Approximate k-means Clustering through Random Projections

by

Elena-Mădălina Persu

A.B., Harvard University (2013)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Signature redacted

Author ....
Department of Electrical Engineering and Computer Science
May 20, 2015

Signature redacted

Certified by..............................
Ankur Moitra
Assistant Professor of Applied Mathematics
Thesis Supervisor

Signature redacted

Accepted by..............................
Leslie A. Kolodziesjki
Chair, Department Committee on Graduate Students

Approximate k-means Clustering through Random Projections

by

Elena-Mădălina Persu

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

Using random row projections, we show how to approximate a data matrix $\mathbf{A}$ with a much smaller sketch $\tilde{\mathbf{A}}$ that can be used to solve a general class of *constrained k-rank approximation* problems to within $(1+\epsilon)$ error. Importantly, this class of problems includes k-means clustering. By reducing data points to just $O(k)$ dimensions, our methods generically accelerate any exact, approximate, or heuristic algorithm for these ubiquitous problems.

For k-means dimensionality reduction, we provide $(1+\epsilon)$ relative error results for random row projections which improve on the $(2+\epsilon)$ prior known constant factor approximation associated with this sketching technique, while preserving the number of dimensions. For k-means clustering, we show how to achieve a $(9+\epsilon)$ approximation by Johnson-Lindenstrauss projecting data points to just $O(\log k/\epsilon^2)$ dimensions. This gives the first result that leverages the specific structure of k-means to achieve dimension independent of input size and sublinear in k.

Thesis Supervisor: Ankur Moitra
Title: Assistant Professor of Applied Mathematics

Acknowledgments

First and foremost I would like to thank my adviser, Ankur Moitra, whose support has been invaluable during my first years in graduate school at MIT. I am very grateful for your excelent guidance, caring, engagement, patience, and for providing me with the right set of tools throughout this work. I am excited for the road that lies ahead.

This thesis is based on a joint collaboration with Michael Cohen, Christopher Musco, Cameron Musco and Sam Elder [CEM$^+$14]. I am very thankful to work with such talented individuals.

Last but definitely not least, I want to express my deepest gratitude to my beloved parents, Camelia and Ion. I owe much of my academic success to their continuous encouragement and support.

Contents

List of Tables

Chapter 1

Introduction

Dimensionality reduction has received considerable attention in the study of fast linear algebra algorithms. The goal is to approximate a large matrix $\mathbf{A}$ with a much smaller *sketch* $\tilde{\mathbf{A}}$ such that solving a given problem on $\tilde{\mathbf{A}}$ gives a good approximation to the solution on $\mathbf{A}$. This can lead to faster runtimes, reduced memory usage, or decreased distributed communication. Methods such as random sampling and Johnson-Lindenstrauss projection have been applied to a variety of problems including matrix multiplication, regression, and low rank approximation [HMT11, Mah11].

Similar tools have been used for accelerating k-means clustering. While exact k-means clustering is NP-hard [ADHP09, MNV09], effective heuristics and provably good approximation algorithms are known [Llo82, KMN$^+$02, KSS04, AV07, HPK07]. Dimensionality reduction seeks to generically accelerate any of these algorithms by reducing the dimension of the data points being clustered. In this thesis, given a data matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, where the rows of $\mathbf{A}$ are d-dimensional data points, the goal is to produce a sketch $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times d'}$, where $d' << d$. Typically, since k is considered to be very small with respect to input size, we seek to achieve d' which is a function of just k and an error parameter, and not of n or d.

We start by noting that the k-means clustering problem can be viewed as a special case of a general *constrained k-rank approximation* problem [DFK$^+$04], which also includes problems related to sparse and nonnegative PCA [PDK13, YZ13, APD14]. Then, following the coreset definitions of [FSS13], we introduce the concept of a *projection-cost preserving sketch*, an approximation where the sum of squared distances of $\tilde{\mathbf{A}}$'s columns from any k-dimensional subspace (plus a fixed constant independent of the subspace) is multiplicatively close to that of $\mathbf{A}$. This ensures that the cost of any k-rank projection of $\mathbf{A}$ is well approximated by $\tilde{\mathbf{A}}$ and thus, we can solve the general constrained k-rank approximation problem approximately for $\mathbf{A}$ using $\tilde{\mathbf{A}}$.

Next, we give a simple and efficient approach for obtaining projection-cost preserving sketches with $(1+\epsilon)$ relative error. All of these techniques simply require multiplying by a random projection matrix. These methods have well developed implementations, are robust, and can be accelerated for sparse or otherwise structured data. As such, we do not focus heavily on specific implementations or runtime analysis. We do emphasize that our proofs are amenable to approximation and acceleration in the underlying sketching techniques – for example, it is possible to use sparse Johnson-Lindenstrauss embeddings.

In addition to the applications in this paper, we hope that projection-cost preserving sketches will be useful in developing future randomized matrix algorithms. They relax the guarantee of *subspace embeddings*, which have received significant attention in recent years [Sar06, CW13]. Subspace embedding sketches require that $\|\mathbf{x}\tilde{\mathbf{A}}\| \approx \|\mathbf{x}\mathbf{A}\|$ simultaneously for all $\mathbf{x}$, which in particular implies that $\tilde{\mathbf{A}}$ preserves the cost of *any* column projection of $\mathbf{A}$[1]. However, in general such an $\tilde{\mathbf{A}}$ will require at least $\Theta(\text{rank}(\mathbf{A}))$ columns. On the other hand, our projection-cost preserving sketches only work for projections with rank at most k, but only require $O(k)$ columns.

1.1 Summary of Results

In Table 1.1 we summarize our dimensionality reduction results, showing how projection-cost preserving sketches were obtained. We note how many dimensions (columns) are required for a sketch $\tilde{\mathbf{A}}$ that achieves $(1+\epsilon)$ error. We compare to prior work, most of which focuses on constructing sketches for k-means clustering, but applies to general constrained k-rank

[1] $\|(\mathbf{I}-\mathbf{P})\mathbf{A}\|_F \approx \|(\mathbf{I}-\mathbf{P})\tilde{\mathbf{A}}\|_F$ for any projection matrix $\mathbf{P}$.

approximation as well.

	Previous Work			**Our Results**		
Technique	**Reference**	**Dimensions**	**Error**	**Theorem**	**Dimensions**	**Error**
Random Projection	[BZD10]	$O(k/\epsilon^2)$	$2+\epsilon$	Thm 6	$O(k/\epsilon^2)$	$1+\epsilon$
				Thm 7	$O(\log k/\epsilon^2)$	$9+\epsilon$ †

Table 1.1: Summary of results.

Our random projection results are based on a unified proof technique that relies on a reduction to a spectral approximation problem. The approach allows us to tighten and generalize a fruitful line of work in [BMD09, BZD10, BZMD15, BMI13], which were the first papers to address dimensionality reduction for k-means using random projection. They inspired our general proof technique.

Specifically, we show that a $(1+\epsilon)$ error projection-cost preserving sketch can be obtained by randomly projecting $\mathbf{A}$'s rows to $O(k/\epsilon^2)$ dimensions – i.e., multiplying on the right by a Johnson-Lindenstrauss matrix with $O(k/\epsilon^2)$ columns. Our results improve on constant factor bounds in [BMD09, BZD10, BMI13, BZMD15] which shows that an $O(k/\epsilon^2)$ dimension random projections can give $(2+\epsilon)$ error for k-means clustering.

Finally, for general constrained k-rank approximation, it is not possible to reduce to dimension below $\Theta(k)$. However, we conclude by showing that it *is possible* to do better for k-means clustering by leveraging the problem's specific structure. Specifically, randomly projecting to $O(\log k/\epsilon^2)$ dimensions is sufficient to obtain a $(9+\epsilon)$ approximation to the optimal clustering. This gives the first k-means sketch with dimension independent of the input size and sublinear in k. It is simple to show via the standard Johnson-Lindenstrauss lemma that $O(\log n/\epsilon^2)$ dimension projections yield $(1+\epsilon)$ error, also specifically for k-means [BZMD15]. Our results offers significantly reduced dimension and we are interested in knowing whether our $(9+\epsilon)$ error bound can be improved.

† k-means clustering only. ‡ k-rank approximation only.

Chapter 2

Preliminaries

2.1 Linear Algebra Basics

For any n and d, consider a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$. Let $r = \text{rank}(\mathbf{A})$. Using a singular value decomposition, we can write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n \times r}$ and $\mathbf{V} \in \mathbb{R}^{d \times r}$ have orthogonal columns (the left and right singular vectors of $\mathbf{A}$), and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is a positive diagonal matrix containing the singular values of $\mathbf{A}$: $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$. The pseudoinverse of $\mathbf{A}$ is given by $\mathbf{A}^+ = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top$.

