Dimension reduction in projective clustering

Rafael Zuolo Coppini Lima

Thesis presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
in partial fulfillment
of the requirements
for the degree of
Master of Science

Program: Ciência da Computação
Advisor: Prof. Dr. Yoshiharu Kohayakawa

During the development of this work the author
was supported by CAPES (Finance Code 001)

São Paulo
June 22, 2022

Dimension reduction
in projective clustering

Rafael Zuolo Coppini Lima

This version of the thesis includes the corrections and modifications suggested by the Examining Committee during the defense of the original version of the work, which took place on June 22, 2022.

A copy of the original version is available at the Institute of Mathematics and Statistics of the University of São Paulo.

Examining Committee:

Prof. Dr. Yoshiharu Kohayakawa (advisor) – IME-USP
Prof. Dr. Eduardo Sany Laber – PUCRJ
Prof. Dr. Felipe Miguel Pait – EP - USP

Resumo

Rafael Zuolo Coppini Lima. **Redução de dimensão para agrupamento projetivo**. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022.

A dimensão dos dados pode ser uma barreira para a eficiência de algoritmos (Nelson, 2020) principalmente em razão da chamada "maldição da dimensão", que impõe dependências exponenciais na dimensão para a complexidade de tempo e/ou espaço dos algoritmos para alguns problemas. Este é o caso, por exemplo, do problema do vizinho mais próximo (Har-Peled, Indyk e Motwani, 2012). É natural então estudar aproximações de soluções dos problemas e formas de reduzir a dimensão das instâncias para tentar quebrar essa maldição. Nosso objetivo é escrever uma dissertação sobre um esquema de redução de dimensão para *clustering (agrupamento) sob a métrica* ℓ_2^2, pondo foco em um esquema de aproximação para um caso particular do problema anterior, chamado *projective clustering* (agrupamento projetivo). A redução de dimensão é feita combinando técnicas aleatorizadas, como o Lema de Johnson e Lindenstrauss, e determinísticas, como a decomposição em valores singulares. Obtém-se uma $(1 + \varepsilon)$-aproximação para o problema do agrupamento projetivo, polinomial no número de pontos e na dimensão. Esta dissertação terá como referências principais quatro artigos: Sarlós, 2006, Feldman, Schmidt e Sohler, 2020, Pratap e Sen, 2018 e Deshpande, Rademacher, Vempala e Wang, 2006. Os resultados apresentados na dissertação serão ou os originais ou versões modificadas, incorporando aprimoramentos recentes.

Palavras-chave: Clustering. Agrupamento projetivo. Redução de dimensão. Aproximação. Decomposição em valores singulares. Lema de Johnson e Lindenstrauss.

Abstract

Rafael Zuolo Coppini Lima. **Dimension reduction in projective clustering**. Thesis (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2022.

The high dimensionality of data may be a barrier to algorithmic efficiency (Nelson, 2020), mainly because of the well known "curse of dimensionality" which imposes exponential time and/or memory complexity for algorithms, such as the *nearest neighbour* problem (Har-Peled, Indyk, and Motwani, 2012). It is natural then to search for ways to break the curse by relaxing the problem with approximate versions and by finding good ways to reduce the dimension of data. Our objective is to write a dissertation about a dimension reduction scheme for *clustering under ℓ_2^2 metric*, with a focus on an approximation scheme for a particular case of this problem, called *projective clustering*. The dimension reduction is achieved by combining randomized techniques, such as the Johnson and Lindenstrauss Lemma, and deterministic techniques, such as the singular value decomposition. The result is an $(1+\varepsilon)$-approximation for projective clustering that is polynomial in the number of data points and the dimension of the space. This dissertation will have as main references four papers: Sarlós, 2006, Feldman, Schmidt, and Sohler, 2020, Pratap and Sen, 2018 and Deshpande, Rademacher, Vempala, and Wang, 2006. The results presented in the dissertation will be either the original or modified versions that incorporate current improvements.

Keywords: Clustering. Projective Clustering. Dimension reduction. Approximation. Singular value decomposition. Johnson-Lindenstrauss lemma.

Contents

Chapter 1

Introduction

In many areas of science, research depends on the analysis of large amounts of high-dimensional data. But many algorithms suffer from the well known "curse of dimensionality" and are inefficient when the dimension of the space is too big. One example of this phenomenon is the currently known algorithms for the nearest neighbour search problem. See Har-Peled, Indyk, and Motwani (2012).

One way to circumvent this barrier is to make the problems easier by relaxing the necessity of an exact and deterministic answer. This can be enough to break the curse and enable practical applications that scale well with the dimension and the number of data points. See Matoušek (2013, Chap. 2, Section 6).

It makes sense then to study *dimension reduction* as a way to find approximations. Given a data set for a problem where dimensionality is a barrier for algorithmic efficiency, we want to find another data set called *sketch* that is contained in a subspace with lower dimension. The desired property is that solving the problem for the sketch is more efficient and would give a "good" approximation for the original data set with high probability.

In this dissertation we will study how to find approximations for clustering problems under the ℓ_2^2-metric via dimension reduction. We will present a randomized algorithm that with high probability finds sketches contained in subspaces of dimension independent of the number of points and of the original dimension of the space with time complexity linear in those parameters. This sketch is generic enough to be useful for any clustering problem under the ℓ_2^2-metric, which includes the k-means clustering problem and the projective clustering problem. We will also show an example where we "break the curse of dimensionality", that is, we use the developed sketch to improve the time complexity of an approximation algorithm for linear projective clustering from an exponential dependence in the dimension to a polynomial one. This last result is mostly of theoretical interest, since the degree of this polynomial is rather large.

Chapter 2

Preliminaries

In this chapter we will present our notation, definitions, and preliminary results and theory that will be used throughout this work.

2.1 Definitions

The letters n, d, r, i and j will always denote non-negative integers. Usually n will denote the cardinality of a non-empty finite set or multiset, and d will denote the dimension of a real vector space. We will adopt the convention that for any two sets A and B, if $A \subset B$, then we may have $A = B$.

Matrix notation: The set of real $d \times n$ matrices will be denoted by $\mathbb{R}^{d\times n}$. It will be useful to visualize the matrices as multisets of vectors of $\mathbb{R}^d$. For example, if $A \in \mathbb{R}^{d\times n}$ then the i^{th} column a_i of A can be seen as a vector of $\mathbb{R}^d$. We will abuse the notation and write $a \in A$ to say that a is a column of A seen as a vector, and $a_i \in A$ to say that a_i is the i^{th} column of A seen as a vector. Since matrices may have two or more equal columns, it can happen that for a_i and $a_j \in A$, we have $i \neq j$ but $a_i = a_j$.

When convenient and to not pollute the notation we will treat vectors as column matrices. For example, suppose that u and $v \in \mathbb{R}^d$. The usual real inner product denoted as $\langle u, v\rangle$ can be written as $u^T v$, where u^T is seen as an $1 \times d$ matrix and v as an $d \times 1$ matrix.

Definition 1. *Let $C \subset \mathbb{R}^d$ be a non-empty set and let $p \in \mathbb{R}^d$ be a vector. Then*

$$\operatorname{dist}(p, C) := \inf\{\|p - c\| : c \in C\},$$

where $\|\cdot\|$ is the usual Euclidean norm.

Definition 2. *Let $C \subset \mathbb{R}^d$ be a non-empty set and let $A \in \mathbb{R}^{d\times n}$ be a matrix. Then*

$$\operatorname{dist}^2(A, C) := \sum_{a\in A} (\operatorname{dist}(a, C))^2.$$

Definition 3. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix. The subspace of $\mathbb{R}^d$ spanned by the columns of A will be denoted as* span(A).

Definition 4. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and let a_{ij} be the row i and column j entry of A. The* Frobenius norm *of A is*

$$\|A\|_F := \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{n} a_{ij}^2}.$$

Note that if we see each column of A as a $\mathbb{R}^d$ vector we have

$$\|A\|_F^2 = \sum_{a\in A} \|a\|^2.$$

Note also that by seeing vectors as $d \times 1$ matrices, we have for any $p \in \mathbb{R}^d$

$$\|p\|_F = \|p\|.$$

Definition 5. *We say that a matrix $P \in \mathbb{R}^{d\times j}$ has* orthonormal columns *when the set $\{p_1, \ldots, p_j\}$ of columns of P all have norm one and P^TP is equal to the $j \times j$ identity matrix . An* orthogonal matrix *is a square matrix with orthonormal columns.*

Definition 6 (Orthogonal projection)**.** *Let $v \in \mathbb{R}^d$ be a vector and let L be a subspace of $\mathbb{R}^d$. Then $\pi_L(v) \in \mathbb{R}^d$ is the* orthogonal projection of v onto L.

Definition 7 (Orthogonal projection of a matrix)**.** *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and let L be a subspace of $\mathbb{R}^d$. The* orthogonal projection of A into L *is the matrix $\pi_L(A) \in \mathbb{R}^{d\times n}$ where the i^{th} column of $\pi_L(A)$ is the orthogonal projection of the i^{th} column of A into L.*

Now we will state some facts and results that will be useful later. We will omit the proofs which are straightforward.

Fact 8. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, let L be a subspace of $\mathbb{R}^d$ of dimension j and let $P \in \mathbb{R}^{d\times j}$ be a matrix with orthonormal columns that spans L. Then the orthogonal projection $\pi_L(A)$ equals PP^TA. Also note that*

$$\text{dist}^2(A, L) = \|A - PP^TA\|_F^2.$$

Definition 9. *Let A and $B \in \mathbb{R}^{d\times n}$ be matrices. We will abuse the notation and adopt that $\pi_B(A)$ means the same as $\pi_{\text{span}(B)}(A)$.*

Note that if B has orthonormal columns, then $\pi_B(A) = BB^TA$. When it is clear from the context we will call matrices like BB^T *projector matrices* to span(B).

Fact 10 (Pythagoras' Theorem)**.** *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, let L be a subspace of $\mathbb{R}^d$ of dimension j and let $P \in \mathbb{R}^{d\times j}$ with orthonormal columns that spans L. Then it follows from the Pythagoras' Theorem that*

$$\|A\|_F^2 = \|PP^TA\|_F^2 + \|A - PP^TA\|_F^2.$$

Fact 11 (Trace representation). *For any real matrix A, we have that*

$$\|A\|_F^2 = \mathrm{Tr}(A^T A)$$

and

$$\|A\|_F^2 = \|A^T\|_F^2.$$

Fact 12 (Cyclic property of the Trace function). *For any matrices $A \in \mathbb{R}^{d\times n}$ and $B \in \mathbb{R}^{n\times d}$, we have*

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA).$$

Fact 13 (Unitarily invariant norm). *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, let $P \in \mathbb{R}^{d\times d}$ and $Q \in \mathbb{R}^{n\times n}$ be orthogonal matrices and let $S \in \mathbb{R}^{s\times d}$ be a matrix with orthonormal columns and let $R \in \mathbb{R}^{n\times r}$ be such that R^T has orthonormal columns. Then*

$$\|A\|_F^2 = \|PA\|_F^2 = \|AQ\|_F^2 = \|SA\|_F^2 = \|AR\|_F^2.$$

Proof. Note that $\|A\|_F^2 = \mathrm{Tr}(A^T A)$ and for any unitary matrix P, we have that $P^T P = PP^T = I$. Also if S has orthonormal columns, then $S^T S = I$ and $RR^T = I$. We have that

$$\|PA\|_F^2 = \mathrm{Tr}(A^T P^T PA) = \mathrm{Tr}(A^T A) = \|A\|_F^2$$

and

$$\|SA\|_F^2 = \mathrm{Tr}(A^T S^T SA) = \mathrm{Tr}(A^T A) = \|A\|_F^2.$$

Similarly we have

$$\|AQ\|_F^2 = \|Q^T A^T\|_F^2 = \|A\|_F^2$$

and

$$\|AR\|_F^2 = \|R^T A^T\|_F^2 = \|A\|_F^2,$$

by the same argument. ∎

Definition 14 (Big-O notation). *For two functions $f, g : \mathbb{R} \to \mathbb{R}$, we say that $f(x) = O(g(x))$ if there exist constants K and x_0 such that $|f(x)| \le K|g(x)|$ for all $x \ge x_0$. We say that $f(x) = \Omega(g(x))$ if $g(x) = O(f(x))$. Finally, we say that $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $f(x) = \Omega(g(x))$.*

When we use this notation the variable x and the functions f and g will be implicit from the context. Note that the domain of the functions can be changed to $\mathbb{Z}$ or to $[a, \infty)$, for any real a, and the notation will still be well defined.

Model of Computation: To study the complexity of algorithms we will adopt the *real random access machine* model, proposed by Shamos (1978). It is similar to the usual RAM model, but we allow the storage and computation of real numbers. We assume that all arithmetic operations $(+, -, \times, /, \log, \sin, \dots)$ and comparisons $(<, \le, =, \ge, >)$ between two reals can be computed in constant time. Also we will assume that reading and writing reals can be done in constant time. For example, reading a matrix $A \in \mathbb{R}^{d\times n}$ and computing the product Av for some vector $v \in \mathbb{R}^n$ can be realized with $O(nd)$ operations and with "infinite precision", i.e., we always obtain the exact answer. Since each operation takes constant

time we have that this product takes $O(nd)$ time. Storing a real number will require $O(1)$ storage locations, therefore to store the matrix A we can use $O(nd)$ memory space.

The bounds we will obtain for the time complexity of algorithms will therefore reflect the number of operations in this idealized model. We will ignore questions as to how represent, read and operate with real numbers in finite time.

2.2 Best-fit subspace and singular value decomposition

Let $A \in \mathbb{R}^{d\times n}$ be a matrix. In this section we will study the problem of finding a subspace V of given dimension $j \leq d$ that minimizes $\text{dist}^2(A, V)$ and its relation to matrix decomposition.

Definition 15 (best-fit linear j-subspace problem)**.** *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and let $j \leq d$ be a positive integer. The* best-fit linear j-subspace *problem in $\mathbb{R}^d$ is to find a linear subspace V of $\mathbb{R}^d$ of dimension j such that*

$$\text{dist}^2(A, V) = \min \left\{ \text{dist}^2(A, L) \,:\, L \subset \mathbb{R}^d \text{ is a subspace of dimension } j \right\}.$$

We emphasize that the subspace is linear, since we could consider a similar more general problem where we want to find the best-fit *affine* subspace.

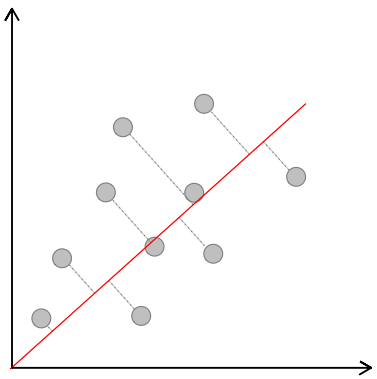


Figure 2.1: *Example of best-fit linear* 1*-subspace in* $\mathbb{R}^2$.

The *singular value decomposition* (SVD) of A is a matrix decomposition that comes naturally when we try to solve the best-fit linear j-subspace problem. Every matrix admits a singular value decomposition. Suppose that $\text{rank}(A) = r$. There exist matrices $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{n\times r}$ with orthonormal columns and a diagonal matrix $\Sigma \in \mathbb{R}^{r\times r}$ with positive entries such that $A = U\Sigma V^T$.

Let $u_1, \ldots, u_r$ and $v_1, \ldots, v_r$ be the columns of U and V, respectively. Let $\Sigma = \text{Diag}\{\sigma_1, \ldots, \sigma_r\}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. We call the vectors u_i *left singular vectors*, the v_i *right singular vectors* and the σ_i *singular values*. We will show that the sequence $\sigma_1, \ldots, \sigma_r$ of singular values is uniquely determined by A.

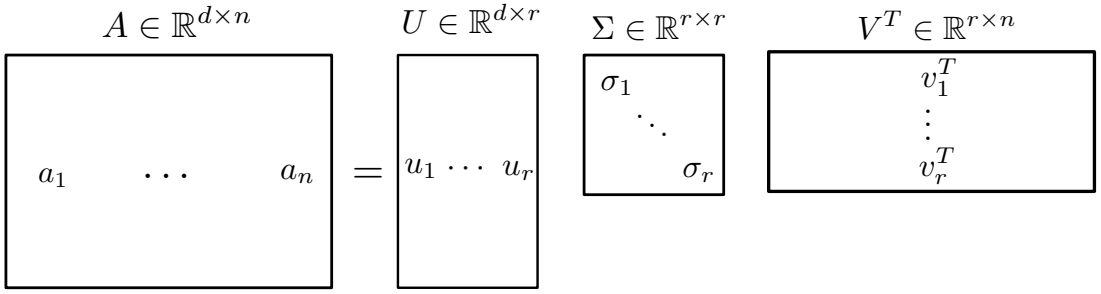


Figure 2.2: *A SVD decomposition of an $d \times n$ matrix of rank r.*

Notation 16. *Let A be a matrix and $1 \leq i \leq \text{rank}(A)$ an integer. The i^{th} singular value of A is denoted $\sigma_i(A)$.*

To better visualize the singular value decomposition, we will define the left singular vectors and singular values and its relation to the problem of finding a best-fit linear j-subspace. Let us start with $j = 1$. Finding the best-fit linear 1-subspace is equivalent to finding a norm-one vector u that minimizes

$$\|A - uu^T A\|_F^2, \tag{2.1}$$

because equation (2.1) equals $\text{dist}^2(A, V)$, where V is the subspace spanned by u. From Fact 10 (Pythagoras) it follows that minimizing the above expression is equivalent to maximizing

$$\|uu^T A\|_F^2 = \|u^T A\|_F^2.$$

Definition 17. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix. A* first left singular vector *of A is any vector $u_1 \in \mathbb{R}^d$ such that*

$$u_1 := \text{argmax} \left\{ \|u^T A\|_F : \|u\| = 1 \right\},$$

and the first singular value *is*

$$\sigma_1 := \|u_1^T A\|_F.$$

We say that u_1 is a left singular vector associated with the singular value σ_1.

Note that u_1 is not uniquely defined since $-u_1$ would also be a valid choice. To define the subsequent singular vectors and values, we can use a greedy approach. A second left singular vector u_2 maximizes the same expression, but with the restriction that it must be orthogonal to u_1.

$$u_2 := \text{argmax} \left\{ \|u^T A\|_F : \|u\| = 1,\ u \perp u_1 \right\}.$$

The second singular value is

$$\sigma_2 := \|u_2^T A\|_F.$$

We can define all singular values and vectors this way, where an i^{th} left singular vector u_i is a norm one vector that is orthogonal to all lower singular vectors and maximizes $\|u^T A\|_F$.

The associated singular value is $\|u_i^T A\|_F$. We stop when $i = \text{rank}(A)$, since for $i > \text{rank}(A)$, the set $\{u_1, \dots, u_r\}$ spans the same subspace as span(A), and thus $\|u^T A\|_F$ would be zero for all vectors orthogonal to $u_1, \dots, u_r$. An interesting fact is that for all $j = 1, \dots, r$, the subspace spanned by $u_1, \dots, u_j$ is a best-fit j-subspace. The proof of this fact can be found on Blum, Hopcroft, and Kannan (2020, Chap. 3, Sec. 3).

Having defined the left singular vector and singular values, we can define the right singular vectors in the following way: if u_i is a i^{th} left singular vector and σ_i is the i^{th} singular value, the i^{th} right singular vector v_i is

$$v_i := \frac{1}{\sigma_i} A^T u_i.$$

Note that by definition $\|v_i\| = 1$. The proof that $v_i \perp v_j$ for all $i \neq j$ can be found in Blum, Hopcroft, and Kannan (2020, Chap. 3, Sec. 6).

Having the definitions of U, Σ and V, we just need to show that $A = U\Sigma V^T$. One way of doing this is by rewriting $U\Sigma V^T$ as

$$U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T, \tag{2.2}$$

and seeing that for every vector $x \in \mathbb{R}^d$, we have $x^T A = x^T U\Sigma V^T$. The details can be found in Blum, Hopcroft, and Kannan (2020, Chap. 3, Sec 3).

The singular values of a matrix can be useful to determine how much a matrix can change the norm of a vector. For every $x \in \mathbb{R}^d$ we have

$$\left\|A^T x\right\| \leq \sigma_1(A) \|x\|.$$

This follows by the definition of $\sigma_1(A)$.

If the matrix A is full rank, that is, if $r = d$, then we have that the last singular value $\sigma_d(A)$ is a lower bound for the norm of $A^T x$.

$$\left\|A^T x\right\| \geq \sigma_d(A) \|x\|.$$

This follows by noting that $\left\|A^T x\right\|^2 = \left\|V^T A x\right\|^2 = \left\|\Sigma U^T x\right\|^2$ and by using that $\sigma_d(A) \leq \sigma_i(A)$ for every $i = 1, \dots, d$. Hence we can write

$$\left\|\Sigma U^T x\right\|^2 = \sum_{i=1}^{d} \sigma_i(A)^2 \langle u_i, x\rangle^2 \geq \sigma_d(A)^2 \sum_{i=1}^{d} \langle u_i, x\rangle^2 = \sigma_d(A)^2 \|x\|^2.$$

2.2.1 Best approximation in the Frobenius norm

Let us take another look at expression (2.2). Suppose that $k \leq r$ is an integer. We define

$$A_k := \sum_{i=1}^{k} \sigma_i u_i v_i^T, \tag{2.3}$$

the matrix obtained by the truncation of the sum (2.2) in the k^{th} parcel. Note that $\text{rank}(A_k) = k$.

Fact 18. *The matrix A_k is the orthogonal projection of A to the subspace spanned by the first k left singular vectors $u_1, \ldots, u_k$.*

This can be seen by making $U^{(k)} \in \mathbb{R}^{d \times k}$ the matrix with columns $u_1, \ldots, u_k$ and verifying that $U^{(k)}(U^{(k)})^T A = A_k$.

Fact 19. *The matrix A_k is a* best rank k approximation *of A in the Frobenius norm, that is, for any matrix $B \in \mathbb{R}^{d \times n}$ with rank at most k we have*

$$\|A - A_k\|_F \leq \|A - B\|_F. \tag{2.4}$$

Fact 19 follows from Fact 18 and from $\text{span}(U^{(k)})$ being a best-fit subspace of dimension k. More details and proofs can be found in Blum, Hopcroft, and Kannan (2020, Chap. 3, Sec 5).

The matrix A_k is a best rank k approximation not only in the Frobenius norm, but also on any unitarily invariant norm. See Horn and Johnson (2013).

2.2.2 Finding the SVD in polynomial time

The singular value decomposition is useful in practice because it can be found in polynomial time. Some algorithms find just the singular values, while others find the pairs of singular values and singular vectors (either right or left). Here we present a naive algorithm that finds pairs of singular values and left singular vectors by reducing this problem to finding pairs of eigenvalues and eigenvectors.

Fact 20. *Let $A \in \mathbb{R}^{d \times n}$ be a matrix of rank r. The first r eigenvalues $\lambda_1, \ldots, \lambda_r$ of the matrix AA^T are the square of the singular values $\sigma_1^2, \ldots, \sigma_r^2$ of A, and the remaining eigenvalues are zero. The eigenvectors associated with the positive eigenvalues are the left singular vectors associated with the corresponding singular value.*

Proof. Let $U\Sigma V^T$ be a singular value decomposition of A. Then

$$AA^T = U\Sigma V^T V \Sigma U^T = U\Sigma^2 U^T.$$

Since the columns $u_1, \ldots, u_r$ of U are orthonormal, we have that for all $i = 1, \ldots, r$

$$AA^T u_i = U\Sigma^2 U^T u_i = \sigma_i^2 u_i,$$

hence u_i is an eigenvector of AA^T associated with σ_i^2. It follows that the space spanned by the eigenvectors associated with $\lambda_i = \sigma_i^2$ is the same as the left singular vectors associated with the singular value σ_i.

If we complete the set $u_1, \dots, u_r$ to an orthonormal basis of $\mathbb{R}^d$, we can see that for all $i = r+1, \dots, d$

$$AA^T u_i = U\Sigma^2 U^T u_i = 0,$$

which means that the remaining eigenvalues are zero and the null-space of AA^T is equal to the null-space of U. ∎

The fact above shows us that finding singular values and vectors can be reduced to the problem of finding eigenvalues and eigenvectors, therefore a simple algorithm to find the SVD of a matrix A is, assuming $d \leq n$, to compute AA^T and then find its eigen-pairs, which can be done in time $O(d^3)$ (Pan, Chen, Zheng, *et al.*, 1999). Note that the whole process takes time $O(nd^2)$. For simplicity we will assume that the SVD can be computed exactly and in time $O(nd^2)$.

The method we presented here to compute the SVD is not what is used in practice, since it may have rounding problems and other issues. A discussion about this topic can be found in Trefethen and Bau (1997, Lecture 31).

Fact 20 let us derive some useful relations between the SVD, the Frobenius norm, and the trace function.

Fact 21. *For any matrix $A \in \mathbb{R}^{d\times n}$, we have*

$$\|A\|_F^2 = \operatorname{Tr}(AA^T) = \sum_{i=1}^{d} \lambda(AA^T) = \sum_{i=1}^{d} \sigma^2(A).$$

2.3 The Johnson-Lindenstrauss Lemma

The Johnson-Lindenstrauss Lemma (Johnson and Lindenstrauss, 1984) is the following surprising result.

Theorem 22 (Johnson-Lindenstrauss Lemma)**.** *There exists a constant κ such that for any set A of n points in $\mathbb{R}^d$, any $\varepsilon \in (0,1)$ fixed and all integers $r \geq \kappa\varepsilon^{-2}\log n$ there exists a linear function $f : \mathbb{R}^d \to \mathbb{R}^r$ such that for every pair $x, y \in A$ we have*

$$(1-\varepsilon)\|x-y\|^2 \leq \|f(x)-f(y)\|^2 \leq (1+\varepsilon)\|x-y\|^2. \tag{2.5}$$

Intuitively, the theorem says that a space of dimension $O(\varepsilon^{-2}\log n)$ is enough to represent n points such that their pairwise Euclidean distances are "almost" preserved[1]. The functions from this theorem are also used for efficient data storage and for speeding up algorithms with exponential run-time complexity on the dimension (Achlioptas, 2003).

[1] Note that the inequalities in (2.5) imply $(1-\varepsilon)\|x-y\| \leq \|f(x)-f(y)\| \leq (1+\varepsilon)\|x-y\|$.

