Analysis of Incomplete Data and an Intrinsic-Dimension Helly Theorem

Jie Gao * Michael Langberg † Leonard J. Schulman†

Abstract

The analysis of incomplete data is a long-standing challenge in practical statistics. When, as is typical, data objects are represented by points in $\mathbb{R}^d$, incomplete data objects correspond to affine subspaces (lines or Δ-flats). With this motivation we study the problem of finding the *minimum intersection radius* $r(\mathcal{L})$ of a set of lines or Δ-flats $\mathcal{L}$: the least r such that there is a ball of radius r intersecting every flat in $\mathcal{L}$. Known algorithms for finding the minimum enclosing ball for a point set (or clustering by several balls) do not easily extend to higher-dimensional flats, primarily because "distances" between flats do not satisfy the triangle inequality. In this paper we show how to restore geometry (i.e., a substitute for the triangle inequality) to the problem, through a new analog of Helly's theorem. This "intrinsic-dimension" Helly theorem states: for any family $\mathcal{L}$ of Δ-dimensional convex sets in a Hilbert space, there exist $\Delta + 2$ sets $\mathcal{L}' \subseteq \mathcal{L}$ such that $r(\mathcal{L}) \leq 2r(\mathcal{L}')$. Based upon this we present an algorithm that computes a $(1+\varepsilon)$-core set $\mathcal{L}' \subseteq \mathcal{L}$, $|\mathcal{L}'| = O(\Delta^4/\varepsilon^2)$, such that the ball centered at a point c with radius $(1+\varepsilon)r(\mathcal{L}')$ intersects every element of $\mathcal{L}$. The running time of the algorithm is $O(n^{\Delta+1} d \operatorname{poly}(1/\varepsilon))$. For the case of lines or line segments ($\Delta = 1$), the (expected) running time of the algorithm can be improved to $O(nd \operatorname{poly}(1/\varepsilon))$. We note that the size of the core set depends only on the dimension of the input objects and is independent of the input size n and the dimension d of the ambient space.

1 Introduction

One of the great challenges in computational theory is the extraction of patterns from massive and high-dimensional data sets. A common difficulty associated with such data sets is that entries are incomplete—a few questions are left blank on a questionnaire; weather records for a region omit the figures for one weather station for a short period because of a malfunction; stock exchange data is absent for one stock on one day because of a trading suspension; and so forth. How should we process the partial data? Statisticians approach this problem in a variety of ways: deleting incomplete entries; filling in incomplete entries based on the most similar complete entry ("hot deck imputation"); filling in incomplete entries with the sample mean ("mean substitution"); or using a learning algorithm or criterion (EM, max likelihood) to infer a missing entry [15]. All of these are attempts to cope concurrently with two difficulties: (1) The statistical relationship between the present and missing data is usually not known. This precludes a universal answer to which approach is most statistically sound. (2) There is a combinatorial explosion inherent in trying out all candidate assignments to the missing values. The present paper offers a new approach to the problem of incomplete data: an approach rooted in the geometry of the data set.

From the computational point of view, a data item with d representative features is typically represented by a point in $\mathbb{R}^d$, each dimension corresponding to a feature; frequently one obtains good results by approximating the similarity of two items by their Euclidean distance, after choosing a good scaling of the axes. The most elementary form of data analysis for such a data set is to find the smallest ball that approximates the data set, whether in terms of the sum of distances to the center of the ball, maximum distance to the center, etc. An immediate generalization is to find a small number of balls (a "clustering" of the data) which between them cover the points (different interpretations of "cover" lead to the well-known k-median problem, k-center problem, etc.). There is copious work on these problems in the machine learning and algorithms literature.

A data item that is lacking information about one or more features corresponds to a line or a flat in $\mathbb{R}^d$, whose dimension is the number of missing features. There is no difficulty in assigning a distance between two such flats; it is simply the distance between the nearest points on the flats. So we can seek to cluster the flats so as to minimize some objective function. Right away there is a major difficulty: *"distances" between flats do not satisfy the triangle inequality.* The problem is not that the triangle inequality is slightly violated, but that *no* relaxation of it holds. No matter how far apart lines a and c are, there is always a line b that intersects both. This problem defeats many existing algorithmic approaches for "clustering"-type tasks, and for good reason—the geometry seems, in a genuine sense, to be absent. What does it mean to cluster lines, if "a resembles b" and "b resembles c" imply nothing at all about a resembling c?

*Department of Computer Science, Stony Brook University, Stony Brook, NY, 11794. Email: jgao@cs.sunysb.edu. Work was done when the author was with Center for the Mathematics of Information, California Institute of Technology.

†Department of Computer Science, California Institute of Technology, Pasadena, CA 91125. Email: {mikel,schulman}@caltech.edu. M. Langberg is supported in part by NSF grant CCF-0346991. L. Schulman is supported in part by an NSF ITR and the Okawa Foundation.

In this paper we initiate work on data analysis for convex sets of low dimension inside an ambient space of possibly high dimension. Specifically, we assume each input object is a convex subset of a flat of dimension Δ within a Hilbert space. For the existence theorems, this space may be infinite-dimensional, while for the algorithmic statements, we take it to be $\mathbb{R}^d$ for a value of d that we shall consider to be much higher than Δ. Our measure of the similarity amongst a collection of convex sets $\mathcal{L}$ is the *minimum intersection radius* $r(\mathcal{L})$, the least r such that there is a ball of radius r intersecting every flat in $\mathcal{L}$. The center of the optimum ball (selected arbitrarily in the degenerate case that it is not unique) is termed the *minimum intersection center* and denoted $c(\mathcal{L})$. Intuitively this center is the best explanation of the incomplete input data; the minimum intersection radius captures how well this center fits the data, in the sense that every incomplete data item can be *completed* to a point within that distance of the center. Significantly, therefore, our model makes a functional prediction for reconstructing missing data: among all points of the flat, use that which is closest to $c(\mathcal{L})$. Thus in addition to the role of our model in learning aggregate properties of the data set, it also provides an inference mechanism about the missing features of individual records.

Our approach can also be described under the shape fitting framework [1], where one asks for a shape, a point in our case, that best fits the input set, lines or flats, under some criterion.

The core of our contribution is to show a way to restore geometry to the problem of analyzing incomplete data, in spite of the failure of the triangle inequality. This restoration goes through a variant of Helly's theorem. Suppose we blow up each line or flat ℓ to a cylinder or a slab that encloses all the points within distance r from ℓ. Helly's theorem says that if every $d+1$ of these have a common intersection, then all of them have a common intersection. In other words, Helly's theorem restores geometry because if every subcollection of $d+1$ out of the n "data flats" are within distance r of some "explanation point", then all of the n lines are within distance r of an explanation point.

As it stands, however, this chain of reasoning is too weak. The dimension of the ambient space, d, is typically of the order of hundreds or thousands, much larger than the maximum dimension Δ of the individual data items. We redress this gap by developing a version of Helly's theorem that takes into account the low dimension of the sets involved. Beginning with the case of lines ($\Delta = 1$), we show that if every 3 of the n "data lines" are within distance r of some "explanation point" then all of the n lines are within distance $2r$ of some explanation point. Notice that we are now free of the "extrinsic" dimension of the ambient space, and depend only on the intrinsic dimensionality, $\Delta = 1$, of the data sets. This result can be extended to any Δ-dimensional convex objects $\mathcal{L}$, for any $0 \leq \Delta \leq d$, as follows: if every subset of $\Delta+2$ convex objects of dimension at most Δ in a Hilbert space are within distance r of some point, then all of the objects are within distance $2r$ of some point. This result is optimal in the sense that there exist configurations in which any $\Delta + 1$ convex objects of dimension at most Δ in $\mathbb{R}^d$ have a minimum intersection radius that is strictly smaller than $1/2$ of that of $\mathcal{L}$. We call this result the intrinsic-dimension Helly theorem:

THEOREM 1.1. (INTRINSIC-DIMENSION HELLY THEOREM) *For any n convex sets of dimension at most Δ in a Hilbert space, $\mathcal{L} = \{\ell_1, \ell_2, \cdots, \ell_n\}$, there exist $\Delta + 2$ sets $\mathcal{L}' \subset \mathcal{L}$ such that $r(\mathcal{L}) \leq 2r(\mathcal{L}')$.*

Note that when $r = 0$ (*i.e.*, when the sets of $\mathcal{L}$ intersect), Theorem 1.1 directly generalizes Helly's theorem (except that it is weaker by 1 in the case that the ambient space is of finite dimension d, and $\Delta = d$). The implications of the theorem to the analysis of incomplete data are immediate: given a collection $\mathcal{L}$ of n convex sets of dimension at most Δ, a 2-approximation of the minimum intersection radius of $\mathcal{L}$ results from enumerating all subsets $\mathcal{L}'$ of $\mathcal{L}$ of size $\Delta + 2$ and determining the largest $r(\mathcal{L}')$. Actually as will be seen later, the implicit $O(n^{\Delta+2} d \operatorname{poly}(\Delta))$-runtime can be replaced by an algorithm with expected running time $2^{O(\Delta \log \Delta)} nd$ and a suitable center for $\mathcal{L}$ (not generally equal to $c(\mathcal{L}')$) is identified as part of the same process.

