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 [BFKV 99], 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 is asymptotically the best possible, as shown by a matching lower bound. 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[BG 97, DG 93]. 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 [BFKV 99]. 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.

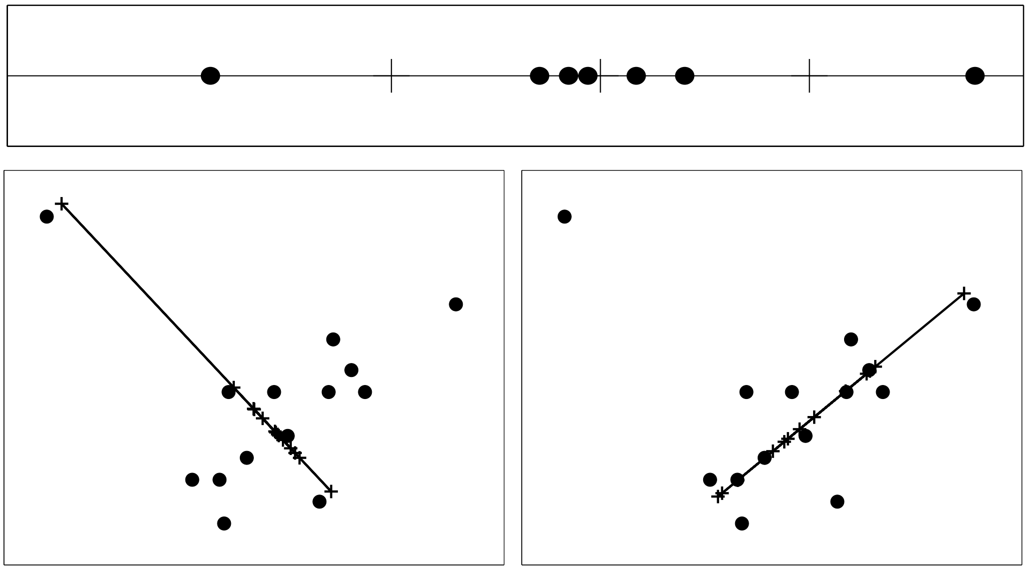

Figure 1: Defining Outliers

This definition of an outlier in $\mathcal{R}^n$ has a long history in statistics and machine learning. An equivalent definition using terminology from the field of machine learning is "a point is a γ^2-outlier if it has Mahalanobis distance greater than γ." A statistician might say "after normalizing by the covariance of the data, the point is more than γ away from the origin." The constructive procedure for identifying outliers in section 2 shows the equivalence of our definition to these two other definitions.

The first problem we address is the following: does there exist a small subset of (a point set) 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 P, the remaining set has no β-outliers (*with respect to the remaining set*)?

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 has been used to speed up random walks in [LKS 95]. 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 $\mathcal{Z}_b^n$ denote the set of n-dimensional b-bit integers, $\{1, \ldots, 2^b\}^n$. In place of the point set P in the discussion above we have any probability distribution μ on $\mathcal{Z}_b^n$. For a probability distribution μ, let $\mu(S)$ denote the probability of a subset of space S.

Theorem 1 (Outlier Removal over Integer Support) [1] *Let μ be a probability distribution on $\mathcal{Z}_b^n$. 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 theorem 1 (section 4) is constructive. Before proving theorem 1, we will prove a similar theorem about distributions with arbitrary support (theorem 2, section 3). Although the hypothesis on the support of the distribution in theorem 2 is much weaker, we need an additional assumption. The proofs of theorems 1 and 2 make use of the same principal idea.

In section 2, we describe (two variants of) an algorithm for outlier removal. The theorems can be proven using either variant. Although the theorems are not obvious, the algorithm is extremely simple. To convince the reader of this, we include a matlab implementation of the algorithm in section 11.

For a point set with m points ($m > n$) the algorithm runs in $O(m^2 n)$ time. In section 5 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 $\tilde{O}(\frac{b^2 n^5}{\epsilon \delta^4})$ for accuracy $(1 + \delta)$.

One variant of our algorithm is identical to the algorithm of [BFKV 99], the immediate inspiration for our work. The bound on β in theorems 1 and 2 improves on the previous best bound of $O(\frac{n^7 b}{\epsilon})$ given in [BFKV 99]. 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 6 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 the complexity of achieving the best possible β for any particular distribution. Given a distribution

[1] An early version of this work[DV 01] claimed a slightly different version of theorem 1 with an insufficiently strong hypothesis.

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. We show this question to be NP-hard even in the one-dimensional case by a reduction to subset-sum. In section 7, 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 8. Let μ be a probability distribution on Z_b^n. 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 $(\frac{1-\epsilon}{1-3\epsilon})$-approximation algorithm for the corresponding optimization problem.

In section 9, we prove a theorem describing a connection between outlier removal and robust statistics. In section 10, we conclude by proving some technical properties of matrices that are used elsewhere in the paper.

2 Algorithms for Outlier Removal

The first question we address is that of detecting outliers. Since our definition of a γ^2-outlier involves all directions, it might not be obvious that this can be done in finite time.

In order to detect outliers, we use a linear transformation. Let $M = \mathbf{E}[xx^T]$ where x is a sample drawn according to the probability distribution μ. If 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* position [LKS 95], and we will refer to the transformation as *rounding*. Such transformations have previously been used in the design of algorithms to make geometric random walks more efficient [LKS 97]. If M does not have full rank, it is still positive semi-definite, and we instead round μ in the span of M. For those familiar with the definitions of Mahalanobis distance or normalizing by the covariance of the data set, this transformation shows the equivalence between our definition of an outlier and these two other definitions.

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, remove them; stop when there are no outliers. In the description below, μ is the given distribution and $\beta = \gamma^2$, where the exact value of β is specified in the proofs of theorems 1 and 2.

Algorithm 1 (Restriction to Ellipsoids):

1. Round μ. If there exists x such that $|x| > \gamma$, let $S = \{x : |x| \leq \gamma\}$. Retain only points in S.
2. Repeat until the condition is not met.

Algorithm 1 is identical to the outlier removal algorithm of [BFKV 99]. 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. Round 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. Repeat until the condition is not met.

3 Outlier Removal over Arbitrary Support

We will prove the following theorem about outlier removal over a distribution with arbitrary support before proceeding to prove theorem 1. We refer to conditions (I, II) in the hypothesis of theorem 2 as the *full-dimensional* condition. In theorem 1 we will remove this condition, replacing it only by a condition on the support of the distribution.

Theorem 2 (Outlier Removal over Arbitrary Support) *Let μ be a probability distribution on $\mathcal{R}^n$ satisfying*

$$\begin{array}{llcl} \text{(I) } \forall \text{ unit vector } \hat{w}, & \int (\hat{w}^T x)^2 d\mu & \leq & R^2 \\ \text{(II) } \forall \text{ unit vector } \hat{w}, \quad \forall S : \mu(S) \geq 1 - \bar{\epsilon}, & \int_S (\hat{w}^T x)^2 d\mu & \geq & r^2 \end{array}$$

Then for every ϵ such that $0 < \epsilon \leq \bar{\epsilon}$, there exists S and

$$\beta = O\left(\frac{n}{\epsilon} \ln \frac{R}{r}\right)$$

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 $w \in \mathcal{R}^n$

To prove the theorem, we analyze the set S returned by either algorithm. This set S is clearly β-outlier free. 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.

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

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

In other words, M_S is the M obtained after restricting μ to S (zeroing out points outside of S, not renormalizing the distribution). We denote this restricted probability distribution directly by μ_S. Throughout this chapter, μ_S will denote a restriction of μ to the subset of space S, never a new and unrelated distribution. The useful property attained by rounding with respect to μ_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 2 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. Let $x \in \mu_S$ denote $x \in S : \mu(x) > 0$, and let $span(\mu_S)$ denote the span of $\{x \in \mu_S\}$.

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 singular vectors of A. The axis lengths of $W(M)$ and $E(M)$ are given by the singular values of A^{-1} and A, and so they are reciprocals. 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 of μ_S. 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 ellipsoid 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 $e^{\gamma^2 p/2}$. This concludes the proof of lemma 1. ■

Note that if we desire to apply the lemma to analyze the result of a later iteration of Algorithm 2, where μ_T goes to $\mu_{T \cap S}$, we simply replace the starting identity by

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

The analysis and conclusion remain the same.

Lemma 2 (Dual Volume Growth) *Let μ be a distribution satisfying*

$$\text{(I) } \forall \text{ unit vector } \hat{w}, \qquad \int (\hat{w}^T x)^2 d\mu \leq R^2$$
$$\text{(II) } \forall \text{ unit vector } \hat{w}, \quad \forall S : \mu(S) \geq 1 - \bar{\epsilon}, \quad \int_S (\hat{w}^T x)^2 d\mu \geq r^2$$

For any S^*, let μ_{S^*} be the restriction of μ to S^*. Assume $\mu(S^*) \geq 1 - \bar{\epsilon}$. Then

$$Vol(W(M)) \geq \frac{f(n)}{R^n}$$
$$Vol(W(M_{S^*})) \leq \frac{f(n)}{r^n}$$

Proof: First we lower bound the initial dual volume, $Vol(W(M))$. Consider any vector v of length at most $1/R$. We have

$$v^T M v = \mathbf{E}[(v^T x)^2] = \int (v^T x)^2 d\mu \leq (v^2 R^2) \leq 1$$

so v belongs to the dual ellipsoid. Thus the dual ellipsoid initially has volume at least $f(n)/R^n$.

Next we upper bound $Vol(W(M_{S^*}))$. Consider any vector v of length more than $1/r$. Then

$$v^T M_{S^*} v = \int_{S^*} (v^T x)^2 d\mu \geq (v^2 r^2) > 1$$

Thus v is not in $W(M_{S^*})$, and thus the ultimate volume of the dual ellipsoid is no more than the volume of the sphere of radius $1/r$, yielding the claimed upper bound. ■

In the proof of theorem 2 below, μ_{S^*} will be the final distribution resulting from application of either algorithm. Using lemmas 1 and 2, we prove that Algorithm 2 terminates with $S = S^*$ satisfying theorem 2.

