Optimal Outlier Removal in High-Dimensional Spaces

John Dunagan[*]   Santosh Vempala[*]

Abstract

We study the problem of finding an outlier-free subset of a set of points (or a probability distribution) in n-dimensional Euclidean space. As in [1], a point x is defined to be a β-outlier if there exists some direction w in which its squared distance from the mean along w is greater than β times the average squared distance from the mean along w. Our main theorem is that for any $\epsilon > 0$, there exists a $(1-\epsilon)$ fraction of the original distribution that has no $O(\frac{n}{\epsilon}(b+\log\frac{n}{\epsilon}))$-outliers, improving on the previous bound of $O(n^7 b/\epsilon)$. This bound is shown to be nearly the best possible. The theorem is constructive, and results in a $\frac{1}{1-\epsilon}$ approximation to the following optimization problem: given a distribution μ (i.e. the ability to sample from it), and a parameter $\epsilon > 0$, find the minimum β for which there exists a subset of probability at least $(1-\epsilon)$ with no β-outliers.

1 Introduction

The term "outlier" is a familiar one in many contexts. Statisticians have several notions of outliers. Typically they quantify how far the outlier is from the rest of the data, e.g. the difference between the outlier and the mean or the difference between the outlier and the closest point in the rest of the data. In addition, this difference might be normalized by some measure of the "scatter" of the set, e.g. the range or the standard deviation. Data points that are outside some threshold are labelled outliers.

Identifying outliers is a fundamental and ubiquitous problem. The outliers in a data set might represent experimental error, in which case it would be desirable to remove them. They could affect the performance of a computer program, by slowing down or even misleading an algorithm; machine learning is an area where outliers in the training data could cause an algorithm to find a wayward hypothesis. Even from a purely theoretical standpoint, removing outliers could lead to simpler mathematical models, or the outliers themselves might constitute the phenomenon of interest.

How does one find outliers? To address this question we have to first answer another: *what precisely is an outlier?* In this paper we will assume that the data consists of points (or a distribution) in n-dimensional Euclidean space. In the one-dimensional case, one could use one of the definitions alluded to above, viz. a point is an outlier if its distance from the mean is greater than some factor times the standard deviation. In figure 1, the top data set depicts this definition: the data points are the solid circles, and the mean, along with

[*] Department of Mathematics, MIT, Cambridge MA, 02139. Email: {jdunagan, vempala}@math.mit.edu
Supported in part by NSF Career award CCR-9875024.

the mean plus or minus one standard deviation, are the hash marks. The leftmost point is 1.86 standard deviations away from the mean.

The following generalization to higher dimensions was used in [1]. Let P be a set of points in $\mathcal{R}^n$. A point x in P is called a β-outlier if there exists a vector w such that the squared length of x along w is more than β times the average squared length of P along w, i.e. if

$$(w^T x)^2 > \beta \mathbf{E}_{x \in P}[(w^T x)^2]$$

Note that $(w^T x)^2$ is the squared distance along w from the origin. In figure 1, the bottom two pictures show how different points may be the furthest outliers for different choices of w. In each graph, the solid circles are the data points, the line is the direction w, and the hash marks along the line are the projections of the data points onto the line. The first problem we address is the following: does there exist a small subset of P whose removal ensures that the remaining set has no outliers? More precisely, what is the smallest β such that on removing a subset consisting of at most an ϵ fraction of the points, the remaining set has no β-outliers (*with respect to the remaining set*)?

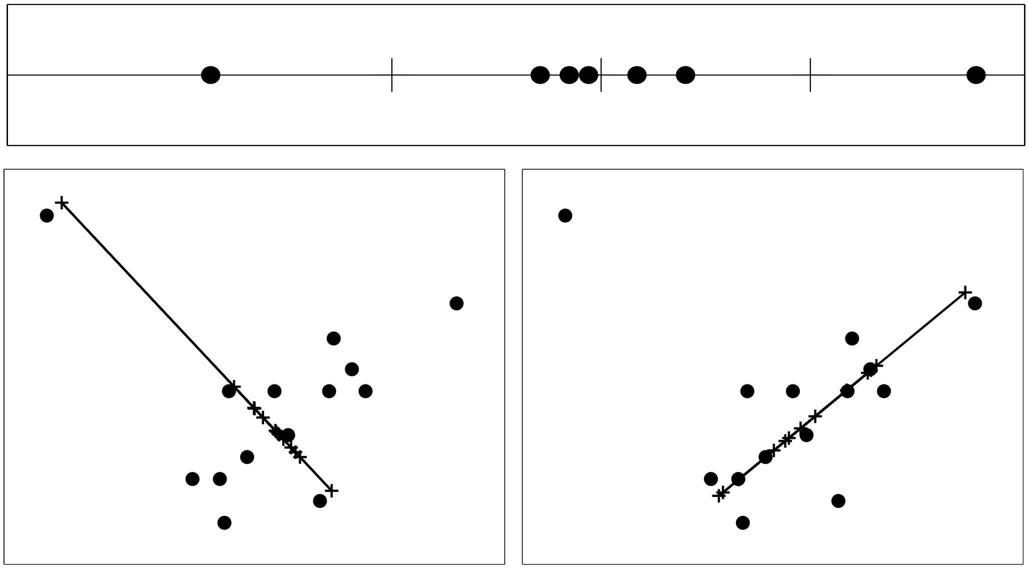

Figure 1: Defining Outliers

A natural approach is to find all β-outliers in the set and remove them. This can be done by first applying a linear transformation (described in section 2) that results in the average squared length of the distribution being 1 along every unit vector (the so-called *isotropic* position). Isotropic position was used in [4] to show that any convex set K in isotropic position contains a unit ball and is contained in a ball of radius n. Bringing a distribution into isotropic position allows us to identify outliers easily. Now a point that is a β-outlier simply has squared length more than β. The main difficulty is that the remaining set might still have outliers — it is possible that points that were previously not outliers now become outliers. Can this happen repeatedly and force us to throw out most of the set?

Our main result is that the answer to this question is "no" for a surprisingly small value of β. We present it below in a more general framework. Let $\mathcal{R}_b^n = \{0\} \cup \{x \in \mathcal{R}^n : 2^{-b} \leq |x| \leq 2^b\}$, i.e. the subset of n-dimensional space outside of a very small ball and inside a very large ball, plus the origin. In place of the point set P in the discussion above we have any probability distribution μ on $\mathcal{R}_b^n$. Note that $\mathcal{R}_b$ contains the set of all b-bit rationals, an object of frequent interest in theoretical computer science. For a probability distribution μ, let $\mu(S)$ denote the probability of a subset of space S.

Theorem 1 (Main Theorem) *Let μ be a probability distribution on $\mathcal{R}_b^n$. Then for every $\epsilon > 0$, there exists S and*

$$\beta = O\left(\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})\right)$$

such that
(i) $\mu(S) \geq 1 - \epsilon$
(ii) $\max\{(w^T x)^2 : x \in S\} \leq \beta \mathbf{E}[(w^T x)^2 : x \in S]$ *for all* $w \in \mathcal{R}^n$

The proof of the theorem (section 3) is constructive. In section 2 we describe (two variants of) an algorithm for outlier removal. The theorem can be proven using either variant. Although the theorem is not obvious, the algorithm is extremely simple. To convince the reader of this, we include a matlab implementation of the algorithm in section 8.

For a point set with m points ($m > n$) the algorithm runs in $O(m^2 n)$ time. In section 4 we show that the algorithm can also be used on an unknown distribution if it is allowed to draw random samples from the distribution. The number of samples required is $\tilde{O}(\frac{n^2 b}{\epsilon})$ and the running time is $\tilde{O}(\frac{b^2 n^5}{\epsilon^2})$.

Our algorithm is similar to the algorithm of [1], the immediate inspiration for our work. The bound on β in the theorem improves on the previous best bound of $O(\frac{n^7 b}{\epsilon})$ given in [1]. There it was used as a crucial component in the first polytime algorithm for learning linear threshold functions in the presence of random noise. Due to the high value of β, the bound on the running time of the learning algorithm, although polynomial, is a somewhat prohibitive $\tilde{O}(n^{28})$. In contrast, our theorem implies an improved bound of $\tilde{O}(n^5)$ for learning linear thresholds from arbitrary distributions in the presence of random noise. Further, our bound on β is asymptotically the best possible. This is shown in section 5 by an example where for any $\epsilon < \frac{1}{2}$, a bound on β better than $\Omega(\frac{n}{\epsilon}(b - \log\frac{1}{\epsilon}))$ is not possible.

Our main theorem gives an extremal bound on β. A natural follow-up question is whether one can achieve the best possible β for any particular distribution. Given a distribution μ and a parameter ϵ, we want to find a subset of probability at most ϵ whose removal leaves an outlier-free set with the smallest possible β. This question can be shown to be NP-hard even in the one-dimensional case by a reduction to subset-sum. In section 6 we prove that our algorithm achieves a $(\frac{1}{1-\epsilon})$-approximation to the best possible β for any given ϵ.

In some cases it may be desirable to translate the data set so that the origin coincides with the mean, rather than having a fixed origin. We prove the following corollary for standard deviations from the mean in section 7. Let μ be a probability distribution on I_b^n, where $I_b = \{\frac{p}{q} : |p|, |q| \in \{0, 1, 2, ..., 2^b - 1\}, q \neq 0\}$, i.e. I_b^n is the set of all n-dimensional vectors of b-bit rationals. Then for any $\epsilon > 0$, there exists a $(1 - \epsilon)$ fraction of the distribution such that along every direction, no point is further away from the mean than $O(\sqrt{\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})})$ standard deviations in that direction. We also give a $\left(\frac{1-\epsilon}{1-2\epsilon}\right)^2$-approximation algorithm for the corresponding optimization problem.