Next, we provide a method to achieve an approximation ratio of $1 + \varepsilon$ (for any $\varepsilon > 0$) for the minimum intersection radius. A subset $\mathcal{L}' \subseteq \mathcal{L}$ is said to be an α-core set, with respect to $r(\mathcal{L})$, if the minimum intersection radius $r(\mathcal{L}')$ of $\mathcal{L}'$ approximates $r(\mathcal{L})$ within a multiplicative factor of α. Theorem 1.1 says that when $\mathcal{L}$ is a set of lines in d-dimensional Euclidian space, one can find a 2-core set $\mathcal{L}'$ of size 3; and in general, if $\mathcal{L}$ consists of Δ-dimensional convex sets, there exists a 2-core set $\mathcal{L}'$ of size $\Delta + 2$. For general values of $\alpha = 1 + \varepsilon$ and for $\mathcal{L}$ consisting of Δ-dimensional flats in $\mathbb{R}^d$, we show that for any $\varepsilon > 0$ there exists a $(1 + \varepsilon)$-core set of size $O(\Delta^4/\varepsilon^2)$. Here and throughout the paper we assume that each convex set can be represented by a constant number of constraints. Such a core set can be found in time $O(n^{\Delta+1} d \operatorname{poly}(1/\varepsilon))$. For the case of lines ($\Delta = 1$), the running time of the algorithm can be improved to $O(nd \operatorname{poly}(1/\varepsilon))$. Notice that the size of the $(1 + \varepsilon)$-core set only depends on ε and Δ, and is independent of the total input n or the dimension of the ambient space.

To the best of our knowledge, this is the first work to address core sets for collections $\mathcal{L}$ that consist of Δ-dimensional convex sets. We summarize the core set result by the following two theorems:

THEOREM 1.2. ($(1+\varepsilon)$-CORE SET FOR LINE SEGMENTS) *Let $\varepsilon > 0$. Let $\mathcal{L}$ be a set of lines or line segments*

$\{\ell_1, \ldots, \ell_n\}$ *in* $\mathbb{R}^d$. *There exist a subset* $\mathcal{L}' \subseteq \mathcal{L}$ *of size* $O(1/\varepsilon^2)$ *such that* $r(\mathcal{L}')(1+\varepsilon) \geq r(\mathcal{L})$. *The set* $\mathcal{L}'$ *and a center* c, *such that the ball centered at* c *of radius* $(1+\varepsilon)r(\mathcal{L}')$ *intersects all lines or line segments in* $\mathcal{L}$, *can be found in expected time* $O(nd\,\mathrm{poly}(1/\varepsilon))$.

THEOREM 1.3. ($(1+\varepsilon)$-CORE SETS) *Let* $\varepsilon > 0$. *Let* $\mathcal{L}$ *be a set of convex sets of dimension* $\leq \Delta$, $\mathcal{L} = \{\ell_1, \ldots, \ell_n\}$, *in* $\mathbb{R}^d$. *There exist a subset* $\mathcal{L}' \subseteq \mathcal{L}$ *of size* $O(\Delta^4/\varepsilon^2)$ *such that* $r(\mathcal{L}')(1+\varepsilon) \geq r(\mathcal{L})$. *The set* $\mathcal{L}'$ *and a center* c, *such that the ball centered at* c *of radius* $(1+\varepsilon)r(\mathcal{L}')$ *intersects all sets in* $\mathcal{L}$, *can be found in time* $O(n^{\Delta+1}d\,\mathrm{poly}(1/\varepsilon))$.

As described above, the main focus of this paper is the near-optimal representation of a set of incomplete data entries (Δ-flats) by a single ball of minimal radius. Naturally, this is only the first step toward a more comprehensive theory that should provide algorithms for *clustering* incomplete data entries by providing several balls of small radius, at least one of which intersects each of the Δ-flats. It is easy to see that minimizing this radius is NP-hard, from the NP-hardness of the k-center problem for points. Our work already implies an initial result in this area: using our core set method, there is a straightforward $\widetilde{O}(n^k)$-time algorithm to obtain a 2-approximate k-clustering of n lines. Due to space limitations, the details are omitted from this extended abstract and will appear in the full version of the paper.

1.1 Related work Clustering and shape fitting problems on points have been actively studied in recent years. One of the powerful techniques is to devise a core set, i.e., a small subset of representative points S' of S such that the optimization problems on S' is a good approximation to the optimal solution on S [1]. Precisely, a subset S' is a $(1+\varepsilon)$-core set of S if $(1+\varepsilon)\mu(S') \geq \mu(S)$, where μ is a monotonic measure function. Agarwal *et al.* provided a framework for computing a $(1+\varepsilon)$-core set for a set of points S in $\mathbb{R}^d$ with respect to many measure functions that depend on the extent of the point set, such as diameter, width, radius of the minimum enclosing ball, and volume of the smallest enclosing box [2]. The basic idea is to find a subset of points of size $O(1/\varepsilon^{O(d)})$ whose convex hull approximates the convex hull of S. For some of the problems such as the minimum enclosing ball or ellipsoid, there is an incremental algorithm that computes a $(1+\varepsilon)$-core set of size that depends only on ε [8, 7, 13, 14]. Thus one can apply brute-forth algorithms on the small core set S' and obtain efficient approximation algorithms for the optimization problems on S. Indeed, many geometric optimization problems such as minimum enclosing ball, k-clustering, and various shape fitting problems can be solved efficiently by using a small core set [4, 8, 9, 10, 11, 13, 14]. However, to the best of our knowledge, no work has been done on devising a core set for lines or flats with respect to a natural quality measure.

The study of core sets for points can not be directly applied to core sets for lines or flats. For a set of lines or flats, there is no natural definition of "convex hull". Our core set algorithms for lines or flats are more related with the incremental core set algorithm for points S in $\mathbb{R}^d$ with respect to the radius of the minimum enclosing ball [8], which is described as follows. The algorithm starts with S' being a pair of furthest away points and computes the minimum enclosing ball of S'. If all the points are included in the minimum enclosing ball enlarged by a factor of $(1+\varepsilon)$, then S' is a core set. Otherwise, a point outside the enlarged ball is added to S'. It can be shown that for each step, the radius of the minimum enclosing ball of S' is increased by a factor of $1 + O(\varepsilon^2)$. After $O(1/\varepsilon^2)$ steps, the algorithm terminates. However, there is a major difficulty to apply this algorithm for a set of lines $\mathcal{L}$: there is a situation where by adding each extra line, the minimum intersection radius of the current subset $\mathcal{L}'$ stays the same but the minimum intersection radius of $\mathcal{L}'$, $r(\mathcal{L}')$, is still far away from the real value $r(\mathcal{L})$. A substantial part of this paper is devoted to showing that a carefully selected set of two lines (or $\Delta+1$, more generally, Δ-flats) can improve the minimum intersection radius substantially.

We also note that there has been work on "clustering points with lines" [3, 5, 10], where one finds a set of lines $\mathcal{L}$ such that the set of cylinders with radius r and axis as the lines of $\mathcal{L}$ covers all the input points S. The problem we study in this paper can be phrased as "clustering lines with a point". There does not exist an obvious connection between these two problems as a natural duality does not exist.

1.2 Organization The remainder of the paper is organized as follows. We start with a few preliminaries in Section 2. In Section 3 we present the proof of Theorem 1.1. In Section 4 we present the proof of Theorem 1.2. The proof of Theorem 1.3 is very similar to that of Theorem 1.2, and is sketched in Section 5. Due to space limitations the proof of some of our claims are omitted.

2 Preliminaries, definitions and notation

We denote by $B_r(c)$ a ball centered at a center c with radius r in $\mathbb{R}^d$. We denote by $d(.,.)$ the Euclidean distance function. The distance between two points p, q is also written as $|pq|$.

DEFINITION 2.1. *A* Δ*-flat in a Hilbert space is a* Δ*-dimensional affine subspace. The dimension of a convex set in a Hilbert space is the least dimension of any flat containing it.*

DEFINITION 2.2. *The* minimum intersection ball $B(\mathcal{L})$ *of a collection of convex sets* $\mathcal{L}$ *in a Hilbert space is defined to be (one of) the minimum radius balls that intersects all the sets. The center of the minimum intersection ball is called the* minimum intersection center, *denoted as* $c(\mathcal{L})$.

