Sampling-Based Algorithms for Dimension Reduction

by

Amit Jayant Deshpande
B.Sc., Chennai Mathematical Institute, 2002

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2007

Author .....
Department of Mathematics
April 11, 2007

Certified by .....
Santosh S. Vempala
Associate Professor of Applied Mathematics
Thesis co-supervisor

Certified by .....
Daniel A. Spielman
Professor of Computer Science, Yale University
Thesis co-supervisor

Accepted by .....
Alar Toomre
Chairman, Applied Mathematics Committee

Accepted by .....
Pavel I. Etingof
Chairman, Department Committee on Graduate Students

Sampling-Based Algorithms for Dimension Reduction

by

Amit Jayant Deshpande

Submitted to the Department of Mathematics
on April 11, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Can one compute a low-dimensional representation of any given data by looking only at its small sample, chosen cleverly on the fly?

Motivated by the above question, we consider the problem of low-rank matrix approximation: given a matrix $A \in \mathbb{R}^{m \times n}$, one wants to compute a rank-k matrix (where $k \ll \min\{m, n\}$) nearest to A in the Frobenius norm (also known as the Hilbert-Schmidt norm). We prove that using a sample of roughly $O(k/\epsilon)$ rows of A one can compute, with high probability, a $(1+\epsilon)$-approximation to the nearest rank-k matrix. This gives an algorithm for low-rank approximation with an improved error guarantee (compared to the additive $\epsilon \|A\|_F^2$ guarantee known earlier from the work of Frieze, Kannan, and Vempala) and running time $O(Mk/\epsilon)$, where M is the number of non-zero entries of A. The proof is based on two sampling techniques called adaptive sampling and volume sampling, and some linear algebraic tools.

Low-rank matrix approximation under the Frobenius norm is equivalent to the problem of finding a low-dimensional subspace that minimizes the sum of squared distances to given points. The general subspace approximation problem asks one to find a low-dimensional subspace that minimizes the sum of p-th powers of distances (for $p \geq 1$) to given points. We generalize our sampling techniques and prove similar sampling-based dimension reduction results for subspace approximation. However, the proof is geometric.

Thesis co-supervisor: Santosh S. Vempala
Title: Associate Professor of Applied Mathematics

Thesis co-supervisor: Daniel A. Spielman
Title: Professor of Computer Science, Yale Unversity

To dear *aai* and *baba,*

Acknowledgments

I am greatly indebted to my advisors Daniel Spielman and Santosh Vempala for their guidance, encouragement, and patience during all these years. Both of them showed tremendous faith in me even when my progress was slow. In Spring'05, Santosh taught a wonderful course on Spectral Algorithms and Representations; most of my thesis has germinated from some questions that he posed in the class. I thank my collaborators Luis Rademacher, Grant Wang, and Kasturi Varadarajan, whose contributions have been pivotal in my research. Luis has been more than just a collaborator. I thank him for all the conversations, lunches, dinners, squash, sailing, and travel that we enjoyed together. I thank Piotr Indyk, Daniel Kleitman, and Peter Shor for agreeing to be on my thesis committee. (Finally, it was Peter Shor who attended my thesis defense and signed the thesis approval form.)

My five years at MIT, including a semester spent at Georgia Tech, were made memorable by people around me: teachers, colleagues, friends. I thank my professors Daniel Kleitman, Madhu Sudan, Michel Goemans, and Rom Pinchasi for the wonderful courses they taught. I thank Subhash Khot and Prahladh Harsha for playing the roles of my mentors from time to time. I thank my roommate Shashi, my occasional cooking and chit-chat buddies – Ajay, Pavithra, Kripa, Sreekar, Punya, Shubhangi, ... my friends from CMI – Tejaswi, Krishna, Baskar, Debajyoti, Sourav, ... my friends from the skit group – Vikram, Raghavendran, Pranava, ... and many others who made my everyday life enjoyable. Words are insufficient to express my gratitude towards my music teacher Warren Senders, who introduced me to the hell of frustration as well as the heaven of enlightenment called Hindustani music!

I thank Jaikumar Radhakrishnan, Ramesh Hariharan, Narayan Kumar, K. V. Subrahmanyam, and Meena Mahajan, my professors during undergrad, for my upbringing in Theory. I thank my teachers from school and math olympiad camps for cultivating my interest in mathematics. I thank my relatives – cousins, aunts, uncles – and friends for asking me every year, "How many more?" and providing the external force required by Newton's first law of graduation!

Above all, I thank my grandparents and my extended family, who instilled in me the liking for an academic career. I thank my parents for always giving me the freedom to choose my own path. I couldn't have reached this far if they hadn't walked me through my very early steps. This thesis is dedicated to them.

Contents

Chapter 1

Introduction

1.1 Sampling-based dimension reduction

Large or massive data sets can often be visualized as sets of points in a high-dimensional Euclidean space. Typical examples of large data sets include document-term matrix, DNA microarray data, PageRank matrix etc. Most of these examples store information about some n objects, and some m features for each of them. Hence, we can think of a large data set as a set of points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, where both m and n are large, and a_i's correspond to the feature vectors. Many of these features may be correlated, causing much redundancy in the data and in addition, there may be some noise.

In data-mining, statistics, and clustering, we want to remove this redundancy and noise in large data sets to make sure that our algorithms do not suffer from this false high-dimensionality of data. Hence, we want to find a low-dimensional representation of given large data set. One way to achieve this is by dimension reduction, i.e., by embedding the points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$ into a low-dimensional Euclidean space, say $V \simeq \mathbb{R}^k$, where $k \ll m, n$. For simplicity, we will consider V to be a k-dimensional linear subspace of $\mathbb{R}^n$, and our embedding to be the orthogonal projection onto V. This brings forth a few obvious questions:

1. How to measure the "goodness" of subspace V? In other words, when is a k-dimensional linear subspace V the best fit for our data points $a_1, a_2, \ldots, a_m$?

2. Can we find the optimal subspace V^* under this measure?

3. How good is a random subspace V?

These questions gave rise to spectral projection and random projection, which are the two most popular dimension reduction techniques in practice. However for certain problems, random projections are not quite effective as they are data-independent, and spectral projections are either futile or computationally expensive. In this thesis, we will explore this gray area where we need data-dependent dimension reduction techniques. Here are a few important questions to begin with:

1. Can one compute an approximately optimal subspace by looking only at a small sample S of the given data-points $a_1, a_2, \ldots, a_m$?

2. What distributions are useful? How to do the sampling?

3. What does it mean to have $V = \text{span}(S)$, for a small sample S, approximately optimal?

Finding low-dimensional representations of large data sets using their small sample is what we call sampling-based dimension reduction. Some obvious advantages of sampling-based dimension reduction are as follows:

1. It maintains sparsity of the data, i.e., if original data $a_1, a_2, \ldots, a_m$ is sparse, then the description of V as $\text{span}(S)$ is also sparse.

2. $\text{span}(S)$ gives, in some sense, a set of "more significant" features and the other features behave roughly like linear combinations of features in S. This is useful to extract correlations.

3. It gives pass-efficient algorithms, i.e., we can read data in a few sequential passes and come up with its low-dimensional representation using a sample picked on the fly.

1.2 Overview

In Chapter 2, we consider the problem of low-rank matrix approximation under the Frobenius norm. The measure of "goodness" of a subspace V here is the least squared error, i.e., we want a k-dimensional linear subspace V that minimizes

$$\left(\sum_{i=1}^{m} d(a_i, V)^2\right)^{1/2}.$$

In other words, given a matrix $A \in \mathbb{R}^{m \times n}$, we want to find another matrix $B \in \mathbb{R}^{m \times n}$ of rank at most k that minimizes the Frobenius norm,

$$\|A - B\|_F = \left(\sum_{i,j} (A_{ij} - B_{ij})^2\right)^{1/2}.$$

We will prove that using a sample S of $\tilde{O}(k/\epsilon)$ points from $a_1, a_2, \ldots, a_m$, one can find a k-dimensional linear subspace $V \subseteq \text{span}(S)$ which, in expectation, gives $(1+\epsilon)$-approximation to the optimal. This leads to a randomized $(1+\epsilon)$-approximation algorithm for low-rank matrix approximation that requires $O(k \log k)$ passes over the data and runs in time $\tilde{O}(Mk/\epsilon)$ – linear in M, the number of non-zero entries of A – outperforming spectral and random projection techniques. In the process, we improve upon a previous result of Frieze, Kannan, and Vempala [15], and generalize their

squared-length sampling scheme in two different ways: adaptive sampling and volume sampling. Our proofs use linear algebraic tools, and combine these two generalizations to get the final result. These results appeared in joint papers with Luis Rademacher, Santosh Vempala, and Grant Wang [11, 13].

In Chapter 3, we consider a generalization of low-rank matrix approximation called subspace approximation, where the measure of "goodness" of V is L_p-error, i.e., we want a k-dimensional linear subspace V that minimizes

$$\left(\sum_{i=1}^{m} d(a_i, V)^p\right)^{1/p}.$$

where $p \geq 1$. We extend our adaptive and volume sampling techniques to prove that using a sample S of $\tilde{O}\left(k^2(k/\epsilon)^{p+1}\right)$ points from $a_1, a_2, \ldots, a_m$, we get a randomized bi-criteria approximation, i.e., span(S) gives a $(1+\epsilon)$-approximation to the optimal k-dimensional subspace, with high probability. This work appeared in a joint paper with Kasturi Varadarajan [12].

In Chapter 4, we will see some applications of our sampling-based techniques and a few related problems.

1.3 Related work

Since the result of Frieze, Kannan, and Vempala [15], there has been a lot of work on pass-efficient algorithms for low-rank matrix approximation. All of these are randomized algorithms that read the given large matrix sequentially in a small number of passes, and produce an approximately optimal low-rank approximation. Most notable among these are the results of Achlioptas and McSherry [3], Drineas, Kannan, and Mahoney [8], which achieved better additive approximations than [15]. Har-Peled [17] and Drineas, Mahoney, and Muthukrishnan [9], independent of our work, gave linear time algorithms for low-rank matrix approximation with a multiplicative guarantee (although with more number of passes over the data). Finally, a recent result of Sarlos [21] gave a 2-pass algorithm for low-rank matrix approximation, although his algorithm is not sampling-based.

The work of Shyamalkumar and Varadarajan [22] has been a precursor to the results in Chapter 3. The ideas therein were inspired by the earlier works of Har-Peled and Varadarajan [18, 19].

Chapter 2

Low-Rank Matrix Approximation

In this chapter, we consider the problem of low-rank matrix approximation under the Frobenius norm: Given $A \in \mathbb{R}^{m \times n}$ and an integer k, we want to find another matrix $B \in \mathbb{R}^{m \times n}$ of rank at most k that minimizes the Frobenius norm of their difference, i.e.,

$$\|A - B\|_F = \left(\sum_{i,j} (A_{ij} - B_{ij}) \right)^{1/2}.$$

It is easy to see that low-rank matrix approximation under the Frobenius norm is equivalent to the following problem: Given points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, we want to find a k-dimensional linear subspace V that minimizes the least squared error, or equivalently the L_2-error,

$$\left(\sum_{i=1}^{m} d(a_i, V)^2 \right)^{1/2}.$$

To see the equivalence of these two formulations, we can think of the rows of A as $a_1, a_2, \ldots, a_m$, the rows of B as $b_1, b_2, \ldots, b_m$, and $\text{span}(b_1, b_2, \ldots, b_m) \subseteq V$.

2.1 Singular value decomposition (SVD)

How to find the optimal subspace V^* when its "goodness" is measured by the least squared error? It is known that the optimal subspace can be computed using Singular Value Decomposition (SVD) as follows.

Any $A \in \mathbb{R}^{m \times n}$ has a decomposition given by

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T,$$

where r is the rank of A and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$ are called singular values of A; $u_1, u_2, \ldots, u_r \in \mathbb{R}^m$ and $v_1, v_2, \ldots, v_r \in \mathbb{R}^n$ are orthonormal sets of vectors called as left and right singular vectors of A, respectively. This is called Singular Value Decomposition (SVD) of A and it can be computed in time $O(\min\{mn^2, m^2n\})$. It

follows that $Av_i = \sigma_i u_i$ and $A^T u_i = \sigma_i v_i$, for $1 \leq i \leq r$, and

$$\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2.$$

In other words, $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2 \geq 0$ are the eigenvalues of AA^T.

Proposition 1. *Let $A \in \mathbb{R}^{m \times n}$, then an optimal rank-$k$ approximation to A under the Frobenius norm is given by*

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

Or in other words, $V^* = \text{span}(v_1, v_2, \ldots, v_k)$ is an optimal k-dimensional linear subspace under the least squared error measure. Moreover, the minimum error of rank-k approximation is

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2.$$

Proof. See Theorem 8.2 in [23]. □

Even though SVD runs in polynomial time, with the emergence of large data sets or large matrices in practical applications, it has become important to design faster algorithms for low-rank matrix approximation. A first step in this direction is to perform dimension reduction. We consider the simplest form of dimension reduction: projecting the given points $a_1, a_2, \ldots, a_m$ onto a low-dimensional subspace $W \subseteq \mathbb{R}^n$. Let A be a matrix with rows are $a_1, a_2, \ldots, a_m$, and $\pi_W(A)$ be the matrix whose rows are $\pi_W(a_1), \pi_W(a_2), \ldots, \pi_W(a_m)$, i.e., projections of $a_1, a_2, \ldots, a_m$ onto W. We want a subspace W with the guarantee that finding low-rank matrix approximation of $\pi_W(A)$ gives an approximately good answer to the low-rank approximation of A. If we can efficiently find such a subspace W, our algorithm is simple: find W, compute $\pi_W(A)$, use SVD to find the best rank-k approximation to $\pi_W(A)$. Hence, the running time of our algorithm, apart from the time required to find W, is

$$O\left(M \cdot \dim(W) + \min\{m \cdot \dim(W)^2, m^2 \cdot \dim(W)\}\right),$$

where M is the number of non-zero entries of A. This is effectively $O\left(M \dim(W)/\epsilon\right)$ and it outperforms SVD if $\dim(W)$ is small and if we can find W efficiently.