2 Algorithms for Outlier Removal

We state our algorithms for probability distributions over a restriction of $\mathcal{R}_b^n$ for ease of exposition. We remove this restriction in a single step at the end of section 3. Let $J_b = \{0\} \cup [2^{-b}, 2^b]$ and $J_b^n = \{x \in \mathcal{R}^n | \quad x_i \in J_b, \quad i = 1, 2, \ldots, n\}$, i.e. the subset of n-dimensional space with coordinates in J_b. Note that the coordinate axes play a special role in J_b^n. We assume for the remainder of this section that the probability distribution we are interested in is over J_b^n.

In order to detect outliers, we use a linear transformation. Let $M = \mathbf{E}[xx^T]$ where x is drawn according to the probability distribution μ. Since M is positive definite, there exists a matrix A such that $M = A^2$. Consider the transformed space $z = A^{-1}x$. This transformation preserves outliers: if z is a β-outlier in direction w in the transformed space, the corresponding $x = Az$ is a β-outlier in direction $w' = A^{-1}w$ in the untransformed space, and vice versa. The transformed distribution is in *isotropic* [4] position, and we will refer to the transformation as *centering*. Such transformations have previously been used to make geometric random walks more efficient [3]. If M does not have full rank, it is still positive semi-definite, and we instead center μ in the span of M.

For an isotropic distribution, any point x that is an outlier for some direction w is also an outlier in the direction x. This follows from the fact that an isotropic distribution has $\mathbf{E}[(w^T x)^2] = 1$ for every w such that $|w| = 1$, and that the projection of the point x on to a direction w is greatest when $w = x/|x|$. Thus outlier identification is easy for isotropic distributions.

The first algorithm has the following simple form: while there are β-outliers, throw them out; if at any point we are very close to a lower dimensional subspace, drop to the lower dimensional subspace. Stop when there are no outliers. In the description below, μ is the given distribution, c is an absolute constant and $\beta = \gamma^2 = c(\frac{n}{\epsilon}(b + \log \frac{n}{\epsilon}))$.

Algorithm 1 (Restriction to Ellipsoids):

1. Center μ. If there exists x such that $|x| > \gamma$, let $S = \{x : |x| \leq \gamma\}$. Retain only points in S.
2. Transform back to the original space. If there is some coordinate axis i such that $\Pr[x_i \geq 2^{-b}] \leq \frac{\epsilon}{3n}$, throw out all points x with $x_i \geq 2^{-b}$.
3. Repeat until neither of the above conditions is met.

The following variant of the above algorithm will be significantly easier to analyze. Whereas in the previous algorithm, we removed outliers in every direction in one step, in Algorithm 2, we only remove outliers in one direction per step.

Algorithm 2 (Restriction to Slabs):

1. Center μ. If there exists a unit vector w such that $\max\{(w^T x)^2\} > \gamma^2$, let $S = \{x : (w^T x)^2 \leq \gamma^2\}$. Retain only points in S.
2. Transform back to the original space. If there is some coordinate axis i such that $\Pr[x_i \geq 2^{-b}] \leq \frac{\epsilon}{3n}$, throw out all points x with $x_i \geq 2^{-b}$.
3. Repeat until neither of the above conditions is met.

3 Proof of the Main Theorem

When either algorithm terminates, we clearly have a β-outlier free subset. It remains to show that we do not discard too much of the distribution. The main idea of the proof is to show that in every step the volume of an associated dual ellipsoid increases. By bounding the total growth of the dual ellipsoid volume over the course of the algorithm, we will deduce that no more than a certain fraction of the original probability mass is thrown out before the algorithm terminates. Special care will be taken to deal with the possibility that at some step the distribution μ becomes concentrated on a subspace of lower dimension.

Towards this end, we will need some definitions. For a matrix M such that $M = A^2$ define the ellipsoids $E(M)$ and $W(M)$ as

$$E(M) = \{x : |A^{-1}x| \leq 1\} \quad \text{and} \quad W(M) = \{x : |Ax| \leq 1\}.$$

We will refer to $E(M)$ and $W(M)$ as the primal inertial ellipsoid and the dual ellipsoid respectively. For any subset S of $\mathcal{R}^n$, we denote by M_S the matrix given by

$$M_S = \mathbf{E}[xx^T : x \in S] \Pr[x \in S] = \sum_{x \in S} \mu(x) x x^T$$

In other words, M_S is the matrix obtained by restricting μ to S (zeroing out points outside of S). We denote this restricted probability distribution directly by $\mu_{|S}$. The useful property attained by centering with respect to $\mu_{|S}$ (the restriction of the original distribution to S) is that

$$\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S] = 1$$

for every unit vector w, where the expectation and probability are with respect to x drawn from μ. We will actually prove theorem 1 with $\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]$ in place of $\mathbf{E}[(w^T x)^2]$. Note that this is a stronger statement than the original theorem.

We will also need the following elementary facts about ellipsoids: the volume of a full-dimensional ellipsoid is given by the product of the axis lengths times the volume of the unit ball, which we will denote by $f(n)$. The ellipsoid $\{x : |A^{-1}x| \leq 1\}$ has axes given by the eigenvectors of A; It follows that $Vol(W(M))Vol(E(M)) = (f(n))^2$, a function solely of the dimension.

Lemma 1 relates the dual volume growth to the loss of probability mass, and lemma 2 upper bounds the total dual volume growth.

Lemma 1 (Restriction to a Slab) *Let γ be fixed, and let μ be a full-dimensional isotropic distribution. Suppose $\exists w, |w| = 1$ such that*

$$\max\{(w^T x)^2\} > \gamma^2 \mathbf{E}[(w^T x)^2]$$

Let $S = \{x : (w^T x)^2 \leq \gamma^2\}$ and $p = \Pr[x \notin S]$. Then

$$Vol(W(M_S)) \geq e^{p\gamma^2/2} Vol(W(M))$$

Proof: Let $a^2 = \mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]$. Starting from the identity

$$\mathbf{E}[(w^T x)^2] = \mathbf{E}_{x \in S}[(w^T x)^2] \Pr[x \in S] + \mathbf{E}_{x \notin S}[(w^T x)^2] \Pr[x \notin S]$$

and using that $(w^T x)^2 \geq \gamma^2$ for all x not in S, we get that $1 \geq a^2 + \gamma^2 p$, which implies

$$a^2 \leq 1 - \gamma^2 p \leq e^{-\gamma^2 p}$$

We now construct a vector w' of length $1/a$ belonging to the dual ellipsoid. Letting $w' = w/a$ suffices since w is a unit vector by assumption and

$$a^2 = \mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S] = w^T M_S w$$

$$\Rightarrow \quad 1 = w'^T M_S w' \quad \Rightarrow \quad w' \in W(M_S)$$

We also show that every $v \in W(M)$ also belongs to $W(M_S)$. We have

$$M_S = M - \sum_{x \notin S} \mu(x) x x^T.$$

Hence,

$$v^T M_S v = v^T M v - \sum_{x \notin S} \mu(x) v^T x x^T v$$

$$= v^T M v - \sum_{x \notin S} \mu(x) (v^T x)^2 \leq v^T M v \leq 1$$

implying that $v \in W(M_S)$ (the last step is from the assumption that $v \in W(M)$). The length of a point on the boundary of an ellipsoid lower bounds the length of the longest axis. Since at least one axis of the dual has length $1/a$, and all the other axes have length at least 1, $Vol(W(M_S)) \geq (1/a) f(n)$ while $Vol(W(M)) = f(n)$, implying the dual volume grows by at least a factor of $e^{\gamma^2 p/2}$. If the dual ellipsoid has infinite volume after this iteration, then the statement is still true because the dual ellipsoid had finite volume at the beginning of the iteration. This concludes the proof of lemma 1. ■

Lemma 2 (Dual Volume Growth) *Let μ be an initial full-dimensional distribution, and let $\mu_{|_{S^*}}$ be the final distribution resulting from application of either algorithm. Assume $\mu_{|_{S^*}}$ is full-dimensional. Let $L = (b + \log \frac{n}{\epsilon})$. Then*

$$Vol(W(M)) \geq 2^{-nL}$$

$$Vol(W(M_{S^*})) \leq 2^{nL}$$

Proof: First we lower bound the initial dual volume, $Vol(W(M))$. Consider any vector v of length at most $2^{-b}/\sqrt{n}$. Since the longest possible x belonging to μ is of length at most $2^b\sqrt{n}$, we have that $v^T M v = \mathbf{E}[(v^T x)^2] \leq 1$ and so v belongs to the dual ellipsoid. Thus the dual ellipsoid initially has volume at least $(2^{-b}/n)^n = 2^{n(-b-\log n)} \geq 2^{-nL}$ (using the inscribed cube to lower bound the volume of the ball).