Most proofs of Theorem 22 rely on the probabilistic method. We draw a random function from some distribution in the space of linear functions, and bound the probability for this function to satisfy the theorem. If this probability is positive then there exists at least one function that satisfies Theorem 22. A relevant distribution is given by the following lemma, usually called *random projection lemma*.

Lemma 23 (Random projection lemma)**.** *Suppose $T \in \mathbb{R}^{r\times d}$ is a random matrix where each entry t_{ij} of T is an independent random variable that assumes values uniformly in $\{+1, -1\}$. Let $S \in \mathbb{R}^{r\times d}$ be defined as*

$$S = \frac{1}{\sqrt{r}}T.$$

Then for all $v \in \mathbb{R}^d$ and $\varepsilon \in (0, 1)$ we have

$$\mathbb{P}\left[\|Sv\|^2 > (1+\varepsilon)\|v\|^2\right] < e^{-r\varepsilon^2/12},$$
$$\mathbb{P}\left[\|Sv\|^2 < (1-\varepsilon)\|v\|^2\right] < e^{-r\varepsilon^2/12}.$$

For now on we will call random $r \times d$ matrices where each entry is an independent uniform $\{+1, -1\}$ random variable *Johnson-Lindenstrauss matrix*, and we will call random $r \times d$ matrices where each entry is an independent uniform $\{+1/\sqrt{r}, -1/\sqrt{r}\}$ random variable *normalized Johnson-Lindenstrauss matrix*. For example, in Lemma 23 the matrix T is a Johnson-Lindenstrauss matrix and the matrix S is a normalized Johnson-Lindenstrauss matrix.

The proof of Lemma 23 can be found in Achlioptas (2003, Sec. 5). But the idea behind it is the following: if we look at the square of each coordinate of the vector Tv, we will note that its mean is equal to $\|v\|^2$. Thus the random variable $\|Sv\|^2$ is the sum of independent random variables with mean $\|v\|^2$. This means that we can apply a Chernoff-like bound to obtain the result.

Proof of Theorem 22. Let r be a positive integer and let $S \in \mathbb{R}^{r\times d}$ be a normalized Johnson-Lindenstrauss matrix. Define $f : \mathbb{R}^d \to \mathbb{R}^r$ as $f(x) = Sx$. To prove that this functions works, we need to show that it avoids the following bad events:

$$\|f(x-y)\|^2 = \|f(x) - f(y)\|^2 > (1+\varepsilon)\|x-y\|^2$$

and

$$\|f(x-y)\|^2 = \|f(x) - f(y)\|^2 < (1-\varepsilon)\|x-y\|^2.$$

By Lemma 23 we have that for every pair of points $x, y \in A$ the probability that each of the above bad events happen is at most $e^{-r\varepsilon^2/12}$.

By the union bound the probability that, for at least one pair of points, at least one of the bad events happens is

$$\mathbb{P}\left[\exists x, y : \left|\|f(x) - f(y)\|^2 - \|x-y\|^2\right| > \varepsilon\|x-y\|^2\right] < 2n^2 e^{-r\varepsilon^2/12}. \tag{2.6}$$

Fix $\kappa = 37$. If $r \geq 37\varepsilon^{-2} \log n$ then

$$e^{-r\varepsilon^2/12} \leq \frac{1}{n^{37/12}},$$

thus we can bound expression (2.6) with

$$2n^2 e^{-r\varepsilon^2/12} \leq 2n^2 \frac{1}{n^{37/12}} < 1.$$

We conclude that the probability in (2.6) is strictly smaller than one, therefore there exists at least one function f that avoids the bad events for all pairs of points. ∎

Remark: *The proof we presented here is not optimal in the constant κ. Achlioptas (2003) uses a slightly modified bound on the probability of the random projection lemma (but with the same random matrix) and obtains the following bound for r:*

$$r \geq \frac{4 + 2\beta}{\varepsilon^2/2 - \varepsilon^3/3} \log n,$$

where β can assume any positive value. Also the author states that the probability of a matrix randomly chosen by Lemma 23 to satisfy Theorem 22 is $1 - n^{-\beta}$. Therefore for all $\delta \in (0, 1)$, if $\beta = 2\log(1/\delta) \geq \log(1/\delta)/\log n$, then the probability of success will be at least $1 - \delta$.

The distribution defined in Lemma 23 is not the only distribution with that property. We can prove a result similar to Lemma 23 if we instead use random matrices where each entry is an independent normalized Gaussian, or a sparser random matrix where each entry is a random variable that is zero with probability 2/3, is $\sqrt{3/r}$ with probability 1/6 and $-\sqrt{3/r}$ with probability 1/6 (see Achlioptas, 2003, Theorem 1.1).

Definition 24 (Johnson-Lindenstrauss Transform)**.** *Let $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$ be fixed. Suppose that $r < d$ are positive integers and that $\boldsymbol{D}_{d,r}$ is a probability distribution over the space of linear functions with domain $\mathbb{R}^d$ and codomain $\mathbb{R}^r$. We say that $\boldsymbol{D}_{d,r}$ is a* Johnson-Lindenstrauss Transform *with parameters ε, δ, and n or $JLT(\varepsilon, \delta, n)$ for short if for any set $A \subset \mathbb{R}^d$ of n points with probability at least $1 - \delta$ we have that a function f drawn from $\boldsymbol{D}_{d,r}$ satisfies for every pair $x, y \in A$*

$$(1 - \varepsilon)\|x - y\|^2 \leq \|f(x) - f(y)\|^2 \leq (1 + \varepsilon)\|x - y\|^2.$$

Note that if $\boldsymbol{D}_{d,r}$ is a $JLT(\varepsilon, \delta, n)$, then $\boldsymbol{D}_{d,r}$ is also a $JLT(\varepsilon, \delta, n')$ for all positive $n' \leq n$. For convenience we will abuse the notation and define random matrices as if they were a probability distribution over linear functions and also as if they were linear functions.

Definition 25. *Fix $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$. Suppose that $r < d$ is a positive integer. We say that a random matrix S is a* Johnson-Lindenstrauss Transform *with parameters ε, δ, and n or $JLT(\varepsilon, \delta, n)$ for short if for any set $A \subset \mathbb{R}^d$ of n points with probability at least $1 - \delta$ for*

every pair $x, y \in A$ *we have*

$$(1-\varepsilon)\|x-y\|^2 \leq \|Sx-Sy\|^2 \leq (1+\varepsilon)\|x-y\|^2.$$

Note that a $r \times d$ normalized Johnson-Lindenstrauss matrix is a $JLT(\varepsilon, \delta, n)$ for some $r = O(\varepsilon^{-2}\log(1/\delta)\log n)$.

2.3.1 Consequences of the Johnson and Lindenstrauss Lemma

Theorem 22 has interesting and useful consequences. The Johnson-Lindenstrauss Transform not only "almost" preserves distances between points in a finite set, but also "almost" preserves structures such as angles between points in a finite set.

Corollary 26. *Let* $\varepsilon \in (0,1)$ *and* $\delta \in (0,1)$ *be fixed. If* f *is drawn from a* $JLT(\varepsilon, \delta, n+1)$ *then for any* $A \subset \mathbb{R}^d$ *set of* n *points, with probability at least* $1-\delta$ *for all* $x, y \in A$ *we have*

$$\langle x, y\rangle - \varepsilon\|x\|\|y\| \leq \langle f(x), f(y)\rangle \leq \langle x, y\rangle + \varepsilon\|x\|\|y\|. \tag{2.7}$$

Proof. Let $A' = A \cup \{0\}$. Since f is drawn from a $JLT(\varepsilon, \delta, n+1)$ and $0 \in A'$, we have that for all $x, y \in A'$

$$(1-\varepsilon)\|x-y\|^2 \leq \|f(x)-f(y)\|^2 \leq (1+\varepsilon)\|x-y\|^2, \tag{2.8}$$

and

$$(1-\varepsilon)\|x+y\|^2 \leq \|f(x)+f(y)\|^2 \leq (1+\varepsilon)\|x+y\|^2. \tag{2.9}$$

Inequality (2.8) follows the definition of f. The proof that f also satisfies inequality (2.9) can be found in Fact 60 in Appendix A.

Inequalities (2.7) are true for x or $y = 0$. Suppose that $x, y \in A'$ are not the null vector. Let us prove the case when $\|x\| = \|y\| = 1$. By the parallelogram rule in real inner product spaces, we have

$$\begin{aligned} 4\langle f(x), f(y)\rangle &= \|f(x)+f(y)\|^2 - \|f(x)-f(y)\|^2 \\ &\leq (1+\varepsilon)\|x+y\|^2 - (1-\varepsilon)\|x-y\|^2 \qquad \text{from definition of } JLT(\varepsilon, \delta, n+1). \end{aligned}$$

Now using that the Euclidean norm is induced by the inner product, we have

$$4\langle f(x), f(y)\rangle \leq (1+\varepsilon)(\|x\|^2 + \|y\|^2 + 2\langle x, y\rangle) - (1-\varepsilon)(\|x\|^2 + \|y\|^2 - 2\langle x, y\rangle).$$

After rearranging and cancelling some terms we obtain

$$4\langle f(x), f(y)\rangle \leq 4\langle x, y\rangle + 4\varepsilon.$$

For the case in which $\|x\|$ or $\|y\|$ is not equal to one, we can use the linearity of f and

apply the same argument to

$$\|x\|\|y\| \left\langle f\left(\frac{x}{\|x\|}\right), f\left(\frac{y}{\|y\|}\right)\right\rangle$$

to obtain (2.7). The lower bound argument is similar. ∎

A useful result that is not directly derived from the Johnson and Lindenstrauss Lemma but from the random projection lemma is the following result from Alon, Gibbons, Matias, and Szegedy (2002).

Lemma 27. *Let $\varepsilon \in (0,1)$ be fixed and let $S \in \mathbb{R}^{\lceil \varepsilon^{-2} \rceil \times d}$ be a normalized Johnson-Lindenstrauss matrix. Then for every $x, y \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\langle Sx, Sy\rangle\right] = \langle x, y\rangle$$

and

$$\mathrm{Var}\left[\langle Sx, Sy\rangle\right] \leq 2\varepsilon^2 \|x\|^2 \|y\|^2 .$$

A proof for Lemma 27 can be found in the appendix A. We will use this lemma and the corollary above it to prove a result about approximations for matrix multiplication.

Lemma 28 (Sarlós, 2006)**.** *Let $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{d\times m}$ be matrices, and let $\varepsilon \in (0,1)$ be fixed. The following statements are true:*

1. *If a matrix S is a $JLT(\varepsilon, \delta, n+m+1)$ then with probability at least $1-\delta$ we have*

$$\left\|AB - AS^TSB\right\|_F \leq \varepsilon \|A\|_F \|B\|_F .$$

2. *If a matrix $S \in \mathbb{R}^{\lceil \varepsilon^{-2} \rceil \times d}$ is a normalized Johnson-Lindenstrauss matrix then*

$$\mathbb{E}\left[AS^TSB\right] = AB$$

and

$$\mathbb{E}\left[\left\|AB - AS^TSB\right\|_F^2\right] \leq 2\varepsilon^2 \|A\|_F^2 \|B\|_F^2 ,$$

where the mean $\mathbb{E}[X]$ of a random matrix X is the matrix where the entry $\mathbb{E}[X]_{ij}$ is $\mathbb{E}\left[X_{ij}\right]$.

Proof. We will begin with item 1. Let a_i be the i^{th} row of A and b_i be the i^{th} column of B. The entry in the i^{th} column and j^{th} row of the matrix $AB - AS^TSB$ is $\left\langle a_i, b_j\right\rangle - \left\langle Sa_i, Sb_j\right\rangle$. If we consider the set $C = \{a_1, \ldots, a_n, b_1, \ldots, b_m\}$, by Corollary 26 we have that with probability at least $1-\delta$

$$\left|\left\langle a_i, b_j\right\rangle - \left\langle Sa_i, Sb_j\right\rangle\right| \leq \varepsilon \|a_i\| \|b_j\| , \tag{2.10}$$

and by applying the definition of Frobenius norm we have

$$\begin{aligned}\left\|AB - AS^TSB\right\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left\langle a_i, b_j\right\rangle - \left\langle Sa_i, Sb_j\right\rangle\right)^2 \\ &\leq \sum_{i=1}^{n}\sum_{j=1}^{m}\varepsilon^2\|a_i\|^2\|b_j\|^2 \\ &= \varepsilon^2\|A\|_F^2\|B\|_F^2,\end{aligned}$$

finishing the proof of item 1.

For item 2, note that from Lemma 27 we have that

$$\mathbb{E}\left[\left\langle Sa_i, Sb_j\right\rangle\right] = \left\langle a_i, b_j\right\rangle$$

and thus $\mathbb{E}\left[AS^TSB\right] = AB$. Further,

$$\begin{aligned}2\varepsilon^2\|a_i\|^2\|b_j\|^2 &\geq \operatorname{Var}\left[\left\langle Sa_i, Sb_j\right\rangle\right] && \text{by Lemma 27} \\ &= \mathbb{E}\left[\left(\left\langle a_i, b_j\right\rangle - \left\langle Sa_i, Sb_j\right\rangle\right)^2\right] && \text{by variance definition.}\end{aligned}$$

By applying the Frobenius norm definition in the same way we did for item 1 we conclude that $\mathbb{E}\left[\left\|AB - AS^TSB\right\|_F^2\right] \leq 2\varepsilon^2\|A\|_F^2\|B\|_F^2$. ∎

Combining the first statement of Lemma 28 with the fact that an $r \times d$ normalized Johnson-Lindenstrauss matrix is a $JLT(\varepsilon, \delta, m+n+1)$ for some $r = O(\varepsilon^{-2}\log(1/\delta)\log(m+n+1))$ give an algorithm to accelerate matrix multiplication. Usually the time complexity to compute AB naively is $O(mnd)$, but by computing $\widetilde{A} := AS^T$ and $\widetilde{B} := SB$ first, and then computing $\widetilde{A}\widetilde{B}$ takes time $O((mn + dm + dn)\varepsilon^{-2}\log(1/\delta)\log(m+n+1))$ and give us a matrix that with probability at least $1-\delta$ is close to AB in the Frobenius norm.

Chapter 3

The (ℓ_2^2, C)-clustering problem

Let C be a non-empty family of non-empty sets of $\mathbb{R}^d$ and let $A \in \mathbb{R}^{d \times n}$ be a matrix. The (ℓ_2^2, C)-*clustering* problem in $\mathbb{R}^d$ is to find a set $C \in \mathcal{C}$ that minimizes $\text{dist}^2(A, C)$. We say that A an *instance* of the problem. The sets $C \in \mathcal{C}$ are called *solutions*. We say that a solution C^* is an *optimal solution* if $\text{dist}^2(A, C^*) \leq \text{dist}^2(A, C)$ for every $C \in \mathcal{C}$. Note that it is possible to have multiple distincts optimal solutions. When the instance A is implicit or is known from the context, we will call the value $\text{dist}^2(A, C)$ *cost* of C. For example, if C^* is an optimal solution then the cost of C^* is less than or equal to the cost of every other possible solution.

This problem is a general formulation for clustering problems under ℓ_2^2 metric. We will state some examples of known clustering problems that fit this formulation.

k-means clustering

The *k-means clustering* problem is to find a set C of k distinct points $\{c_1, \ldots, c_k\}$ called *centers* such that $\text{dist}^2(A, \{c_1, \ldots, c_k\})$ is minimized.

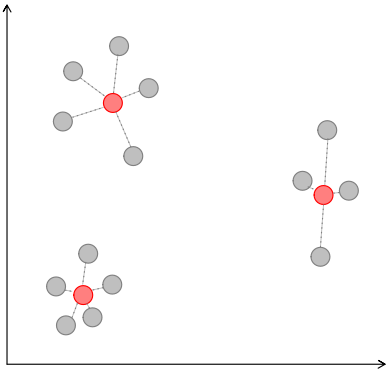


Figure 3.1: *An example of* 3*-means clustering in* $\mathbb{R}^2$*. The grey dots form the instance and the red dots are the centers.*

Note that k-means clustering problem is an $(\ell_2^2, \binom{\mathbb{R}^d}{k})$-clustering, where

$$\binom{\mathbb{R}^d}{k} := \{W \subset \mathbb{R}^d \,:\, |W| = k\}.$$

It is known that this problem is NP-hard for $k = 2$ (see Aloise, Deshpande, Hansen, and Popat, 2009). If the dimension d of the space is constant, then k-means can be solved in $O(n^{dk+1})$ time (see Inaba, Katoh, and Imai, 1994).

Best-fit j-subspace

Remembering definition 15, the *best-fit linear j-subspace* problem in $\mathbb{R}^d$ is to find a linear subspace V of dimension j of $\mathbb{R}^d$ that minimizes $\text{dist}^2(A, V)$. There is also a variant of this problem that consist of searching for *affine* subspaces of dimension j instead.

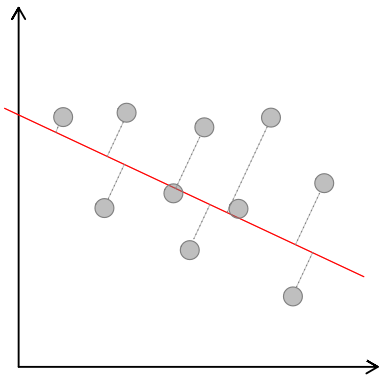


Figure 3.2: *Example of best-fit affine* 1*-subspace. The red line is one solution.*

We can see that best-fit linear j-subspace is a (ℓ_2^2, C)-clustering by taking C to be the set $\mathcal{V}_j$ of all subspaces of dimension j. Best-fit affine j-subspace also fits the general description by taking C to be the set $\mathcal{A}_j$ of all affine subspaces of dimension j.

Projective clustering

Let k and j be positive integers with $j < d$. The *linear j-subspace k-clustering* problem is to find a set of k distinct linear j-subspaces $\{V_1, \ldots, V_k\}$ of $\mathbb{R}^d$ that minimizes $\text{dist}^2(A, \bigcup_{i=1}^k V_i)$. This problem and a variant with affine subspaces instead of linear are both known as *projective clustering*.

We can see that both affine and linear j-subspace k-clustering problems are a particular case of the (ℓ_2^2, C)-clustering problem where

$$C := \left\{ \bigcup_{i=1}^{k} V_i \,:\, \{V_1, \ldots, V_k\} \in \binom{\mathcal{V}_j}{k} \right\}$$

for the linear case. For the affine case we just exchange $\binom{\mathcal{V}_j}{k}$ for $\binom{\mathcal{A}_j}{k}$. This problem is a generalization of the two previous problems. Note that affine 0-subspace k-clustering problem is equivalent to k-means clustering problem and that linear j-subspace 1-clustering problem is equivalent to the best-fit linear j-subspace problem.

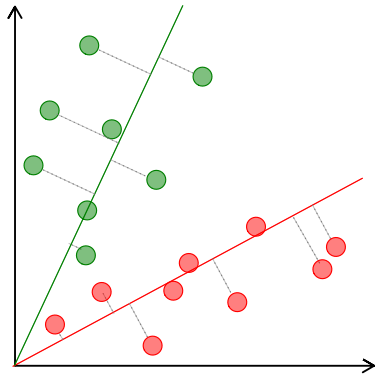


Figure 3.3: *Example of linear* 1*-subspace* 2*-clustering in* $\mathbb{R}^2$*. The dots form the instance. The green dots are closer to the green subspace, and the red dots are closer to the red subspace.*

A result from Megiddo and Tamir (1981) states that the problem of deciding if k lines are enough to cover a set of points in $\mathbb{R}^2$ is NP-complete. This implies that affine j-subspace k-clustering problem in $\mathbb{R}^2$ for $j = 1$ and $d = 2$ is NP-hard to approximate for any multiplicative approximation factor, since any α-approximation would decide the problem of covering points with lines.

3.1 Low dimensional representation

The three problems stated above have an interesting property: Let $A \in \mathbb{R}^{d\times n}$ be an instance, but suppose that the rank of A is $d' < d$. We can define another (ℓ_2^2, C')-clustering problem in $\mathbb{R}^{d'}$ and another instance $A' \in \mathbb{R}^{d'\times n}$ such that any solution for A' can be mapped to a solution for A, and an optimal solution for A' has the same cost as an optimal solution of A. We will show this for the best-fit linear j-subspace problem, but the proof strategy is the same for the other two problems.

Theorem 29 (Low dimensional representation)**.** *Let* $A \in \mathbb{R}^{d\times n}$ *be an instance of best-fit linear* j*-subspace problem in* $\mathbb{R}^d$*, and suppose that* $j < \operatorname{rank}(A) = d' < d$*. Then there exists an instance* $A' \in \mathbb{R}^{d'\times n}$ *of best-fit linear* j*-subspace problem in* $\mathbb{R}^{d'}$ *that satisfies the following:*

1. *There exists an isometric embedding* $f : \mathbb{R}^{d'} \to \mathbb{R}^d$ *such that if* C' *is a solution for* A' *then* $C := f(C')$ *is a solution for* A *and*

$$\operatorname{dist}^2(A', C') = \operatorname{dist}^2(A, C).$$

2. *For every solution* C *for* A *there exists a solution* C' *for* A' *such that*

$$\operatorname{dist}^2(A', C') \leq \operatorname{dist}^2(A, C).$$

Proof. Since $\operatorname{rank}(A) = d'$, there exists a matrix $U \in \mathbb{R}^{d \times d'}$ with orthonormal columns such that $\operatorname{span}(U) = \operatorname{span}(A)$. We claim that

$$A' := U^T A$$

is our desired instance of best-fit j-subspace problem in $\mathbb{R}^{d'}$.

Let us start with the proof of item 1. Since $\operatorname{span}(U) = \operatorname{span}(A)$, and since U has orthonormal columns, by Fact 8 we have that $UU^T A$ is an orthogonal projection to $\operatorname{span}(U)$, therefore $UU^T A = \pi_{\operatorname{span}(U)}(A) = \pi_{\operatorname{span}(A)}(A) = A$, which implies that $UA' = A$. From this fact it would be natural to define the isometric embedding f as $f(x) = Ux$. The function f is an isometric embedding because it maps the canonical base of $\mathbb{R}^{d'}$ into the orthonormal columns of U.

Now we will show that this definition of A' and f works. Let C' be a solution for A'. Remember that

$$\operatorname{dist}^2(A', C') = \|A' - \pi_{C'}(A')\|_F^2 = \sum_{a' \in A'} \|a' - \pi_{C'}(a')\|^2.$$

Let $B' \in \mathbb{R}^{d' \times j}$ be a matrix with orthonormal columns that span C'. For every vector $x' \in \mathbb{R}^{d'}$ we have

$$\pi_{C'}(x') = B'B'^T x'$$

and we also have that

$$\pi_{C'}(A') = B'B'^T A'.$$

To find $C := f(C')$, it suffices to find a basis $B \in \mathbb{R}^{d \times j}$ for C. For each column b'_i of B' we define the i^{th} column of B as $b_i := f(b'_i)$, so we have

$$B := UB'.$$

The columns of B form a set of orthonormal vectors: we just need to see that $(UB')^T UB'$ is equal to the identity matrix.

Now we use the fact that the Frobenius norm is invariant under unitary transformations to show that

$$\begin{aligned} \operatorname{dist}^2(A', C') &= \|A' - \pi_{C'}(A')\|_F^2 \\ &= \|A' - B'B'^T A'\|_F^2 && \text{by the definition of projection,} \\ &= \|UA' - UB'B'^T A'\|_F^2 && \text{by the unitary invariant.} \end{aligned}$$

Now using that $A = UA'$ and $B = UB'$ we obtain

$$\|UA' - UB'B'^T A'\|_F^2 = \|A - BB'^T A'\|_F^2.$$

Noting that U^TU equals the $d' \times d'$ identity matrix, we have

$$\|A - BB'^TA'\|_F^2 = \|A - BB'^TU^TUA'\|_F^2.$$

Using that $B^T = (UB')^T = B'^TU^T$ and that $A = UA'$, we have

$$\|A - BB'^TU^TUA\|_F^2 = \|A - BB^TA\|_F^2.$$

Finally, we conclude that

$$\text{dist}^2(A', C') = \|A - BB^TA\|_F^2 = \|A - \pi_C(A)\|_F^2 = \text{dist}^2(A, C).$$

Before proving item 2 we will need the following fact:

Fact 30. *Let $A \in \mathbb{R}^{d\times n}$ and $B \in \mathbb{R}^{d\times j}$ be matrices. Suppose that B has orthonormal columns. Then*

$$\|A\|_F^2 \geq \|B^TA\|_F^2.$$

Proof. Let $B' \in \mathbb{R}^{d\times d}$ be an orthogonal matrix that agrees with B in the first j columns. Since B' is orthogonal, B'^T is orthogonal, hence we have

$$\|A\|_F^2 = \|B'^TA\|_F^2 \geq \|B^TA\|_F^2,$$

since in the Frobenius norm we add the square of the entries, and in $\|B^TA\|_F^2$ we are adding fewer entries. ∎

Now we are ready to prove item 2. Let C be a solution for A, let $B \in \mathbb{R}^{d\times j}$ be a matrix with orthonormal columns that spans C and define $B' := U^TB$. We claim that $L := \text{span}(B')$ is a set in $\mathbb{R}^{d'}$ such that $\text{dist}^2(A', L) \leq \text{dist}^2(A, C)$.