Proof of Theorem 2: Let $\beta = 4\frac{n}{\epsilon}(\ln\frac{R}{r}+1)$. 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 μ_{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 μ_{S^*}. We thus achieve a β-outlier free subset with

$$\beta = \gamma^2 = 4\frac{n}{\epsilon}(\ln\frac{R}{r}+1)$$

It now remains to show that $\epsilon' \leq \epsilon$, i.e. that we do not throw out more of the probability mass than claimed. 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

$$e^{\frac{\gamma^2}{2}\sum p_i} \leq (\frac{R}{r})^n = e^{n\ln\frac{R}{r}}$$

$$\Rightarrow \frac{1}{2}\gamma^2\epsilon' = \frac{1}{2}(4\frac{n}{\epsilon}\ln\frac{R}{r})\epsilon' \leq n\ln\frac{R}{r}$$

$$\Rightarrow \epsilon' \leq \epsilon/2$$

The one remaining catch is showing that $\epsilon' \leq \bar{\epsilon}$, since we relied on this in applying lemma 2 above. By slight overloading of notation, we let ϵ' denote the cumulative probability mass that has been removed at any point during the algorithm. Suppose for the purpose of establishing a contradiction that in iteration j, $\epsilon' \leq \bar{\epsilon}$, but then in iteration $j+1$, $\epsilon' > \bar{\epsilon}$. Then on step j, we can apply lemma 2, and from the analysis above, $\epsilon' \leq \epsilon/2 \leq \bar{\epsilon}/2$. However, in any single iteration, the maximum probability mass the algorithm might throw out is $1/\gamma^2$, as can be seen from the proof of lemma 1:

$$a^2 \leq 1-\gamma^2 p \quad\Rightarrow\quad 0 \leq 1-\gamma^2 p \quad\Rightarrow\quad p \leq 1/\gamma^2$$

Thus in one step ϵ' increase by at most $\epsilon/[4n(\ln(R/r)+1)] \leq \frac{\bar{\epsilon}}{2}$, and so on step $j+1$, we still have $\epsilon' \leq \bar{\epsilon}$. This concludes the proof of theorem 2. ■

We now give an alternate proof of theorem 2 using the construction given by Algorithm 1. We begin by proving an analogue to lemma 1.

Lemma 3 (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 averaging over rotations of 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 3.

■

Finally, we prove that Algorithm 1 terminates with S satisfying theorem 2.

Proof of Theorem 2: As in the proof of theorem 2 using Algorithm 2, let $\beta = 4\frac{n}{\epsilon}(\ln \frac{R}{r}+1)$. Lemma 2 still holds. The rate of increase in the dual volume as we throw out probability mass (lemma 3) is the same as before (lemma 1). The only thing we need to address is what we called "the one remaining catch" in the proof using Algorithm 2. Our bound on the amount of probability mass that can be thrown out in a single step is no longer $1/\gamma^2$, but is now n/γ^2. However, $n/\gamma^2 = \epsilon/[4(\ln(R/r)+1)] \leq \frac{\epsilon}{2}$ just as before. This concludes the analysis of Algorithm 1. ■

The following connection shows that the success of either algorithm implies that they both succeed. If our criterion for a point x to be a β-outlier in a direction w were instead that

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

then Algorithms 1 and 2 both throw out the exact same points, and so must yield the same bound on β as a function of ϵ. To see this, note that any β-outlier under this definition remains a β-outlier as further points are removed, and so will have to be removed itself eventually. Also, no point is ever removed unless it currently is a β-outlier. Thus the two algorithms throw out exactly the same set of points in the end under this alternative definition of an outlier. In section 7, we develop this observation into an approximation algorithm for the problem of outlier removal using the standard definition of a β-outlier (not this alternative definition).

We pause to stress what we have gained by allowing some points of the distribution to be removed. If we force $\epsilon = 0$, then even under the hypothesis of theorem 2, β may be unbounded. Even a radially symmetric distribution satisfying the hypothesis with support in $\{B_R \setminus B_{r\sqrt{n}}\}$, where B_R denotes the ball of radius R, might have β as large as

$$\beta = \frac{R^2}{r^2}$$

By allowing $\epsilon > 0$, we have achieved

$$\beta = 4\frac{n}{\epsilon}(\ln \frac{R}{r} + 1)$$

4 Outlier Removal over Discrete Support

While theorem 2 might suffice for many applications, it is indeed possible that during outlier removal on an arbitrary set, the full-dimensional condition might be violated (indeed, the dimensionality of the remaining set might decrease). In this section we prove the following theorem, which shows that for distributions over integers, the full-dimensional condition is entirely unnecessary.

Theorem 1 (Outlier Removal over Discrete Support) *Let μ be a probability distribution on $\mathcal{Z}_b^n$. 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 this theorem presents two difficulties that were not present in the proof of theorem 2. First, μ might initially lie entirely on a lower-dimensional subspace, or μ might lie on a lower-dimensional subspace after the removal of a few points. Secondly, even if the distribution does not lie on a lower-dimensional subspace, we do not have the same lower bound on the smallest singular value of the distribution (singular value of the matrix M associated with μ). While we insisted in the hypothesis of theorem 2 that the smallest singular value must be at least $1/r$, which will be roughly equivalent to 2^{-b} in the discrete case, it may be that the smallest singular value is actually 2^{-nb}, as the following example makes clear.

Example 1 *Let $B = 2^b$, and let each row of the matrix below represent a point in space. Denote the first $n-1$ rows by $\{v_i\}_{i=1}^{n-1}$ and denote the last row by p.*

$$\begin{bmatrix} B & -1 & & & \\ & B & -1 & & \\ & & B & -1 & \\ & & & \cdots & \\ 1 & & & & \end{bmatrix}$$

This set of points is clearly full-dimensional, and in most directions the singular values are on the order of B. However, in the direction $w = [B^{-n}, B^{-n+1}, \ldots, B^{-1}, 1]$, we find that $(w^T v_i)^2 = 0$ while $(w^T p)^2 = B^{-n} = 2^{-nb}$. Since $w > 1$, the singular value is actually slightly less than 2^{-nb}.

Example 1 shows that even disregarding issues of the distribution not being full-dimensional, we *cannot use theorem 2* to treat the distribution with integer support unless we are willing to settle for $\beta = \tilde{O}(\frac{n^2 b}{\epsilon})$. In extending our techniques to prove theorem 1, we will show that although one singular value may be small, they are not all small simultaneously in an appropriate *amortized* sense.

The first thing we shall define is a potential function that generalizes the dual ellipsoid volume we used in the proof of theorem 2. This potential function will account for the distribution μ being concentrated on a lower dimensional subspace, or even the possibility that μ is simply quite close to a lower dimensional distribution. We begin by defining the α-*core* of a distribution to be that subset of the distribution which lies on a subspace spanned by every large subset of the distribution. It will help to define the indicator function of E to be

$$\chi^E = \begin{cases} 1 & \text{if } E \text{ is true} \\ 0 & \text{if } E \text{ is false} \end{cases}$$

where E is a logical statement. The α-core is then given by

Definition 1 (α-core) *Define the α-core of μ_S to be μ_T, where $T \subset S$ is chosen to be maximum such that*

$$\forall w \in span(\mu_T) \text{ such that } w \neq 0, \quad \sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \alpha$$

We now establish some characteristics of the α-core, including that the α-core is well-defined.

Lemma 4 (Characterization of α-core)

(i) For any μ_T,

$$\forall w \in span(\mu_T) \text{ such that } w \neq 0, \quad \sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \alpha \qquad (a)$$

if and only if

$$\forall w, \quad w^T x \neq 0 \text{ for some } x \in \mu_T \ \Rightarrow \sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \alpha \qquad (b)$$

(ii) $Q \subset S \implies \alpha\text{-core}(\mu_Q) \subset \alpha\text{-core}(\mu_S)$

(iii) $Q \subset S \implies \alpha\text{-core}(\mu_Q) = \alpha\text{-core}(\mu_Q \cap \alpha\text{-core}(\mu_S))$

(iv) Suppose that $\mu_T = \alpha\text{-core}(\mu_S)$, and that $dim(span(\mu_S)) = k$, $dim(span(\mu_T)) = k'$. Then $\mu(S \setminus T) \leq (k - k')\alpha$

Proof: We first establish (i). Let μ_T be arbitrary. Assume (b) does not hold, i.e., there exists $x \in \mu_T$ and a direction w such that $w^T x \neq 0$ and $\sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha$. Writing

$$w = w_1 + w_2, \quad w_1 \in span(\mu_T), \quad w_2 \perp span(\mu_T)$$

we find $w^T x = w_1^T x \neq 0$, while $\sum_{x \in \mu_T} \chi^{\{w'^T x \neq 0\}} \mu(x) = \sum_{x \in \mu_T} \chi^{\{w_1^T x \neq 0\}} \mu(x) < \alpha$. Since $w_1 \in span(\mu_T)$, (a) does not hold.

Now suppose that (b) does hold. If $w \in span(\mu_T)$, then $w^T x \neq 0$ for some $x \in \mu_T$, and hence (b) implies that $\sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \alpha$, so (a) holds too.

To show (ii), we give an algorithm for constructing α-core(μ_S):

1. If there exists $x \in \mu_S$ and a direction w such that $w^T x \neq 0$ but $\sum_{x \in \mu_S} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha$, remove x from μ_S.
2. Repeat until there does not exists such an x.

To argue the correctness of this algorithm, it suffices to show that if x meets the criterion of step 1, then x cannot be in any μ_R, $R \subset S$ such that $\forall w, w^T x \neq 0 \implies \sum_{x \in \mu_R} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha$. But this is obvious, since the w associated with x in step 1 satisfies $w^T x \neq 0$ and yet $\sum_{x \in \mu_R} \chi^{\{w^T x \neq 0\}} \mu(x) \leq \sum_{x \in \mu_S} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha$. Therefore any x identified in step 1 cannot be in α-core(μ_S). Since the algorithm stops when it has arrived at μ_T satisfying (b), and no point has been removed that could be in α-core(μ_S), $\mu_T = \alpha$-core(μ_S). This establishes that the α-core is well-defined.

Now consider the order in which points are identified in step 1 when the algorithm is applied to μ_S. Considering points in the same order (and omitting points that are in μ_S but not in μ_Q), the algorithm run on μ_Q would always remove the points as well, simply because $\sum_{x \in \mu_{Q'}} \chi^{\{w^T x \neq 0\}} \mu(x) \leq \sum_{x \in \mu_{S'}} \chi^{\{w^T x \neq 0\}} \mu(x)$, where Q' and S' are Q and S minus the points that the algorithm has removed prior to the iteration under consideration.

We prove (iii) using (ii).

$$
\begin{aligned}
\alpha\text{-core}(\mu_Q) &\subset \alpha\text{-core}(\mu_S) \Rightarrow \\
\mu_Q \cap \alpha\text{-core}(\mu_Q) &\subset \mu_Q \cap \alpha\text{-core}(\mu_S) \Rightarrow \\
\alpha\text{-core}(\mu_Q \cap \alpha\text{-core}(\mu_Q)) &\subset \alpha\text{-core}(\mu_Q \cap \alpha\text{-core}(\mu_S))
\end{aligned}
$$

We note that α-core$(\mu_Q \cap \alpha$-core$(\mu_Q)) = \alpha$-core(μ_Q). Now

$$
\begin{aligned}
\mu_Q \cap \alpha\text{-core}(\mu_S) &\subset \mu_Q \Rightarrow \\
\alpha\text{-core}(\mu_Q \cap \alpha\text{-core}(\mu_S)) &\subset \alpha\text{-core}(\mu_Q)
\end{aligned}
$$

Combining these yields α-core$(\mu_Q) = \alpha$-core$(\mu_Q \cap \alpha$-core$(\mu_S))$.