2.1.1 Notation

Given a subspace $W \subseteq \mathbb{R}^n$, we denote the orthogonal projection onto W by $\pi_W(\cdot)$. For a matrix $A \in \mathbb{R}^{m \times n}$ with rows $a_1, a_2, \ldots, a_m$, we use $\pi_W(A)$ to denote to matrix whose rows are $\pi_W(a_1), \pi_W(a_2), \ldots, \pi_W(a_m)$. We use $\pi_{W,k}(A)$ to denote the best rank-k approximation to $\pi_W(A)$. In other words, it is the best rank-k approximation to A whose rows lie in W.

When $W = \text{span}(S)$, as in the case of sampling-based dimension reduction, we use $\pi_{\text{span}(S),k}(A)$ to denote the best rank-k approximation to A whose rows lie in $\text{span}(S)$.

2.2 Random projection

A quick way to find the subspace $W \subseteq \mathbb{R}^n$ desired in the previous section, is to use a random subspace of $\mathbb{R}^n$. Using Johnson-Lindenstrauss Lemma [20], one can show the following.

Proposition 2. *Let W be a random linear subspace of $\mathbb{R}^n$ of dimension $O(\log m/\epsilon^2)$. Then*

$$\mathbb{E}_W \left[\|A - \pi_{W,k}(A)\|_F^2 \right] \leq \|A - A_k\|_F^2 + \epsilon \|A\|_F^2 .$$

Proof. See Theorem 8.5 in [23]. □

From Proposition 2, we have

$$\mathbb{E}_W \left[\|A - \pi_{W,k}(A)\|_F^2 - \|A - A_k\|_F^2 \right] \leq \epsilon \|A\|_F^2 .$$

Therefore, using Markov's inequality, with probability at least 3/4 we have

$$\|A - \pi_{W,k}(A)\|_F^2 \leq \|A - A_k\|_F^2 + 4\epsilon \|A\|_F^2 .$$

This gives randomized algorithm that computes, with high probability, an additive $4\epsilon \|A\|_F^2$ approximation to A_k in time $O(M \log m/\epsilon^2)$, where M is the number of non-zero entries of A.

2.3 Frieze-Kannan-Vempala algorithm

The work on sampling techniques for low-rank matrix approximation was initiated by a result of Frieze, Kannan, and Vempala [15]. They showed that if S is an i.i.d. sample of $O(k/\epsilon)$ rows of A picked from the *squared-length distribution*, then $\text{span}(S)$ contains a k-dimensional subspace V whose expected error is within at most $\epsilon \|A\|_F^2$ of the optimum.

Theorem 3. *Let $A \in \mathbb{R}^{m \times n}$ and S be an i.i.d. sample of s rows of A picked from the following distribution:*

$$P_i = \Pr(\textit{picking } a_i) \propto \|a_i\|^2 ,$$

then

$$\mathbb{E}_S \left[\left\| A - \pi_{\text{span}(S),k}(A) \right\|_F^2 \right] \leq \|A - A_k\|_F^2 + \frac{k}{s} \|A\|_F^2 .$$

Hence using $s = k/\epsilon$, we get an additive $\epsilon \|A\|_F^2$ approximation.

This leads to a randomized algorithm that computes, with high probability, an additive $\epsilon \|A\|_F^2$ approximation to A_k in time $O(Mk/\epsilon)$. This gives an improvement over the random projection method. The result of Frieze, Kannan, and Vempala [15]

was further improved in the subsequent works by Drineas, Frieze, Kannan, Vempala, and Vinay [7] and Drineas, Kannan, and Mahoney [8], but all these results gave an additive approximation guarantee.

Our goal is to improve upon the Frieze-Kannan-Vempala algorithm. Here are a few observations about where the squared-length sampling scheme fails, so that we can think of some ways of fixing it.

1. In general, $\epsilon \|A\|_F^2$ can be arbitrarily large compared to $\|A - A_k\|_F^2$. For example, consider the matrix A of all 1's. It has a rank-1 approximation with zero error but $\epsilon \|A\|_F^2 = \epsilon mn$ is very large. So we would like a multiplicative $(1+\epsilon)$-approximation instead of an additive $\epsilon \|A\|_F^2$ approximation. However, when A is all 1's matrix, any row gives a good approximation using its span. So is it just the analysis of the squared-length sampling that is weak or do we really need a new sampling technique?

2. Let $e_1, e_2, \ldots, e_n$ be the standard basis of $\mathbb{R}^n$. Imagine that our set of points $a_1, a_2, \ldots, a_m$ is as follows:

$$\begin{aligned} a_1 &= e_1, \\ a_i &= e_2 + v_i, \text{ for } i = 2, 3, \ldots, m, \end{aligned}$$

   where $v_i \in \text{span}(e_3, \ldots, e_n)$ of very small length, say $\|v_i\| \leq \epsilon$. Thus, we have a lot of points $a_2, a_3, \ldots, a_m$ clustered around e_2, and a single point $a_1 = e_1$. The best 2-dimensional subspace should have error smaller than the error of $\text{span}(e_1, e_2)$, which is $\epsilon^2(m-1)$. Whereas, if we pick a sample S of points by squared-length sampling, most likely we will miss the lonely point a_1, which means the error for a_1 using subspace in $\text{span}(S)$ will be at least 1. This is much larger than $\epsilon^2(m-1)$, for small ϵ. So we need to modify the squared-length sampling scheme so as to avoid the situation mentioned above.

2.4 Adaptive sampling

From the example given in the previous section where the squared-length sampling fails, we see that a sample picked according to squared-length distribution may miss some of the important points. Therefore, it may be better to pick some additional points to capture the directions that are orthogonal to the span of our current sample. This gives rise to the first generalization of squared-length sampling called *adaptive sampling*. In adaptive sampling, we start with an initial subspace V, pick a sample S of points with probabilities proportional to their squared distances from V, and use $\text{span}(V \cup S)$ as our next subspace. Here is the analysis of one round of adaptive sampling.

Theorem 4. *Let $A \in \mathbb{R}^{m \times n}$, and $V \subseteq \mathbb{R}^n$ be any linear subspace. Let $E = A - \pi_V(A)$. Let S be an i.i.d. sample of s rows of A picked from the following distribution:*

$$P_i = \Pr(\textit{picking } a_i) \propto d(a_i, V)^2 = \|a_i - \pi_V(a_i)\|^2.$$

Then

$$\mathbb{E}_S\left[\left\|A - \pi_{\text{span}(V\cup S),k}(A)\right\|_F^2\right] \leq \|A - A_k\|_F^2 + \frac{k}{s}\|E\|_F^2.$$

Remark 5. *Notice that using $V = 0$, Theorem 4 implies Theorem 3 as its special case.*

Proof. This proof is a small modification of the original proof of Theorem 3 by Frieze, Kannan, and Vempala [15]. We define vectors $w_1, w_2, \ldots, w_k \in \text{span}(V \cup S)$ such that $W = \text{span}(\{w_1, \ldots, w_k\})$ and show that W is a good approximation to $\text{span}(\{v_1, v_2, \ldots, v_k\})$ in the sense that

$$\mathbb{E}_S\left[\|A - \pi_W(A)\|_F^2\right] \leq \|A - A_k\|_F^2 + \frac{k}{s}\|E\|_F^2. \tag{2.1}$$

Recall that $A_k = \pi_{\text{span}\{v_1,\ldots,v_k\}}(A)$, i.e., $\text{span}\{v_1, \ldots, v_k\}$ is the optimal subspace upon which to project. Proving (2.1) proves the theorem, since $W \subseteq \text{span}(V \cup S)$ and $\dim(W) \leq k$.

For $1 \leq j \leq k$, define random vectors $x_j, w_j \in \mathbb{R}^n$ as follows.

$$x_j = \frac{1}{s}\sum_{i\in S}\frac{(u_j)_i}{P_i}\left(a_i - \pi_V(a_i)\right),$$

and

$$w_j = \pi_V(A)^T u_j + x_j.$$

Since the elements of S are i.i.d., one can think of x_j as $x_j = \frac{1}{s}\sum_{l=1}^{s} x_j^{(l)}$, where $x_j^{(l)} \in \mathbb{R}^n$ are i.i.d. random vectors defined as follows.

$$x_j^{(l)} = \frac{(u_j)_i}{P_i}\left(a_i - \pi_V(a_i)\right), \text{ with probability } P_i.$$

Notice that when we take coordinate-wise expectation of x_j, we get

$$\begin{aligned}\mathbb{E}_S[x_j] &= \frac{1}{s}\sum_{l=1}^{s}\mathbb{E}\left[x_j^{(l)}\right] \\ &= \sum_{i=1}^{m}(u_j)_i\left(a_i - \pi_V(a_i)\right) \\ &= E^T u_j\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_S[w_j] &= \pi_V(A)^T u_j + E^T u_j \\ &= A^T u_j \\ &= \sigma_j v_j.\end{aligned}$$

Therefore, in expectation, W gives the optimal k-dimensional subspace that we are

looking for. Now we can prove a bound on the second central moment of each w_j as follows.

$$\begin{aligned}\mathbb{E}_S\left[\|w_j - \sigma_j v_j\|^2\right] &= \mathbb{E}_S\left[\pi_V(A)^T u_j + x_j - A^T u_j\right] \\ &= \mathbb{E}_S\left[x_j - E^T u_j\right] \\ &= \mathbb{E}_S\left[\|x_j\|^2\right] - 2\mathbb{E}_S\left[x_j \cdot E^T u_j\right] + \left\|E^T u_j\right\|^2 \\ &= \mathbb{E}_S\left[\|x_j\|^2\right] - \left\|E^T u_j\right\|^2. \end{aligned} \tag{2.2}$$

We evaluate the first term in (2.2),

$$\begin{aligned}\mathbb{E}_S\left[\|x_j\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{s}\sum_{l=1}^{s} x_j^{(l)}\right\|^2\right] \\ &= \frac{1}{s^2}\,\mathbb{E}\left[\left\|\sum_{l=1}^{s} x_j^{(l)}\right\|^2\right] \\ &= \frac{1}{s^2}\sum_{l=1}^{s}\mathbb{E}\left[\left\|x_j^{(l)}\right\|^2\right] + \frac{2}{s^2}\sum_{1\le l_1<l_2\le s}\mathbb{E}\left[x_j^{(l_1)}\cdot x_j^{(l_2)}\right] \\ &= \frac{1}{s^2}\sum_{l=1}^{s}\mathbb{E}\left[\left\|x_j^{(l)}\right\|^2\right] + \frac{2}{s^2}\binom{s}{2}\mathbb{E}\left[x_j^{(l_1)}\right]\cdot\mathbb{E}\left[x_j^{(l_2)}\right] \\ &= \frac{1}{s^2}\sum_{l=1}^{s}\mathbb{E}\left[\left\|x_j^{(l)}\right\|^2\right] + \frac{s-1}{s}\left\|E^T u_j\right\|^2. \end{aligned} \tag{2.3}$$

In (2.3) we used that $x_j^{(l_1)}$ and $x_j^{(l_2)}$ are i.i.d. From (2.2) and (2.3) we have that

$$\mathbb{E}_S\left[\|w_j - \sigma_j v_j\|^2\right] = \frac{1}{s^2}\sum_{l=1}^{s}\mathbb{E}\left[\left\|x_j^{(l)}\right\|^2\right] - \frac{1}{s}\left\|E^T u_j\right\|^2.$$

The definition of P_i gives us

$$\mathbb{E}\left[\left\|x_j^{(l)}\right\|^2\right] = \sum_{i=1}^{m} P_i \frac{\|(u_j)_i\,(a_i - \pi_V(a_i))\|}{P_i^2} \le \|E\|_F^2.$$

Thus, we get a bound on the second central moment of w_j as follows.

$$\mathbb{E}_S\left[\|w_j - \sigma_j v_j\|^2\right] \le \frac{1}{s}\|E\|_F^2. \tag{2.4}$$

With this bound in hand, we can complete the proof. Let $y_j = w_j/\sigma_j$ for $1 \le j \le k$, and consider the matrix $F = A\sum_{i=1}^{k} v_i y_i^T$. The row space of F is contained in $W = \mathrm{span}\,(\{w_1, \ldots, w_k\})$. Therefore, $\|A - \pi_W(A)\|_F^2 \le \|A - F\|_F^2$. We will use F to bound the error $\|A - \pi_W(A)\|_F^2$.

By decomposing $A - F$ along the left singular vectors $u_1, \ldots, u_r$, we can use the inequality (2.4) to bound $\|A - F\|_F^2$:

$$\begin{aligned}
\mathbb{E}_S\left[\|A - \pi_W(A)\|_F^2\right] \leq \mathbb{E}_S\left[\|A - F\|_F^2\right] &= \sum_{i=1}^{r} \mathbb{E}_S\left[\left\|(A-F)^T u_i\right\|_F^2\right] \\
&= \sum_{i=1}^{k} \mathbb{E}_S\left[\|\sigma_i v_i - w_i\|^2\right] + \sum_{i=k+1}^{r} \sigma_i^2 \\
&\leq \frac{k}{s}\|E\|_F^2 + \|A - A_k\|_F^2 . \qquad (2.5)
\end{aligned}$$

□

In general, one can have multiple rounds of adaptive sampling. In each step, we pick a new sample of points with probabilities proportional to their squared distance from the span of our initial subspace and the current sample. Here is the analysis of t rounds of adaptive sampling.

t-round Adaptive Sampling

Input: $A \in \mathbb{R}^{m \times n}$, integer $k \leq m$, initial subspace $V \supseteq \mathbb{R}^n$.
Output: a sample $S = S_1 \cup S_2 \cup \cdots \cup S_t$ of $\sum_{i=1}^{t} s_i$ rows of A, where $|S_i| = s_i$.