From the fact 30 we have

$$\text{dist}^2(A, C) = \|A - BB^TA\|_F^2 \geq \|U^TA - U^TBB^TA\|_F^2.$$

Now, using that $A = UA'$, we have

$$\begin{aligned}\|U^TA - U^TBB^TA\|_F^2 &= \|A' - U^TB(U^TB)^TA'\|_F^2 \\ &= \|A' - B'B'^TA'\|_F^2 \qquad \text{by the definition of } B'.\end{aligned}$$

Note that for all $x \in \mathbb{R}^{d'}$ the vector $B'B'^Tx$ lies in L, which implies

$$\|x - B'B'^Tx\| \geq \|x - \pi_L(x)\|,$$

since the orthogonal projection is the closest vector of L to x. So we have

$$\|A' - B'B'^TA'\|_F^2 \geq \|A' - \pi_L(A')\|_F^2 = \text{dist}^2(A', L).$$

We conclude that $\text{dist}^2(A', L) \leq \text{dist}^2(A, C)$. The final step is to note that L is a subspace with dimension at most j. If the dimension of L is j, we are done. If not, we can find any base of L and complete it such that it spans a subspace C' of dimension j. Since $L \subset C'$, we have

$$\text{dist}^2(A', L) \geq \text{dist}^2(A', C').$$

We have found the solution C' for A' with the desired property. ∎

Corollary 31. *Under the same hypothesis of Theorem 29, if C'^* is an optimal solution for A' then $f(C'^*)$ is an optimal solution for A, where f is the isometric embedding given by (1) in Theorem 29.*

Proof. Using Theorem 29 (1), we have that

$$\text{dist}^2(A, f(C'^*)) = \text{dist}^2(A', C'^*).$$

Suppose that C^* is an optimal solution for A. Then we have that

$$\text{dist}^2(A, C^*) \leq \text{dist}^2(A, f(C'^*)).$$

Using Theorem 29 (2), there exists a solution C' for A' such that

$$\text{dist}^2(A', C') \leq \text{dist}^2(A, C^*).$$

Using that C'^* is optimal, we have

$$\text{dist}^2(A', C'^*) \leq \text{dist}^2(A', C').$$

Joining the above inequalities we arrive at

$$\text{dist}^2(A, f(C'^*)) = \text{dist}^2(A', C'^*) \leq \text{dist}^2(A, C^*) \leq \text{dist}^2(A, f(C'^*)),$$

and we have that $\text{dist}^2(A, f(C'^*)) = \text{dist}^2(A, C^*)$ which implies that $f(C'^*)$ is an optimal for A too. ∎

An analogous of Theorem 29 can be proved for the k-means problem and for the projective clustering problem.

3.2 Sketch definition and dimension reduction

Now that we defined our problem, we can start to think about how to approximate it.

Definition 32. *Let $A \in \mathbb{R}^{d \times n}$ be an instance of the (ℓ_2^2, C)-clustering problem and let $\alpha \geq 1$ be a real. Let C^* be an optimal solution for A. We say that a solution $\widetilde{C}$ is an α-*approximation

for A if

$$\operatorname{dist}^2(A, \widetilde{C}) \leq \alpha \operatorname{dist}^2(A, C^*).$$

The real α is called approximation factor.

There are multiple strategies we can try to find an approximation for the (ℓ_2^2, C)-clustering problem. For example, instead of the family $\mathcal{C}$, we could consider just a subset of $\mathcal{C}$. Another approach is to find an instance $\widetilde{A}$ that is a "sketch" of the original instance A such that it would be "easier" to solve for the sketch, and any "good" approximation for $\widetilde{A}$ should give a good approximation for A.

In this section we will focus on the second approach. What would be a suitable and precise definition of sketch for an instance? Given the idea in Theorem 29 we require that the sketch must have low rank and that its optimal solutions must be an $(1+\varepsilon)$-approximation for A.

Obtaining low rank representations of A can be achieved by projecting it orthogonally to a low dimension subspace, and hence a candidate for the sketch would be $\widetilde{A} := BB^T A$ for some matrix B with orthonormal columns. To control the cost and guarantee that the sketch will give a good approximation, what naturally is desired is that $\widetilde{A}$ has the following property for ever $C \in \mathcal{C}$:

$$(1-\varepsilon)\operatorname{dist}^2(A, C) \leq \operatorname{dist}^2(\widetilde{A}, C) \leq (1+\varepsilon)\operatorname{dist}^2(A, C). \tag{3.1}$$

But the following fact shows us that this definition needs to be adjusted:

Fact 33. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, let V be a subspace of dimension j and let $C \subset V$ be a non-empty set. Suppose that $B \in \mathbb{R}^{d\times j}$ has orthonormal columns that spans V. Then*

$$\operatorname{dist}^2(A, C) = \|A - BB^T A\|_F^2 + \operatorname{dist}^2(BB^T A, C).$$

Proof. Apply the Pythagoras' Theorem to each point $a \in A$. Since $C \subset V$, we have that

$$(\operatorname{dist}(a, C))^2 = \|a - BB^T a\|^2 + (\operatorname{dist}(BB^T a, C))^2.$$

Adding this expression for all points we obtain what we desire. ∎

The fact above implies that the cost of all solutions in the space spanned by the columns of B for $\widetilde{A}$ will have an extra additive term of $\|A - BB^T A\|_F^2$ compared to the cost for A. This means that the usual definition in expression (3.1) will not usually work. Thus we give the following adjusted definition of sketch:

Definition 34 (ε-sketch for (ℓ_2^2, C)-clustering)**.** *Let $A \in \mathbb{R}^{d\times n}$ be an instance of the (ℓ_2^2, C)-clustering problem and let $\varepsilon \in (0,1)$ be fixed. We say that a matrix $\widetilde{A} \in \mathbb{R}^{d\times n}$ is an* ε-sketch *for A if there exists a non-negative constant $\Delta = \Delta(A, \mathcal{C}, \varepsilon)$ such that for every solution $C \in \mathcal{C}$ we have*

$$(1-\varepsilon)\operatorname{dist}^2(A, C) \leq \operatorname{dist}^2(\widetilde{A}, C) + \Delta \leq (1+\varepsilon)\operatorname{dist}^2(A, C). \tag{3.2}$$

Note that A is an ε-sketch for A for all $\varepsilon \in (0,1)$ and $\Delta = 0$. It is straightforward to deduce that for any $\alpha \geq 1$ and $\varepsilon \in (0, 1/3)$, a solution that is an α-approximation for the sketch $\widetilde{A}$ will be an $\alpha(1+3\varepsilon)$-approximation[1] for A. The proof is given later, in the proof of Fact 59. The argument is similar to the proof of Corollary 31.

Definition 34 is still strong in the sense that we impose the inequalities (3.2) for every $C \in \mathcal{C}$. But to just obtain an $(1+\varepsilon)$-approximation we can require less from $\widetilde{A}$:

Definition 35 (weak ε-sketch for $(\ell_2^2, \mathcal{C})$-clustering)**.** *Let $A \in \mathbb{R}^{d\times n}$ be an instance of the $(\ell_2^2, \mathcal{C})$-clustering problem and let $\varepsilon \in (0,1)$ be fixed. We say that $\widetilde{A} \in \mathbb{R}^{d\times n}$ is a* weak ε-sketch *for A if for every optimal solution $\widetilde{C}^*$ for $\widetilde{A}$ and any optimal solution C^* for A we have*

$$\operatorname{dist}^2(A, \widetilde{C}^*) \leq (1+\varepsilon)\operatorname{dist}^2(A, C^*). \tag{3.3}$$

Note that if $\widetilde{A}$ is an ε-sketch for A, then it is a weak 3ε-sketch if ε is small enough.

The result of Corollary 31 together with the definition of sketch give us the concept of approximation by *dimension reduction*. Our task now will be to find low-rank sketches of A, since solving the low dimensional sketch intuitively should be more efficient. In later chapters we will show that, indeed, solving the problems in low dimension is easier for linear projective clustering.

Definition 36 (Dimension reduction scheme)**.** *A* dimension reduction scheme *for the $(\ell_2^2, \mathcal{C})$-clustering problem in $\mathbb{R}^d$ is any scheme that takes an instance $A \in \mathbb{R}^{d\times n}$ and a parameter $\varepsilon \in (0,1)$ as input and outputs another instance $\widetilde{A} \in \mathbb{R}^{d\times n}$ such that $\operatorname{rank}(\widetilde{A}) < d$ and $\widetilde{A}$ is either an ε-sketch for A or a weak ε-sketch for A.*

We will present a dimension reduction scheme for the general case of $(\ell_2^2, \mathcal{C})$-clustering. The main tools we will use for the dimension reduction scheme will be the Johnson-Lindenstrauss Transform and the SVD.

Remark: *Most of the dimension reduction schemes we will present outputs an instance $\widetilde{A}$ with rank that depends on the family $\mathcal{C}$ and on the value of ε. This means that, although in principle any value in the range of $(0,1)$ is possible for ε, we cannot have ε arbitrary close to zero, as it could happen that the resulting rank of $\widetilde{A}$ would be higher than the dimension d of the space and we would not have a dimension reduction scheme. For example, if a scheme for k-means clustering in $\mathbb{R}^d$ outputs a sketch of rank $\lceil \varepsilon^{-2}k \rceil$, then it is clear that we must have $\varepsilon \geq \sqrt{k/d}$, else the rank of the sketch would be at least d.*

[1] Note that if $\varepsilon \in (0, 1/3)$ then $(1+\varepsilon)/(1-\varepsilon) \leq 1+3\varepsilon$.

Chapter 4

Approximation by dimension reduction

In this chapter we will present the main results of approximation by dimension reduction for the (ℓ_2^2, C)-clustering problem. In Section 4.1 we will study a weak ε-sketch for the best-fit linear subspace problem developed by Sarlós (2006). In Section 4.2 we will present sufficient conditions for a low rank matrix to be an ε-sketch for the (ℓ_2^2, C)-clustering problem. This result was first published by Feldman, Schmidt, and Sohler (2020) using the singular value decomposition, but Pratap and Sen (2018) improved it with the weak ε-sketch of Sarlós.

4.1 Random dimension reduction scheme for the best-fit linear j-subspace problem

We will return to the best-fit linear j-subspace problem introduced in Section 2.2, where we talked about how it is associated with the singular value decomposition. We showed that the SVD can be computed in time $O(nd^2)$ for any instance $A \in \mathbb{R}^{d \times n}$, assuming without loss of generality $d \leq n$.

But if $j \ll \min\{n, d\}$, we can find an approximation for the best-fit linear j-subspace problem more efficiently. Sarlós developed a randomized algorithm based on the Johnson-Lindenstrauss lemma that outputs a matrix $\widetilde{A}$ of rank $r = O(j/\varepsilon + j \log j)$ that with probability at least $1/2$ is a weak ε-sketch for A.

Theorem 37 (Sarlós, 2006). *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, let $j < \min\{d, n\}$ be an integer and let $\varepsilon \in (0,1)$ be fixed. There exists an integer $r = \Theta(\varepsilon^{-1}j + j\log j)$ such that if S is a $r \times n$ normalized Johnson-Lindenstrauss matrix then with probability at least $1/2$ we have*

$$\|A - \pi_{AS^T}(A)_j\|_F \leq (1+\varepsilon)\|A - A_j\|_F.$$

Computing $\pi_{AS^T}(A)_j$ can be done with two readings of the matrix A and in time $O(ndr + (n+d)r^2)$.

In other words, the theorem says that with probability at least $1/2$ the matrix $\pi_{AS^T}(A)$ is a weak 3ε-sketch for A for the problem of finding the best-fit linear j-subspace. A simple algorithm for computing $\pi_{AS^T}(A)$ is to generate an appropriate matrix S, compute the multiplication AS^T, find a basis for $\text{span}(AS^T)$ and project A orthogonally using this basis.

The time complexity analysis of this algorithm is straightforward. The product AS^T can be done in time $O(ndr)$ and requires one read of A. To realize the orthogonal projection, it is sufficient to apply the Gram-Schmidt process to find an orthonormal basis of $\text{span}(AS^T)$. This can be done in time $O(dr^2)$. Finally, we read the matrix A again to realize the projection, which can be done in time $O(ndr)$. We end up with an $r \times n$ matrix, and thus calculating the SVD of this matrix will take time $O(nr^2)$. Therefore the time needed to find $\pi_{AS^T}(A)_j$ is $O(ndr + (n+d)r^2)$.

To prove Theorem 37, we are going to need more tools and results.

Preliminary results for Theorem 37

Another property of the Johnson-Lindenstrauss Transform is about "almost" preserving subspaces. We can use these distributions to map any subspace L of $\mathbb{R}^d$ of dimension j into a space of dimension $O(j\log(j/\varepsilon)/\varepsilon^2)$ and "almost" preserve the norm of every vector of L with high probability. This seems counter-intuitive, since Theorem 22 is for finite sets only. But note that the dimension of the codomain space is higher than the dimension of L. It must be, else there would always be non-null vectors of L in the kernel of the mapping.

Lemma 38 (Sarlós, 2006). *Let $L \subset \mathbb{R}^d$ be a subspace of dimension j, and let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ be fixed. If f is drawn from some $JLT(\varepsilon/4, \delta, O((3j/\varepsilon)^j))$, then with probability at least $1-\delta$ for all $v \in L$ we have*

$$(1-\varepsilon)\|v\| \leq \|f(v)\| \leq (1+\varepsilon)\|v\|.$$

We say that such an f is a subspace ε-embedding.

It follows that an $r \times d$ normalized Johnson-Lindenstrauss Transform is a subspace ε-embedding for some $r = O(\varepsilon^{-2}j\log(j/\varepsilon))$. The intuition of the proof of this lemma is to consider a finite set H' such that every point in the unit sphere is close enough to a point of H' and use Theorem 22 in H'.

Definition 39 (δ-fine grid). *Fix $\delta \in (0,1)$. Let $e_1, \ldots, e_j$ be the vectors of the canonical base*

of $\mathbb{R}^j$. *The set*

$$H := \left\{ h \in \mathbb{R}^j \,:\, \forall i = 1, \dots, j, \forall c_i \in \mathbb{Z} \cap [-\delta^{-1}, \delta^{-1}], h = \sum_{i=1}^{j} \delta c_i e_i \right\}.$$

is called δ-fine grid *of dimension* j.

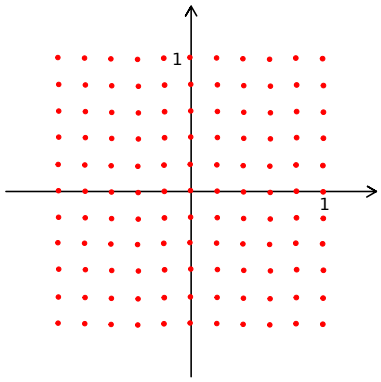


Figure 4.1: *An example of a* $1/5$*-fine grid of dimension 2. The red dots are the elements of the grid.*

Proof of Lemma 38. To facilitate the calculations, we will use ε instead of $\varepsilon/4$ in the hypothesis. When the proof is done the result will follow by re-scaling ε.

Since f is linear we have that $\|f(0)\| = 0$, and for every non-null $x \in L$ we have that $f(x) = \|x\| f(x/\|x\|)$, thus we just need to prove the lemma for the case $\|x\| = 1$. Let $B \in \mathbb{R}^{d \times j}$ with orthonormal columns $b_1, \dots, b_j$ that spans L. Let $\delta = \min\{\sqrt{\varepsilon}/j, \varepsilon/\sqrt{j}\}$ and let H be a δ-fine grid of dimension j. Fix $H' = \{Bh | h \in H\}$. Note that $|H'| = O((3j/\varepsilon)^j)$.

We will apply Corollary 26 in the set $H' \bigcup \{b_1, \dots, b_j\}$. Note that any vector of L can be written as $Bw = \sum_{p=1}^{j} \alpha_p b_p$, for some $w = (\alpha_1, \dots, \alpha_j) \in \mathbb{R}^j$, therefore for any $Bw \in L$

$$\begin{aligned}
\|f(Bw)\|^2 &= \langle f(Bw), f(Bw) \rangle \\
&= \sum_{p=1}^{j} \sum_{q=1}^{j} \left\langle \alpha_p f(b_p), \alpha_q f(b_q) \right\rangle && \text{by linearity} \\
&\leq \sum_{p=1}^{j} \sum_{q=1}^{j} |\alpha_p||\alpha_q| \left(\left\langle b_p, b_q \right\rangle + \varepsilon \right) && \text{by Corollary 26.}
\end{aligned}$$

Note that $\left\langle b_p, b_q \right\rangle$ is zero for $p \neq q$. This means that

$$\|f(Bw)\|^2 \leq \|w\|^2 + \varepsilon \sum_{p=1}^{j} \sum_{q=1}^{j} |\alpha_p||\alpha_q| = \|w\|^2 + \varepsilon \|w\|_1^2, \tag{4.1}$$

where the norm 1 of a vector w is given by

$$\|w\|_1 = \sum_{i=1}^{j} \alpha_i.$$

Remember Fact 13, which states that if a matrix B has orthonormal columns, then $\|w\| = \|Bw\|$. So if $\|w\|_1$ is small enough, we have that $\|f(Bw)\|$ is close to $\|Bw\|$.

Now let $v \in \mathbb{R}^j$ with $\|v\| = 1$. Each coordinate of v is in the interval $[-1, 1]$. This implies that there exists an $h \in H$ such that each coordinate of h differs in at most δ from v and $\|h\| \leq 1$. Therefore

$$\|v - h\| \leq \delta\sqrt{j} \leq \varepsilon \tag{4.2}$$

and

$$\|v - h\|_1 \leq \delta j \leq \sqrt{\varepsilon}. \tag{4.3}$$

By the triangle inequality in $\|(v - h) + h\|$ it follows

$$1 - \varepsilon \leq \|h\| = \|Bh\|.$$

Since f is drawn from some $JLT(\varepsilon, \delta, O((3j/\varepsilon)^j))$, we can assume

$$(1 - \varepsilon)^2 \leq (1 - \varepsilon)\|Bh\| \leq \|f(Bh)\| \leq (1 + \varepsilon)\|Bh\| = (1 + \varepsilon). \tag{4.4}$$

Let x be an arbitrary vector of the unity sphere of L. Then exists an $v \in \mathbb{R}^j$ with $\|v\| = 1$ such that $x = Bv$. Let $w = v - h$ and apply inequality (4.1). We obtain

$$\|f(B(v - h))\|^2 \leq \|v - h\|^2 + \varepsilon\|v - h\|_1^2.$$

Applying inequalities (4.2) and (4.3) we have

$$\|v - h\|^2 + \varepsilon\|v - h\|_1^2 \leq \varepsilon^2 + \varepsilon^2 = 2\varepsilon^2. \tag{4.5}$$

By linearity, triangle inequality and the right-hand side of inequality (4.4) we have

$$\|f(Bv)\| \leq \|f(B(v - h))\| + \|f(Bh)\| \leq \sqrt{2}\varepsilon + 1 + \varepsilon \leq 1 + 3\varepsilon.$$

The lower bound can be obtained by applying the triangle inequality in the vector $\|f(Bh)\| = \|f(Bv) + f(B(h - v))\|$ and using the left-hand side of inequality (4.4) to get

$$\|f(Bv)\| \geq \|f(Bh)\| - \|f(B(v - h))\| \geq (1 - \varepsilon)^2 - \sqrt{2}\varepsilon \geq 1 - 4\varepsilon,$$

since for $\varepsilon \in (0, 1)$ we have $1 - 2\varepsilon \leq (1 - \varepsilon)^2$. Re-scaling the ε finishes the proof. ∎

Corollary 40 (Sarlós, 2006)**.** *Let $U \in \mathbb{R}^{d \times j}$ be a matrix with orthonormal columns and let $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$ be fixed. Suppose that S is an $r \times d$ normalized Johnson-Linden-*

strauss matrix for some $r = O(\varepsilon^{-2} j \log(j/\varepsilon) \log(1/\delta))$. *Then with probability at least* $1 - \delta$ *for all* $i = 1, \ldots, j$ *we have*

$$|1 - \sigma_i(SU)| \leq \varepsilon.$$

Proof. From Lemma 38 we have that the function $v \mapsto Sv$ is a subspace ε-embedding for span(U). So for all norm-one vector v of $\mathbb{R}^d$

$$1 - \varepsilon = (1 - \varepsilon) \|Uv\| \leq \|SUv\| \leq (1 + \varepsilon) \|Uv\| = 1 + \varepsilon.$$

The result follows by the definition of singular value. ∎

The next preliminary result is an application of Theorem 22 and subspaces ε-embeddings in linear ℓ_2 regression.

Definition 41 (linear ℓ_2 regression)**.** *Let* $A \in \mathbb{R}^{d \times n}$ *be a matrix and let* $b \in \mathbb{R}^d$ *be fixed. The* linear ℓ_2 regression problem *is to find at least one vector* $x^* \in \mathbb{R}^n$ *such that for all* $x \in \mathbb{R}^n$ *we have*

$$\|b - Ax^*\| \leq \|b - Ax\|.$$

In other words, we want to find the linear combination of the columns of A that is the closest to b in the Euclidean norm.

The idea to find x^* is straightforward. Remember that the set of all linear combinations of the columns of A is a subspace of $\mathbb{R}^d$, and the point of this subspace that is closest to b is the orthogonal projection of b. Let $U\Sigma V^T$ be a SVD of A. The projection of b to span(A) can be written as $UU^T b$. This means that we can reformulate the problem as to find a vector $x \in \mathbb{R}^n$ such that

$$Ax = UU^T b.$$

We know that such an x exists since $UU^T b$ lies in the column span of A. Note that if we rewrite UU^T as

$$UU^T = U\Sigma\Sigma^{-1}U^T = U\Sigma V^T V\Sigma^{-1}U^T$$

we obtain

$$Ax = A\left(V\Sigma^{-1}U^T\right) b,$$

thus a vector that satisfy the above expression is

$$x^* := V\Sigma^{-1}U^T b.$$

Define $A^\dagger := V\Sigma^{-1}U^T$. We call the matrix $A^\dagger$ the *Moore-Penrose generalized inverse* of A. It follows that a solution can be written as $x^* = A^\dagger b$.

Remark: *This generalized inverse has some interesting properties. By the above discussion we can see that the orthogonal projection of a vector* b *to* span(A) *is* $AA^\dagger b$. *Also it is a direct calculation to show that* $A^T AA^\dagger = A^T$ *and* $AA^\dagger A = A$.

The normalized Johnson-Lindenstrauss matrix can be used to find an approximation for the linear ℓ_2 regression problem.

Theorem 42 (Sarlós, 2006). *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and let $b \in \mathbb{R}^d$ and $\varepsilon \in (0,1)$ be fixed. Let $\mathcal{Z} = \min_{x\in\mathbb{R}^n} \|b - Ax\| = \|b - Ax^*\|$, where $x^* = A^\dagger b$. Let $S \in \mathbb{R}^{s\times d}$ be a normalized Johnson-Lindenstrauss matrix for some positive integer $s = O(\varepsilon^{-1} n \log n)$. Finally, let $\widetilde{\mathcal{Z}} = \min_{x\in\mathbb{R}^n} \|Sb - SAx\| = \|Sb - SA\widetilde{x}^*\|$, where $\widetilde{x}^* = (SA)^\dagger Sb$. Then with probability at least $1/3$ we have*

$$\|b - A\widetilde{x}^*\| \leq (1+\varepsilon)\mathcal{Z}.$$

Note that Sb and SA act similarly to a weak ε-sketch, since its optimal solution is an approximation for the original problem.

Proof. Let $A = U\Sigma V^T$ be a SVD of A. Suppose that $r = \operatorname{rank}(A)$. Since $\operatorname{span}(U) = \operatorname{span}(A)$, there exists vectors α and $\beta \in \mathbb{R}^r$ such that

$$Ax^* = U\alpha$$

and

$$A\widetilde{x}^* - Ax^* = U\beta.$$

Let $w \in \mathbb{R}^d$ be a vector such that $w = b - Ax^*$. Note that w is orthogonal to $\operatorname{span}(A)$ and that $\|w\| = \mathcal{Z}$. Therefore

$$\begin{aligned} \|b - A\widetilde{x}^*\|^2 &= \|b - Ax^* + Ax^* - A\widetilde{x}^*\|^2 && (4.6)\\ &= \|w - U\beta\|^2 && (4.7)\\ &= \mathcal{Z}^2 + \|\beta\|^2 && (4.8)\end{aligned}$$

This means that in order to bound $\|b - A\widetilde{x}^*\|$ we need to bound $\|\beta\|$. Observe that

$$SU(\alpha+\beta) = SA\widetilde{x}^*, \tag{4.9}$$

and from the definition of $\widetilde{x}^*$ we have

$$SA\widetilde{x}^* = SA(SA)^\dagger Sb = \pi_{SA}(Sb).$$

Since $b = w + Ax^* = w + U\alpha$ we have

$$\pi_{SA}(Sb) = \pi_{SA}(SU\alpha + Sw).$$

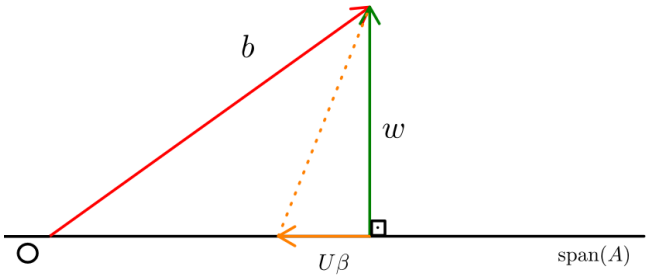
Since $\operatorname{span}(A) = \operatorname{span}(U)$, the orthogonal projector π_{SA} is equal to π_{SU}, which implies

$$\pi_{SA}(SU\alpha + Sw) = \pi_{SU}(SU\alpha + Sw) = SU\alpha + \pi_{SU}(Sw).$$

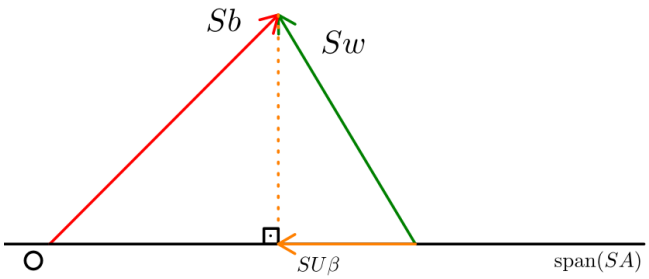
From equation (4.9), we have

$$SU\beta = \pi_{SU}(Sw).$$

Remember that we can write $\pi_{SU}(Sw)$ as $(SU)(SU)^\dagger Sw$. Using the property of the



(a) *The vector w (green) is orthogonal to* span(A).