$\mathbf{A}$'s squared Frobenius norm is given by $\|\mathbf{A}\|_F^2 = \sum_{i,j} \mathbf{A}_{i,j}^2 = \text{tr}(\mathbf{A}\mathbf{A}^\top) = \sum_i \sigma_i^2$. Its spectral norm is given by $\|\mathbf{A}\|_2 = \sigma_1$. Let $\mathbf{\Sigma}_k$ be $\mathbf{\Sigma}$ with all but its largest k singular values zeroed out. Let $\mathbf{U}_k$ and $\mathbf{V}_k$ be $\mathbf{U}$ and $\mathbf{V}$ with all but their first k columns zeroed out. For any $k \leq r$, $\mathbf{A}_k = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^\top = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\top$ is the closest rank k approximation to $\mathbf{A}$ for any unitarily invariant norm, including the Frobenius norm and spectral norm [Mir60]. That is,

$$\|\mathbf{A} - \mathbf{A}_k\|_F = \min_{\mathbf{B}|\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F \text{ and}$$

$$\|\mathbf{A} - \mathbf{A}_k\|_2 = \min_{\mathbf{B}|\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_2.$$

We often work with the remainder matrix $\mathbf{A} - \mathbf{A}_k$ and label it $\mathbf{A}_{r \setminus k}$.

For any two matrices $\mathbf{M}$ and $\mathbf{N}$, $\|\mathbf{M}\mathbf{N}\|_F \leq \|\mathbf{M}\|_F\|\mathbf{N}\|_2$ and $\|\mathbf{M}\mathbf{N}\|_F \leq \|\mathbf{N}\|_F\|\mathbf{M}\|_2$. This property is known as *spectral submultiplicativity.* It holds because multiplying by a matrix can scale each row or column, and hence the Frobenius norm, by at most the matrix's spectral norm. Submultiplicativity implies that multiplying by an orthogonal projection matrix (which only has singular values of 0 or 1) can only decrease Frobenius norm, a fact that we will use repeatedly.

If $\mathbf{M}$ and $\mathbf{N}$ have the same dimensions and $\mathbf{M}\mathbf{N}^\top = \mathbf{0}$ then $\|\mathbf{M}+\mathbf{N}\|_F^2 = \|\mathbf{M}\|_F^2 + \|\mathbf{N}\|_F^2$. This matrix Pythagorean theorem follows from the fact that $\|\mathbf{M}+\mathbf{N}\|_F^2 = \text{tr}((\mathbf{M}+\mathbf{N})(\mathbf{M}+\mathbf{N})^\top)$. As an example, note that since $\mathbf{A}_k$ is an orthogonal projection of $\mathbf{A}$ and $\mathbf{A}_{r\setminus k}$ is its residual, $\mathbf{A}_k\mathbf{A}_{r\setminus k}^\top = \mathbf{0}$. Thus, $\|\mathbf{A}_k\|_F^2 + \|\mathbf{A}_{r\setminus k}\|_F^2 = \|\mathbf{A}_k + \mathbf{A}_{r\setminus k}\|_F^2 = \|\mathbf{A}\|_F^2$.

For any two symmetric matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{n\times n}$, $\mathbf{M} \preceq \mathbf{N}$ indicates that $\mathbf{N} - \mathbf{M}$ is positive semidefinite – that is, it has all nonnegative eigenvalues and $\mathbf{x}^\top(\mathbf{N}-\mathbf{M})\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. We use $\lambda_i(\mathbf{M})$ to denote the i^{th} largest eigenvalue of $\mathbf{M}$ *in absolute value.*

Finally, we often use $\mathbf{P}$ to denote an orthogonal projection matrix, which is any matrix that can be written as $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ where $\mathbf{Q}$ is a matrix with orthonormal columns. Multiplying a matrix by $\mathbf{P}$ on the left will project its columns to the column span of $\mathbf{Q}$. If $\mathbf{Q}$ has just k columns, the projection has rank k. Note that $\mathbf{B}^* = \mathbf{P}\mathbf{A}$ minimizes $\|\mathbf{A} - \mathbf{B}\|_F$ amongst all matrices $\mathbf{B}$ whose columns lie in the column span of $\mathbf{Q}$ [Woo14].

2.2 Constrained Low Rank Approximation

To develop sketching algorithms for k-means clustering, we show that the problem reduces to a general constrained low rank approximation objective. Consider a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and any set S of rank k orthogonal projection matrices in $\mathbb{R}^{n\times n}$. We want to find

$$\mathbf{P}^* = \operatorname*{arg\,min}_{\mathbf{P}\in S} \|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2. \tag{2.1}$$

We often write $\mathbf{Y} = \mathbf{I}_{n\times n} - \mathbf{P}$ and refer to $\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2 = \|\mathbf{Y}\mathbf{A}\|_F^2$ as the *cost* of the projection $\mathbf{P}$.

When S is the set of all rank k orthogonal projections, this problem is equivalent to finding the optimal rank k approximation for $\mathbf{A}$, and is solved by computing $\mathbf{U}_k$ using an SVD algorithm and setting $\mathbf{P}^* = \mathbf{U}_k\mathbf{U}_k^\top$. In this case, the cost of the optimal projection is $\|\mathbf{A} - \mathbf{U}_k\mathbf{U}_k^\top\mathbf{A}\|_F^2 = \|\mathbf{A}_{r\setminus k}\|_F^2$. As the optimum cost in the unconstrained case, $\|\mathbf{A}_{r\setminus k}\|_F^2$ is a universal lower bound on $\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2$.

2.3 k-Means Clustering as Constrained Low Rank Approximation

Formally, k-means clustering asks us to partition n vectors in $\mathbb{R}^d$, $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$, into k cluster sets, $\{C_1, \dots, C_k\}$. Let $\boldsymbol{\mu}_i$ be the centroid of the vectors in C_i. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a data matrix containing our vectors as rows and let $C(\mathbf{a}_j)$ be the set that vector $\mathbf{a}_j$ is assigned to. The goal is to minimize the objective function

$$\sum_{i=1}^{k} \sum_{\mathbf{a}_j\in C_i} \|\mathbf{a}_j - \boldsymbol{\mu}_i\|_2^2 = \sum_{j=1}^{n} \|\mathbf{a}_j - \boldsymbol{\mu}_{C(\mathbf{a}_j)}\|_2^2.$$

To see that k-means clustering is an instance of general constrained low rank approximation, we rely on a linear algebraic formulation of the k-means objective that has been used critically in prior work on dimensionality reduction for the problem (see e.g. [BMD09]).

For a clustering $C = \{C_1, \dots, C_k\}$, let $\mathbf{X}_C \in \mathbb{R}^{n\times k}$ be the *cluster indicator matrix*, with $\mathbf{X}_C(i,j) = 1/\sqrt{|C_j|}$ if $\mathbf{a}_i$ is assigned to C_j. $\mathbf{X}_C(i,j) = 0$ otherwise. Thus, $\mathbf{X}_C\mathbf{X}_C^\top\mathbf{A}$ has its i^{th} row equal to $\boldsymbol{\mu}_{C(\mathbf{a}_i)}$, the center of $\mathbf{a}_i$'s assigned cluster. So we can express the k-means objective function as:

$$\|\mathbf{A} - \mathbf{X}_C\mathbf{X}_C^\top\mathbf{A}\|_F^2 = \sum_{j=1}^{n} \|\mathbf{a}_j - \boldsymbol{\mu}_{C(\mathbf{a}_j)}\|_2^2.$$

By construction, the columns of $\mathbf{X}_C$ have disjoint supports and so are orthonormal vectors. Thus $\mathbf{X}_C\mathbf{X}_C^\top$ is an orthogonal projection matrix with rank k, and k-means is just the constrained low rank approximation problem of (2.1) with S as the set of all possible cluster projection matrices $\mathbf{X}_C\mathbf{X}_C^\top$.

While the goal of k-means is to well approximate each *row* of $\mathbf{A}$ with its cluster center, this formulation shows that the problem actually amounts to finding an optimal rank k subspace for approximating the *columns* of $\mathbf{A}$. The choice of subspace is constrained because it must be spanned by the columns of a cluster indicator matrix.

Chapter 3

Projection-Cost Preserving Sketches

We hope to find an approximately optimal constrained low rank approximation (2.1) for $\mathbf{A}$ by optimizing $\mathbf{P}$ (either exactly or approximately) over a sketch $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times d'}$ with $d' \ll d$. This approach will certainly work if the cost $\|\tilde{\mathbf{A}} - \mathbf{P}\tilde{\mathbf{A}}\|_F^2$ approximates the cost of $\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2$ for *any* $\mathbf{P} \in S$. An even stronger requirement is that $\tilde{\mathbf{A}}$ approximates projection-cost for all rank k projections (of which S is a subset). We call such an $\tilde{\mathbf{A}}$ a *projection-cost preserving sketch.* This definition is equivalent to the (k, ϵ)-coresets of [FSS13] (see their Definition 2).