Next we upper bound $Vol(W(M_{S^*}))$. Consider any vector v of length $|v|$. Then v has length at least $|v|/\sqrt{n}$ on some coordinate axis i. By the property that $\mu_{|_{S^*}}$ is full-dimensional, the test condition in step 2 of either algorithm was never satisfied, and thus at least an $\epsilon/3n$ fraction of the points x have value at least 2^{-b} on this coordinate axis. This yields that

$$v^T M v = \mathbf{E}[(v^T x)^2] \geq (|v|^2/n)(2^{-2b}) \Pr[x_i \geq 2^{-b}] \geq (|v|^2/n)(2^{-2b})(\epsilon/3n)$$

For v to be in the dual ellipsoid, we must have $v^T M v \leq 1 \leftrightarrow |v| \leq 2^{b+\log n+1/2\log 3/\epsilon}$. Thus the ultimate volume of the dual ellipsoid is no more than $2^{n(b+\log n+1/2\log 3/\epsilon)} \leq 2^{nL}$ (using the containing cube to upper bound the volume of the ball). ■

Using lemmas 1 and 2, we prove that Algorithm 2 terminates with S satisfying theorem 1.

Proof: Suppose that the algorithm terminates with subset S^* after having thrown out no more than ϵ' of the original probability mass. Then we have that for every w,

$$\max\{(w^T x)^2 : x \in S^*\} \leq \gamma^2 \mathbf{E}[(w^T x)^2 : x \in S^*] \Pr[x \in S^*]$$

We remind the reader again that normalizing $\mu_{|S^*}$ so that it is a probability distribution on points from μ, rather than with points outside of S^* replaced by zeros, increases the right-hand side of this inequality by the factor $1/\mu(S^*)$, but does not increase the left-hand side. Thus the inequality will still be true even if we normalize $\mu_{|S^*}$. We thus achieve

$$\beta = \gamma^2 = c\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})$$

It now remains to show that $\epsilon' \leq \epsilon$, i.e. that we do not throw out more of the probability mass than claimed. First suppose $\mu_{|S^*}$ is full dimensional. Let $L = (b + \log\frac{n}{\epsilon})$.

Suppose that during the i^{th} iteration of the algorithm through step 1, a p_i fraction of the original points are thrown out. Then the total amount thrown out is $\sum p_i$. By lemma 1, the total amount of dual volume increase is $\prod_i e^{p_i\gamma^2/2} = e^{\frac{\gamma^2}{2}\sum p_i}$. Comparing this to our bound on the total increase in the dual volume from lemma 2 yields

$$e^{\frac{\gamma^2}{2}\sum p_i} \leq 2^{2nL}$$

For our choice of γ^2 with $c = 36$, we have that $\sum p_i \leq \epsilon/3$.

We now extend the proof to the case that the final distribution is not full-dimensional. This drop in dimension is the main issue of the proof once we have lemmas 1 and 2. Suppose that at some step of the algorithm, we move from M of dimension k to M' of dimension k'. We then want to restrict ourselves for the rest of the algorithm to the k'-dimensional subspace spanned by M'. That is, both $W(M')$ and $E(M')$ are k'-dimensional objects. We denote the fact that we are now considering the volume of a lower-dimensional object by adding a subscript to our $Vol(\cdot)$ function. We show two things

(a) Our upper bound on the total volume of the dual ellipsoid decreases by $2^{(k-k')L}$

(b) $\frac{Vol_k(W(M))}{Vol_{k'}(W(M'))} \leq 2^{(k-k')L}$

Thus the amount by which we are away from our upper bound on the total growth of the dual volume cannot have increased. (a) is immediate from the way we calculated the upper bound on the final dual volume. (b) will be proved shortly. We now conclude the argument assuming (b). Every time μ drops in dimension (i.e., the rank of M decreases), we may not have made any progress in increasing the dual volume, and thus we cannot apply lemma 1 to the probability mass thrown out in this step. However, taking $\gamma^2 \geq 3n/\epsilon$ yields that we won't throw out more than an $\epsilon/(3n)$ fraction of the total probability mass on any one step. This follows from

$$\Pr[x \notin S] = \Pr[(w^T x)^2 \geq \gamma^2] \leq \frac{\mathbf{E}[(w^T x)^2]}{\gamma^2} = \frac{1}{\gamma^2}$$

Since our choice for γ^2 already satisfied the above criterion, we conclude that every time we drop in dimension, we throw out no more than an $\epsilon/(3n)$ fraction of the probability mass. Since we might also drop in dimension by explicitly throwing out up to an $\epsilon/(3n)$ fraction of the probability mass (the criterion for step 2 of Algorithm 2), we can upper bound the total amount of probability mass thrown out without any increase in the dual by $\frac{2\epsilon}{3n} \times$ (total number of dimensions dropped) $\leq \frac{2\epsilon}{3}$. Combining this with $\sum p_i \leq \epsilon/3$, we see that we throw out no more than ϵ probability mass over the life of the algorithm. Thus $\epsilon' \leq \epsilon$, and our final bound on β is

$$\beta \leq \frac{36n}{\epsilon}(b + \log \frac{n}{\epsilon})$$

To prove property (b) we use the following simple lemma about ellipsoids.

Lemma 3 (Ellipsoid Slices) *Consider an n-dimensional ellipsoid E with axis lengths $a_1 \geq \ldots a_n$. Now take any k-dimensional slice C through the center of E. Then*

$$Vol_k(C) \geq f(k) \prod_{i=n-k+1}^{n} a_i.$$

(b) is a corollary of lemma 3 because no axis of the dual ellipsoid has length more than 2^L, and the other axes of $W(M)$ can only grow when we throw out probability mass from μ. It only helps that $f(n)$ is monotonically decreasing in n.

Proof of lemma 3. The main tool is the Courant-Fischer Theorem [5].

Theorem 2 (Courant-Fisher) *Let A be a real symmetric $n \times n$ matrix, λ_i the i^{th} eigenvalue, $\lambda_1 \geq \ldots \lambda_n$. Then*

$$\lambda_i = \min_{U} \max_{x \in U, x \neq 0} \frac{x^T A x}{x^T x}$$

where the minimum is over all $(n-i+1)$-dimensional subspaces U.

Since the k dimensional slice is a subspace U with $i = n-k+1$, and the axis lengths of $W(M)$ are given by the eigenvalues of A, we find that the longest axis of the sliced ellipsoid has length at least a_{n-k+1}. Applying the same argument to the $k-1$ dimensional subspace of C perpendicular to the longest axis of the sliced ellipsoid, we find that the next longest axis has length at least a_{n-k+2}. Applying the argument to the remainder of the axes concludes the proof of lemma 3. This also concludes the proof of the main theorem. ■

We now give an alternate proof of the main theorem using the construction given by Algorithm 1. We begin by proving an analogue to lemma 1.

Lemma 4 (Restriction to an Ellipsoid) *Let γ be fixed, and let μ be a full-dimensional isotropic distribution. Let $S = \{x : (x^T x) \leq \gamma^2\}$ and $p = \Pr[x \notin S]$. Then*

$$Vol(W(M_S)) \geq e^{p\gamma^2/2} Vol(W(M))$$

Proof: First we establish the tradeoff for a radially symmetric distribution, and then we show that a radially symmetric distribution is the worst case for the tradeoff we want.

Let μ' be a radially symmetric distribution, and define M', S, and p as above. We then calculate the increase in $Vol(W(M'))$. Let $a^2 = \mathbf{E}_{\mu'}[(w^T x)^2 : x \in S] \Pr[x \in S]$ for any $w, |w| = 1$. From the center of an n-dimensional sphere of radius γ, the projection of the sphere on to any direction is sharply concentrated around $\gamma/\sqrt{n}$, and the squared expectation is exactly γ^2/n. Using the identity

$$\mathbf{E}[(w^T x)^2] = \mathbf{E}_{x \notin S}[(w^T x)^2] \Pr[x \notin S] + \mathbf{E}_{x \in S}[(w^T x)^2] \Pr[x \in S]$$

as in the proof of lemma 1, but now for any w, we deduce $1 \geq a^2 + \gamma^2 p/n$, and thus

$$a^n \leq \left(1 - \frac{\gamma^2 p}{n}\right)^{n/2} \leq e^{-\gamma^2 p/2}$$

As in the proof of lemma 1, we observe that $W(M'_S)$ includes a vector of length $1/a$ in the direction of w. Since this is now true for every w, the dual ellipsoid volume increases by at least a factor of $(1/a)^n$. This shows that in the case of a radially symmetric distribution,

$$Vol(W(M_S)) \geq e^{p\gamma^2/2} Vol(W(M))$$

Now we show that a radially symmetric distribution is the worst case for the tradeoff we want. Suppose there were some isotropic, full-dimensional distribution μ for which the statement of the lemma was not true. We construct a new isotropic, full-dimensional and radially symmetric distribution μ' for which the statement is also false.

We begin by noting that every point thrown out from μ is also thrown out from any rotation of μ – this just follows from the fact that μ is isotropic. Let μ' be the expectation of μ under a random rotation. That is, μ' is a radially symmetric distribution such that the probability of choosing x from μ' at distance less than r from the origin is exactly the same as the probability of choosing x from μ at distance less than r from the origin, for every r. Let M' correspond to μ'.

Consider an axis direction w_i of $E(M_S)$, $|w_i| = 1$. We have $a_i^2 = \mathbf{E}[(w_i^T x)^2 : x \in S] \Pr[x \in S]$. For $E(M'_S)$, denote the axis length for any axis (also just the radius of $E(M'_S)$) by $\bar{a}$. We find from the construction of μ' that

$$\bar{a}^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[(w_i^T x)^2 : x \in S] \Pr[x \in S] = \frac{1}{n} \sum_{i=1}^{n} a_i^2$$

One way to visualize this equality is to take μ and simply consider $\tilde{\mu}$ achieved by rotating the axes of μ onto the other axes of μ; since this is a discrete set of rotations, it is clear that the squared axis lengths of $\tilde{\mu}$ are just the arithmetic averages of the squared axis lengths of μ. Then we can make $\tilde{\mu}$ into μ' by taking a continuous set of rotations, without affecting the axis lengths from $\tilde{\mu}$.

