# High-Dimensional Shape Fitting in Linear Time

Sariel Har-Peled* Kasturi R. Varadarajan†

27/05/2003 12:40

**Abstract**

Let $P$ be a set of $n$ points in $\mathbb{R}^d$. The *radius* of a $k$-dimensional flat $\mathcal{F}$ with respect to $P$, which we denote by $\mathcal{RD}(\mathcal{F}, P)$, is defined to be $\max_{p \in P} \operatorname{dist}(\mathcal{F}, p)$, where $\operatorname{dist}(\mathcal{F}, p)$ denotes the Euclidean distance between $p$ and its projection onto $\mathcal{F}$. The $k$-flat radius of $P$, which we denote by $R_k^{opt}(P)$, is the minimum, over all $k$-dimensional flats $\mathcal{F}$, of $\mathcal{RD}(\mathcal{F}, P)$. We consider the problem of computing $R_k^{opt}(P)$ for a given set of points $P$. We are interested in the high-dimensional case where $d$ is a part of the input and not a constant. This problem is NP-hard even for $k = 1$. We present an algorithm that, given $P$ and a parameter $0 < \varepsilon \leq 1$, returns a $k$-flat $\mathcal{F}$ such that $\mathcal{RD}(\mathcal{F}, P) \leq (1+\varepsilon) R_k^{opt}(P)$. The algorithm runs in $O(ndC_{\varepsilon,k})$ time, where $C_{\varepsilon,k}$ is a constant that depends only on $\varepsilon$ and $k$. Thus the algorithm runs in time linear in the size of the point set and is a substantial improvement over previous known algorithms, whose running time is of the order of $dn^{O(k/\varepsilon^c)}$, where $c$ is an appropriate constant.

## 1 Introduction

Let $P$ be a set of $n$ points in $\mathbb{R}^d$. The *radius* of a $k$-dimensional flat $\mathcal{F}$ with respect to $P$, which we denote by $\mathcal{RD}(\mathcal{F}, P)$, is defined to be $\max_{p \in P} \operatorname{dist}(\mathcal{F}, p)$, where $\operatorname{dist}(\mathcal{F}, p)$ denotes the Euclidean distance between $p$ and its projection onto $\mathcal{F}$. The $k$-flat radius of $P$, which we denote by $R_k^{opt}(P)$, is the minimum, over all $k$-dimensional flats $\mathcal{F}$, of $\mathcal{RD}(\mathcal{F}, P)$. (See Section 2 for detailed definitions.) Thus $R_0^{opt}(P)$ is the radius of the min-enclosing ball of $P$, $R_1^{opt}(P)$ is the radius of the min-enclosing cylinder of $P$, and so on. Informally, the $k$-flat radius of $P$ measures how well the point set $P$ can be approximated by an affine subspace of dimension $k$. Computing the $k$-flat radius of a point set is a fundamental problem in computational convexity and has applications in data mining, statistics, and clustering scenarios [GK93, GK94, HV02].

The problem of computing the $k$-flat radius of a point set has received considerable attention in the computational geometry literature. A classic result is that the min-enclsoing ball ($R_0^{opt}(P)$) of a point set $P$ can be computed in linear time when the dimension is fixed. For $k \geq 1$, the $k$-flat radius of a set $P$ of $n$ points can be computed exactly in polynomial time in fixed dimension [FKS96]. It can also be approximated to within a factor of $(1+\varepsilon)$, for any $\varepsilon > 0$, in $O(n + (1/\varepsilon)^c)$ time, where $c$ is a constant that depends only on $d$ and $k$ [BH01, HV01]. Thus the problem is reasonably well understood when the dimension $d$ is taken to be a fixed constant. These algorithms are not satisfactory when the dimension is large as their running times depend exponentially on the dimension. In the rest of this section, we are interested in efficient algorithms when the dimension $d$ can be as large as $n$.

*Department of Computer Science, DCL 2111; University of Illinois; 1304 West Springfield Ave.; Urbana, IL 61801; USA; `http://www.uiuc.edu/~sariel/`; `sariel@uiuc.edu`. Work on this paper was partially supported by a NSF CAREER award CCR-0132901.