Definition 1 (Rank k Projection-Cost Preserving Sketch with Two-sided Error). $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times d'}$ *is a rank k projection-cost preserving sketch of $\mathbf{A} \in \mathbb{R}^{n \times d}$ with error $0 \leq \epsilon < 1$ if, for all rank k orthogonal projection matrices $\mathbf{P} \in \mathbb{R}^{n \times n}$,*

$$(1 - \epsilon)\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2 \leq \|\tilde{\mathbf{A}} - \mathbf{P}\tilde{\mathbf{A}}\|_F^2 + c \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2,$$

for some fixed non-negative constant c that may depend on $\mathbf{A}$ and $\tilde{\mathbf{A}}$ but is independent of $\mathbf{P}$.

3.1 Application to Constrained Low Rank Approximation

It is straightforward to show that a projection-cost preserving sketch is sufficient for approximately optimizing (2.1), our constrained low rank approximation problem.

Lemma 2 (Low Rank Approximation via Projection-Cost Preserving Sketches). *For any* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and any set* S *of rank* k *orthogonal projections, let* $\mathbf{P}^* = \arg\min_{\mathbf{P}\in S} \|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2$. *Accordingly, for any* $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times d'}$, *let* $\tilde{\mathbf{P}}^* = \arg\min_{\mathbf{P}\in S} \|\tilde{\mathbf{A}} - \mathbf{P}\tilde{\mathbf{A}}\|_F^2$. *If* $\tilde{\mathbf{A}}$ *is a rank* k *projection-cost preserving sketch for* $\mathbf{A}$ *with error* ϵ, *then for any* $\gamma \geq 1$, *if* $\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}\tilde{\mathbf{A}}\|_F^2 \leq \gamma \|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}^*\tilde{\mathbf{A}}\|_F^2$

$$\|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F^2 \leq \frac{(1+\epsilon)}{(1-\epsilon)} \cdot \gamma \|\mathbf{A} - \mathbf{P}^*\mathbf{A}\|_F^2.$$

That is, if $\tilde{\mathbf{P}}$ is an (approximately) optimal solution for $\tilde{\mathbf{A}}$, then it is also approximately optimal for $\mathbf{A}$.

Proof. By optimality of $\tilde{\mathbf{P}}^*$ for $\tilde{\mathbf{A}}$, $\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}^*\tilde{\mathbf{A}}\|_F^2 \leq \|\tilde{\mathbf{A}} - \mathbf{P}^*\tilde{\mathbf{A}}\|_F^2$ and thus,

$$\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}\tilde{\mathbf{A}}\|_F^2 \leq \gamma \|\tilde{\mathbf{A}} - \mathbf{P}^*\tilde{\mathbf{A}}\|_F^2. \tag{3.1}$$

Furthermore, since $\tilde{\mathbf{A}}$ is projection-cost preserving, the following two inequalities hold:

$$\|\tilde{\mathbf{A}} - \mathbf{P}^*\tilde{\mathbf{A}}\|_F^2 \leq (1+\epsilon)\|\mathbf{A} - \mathbf{P}^*\mathbf{A}\|_F^2 - c, \tag{3.2}$$

$$\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}\tilde{\mathbf{A}}\|_F^2 \geq (1-\epsilon)\|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F^2 - c. \tag{3.3}$$

Combining (3.1),(3.2), and (3.3), we see that:

$$\begin{aligned} (1-\epsilon)\|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F^2 - c &\leq (1+\epsilon) \cdot \gamma \|\mathbf{A} - \mathbf{P}^*\mathbf{A}\|_F^2 - \gamma c \\ \|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F^2 &\leq \frac{(1+\epsilon)}{(1-\epsilon)} \cdot \gamma \|\mathbf{A} - \mathbf{P}^*\mathbf{A}\|_F^2, \end{aligned}$$

where the final step is simply the consequence of $c \geq 0$ and $\gamma \geq 1$. □

For any $0 \leq \epsilon' < 1$, to achieve a $(1+\epsilon')\gamma$ approximation with Lemma 2, we just need to set $\epsilon = \frac{\epsilon'}{2+\epsilon'} \geq \frac{\epsilon'}{3}$.

3.2 Sufficient Conditions

With Lemma 2 in place, we seek to characterize what sort of sketch suffices for rank k projection-cost preservation. We discuss sufficient conditions that will be used throughout the remainder of the paper. Before giving the full technical analysis, it is helpful to overview our general approach and highlight connections to prior work.

Using the notation $\mathbf{Y} = \mathbf{I}_{n \times n} - \mathbf{P}$, we can rewrite the guarantees for Definition 1 as:

$$(1-\epsilon)\operatorname{tr}(\mathbf{Y}\mathbf{A}\mathbf{A}^\top\mathbf{Y}) \leq \operatorname{tr}(\mathbf{Y}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top\mathbf{Y}) + c \leq (1+\epsilon)\operatorname{tr}(\mathbf{Y}\mathbf{A}\mathbf{A}^\top\mathbf{Y}). \tag{3.4}$$

Thus, in approximating $\mathbf{A}$ with $\tilde{\mathbf{A}}$, we are really attempting to approximate $\mathbf{A}\mathbf{A}^\top$.

Furthermore, sketching through random row projections is linear – i.e. we can always write $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R}$. Suppose our sketching dimension is $m = O(k)$. For a Johnson-Lindenstrauss random projection, $\mathbf{R}$ is a $d \times m$ random matrix. So, our goal is to show:

$$\operatorname{tr}(\mathbf{Y}\mathbf{A}\mathbf{A}^\top\mathbf{Y}) \approx \operatorname{tr}(\mathbf{Y}\mathbf{A}\mathbf{R}\mathbf{R}^\top\mathbf{A}^\top\mathbf{Y}) + c.$$

A common trend in prior work has been to attack this analysis by splitting $\mathbf{A}$ into separate orthogonal components [DFK$^+$04, BZMD15]. In particular, previous results note that $\mathbf{A} = \mathbf{A}_k + \mathbf{A}_{r\setminus k}$ and implicitly compare

$$\begin{aligned}\operatorname{tr}(\mathbf{Y}\mathbf{A}\mathbf{A}^\top\mathbf{Y}) &= \operatorname{tr}(\mathbf{Y}\mathbf{A}_k\mathbf{A}_k^\top\mathbf{Y}) + \operatorname{tr}(\mathbf{Y}\mathbf{A}_{r\setminus k}\mathbf{A}_{r\setminus k}^\top\mathbf{Y}) + \operatorname{tr}(\mathbf{Y}\mathbf{A}_k\mathbf{A}_{r\setminus k}^\top\mathbf{Y}) + \operatorname{tr}(\mathbf{Y}\mathbf{A}_{r\setminus k}\mathbf{A}_k^\top\mathbf{Y}) \\ &= \operatorname{tr}(\mathbf{Y}\mathbf{A}_k\mathbf{A}_k^\top\mathbf{Y}) + \operatorname{tr}(\mathbf{Y}\mathbf{A}_{r\setminus k}\mathbf{A}_{r\setminus k}^\top\mathbf{Y}) + 0 + 0,\end{aligned}$$

to

$$\operatorname{tr}(\mathbf{YARR}^\top\mathbf{A}^\top\mathbf{Y}) = \operatorname{tr}(\mathbf{YA}_k\mathbf{RR}^\top\mathbf{A}_k^\top\mathbf{Y}) + \operatorname{tr}(\mathbf{YA}_{r\setminus k}\mathbf{RR}^\top\mathbf{A}_{r\setminus k}^\top\mathbf{Y})$$
$$+ \operatorname{tr}(\mathbf{YA}_k\mathbf{RR}^\top\mathbf{A}_{r\setminus k}^\top\mathbf{Y}) + \operatorname{tr}(\mathbf{YA}_{r\setminus k}\mathbf{RR}^\top\mathbf{A}_k^\top\mathbf{Y}).$$

We adopt this same general technique, but make the comparison more explicit and analyze the difference between each of the four terms separately. In Lemma 3, the allowable error in each term will correspond to $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, and $\mathbf{E}_4$, respectively.

Additionally, our analysis generalizes the approach by splitting $\mathbf{A}$ into a wider variety of orthogonal pairs. Our random projection results split $\mathbf{A} = \mathbf{A}_{2k} + \mathbf{A}_{r\setminus 2k}$, while our $O(\log k)$ result for k-means clustering splits $\mathbf{A} = \mathbf{P}^*\mathbf{A} + (\mathbf{I} - \mathbf{P}^*)\mathbf{A}$ where $\mathbf{P}^*$ is the optimal k-means projection matrix for $\mathbf{A}$.

3.3 Characterization of Projection-Cost Preserving Sketches

Next we formally analyze what sort of error, $\mathbf{E} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top - \mathbf{A}\mathbf{A}^\top$, is permissible for a projection-cost preserving sketch.

