698–714, April 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.226.

Gruber, P. M. *Aspects of approximation of convex bodies, Handbook of Convex Geometry*, volume A, pp. 319–345. North-Holland, Amsterdam, 1993.

Khachiyan, L., Boros, E., Borys, K., Elbassioni, K., and Gurvich, V. Generating all vertices of a polyhedron is hard. *Discrete & Computational Geometry*, 39(1):174–190, 2008. doi: 10.1007/s00454-008-9050-5. URL https://doi.org/10.1007/s00454-008-9050-5.

Recht, B., Re, C., Tropp, J., and Bittorf, V. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pp. 1214–1222, 2012.

Rong, G. and Zou, J. Intersecting faces: Non-negative matrix factorization with new guarantees. In *ICML*, pp. 2295–2303, 2015.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. ISSN 01621459. URL http://www.jstor.org/stable/27639773.

Vavasis, S. A. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

Supplementary Material: Finding k in Latent k polytope

A. Missing proofs of Section 3.2

We first give the proof of Claim 4, which is restated here.

Claim 4. *Let W_0 denote the convex hill of the points $w_1, \ldots, w_{k-1}$. Then,* $\mathsf{dist}(\mathrm{K}, W_0) \leq \rho$.

Proof. As in the upper bound proof, it is enough to show that for every $\ell \in [k]$, $d(M_{.,\ell}, W_0) \leq \rho$, and that for $w_j \in W_0$, $d(w_j, M) \leq \rho$.

Fix an index $\ell \in [k]$. We know by (2) that $|M_{.,\ell} - P_{.,S_\ell}| \leq \frac{\rho}{30}$, and Claim 2 implies that $|A_{.,S_\ell} - P_{.,S_\ell}| \leq \frac{\sigma}{\sqrt{\delta}}$. Therefore, $|M_{.,\ell} - A_{.,S_\ell}| \leq \rho/3$ (using (3)). Finally, the fact that $\mathsf{dist}(H, W) \leq \rho/3$ implies that there is a point $w \in W$ such that $|A_{.,S_\ell} - w| \leq \rho/3$. It follows that $|M_{.,\ell} - w| \leq 2\rho/3$..

Now we show the other direction. Fix an index $\ell \in [k-1]$. By (4) and the upper bound result that $\mathsf{dist}(H, H') \leq \rho/3$, it follows that $d(w_\ell, H') \leq 2\rho/3$. Therefore there exist non-negative $\lambda_{\ell'}, \ell' \in [k]$, adding up to 1 such that $|w_\ell - \sum_{\ell' \in [k]} \lambda_{\ell'} A_{.,S_{\ell'}}| \leq 2\rho/3$. Condition (2) and Claim 2 now imply that

$$|w_\ell - \sum_{\ell' \in [k]} \lambda_{\ell'} M_{.,\ell'}| \leq 2\rho/3 + \frac{\rho}{30} + \frac{\sigma}{\delta} \overset{(3)}{\leq} \rho.$$

Thus, $d(w_\ell, M) \leq \rho$. □

B. Missing proofs of Section 4

We prove Lemma 10.

Lemma 10. *Suppose the conditions (8), (9) and (10) stated in Theorem 9 are satisfied by the input data. Then* $\mathtt{ICR}(6\sigma/\sqrt{\delta}, \widetilde{\mathbf{A}}) = k$.

Proof. We first show that $\mathtt{ICR}(6\sigma/\delta, \widetilde{\mathbf{A}}) \leq k$. The proof is very similar to the argument in Section 3.3. Fix an $R \in \mathcal{R}_\delta$. Since $P_{.,R} \in K$, it can be written as a convex combination $\sum_{\ell \in [k]} \lambda_\ell M_{.,\ell}$ of the vertices of K. Then, we have

$$\begin{aligned} \left| A_{.,R} - \sum_{\ell \in [k]} \lambda_\ell A_{.,S_\ell} \right| &\leq |A_{.,R} - P_{.,R}| + \sum_\ell \lambda_\ell |M_{.,\ell} - P_{.,S_\ell}| \\ &\quad + \sum_\ell \lambda_\ell |A_{.,S_\ell} - P_{.,S_\ell}| \\ &\leq \frac{2\sigma}{\sqrt{\delta}} + \frac{4\sigma}{\sqrt{\delta}} \leq \frac{6\sigma}{\sqrt{\delta}}, \end{aligned}$$

since the first and third term (in the RHS of the first inequality) are each at most $\sigma/\sqrt{\delta}$ by the Claim 2 and the second term is at most $\frac{4\sigma}{\sqrt{\delta}}$ by (9). This shows that $\mathtt{ICR}(6\sigma/\sqrt{\delta}, \widetilde{\mathbf{A}}) \leq k$.

We now show the lower bound $\mathtt{ICR}(6\sigma/\delta, \widetilde{\mathbf{A}}) \geq k$. Suppose for the sake of contradiction that there exist $k-1$ points $w_1, \ldots, w_{k-1}$ such that

$$\mathsf{dist}(\mathsf{CH}(\widetilde{\mathbf{A}}), \mathsf{CH}(\{w_1, \ldots, w_{k-1}\})) \leq 6\sigma/\sqrt{\delta}.$$

The following claim is similar to Claim 4. Let W_0 denote $\mathsf{CH}(\{w_1, \ldots, w_{k-1}\})$.

Claim 21. $\mathsf{dist}(\mathrm{K}, W_0) \leq 17\sigma/\sqrt{\delta}$.

Proof. As in the upper bound proof, it is enough to show that for every $\ell \in [k]$, $d(M_{.,\ell}, W_0) \leq \rho$, and that for $w_j \in W_0$, $d(w_j, M) \leq \rho$.

Fix an index $\ell \in [k]$. We know by (9) that $|M_{.,\ell} - P_{.,S_\ell}| \leq \frac{4\sigma}{\sqrt{\delta}}$, and Claim 2 implies that $|A_{.,S_\ell} - P_{.,S_\ell}| \leq \frac{\sigma}{\sqrt{\delta}}$. Therefore, $|M_{.,\ell} - A_{.,S_\ell}| \leq \frac{5\sigma}{\sqrt{\delta}}$ (using (10)). Finally, the fact that $\mathsf{dist}(\mathsf{CH}(\widetilde{\mathbf{A}}), W) \leq 6\sigma/\sqrt{\delta}$ implies that there is a point $w \in W$ such that $|A_{.,S_\ell} - w| \leq 6\sigma/\sqrt{\delta}$. It follows that $|M_{.,\ell} - w| \leq 11\sigma/\sqrt{\delta}$.