To see (iv), construct μ_T from μ_S in the following greedy manner. If $dim(span(\mu_T)) < dim(span(\mu_S))$, then

$$
\exists w \in span(\mu_S) \quad \text{such that} \sum_{x \in \mu_S} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha
$$

If w were not $\perp$ to $span(\mu_T)$, we could write

$$
w = w_1 + w_2, \quad w_1 \perp span(\mu_T), \quad w_2 \in span(\mu_T)
$$

and then argue $\sum_{x \in \mu_S} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) = \sum_{x \in \mu_T} \chi^{\{w_2^T x \neq 0\}} \mu(x) \geq \alpha$. Hence $w \perp span(\mu_T)$. Remove every point $x \in \mu_S$ such that $w^T x \neq 0$ (a less than α fraction of the total probability mass), and note that this causes $dim(span(\mu_S))$ to drop by at least 1. Therefore this construction can be iterated at most $(k - k')$ times, and hence $\mu(S \setminus T) \leq (k - k')\alpha$. ■

We can now define the potential ϕ of a distribution (or a subset of a distribution).

Definition 2 (Potential Function: ϕ) *Let μ_T be the α-core of μ_S. Let $\phi(\mu_S)$ be $Vol(W(M_T))$, the volume of the dual ellipsoid of μ_T. If μ_T is not full dimensional, but instead lies in a space of dimension k, let $\phi(\mu_S)$ be $Vol_k(W(M_T))$, the k-dimensional volume (within the span of μ_T) of the dual ellipsoid of μ_T.*

We now prove upper and lower bounds on $\phi(\mu_S)$, analogous to lemma 2, for the case that the α-core of μ_S is full-dimensional. Although a tighter version of this lemma may be possible, the analysis here is sufficient to show the asymptotic result of theorem 1.

Lemma 5 (Bounds on ϕ) *Denote the α-core of μ_S by μ_T and suppose that μ_T is full-dimensional (and hence $\mu_T = \mu_S$). Then*

$$\phi(\mu_S) \geq (2^b\sqrt{n})^{-n} f(n)$$

$$\phi(\mu_S) \leq (n/\alpha)^n f(n)$$

Proof: We lower bound ϕ by showing that for any vector v satisfying $|v| \leq \frac{2^{-b}}{\sqrt{n}}$, v is in the dual ellipsoid. Using that no element x of μ has length greater than $\sqrt{n}2^b$, we find that

$$v^T M_S v = \sum_{x \in \mu_S} (x^T v)^2 \mu(x) \leq \sum_{x \in \mu_S} x^2 v^2 \mu(x) \leq 1$$

The claimed lower bound now follows from the fact that $W(M_S)$ contains a ball of radius $\frac{2^{-b}}{\sqrt{n}}$.

To upper bound $\phi(\mu_S)$, we will use that μ_T is full-dimensional. Because the volume of an ellipse is equal to the product of the axis lengths times a factor that depends only on the dimension, we have that $\phi(\mu_S) = f(n)/Det(M_S)$ where $M_S = \sum_{x \in \mu_S} xx^T \mu(x)$. We now show that we can decompose M_S into a set of simpler components plus some extra points,

$$M_S = \sum_i \lambda_i M_i + \sum_y yy^T$$

where each M_i is a positive definite nxn matrix of integers and $\sum \lambda_i \geq \alpha/n,\ \lambda_i \geq 0$.

To see this decomposition, begin by picking any point $x_1 \in \mu_S$. Now pick any point $x_2 \in \mu_S$ such that $x_2 \notin span(x_1)$. Now pick any point $x_3 \notin span(x_1, x_2)$. Continuing, we can always make such a choice by considering any direction w perpendicular to the span of the previous points — any point with non-zero inner product with this w, guaranteed to exist by the definition of α-core, lies off the span of the previous points. This first set of points $\{x_j\}_{j=1}^n$ yields $M_1 = \sum_j x_j x_j^T$ with $\lambda_1 = \min_j \mu(x_j)$. To form M_2, we must restrict ourselves to picking points from $\{\mu_S \setminus \lambda_1 M_1\}$ (using slight overloading of notation). By the definition of α-core, as long as $\sum \lambda_i < \alpha/n$, we will always be able to form a new M_i because we have subtracted off less than an α fraction of the probability mass from the distribution thus far. The process can be seen to terminate in a finite number of steps because the support of μ_S is initially a finite number of points, and at every step the cardinality of the support decreases by at least one. The $\{y\}$ which we referred to as "extra points" above are simply the points remaining in μ_S when this operation, having formed a sufficient number of M_i, comes to an end.

Note that each matrix M_i satisifies $Det(M_i) \geq 1$ because it is the sum of the products of many integer terms and it is positive (because M_i is positive definite). We now show that we may ignore the y terms in establishing a lower bound for $Det(M_S)$. Another consequence of M_i being positive definite is that $M_i = A_i A_i^T$ for some A_i. Since the determinant of M_S is the product of the eigenvalues, and each eigenvalue e_j is equal to $\sum_i \lambda_i (A_i^T w_j)^2 + \sum_y (y^T w_j)^2$ for some unit vector w_j, $Det(\sum_i \lambda_i A_i A_i^T) \leq Det(\sum_i \lambda_i A_i A_i^T + \sum yy^T)$.

We have from fact 2 in section 10 (and since the geometric mean is at least the min) that for $\sum \lambda_i' = 1$,

$$Det(\sum_i \lambda_i' M_i) \geq \min_i \{Det(M_i)\}$$

The last step is to write $Det(\xi M) = \xi^n Det(M)$, which implies $Det(M_S) \geq (\alpha/n)^n$. This yields the claimed upper bound on $\phi(\mu_S)$. ■

Note that lemma 5 implies that the log of the ratios between the upper and lower bounds on ϕ is at most $n(b + 1.5(\log \frac{n}{\alpha}))$. (The relevant setting of α for the proof of theorem 1 will be $\alpha = \epsilon/(3n)$.) This compares favorably with the corresponding ratio in the continuous case, $n \ln \frac{R}{r}$, and suggests that we have not introduced much slack while extending our techniques to amortize over the singular values.

We now address the issue of dimension dropping. We refer to *non-monotone growth* in the title of lemma 6 because now ϕ may drop when we remove some of the distribution. To see this, consider example 1 again: ϕ is initially about $f(n)$, but after removing the point p, ϕ becomes roughly $2^{-b(n-1)} f(n)$. In the proof of theorem 2, we bounded the drop in probability mass by bounding the increase in the volume of the dual ellipsoid. Because ϕ may decrease greatly during the course of the algorithm (when the α-core drops in dimension), a bound on ϕ's final value is no longer enough to bound the drop in probability mass. Happily, we can still bound the growth of ϕ in the following sense:

Lemma 6 (Non-Monotone Growth of ϕ) *Over the course of either algorithm on distribution μ, let $(\Delta\phi)_i$ denote the relative increase in ϕ while α-core(μ) spans a subspace of dimension i (or 1 if α-core(μ) is never concentrated on a subspace of dimension i). Then*

$$\prod_i (\Delta\phi)_i \leq 2^{n(b + 3\log\frac{n}{\alpha} + 1)}$$

Proof: Suppose that initially the α-core of μ is full-dimensional, and that $\prod_i (\Delta\phi)_i = V$. Under a simplifying assumption, we construct a distribution μ' such that the α-core of μ' is full-dimensional and $\phi(\mu')/\phi(\mu) \geq V$. (If the result of applying the oulier removal algorithm to μ is μ_S that has full-dimensional α-core, then $\mu' = \mu_S$ and there is nothing to do.) By lemma 5, $\phi(\mu')$ and $\phi(\mu)$ cannot differ by a factor of more than $2^{n(b+1.5\log\frac{n}{\alpha})}$, and thus this suffices to prove the bound on V. We then remove the simplifying assumption. We defer the issue that the α-core of μ might not initially be full-dimensional to the very end of the proof.

Suppose that the algorithm goes from μ_R of dimension $(i+1)$ to μ_S of dimension i, and then runs for a while to produce μ_T (still of dimension i). The simplifying assumption we mentioned above is that the dimension of the α-core has only fallen by 1 on this step. For ease of exposition, assume that each distribution is equal to its α-core. This is without loss of generality because ϕ is defined in terms of the α-core, and so the points outside the α-core are irrelevant for this lemma. We will construct $\mu'_{S'}$ and $\mu'_{T'}$ of dimension $(i+1)$ such that $\phi(\mu'_{S'}) \geq \phi(\mu_R)$ and $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = \frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$. Then we will have

$$\phi(\mu'_{T'}) = \frac{\phi(\mu_T)}{\phi(\mu_S)} \phi(\mu'_{S'}) \geq (\Delta\phi)_i \phi(\mu_R)$$

Applying this construction iteratively over all the dimensions yields μ' of dimension n satisfying $\phi(\mu')/\phi(\mu) \geq V$.

Let us now construct $\mu'_{S'}$ and $\mu'_{T'}$. Define $p_j = \mu(x_j)$ for all $x_j \in \mu_{R\setminus S}$ and let $P = \sum_j p_j$. Then

$$M_R = M_S + \sum_j p_j x_j x_j^T = \sum_i \frac{p_j}{P}(M_S + P x_j x_j^T)$$

Define X_j to be $(M_S + P x_j x_j^T)$. By fact 2 (section 10),

$$Det(\sum \lambda_j X_j) \geq \min\{Det(X_j)\}, \quad \sum \lambda_j = 1$$

there exists j such that $Det(X_j) \leq Det(M_R)$. Denote this particular x_j by x, and let

$$\mu'_{S'} = \{\mu_S + x \text{ with weight } \alpha\}$$

$$\mu'_{T'} = \{\mu_T + x \text{ with weight } \alpha\}$$

Note that $P \geq \alpha$, and so $Det(M'_{S'}) \leq Det(X_j)$. Thus $\phi(\mu'_{S'}) \geq \phi(\mu_R)$.

We now show $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = \frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$. Rotate the distributions so that span(μ_S) and span(μ_T) are equal to the first i coordinate axes, and x lies in the span of the first $i+1$ coordinate axes. Denote the vector formed from the first i coordinates of x by $x[1\ldots i]$, and the $(i+1)^{st}$ coordinate of x by $x[i+1]$. Then the distance of x to span(μ_S) is just $x[i+1]$, and this is also the distance of x to span(μ_T). We have $\phi(\mu_T) = f(i)/Det(M_T)$, while

$$\phi(\mu'_{T'}) = f(i+1)/Det\left(\begin{bmatrix} M_T & 0 \\ 0 & 0 \end{bmatrix} + \alpha^2 \begin{bmatrix} x[1\ldots i]x[1\ldots i]^T & x[i+1]x[1\ldots i] \\ x[i+1]x[1\ldots i]^T & x[i+1]^2 \end{bmatrix}\right)$$

where the upper left matrix block $(M_T + \alpha^2 x[1\ldots i]x[1\ldots i]^T)$ is ixi. For any matrix A, subtracting a scalar multiple of some row of A from another row of A does not change the determinant of A. To calculate $\phi(\mu'_{T'})$, we subtract $x[l]/x[i+1]$ times the last row of the matrix from the l^{th} row for every $l \leq i$. This yields

$$Det\left(\begin{bmatrix} M_T & 0 \\ 0 & 0 \end{bmatrix} + \alpha^2 \begin{bmatrix} 0 & 0 \\ x[i+1]x[1\ldots i]^T & x[i+1]^2 \end{bmatrix}\right) = Det(M_T)\alpha^2 x[i+1]^2$$

Therefore $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = (\alpha x[i+1])^{-2}\frac{f(i+1)}{f(i)}$. An identical calculation yields an identical result for $\frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$. This shows that $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = \frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$.