Lemma 3. $\tilde{\mathbf{A}}$ *is a rank k projection-cost preserving sketch with two-sided error ϵ (i.e. satisfies Definition 1) as long as we can write $\tilde{\mathbf{C}} = \mathbf{C} + \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 + \mathbf{E}_4$ where*

1. *$\mathbf{E}_1$ is symmetric and $-\epsilon_1\mathbf{C} \preceq \mathbf{E}_1 \preceq \epsilon_1\mathbf{C}$*

2. *$\mathbf{E}_2$ is symmetric, $\sum_{i=1}^k |\lambda_i(\mathbf{E}_2)| \leq \epsilon_2\|\mathbf{A}_{r\setminus k}\|_F^2$, and $\operatorname{tr}(\mathbf{E}_2) \leq \epsilon_2'\|\mathbf{A}_{r\setminus k}\|_F^2$*

3. *The columns of $\mathbf{E}_3$ fall in the column span of $\mathbf{C}$ and $\operatorname{tr}(\mathbf{E}_3^\top\mathbf{C}^+\mathbf{E}_3) \leq \epsilon_3^2\|\mathbf{A}_{r\setminus k}\|_F^2$*

4. *The rows of $\mathbf{E}_4$ fall in the row span of $\mathbf{C}$ and $\operatorname{tr}(\mathbf{E}_4\mathbf{C}^+\mathbf{E}_4^\top) \leq \epsilon_4^2\|\mathbf{A}_{r\setminus k}\|_F^2$*

and $\epsilon_1 + \epsilon_2 + \epsilon_2' + \epsilon_3 + \epsilon_4 = \epsilon$. Specifically, referring to the guarantee in Equation 3.4, we show that for any rank k orthogonal projection $\mathbf{P}$ and $\mathbf{Y} = \mathbf{I} - \mathbf{P}$,

$$(1-\epsilon)\operatorname{tr}(\mathbf{YCY}) \leq \operatorname{tr}(\mathbf{Y}\tilde{\mathbf{C}}\mathbf{Y}) - \min\{0, \operatorname{tr}(\mathbf{E}_2)\} \leq (1+\epsilon)\operatorname{tr}(\mathbf{YCY}).$$

To get intuition about this lemma note that since $\mathbf{P}$ is a rank k projection, any projection dependent error at worst depends on the largest k eigenvalues of our error matrix. Since the cost of any rank k projection is at least $\|\mathbf{A}_{r\setminus k}\|_F^2$ we need restrictions having $\epsilon\|\mathbf{A}_{r\setminus k}\|_F^2$ as upper bounds on the traces of the different error types to achieve relative error approximation.

Proof. By linearity of the trace note that

$$\operatorname{tr}(\mathbf{Y}\tilde{\mathbf{C}}\mathbf{Y}) = \operatorname{tr}(\mathbf{YCY}) + \operatorname{tr}(\mathbf{YE}_1\mathbf{Y}) + \operatorname{tr}(\mathbf{YE}_2\mathbf{Y}) + \operatorname{tr}(\mathbf{YE}_3\mathbf{Y}) + \operatorname{tr}(\mathbf{YE}_4\mathbf{Y}). \tag{3.5}$$

We handle each error term separately. Starting with $\mathbf{E}_1$, note that $\operatorname{tr}(\mathbf{YE}_1\mathbf{Y}) = \sum_{i=1}^n \mathbf{y}_i^\top \mathbf{E}_1 \mathbf{y}_i$ where $\mathbf{y}_i$ is the i^{th} column (equivalently row) of $\mathbf{Y}$. So, by the spectral bounds on $\mathbf{E}_1$

$$-\epsilon_1 \operatorname{tr}(\mathbf{YCY}) \leq \operatorname{tr}(\mathbf{YE}_1\mathbf{Y}) \leq \epsilon_1 \operatorname{tr}(\mathbf{YCY}). \tag{3.6}$$

For $\mathbf{E}_2$, note that by the cyclic property of the trace and the fact that $\mathbf{Y}^2 = \mathbf{Y}$ since $\mathbf{Y}$ is a projection matrix we have

$$\begin{aligned}\operatorname{tr}(\mathbf{YE}_2\mathbf{Y}) &= \operatorname{tr}(\mathbf{Y}^2\mathbf{E}_2) \\ &= \operatorname{tr}(\mathbf{YE}_2) \\ &= \operatorname{tr}(\mathbf{E}_2) - \operatorname{tr}(\mathbf{PE}_2).\end{aligned} \tag{3.7}$$

Since $\mathbf{E}_2$ is symmetric, let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be the eigenvectors of $\mathbf{E}_2$, and write

$$\begin{aligned}\mathbf{E}_2 &= \sum_{i=1}^r \lambda_i(\mathbf{E}_2)\mathbf{v}_i\mathbf{v}_i^\top \text{ and thus} \\ \operatorname{tr}(\mathbf{PE}_2) &= \sum_{i=1}^r \lambda_i(\mathbf{E}_2) \operatorname{tr}(\mathbf{Pv}_i\mathbf{v}_i^\top).\end{aligned} \tag{3.8}$$

Note that

$$|\operatorname{tr}(\mathbf{P}\mathbf{E}_2)| = \left|\sum_{i=1}^{r} \lambda_i(\mathbf{E}_2)\operatorname{tr}(\mathbf{P}\mathbf{v}_i\mathbf{v}_i^\top)\right| \leq \sum_{i=1}^{r} |\lambda_i(\mathbf{E}_2)|\operatorname{tr}(\mathbf{P}\mathbf{v}_i\mathbf{v}_i^\top).$$

For all i, $0 \leq \operatorname{tr}(\mathbf{P}\mathbf{v}_i\mathbf{v}_i^\top) \leq \|\mathbf{v}_i\|_2^2 \leq 1$ and $\sum_{i=1}^{r}\operatorname{tr}(\mathbf{P}\mathbf{v}_i\mathbf{v}_i^\top) \leq \operatorname{tr}(\mathbf{P}) = k$. Thus, $\sum_{i=1}^{r}|\lambda_i(\mathbf{E}_2)|\operatorname{tr}(\mathbf{P}\mathbf{v}_i\mathbf{v}_i^\top)$ is minimized when $\operatorname{tr}(\mathbf{P}\mathbf{v}_i\mathbf{v}_i^\top) = 1$ for $\mathbf{v}_1, \ldots, \mathbf{v}_k$, the eigenvectors corresponding to $\mathbf{E}_2$'s largest magnitude eigenvalues. Combined with our requirement that $\sum_{i=1}^{k}|\lambda_i(\mathbf{E}_2)| \leq \epsilon_2\|\mathbf{A}_{r\setminus k}\|_F^2$, we see that $|\operatorname{tr}(\mathbf{P}\mathbf{E}_2)| \leq \epsilon_2\|\mathbf{A}_{r\setminus k}\|_F^2$. Accordingly,

$$\begin{aligned}
\operatorname{tr}(\mathbf{E}_2) - \epsilon_2\|\mathbf{A}_{r\setminus k}\|_F^2 &\leq \operatorname{tr}(\mathbf{Y}\mathbf{E}_2\mathbf{Y}) \leq \operatorname{tr}(\mathbf{E}_2) + \epsilon_2\|\mathbf{A}_{r\setminus k}\|_F^2 \\
\min\{0, \operatorname{tr}(\mathbf{E}_2)\} - \epsilon_2\|\mathbf{A}_{r\setminus k}\|_F^2 &\leq \operatorname{tr}(\mathbf{Y}\mathbf{E}_2\mathbf{Y}) \leq \min\{0, \operatorname{tr}(\mathbf{E}_2)\} + (\epsilon_2 + \epsilon_2')\|\mathbf{A}_{r\setminus k}\|_F^2 \\
\min\{0, \operatorname{tr}(\mathbf{E}_2)\} - (\epsilon_2 + \epsilon_2')\operatorname{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y}) &\leq \operatorname{tr}(\mathbf{Y}\mathbf{E}_2\mathbf{Y}) \leq \min\{0, \operatorname{tr}(\mathbf{E}_2)\} + (\epsilon_2 + \epsilon_2')\operatorname{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y}).
\end{aligned} \tag{3.9}$$

The second step follows from the trace bound on $\mathbf{E}_2$. The last step follows from recalling that $\|\mathbf{A}_{r\setminus k}\|_F^2$ is a universal lower bound on $\operatorname{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y})$.