(b) *The vector Sw (green) may not be orthogonal to* span(SA). *Its orthogonal projection equals $SU\beta$ (orange).*

Figure 4.2: *Visualization of the relation between $SU\beta$ and Sw.*

Moore-Penrose generalized inverse that for any matrix A we have $A^T AA^\dagger = A^T$, we obtain

$$(SU)^T SU\beta = (SU)^T Sw. \tag{4.10}$$

Fact 43. *There exists some $s = O(\varepsilon^{-1} n \log n)$ such that with probability at least $2/3$ we have*

$$\sigma_i((SU)^T(SU)) = \sigma_i^2(SU) \geq \frac{1}{\sqrt{2}}.$$

Indeed, by choosing the constants of s correctly we can apply Corollary 40 with $\varepsilon' = 1 - 2^{-1/4}$ and $\delta = 1/3$. This fact implies that with probability at least $2/3$

$$\frac{\|\beta\|^2}{2} \leq \left\|(SU)^T(SU)\beta\right\|^2 = \left\|U^T S^T Sw\right\|^2. \tag{4.11}$$

With the same argument of choosing the constants of s correctly, we observe that S is also a $JLT(\sqrt{\varepsilon/(2n)}, 1/3, n+2)$, and thus we can apply Lemma 28.1 to U^T and w with $\varepsilon'' = \sqrt{\varepsilon/(2n)}$ and $\delta = 1/3$ to conclude that with probability at least $2/3$ we have

$$\begin{aligned} \left\|U^T S^T Sw\right\|^2 &= \left\|U^T w - U^T S^T Sw\right\|^2 \\ &\leq \frac{\varepsilon}{n} \left\|U^T\right\|_F^2 \|w\|^2 \\ &\leq \varepsilon \|w\|^2 = \varepsilon \mathcal{Z}^2. \end{aligned} \tag{4.12}$$

Using the union bound in the events (4.11) and (4.12), we have that with probability at least $1/3$

$$\|\beta\|^2 \leq 2\varepsilon\mathcal{Z}^2,$$

and thus by equation (4.8) we have

$$\|b - A\widetilde{x}^*\| \leq \sqrt{1+2\varepsilon}\mathcal{Z} \leq (1+\varepsilon)\mathcal{Z},$$

finishing the proof. ∎

Now we are ready to prove Theorem 37. Before presenting the full proof, we will first give an intuition of each step. The proof starts by changing the projector matrix π_{AS^T} to a weaker one. This will allow us to manipulate the inequality of the theorem and reduce part of it as a linear ℓ_2 regression problem. We can then use a similar argument of the proof of Theorem 42 by applying Corollary 40 with $\delta = 1/4$ and $\varepsilon = 1 - 2^{-1/4}$, requiring the hypothesis of $r = \Omega(j \log j)$. Next we will apply Lemma 28.2, but instead of ε we will use $\sqrt{\varepsilon/j}$, requiring the hypothesis of $r = \Omega(\varepsilon/j)$. This justify the value of $\Theta(\varepsilon/j + j\log j)$ for r.

proof of Theorem 37. Let $A = U\Sigma V^T$ be a SVD of A, and let $\rho = \operatorname{rank}(A)$. We can assume that $j < \rho$, otherwise $A_j = A$. Let $P = (\pi_{AS^T}(A_j))(\pi_{AS^T}(A_j))^\dagger$ be a projector matrix to $\operatorname{span}(\pi_{AS^T}(A_j))$. Since $\operatorname{rank}(\pi_{AS^T}(A_j))$ is at most j, and $\pi_{AS^T}(A)_j$ is a best rank j approximation of $\pi_{AS^T}(A)$, we have that

$$\left\|A - \pi_{AS^T}(A)_j\right\|_F^2 \leq \|A - PA\|_F^2 .$$

Now we will bound the value of $\|A - PA\|_F^2$. By the unitary invariance of the Frobenius norm (Fact 13) we have

$$\|A - PA\|_F^2 = \left\|U\Sigma V^T - PU\Sigma V^T\right\|_F^2 = \|U\Sigma - PU\Sigma\|_F^2 .$$

Let $U^{(j)} \in \mathbb{R}^{d\times j}$ be the matrix that agrees with the first j columns of U and $U^{(\rho-j)} \in \mathbb{R}^{d\times(\rho-j)}$ be the matrix that agrees with the last $\rho - j$ columns of U. Analogously, let $V^{(j)}$ and $V^{(\rho-j)}$ be the matrices that agrees with the first j and last $\rho - j$ columns of V, respectively.

Let $\Sigma^{(j)} \in \mathbb{R}^{j\times j}$ be the matrix with diagonal elements that agrees with the first j elements of the diagonal of Σ. Let $\Sigma^{(\rho-j)} \in \mathbb{R}^{(\rho-j)\times(\rho-j)}$ be the diagonal matrix that agrees with the last $\rho - j$ elements of the diagonal of Σ. By Pythagoras' Theorem we have that

$$\|U\Sigma - PU\Sigma\|_F^2 = \left\|U^{(j)}\Sigma^{(j)} - PU^{(j)}\Sigma^{(j)}\right\|_F^2 + \left\|U^{(\rho-j)}\Sigma^{(\rho-j)} - PU^{(\rho-j)}\Sigma^{(\rho-j)}\right\|_F^2 .$$

Note that orthogonal projections never increase the norm of vectors. Hence we can

bound the last term as follows:

$$\begin{aligned}\left\|U^{(\rho-j)}\Sigma^{(\rho-j)} - PU^{(\rho-j)}\Sigma^{(\rho-j)}\right\|_F^2 &= \left\|(I-P)\left(U^{(\rho-j)}\Sigma^{(\rho-j)}\right)\right\|_F^2 \\ &\leq \left\|U^{(\rho-j)}\Sigma^{(\rho-j)}\right\|_F^2 \quad \text{since } I-P \text{ is an orthogonal projector} \\ &= \left\|A - A_j\right\|_F^2.\end{aligned}$$

To finish the proof is enough to show that $\left\|U^{(j)}\Sigma^{(j)} - PU^{(j)}\Sigma^{(j)}\right\|_F^2 \leq 2\varepsilon \left\|A - A_j\right\|^2$ with probability at least $1/2$, since combining it with the previous equations we arrive at

$$\|A - PA_j\|_F \leq \sqrt{1+2\varepsilon}\left\|A - A_j\right\|_F \leq (1+\varepsilon)\left\|A - A_j\right\|_F.$$

Note that by unitarily invariance we have

$$\begin{aligned}\left\|U^{(j)}\Sigma^{(j)} - PU^{(j)}\Sigma^{(j)}\right\|_F^2 &= \left\|U^{(j)}\Sigma^{(j)}V^{(j)T} - PU^{(j)}\Sigma^{(j)}V^{(j)T}\right\|_F^2 \\ &= \left\|A_j - PA_j\right\|_F^2.\end{aligned}$$

Remember that $PA_j = (AS^T)(AS^T)^\dagger A_j$ is an orthogonal projection of A_j to $\text{span}(AS^T)$. This means that

$$\begin{aligned}\left\|A_j - PA_j\right\|_F^2 &\leq \left\|A_j - (AS^T)(A_jS^T)^\dagger A_j\right\|_F^2 \\ &= \left\|A_j^T - A_j^T(SA_j^T)^\dagger(SA^T)\right\|_F^2.\end{aligned}$$

Let us study the following linear ℓ_2 regression problems: Let $b_i \in \mathbb{R}^n$ denote the i^{th} column of A^T, and $b_{(j),i}$ the i^{th} column of A_j^T. For $i = 1, \ldots, d$, find x_i^* and $\widetilde{x}_i^* \in \mathbb{R}^d$ such that they minimize

$$\left\|b_i - A_j^T x_i^*\right\|$$

and

$$\left\|Sb_i - SA_j^T\widetilde{x}_i^*\right\|.$$

We know from previous discussions that the minimizers for the first problem are the orthogonal projections of b_i to $\text{span}(A_j^T)$, and from the SVD theory they are $b_{(j),i}$. From the proof of Theorem 42 we can define vectors $\beta_1, \ldots, \beta_d \in \mathbb{R}^j$ and $w_1, \ldots, w_d \in \mathbb{R}^n$ such that for all $i = 1, \ldots, d$ such that

$$V^{(j)}\beta_i = A_j^T\widetilde{x}_i^* - A_j^T x_i^* = A_j^T\widetilde{x}_i^* - b_{(j),i},$$

$$w_i = b_i - A_j^T x_i^* = b_i - b_{(j),i}$$

and

$$V^{(j)T}S^TSV^{(j)}\beta_i = V^{(j)T}S^TSw_i.$$

From the results about linear ℓ_2 regression we know that $\widetilde{x}_i^* = (SA_j^T)^\dagger Sb_i$. From the

definition of Frobenius norm and its unitary invariance we can deduce that

$$\begin{aligned}\sum_{i=1}^{d} \|\beta_i\|^2 &= \sum_{i=1}^{d} \left\|V^{(j)}\beta_i\right\|^2 \\ &= \sum_{i=1}^{d} \left\|A_j^T (SA_j^T)^\dagger Sb_i - b_{(j),i}\right\|^2 \\ &= \left\|A_j^T - A_j^T (SA_j^T)^\dagger SA^T\right\|_F^2 .\end{aligned}$$

Similarly to the argument we used to obtain inequality (4.11), we can take $r = \Omega(j \log j)$ such that it is enough to apply Corollary 40 in the matrix $V^{(j)}$ with $\delta = 1/4$ and with $\varepsilon' = 1 - 2^{-1/4}$, obtaining with probability at least $3/4$ that for all $k = 1, \dots, j$ we have

$$2^{-1/4} \leq \sigma_k(SV^{(j)}),$$

and hence

$$\frac{1}{\sqrt{2}} \leq \sigma_k^2(SV^{(j)}) = \sigma_k(V^{(j)T} S^T SV^{(j)}).$$

This implies that with probability at least $3/4$ for every vector $y \in \mathbb{R}^j$ we have

$$\frac{\|y\|^2}{2} \leq \left\|V^{(j)T} S^T SV^{(j)} y\right\|^2, \tag{4.13}$$

and in particular, for all $i = 1, \dots, d$ we have

$$\frac{\|\beta_i\|^2}{2} \leq \left\|V^{(j)T} S^T SV^{(j)} \beta_i\right\|^2 = \left\|V^{(j)T} S^T Sw_i\right\|^2 . \tag{4.14}$$

Now again with an argument similar to the one we used to obtain inequality (4.12), we bound the following expression:

$$\mathbb{E}\left[\sum_{i-1}^{d} \left\|V^{(j)T} S^T Sw_i\right\|^2\right] = \sum_{i=1}^{d} \mathbb{E}\left[\left\|V^{(j)T} S^T Sw_i\right\|^2\right].$$

Again taking $r = \Omega(j/\varepsilon)$ big enough, we can apply the second statement of Lemma 28 with $\varepsilon' = \sqrt{\varepsilon/(2j)}$ to $(V^{(j)})^T$ and w_i, obtaining

$$\begin{aligned}\sum_{i=1}^{d} \mathbb{E}\left[\left\|V^{(j)T} S^T Sw_i\right\|^2\right] &\leq \sum_{i=1}^{d} \varepsilon \|w_i\|^2 \\ &= \varepsilon \sum_{i=1}^{d} \left\|b_i - b_{(j),i}\right\|^2 = \varepsilon \left\|A^T - A_j^T\right\|_F^2 .\end{aligned} \tag{4.15}$$

Applying Markov Inequality to inequality (4.15) we have that

$$\mathbb{P}\left[\sum_{i=1}^{d}\left\|V^{(j)T}S^TSw_i\right\|^2 \geq 4\varepsilon\left\|A^T - A_j^T\right\|_F^2\right] \leq \frac{\sum_{i=1}^{d}\mathbb{E}\left[\left\|V^{(j)T}S^TSw_i\right\|^2\right]}{4\varepsilon\left\|A^T - A_j^T\right\|_F^2}$$
$$\leq \frac{\varepsilon\left\|A^T - A_j^T\right\|_F^2}{4\varepsilon\left\|A^T - A_j^T\right\|_F^2} = \frac{1}{4}. \tag{4.16}$$

Finally by using the union bound on equations (4.14) and (4.16), with probability at least $1/2$ we have that

$$\left\|U^{(j)} - PU^{(j)}\Sigma^{(j)}\right\|_F^2 \leq \sum_{i=1}^{d}\|\beta_i\|^2 \leq 8\varepsilon\left\|A^T - A_j^T\right\|_F^2.$$

Rescaling ε give the desired result. ∎

With Theorem 37 we have obtained a randomized algorithm to find a weak ε-sketch for the best-fit linear j-subspace problem. We will use this result next chapter to present an algorithm that finds an ε-sketch for the more general (ℓ_2^2, C)-clustering problem.

Remark: *The probability of success of Theorem 37 can be boosted to* $1-\delta$*, for any* $\delta \in (0,1)$*. Note that by Fact 10 (Pythagoras' Theorem) we have*

$$\|A\|_F^2 = \left\|A - \pi_{AS^T}(A)_j\right\|_F^2 + \left\|\pi_{AS^T}(A)_j\right\|_F^2,$$

therefore if we run $\log_2(1/\delta)$ *independent instances of* S *and choose the one that maximizes* $\left\|\pi_{AS^T}(A)_j\right\|_F^2$*, we will have a probability of success of* $1-\delta$*. Indeed, for this procedure to give us a wrong answer, all of the* $\log_2(1/\delta)$ *independent instances must yield the wrong answer, and this happens with probability* $2^{-\log_2 1/\delta} = \delta$*.*

4.2 Finding an ε-sketch for the (ℓ_2^2, C)-clustering problem

In this section we will present an ε-sketch for the (ℓ_2^2, C)-clustering problem. This result is a slight improvement on the work of Pratap and Sen, since the authors only dealt with the case for the projective clustering problem. The only restriction we require on the family C is in the dimensionality of its elements, as we will define later. Otherwise this sketch works for any C, thus it can be applied for the projective clustering problem and others.

This sketch was first developed by Feldman, Schmidt, and Sohler, but Pratap and Sen improved it by applying the weak ε-sketch of Sarlós and using other results from Cohen *et al.* (2015).

Definition 44. *Let* C *be a non-empty family of non-empty subsets of* $\mathbb{R}^d$ *and let* $m \leq d$ *be a positive integer. We say that* C *is an* m-dimensional family *if for every* $C \in C$ *there exists a*

linear subspace $L(C) \subset \mathbb{R}^d$ of dimension m that contains C.

For example, in the k-means clustering problem in $\mathbb{R}^d$ the set $\mathcal{C}$ is a k-dimensional family if $k \leq d$, since for each set C of k points the subspace spanned by C has dimension at most k, and thus can be completed to some subspace L of dimension k. Similarly in the linear j-subspace k-clustering problem if $jk \leq d$ then the set $\mathcal{C}$ is a jk-dimensional family.

The main result we are going to present in this section is the following theorem.

Theorem 45 (based on Pratap and Sen, 2018)**.** *Let $A \in \mathbb{R}^{d\times n}$ be an instance of the $(\ell_2^2, \mathcal{C})$-clustering problem where $\mathcal{C}$ is an m-dimensional family and $m < \min\{d, n\}$. Let $\varepsilon \in (0,1)$ and let $s = \lceil 8\varepsilon^{-2}m \rceil$ be fixed. If a matrix $\widetilde{A}^T \in \mathbb{R}^{n\times d}$ is an orthogonal projection of A^T to some subspace of dimension s of $\mathbb{R}^n$ and satisfies*

$$\left\|A - \widetilde{A}\right\|_F^2 \leq \left(1 + \frac{\varepsilon^2}{8}\right) \|A - A_s\|_F^2 \tag{4.17}$$

then $\widetilde{A}$ is an ε-sketch for A with constant $\Delta = \|A - A_s\|_F^2$.

We are going to postpone the proof of Theorem 45, because we need some preliminary results.

Note that taking $\widetilde{A} = A_s$ trivially satisfies equation (4.17) and Theorem 45 hypothesis, hence A_s is an ε-sketch for A. Therefore a simple algorithm for finding an ε-sketch is to calculate the SVD of A. But this takes time $O(nd^2)$, and as we saw in the previous section this can be improved with Theorem 37 since it is sufficient to find an approximation of it. We present the improved algorithm bellow.

Algorithm 1: A randomized dimension reduction scheme for $(\ell_2^2, \mathcal{C})$-clustering when $\mathcal{C}$ is m-dimensional

Input: An instance $A \in \mathbb{R}^{d\times n}$ and a parameter $\varepsilon \in (0,1)$.
Output: A matrix that with probability at least $1/2$ is an ε-sketch for A.

1 Let $s := \lceil 8\varepsilon^{-2}m \rceil$ be given by Theorem 45;
2 Let $r := \Theta(\varepsilon^{-2}s + s\log s)$ be given by Theorem 37;
3 Let $S \in \mathbb{R}^{r\times d}$ be a normalized Johnson-Lindenstrauss matrix;
4 Compute $\widetilde{A}^T = \pi_{A^T S^T}(A^T)_s$;
5 Return $\widetilde{A}$;

The correctness of this algorithm follows directly from Theorem 37 and 45. Theorem 37 guarantees that the matrix $\widetilde{A}$ computed on line 5 satisfies equation (4.17) with probability at least $1/2$. Theorem 45 guarantees that such an $\widetilde{A}$ is an ε-sketch for A. The time complexity follows from Theorem 37 too, which states that $\widetilde{A}$ can be computed in time

$$\begin{aligned} O(ndr + (n+d)r^2) &= O(nd(s\varepsilon^{-2} + s\log s) + (n+d)(\varepsilon^{-2}s + s\log s)^2) \\ &= O(nd(\varepsilon^{-4}m + (\varepsilon^{-2}m\log(\varepsilon^{-2}m))) + (n+d)(\varepsilon^{-8}m^2 + (\varepsilon^{-4}m^2\log^2(\varepsilon^{-2}m)))). \end{aligned}$$

The probability of success of this algorithm can be boosted to $1-\delta$, for any $\delta \in (0,1)$, by running $\log_2(1/\delta)$ independent copies of the matrix S and choosing the one that

maximizes $\left\|\pi_{A^T S^T}(A^T)\right\|_F$, as we saw in the previous section.

To prove Theorem 45 we will require some technical results. The first is a weak triangle inequality between points and sets. The proof of it is in appendix A.

Lemma 46 (Weak triangle inequality)**.** *Let $p, q \in \mathbb{R}^d$ be fixed and let $C \subset \mathbb{R}^d$ be a non-empty set. For all $\varepsilon \in (0,1)$ we have*

$$\left|\operatorname{dist}^2(p,C) - \operatorname{dist}^2(q,C)\right| \leq \varepsilon \operatorname{dist}^2(p,C) + \frac{2}{\varepsilon}\|p-q\|^2 \tag{4.18}$$

The second result states that for every subspace of dimension m the Frobenius norm of the projection of A to this subspace is relatively close to the Frobenius norm of the projection of $\widetilde{A}$ to the same subspace.

Lemma 47. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and suppose that $m < \min\{d,n\}$ is a positive integer. Fix $\varepsilon' \in (0,1)$ and $s = \lceil \varepsilon'^{-1} m \rceil$. Suppose that $\widetilde{A}^T \in \mathbb{R}^{n\times d}$ is a matrix that is an orthogonal projection of A^T to some subspace of dimension s of $\mathbb{R}^n$ and satisfies*

$$\left\|A - \widetilde{A}\right\|_F^2 \leq (1+\varepsilon')\|A - A_s\|_F^2 . \tag{4.19}$$

For every matrix $X \in \mathbb{R}^{d\times m}$ with orthonormal columns and matrix $Y \in \mathbb{R}^{d\times(d-m)}$ with orthonormal columns such that $\operatorname{span}(Y)$ is the orthogonal complement of $\operatorname{span}(X)$ the following inequalities are true:

$$0 \leq \left\|X^T A\right\|_F^2 - \left\|X^T \widetilde{A}\right\|_F^2 \leq 2\varepsilon' \left\|Y^T A\right\|_F^2 \tag{4.20}$$

and

$$\left\|XX^T A - XX^T \widetilde{A}\right\|_F^2 \leq 2\varepsilon' \left\|Y^T A\right\|_F^2 . \tag{4.21}$$

Proof. We will start proving (4.20) by expressing it in the trace function form.

$$\left\|X^T A\right\|_F^2 - \left\|X^T \widetilde{A}\right\|_F^2 = \operatorname{Tr}(A^T XX^T A) - \operatorname{Tr}(\widetilde{A}^T XX^T \widetilde{A}).$$

By the cyclic propriety of the trace function, we have

$$\operatorname{Tr}(A^T XX^T A) - \operatorname{Tr}(\widetilde{A}^T XX^T \widetilde{A}) = \operatorname{Tr}(XX^T AA^T) - \operatorname{Tr}(XX^T \widetilde{A}\widetilde{A}^T).$$

The trace function is linear, therefore

$$\operatorname{Tr}(XX^T AA^T) - \operatorname{Tr}(XX^T \widetilde{A}\widetilde{A}^T) = \operatorname{Tr}(XX^T(AA^T - \widetilde{A}\widetilde{A}^T)).$$

Let $M = AA^T - \widetilde{A}\widetilde{A}^T$. We will show that M is equal to $(A-\widetilde{A})(A-\widetilde{A})^T$, and thus is

positive semi-definite. Note that

$$\begin{aligned} M &= AA^T - \widetilde{A}\widetilde{A}^T = AA^T - (A - A + \widetilde{A})(A - A + \widetilde{A})^T \\ &= AA^T - (A - (A - \widetilde{A}))(A^T - (A - \widetilde{A})^T) \\ &= AA^T - AA^T + A(A - \widetilde{A})^T + (A - \widetilde{A})A^T - (A - \widetilde{A})(A - \widetilde{A})^T \\ &= (A - \widetilde{A} + \widetilde{A})(A - \widetilde{A})^T + (A - \widetilde{A})(A - \widetilde{A} + \widetilde{A})^T - (A - \widetilde{A})(A - \widetilde{A})^T \\ &= (A - \widetilde{A})(A - \widetilde{A})^T + \widetilde{A}(A - \widetilde{A})^T + (A - \widetilde{A})(A - \widetilde{A})^T + (A - \widetilde{A})\widetilde{A}^T - (A - \widetilde{A})(A - \widetilde{A})^T. \end{aligned} \tag{4.22}$$

Note that the column span of $\widetilde{A}^T$ is orthogonal to the column span of $(A - \widetilde{A})^T$, which means that $\widetilde{A}(A - \widetilde{A})^T = (A - \widetilde{A})\widetilde{A}^T = 0$, and from equation (4.22) we get

$$M = (A - \widetilde{A})(A - \widetilde{A})^T.$$

Next we will show that the left-hand side of (4.21) is also equal to $\mathrm{Tr}(XX^TM)$. By the relation between the Frobenius norm and the trace function, we have

$$\begin{aligned} \left\|XX^TA - XX^T\widetilde{A}\right\|_F^2 &= \left\|XX^T(A - \widetilde{A})\right\|_F^2 \\ &= \mathrm{Tr}((A - \widetilde{A})^TXX^TXX^T(A - \widetilde{A})). \end{aligned}$$

By the cyclic property of the trace and by the fact that X has orthonormal columns, we have

$$\mathrm{Tr}((A - \widetilde{A})^TXX^T(A - \widetilde{A})) = \mathrm{Tr}(XX^T(A - \widetilde{A})(A - \widetilde{A})^T).$$

We claim that $\mathrm{Tr}(XX^TM) = \left\|XX^T(A - \widetilde{A})\right\|_F^2 \leq \left\|(A - \widetilde{A})_m\right\|_F^2$. By the SVD theory, we have that

$$\left\|(A - \widetilde{A}) - (A - \widetilde{A})_m\right\|_F^2 \leq \left\|(A - \widetilde{A}) - XX^T(A - \widetilde{A})\right\|_F^2,$$

since $\mathrm{rank}(XX^T(A - \widetilde{A}))$ is at most m. By Fact 10 (Pythagoras) it follows that

$$\left\|(A - \widetilde{A})\right\|_F^2 - \left\|(A - \widetilde{A})_m\right\|_F^2 \leq \left\|(A - \widetilde{A})\right\|_F^2 - \left\|XX^T(A - \widetilde{A})\right\|_F^2,$$

which implies

$$\left\|XX^T(A - \widetilde{A})\right\|_F^2 \leq \left\|(A - \widetilde{A})_m\right\|_F^2.$$

Now we need to prove that $\left\|(A - \widetilde{A})_m\right\|_F^2 \leq 2\varepsilon' \left\|A - A_m\right\|_F^2$. We start with the following fact.

Fact 48. *Let A and $B \in \mathbb{R}^{d\times n}$ be matrices. Then* $\mathrm{rank}(A + B)$ *is at most* $\mathrm{rank}(A) + \mathrm{rank}(B)$.

This fact follows by noting that for any vector $v \in \mathbb{R}^d$ we have $(A + B)v = Av + Bv$, therefore if α is a basis for $\mathrm{span}(A)$ and β is a basis for $\mathrm{span}(B)$, then $\mathrm{span}(A + B)$ is

contained in the subspace spanned by $\alpha \cup \beta$.

By the fact above, we have that the rank of $\widetilde{A} + (A - \widetilde{A})_m$ is at most $s + m$, thus by the property of best rank $s + m$ approximation in the Frobenius norm property of the SVD we have

$$\|A - A_{s+m}\|_F^2 \leq \left\|A - (\widetilde{A} + (A - \widetilde{A})_m)\right\|_F^2.$$

Using Fact 10 (Pythagoras) and that $\|B\|_F = \left\|B^T\right\|_F$ for any matrix B we have

$$\left\|A - (\widetilde{A} + (A - \widetilde{A})_m)\right\|_F^2 = \left\|A - \widetilde{A}\right\|_F^2 - \left\|(A - \widetilde{A})_m\right\|_F^2.$$

Reordering and using that $\widetilde{A}$ satisfies equation (4.19) give

$$\left\|(A - \widetilde{A})_m\right\|_F^2 \leq \left\|A - \widetilde{A}\right\|_F^2 - \|A - A_{s+m}\|_F^2 \leq (1 + \varepsilon') \|A - A_s\|_F^2 - \|A - A_{s+m}\|_F^2. \tag{4.23}$$

Remember that $\|A\|_F^2 = \sum_{i=1}^{\operatorname{rank}(A)} \sigma_i^2(A)$. From the definition of A_{s+m} and A_s it follows that

$$\|A - A_{s+m}\|_F^2 = \sum_{i=s+m+1}^{\operatorname{rank}(A)} \sigma_i^2(A)$$

and

$$\|A - A_s\|_F^2 = \sum_{i=s+1}^{\operatorname{rank}(A)} \sigma_i^2(A),$$

thus expression (4.23) can be rearranged to

$$\left\|(A - \widetilde{A})_m\right\|_F^2 \leq \varepsilon' \|A - A_s\|_F^2 + \sum_{i=s+1}^{s+m} \sigma_i^2(A). \tag{4.24}$$

The singular values are positive and non-increasing, thus the last m terms of the sum $\sum_{i=m+1}^{s+m} \sigma_i^2(A)$ are the smallest in a sum with $s = \lceil \varepsilon'^{-1} m \rceil$ terms. Therefore

$$\sum_{i=m+1}^{s+m} \sigma_i^2(A) \geq \varepsilon'^{-1} \sum_{i=s+1}^{s+m} \sigma_i^2(A).$$

Applying this in expression (4.24) we have

$$\begin{aligned}\left\|(A - \widetilde{A})_m\right\|_F^2 &\leq \varepsilon' \|A - A_s\|_F^2 + \varepsilon' \sum_{i=m+1}^{s+m} \sigma_i^2(A) \\ &\leq \varepsilon' \|A - A_s\|_F^2 + \varepsilon' \sum_{i=m+1}^{\operatorname{rank}(A)} \sigma_i^2(A).\end{aligned}$$

Finally we use that $\|A - A_s\|_F \leq \|A - A_m\|_F$ since $\text{rank}(A_s) \geq \text{rank}(A_m)$ to obtain

$$\begin{aligned}\left\|(A - \widetilde{A})_m\right\|_F^2 &\leq 2\varepsilon' \|A - A_m\|_F^2 \\ &\leq 2\varepsilon' \left\|A - XX^T A\right\|_F^2 = 2\varepsilon' \left\|Y^T A\right\|_F^2 .\end{aligned}$$

And since

$$\begin{aligned}\left\|(A - \widetilde{A})_m\right\|_F^2 &= \sum_{i=1}^{m} \lambda_i(M) \\ &\geq \text{Tr}(XX^T M) = \text{Tr}(XX^T(AA^T - \widetilde{A}\widetilde{A}^T)) \\ &= \left\|X^T A\right\|_F^2 - \left\|X^T \widetilde{A}\right\|_F^2 = \left\|XX^T A - XX^T \widetilde{A}\right\|_F^2 ,\end{aligned}$$

we have finished the proof. ∎

The last result we need is related to the previous lemma.