We now remove the simplifying assumption and extend this construction to the case that at some step the α-core falls in dimension by more than 1. If μ_R and μ_S differ by k dimensions, we construct $\mu'_{S'}$ by adjoining k points from $\mu_{R\setminus S}$, each with weight α/k. We now show how to find these points. Since μ_R is an α-core, and $span(\mu_S)$ is a subspace of k dimensions less, we can use the construction of lemma 5 to write

$$M_R = M_S + \sum_i \lambda_i A_i A_i^T + \sum y y^T, \qquad \sum_i \lambda_i = \Lambda \geq \frac{\alpha}{k}$$

where each A_i is a set of k points such that $span(\{\mu_S + A_i\}) = span(\mu_R)$. As above,

$$Det(M_R) \geq Det(M_S + \sum_i \lambda_i A_i A_i^T) \geq \min_i\{Det(M_S + \Lambda A_i A_i^T)\} \geq \min_i\{Det(M_S + \frac{\alpha}{k} A_i A_i^T)\}$$

Let A denote the A_i realizing this minimum and let

$$\mu'_{S'} = \{\mu_S + A \text{ with weight } \frac{\alpha}{k}\}$$

$$\mu'_{T'} = \{\mu_T + A \text{ with weight } \frac{\alpha}{k}\}$$

We have $\phi(\mu'_{S'}) \geq \phi(\mu_R)$ by construction. It remains to show $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = \frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$. We do this by by showing that the previous calculation (giving this fact under the simplifying assumption) can be repeated k times. Let ${\mu'}^{(l)}_{S'}$ denote $\{\mu_S +$ first l points of A with weight $\frac{\alpha}{k}\}$ and define ${\mu'}^{(l)}_{T'}$ similarly. Then the previous calculation yields

$$\frac{\phi({\mu'}^{(1)}_{T'})}{\phi(\mu_T)} = \frac{\phi({\mu'}^{(1)}_{S'})}{\phi(\mu_S)}, \quad \frac{\phi({\mu'}^{(2)}_{T'})}{\phi({\mu'}^{(1)}_{T'})} = \frac{\phi({\mu'}^{(2)}_{S'})}{\phi({\mu'}^{(1)}_{S'})}, \quad \ldots \quad \Rightarrow \frac{\phi({\mu'}^{(k)}_{T'})}{\phi({\mu'}_{T'})} = \frac{\phi({\mu'}^{(k)}_{S'})}{\phi({\mu'}_{S'})}$$

This concludes the construction of μ' which is full-dimensional and at least an α/n-core.

We now turn to the case that μ is initially only k dimensional, where $k < n$. In this case, we adjoin any $(n-k)$ points, each with weight α, to form $\bar{\mu}$, where $\bar{\mu}$ is full-dimensional. Then $\bar{\mu}$ may not be a probability distribution, but it has total weight at most $1 + n\alpha$, and so

$$\phi(\bar{\mu}) \geq (2^b \sqrt{n}\sqrt{1+n\alpha})^{-n} f(n)$$

by the same construction as in the lower bound of lemma 5. The iterative construction above (without the simplifying assumption) yields $\bar{\mu}'$ such that $\phi(\bar{\mu}') \leq (n^2/\alpha)^n f(n)$, and so

$$\phi(\bar{\mu}')/\phi(\bar{\mu}) \leq (\frac{n^2}{\alpha} 2^b \sqrt{n + n^2\alpha})^n \leq (2n^3 2^b/\alpha)^n \leq 2^{n(b+3\log\frac{n}{\alpha}+1)}$$

This concludes the proof of the lemma. ■

We now prove that Algorithm 2 applied to a distribution μ over the b-bit integers yields S satifying theorem 1.

Proof of Theorem 1: Let $\alpha = \epsilon/(3n)$ and let $\beta = \gamma^2 = 6\frac{n}{\epsilon}(b + 4\log\frac{n}{\epsilon} + 3)$. The only time that the drop in probability mass due to action by the algorithm does not lead to an increase in ϕ is when either the algorithm causes the dimension of the α-core to drop, or the algorithm removes probability mass that lies outside the α-core.

We first consider the fraction of the distribution that is not part of the α-core. Initially, $dim(span(\mu))$ is at most n. Suppose $dim(span(\alpha\text{-core}(\mu))) = k$. Then by lemma 4:(iv), at most an $\alpha(n-k)$ fraction of μ lies outside of the α-core of μ. Points only leave the α-core when they are removed by the algorithm, or when the algorithm takes μ_S to $\mu_{S'}$ in one step and

$$dim(\alpha\text{-core}(\mu_S)) = k_1, \quad dim(\alpha\text{-core}(\mu_{S'})) = k_2, \quad k_2 < k_1$$

In the latter case, the probability mass lost from the α-core (and not removed by the algorithm) is given by

$$\{\mu_{S'} \cap \alpha\text{-core}(\mu_S)\} \setminus \{\alpha\text{-core}(\mu_{S'})\}$$

Since $S' \subset S$, by lemma 4:(iii), this is the same as

$$\{\mu_{S'} \cap \alpha\text{-core}(\mu_S)\} \setminus \{\alpha\text{-core}(\mu_{S'} \cap \alpha\text{-core}(\mu_S))\}$$

and by lemma 4:(iv) this is no more than $\alpha(k_1 - k_2)$. Since the cumulative drop in dimension of the α-core is no more than n dimensions, no more than an αn fraction of the distribution ever leaves the α-core (without being removed by the algorithm) over the course of the algorithm.

We now bound the amount of the distribution removed by the algorithm on steps in which the α-core drops in dimension. In the proof of theorem 2, we showed that in any single step, Algorithm 2 throws out no more than a $1/\gamma^2$ fraction of the distribution. Since there are no more than n steps where the α-core drops in dimension, we throw out no more than an n/γ^2 fraction in this way. This yields that at most an $n\alpha + n/\gamma^2 \leq 2\epsilon/3$ fraction of the probability mass that we throw away does not contribute to increasing ϕ.

We now proceed exactly as we did in the proof of theorem 2 for Algorithm 2. Every time we remove p_i of the probability mass from the α-core and the α-core does not drop in dimension, we have from lemma 1 that ϕ must increase by $e^{p_i\gamma^2/2}$. If we throw out an ϵ' fraction of μ, at most a $2\epsilon/3$ fraction does not contribute to ϕ increasing, so by application of lemma 1

$$\prod(\Delta\phi)_i \geq e^{\frac{\gamma^2}{2}(\epsilon' - \frac{2\epsilon}{3})}$$

Lemma 6 then yields

$$2^{n(b+4\log\frac{n}{\epsilon}+3)} \geq e^{\frac{\gamma^2}{2}(\epsilon' - \frac{2\epsilon}{3})}$$

$$\Rightarrow n(b + 4\log\frac{n}{\epsilon} + 3) \geq \frac{\gamma^2}{2}(\epsilon' - \frac{2\epsilon}{3})$$

$$\Rightarrow 1 \geq \frac{3}{\epsilon}(\epsilon' - \frac{2\epsilon}{3}) \Rightarrow \epsilon' \leq \epsilon$$

This concludes the proof of theorem 1 using Algorithm 2. ■

We now prove theorem 1 using Algorithm 1. As we noted previously, this may be obtained as a corollary of the success of Algorithm 2, but a direct proof raises an additional issue that we explore below. The resolution of this issue leads to a bound on β with smaller leading constant.

Proof of Theorem 1: Let $\alpha = \epsilon/(3n)$ and let $\beta = \gamma^2 = 3\frac{n}{\epsilon}(b + 4\log\frac{n}{\epsilon} + 3)$. The only new issue is bounding the amount of probability mass removed by the algorithm on steps in which the α-core drops in dimension. We might remove up to an n/γ^2 fraction in a single step, but our asymptotic bound would not stand up if we could remove up to an n^2/γ^2 fraction of the probability mass over the course of the algorithm.

Suppose the α-core falls by k dimensions in one step of the algorithm. Rather than considering all the points outside $S = \{x : |x| \leq \gamma \text{ after rounding}\}$ as being removed at once, imagine instead that the probability mass on every point is uniformly decreased. Then, ϕ increases continuously except for at most k discrete time steps, when the dimension of the α-core drops. Apart from the steps on which the α-core drops, ϕ increases as a function of the probability mass removed exactly as implied by lemma 3. Every time the α-core drops by i dimensions, at most an $i\alpha$ amount of probability mass leaves the α-core (by lemma 4:(iv)). Therefore at most a $k\alpha$ amount of probability mass is removed without an increase in ϕ. In this thought experiment, no probability mass in the α-core is removed by the algorithm without an increase in ϕ. Thus at most an $n\alpha = \epsilon/3$ amount of probability mass is removed without an increase in ϕ. We apply lemma 3 and lemma 6 as before to obtain

$$\prod(\Delta\phi)_i \geq e^{\frac{\gamma^2}{2}(\epsilon' - \frac{\epsilon}{3})}$$

$$\Rightarrow 2^{n(b+4\log\frac{n}{\epsilon}+3)} \geq e^{\frac{\gamma^2}{2}(\epsilon' - \frac{\epsilon}{3})}$$

$$\Rightarrow n(b + 4\log\frac{n}{\epsilon} + 3) \geq \frac{\gamma^2}{2}(\epsilon' - \frac{\epsilon}{3})$$

$$\Rightarrow 1 \geq \frac{3}{2\epsilon}(\epsilon' - \frac{\epsilon}{3}) \Rightarrow \epsilon' \leq \epsilon$$

This concludes the proof of theorem 1 using Algorithm 1. ■

5 Time and Sample Complexity

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

5.1 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 round 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. 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 rounding and looking for outliers. In the case that a single data point is throw out, rounding the distribution can be done more efficiently. If the distribution is initially isotropic, and v of probability p is removed, then $M' = I - pvv^T$ gives the new inertial ellipsoid. We can factor ${M'}^{-1}$ symbolically as

$${M'}^{-1} = BB^T = \left(I - \left(1 - \frac{1}{\sqrt{1 - v^Tvp}}\right)\frac{vv^T}{v^Tv}\right)^2$$

where we have chosen B to be symmetric. To verify this calculation, note that

$$BM'B^T = (I - bvv^T)(I - pvv^T)(I - bvv^T) = [I - (2b - b^2v^2)vv^T][I - pvv^T]$$

where $b = \frac{1}{v^2}\left(1 - \frac{1}{\sqrt{1-v^2p}}\right)$ and we have used that the matrices commute. We calculate

$$2b - b^2v^2 = \frac{1}{v^2}\left((2 - \frac{2}{\sqrt{1-pv^2}}) - (1 - \frac{2}{\sqrt{1-pv^2}} + \frac{1}{1-pv^2})\right)$$

$$= \frac{1}{v^2}\left(1 - \frac{1}{1-pv^2}\right) = \frac{-p}{1-pv^2}$$

Plugging this in completes the verification

$$[I - (2b - b^2v^2)vv^T][I - pvv^T] = [I + \frac{p}{1-pv^2}vv^T][I - pvv^T]$$

$$= [I + (\frac{p}{1-pv^2} - p - \frac{p^2v^2}{1-pv^2})vv^T] = I$$

If the old distribution was $\{x\}$, the new isotropic distribution is $\{Bx\}$, where our formula for B yields

$$Bx = x - \left(1 - \frac{1}{\sqrt{1 - v^T vp}}\right)\frac{v(v^T x)}{v^T v}$$

which is computable in time $O(n)$ for any point x. Another explanation for this formula is that we are just correcting the inertial ellipsoid in the direction of v; this type of update step is sometimes referred to as a *rank-1 update*. Using this observation, we can compute M from scratch once $(O(mn^2))$, round the distribution from scratch once $(O(n^3 + mn^2))$, and then find an outlier $(O(mn))$ and reround using our formula above $(O(mn))$ a total of at most m times. This yields the improved time bound of $O(m^2n + mn^2 + n^3)$. If we throw away less than an ϵ fraction of the point set, the time bound is just $O(\epsilon m^2 n + mn^2 + n^3)$.