Next, we note that, since $\mathbf{E}_3$'s columns fall in the column span of $\mathbf{C}$, $\mathbf{C}\mathbf{C}^+\mathbf{E}_3 = \mathbf{E}_3$. Thus,

$$\operatorname{tr}(\mathbf{Y}\mathbf{E}_3\mathbf{Y}) = \operatorname{tr}(\mathbf{Y}\mathbf{E}_3) = \operatorname{tr}\left((\mathbf{Y}\mathbf{C})\mathbf{C}^+(\mathbf{E}_3)\right).$$

$\langle\mathbf{M}, \mathbf{N}\rangle = \operatorname{tr}(\mathbf{M}\mathbf{C}^+\mathbf{N}^\top)$ is a semi-inner product since $\mathbf{C} = \mathbf{A}\mathbf{A}^\top$, and therefore also $\mathbf{C}^+$, is positive semidefinite. Thus, by the Cauchy-Schwarz inequality,

$$\left|\operatorname{tr}\left((\mathbf{Y}\mathbf{C})\mathbf{C}^+(\mathbf{E}_3)\right)\right| \leq \sqrt{\operatorname{tr}(\mathbf{Y}\mathbf{C}\mathbf{C}^+\mathbf{C}\mathbf{Y}) \cdot \operatorname{tr}(\mathbf{E}_3^\top\mathbf{C}^+\mathbf{E}_3)} \leq \epsilon_3\|\mathbf{A}_{r\setminus k}\|_F \cdot \sqrt{\operatorname{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y})}.$$

Since $\sqrt{\operatorname{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y})} \geq \|\mathbf{A}_{r\setminus k}\|_F$, we conclude that

$$|\operatorname{tr}(\mathbf{Y}\mathbf{E}_3\mathbf{Y})| \leq \epsilon_3 \cdot \operatorname{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y}). \tag{3.10}$$

For $\mathbf{E}_4$ we make a symmetric argument.

$$\left|\mathrm{tr}(\mathbf{Y}\mathbf{E}_4\mathbf{Y})\right| = \left|\mathrm{tr}\left((\mathbf{E}_4)\mathbf{C}^{+}(\mathbf{C}\mathbf{Y})\right)\right| \leq \sqrt{\mathrm{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y}) \cdot \mathrm{tr}(\mathbf{E}_4\mathbf{C}^{+}\mathbf{E}_4^{\top})} \leq \epsilon_4 \cdot \mathrm{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y}). \quad (3.11)$$

Finally, combining equations (3.5), (3.6), (3.9), (3.10), and (3.11) and recalling that $\epsilon_1 + \epsilon_2 + \epsilon_2' + \epsilon_3 + \epsilon_4 = \epsilon$, we have:

$$(1-\epsilon)\,\mathrm{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y}) \leq \mathrm{tr}(\mathbf{Y}\tilde{\mathbf{C}}\mathbf{Y}) - \min\{0, \mathrm{tr}(\mathbf{E}_2)\} \leq (1+\epsilon)\,\mathrm{tr}(\mathbf{Y}\mathbf{C}\mathbf{Y}).$$

□

Chapter 4

Reduction to Spectral Norm Matrix Approximation

To prove our random projection results, we rely on a reduction from the requirements of Lemma 3 to *spectral norm matrix approximation.* Recall that for random projection, we can always write $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R}$, where $\mathbf{R} \in \mathbb{R}^{d\times m}$ is a random Johnson-Lindenstrauss matrix. In order to simplify our proofs we wish to construct a new matrix $\mathbf{B}$ such that, along with a few other conditions,

$$\|\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top\|_2 < \epsilon$$

implies that $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R}$ satisfies the conditions of Lemma 3. Specifically we show:

Lemma 4. *Suppose that, for $m \leq 2k$, we have some $\mathbf{Z} \in \mathbb{R}^{d\times m}$ with orthonormal columns satisfying $\|\mathbf{A}-\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_F^2 \leq 2\|\mathbf{A}_{r\setminus k}\|_F^2$ and $\|\mathbf{A}-\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_2^2 \leq \frac{2}{k}\|\mathbf{A}_{r\setminus k}\|_F^2$. Set $\mathbf{B} \in \mathbb{R}^{(n+m)\times d}$ to have $\mathbf{B}_1 = \mathbf{Z}^\top$ as its first m rows and $\mathbf{B}_2 = \frac{\sqrt{k}}{\|\mathbf{A}_{r\setminus k}\|_F} \cdot (\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top)$ as its remaining n rows. Then $1 \leq \|\mathbf{B}\mathbf{B}^\top\|_2 \leq 2$, $\mathrm{tr}(\mathbf{B}\mathbf{B}^\top) \leq 3k$, and $\mathrm{tr}(\mathbf{B}_2\mathbf{B}_2^\top) \leq 2k$. Furthermore, if*

$$\|\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top\|_2 < \epsilon \tag{4.1}$$

and

$$\operatorname{tr}(\mathbf{B}_2\mathbf{R}\mathbf{R}^\top\mathbf{B}_2^\top) - \operatorname{tr}(\mathbf{B}_2\mathbf{B}_2^\top) \leq \epsilon k, \tag{4.2}$$

then $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R}$ *satisfies the conditions of Lemma 3 with error* 6ϵ.

Note that the construction of $\mathbf{B}$ is actually just an approach to splitting $\mathbf{A}$ into orthogonal pairs as described in Section 3.2. The conditions on $\mathbf{Z}$ ensure that $\mathbf{A}\mathbf{Z}\mathbf{Z}^\top$ is a good low rank approximation for $\mathbf{A}$ in both the Frobenius norm and spectral norm sense. We could simply define $\mathbf{B}$ with $\mathbf{Z} = \mathbf{V}_{2k}$, the top right singular vectors of $\mathbf{A}$. In fact, this is all we need for our random projection result, but we keep the lemma in this more general form. Also note that $\operatorname{tr}(\mathbf{B}_2\mathbf{B}_2^\top) = \frac{k}{\|\mathbf{A}_{r\setminus k}\|_F^2}\|\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_F^2 \leq (1+\epsilon)k$. So equation (4.2) is in essence just ensuring that $\mathbf{R}$ preserves the trace of this block of $\mathbf{B}\mathbf{B}^\top$ to relative error.

Proof. We first show that $1 \leq \|\mathbf{B}\mathbf{B}^\top\|_2 \leq 2$. Notice that $\mathbf{B}_1\mathbf{B}_2^\top = \mathbf{0}$, so $\mathbf{B}\mathbf{B}^\top$ is a block diagonal matrix with an upper left block equal to $\mathbf{B}_1\mathbf{B}_1^\top = \mathbf{I}$ and lower right block equal to $\mathbf{B}_2\mathbf{B}_2^\top$. The spectral norm of the upper left block is 1. By our spectral norm bound on $\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top$, $\|\mathbf{B}_2\mathbf{B}_2^\top\|_2 \leq \frac{2}{k}\|\mathbf{A}_{r\setminus k}\|_F^2\frac{k}{\|\mathbf{A}_{r\setminus k}\|_F^2} = 2$, giving us the upper bound for $\mathbf{B}\mathbf{B}^\top$. Additionally, $\operatorname{tr}(\mathbf{B}_2\mathbf{B}_2^\top) \leq \frac{k}{\|\mathbf{A}_{r\setminus k}\|_F^2}\|\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_F^2 \leq 2k$ by our Frobenius norm condition on $\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top$. Finally, $\operatorname{tr}(\mathbf{B}\mathbf{B}^\top) = \operatorname{tr}(\mathbf{B}_1\mathbf{B}_1^\top) + \operatorname{tr}(\mathbf{B}_2\mathbf{B}_2^\top) \leq 3k$.

We now proceed to the main reduction. Start by setting $\mathbf{E} = \tilde{\mathbf{C}} - \mathbf{C} = \mathbf{A}\mathbf{R}\mathbf{R}^\top\mathbf{A}^\top - \mathbf{A}\mathbf{A}^\top$. Now, choose $\mathbf{W}_1 \in \mathbb{R}^{n\times(n+m)}$ such that $\mathbf{W}_1\mathbf{B} = \mathbf{A}\mathbf{Z}\mathbf{Z}^\top$. Note that $\mathbf{W}_1$ has all columns other than its first m as zero, since reconstructing $\mathbf{A}\mathbf{Z}\mathbf{Z}^\top$ only requires recombining rows of $\mathbf{B}_1 = \mathbf{Z}^\top$. Set $\mathbf{W}_2 \in \mathbb{R}^{n\times(n+m)}$ to have its first m columns zero and its next n columns as the $n \times n$ identity matrix multiplied by $\frac{\|\mathbf{A}_{r\setminus k}\|_F}{\sqrt{k}}$. This insures that $\mathbf{W}_2\mathbf{B} = \frac{\|\mathbf{A}_{r\setminus k}\|_F}{\sqrt{k}}\mathbf{B}_2 = \mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top$. So, $\mathbf{A} = \mathbf{W}_1\mathbf{B} + \mathbf{W}_2\mathbf{B}$ and we can rewrite:

$$\begin{aligned}\mathbf{E} = (&\mathbf{W}_1\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_1^\top - \mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_1^\top) + (\mathbf{W}_2\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_2^\top - \mathbf{W}_2\mathbf{B}\mathbf{B}^\top\mathbf{W}_2^\top) + \\ &(\mathbf{W}_1\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_2^\top - \mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_2^\top) + (\mathbf{W}_2\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{B}\mathbf{B}^\top\mathbf{W}_1^\top)\end{aligned}$$