Theorem 49. *Under the same hypothesis of Lemma 47 we have that*

$$\left| \left\|Y^T \widetilde{A}\right\|_F^2 + \|A - A_s\|_F^2 - \left\|Y^T A\right\|_F^2 \right| \leq 2\varepsilon' \left\|Y^T A\right\|_F^2 .$$

Proof. By Fact 10 (Pythagoras) we have

$$\|A - A_s\|_F^2 = \|A\|_F^2 - \|A_s\|_F^2$$

and

$$\left\|Y^T A\right\|_F^2 + \left\|X^T A\right\|_F^2 = \|A\|_F^2 ,$$

therefore

$$\begin{aligned}\left| \left\|Y^T \widetilde{A}\right\|_F^2 + \|A - A_s\|_F^2 - \left\|Y^T A\right\|_F^2 \right| &= \left| \left\|\widetilde{A}\right\|_F^2 - \left\|X^T \widetilde{A}\right\|_F^2 + \|A\|_F^2 - \|A_s\|_F^2 - \|A\|_F^2 + \left\|X^T A\right\|_F^2 \right| \\ &= \left| \left\|\widetilde{A}\right\|_F^2 - \|A_s\|_F^2 - \left\|X^T \widetilde{A}\right\|_F^2 + \left\|X^T A\right\|_F^2 \right| . \qquad (4.25)\end{aligned}$$

Because A_s is a best rank s approximation of A in the Frobenius norm, we have $\left\|\widetilde{A}\right\|_F^2 \leq \|A_s\|_F^2$, and thus

$$\left\|\widetilde{A}\right\|_F^2 - \|A_s\|_F^2 \leq 0. \qquad (4.26)$$

From the hypothesis, the matrix $\widetilde{A}$ satisfies expression (4.19), and by Pythagoras we can write

$$\|A\|_F^2 - \left\|\widetilde{A}\right\|_F^2 \leq \|A\|_F^2 - \|A_s\|_F^2 + \varepsilon' \|A - A_s\|_F^2 ,$$

which implies

$$-\varepsilon' \|A - A_s\|_F^2 \leq \left\|\widetilde{A}\right\|_F^2 - \|A_s\|_F^2 .$$

Using that A_s is a best rank $\lceil \varepsilon'^{-1} m \rceil \geq m$ approximation of A and XX^TA is of rank at most m, we obtain

$$-\varepsilon' \|A - A_s\|_F^2 \geq -\varepsilon' \left\|A - XX^TA\right\|_F^2 = -\varepsilon' \left\|Y^TA\right\|_F^2.$$

Combining this with inequality (4.20) of Lemma 47 we can conclude that the expression inside the modulus in (4.25) is at least $-\varepsilon' \left\|Y^TA\right\|_F^2$ and at most $2\varepsilon' \left\|Y^TA\right\|_F^2$, finishing the proof. ∎

We are now ready to prove Theorem 45.

Proof of Theorem 45. To prove that $\widetilde{A}$ is an ε-sketch of A, we will show that $\Delta = \|A - A_s\|_F^2$ is a constant with the property that for every $C \in \mathcal{C}$

$$\left|\operatorname{dist}^2(\widetilde{A}, C) + \Delta - \operatorname{dist}^2(A, C)\right| \leq \varepsilon \operatorname{dist}^2(A, C). \tag{4.27}$$

Let $C \in \mathcal{C}$ be any solution. Since $\mathcal{C}$ is a m-dimensional family, there exists a subspace L of dimension m such that $C \subset L$. Let $L^\perp$ be the orthogonal complement of L. Suppose that $X \in \mathbb{R}^{d \times m}$ has orthonormal columns that span L and $Y \in \mathbb{R}^{d \times (d-m)}$ has orthonormal columns that span $L^\perp$.

Since $C \subset L$, by Fact 10 (Pythagoras) we have

$$\operatorname{dist}^2(A, C) = \left\|Y^TA\right\|_F^2 + \operatorname{dist}^2(XX^TA, C). \tag{4.28}$$

Applying this to the left-hand side of inequality (4.27) give

$$\left|\operatorname{dist}^2(\widetilde{A}, C) + \|A - A_s\|_F^2 - \operatorname{dist}^2(A, C)\right| \tag{4.29}$$

$$= \left|\left\|Y^T\widetilde{A}\right\|_F^2 + \operatorname{dist}^2(XX^T\widetilde{A}, C) + \|A - A_s\|_F^2 - \left\|Y^TA\right\|_F^2 - \operatorname{dist}^2(XX^TA, C)\right| \tag{4.30}$$

$$\leq \underbrace{\left|\left\|Y^T\widetilde{A}\right\|_F^2 + \|A - A_s\|_F^2 - \left\|Y^TA\right\|_F^2\right|}_{\text{first term}} + \underbrace{\left|\operatorname{dist}^2(XX^T\widetilde{A}, C) - \operatorname{dist}^2(XX^TA, C)\right|}_{\text{second term}} \tag{4.31}$$

To bound the first term, we use Theorem 49 with $\varepsilon' = \varepsilon^2/8$, obtaining

$$\left|\left\|Y^T\widetilde{A}\right\|_F^2 + \|A - A_s\|_F^2 - \left\|Y^TA\right\|_F^2\right| \leq 2\frac{\varepsilon^2}{8}\left\|Y^TA\right\|_F^2. \tag{4.32}$$

To bound the second term, we start applying Lemma 46, the weak triangle inequality, in each column of XX^TA and $XX^T\widetilde{A}$, obtaining

$$\left|\operatorname{dist}^2(XX^T\widetilde{A}, C) - \operatorname{dist}^2(XX^TA, C)\right| \leq \varepsilon \operatorname{dist}^2(XX^TA, C) + \frac{2}{\varepsilon}\left\|XX^TA - XX^T\widetilde{A}\right\|_F^2. \tag{4.33}$$

Applying inequality (4.21) of Lemma 47 with $\varepsilon' = \varepsilon^2/8$, we obtain

$$\varepsilon \operatorname{dist}^2(XX^TA, C) + \frac{2}{\varepsilon}\left\|XX^TA - XX^T\widetilde{A}\right\|_F^2 \leq \varepsilon \operatorname{dist}^2(XX^TA, C) + \frac{2}{\varepsilon}\cdot 2\frac{\varepsilon^2}{8}\left\|Y^TA\right\|_F^2. \tag{4.34}$$

Combining inequalities (4.32) and (4.34), and using that for any $\varepsilon \in (0,1)$ the expression $\varepsilon^2/4 + \varepsilon/2$ is at most ε we get

$$\begin{aligned}\left|\operatorname{dist}^2(\widetilde{A}, C) + \|A - A_s\|_F^2 - \operatorname{dist}^2(A, C)\right| &\leq \left(\frac{\varepsilon^2}{4} + \frac{\varepsilon}{2}\right)\left\|Y^TA\right\|_F^2 + \varepsilon \operatorname{dist}^2(XX^TA, C)\\ &\leq \varepsilon\left\|Y^TA\right\|_F^2 + \varepsilon \operatorname{dist}^2(XX^TA, C) = \varepsilon \operatorname{dist}^2(A, C),\end{aligned}$$

finishing the proof. ∎

Remark: *In this chapter we have presented a random dimension reduction scheme for the (ℓ_2^2, C)-clustering problem that can be found in linear time in the dimension d and in the number of points n. An important aspect of the sketch we obtain is that it lies in a subspace of dimension that depends only on the dimensionality of the set C and on the parameter ε. Therefore some algorithms that have complexity exponential in d can be modified to become an $(1+\varepsilon)$-approximation with a polynomial dependence in d.*

Chapter 5

Application to projective clustering

In this chapter we will present an approximation scheme for the linear j-subspace k-clustering problem in $\mathbb{R}^d$ with polynomial time complexity for n, the number of points, but exponential in d. Then we will show an application of Algorithm 1 to obtain a polynomial time random approximation scheme (PRAS) that is polynomial for both n and d for this problem. This result is of theoretical application mostly, since the degree of the polynomial is rather large.

5.1 A $(1+\varepsilon)$-approximation for projective clustering

The work we present is based in a result of Deshpande, Rademacher, Vempala, and Wang (2006). The authors claim that they developed a polynomial time approximation scheme (PTAS) for the linear j-subspace k-clustering problem. Given an upper bound B for the cost of an optimal solution, their scheme returns in polynomial time in n and d a solution with cost at most $(1+\varepsilon)B$.

The main idea of their scheme is to consider a special finite subset $C' \subset C$ of solutions, and then compute the cost of every solution spanned by this set and report the one with lowest cost. This special finite set is called *δ-net*.

Definition 50 (δ-net). *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, and let $\delta > 0$ and $\mathcal{R} > 0$ be fixed. We say that a set $D \subset \mathbb{R}^d$ is a* δ-net *with radius $\mathcal{R}$ for A if for every $x \in \mathbb{R}^d$ such that* $\operatorname{dist}(x, A) \leq \mathcal{R}$ *there exists an $y \in D$ such that $\|x - y\| \leq \delta$.*

To explicitly build a δ-net of radius $\mathcal{R}$ for a finite set $A \subset \mathbb{R}^d$ with $|A| = n$, we could start with a $(\delta\mathcal{R}^{-1}d^{-1/2})$-fine grid H of dimension d. The scaled up set

$$H' = \{\mathcal{R}h \,:\, \forall h \in H\}$$

will be a δ-net for the origin. By putting copies of this set around each point of A we obtain

the δ-net. More precisely, the set

$$A + H' = \left\{ a + h' \ : \ \forall a \in A, \forall h' \in H' \right\}$$

will be a δ-net for A with cardinality

$$\left| A + H' \right| = O\left(n \left(\frac{3\mathcal{R}\sqrt{d}}{\delta} \right)^d \right).$$

A proof that H' is a δ-net for the origin can be found in Fact 61 at Appendix A. This construction has the property that $A \subset D$, which is necessary for the next algorithm.

Algorithm 2: An approximation for the linear j-subspace k-clustering

Input: A matrix $A \in \mathbb{R}^{d \times n}$, a real $\varepsilon \in (0, 1)$ and a real $B > 0$.
Output: A set $F_1, \ldots, F_k$ of k subspaces of dimension j.

1 Set

$$\delta \ := \frac{\varepsilon \sqrt{B}}{8jk\sqrt{n}};$$

2 Set

$$\mathcal{R} \ := \sqrt{B} + 2\delta j;$$

3 Let D be a δ-net with radius $\mathcal{R}$ for A that contains A;
4 For each choice of k subspaces $F_1, \ldots, F_k$ of dimension j, each one spanned by j points of D, compute $\text{dist}^2(A, \bigcup_{i=1}^{k} F_i)$;
5 Return the subspaces $F_1, \ldots, F_k$ with lowest cost;

This algorithm is just a brute-force. We are searching the optimal set of subspaces in a finite family (the subspaces spanned by the points of the δ-net D), and we do this by computing the cost of all of them and returning one with minimum cost. Bounding the time complexity is straightforward.

The number of solutions spanned by D is at most

$$\binom{\binom{|D|}{j}}{k} \leq (|D|)^{jk}.$$

This is because to span a subspace of dimension j we must choose a set with at least j points from D, and to obtain one solution we must choose k of these sets at a time. Since it is possible to build a δ-net of radius $\mathcal{R}$ for A with cardinality

$$O\left(n \left(\frac{3\mathcal{R}\sqrt{d}}{\delta} \right)^d \right),$$

the number of solutions the algorithm must test is at most

$$O\left(n^{jk}\left(\frac{3\mathcal{R}\sqrt{d}}{\delta}\right)^{jkd}\right).$$

Finding an orthonormal basis of an subspace takes time $O(j^2d)$, if we use Gram-Schmidt. Computing the cost of an solution takes time $O(jknd + j^2kd)$, since we need to find an orthonormal basis for every subspace of the solution and we need to project every point of A into all of the k subspaces. Doing this for every solution takes time

$$O\left((jknd + j^2kd)n^{jk}\left(\frac{3\mathcal{R}\sqrt{d}}{\delta}\right)^{jkd}\right).$$

Substituting the values of δ and $\mathcal{R}$, we find that

$$\frac{\mathcal{R}}{\delta} \leq \frac{10jk\sqrt{n}}{\varepsilon}.$$

The final bound is

$$O\left((jknd + j^2kd)n^{jk}\left(\frac{30jk\sqrt{nd}}{\varepsilon}\right)^{jkd}\right) = O\left(\left(\frac{jknd}{\varepsilon}\right)^{3jkd}\right).$$

The following theorem guarantee that the solution returned by Algorithm 2 is correct.

Theorem 51 (Deshpande, Rademacher, Vempala, and Wang, 2006)**.** *Let (A, ε, B) be an input for Algorithm 2. Suppose that the cost of an optimal solution for the instance A is B^*. If $B \geq B^*$, then the solution given by Algorithm 2 has cost at most $B^* + \varepsilon B$.*

This theorem implies that for $B \geq B^*$ Algorithm 2 is a $(1 + \varepsilon(B/B^*))$-approximation for the linear j-subspace k-clustering problem. To prove Theorem 51 we will show that for every subspace in an optimal solution a δ-net with "sufficient" radius will contain points that spans a subspace with cost similar to the optimal. This notion of "sufficient" radius is given by the following lemma.

Lemma 52 (Deshpande, Rademacher, Vempala, and Wang, 2006)**.** *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and let $\delta > 0$ be fixed. Suppose that D is a δ-net for A with radius $\mathcal{R}$. For every subspace $W \subset \mathbb{R}^d$ of dimension j, if*

$$\mathcal{R} \geq \sqrt{\operatorname{dist}^2(A, W)} + 2\delta j$$

then there exists a subspace F of dimension j of $\mathbb{R}^d$ spanned by j points of D such that

$$\operatorname{dist}^2(A, F) \leq \operatorname{dist}^2(A, W) + 4j^2n\delta^2 + 4j\delta\sum_{a\in A}\operatorname{dist}(a, W). \tag{5.1}$$

To prove this lemma we first need to define a special linear function called *rotation*.

Let u and v be non-null vectors of $\mathbb{R}^d$. Suppose that u and v are linearly independent. Let $u' := u/\|u\|$ and $v' := v/\|v\|$. Let P be the subspace spanned by u and v and let $P^\perp$ its orthogonal complement. We say that P is the *plane of rotation* defined by u and v. Let $B \in \mathbb{R}^{d\times d-2}$ be a matrix with orthonormal columns that spans $P^\perp$. There exists norm one vectors u^* and v^* such that B_u and B_v are $d \times d$ matrices that agrees with B in the first $d-2$ columns and B_u has u' and u^* respectively as the last two columns and B_v has v' and v^* respectively as the last two columns. Its straightforward to show that if we add the restriction $\det(B_u) = \det(B_v) = 1$ then u^* and v^* are uniquely determined, thus we will impose that $\det(B_u) = \det(B_v) = 1$.

Any linear function can be defined by its application on a basis of the space. Thus let $f : \mathbb{R}^d \to \mathbb{R}^d$ be such that $f(b) = b$ for every $b \in B$, and $f(u') = v'$ and $f(u^*) = v^*$. Since f is linear, there exists a matrix $R \in \mathbb{R}^{d\times d}$ such that for every $x \in \mathbb{R}^d$ we have $f(x) = Rx$. We say that the matrix R is an (u, v)-*rotation matrix*. A nice property of this matrix is that it is unitary, i.e., for every vector x we have $\|x\| = \|Rx\|$. Also, since $\det(B_u) = \det(B_v)$, we have $\det(R) = 1$. This matrix R is usually known as *proper rotation matrix*. More details of this type of matrix can be found in Horn and Johnson, 2013, Chap. 2.

Note that we don't necessarily need u and v to be linearly independent. In case they are linearly dependent we will define R as the $d \times d$ identity matrix.

Fact 53. *For any $x \in \mathbb{R}^d$ we have*

$$\|x - Rx\| = \|\pi_P(x) - R\pi_P(x)\|.$$

This fact follows directly from the definition of R.

Fact 54. *For any $x \in \mathbb{R}^d$ and non-negative real δ such that* $\operatorname{dist}(x, P^\perp) \leq \delta$ *we have*

$$\|x - Rx\| \leq 2\delta.$$

Proof. If R is the identity matrix the fact is trivial. Suppose that u and v are linearly independent. Since B_u is a base of $\mathbb{R}^d$, there exists reals $\alpha, \beta, \gamma_1, \ldots, \gamma_{d-2}$ such that

$$x = \alpha u' + \beta u^* + \sum_{i=1}^{d-2} \gamma_i b_i.$$

This means that we can write the projection of x as $\pi_P(x) = \alpha u' + \beta u^*$ and $\pi_{P^\perp}(x) = \sum_{i=1}^{d-2} \gamma_i b_i$. Using this with the triangle inequality we get

$$\|x - Rx\| \leq \|x - \pi_{P^\perp}(x)\| + \|\pi_{P^\perp}(x) - Rx\| \leq \delta + \left\| \sum_{i=1}^{d-2} \gamma_i b_i - \alpha v' - \beta v^* - \sum_{i=1}^{d-2} \gamma_i b_i \right\|,$$

which can be simplified to

$$\delta + \|R\pi_P(x)\|.$$

Since R is unitary we have $\|R\pi_P(x)\| = \|\pi_P(x)\|$. Using that $\text{dist}(x, P^\perp) = \|\pi_P(x)\|$ we obtain $\|x - Rx\| \leq 2\delta$. ∎

With this tool we can now prove Lemma 52.

Proof. The subspace F will be constructed inductively. We start with $F_0 = W$ and find subspaces $F_1, \ldots, F_j$ such that F_i is the result of an application of an (u_i, v_i)-rotation to F_{i-1}, for some u_i and v_i, and includes a new point from the δ-net.

It is sufficient to show that every subspace have the property that for every point $a \in A$ we have

$$\text{dist}(a, F_i) \leq \text{dist}(a, F_{i-1}) + 2\delta, \tag{5.2}$$

as this implies $\text{dist}(a, F) \leq \text{dist}(a, W) + 2j\delta$. Taking the square of both sides of the inequality and adding for all $a \in A$ give us inequality (5.1).

We will begin with the base case. Suppose that $i = 1$. If F_0 contains at least one non-null point from D, we are done. If not, let $a_1^* = \text{argmax}\left\{\|\pi_{F_0}(a)\| : a \in A\right\}$ and let $g_1 = \text{argmin}\left\{\|\pi_{F_0}(a_1^*) - g\| : g \in D\right\}$. Suppose that $g_1 \neq \vec{0}$ and $\pi_{F_0}(a_1^*) \neq \vec{0}$. We will deal with the null case later. We can define R_1 as being the $(\pi_{F_0}(a_1^*), g_1)$-rotation matrix. Let $F_1 = R_1 F_0$. To prove the lemma for $i = 1$ is enough to show that for every $a \in A$ we have

$$\|\pi_{F_0}(a) - R_1\pi_{F_0}(a)\| \leq 2\delta, \tag{5.3}$$

since we can deduce the following bound:

$$\text{dist}(a, F_1) = \|a - \pi_{F_1}(a)\| \leq \|a - R_1\pi_{F_0}(a)\|,$$

as $\pi_{F_1}(a)$ is the closest point from F_1 to a and $R_1\pi_{F_0}(a) \in F_1$. Applying the triangle inequality we obtain

$$\begin{aligned}\|a - R_1\pi_{F_0}(a)\| &\leq \|a - \pi_{F_0}(a)\| + \|\pi_{F_0}(a) - R_1\pi_{F_0}(a)\| \\ &= \text{dist}(a, F_0) + \|\pi_{F_0}(a) - R_1\pi_{F_0}(a)\|.\end{aligned}$$

To prove (5.3) we observe that $\|\pi_{F_0}(a) - R_1\pi_{F_0}(a)\|$ is the distance between a point and its rotation, thus it is maximized when $\|\pi_{F_0}(a)\|$ is maximized, and by construction this happens when $a = a_1^*$. Now we apply the triangle inequality to

$$\|\pi_{F_0}(a_1^*) - R_1\pi_{F_0}(a_1^*)\| \leq \|\pi_{F_0}(a_1^*) - g_1\| + \|g_1 - R_1\pi_{F_0}(a_1^*)\|.$$

The rotation applied to $\pi_{F_0}(a_1^*)/\|\pi_{F_0}(a_1^*)\|$ is $g_1/\|g_1\|$, by the definition of $(\pi_W(a_1^*), g_1)$-rotation, therefore g_1 is equal to $R_1\pi_{F_0}(a_1^*)$ scaled by a positive factor, hence

$$\|\pi_{F_0}(a_1^*) - g_1\| + \|g_1 - R_1\pi_{F_0}(a_1^*)\| = \|\pi_{F_0}(a_1^*) - g_1\| + \big|\,\|g_1\| - \|R_1\pi_{F_0}(a_1^*)\|\,\big|.$$

Now we use that R_1 is unitary and another triangle inequality to obtain

$$\begin{aligned}\|\pi_{F_0}(a_1^*) - g_1\| + \big|\|g_1\| - \|R_1\pi_{F_0}(a_1^*)\|\big| &= \|\pi_{F_0}(a_1^*) - g_1\| + \big|\|g_1\| - \|\pi_{F_0}(a_1^*)\|\big| \\ &\le \|\pi_{F_0}(a_1^*) - g_1\| + \|g_1 - \pi_{F_0}(a_1^*)\| = 2\|\pi_{F_0}(a_1^*) - g_1\|.\end{aligned}$$

By the choice of $\mathcal{R}$, we have that the distance between F_0 and a_1^* is at most $\mathcal{R}$. Since D is a δ-net, the distance $\|\pi_{F_0}(a_1^*) - g_1\|$ is at most δ. Therefore for every $a \in A$ we have

$$\|\pi_{F_0}(a) - R_1\pi_{F_0}(a)\| \le 2\delta.$$

The case where $\pi_{F_0}(a_1^*) = \vec{0}$ implies that for every $a \in A$ we have $\pi_{F_0}(a) = \vec{0}$, therefore any rotation R_1 such that R_1F_0 contains some non-null $g \in D$ will work, since $\|\pi_{F_0}(a) - R_1\pi_{F_0}(a)\| = 0 \le 2\delta$.

The case where $g_1 = \vec{0}$ implies that for every $a \in A$ the value $\|\pi_{F_0}(a)\|$ is at most δ. For any plane of rotation P, since $\vec{0} \in P^\perp$ we have $\operatorname{dist}(\pi_{F_0}(a), P^\perp) \le \delta$. Therefore we can let g_1 be any non-null $g \in D$ since from Fact 54 any $(\pi_{F_0}(a_1^*), g_1)$-rotation matrix R_1 will have the property

$$\|a - R_1\pi_{F_0}(a)\| \le 2\delta.$$

Now we will prove the induction. Suppose that $1 < i \le j$, let $\mathcal{G}_1 := \{g_1\}$ and for all $1 < k < i$ let $\mathcal{G}_k := \mathcal{G}_{k-1} \cup \{g_k\}$. Define $G_i := \operatorname{span}(\mathcal{G}_i)$. We want $G_i \subset F_i$, thus we will make each rotation R_i such that its plane of rotation is orthogonal to G_{i-1}.

Let $a_i^* := \operatorname{argmax}\left\{\left\|\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a))\right\| : a \in A\right\}$, and let $g_i := \operatorname{argmin}\left\{\left\|\pi_{F_{i-1}}(a_i^*) - g\right\| : g \in D\right\}$. We will show that $\left\|\pi_{F_{i-1}}(a_i^*) - g_i\right\| \le \delta$ by proving that $\left\|a_i^* - \pi_{F_{i-1}}(a_i^*)\right\| \le \mathcal{R}$. This implies that such $g_i \in D$ exist by the δ-net definition.

