Sampling-Based Dimension Reduction for Subspace Approximation

Amit Deshpande*
MIT
amitd@mit.edu

Kasturi Varadarajan†
Univ. of Iowa
kvaradar@cs.uiowa.edu

ABSTRACT

We give a randomized bi-criteria algorithm for the problem of finding a k-dimensional subspace that minimizes the L_p-error for given points, i.e., p-th root of the sum of p-th powers of distances to given points, for any $p \geq 1$. Our algorithm runs in time $\tilde{O}\left(mn \cdot k^3(k/\epsilon)^{p+1}\right)$ and produces a subset of size $\tilde{O}\left(k^2(k/\epsilon)^{p+1}\right)$ from the given points such that, with high probability, the span of these points gives a $(1+\epsilon)$-approximation to the optimal k-dimensional subspace. We also show a dimension reduction type of result for this problem where we can efficiently find a subset of size $\tilde{O}\left(k^{p+3} + (k/\epsilon)^{p+2}\right)$ such that, with high probability, their span contains a k-dimensional subspace that gives $(1+\epsilon)$-approximation to the optimum. We prove similar results for the corresponding projective clustering problem where we need to find multiple k-dimensional subspaces.

Categories and Subject Descriptors

F.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

General Terms

Algorithms,Theory

Keywords

subspace approximation, projective clustering

1. INTRODUCTION

Low-dimensional representations of massive data sets are often important in data mining, statistics, and clustering. We consider the problem of *subspace approximation*, i.e., 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 $p \geq 1$. We also consider the corresponding *projective clustering* problem where instead of one k-dimensional subspace we want to find s of them such that the p-th root of the sum of of the p-th powers of distances from each a_i to its nearest subspace is minimized.

The $p = 2$ case for subspace approximation (also known as *low-rank matrix approximation*) is well studied because a k-dimensional subspace that minimizes the sum of squared distances is spanned by the top k right singular vectors of a matrix $A \in \mathbb{R}^{m \times n}$ (with rows $a_1, a_2, \ldots, a_m$), and can be computed in time $O(\min\{mn^2, m^2n\})$ using Singular Value Decomposition (SVD). Some recent work on $p = 2$ case [1, 2, 3, 4, 5, 9, 12], initiated by a result due to Frieze, Kannan, and Vempala [7], has focused on algorithms for computing a k-dimensional subspace that gives $(1+\epsilon)$-approximation to the optimum in time $O(mn \cdot \text{poly}(k, 1/\epsilon))$, i.e., linear in the number of co-ordinates we store. Most of these algorithms, with the exception of [1, 12], depend on subroutines that sample $\text{poly}(k, 1/\epsilon)$ points from given $a_1, a_2, \ldots, a_m$ with the guarantee that, with high probability, their span contains a k-dimensional subspace that gives $(1+\epsilon)$-approximation to the optimum.

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 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 [13].

2. OUR RESULTS

We state our problems once again.

Subspace Approximation Problem:

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^*.

*Most of this work was done while the author was visiting College of Computing, Georgia Institute of Technology.

†The author was partially supported by NSF CAREER award CCR 0237431

STOC'07, June 11–13, 2007, San Diego, California, USA.

Subspace Projective Clustering:

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

We now state our results and relate them to other relevant results:

1. We first obtain a bi-criteria result: 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:

   (a) In Section 3, we prove that the span of k points picked using *volume sampling* has expected error $(k+1)$ times the optimum. Since we do not know how to do volume sampling exactly in an efficient manner, Section 3.2 describes an efficient procedure to implement volume sampling approximately with a weaker multiplicative guarantee of $k! \cdot (k+1)$.

   (b) In Section 4, 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.

   (c) 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 5, 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 this 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.

   (d) 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.