Now we show the other direction. Fix an index $\ell \in [k-1]$. Since $\mathsf{dist}(W_0, \mathsf{CH}(\widetilde{\mathbf{A}})) \leq 6\sigma/\sqrt{\delta}$, $d(w_\ell, \mathsf{CH}(\widetilde{\mathbf{A}})) \leq 6\sigma/\sqrt{\delta}$. The argument used above for proving upper bound shows that $d(\mathsf{CH}(\widetilde{\mathbf{A}}), \mathsf{CH}(\{A_{.,S_1}, \ldots, A_{.,S_k}\}) \leq 6\sigma/\sqrt{\delta}$. It follows that $d(w_\ell, \mathsf{CH}(\{A_{.,S_1}, \ldots, A_{.,S_k}\}) \leq 12\sigma/\sqrt{\delta}$. Therefore there exist non-negative $\lambda_{\ell'}, \ell' \in [k]$, adding up to 1 such that $|w_\ell - \sum_{\ell' \in [k]} \lambda_{\ell'} A_{.,S_{\ell'}}| \leq 12\sigma/\sqrt{\delta}$. Condition (9) and Claim 2 now imply that

$$|w_\ell - \sum_{\ell' \in [k]} \lambda_{\ell'} M_{.,\ell'}| \leq \frac{12\sigma}{\sqrt{\delta}} + \frac{4\sigma}{\sqrt{\delta}} + \frac{\sigma}{\sqrt{\delta}} = \frac{17\sigma}{\sqrt{\delta}}.$$

This proves the desired result. □

Now observe that if we set $\rho := 17\sigma/\sqrt{\delta}$, then the RHS of (8) is at least (using (10))

$$\frac{30\sigma}{\delta^2} \geq 7\rho,$$

since $\delta \leq 0.5$ (assuming $k \geq 2$). So the conditions in Theorem 5 hold with $\mathbf{R} = \mathbf{M}, W = \{w_1, \ldots, w_{k-1}\}$ and ρ as defined above. But then Theorem 5 implies that $|W| \geq k$, a contradiction. This proves that $\texttt{ICR}(6\sigma/\delta, \widetilde{\mathbf{A}}) \geq k$.

We have shown that $\texttt{ICR}(6\sigma/\delta, \widetilde{\mathbf{A}})$ has to be at least and at most k, and so, must be equal to k. ☐

C. Applications to Exact NMF

In this section, we show that the notion of ICR in the context of NMF reveals the inner dimension of the data matrix.

As mentioned in the introduction, the data matrix $\mathbf{A}$ consists of non-negative entries, and we seek a factorization into non-negative matrices $\mathbf{M}, \mathbf{W}$ of the form (subscripts denote dimensions of the matrices):

$$\mathbf{A}_{d\times n} = \mathbf{M}_{d\times k}\mathbf{W}_{k\times n}.$$

The parameter k above is called the inner dimension of the factorization $\mathbf{A} = \mathbf{M}\mathbf{W}$. Note that neither the factorization, nor, even the value of k is unique. For example, there are always the trivial NMFs: $\mathbf{A} = \mathbf{A}\mathbf{I} = \mathbf{I}\mathbf{A}$ which have inner dimensions n, d respectively. In what follows, we motivate our definition of ICR by showing that in the ideal situation of exact NMF, with a condition called separability (see below), ICR yields the correct inner dimension. We first show that the columns of the above matrices can be assumed to be normalized to 1.

Fact 22. *In seeking non-neg factorization of* $\mathbf{A}$ *into* $\mathbf{M}\mathbf{W}$*, we may assume without loss of generality that each column of* $\mathbf{A}, \mathbf{W}, \mathbf{M}$ *sums to 1.*

Proof. Suppose $\mathbf{A} = \mathbf{M}\mathbf{W}$. Let $\Delta_\ell := \sum_{i=1}^d M_{i,\ell}$ denote the sum of all the entries in the ℓ^{th} column of $\mathbf{M}$. Dividing ℓ^{th} column of $\mathbf{M}$ and multiplying ℓ^{th} row of $\mathbf{W}$ by Δ_ℓ leaves the product $\mathbf{M}\mathbf{W}$ unchanged.

The above transformation ensures that each column of $\mathbf{M}$ adds to 1. Now, for each $j \in [n]$, the sums of entries in the columns $A_{\cdot,j}$ and $W_{\cdot,j}$ are identical, call this quantity γ_j. By dividing all the entries in the j^{th} column of $\mathbf{A}$ and $\mathbf{W}$ by α_j ensures that the column sums in both $\mathbf{A}$ and $\mathbf{W}$ are also 1. ☐

We shall call a NMF $\mathbf{A} = \mathbf{M}\mathbf{W}$ *normalized* if the columns of each of these matrices sum to 1. The above claim shows that we can assume this condition without loss of generality. Henceforth, we shall seek such normalized NMF only. Once normalized, a column of $\mathbf{A}$ is a non-negative combination of other columns iff it is also a convex combination.

We now define the notion of separability.

Definition 7. *An NMF* $\mathbf{A} = \mathbf{M}\mathbf{W}$ *of inner dimension* k *is said to be separable if each unit vector* $e_1, e_2, \ldots, e_k \in \Re^k$ *is a column of* $\mathbf{W}$.

The result below presents a situation where the notion of ICR is more appropriate than non-negative factorization. For sake of brevity, we shall use $\texttt{ICR}(\mathbf{B})$ to denote $\texttt{ICR}(0, \mathbf{B})$ for a matrix $\mathbf{B}$.