† Department of Computer Science, University of Iowa, `kvaradar@cs.uiowa.edu`

It is well-known that the minimum enclosing ball of a set of points can be computed in polynomial time; see for instance the paper by Gritzmann and Klee [GK93]. Megiddo [Meg90] shows that the problem of determining whether there is a line that intersects a set of balls is NP-hard. In his reduction, the balls have the same radius, which implies that computing the radius $R_1^{opt}(P)$ of the min-enclosing cylinder of a set of points $P$ is NP-hard. Bădoiu et al. [BHI02] give a poly-time algorithm that computes a $(1+\varepsilon)$-approximation, for any $\varepsilon > 0$, of the minimum enclosing cylinder ($R_1^{opt}(P)$) of a set of points. Har-Peled and Varadarajan [HV02] give a poly-time algorithm that computes a $(1+\varepsilon)$-approximation, for any $\varepsilon > 0$, to $R_k^{opt}(P)$ whenever $k$ is a fixed constant. Their algorithm runs in $O(dn^{k/\varepsilon^5})$ time. These results show that the $k$-flat radius can be efficiently approximated for small $k$.

The problem seems to become harder when $k$ becomes large. Bodlaender et al. [BGKvL90] show that computing the width, $R_{d-1}^{opt}(P)$, of a point set is NP-hard. Gritzmann and Klee [GK93] show that it is NP-hard to compute the width of even a $d$-dimensional simplex (that is, a set of $d+1$ points). They also show that it is NP-hard to compute $R_k^{opt}(P)$ for small point sets (consisting of $2d$ points) as long as $k \geq c \cdot d$, for any fixed $0 < c < 1$. Recently, Brieden [Bri02] showed that is NP-hard to approximate the width of a point set to within *any* constant factor. Varadarajan et al. [VVZ02] show that Brieden's result can be strengthened as follows: there exists a constant $\delta > 0$ such that it is NP-hard to approximate the $k$-flat radius of a set of $n$ points to within a factor of $(\log n)^\delta$ whenever $k \geq d^\varepsilon$, for any constant $\varepsilon > 0$. Turning to upper bounds, the papers of Nesterov [Nes98] and Nemirovski et al. [NRT99] imply an $O(\sqrt{\log n})$-approximation for the width ($R_{d-1}^{opt}$) of a point set in polynomial time. Varadarajan et al. [VVZ02] describe a poly-time algorithm that approximates the $k$-flat radius of a point set to within a factor of $\sqrt{\log n \cdot \log d}$ (for any $k$). We refer the reader to this paper for further details on approximating the $k$-flat radius for large $k$.

Our current paper focusses on the case when $k$ is small and significantly improves upon our previous work [HV02]. In that paper, we presented an algorithm that, given a point set $P$ with $n$ points and a parameter $0 < \varepsilon \leq 1$, computes a $k$-flat whose radius with respect to $P$ is at most $(1+\varepsilon)R_k^{opt}(P)$ in $dn^{O(k/\varepsilon^5 \log(1/\varepsilon))}$ time. The basic algorithm of [HV02] is based on proving the existence of a small set of points, called a core set, such that the affine subspace spanned by the points contains a $k$-flat that is near-optimal. The size of the core set depends on $k$ and $\varepsilon$ but not on the dimension $d$. Now, by brute force enumeration on all such small sets, the algorithms generates all possible candidate core sets. For each candidate core set, the algorithm finds the best $k$-flat in the affine subspace spanned by the core set; this, being a "low-dimensional" problem, is feasible. The algorithm then returns the best $k$-flat overall. The nice property of coresets is that they are deus ex machina – their existence immediately implies relatively efficient approximation algorithms that extends immediately to fitting multiple-flats with outliers. On the negative side, these techniques have relatively bad dependency on $n$. In this paper, we develop a new technique that that bypasses the enumeration of core sets altogether, and present direct algorithms for computing a good $k$-flat. In particular, our new algorithms have linear dependency on the number of input points and the dimension.