We now consider the volume of $E(M'_S)$. We have

$$Vol(E(M'_S)) = f(n) \prod_{i=1}^{n} \bar{a} = f(n) \left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} a_i^2}\right)^n \geq f(n) \prod_{i=1}^{n} a_i = Vol(E(M_S)$$

using the arithmetic mean-geometric mean inequality. This implies that $Vol(W(M_S)) \geq Vol(W(M'_S))$. This concludes the proof of lemma 4.

■

Finally, we prove that Algorithm 1 terminates with S satisfying theorem 1.

Proof: Lemma 2 still holds. The rate of increase in the dual volume as we throw out probability mass when the dimension remains constant is clearly still good as well. However, our bound on the amount of probability mass p that can be thrown out in a single step is no longer $\epsilon/(3n)$. Instead, we have from the proof of lemma 4 only that
$0 \leq 1 - p\gamma^2/n$ which only leads to $p = O(\epsilon)$ for our chosen value of γ^2. Thus, the analysis for Algorithm 2 does not immediately apply.

To analyze Algorithm 2, we argued that any drop in the dual on the up to n possible steps in which the dimension dropped could not lead to more than an ϵ overall drop in the probability mass. Thus we were able to just argue that the maximum amount by which the dual volume still might grow couldn't increase in this step. To successfully analyze Algorithm 1, we must prove the stronger statement that if we throw out a lot of probability mass, and the dimension drops by only a small amount, then we still make significant progress on the overall growth of the dual volume. To be precise, letting M, M', p, k, k', L be as before (in the analysis of Algorithm 2), we prove

(c) $\frac{Vol_{k'}(W(M'))}{Vol_k(W(M))} \geq 2^{-(k-k')(L+1)}e^{p\gamma^2/4}$

(Compare to (b) in the analysis of Algorithm 1.) The idea for our proof of (c) is that any point either has a large component in the subspace that vanishes, or a large component in the subspace that remains — if the dimension drops by only a small amount, there cannot have been too many points with a large component in the subspace that vanishes, and the dual volume growth results from the discarded points with a large component in the subspace that remains. Let W_{remain} be the k'-dimensional space spanned by M', and W_{tossed} be the $(k-k')$-dimensional subspace of $span(M)$ orthogonal to W_{remain}. Let p_1 be the probability that the projection of x onto W_{remain} is at least $\frac{\gamma}{\sqrt{2}}$, and similarly let p_2 be the probability that the projection of x onto W_{tossed} is at least $\frac{\gamma}{\sqrt{2}}$. One of these two events happens for every point x. Thus $p \leq p_1 + p_2$.

Taking any unit vector $w \in W_{tossed}$ and using our favorite equation

$$\mathbf{E}[(w^Tx)^2] = \mathbf{E}_{x \notin S}[(w^Tx)^2]\Pr[x \notin S] + \mathbf{E}_{x \in S}[(w^Tx)^2]\Pr[x \in S],$$

we have that

$$1 \geq 0 + (\frac{\gamma^2}{2}\frac{1}{k-k'})p_1 \Rightarrow p_1 \leq \frac{2(k-k')}{\gamma^2}$$

where the factor $1/(k-k')$ comes from the expected squared projection of a $(k-k')$-dimensional unit vector on to a random direction.

Taking instead a unit vector $w \in W_{remain}$ and letting $\bar{a}^2 = \mathbf{E}[(w^Tx)^2 : x \in S]\Pr[x \in S]$, a similar analysis gives

$$\bar{a}^{k'} \leq \left(1 - \frac{p_2\gamma^2}{2k'}\right)^{k'/2}$$

Since $1/(\bar{a}^{k'})$ is a lower bound on the increase in the volume of the dual ellipsoid in the subspace W_{remain}, and the dual volume drop in moving to that subspace was at most $2^{-(k-k')L}$, we have an increase in the dual volume of at least $2^{-(k-k')L}e^{p_2\gamma^2/4}$. Using $p_2 \geq p - p_1 \geq p - \frac{2(k-k')}{\gamma^2}$ yields that this is $2^{-(k-k')L}e^{p\gamma^2/4}e^{(k-k')/2}$, which then implies (c).

The tradeoff expressed in (c) is slightly weaker than (b) from the analysis of Algorithm 2, but it is independent of whether we drop in dimension on a given iteration or not.

Suppose we throw out probability mass p_i the i^{th} time through step 1 of Algorithm 1. From (c), we have that dropping in dimension only takes away from our progress towards the upper bound on dual volume growth by at most a factor of two per dimension that we drop, and thus we find

$$e^{\frac{\gamma^2}{2}\sum p_i} \leq 2^{2n(L+1)}$$

Our choice of γ^2 with $c = 36$, implies $\sum p_i \leq 2\epsilon/3$. We throw out less than $\epsilon/3$ of the probability mass in step 2 over the life of the algorithm. Thus the entire algorithm throws out no more than an ϵ fraction of the probability mass. This concludes the analysis of Algorithm 1. ■

We now remove the technicality of restricting to J_b^n instead of $\mathcal{R}_b^n$.

Lemma 5 (Choice of Axes) *Let μ be a probability distribution over $\mathcal{R}_b^n$. There exists a rotation of μ, which we denote by μ', a subset of space T, and $B = O(b + \log\frac{n}{\epsilon})$ such that*
(i) $\mu'(T) \geq 1 - \frac{\epsilon}{2}$
(ii) $\mu'_{|_T}$ is a probability distribution over J_B^n

Proof: We construct μ' by randomly rotating μ. T is chosen to be the largest subset of space we may retain and still not have any points with too small a coordinate value in any axis (so T equals J_B^n, the set that we want $\mu'_{|_T}$ to be over).

Restricting to T consists of throwing away any points with value less than 2^{-B} along any axis. The expected amount of probability mass thrown away is then at most n times the amount thrown away by restricting along one axis (i.e., removing all points within the slab of width 2^{-B+1} centered at the origin and perpendicular to the one axis.)

Consider a single point $x \in R_b^n, x \neq 0$. The following bound on the probability of a small projection in the direction of a random unit vector is proved in [2].

$$\Pr_{w\in S_n}[(w^Tx)^2 \leq \frac{x^2}{nC}] \leq \frac{4}{C}$$

Let $C = \frac{16n}{\epsilon}$ and $B = b + \log\frac{16n^2}{\epsilon}$. Since $|x| \geq 2^{-b}$, we have for the chosen value of C that $2^{-2B} \leq \frac{x^2}{nC}$. As we only throw out the point if it's projection is less than 2^{-2B}, this choice of B implies that the probability of throwing out a point is at most $\frac{\epsilon}{4n}$ for one axis. Thus with probability at least a half, we throw out no more than an $\frac{\epsilon}{2}$ fraction of the distribution after considering all n axes. ■

Let us now consider the bound on β we achieve when our initial distribution is not over J_b^n, but rather over R_b^n. We apply the transformation in lemma 5 as an initial step, and then apply the algorithm as before. Our upper bound on β does increase, but only by a constant factor. This is shown in our concluding calculation, where we use $\frac{\epsilon}{2}, B$ in place of ϵ, b.

$$\beta = O\left(\frac{n}{\epsilon/2}(B + \log\frac{n}{\epsilon})\right) = O\left(\frac{2n}{\epsilon}(b + \log\frac{16n^2}{\epsilon} + \log\frac{n}{\epsilon})\right) = O\left(\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})\right).$$

4 Efficiency

In this section we describe polynomial time versions of both algorithms. The computational model is to allow multiplications and additions in unit time.

4.1 Point sets

Suppose the distribution μ is specified explicitly as a set of m points with weights corresponding to probabilities. Then we can achieve exactly the stated value of β with either algorithm deterministically. The running time for either algorithm is given by the time to compute M ($O(mn^2)$), the time to center the distribution ($O(n^3 + mn^2)$), the time to find an outlier ($O(mn)$), and the need to repeat the whole process up to m times. The amount of time spent in step 2 of either algorithm is negligible. This yields a time bound of $O(m^2n^2 + mn^3)$.

In the above discussion we made the worst case assumption that only one data point was thrown out in each iteration of centering and looking for outliers. In the case that a single data point is throw out, centering the distribution can be done more efficiently. If the distribution is initially isotropic, and v of probability p is removed, then the new isotropic distribution is achieved by replacing each vector u by $u - \left(1 - \sqrt{1 - v^2 p}\right)\frac{(u^T v)v}{v^2}$ An intuitive explanation for this formula is that we are just correcting the inertial ellipsoid in the direction of v. Using this observation, we compute M from scratch once ($O(mn^2)$), center the distribution from scratch once ($O(n^3 + mn^2)$), and then find an outlier ($O(mn)$) and recenter using our formula above ($O(mn)$) a total of at most m times. This yields the improved time bound of $O(m^2n + mn^2 + n^3)$.

4.2 Arbitrary distributions

The more interesting problem is where we are not given μ explicitly, but rather only the ability to sample from μ. The outlier-free restriction of μ will be specified as the part of μ contained in an ellipsoid. The algorithm for distributions is:

1. Apply lemma 5 (if necessary) to get a set of "clean" axes.
2. Get a set $P = \{x_1, \dots, x_m\}$ of m samples from μ.
3. Run the outlier removal algorithm algorithm on the discrete point set P with parameter Γ^2.
4. Let P' be the outlier free subset of P. Then the outlier-free restriction of μ is given by $(1 + \delta)^2\Gamma^2 E(\bar{M})$, where $\bar{M} = \frac{1}{m}\sum_{x_i \in P'} x_i x_i^T$ and $\delta > 0$ is an accuracy parameter.