The radius of the minimum intersection ball is called the minimum intersection radius, *denoted as* $r(\mathcal{L})$. *Namely,* $B(\mathcal{L}) = B_{r(\mathcal{L})}(c(\mathcal{L}))$.

DEFINITION 2.3. ($(1+\varepsilon)$-CORE SET) *Let $\mathcal{L}$ be a set of convex sets in a Hilbert space. A subset $\mathcal{L}'$ of $\mathcal{L}$ is said to be a $(1+\varepsilon)$-core set w.r.t. the minimum intersection radius of $\mathcal{L}$ if $r(\mathcal{L}) \leq (1+\varepsilon)r(\mathcal{L}')$.*

We begin by noting that the minimum intersection radius and center of $\mathcal{L}$ can be found in polynomial time up to an absolute error δ using convex programming. The proof of the following lemma appears in the Appendix.

LEMMA 2.1. *Let $\mathcal{L}$ be a set of n convex sets with dimension at most Δ in $\mathbb{R}^d$. $c(\mathcal{L})$ and $r(\mathcal{L})$ can be computed to an absolute precision $\delta > 0$ in time $O(\sqrt{n}(d^3 + d^2 n\Delta)\log(n/\delta))$.*

3 Intrinsic-dimension Helly theorem

We now prove Theorem 1.1.

Proof. For each $(\Delta+2)$-tuple $\mathbf{i} = \{i_1, \ldots, i_{\Delta+2}\}$ in $\{1, \ldots, n\}$, let $B_{\mathbf{i}}$ be the minimum intersection ball of the subset $\mathcal{L}_{\mathbf{i}} = \{\ell_{i_1}, \ldots, \ell_{i_{\Delta+2}}\}$ centered at point $c_{\mathbf{i}}$, and let $r_{\mathbf{i}}$ be the radius of $B_{\mathbf{i}}$.

Let the largest radius among the $r_{\mathbf{i}}$'s be r. Now we claim that we can find a ball of radius $2r$ that intersects all the sets in $\mathcal{L}$. Consider the set ℓ_1 and denote by I_j the set of points on ℓ_1 with distance no more than $2r$ from ℓ_j. That is, $I_j = \{p \in \ell_1 | d(p, \ell_j) \leq 2r\}$. I_j is convex, since I_j is the intersection of two convex objects, the set ℓ_1 and the set of points of distance $2r$ from ℓ_j. Notice that I_j is a convex set of dimension Δ in ℓ_1.

Consider any $(\Delta+1)$-tuple $\mathbf{i}' = \{i_1, \ldots, i_{\Delta+1}\}$ in $\{2, \ldots, n\}$. Now we claim that the corresponding $\Delta+1$ sets $\{I_{i_1}, \ldots, I_{i_{\Delta+1}}\}$ have non-empty intersection. Let $\mathbf{i}$ be the $(\Delta+2)$-tuple obtained from $\mathbf{i}'$ by adding an additional index of value 1, namely $\mathbf{i} = \{1, i_1, \ldots, i_{\Delta+1}\}$. By the above discussion we have that $r_{\mathbf{i}} \leq r$. Let $c_{\mathbf{i}}$ be the center of the minimum intersection ball $B_{\mathbf{i}}$ of $\mathcal{L}_{\mathbf{i}}$. Let $c'_{\mathbf{i}}$ be the point on the set ℓ_1 that is closest to $c_{\mathbf{i}}$. As the point $c_{\mathbf{i}}$ is within distance r from all sets in $\mathcal{L}_{\mathbf{i}}$, we have that the point $c'_{\mathbf{i}}$ is within distance $2r$ from all the sets in $\mathcal{L}_{\mathbf{i}}$. This implies that $c'_{\mathbf{i}}$ is in I_j for all $j \in \mathbf{i}'$. See Figure 1.

Since for any $(\Delta+1)$-tuple $\mathbf{i}' = \{i_1, \ldots, i_{\Delta+1}\}$ the corresponding $\Delta+1$ convex sets I_j have non-empty intersection, and these sets are embedded in the Δ dimensional set ℓ_1, by Helly's Theorem [12] all the sets I_j, $2 \leq j \leq n$ have a non-empty intersection. Now let o be a point in $\bigcap_j I_j$, $2 \leq j \leq n$. The ball centered at o with radius $2r$ intersects all n sets of $\mathcal{L}$.

The case of $r = 0$ is closest in form to the original Helly theorem:

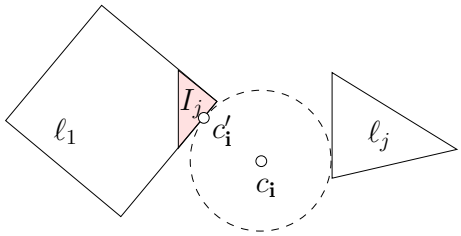

Figure 1: Proof of the reduction from the minimum intersection radius of n sets to that of $\Delta+2$ sets.

COROLLARY 3.1. *For any n convex sets of dimension at most Δ in a Hilbert space, $\mathcal{L} = \{\ell_1, \ell_2, \cdots, \ell_n\}$, if every $\Delta+2$ sets in $\mathcal{L}$ intersect, then all sets in $\mathcal{L}$ intersect.*

THEOREM 3.1. (OPTIMALITY) *For any Δ, there exist a set of n convex sets $\mathcal{L} = \{\ell_1, \ldots, \ell_n\}$ such that for every subset $\mathcal{L}' \subseteq \mathcal{L}$ of size less than $\Delta+2$ it holds that $r(\mathcal{L}) > 2r(\mathcal{L}')$.*

Proof. For any Δ, consider the $(\Delta+1)$-dimensional simplex. Namely, the set $\Omega_{\Delta+1} = \{(p_1, \ldots, p_{\Delta+2}) \mid \sum_{i=1}^{\Delta+2} p_i = 1\}$. Our sets ℓ_i will be subsets of $\Omega_{\Delta+1}$ of dimension Δ. $|\mathcal{L}| = \Delta+2$. For $i = 1$ to $\Delta+2$, define $\ell_i = \{(p_1, \ldots, p_{\Delta+2}) \in \Omega_{\Delta+1} \mid p_i = 0\}$. It is not hard to verify that every subset $\mathcal{L}' \subseteq \mathcal{L}$ of size $\Delta+1$ has a non-empty intersection — if $\ell_i \notin \mathcal{L}'$ then the unit vector with '1' in the i'th coordinate is in $\mathcal{L}'$. Thus $r(\mathcal{L}') = 0$. However, the sets in $\mathcal{L}$ do not have a common intersection. $r(\mathcal{L}) > 0$.

Since this construction lies in $\mathbb{R}^{\Delta+1}$, the only choice of parameters for which Theorem 1.1 is not tight is the obvious case of $\Delta = d$.

Theorem 1.1 indicates a straightforward algorithm to find a 2-approximation to the minimum intersection radius by simply taking the maximum radius of all $(\Delta+2)$-tuples of sets that include ℓ_1. One can improve the running time of this naive algorithm (and actually give an alternative proof to Theorem 1.1) using the methodology of 'LP-type' programming (e.g. [16]) to $2^{O(\Delta \log \Delta)} nd$. The proof of the following lemma is given in the Appendix.

LEMMA 3.1. *Let $\mathcal{L}$ be a set of n convex sets with dimension at most Δ in $\mathbb{R}^d$. Let $\ell \in \mathcal{L}$. The minimum radius ball $B_r(c)$ that covers the sets $\mathcal{L}$ with center c on ℓ, and a set $\mathcal{L}' = \{l_1, \ldots, l_{\Delta+2}\} \subseteq \mathcal{L}$ for which $r(\mathcal{L}) \leq 2r(\mathcal{L}')$ can be found in expected time $2^{O(\Delta \log \Delta)} nd$.*

4 $(1+\varepsilon)$-core set for lines or line segments

The proof of Theorem 1.2 we present shortly strongly builds upon the notion of a $(1+\varepsilon)$-*approximate intersection center*, $\varepsilon > 0$. In what follows we define $(1+\varepsilon)$-approximate intersection centers, and state Theorem 4.1 which addresses a

certain property of these centers. We then prove Theorem 1.2 based on Theorem 4.1. In what follows we will assume that $\mathcal{L}$ is a set of n lines. The general case in which $\mathcal{L}$ also includes line segments is proven in a very similar manner and is given in detail in the Appendix.