Lemma 23. *Suppose* $\mathbf{A}$ *has a normalized, separable factorization* $\mathbf{M}\mathbf{W}$ *with inner dimension* k *and also* $\texttt{ICR}(0, \mathbf{M}) = k$*. Then,* $\texttt{ICR}(0, \mathbf{A})$ *equals* k*.*

Further, for each $k \geq 4$*, there is an* $\mathbf{A}$ *with NNR*$(\mathbf{A}) = 3$*, but there is a normalized, separable NMF of* $\mathbf{A}$ *of inner dimension* k*.*

Remark: Note that $\texttt{ICR}(\mathbf{M}) = k$ is without loss of generality, since we are assuming that the columns of $\mathbf{M}$ are extreme points of $\mathsf{CH}(\mathbf{M})$. We also note that the condition of Separability was introduced in (Donoho & Stodden, 2003b) especially to get at the "correct" NMF. But for getting the correct NMF, they also needed additional conditions besides separability. What the simple Lemma here shows is that indeed separability alone is sufficient to pin down k (not necessarily the factorization) in the ideal situation of exact NMF. In the more realistic approximate-NMF situation, Theorem 1 shows a similar result, but the proof is not nearly that simple.

Proof. $\texttt{ICR}(\mathbf{A})$ is at most k since there is a separable factorization of inner dimension k. Supposed for sake of contradiction that $\texttt{ICR}(\mathbf{A}) \leq k-1$. So, there are $k-1$ columns of $\mathbf{A}$ whose convex hull contains all the columns of $\mathbf{A}$. This implies the number of extreme points of $\mathsf{CH}(\mathbf{A})$ is at most $k-1$. But $\mathsf{CH}(\mathbf{A}) = \mathsf{CH}(\mathbf{M})$ since the factorization is separable and since $\texttt{ICR}(\mathbf{M}) = k$, the number of extreme points of $\mathsf{CH}(\mathbf{M})$ is k producing a contradiction. This proves the first assertion in the lemma.

For the second assertion in the lemma, consider the following example. We set $d = 2$, $k = 100$ and the columns of $\mathbf{M}$ are the 100 vertices of a regular 100-polygon of side 1 in the plane whose center is $(10^6, 10^6)$. It is contained in the triangle of side 10000 centered at $(10^6, 10^6)$, so NNR is at most 3. ☐

D. Applications to Inexact NMF

In this section, we assume that the latent data matrix $\mathbf{P}$ can be written as $\mathbf{M}\mathbf{W}$, but the actual data matrix $\mathbf{A}$ is a perturbation of $\mathbf{P}$. In this sense, NMF is a special case of LkP. We prove Theorem 9 in several parts. We first show (in Corollary 25) that under suitable assumptions, Algorithm 1 finds the correct value of k. Subsequently, in Theorem 26, we show that Algorithm 2 uses this value of k

to find approximations to the vertices of $\mathbf{M}$.

We now restate Lemma 19 formally and prove it.

Lemma 24. *Suppose the Dominant Features assumption as stated below holds:*

$$\exists T_1, T_2, \ldots, T_k \textit{ disjoint } : \forall \ell, \sum_{i \in T_\ell} M_{i,\ell} \geq 4 \sum_{i \in T_\ell} \sum_{\ell' \neq \ell} M_{i,\ell'} \tag{13}$$

$$\forall \ell, \frac{\sum_{i \in T_\ell} M_{i,\ell}}{\sqrt{|T_\ell|}} \geq 8\delta \cdot \textit{Max}_{\ell'} |M_{\cdot,\ell'}| \tag{14}$$

Then, condition (8) *of Theorem* 9 *holds.*

Remark We point out that the Dominant Features assumption is reasonable. But first, we bring the reader's attention to the fact that there are two parts of this assumption; the full assumption is stated above and an abbreviated version in the main text. The first part says that summed over the dominant features of each ℓ, the other columns have a smaller sum. Separability like assumptions, would make the other sums 0, so what we have here is a weaker requirement. For the second condition, note that if $\sum_{i \in T_\ell} M_{i,\ell}$ is a constant fraction of $\text{Max}_\ell \sum_{i=1}^d M_{i,\ell}$ and $|T_\ell| \in O(1/\delta^2)$, then, indeed the second assumption holds. The setting $|T_\ell| = O(1/\delta^2)$ is reasonable – we expect a small number of important features. In the special case of Topic Modelling, this says a small number of words together have a constant fraction frequency which is reasonable.

Proof. Let $|T_1| + |T_2| + \cdots + |T_k| = m$. Reorder $i = 1, 2, \ldots, d$ so that the first $|T_1|$ are from T_1, the next $|T_2|$ are from T_2 etc. and so on until T_k and the last $d - m$ don't belong to any T_ℓ. Now, let $\widehat{\mathbf{M}}$ be the $m \times k$ matrix obtained by deleting the last $d - m$ rows from $\mathbf{M}$. It suffices to prove that

$$\left| \mathbf{proj}(\widehat{M}_{\cdot,\ell}, \mathbf{Null}(\widehat{M}_{\cdot,\ell'}, \ell' \neq \ell)) \right| \geq \delta |M_{\cdot,\ell}|. \tag{15}$$

This can be seen by noting that

$$\begin{aligned} &|\mathbf{proj}(M_{\cdot,\ell}, \mathbf{Null}(M_{\cdot,\ell'}, \ell' \neq \ell))| \\ &\quad = \text{Max}_{u \in \mathbf{Null}(M_{\cdot,\ell'}, \ell' \neq \ell), |u|=1} |u \cdot M_{\cdot,\ell}|, \end{aligned} \tag{16}$$

and taking u to be the vector obtained from u' by appending $d - m$ zeros, where, $u' = \mathbf{proj}(\widehat{M}_{\cdot,\ell}, \mathbf{Null}(\widehat{M}_{\cdot,\ell'}, \ell' \neq \ell))$, normalized to length 1. Now, we focus on proving (15).