Given a set $P$ of $n$ points in $d$ dimensions ($d$ might be as large as $n$) and a parameter $0 < \varepsilon \leq 1$, we present efficient algorithms for the following problems:

- **Minimum Radius Line.** In Section 3, we present an approximation algorithm that returns a line $\ell$ such that $\mathcal{RD}(\ell, P) \leq (1+\varepsilon)R_1^{opt}(P)$. The running time of the new algorithm is $O(ndC_\varepsilon)$, where $C_\varepsilon = \exp\left(\frac{1}{\varepsilon^3}\log^2\frac{1}{\varepsilon}\right)$.

  This is a substantial improvement over the previously fastest algorithms of [BC02, BHI02, KMY03, HV02], which all had running times of the form $dn^{O(1/\varepsilon^c)}$, where $c$ is an appropriate constant.

It is natural to ask if there is an algorithm for solving this problem with a running time that depends polynomially on $n$, $d$, and $1/\varepsilon$. We prove in Section 4 that such an algorithm does not exist unless P = NP. We do this by showing that the NP-hardness reduction of Megiddo can be modified so that it yields an appropriate hardness of approximation result.

- **Minimum Radius $k$-flat.** In Section 5, we generalize the algorithm for a line to compute a $k$-flat $\mathcal{F}$ such that $\mathcal{RD}(\mathcal{F}, P) \leq (1+\varepsilon)R_k^{opt}(P)$. The running time of the new algorithm is $n \cdot d \cdot \exp\left(\frac{e^{O(k^2)}}{\varepsilon^{2k+1}} \log^2 \frac{1}{\varepsilon}\right)$. This substantially improves over the previous fastest algorithm [BC02, KMY03, HV02] that has running time $dn^{O(k/\varepsilon^c)}$, where $c$ is an appropriate constant. It is quite likely that the hardness result for $k = 1$ implies a hardness result for any $k \geq 1$, but we have not attempted to prove this rigorously.

The main result of this paper is the algorithm for the case $k = 1$, which extends naturally to the case $k > 1$. We describe the ideas behind this algorithm. Let $\ell_{opt}$ denote an optimal line for the point set $P$, and $\ell$ be any line. If $\ell$ is not nearly as good as $\ell_{opt}$, then the 2-flat $h$ containing $\ell$ and the point in $P$ maximizing $\mathrm{dist}(\ell, p)$ has a line $\ell'$ that is "significantly closer" to $\ell_{opt}$ than $\ell$. Moreover, we can compute a small number of candidate lines on $h$ one of which is guaranteed to contain $\ell'$. By trying all the candidates, we are guaranteed to have a line that is significantly closer to $\ell_{opt}$ than $\ell$. By repeating this process a small number of times, we either stumble upon a line that is near-optimal, or we end up with a line that is very close to $\ell_{opt}$ and is therefore near-optimal. To realize this vision, we need an appropriate notion of "closeness" between lines and corresponding machinery to argue about convergence; the provision of these is the most interesting contribution of this paper.

# 2 Preliminaries

**Definition 2.1** Given $j$ points $v_1, \ldots, v_j \in \mathbb{R}^d$, the linear subspace they span, is denoted by $\mathrm{span}(v_1, \ldots, v_j) = \left\{ v \,\middle|\, v = \sum_{i=1}^{j} a_i v_i \text{ for any } a_1, \ldots, a_j \in \mathbb{R} \right\}$.

Given $j+1$ points $p_1, \ldots, p_{j+1}$ in $\mathbb{R}^d$, the *affine space* spanned by them is

$$\mathrm{affine}(p_1, \ldots, p_{j+1}) = \left\{ v \,\middle|\, v = \sum_{i=1}^{j+1} a_i v_i, \text{ for any } a_1, \ldots, a_{j+1} \in \mathbb{R} \text{ such that } \sum_{i=1}^{j+1} a_i = 1 \right\}.$$

Alternatively,

$$\mathrm{affine}(p_1, \ldots, p_{j+1}) = p_1 + \mathrm{span}(p_2 - p_1, p_3 - p_1, \ldots, p_{j+1} - p_1).$$

We refer to an affine subspace of dimension $k$ as a *$k$-flat.* When the dimension is not relevant, we simply refer to an affine subspace as a flat. See [Ede87] for more on these definitions.