DEFINITION 4.1. *A $(1+\varepsilon)$-approximate intersection ball of a set $\mathcal{L}$ in d-dimensional Euclidean space is a ball of radius $(1+\varepsilon)r(\mathcal{L})$ that intersects all sets in $\mathcal{L}$. The center of a $(1+\varepsilon)$-approximate intersection ball is called a $(1+\varepsilon)$-approximate intersection center. We denote by $C_\varepsilon(\mathcal{L})$ the set of $(1+\varepsilon)$-approximate intersection centers of $\mathcal{L}$.*

OBSERVATION 4.1. *For any set $\mathcal{L}$, $C_\varepsilon(\mathcal{L})$ is convex.*

DEFINITION 4.2. *Let ℓ be a line in $\mathbb{R}^d$. A cylinder of radius r with axis ℓ in $\mathbb{R}^d$ is defined as the set of points in $\mathbb{R}^d$ which are of distance at most r from ℓ.*

THEOREM 4.1. *$C_\varepsilon(\mathcal{L})$ is included in a cylinder of radius $25\sqrt{\varepsilon}r(\mathcal{L})$ with axis parallel to one of the sets $\ell_i \in \mathcal{L}$. Moreover this axis passes through $c(\mathcal{L})$.*

The proof of Theorem 4.1 is based on Lemma 4.1 4.2 4.3. The proofs of the Lemmas and Theorem 4.1 are non-trivial and rather technical; they appear in the Appendix.

LEMMA 4.1. *Suppose c is a minimum intersection center of a set of lines $\mathcal{L}$ and p is a $(1+\varepsilon)$-approximate intersection center, $|cp| \geq \alpha\sqrt{\varepsilon}r(\mathcal{L})$, then there exists a line $\ell \in \mathcal{L}$ such that the angle between ℓ and the interval cp is bounded by $\arcsin(\sqrt{2+\varepsilon}/\alpha)$.*

LEMMA 4.2. *For a set of $\Delta+1$ orthogonal vectors $\{v_1, v_2, \cdots, v_{\Delta+1}\}$ and a Δ-flat ℓ in $\mathbb{R}^{\Delta+1}$, there must be a vector v_j such that the angle between v_j and ℓ is at least $\arcsin(1/\sqrt{\Delta+1})$.*

LEMMA 4.3. *$C_\varepsilon(\mathcal{L})$ does not include a 2-dimensional square with side length $5\sqrt{\varepsilon}r(\mathcal{L})$.*

With Theorem 4.1, we now prove Theorem 1.2. The idea is similar with the construction of a $(1+\varepsilon)$-core set P' for a set of points P in $\mathbb{R}^d$ such that the radius of the minimum enclosing ball of P is bounded by $(1+\varepsilon)$ times that of P' [8]. The basic idea in [8] is to add a point not covered by the minimum enclosing ball of the current core set such that the minimum radius is increased substantially. However, a direct application of this idea does not work for the case of lines. One can find a scenario where adding a line can not improve the minimum intersection radius. We will show that a careful selection of two lines can always increase the minimum intersection radius by a substantial factor.

Proof. [Theorem 1.2] Let $\varepsilon > 0$. Throughout this proof we assume that ε is sufficiently small. In what follows we present an algorithm for finding $\mathcal{L}'$. Our algorithm is greedy and strongly builds upon Theorem 4.1. For a set $\mathcal{L}'$, let ℓ' be the axis of the cylinder of radius $\frac{\varepsilon}{2}r(\mathcal{L}')$ which contains the collection of $(1+\varepsilon^2/50^2)$-approximate intersection centers of $\mathcal{L}'$.

Roughly speaking, the main idea of our algorithm is as follows. We start out by picking a subset of $\mathcal{L}'$ of size 3 according to Lemma 3.1. For these lines it holds that $\alpha r(\mathcal{L}') \geq r(\mathcal{L})$ where $\alpha = 2$. This is a good starting point, but we still need to reduce the value α above to $(1+\varepsilon)$. We do this in a series of steps. In each step, a line or two are added to $\mathcal{L}'$ and α reduces by a factor of $(1-\frac{1}{2}\varepsilon^2/50^2)$. Hence, after $O(1/\varepsilon^2)$ such steps we are in a situation in which $(1+\varepsilon)r(\mathcal{L}') \geq r(\mathcal{L})$ and we are done. The second part of the theorem (regarding efficiency matters) will follow from the detailed description of the algorithm.

We first focus on an iteration of the algorithm. Let $\mathcal{L}'$ be the subset defined by the algorithm so far. Let $c = c(\mathcal{L}')$ be the minimum intersection center of $\mathcal{L}'$, and let ℓ' be the axis of the cylinder of radius $\frac{\varepsilon}{2}r(\mathcal{L}')$ which contains the collection of $(1+\varepsilon^2/50^2)$-approximate intersection centers of $\mathcal{L}'$. Recall from Theorem 4.1 that ℓ' passes through c and is parallel to one of the sets in $\mathcal{L}'$.

If the ball centered at c of radius $(1+\varepsilon)r(\mathcal{L}')$ intersects $\mathcal{L}$, halt and output the set $\mathcal{L}'$. Otherwise, if there exists a line $\ell \in \mathcal{L}$ such that $r(\mathcal{L}' \cup \{\ell\}) \geq (1+\varepsilon^2/50^2)r(\mathcal{L}')$ add ℓ to $\mathcal{L}'$ and proceed in an additional iteration of the algorithm (see remark at the end of the proof).

We are now in a situation that for every line $\ell \in \mathcal{L}$ the radius of the minimum intersection ball of $\mathcal{L}' \cup \{\ell\}$ is very close to the radius of the minimum intersection ball of $\mathcal{L}'$, namely, $r(\mathcal{L}' \cup \{\ell\}) < (1+\varepsilon^2/50^2)r(\mathcal{L}')$, this implies that the center of the minimum intersection ball of $\mathcal{L}' \cup \{\ell\}$ is in the $\frac{\varepsilon}{2}r(\mathcal{L}')$ cylinder around ℓ'. In this case, we use the axis ℓ' to find a *pair* of lines that when added to $\mathcal{L}'$ will increase $r(\mathcal{L}')$ substantially. For each line $\ell_i \in \mathcal{L} \setminus \mathcal{L}'$ we now compute a certain interval I_i on ℓ'. Namely, we define I_i to be the set of points x on ℓ' such that the ball of radius $(1+\varepsilon)r(\mathcal{L}')$ centered at x intersects the sets in $\mathcal{L}' \cup \{\ell_i\}$. It is not hard to verify that for each line ℓ_i this interval is not empty. Indeed, consider the minimum intersection center c_i^* of $\mathcal{L}' \cup \{\ell_i\}$. As $r(\mathcal{L}' \cup \{\ell\}) < (1+\varepsilon^2/50^2)r(\mathcal{L}')$, i.e., c_i^* is a $(1+\varepsilon^2/50^2)$-approximate center of $\mathcal{L}'$, it follows from Theorem 4.1 that the distance of c_i^* from ℓ' is at most $\frac{\varepsilon}{2}r(\mathcal{L}')$. Consider the projection of c_i^* onto the line ℓ'. Denote this projection by c_i'. It now follows that the ball of radius $(1+\varepsilon/2+\varepsilon^2/50^2)r(\mathcal{L}') \leq (1+\varepsilon)r(\mathcal{L}')$ centered at c_i' covers $\mathcal{L}' \cup \{\ell_i\}$, which implies that $c_i' \in I_i$.

If for all pairs of lines ℓ_i and ℓ_j the corresponding intervals intersect, then by Helly theorem, there is a point c' in the intersection of all the intervals. This implies that the

ball of radius $(1+\varepsilon)r(\mathcal{L}')$ centered at c' covers all the lines and we may halt the algorithm and output the set $\mathcal{L}'$.

Finally, if there are two lines ℓ_i and ℓ_j with corresponding intervals that do not intersect, then we claim that $r(\mathcal{L}' \cup \{\ell_i, \ell_j\}) \geq (1+\varepsilon^2/50^2)r(\mathcal{L}')$ and we may add both ℓ_i and ℓ_j to $\mathcal{L}'$ and proceed in an additional iteration of the algorithm (see remark at end of proof). Assume for contradiction that $r(\mathcal{L}' \cup \{\ell_i, \ell_j\}) < (1+\varepsilon^2/50^2)r(\mathcal{L}')$ and let c^* be the minimum intersection center of $\mathcal{L}' \cup \{\ell_i, \ell_j\}$. As the ball of radius $(1+\varepsilon^2/50^2)r(\mathcal{L}')$ centered at c^* also covers $\mathcal{L}'$ it follows that the distance of c^* from ℓ' is at most $\frac{\varepsilon}{2}r(\mathcal{L}')$. As before, consider the projection of c^* onto the line ℓ'. Denote this projection by c'. It now follows that the ball of radius $(1+\frac{\varepsilon}{2}+\varepsilon^2/50^2)r(\mathcal{L}') \leq (1+\varepsilon)r(\mathcal{L}')$ centered at c' covers $\mathcal{L}' \cup \{\ell_i, \ell_j\}$. This implies that the point c' is in the intervals I_i and I_j corresponding to ℓ_i and ℓ_j, which is a contradiction.