We consider each term of this sum separately, showing that each corresponds to one of the allowed error terms from Lemma 3. Set $\mathbf{E}_1 = (\mathbf{W}_1\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_1^\top - \mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_1^\top)$. Clearly $\mathbf{E}_1$ is symmetric. If, as required, $\|\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top\|_2 < \epsilon$, $-\epsilon\mathbf{I} \preceq (\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top) \preceq \epsilon\mathbf{I}$ so $-\epsilon\mathbf{W}_1\mathbf{W}_1^\top \preceq \mathbf{E}_1 \preceq \epsilon\mathbf{W}_1\mathbf{W}_1^\top$. Furthermore, $\mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_1^\top = \mathbf{A}\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{A}^\top \preceq \mathbf{A}\mathbf{A}^\top = \mathbf{C}$. Since $\mathbf{W}_1$ is all zeros except in its first m columns and since $\mathbf{B}_1\mathbf{B}_1^\top = \mathbf{I}$, $\mathbf{W}_1\mathbf{W}_1^\top = \mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_1^\top$. This gives us:

$$\mathbf{W}_1\mathbf{W}_1^\top = \mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_1^\top \preceq \mathbf{C}. \tag{4.3}$$

So overall we have:

$$-\epsilon\mathbf{C} \preceq \mathbf{E}_1 \preceq \epsilon\mathbf{C}, \tag{4.4}$$

satisfying the error conditions of Lemma 3.

Next, set $\mathbf{E}_2 = (\mathbf{W}_2\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_2^\top - \mathbf{W}_2\mathbf{B}\mathbf{B}^\top\mathbf{W}_2^\top)$. Again, $\mathbf{E}_2$ is symmetric and

$$\operatorname{tr}(\mathbf{E}_2) = \frac{\|\mathbf{A}_{r\setminus k}\|_F^2}{k}\operatorname{tr}(\mathbf{B}_2\mathbf{R}\mathbf{R}^\top\mathbf{B}_2^\top - \mathbf{B}_2\mathbf{B}_2^\top) \leq \epsilon\|\mathbf{A}_{r\setminus k}\|_F^2 \tag{4.5}$$

by condition (4.2). Furthermore,

$$\begin{aligned}\sum_{i=1}^{k}|\lambda_i(\mathbf{E}_2)| &\leq k\cdot|\lambda_1(\mathbf{E}_2)|\\ &\leq k\cdot\frac{\|\mathbf{A}_{r\setminus k}\|_F^2}{k}|\lambda_1(\mathbf{B}_2\mathbf{R}\mathbf{R}^\top\mathbf{B}_2^\top - \mathbf{B}_2\mathbf{B}_2^\top)|\\ &\leq \|\mathbf{A}_{r\setminus k}\|_F^2\cdot|\lambda_1(\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top)|\\ &\leq \epsilon\|\mathbf{A}_{r\setminus k}\|_F^2 \end{aligned}\tag{4.6}$$

by condition (4.1). So $\mathbf{E}_2$ also satisfies the conditions of Lemma 3.

Next, set $\mathbf{E}_3 = (\mathbf{W}_1\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_2^\top - \mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_2^\top)$. The columns of $\mathbf{E}_3$ are in the column span of $\mathbf{W}_1\mathbf{B} = \mathbf{A}\mathbf{Z}\mathbf{Z}^\top$, and so in the column span of $\mathbf{C}$. Now:

$$\mathbf{E}_3^\top \mathbf{C}^+ \mathbf{E}_3 = \mathbf{W}_2(\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top)\mathbf{W}_1^\top \mathbf{C}^+ \mathbf{W}_1(\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top)\mathbf{W}_2^\top.$$

$\mathbf{W}_1\mathbf{W}_1^\top \preceq \mathbf{C}$ by (4.3), so $\mathbf{W}_1^\top \mathbf{C}^+ \mathbf{W}_1 \preceq \mathbf{I}$. So:

$$\mathbf{E}_3^\top \mathbf{C}^+ \mathbf{E}_3 \preceq \mathbf{W}_2(\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top)^2 \mathbf{W}_2^\top$$

which gives:

$$\|\mathbf{E}_3^\top \mathbf{C}^+ \mathbf{E}_3\|_2 \leq \|\mathbf{W}_2(\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top)^2 \mathbf{W}_2^\top\|_2 \leq \frac{\|\mathbf{A}_{r\setminus k}\|_F^2}{k} \|(\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top - \mathbf{B}\mathbf{B}^\top)^2\|_2 \leq \epsilon^2 \frac{\|\mathbf{A}_{r\setminus k}\|_F^2}{k}$$

by condition (4.1). Now, $\mathbf{E}_3$ and hence $\mathbf{E}_3^\top \mathbf{C}^+ \mathbf{E}_3$ only have rank $m \leq 2k$ so

$$\operatorname{tr}(\mathbf{E}_3^\top \mathbf{C}^+ \mathbf{E}_3) \leq 2\epsilon^2 \|\mathbf{A}_{r\setminus k}\|_F^2. \tag{4.7}$$

Finally, we set $\mathbf{E}_4 = (\mathbf{W}_2\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{B}\mathbf{B}^\top\mathbf{W}_1^\top) = \mathbf{E}_3^\top$ and thus immediately have:

$$\operatorname{tr}(\mathbf{E}_4 \mathbf{C}^+ \mathbf{E}_4^\top) \leq 2\epsilon^2 \|\mathbf{A}_{r\setminus k}\|_F^2. \tag{4.8}$$

Together, (4.4), (4.5), (4.6), (4.7), and (4.8) ensure that $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R}$ satisfies Lemma 3 with error $3\epsilon + 2\sqrt{2}\epsilon \leq 6\epsilon$.

□

Chapter 5

Sketches through Random Row Projections

The reduction in Lemma 4 reduces the problem of finding a projection-cost preserving sketch to well understood matrix sketching guarantees – subspace embedding (4.1) and trace preservation (4.2). A variety of known sketching techniques achieve the error bounds required, including several families of *subspace embedding* matrices which are referred to as Johnson-Lindenstrauss or random projection matrices throughout this paper. Note that, to better match previous writing in this area, the matrix $\mathbf{M}$ given below will correspond to the transpose of $\mathbf{B}$ in Lemma 4.

Lemma 5. *Let* $\mathbf{M}$ *be a matrix with* q *rows,* $\|\mathbf{M}^\intercal\mathbf{M}\|_2 \leq 1$*, and* $\frac{\operatorname{tr}(\mathbf{M}^\intercal\mathbf{M})}{\|\mathbf{M}^\intercal\mathbf{M}\|_2} \leq k$*. Suppose* $\mathbf{R}$ *is a sketch drawn from any of the following probability distributions of matrices. Then, for any* $\epsilon < 1$ *and* $\delta < 1/2$*,* $\|\mathbf{M}^\intercal\mathbf{R}^\intercal\mathbf{R}\mathbf{M} - \mathbf{M}^\intercal\mathbf{M}\|_2 \leq \epsilon$ *and* $\left|\operatorname{tr}(\mathbf{M}^\intercal\mathbf{R}^\intercal\mathbf{R}\mathbf{M}) - \operatorname{tr}(\mathbf{M}^\intercal\mathbf{M})\right| \leq \epsilon k$ *with probability at least* $1 - \delta$*.*

1. $\mathbf{R} \in \mathbb{R}^{d' \times q}$ *a dense Johnson-Lindenstrauss matrix with* $d' = O\left(\frac{k+\log(1/\delta)}{\epsilon^2}\right)$*, where each element is chosen independently and uniformly as* $\pm\sqrt{1/d'}$ *[Ach03]. Additionally, the same matrix family except with elements only* $O(\log(k/\delta))$*-independent [CW09].*

2. $\mathbf{R} \in \mathbb{R}^{d' \times q}$ *a fully sparse embedding matrix with* $d' = O\left(\frac{k^2}{\epsilon^2\delta}\right)$*, where each column has*

a single ±1 in a random position (sign and position chosen uniformly and independently). Additionally, the same matrix family with position and sign determined by a 4-independent hash function [CW13, MM13, NN13].

3. **R** *an OSNAP sparse subspace embedding matrix [NN13].*

Lemma 5 requires that $\mathbf{M}$ has *stable rank* $\frac{\|\mathbf{M}\|_F^2}{\|\mathbf{M}\|_2^2} \leq k$. It is well known that if $\mathbf{M}$ has *rank* $\leq k$, the $\|\mathbf{M}^\top\mathbf{R}^\top\mathbf{R}\mathbf{M} - \mathbf{M}^\top\mathbf{M}\|_2 \leq \epsilon$ bound holds for all of the above families because they are all subspace embedding matrices. It can be shown that the relaxed stable rank guarantee is sufficient as well [CNW14]. We include an alternative proof that gives a slightly worse δ dependence for some constructions but does not rely on these stable rank results.