If we specialize our analysis to $\mathcal{Z}_b^n$ and the case that the distribution has full-dimensional α-core throughout the algorithm, we can obtain a running time with a different dependance on the relevant parameters. Suppose that on some step of Algorithm 1 with parameter β we remove all β-outliers and ϕ (equivalently, the dual ellipsoid volume) increases by a factor of no more than $(1+\delta)$ — then the remaining data set is $(1+\delta)\beta$-outlier free. Because we may have removed many points, we cannot use the technique just developed above, and our time bound is $O(mn^2 + n^3)$ per iteration. However, by our upper and lower bounds on ϕ, there are at most $\log_{(1+\delta)} 2^{\tilde{O}(nb)} = \tilde{O}(\frac{nb}{\delta})$ iterations where ϕ increases by $(1+\delta)$ or more. The final bound on the running time is then $\tilde{O}(\frac{mn^3b+n^4b}{\delta})$ to obtain a $(1+\delta)\beta$-outlier free set.

5.2 Arbitrary distributions

Now suppose that we are not given μ explicitly, but rather only the ability to sample from μ. For ease of exposition, we will refer only to the case that the support of μ is in $\mathcal{Z}_b^n$. The outlier-free restriction of μ will be specified as the part of μ contained in an ellipsoid. The algorithm for distributions is:

1. Get a set $P = \{x_1, \ldots, x_m\}$ of m samples from μ.
2. Run the outlier removal algorithm on the discrete point set P with parameter Γ^2.
3. Let P' be the outlier-free subset of P. Then the outlier-free restriction of P is given by $\Gamma E(M')$, where $M' = \frac{1}{m}\sum_{x_i \in P'} x_i x_i^T$. The outlier-free restriction of μ is given by $(1+\delta)\Gamma E(M')$, where $\delta \in (0, 1/4)$ is an accuracy parameter.

The main theorem of this section is the following.

Theorem 3 (Sample Complexity) *Let*

$$m = O\left(\frac{\gamma^2}{\delta^2}\left(n \log \frac{n}{\delta} + \log \frac{n(b+\log n)}{\delta}\right)\right) = \tilde{O}\left(\frac{n\gamma^2}{\delta^2}\right)$$

Then with high probability, either outlier removal algorithm run with parameter $\Gamma^2 = (1+\delta)^2\gamma^2$ returns an ellipsoid $T = \Gamma E(M')$ satisfying
(i) $\mu((1+\delta)T) \geq 1-\epsilon$
(ii) $(1+\delta)T$ has no $(1+\delta)^{O(1)}\gamma^2$-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 μ_S has no γ^2-outliers. The statement "μ_S has no γ^2-outliers", or simply "S has no γ^2-outliers" (since μ is implicit), is exactly that

$$\forall w, \quad \max\{(w^T x)^2 : x \in S\} \leq \gamma^2 \mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S] = \gamma^2 \sum_{x \in S} (w^T x)^2 \mu(x)$$

The max is not over $x \in \mu_S$, but rather $x \in S$. This is an important subtlety. Since S and T constructed by the algorithm 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 $S \subseteq T$. If we had instead defined the max over $x \in \mu_S$, we would only be able to conclude that $\mu(S \setminus T) = 0$ (i.e., although S might not be a subset of T, μ does not assign positive probability to any point in S that lies outside of T); this alternative definition would have increased the length of the proof.

We know that $\gamma^2 = \tilde{O}(\frac{bn}{\epsilon})$ is always achievable, but in some cases we may do better. 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 our working estimate of M, $\bar{M}$, 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 $(1+\delta)^{O(1)}\gamma^2$-outlier-free subset in space, where γ^2 is achieved by the deterministic version of the algorithm. In lemma 7 we show this for a particular direction in a particular iteration. In lemma 8 we extend this to all iterations, and in the proof of theorem 3 we extend this to all directions and all iterations, at every step bounding the sample complexity.

Lemma 7 (Outlier Detection, One Iteration) *Fix a direction w. Let S be a subset of space. Let our number of samples be*

$$m = O(\frac{\gamma^2}{\delta^2})$$

and consider the sample distances in direction w given by $\{w^T x_i\}$. Let y denote the true variance of S and $\bar{y}$ denote the sample variance,

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

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

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 is assumed to be γ^2-outlier-free in the hypotheses of both (i) and (ii). Lemma 8 will not rely upon part (ii) explicitly, but it will involve a similar argument.

Let X_i be the random variable representing the squared distance of x_i along the direction w, $X_i = (w^T x_i)^2$, or 0 if $x_i \notin S$. Without loss of generality, assume $\max\{(w^T x)^2 : x \in S\} = 1$ (by an appropriate scaling). First we show (i). Since μ_S has no γ^2-outliers, we have $y \ge \frac{1}{\gamma^2}$. Applying the Chernoff bound to determine the probability that $\bar{y}$ is not a good estimate for y, we have

$$\Pr[|m\bar{y} - my| \ge \delta m y] \le e^{-\delta^2 m y / 3}$$

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 $y \ge \frac{1}{\gamma^2}$, and we would have found $\bar{y}$ to be an accurate estimate by the analysis in the previous paragraph. In this case, $(1-\delta) y \le \bar{y} \Rightarrow y \le (1+\delta)^2 \bar{y}$, and S has no γ^2-outliers implies $\max\{(w^T x)^2 : x \in S\} \le \gamma^2 y \le \Gamma^2 \bar{y}$. This then implies $S \subseteq T$. ∎

Lemma 8 (Outlier Detection, Many Iterations) *Fix w. Assume S is full-dimensional. Let*

$$m = O\left(\frac{\gamma^2}{\delta^2} \log \frac{n(b + \log n)}{\delta}\right) = \tilde{O}\left(\frac{\gamma^2}{\delta^2}\right)$$

Then with constant probability either outlier removal algorithm restricted to w with parameter Γ^2 produces a subset of space

$$T = \{x : (w^T x)^2 \le t\}$$

for some value t such that
(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 version 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 $\bar{y}$ as in lemma 7. By lemma 7 part (i), we have that $\bar{y}$ 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.

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 $\alpha \in [t, (1+\delta)t]$ by the same reasoning as in the preceding paragraph. For every t, we will show that for some nearby (within a factor of $(1+\delta)$) value of α, we estimate the sample variance of the restriction of μ to T_α with sufficient accuracy. We proceed to analyze what values of α we need to consider.

Assume without loss of generality that w is a unit vector. An easy upper bound on $\max(w^T x)^2$ is $2^b\sqrt{n}$. To develop a lower bound, we will need to use the assumption that S is full-dimensional. For any μ_S, we can write $\max(w^T x)^2 \geq \mathbf{E}[(w^T x)^2]$. By decomposing μ_S in the manner of lemma 5, we can obtain the stronger statement $\max(w^T x)^2 \geq \mathbf{E}_{\{y_j\}_{j=1}^n}[(w^T y_j)^2]$ where the probability distribution on the $\{y_j\}$ is uniform and the $\{y_j\}$ are full-dimensional. The term $\mathbf{E}[(w^T y_j)^2]$ is lower bounded by the smallest singular value of the $\{y_j\}$. We have previously shown that the product of the singular values of such a distribution is at least n^{-2n}. Since no individual singular value is more than $2^b\sqrt{n}$, we have that the smallest is at least $n^{-2n}\, 2^{n(b+.5\log n)} = 2^{\tilde{O}(nb)}$. Therefore we can restrict our attention to $\alpha = (1+\delta)^k$ for k an integer and union bound over the at most $\log_{(1+\delta)} 2^{\tilde{O}(nb)} = O(\frac{n(b+\log n)}{\delta})$ 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 with good probability our sample shows T_α to have at least a $(1+\delta)^2\Gamma^2$-outlier. Let X_i be the random variable representing the squared distance of x_i along the direction w, $X_i = (w^T x_i)^2$, or zero if $x_i \notin T_\alpha$. Without loss of generality, assume $\alpha = 1$. Define y and $\bar{y}$ as in lemma 7 (but with T_α in place of S). Then by assumption on T_α, $y = \mathbf{E}[X_i] \leq \frac{1}{(1+\delta)^6\Gamma^2}$. The condition that our samples show T_α to have at least a $(1+\delta)^2\Gamma^2$-outlier is $\bar{y} = \frac{1}{m}\sum X_i \leq \frac{1}{(1+\delta)^2\Gamma^2}$. We apply the Chernoff bound,

$$\Pr[\bar{y} > (1+\Delta)y] < e^{-\Delta^2 my/3}$$

where we have stated the Chernoff bound for the case that $\Delta < 1$. Let $\Delta = \frac{1}{y(1+\delta)^2\Gamma^2} - 1$ (this yields the event that $\bar{y} > \frac{1}{(1+\delta)^2\Gamma^2}$ in our probability calculation). If $\Delta < 1$, then

$$\Delta^2 y \quad = \quad \left(\frac{1}{(1+\delta)^2\Gamma^2} - y\right)^2 \frac{1}{y} \quad \geq \quad \left(\frac{4\delta}{(1+\delta)^6\Gamma^2}\right)^2 \frac{1}{y} \quad \geq \quad \frac{\delta^2}{\Gamma^2}$$

and the probability we do not correctly identify the furthest outlier is at most $e^{-\Delta^2 my/3} = O(1)$ for $m = O(\frac{\Gamma^2}{\delta^2})$. If $\Delta \geq 1$, then

$$\Delta y = \frac{1}{(1+\delta)^2\Gamma^2} - y \geq \frac{\delta}{\Gamma^2}$$

and the applicable alternate form of the Chernoff bound

$$\Pr[\bar{y} > (1+\Delta)y] < e^{-\Delta my/3}$$

yields that $e^{-\Delta my/3} = O(1)$ for the same setting of m.

Since there are only $O(\frac{n(b+\log n)}{\delta})$ different values of α to consider, $m = O(\frac{\Gamma^2}{\delta^2} \log \frac{n(b+\log n)}{\delta})$ 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 7 and 8 from a fixed direction to all directions and argue the correctness of the entire algorithm by proving theorem 3.

Proof of Theorem 3: Let S be the ellipsoid found by the deterministic algorithm (i.e. the outlier-free subset of points lies in this ellipsoid). Assume initially that S is full-dimensional. Rather than considering the original space, consider the transformed space where S is the unit sphere.