By applying the triangle inequality and that the closest vector to a_i^* in a subspace is the orthogonal projection multiple times we obtain

$$\begin{aligned}\left\|a_i^* - \pi_{F_{i-1}}(a_i^*)\right\| &\le \left\|a_i^* - R_{i-1}\pi_{F_{i-2}}(a_i^*)\right\| \\ &\le \left\|a_i^* - \pi_{F_{i-2}}(a_i^*)\right\| + \left\|\pi_{F_{i-2}}(a_i^*) - R_{i-1}\pi_{F_{i-2}}(a_i^*)\right\| \\ &\le \left\|a_i^* - \pi_{F_0}(a_i^*)\right\| + \sum_{k=0}^{i-2}\left\|\pi_{F_k}(a_i^*) - R_{k+1}\pi_{F_k}(a_i^*)\right\|.\end{aligned}$$

Note that

$$\left\|a_i^* - \pi_{F_0}(a_i^*)\right\| \le \sqrt{\sum_{a\in A}(\operatorname{dist}(a, W))^2} = \sqrt{\operatorname{dist}^2(A, W)},$$

and by the induction hypothesis for every $k = 0, \dots, i-2$ and for every $a \in A$ we have

$$\left\|\pi_{F_k}(a) - R_{k+1}\pi_{F_k}(a)\right\| \le 2\delta,$$

therefore

$$\left\|a_i^* - \pi_{F_{i-1}}(a_i^*)\right\| \le \sqrt{\operatorname{dist}^2(A, W)} + 2j\delta \le \mathcal{R}.$$

We have proved that $\left\|\pi_{F_{i-1}}(a_i^*) - g_i\right\| \leq \delta$. Now the natural step is to define R_i as an $(\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*)), \pi_{G_{i-1}^\perp}(g_i))$-rotation, but this rotation may not be well defined. We must consider three cases:

1. When $\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*)) = \vec{0}$;
2. When $\pi_{G_{i-1}^\perp}(g_i) = \vec{0}$;
3. When R_i is well defined as an $(\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*)), \pi_{G_{i-1}^\perp}(g_i))$-rotation;

In the first two cases, let $g^* \in D$ be such that $\pi_{G_{i-1}^\perp}(g^*) \neq \vec{0}$ and let $x \in F_{i-1}$ be such that $\pi_{G_{i-1}^\perp}(x) \neq \vec{0}$. We define R_i as an $(\pi_{G_{i-1}^\perp}(x), \pi_{G_{i-1}^\perp}(g^*))$-rotation. This rotation guarantees that $F_i = R_i F_{i-1}$ contains $\mathcal{G}_{i-1} \cup g^*$.

For case 1 note that it implies $\pi_{F_{i-1}}(a) \in G_{i-1}$ for every $a \in A$. Therefore $\pi_{F_{i-1}}(a) = \pi_{F_i}(a)$ and $\left\|\pi_{F_{i-1}}(a) - R_i\pi_{F_{i-1}}(a)\right\| = 0 \leq 2\delta$.

For case 2 note that it implies $g_i \in G_{i-1}$, and thus $\operatorname{dist}(\pi_{F_{i-1}}(a), G_{i-1}) \leq \delta$ for every $a \in A$. By Fact 54 we obtain $\left\|\pi_{F_{i-1}}(a) - R_i\pi_{F_{i-1}}(a)\right\| \leq 2\delta$.

For the last case we can bound (5.2) similarly to the base case. It is enough to bound

$$\left\|\pi_{F_{i-1}}(a) - R_i\pi_{F_{i-1}}(a)\right\|$$

as we can use the triangle inequality to achieve (5.1). Also, since the rotation does not change any vector in the subspace G_{i-1} then

$$\left\|\pi_{F_{i-1}}(a) - R_i\pi_{F_{i-1}}(a)\right\| = \left\|\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a)) - R_i\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a))\right\|,$$

The distance between a vector and its rotation is maximized when its norm is maximized, thus for every $a \in A$ we have

$$\left\|\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a)) - R_i\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a))\right\| \leq \left\|\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*)) - R_i\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*))\right\|.$$

By the triangle inequality we have

$$\left\|\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*)) - R_i\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*))\right\| \leq \left\|\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*)) - \pi_{G_{i-1}^\perp}(g_i)\right\| + \left\|\pi_{G_{i-1}^\perp}(g_i) - R_i\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*))\right\|. \tag{5.4}$$

The vector $\pi_{G_{i-1}^\perp}(g_i)$ is equal to $R_i\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*))$ scaled by a positive factor, thus the right-hand side of (5.4) is equal to

$$\left\|\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*)) - \pi_{G_{i-1}^\perp}(g_i)\right\| + \left|\left\|\pi_{G_{i-1}^\perp}(g_i)\right\| - \left\|R_i\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*))\right\|\right| \tag{5.5}$$

and since R_i is unitary, and by the triangle inequality we obtain that (5.5) is at most

$$\begin{aligned}\left\|\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*)) - \pi_{G_{i-1}^\perp}(g_i)\right\| + \left\|\pi_{G_{i-1}^\perp}(g_i) - \pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*))\right\| &= 2\left\|\pi_{G_{i-1}^\perp}(\pi_{F_{i-1}}(a_i^*)) - \pi_{G_{i-1}^\perp}(g_i)\right\| \\ &\leq 2\left\|\pi_{F_{i-1}}(a_i^*) - g_i\right\| \\ &\leq 2\delta,\end{aligned}$$

since the distance between the projection of two points is at most their original distance and $\operatorname{dist}(\pi_{F_{i-1}(a_i^*)}, A) \leq \mathcal{R}$. ∎

Now we can use this lemma to finish the proof of the correctness of Algorithm 2 by proving Theorem 51.

Proof of Theorem 51. Let $C^* = \{W_0, \ldots, W_k\}$ be an optimal solution with cost equals to B^*. Suppose that $B^* \leq B$. Consider a partition of $A = S_1 \cup \cdots \cup S_k$ with the property that

$$S_i = \left\{ a \in A \,:\, \forall\, i' \neq i, \operatorname{dist}(a, W_i) \leq \operatorname{dist}(a, W_{i'}) \right\}.$$

Notice that

$$\sum_{i=1}^{k} \operatorname{dist}^2(S_i, W_i) = \operatorname{dist}^2(A, C^*) \leq B$$

and that the δ-net D of radius $\mathcal{R}$ for A is also a δ-net of radius $\mathcal{R}$ for S_i, for every $i = 1, \ldots, k$. By our choice of $\mathcal{R}$, we have that for every $i = 1, \ldots, k$

$$\begin{aligned} \sqrt{\operatorname{dist}^2(S_i, W_i)} &\leq \sqrt{\operatorname{dist}^2(A, C^*)} \\ &\leq \sqrt{B} + 2\delta j \leq \mathcal{R}. \end{aligned}$$

Therefore by Lemma 52 there exists a subspace F_i spanned by j points of D such that

$$\operatorname{dist}^2(S_i, F_i) \leq \operatorname{dist}^2(S_i, W_i) + 4j^2\delta^2 n + 4j\delta \sum_{a \in S_i} \operatorname{dist}(a, W_i). \tag{5.6}$$

We can bound the last sum of the right-hand side of (5.6) with Cauchy-Swartz inequality to obtain

$$\sum_{a \in S_i} \operatorname{dist}(a, W_i) \leq \sqrt{n}\sqrt{\operatorname{dist}^2(S_i, W_i)}.$$

This together with our choice of $\delta = \varepsilon\sqrt{B}/(8jk\sqrt{n})$ give

$$\begin{aligned} \operatorname{dist}^2(S_i, F_i) &\leq \operatorname{dist}^2(S_i, W_i) + \frac{4\varepsilon^2 B}{8^2 k^2} + \frac{4\varepsilon\sqrt{\operatorname{dist}^2(S_i, W_i)}\sqrt{B}}{8k} \\ &\leq \operatorname{dist}^2(S_i, W_i) + \frac{\varepsilon B}{2k} + \frac{\varepsilon B}{2k} \\ &\leq \operatorname{dist}^2(S_i, W_i) + \frac{\varepsilon B}{k}. \end{aligned}$$

Let $\widetilde{C} = \{F_1, \ldots, F_k\}$. The cost of $\widetilde{C}$ can be bounded as

$$\begin{aligned}\operatorname{dist}^2(A,\widetilde{C}) &\leq \sum_{i=1}^{k} \operatorname{dist}^2(S_i,F_i)\\ &\leq \sum_{i=1}^{k} \operatorname{dist}^2(S_i,W_i) + \frac{\varepsilon B}{2k}\\ &= \operatorname{dist}^2(A,C^*) + \varepsilon B = B^* + \varepsilon B.\end{aligned}$$

Algorithm 2 will enumerate all choices of subspaces spanned by D. At some point it will hit $F_1,\dots,F_k$, and compute its cost. Therefore the solution returned will have cost at most $\operatorname{dist}^2(A,\widetilde{C}) \leq B^* + \varepsilon B$. ∎

The approximation we obtain from Algorithm 2 depends on the parameter B, and it is not a $(1+\varepsilon)$-approximation. To be more precise, this algorithm give a $(1+\varepsilon(B/B^*))$-approximation, since $B^* + \varepsilon B = (1+\varepsilon(B/B^*))B^*$.

To obtain a $(1+\varepsilon)$-approximation, we will need to study some properties of this algorithm to improve it. The first property of Algorithm 2 we note is that if the δ-net D contains the instance A, and the cost of an optimal solution is zero, then it will return an optimal solution. This is clear since if the cost of a solution is zero, it must be optimal and it must be that A is contained in it. Therefore the algorithm will enumerates the elements of A and test the subspace that contains $\operatorname{span}(A)$, obtaining the solution of cost zero.

The second property is that Algorithm 2 always returns a viable solution for ever parameter $\varepsilon \in (0,1)$ and $B > 0$, since for every $\delta > 0$ and $\mathcal{R} > 0$ we can build a δ-net of radius $\mathcal{R}$. The algorithm will then enumerate the solutions spanned by it and return the one with minimum cost. This will be useful for identifying lower bounds of the cost of an optimal solution.

Corollary 55. *Suppose that $A \in \mathbb{R}^{d\times n}$ is an instance of linear j-subspace k-clustering problem, the cost of an optimal solution is B^*, and $\varepsilon \in (0,1)$ is fixed. Suppose also that $B > 0$ is fixed, and that the cost of the solution returned by Algorithm 2 with input (A,ε,B) is $\widetilde{B}$. If*

$$\widetilde{B} > (1+\varepsilon)B,$$

then $B < B^*$. When this happens we say that B is a certified lower bound *for the pair (A,ε).*

Proof. By Theorem 51, if $B \geq B^*$, then the cost of the returned solution $\widetilde{B}$ will be at most $B^* + \varepsilon B \leq (1+\varepsilon)B$. Since $\widetilde{B} > (1+\varepsilon)B$, we must have $B < B^*$. ∎

Corollary 55 allows us to know when the cost of the solution we tried was too low. If we run Algorithm 2 with input (A,ε,B) and $(1+\varepsilon)B < B^*$, then B will be a certified lower bound for (A,ε) since we cannot have a solution with cost lower than B^*. This can be used as a parameter to evaluate B; if the cost of the solution returned is higher than $(1+\varepsilon)B$, then B is a certified lower bound for (A,ε). Else B is an upper bound for B^* and we can keep the solution returned. Designing a binary search algorithm to find a $(1+\varepsilon)$-approximation is now straightforward, as we show in Algorithm 3.

Algorithm 3: A $(1+\varepsilon)$-approximation for linear j-subspace k-clustering

Input: A matrix $A \in \mathbb{R}^{d \times n}$, a parameter $\varepsilon \in (0,1)$.

Output: A set $C := \{F_1, \ldots, F_k\}$ of k subspaces of dimension j.

Set B as any upper bound for the cost of an optimal solution ;

/* A valid initialization value for B would be $\|A\|_F^2$ */

Set C as the solution returned by Algorithm 2 with input $(A, \varepsilon/3, B)$;

Set $\beta := \text{dist}^2(A, C)$;

if $\beta = 0$ **then**

    Return C;

/* Start the loop to find a lower bound for the cost B^* of an optimal solution */

Set $\widetilde{C} := C$;

Set $B_0 := B$;

Set $i := 0$;

while $\beta \leq (1+\varepsilon/3)B_i$ **do**

    $i := i + 1$;

    $C := \widetilde{C}$; // Note that $\text{dist}^2(A, C) \leq (1+\varepsilon/3)B_{i-1}$

    $B_i := B_{i-1}/2$;

    Update $\widetilde{C}$ with the solution returned by Algorithm 2 with input $(A, \varepsilon/3, B_i)$;

    Update β with the cost of $\widetilde{C}$;

/* Note that B_i is a lower bound for B^*. Now we start the binary search. */

Set $lo := B_i$;

Set $hi := B_{i-1}$;

Set $mid := (hi + lo)/2$;

while $hi - lo > (\varepsilon/3)B_i$ **do**

    Update $\widetilde{C}$ with the solution returned by Algorithm 2 with input $(A, \varepsilon/3, mid)$;

    Update β with $\text{dist}^2(A, \widetilde{C})$;

    if $\beta \leq (1+\varepsilon/3)mid$ **then**

        $hi := mid$;

        $C := \widetilde{C}$; // Note that $\text{dist}^2(A, C) \leq (1+\varepsilon/3)hi$

    else

        Update lo with mid;

    $mid := (hi + lo)/2$;

Return C;

To obtain an upper bound for the time complexity of Algorithm 3 we will first bound the number of executions of the loops on lines 9 and 18, since these equals the number of times Algorithm 2 is executed.

Fact 56. *The while-loop on line 9 is executed at most* $\log_2(B/B^*) + 2$ *times.*

Proof. Each time this loop is executed, the value we are guessing for the lower bound of B^* is halved. Let i be the number of executions of the loop such that $(1+\varepsilon)B_i < B^* \leq (1+\varepsilon)B_{i-1}$. Then it is guaranteed that the condition of the while-loop on line 9 will have already been violated at step i. Using that $\varepsilon \in (0,1)$, we can conclude that

$$B^* \leq (1+\varepsilon)B_{i-1} < 2B_{i-1} = B/2^{i-2},$$

therefore

$$\log_2\left(\frac{B}{B^*}\right) \geq i-2$$

and thus the number of executions of the loop will be at most $\log_2(B/B^*) + 2$. ∎

Fact 57. *The while-loop on line 18 is executed at most* $\log_2(3/\varepsilon) + 1$ *times.*

Proof. The binary search starts with an interval of length $B_{i-1} - B_i$, and ends when the interval has length at most $(\varepsilon/3)B_i$. Also note that by definition $B_i = B_{i-1}/2$, hence $B_{i-1} - B_i = B_i$. At each step the search interval is halved, therefore if after exactly s steps the search interval shrink to at most $(\varepsilon/3)B_i$ we have

$$\frac{B_{i-1} - B_i}{2^s} = \frac{B_i}{2^s} \leq (\varepsilon/3)B_i \leq \frac{B_i}{2^{s-1}}.$$

From the right-hand side inequality we obtain

$$s - 1 \leq \log_2\left(\frac{3B_i}{\varepsilon B_i}\right) = \log_2\left(\frac{3}{\varepsilon}\right).$$

Rearranging we obtain the desired bound. ∎

The time complexity of Algorithm 3 follows directly from Fact 56 and 57:

$$O\left(\left(\frac{jknd}{\varepsilon}\right)^{3jkd} \log\left(\frac{B}{\varepsilon B^*}\right)\right).$$

This value depends on the relationship between B and B^*. In line 1 if the value B is initialized such that B/B^* is at most an exponential in n, we would have a polynomial time complexity in n. It remains to show that the solution returned by Algorithm 3 is correct.

Lemma 58. *The solution C returned by Algorithm 3 with input (A, ε) is a $(1+\varepsilon)$-approximation for A.*

Proof. Let B^* be the cost of an optimal solution. Algorithm 3 can return a solution in two places. If the returns occurs in line 5, then the solution must be optimal, since it has cost zero, and therefore it trivially is a $(1+\varepsilon)$-approximation. Else, the algorithm continues past line 5. Before the execution of line 9 we have the following relation between C and B_0.

$$\text{dist}^2(A, C) \leq (1+\varepsilon/3)B_0.$$

This follows from Theorem 51, since B is an upper bound of the value B^*. After the i^{th} execution of the while-loop on line 9, the same invariant is true.

$$\text{dist}^2(A, C) \leq (1+\varepsilon/3)B_{i-1}.$$

This follows from the fact that C is updated with $\widetilde{C}$ only when in the previous iteration $\text{dist}^2(A, \widetilde{C}) \leq (1+\varepsilon/3)B_i$ (the value of i is incremented at the start of the loop, thus B_{i-1} is equal to B_i of the previous iteration).

Before the execution of the while-loop on line 18, by Corollary 55 the value lo is a certified lower bound for $(A, \varepsilon/3)$, and by the above argument, we have $\text{dist}^2(A, C) \leq (1+\varepsilon/3)hi$. This will be the invariant during the execution of this loop.

During each execution of the while-loop on line 18, the solution C will be updated with $\widetilde{C}$ only when $\text{dist}^2(A, \widetilde{C}) \leq (1+\varepsilon/3)mid$ and hi is updated with the value mid. Also, the variable lo will be updated with mid only when mid is a certified lower bound for $(A, \varepsilon/3)$.

After the execution of the while-loop on line 18, the algorithm returns the solution C. From the above argument, we have the following inequalities:

$$hi - lo \leq (\varepsilon/3)lo, \tag{5.7}$$
$$\text{dist}^2(A, C) \leq (1+\varepsilon/3)hi, \tag{5.8}$$
$$lo < B^*. \tag{5.9}$$

From inequality (5.7), we have

$$hi \leq (1+\varepsilon/3)lo. \tag{5.10}$$

Applying (5.10) in (5.8) and from the fact that $\varepsilon \in (0, 1)$

$$\text{dist}^2(A, C) \leq (1+\varepsilon/3)^2 lo \leq (1+\varepsilon)lo. \tag{5.11}$$

Applying inequality (5.9) in (5.11) we conclude that $\text{dist}^2(A, C) \leq (1+\varepsilon)B^*$. ∎

Remark: *Even though we have obtained a $(1+\varepsilon)$-approximation scheme for the linear j-subspace k-clustering problem, we cannot claim we have obtained a PTAS, since the time complexity depends on the term $\log(B/\varepsilon B^*)$. This dependence is necessary because our chosen model of computation allows the representation of any real using $O(1)$ of memory space. This allows the existence of a family of instances with constant size that has at least one solution with constant cost but optimal solution arbitrarily small. Therefore we could*

have values of B/B^* arbitrarily large, and Algorithm 3 would not be polynomial in the size of the input. See the appendix B for an example of such a family.

5.2 Faster approximation using ε-sketches

The algorithm we saw finds an approximation for the linear j-subspace k-clustering problem, but its time complexity has an exponential dependence in the dimension of the space. By using our developed low rank ε-sketch, we can mitigate the effect of the dimension at the cost of using a randomized algorithm. The following fact shows how we can do that.

Fact 59. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and let $\varepsilon \in (0, 1/3)$ be fixed. Suppose that $\widetilde{A} \in \mathbb{R}^{d\times n}$ is an ε-sketch for A for the linear j-subspace k-clustering problem and $\alpha > 1$ is a fixed real. If an algorithm $\mathcal{A}$ is an α-approximation for the linear j-subspace k-clustering problem, then the output of $\mathcal{A}(\widetilde{A})$ will be an $\alpha(1 + 3\varepsilon)$-approximation for A.*

Proof. Suppose that C is the output of $\mathcal{A}(A)$, that $\widetilde{C}$ is the output of $\mathcal{A}(\widetilde{A})$, that C^* is an optimal solution for A and $\widetilde{C}^*$ is an optimal solution for $\widetilde{A}$. Since $\mathcal{A}$ outputs an α-approximation for every matrix, we have

$$\operatorname{dist}^2(A, C) \leq \alpha \operatorname{dist}^2(A, C^*).$$

From the ε-sketch property, we have that there exists a non-negative constant Δ such that

$$\operatorname{dist}^2(\widetilde{A}, C^*) + \Delta \leq (1 + \varepsilon) \operatorname{dist}^2(A, C^*) \tag{5.12}$$

and

$$(1 - \varepsilon) \operatorname{dist}^2(A, \widetilde{C}) \leq \operatorname{dist}^2(\widetilde{A}, \widetilde{C}) + \Delta. \tag{5.13}$$

Also since $\widetilde{C}$ is an α-approximation for $\widetilde{A}$, we have that

$$\operatorname{dist}^2(\widetilde{A}, \widetilde{C}) \leq \alpha \operatorname{dist}^2(\widetilde{A}, \widetilde{C}^*) \leq \alpha \operatorname{dist}^2(\widetilde{A}, C^*). \tag{5.14}$$

Therefore from inequalities (5.12) and (5.14), and from the fact that Δ is non-negative we can deduce that

$$\operatorname{dist}^2(\widetilde{A}, \widetilde{C}) + \Delta \ \leq \alpha(\operatorname{dist}^2(\widetilde{A}, C^*) + \Delta) \leq \alpha(1 + \varepsilon) \operatorname{dist}^2(A, C^*).$$

From inequality (5.13) it follows that

$$(1 - \varepsilon) \operatorname{dist}^2(A, \widetilde{C}) \leq \operatorname{dist}^2(\widetilde{A}, \widetilde{C}) + \Delta \leq \alpha(1 + \varepsilon) \operatorname{dist}^2(A, C^*),$$

which implies

$$\operatorname{dist}^2(A, \widetilde{C}) \leq \alpha \frac{(1 + \varepsilon)}{(1 - \varepsilon)} \operatorname{dist}^2(A, C^*).$$

Since $\varepsilon \leq 1/3$, we have that $(1 + \varepsilon)/(1 - \varepsilon) \leq 1 + 3\varepsilon$, finishing the proof. ∎

Obtaining a $(1+\varepsilon)$-approximation for the linear j-subspace k-clustering problem without an exponential dependence on the dimension is now straightforward: we can adjust the input of Algorithm 3 so that it returns a $(1+\varepsilon/3)$-approximation, and we can use an $(\varepsilon/9)$-sketch $\widetilde{A}$. Using that $(1+\varepsilon/3)^2 \leq 1+\varepsilon$ we have that by Fact 59 the output of Algorithm 3 with input $(\widetilde{A}, \varepsilon/3)$ is an $(1+\varepsilon)$-approximation.

What would be the time complexity to find such an approximation with that sketch? We could argue that since $\widetilde{A}$ is a $d \times n$ matrix, the time complexity of Algorithm 3 would be the same as using A in the input. But we have to remember that the rank of $\widetilde{A}$ is proportional to the dimensionality of the problem. Since in the linear j-subspace k-clustering problem the underlying family $\mathcal{C}$ is jk-dimensional, by Theorem 45 we have[1] $\operatorname{rank}(\widetilde{A}) = 648\varepsilon^{-2}jk$. Using the ideas of Section 3.1, we can define a new matrix $\widetilde{A}' \in \mathbb{R}^{\lceil 648\varepsilon^{-2}jk \rceil \times n}$ that is an isometric embedding of the points of $\widetilde{A}$ into $\mathbb{R}^{\lceil 648\varepsilon^{-2}jk \rceil}$, and define a new linear j-subspace k-clustering problem in $\mathbb{R}^{\lceil 648\varepsilon^{-2}jk \rceil}$ such that every solution for $\widetilde{A}'$ can be mapped back to $\mathbb{R}^d$ while preserving the cost for $\widetilde{A}$.

Now Algorithm 3 with input $(\widetilde{A}', \varepsilon/3)$ will be executed in time

$$O\left(\left(\frac{j^2k^2n}{\varepsilon^3}\right)^{3jk\lceil 648\varepsilon^{-2}jk\rceil} \log\left(\frac{B}{\varepsilon B^*}\right)\right).$$

Using Algorithm 1 to find a matrix that with probability at least $1/2$ is the desired ε-sketch $\widetilde{A}$ takes time

$$O(ndj^2k^2\varepsilon^{-8}\log^2(\varepsilon^{-2}jk)).$$

We can find $\widetilde{A}'$ by embedding $\widetilde{A}$ onto $\operatorname{span}(\widetilde{A})$. If $B \in \mathbb{R}^{d \times \lceil 648\varepsilon^{-2}jk \rceil}$ is a matrix with orthonormal columns that spans $\widetilde{A}$, then a matrix that works as $\widetilde{A}'$ is

$$\widetilde{A}' := B^T\widetilde{A}.$$

Finding such a matrix B and computing the product takes time $O(nd\varepsilon^{-4}j^2k^2)$. Given a solution $\widetilde{C}$ in $\mathbb{R}^{\lceil 648\varepsilon^{-2}jk \rceil}$, it is straightforward to see that the solution $B\widetilde{C} \subset \mathbb{R}^d$ will have $\operatorname{dist}^2(\widetilde{A}, B\widetilde{C}) = \operatorname{dist}^2(\widetilde{A}', \widetilde{C})$.

The final time complexity is

$$O\left(\frac{ndj^2k^2\log^2(\varepsilon^{-2}jk)}{\varepsilon^8} + \left(\frac{j^2k^2n}{\varepsilon^3}\right)^{1947\varepsilon^{-2}j^2k^2} \log\left(\frac{B}{\varepsilon B^*}\right)\right),$$

and this algorithm will be correct with probability at least $1/2$, but this can be increased to $1-\delta$ for any $\delta \in (0,1)$, as we saw in Algorithm 1 remark.

We have thus obtained a PRAS for the linear j-subspace k-clustering problem. The probability that $\widetilde{A}$ is an $(\varepsilon/9)$-sketch is at least $1/2$, but we saw that it can be improved to $1-\delta$ for any $\delta \in (0,1)$ by running independent instances of the sketch.

[1] The value 648 comes from $8 \cdot 9^2$, with the factor 8 coming from Theorem 45, and the factor 9^2 coming from the fact we are using an $\varepsilon/9$-sketch.

Chapter 6

Conclusion and further questions

In this work we have presented recent results about approximations for the (ℓ_2^2, C)-clustering problem via dimension reduction. We showed how we can exploit inputs with low rank to obtain faster algorithms, we defined a type of approximation with sketches of the input, and we presented a randomized algorithm that finds a low-rank ε-sketch of the input in linear time with respect to the number of points of the input and the dimension. We have also presented an application of the ε-sketch to find an approximation for the linear j-subspace k-clustering problem, showing how we can improve the time complexity from an exponential dependence in the dimension to a polynomial one.

There are still multiple questions that can be explored about the (ℓ_2^2, C)-clustering problem and our ε-sketch. We saw that, when C is an m-dimensional family, we can efficiently find an ε-sketch of rank $O(\varepsilon^{-2}m)$ for any instance. But for the m-means clustering Cohen *et al.*, 2015 and Makarychev, Makarychev, and Razenshteyn, 2019 obtained better bounds, respectively $O(m/\varepsilon)$ and $O(\varepsilon^{-2}\log(m/\varepsilon))$. Can better upper bounds for the rank of ε-sketches be obtained for the (ℓ_2^2, C)-clustering problem or the linear j-subspace k-clustering problem? It would be interesting also to investigate lower bounds, i. e., if there are instances of the (ℓ_2^2, C)-clustering problem that does not admit an ε-sketch of rank too low.

To obtain our ε-sketches, we required the exact computation of the singular value decomposition of a matrix, therefore our model of computation must compute real numbers. How should Theorem 37 be modified if we only have floating point arithmetic? There are interesting questions we did not explore in the field of numerical analysis.