1. Start with a linear subspace V. Let $E_0 = A - \pi_V(A)$, and $S = \emptyset$.
2. For $j = 1$ to t, do:
   (a) Pick an i.i.d. sample S_j of s_j rows of A from the following distribution:
   $$P_i^{(j-1)} = \Pr(\text{picking } a_i) \propto \|(E_{j-1})_i\|^2 .$$
   (b) $S = S \cup S_j$.
   (c) $E_j = A - \pi_{\text{span}(V \cup S)}(A)$.

We can now prove the following corollary of Theorem 4, for t rounds of adaptive sampling, using induction on the number of rounds.

Corollary 6. *After t rounds of the adaptive sampling procedure described above where we pick a sample $S = S_1 \cup \cdots \cup S_t$, we have*

$$\begin{aligned}
&\mathbb{E}_S\left[\left\|A - \pi_{\text{span}(V \cup S),k}(A)\right\|_F^2\right] \\
&\leq \left(1 + \frac{k}{s_t} + \frac{k^2}{s_t s_{t-1}} + \ldots + \frac{k^{t-1}}{s_t s_{t-1} \ldots s_2}\right)\|A - A_k\|_F^2 + \frac{k^t}{s_t s_{t-1} \ldots s_1}\|E_0\|_F^2 .
\end{aligned}$$

Proof. We prove the theorem by induction on t. The case $t = 1$ is implied by Theorem 4. For the inductive step, using Theorem 4 with $\text{span}(V \cup S_1 \cup \cdots \cup S_{t-1})$ as our initial subspace, we have

$$\mathbb{E}_{S_t}\left[\left\|A - \pi_{\text{span}(V \cup S),k}(A)\right\|_F^2\right] \leq \|A - A_k\|_F^2 + \frac{k}{s_t}\|E_{t-1}\|_F^2 .$$

Combining this inequality with the fact that

$$\|E_{t-1}\|_F^2 = \left\|A - \pi_{\text{span}(V \cup S_1 \cup \cdots \cup S_{t-1})}(A)\right\|_F^2 \leq \left\|A - \pi_{\text{span}(V \cup S_1 \cup \cdots \cup S_{t-1}),k}(A)\right\|_F^2 ,$$

we get

$$\mathbb{E}_{S_t}\left[\left\|A - \pi_{\text{span}(V \cup S),k}(A)\right\|_F^2\right] \leq \|A - A_k\|_F^2 + \frac{k}{s_t}\left\|A - \pi_{\text{span}(V \cup S_1 \cup \cdots \cup S_{t-1}),k}(A)\right\|_F^2 .$$

Finally, taking the expectation over $S_1, \ldots, S_{t-1}$:

$$\begin{aligned}
&\mathbb{E}_{S_1,\ldots,S_t}\left[\left\|A - \pi_{\text{span}(V \cup S),k}(A)\right\|_F^2\right] \\
&\quad \leq \|A - A_k\|_F^2 + \frac{k}{s_t}\mathbb{E}_{S_1,\ldots,S_{t-1}}\left[\left\|A - \pi_{\text{span}(V \cup S_1 \cup \cdots \cup S_{t-1}),k}(A)\right\|_F^2\right]
\end{aligned}$$

and the result follows from the induction hypothesis for $t - 1$. □

Here we restate the corollary in an easier way. Stating it in the following way highlights an important feature of adaptive sampling: the additive error drops exponentially with the number of rounds.

Corollary 7. *Let $V \subseteq \mathbb{R}^n$ be a subspace such that $\|A - \pi_V(A)\|_F^2 \leq \alpha \|A - A_k\|_F^2$, for some $\alpha \geq 0$. Then, with V as our initial subspace and using t rounds of the adaptive sampling described above, where we pick a sample $S = S_1 \cup \cdots \cup S_t$ such that*

$$s_1 = \cdots = s_{t-1} = 2k, \text{ and } s_t = 4k/\epsilon,$$

we have

$$\mathbb{E}_S\left[\left\|A - \pi_{\text{span}(V \cup S),k}(A)\right\|_F^2\right] \leq \left(1 + \frac{\epsilon}{2} + \frac{\epsilon\alpha}{2^{t+1}}\right)\|A - A_k\|_F^2 .$$

Proof. Immediate by substituting the values in Corollary 6. □

Restating the corollary as above, it is clear that if we can find a good initial subspace with multiplicative guarantee of α, then using $O(\log \alpha)$ rounds of adaptive sampling we can get down to a multiplicative $(1+\epsilon)$ guarantee. Does such a subspace V always exist? Can we find it efficiently? We will try to answer these questions in the next section.

2.5 Volume sampling

Here is another way to generalize the squared-length sampling scheme. We sample subsets of rows instead of individual rows. Let S be a subset of k rows of A, and Δ_S be the simplex formed by these rows and the origin. Volume sampling corresponds to the following distribution: we pick subset S with probability equal to

$$P_S = \frac{\text{vol}(\Delta_S)^2}{\sum_{T,|T|=k} \text{vol}(\Delta_T)^2}.$$

In Theorem 8, we will show that for S picked by volume sampling, $V = \text{span}(S)$ gives an expected multiplicative guarantee of $(k+1)$ for rank-k approximation.

Theorem 8. *Let $A \in \mathbb{R}^{m\times n}$ and S be a k-subset of its rows picked from the following distribution:*

$$P_S = \Pr(\textit{picking } S) \propto \text{vol}(\Delta_S)^2,$$

where Δ_S is the simplex formed by points in S with the origin. Then,

$$\mathbb{E}_S\left[\left\|A - \pi_{\text{span}(S)}(A)\right\|_F^2\right] \leq (k+1)\left\|A - A_k\right\|_F^2.$$

Proof. By expanding the expectation we get

$$\begin{aligned}
\mathbb{E}_S\left[\left\|A - \pi_{\text{span}(S)}A\right\|_F^2\right] &= \sum_{S,|S|=k} \frac{\text{vol}(\Delta_S)^2}{\sum_{T,|T|=k}\text{vol}(\Delta_T)^2}\left\|A - \pi_{\text{span}(S)}(A)\right\|_F^2 \\
&= \sum_{S,|S|=k} \frac{\text{vol}(\Delta_S)^2}{\sum_{T,|T|=k}\text{vol}(\Delta_T)^2}\sum_{i=1}^{m} d(a_i, \text{span}(S))^2 \\
&= \frac{\sum_{S,|S|=k}\sum_{i=1}^{m}(k+1)^2\text{vol}(\Delta_{S\cup\{i\}})^2}{\sum_{T,|T|=k}\text{vol}(\Delta_T)^2} \\
&= (k+1)^3\,\frac{\sum_{S,|S|=k+1}\text{vol}(\Delta_S)^2}{\sum_{T,|T|=k}\text{vol}(\Delta_T)^2} \\
&= (k+1)^3\,\frac{(k+1)!^{-2}\sum_{i_1<i_2<\ldots<i_{k+1}}\sigma_{i_1}^2\sigma_{i_2}^2\cdots\sigma_{i_{k+1}}^2}{k!^{-2}\sum_{i_1<i_2<\ldots<i_k}\sigma_{i_1}^2\sigma_{i_2}^2\cdots\sigma_{i_k}^2} \\
&\qquad \text{(using Lemma 9)} \\
&\leq (k+1)\sum_{i=k+1}^{r}\sigma_i^2 \\
&= (k+1)\left\|A - A_k\right\|_F^2 \qquad \text{(by Proposition 1)}
\end{aligned}$$

□

Now we will prove Lemma 9 that expresses the sum of $\text{vol}(\Delta_S)^2$ over all k-subsets S of the rows of A in terms of the singular values of A.

Lemma 9.

$$\sum_{S,|S|=k} \mathrm{vol}(\Delta_S)^2 = \frac{1}{(k!)^2} \sum_{i_1<i_2<\cdots<i_k} \sigma_{i_1}^2 \sigma_{i_2}^2 \cdots \sigma_{i_k}^2.$$

Proof.

$$\sum_{S,|S|=k} \mathrm{vol}(\Delta_S)^2 = \sum_{S,|S|=k} \frac{1}{(k!)^2} \det(A_S A_S^T)$$

where A_S is a submatrix of A given by rows in S. Now consider the char. poly. of AA^T.

$$det(\lambda I - AA^T) = (\lambda - \sigma_1^2)(\lambda - \sigma_2^2)\cdots(\lambda - \sigma_m^2),$$

where $\sigma_1 \geq \ldots \geq \sigma_r \geq \sigma_{r+1} = \ldots = \sigma_m = 0$. Comparing the coefficients of λ^{m-k} from both sides (using Lemma 10 for the LHS), we get

$$\sum_{S,|S|=k} \det(A_S A_S^T) = \sum_{i_1<\ldots<i_k} \sigma_{i_1}^2 \cdots \sigma_{i_k}^2.$$

Therefore,

$$\sum_{S,|S|=k} \mathrm{vol}(\Delta_S)^2 = \frac{1}{(k!)^2} \sum_{i_1<i_2<\cdots<i_k} \sigma_{i_1}^2 \sigma_{i_2}^2 \cdots \sigma_{i_k}^2.$$

□

Following is a simple linear algebraic lemma that can be found in standard textbooks on matrices and determinants (see [1], for example). It expresses the coefficients of the characteristic polynomial of a square matrix M in terms of the determinants of the principal minors of M.

Lemma 10. *Let the characteristic polynomial of $M \in \mathbb{R}^{m\times m}$ be*

$$\det(\lambda I_m - M) = \lambda^m + c_{m-1}\lambda^{m-1} + \cdots + c_0.$$

Then

$$c_k = (-1)^{m-k} \sum_{\substack{B:\ \textit{principal} \\ \textit{k-minor of } AA^T}} \det(B) \qquad \textit{for } 1 \leq k \leq m.$$

The bound proved in Theorem 8 is in fact asymptotically tight as shown in the following proposition.

Proposition 11. *Given any $\epsilon > 0$, there exists a $(k+1) \times (k+1)$ matrix A such that for any k-subset S of rows of A,*

$$\|A - \pi_{S,k}(A)\|_F^2 \geq (1-\epsilon)\ (k+1)\ \|A - A_k\|_F^2.$$

Proof. The tight example consists of a matrix $A \in \mathbb{R}^{(k+1)\times(k+1)}$ whose rows $a_1, a_2, \ldots, a_k$ are the vertices of a regular k-dimensional simplex with sides of length $\sqrt{2}$, lying on the affine hyperplane $\{X_{k+1} = \alpha\}$ in $\mathbb{R}^{k+1}$ and having $p = (0, 0, \ldots, 0, \alpha)$ as their

centroid. For α small enough, the best k dimensional subspace for these points is given by $\{X_{k+1} = 0\}$ and

$$\|A - A_k\|_F^2 = (k+1)\alpha^2.$$

Consider any k-subset of rows from these, say $S = \{a_1, a_2, \ldots, a_k\}$, and let H_S be the linear subspace spanning them. Then,

$$\left\|A - \pi_{\text{span}(S),k}(A)\right\|_F^2 = d(a_{k+1}, H_S)^2.$$

We can express the volume of the simplex Conv $(\bar{0}, a_1, \ldots, a_k)$ in two different ways

$$\text{vol}(\text{Conv}\,(\bar{0}, a_1, \ldots, a_k)) = \alpha \text{vol}(\text{Conv}\,(p, a_1, a_2, \ldots, a_k)),$$

as well as

$$\text{vol}(\text{Conv}\,(\bar{0}, a_1, \ldots, a_k)) = \frac{1}{k+1} d(a_{k+1}, H_S) \text{vol}(\text{Conv}\,(\bar{0}, a_1, \ldots, a_k)).$$

Therefore, we get

$$d(a_{k+1}, H_S) = (k+1)\alpha \cdot \frac{\text{vol}(\text{Conv}\,(p, a_1, a_2, \ldots, a_k))}{\text{vol}(\text{Conv}\,(\bar{0}, a_1, \ldots, a_k))}.$$

Hence, for any $\epsilon > 0$, we can choose α small enough so that

$$d(a_{k+1}, H_S) \geq \sqrt{(1-\epsilon)}(k+1)\alpha.$$

Choose α small enough so that $d(p, H_S) \geq \sqrt{(1-\epsilon)}\alpha$. Now

$$\frac{d(a_{k+1}, H_S)}{d(p, H_S)} = \frac{d(a_{k+1}, \text{Conv}\,(a_1, \ldots, a_k))}{d(p, \text{Conv}\,(a_1, \ldots, a_k))} = k+1,$$

since the points form a simplex and p is their centroid. The claim follows. Hence,

$$\left\|A - \pi_{\text{span}(S),k}(A)\right\|_F^2 = d(a_{k+1}, H_S)^2 \geq (1-\epsilon)\,(k+1)^2\,\alpha^2 = (1-\epsilon)\,(k+1)\,\|A - A_k\|_F^2.$$

□

2.5.1 Intuition from exterior algebra

Here is some intuition from exterior algebra that explains why volume sampling works. Given a matrix $A \in \mathbb{R}^{m \times n}$ with rows $a_1, a_2, \ldots, a_m$, consider a new matrix A' of size $\binom{m}{k} \times \binom{n}{k}$ as follows.

- The rows of A' are indexed by all k-subsets S of the rows of A. For $S = \{a_{i_1}, a_{i_2}, \ldots, a_{i_k}\}$, the corresponding row A'_S is just the wedge product $a_{i_1} \wedge a_{i_2} \wedge \cdots \wedge a_{i_k} \in \bigwedge^k \mathbb{R}^n$, whose construction is described below.
- $a_{i_1} \wedge a_{i_2} \wedge \cdots \wedge a_{i_k}$ is an $\binom{n}{k}$-dimensional vector whose coordinates are indexed

by k-subsets T of the columns of A. The T-th coordinate is given by $\det(A_{S,T})$, where $A_{S,T}$ is the $k \times k$ submatrix of A given by rows in S and columns in T.