We still need to show, given a new set $\mathcal{L}'$ how to find $c(\mathcal{L}')$, $r(\mathcal{L}')$ and the axis ℓ' of $C_{\varepsilon^2/50^2}(\mathcal{L}')$. Computing $c(\mathcal{L}')$ and $r(\mathcal{L}')$ can be done (with sufficient precision) in time $d\,\mathrm{poly}(1/\varepsilon)$ using Lemma 2.1 (here we use the fact that the total dimension of the lines involved in the computation is independent of d). Regarding the line ℓ', by Theorem 4.1 we know that ℓ' is parallel to one of the lines in $\mathcal{L}'$ and passes through $c = c(\mathcal{L}')$. In the upcoming iteration of our algorithm, we may try each and every line in $\mathcal{L}'$ (a constant number) and run the additional iteration with that center. As we halt our algorithm, or proceed to add lines to $\mathcal{L}'$ only if certain conditions hold. We are sure to encounter these conditions once we have chosen the correct line ℓ'.

5 $(1+\varepsilon)$-core set for convex sets of dimension $\leq \Delta$

The algorithm in the previous section can be extended to the general case of convex sets with dimension at most Δ, resulting in the proof of Theorem 1.3. As in the previous section, the proof of Theorem 1.3 uses the notion of a $(1+\varepsilon)$-approximate intersection center. In what follows we state Definition 5.1 and Theorem 5.1. The proofs can be found in the Appendix.

DEFINITION 5.1. *Let ℓ be a convex set with dimension $\leq \Delta$ in $\mathbb{R}^d$. A slab of radius r with axis ℓ in $\mathbb{R}^d$ is defined as the set of points in $\mathbb{R}^d$ which are of distance at most r from ℓ.*

THEOREM 5.1. *The set of $(1+\varepsilon)$-approximate intersection centers of a collection $\mathcal{L}$ of convex sets with dimension at most Δ, $C_\varepsilon(\mathcal{L})$, is included in a Δ-slab of width $\beta\sqrt{\varepsilon(\Delta+1)^3}r(\mathcal{L})$, for some constant β.*

Acknowledgements

The authors would like to thank the anonymous SODA referee for the useful comments on LP-type problems, and Lin Xiao for several helpful discussions on convex programming.

References

[1] P. Agarwal, S. Har-Peled, and K. R. Varadarajan. Geometric approximation via coresets. In *Current Trends in Combinatorial and Computational Geometry*. Cambridge University Press, 2005.

[2] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, 2004.

[3] P. K. Agarwal and C. M. Procopiuc. Approximation algorithms for projective clustering. In *SODA '00: Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*, pages 538–547, Philadelphia, PA, USA, 2000. Society for Industrial and Applied Mathematics.

[4] P. K. Agarwal, C. M. Procopiuc, and K. R. Varadarajan. Approximation algorithms for k-line center. In *ESA '02: Proceedings of the 10th Annual European Symposium on Algorithms*, pages 54–63, London, UK, 2002. Springer-Verlag.

[5] P. K. Agarwal, C. M. Procopiuc, and K. R. Varadarajan. A $(1+\varepsilon)$-approximation algorithm for 2-line-center. *Comput. Geom. Theory Appl.*, 26(2):119–128, 2003.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2003.

[7] M. Bădoiu and K. L. Clarkson. Smaller core-sets for balls. In *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 801–802, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

[8] M. Bădoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 250–257, New York, NY, USA, 2002. ACM Press.

[9] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300, New York, NY, USA, 2004. ACM Press.

[10] S. Har-Peled and K. Varadarajan. Projective clustering in high dimensions using core-sets. In *SCG '02: Proceedings of the eighteenth annual symposium on Computational geometry*, pages 312–318, New York, NY, USA, 2002. ACM Press.

[11] S. Har-Peled and Y. Wang. Shape fitting with outliers. *SIAM J. Computing*, 33:269–285, 2003.

[12] E. Helly. Über Mengen konvexer Körper mit gemeinschaftlichen Punkten. *Jahresbericht Deutsch. Math. Verein.*, 32:175–176, 1923.

[13] P. Kumar, J. S. B. Mitchell, and E. A. Yildirim. Approximate minimum enclosing balls in high dimensions using core-sets. *J. Exp. Algorithmics*, 8:1.1, 2003.

[14] P. Kumar and E. A. Yildirim. Approximating minimum volume enclosing ellipsoids using core sets. *J. Opt. Theo. Appl.*, 2004. to appear.

[15] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[16] J. Matoušek, M. Sharir, and E. Welzl. A subexponential bound for linear programming. *Algorithmica*, 16:498–516, 1996.

6 Appendix

Proof. [Lemma 2.1] We can formulate this problem by convex programming. Assume c is a point in $\mathbb{R}^d$ and r is the intersection radius of c with respect to $\mathcal{L}$. Each convex set ℓ_i has dimension at most Δ and is represented as follows. ℓ_i is in a Δ-dimensional space B_i, which has origin $o^{(i)}$ and is spanned by k unit vectors $\vec{b}_j^{(i)}$, $j = 1, \cdots, \Delta$. Therefore, each point in ℓ_i can be represented by $o^{(i)} + \sum_{j=1}^{\Delta} \lambda_j^{(i)} \vec{b}_j^{(i)}$, where $\lambda_j^{(i)}$ is a scalar. Inside B_i, the convex set ℓ_i is specified by m_i convex constraints $f_j^{(i)}(\lambda^{(i)}) \leq 0$, $j = 1, \cdots, m_i$, where $\lambda^{(i)} = \{\lambda_j^{(i)}\}$. We can find the minimum intersection radius and center of $\mathcal{L}$ by solving the following optimization problem:

$$\begin{array}{ll} & \min r \\ s.t. & r \geq ||c - o^{(i)} - \sum_{j=1}^{\Delta} \lambda_j^{(i)} \vec{b}_j^{(i)}||\,, 1 \leq i \leq n\,; \\ & f_j^{(i)}(\lambda^{(i)}) \leq 0\,, 1 \leq j \leq m_i, 1 \leq i \leq n\,. \end{array}$$

This is a convex optimization problem. The total number of variables is $n\Delta + d + 1$. The total number of constraints is $N + n$, where $N = \sum_i m_i = O(n)$. Thus one can find the solution up to precision $\delta > 0$ in time $O((N + n)(n\Delta + d)^3 \log(n/\delta))$, by using a generic interior point method [6]. A more careful analysis by exploring the sparsity of matrices shows a better bound on the running time $O(\sqrt{n}(d^3 + d^2 n k) \log(n/\delta))$. The details are omitted here.

Proof. [Lemma 3.1] Let $\mathcal{L}$ be a set of n convex sets of dimension Δ in $\mathbb{R}^d$. Let $\ell \in \mathcal{L}$. In what follows we study the problem of finding the minimum radius ball covering $\mathcal{L}$ with center on ℓ. We show that this problem falls in the abstract framework of so called 'LP-type' problems, and can be solved by the randomized algorithm of [16] in expected running time $2^{O(\Delta \log \Delta)} nd$. The algorithm of [16] not only finds the minimum radius ball (say $B_r(c)$ centered at c of radius r) covering $\mathcal{L}$ with center on ℓ, it also returns a subset $\mathcal{L}'$ of $\mathcal{L}$ of size $\Delta + 1$ such that the minimum radius ball covering $\mathcal{L}'$ with center on ℓ is also $B_r(c)$. This implies that $r(\mathcal{L}) \leq 2r(\mathcal{L}' \cup \ell)$. Indeed, $r \geq r(\mathcal{L})$, and $r(\mathcal{L}' \cup \ell) \geq \frac{r(\mathcal{L})}{2}$ otherwise by projecting onto ℓ and using the triangle inequality one could find a ball of radius less than r that covers $\mathcal{L}'$ with center on ℓ.

We now sketch the proof that the problem at hand is an LP-type problem. Throughout our proof we use the notation of [16] freely. Our problem is defined by a couple (H, w) where H is the set of constraints corresponding to each set in $\mathcal{L}$, and w is the function on subsets G of H which returns the minimum radius ball covering sets corresponding to G with center on ℓ (ties broken using a lexicographic order on ℓ). For various technical reasons we alter w as to satisfy *basis regularity* (as described in [16]). To use the framework outlined in [16] it suffices to prove the following claims. Detailed proof is omitted form this extended abstract and will appear in the full version of the paper. Roughly speaking the proofs follow the line of proof used in proving that the minimum enclosing ball of a set of n points is an LP-type problem.