Our analysis showed that Algorithm 3 is not necessarily polynomial in the real RAM model. It would be interesting to find out if this algorithms would still be a $(1+\varepsilon)$-approximation if we only have floating point arithmetic or we only have rational numbers, and what would be the complexity of this algorithm with these restrictions, and in other models of computation.

Appendix A

Additional proofs

Here we prove inequality (2.8) necessary in the proof of Corollary 26, stated in Chapter 2, Section 2.3.

Fact 60. *Fix $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. If f is drawn from a $JLT(\varepsilon, \delta, n)$ then for all set A with $|A| = n$ and $0 \in A$ with probability $1-\delta$ we have that for all pairs $x, y \in A$*

$$(1-\varepsilon)\|x+y\|^2 \leq \|f(x)+f(y)\|^2 \leq (1+\varepsilon)\|x+y\|^2. \qquad \text{(2.9 revisited)}$$

Proof. For any vector v, we have $\|v\|^2 = \langle v, v\rangle$. Applying this to $v = x - y$ and $v = x + y$ we deduce that

$$2\langle x, y\rangle = \|x\|^2 + \|y\|^2 - \|x-y\|^2$$

and

$$2\langle x, y\rangle = -\|x\|^2 - \|y\|^2 + \|x+y\|^2.$$

Now by subtracting one equation from another and rearranging we arrive at

$$\|x+y\|^2 = 2\|x\|^2 + 2\|y\|^2 - \|x-y\|^2.$$

This is true in particular to

$$\|f(x)+f(y)\|^2 = 2\|f(x)\|^2 + 2\|f(y)\|^2 - \|f(x)-f(y)\|^2,$$

and since f is an $JLT(\varepsilon, \delta, n)$, the null vector belongs to A and $f(0) = 0$ Therefore we have

$$\|f(x)+f(y)\|^2 \leq (1+\varepsilon)(2\|x\|^2 + 2\|y\|^2 - \|x-y\|^2) = (1+\varepsilon)\|x+y\|^2.$$

The lower bound is analogous. ∎

Here we prove Lemma 27 stated in Chapter 2, Section 2.3, Subsection 2.3.1.

Lemma 27. *Fix $\varepsilon \in (0,1)$ and let $S \in \mathbb{R}^{\lceil \varepsilon^{-2} \rceil \times d}$ be a normalized Johnson-Lindenstrauss matrix.*

Then for any $x, y \in \mathbb{R}^d$ we have that

$$\mathbb{E}\left[\langle Sx, Sy\rangle\right] = \langle x, y\rangle,$$

and

$$\mathrm{Var}[\langle Sx, Sy\rangle] \leq 2\varepsilon^2 \|x\|^2 \|y\|^2.$$

Proof. For an easier notation let $r = \lceil \varepsilon^{-2} \rceil$. Let s_{ij} be the i^{th} row and j^{th} column entry of S and for any vector x let x_j be its j^{th} coordinate. Note that

$$(Sx)_i = \frac{1}{\sqrt{r}} \sum_{j=1}^{d} s_{ij} x_j$$

and

$$\langle x, y\rangle = \sum_{j=1}^{d} x_j y_j.$$

Then

$$\begin{aligned}\langle Sx, Sy\rangle &= \frac{1}{r} \sum_{i=1}^{r} \left(\sum_{j=1}^{d} s_{ij} x_j \right) \left(\sum_{j'=1}^{d} s_{ij'} y_{j'} \right) \\ &= \frac{1}{r} \sum_{i=1}^{r} \sum_{j=1}^{d} \sum_{j'=1}^{d} s_{ij} s_{ij'} x_j y_{j'}. \end{aligned} \tag{A.1}$$

Now, using the expectancy linearity, we obtain

$$\mathbb{E}\left[\langle Sx, Sy\rangle\right] = \frac{1}{r} \sum_{i=1}^{r} \sum_{j=1}^{d} \sum_{j'=1}^{d} \mathbb{E}[s_{ij} s_{ij'}] x_j y_{j'}, \tag{A.2}$$

and since s_{ij} is independent of $s_{ij'}$ when $j \neq j'$, we have $\mathbb{E}[s_{ij} s_{ij'}] = \mathbb{E}[s_{ij}]\mathbb{E}[s_{ij'}] = 0$. When $j = j'$ we have that $\mathbb{E}[s_{ij} s_{ij'}] = \mathbb{E}[s_{ij}^2] = 1$, and thus

$$\mathbb{E}\left[\langle Sx, Sy\rangle\right] = \frac{1}{r} \sum_{i=1}^{r} \sum_{j=1}^{d} x_j y_j = \langle x, y\rangle. \tag{A.3}$$

To obtain the bound on the variance, we will first bound the second moment $\mathbb{E}\left[\langle Sx, Sy\rangle^2\right]$. We have that

$$\begin{aligned}\langle Sx, Sy\rangle^2 &= \left(\frac{1}{r} \sum_{i=1}^{r} \sum_{j=1}^{d} \sum_{j'=1}^{d} s_{ij} s_{ij'} x_j y_{j'} \right)^2 \\ &= \frac{1}{r^2} \sum_{i=1}^{r} \sum_{i'=1}^{r} \sum_{j=1}^{d} \sum_{j'=1}^{d} \sum_{j''=1}^{d} \sum_{j'''=1}^{d} s_{ij} s_{ij'} s_{i'j''} s_{i'j'''} x_j y_{j'} x_{j''} y_{j'''}. \end{aligned} \tag{A.4}$$

When we apply the linearity of expectancy in (A.4), the main term we need to analyse is

$$\mathbb{E}[s_{ij}s_{ij'}s_{i'j''}s_{i'j'''}]. \tag{A.5}$$

Note that by independence, if any one pair of index ij differs from the other three, then the term above will be zero since $\mathbb{E}[s_{ij}] = 0$. Thus we have that

1. **When** $j = j'$ **and** $j'' = j'''$ the term (A.5) will be one for any i and i';

2. **When** $j = j''$ **and** $j' = j'''$ the term (A.5) will be one if and only if $i = i'$;

3. **When** $j = j'''$ **and** $j' = j''$ the term (A.5) will be one if and only if $i = i'$.

For any other relation of indexes the term (A.5) will be zero. Therefore

$$\mathbb{E}\left[\langle Sx, Sy\rangle^2\right] = \frac{1}{r^2}\sum_{i=1}^{r}\sum_{i'=1}^{r}\sum_{j=1}^{d}\sum_{j''=1}^{d} x_j y_j x_{j''} y_{j''} \tag{A.6}$$

$$+ \frac{1}{r^2}\sum_{i=1}^{r}\sum_{j=1}^{d}\sum_{j'=1}^{d} x_j^2 y_{j'}^2 \tag{A.7}$$

$$+ \frac{1}{r^2}\sum_{i=1}^{r}\sum_{j=1}^{d}\sum_{j'=1}^{d} x_j y_j x_{j'} y_{j'}, \tag{A.8}$$

where parcel (A.6) is due to item 1, parcel (A.7) is due to item 2 and parcel (A.8) is due to item 3.

Each parcel can be bounded. For parcel (A.6) we have

$$\frac{1}{r^2}\sum_{i=1}^{r}\sum_{i'=1}^{r}\sum_{j=1}^{d}\sum_{j''=1}^{d} x_j y_j x_{j''} y_{j''} = \langle x, y\rangle^2 = \mathbb{E}[\langle Sx, Sy\rangle]^2.$$

For parcel (A.7) we have

$$\frac{1}{r^2}\sum_{i=1}^{r}\sum_{j=1}^{d}\sum_{j'=1}^{d} x_j^2 y_j'^2 = \frac{1}{r}\|x\|^2\|y\|^2.$$

For parcel (A.8) by Cauchy-Schwartz we have

$$\frac{1}{r^2}\sum_{i=1}^{r}\sum_{j=1}^{d}\sum_{j'=1}^{d} x_j y_j x_{j'} y_{j'} = \frac{1}{r}\langle x, y\rangle^2 \leq \frac{1}{r}\|x\|^2\|y\|^2.$$

Finally, by the definition of variance we have

$$\begin{aligned}\mathrm{Var}[\langle Sx, Sy\rangle] &= \mathbb{E}[\langle Sx, Sy\rangle^2] - \mathbb{E}[\langle Sx, Sy\rangle]^2 \\ &\leq \frac{2}{r}\|x\|^2\|y\|^2 + \langle x, y\rangle^2 - \langle x, y\rangle^2 \\ &\leq 2\varepsilon^2\|x\|^2\|y\|^2,\end{aligned}$$

completing the proof. ∎

Here we prove Lemma 46, Stated in Chapter 4, Section 4.2.

Lemma 46 (Weak triangle inequality). *Let $p, q \in \mathbb{R}^d$ be fixed and let $C \subset \mathbb{R}^d$ be a non-empty set. For all $\varepsilon \in (0, 1)$, we have*

$$\left|\mathrm{dist}^2(p, C) - \mathrm{dist}^2(q, C)\right| \leq \varepsilon\,\mathrm{dist}^2(p, C) + \frac{2}{\varepsilon}\|p - q\|^2. \qquad \text{(4.18 revisited)}$$

Proof. From factoring the right-hand side of inequality (4.18) we obtain

$$\left|\mathrm{dist}^2(p, C) - \mathrm{dist}^2(q, C)\right| = |\mathrm{dist}(p, C) - \mathrm{dist}(q, C)| \cdot (\mathrm{dist}(p, C) + \mathrm{dist}(q, C)). \qquad \text{(A.9)}$$

From usual metric space triangle inequality we have that for any two points $p, q \in \mathbb{R}^d$ and a set $C \subset \mathbb{R}^d$

$$|\mathrm{dist}(p, C) - \mathrm{dist}(q, C)| \leq \|p - q\|. \qquad \text{(A.10)}$$

Using inequality (A.10) on each factor of inequality (A.9) we obtain

$$\begin{aligned}&|\mathrm{dist}(p, C) - \mathrm{dist}(q, C)| \cdot (\mathrm{dist}(p, C) + \mathrm{dist}(q, C)) \\ &\leq \|p - q\|\,(\mathrm{dist}(p, C) + \mathrm{dist}(p, C) - \mathrm{dist}(p, C) + \mathrm{dist}(q, C) + \|p - q\|) && \text{(A.11)} \\ &\leq \|p - q\|\,(2\,\mathrm{dist}(p, C) + \|p - q\|) && \text{(A.12)} \\ &= 2\,\|p - q\|\,\mathrm{dist}(p, C) + \|p - q\|^2. && \text{(A.13)}\end{aligned}$$

Now using that $\varepsilon \neq 0$ we have expression (A.13) equals

$$2\,\mathrm{dist}(p, C)\sqrt{\varepsilon}\frac{1}{\sqrt{\varepsilon}}\|p - q\| + \|p - q\|^2. \qquad \text{(A.14)}$$

Finally we use that for any two reals a and b it is true that $2ab \leq a^2 + b^2$ to bound

$$\begin{aligned}2\,\mathrm{dist}(p, C)\sqrt{\varepsilon}\frac{1}{\sqrt{\varepsilon}}\|p - q\| + \|p - q\|. &\leq \varepsilon\,\mathrm{dist}^2(p, C) + \frac{1}{\varepsilon}\|p - q\|^2 + \|p - q\|^2 \\ &\leq \varepsilon\,\mathrm{dist}^2(p, C) + \frac{2}{\varepsilon}\|p - q\|,\end{aligned}$$

finishing the proof. ∎

Here we construct a δ-net for the origin, necessary for Algorithm 2 in Chapter 5, Section 5.1.

Fact 61. *For every $\mathcal{R} > 0$ and every $\delta > 0$, if $H \subset \mathbb{R}^d$ is a $(\delta\mathcal{R}^{-1}d^{-1/2})$-fine grid of dimension d then*

$$H' = \{\mathcal{R}h \,:\, \forall h \in H\}$$

is a δ-net for the origin.

Proof. Let $x \in \mathbb{R}^d$ such that $\|x\| \leq \mathcal{R}$. We will find an $y \in H'$ that is close enough to x. Suppose that $e_1, \ldots, e_d$ are the vectors of the canonical base. There exists reals $\alpha_1, \ldots, \alpha_d$ such that

$$x = \sum_{i=1}^{d} \alpha_i e_i.$$

Note that for every $i = 1, \ldots, d$ we have $|\alpha_i| \leq \mathcal{R}$. We will now define each coordinate of a vector $y \in H'$ such that $\|x - y\| \leq \delta$. For every $i = 1, \ldots, d$, let

$$c_i := \left\lfloor \alpha_i \frac{\sqrt{d}}{\delta} \right\rfloor.$$

From the fact that $|\alpha_i| \leq \mathcal{R}$, we conclude that $c_i \in [-d^{1/2}\mathcal{R}\delta^{-1}, d^{1/2}\mathcal{R}\delta^{-1}] \cap \mathbb{Z}$. Therefore the vector

$$y := \sum_{i=1}^{d} \frac{\delta}{\sqrt{d}} c_i e_i$$

is a member of H'. Now we bound the distance $\|x - y\|$.

$$\begin{aligned}
\|x - y\| &= \sqrt{\sum_{i=1}^{d} \left(\alpha_i - \frac{\delta}{\sqrt{d}} \left\lfloor \alpha_i \frac{\sqrt{d}}{\delta} \right\rfloor \right)^2} \\
&= \sqrt{\sum_{i=1}^{d} \left(\frac{\delta}{\sqrt{d}} \left(\alpha_i \frac{\sqrt{d}}{\delta} - \left\lfloor \alpha_i \frac{\sqrt{d}}{\delta} \right\rfloor \right) \right)^2},
\end{aligned}$$

and from the definition of floor function we can conclude that

$$\sqrt{\sum_{i=1}^{d} \left(\frac{\delta}{\sqrt{d}} \left(\alpha_i \frac{\sqrt{d}}{\delta} - \left\lfloor \alpha_i \frac{\sqrt{d}}{\delta} \right\rfloor \right) \right)^2} \leq \sqrt{\sum_{i=1}^{d} \left(\frac{\delta}{\sqrt{d}} \right)^2} \leq \delta,$$

and thus $\|x - y\| \leq \delta$. ∎

Appendix B

Constant size family example

Here we show an example of a family of instances of constant size in our model of computation for Algorithm 3 that takes an arbitrarily large amount of time to execute.

Let p be a positive integer and let $A^{(p)} \in \mathbb{R}^{2\times 4}$ be a matrix with columns representing the vectors $a_1 = (\sqrt{2}, 0)$, $a_2 = (0, \sqrt{2})$, $a_3 = (p^{-2}, p^{-2})$ and $a_4 = (p^{-2}, -p^{-2})$. Let L_1 be the subspace spanned by a_1, let L_2 be the subspace spanned by a_2, let L_3 be the subspace spanned by a_3 and let L_4 be the subspace spanned by a_4.

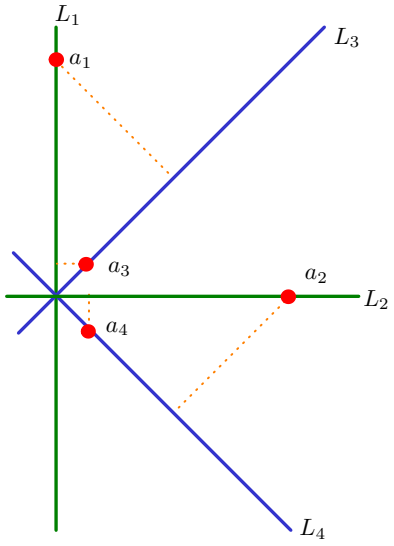


Figure B.1: *A visualisation of the instance A and the solutions. The red dots represent the instance. The green lines represent the solution C_1. The red lines represent the solution C_2.*

Let $C_1 := L_1 \cup L_2$ and $C_2 := L_3 \cup L_4$. Figure B.1 illustrate the instance and the solutions. We have that

$$\text{dist}^2(A, C_1) = (\text{dist}(a_3, L_1))^2 + (\text{dist}(a_4, L_2))^2 = 2p^{-1},$$

and

$$\text{dist}^2(A, C_2) = (\text{dist}(a_1, L_3))^2 + (\text{dist}(a_2, L_4))^2 = 2.$$

Let B^* be the optimal cost for A and let $B := \text{dist}^2(A, C_2)$. We can see that for any positive value of p we have

$$\frac{B}{B^*} \geq p.$$

This means that for any fixed $\varepsilon \in (0, 1)$ if we run Algorithm 3 with input $(A^{(p)}, B, \varepsilon)$ we could have an arbitrarily large execution time with an input of constant size, since p can be arbitrarily large and our model of computation can store any real consuming $O(1)$ memory space.

Appendix C

Quick reference for theorems, lemmas, corollaries, facts and definitions

Definition 1. *Let $C \subset \mathbb{R}^d$ be a non-empty set and let $p \in \mathbb{R}^d$ be a vector. Then*

$$\operatorname{dist}(p, C) := \inf\{\|p - c\| : c \in C\},$$

where $\|\cdot\|$ is the usual Euclidean norm.

Definition 2. *Let $C \subset \mathbb{R}^d$ be a non-empty set and let $A \in \mathbb{R}^{d\times n}$ be a matrix. Then*

$$\operatorname{dist}^2(A, C) := \sum_{a\in A}(\operatorname{dist}(a, C))^2.$$

Definition 3. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix. The subspace of $\mathbb{R}^d$ spanned by the columns of A will be denoted as* $\operatorname{span}(A)$

Definition 4. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and let a_{ij} be the row i and column j entry of A. The* Frobenius norm *of A is*

$$\|A\|_F := \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{n} a_{ij}^2}.$$

Definition 5. *We say that a matrix $P \in \mathbb{R}^{d\times j}$ have* orthonormal columns *when the set $\{p_1, \ldots, p_j\}$ of columns of P are linearly independent and all have norm one.*

Definition 6 (Orthogonal projection)**.** *Let $v \in \mathbb{R}^d$ be a vector and let L be a subspace of $\mathbb{R}^d$. Then $\pi_L(v) \in \mathbb{R}^d$ is the* orthogonal projection of v onto L.

Definition 7 (Orthogonal projection of a matrix)**.** *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and let L be a subspace of $\mathbb{R}^d$. The* orthogonal projection of A into L *is the matrix $\pi_L(A) \in \mathbb{R}^{d\times n}$ where the i^{th} column of $\pi_L(A)$ is the orthogonal projection of the i^{th} column of A into L.*

Fact 8. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, let L be a subspace of $\mathbb{R}^d$ of dimension j and let $P \in \mathbb{R}^{d\times j}$ be a matrix with orthonormal columns that spans L. Then the orthogonal projection $\pi_L(A)$ equals PP^TA. Also note that*

$$\text{dist}^2(A, L) = \|A - PP^TA\|_F^2.$$

Definition 9. *Let A and $B \in \mathbb{R}^{d\times n}$ be matrices. We will abuse the notation and adopt that $\pi_B(A)$ means the same as $\pi_{\text{span}(B)}(A)$.*

Fact 10 (Pythagoras Theorem)**.** *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, let L be a subspace of $\mathbb{R}^d$ of dimension j and let $P \in \mathbb{R}^{d\times j}$ with orthonormal columns that spans L. Then it follows from the Pythagoras Theorem that*

$$\|A\|_F^2 = \|PP^TA\|_F^2 + \|A - PP^TA\|_F^2.$$

Fact 11 (Trace representation)**.** *For any real matrix A, we have that*

$$\|A\|_F^2 = \text{Tr}(A^TA)$$

and

$$\|A\|_F^2 = \|A^T\|_F^2.$$

Fact 12 (Cyclic property of the Trace function)**.** *For any matrices $A \in \mathbb{R}^{d\times n}$ and $B \in \mathbb{R}^{n\times d}$, we have*

$$\text{Tr}(AB) = \text{Tr}(BA).$$

Fact 13 (Unitarily invariant norm)**.** *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, let $P \in \mathbb{R}^{d\times d}$ and $Q \in \mathbb{R}^{n\times n}$ be orthogonal matrices and let $S \in \mathbb{R}^{s\times d}$ be a matrix with orthonormal columns and let $R \in \mathbb{R}^{n\times r}$ be such that R^T has orthonormal columns. Then*

$$\|A\|_F^2 = \|PA\|_F^2 = \|AQ\|_F^2 = \|SA\|_F^2 = \|AR\|_F^2.$$

Definition 14 (Big-O notation)**.** *For two functions $f, g : \mathbb{R} \to \mathbb{R}$, we say that $f(x) = O(g(x))$ if there exist constants K and x_0 such that $|f(x)| \le K|g(x)|$ for all $x \ge x_0$. We say that $f(x) = \Omega(g(x))$ if $g(x) = O(f(x))$. Finally, we say that $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $f(x) = \Omega(g(x))$.*

Definition 15 (best-fit linear j-subspace problem)**.** *Let $A \in \mathbb{R}^{d\times n}$. The* best-fit linear j-subspace *problem in $\mathbb{R}^d$ is to find a linear subspace V of dimension j of $\mathbb{R}^d$ that minimizes* $\text{dist}^2(A, V)$.

Notation 16. *Let A be a matrix and $1 \le i \le \text{rank}(A)$ an integer. The i^{th} singular value of A is denoted $\sigma_i(A)$.*

Definition 17. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix. A* first left singular vector *of A is any vector $u_1 \in \mathbb{R}^d$ such that*

$$u_1 := \text{argmax}\left\{\|u^TA\|_F : \|u\| = 1\right\},$$

and the first singular value *is*

$$\sigma_1 := \|u_1^T A\|_F.$$

We say that u_1 is the left singular vector associated with the singular value σ_1.

Fact 18. *The matrix A_k is the orthogonal projection of A to the subspace spanned by the first k left singular vectors $u_1, \dots, u_k$.*

Fact 19. *The matrix A_k is a* best rank k approximation *of A in the Frobenius norm, that is, for any matrix $B \in \mathbb{R}^{d\times n}$ with rank at most k we have*

$$\|A - A_k\|_F \leq \|A - B\|_F. \tag{2.4}$$

Fact 20. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix of rank r. The first r eigenvalues $\lambda_1, \dots, \lambda_r$ of the matrix AA^T are the square of the singular values $\sigma_1^2, \dots, \sigma_r^2$ of A, and the remaining eigenvalues are zero. The eigenvectors associated with the eigenvalues $\lambda_i = \lambda_{i+1} = \cdots = \lambda_j$ span the same subspace as any sequence $u_i, u_{i+1}, \dots, u_j$ of left singular vectors associated with the singular values $\sigma_i = \sigma_{i+1} = \cdots = \sigma_j$.*

Fact 21. *For any matrix $A \in \mathbb{R}^{d\times n}$, we have*

$$\|A\|_F^2 = \mathrm{Tr}(AA^T) = \sum_{i=1}^{d} \lambda(AA^T) = \sum_{i=1}^{d} \sigma^2(A).$$

Theorem 22 (Johnson-Lindenstrauss Lemma)**.** *There exists a constant κ such that for any set A of n points in $\mathbb{R}^d$, any $\varepsilon \in (0,1)$ fixed and all integers $r \geq \kappa\varepsilon^{-2}\log n$ there exists a linear function $f : \mathbb{R}^d \to \mathbb{R}^r$ such that for every pair $x, y \in A$ we have*

$$(1-\varepsilon)\|x-y\|^2 \leq \|f(x)-f(y)\|^2 \leq (1+\varepsilon)\|x-y\|^2. \tag{2.5}$$

Lemma 23 (Random projection lemma)**.** *Suppose $T \in \mathbb{R}^{r\times d}$ is a random matrix where each entry t_{ij} of T is an independent random variable that assumes values uniformly in $\{+1, -1\}$. Let $S \in \mathbb{R}^{r\times d}$ be defined as*

$$S = \frac{1}{\sqrt{r}} T.$$

Then for all $v \in \mathbb{R}^d$ and $\varepsilon \in (0,1)$ we have

$$\mathbb{P}\left[\|Sv\|^2 > (1+\varepsilon)\|v\|^2\right] < e^{-r\varepsilon^2/12},$$
$$\mathbb{P}\left[\|Sv\|^2 < (1-\varepsilon)\|v\|^2\right] < e^{-r\varepsilon^2/12}.$$

Definition 24 (Johnson-Lindenstrauss Transform)**.** *Let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ be fixed. Suppose that $r < d$ are positive integers and that $\mathcal{D}_{d,r}$ is a probability distribution over the space of linear functions with domain $\mathbb{R}^d$ and codomain $\mathbb{R}^r$. We say that $\mathcal{D}_{d,r}$ is a* Johnson-Lindenstrauss Transform *with parameters ε, δ, and n or $JLT(\varepsilon, \delta, n)$ for short if for any set $A \subset \mathbb{R}^d$ of n points with probability at least $1-\delta$ we have that a function f drawn*

from $\mathcal{D}_{d,r}$ satisfies for every pair $x, y \in A$

$$(1-\varepsilon)\|x-y\|^2 \leq \|f(x)-f(y)\|^2 \leq (1+\varepsilon)\|x-y\|^2.$$

Definition 25. *Fix $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Suppose that $r < d$ are positive integers. We say that a random matrix S is a* Johnson-Lindenstrauss Transform *with parameters ε, δ, and n or $JLT(\varepsilon, \delta, n)$ for short if for any set $A \subset \mathbb{R}^d$ of n points with probability at least $1-\delta$ for every pair $x, y \in A$ we have*

$$(1-\varepsilon)\|x-y\|^2 \leq \|Sx-Sy\|^2 \leq (1+\varepsilon)\|x-y\|^2.$$

Corollary 26. *Let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ be fixed. If f is drawn from a $JLT(\varepsilon, \delta, n+1)$ then for any $A \subset \mathbb{R}^d$ set of n points, with probability at least $1-\delta$ for all $x, y \in A$ we have*

$$\langle x, y\rangle - \varepsilon\|x\|\|y\| \leq \langle f(x), f(y)\rangle \leq \langle x, y\rangle + \varepsilon\|x\|\|y\|. \tag{2.7}$$

Lemma 27. *Let $\varepsilon \in (0,1)$ be fixed and let $S \in \mathbb{R}^{\lceil \varepsilon^{-2}\rceil \times d}$ be a normalized Johnson-Lindenstrauss matrix. Then for every $x, y \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\langle Sx, Sy\rangle\right] = \langle x, y\rangle$$

and

$$\mathrm{Var}\left[\langle Sx, Sy\rangle\right] \leq 2\varepsilon^2 \|x\|^2 \|y\|^2 .$$

Lemma 28 (Sarlós, 2006)**.** *Let $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{d\times m}$ be matrices, and let $\varepsilon \in (0,1)$ be fixed. The following statements are true:*

1. *If a matrix S is a $JLT(\varepsilon, \delta, n+m+1)$ then with probability at least $1-\delta$ we have*

$$\left\|AB - AS^TSB\right\|_F \leq \varepsilon \|A\|_F \|B\|_F .$$

2. *If a matrix $S \in \mathbb{R}^{\lceil \varepsilon^{-2}\rceil \times d}$ is a normalized Johnson-Lindenstrauss matrix then*

$$\mathbb{E}\left[AS^TSB\right] = AB$$

and

$$\mathbb{E}\left[\left\|AB - AS^TSB\right\|_F^2\right] \leq 2\varepsilon^2 \|A\|_F^2 \|B\|_F^2 ,$$

where the mean $\mathbb{E}[X]$ of a random matrix X is the matrix where the entry $\mathbb{E}[X]_{ij}$ is $\mathbb{E}\left[X_{ij}\right]$.