- For a k-subset S, let A_S be the $k \times n$ submatrix given by the rows in S. Some easy observations show that for any two k-subsets S_1 and S_2,

$$\langle A'_{S_1}, A'_{S_2} \rangle = \det(A_{S_1} A_{S_2}^T),$$

and

$$\|A'_S\|^2 = \det(A_S A_S^T) = (k!)^2 \mathrm{vol}(\Delta_S)^2.$$

- The singular values of A' turn out to be products of k singular values of A, e.g., $\sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_k}$ for $i_1 < i_2 < \cdots < i_k$. And the right singular vectors are wedge products of k right singular vectors of A, e.g., $v_{i_1} \wedge v_{i_2} \wedge \cdots v_{i_k}$, for $i_1 < i_2 < \cdots < i_k$ (whose construction is similar to that of $a_{i_1} \wedge a_{i_2} \wedge \cdots \wedge a_{i_k}$ described above).

Hence, the topmost singular value of A' is $\sigma_1\sigma_2\cdots\sigma_k$ and its corresponding right singular vector is $v_1 \wedge v_2 \wedge \cdots v_k$. Moreover, squared-length sampling on the rows of A' is equivalent to volume sampling on k-subsets of rows of A. Thus, doing volume sampling to find the optimal k-dimensional subspace $\mathrm{span}(\{v_1, v_2, \ldots, v_k\})$ for the points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$ is equivalent to doing squared-length sampling to find the optimal 1-dimensional subspace $\mathrm{span}(v_1 \wedge v_2 \wedge \cdots \wedge v_k)$ for the points $\{A'_S \in \mathbb{R}^{\binom{n}{k}} : |S| = k\}$.

2.5.2 Approximate volume sampling

Here we give an algorithm for approximate volume sampling. In brief, we run a k-round adaptive sampling procedure, picking one row in each round.

Approximate Volume Sampling

1. Initialize $S = \emptyset$. While $|S| < k$ do:
   (a) Pick a point a_i from the following distribution:
   $$\Pr(\text{picking } a_i) \propto d(a_i, \mathrm{span}(S))^2.$$
   (b) $S = S \cup \{a_i\}$.
2. Output the k-subset S.

Next we show that the above procedure gives an approximate implementation of volume sampling.

Proposition 12. *Let P_S be the probability of picking a k-subset S according to volume sampling. Then the k-round adaptive procedure mentioned above picks a subset S with probability $\tilde{P}_S$ such that*

$$\tilde{P}_S \leq k! \cdot P_S$$

Proof. Let $S = \{a_{i_1}, a_{i_2}, \ldots, a_{i_k}\}$ be a subset of k rows, and let $\tau \in \Pi_k$, the set of all permutations of $\{i_1, i_2, \ldots, i_k\}$. By notation $H_{\tau,t}$, we denote the linear subspace $\text{span}(a_{\tau(i_1)}, a_{\tau(i_2)}, \ldots, a_{\tau(i_t)})$, and by $d(a_i, H_{\tau,t})$ we denote the orthogonal distance of point a_i from this subspace. Our adaptive procedure picks a subset S with probability equal to

$$\begin{aligned}
\tilde{P}_S &= \sum_{\tau\in\Pi_k} \frac{\left\|a_{\tau(i_1)}\right\|^2}{\|A\|_F^2} \frac{d(a_{\tau(i_2)}, H_{\tau,1})^2}{\sum_{i=1}^m d(a_i, H_{\tau,1})^2} \cdots \frac{d(a_{\tau(i_k)}, H_{\tau,k-1})^2}{\sum_{i=1}^m d(a_i, H_{\tau,k-1})^2} \\
&\leq \frac{\sum_{\tau\in\Pi_k} \left\|a_{\tau(i_1)}\right\|^2 d(a_{\tau(i_2)}, H_{\tau,1})^2 \cdots d(a_{\tau(i_k)}, H_{\tau,k-1})^2}{\|A\|_F^2 \|A - A_1\|_F^2 \cdots \|A - A_{k-1}\|_F^2} \\
&= \frac{\sum_{\tau\in\Pi_k} (k!)^2 \text{vol}(\Delta(S))^2}{\|A\|_F^2 \|A - A_1\|_F^2 \cdots \|A - A_{k-1}\|_F^2} \\
&= \frac{(k!)^3 \text{ vol}(\Delta(S))^2}{\sum_{i=1}^m \sigma_i^2 \sum_{i=2}^m \sigma_i^2 \cdots \sum_{i=k}^m \sigma_i^2} \\
&\leq \frac{(k!)^3 \text{ vol}(\Delta(S))^2}{\sum_{1\leq i_1<i_2<\ldots<i_k\leq m} \sigma_{i_1}^2 \sigma_{i_2}^2 \cdots \sigma_{i_k}^2} \\
&= \frac{k! \text{ vol}(\Delta(S))^2}{\sum_{T:|T|=k} \text{vol}(\Delta(T))^2} \qquad \text{(using Lemma 9)} \\
&= k! \cdot P_S
\end{aligned}$$

□

Now we will show why it suffices to have just the approximate implementation of volume sampling. If we sample subsets S with probabilities $\tilde{P}_S$ instead of P_S, we get an analog of Theorem 8 with a weaker multiplicative approximation.

Proposition 13. *If we sample a subset S of k rows using the k-round adaptive sampling procedure mentioned above, then*

$$\mathbb{E}_S \left[\left\|A - \pi_{\text{span}(S)}(A)\right\|_F^2\right] \leq (k+1)! \ \|A - A_k\|_F^2 .$$

Proof. Since we are picking a subset S with probability $\tilde{P}_S$ the expected error is

$$\begin{aligned}
\mathbb{E}_S \left[\left\|A - \pi_{\text{span}(S)}(A)\right\|_F^2\right] &= \sum_{S,|S|=k} \tilde{P}_S \left\|A - \pi_{\text{span}(S)}(A)\right\|_F^2 \\
&\leq k! \sum_{S,|S|=k} P_S \left\|A - \pi_{\text{span}(S)}(A)\right\|_F^2 \\
&\leq k! \ (k+1) \|A - A_k\|_F^2 \qquad \text{(using Theorem 8)} \\
&= (k+1)! \ \|A - A_k\|_F^2
\end{aligned}$$

□

2.6 Fast algorithm for low-rank approximation

This section has two parts. In Subsection 2.6.1, we will assume the existence of k rows whose span gives a $(k+1)$-approximation for rank-k approximation, as guaranteed by volume sampling. We will use the span of these k rows as our initial subspace and then use $O(\log k)$ rounds of adaptive sampling to bring the multiplicative error down to $(1+\epsilon)$. We will prove that, for any real matrix, there exist $O(k/\epsilon + k\log k)$ rows whose span contains the rows of another rank-k matrix that has error at most $(1+\epsilon)$ times the error of the optimal rank-k approximation.

However our inability to do volume sampling exactly leads to a weaker algorithmic result that uses $O(k/\epsilon + k^2\log k)$ rows instead. We will discuss this in Subsection 2.6.2.

2.6.1 Existence result

Theorem 14. *Any $m \times n$ matrix A contains a subset S of $4k/\epsilon + 2k\log(k+1)$ rows such that there is a matrix $\tilde{A}$ of rank at most k whose rows lie in* $\mathrm{span}(S)$ *and*

$$\left\|A - \tilde{A}\right\|_F^2 \leq (1+\epsilon)\left\|A - A_k\right\|_F^2.$$

Proof. From Theorem 8, we know that there exists a subset S_0 of k rows of A such that

$$\left\|A - \pi_{\mathrm{span}(S_0)}(A)\right\|_F^2 \leq (k+1)\left\|A - A_k\right\|_F^2.$$

Let $V = \mathrm{span}(S_0)$, $t = \log(k+1)$, $\alpha = k+1$ in Corollary 7, we know that there exist subsets $S_1, \ldots, S_t$ of rows with sizes $s_1 = \ldots = s_{t-1} = 2k$ and $s_t = 4k/\epsilon$, respectively, such that

$$\begin{aligned}\left\|A - \pi_{\mathrm{span}(V \cup S_1 \cup \ldots \cup S_t),k}(A)\right\|_F^2 &\leq \left(1 + \frac{\epsilon}{2} + \frac{\epsilon}{2}\right)\left\|A - A_k\right\|_F^2 \\ &= (1+\epsilon)\left\|A - A_k\right\|_F^2\end{aligned}$$

Therefore, for $S = S_0 \cup S_1 \cup \ldots \cup S_t$ we have

$$|S| \leq \sum_{j=0}^{t} |S_j| = k + 2k(\log(k+1) - 1) + \frac{4k}{\epsilon} \leq \frac{4k}{\epsilon} + 2k\log(k+1)$$

and

$$\left\|A - \pi_{\mathrm{span}(S'),k}(A)\right\|_F^2 \leq (1+\epsilon)\left\|A - A_k\right\|_F^2.$$

□

2.6.2 Fast algorithm for low-rank approximation

In this section we describe an algorithm that given a matrix $A \in \mathbb{R}^{m\times n}$, finds another matrix $\tilde{A}_k$ of rank at most k such that $\left\|A - \tilde{A}_k\right\|_F^2 \leq (1+\epsilon)\left\|A - A_k\right\|_F^2$. The algorithm

has two phases. In the first phase, we pick a subset of k rows using the approximate volume sampling procedure described in Subsection 2.5.2. In the second phase, we use the span of these k rows as our initial subspace and perform $(k+1)\log(k+1)$ rounds of adaptive sampling. The rows chosen are all from the original matrix A.

Linear Time Low-Rank Matrix Approximation

Input: $A \in \mathbb{R}^{m \times n}$, integer $k \leq m$, error parameter $\epsilon > 0$.
Output: $\tilde{A}_k \in \mathbb{R}^{m \times n}$ of rank at most k.

1. Pick a subset S_0 of k rows of A using the approximate volume sampling procedure described in Subsection 2.5.2. Compute an orthonormal basis $\mathcal{B}_0$ of $\text{span}(S_0)$.
2. Initialize $V = \text{span}(S_0)$. Fix parameters as $t = (k+1)\log(k+1)$, $s_1 = s_2 = \ldots = s_{t-1} = 2k$, and $s_t = 16k/\epsilon$.
3. Pick subsets of rows $S_1, S_2, \ldots, S_t$, using t-round adaptive sampling procedure described in Subsection 2.4. After round j, extend the previous orthonormal basis $\mathcal{B}_{j-1}$ to an orthonormal basis $\mathcal{B}_j$ of $\text{span}(S_0 \cup S_1 \cup \ldots \cup S_j)$.
4. $S = \bigcup_{j=0}^{t} S_j$, and we have an orthonormal basis $\mathcal{B}_t$ of $\text{span}(S)$.
5. Compute $h_1, h_2, \ldots, h_k$, the top k right singular vectors of $\pi_{\text{span}(S)}(A)$.
6. Output matrix $\tilde{A}_k = \pi_{\text{span}(h_1,\ldots,h_k)}(A)$, written in the standard basis.

Here are some details about the implementations of these steps.

In Step 1, we use the k-round adaptive procedure for approximate volume sampling. In the j-th round of this procedure, we sample a row and compute its component v_j orthogonal to the span of the rows picked in rounds $1, 2, \ldots, j-1$. The residual squared lengths of the rows are computed using $\left\|E_j^{(i)}\right\|^2 = \left\|E_{j-1}^{(i)}\right\|^2 - A^{(i)} \cdot v_j$, and $\|E_j\|_F^2 = \|E_{j-1}\|_F^2 - \|Av_j\|^2$. In the end, we have an orthonormal basis $\mathcal{B}_0 = \{v_1/\|v_1\|, \ldots, v_k/\|v_k\|\}$.

In Step 3, there are $(k+1)\log(k+1)$ rounds of adaptive sampling. In the j-th round, we extend the orthonormal basis from $\mathcal{B}_{j-1}$ to $\mathcal{B}_j$ by Gram-Schmidt orthonormalization. We compute the residual squared lengths of the rows $\left\|E_j^{(i)}\right\|^2$, as well as the total, $\|E_j\|_F^2$, by subtracting the contribution $\pi_{\text{span}(B_j \setminus B_{j-1})}(A)$ from the values that they had during the previous round.

Each round in Steps 1 and 3 can be implemented using 2 passes over A: one pass to figure out the sampling distribution, and an another one to sample a row (or a subset of rows) according to this distribution. So Steps 1 and 3 require $2(k+1)\log(k+1)+2k$ passes.

Finally, in Step 5, we compute $\pi_{\text{span}(S)}(A)$ in terms of basis $\mathcal{B}_t$ using one pass (now we have an $m \times O(k/\epsilon + k^2 \log k)$ matrix), and we compute its top k right singular vectors using SVD. In Step 6, we rewrite them in the standard basis and project matrix A onto their span, which requires one additional pass.

So the total number of passes is $2(k+1)(\log(k+1)+1)$.

Theorem 15. *For any given $A \in \mathbb{R}^{m\times n}$, the above algorithm outputs, with probability at least $3/4$, another matrix $\tilde{A}_k$ of rank at most k such that*

$$\left\|A - \tilde{A}_k\right\|_F^2 \leq (1+\epsilon)\left\|A - A_k\right\|_F^2.$$

Moreover, the algorithm takes

$$O\left(M\left(\frac{k}{\epsilon} + k^2 \log k\right) + (m+n)\left(\frac{k^2}{\epsilon^2} + \frac{k^3 \log k}{\epsilon} + k^4 \log^2 k\right)\right)$$

time, $O\left(\min\{m,n\}(\frac{k}{\epsilon} + k^2 \log k)\right)$ space, and can be implemented using only $O(k \log k)$ passes over A.