Let

$$\gamma_\ell = \sum_{i \in T_\ell} M_{i,\ell}.$$

We now define a $k \times m$ matrix $\mathbf{M}$, where the columns of $\mathbf{B}$ are indexed by the sets $T_1, \ldots, T_k$ in this order (note that $|T_1| + \ldots + |T_k| = m$). For each $\ell \in [k]$, the entries $B_{\ell,j} = 1/\gamma_\ell$, for all $j \in T_\ell$. Rest of the entries of $\mathbf{B}$ are 0. It is easy to see that

$$\mathbf{B}\widehat{\mathbf{M}} = \mathbf{I}_{k \times k} - \Delta, \text{ where,}$$

$$\Delta_{\ell,\ell'} = \begin{cases} 0 & \text{for } \ell' = \ell \\ -\frac{1}{\gamma_\ell} \sum_{i \in T_\ell} M_{i,\ell'} & \text{for } \ell \neq \ell'. \end{cases}.$$

For an $s \times t$ matrix $\mathbf{D}$ and $\ell \in [s]$, let $\rho_\ell(\mathbf{D})$ denote $\sum_{\ell'=1}^t |D_{\ell,\ell'}|$. The first Dominant Features Assumption (13) implies that $\rho_\ell(\Delta) \leq 1/4$. So the matrix $\mathbf{I} - \Delta$ is diagonal dominant and hence invertible. Let $\mathbf{C}$ denote $(\mathbf{I} - \Delta)^{-1}$. Then, by power series expansion, we have

$$\mathbf{C} = (I + \Delta + \Delta^2 + \cdots).$$

It is easy to see that $\rho_\ell(\Delta^r) \leq 1/4^r$ by simple induction. So we have

$$\rho_\ell(\mathbf{C}) \leq 1 + (1/4) + (1/4^2) + \ldots = 4/3. \tag{17}$$

Now, we have

$$\mathbf{C}\mathbf{B}\widehat{\mathbf{M}} = \mathbf{I}.$$

Consider

$$u = (\mathbf{CB})_{\ell,\cdot} / |(\mathbf{CB})_{\ell,\cdot}|$$

and apply (16) (with $\widehat{\mathbf{M}}$ instead of $\mathbf{M}$) to get

$$\begin{aligned} &\left| \mathbf{proj}(\widehat{M}_{\cdot,\ell}, \mathbf{Null}(\widehat{M}_{\cdot,\ell'}, \ell' \neq \ell)) \right| \geq u \cdot \widehat{M}_{\cdot,\ell} \\ &= \frac{1}{|(\mathbf{CB})_{\ell,\cdot}|} \geq \frac{1}{|\mathbf{C}_{\ell,\cdot}| \cdot ||\mathbf{B}||} \overset{(17)}{\geq} \frac{3}{4} \cdot \min_{\ell'} \frac{\gamma_{\ell'}}{\sqrt{|T_{\ell'}|}} \geq 2\delta |M_{\cdot,\ell}|, \end{aligned}$$

by the second Dominant Features assumption (14). This completes the proof of the lemma. ☐

As a corollary we get the following result.

Corollary 25. *Assuming conditions* (9) *and* (10) *and also the Dominant Features assumption, Algorithm 1 finds k in polynomial time.*

Proof. Lemma 24 shows that all the conditions in the statement of Theorem 9 are met. Therefore, the result follows from Theorem 9. In order to complete the proof of Thereom 20, it remains to analyze Algorithm 2, which we do next. ☐

D.1. Analysis of Algorithm 2

The following result states the guarantees obtained by Algorithm 2.

Theorem 26. *Suppose that the input data satisfies the following assumptions. There exists a $\delta \in (0, 1)$ such that*

- **Well-Separatedness**: *Every vertex $M_{\cdot,\ell}$ of K satisfies the following condition:*

$$|\mathbf{proj}\left(M_{\cdot,\ell}, \mathbf{Null}\left(\mathbf{M} \setminus M_{\cdot,\ell}\right)\right)| \geq \delta \max_{\ell'} |M_{\cdot,\ell'}|. \tag{18}$$

- **Proximity**: *For all $\ell \in [k]$, let S_ℓ denote the following set of indices:*

$$S_\ell := \{j : |M_{\cdot,\ell} - P_{\cdot,j}| \leq \frac{4\sigma}{\sqrt{\delta}}\}. \tag{19}$$

Then $|S_\ell| \geq \delta n$.

- **Spectrally Bounded Perturbations**: *For every $\ell \in [k]$*

$$\sigma \leq \delta^{13} |M_{\cdot,\ell}|/10000. \tag{20}$$

Let δ_1 denote $\frac{150k^4\sigma}{\delta^{1.5}}$. Given k, Algorithm 2 finds, with high probability, k points each of which is within distance δ_1 of a unique vertex of $\mathbf{M}$.

In the rest of this section, we prove this theorem. The algorithm will maintain the following invariants at the beginning of each iteration $r, r \leq k$:

There are r distinct indices $\ell_1, \ell_2, \ldots, \ell_r \in [k]$ such that for all $t = 1, \ldots, r$

$$|\hat{A}_{\cdot,R_t} - M_{\cdot,\ell_t}| \leq \delta_1. \tag{21}$$

We prove this statement by induction on r. When $r = 0$, the statement follows trivially. Now assume that the above statement holds at the beginning of iteration $r, r < k$. We now show that $\hat{A}_{\cdot,r+1}$ will be close to a new vertex of K. We give some notation first. Let $\widetilde{\mathbf{M}}$ and $\widetilde{\mathbf{A}}$ denote the following $d \times r$ matrices:

$$\widetilde{\mathbf{M}} = (M_{\cdot,\ell_1} \mid M_{\cdot,\ell_2} \mid \ldots |M_{\cdot,\ell_r})$$
$$\widetilde{\mathbf{A}} = (\hat{A}_{\cdot,R_1} \mid \hat{A}_{\cdot,R_2} \mid \ldots |\hat{A}_{\cdot,R_r})$$

The invariant condition (21) implies that

$$||\widetilde{\mathbf{M}} - \widetilde{\mathbf{A}}|| \leq \delta_1 \sqrt{k} \tag{22}$$

We now give lower bounds on the singular values of $\widetilde{\mathbf{M}}$.

Lemma 27.

$$s_r(\widetilde{\mathbf{M}}) \geq s_k(\mathbf{M}) \geq \frac{10000\sigma}{\delta^{12}\sqrt{k}}.$$

Proof.

$$s_r(\widetilde{\mathbf{M}}) = \min_{|x|=1} |\widetilde{\mathbf{M}}x| \geq \min_{|y|=1} |\mathbf{M}y| = s_k(\mathbf{M}).$$

Now, $s_k(\mathbf{M})$ is $\min_{x:|x|=1} |\mathbf{M}x|$. Consider any unit vector $x \in \Re^k$. There must be an index ℓ such that $|x_\ell| \geq \frac{1}{\sqrt{k}}$. But then $|\mathbf{M} \cdot x| \geq |\mathbf{proj}(\mathbf{M} \cdot x, \mathbf{Null}(\mathbf{M} \setminus M_{\cdot,\ell})| = |x_\ell||\mathbf{proj}(M_{\cdot,\ell}, \mathbf{Null}(\mathbf{M} \setminus M_{\cdot,\ell})| \geq \frac{\delta}{\sqrt{k}} \cdot |M_{\cdot,\ell}|$, by (18). The result now follows from (20). □

We now show the following consequence of **Well-Separatedness**.