**Definition 2.2** For a flat $\mathcal{F}$ in $\mathbb{R}^d$ and a point $y \in \mathbb{R}^d$, let $\mathrm{proj}(\mathcal{F}, y)$ denote the projection of $y$ onto $\mathcal{F}$. Namely $\mathrm{proj}(\mathcal{F}, y) = \mathrm{argmin}_{x \in \mathcal{F}} \|xy\|$. For a set point $Y \in \mathbb{R}^d$, let $\mathrm{proj}(\mathcal{F}, Y) = \{\mathrm{proj}(\mathcal{F}, y) \mid y \in Y\}$.

For a flat $\mathcal{F}$ and a point $y$, we let $\mathrm{dist}(\mathcal{F}, y)$ denote the distance $\|y\,\mathrm{proj}(\mathcal{F}, y)\|$ of $y$ from its projection onto $\mathcal{F}$. We sometimes refer to $\mathrm{dist}(\mathcal{F}, y)$ as the distance of $y$ from $\mathcal{F}$. For a flat $\mathcal{F}$ and point set $P \subseteq \mathbb{R}^d$, the *radius* of $\mathcal{F}$ with respect to $P$ is $\mathcal{RD}(\mathcal{F}, P) = \max_{p \in P} \mathrm{dist}(\mathcal{F}, p)$.

For $k \geq 0$, let $R_k^{opt}(P)$ denote the radius of a $k$-flat which minimizes the radius with respect to $P$. Formally, $R_k^{opt}(P) = \min_{\mathcal{F} \in FLT_k} \mathcal{RD}(\mathcal{F}, P)$, where $FLT_k$ is the set of all $k$-flats in $\mathbb{R}^d$.

Given a $k$-flat $\mathcal{F}$, a point $p$ on $\mathcal{F}$, and positive real numbers $0 < \beta \le \alpha$, an $(\alpha, \beta)$-*net around* $p$ *on* $\mathcal{F}$ is a set $S$ of points on $\mathcal{F}$ such that for any point $q$ on $\mathcal{F}$ such that $\|qp\| \le \alpha$, there is a point $s \in S$ such that $\|qs\| \le \beta$. It is well-known that there exists a net $S$ of size $\exp(O(k \log(\alpha/\beta)))$ [Mat02], and it is easy to see that a larger net of size $\exp(O(k \log(k\alpha/\beta)))$ can be computed in $\exp(O(k \log(k\alpha/\beta)))$ time.

## 2.1 Distance Function

**Definition 2.3** Given a $j$-flat $\mathcal{G}$, the *distance function* $\mathrm{d}_{\mathcal{G}}(x)$, for $x \in \mathbb{R}^d$, is the distance of $x$ from its projection onto $\mathcal{G}$; namely $\mathrm{d}_{\mathcal{G}}(x) = \|x \,\mathrm{proj}(\mathcal{G}, x)\|$.

Let $\ell$ be a line in $\mathbb{R}^d$ that is parameterized by $\{a_1 + ta_2 | t \in \mathbb{R}\}$, where $a_1, a_2 \in \mathbb{R}^d$ and $||a_2|| = 1$. Let $\mathrm{d}(t)$ denote the distance of the point $\ell(t) = a_1 + ta_2$ on $\ell$ from $\mathcal{G}$, that is, $\mathrm{d}(t) = \mathrm{d}_{\mathcal{G}}(t)$. Then $\mathrm{d}(t)$ is of the form $\mathrm{d}(t) = \sqrt{at^2 + bt + c}$, where $a, b, c$ are appropriate constants that depend on $\ell$ (specifically $a_1, a_2$) and $\mathcal{G}$. Moreover, the constants $a$, $b$, and $c$ are such that $at^2 + bt + c$ is non-negative for every $t \in \mathbb{R}$.