Since $\|\mathbf{M}^\top\mathbf{M}\|_2 \leq 1$, our stable rank requirement ensures that $\mathrm{tr}(\mathbf{M}^\top\mathbf{M}) = \|\mathbf{M}\|_F^2 \leq k$. Thus, the $\left|\mathrm{tr}(\mathbf{M}^\top\mathbf{R}^\top\mathbf{R}\mathbf{M}) - \mathrm{tr}(\mathbf{M}^\top\mathbf{M})\right| \leq \epsilon k$ bound holds as long as $|\|\mathbf{R}\mathbf{M}\|_F^2 - \|\mathbf{M}\|_F^2| \leq \epsilon\|\mathbf{M}\|_F^2$. This Frobenius norm bound is standard for embedding matrices and can be proven via the JL-moment property (see Lemma 2.6 in [CW09] or Problem 2(c) in [Nel13]). For family *1*, a proof of the required moment bounds can be found in Lemma 2.7 of [CW09]. For family *2* see Remark 23 in [KN14]. For family *3* see Section 6 in [KN14].

To apply the matrix families from Lemma 5 to Lemma 4, we first set $\mathbf{M}$ to $\frac{1}{2}\mathbf{B}^\top$ and use the sketch matrix $\mathbf{R}^\top$. Applying Lemma 5 with $\epsilon' = \epsilon/4$ gives requirement (4.1) with probability $1-\delta$. Note that for all the above families of random matrices, (4.2) follows from applying Lemma 5 separately with $\mathbf{M} = \frac{1}{2}\mathbf{B}_2^\top$ and $\epsilon' = \epsilon/4$.

5.1 Projection-cost preserving sketches from random projection matrices

Since all the matrix families listed are oblivious (do not depend on $\mathbf{M}$) we can apply Lemma 4 with any suitable $\mathbf{B}$, including the one coming from the exact SVD with $\mathbf{Z} = \mathbf{V}_{2k}$. Note that $\mathbf{B}$ *does not* need to be computed at all to apply these oblivious reductions – it is purely for the analysis. This gives our main random projection result:

Theorem 6. *Let $\mathbf{R} \in \mathbb{R}^{d' \times d}$ be drawn from any of the first three matrix families from Lemma 5. Then, for any matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, with probability at least $1 - O(\delta)$, $\mathbf{A}\mathbf{R}^\top$ is a rank k projection-cost preserving sketch of $\mathbf{A}$ (i.e. satisfies Definition 1) with error $O(\epsilon)$.*

Family *1* gives oblivious reduction to $O(k/\epsilon^2)$ dimensions, while family *2* achieves $O(k^2/\epsilon^2)$ dimensions with the advantage of being faster to apply to $\mathbf{A}$, especially when our data is sparse. Family *3* allows a tradeoff between output dimension and computational cost.

A simple proof of Theorem 6 can be obtained that avoids work in [CNW14] and only depends on more well establish Johnson-Lindenstrauss properties. We set $\mathbf{Z} = \mathbf{V}_k$ and bound the error terms from Lemma 4 directly (without going through Lemma 5). The bound on $\mathbf{E}_1$ (4.4) follows from noting that $\mathbf{W}_1\mathbf{B} = \mathbf{A}\mathbf{V}_k\mathbf{V}_k^\top$ only has rank k. Thus, we can apply the fact that families *1*, *2*, and *3* are subspace embeddings to claim that $\operatorname{tr}(\mathbf{W}_1\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_1^\top - \mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_1^\top) \leq \epsilon \operatorname{tr}(\mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_1^\top)$.

The bound on $\mathbf{E}_2$ (4.6) follows from first noting that, since we set $\mathbf{Z} = \mathbf{V}_k$, $\mathbf{E}_2 = (\mathbf{A}_{r\setminus k}\mathbf{R}\mathbf{R}^\top\mathbf{A}_{r\setminus k}^\top - \mathbf{A}_{r\setminus k}\mathbf{A}_{r\setminus k}^\top)$. Applying Theorem 21 of [KN14] (approximate matrix multiplication) along with the referenced JL-moment bounds for our first three families gives $\|\mathbf{E}_2\|_F \leq \frac{\epsilon}{\sqrt{k}}\|\mathbf{A}_{r\setminus k}\|_F^2$. Since $\sum_{i=1}^k |\lambda_i(\mathbf{E}_2)| \leq \sqrt{k}\|\mathbf{E}_2\|_F$, (4.6) follows. Note that (4.5) did not require the stable rank generalization, so we do not need any modified analysis.

Finally, the bounds on $\mathbf{E}_3$ and $\mathbf{E}_4$, (4.7) and (4.8), follow from the fact that:

$$\operatorname{tr}(\mathbf{E}_3^\top\mathbf{C}^+\mathbf{E}_3) = \|\mathbf{\Sigma}^{-1}\mathbf{U}^\top(\mathbf{W}_1\mathbf{B}\mathbf{R}\mathbf{R}^\top\mathbf{B}^\top\mathbf{W}_2^\top - \mathbf{W}_1\mathbf{B}\mathbf{B}^\top\mathbf{W}_2^\top)\|_F^2 = \|\mathbf{V}_k\mathbf{R}\mathbf{R}^\top\mathbf{A}_{r\setminus k}^\top\|_F^2 \leq \epsilon^2\|\mathbf{A}_{r\setminus k}\|_F^2$$

again by Theorem 21 of [KN14] and the fact that $\|\mathbf{V}_k\|_F^2 = k$. In both cases, we apply the approximate matrix multiplication result with error $\epsilon/\sqrt{k}$. For family *1*, the required moment bound needs a sketch with dimension $O\left(\frac{k\log(1/\delta)}{\epsilon^2}\right)$ (see Lemma 2.7 of [CW09]). Thus, our alternative proof slightly increases the δ dependence stated in Lemma 5.

Chapter 6

Constant Factor Approximation with $O(\log k)$ Dimensions

In this section we show that randomly projecting $\mathbf{A}$ to just $O(\log k/\epsilon^2)$ dimensions using a Johnson-Lindenstrauss matrix is sufficient for approximating k-means up to a factor of $(9+\epsilon)$. To the best of our knowledge, this is the first result achieving a constant factor approximation using a sketch with data dimension independent of the input size (n and d) and sublinear in k. This result opens up the interesting question of whether it is possible to achieve a $(1+\epsilon)$ relative error approximation to k-means using just $O(\log k)$ rather than $O(k)$ dimensions. Specifically, we show:

Theorem 7. *For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, any $0 \leq \epsilon < 1$, and $\mathbf{R} \in \mathbb{R}^{O\left(\frac{\log(k/\delta)}{\epsilon^2}\right) \times d}$ drawn from a Johnson-Lindenstrauss distribution, let $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R}^\top$. Let S be the set of all k-cluster projection matrices, let $\mathbf{P}^* = \arg\min_{\mathbf{P} \in S} \|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2$, and let $\tilde{\mathbf{P}}^* = \arg\min_{\mathbf{P} \in S} \|\tilde{\mathbf{A}} - \mathbf{P}\tilde{\mathbf{A}}\|_F^2$. With probability $1-\delta$, for any $\gamma \geq 1$, and $\tilde{\mathbf{P}} \in S$, if $\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}\tilde{\mathbf{A}}\|_F^2 \leq \gamma\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}^*\tilde{\mathbf{A}}\|_F^2$:*

$$\|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F^2 \leq (9+\epsilon) \cdot \gamma \|\mathbf{A} - \mathbf{P}^*\mathbf{A}\|_F^2.$$

In other words, if $\tilde{\mathbf{P}}$ is a cluster indicator matrix (see Section 2.3) for an approximately optimal clustering of $\tilde{\mathbf{A}}$, then the clustering is also within a constant factor of optimal for $\mathbf{A}$.

Note that there are a variety of distributions that are sufficient for choosing $\mathbf{R}$. For example, we may use the dense Rademacher matrix distribution of family *1* of Lemma 5, or a sparse family such as those given in [KN14].

To achieve the $O(\log k/\epsilon^2)$ bound, we must focus specifically on k-means clustering – it is clear that projecting to $< k$ dimensions is insufficient for solving general constrained k-rank approximation as $\tilde{\mathbf{A}}$ will not even have rank k. Additionally, other sketching techniques than random projection do not work when $\tilde{\mathbf{A}}$ has fewer than O(k) columns. Consider clustering the rows of the $n \times n$ identity into n clusters, achieving cost 0. An SVD projecting to less than $k = n - 1$ dimensions or column selection technique taking less than $k = n - 1$ columns will leave at least two rows in $\tilde{\mathbf{A}}$ with all zeros. These rows may be clustered together when optimizing the k-means objective for $\tilde{\mathbf{A}}$, giving a clustering with cost > 0 for $\mathbf{A}$ and hence failing to achieve multiplicative error.