2. Our next result is a dimension reduction for the subspace approximation problem: We describe an algorithm that runs in $mn \cdot \text{poly}(\frac{k}{\epsilon})$ time and returns a subspace C of dimension $\tilde{O}\left(k^{p+3} + (k/\epsilon)^{p+2}\right)$ that, with probability at least 1/2, is guaranteed to contain a k-subspace whose error is at most $(1+\epsilon)H_k^*$, for any $\epsilon > 0$. This kind of result was known for the case $p = 2$, but not for the case $p = 1$. (For the special case $k = 1$, it was implicit in [13]; however, that approach does not generalize to larger k.) Its importance is precisely in its being a dimension reduction result – algorithms developed for the subspace approximation problem in low or 'fixed' dimension, which were designed to optimize the dependence on the number of points but not the dimension, can be plugged in to obtain algorithms with very good dependence on the dimension. Approximation algorithms for the k-subspace approximation problem in fixed dimension are near linear in the number of points but exponential in the dimension [8] – plugging these in yields algorithms whose running time is comparable to but not significantly better than the $O(mn2^{\text{poly}(k/\epsilon)})$ algorithm of [13] for $p = 1$. Note that the dimension reduction can be seen as reducing to a constrained instance of the problem in $\dim(C)+1$ dimensions.

   The result is obtained by first using the previous bi-criteria result to obtain a subspace V of dimension $\tilde{O}(k^{p+3})$ that gives a 2-approximation to the optimal k-dimensional subspace. Assuming without loss of generality that V has dimension at least k, the algorithm of Section 6 uses adaptive sampling to pick $\tilde{O}\left((k/\epsilon)^{p+2}\right)$ points so that the span of V with these new points contains a k-dimensional subspace that gives a $(1+\epsilon)$ approximation to the optimum.

   The algorithm of [13], that runs in $O(mn2^{\text{poly}(k/\epsilon)})$ time for $p = 1$ and returns with probability at least 1/2 a nearly optimal k-subspace, works by first finding a line ℓ that lies in a nearly optimal k-subspace, then a 2-subspace B that lies in a nearly optimal k-subspace, and so on till it finds a nearly optimal k-subspace. The authors of [13] show that the span of a sample A_1 of $O(\text{poly}(k/\epsilon))$ points contains with high probability such a line ℓ, provided the input points are sampled in proportion to the norms. However, the algorithm needs ℓ and not just A_1 – this is because the next stage for finding B needs to sample based on distances from ℓ. So they guess ℓ, but the guess works with a probability that is only inversely proportional to $2^{\text{poly}(k/\epsilon)}$. This is why their sampling technique is inadequate for obtaining our dimension reduction result.

   We now illustrate how we harness the power of adaptive sampling. Consider the case when $k = 1$, and let ℓ denote the optimal solution, and V a subspace of small dimension whose error is within a constant factor of that of ℓ. Let $\hat{\ell}$ denote the projection of ℓ onto V – this can be viewed as V's proxy for ℓ. It can be seen that the error of $\hat{\ell}$ is within a constant factor of that of ℓ. But suppose that $\hat{\ell}$ is not quite good enough, that is, the error of $\hat{\ell}$ is at least $(1+\epsilon)$ times that of ℓ. We would like an input point a that is a witness to this – it must satisfy $d(a, \hat{\ell}) > (1 + \epsilon/2) d(a, \hat{\ell})$. Such a point would enlarge V so that the resulting subspace is closer to ℓ than V. How can we find a witness given that we know only V and not ℓ or $\hat{\ell}$? The observation is that adaptive sampling, that is, sampling according to distances from V, yields a witness with probability $\Omega(\epsilon)$. It is via this observation that we combine adaptive sampling with the analysis techniques in [13] to get our dimension reduction result.

3. The usefulness of the adaptive sampling approach and the flexibility of our analysis are perhaps best demonstrated by our result for dimension reduction for projective clustering. In Section 7, we describe a randomized algorithm that runs in $O(mn \cdot \text{poly}(\frac{ks}{\epsilon}))$ time and returns a subspace spanned by $\text{poly}(\frac{ks}{\epsilon})$ points that is guaranteed, with probability at least $1/2$, to contain s k-subspaces whose union is a $(1+\epsilon)$-approximation to the optimum H^*. To our knowledge, such a dimension reduction result is not known for the projective clustering problem for any p, including the cases $p = 2$ and $p = \infty$. Previous results for the cases $p = 1, 2, \infty$ [10, 4, 13] only showed the existence of such a subspace spanned by $\text{poly}(\frac{ks}{\epsilon})$ points – the algorithm for finding the subspace enumerated all subsets of $\text{poly}(\frac{ks}{\epsilon})$ points. Our dimension reduction result, combined with the recent fixed-dimensional result of [6], yields an $O(mn \cdot \text{poly}(\frac{s}{\epsilon}) + m(\log m)^{f(s/\epsilon)})$ time algorithm for the projective clustering problem with $k = 1$. For lack of space, we do not elaborate on this application to the $k = 1$ case here.