**Lemma 2.4** *Let $\mathrm{d}_{\mathcal{G}}(\cdot)$ be the distance function to a flat $\mathcal{G}$. Let $x, y \in \mathbb{R}^d$ be any two points.*

*(i) For any $0 \le \beta \le 1$, $\mathrm{d}_{\mathcal{G}}(\beta x + (1-\beta)y) \le \beta \mathrm{d}_{\mathcal{G}}(x) + (1-\beta)\mathrm{d}_{\mathcal{G}}(y)$. That is, $\mathrm{d}_{\mathcal{G}}(\cdot)$ is convex.*

*(ii) Let $\ell$ be the line through $x$ and $y$, and $w$ and $z$ be any two points on $\ell$. Then*

$$\mathrm{d}_{\mathcal{G}}(z) \le \mathrm{d}_{\mathcal{G}}(w) + 2\frac{||wz||}{||xy||} \max(\mathrm{d}_{\mathcal{G}}(x), \mathrm{d}_{\mathcal{G}}(y)).$$

*Proof:* Suppose that the $\ell$ through $x$ and $y$ is parameterized by $\{x + ta_1 | t \in \mathbb{R}\}$, where $a_1 \in \mathbb{R}^d$ is the unit vector $\frac{1}{||xy||}(y - x)$. Let $\mathrm{d}(t)$ denote the distance of the point $\ell(t) = x + ta_1$ on $\ell$ from $\mathcal{G}$, that is, $\mathrm{d}(t) = \mathrm{d}_{\mathcal{G}}(t)$. Then $\mathrm{d}(t)$ is of the form $\mathrm{d}(t) = \sqrt{at^2 + bt + c}$, where $a, b, c$ are appropriate constants that depend on $\ell$ (specifically $x, y$) and $\mathcal{G}$. Moreover, the constants $a$, $b$, and $c$ are such that $at^2 + bt + c$ is non-negative for every $t \in \mathbb{R}$. This implies that $a \ge 0$. Let $\alpha = \mathrm{argmin}_t \, \mathrm{d}(t)$. Clearly, $\mathrm{d}(t)$ can be written as $\mathrm{d}(t) = \sqrt{a(t-\alpha)^2 + d}$, where $d \ge 0$ is an appropriate constant.

(i) This is well-known and so we skip the proof.

(ii) Consider the function $g(t) = \sqrt{a(t-\alpha)^2} = \sqrt{a}\,|t - \alpha|$. It is easy to verify that $g(t) \le \mathrm{d}(t)$ for all $t \in \mathbb{R}$ and $\mathrm{d}'(t) \le g'(t)$, for all $t \in \mathbb{R} \setminus \{\alpha\}$.

Let $U = \max(\mathrm{d}_{\mathcal{G}}(x), \mathrm{d}_{\mathcal{G}}(y))$. Because of the convexity established in part (i), we have $\mathrm{d}(t) \le U$ for any $t \in [t_x, t_y]$. Thus, $g(t) \le U$, for all $t \in [t_x, t_y]$. And $|g'(t)| = \sqrt{a}$, for all $t \in \mathbb{R} \setminus \{\alpha\}$. However, the value of $\sqrt{a}$ is maximized when $\alpha$ is the middle point of $[t_x, t_y]$, and $g(t_x) = g(t_y) = U$. In this case, $a = (2U/(t_y - t_x))^2$, and $|g'(t)| = 2U/(t_y - t_x)$, for all $t \in \mathbb{R} \setminus \{\alpha\}$. This implies $\mathrm{d}'(t) \le g'(t) \le 2U/(t_y - t_x)$, for all $t \in \mathbb{R}$.

Suppose that $\ell(t_w) = w$ and $\ell(t_z) = z$. Because of the bound on $\mathrm{d}'(t)$, we have

$$\mathrm{d}(t_z) \le \mathrm{d}(t_w) + \left(\max_t \mathrm{d}'(t)\right) |t_z - t_w| \le \mathrm{d}(t_w) + \frac{2U}{t_y - t_x} \cdot |t_z - t_w|,$$

which completes the proof. ■

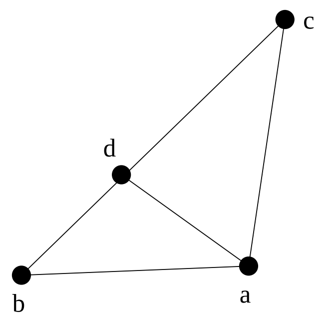


Figure 1: The traingles bac and bda are similar.