Proof. We begin with a proof of correctness. After the first phase of approximate volume sampling, using Proposition 13, we have

$$\mathbb{E}_{S_0}\left[\left\|A - \pi_{\text{span}(S_0)}(A)\right\|_F^2\right] \leq (k+1)!\ \left\|A - A_k\right\|_F^2.$$

Now using $V = \text{span}(S_0)$, $t = (k+1)\log(k+1)$, $s_t = 16k/\epsilon$, $s_{t-1} = \ldots = s_1 = 2k$ in Corollary 6 we get that

$$\begin{aligned}
&\mathbb{E}_{S_1,\ldots,S_t}\left[\left\|A - \pi_{\text{span}(S),k}(A)\right\|_F^2\right] \\
&\quad\leq \left(1 + \frac{\epsilon}{16} + \frac{\epsilon}{32} + \ldots\right)\left\|A - A_k\right\|_F^2 + \frac{\epsilon}{2^{t+3}}\ \left\|A - \pi_{\text{span}(S_0)}(A)\right\|_F^2 \\
&\quad\leq \left(1 + \frac{\epsilon}{8}\right)\ \left\|A - A_k\right\|_F^2 + \frac{\epsilon}{8\cdot 2^t}\ \left\|A - \pi_{\text{span}(S_0)}(A)\right\|_F^2.
\end{aligned}$$

Now taking expectation over S_0 we have

$$\begin{aligned}
&\mathbb{E}_{S_0,\ldots,S_t}\left[\left\|A - \pi_{\text{span}(S),k}(A)\right\|_F^2\right] \\
&\quad\leq \left(1 + \frac{\epsilon}{8}\right)\ \left\|A - A_k\right\|_F^2 + \frac{\epsilon}{8\cdot 2^t}\ \mathbb{E}_{S_0}\left[\left\|A - \pi_{\text{span}(S_0)}(A)\right\|_F^2\right] \\
&\quad\leq \left(1 + \frac{\epsilon}{8}\right)\ \left\|A - A_k\right\|_F^2 + \frac{\epsilon}{8\cdot 2^t}\ (k+1)!\ \left\|A - A_k\right\|_F^2 \\
&\quad\leq \left(1 + \frac{\epsilon}{8}\right)\ \left\|A - A_k\right\|_F^2 + \frac{\epsilon}{8\cdot 2^t}\ (k+1)^{(k+1)}\ \left\|A - A_k\right\|_F^2 \\
&\quad\leq \left(1 + \frac{\epsilon}{8}\right)\ \left\|A - A_k\right\|_F^2 + \frac{\epsilon}{8}\ \left\|A - A_k\right\|_F^2 \\
&\quad= \left(1 + \frac{\epsilon}{4}\right)\left\|A - A_k\right\|_F^2.
\end{aligned}$$

This means

$$\mathbb{E}_{S_0,\ldots,S_t}\left[\left\|A-\pi_{\mathrm{span}(S),k}(A)\right\|_F^2-\|A-A_k\|_F^2\right]\leq\frac{\epsilon}{4}\|A-A_k\|_F^2.$$

Therefore, using Markov's inequality, with probability at least 3/4 the algorithm gives a matrix $\tilde{A}_k=\pi_{\mathrm{span}(S),k}(A)$ satisfying

$$\left\|A-\tilde{A}_k\right\|_F^2\leq(1+\epsilon)\|A-A_k\|_F^2.$$

Now let us analyze its complexity.

Step 1 has k rounds of adaptive sampling. In each round, the matrix-vector multiplication requires $O(M)$ time and storing vector v_j requires $O(n)$ space. So overall, Step 1 takes $O(Mk+nk)$ time, $O(nk)$ space.

Step 3 has $2(k+1)\log(k+1)$ rounds of adaptive sampling. The j-th round (except for the last round), involves Gram-Schmidt orthonormalization of $2k$ vectors in $\mathbb{R}^n$ against an orthonormal basis of size at most $(2j+1)k$, which takes time $O(njk^2)$. Computing $\pi_{\mathrm{span}(B_j\setminus B_{j-1})}(A)$ for updating the values $\left\|E_j^{(i)}\right\|^2$ and $\|E_j\|_F^2$ takes time $O(Mk)$. Thus, the total time for the j-th round is $O(Mk+njk^2)$. In the last round, we pick $O(k/\epsilon)$ rows. The Gram-Schmidt orthonormalization of these $O(k/\epsilon)$ vectors against an orthonormal basis of $O(k^2\log k)$ vectors takes $O(nk^3\log k/\epsilon)$ time; storing this basis requires $O(nk/\epsilon+nk^2\log k)$ space. So overall, Step 3 takes $O\left(Mk^2\log k+n(k^3\log k/\epsilon+k^4\log^2 k)\right)$ time and $O(nk/\epsilon+nk^2\log k)$ space (to store the basis $\mathcal{B}_t$).

In Step 5, projecting A onto $\mathrm{span}(S)$ takes $O\left(M(k/\epsilon+k^2\log k)\right)$ time. Now we have $\pi_{\mathrm{span}(S)}(A)$ in terms of our basis $\mathcal{B}_t$ (which is a $m\times O(k^2\log k+k/\epsilon)$ matrix) and computation of its top k right singular vectors takes time $O\left(m(k/\epsilon+k^2\log k)^2\right)$.

In Step 6, rewriting $h_1,h_2,\ldots,h_k$ in terms of the standard basis takes time $O\left(n(k^3\log k+k^2/\epsilon)\right)$. And finally, projecting the matrix A onto $\mathrm{span}(h_1,\ldots,h_k)$ takes time $O(Mk)$.

Putting it all together, the algorithm takes

$$O\left(M\left(\frac{k}{\epsilon}+k^2\log k\right)+(m+n)\left(\frac{k^2}{\epsilon^2}+\frac{k^3\log k}{\epsilon}+k^4\log^2 k\right)\right)$$

time and $O\left(\min\{m,n\}(k/\epsilon+k^2\log k)\right)$ space (since we can do the same with columns instead of rows), and $O(k\log k)$ passes over the data. □

This algorithm can be made to work with high probability, by running independent copies of the algorithm in each pass and taking the best answer found at the end. The overhead to get a probability of success of $1-\delta$ is $O(\sqrt{\log(1/\delta)})$.

2.7 Lower bound for low-rank approximation

Here we show a lower bound of $\Omega(k/\epsilon)$ for rank-k approximation using a subset of rows. Thus, our sampling-based existence result in Theorem 14 is almost tight.

Proposition 16. *Given $\epsilon > 0$ and n large enough so that $n\epsilon \geq 2$, there exists an $n \times (n+1)$ matrix A such that for any subset S of its rows with $|S| \leq 1/2\epsilon$,*

$$\left\|A - \pi_{\text{span}(S),1}(A)\right\|_F^2 \geq (1+\epsilon)\left\|A - A_1\right\|_F^2$$

Proof. Let $e_1, e_2, \ldots, e_{n+1}$ be the standard basis for $\mathbb{R}^{n+1}$, considered as rows. Consider the $n \times (n+1)$ matrix A, whose i-th row is given by $A^{(i)} = e_1 + \epsilon\, e_{i+1}$, for $i = 1, 2, \ldots, n$. The best rank-1 approximation for this is A_1, whose i-th row is given by$A_1^{(i)} = e_1 + \sum_{i=1}^n \frac{1}{n} e_{i+1}$. Therefore,

$$\|A - A_1\|_F^2 = \sum_{i=1}^n \left\|A^{(i)} - A_1^{(i)}\right\|^2 = n\left(\frac{(n-1)^2\epsilon^2}{n^2} + (n-1)\frac{\epsilon^2}{n^2}\right) = (n-1)\epsilon^2.$$

Now let S be any subset of the rows with $|S| = s$. It is easy to see that the best rank-1 approximation for A in the span of S is given by $\pi_{\text{span}(S),1}(A)$, whose i-th row is given by $\pi_{\text{span}(S),1}(A)^{(i)} = e_1 + \frac{\epsilon}{s}\sum_{i\in S} e_{i+1}$, for all i (because it has to be a symmetric linear combination of them). Hence,

$$\begin{aligned}
\left\|A - \pi_{\text{span}(S),1}(A)\right\|_F^2 &= \sum_{i\in S}\left\|A^{(i)} - \pi_{\text{span}(S),1}(A)^{(i)}\right\|^2 + \sum_{i\notin S}\left\|A^{(i)} - \pi_{\text{span}(S),1}(A)^{(i)}\right\|^2 \\
&= s\left(\frac{(s-1)^2\epsilon^2}{s^2} + (s-1)\frac{\epsilon^2}{s^2}\right) + (n-s)\left(s\,\frac{\epsilon^2}{s^2} + \epsilon^2\right) \\
&= \frac{(s-1)^2\epsilon^2}{s} + \frac{(s-1)\epsilon^2}{s} + \frac{n\epsilon^2}{s} + n\epsilon^2 - \epsilon^2 - s\epsilon^2 \\
&= \frac{n\epsilon^2}{s} + n\epsilon^2 - 2\epsilon^2.
\end{aligned}$$

Now if $s \leq \frac{1}{2\epsilon}$ then $\left\|A - \pi_{\text{span}(S),1}(A)\right\|_F^2 = (1+2\epsilon)n\epsilon^2 - 2\epsilon^2 \geq (1+\epsilon)n\epsilon^2 \geq (1+\epsilon)\left\|A - A_1\right\|_F^2$, for n chosen large enough so that $n\epsilon \geq 2$. □

Now we will try to extend this lower bound for relative rank-k approximation.

Proposition 17. *Given $\epsilon > 0$, k, and n large enough so that $n\epsilon \geq 2k$, there exists a $kn \times k(n+1)$ matrix B such that for any subset S of its rows with $|S| \leq k/2\epsilon$,*

$$\left\|B - \pi_{\text{span}(S),k}(A)\right\|_F^2 \geq (1+\epsilon)\left\|B - B_k\right\|_F^2.$$

Proof. Consider B to be a $kn \times k(n+1)$ block-diagonal matrix with k blocks, where each of the blocks is equal to A defined as in Proposition 16 above. It is easy to see that

$$\|B - B_k\|_F^2 = k\left\|A - A_1\right\|_F^2.$$

Now pick any subset S of rows with $|S| \leq \frac{k}{2\epsilon}$. Let S_i be the subset of rows taken from the i-th block, and let $|S_i| = \frac{k}{2\epsilon_i}$. We know that $\sum_{i=1}^{k} |S_i| = \sum_{i=1}^{k} \frac{k}{2\epsilon_i} \leq \frac{k}{2\epsilon}$, and hence $n\epsilon_i \geq n\epsilon \geq 2$.

Therefore,

$$\begin{aligned}
\left\|B - \pi_{\mathrm{span}(S),k}(B)\right\|_F^2 &= \sum_{i=1}^{k} \left\|A - \pi_{\mathrm{span}(S_i),1}(A)\right\|_F^2 \\
&\geq \sum_{i=1}^{k} (1 + \frac{\epsilon_i}{k}) \left\|A - A_1\right\|_F^2 \qquad \text{(using Proposition 16)} \\
&= (k + \frac{\sum_{i=1}^{k} \epsilon_i}{k}) \left\|A - A_1\right\|_F^2 \\
&\geq (k + \frac{k}{\sum_{i=1}^{k} 1/\epsilon_i}) \left\|A - A_1\right\|_F^2 \qquad \text{(by A.M.-H.M. inequality)} \\
&\geq (k + k\epsilon) \left\|A - A_1\right\|_F^2 \\
&= k(1 + \epsilon) \left\|A - A_1\right\|_F^2 \\
&= (1 + \epsilon) \left\|B - B_k\right\|_F^2 .
\end{aligned}$$

□

Chapter 3

Subspace Approximation

Subspace approximation problem is a generalization of low-rank matrix approximation. Here we want to find a k-dimensional linear subspace that minimizes the sum of p-th powers of distances to given points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, for $p \geq 1$, i.e., Given points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$ and $k > 0$, we want to find a k-dimensional linear subspace H that minimizes the L_p-error

$$\left(\sum_{i=1}^{m} d(a_i, H)^p \right)^{\frac{1}{p}}.$$

We denote an optimal subspace by H_k^*.

When $p \neq 2$ we have neither the luxury of a tool like SVD, nor any simple description of an optimal subspace (such as the span of top few right singular vectors). We will show that one can get around this difficulty by generalizing and modifying some of the sampling techniques used in low-rank matrix approximation. Our proofs are of geometric nature though, significantly different from the linear algebraic tools used in low-rank matrix approximation. For a recent review of related work on the subspace approximation problem, including the cases $p = 2$ and $p = \infty$ (where we want a subspace that minimizes the maximum distance to the points), we refer the reader to [22].

We obtain a bi-criteria algorithm for subspace approximation: a randomized algorithm that runs in $\tilde{O}\left(mn \cdot k^3(k/\epsilon)^{p+1}\right)$ time and finds a $\tilde{O}\left(k^2(k/\epsilon)^{p+1}\right)$-dimensional subspace whose error is, with a probability of at least $1/2$, at most $(1+\epsilon)$ times the error of an optimal k-dimensional subspace, (Note: We use the notation $\tilde{O}(\cdot)$ to hide small polylog$(k, 1/\epsilon)$ factors for the convenience of readers.) We obtain our results in several steps, using techniques that we believe are of interest:

1. In Section 3.1, we prove that the span of k points picked using *volume sampling* has expected error at most $(k+1)$ times the optimum. Since we do not know how to do volume sampling exactly in an efficient manner, Section 3.1.2 describes an efficient procedure to implement volume sampling approximately with a weaker multiplicative guarantee of $k! \cdot (k+1)$.

2. In Section 3.2, we show how sampling points proportional to their lengths (or distances from the span of current sample) can be used to find a $\tilde{O}\left(k(k/\epsilon)^{p+1}\right)$-dimensional subspace that gives an additive $\epsilon \left(\sum_{i=1}^{m} \|a_i\|^p\right)^{1/p}$ approximation to an optimal k-dimensional subspace.

3. We call this method of picking new points with probabilities proportional to their distances from the span of current sample as *adaptive sampling*. In Section 3.3, we show that if we start with an initial subspace V, then using *adaptive sampling* we can find $\tilde{O}\left(k(k/\epsilon)^{p+1}\right)$ additional points so that the span of V with these additional points gives an additive $\epsilon \left(\sum_{i=1}^{m} d(a_i, V)^p\right)^{1/p}$ approximation to an optimal k-dimensional subspace. Moreover, using t rounds of the above procedure, this additive error is brought down to $\epsilon^t \left(\sum_{i=1}^{m} d(a_i, V)^p\right)^{1/p}$. The ideas used in this section are adaptations of previous work for the $p = 2$ case.