3. 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 $(k+1)$-approximation using k points

THEOREM 1. *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.

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

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 2 gives

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

Hence, taking p-th power we have

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

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

$$\begin{aligned}
&\sum_{S,|S|=k+1} \text{vol}(\Delta_S)^p \\
&\leq \frac{1}{(k+1)} \sum_{i=1}^m \sum_{T,|T|=k} d(a_i, V_{T\cup\{i\}})^p\ \text{vol}(\Delta_T)^p \\
&\leq \frac{1}{(k+1)} \sum_{i=1}^m \sum_{T,|T|=k} d(a_i, H_k^*)^p\ \text{vol}(\Delta_T)^p \\
&= \frac{1}{(k+1)} \left(\sum_{i=1}^m d(a_i, H_k^*)^p\right) \left(\sum_{T,|T|=k} \text{vol}(\Delta_T)^p\right), \qquad (2)
\end{aligned}$$

where in the second inequality, the fact that $d(a_i, V_{T\cup\{i\}}) \leq d(a_i, H_k^*)$ 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 (1) and (2) we get

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

□

LEMMA 2. *Let $S \subseteq [m]$ be a $(k+1)$-subset and V be any k-dimensional linear subspace of H_S. Then*

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

PROOF. W.l.o.g. we can identify H_S with $\mathbb{R}^{k+1}$ and the subspace V with $\text{span}(\{e_2, e_3, \ldots, e_{k+1}\})$, 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

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

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

3.2 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 3. *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 $\text{span}(\{A^{(\tau(1))}, A^{(\tau(2))}, \ldots, A^{(\tau(j))}\})$.

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

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 (2) (replacing $k+1$ with k) to get

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

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

4. 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^*.

4.1 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 $H_l = \text{span}(H' \cup l)$. 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 $H_l = \text{span}(H' \cup l)$, where H' is any $(k-1)$-dimensional subspace of H.

LEMMA 4. *Let S be a sample of $O\left((2k/\epsilon)^p \, (k/\epsilon) \log(k/\epsilon)\right)$ i.i.d. points from $a_1, a_2, \ldots, a_m$ 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

$$\begin{aligned}
\left(\sum_{i=1}^m d(a_i, H_{l_j})^p \right)^{\frac{1}{p}} &> \left(\sum_{i=1}^m d(a_i, H)^p \right)^{\frac{1}{p}} \\
&\quad + \frac{\epsilon}{k} \left(\sum_{i=1}^m \|a_i\|^p \right)^{\frac{1}{p}}. \qquad (3)
\end{aligned}$$

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

$$\sum_{i \in BAD} \|a_i\|^p > \left(\frac{\epsilon}{2k} \right)^p \sum_{i=1}^m \|a_i\|^p. \qquad (4)$$

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).

Inequality (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 12, there exists a line l' in $\text{span}(\{a_i\} \cup l_j)$ 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) fails to hold, or (b) the inequality (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) 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) 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

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

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

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

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. □

4.2 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 5. *The above algorithm returns a subset $S \subseteq [m]$ of $O\left(k\cdot(2k/\delta)^p(k/\delta)\log(k/\delta)\right)$ points such that*

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

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

PROOF. For a start, let us only look at step 2. From Lemma 4, 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 4 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

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

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 (1) $\dim(F_k) \leq 1$, (2) $\dim(F_k \cap \pi_{k-1}(H_{S_k})) \geq \min\{\dim(F_k), 1\}$, where π_t denotes projection to the orthogonal complement of $H_{S_1 \cup \cdots \cup S_t}$, and (3)

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

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,

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

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

5. 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.

5.1 Exponential drop in additive error

PROPOSITION 6. *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 5 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 7. *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 6 in t rounds by induction. □

5.2 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 8 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 8. *Let $V = \text{span}(S_0)$, where S_0 is a k-subset of rows picked by Approximate Volume Sampling procedure (see Subsection 3.2), $t = O(k \log k)$, and $S_1, S_2, \ldots, S_t$ as in Theorem 7. 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 7. □

6. DIMENSION REDUCTION FOR SUBSPACE APPROXIMATION

Dimension Reduction
Input: $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, $k > 0$, and a subspace V of dimension at least k.
Output: a subset $S \subseteq [m]$ of $O\left((k/\epsilon)^p \cdot k^2/\epsilon \cdot \log(k/\epsilon)\right)$.