Claim 28. *Let ℓ, ℓ' be two distinct indices in $[k]$, such that $\ell, \ell' \notin \{\ell_1, \ldots, \ell_r\}$. Then,*

$$|\mathbf{proj}(M_{\cdot,\ell} - M_{\cdot,\ell'}, \mathbf{Null}(\widetilde{\mathbf{M}})| \geq \delta \cdot Max_{\ell''} |M_{\cdot,\ell''}|.$$

Proof. By definition,

$$|\mathbf{proj}(M_{\cdot,\ell} - M_{\cdot,\ell'}, \mathbf{Null}(\widetilde{\mathbf{M}})| = \mathrm{Min}_x |M_{\cdot,\ell} - M_{\cdot,\ell'} - \widetilde{\mathbf{M}}x|,$$

where the minimum is taken over all vectors x. Now, for any vector x, $M_{\cdot,\ell'} - \widetilde{\mathbf{M}}x$ lies in the span of $\{M_{\cdot,\ell''} : \ell'' \neq \ell, \ell'' \in [k]\}$. Hence, the expression in the RHS above is at least $|\mathbf{proj}(M_{\cdot,\ell}, \mathbf{Null}(\mathbf{M} \setminus M_{\cdot,\ell})|$. The desired result now follows from (18). □

The following property of $\widetilde{\mathbf{A}}$ is easy to see:

Claim 29. *The matrix $\widetilde{\mathbf{A}}$ is a full rank matrix. Further,*

$$s_r(\widetilde{\mathbf{A}}) \geq \frac{5000\sigma}{\delta^{12}\sqrt{k}}$$

Proof. Consider an index $j \leq r$. Since $\hat{A}_{\cdot,R_j}$ is not in $\mathbf{Span}(\hat{A}_{\cdot,R_1}, \ldots, \hat{A}_{\cdot,R_{j-1}}$, $\widetilde{\mathbf{A}}$ is full rank. Now,

$$s_r(\widetilde{\mathbf{A}}) \geq s_r(\widetilde{\mathbf{M}}) - ||\widetilde{\mathbf{A}} - \widetilde{\mathbf{M}}|| \overset{(22)}{\geq} s_k(\mathbf{M}) - \delta_1\sqrt{k},$$

where the second inequality follows from the fact that $\widetilde{\mathbf{M}}$ is obtained from $\mathbf{M}$ by removing some of the columns. The desired result now follows from the definition of δ_1 and Lemma 27. □

Recall the definition of the subspace U_r in Algorithm 2: $V \cap \mathbf{Null}(\hat{A}_{\cdot,R_1}, \ldots, \hat{A}_{\cdot,R_r})$. The subspace V is supposed contain the span of $\mathbf{M}$ and $\mathbf{Null}(\hat{A}_{\cdot,R_1}, \ldots, \hat{A}_{\cdot,R_r})$ is supposed to be a close approximation to $\mathbf{Null}(M_{\cdot,\ell_1}, \ldots, M_{\cdot,\ell_r})$. Therefore, we would expect the subspace U_r to contain a close approximation to $\mathbf{Span}(\mathbf{M}) \cap \mathbf{Null}(\widetilde{\mathbf{M}})$. The following key lemma formalises this intuition.

Lemma 30. *For every unit vector $x \in \mathbf{Span}(\mathbf{M}) \cap \mathbf{Null}(\widetilde{\mathbf{M}})$, there is a $y \in U_r$ such that*

$$|x - y| \leq 1/1600.$$

Proof. Let $S_{\ell_1}, \ldots, S_{\ell_r}$ be as guaranteed by condition (19) in the **Proximity** assumption. Define

$$\mathbf{B} = (\hat{A}_{\cdot,S_1} \mid \hat{A}_{\cdot,S_2} \mid \ldots |\hat{A}_{\cdot,S_k}).$$

Consider a unit vector $x \in \mathbf{Span}(\mathbf{M}) \cap \mathbf{Null}(\widetilde{\mathbf{M}})$. So $x = \mathbf{M}w$ for some vector w. It follows that

$$1 = |x| \geq s_k(\mathbf{M})|w|.$$

Therefore,

$$|\mathbf{M}w - \mathbf{B}w| \leq ||\mathbf{M} - \mathbf{B}||\ |w| \leq ||\mathbf{M} - \mathbf{B}||/s_k(\mathbf{M}). \tag{23}$$

Inequality (9) implies that

$$||\mathbf{M} - \mathbf{B}|| \leq \frac{4\sigma\sqrt{k}}{\sqrt{\delta}}.$$

Lemma 27 implies that $s_k(\mathbf{M}) \geq \frac{10000\sigma}{\delta^{12}\sqrt{k}}$. Therefore, using (23), we get

$$|\mathbf{M}w - \mathbf{B}w| \leq k\delta^{11.5}/2500. \tag{24}$$

We are trying to prove that there is a point in U_r close to $x = \mathbf{M}w$. Inequality (24) shows that it is enough to find a point in U_r which is close to $\mathbf{B}w$. A natural candidate for this point is the projection of $\mathbf{B}w$ onto U_r; note that $\mathbf{B}w \in V$, so its projection onto U_r is obtained by subtracting its component in $\mathbf{Span}(\widetilde{\mathbf{A}})$, namely, the component is (Claim 29 implies that $\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}$ is invertible):

$$y = \mathbf{B}w - \widetilde{\mathbf{A}}(\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{A}}^T\mathbf{B}w.$$

Now,

$$\begin{aligned}|y - x| &= |y - \widetilde{\mathbf{M}}w| \leq |y - \mathbf{B}w| + |\widetilde{\mathbf{M}}w - \mathbf{B}w| \\ &\overset{(24)}{\leq} |\widetilde{\mathbf{A}}(\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{A}}^T\mathbf{B}w| + \frac{k\delta^{11.5}}{2500}\end{aligned} \tag{25}$$