4. Using $O(k \log k)$ rounds of the above procedure on the initial subspace V obtained by *approximate volume sampling* (from Procedure 1 above), we get our bi-criteria result.

3.1 Volume sampling

In this section, we show how to find a k-subset of the given points such that their span gives a crude but reasonable approximation to the optimal k-dimensional subspace H_k^* that minimizes the sum of p-th powers of distances to the given points.

For any subset $S \subseteq [m]$, we define H_S to be the linear subspace, $\text{span}(\{a_i : i \in S\})$, and Δ_S to be the simplex, $\text{Conv}(\{\bar{0}\} \cup \{a_i : i \in S\})$. By *volume sampling* k-subsets of $[m]$, we mean sampling from the following probability distribution:

$$\Pr(\text{picking } S) = P_S = \frac{\text{vol}(\Delta_S)^p}{\sum_{T,|T|=k} \text{vol}(\Delta_T)^p}.$$

3.1.1 $(k+1)$-approximation using k points

Theorem 18. *For any $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, if we pick a random k-subset $S \subseteq [m]$ by volume sampling then*

$$\mathbb{E}_S \left[\sum_{i=1}^{m} d(a_i, H_S)^p \right] \leq (k+1)^p \sum_{i=1}^{m} d(a_i, H_k^*)^p.$$

Proof.

$$\begin{aligned}
\mathbb{E}_S\left[\sum_{i=1}^{m} d(a_i, H_S)^p\right] &= \sum_{S,|S|=k} \frac{\mathrm{vol}(\Delta_S)^p}{\sum_{T,|T|=k}\mathrm{vol}(\Delta_T)^p} \sum_{i=1}^{m} d(a_i, H_S)^p \\
&= \frac{\sum_{S,|S|=k}\sum_{i=1}^{m}(k+1)^p\ \mathrm{vol}(\Delta_{S\cup\{i\}})^p}{\sum_{T,|T|=k}\mathrm{vol}(\Delta_T)^p} \\
&= \frac{(k+1)^{p+1}\sum_{S,|S|=k+1}\mathrm{vol}(\Delta_S)^p}{\sum_{T,|T|=k}\mathrm{vol}(\Delta_T)^p} \qquad (3.1)
\end{aligned}$$

For any $(k+1)$-subset S, let V_S denote an arbitrary but fixed k-dimensional linear subspace of H_S containing the projection of H_k^* on to H_S. Now for any $(k+1)$-subset S, Lemma 19 gives

$$\mathrm{vol}(\Delta_S) \le \frac{1}{(k+1)}\sum_{i\in S} d(a_i, V_S)\ \mathrm{vol}\left(\Delta_{S\setminus\{i\}}\right).$$

Hence, taking p-th power we have

$$\begin{aligned}
\mathrm{vol}(\Delta_S)^p &\le \frac{1}{(k+1)^p}\left(\sum_{i\in S} d(a_i, V_S)\ \mathrm{vol}\left(\Delta_{S\setminus\{i\}}\right)\right)^p \\
&\le \frac{1}{(k+1)^p}\ (k+1)^{p-1}\sum_{i\in S} d(a_i, V_S)^p\ \mathrm{vol}\left(\Delta_{S\setminus\{i\}}\right)^p \\
&\qquad\qquad \text{(by Hölder's inequality)} \\
&\le \frac{1}{(k+1)}\sum_{i\in S} d(a_i, V_S)^p\ \mathrm{vol}\left(\Delta_{S\setminus\{i\}}\right)^p
\end{aligned}$$

Summing up over all subsets S of size $(k+1)$ we get

$$\begin{aligned}
\sum_{S,|S|=k+1}\mathrm{vol}(\Delta_S)^p &\le \frac{1}{(k+1)}\sum_{i=1}^{m}\sum_{T,|T|=k} d(a_i, V_{T\cup\{i\}})^p\ \mathrm{vol}(\Delta_T)^p \\
&\le \frac{1}{(k+1)}\sum_{i=1}^{m}\sum_{T,|T|=k} d(a_i, H_k^*)^p\ \mathrm{vol}(\Delta_T)^p \\
&= \frac{1}{(k+1)}\left(\sum_{i=1}^{m} d(a_i, H_k^*)^p\right)\left(\sum_{T,|T|=k}\mathrm{vol}(\Delta_T)^p\right), \qquad (3.2)
\end{aligned}$$

where in the second inequality, the fact that $d(a_i, V_{T\cup\{i\}}) \le d(a_i, H_k^*)$ is because $a_i \in H_{T\cup\{i\}}$ and $V_{T\cup\{i\}}$ contains the projection of H_k^* on to $H_{T\cup\{i\}}$. Finally, combining

equations (3.1) and (3.2) we get

$$\mathbb{E}_S\left[\sum_{i=1}^{m} d(a_i, H_S)^p\right] \leq (k+1)^p \sum_{i=1}^{m} d(a_i, H_k^*)^p.$$

□

Now we will prove Lemma 19, which replaces Lemma 9 in the previous chapter, to complete the proof of our generalized volume sampling theorem.

Lemma 19. *Let $S \subseteq [m]$ be a $(k+1)$-subset and V be any k-dimensional linear subspace of H_S. Then*

$$\mathrm{vol}(\Delta_S) \leq \frac{1}{(k+1)} \sum_{i \in S} d(a_i, V)\ \mathrm{vol}(\Delta_{S\setminus\{i\}}).$$

Proof. W.l.o.g. we identify H_S with $\mathbb{R}^{k+1}$ and the k-dimensional subspace V with $\mathrm{span}(\{e_2, e_3, \ldots, e_{k+1}\})$, where the vectors $\{e_1, e_2, \ldots, e_{k+1}\}$ form an orthonormal basis of $\mathbb{R}^{k+1}$. Let $A_S \in \mathbb{R}^{(k+1)\times(k+1)}$ be a matrix with rows $\{a_i : i \in S\}$ written in the above basis, and let C_{ij} denote its submatrix obtained by removing row i and column j. For any k-subset $T \subseteq S$, let Δ'_T be the projection of Δ_T onto V. Then

$$\begin{aligned}
\mathrm{vol}(\Delta_S) &= \frac{1}{(k+1)!}\ |\det(A_S)| \\
&= \frac{1}{(k+1)!}\ \left|\sum_{i \in S} (-1)^{i+1} (A_S)_{i1} \det(C_{i1})\right| \\
&\leq \frac{1}{(k+1)} \sum_{i \in S} |(A_S)_{i1}| \cdot \frac{1}{k!}\ |\det(C_{i1})| \\
&= \frac{1}{(k+1)} \sum_{i \in S} d(a_i, V)\ \mathrm{vol}(\Delta'_{S\setminus\{i\}}) \\
&\leq \frac{1}{(k+1)} \sum_{i \in S} d(a_i, V)\ \mathrm{vol}(\Delta_{S\setminus\{i\}}),
\end{aligned}$$

since $\mathrm{vol}(\Delta'_{S\setminus\{i\}}) \leq \mathrm{vol}(\Delta_{S\setminus\{i\}})$. □

3.1.2 Approximate volume sampling

Here we describe a simple iterative procedure to do volume sampling approximately.

Approximate Volume Sampling

1. Initialize $S = \emptyset$. While $|S| < k$ do:
   (a) Pick a point from the following distribution:
   $$\Pr(\text{picking } a_i) \propto d(a_i, H_S)^p.$$
   (b) $S = S \cup \{i\}$.
2. Output the k-subset S.

Theorem 20. *Let $\tilde{P}_S$ denote the probability with which the above procedure picks a k-subset S. Then*

$$\tilde{P}_S \leq (k!)^p \cdot P_S,$$

where P_S is the true volume sampling probability of S. Thus,

$$\mathbb{E}_S \left[\sum_{i=1}^{m} d(a_i, H_S)^p \right] \leq (k!)^p \cdot (k+1)^p \sum_{i=1}^{m} d(a_i, H_k^*)^p,$$

where the expectation is over the distribution $\tilde{P}_S$. This implies that

$$\mathbb{E}_S \left[\left(\sum_{i=1}^{m} d(a_i, H_S)^p \right)^{\frac{1}{p}} \right] \leq k! \cdot (k+1) \left(\sum_{i=1}^{m} d(a_i, H_k^*)^p \right)^{\frac{1}{p}}$$

Proof. W.l.o.g., let $S = \{1, 2, \ldots, k\}$, and let Π_k be the set of all permutations of $\{1, 2, \ldots, k\}$. For any $\tau \in \Pi_k$, we also use $H_\tau^{(j)}$ to denote $\text{span}(\{A^{(\tau(1))}, A^{(\tau(2))}, \ldots, A^{(\tau(j))}\})$.

$$\begin{aligned}
\tilde{P}_S &= \sum_{\tau \in \Pi_k} \frac{\|a_{\tau(1)}\|^p}{\sum_{i=1}^m \|a_i\|^p} \frac{d(a_{\tau(2)}, H_\tau^1)^p}{\sum_{i=1}^m d(a_i, H_\tau^1)^p} \cdots \frac{d(a_{\tau(k)}, H_\tau^{k-1})^p}{\sum_{i=1}^m d(a_i, H_\tau^{k-1})^p} \\
&\leq |\Pi_k| \frac{(k!)^p \ \text{vol}(\Delta_S)^p}{\sum_{i=1}^m \|a_i\|^p \cdot \sum_{i=1}^m d(a_i, H_1^*)^p \cdots \sum_{i=1}^m d(a_i, H_{k-1}^*)^p} \\
&= P_S \cdot \frac{(k!)^{p+1} \ \sum_{S, |S|=k} \text{vol}(\Delta_S)^p}{\sum_{i=1}^m \|a_i\|^p \cdot \sum_{i=1}^m d(a_i, H_1^*)^p \cdots \sum_{i=1}^m d(a_i, H_{k-1}^*)^p}.
\end{aligned}$$

Therefore,

$$\frac{\tilde{P}_S}{P_S} \leq \frac{(k!)^{p+1} \ \sum_{S, |S|=k} \text{vol}(\Delta_S)^p}{\sum_{i=1}^m \|a_i\|^p \cdot \sum_{i=1}^m d(a_i, H_1^*)^p \cdots \sum_{i=1}^m d(a_i, H_{k-1}^*)^p}.$$

Now we claim the following, which completes the proof.

Claim:

$$\frac{k! \ \sum_{S, |S|=k} \text{vol}(\Delta_S)^p}{\sum_{i=1}^m \|a_i\|^p \cdot \sum_{i=1}^m d(a_i, H_1^*)^p \cdots \sum_{i=1}^m d(a_i, H_{k-1}^*)^p} \leq 1.$$

Now we will prove the above claim using induction on k. The $k = 1$ case is obvious. For $k > 1$, we can proceed as for equation (3.2) (replacing $k+1$ with k) to get

$$\begin{aligned}
&\frac{k!\ \sum_{S,|S|=k} \mathrm{vol}(\Delta_S)^p}{\sum_{i=1}^m \|a_i\|^p \cdot \sum_{i=1}^m d(a_i, H_1^*)^p \cdots \sum_{i=1}^m d(a_i, H_{k-1}^*)^p} \\
&\leq \frac{(k-1)! \left(\sum_{T,|T|=k-1} \mathrm{vol}(\Delta_T)^p\right) \left(\sum_{i=1}^m d(a_i, H_{k-1}^*)^p\right)}{\sum_{i=1}^m \|a_i\|^p \cdot \sum_{i=1}^m d(a_i, H_1^*)^p \cdots \sum_{i=1}^m d(a_i, H_{k-1}^*)^p} \\
&\leq \frac{(k-1)!\ \sum_{T,|T|=k-1} \mathrm{vol}(\Delta_T)^p}{\sum_{i=1}^m \|a_i\|^p \cdot \sum_{i=1}^m d(a_i, H_1^*)^p \cdots \sum_{i=1}^m d(a_i, H_{k-2}^*)^p} \\
&\leq 1,
\end{aligned}$$

by induction hypothesis for the $(k-1)$ case. □

3.2 Additive approximation

We prove bounds on the subspaces that we find in terms of any k-subspace H of $\mathbb{R}^n$, which therefore, also hold for the optimal subspace H_k^*.

3.2.1 Finding a close line

Given any k-dimensional subspace H and a line l, we define H_l as follows. If l is not orthogonal to H, then its projection onto H is a line, say l'. Let H' be the $(k-1)$-dimensional subspace of H that is orthogonal to l'. Then we define $H_l = \mathrm{span}(H' \cup l)$. In short, H_l is a rotation of H so as to contain line l. In case when l is orthogonal to H, we define $H_l = \mathrm{span}(H' \cup l)$, where H' is any $(k-1)$-dimensional subspace of H.

Lemma 21. *Let S be a sample of $O\left((2k/\epsilon)^p (k/\epsilon) \log(k/\epsilon)\right)$ i.i.d. points from the set of given points $a_1, a_2, \ldots, a_m$, using the following distribution:*

$$\Pr(\textit{picking } a_i) \propto \|a_i\|^p$$

then, with probability at least $1 - (\epsilon/k)^{k/\epsilon}$, H_S contains a line l such that

$$\left(\sum_{i=1}^m d(a_i, H_l)^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^m d(a_i, H)^p\right)^{\frac{1}{p}} + \frac{\epsilon}{k}\left(\sum_{i=1}^m \|a_i\|^p\right)^{\frac{1}{p}},$$

where H_l is defined as above.
Remark: *It means that there exists a k-dimensional subspace H_l, within an additive error of the optimal, that intersects H_S in at least one dimension.*

Proof. Let l_1 be the line spanned by the first point in our sample, and let θ_1 be its angle with H. In general, let l_j be the line in the span of the first j sample points that makes the smallest angle with H, and let θ_j denote this smallest angle.