Claim 1: the *combinatorial dimension* of (H, w) is $\Delta + 1$. This follows essentially from the argument that in the Δ dimensional flat ℓ the minimum enclosing ball of n points has an exact coreset of size $\Delta + 1$. **Claim 2**: *Monotonicity* and *Locality*, follow by the definition of w. **Claim 3**: *Violation test* and *Basis Computation*. We present an algorithm which given a set of constraints G of size $\leq \Delta + 2$, finds the value of $w(G)$ along with a *basis* for G. Here we use the fact that finding the minimum covering ball of k convex sets in k^2 dimensions can be done in time $exp(O(k \log k))$ (see for example [16]), we also need a few additional ideas that tie this problem with our basis computation problem.

Proof. [Lemma 4.1] We take a point t on cp that moves infinitesimally away from c towards p. We argue that there must be a line $\ell \in \mathcal{L}$ such that $d(c, \ell) = r(\mathcal{L})$ and the distance from a point t on cp to ℓ is non-decreasing, when t moves infinitesimally from c to p on cp. Otherwise, the distances from t to all the lines in $\mathcal{L}$ are strictly less than $r(\mathcal{L})$ when t moves infinitesimally away from c. This contradicts with the fact that $r(\mathcal{L})$ is the minimum intersection radius.

Now suppose for the line $\ell \in \mathcal{L}$, the distance from t to ℓ stays the same, when t moves infinitesimally away from c. Then it must be that cp is parallel with ℓ. Therefore the angle between ℓ and cp is zero. The claim is true.

If the distance from t to ℓ is strictly increasing when t moves infinitesimally away from c, the distance $d(t, \ell)$ is monotonically increasing as t moves linearly from c to p. Now we bound the angle between cp and ℓ as follows. The point p is a $(1 + \varepsilon)$-approximate intersection center, $r(\mathcal{L}) < d(p, \ell) \leq (1+\varepsilon) r(\mathcal{L})$. We denote by ℓ' the line that is parallel with ℓ and goes through center c. Let q be the point on ℓ for which pq perpendicular to ℓ, $|pq| = d(p, \ell) \leq (1 + \varepsilon) \cdot r(\mathcal{L})$. Let q' be the point on ℓ' for which qq' is perpendicular to ℓ', $|qq'| = d(c, \ell) = r(\mathcal{L})$. Finally, let s be the point on ℓ for which cs is perpendicular to ℓ, $|cs| = d(c, \ell) = r(\mathcal{L})$. See Figure 2.

Let H be the hyperplane with normal cs that passes through c, and let S be the (closed) half space defined by H that does not include ℓ. The segment cp (not including c) is either entirely contained in S or entirely contained in the complement of S. We now claim the cp is in S. Consider the ball B of radius $r(\mathcal{L})$ around s. The point c is on the boundary of both B and S. And the segment cp does not intersect the interior of B. Otherwise there would be a point on cp of distance less than $r(\mathcal{L})$ to s. This can only happen if cp is in S.

If p is in S, the inner angle of the triangle $\triangle pq'q$ at

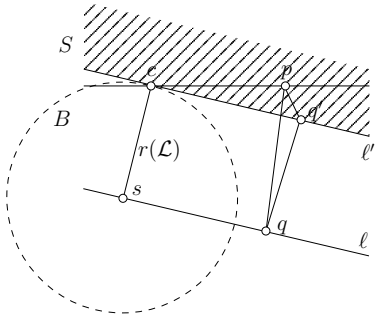

Figure 2: If the distance between an approximate center p to the minimum center c is more than $\alpha\sqrt{\varepsilon}\cdot r(\mathcal{L})$, then there exists a line whose angle to line cp is bounded by $\arcsin(\sqrt{2+\varepsilon}/\alpha)$.

vertex q', $\angle pq'q \geq \pi/2$. Thus,

$$\begin{aligned}|pq'| &\leq \sqrt{|pq|^2 - |qq'|^2} \\ &\leq r(\mathcal{L})\sqrt{(1+\varepsilon)^2 - 1} = r(\mathcal{L})\sqrt{2\varepsilon+\varepsilon^2}.\end{aligned}$$

On the other hand, ℓ' is perpendicular to both the line qq' and the line pq. Thus ℓ' is perpendicular to the plane defined by the triangle $pq'q$. This implies that $\angle pq'c = \pi/2$. Therefore

$$\sin\angle pcq' = \frac{|pq'|}{|pc|} \leq \frac{\sqrt{2\varepsilon+\varepsilon^2}\cdot r(\mathcal{L})}{\alpha\sqrt{\varepsilon}\cdot r(\mathcal{L})} = \sqrt{2+\varepsilon}/\alpha.$$

Thus the angle between ℓ and line cp is $\angle pcq' \leq \arcsin(\sqrt{2+\varepsilon}/\alpha)$.

Proof. [Lemma 4.2] Without loss of generality, we can assume that v_i's are the unit vectors along the $k+1$ axes and ℓ passes through the origin. We take the unit normal vector v of ℓ. Then the angle between v_j and ℓ, denoted by θ_j, is $\pi/2 - \theta'_j$, where θ'_j is the angle between vectors v_j and v. In order to minimize $\max_i \theta_i = \max_i \arcsin(\sin\theta_i) = \max_i \arcsin(\cos\theta'_i) = \max_i \arcsin(v_i \cdot v)$, one chooses $v = (1,1,\cdots,1)/\sqrt{k+1}$. Thus we have $\max_i \theta_i \geq \arcsin(1/\sqrt{k+1})$.

Proof. [Lemma 4.3] Assume that there is a square R with side length $5\sqrt{\varepsilon}r(\mathcal{L})$ inside $C_\varepsilon(\mathcal{L})$. Denote by u the center of R, and a, b, c, d the four points on the boundary such that ua, ub, uc, ud are perpendicular to the four sides of R respectively, see Figure 3. $|ua| = |ub| = |uc| = |ud| = 5\sqrt{\varepsilon}r(\mathcal{L})/2$.

Now we claim that for any line $\ell \in \mathcal{L}$, $d(u,\ell) < r(\mathcal{L})$. Assume otherwise, there is a line $\ell \in \mathcal{L}$ such that $d(u,\ell) \geq r(\mathcal{L})$. We observe that when we move a point t continuously from u to a or from u to c, at least in one case the distance $d(t,\ell)$ is monotonically increasing. Similarly for nodes b, d. We take such two vectors, say $\vec{ua}, \vec{ub}$. The vectors $\vec{ua}, \vec{ub}$ are perpendicular to each other. Furthermore, a, b are both inside $C_\varepsilon(\mathcal{L})$, so $d(a,\ell) \leq (1+\varepsilon)r(\mathcal{L})$, $d(b,\ell) \leq (1+\varepsilon)r(\mathcal{L})$.

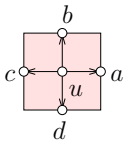

Figure 3: The set of $(1+\varepsilon)$-approximate centers $C_\varepsilon(\mathcal{L})$ does not include a 2-dimensional square with side length $12\sqrt{\varepsilon}\cdot r(\mathcal{L})$.

Also $|ua| = |ub| = 5\sqrt{\varepsilon}r(\mathcal{L})$. By a similar argument as in Lemma 4.1, the angle from ℓ to the line ua (ub) is at most $\arcsin(2\sqrt{2+\varepsilon}/5) < \pi/4$, a contradiction.

Thus, $d(u,\ell) < r(\mathcal{L})$, for any line $\ell \in \mathcal{L}$. We conclude that u is an intersection center of $\mathcal{L}$ with radius less than $r(\mathcal{L})$. This gives a contradiction to the definition of $r(\mathcal{L})$.

Proof. [Theorem 4.1] We first study the case where the set of minimum intersection centers is included in a finite radius ball centered at the origin. Then the set of approximate intersection centers is also included in a finite radius ball. Let c be a minimum intersection center. Denote by p the approximate center such that $|cp|$ is maximal. If $|cp| \leq 20\sqrt{\varepsilon}r(\mathcal{L})$, then the approximate centers are inside a ball with radius $20\sqrt{\varepsilon}r(\mathcal{L})$ centered at c, and thus included in any cylinder with axis through c and radius $20\sqrt{\varepsilon}r(\mathcal{L})$. The claim is true.