It remains to bound the first term above. Now,

$$\begin{aligned}|\widetilde{\mathbf{A}}(\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{A}}^T\mathbf{B}w| &\leq \left|\widetilde{\mathbf{A}}(\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{A}}^T(\mathbf{B}w - \mathbf{M}w)\right| \\ &\quad + \left|\widetilde{\mathbf{A}}(\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{A}}^T\mathbf{M}w\right| \\ &\leq |\mathbf{B}w - \mathbf{M}w| \\ &\quad + |\widetilde{\mathbf{A}}(\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}})^{-1}(\widetilde{\mathbf{A}}^T - \widetilde{\mathbf{M}}^T)\mathbf{M}w| \\ &\leq |\mathbf{B}w - \mathbf{M}w| + \frac{1}{s_r(\hat{\mathbf{A}})}||\widetilde{\mathbf{A}} - \widetilde{\mathbf{M}}||.\end{aligned}$$

The second inequality above uses the fact that $\widetilde{\mathbf{A}}(\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{A}}$ is a projection matrix, and hence has spectral norm at most 1; and $\widetilde{\mathbf{M}}^T\mathbf{M}w = \widetilde{\mathbf{M}}^t x = 0$ since $x \in \mathbf{Null}(\widetilde{\mathbf{M}})$. The third inequality uses the fact that $||\widetilde{\mathbf{A}}(\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}})^{-1}|| = \frac{1}{s_r(\widetilde{\mathbf{A}})}$ and $|\mathbf{M}w| = |x| = 1$. The invariant condition (23) implies that $||\widetilde{\mathbf{M}} - \widetilde{\mathbf{A}}|| \leq \delta_1\sqrt{k}$. Combining this with Claim 29 and (24), we see that the RHS above is at most

$$\frac{k\delta^{11.5}}{5000} + \frac{\delta_1\delta^{12}k}{10000\sigma}.$$

The desired result now follows from the definition of δ_1. $\square$

Let u_r be the vector selected during iteration r of Algorithm 2.

Lemma 31. *With probability at least $1 - (c/k^{3/2})$, the following hold:*

$$\begin{aligned}&\forall \ell, \ell' \notin \{\ell_1, \ell_2, \ldots, \ell_r\}, \ell \neq \ell' : \\ &\quad |u_r \cdot (M_{\cdot,\ell} - M_{\cdot,\ell'})| \geq \frac{0.097\delta}{k^4} Max_{\ell''}|M_{\cdot,\ell''}|. \\ &\forall \ell \notin \{\ell_1, \ell_2, \ldots, \ell_r\} : \\ &\quad |u_r \cdot (M_{\cdot,\ell})| \geq \frac{0.09989\delta}{k^4} Max_{\ell'}|M_{\cdot,\ell'}|.\end{aligned}$$

Proof. We can write

$$M_{\cdot,\ell} = \underbrace{\mathbf{proj}(M_{\cdot,\ell}, \mathbf{Null}(\widetilde{\mathbf{M}}))}_{q_\ell} + \underbrace{\mathbf{proj}(M_{\cdot,\ell}, \mathbf{Span}(\widetilde{\mathbf{M}}))}_{p_\ell = \widetilde{\mathbf{M}}w^{(\ell)}},$$

where we use the fact that q_ℓ can be written as $M_{\cdot,\ell} - \widetilde{\mathbf{M}}w^{(\ell)}$ for some $w^{(\ell)}$.

Since $|p_\ell| \leq |M_{\cdot,\ell}|$, Lemma 27 implies:

$$|w^{(\ell)}| \leq |p_\ell|/s_r(\widetilde{\mathbf{M}}) \leq \frac{|M_{\cdot,\ell}|\delta^{12}\sqrt{k}}{10000\sigma} \tag{26}$$

Recall that u_r is a random unit length vector in subspace U_r. Now,

$$\begin{aligned}u_r \cdot M_{\cdot,\ell} &= u_r \cdot q_\ell + u_r^T\widetilde{\mathbf{M}}w^{(\ell)} \\ &= u_r \cdot \mathbf{proj}(q_\ell, U_r) + u_r^T(\widetilde{\mathbf{M}} - \widetilde{\mathbf{A}})w^{(\ell)},\end{aligned}$$

since $u_r^T\widetilde{\mathbf{A}} = 0$. So,

$$\begin{aligned}&|u_r \cdot M_{\cdot,\ell} - u_r \cdot \mathbf{proj}(q_\ell, U_r)| \leq ||(\widetilde{\mathbf{M}} - \widetilde{\mathbf{A}})w^{(\ell)}|| \\ &\quad \leq ||\widetilde{\mathbf{M}} - \widetilde{\mathbf{A}}|||w^{(\ell)}| \leq \delta^{10.5}k^5|M_{\cdot,\ell}|/60 \leq \delta^{5.5}|M_{\cdot,\ell}|/60,\end{aligned} \tag{27}$$

using (22) and (26). Similarly, for $\ell' \neq \ell$,

$$u_r \cdot (M_{\cdot,\ell} - M_{\cdot,\ell'}) = u_r \cdot \mathbf{proj}(q_\ell - q_{\ell'}, U) + u_r^T\widetilde{\mathbf{M}}(w^{(\ell)} - w^{(\ell')}).$$

Since $u_r^T \widetilde{A} = 0$,

$$|u_r \cdot (M_{\cdot,\ell} - M_{\cdot,\ell'}) - u_r \cdot \mathbf{proj}(q_\ell - q_{\ell'}, U)| \\ \leq |u_r^T(\widetilde{\mathbf{M}} - \widetilde{\mathbf{A}})(w^{(\ell)} - w^{(\ell')})|$$

Arguing as in (27),

$$|u_r \cdot (M_{\cdot,\ell} - M_{\cdot,\ell'}) - u_r \cdot \mathbf{proj}(q_\ell - q_{\ell'}, U)| \\ \leq ||\widetilde{\mathbf{M}} - \widetilde{\mathbf{A}}|||w^{(\ell)} - w^{(\ell')}| \\ \leq \quad \delta^{5.5}|M_{\cdot,\ell} - M_{\cdot,\ell'}|/60. \tag{28}$$