The main theorem of this section is the following.

Theorem 3 (Sample Complexity) *Let* $m = (\frac{\Gamma^2}{\delta^2}(n \log \frac{n}{\delta} + \log \frac{b+\log\frac{n}{\epsilon}}{\delta}))$*. With high probability, either outlier removal algorithm run with parameter* $\Gamma^2 = (1+\delta)^2\gamma^2$ *returns a set* T *satisfying*
(i) $\mu((1+\delta)^2 T) \geq 1-\epsilon$
(ii) $(1+\delta)^2 T$ *has no* $(1+O(\delta))\gamma^2$*-outliers*
where (γ^2, ϵ) *is achieved by the deterministic omniscient algorithm (omniscient in that it knows the distribution exactly).*

For the remainder of this section, assume that the deterministic omniscient algorithm with parameter γ^2 finds a subset S such that $\mu(S) \geq 1-\epsilon$, and $\mu_{|S}$ has no γ^2-outliers. The statement "$\mu_{|S}$ has no γ^2-outliers", or simply "S has no γ^2-outliers" (since μ is implicit), is exactly that $\forall w, \max\{(w^T x)^2 : x \in S\} \leq \gamma^2 \mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]$. Since S and T are always convex, whenever we have $\forall w, \max\{(w^T x)^2 : x \in S\} \leq \max\{(w^T x)^2 : x \in T\}$, we will be able to conclude that $S \subseteq T$ (or at least $\mu(T \setminus S) = 0$ which is just as good). We know that $\gamma^2 = \tilde{O}(\frac{bn}{\epsilon})$ is always achievable, but in some cases we may do better, and our bound on running time is proved for arbitrary values of γ^2.

Suppose that at some step we can estimate $E(M)$ to within $1 \pm \delta$ in every direction. Let $\Gamma^2 = (1+\delta)^2\gamma^2$. Then every point that we perceive to be a Γ^2-outlier will be at least a γ^2-outlier with respect to the true distribution, and so removing them does not throw away any point that the deterministic algorithm keeps. Similarly, if we perceive the distribution to have no Γ^2-outliers, the true distribution will have no $(1+\delta)^2\Gamma^2$-outliers. Before removing outliers, we may not have that $\bar{M}$ (our working estimate of M) is within $1 \pm \delta$ of M. However, whenever we are wrong by more than $1+\delta$, there is some true outlier with respect to the original distribution that we throw out even using our flawed estimate $\bar{M}$. This line of reasoning (made rigorous) will allow us to find a $(1+O(\delta))\gamma^2$-outlier-free subset in space, where γ^2 is the parameter achieved by the deterministic version of the algorithm. In lemma 6 we show this for a particular direction in a particular iteration. In lemma 7 we extend this to all iterations, and finally in the proof of theorem 3 we extend this to all directions and all iterations. We also show $m = O(\frac{\Gamma^2}{\delta^2}(n \log \frac{n}{\delta} + \log \frac{b+\log\frac{n}{\epsilon}}{\delta}))$ samples suffice to achieve our stated goal of a $(1+O(\delta))\gamma^2$-outlier free set with high probability.

Lemma 6 (Outlier Detection, One Iteration) *Fix a direction* w*. Let* S *be a subset of space. Let our number of samples be* $m = O(\gamma^2/\delta^2)$*, and consider the sample distances in direction* w *given by* $\{w^T x_i\}$*. Let* $y = \mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]$ *and* $\bar{x}$ *be the sample variance,*

$$\bar{x} = \frac{1}{m} \sum_{x_i \in S} (w^T x_i)^2.$$

Then with constant probability
(i) $\max\{(w^T x)^2 : x \in S\} \leq \gamma^2 y \Rightarrow (1-\delta)y \leq \bar{x} \leq (1+\delta)y.$
(ii) $\max\{(w^T x)^2 : x \in S\} \leq \gamma^2 y$ *and* $T = \{x : (w^T x)^2 \leq \Gamma^2 \bar{x}\} \Rightarrow S \subset T.$

Proof: Property (i) says that we do correctly estimate the variance of an outlier-free restriction of the distribution, and property (ii) assures us that any outlier-free restriction of the distribution has no probability mass past Γ^2 times the sample variance (i.e., we can always safely throw away probability mass using the sample variance). Both claims are for a fixed direction w . Note that S found by the deterministic omniscient algorithm satisfies the conditions of both (i) and (ii).

Let X_i be the random variable representing the squared distance of x_i along the direction w, $X_i = (w^T x_i)^2$. Without loss of generality, assume $\max\{(w^T x)^2 : x \in S\} = 1$ (by an appropriate scaling). First we show (i). Since $\mu_{|_S}$ has no γ^2-outliers, we have $y \geq \frac{1}{\gamma^2}$. Applying the Chernoff bound to determine the probability that $\bar{x}$ is not a good estimate for y, we have

$$\Pr[|m\bar{x} - my| \geq \delta my] \leq e^{-\delta^2 my/3}$$

This occurs with constant probability for $m = O(\frac{\gamma^2}{\delta^2})$.

Now we show (ii). Let T be as above, and again assume $\max\{(w^T x)^2 : x \in S\} = 1$ without loss of generality. If S has no γ^2-outliers, then $y \geq \frac{1}{\gamma^2}$, and we would have found $\bar{x}$ to be an accurate estimate by the analysis in the previous paragraph. In this case, $\bar{x} \geq \frac{y}{1+\delta}$, and S has no γ^2-outliers implies $\max\{(w^T x)^2 : x \in S\} \leq \gamma^2 y \leq \Gamma^2 \bar{x}$. This then implies $S \subseteq T$. ■

Lemma 7 (Outlier Detection, Many Iterations) *Fix w. Let $m = O(\frac{\gamma^2}{\delta^2} \log \frac{b + \log \frac{n}{\epsilon}}{\delta})$ Then either outlier removal algorithm restricted to w with parameter Γ^2 produces a subset of space $T = \{x : (w^T x)^2 \leq t\}$ (for some value t) such that, with constant probability,*
(i) For any subset of space S that has no γ^2-outliers along w, $S \subseteq T$.
(ii) $(1+\delta)T$ has no $(1+\delta)^8\gamma^2$-outliers along w.

Proof: By "either outlier removal algorithm restricted to w", we simply mean the one-dimensional versions of the two algorithms. Consider S achieved by the deterministic omniscient version of the algorithm (restricted to w). Since our outlier removal algorithm only throws away probability mass when necessary, this S is the largest possible restriction that is γ^2-outlier free. Define y and $\bar{x}$ as in lemma 6. By lemma 6, we have that $\bar{x}$ is a good approximation to y. This ensures that with good probability, we identify S as Γ^2-outlier free, and so (i) is proved. It remains to show that, if our algorithm for some reason chooses a substantially larger set T, then $(1+\delta)T$ has no $(1+\delta)^8\gamma^2$-outliers.

Define $T_\alpha = \{x : (w^T x)^2 \leq \alpha\}$. Suppose $\exists\alpha$ such that T_α has no Γ^2-outliers. Then $T_{(1+\delta)\alpha}$ has no $(1+\delta)^2\Gamma^2$-outliers. This follows from the fact that $\max\{(w^T x)^2 : x \in T_{(1+\delta)\alpha}\} \leq (1+\delta)^2 \max\{(w^T x)^2 : x \in T_\alpha\}$ and $\mathbf{E}[(w^T x)^2 : x \in T_\alpha] \Pr[x \in T_\alpha]$ is a monotonically increasing function of α. This analysis holds whether we are considering the true underlying distribution or just the samples.

Suppose we estimate that some set $T = T_t$ has no Γ^2-outliers (in which case the algorithm might return T as an answer). Then our sample also leads us to calculate that T_α has no $(1+\delta)^2\Gamma^2$-outliers for $\alpha \in [t, (1+\delta)t]$. For every t, we will show that for some nearby (within a factor of $(1+\delta)$) value of α, we correctly estimate the sample variance on the restriction of μ to T_α. Since the range of possible values for t is at most $2^{2(b+\log\frac{n}{\epsilon})}$, we can take every value of $t = (1+\delta)^k$ for some integral k and union bound over the at most $\log_{1+\delta} 2^{2(b+\log\frac{n}{\epsilon})} = O(\frac{b+\log\frac{n}{\epsilon}}{\delta})$ possible values for k.

We now show that if we estimate T_α to have no $(1+\delta)^2\Gamma^2$-outliers, then with good probability T_α actually has no

$(1+\delta)^6\Gamma^2$-outliers with respect to the true distribution, and by our reasoning above, since there is an α within $(1+\delta)$ of t, $T_{(1+\delta)t}$ is $(1+\delta)^8\Gamma^2$ outlier free.