1. Initialize $S = \emptyset$. While $|S| < O\left((20k/\epsilon)^p \cdot k^2/\epsilon \cdot \log(k/\epsilon)\right)$ do:
   (a) Pick a point a_i from the following distribution:
   $$\Pr(\text{picking } a_i) \propto d(a_i, \text{span}(V \cup H_S))^p.$$
   (b) $S \leftarrow S \cup \{a_i\}$.
2. Output S.

THEOREM 9. *Using a subspace* V *of dimension at least* k *with the guarantee*

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

the above algorithm finds, with probability that is at least $1 - (\epsilon/2k)^{2k^2/\epsilon}$, S *such that* $\text{span}(V \cup H_S)$ *contains a* k*-dimensional subspace* H' *satisfying*

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

PROOF. Let $\delta = \frac{\epsilon}{2k}$. For simplicity, we divide the steps of our algorithm into phases. Phase j, for $0 \leq j \leq k$, means that for the current sample S, there exists a k-dimensional subspace F_j such that $\dim(F_j \cap \text{span}(V \cup H_S)) \geq j$ and

$$\left(\sum_{i=1}^m d(a_i, F_j)^p\right)^{\frac{1}{p}} \leq (1+\delta)^j\left(\sum_{i=1}^m d(a_i, H_k^*)^p\right)^{\frac{1}{p}}.$$

So once a step is in phase j, all the steps following it must be in phase j', for some $j' \geq j$. Reaching phase k implies that we are done because then $F_k \subseteq \text{span}(V \cup H_S)$ and

$$\left(\sum_{i=1}^m d(a_i, F_k)^p\right)^{\frac{1}{p}} \leq (1+\delta)^k \left(\sum_{i=1}^m d(a_i, H_k^*)^p\right)^{\frac{1}{p}} \leq (1+\epsilon)\left(\sum_{i=1}^m d(a_i, H_k^*)^p\right)^{\frac{1}{p}}.$$

At the beginning of the algorithm, say $\dim(V \cap H_k^*) = j$. Then we attempt to execute the first step of the algorithm in phase j by taking $F_j = H_k^*$.

Consider the situation when we are attempting to execute the first step in phase j. Let us call $G = F_j \cap \text{span}(V \cup H_S)$; G will be a j-dimensional subspace. Let F_j^o and V^o be the orthogonal complements of G in F_j and $\text{span}(V \cup H_S)$, respectively. Let l be the line in F_j^o that makes the smallest angle with V^o, and l^o be the line in V^o that makes this angle with l. This smallest angle must be positive because we are trying to execute in phase j. Let $\hat{F}^o$ be the rotation of F_j^o so as to contain l^o, and $\hat{F}$ be the k-dimensional subspace given by $\text{span}(\hat{F}^o \cup G)$. Note that $\dim(\hat{F} \cap \text{span}(V \cup H_S)) = j'$, for some $j' > j$. If $\left(\sum_{i=1}^m d(a_i, \hat{F})^p\right)^{1/p} \leq (1+\delta)\left(\sum_{i=1}^m d(a_i, F_j)^p\right)^{1/p}$, then we do not execute in phase j but attempt to execute it in phase j' with $F_{j'} = \hat{F}$.

Now consider the situation after zero or more steps have executed in phase j, when we may have added a few dimensions to get our new $\text{span}(V \cup H_S)$. Let l be the line in F_j^o that is closest to the new V^o, i.e., orthogonal complement of the old G in the new span $\text{span}(V \cup H_S)$, and α_j' be the sine of its angle to the new V^o, i.e., there exists a line l^o in V^o such that α_j' is the sine of the angle between l and l^o. There are some cases:

1. $\alpha_j' = 0$ means that $\dim(F_j \cap \text{span}(V \cup H_S)) = j'$, for some $j' > j$ and we will attempt to execute the next step in phase j' with $F_{j'} = F_j$.

2. $\alpha_j' > 0$. As before, let $\hat{F}^o$ be the rotation of F_j^o so as to contain l^o, and $\hat{F}$ be the k-dimensional subspace given by $\text{span}(\hat{F}^o \cup G)$.
   (a) If it is the case that $\left(\sum_{i=1}^m d(a_i, \hat{F})^p\right)^{1/p} \leq (1+\delta)\left(\sum_{i=1}^m d(a_i, F_j)^p\right)^{1/p}$, then as before we consider the next step in some phase $j' > j$ with $F_{j'} = \hat{F}$.
   (b) Otherwise, we consider the next step in phase j itself.