Consider the $(j+1)$-th sample point for some $j \geq 1$, and assume that

$$\left(\sum_{i=1}^{m} d(a_i, H_{l_j})^p\right)^{\frac{1}{p}} > \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \frac{\epsilon}{k}\left(\sum_{i=1}^{m} \|a_i\|^p\right)^{\frac{1}{p}}. \tag{3.3}$$

Define BAD $= \{i \ : \ d(a_i, H_{l_j}) > \left(1+\frac{\epsilon}{2k}\right) d(a_i, H)$ and GOOD $= [m] \setminus$ BAD. We claim that

$$\sum_{i \in BAD} \|a_i\|^p > \left(\frac{\epsilon}{2k}\right)^p \sum_{i=1}^{m} \|a_i\|^p. \tag{3.4}$$

Because, otherwise, using Minkowski's inequality, the triangle inequality for the L_p norm,

$$\begin{aligned}
\left(\sum_{i=1}^{m} d(a_i, H_{l_j})^p\right)^{1/p} &\leq \left(\sum_{i \in \text{GOOD}} d(a_i, H_{l_j})^p\right)^{1/p} + \left(\sum_{i \in \text{BAD}} d(a_i, H_{l_j})^p\right)^{1/p} \\
&\leq \left(1+\frac{\epsilon}{2k}\right)\left(\sum_{i \in \text{GOOD}} d(a_i, H)^p\right)^{1/p} + \left(\sum_{i \in \text{BAD}} \|a_i\|^p\right)^{1/p} \\
&\leq \left(1+\frac{\epsilon}{2k}\right)\left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{1/p} + \frac{\epsilon}{2k}\left(\sum_{i=1}^{m} \|a_i\|^p\right)^{1/p} \\
&\leq \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{1/p} + \frac{\epsilon}{k}\left(\sum_{i=1}^{m} \|a_i\|^p\right)^{1/p},
\end{aligned}$$

contradicting our assumption about H_{l_j} as in equation (3.3).

Inequality (3.4) implies that with probability at least $(\epsilon/2k)^p$ we pick as our $(j+1)$-th point a_i with $i \in$ BAD and by definition

$$d(a_i, H_{l_j}) \geq \left(1+\frac{\epsilon}{2k}\right) d(a_i, H).$$

Now, by Lemma 26, there exists a line l' in span$(\{a_i\} \cup l_j)$ such that the sine of the angle that l' makes with H is at most $(1-\epsilon/4k)\sin\theta_j$. This implies that

$$\sin\theta_{j+1} \leq \left(1-\frac{\epsilon}{4k}\right)\sin\theta_j.$$

Let us call the $(j+1)$-th sample a success if either (a) the inequality (3.3) fails to hold, or (b) the inequality (3.3) holds but $\sin\theta_{j+1} \leq (1-\epsilon/4k)\sin\theta_j$. We conclude that the probability that the $(j+1)$-th sample is a success is at least $(\epsilon/2k)^p$.

Let N denote the number of times our algorithm samples, and suppose that there are $\Omega\left((k/\epsilon)\log(k/\epsilon)\right)$ successes among the samples $2, \ldots, N$. If inequality (3.3) fails

to hold for some $1 \leq j \leq N-1$, then H_S contains a line, namely l_j, that satisfies the inequality claimed in the Lemma. Let us assume that the inequality (3.3) holds for every $1 \leq j \leq N-1$. Clearly, we have $\sin\theta_{j+1} \leq \sin\theta_j$ for each $1 \leq j \leq N-1$ and furthermore we have $\sin\theta_{j+1} \leq (1-\epsilon/4k)\sin\theta_j$ if the $(j+1)$-th sample is a success. Therefore

$$\sin\theta_N \leq \left(1-\frac{\epsilon}{4k}\right)^{\Omega((k/\epsilon)\log(k/\epsilon))} \sin\theta_0 \leq \frac{\epsilon}{k}.$$

Now using Minkowski's inequality we have

$$\left(\sum_{i=1}^{m} d(a_i, H_{l_N})^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{m} d(\bar{a}_i, a_i')^p\right)^{\frac{1}{p}},$$

where $\bar{a}_i$ is the projection of a_i onto H, and a_i' is the projection of $\bar{a}_i$ onto H_{l_N}. But $d(\bar{a}_i, a_i') \leq \sin\theta_N \|a_i\|$, which implies

$$\left(\sum_{i=1}^{m} d(a_i, H_{l_N})^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \frac{\epsilon}{k}\left(\sum_{i=1}^{m} \|a_i\|^p\right)^{\frac{1}{p}}.$$

Thus H_S contains the line l_N that satisfies the inequality claimed in the Lemma.

Our algorithm samples $O\left((2k/\epsilon)^p(k/\epsilon)\log(k/\epsilon)\right)$ times, and the probability that a sample is a success is at least $(\epsilon/2k)^p$. Using the Chernoff inequality with some care, we conclude that with a probability of at least $1-(\epsilon/k)^{k/\epsilon}$, there are at least $\Omega\left((k/\epsilon)\log(k/\epsilon)\right)$ successes among the samples $2,\ldots,N$. This completes the proof. □

3.2.2 From line to subspace

Additive Approximation
Input: $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, $k > 0$.
Output: a subset $S \subseteq [m]$ of $\tilde{O}\left(k \cdot (k/\epsilon)^{p+1}\right)$ points.

1. Repeat the following $O(k\log k)$ times and pick the best sample S amongst all that minimizes $\sum_{i=1}^{m} d(a_i, H_S)^p$.
2. Initialize $S = S_0 = \emptyset$, $\delta = \epsilon/\log k$. For $t = 1$ to k do:
   (a) Pick a sample S_t of $O\left((2k/\delta)^p(k/\delta)\log(k/\delta)\right)$ points from the following distribution:
   $$\Pr(\text{picking } a_i) \propto d(a_i, H_S)^p.$$
   (b) $S \leftarrow S \cup S_t$.

Theorem 22. *The above algorithm returns a subset $S \subseteq [m]$ of the given points such that*

$$|S| = O\left(k \cdot (2k/\delta)^p(k/\delta)\log(k/\delta)\right)$$

and

$$\left(\sum_{i=1}^{m} d(a_i, H_S)^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \epsilon \left(\sum_{i=1}^{m} \|a_i\|^p\right)^{\frac{1}{p}}.$$

with probability at least $1 - 1/k$.

Proof. For a start, let us only look at step 2. From Lemma 21, we know that there exists a k-dimensional subspace F_1 such that $\dim(F_1 \cap H_{S_1}) \geq 1$ and

$$\left(\sum_{i=1}^{m} d(a_i, F_1)^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \frac{\delta}{k} \left(\sum_{i=1}^{m} \|a_i\|^p\right)^{\frac{1}{p}},$$

with probability at least

$$1 - \left(\frac{\delta}{k}\right)^{k/\delta}.$$

Let π_1 be the orthogonal projection onto $(H_{S_1})^{\perp}$. Consider a new set of points $\pi_1(a_i)$ and a new subspace $\pi_1(F_1)$ of dimension $j \leq k-1$. Using Lemma 21 for the new points and subspace, we get that there exists a j-dimensional subspace F_2 in $(H_{S_1})^{\perp}$ such that $\dim(F_2 \cap \pi_1(H_{S_2})) \geq \min\{j, 1\}$ and

$$\begin{aligned}
\left(\sum_{i=1}^{m} d(\pi_1(a_i), F_2))^p\right)^{\frac{1}{p}} &\leq \left(\sum_{i=1}^{m} d(\pi_1(a_i), \pi_1(F_1))^p\right)^{\frac{1}{p}} + \frac{\delta}{k-1} \left(\sum_{i=1}^{m} \|\pi_1(a_i)\|^p\right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i=1}^{m} d(a_i, F_1)^p\right)^{\frac{1}{p}} + \frac{\delta}{k-1} \left(\sum_{i=1}^{m} \|a_i\|^p\right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \delta \left(\frac{1}{k} + \frac{1}{k-1}\right) \left(\sum_{i=1}^{m} \|a_i\|^p\right)^{\frac{1}{p}},
\end{aligned}$$

with probability at least

$$\left(1 - \left(\frac{\delta}{k}\right)^{\frac{k}{\delta}}\right) \left(1 - \left(\frac{\delta}{k-1}\right)^{\frac{k-1}{\delta}}\right).$$

Proceeding similarly for k steps, we have a subspace F_k in the orthogonal complement of $H_{S_1 \cup \cdots \cup S_{k-1}}$ such that $\dim(F_k) \leq 1$, $\dim(F_k \cap \pi_{k-1}(H_{S_k})) \geq \min\{\dim(F_k), 1\}$, where π_t denotes projection to the orthogonal complement of $H_{S_1 \cup \cdots \cup S_t}$, and

$$\left(\sum_{i=1}^{m} d(\pi_{k-1}(a_i), F_k)^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \delta \left(\frac{1}{k} + \frac{1}{k-1} + \cdots + 1\right) \left(\sum_{i=1}^{m} \|a_i\|^p\right)^{\frac{1}{p}},$$

with probability at least

$$\left(1-\frac{\delta}{k}\right)\left(1-\frac{\delta}{k-1}\right)\cdots \geq \frac{1-\delta}{k} \geq \frac{1}{2k}.$$

The conditions (1) and (2) imply that $F_k \subseteq \pi_{k-1}(H_{S_k})$. Therefore with $S = S_1 \cup \cdots \cup S_k$, we have $d(a_i, H_S) = \|\pi_k(a_i)\| \leq d(\pi_{k-1}(a_i), \pi_{k-1}(H_{S_k})) \leq d(\pi_{k-1}(a_i), F_k)$, for all i. Hence,

$$\begin{aligned}\left(\sum_{i=1}^{m} d(a_i, H_S)^p\right)^{\frac{1}{p}} &\leq \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \delta\ O(\log k)\ \left(\sum_{i=1}^{m} \|a_i\|^p\right)^{\frac{1}{p}} \\ &= \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \epsilon\ \left(\sum_{i=1}^{m} \|a_i\|^p\right)^{\frac{1}{p}},\end{aligned}$$

with probability at least $1/2k$. Repeating this $O(k \log k)$ times boosts the success probability to $1 - 1/k$. □

3.3 Adaptive sampling

By *adaptive sampling* we mean picking a subset S of points and then sampling new points with probabilities proportional to their distances from H_S. The benefits of doing this were implicit in the previous sections, but here we introduce the most important one: additive error drops exponentially with the number of rounds of adaptive sampling.

3.3.1 Exponential drop in additive error

Proposition 23. *Suppose we have an initial subspace V of $\mathbb{R}^n$. Then we can find a sample S of $\tilde{O}\left(k \cdot (k/\epsilon)^{p+1}\right)$ rows such that*

$$\left(\sum_{i=1}^{m} d(a_i, \text{span}(V \cup H_S))^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}} + \epsilon \left(\sum_{i=1}^{m} d(a_i, V)^p\right)^{\frac{1}{p}},$$

with probability at least $1 - 1/k$.

Proof. Use a new points set $\pi(a_i)$ and a new subspace $\pi(H)$, where $\pi(\cdot)$ is orthogonal projection onto $V^{\perp}$. Now using Theorem 22 we get

$$\left(\sum_{i=1}^{m} d(\pi(a_i), \pi(H_S))^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{m} d(\pi(a_i), \pi(H))^p\right)^{\frac{1}{p}} + \epsilon \left(\sum_{i=1}^{m} \|\pi(a_i)\|^p\right)^{\frac{1}{p}}.$$

And the proof follows by using

$$d(a_i, \text{span}(V \cup H_S)) \leq d(\pi(a_i), \pi(H_S)), \text{ for all } i.$$

□

Theorem 24. *Suppose we have an initial subspace V of $\mathbb{R}^n$. Then using t rounds of adaptive sampling we can find subsets $S_1, S_2, \ldots, S_t \subseteq [m]$ with*

$$|S_1 \cup S_2 \cup \cdots \cup S_t| = \tilde{O}\left(tk \cdot (k/\epsilon)^{p+1}\right),$$

such that

$$\left(\sum_{i=1}^{m} d(a_i, \text{span}(V \cup H_{S_1 \cup \cdots \cup S_t}))^p\right)^{\frac{1}{p}} \leq \frac{1}{1-\epsilon}\left(\sum_{i=1}^{m}(a_i, H)^p\right)^{\frac{1}{p}} + \epsilon^t\left(\sum_{i=1}^{m} d(a_i, V)^p\right)^{\frac{1}{p}},$$

with probability at least $(1 - 1/k)^t$.

Proof. using Proposition 23 in t rounds by induction. □

3.3.2 Combining volume and adaptive sampling

We can combine volume sampling and adaptive sampling to give a bi-criteria algorithm for subspace approximation. The algorithm (implicit in Theorem 25 below) finds a
$\tilde{O}\left(k^2(k/\epsilon)^{p+1}\right)$-dimensional subspace whose error is at most $(1+\epsilon)$ times the error of the best k-dimensional subspace.

Theorem 25. *Let $V = \text{span}(S_0)$, where S_0 is a k-subset of rows picked by Approximate Volume Sampling procedure (see Subsection 3.1.2), $t = O(k \log k)$, and $S_1, S_2, \ldots, S_t$ as in Theorem 24. Then*

$$\left(\sum_{i=1}^{m} d(a_i, H_{S_0 \cup \cdots \cup S_t})^p\right)^{\frac{1}{p}} \leq (1+\epsilon)\left(\sum_{i=1}^{m} d(a_i, H)^p\right)^{\frac{1}{p}},$$

with probability $1/k$. Repeating $O(k)$ times we can boost this success probability to $3/4$, and the subset we find is of size

$$|S_0 \cup S_1 \cup \ldots \cup S_t| = \tilde{O}\left(k^2(k/\epsilon)^{p+1}\right).$$

Computation of these subsets takes time effectively $\tilde{O}\left(mn \cdot k^3(k/\epsilon)^{p+1}\right)$.