### 2.2 Triangles

We now state a simple lemma that will play an important role in this paper. See Figure 1

**Lemma 2.5** *Let $a$, $b$, and $c$ be points in $\mathbb{R}^d$ such that $\|ab\| \leq r$ and $\|bc\| \geq (1+\varepsilon)r$, where $r \geq 0$ and $0 < \varepsilon \leq 1$. Then there is a point $d$ on the segment $bc$ such that $\|ad\| \leq (1-\varepsilon/2)\|ac\|$.*

*Proof:* Let $\rho = \|ab\| / \|bc\|$, and let $d$ be the point placed on segment $bc$ at distance $\rho \|ab\|$ from $b$. It is easy to see that $\triangle bac$ is similar to $\triangle bda$, with a scaling factor of $\rho$. Thus

$$\|ad\| = \rho \|ac\| \leq \frac{1}{1+\varepsilon} \|ac\| \leq (1-\varepsilon/2) \|ac\| .$$

■

### 2.3 Rotations

Let $\mathcal{F}$ be a $k$-dimensional flat in $\mathbb{R}^d$, $\mathcal{G}$ a $j$-dimensional flat that lies in $\mathcal{F}$, where $j < k$, and $p$ be any point in $\mathbb{R}^d$. We define $\mathrm{Rot}(\mathcal{F}, \mathcal{G}, p)$, the rotation of $\mathcal{F}$ around $\mathcal{G}$ that passes through $p$, in the following way. If $p \in \mathcal{F}$, $\mathrm{Rot}(\mathcal{F}, \mathcal{G}, p)$ is $\mathcal{F}$ itself. Otherwise, let $p' = \mathrm{proj}(\mathcal{F}, p)$ and $p'' = \mathrm{proj}(\mathcal{G}, p') = \mathrm{proj}(\mathcal{G}, p)$. If $p' = p''$, we let $\mathcal{H}$ be any $(k-1)$-flat that contains $\mathcal{G}$; otherwise, we let $\mathcal{H}$ be the $(k-1)$-flat contained in $\mathcal{F}$ that passes through $p''$ and is orthogonal to the vector $\overrightarrow{p''p'}$. It is easy to see that in either case $\mathcal{H}$ contains $\mathcal{G}$. We define $\mathrm{Rot}(\mathcal{F}, \mathcal{G}, p)$ to be the $k$-flat that contains $\mathcal{H}$ and $p$. See Figure 2

**Lemma 2.6** *Let $\mathcal{F}$ be a $k$-flat in $\mathbb{R}^d$, $\mathcal{G}$ a $j$-flat that lies in $\mathcal{F}$, where $j < k$, and $p$ be any point in $\mathbb{R}^d$ not on $\mathcal{G}$. For any $q \in \mathbb{R}^d$, we have*

$$\mathrm{dist}(Rot(\mathcal{F}, \mathcal{G}, p), q) \leq \mathrm{dist}(\mathcal{F}, q) + \frac{\mathrm{dist}(\mathcal{G}, q)}{\mathrm{dist}(\mathcal{G}, p)} \mathrm{dist}(\mathcal{F}, p).$$

*Proof:* The claim follows readily when $p \in \mathcal{F}$, since in that case $\mathrm{Rot}(\mathcal{F}, \mathcal{G}, p) = \mathcal{F}$. Otherwise, let $p' = \mathrm{proj}(\mathcal{F}, p)$ and $p'' = \mathrm{proj}(\mathcal{G}, p') = \mathrm{proj}(\mathcal{G}, p)$. Let $\mathcal{H}$ be the $(k-1)$-flat in the definition of $\mathrm{Rot}(\mathcal{F}, \mathcal{G}, p)$; that is, $\mathcal{H}$ is the intersection of $\mathrm{Rot}(\mathcal{F}, \mathcal{G}, p)$ with $\mathcal{F}$. It is easy to see that $p'' = \mathrm{proj}(\mathcal{H}, p') = \mathrm{proj}(\mathcal{H}, p)$.