Once we attempt to execute a step in phase k, then all subsequent steps will simply execute in phase k. Thus we have completely classified all the steps of our algorithm into $(k+1)$ phases. Now we will show that the algorithm succeeds, i.e., it executes some step in phase k, with high probability. To do this, we need to show that each phase contains few steps. Let us call a step of the algorithm *good* if (i) either the step executes in phase k, or (ii) the step executes in some phase $j < k$ and the point a_i sampled in the step has the property that $d(a_i^o, \hat{F}^o) > (1+\delta/2)d(a_i^o, F_j^o)$, where for any point a_i, a_i^o denotes the projection of a_i into the orthogonal complement of G.

Consider some phase $j < k$ in which we execute one or more steps. We bound the number of good steps in phase j. Let us use α_j to denote the sine of the angle between ℓ^o and ℓ before the execution of the first step in the phase, and α_j' to denote the same quantity at any subsequent point in the phase. We first bound α_j. Let $\bar{a}_i$ denote the projection of a_i onto F_j, and $\bar{a}_i^o$ denote the projection of $\bar{a}_i$ into the orthogonal complement of G. Focussing on the beginning of phase j, we have

$$
\begin{aligned}
&\alpha_j \left(\sum_{i=1}^{m} \|\bar{a}_i^o\|^p\right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i=1}^{m} d(\bar{a}_i^o, V^o)^p\right)^{\frac{1}{p}} \\
&= \left(\sum_{i=1}^{m} d(\bar{a}_i, \text{span}(V \cup H_S))^p\right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i=1}^{m} d(\bar{a}_i, a_i)^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{m} d(a_i, \text{span}(V \cup H_S))^p\right)^{\frac{1}{p}} \\
&\qquad \text{(by Minkowski's inequality)} \\
&= \left(\sum_{i=1}^{m} d(a_i, F_j)^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{m} d(a_i, \text{span}(V \cup H_S))^p\right)^{\frac{1}{p}} \\
&\leq \left(2 + (1+\delta)^j\right) \left(\sum_{i=1}^{m} d(a_i, H_k^*)^p\right)^{\frac{1}{p}} \\
&\leq 4 \left(\sum_{i=1}^{m} d(a_i, H_k^*)^p\right)^{\frac{1}{p}}, \qquad (5)
\end{aligned}
$$

where in the penultimate inequality we use $\delta < 1/2k$ and our initial assumption about V.

If a step in phase j is good, then by Lemma 12, there is a line in $\text{span}(a_i^o, l^o)$ for which the sine of its angle with F_j^o is at most $(1 - \delta/4)$ times the value of α_j' before the step. (Here a_i is the point that is sampled in the good step.) That is, the value of α_j' after the step is at most $(1 - \delta/4)$ times its previous value.

Hence, if we encounter $O(1/\delta \log 1/\delta)$ good steps in phase j, then after these steps we have $\alpha_j' \leq (\delta/4)\alpha_j$. Hence,

$$
\begin{aligned}
&\left(\sum_{i=1}^{m} d(a_i, \hat{F})^p\right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i=1}^{m} d(a_i, \bar{a}_i)^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{m} d(\bar{a}_i, \hat{F})^p\right)^{\frac{1}{p}} \\
&\qquad \text{(by Minkowski's inequality)} \\
&\leq \left(\sum_{i=1}^{m} d(a_i, F_j)^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{m} d(\bar{a}_i, \hat{F})^p\right)^{\frac{1}{p}} \\
&= \left(\sum_{i=1}^{m} d(a_i, F_j)^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{m} d(\bar{a}_i^o, \hat{F}^o)^p\right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i=1}^{m} d(a_i, F_j)^p\right)^{\frac{1}{p}} + \alpha_j' \left(\sum_{i=1}^{m} \|\bar{a}_i^o\|^p\right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i=1}^{m} d(a_i, F_j)^p\right)^{\frac{1}{p}} + \frac{\delta}{4}\, \alpha_j \left(\sum_{i=1}^{m} \|\bar{a}_i^o\|^p\right)^{\frac{1}{p}} \\
&\leq (1+\delta) \left(\sum_{i=1}^{m} d(a_i, F_j)^p\right)^{\frac{1}{p}},
\end{aligned}
$$

where in the last inequality we used equation (5). This implies that the next step will be in phase j', for some $j' > j$, according to our case analysis of phases. We conclude that a phase will not see more than $O(1/\delta \log 1/\delta)$ good steps.