If $|cp| > 20\sqrt{\varepsilon}r(\mathcal{L})$, we claim that the cylinder with axis cp and radius $20\sqrt{\varepsilon}r(\mathcal{L})$ includes $C_\varepsilon(\mathcal{L})$. Assume otherwise, there must be an approximate center v with distance more than $20\sqrt{\varepsilon}r(\mathcal{L})$ away from the line cp. Denote by q the reflection point of p on the line cp, $|cq| = |cp|$. Since p is the furthest away approximate center from c, thus the projections of the other approximate centers on the line cp fall inside line segment pq. Now we consider the triangle $\triangle vcp$. By the convexity of $C_\varepsilon(\mathcal{L})$, all the points inside $\triangle vcp$ are $(1+\varepsilon)$-approximate centers. We claim that there must be a 2-dimensional square R with side length $5\sqrt{\varepsilon}r(\mathcal{L})$ inside $\triangle vcp$. Take the middle point of the line segment cp, denoted as s. $|cs| > 10\sqrt{\varepsilon}r(\mathcal{L})$. Take the point s' on line segments cv and vp such that the projection of s' on the line cp is s. Now we argue that a square of side length $5\sqrt{\varepsilon}r(\mathcal{L})$ with s as one corner and a portion of line segment sc as one of the sides must be completely inside $\triangle cvp$. If the projection s of v on the line cp lies on the segment cp, as shown in Figure 4 (i), then the length of ss' is at least $10\sqrt{\varepsilon}r(\mathcal{L})$. Thus $\triangle css'$ must have R completely inside. If the projection s of v on the line cp lies on the segment cq, as shown in Figure 4 (ii), the length of ss' is at least $5\sqrt{\varepsilon}r(\mathcal{L})$. Again it is not hard to verify that the square R is completely inside $\triangle cvp$.

Thus we can find a square R inside $\triangle vcp$ with side length $5\sqrt{\varepsilon}r(\mathcal{L})$, as shown in Figure 4. This implies a contradiction by Lemma 4.3. Thus the cylinder with axis cp and radius $20\sqrt{\varepsilon}r(\mathcal{L})$ includes all the approximate centers.

For a line $\ell_i \in \mathcal{L}$, we find the line ℓ'_i that is parallel with

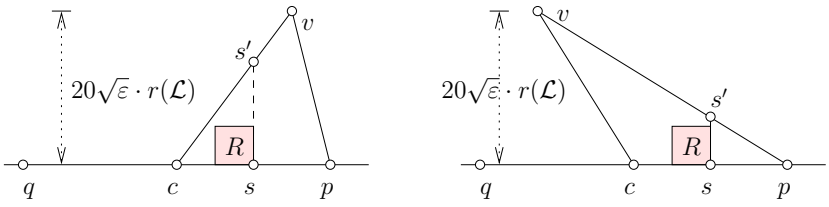

Figure 4: If $|cp| > 20\sqrt{\varepsilon}\cdot r(\mathcal{L})$ and the cylinder with axis cp and radius $20\sqrt{\varepsilon}\cdot r(\mathcal{L})$ does not include all the approximate centers, we can find a square R inside $\triangle vcp$ with side length $5\sqrt{\varepsilon}\cdot r(\mathcal{L})$.

ℓ_i and goes through c. We claim that one of the cylinders with axis ℓ_i' and radius $25\sqrt{\varepsilon}r(\mathcal{L})$ includes $C_\varepsilon(\mathcal{L})$. Recall that q is the reflection point of p on the line cp and the projections of the other approximate centers on the line cp fall inside line segment pq. Let α satisfy $|cp| > \alpha\sqrt{\varepsilon}r(\mathcal{L})$, by Lemma 4.1 there is a line $\ell \in \mathcal{L}$ such that the angle between ℓ and the line cp is $\arcsin(\sqrt{2+\varepsilon}/\alpha)$. Take ℓ' to be the line through c which is parallel with ℓ. The distance from p to ℓ' is at most $r(\mathcal{L})\sqrt{2\varepsilon+\varepsilon^2}$. Thus the distance from any point on the line segment pq to line ℓ' is no more than $r(\mathcal{L})\sqrt{2\varepsilon+\varepsilon^2}$, by simple geometry. Take any approximate center t'', assume its projection to line cp is t, and the projection of t on line ℓ' is t'. The distance from t'' to ℓ' is

$$\begin{aligned} d(t'',\ell') &\le d(t'',t)+d(t,t') \\ &\le 20\sqrt{\varepsilon}r(\mathcal{L})+r(\mathcal{L})\sqrt{2\varepsilon+\varepsilon^2} \\ &\le 25\sqrt{\varepsilon}r(\mathcal{L}). \end{aligned}$$

Thus the cylinder with axis ℓ' and radius $25\sqrt{\varepsilon}r(\mathcal{L})$ includes the collection of $(1+\varepsilon)$-approximate centers.

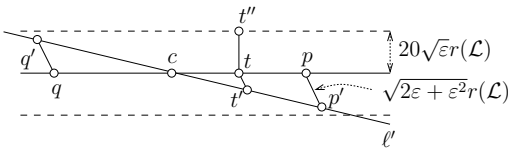

Figure 5: There is a cylinder with axis parallel with one of the lines in $\mathcal{L}$ and radius $25\sqrt{\varepsilon}r(\mathcal{L})$ that includes the collection of $(1+\varepsilon)$-approximate intersection centers.

If the set of minimum intersection centers is not included in a finite radius ball, then all the lines are parallel, and the cylinder that includes all the approximate centers must have a axis parallel with the lines in $\mathcal{L}$ (otherwise it can not cover the minimum intersection centers). Thus if there is an approximate center v outside the cylinder with radius $25\sqrt{\varepsilon}r(\mathcal{L})$, by a similar argument as above we can find a square with side length more than $5\sqrt{\varepsilon}r(\mathcal{L})$ completely inside $C_\varepsilon(\mathcal{L})$. By using Lemma 4.3 we have a contradiction. Thus the cylinder with axis that equals the line of the minimum intersection centers and radius $25\sqrt{\varepsilon}r(\mathcal{L})$ includes all the approximate centers.

Proof. [Theorem 5.1] The proof is similar to the proof of Theorem 4.1 and is based on the analogs of the Lemmas 4.1, 4.3, stated as follows. We first start with the case of line segments.

LEMMA 6.1. *Suppose c is a minimum intersection center of a set $\mathcal{L}$ of lines or line segments and p is a $(1+\varepsilon)$-approximate intersection center, $|cp| \ge \alpha\sqrt{\varepsilon}r(\mathcal{L})$, $\alpha > 2\sqrt{2+\varepsilon}$, then there exists a line or line segment $\ell \in \mathcal{L}$ such that the angle between ℓ and the interval cp is bounded by $\arcsin(\sqrt{2+\varepsilon}/(\alpha-2\sqrt{2+\varepsilon}))$.*

Proof. For the case of line segments, the proof is basically the same as Lemma 4.1. We can assume without loss of generality that the closest point s (q) to c (p) on ℓ is not the endpoints of the line segment. If otherwise, we can find a point $\hat{c}$ on the line segment cp such that $\hat{c}$ is the furthest point from c on cp such that $\hat{c}$'s closest point on ℓ is still s. Notice that $r(\mathcal{L}) < d(\hat{c},\ell) \le (1+\varepsilon)r(\mathcal{L})$. Since the distance $d(t,\ell)$ is monotonically increasing, as t moves from c to p, the inner angle of $\triangle\hat{c}cw$ at c is at least $\pi/2$. Thus $|c\hat{c}| \le \sqrt{(1+\varepsilon)^2-1}r(\mathcal{L}) = \sqrt{(2+\varepsilon)\varepsilon}r(\mathcal{L})$. Similarly, one can find the corresponding point $\hat{p}$ on $\hat{c}p$. $|p\hat{p}| \le \sqrt{(2+\varepsilon)\varepsilon}r(\mathcal{L})$. Thus $|\hat{c}\hat{p}| \ge (\alpha-2\sqrt{2+\varepsilon})\sqrt{\varepsilon}r(\mathcal{L})$. Now we follow the arguments in the case of lines. The angle between ℓ and cp is bounded by $\arcsin(\sqrt{2+\varepsilon}/(\alpha-2\sqrt{2+\varepsilon}))$.

LEMMA 6.2. *Suppose c is a minimum intersection center of a set $\mathcal{L}$ of convex sets with dimension at most Δ and p is a $(1+\varepsilon)$-approximate intersection center, $|cp| \ge \alpha\sqrt{\varepsilon}r(\mathcal{L})$, $\alpha > 2\sqrt{2+\varepsilon}$, then there exists a convex set $\ell \in \mathcal{L}$ such that the angle between ℓ and the interval cp is bounded by $\arcsin(\sqrt{2+\varepsilon}/(\alpha-2\sqrt{2+\varepsilon}))$.*

Proof. The proof is very similar to the proof of Lemma 4.1 and 6.1. There is a convex set $\ell \in \mathcal{L}$ such that $d(c,\ell) = r(\mathcal{L})$, $d(p,\ell) \le (1+\varepsilon)r(\mathcal{L})$ and $d(t,\ell) > 1$ monotonically non-decreasing as t moves from c to p. Let s be the point on ℓ

closest to c and q the point on ℓ closest to p. Now we know that $|cs| = r(\mathcal{L})$ and $|pq| \leq (1+\varepsilon)r(\mathcal{L})$.