Consider the many directions w given by a δ'-grid in the unit cube, $\delta' = \frac{\delta}{6n}$. We form this grid by choosing every w such that the coordinates of w lie in $\{0, \delta', 2\delta', \ldots, 1\}$. By our choice of m, we can apply lemma 8 part (i) to each of these $(\frac{6n}{\delta})^n$ directions simultaneously and then union bound. Then with good probability, for every w in the δ'-grid, $\max\{(w^T x)^2 : x \in T\} \geq \max\{(w^T x)^2 : x \in S\}$ (i.e., in this direction T contains S). We now show that for an arbitrary direction w, $(1+\delta)T$ contains S.

Consider an arbitrary unit vector w. By rounding every coordinate of w up or down to an integer multiple of δ' we obtain a point on the δ'-grid. The set of all possible such roundings forms a box surrounding w, and some (not neccessarily unique) subset of n of these points, which we denote $\{w_i\}$, satisfy that w is in the convex cone of the $\{w_i\}$. Since w is a unit vector, each w_i has length $|w_i| \in (1 \pm \delta'\sqrt{n})$, and so $\hat{w}_i = w_i/|w_i|$ is within $2\delta'\sqrt{n}$ of w. Define $T(y)$ to be the distance to the boundary of T along the direction y. Since T is convex and $T(w_i) \geq 1$, the quantity $T(w)$ is lower bounded by the minimum distance of points on the convex hull of the $\{\hat{w}_i\}$ to the origin. Since w is within $2\delta'\sqrt{n}$ of each $\{\hat{w}_i\}$, so is the projection of w to their convex hull. Since the point on the convex hull is at most $2\delta'\sqrt{n}$ away from $\hat{w}_i$ for any i, $T(w) \geq 1 - 2\delta'\sqrt{n} \geq 1 - \delta/3$. Since S is within 1 of the origin everywhere, $(1+\delta)T$ contains S. This concludes the proof of (i).

Now consider (ii). Since $S \subset (1+\delta)T$, T is full-dimensional as well. 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 8 part (ii). As before, consider the transformed space in which $(1+\delta)T$ is the unit sphere. Let $R = E(M_T)$ be the actual inertial ellipsoid of μ_T. Let w be an arbitrary unit vector and define $\{w_i\}$ as before. We have that $R(w_i) \geq \frac{1}{(1+\delta)^8\Gamma^2}$ and we reason as above that $R(w) \geq \frac{1-\delta/3}{(1+\delta)^8\Gamma^2} \geq \frac{1}{(1+\delta)^9\Gamma^2}$. Therefore $(1+\delta)T$ is $(1+\delta)^9\Gamma^2$-outlier-free.

We now remove the assumption that S is full-dimensional. Suppose S is not full-dimensional, but rather spans a subspace ζ. It suffices to consider $w \in \zeta$. For such a w, the projection of the associated $\{w_i\}$ to ζ yields $\{w_i'\}$ that are within $\delta'\sqrt{n}$ of the $\{w_i\}$ (because they don't move further than the distance to w upon projection). We can compare the $\{w_i'\}$ and w just as we did the $\{w_i\}$ and w previously. Because the max along w_i' is within a factor $(1-\delta/3)$ of the max along w_i, and the max along w_i' was lower bounded in lemma 8, the max along w_i is similarly lower bounded even though $w_i \notin \zeta$ (the change in the lower bound on the max is asymptotically negligible). Therefore we can apply lemma 8 part (i) to w_i. Thus $T(w_i) \geq 1-\delta/3$, and so $T(w) \geq 1-\frac{2\delta}{3}$. Thus $(1+\delta)T$ contains S. This establishes part (i).

We can extend the proof of part (ii) to the case that T is not full-dimensional in an identical manner. This concludes the proof of theorem 3. ■

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

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 from section 5.1 yields that the algorithm runs in time $\tilde{O}(\frac{b^2n^5}{\epsilon\delta^4})$, which is the bound we referred to in the introduction. In this time we achieve a $(1+\delta)^{O(1)} = 1+O(\delta)$ approximation to the optimal value of β. ■

We now pose a related problem: Suppose that we are not given the parameter γ^2, but rather only ϵ, and asked to find the appropriate γ^2. Lemma 9 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 $\log_{(1+\delta)} \tilde{O}(\frac{bn}{\epsilon}) = O(\frac{\log(\frac{bn}{\epsilon})}{\delta})$ 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 T satisfying parameters $((1+O(\delta))\gamma^2, (1+O(\delta))\epsilon)$. Our asymptotic running time is still $\tilde{O}(\frac{b^2n^5}{\epsilon\delta^4})$.

Lemma 9 (Probability Mass Location) *Let E be an ellipsoid. Let our number of samples $\{x_i\}_{i=1}^m$ be $m = O(\frac{1}{\epsilon\delta^2})$. Then with constant probability, if we estimate a $(1+\delta)\epsilon$ fraction of our samples to be outside of E, at most a $(1+\delta)^2\epsilon$ fraction is outside of E, and at least an ϵ fraction is outside of E.*

Proof: Round E. Let Y_i be a random variable, $Y_i = 1$ iff $x_i^2 > 1$. Let $\bar{y} = \frac{1}{m}\sum Y_i$ and $y = \mathbf{E}[y_i]$. The event that we estimate a $(1+\delta)\epsilon$ fraction of the sample to be outside E when less than an ϵ fraction truly lies outside E, is $y < \epsilon$, $\bar{y} \geq (1+\delta)\epsilon$. We can upper bound the probability of this event using the Chernoff bound

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

where $\Delta = \frac{(1+\delta)\epsilon}{y} - 1$. Then

$$\Delta^2 y = (\frac{(1+\delta)\epsilon}{y} - 1)((1+\delta)\epsilon - y) \geq (\frac{(1+\delta)\epsilon}{\epsilon} - 1)((1+\delta)\epsilon - \epsilon) = \delta^2\epsilon$$

and so the upper bound on the probability is constant for $m = O(\frac{1}{\epsilon\delta^2})$. If $\Delta \geq 1$, in which case the alternate form of the Chernoff bound is applicable, we find $\Delta y \geq \delta\epsilon$, and so the number of samples is still sufficient.

A similar calculation for the event that $y > (1+\delta)^2\epsilon$, $\bar{y} \leq (1+\delta)\epsilon$ using

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

involves setting $\Delta = 1 - \frac{(1+\delta)\epsilon}{y}$, which yields

$$\Delta^2 y = (y - (1+\delta)\epsilon)(1 - \frac{(1+\delta)\epsilon}{y}) \geq \delta^2\epsilon$$

and similarly for the alternate form of the Chernoff bound if $\Delta \geq 1$. Therefore the probability of significantly underestimating the amount of probability mass outside E is at most a constant for the same value of m. ■

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 on Z_b^n (after removing at most an ϵ fraction) and thus bring it into nearly isotropic position.

6 A Matching Lower Bound

We show that for any $\epsilon < 1/4$ there exists a distribution μ with support $\subset \mathcal{Z}_b^n$ 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 theorem 1 as asymptotically optimal.

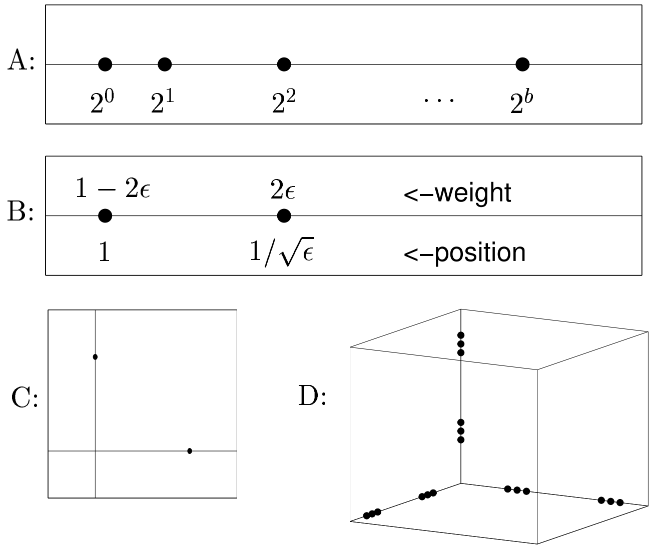

Figure 2: Lower Bound Constructions

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 $\{2^0, 2^1, \dots 2^b\}$. 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), $\max\{x^2 : x \in S\} = 2^{2k}$, while $\mathbf{E}[x^2 : x \in S] \leq \frac{2^0 + \dots 2^{2k}}{(b+1)(1-\epsilon)} = O(\frac{2^{2k}}{b})$. Since $\beta = \frac{\max\{\cdot\}}{\mathbf{E}[\cdot]}$, we find $\beta = \Omega(b)$.

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 $\max\{(w^T x)^2 : x \in S\} = 1$, $\mathbf{E}[(w^T x)^2 : x \in S] \leq \frac{4}{n}$, 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 interesting choices to ones that treat each axis symmetrically. For the purpose of establishing a contradiction, suppose that it helped to treat the different axes differently. 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 — to see this, note that the rounding transformation need only scale the axes, the maximizing w after rounding is in the direction of some point (i.e., along an axis), and therefore the maximizing w before rounding is also along an axis. Let μ_1 be a distribution concentrated on the axes and symmetric on each axis 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 such that this $\bar{\beta}$ is achievable, and the only S achieving $\bar{\beta}$ is asymmetric. Let axis i be an axis that a 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

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

$$\leq \frac{2^{2b'-1}}{2b'} + \frac{2^{2k+1}}{b'} \leq \frac{2^{2k+2}}{b'}$$

yields that $\bar{\beta} = \frac{\max[\cdot]}{\mathbf{E}[\cdot]} = \frac{\max[\cdot]}{\mathbf{E}[\cdot]\Pr[\cdot]}\Pr[\cdot] \geq \frac{b'}{4\epsilon}(1-\epsilon) \geq \frac{b'}{8\epsilon}$ for the one-dimensional case. Thus our lower bound in the n-dimensional case is

$$\bar{\beta} \geq \frac{n}{16\epsilon}(b - \log\frac{1}{\epsilon})$$

7 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 show this problem to be NP-hard even for one-dimensional data by a reduction from the subset-sum problem. We then 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.

The subset-sum problem is: given $p_i \in (0,1), i \in \{1, ...n\}$, find I maximizing $\sum_{i \in I} p_i$ subject to the constraint that $\sum_{i \in I} p_i \leq 1$. To form a corresponding instance (μ, ϵ) of the outlier removal problem, let $P = \sum_i p_i, \epsilon = \frac{1}{2P}$, and let μ be given by

- a point at 1 with probability mass $\frac{1}{2}$
- $\forall i$, a point at 0 with probability mass $p'_i = \frac{p_i}{2P}$

Let S be a possible solution to this instance of the outlier removal problem. Since $P > 1$ (otherwise the subset-sum problem is trivial), the point at 1 cannot be removed, and hence $\max_{x \in S} = 1$. If we remove probability mass ϵ' of the points at 0, $\mathbf{E}[x^2 : x \in S] = \frac{(1)\frac{1}{2} + (0)(\frac{1}{2} - \epsilon')}{1 - \epsilon'} = \frac{1}{2 - 2\epsilon'}$. Thus the ratio $\frac{\max[\cdot]}{\mathbf{E}[\cdot]} = 2 - 2\epsilon'$, and minimizing this subject to $\epsilon' \leq \epsilon$ is exactly the problem of finding the optimal solution I to the subset-sum problem.