Our algorithm executes $N = O\left((10/\delta)^p \cdot k/\delta \cdot \log(1/\delta)\right)$ steps. The event that it fails to reach phase k in these many steps implies the event that it had less than $O(k/\delta \log 1/\delta)$ good steps in its entire execution. From Lemma 10, we know that a step is good with probability at least $(\delta/10)^p$. Thus the probability that the algorithm fails to reach phase k in N steps is bounded by $\delta^{k/\delta}$.

Therefore, with probability at least $1 - \delta^{k/\delta}$, in the end $\text{span}(V \cup H_S)$ contains a subspace H' of dimension k such that

$$
\begin{aligned}
\left(\sum_{i=1}^{m} d(a_i, H')^p\right)^{\frac{1}{p}} &\leq (1+\delta)^k \left(\sum_{i=1}^{m} d(a_i, H_k^*)^p\right)^{\frac{1}{p}} \\
&\leq (1+\epsilon) \left(\sum_{i=1}^{m} d(a_i, H_k^*)^p\right)^{\frac{1}{p}}.
\end{aligned}
$$

□

LEMMA 10. *Suppose that the current step of our algorithm is in phase $j < k$. Then with probability at least $(\delta/10)^p$, the point a_i sampled in the step has the property that $d(a_i^o, \hat{F}^o) > (1 + \delta/2) d(a_i^o, F_j^o)$.*

PROOF. We must have

$$
\left(\sum_{i=1}^{m} d(a_i, \hat{F})^p\right)^{\frac{1}{p}} > (1+\delta) \left(\sum_{i=1}^{m} d(a_i, F_j)^p\right)^{\frac{1}{p}}, \qquad (6)
$$

according to our case analysis of phases. We call a point a_i "witness" if

$$
d(a_i, \hat{F}) > \left(1 + \frac{\delta}{2}\right) d(a_i, F_j).
$$

Let $W \subseteq [m]$ correspond to the set of all "witness" points. We claim that

$$
\begin{aligned}
&\left(\sum_{i \in W} d(a_i, \text{span}(V \cup H_S))^p\right)^{\frac{1}{p}} \\
&\geq \frac{\delta}{10} \left(\sum_{i=1}^{m} d(a_i, \text{span}(V \cup H_S))^p\right)^{\frac{1}{p}},
\end{aligned}
$$

for the current sample S, that is, with probability at least $(\delta/10)^p$ our algorithm picks a "witness" point a_i in the next step. Suppose this is not the case. Then, let h_i be the projection of a_i onto $\text{span}(V \cup H_S)$.

We have $d(a_i, \hat{F}) \leq \left(1 + \frac{\delta}{2}\right) d(a_i, F_j)$ for $i \in [m] \setminus W$, and for $i \in W$, we have

$$
\begin{aligned}
d(a_i, \hat{F}) &\leq d(a_i, h_i) + d(h_i, \hat{F}) \\
&\leq d(a_i, h_i) + d(h_i, F_j) \\
&\text{(because for any } h \in \text{span}(V \cup H_S),\ d(h, \hat{F}) \leq d(h, F_j)) \\
&\leq 2d(a_i, h_i) + d(a_i, F_j) \\
&\leq \left(1 + \frac{\delta}{2}\right) d(a_i, F_j) + 2d(a_i, h_i).
\end{aligned}
$$

Using these with Minkowski's inequality, we get

$$\left(\sum_{i=1}^{m} d(a_i,\hat{F})^p\right)^{\frac{1}{p}}$$
$$\leq \left(1+\frac{\delta}{2}\right)\left(\sum_{i=1}^{m} d(a_i,F_j)^p\right)^{\frac{1}{p}} + 2\left(\sum_{i\in W} d(a_i,h_i)^p\right)^{\frac{1}{p}}$$
$$\leq \left(1+\frac{\delta}{2}\right)\left(\sum_{i=1}^{m} d(a_i,F_j)^p\right)^{\frac{1}{p}} + \frac{2\delta}{10}\left(\sum_{i=1}^{m} d(a_i,\mathrm{span}(V\cup H_S))^p\right)^{\frac{1}{p}}$$
$$\leq \left(1+\frac{\delta}{2}\right)\left(\sum_{i=1}^{m} d(a_i,F_j)^p\right)^{\frac{1}{p}} + \frac{2\delta}{5}\left(\sum_{i=1}^{m} d(a_i,H_k^*)^p\right)^{\frac{1}{p}}$$
(by initial assumption on V in Theorem 9)
$$\leq (1+\delta)\left(\sum_{i=1}^{m} d(a_i,F_j)^p\right)^{\frac{1}{p}},$$

which is a contradiction to our assumption (see equation (6)).