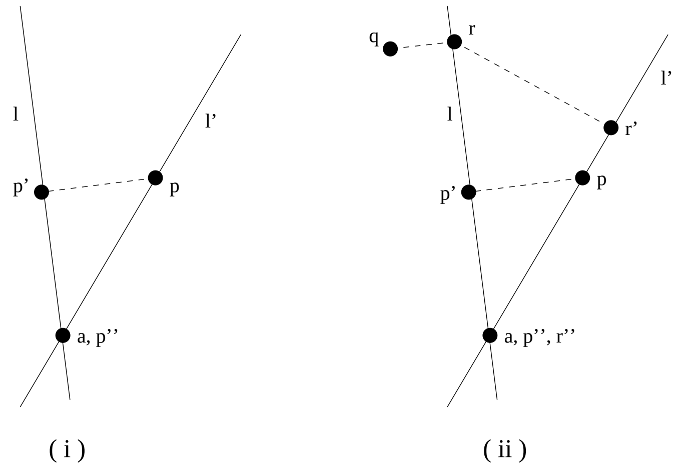


Figure 2: (i) $l' = \mathrm{Rot}(l, a, p)$ for a line $l$, a point $a$ on $l$, and a point $p$. (ii) In the proof of Lemma 2.6, triangles $pp'p''$ and $rr'r''$ are similar.

Let $r = \mathrm{proj}(\mathcal{F}, q)$. If $r$ lies on $\mathcal{H}$, then the lemma follows because $\mathrm{dist}(\mathrm{Rot}(\mathcal{F}, \mathcal{G}, p), q) \leq \mathrm{dist}(\mathcal{H}, q) = \mathrm{dist}(\mathcal{F}, q)$. If $r$ does not lie on $\mathcal{H}$, let $r' = \mathrm{proj}(\mathrm{Rot}(\mathcal{F}, \mathcal{G}, p), r)$ and $r'' = \mathrm{proj}(\mathcal{H}, r) = \mathrm{proj}(\mathcal{H}, r')$. It is easy to verify that $\triangle pp'p''$ and $\triangle rr'r''$ are similar. (See Figure 2.) It follows that

$$||rr'|| = \frac{||rr''||}{||pp''||}||pp'|| = \frac{\mathrm{dist}(\mathcal{H}, r)}{\mathrm{dist}(\mathcal{G}, p)}\,\mathrm{dist}(\mathcal{F}, p) \leq \frac{\mathrm{dist}(\mathcal{G}, r)}{\mathrm{dist}(\mathcal{G}, p)}\,\mathrm{dist}(\mathcal{F}, p).$$

We conclude that

$$\mathrm{dist}(\mathrm{Rot}(\mathcal{F}, \mathcal{G}, p), q) \leq ||qr|| + ||rr'|| = \mathrm{dist}(\mathcal{F}, q) + \frac{\mathrm{dist}(\mathcal{G}, q)}{\mathrm{dist}(\mathcal{G}, p)}\,\mathrm{dist}(\mathcal{F}, p).$$

■

# 3 Minimum Radius Cylinder

In this section, we present an efficient algorithm that given a set $P \subseteq \mathbb{R}^d$ of $n$ points and a parameter $0 < \varepsilon \leq 1$ computes a line $\ell$ such that $\mathcal{RD}(\ell, P) \leq (1+\varepsilon)R_1^{opt}(P)$. It will be convenient to first describe the algorithm assuming the presence of an appropriate oracle (to be described below). At any time during its execution, the algorithm can present the oracle with a set of possible choices, and the oracle simply picks the right one.

**The Algorithm.** Let $\ell_{opt}$ denote an optimal line for the input point set $P$, that is, $\mathcal{RD}(\ell_{opt}, P) = R_1^{opt}(P)$. Let $R^{opt} = R_1^{opt}(P)$. For any $p \in P$, let $p' = \mathrm{proj}(\ell_{opt}, p)$ denote the projection of $p$ onto $\ell_{opt}$. Let $P' = \{p' \mid p \in P\}$. Let $\mathcal{I} = \mathcal{CH}(P')$ denote the convex hull of $P'$. Clearly, $\mathcal{I}$ is the line segment joining $u'$ and $v'$, for some $u, v \in P$. The length $L$ of the segment $\mathcal{I}$ is clearly at most $\mathrm{diam}(P)$. We need the following lemma, whose proof is presented at the end of Section 3.