We now prove a lemma that enables the approximation result.

Lemma 10 (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 5, 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 10. 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. As before, the running time is $O(m^2 n)$ for $m > n$.

The outlier removal algorithm 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 one at a time, and then look back to see the lowest value of β achieved by any $\epsilon' < \epsilon$.

8 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 $\mathcal{Z}_b^n$. 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. The appropriately modified outlier removal algorithm for constructing S is simply to translate the data set so that the origin coincides with the mean before each removal step. We can easily show that translating μ so that the origin coincides with the mean never decreases the volume of the dual ellipsoid of μ. We then explain how a variation on our potential function ϕ, and the upper and lower bounds on ϕ, imply that the modified algorithm does not throw out more than an ϵ fraction of the data set.

To analyze the volume of the dual ellipsoid, consider a fixed direction w and let $\frac{1}{r^2} = \mathbf{E}[(w^T x)^2]$ (r is the length of the dual ellipsoid in this direction). If we translate the origin to a value z along w, then have $\frac{1}{r^2} = \mathbf{E}[(w^T(x-z))^2]$. Single variable calculus shows that the value maximizing r is $z = \mathbf{E}[w^T x/|w|]$, which is just the mean. Thus translating our origin to $\bar{x}$ maximizes the length of the dual ellipsoid in every direction simultaneously. Thus the tradeoff between drop in probability mass and growth of the dual ellipsoid shown in lemmas 1 and 3 also holds for the modified algorithm.

To describe our modified ϕ, we need to define the *α-affine-core* of a distribution μ_S to be μ_T where $T \subset S$ is chosen to be maximum subject to the requirement that the affine hull of $\{\mu_T$ minus an α fraction of $\mu_T\}$ is not of lower dimension than the affine hull of μ_T for any choice of the α fraction. Under this definition, an appropriately modified version of lemma 4 is still true. Define $\phi'(\mu_S)$ to be an appropriately modified ϕ, $\phi'(\mu_S) = Vol(W(M_T))$ where μ_T is the α-affine-core of μ_S. We now explain how to derive upper and lower bounds on ϕ' analogous to lemma 5 in the case that the α-affine-core of μ_S is full-dimensional.

The lower bound is immediately implied by the argument above that translating the origin to the mean does not decrease the dual volume. To derive the upper bound, consider a set A of $n+1$ points $\{a_i\}$ whose affine hull is full-dimensional. In lemma 5, we argued that $Det(AA^T)$ was a positive integer, not zero by choice of A, and thus $Det(AA^T) \geq 1$. Letting $\bar{a} = \frac{1}{n+1}\sum_i^{n+1} a_i$, we must lower bound $Det(\sum_i^{n+1}(a_i - \bar{a})(a_i - \bar{a})^T)$. Writing

$$(n+1)^{2n} Det(\sum_i^{n+1}(a_i - \bar{a})(a_i - \bar{a})^T) = Det(\sum_i^{n+1}((n+1)a_i - (n+1)\bar{a})((n+1)a_i - (n+1)\bar{a})^T)$$

we have that the second term is the positive non-zero determinant of an integer matrix, and hence the original determinant is at least $\frac{1}{(n+1)^{2n}}$. Because the origin corresponding to the mean of a set of points maximizes the dual volume, this bound holds for all possibilities for the origin. The upper bound on ϕ' is then $(\frac{n}{\alpha})^n (n+1)^{2n} f(n)$.

To prove a statement analogous to lemma 6 for the cumulative drop in ϕ', we revisit the construction of $\mu'_{T'}, \mu'_{S'}$ from μ_R, μ_S, μ_T. Define these objects just as in the proof of lemma 6. We have that $Det(M'_{S'}) \leq Det(M_R)$ when the origin is the mean of μ_R, and so $\phi'(\mu'_{S'}) \geq \phi'(\mu_R)$ to at least the same extent. We now calculate $\frac{\phi'(\mu'_{T'})}{\phi'(\mu_T)}$. Letting the origin correspond to the mean of μ_T, we have $\phi'(\mu_T) = f(i)/Det(M_T)$ where $M_T = \sum_{y\in T} yy^T\mu(y)$. The mean of $\mu'_{T'}$ is given by $\bar{x} = \frac{\alpha x}{\alpha + \mu(T)}$. Then

$$M'_{T'} = \sum_{y\in T}(y-\bar{x})(y-\bar{x})^T\mu(y) + \alpha^2(x-\bar{x})(x-\bar{x})^T =$$

$$\left(\sum_{y\in T} yy^T\mu(y) - \sum_{y\in T}\bar{x}y^T\mu(y) - \sum_{y\in T} y\bar{x}^T\mu(y) + \bar{x}\bar{x}^T\mu(T)\right) + \alpha^2(x-\bar{x})(x-\bar{x})^T =$$

$$\sum_{y\in T} yy^T\mu(y) + \bar{x}\bar{x}^T\mu(T) + \alpha^2(x-\bar{x})(x-\bar{x})^T = \sum_{y\in T} yy^T\mu(y) + \frac{\mu(T)^2\alpha^2 + \mu(T)\alpha^2}{(\mu(T)+\alpha)^2}xx^T$$

Performing the same analysis using Gaussian elimination as we did previously and then computing the ratio yields

$$\frac{\phi'(\mu'_{T'})}{\phi'(\mu_T)} = \frac{f(i+1)}{f(i)}\frac{(\mu(T)+\alpha)^2}{\mu(T)\alpha^2(1+\mu(T))x[i+1]^2}$$

$$\Rightarrow \frac{\phi'(\mu'_{T'})}{\phi'(\mu_T)}\frac{\phi'(\mu_S)}{\phi'(\mu'_{S'})} = \frac{(\mu(T)+\alpha)^2}{\mu(T)\alpha^2(1+\mu(T))}\frac{\mu(S)\alpha^2(1+\mu(S))}{(\mu(S)+\alpha)^2}$$

We will now assume that we never remove more than an ϵ fraction of the probability mass. This is not circular reasoning — just as in the proof of theorem 2 using algorithm 2, the upper bound on ϕ' under this assumption will imply that we never remove more than an $\epsilon/2$ fraction of the probability mass, and since we never remove more than an $\epsilon/2$ fraction on any one step, the assumption will always hold. Using this assumption, we calculate

$$\frac{(\mu(T)+\alpha)^2}{(\mu(S)+\alpha)^2} \geq \frac{(1-\epsilon)^2}{1^2} \geq \frac{1}{4}, \quad \frac{\mu(S)\alpha^2(1+\mu(S))}{\mu(T)\alpha^2(1+\mu(T))} \geq 1$$

Multiplying these factors together over the at most n steps in the iterative construction yields an additional cumulative factor of at most 2^{2n}, which is negligible. Combining this bit of additional slack with the new bound on ϕ' in the full dimensional case and the possibility that we only have an (α/n)-affine-core (as at the end of the proof of lemma 6), we finally arrive at a bound on the total cumulative drop in ϕ' of

$$2^{n(b+3\log\frac{n}{\alpha}+3)}$$

This immediately implies the claimed value for β in corollary 2. ■

We now show that the $\frac{1}{1-\epsilon}$-approximation algorithm of section 7 naturally extends to a $\left(\frac{1-\epsilon}{1-3\epsilon}\right)$-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 10.

Lemma 11 (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 $\beta\frac{1-3\epsilon}{1-\epsilon}$-outlier with respect to any subset S satisfying $\mu(S) \geq 1-\epsilon$.*

Proof: As in the proof of lemma 10, consider a unit vector 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 10 will result from the mean possibly moving closer to x_0 after removing other points $\{x_i\}$. Without loss of generality, let the mean of μ be the origin, and let $\mathbf{E}[(w^T x)^2] = 1$.

Suppose that to reach S we remove points $\{x_i\}$ of total probabilty mass $\epsilon' \leq \epsilon$. Then

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

$$\Rightarrow \mathbf{E}[(w^T x)^2 : x \in S] = (1 - \sum_i (w^T x_i)^2 \mu(x_i))/(1-\epsilon')$$

We calculate the new mean as

$$\mathbf{E}[(w^T x) : x \in S] \Pr[x \in S] = 0 - \sum_i (w^T x_i)\mu(x_i)$$

$$\Rightarrow \mathbf{E}[(w^T x) : x \in S] = (0 - \sum_i (w^T x_i)\mu(x_i))/(1-\epsilon')$$

Therefore the new distance of x_0 to the mean is $\left(\gamma - (0 - \sum_i (w^T x_i)\mu(x_i))/(1-\epsilon')\right)$. We calculate

$$\gamma'^2 = \frac{\text{distance}^2}{\text{variance}} = \frac{\left(\gamma + \frac{\sum_i (w^T x_i)\mu(x_i)}{1-\epsilon'}\right)^2}{\left(\frac{1-\sum_i (w^T x_i)^2 \mu(x_i)}{1-\epsilon'}\right)} = \frac{((1-\epsilon')\gamma + \sum_i (w^T x_i)\mu(x_i))^2}{(1-\epsilon')(1-\sum_i (w^T x_i)^2 \mu(x_i))}$$

Let $\bar{x} = \frac{\sum w^T x_i \mu(x_i)}{\epsilon'}$, the average of the points. Then removing $\bar{x}$ of weight ϵ' changes the numerator by the same amount, and $\bar{x}^2 \epsilon' \leq \sum (w^T x_i)^2 \mu(x_i)$, so the denominator cannot decrease. The derivation of $\bar{x}^2 \epsilon \leq \sum x_i^2 \mu(x_i)$ follows from

$$(\sum \lambda_i x_i)^2 \leq \sum \lambda_i x_i^2, \quad \sum \lambda_i = 1, \quad \lambda_i \geq 0$$

which follows from

$$(\frac{1}{2}a + \frac{1}{2}b)^2 \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$$

along the same lines that fact 2 follows from fact 1 in section 10. Now we have shown we may consider removing only a single point $\bar{x}$ of weight ϵ' in order to lower bound γ'^2. We may view this as a constrained maximization problem over $\bar{x}$, with constraints $|\bar{x}| \leq \gamma$, and $\bar{x}^2 \epsilon' \leq 1$. The expression for $f(\bar{x}) = \gamma'^2$ is

$$\gamma'^2 = \frac{((1-\epsilon')\gamma + \epsilon'\bar{x})^2}{(1-\epsilon')(1-\epsilon'\bar{x}^2)}$$

If the constraint $\bar{x}^2\epsilon' \leq 1$ were tight, then the variance of the distribution after removing $\bar{x}$ would be 0, which would imply $\gamma'^2 = 1$. If the constraint $|\bar{x}| \leq \gamma$ were tight, we would have

$$\gamma'^2 = \frac{((1-\epsilon')\gamma - \gamma\epsilon')^2}{(1-\epsilon')(1-\gamma^2\epsilon')} = \gamma^2 \frac{(1-2\epsilon')^2}{1-\epsilon'} \frac{1}{1-\gamma^2\epsilon'} \geq \gamma^2 \left(\frac{1-2\epsilon'}{1-\epsilon'}\right)^2 \geq \gamma^2 \left(\frac{1-3\epsilon'}{1-\epsilon'}\right)$$