Therefore, with probability at least $(\delta/10)^p$, the point a_i picked in the next step is a "witness" point. This means

$$\begin{aligned} d(a_i^o,\hat{F}^o) &= d(a_i,\hat{F}) \\ &> \left(1+\frac{\delta}{2}\right) d(a_i,F_j) \\ &= \left(1+\frac{\delta}{2}\right) d(a_i^o,F_j^o). \end{aligned}$$

□

7. DIMENSION REDUCTION FOR PROJECTIVE CLUSTERING

Let $A=\{a_1,\ldots,a_m\}\subseteq\mathbb{R}^n$ be the set of input points, and $k,s>0$ be integer parameters. We wish to find subspaces $H[1],\ldots,H[s]$ that minimize $\left(\sum_{i=1}^m d(a_i,H)^p\right)^{1/p}$ where, H denotes $H[1]\cup\cdots\cup H[s]$. Let $H^*[1],\ldots,H^*[s]$ denote the optimal set of subspaces, and H^* denote their union.

Dimension Reduction for Projective Clustering
Input: $a_1,a_2,\ldots,a_m\in\mathbb{R}^n$, $k,s>0$, and a subspace V of dimension at least k.
Output: a subset $S\subseteq[m]$ of size $\tilde{O}\left(\left(\frac{k^2}{\epsilon}\right)^p\frac{k^4 s}{\epsilon^2}\right)$

1. Initialize $S=\emptyset$. Until $|S|<\tilde{O}\left(\left(\frac{k^2}{\epsilon}\right)^p\frac{k^4 s}{\epsilon^2}\right)$ do:
   (a) Pick a point a_i from the following distribution:
   $$\Pr(\text{picking } a_i)\propto d(a_i,\mathrm{span}(V\cup H_S)).$$
   (b) $S\leftarrow S\cup\{a_i\}$.
2. Output S.

THEOREM 11. *Using a subspace V of dimension at least k with the guarantee*

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

the above algorithm finds, with probability at least $1-1/4ks$, S such that $\mathrm{span}(V\cup H_S)$ contains a s 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 $H'[1]\cup\ldots\cup H'[s]$.

PROOF. At a high level, the proof is analogous to that of Theorem 9 but is omitted for lack of space. □

Acknowledgements We thank Santosh Vempala and Sariel Har-Peled for various comments and suggestions.

8. REFERENCES

[1] D. Achlioptas, F. McSherry. Fast Computation of Low Rank Approximations. Proc. of the 33rd ACM Symposium on Theory of Computing (STOC), 2001.

[2] P. Drineas, R. Kannan, M. Mahoney. Fast Monte Carlo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix. Yale University Technical Report, YALEU/DCS/TR-1270, 2004.

[3] P. Drineas, M. Mahoney, S. Muthukrishnan. Polynomial time algorithm for column-row based relative error low-rank matrix approximation. DIMACS Technical Report 2006-04, 2006.

[4] A. Deshpande, L. Rademacher, S. Vempala, G. Wang. Matrix Approximation and Projective Clustering via Volume Sampling. Proc. of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA), 2006.

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

[6] 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.

[7] A. Frieze, R. Kannan, S. Vempala. Fast Monte-Carlo Algorithms for Finding Low-Rank Approximations. Proc. of IEEE Symposium on Foundations of Computer Science (FOCS), 1998.

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

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

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

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

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

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

APPENDIX

A. ANGLE-DROP LEMMA

LEMMA 12. *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 [13], 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 [13]) 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.

$$\alpha(l'') \leq \frac{|eq|}{|oe|} \leq \left(1 - \frac{\delta}{4}\right)\frac{|ee'|}{|oe|} \leq \left(1 - \frac{\delta}{4}\right)\alpha(l'), \quad (7)$$

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'. □