Proof. Immediate from Theorem 24. □

3.4 Angle-drop lemma

Lemma 26. *Let F be a k-subspace in $\mathbb{R}^n$ for some $k > 0$, l' be any line, $\alpha(l')$ the sine of the angle that l' makes with F, l the projection of l' onto F (if $\alpha(l') = 1$ then take l to be any line in F), E the orthogonal complement of l in F, and $\hat{F}$ the subspace spanned by E and l'. That is, $\hat{F}$ is the rotation of F so as to contain l'. Suppose that $a \in \mathbb{R}^n$ is such that $d(a, \hat{F}) > (1+\delta/2)d(a, F)$. Then there is a line l'' in the subspace spanned by l' and a such that $\alpha(l'')$, the sine of the angle made by l'' with F, is at most $(1 - \frac{\delta}{4})\alpha(l')$.*

Proof. The proof is from [22], and is presented here for completeness. Let $\pi_E(\cdot)$ denote the projection onto E. Note that $\pi_E(l')$ is just the origin o. Let $\bar{a}$ denote the projection of a onto F, and a' the projection of $\bar{a}$ onto $\hat{F}$. Since $d(a, \hat{F}) > (1 + \delta/2)d(a, F)$, we have $|aa'| > (1 + \delta/2)|a\bar{a}|$. Elementary geometric reasoning about the triangle $\triangle aa'\bar{a}$ (see for example Lemma 2.1 of [22]) tells us that there is a point s on the segment $\overline{a'a}$ such that $|\bar{a}s| \leq (1 - \delta/4)|\bar{a}a'|$.

Let $\hat{a} = \pi_E(a) = \pi_E(\bar{a}) = \pi_E(a')$. We verify that the point $q' = a' - \hat{a}$ lies on the line l'. Considering $\triangle aa'q'$, and recalling that s lies on $\overline{a'a}$, we see that there is a point q on the segment $\overline{q'a}$ such that $q - s$ is a scaling of $-\hat{a}$. (If $\hat{a} = o$, q' and q degenerate to a' and s respectively.) Let e be the point on the line $\{\bar{a} - t\hat{a} | t \in \mathbb{R}\}$ closest to q. (If $\hat{a} = o$, then $e = \bar{a}$.) It is easy to verify that $|eq| \leq |\bar{a}s|$ since $\bar{a}$ and s are on lines parallel to $-\hat{a}$ and $|eq|$ is the distance between these lines. Finally, let e' be the projection of e onto $\hat{F}$. Since e is a translation of $\bar{a}$ by a vector that is scale of $-\hat{a}$ and which therefore lies in $\hat{F}$, we have $|\bar{a}a'| = |ee'|$. So we have

$$|eq| \leq |\bar{a}s| \leq \left(1 - \frac{\delta}{4}\right)|\bar{a}a'| = \left(1 - \frac{\delta}{4}\right)|ee'|.$$

We take l'' to be the line through q. Note that l'' indeed lies in the span of l' and a. To bound $\alpha(l'')$, it is enough to bound the sine of the angle between l'' and $l(e)$, the line through e, since e lies on F.

$$\alpha(l'') \leq \frac{|eq|}{|oe|} \leq \left(1 - \frac{\delta}{4}\right)\frac{|ee'|}{|oe|} \leq \left(1 - \frac{\delta}{4}\right)\alpha(l'), \tag{3.5}$$

where the last inequality can be seen from the facts that e lies on F, e' is the projection of e onto $\hat{F}$, and $\hat{F}$ is the rotation of F through ℓ'. □

Chapter 4

Applications and Future Directions

The existence results and algorithms for sampling-based dimension reduction from the previous chapters have several interesting applications, some of which will be discussed here. We will also see a few related problems.

4.1 Projective clustering

Projective clustering is a generalization of subspace approximation where we want to find multiple low-dimensional subspaces that fit our data, instead of a single subspace. Here is a statement of the problem in a very general form; we call it as *subspace projective clustering.*

Given points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, $k, s > 0$, and $p \geq 1$, we want to find k-dimensional linear subspaces $H[1], H[2], \ldots, H[s]$ that minimize the error $(\sum_{i=1}^{m} d(a_i, H)^p)^{\frac{1}{p}}$, where H denotes $H[1] \cup H[2] \cup \cdots \cup H[s]$. Let $H^*[1], \ldots, H^*[s]$ denote the optimal subspaces and let H^* denote their union $H^*[1] \cup \cdots \cup H^*[s]$.

There has been a lot of work on various special cases of the projective clustering [2, 5, 4]. We state here a result from [11] which gives a PTAS for the special case $p = 2$, when k, s are taken to be fixed constants.

Theorem 27. *Given m points in $\mathbb{R}^n$ and parameters B and ϵ, in time*

$$n\left(\frac{m}{\epsilon}\right)^{O(sk^3/\epsilon)}$$

we can find a solution to the projective clustering problem which is of cost at most $(1+\epsilon)B$ provided there is a solution of cost B.

For $p \geq 1$, Kasturi Varadarajan [12] proved a sampling-based dimension reduction for subspace approximation can be extended to get a sampling-based dimension reduction result for projective clustering as follows.

Theorem 28. *Given points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$ and a subspace V of dimension at least k with the guarantee*

$$\left(\sum_{i=1}^{m} d(a_i, V)^p\right)^{1/p} \leq 2\left(\sum_{i=1}^{m} d(a_i, H^*)^p\right)^{1/p},$$

one can find a sample S of

$$\tilde{O}\left(\frac{sk^{2(p+2)}}{\epsilon^{p+1}}\right)$$

additional points from $a_1, a_2, \ldots, a_m$ such that, with probability at least $1 - 1/4ks$, $\mathrm{span}(V \cup H_S)$ contains k-dimensional subspaces $H'[1], \ldots, H'[s]$ satisfying

$$\left(\sum_{i=1}^{m} d(a_i, H')^p\right)^{1/p} \leq (1+\epsilon)\left(\sum_{i=1}^{m} d(a_i, H^*)^p\right)^{1/p},$$

where H' denotes $H'[1] \cup \ldots \cup H'[s]$.

4.2 Over finite fields

Low-rank matrix approximation of matrices over finite fields is quite different from low-rank matrix approximation of real matrices. We consider the following version of the problem. Given a matrix $A \in \mathbb{F}_2^{m \times n}$, find another matrix $B \in \mathbb{F}_2^{m \times n}$ such that $d_H(A, B)$ is minimized, where $d_H(\cdot, \cdot)$ is the entry-wise Hamming distance between A and B.

Unlike the low-rank approximation of real matrices, this problem turns out to be NP-hard as shown in Proposition 29 below.

Proposition 29. *The problem of finding the best rank-k approximation to $A \in \mathbb{F}_2^{m \times n}$ under the Hamming distance is as hard as the Nearest Codeword Problem. Moreover, the reduction preserves the approximation factor.*

Proof. Given an instance of the Nearest Codeword Problem: a generator matrix $G \in \mathbb{F}_2^{l \times k}$ and a codeword $y \in \mathbb{F}_2^l$, we reduce it to a low-rank matrix approximation problem as follows. We construct an m by n matrix A, where $m = k(l+1)+1$ and $n = l$, whose first row is y and the others are $(l+1)$ copies of $G_1^T, G_2^T, \ldots, G_k^T$, where $G_1, G_2, \ldots, G_k$ are the columns of G.

Let the nearest codeword to y be $x \in \mathrm{span}(G_1^T, G_2^T, \ldots, G_k^T)$. Also let A_k be the best rank-k approximation to A, and let S be the span of the rows of A_k. We claim that $S = \mathrm{span}(G_1, G_2, \ldots, G_k)$ and therefore, $d_H(A, A_k) = d_H(x, y)$. Otherwise, if $S \neq \mathrm{span}(G_1^T, G_2^T, \ldots, G_k^T)$, then A_k makes at least one error for each block of $G_1^T, G_2^T, \ldots, G_k^T$, which implies that $d_H(A, A_k) \geq l+1$. But using $\mathrm{span}(G_1^T, G_2^T, \ldots, G_k^T)$ we can get a rank-k approximation $\tilde{A}$ with error $d_H(A, \tilde{A}) = d_H(x, y) \leq l$, which is a contradiction to A_k being the best rank-k approximation. □

We know that the Nearest Codeword Problem is hard to approximate within any constant factor, unless P=NP [10]. So the above reduction implies that there is no constant factor approximation algorithm for low-rank matrix approximation over finite fields, unless P=NP. Although some sampling-based dimension reduction results can be shown for this problem, they do not lead to efficient algorithms. It would be interesting to show hardness of some special cases (even rank-1 matrix approximation of matrices over $\mathbb{F}_2$) and find efficient algorithms for low-rank matrix approximation over finite fields. This problem is similar in nature to the matrix rigidity problem [6] and may have other important consequences.

4.3 Missing entries

Another interesting problem is to consider low-rank matrix approximation of matrices with missing entries, where the error of approximation is measures only over the entries that are present. For example, given $A \in \mathbb{R}^{m \times n}$, let $M \subseteq [m] \times [n]$ be the set of its missing entries. We want to find another matrix $B \in \mathbb{R}^{m \times n}$ of rank at most k such that

$$\|A - B\|_F^2 = \sum_{(i,j) \notin M} (A_{ij} - B_{ij})^2$$

is minimized.

Another way of looking at this problem is as follows. Given affine subspaces $A_1, A_2, \ldots, A_m \subseteq \mathbb{R}^n$, we want to find a k-dimensional subspace V that minimizes

$$\sum_{i=1}^{m} d(A_i, V)^2.$$

This is more general that the missing entries problem as the subspaces may not be axis-aligned. Luis Rademacher pointed out to me that the optimal solution for these problems, as stated above, may not always exist, i.e., there may exist a sequence of k-dimensional subspaces with arbitrarily small errors but no single k-dimensional subspace with zero error. So sampling-based multiplicative approximation results do not make sense. However, many of the additive approximation results (e.g., Theorem 22, Proposition 23, and Theorem 24) still hold with little modifications in their proofs. Proving hardness results and finding efficient algorithms for low-rank matrix approximation with missing entries is a promising direction for future work.

Bibliography

[1] A. C. Aitken, Determinants and Matrices, University Mathematical Texts, Oliver and Boyd, 1939.

[2] R. Agarwal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. Data Mining and Knowledge Discovery, 11, 2005, pp. 5-33.

[3] D. Achlioptas, F. McSherry. Fast Computation of Low-Rank Matrix Approximations. Journal of the ACM (JACM), 54 , Vol. 2, Article 9, 2007.

[4] P. Agarwal and N. Mustafa: k-means projective clustering. In Proc. of Principles of Database Systems (PODS'04), ACM Press, 2004, pp. 155-165.

[5] P. Agarwal, C. Procopiuc, and K. Varadarajan. Approximation algorithms for a k-line center. Algorithmica, 42, 2005, pp. 221-230.

[6] B. Codenotti. Matrix Rigidity, Linear Algebra and its Applications, 304 (1-3), 2000, pp. 181-192.

[7] P. Drineas, A. Frieze, R. Kannan, S. Vempala, V. Vinay. Clustering Large Graphs via the Singular Value Decomposition. Machine Learning, 56, 2004, pp. 9-33.

[8] P. Drineas, R. Kannan, and M. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. SIAM J. on Computing, 36, 2006, pp. 158-183.

[9] P. Drineas, M. Mahoney, S. Muthukrishnan. Polynomial time algorithm for column-row based relative error low-rank matrix approximation. Proc. of the 10th International Workshop on Randomization and Computation (RANDOM), 2006.

[10] I. Dumer, D. Micciancio, M. Sudan. Hardness of approximation the minimum distance of a linear code, IEEE Transactions on Information Theory, 49 (1), 2003, pp. 22-37.

[11] A. Deshpande, L. Rademacher, S. Vempala, G. Wang. Matrix Approximation and Projective Clustering via Volume Sampling. Theory of Computing, Vol.2, Article 12, 2006, pp. 225-247.

[12] A. Deshpande, K. Varadarajan. Sampling-Based Dimension Reduction for Subspace Approximation. To appear in the Proc. of ACM Symposium on Theory of Computing (STOC), 2007.

[13] A. Deshpande, S. Vempala. Adaptive Sampling and Fast Low-Rank Matrix Approximation. Proc. of the 10th International Workshop on Randomization and Computation (RANDOM), 2006.

[14] D. Feldman, A. Fiat, and M. Sharir. Coresets for weighted facilities and their applications. Proc. of IEEE Symposium on Foundations of Computer Science (FOCS), 2006.

[15] A. Frieze, R. Kannan, S. Vempala. Fast Monte Carlo algorithms for finding low-rank approximations. Journal of the ACM, 51, 2004, pp. 1025-1041.

[16] S. Har-Peled. How to get close to the median shape. Proc. of ACM Symposium on Computational Geometry (SOCG), 2006.

[17] S. Har-Peled. Low-Rank Matrix Approximation in Linear Time. manuscript.

[18] S. Har-Peled and K. Varadarajan. Projective clustering in high dimensions using core-sets. Proc. of ACM Symposium on Computational Geometry (SOCG), 2002, pp. 312-318.

[19] S. Har-Peled and K. Varadarajan. High-Dimensional Shape Fitting in Linear Time. Discrete & Computational Geometry, 32(2), 2004, pp. 269-288.

[20] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. Contemp. Math. 26, 1984, pp. 189-206.

[21] T. Sarlos. Improved Approximation Algorithms for Large Matrices via Random Projections. Proc. of IEEE Symposium on Foundations of Computer Science (FOCS), 2006.

[22] N. D. Shyamalkumar and K. Varadarajan. Efficient Subspace Approximation Algorithms. Proc. of ACM-SIAM Symposium on Discrete Algorithms (SODA), 2007.

[23] S. Vempala. The Random Projection Method. DIMACS Series in Discrete Math and Theoretical Computer Science, Vol. 65, 2004.