Now, u_r is a random unit length vector in U_r, and $\mathbf{proj}(q_\ell, U_r), \mathbf{proj}(q_\ell - q_{\ell'}, U_r), \ell, \ell' \in [k]$ are fixed vectors in U_r (and the choice of u_r doesn't dependent on them). Consider the following event $\mathcal{E}$:

$$\mathcal{E} : \forall \ell : |u_r \cdot \mathbf{proj}(q_\ell, U_r)| \geq \frac{1}{10k^4}|\mathbf{proj}(q_\ell, U_r)| \text{ AND}$$

$$\forall \ell \neq \ell' : |u_r \cdot \mathbf{proj}(q_\ell - q_{\ell'}, U_r)| \geq \frac{1}{10k^4}|\mathbf{proj}(q_\ell - q_{\ell'}, U)|.$$

The negation of $\mathcal{E}$ is the union of at most k^2 events (for each ℓ and each ℓ, ℓ') and each of these has a failure probability of at most $1/10k^{3.5}$ (since the $k-1$ volume of $\{x \in U_r : u_r \cdot x = 0\}$ is at most $\sqrt{k}$ times the volume of the unit ball in U). Thus, we have:

$$\mathsf{Pr}(\mathcal{E}) \geq 1 - \frac{1}{10k^{1.5}}. \tag{29}$$

We pay the failure probability and assume from now on that $\mathcal{E}$ holds.

By Lemma 30, we have that there is a $q'_\ell \in U_r$ with $|q'_\ell - q_\ell| \leq |q_\ell|/1600$ which implies (recall $k \geq 2$):

$$|q_\ell - \mathbf{proj}(q_\ell, U_r)| \leq \frac{|q_\ell|}{1600} \implies \\ |\mathbf{proj}(q_\ell, U_r)| \geq .9999|q_\ell|. \tag{30}$$

So, under $\mathcal{E}, \forall \ell \notin \{\ell_1, \ell_2, \ldots, \ell_r\}$

$$|u_r \cdot \mathbf{proj}(q_\ell, U_r)| \geq |\mathbf{proj}(q_\ell, U_r)|\frac{1}{10k^4} \\ \geq \frac{.09999|q_\ell|}{k^4} \geq \frac{.09999\delta|M_{\cdot,\ell}|}{k^4}, \tag{31}$$

since $|q_\ell| \geq |\mathbf{proj}(M_{\cdot,\ell}, \mathbf{Null}(\mathbf{M} \setminus M_{\cdot,\ell}))| \geq \delta|M_{\cdot,\ell}|$ by (8).

By (27) and (31), $\forall \ell \notin \{\ell_1, \ell_2, \ldots, \ell_r\}$,

$$|u_r \cdot M_{\cdot,\ell}| \geq |u_r \cdot \mathbf{proj}(q_\ell, U_r)| - \frac{\delta^{5.5}|M_{\cdot,\ell}|}{60} \geq \frac{.09989\delta|M_{\cdot,\ell}|}{k^4}, \tag{32}$$

proving the second assertion of the Lemma.

Now we prove the first assertion. For $\ell \notin \{\ell_1, \ell_2, \ldots, \ell_r\}$ and $\ell' \notin \{\ell, \ell_1, \ell_2, \ldots, \ell_r\}$, by (28),

$$|u_r \cdot (M_{\cdot,\ell} - M_{\cdot,\ell'})| \geq |u_r \cdot \mathbf{proj}(q_\ell - q_{\ell'}, U_r)| - \frac{\delta^{5.5}|M_{\cdot,\ell} - M_{\cdot,\ell'}|}{60} \\ \geq \frac{1}{10k^4}|\mathbf{proj}(q_\ell - q_{\ell'}, U_r)| - \frac{\delta^{5.5}|M_{\cdot,\ell} - M_{\cdot,\ell'}|}{60} \quad \text{by } \mathcal{E}.$$

By Lemma 30, there exists a vector $x \in U_r$, such that $|x - (q_\ell - q_{\ell'})| \leq \frac{|q_\ell - q_{\ell'}|}{1600}$. Thus, we have

$$|\mathbf{proj}(q_\ell - q_{\ell'}, U_r)| \geq .99|q_\ell - q_{\ell'}| \geq .99\delta \text{Max}_{\ell''}|M_{\cdot,\ell''}|,$$

by Claim 28. This finishes the proof of the first assertion and of the Lemma. □

Claim 32. *For all $S \subseteq [n]$,*

$$|\hat{A}_{\cdot,S} - P_{\cdot,S}| \leq \frac{2\sigma\sqrt{n}}{\sqrt{|S|}}.$$

Proof. Let $\hat{\mathbf{A}}$ denote the matrix whose columns are $\hat{A}_{\cdot,1}, \ldots, \hat{A}_{\cdot,n}$. Since $\mathbf{P}$ has rank at most $1/\delta$, it follows that

$$||\hat{\mathbf{A}} - \mathbf{P}|| \leq ||\hat{\mathbf{A}} - \mathbf{A}|| + ||\mathbf{A} - \mathbf{P}|| \leq 2||\mathbf{A} - \mathbf{P}|| = 2\sigma\sqrt{n}.$$

Let $\mathbf{1}_S$ be the indicator vector for the subset S.

$$|\hat{A}_{\cdot,S} - P_{\cdot,S}| = \frac{1}{|S|}|(\hat{\mathbf{A}} - \mathbf{P})\mathbf{1}_S| \leq \frac{1}{|S|}||\hat{\mathbf{A}} - \mathbf{P}|| \cdot |\mathbf{1}_S| \leq \frac{2\sigma\sqrt{n}}{\sqrt{|S|}}.$$

□

Lemma 33. *Define ℓ by:*

$$\ell = \begin{cases} \arg\max_{\ell'} u_r \cdot M_{\cdot,\ell'} & \text{if } u_r \cdot \hat{A}_{\cdot,R_{r+1}} \geq 0 \\ \arg\min_{\ell'} u_r \cdot M_{\cdot,\ell'} & \text{if } u_r \cdot \hat{A}_{\cdot,R_{r+1}} < 0 \end{cases}.$$

Then, $\ell \notin \{\ell_1, \ell_2, \ldots, \ell_r\}$ and

$$|\hat{A}_{\cdot,R_{r+1}} - M_{\cdot,\ell}| \leq \delta_1.$$

Proof. **Case 1** $u_r \cdot \hat{A}_{\cdot,R_{r+1}} \geq 0$.