We do this by showing that if T_α has a $(1+\delta)^6\Gamma^2$-outlier, then our sample shows T_α to have at least a $(1+\delta)^2\Gamma^2$-outlier with good probability. Let X_i be the random variable representing the squared distance of x_i along the direction w, $X_i = (w^T x_i)^2$. Without loss of generality, assume $\alpha = 1$. Define y and $\bar{x}$ as in lemma 6 (but with T_α in place of S). Then by assumption on T_α, $y \le \frac{1}{(1+\delta)^6\Gamma^2}$. The condition that our samples show T_α to have at least a $(1+\delta)^2\Gamma^2$-outlier is $\frac{1}{m}\sum_{x_i \in T_\alpha} X_i \le \frac{1}{(1+\delta)^2\Gamma^2}$. We apply the Chernoff bound,

$$\Pr[\sum X_i \ge (1+\Delta)m\mathbf{E}[X_i]] \le e^{-\Delta^2 m\mathbf{E}[X_i]/3}$$

where we have stated the Chernoff bound for the case that $\Delta < 1$. Let $\Delta = \frac{1}{\mathbf{E}[X_i](1+\delta)^2\Gamma^2} - 1$ (this yields the correct event in our probability calculation). If $\Delta < 1$, then $\Delta^2\mathbf{E}[X_i] \ge \frac{\delta^2}{\Gamma^2}$, and the probability we do not correctly identify the furthest outlier is at most $e^{-\Delta^2 m\mathbf{E}[X_i]/3} = O(1)$ for $m = O(\frac{\Gamma^2}{\delta^2})$. If $\Delta \ge 1$, then $\Delta\mathbf{E}[X_i] \ge \frac{\delta}{\Gamma^2}$, and the applicable alternate form of the Chernoff bound yields that $\Pr[\sum X_i \le m/\gamma^2]$ is at most $e^{-\Delta m\mathbf{E}[X_i]/3} = O(1)$ for the same setting of m.

Since there are only $O(\frac{b+\log\frac{n}{\epsilon}}{\delta})$ different values of α to consider, $m = O(\frac{\Gamma^2}{\delta^2}\log\frac{b+\log\frac{n}{\epsilon}}{\delta})$ allows us to union bound over all the possible values of α. This shows that with constant probability, if we estimate T to have no Γ^2-outliers (in which case our algorithm might return T), then $(1+\delta)T$ has no $(1+\delta)^8\Gamma^2$-outliers. This implies (ii). ■

We extend the analysis of lemmas 6 and 7 from a fixed direction to all directions and argue the correctness of the entire algorithm by proving theorem 3.

Proof: Let S be the ellipsoid found by the deterministic algorithm (i.e. the outlier free subset of points lies in this ellipsoid). Rather than considering the original space, consider the transformed space where S (not $E(M_S)$) is the unit sphere. Consider the many directions w given by a δ'-grid over the sphere, $\delta' = \frac{\delta}{n}$. We form this grid by choosing every w such that the coordinates of w lie in $\{0, \frac{\delta}{n}, \frac{2\delta}{n}, \ldots, 1\}$. By our choice of m, we can apply lemma 7 part (i) to each of these $(\frac{n}{\delta})^n$ directions simultaneously. We then have that for every w in the δ'-grid, $\max\{(w^T x)^2 : x \in T\} \ge \max\{(w^T x)^2 : x \in S\}$ (i.e., in this direction T contains S). We now show that for an arbitrary direction w, $(1+\delta)^2 T$ contains S.

Consider an arbitrary unit vector w. For every axis direction i, there are some vectors w_i^1 and w_i^2 in the δ'-grid that are above and below w, but within distance δ/n. Since $T = \Gamma^2 E(\bar{M}_T)$ is convex, the minimum distance of T from the origin between w_i^1 and w_i^2 is given by $1 - \frac{\delta}{n}$. Bounding over the maximum decrease in every axis direction gives that T is at least distance $1-\delta$ from the origin in direction w. Since S is within 1 of the origin everywhere, we have that $(1+\delta)T$ contains S, and therefore $(1+\delta)^2 T$ also contains S. This concludes the proof of (i).

Now consider (ii). For every w in our δ'-grid, we have that $(1+\delta)T$ is $(1+\delta)^8\Gamma^2$-outlier free along w by lemma 7 part (ii). Consider an arbitrary unit vector w not in the δ'-grid. Denote w's n nearest neighbors within the δ'-grid by $\{w_i\}$. Let w_i' be the vector in direction w_i with length given by $\mathbf{E}[(w_i^T x)^2 : x \in (1+\delta)T]\Pr[x \in (1+\delta)T]$. Then w' defined similarly is bounded away from the origin by the hyperplane formed by the $\{w_i'\}$ — this follows from the convexity of $E(M_{(1+\delta)T})$. Combining this and the spacing of the δ'-grid as in the previous paragraph, we find that the maximum drop in moving to w from w_i is at most $(1-\delta)$, i.e., $\mathbf{E}[(w^T x)^2 : x \in (1+\delta)T] \geq (1-\delta)\min_i \mathbf{E}[(w_i^T x)^2 : x \in (1+\delta)T]$.

Since $(1+\delta)T$ is $(1+\delta)^8\Gamma^2$-outlier free along w, $(1+\delta)T \subset H_w$, where H_w is the halfspace (really a slab) corresponding to w, $H_w = \{x : (w^T x)^2 \leq (1+\delta)^8\Gamma^2\mathbf{E}[(w^T x)^2 : x \in (1+\delta)T]\Pr[x \in (1+\delta)T]\}$. Let $H = \bigcap_w H_w$, the intersection of all the half-spaces. Then by the same argument using the convexity of H and the spacing of the δ'-grid, for an arbitrary w, $\max\{(w^T x)^2 : x \in H\} \leq (1+\delta)\min_i \max\{(w_i^T x)^2 : x \in H\}$. Our simultaneous lower bound on the expectation for an arbitrary w and upper bound on the maximum for that same w yield that H is $(1+\delta)^{10}\Gamma^2$-outlier free.

Using the same (δ'-grid) reasoning, we find that $(1+\delta)^2 T$ contains H, and therefore $(1+\delta)^2 T$ is $(1+\delta)^{12}\Gamma^2$-outlier free. ■

Before stating the corollary for the running time, we mention that we still have not shown that step 2 of either algorithm can be carried out with high probability. This will be done in our final lemma of this section, lemma 8. Assuming lemma 8, we have the following bound on the running time.

Corollary 1 (Running Time) *The algorithm runs in time $\tilde{O}(\frac{b^2 n^5}{\epsilon^2\delta^4})$.*

Proof: We have from section 3 that $\beta = \gamma^2$ is at most $\tilde{O}(bn/\epsilon)$, and so we never need more than $m = \tilde{O}(\frac{bn^2}{\epsilon\delta^2})$ samples. Plugging in this value for m to our bounds in the discussion of running time at the beginning of section 4 yields that our entire algorithm runs in time $\tilde{O}(\frac{b^2 n^5}{\epsilon^2\delta^4})$, which is the bound we referred to in the introduction. In this time we achieve a $1+O(\delta)$ approximation to the optimal value of β. ■

We now show that we can carry out step 2 of either algorithm with high probability. Additionally, we solve the following problem. Suppose that we are not given the parameter γ^2, but rather only ϵ, and asked to find the appropriate γ^2. Lemma 8 will show that we can at any point determine within a factor of $(1+\delta)$ how much of the probability mass is within a fixed ellipsoid. Since $\gamma^2 \in [1, \tilde{O}(\frac{bn}{\epsilon})]$, there are at most $\log_{1+\delta}\tilde{O}(\frac{bn}{\epsilon}) = \frac{\log\frac{bn}{\epsilon}}{\delta}$ values of γ^2 to consider (with a loss of at most a factor of $(1+\delta)$ in the value we find for γ^2). Therefore we can simply try them all, estimating for each one whether this γ^2 requires us to throw away more than a $(1+\delta)\epsilon$ fraction of the distribution.

Thus, if the parameters (γ^2, ϵ) are achievable for the deterministic algorithm, and we are only given ϵ, we can find a subset of space space T' satifying parameters $((1+O(\delta))\gamma^2, (1+O(\delta))\epsilon)$. Our asymptotic running time increases to $\tilde{O}(\frac{b^2 n^5}{\epsilon^2\delta^4} + \frac{1}{\epsilon\delta^3})$, which is an increase of no more than a constant.

Lemma 8 (Probability Mass Location)
(i) Fix a direction w. Let our number of samples be $m = O(\frac{1}{\epsilon\delta^2})$, and consider the squared sample distances in direction w given by $\{(w^T x_i)^2\}$. Let $\bar{y}$ be the greatest y such that a $(1+\delta)\epsilon$ fraction of our samples are above y. Then with constant probability, at least an ϵ fraction of the distribution has squared distance along w of at least $\bar{y}$, and at most a $(1+\delta)^2\epsilon$ fraction has squared distance along w greater than $\bar{y}$.

(ii) Let E be an ellipsoid. Let our number of samples be $m = O(\frac{1}{\epsilon\delta^2})$. Then if we estimate a $(1+\delta)\epsilon$ fraction of our samples to be outside of E, with constant probability at most a $(1+\delta)^2\epsilon$ fraction is outside of E, and at least an ϵ fraction is outside of or on E.

Proof: First we show (i). Let $y_{high} = \{y : \Pr[(w^T x_i)^2 \geq y] = \epsilon\}$. Let Y_i be a random variable, $Y_i = 1$ iff $(w^T x_i)^2 > y_{high}$. The event that $\bar{y} > y_{high}$ is the same as $\sum Y_i \geq m(1+\delta)\epsilon$. We can upper bound the probability of this event using the Chernoff bound

$$\Pr[\sum Y_i \geq m(1+\delta)\mathbf{E}[Y_i]] \leq e^{-\delta^2 m \mathbf{E}[Y_i]/3}$$

which is constant for $m = O(\frac{1}{\epsilon\delta^2})$. A similar calculation with $y_{low} = \{y : \Pr[(w^T x_i)^2 \geq y] = (1+\delta)^2\epsilon\}$ shows that the probability of $\bar{y} < y_{low}$ is at most a constant for the same value of m.