Theorem 29 (Low dimensional representation)**.** *Let $A \in \mathbb{R}^{d\times n}$ be an instance of best-fit linear j-subspace problem in $\mathbb{R}^d$, and suppose that $j < \mathrm{rank}(A) = d' < d$. Then there exist an instance $A' \in \mathbb{R}^{d'\times n}$ of best-fit linear j-subspace problem in $\mathbb{R}^{d'}$ that satisfy the following:*

1. *There exist an isometric embedding $f : \mathbb{R}^{d'} \to \mathbb{R}^d$ such that if C' is a solution for A'*

then $C := f(C')$ *is a solution for* A *and*

$$\operatorname{dist}^2(A', C') = \operatorname{dist}^2(A, C).$$

2. *For every solution* C *for* A *there exist a solution* C' *for* A' *such that*

$$\operatorname{dist}^2(A', C') \leq \operatorname{dist}^2(A, C).$$

Fact 30. *Let* $A \in \mathbb{R}^{d\times n}$ *and* $B \in \mathbb{R}^{d\times j}$ *be matrices. Suppose that* B *has orthonormal columns. Then*

$$\|A\|_F^2 \geq \|B^T A\|_F^2.$$

Corollary 31. *Under the same conditions of Theorem 29, we have that if* C'^* *is an optimal solution for* A'*, then* $f(C'^*)$ *is an optimal solution for* A*, where* f *is the isometric embedding given by* (1) *of Theorem 29.*

Definition 32. *Let* $A \in \mathbb{R}^{d\times n}$ *be a instance of* (ℓ_2^2, C)*-clustering problem and let* $\alpha \geq 1$ *be a real. Let* C^* *be an optimal solution for* A*. We say that a solution* $\widetilde{C}$ *is an* α-approximation *for* A *if*

$$\operatorname{dist}^2(A, \widetilde{C}) \leq \alpha \operatorname{dist}^2(A, C^*).$$

The real α *is called* approximation factor.

Fact 33. *Let* $A \in \mathbb{R}^{d\times n}$ *be a matrix, let* V *be a subspace of dimension* j *and let* $C \subset V$ *be a non-empty set. Suppose that* $B \in \mathbb{R}^{d\times j}$ *has orthonormal columns that spans* V*. Then*

$$\operatorname{dist}^2(A, C) = \|A - BB^T A\|_F^2 + \operatorname{dist}^2(BB^T A, C).$$

Definition 34 (ε-sketch for (ℓ_2^2, C)-clustering)**.** *Let* $A \in \mathbb{R}^{d\times n}$ *be an instance for* (ℓ_2^2, C)*-clustering problem and let* $\varepsilon \in (0, 1)$ *be fixed. We say that a matrix* $\widetilde{A} \in \mathbb{R}^{d\times n}$ *is an* ε-sketch *for* A *if there exist a non-negative constant* $\Delta = \Delta(A, C, \varepsilon)$ *such that for every solution* $C \in C$ *we have*

$$(1 - \varepsilon)\operatorname{dist}^2(A, C) \leq \operatorname{dist}^2(\widetilde{A}, C) + \Delta \leq (1 + \varepsilon)\operatorname{dist}^2(A, C). \tag{3.2}$$

Definition 35 (weak ε-sketch for (ℓ_2^2, C)-clustering)**.** *Let* $A \in \mathbb{R}^{d\times n}$ *be an instance for* (ℓ_2^2, C)*-clustering problem and let* $\varepsilon \in (0, 1)$ *be fixed. We say that* $\widetilde{A} \in \mathbb{R}^{d\times n}$ *is a* weak ε-sketch *for* A *if for every optimal solution* $\widetilde{C}^*$ *for* $\widetilde{A}$ *and any optimal solution* C^* *for* A *we have*

$$\operatorname{dist}^2(A, \widetilde{C}^*) \leq (1 + \varepsilon)\operatorname{dist}^2(A, C^*). \tag{3.3}$$

Definition 36 (Dimension reduction scheme)**.** *A* dimension reduction scheme *for* (ℓ_2^2, C)*-clustering problem in* $\mathbb{R}^d$ *is any scheme that takes an instance* $A \in \mathbb{R}^{d\times n}$ *and a parameter* $\varepsilon \in (0, 1)$ *as input and outputs another instance* $\widetilde{A} \in \mathbb{R}^{d\times n}$ *such that* $\operatorname{rank}(\widetilde{A}) < d$ *and* $\widetilde{A}$ *is either an* ε*-sketch for* A *or a weak* ε*-sketch for* A*.*

Theorem 37 (Sarlós, 2006)**.** *Let* $A \in \mathbb{R}^{d\times n}$ *be a matrix, let* $j < \min\{d, n\}$ *be an integer and let* $\varepsilon \in (0, 1)$ *be fixed. There exists an integer* $r = \Theta(\varepsilon^{-1} j + j \log j)$ *such that if* S *is a* $r \times n$

normalized Johnson-Lindenstrauss matrix then with probability at least $1/2$ *we have*

$$\|A - \pi_{AS^T}(A)_j\|_F \leq (1+\varepsilon)\|A - A_j\|_F.$$

Computing $\pi_{AS^T}(A)_j$ *can be done with two readings of the matrix* A *and in time* $O(ndr + (n+d)r^2)$.

Lemma 38 (Sarlós, 2006)**.** *Let* $L \subset \mathbb{R}^d$ *be a subspace of dimension* j*, and let* $\varepsilon \in (0,1)$ *and* $\delta \in (0,1)$ *be fixed. If* f *is drawn from some* $JLT(\varepsilon/4, \delta, O((j/\varepsilon)^j))$*, then with probability at least* $1-\delta$ *for all* $v \in L$ *we have*

$$(1-\varepsilon)\|v\| \leq \|f(v)\| \leq (1+\varepsilon)\|v\|.$$

We say that such an f *is a* subspace ε-embedding.

Definition 39 (δ-fine grid)**.** *Fix* $\delta \in (0,1)$*. Let* $e_1, \ldots, e_j$ *be the vectors of the canonical base of* $\mathbb{R}^j$*. The set*

$$H := \left\{ h \in \mathbb{R}^j \,:\, \forall i = 1, \ldots, j, \forall c_i \in \mathbb{Z} \cap [-\delta^{-1}, \delta^{-1}], h = \sum_{i=1}^{j} \delta c_i e_i \right\}.$$

is called δ-fine grid *of dimension* j.

Corollary 40 (Sarlós, 2006)**.** *Let* $U \in \mathbb{R}^{d\times j}$ *be a matrix with orthonormal columns and let* $\varepsilon \in (0,1)$ *and* $\delta \in (0,1)$ *be fixed. Suppose that* S *is an* $r \times d$ *normalized Johnson-Lindenstrauss matrix for some* $r = O(\varepsilon^{-2} j \log(j/\varepsilon)\log(1/\delta))$*. Then with probability at least* $1-\delta$ *for all* $i = 1, \ldots, j$ *we have*

$$|1 - \sigma_i(SU)| \leq \varepsilon.$$

Definition 41 (linear ℓ_2 regression)**.** *Let* $A \in \mathbb{R}^{d\times n}$ *be a matrix and let* $b \in \mathbb{R}^d$ *be fixed. The* linear ℓ_2 regression problem *is to find at least one vector* $x^* \in \mathbb{R}^n$ *such that for all* $x \in \mathbb{R}^n$ *we have*

$$\|b - Ax^*\| \leq \|b - Ax\|.$$

Theorem 42 (Sarlós, 2006)**.** *Let* $A \in \mathbb{R}^{d\times n}$ *be a matrix and let* $b \in \mathbb{R}^d$ *and* $\varepsilon \in (0,1)$ *be fixed. Let* $\mathcal{Z} = \min_{x\in\mathbb{R}^n} \|b - Ax\| = \|b - Ax^*\|$*, where* $x^* = A^\dagger b$*. Let* $S \in \mathbb{R}^{s\times d}$ *be a normalized Johnson-Lindenstrauss matrix for some positive integer* $s = O(\varepsilon^{-1} n \log n)$*. Finally, let* $\widetilde{\mathcal{Z}} = \min_{x\in\mathbb{R}^n} \|Sb - SAx\| = \|Sb - SA\widetilde{x}^*\|$*, where* $\widetilde{x}^* = (SA)^\dagger Sb$*. Then with probability at least* $1/3$ *we have*

$$\|b - A\widetilde{x}^*\| \leq (1+\varepsilon)\mathcal{Z}.$$

Fact 43. *There exists some* $s = O(\varepsilon^{-1} n \log n)$ *such that with probability at least* $2/3$ *we have*

$$\sigma_i((SU)^T(SU)) = \sigma_i^2(SU) \geq \frac{1}{\sqrt{2}}.$$

Definition 44. *Let* $\mathcal{C}$ *be a non-empty set of non-empty subsets of* $\mathbb{R}^d$*, and let* $m \leq d$ *be a positive integer. We say that* $\mathcal{C}$ *is an* m-dimensional family *if for every* $C \in \mathcal{C}$ *there exists a*

linear subspace $L(C) \subset \mathbb{R}^d$ of dimension m that contains C.

Theorem 45 (based on Pratap and Sen, 2018). *Let $A \in \mathbb{R}^{d\times n}$ be an instance of the (ℓ_2^2, C)-clustering problem where C is an m-dimensional family and $m < \min\{d, n\}$. Let $\varepsilon \in (0,1)$ be fixed. There exists an integer $s = \lceil 8\varepsilon^{-2}m \rceil$ such that if a matrix $\widetilde{A}^T \in \mathbb{R}^{n\times d}$ is an orthogonal projection of A^T to some subspace of dimension s of $\mathbb{R}^n$ and satisfies*

$$\left\|A - \widetilde{A}\right\|_F^2 \leq \left(1 + \frac{\varepsilon^2}{8}\right) \|A - A_s\|_F^2 \tag{4.17}$$

then $\widetilde{A}$ is an ε-sketch for A with constant $\Delta = \|A - A_s\|_F^2$.

Lemma 46 (Weak triangle inequality). *Let $p, q \in \mathbb{R}^d$ be fixed and let $C \subset \mathbb{R}^d$ be a non-empty set. For all $\varepsilon \in (0,1)$ we have*

$$\left|\operatorname{dist}^2(p, C) - \operatorname{dist}^2(q, C)\right| \leq \varepsilon \operatorname{dist}^2(p, C) + \frac{2}{\varepsilon}\|p - q\|^2 \tag{4.18}$$

Lemma 47. *Let $A \in \mathbb{R}^{d\times n}$ be a matrix and suppose that $m < \min\{d, n\}$ is a positive integer. Fix $\varepsilon' \in (0,1)$ and $s = \lceil \varepsilon'^{-1}m \rceil$. Suppose that $\widetilde{A}^T \in \mathbb{R}^{n\times d}$ is a matrix that is an orthogonal projection of A^T to some subspace of dimension s of $\mathbb{R}^n$ and satisfies*

$$\left\|A - \widetilde{A}\right\|_F^2 \leq (1 + \varepsilon')\|A - A_s\|_F^2 . \tag{4.19}$$

For every matrix $X \in \mathbb{R}^{d\times m}$ with orthonormal columns and matrix $Y \in \mathbb{R}^{d\times(d-m)}$ with orthonormal columns such that $\operatorname{span}(Y)$ is the orthogonal complement of $\operatorname{span}(X)$ the following inequalities are true:

$$0 \leq \left\|X^T A\right\|_F^2 - \left\|X^T \widetilde{A}\right\|_F^2 \leq 2\varepsilon' \left\|Y^T A\right\|_F^2 \tag{4.20}$$

and

$$\left\|XX^T A - XX^T \widetilde{A}\right\|_F^2 \leq 2\varepsilon' \left\|Y^T A\right\|_F^2 . \tag{4.21}$$

Fact 48. *Let A and B be $d \times n$ matrices. Then $\operatorname{rank}(A + B)$ is at most $\operatorname{rank}(A) + \operatorname{rank}(B)$.*

Theorem 49. *Under the same hypothesis of Lemma 47 we have that*

$$\left|\left\|Y^T \widetilde{A}\right\|_F^2 + \|A - A_s\|_F^2 - \left\|Y^T A\right\|_F^2\right| \leq 2\varepsilon' \left\|Y^T A\right\|_F^2 .$$

Definition 50 (δ-net). *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, and let $\delta > 0$ and $\mathcal{R} > 0$ be fixed. We say that a set $D \subset \mathbb{R}^d$ is a δ-net with radius $\mathcal{R}$ for A if for every $x \in \mathbb{R}^d$ such that $\operatorname{dist}(x, A) \leq \mathcal{R}$ there exists an $y \in D$ such that $\|x - y\| \leq \delta$.*

Theorem 51 (Deshpande, Rademacher, Vempala, and Wang, 2006). *Let (A, ε, B) be an input for Algorithm 2. Suppose that the cost of an optimal solution for linear j-subspace k-clustering for the instance A is B^*. If $B \geq B^*$, then the solution given by Algorithm 2 have cost at*

most $B^* + \varepsilon B$.

Lemma 52 (Deshpande, Rademacher, Vempala, and Wang, 2006)**.** *Let* $A \in \mathbb{R}^{d\times n}$ *be a matrix and let* $\delta > 0$ *be fixed. Suppose that* D *is a* δ*-net for* A *with radius* $\mathcal{R}$*. For every subspace* $W \subset \mathbb{R}^d$ *of dimension* j*, if*

$$\mathcal{R} \geq \sqrt{\operatorname{dist}^2(A, W)} + 2\delta j$$

then there exists a subspace F *of dimension* j *of* $\mathbb{R}^d$ *spanned by* j *points of* D *such that*

$$\operatorname{dist}^2(A, F) \leq \operatorname{dist}^2(A, W) + 4j^2 n\delta^2 + 4j\delta \sum_{a\in A} \operatorname{dist}(a, W). \tag{5.1}$$

Fact 53. *For any* $x \in \mathbb{R}^d$ *we have*

$$\|x - Rx\| = \|\pi_P(x) - R\pi_P(x)\|.$$

Fact 54. *For any* $x \in \mathbb{R}^d$ *and non-negative real* δ *such that* $\operatorname{dist}(x, P^\perp) \leq \delta$ *we have*

$$\|x - Rx\| \leq 2\delta.$$

Corollary 55. *Suppose that* $A \in \mathbb{R}^{d\times n}$ *is an instance of linear* j*-subspace* k*-clustering, the cost of an optimal solution is* B^**, and* $\varepsilon \in (0, 1)$ *is fixed. Suppose also that* $B > 0$ *is fixed, and that the cost of the solution returned by Algorithm 2 with input* (A, ε, B) *is* $\widetilde{B}$*. If*

$$\widetilde{B} > (1 + \varepsilon)B,$$

then $B < B^*$*. When this happens we say that* B *is a* certified lower bound *for the pair* (A, ε)*.*

Fact 56. *The while-loop on line 9 is executed at most* $\log_2(B/B^*) + 2$ *times.*

Fact 57. *The while-loop on line 18 is executed at most* $\log_2(3/\varepsilon) + 1$ *times.*

Lemma 58. *The solution* C *returned by Algorithm 3 with input* (A, ε) *is a* $(1 + \varepsilon)$*-approximation for* A*.*

Fact 59. *Let* $A \in \mathbb{R}^{d\times n}$ *be a matrix and let* $\varepsilon \in (0, 1/3)$ *be fixed. Suppose that* $\widetilde{A} \in \mathbb{R}^{d\times n}$ *is an* ε*-sketch for* A *for the linear* j*-subspace* k*-clustering problem and* $\alpha > 1$ *is a fixed real. If an algorithm* $\mathcal{A}$ *outputs an* α*-approximation for every input, then the output of* $\mathcal{A}(\widetilde{A})$ *will be a* $(1 + 3\varepsilon)\alpha$*-approximation for* A*.*

Appendix D

Quick reference for algorithms

Algorithm 1: A randomized dimension reduction scheme for (ℓ_2^2, C)-clustering when C is m-dimensional

Input: An instance $A \in \mathbb{R}^{d \times n}$ and a parameter $\varepsilon \in (0, 1)$.
Output: A matrix that with probability at least $1/2$ is an ε-sketch for A.

1 Let $s := \Theta(\varepsilon^{-2} m)$ be given by Theorem 45;
2 Let $r := \Theta(\varepsilon^{-2} s + s \log s)$ be given by Theorem 37;
3 Let $S \in \mathbb{R}^{r \times d}$ be a normalized Johnson-Lindenstrauss matrix;
4 Compute $\widetilde{A^T} = \pi_{A^T S^T}(A^T)_s$;
5 Return $\widetilde{A}$;

Algorithm 2: An approximation for linear j-subspace k-clustering

Input: A matrix $A \in \mathbb{R}^{d \times n}$, a real $\varepsilon \in (0, 1)$ and a real $B > 0$.
Output: A set $F_1, \ldots, F_k$ of k subspaces of dimension j.

1 Set

$$\delta := \frac{\varepsilon \sqrt{B}}{8jk\sqrt{n}};$$

2 Set

$$\mathcal{R} := \sqrt{B} + 2\delta j;$$

3 Let D be a δ-net with radius $\mathcal{R}$ for A that contains A;
4 For each choice of k subspaces $F_1, \ldots, F_k$ of dimension j, each one spanned by j points of D, compute $\operatorname{dist}^2(A, \bigcup_{i=1}^{k} F_i)$;
5 Return the subspaces $F_1, \ldots, F_k$ with lowest cost;

Algorithm 3: A $(1+\varepsilon)$-approximation for linear j-subspace k-clustering

Input: A matrix $A \in \mathbb{R}^{d \times n}$, a parameter $\varepsilon \in (0,1)$.
Output: A set $C := \{F_1, \ldots, F_k\}$ of k subspaces of dimension j.

Set B as any upper bound for the cost of an optimal solution;
/* A valid initialization value for B would be $\|A\|_F^2$ */
Set C as the solution returned by Algorithm 2 with input $(A, \varepsilon/3, B)$;
Set $\beta := \text{dist}^2(A, C)$;
if $\beta = 0$ **then**
    Return C;
/* Start the loop to find a lower bound for the cost B^* of an optimal solution */
Set $\widetilde{C} := C$;
Set $B_0 := B$;
Set $i := 0$;
while $\beta \leq (1+\varepsilon/3)B_i$ **do**
    $i := i + 1$;
    $C := \widetilde{C}$; // Note that $\text{dist}^2(A, C) \leq (1+\varepsilon/3)B_{i-1}$
    $B_i := B_{i-1}/2$;
    Update $\widetilde{C}$ with the solution returned by Algorithm 2 with input $(A, \varepsilon/3, B_i)$;
    Update β with the cost of $\widetilde{C}$;
/* Note that B_i is a lower bound for B^*. Now we start the binary search. */
Set $lo := B_i$;
Set $hi := B_{i-1}$;
Set $mid := (hi + lo)/2$;
while $hi - lo > (\varepsilon/3)B_i$ **do**
    Update $\widetilde{C}$ with the solution returned by Algorithm 2 with input $(A, \varepsilon/3, mid)$;
    Update β with $\text{dist}^2(A, \widetilde{C})$;
    if $\beta \leq (1+\varepsilon/3)mid$ **then**
        $hi := mid$;
        $C := \widetilde{C}$; // Note that $\text{dist}^2(A, C) \leq (1+\varepsilon/3)hi$
    else
        Update lo with mid;
    $mid := (hi + lo)/2$;
Return C;

References

[Achlioptas 2003] Dimitris Achlioptas. "Database-friendly random projections: Johnson-Lindenstrauss with binary coins". In: vol. 66. 4. Special issue on PODS 2001 (Santa Barbara, CA). 2003, pp. 671–687. doi: 10.1016/S0022-0000(03)00025-4. url: https://doi.org/10.1016/S0022-0000(03)00025-4 (cit. on pp. 9–11).

[Aloise, Deshpande, Hansen, and Popat 2009] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. "NP-hardness of Euclidean sum-of-squares clustering". In: *Machine Learning* 75.2 (2009), pp. 245–248. doi: 10.1007/s10994-009-5103-0 (cit. on p. 16).

[Alon, Gibbons, Matias, and Szegedy 2002] Noga Alon, Phillip B. Gibbons, Yossi Matias, and Mario Szegedy. "Tracking join and self-join sizes in limited storage". In: vol. 64. 3. Special issue on PODS 1999 (Philadelphia, PA). 2002, pp. 719–747. doi: 10.1006/jcss.2001.1813. url: https://doi.org/10.1006/jcss.2001.1813 (cit. on p. 13).

[Blum, Hopcroft, and Kannan 2020] Avrim Blum, John Hopcroft, and Ravi Kannan. *Foundations of Data Science*. 2020. isbn: 978-1108485067. url: https://www.cs.cornell.edu/jeh/book.pdf (cit. on pp. 7, 8).

[Cohen *et al.* 2015] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Mădălina Persu. "Dimensionality reduction for k-means clustering and low rank approximation". In: *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*. ACM, New York, 2015, pp. 163–172. doi: 10.1145/2746539.2746569 (cit. on pp. 33, 55).

[Deshpande, Rademacher, Vempala, and Wang 2006] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. "Matrix approximation and projective clustering via volume sampling". In: *Theory Comput.* 2 (2006), pp. 225–247. doi: 10.4086/toc.2006.v002a012. url: https://doi.org/10.4086/toc.2006.v002a012 (cit. on pp. v, vi, 41, 43, 69, 70).

[Feldman, Schmidt, and Sohler 2020] Dan Feldman, Melanie Schmidt, and Christian Sohler. "Turning big data into tiny data: constant-size coresets for k-means, PCA, and projective clustering". In: *SIAM J. Comput.* 49.3 (2020), pp. 601–657. issn: 0097-5397. doi: 10.1137/18M1209854. url: https://doi.org/10.1137/18M1209854 (cit. on pp. v, vi, 23, 33).

[Har-Peled, Indyk, and Motwani 2012] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. "Approximate nearest neighbor: towards removing the curse of dimensionality". In: *Theory Comput.* 8 (2012), pp. 321–350. DOI: 10.4086/toc.2012.v008a014. URL: https://doi.org/10.4086/toc.2012.v008a014 (cit. on pp. v, vi, 1).

[Horn and Johnson 2013] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Second. Cambridge University Press, Cambridge, 2013, pp. xviii+643. ISBN: 978-0-521-54823-6 (cit. on pp. 8, 44).

[Inaba, Katoh, and Imai 1994] Mary Inaba, Naoki Katoh, and Hiroshi Imai. "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering". In: *Proceedings of the tenth Annual Symposium on Computational Geometry*. 1994, pp. 332–339. DOI: 10.1145/177424.178042 (cit. on p. 16).

[Johnson and Lindenstrauss 1984] William B. Johnson and Joram Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space". In: *Conference in modern analysis and probability (New Haven, Conn., 1982)*. Vol. 26. Contemp. Math. Amer. Math. Soc., Providence, RI, 1984, pp. 189–206. DOI: 10.1090/conm/026/737400. URL: https://doi.org/10.1090/conm/026/737400 (cit. on p. 9).

[Makarychev, Makarychev, and Razenshteyn 2019] Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. "Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering". In: *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2019, pp. 1027–1038. DOI: 10.1145/3313276.3316350. URL: https://doi.org/10.1145/3313276.3316350 (cit. on p. 55).

[Matoušek 2013] Jiri Matoušek. *Lecture notes on metric embeddings*. Tech. rep. ETH Zürich, 2013 (cit. on p. 1).

[Megiddo and Tamir 1981] Nimrod Megiddo and Arie Tamir. "On the complexity of locating linear facilities in the plane". In: *Oper. Res. Lett.* 1.5 (1981), pp. 194–197. ISSN: 0167-6377. DOI: 10.1016/0167-6377(82)90039-6. URL: https://doi.org/10.1016/0167-6377(82)90039-6 (cit. on p. 17).

[Nelson 2020] Jelani Nelson. *Dimensionality Reduction in Euclidean Space*. Nov. 2020. DOI: 10.1090/noti2166. URL: https://www.ams.org/journals/notices/202010/rnoti-p1498.pdf (cit. on pp. v, vi).

[Pan, Chen, Zheng, *et al.* 1999] Victor Y Pan, Z Chen, Ailong Zheng, *et al.* "The complexity of the algebraic eigenproblem". In: *STOC '99: Proceedings of the thirty-first Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, New York, NY, United States, 1999, pp. 507–516. DOI: 10.1145/301250.301389 (cit. on p. 9).

[Pratap and Sen 2018] Rameshwar Pratap and Sandeep Sen. "Faster coreset construction for projective clustering *via* low-rank approximation". In: *Combinatorial algorithms*. Vol. 10979. Lecture Notes in Comput. Sci. Springer, Cham, 2018, pp. 336–348. doi: 10.1007/978-3-319-94667-2_28. url: https://doi.org/10.1007/978-3-319-94667-2_28 (cit. on pp. v, vi, 23, 33, 34, 69).

[Sarlós 2006] Tamás Sarlós. "Improved approximation algorithms for large matrices via random projections". In: *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE. 2006, pp. 143–152. doi: 10.1109/FOCS.2006.37 (cit. on pp. v, vi, 13, 23, 24, 26, 28, 33, 66–68).

[Shamos 1978] Michael Ian Shamos. "Computational geometry". PhD thesis. 1978 (cit. on p. 4).

[Trefethen and Bau 1997] Lloyd N. Trefethen and David Bau III. *Numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997, pp. xii+361. isbn: 0-89871-361-7. doi: 10.1137/1.9780898719574. url: https://doi.org/10.1137/1.9780898719574 (cit. on p. 9).

Index