Let

$$\ell = \arg\max_{\ell'} u_r \cdot M_{\cdot,\ell'}.$$

We claim that $\ell \notin \{\ell_1, \ell_2, \ldots, \ell_r\}$. Suppose for contradiction, $\ell \in \{\ell_1, \ell_2, \ldots, \ell_r\}$; wlg, say $\ell = \ell_1$. Then, by induction hypothesis (21), we have that $|\hat{A}_{\cdot,R_1} - M_{\cdot,\ell_1}| \leq \delta_1$ and so, $u_r \cdot M_{\cdot,\ell_1} \leq u_r \cdot \hat{A}_{\cdot,R_1} + \delta_1 = \delta_1$ (since $u_r \in U_r$ and so $u_r \perp \hat{A}_{\cdot,R_1}$). So, for all ℓ', $u_r \cdot M_{\cdot,\ell'} \leq u_r \cdot M_{\cdot,\ell_1} \leq \delta_1$.

So, for all $R \subseteq [n]$, $P_{\cdot,R}$ which is in CH($\mathbf{M}$), satisfies $u_r \cdot P_{\cdot,R} \leq \delta_1$. So, by Claim 32,

$$u_r \cdot \hat{A}_{\cdot,R_{r+1}} \leq u_r \cdot P_{\cdot,R_{r+1}} + (2\sigma/\sqrt{\delta}) \leq \delta_1 + (2\sigma/\sqrt{\delta}) \leq 2\delta_1 \tag{33}$$

But for any $t \notin \{\ell_1, \ell_2, \ldots, \ell_r\}$, we have with S_t as in (9),

$$\begin{aligned} |u_r \cdot \hat{A}_{\cdot,S_t}| &\geq |u_r \cdot P_{\cdot,S_t}| - (2\sigma/\sqrt{\delta}), \text{ by Claim 32} \\ &\geq |u_r \cdot M_{\cdot,t}| - (6\sigma/\sqrt{\delta}) \qquad (34) \\ &\geq .09989\delta|M_{\cdot,\ell}|/(k^4) - 6\sigma/\sqrt{\delta} \end{aligned}$$

where the last inequality uses Lemma 31. So, $u_r \cdot \hat{A}_{\cdot,R_{r+1}}$ (which maximizes $u_r \cdot \hat{A}_{\cdot,R}$ over all $R, |R| = \delta n$) must be at least $\frac{\delta|M_{\cdot,\ell}|}{11k^4} - \frac{6\sigma}{\sqrt{\delta}}$ contradicting (33) by applying (10). So, $\ell \notin \{\ell_1, \ell_2, \ldots, \ell_r\}$ and we have by Lemma 31,

$$u_r \cdot M_{\cdot,\ell} \geq \frac{.09989\alpha \text{Max}_{\ell'}|M_{\cdot,\ell'}|}{k^4}. \tag{35}$$

Also for $\ell' \notin \{\ell_1, \ldots, \ell_r\}, \ell' \neq \ell$, Lemma 31 implies that

$$u_r \cdot M_{\cdot,\ell'} \leq u_r \cdot M_{\cdot,\ell} - \frac{.097\delta}{k^4}\text{Max}_{\ell''}|M_{\cdot,\ell''}|. \tag{36}$$

Now, for $\ell' \in \{\ell_1, \ell_2, \ldots, \ell_r\}$, wlg, say $\ell' = \ell_1$, we have noting that $|\hat{A}_{\cdot,R_1} - M_{\cdot,\ell_1}| \leq \delta_1$ by (21):

$$\begin{aligned} u_r \cdot M_{\cdot,\ell_1} &\leq u_r \cdot \hat{A}_{\cdot,R_1} + \delta_1 \\ &\overset{(35)}{\leq} u_r \cdot M_{\cdot,\ell} - \frac{.09989\delta \text{Max}_{\ell'}|M_{\cdot,\ell'}|}{k^4} + \delta_1 \\ &\leq u_r \cdot M_{\cdot,\ell} - \frac{.097\delta}{k^4}\text{Max}_{\ell''}|M_{\cdot,\ell''}|, \end{aligned}$$

where the last inequality uses the definition of δ_1.

Now, $P_{\cdot,R_{r+1}}$ is a convex combination of the columns of $\mathbf{M}$; say the convex combination is $P_{\cdot,R_{r+1}} = \mathbf{M}w$. From above, we have:

$$\begin{aligned} &u_r \cdot \hat{A}_{\cdot,R_{r+1}} \leq u_r \cdot P_{\cdot,R_{r+1}} + \frac{2\sigma}{\sqrt{\delta}}, \text{ by Claim 32} \\ &\leq w_\ell(u_r \cdot M_{\cdot,\ell}) + \\ &\quad \sum_{\ell' \neq \ell} \left((u_r \cdot M_{\cdot,\ell}) - \frac{.097\delta}{k^4}\text{Max}_{\ell''}|M_{\cdot,\ell''}| \right) w_{\ell'} \\ &\leq u_r \cdot M_{\cdot,\ell} - \frac{.097\delta}{k^4}\text{Max}_{\ell''}|M_{\cdot,\ell''}|(1 - w_\ell). \qquad (37) \end{aligned}$$

This and (34) imply:

$$(1 - w_\ell)\text{Max}_{\ell''}|M_{\cdot,\ell''}| \leq \frac{62k^4}{\delta}\frac{\sigma}{\sqrt{\delta}}. \tag{38}$$

So,

$$\begin{aligned} |P_{\cdot,R_{r+1}} - M_{\cdot,\ell}| &= \left| (w_\ell - 1)M_{\cdot,\ell} + \sum_{\ell' \neq \ell} w_{\ell'} M_{\cdot,\ell'} \right| \\ &\leq \sum_{\ell' \neq \ell} w_{\ell'}|M_{\cdot,\ell} - M_{\cdot,\ell'}| \\ &\leq 2(1 - w_\ell)\text{Max}_{\ell''}|M_{\cdot,\ell''}| \leq \frac{124k^4}{\delta}\frac{\sigma}{\sqrt{\delta}}. \end{aligned}$$

Now it follows from Claim 32 that $|\hat{A}_{\cdot,R_{r+1}} - M_{\cdot,\ell}| \leq \frac{150k^4}{\delta}\frac{\sigma}{\sqrt{\delta}}$ finishing the proof of the theorem in this case. An exactly symmetric argument proves the theorem in the case when $u \cdot A_{\cdot,S} \leq 0$.

□