The proof of (ii) is identical. By centering E, we can just consider distance of our samples from the origin rather than the distance along a fixed w. ■

One consequence of the theorems in this section is that a sample of size $\tilde{O}(\frac{n^2 b}{\epsilon})$ is enough to estimate the inertial ellipsoid of any distribution (after removing at most an ϵ fraction) and thus bring it into nearly isotropic position.

5 A Matching Lower Bound

We show that for any $\epsilon < 1/2$ there exists a distribution μ such that, for any S satisfying $\mu(S) \geq 1 - \epsilon$, there exists w such that

$$\max\{(w^T x)^2 : x \in S\} \geq \bar{\beta}\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S] \geq \frac{\bar{\beta}}{2}\mathbf{E}[(w^T x)^2 : x \in S]$$

where $\bar{\beta} = \Omega(\frac{n}{\epsilon}(b - \log\frac{1}{\epsilon}))$. Based on the comparison between our upper and lower bounds on β in the case that we can't throw out more than half the distribution

$$O\left(\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})\right) \quad vs. \quad \Omega\left(\frac{n}{\epsilon}(b - \log\frac{1}{\epsilon})\right)$$

we describe our result as asymptotically optimal.

We motivate the construction of the worst case distribution by constructing three simpler distributions, each of which proves a weaker lower bound. The strong lower bound will follow from examining a distribution that is a composite of the three distributions showing the weaker lower bounds.

To prove the first weak lower bound, let μ be the uniform distribution on the one-dimensional points $\{2^0, 2^1, ...2^b\}$. An illustration of this μ is given in figure 2, part A. We claim that for any $\epsilon < \frac{1}{4}$, the best achievable (i.e. smallest) β satisfies $\beta = \Omega(b)$. The proof is simple: suppose the largest data point we keep is 2^k. Then (ignoring the factor w since we are in one dimension), $\max\{x^2 : x \in S\} = 2^k$, while $\mathbf{E}[x^2 : x \in S] \leq \frac{2^0 + ...2^k}{(b+1)(1-\epsilon)} = O(\frac{2^k}{b})$. Since $\beta \geq \frac{\max\{\cdot\}}{\mathbf{E}[\cdot]}$, we find $\beta = \Omega(b)$.

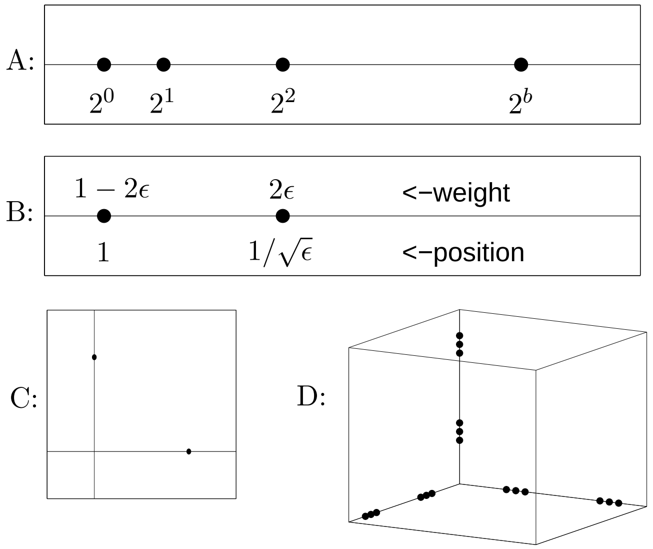

Figure 2: Lower Bound Constructions

To prove the next weak lower bound, we construct a distribution as in figure 2, part B. Let μ be the probability distribution on one-dimensional points given by $\mu(1) = 1-2\epsilon, \mu(\frac{1}{\sqrt{\epsilon}}) = 2\epsilon$. Then for $\epsilon < \frac{1}{4}$, neither point can be thrown away. Thus $\max\{x^2 : x \in S\} = \frac{1}{\epsilon}$, while $\mathbf{E}[x^2 : x \in S] = 3 - 2\epsilon$, yielding $\beta = \Omega(\frac{1}{\epsilon})$.

For the third weak lower bound, we let μ be a distribution on n-dimensional space. In particular, let μ be the uniform distribution on n points, one on each coordinate axis, each one at unit distance from the origin, as illustrated in figure 2, part C. For $\epsilon < \frac{1}{2}$, we do not throw away any points on at least $n/2$ of the axes. Then for w a unit vector along one of the axes where the point is not thrown away, we have $\max\{(w^Tx)^2 : x \in S\} = 1$, $\mathbf{E}[(w^Tx)^2 : x \in S] = \frac{4}{n}$, and thus $\beta = \Omega(n)$.

The composite construction that we use to prove our strong lower bound in illustrated in figure 2, part D. We obtain the composite distribution by taking the distribution of part A, and making two copies that are weighted and translated as the two points are that compose the distribution of part B. We then place a copy of this new one-dimensional distribution along each axis, as in the distribution of part C. We now restate this construction formally and proceed to analyze it.

Fix n, ϵ and $b' = \frac{b}{2} - \frac{1}{4}\log\frac{1}{\epsilon}$. Let μ be a copy of the following distribution along each axis. Let there be $2b'$ points at distances

$$2^0, 2^1, \ldots, 2^{b'-1}, \frac{2^{b'}}{\sqrt{\epsilon}}, \frac{2^{b'+1}}{\sqrt{\epsilon}}, \ldots \frac{2^{2b'-1}}{\sqrt{\epsilon}}$$

and consider the distribution that places a $(1-2\epsilon)$ fraction of the probability mass uniformly on the first b' points and a 2ϵ fraction uniformly on the remaining b' points. This distribution satisifes that the maximum bit length along an axis is $\log\frac{2^{2b'}}{\sqrt{\epsilon}} = b$.

There are many ways of choosing a subset S of this distribution, but we can quickly restrict the set of possible choices. First we show that it never helps to treat the different axes asymmetrically for a distribution that is concentrated on the axes. Suppose that this statement is not true. We begin by noting that for a distribution concentrated on the axes and fixed S, the vector w that maximizes

$$\frac{\max\{(w^T x)^2 : x \in S\}}{\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]}$$

always occurs on an axis. Let μ_1 be a distribution on which it is possible to throw out an ϵ fraction of the distribution and achieve parameter $\bar{\beta}$. Further suppose that this ϵ is the minimum ϵ such that this $\bar{\beta}$ is achievable, and the only S achieving $\bar{\beta}$ is asymmetric. Let axis i be an axis that this maximum outlier occurs on, and suppose that along axis i we throw out an ϵ_i fraction of the total distribution. If $\epsilon_i \leq \epsilon/n$, then let S' be the subset of μ_1 where we throw out the same points along every axis that we threw out along axis i in S. Then we have $\epsilon' = n\epsilon_i \leq \epsilon$, and yet S' achieves $\bar{\beta}$ along each axis, contradicting the assumption that there was no symmetric subset we could throw out achieving the same $(\epsilon, \bar{\beta})$. If $\epsilon_i > \epsilon/n$, then there is some other axis j such that along axis j we throw out an $\epsilon_j < \epsilon_i$ fraction of the probability distribution, but achieving $\bar{\beta}_j \leq \bar{\beta}$ along that axis (i.e. $\max\{x_j : x \in S\} \leq \bar{\beta}_j \mathbf{E}[x_j^2 : x \in S] \Pr[x \in S]$). Constructing S'' by taking S and replacing our choice of points to throw out along axis i with the points thrown out along axis j then yields a contradiction because $\epsilon'' < \epsilon$. Thus we can restrict our attention to S symmetric.

For any direction w along an axis, the projection onto w of any point on the other $n-1$ axes is 0, so we obtain

$$\mathbf{E}[(w^T x)^2] = \frac{1}{n}\mathbf{E}[x^2, \mu \text{ one-dimensional}]$$

We ignore the factor of n for the rest of the proof and restrict our attention to a single coordinate axis. Suppose the furthest point kept by S achieving parameters $(\epsilon, \bar{\beta})$ is the point with exponent k. By our choice of distribution, we cannot have thrown out more than half the points with a $\frac{1}{\sqrt{\epsilon}}$ factor, and so we have $\max\{x^2 : x \in S\} = \frac{2^{2k}}{\epsilon}$, $k > b'$. Calculating the expectation

$$\begin{aligned}\mathbf{E}[x^2 : x \in S] \Pr[x \in S] &\leq \frac{1-2\epsilon}{2b'}(2^0 + 2^2 + \ldots + 2^{2b'-2}) + \frac{2\epsilon}{2b'}\frac{1}{\epsilon}(2^{2b'} + 2^{2b'+2} + \ldots + 2^{2k}) \\ &\leq \frac{2^{2b'-1}}{2b'} + \frac{2^{2k+1}}{b'} \leq \frac{2^{2k+2}}{b'}\end{aligned}$$

yields that $\bar{\beta} \geq \frac{b'}{4\epsilon}$ for the one-dimensional case. Thus our lower bound in the n-dimensional case is

$$\bar{\beta} \geq \frac{n}{8\epsilon}(b - \log\frac{1}{\epsilon})$$

6 An Approximation Algorithm

We showed earlier in the paper that for any distribution μ, and any ϵ we can achieve $\beta = O(\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon}))$. A question that naturally arises is how well we can do on a particular distribution compared to the best possible on that particular distribution. Formally, given μ and ϵ, we seek S minimizing β subject to the constraints that

(i) $\mu(S) \geq 1 - \epsilon$

(ii) $\forall w, \max\{(w^T x)^2 : x \in S\} \leq \beta \mathbf{E}[(w^T x)^2 : x \in S]$

This is really a bicriteria approximation problem with parameters (β, ϵ). Note that in this case, we are looking for the *normalized* probability distribution to be β-outlier free. We exhibit a $(\frac{1}{1-\epsilon}, 1)$-approximation algorithm for this task in the case that we are given the distribution explicitly. If we can only sample from the distribution μ, our algorithm yields a $(\frac{1}{1-\epsilon} + \delta, 1 + \delta)$-approximation for any constant $\delta > 0$ with high probability.