Proof. As mentioned in Section 3.2, the main idea is to analyze an $O(\log k/\epsilon^2)$ dimension random projection by splitting $\mathbf{A}$ in a substantially different way than we did in the analysis of other sketches. Specifically, we split it according to its optimal k clustering and the remainder matrix:

$$\mathbf{A} = \mathbf{P}^*\mathbf{A} + (\mathbf{I} - \mathbf{P}^*)\mathbf{A}.$$

For conciseness, write $\mathbf{B} = \mathbf{P}^*\mathbf{A}$ and $\overline{\mathbf{B}} = (\mathbf{I} - \mathbf{P}^*)\mathbf{A}$. So we have $\mathbf{A} = \mathbf{B} + \overline{\mathbf{B}}$ and $\tilde{\mathbf{A}} = \mathbf{B}\mathbf{R}^\top + \overline{\mathbf{B}}\mathbf{R}^\top$.

By the triangle inequality and the fact that projection can only decrease Frobenius norm:

$$\|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F \leq \|\mathbf{B} - \tilde{\mathbf{P}}\mathbf{B}\|_F + \|\overline{\mathbf{B}} - \tilde{\mathbf{P}}\overline{\mathbf{B}}\|_F \leq \|\mathbf{B} - \tilde{\mathbf{P}}\mathbf{B}\|_F + \|\overline{\mathbf{B}}\|_F. \tag{6.1}$$

Next note that $\mathbf{B}$ is simply $\mathbf{A}$ with every row replaced by its cluster center (in the optimal clustering of $\mathbf{A}$). So $\mathbf{B}$ has just k distinct rows. Multiplying by a Johnson-Lindenstauss matrix with $O(\log(k/\delta)/\epsilon^2)$ columns will preserve the squared distances between all of these k points with probability $1 - \delta$. It is not difficult to see that preserving distances is sufficient to preserve the cost of any clustering of $\mathbf{B}$ since we can rewrite the k-means objection function as a linear function of squared distances alone:

$$\|\mathbf{B} - \mathbf{X}_C\mathbf{X}_C^\top\mathbf{B}\|_F^2 = \sum_{j=1}^{n} \|\mathbf{b}_j - \boldsymbol{\mu}_{C(j)}\|_2^2 = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{\substack{\mathbf{b}_j,\mathbf{b}_k \in C_i \\ j \neq k}} \|\mathbf{b}_j - \mathbf{b}_k\|_2^2.$$

So, $\|\mathbf{B} - \tilde{\mathbf{P}}\mathbf{B}\|_F^2 \leq (1+\epsilon)\|\mathbf{B}\mathbf{R}^\top - \tilde{\mathbf{P}}\mathbf{B}\mathbf{R}^\top\|_F^2$. Combining with (6.1) and noting that square rooting can only reduce multiplicative error, we have:

$$\|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F \leq (1+\epsilon)\|\mathbf{B}\mathbf{R}^\top - \tilde{\mathbf{P}}\mathbf{B}\mathbf{R}^\top\|_F + \|\overline{\mathbf{B}}\|_F.$$

Rewriting $\mathbf{B}\mathbf{R}^\top = \tilde{\mathbf{A}} - \overline{\mathbf{B}}\mathbf{R}^\top$ and again applying triangle inequality and the fact the projection can only decrease Frobenius norm, we have:

$$\begin{aligned}\|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F &\leq (1+\epsilon)\|(\tilde{\mathbf{A}} - \overline{\mathbf{B}}\mathbf{R}^\top) - \tilde{\mathbf{P}}(\tilde{\mathbf{A}} - \overline{\mathbf{B}}\mathbf{R}^\top)\|_F + \|\overline{\mathbf{B}}\|_F \\ &\leq (1+\epsilon)\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}\tilde{\mathbf{A}}\|_F + (1+\epsilon)\|(\mathbf{I} - \tilde{\mathbf{P}})\overline{\mathbf{B}}\mathbf{R}^\top\|_F + \|\overline{\mathbf{B}}\|_F \\ &\leq (1+\epsilon)\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}\tilde{\mathbf{A}}\|_F + (1+\epsilon)\|\overline{\mathbf{B}}\mathbf{R}^\top\|_F + \|\overline{\mathbf{B}}\|_F.\end{aligned}$$

As discussed in Section 5, multiplying by a Johnson-Lindenstrauss matrix with at least $O(\log(1/\delta)/\epsilon^2)$ columns will preserve the Frobenius norm of any fixed matrix up to ϵ error so $\|\overline{\mathbf{B}}\mathbf{R}^\top\|_F \leq (1+\epsilon)\|\overline{\mathbf{B}}\|_F$. Using this and the fact that $\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}\tilde{\mathbf{A}}\|_F^2 \leq \gamma\|\tilde{\mathbf{A}} - \tilde{\mathbf{P}}^*\tilde{\mathbf{A}}\|_F^2 \leq \gamma\|\tilde{\mathbf{A}} - \mathbf{P}^*\tilde{\mathbf{A}}\|_F^2$ we have:

$$\|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F \leq (1+\epsilon)\sqrt{\gamma}\|\tilde{\mathbf{A}} - \mathbf{P}^*\tilde{\mathbf{A}}\|_F + (2+3\epsilon)\|\overline{\mathbf{B}}\|_F.$$

Finally, we note that $\overline{\mathbf{B}} = \mathbf{A} - \mathbf{P}^*\mathbf{A}$ and again apply the fact that multiplying by $\mathbf{R}^\top$ preserves the Frobenius norm of any fixed matrix with high probability. So, $\|\tilde{\mathbf{A}} - \mathbf{P}^*\tilde{\mathbf{A}}\|_F \leq (1+\epsilon)\|\mathbf{A} - \mathbf{P}^*\mathbf{A}\|_F$ and thus:

$$\|\mathbf{A} - \tilde{\mathbf{P}}\mathbf{A}\|_F \leq (3+6\epsilon)\sqrt{\gamma}\|\mathbf{A} - \mathbf{P}^*\mathbf{A}\|_F.$$

Squaring and adjusting ϵ by a constant factor gives the desired result. □

Bibliography

[Ach03] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003. Preliminary version in the 20th Symposium on Principles of Database Systems (PODS).

[ADHP09] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

[APD14] Megasthenis Asteris, Dimitris Papailiopoulos, and Alexandros Dimakis. Nonnegative sparse PCA with provable guarantees. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1728–1736, 2014.

[AV07] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035, 2007.

[BMD09] Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. Unsupervised feature selection for the k-means clustering problem. In *Advances in Neural Information Processing Systems 22 (NIPS)*, pages 153–161, 2009.

[BMI13] Christos Boutsidis and Malik Magdon-Ismail. Deterministic feature selection for k-means clustering. *IEEE Transactions on Information Theory*, 59(9):6099–6110, 2013.

[BZD10] Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 298–306, 2010.

[BZMD15] Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas. Randomized dimensionality reduction for k-means clustering. *IEEE Transactions on Information Theory*, 61(2):1045–1062, Feb 2015.

[CEM$^+$14] Michael Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Mădălina Persu. Dimensionality reduction for k-means clustering and low rank approximation. CoRR, 2014.

[CNW14] Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. Manuscript, 2014.

[CW09] Kenneth Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 205–214, 2009.

[CW13] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 81–90, 2013.

[DFK$^+$04] Petros Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and V Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004. Preliminary version in the 10th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA).

[FSS13] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1434–1453, 2013.

[HMT11] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[HPK07] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discrete and Computational Geometry*, 37(1):3–19, 2007. Preliminary version in the 21st Annual Symposium on Computational Geometry (SCG).

[KMN$^+$02] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for k-means clustering. In *Proceedings of the 18th Annual Symposium on Computational Geometry (SCG)*, pages 10–18, 2002.

[KN14] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1):4, 2014. Preliminary version in the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA).

[KSS04] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 454–462, 2004.

[Llo82] Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[Mah11] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[Mir60] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11:50–59, 1960.

[MM13] Michael W. Mahoney and Xiangrui Meng. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 91–100, 2013.

[MNV09] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is NP-hard. In *Proceedings of the 3rd International Workshop on Algorithms and Computation (WALCOM)*, pages 274–285, 2009.

[Nel13] Jelani Nelson. Cs 229r algorithms for big data, problem set 6. `http://people.seas.harvard.edu/~minilek/cs229r/psets/pset6.pdf`, 2013.

[NN13] Jelani Nelson and Huy L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 117–126, 2013.

[PDK13] Dimitris Papailiopoulos, Alexandros Dimakis, and Stavros Korokythakis. Sparse PCA through low-rank approximations. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 747–755, 2013.

[Sar06] Támas Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

[YZ13] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *The Journal of Machine Learning Research*, 14(1):899–925, 2013.