If s and q are the same point, then the inner angle of $\triangle scp$ at vertex c is at least $\pi/2$. Thus $|cp| \leq \sqrt{|sp|^2 - |cs|^2} \leq \sqrt{(2+\varepsilon)\varepsilon}r(\mathcal{L})$. Thus $\alpha \leq \sqrt{2+\varepsilon}$, which is a contradiction.

If s are q are different, we note that the line segment sq is completely inside ℓ, due to the fact that ℓ is convex. Further, as a point t moves on cp from c to p, the distance between t and the line segment sq is non-decreasing. By Lemma 6.1, the lines cp and sq have a small angle. Thus the angle between cp and ℓ is no more than $\arcsin(\sqrt{2+\varepsilon}/(\alpha - 2\sqrt{2+\varepsilon}))$.

LEMMA 6.3. *$C_\varepsilon(\mathcal{L})$ does not include a $(\Delta+1)$-dimensional cube with side length $\gamma\sqrt{\varepsilon(\Delta+1)}r(\mathcal{L})$, for a constant $\gamma \geq 6\sqrt{2+\varepsilon}$.*

Proof. Suppose otherwise, there is a $(\Delta+1)$-dimensional cube R with side length $\gamma\sqrt{\varepsilon(\Delta+1)}r(\mathcal{L})$ inside $C_\varepsilon(\mathcal{L})$. Denote by u the center of this cube R. Again, we claim that $d(u,\ell) < r(\mathcal{L})$ for any $\ell \in \mathcal{L}$, which contradicts with the definition of $r(\mathcal{L})$. Now suppose that there is a convex set $\ell \in \mathcal{L}$ such that $d(u,\ell) \geq r(\mathcal{L})$, we argue a contradiction. By Lemma 6.2, we can find $\Delta+1$ orthogonal vectors of length $\gamma\sqrt{\varepsilon(\Delta+1)}r(\mathcal{L})/2$ centered at u such that the angle between each vector to the convex set ℓ is no more than $\arcsin(\frac{\sqrt{2+\varepsilon}}{\gamma\sqrt{\Delta+1}/2 - 2\sqrt{2+\varepsilon}}) \leq \arcsin(1/\sqrt{\Delta+1})$. This contradicts with Lemma 4.2.

Finally we show, with the above lemmas we can prove our Theorem. In this final step, given a minimal intersection center c, we find a set of pairs (p_1,q_1), (p_2,q_2), $\ldots$, $(p_{\Delta+1},q_{\Delta+1})$ in the following way. p_1,q_1 are the furthest pairs of points in $C_\varepsilon(\mathcal{L})$ such that the line segment p_1q_1 intersects c. p_2,q_2 are the furthest pairs of points in $C_\varepsilon(\mathcal{L})$ such that the line segment p_2q_2 is perpendicular to the 1-flat spanned by p_1,q_1. Similarly, p_i,q_i are the furthest pairs of points in $C_\varepsilon(\mathcal{L})$ such that the line segment p_iq_i is perpendicular to the $(i-1)$-flat spanned by $\{p_1,q_1,p_2,q_2,\cdots,p_{i-1},q_{i-1}\}$. Define $d_i = |p_iq_i|$. Now we claim that at least for one $1 \leq i \leq \Delta+1$, $d_i \leq \beta\sqrt{\varepsilon(\Delta+1)^3}r(\mathcal{L})$, for some constant β.

Suppose otherwise, $d_i > \beta\sqrt{\varepsilon(\Delta+1)^3}r(\mathcal{L})$ for $1 \leq i \leq \Delta+1$. Now consider the convex polytope P spanned by the points $\{p_1,q_1,p_2,q_2,\cdots,p_{\Delta+1},q_{\Delta+1}\}$. By the convexity of $C_\varepsilon(\mathcal{L})$, all the points in the interior of P are $(1+\varepsilon)$-approximate centers. Thus one can then find a $(\Delta+1)$-dimensional cube with side length $\gamma\sqrt{\varepsilon(\Delta+1)}r(\mathcal{L})$ inside P, with $\gamma \geq 6\sqrt{2+\varepsilon}$. The details are omitted here. By Lemma 6.3 we have a contradiction.

Thus $d_i \leq \beta\sqrt{\varepsilon(\Delta+1)^3}r(\mathcal{L})$ for some i. Therefore $C_\varepsilon(\mathcal{L})$ can be enclosed in a Δ-slab with axis as the flat spanned by p_1q_1, p_2q_2, $\cdots$, $p_{i-1}q_{i-1}$, $p_{i+1}q_{i+1}$, $\cdots$, $p_{\Delta+1}q_{\Delta+1}$ and width $\beta\sqrt{\varepsilon(\Delta+1)^3}r(\mathcal{L})$.

Proof. [Theorem 1.3] The basic idea is the same as in Theorem 1.2. We first focus on the existence of a small size core set. We start with $\Delta+2$ sets $\mathcal{L}' \subseteq \mathcal{L}$ according to Theorem 1.1 such that $r(\mathcal{L}) \leq \alpha \cdot r(\mathcal{L})$, $\alpha = 2$. Let $c = c(\mathcal{L}')$ be the minimum intersection center of $\mathcal{L}'$ and ℓ' be the axis of the slab that contains the collection of $(1 + \frac{\varepsilon^2}{4\beta^2(\Delta+1)^3})$-approximate intersection centers of $\mathcal{L}'$ for some constant β in Theorem 5.1. Define I_i to be a subset of ℓ' such that a point $p \in I_i$ has distance at most $(1+\varepsilon)r(\mathcal{L}')$ away from the sets $\mathcal{L}' \bigcup \{\ell_i\}$, $\ell_i \in \mathcal{L} \setminus \mathcal{L}'$.

If among all I_i, every $\Delta+1$ of them have a non-empty intersection, then $\bigcap_i I_i \neq \emptyset$, by Helly's theorem. Thus the ball with radius $(1+\varepsilon)r(\mathcal{L}')$ centered at a point $c' \in \bigcap_i I_i$ intersects with every set in $\mathcal{L}$ and we are done ($\mathcal{L}'$ is the core set).

If there are $\Delta+1$ sets $\ell_1,\ell_2,\cdots,\ell_{\Delta+1}$ such that their corresponding sets I_j, $j = 1,\cdots,\Delta+1$ do not have a common intersection, then it can be verified in the same way as in Theorem 1.2 that $r(\mathcal{L}' \bigcup \{\ell_1,\cdots,\ell_{\Delta+1}\}) \geq (1 + \frac{\varepsilon^2}{4\beta^2(\Delta+1)^3})r(\mathcal{L}')$. Thus we add all the sets ℓ_j, $j = 1,\cdots,\Delta+1$, to $\mathcal{L}'$ and go to the next iteration.

Thus, for each iteration, at most $\Delta+1$ sets are added to $\mathcal{L}'$ and the value α is decreased by a factor of $(1 - \frac{\varepsilon^2}{2\beta^2(\Delta+1)^3})$. After $O(\Delta^3/\varepsilon^2)$ steps, there are $O(\Delta^4/\varepsilon^2)$ sets in $\mathcal{L}'$ such that $r(\mathcal{L}) \leq (1+\varepsilon)r(\mathcal{L}')$.

This concludes our proof for the existence of the $(1+\varepsilon)$-core set for convex sets of dimension at most Δ. For an algorithm to compute this core set, notice that the axis ℓ' of the slab containing $C_\varepsilon(\mathcal{L}')$ is not known. Thus we will try each $\Delta+1$ tuples of the remaining sets of $\mathcal{L}$ at each iteration. We will terminate our algorithm once no $\Delta+1$ tuple will increase the minimum intersecting radius (significantly). The running time of the algorithm follows from the same arguments as in Theorem 1.2. A center c such that the ball of radius $(1+\varepsilon)r(\mathcal{L}')$ intersects all sets in $\mathcal{L}$ can be found using Lemma 2.1 (convex programming). The careful reader may have noticed that the running time of our algorithm for general Δ is greater than that implied by standard convex programming (Lemma 2.1). Indeed this is the case, however, in our algorithm in addition to returning an approximate center c, we also return the coreset $\mathcal{L}'$ whose existence should be viewed as the main contribution of this Theorem.