If neither constraint is tight, we may solve the unconstrained optimization problem by setting $\frac{df(\bar{x})}{d\bar{x}} = 0$ to find the local maximum, and then evaluating $f(\bar{x})$ at this maximum.

$$f(\bar{x}) = \gamma'^2 = \frac{1}{1-\epsilon} \frac{u(\bar{x})^2}{v(\bar{x})}$$

$$\begin{aligned}
\frac{df(\bar{x})}{d\bar{x}} = \frac{1}{1-\epsilon}\left(\frac{2u(\bar{x})u'(\bar{x})}{v(\bar{x})} - \frac{u(\bar{x})^2 v'(\bar{x})}{v(\bar{x})^2}\right) = 0 \quad &\Rightarrow \\
2v(\bar{x})u'(\bar{x}) - u(\bar{x})v'(\bar{x}) = 0 \quad &\Rightarrow \\
2(1-\epsilon'\bar{x}^2)(\epsilon') - ((1-\epsilon')\gamma + \epsilon'\bar{x})(-2\epsilon'\bar{x}) = 0 \quad &\Rightarrow \\
(1-\epsilon'\bar{x}^2) + ((1-\epsilon')\gamma + \epsilon'\bar{x})\bar{x} = 0 \quad &\Rightarrow \\
1 + \gamma\bar{x} - \epsilon'\gamma\bar{x} = 0 \quad \Rightarrow \quad \bar{x} = -\frac{1}{(1-\epsilon')\gamma}
\end{aligned}$$

$$f(\bar{x}) = \frac{((1-\epsilon')\gamma - \frac{1}{(1-\epsilon')\gamma}\epsilon')^2}{(1-\epsilon')(1-\frac{1}{(1-\epsilon')^2\gamma^2}\epsilon')} = \frac{((1-\epsilon')^2\gamma^2 - \epsilon')^2}{(1-\epsilon')((1-\epsilon')^2\gamma^2 - \epsilon')} =$$

$$\gamma^2 \frac{(1-\epsilon')^2 - \frac{\epsilon'}{\gamma^2}}{(1-\epsilon')} \geq \gamma^2 \left(\frac{1-3\epsilon'}{1-\epsilon'}\right)$$

which proves the lemma. ■

9 A Robust Statistic

In robust statistics, the choice of the median as the quintessential robust statistic is commonly motivated by describing it as a "robust version of the mean." In particular, it is noted that for any data set, the mean of the data set can be changed by an arbitrary amount simply by moving one of the data points to infinity. In contrast, the median does not "go to absurdity," as the literature commonly puts it, until at least half of the data has been so changed by an adversary.

For a one-dimensional data set, define a δ-median to be a point such that at least a δ fraction of the data lies to the left of the point and at least a δ fraction to the right. In n-dimensions, call a point a δ-median if, for every direction w, it satisfies the definition of the one-dimensional δ-median under projection to w.

Using Helly's theorem, one can prove that $\frac{1}{n+1}$-medians exist for any n-dimensional data set (or distribution), and this is best possible. Such a point is called a *centerpoint*. Centerpoints were proposed by Donoho and Gasko[DG 92] as a robust estimator for high-dimensional data. Donoho and Gasko showed centerpoints to have a *high breakdown point*, which is a technical criterion of "robustness" that we shall not discuss further here.

Teng et al [CEMST 93] gave the first polynomial time algorithm for computing an approximate center point (polynomial in n). Their algorithm produces $\Omega(\frac{1}{n^2})$-medians. We show that the algorithm of section 8 produces $\frac{1}{2\gamma^2(1-\epsilon)}$-medians. For a distribution on $\mathcal{Z}_b^n$, this yields $\tilde{\Omega}(\frac{1}{nb})$-medians.

Theorem 4 *Let μ be a distribution, let $\bar{x} = \mathbf{E}[x : x \in S]$, and suppose S satisfies*
(i) $\mu(S) \geq 1 - \epsilon$
(ii) $\max\{(w^T(x-\bar{x}))^2 : x \in S\} \leq \gamma^2 \mathbf{E}[(w^T(x-\bar{x}))^2 : x \in S]$ for all $w \in \mathcal{R}^n$
Then $\bar{x}$ is a $\frac{1}{2\gamma^2(1-\epsilon)}$-median.

Proof: Suppose initially that $\mu(S) = 1$. Without loss of generality, consider a particular direction given by the unit vector w, and assume that $w^T\bar{x} = 0$ and $\mathbf{E}[(w^Tx)^2] = 1$. Since we are restricting our attention to w for the rest of the proof, we may define $y_i = w^Tx_i$. Let $\{y_i\}$ denote the distribution μ on S, and let I denote the index set. We partition I and define $\delta^{\pm}$ via

$$I^- = \{i : x_i < 0\} \qquad I^+ = \{i : x_i \geq 0\}$$
$$\delta^- = \sum_{i \in I^-} \mu(x_i) \qquad \delta^+ = \sum_{i \in I^+} \mu(x_i)$$

Then we have

$$\sum_{i \in I^-} x_i\mu(x_i) + \sum_{i \in I^+} x_i\mu(x_i) = 0 \qquad \sum_{i \in I} x_i^2\mu(x_i) = 1$$

Using that $x_i^2 \leq \gamma|x_i|$, we obtain

$$1 = \sum_{i \in I} x_i^2\mu(x_i) \leq \sum_{i \in I} \gamma|x_i|\mu(x_i) = \gamma(\sum_{i \in I^+} x_i\mu(x_i) - \sum_{i \in I^-} x_i\mu(x_i)) = \gamma(2\sum_{i \in I^+} x_i\mu(x_i)) \leq 2\gamma^2\delta^+$$

From this we conlude that $\delta^+ \geq \frac{1}{2\gamma^2}$, and similary for δ^-. Dropping the assumption that $\mu(S) = 1$ turns our lower bound into $\frac{1}{2\gamma^2(1-\epsilon)}$. ■

10 Some Properties of Matrices

The proof in section 4 relied on fact 2, which we speculate to be well-known. We present the proof of this fact here since it uses techniques that are otherwise not necessary in the rest of section 4.

Fact 1 *For X, Y positive definite*

$$Det((X+Y)/2) \geq \sqrt{Det(X)Det(Y)}$$

Proof: This statement is equivalent to (clearing denominators and squaring twice)

$$Det(XY) \leq Det^2((X+Y)/2)$$

which is equivalent to

$$\begin{aligned}
1 &\leq \frac{Det^2((X+Y)/2)}{Det(XY)} \\
&= Det(\frac{1}{4}(X+Y))Det(X^{-1})Det(X+Y)Det(Y^{-1}) \\
&= Det(\frac{1}{4}(X+Y)(X^{-1})(X+Y)(Y^{-1})) \\
&= Det(\frac{1}{4}(I+YX^{-1})(XY^{-1}+I)) \\
&= Det(\frac{1}{4}(YX^{-1}+2I+XY^{-1})) \\
&= Det(\frac{1}{4}(A+2I+A^{-1}))
\end{aligned}$$

where we let $A = YX^{-1}$ at the very end. Also let $B = \frac{A+2I+A^{-1}}{4}$. We have reduced to the case of showing that $Det(B) \geq 1$. We will show the stronger claim that every eigenvalue of B is at least 1. Consider an arbitrary (eigenvector, eigenvalue)-pair of A, (e, λ). Then

$$Be = \frac{1}{4}(\lambda + 2 + \frac{1}{\lambda})e$$

Since $\frac{1}{4}(\lambda + 2 + \frac{1}{\lambda}) \geq 1$, we have that e is an eigenvector of eigenvalue at least 1 for B (this used that $\lambda \geq 0$, which is true since A is positive definite). Since the eigenvectors of A form an orthonormal basis of the whole space, all of B's eigenvectors are also eigenvectors of A. ■

Fact 2 *For positive definite X_i and $\sum \lambda'_i = 1, \lambda'_i \geq 0$,*

$$Det(\sum_i \lambda'_i X_i) \geq \prod_i Det(X_i)^{\lambda'_i}$$

Proof: This is a straightforward generalization of fact 1.

Suppose first that for each i, λ'_i is exactly equal to $p_i/2^k$ for some integer p_i. In this case, we may apply fact 1 iteratively to find

$$Det(\sum_{i=1}^{2^k} X'_j) \geq \prod_{i=1}^{2^k} Det(X'_j)^{(1/2^k)}$$

Equating p_i of the $\{X'_j\}$ to X_i for each i, we recover fact 2 exactly. For general $\{\lambda'_i\}$, we have that the theorem must hold for any k-bit binary approximation to the λ'_i; fact 2 then follows from standard continuity arguments. ■

References

[BG 97] C. Becker and U. Gather, "The Maximum Asymptotic Bias of Outlier Identifiers," Technical Report TR1998/03, University of Dortmund.

[DG 93] P. Davies and U. Gather, "The identification of multiple outliers," In *Journal of the American Statistical Association*, 88, 1993, p. 782-801.

[BFKV 99] 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.

[DV 01] J. Dunagan and S. Vempala, "Optimal Outlier Removal in High-Dimensional Spaces," In *Proceedings of the 33rd ACM Symposium on the Theory of Computing (STOC '01)*, Crete, 2001, pp627-636.

[LKS 95] L. Lovász, R. Kannan and M. Simonovits, "Isoperimetric problems for convex bodies and a localization lemma," In *Discrete Computational Geometry* 13, 1995, pp541-559.

[LKS 97] 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(1), 1997, pp1-50.

[MY] R. Maronna and V. Yohai, "The behaviour of the Stahel-Donoho robust multivariate estimator," In *Journal of the American Statistical Association* 90(429), pp330-341, 1995.

[CEMST 93] K. L. Clarkson, D. Eppstein, G. L. Miller, C. Sturtivant, and S. Teng, "Approximating center points with iterated radon points," In *Proceedings of the 9th ACM Symposium on Computational Geometry (SOCG '93)*, San Diego, CA, 1993, pp91-98. To appear in *International Journal of Computational Geometry & Applications.*

[DG 92] D. L. Donoho and M. Gasko, "Breakdown properties of location estimates based on halfspace depth and projected outlyingness," In *The Annals of Statistics*, 20(4), 1992, pp1803-1827.

11 An Implementation

Let `X` be an $m \times n$ matrix whose rows are the points of our distribution. Let `m, n, beta, epsilon` be the values for m, n, β, ϵ, and let the boolean variable `done` indicate whether we are finished removing outliers. A complete implementation is given by the following matlab code:

```
%% requires X,m,epsilon,beta
done = 0
while(~done)
  done = 1
  M = cov(X) %% M is the covariance matrix of X
  Y = X/cholinc(sparse(M),'inf') %% Y is the 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
```

As of the Spring of 2002, a java applet illustrating the outlier removal algorithm is available at
`http://theory.lcs.mit.edu/~jdunagan/`