Lemma 9 (Preservation of Outliers) *Let μ be a distribution. Any β-outlier for μ is at least a $\beta(1-\epsilon)$-outlier with respect to any subset S satisfying $\mu(S) \geq 1 - \epsilon$.*

Proof: Let x be a β-outlier in the original distribution. Then for some w, $(w^T x)^2 > \beta \mathbf{E}[(w^T x)^2]$ For any S, we have $\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S] \leq \mathbf{E}[(w^T x)^2]$ and so x satisfies $(w^T x)^2 > \beta(1-\epsilon)\mathbf{E}[(w^T x)^2 : x \in S]$ ■

The approximation algorithm is simply either algorithm described in section 4, with error parameter δ in the case that we are sampling from μ. We could determine the optimal β for a fixed ϵ through a binary search. Suppose the value β_{OPT} is achievable by the restriction of μ to some S satisfying $\mu(S) \geq 1 - \epsilon$. Anytime our algorithm sees a point that is a β'-outlier with respect to the unnormalized distribution, $\beta' > \frac{\beta_{OPT}}{1-\epsilon}$, we know that this cannot be a $(\leq \beta_{OPT})$-outlier under any restriction of μ by lemma 9. So this point will have to be thrown out by the optimal solution. Thus running our algorithm with $\beta = \frac{\beta_{OPT}}{1-\epsilon}$ forces us to throw away no points that the optimal solution does not also throw away. This yields that we achieve a $\frac{1}{1-\epsilon}$-approximation in the case of an explicitly provided distribution. Correctness and running time are clear from the preceding discussions.

There is a more direct method that in fact finds an approximation to β for *every* ϵ in one pass. The algorithms of section 2 can be used to define an *outlier ordering* of a point set, namely, the first point that is an outlier, the second point, etc. Now to approximate the best possible β for a particular value of ϵ we simply remove the initial ϵ fraction of the points in the outlier ordering.

7 Standard Deviations from the Mean

We prove a variant of our theorem that shows we can find a large subset of the original probability distribution where no point is too many standard deviations away from the mean.

Corollary 2 (Standard Deviations from the Mean) *Let μ be a probability distribution on I_b^n. Let S be a subset of space. Denote by $\mu(S)$ the probability that x chosen according to μ is in S. Let $\bar{x} = \mathbf{E}[x : x \in S]$ and $\sigma_w^2 = \mathbf{E}[(w^T(x - \bar{x}))^2 : x \in S]$. Then for every $\epsilon > 0$, there exists S and*

$$\beta = O\left(\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})\right)$$

such that
(i) $\mu(S) \geq 1 - \epsilon$
(ii) $\max\{w^T(x - \bar{x}) : x \in S\} \leq \sqrt{\beta}\sigma_w$ *for all* $w \in \mathcal{R}^n$

Proof: The proof of the corollary is much like the proof of theorem 1, but with two additional steps. In the first step, we show that translating μ so that the origin coincides with the mean does not increase the volume of the primal inertial ellipsoid. In the second step, we show that running the algorithm with our value of b doubled suffices to preserve our dual volume bound. These steps taken together with our previous analysis imply corollary 2.

In order to analyze the volume of the primal inertial ellipsoid, consider the radius r of the primal inertial ellipsoid in a fixed direction w. Implicitly taking all expectations with respect to the restricted probability distribution $\mu_{|S}$, and letting x_w be the value of the projection of x onto w, we have $r^2 = \mathbf{E}[x_w^2]$. Suppose we choose to translate our origin to a value z along w. We then have $r^2 = \mathbf{E}[(x_w - z)^2]$. Single variable calculus shows that the value minimizing r is $z = \mathbf{E}[x_w]$, which is just the mean. Thus translating our origin to $\bar{x}$ minimizes the radius of the primal inertial ellipsoid in every direction simultaneously. Since our analysis in the proof of theorem 1 relied upon showing shrinkage of the primal ellipsoid (growth of the dual ellipsoid), the algorithm that at each step translates the origin to the new mean concludes at least as quickly, throwing out no more probability mass overall.

It may be that at some point our mean is not an element of I_b^n. We show that even in this case our dual volume bound still holds. Suppose that at some step the criterion for step 2 of either algorithm is met for axis i, but with the value of b doubled. Then we throw out all x with $x_i \geq 2^{-2b}$ (where our mean has $x_i = 0$). Not all the remaining elements necessarily have $x_i = 0$, but they do all necessarily have the same x_i value, because two elements of I_b^n cannot have distinct values for a particular coordinate that differ by less than 2^{-2b}. So the dimension does collapse. Our dual volume bound only required the guarantee that in every non-collapsed dimension at least an $\frac{\epsilon}{3n}$ fraction of the points have coordinate value at least 2^{-2b}. This concludes the proof of corollary 2. ■

We now show that the $\frac{1}{1-\epsilon}$-approximation algorithm of section 6 naturally extends to a $\left(\frac{1-\epsilon}{1-2\epsilon}\right)^2$-approximation algorithm in the setting where we measure outlierness with respect to the mean, rather than a fixed origin. To establish this, it suffices to prove the following analogue of lemma 9.

Lemma 10 (Outlier Preservation Variant) *Let μ be a distribution. As in Corollary 2, measure outlierness by squared distance from the mean rather than from a fixed origin. Suppose x_0 is a β-outlier for μ, and no other point is a β'-outlier for $\beta' > \beta$. Then x_0 is at least a $\beta\left(\frac{1-2\epsilon}{1-\epsilon}\right)^2$-outlier with respect to any subset S satisfying $\mu(S) \geq 1 - \epsilon$.*

Proof: As in the proof of lemma 9, consider w such that $(w^T x_0)^2 > \beta \mathbf{E}[(w^T x)^2]$, and let $\beta = \gamma^2$. The difference between this bound and the bound of lemma 9 will result from the mean possibly moving closer to x_0 after removing other points $\{x'\}$. Without loss of generality, let the mean of μ be the origin, and let $\mathbf{E}[(w^T x)^2] = 1$.

Suppose we remove some set of points $\{x'\}$, resulting in a decrease in $\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]$. The mean may also move towards x_0. If the points $\{x'\}$ are different, then the mean would shift by the same amount, while $\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]$ would decrease by less, if we could remove a point of the same probability mass located at the average of the $\{x'\}$. Thus it suffices to consider removing a single point x' for the purposes of our bound.

If we remove a single point x' of probability weight ϵ, we find that the distance of x_0 from the mean goes to $\gamma - \frac{\epsilon x'}{1-\epsilon}$, while $\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]$ becomes $\frac{1-\epsilon x'^2}{1-\epsilon}$. By the assumption that no point x' is initially further away from the mean than x_0 (along w) we can upper bound x' by γ. Then the ratio of $(\gamma - \frac{\epsilon x'}{1-\epsilon})^2$ to $\frac{1-\epsilon x'^2}{1-\epsilon}$ is at least $\gamma^2 \left(\frac{1-2\epsilon}{1-\epsilon}\right)^2$. ■

References

[1] A. Blum, A. Frieze, R. Kannan and S. Vempala, "A Polynomial-Time Algorithm for Learning Noisy Linear Threshold Functions," In *Algorithmica*, 22(1), 1999, pp35-52.

[2] A. Blum, G. Konjevod, R. Ravi, and S. Vempala, "Semi-Definite Relaxations for Minimum Bandwidth and other Vertex-Ordering Problems," In *Proc. of the 30th ACM Symposium on the Theory of Computing*, Dallas, 1998. To appear in Theoretical Computer Science, special issue in honour of Manuel Blum.

[3] L. Lovász, R. Kannan and M. Simonovits, "Random walks and an $O^*(n^5)$ volume algorithm for convex bodies," In *Random Structures and Algorithms* 11, pp1-50.

[4] L. Lovász, R. Kannan and M. Simonovits, "Isoperimetric problems for convex bodies and a localization lemma," In *Discrete Computational Geometry* 13, 1995, pp541-559.

[5] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, 1985, pp179.

8 Implementation

Let X be an $n \times m$ matrix whose columns are the points of our distribution. Let m,beta,epsilon be the values for m, β, ϵ, and let the boolean variable done indicate whether we are finished removing outliers. In the case that X is full dimensional throughout the algorithm (a common case), a complete implementation is given by the following matlab code:

```
done = 0
while(~done)
  done = 1
  M = X*X'/m
  Y = M^(-.5)*X %% Y is isotropic version of X
  for i = 1:m, %% remove current outliers
    if Y(:,i)'*Y(:,i) > beta,  X(:,i)=0, done = 0, end
  end
end
```

Handling dimension dropping adds a few more lines of code.

A java applet illustrating the outlier removal algorithm is available at
http://theory.lcs.mit.edu/~jdunagan/