**Lemma 3.1** *Let $P$ be a set of $n$ points in $\mathbb{R}^d$, $\ell_{opt}$ be a line such that $\mathcal{RD}(\ell_{opt}, P) = R_1^{opt}(P)$, $\mathcal{I}$ be the convex hull of* $\mathrm{proj}(\ell_{opt}, P)$*, $r \geq 0$ be a number such that $R_1^{opt}(P) \leq r \leq 4R_1^{opt}(P)$, $h$ be any 2-flat, and $0 < \beta \leq 1$ be a parameter. We can compute, in $O(nd/\beta^2)$ time, a family of $O(1/\beta^2)$*

*lines on $h$ such that at least one line $\ell$ in the family has the property that for any point $x \in \mathcal{I}$,* $\mathrm{dist}(\ell, x) \le \mathrm{dist}(\mathrm{proj}(h, \ell_{opt}), x) + \beta R_1^{opt}(P)$.

Let $p_\Delta$ be any point of $P$, and let $q_\Delta$ be the furthest point of $P$ from $p$. Let $\ell_0 = \mathrm{affine}(p_\Delta, q_\Delta)$. It is easy to verify that $\mathcal{RD}(\ell_0, P) \le 4R^{opt}$; see Lemma 5.2 for a proof of a more general result. We compute a sequence of lines $\ell_0, \ldots, \ell_{\mathbf{I}}$, where $\mathbf{I} = c\frac{1}{\varepsilon^3}\log\frac{1}{\varepsilon}$ and $c$ is a sufficiently large constant to be determined below. We describe below how the line $\ell_i$ is computed from line $\ell_{i-1}$ in the $i$-th iteration.

In the $i$-th iteration, let $p_i$ be the point of $P$ furthest away from $\ell_{i-1}$. Let $h_i$ be the 2-flat containing $\ell_{i-1}$ and $p_i$ and $\widehat{\ell_i}$ denote the projection of $\ell_{opt}$ onto $h_i$. Using the algorithm of Lemma 3.1 (setting $r = \mathcal{RD}(\ell_0, P)$), we compute a family of $O(1/\delta^2)$ lines on $h_i$ such that at least one line $\ell$ in the family has the property that for any $x \in \mathcal{I}$, $\mathrm{dist}(\ell, x) \le \mathrm{dist}(\widehat{\ell_i}, x) + \delta R^{opt}$, where $\delta = \varepsilon/4\mathbf{I}$. We present the oracle this family of lines, and the oracle picks a line with this property. It can do this by specifying $O(\log 1/\delta)$ bits. We let $\ell_i$ be the line chosen by the oracle.

At the end of the $k$'th iteration, we return the best line from the sequence $\ell_0, \ldots, \ell_{\mathbf{I}}$. That is, we return the line $\ell$ from the sequence that minimizes $\mathcal{RD}(\ell, P)$. We argue below that $\mathcal{RD}(\ell, P) \le (1+\varepsilon)R^{opt}$ for such a line $\ell$. Let us assume the contrary, that is, $\mathcal{RD}(\ell_i, P) > (1+\varepsilon)R^{opt}$, for each $0 \le i \le \mathbf{I}$. We will derive a contradiction.

**Proof of Correctness.** For $0 \le i \le \mathbf{I}$, $\mathrm{d}_i(x) = ||x\,\mathrm{proj}(\ell_i, x)||$ denote the distance between a point $x \in \mathbb{R}^d$ and $\ell_i$. For each $1 \le i \le \mathbf{I}$ and any $x \in \mathcal{I}$, we have

$$\mathrm{d}_i(x) \le \mathrm{dist}(\widehat{\ell_i}, x) + \delta R^{opt} \le \mathrm{d}_{i-1}(x) + \delta R^{opt}. \tag{1}$$

Intuitively, $\mathrm{d}_1, \mathrm{d}_2, \ldots,$ are almost monotonically decreasing functions converging to the zero function on the points in $\mathcal{I}$.

Clearly, $\mathrm{dist}(u, u') \le R^{opt}$. Since $\mathcal{RD}(\ell_0, P) \le 4R^{opt}$, we have $\mathrm{dist}(\ell_0, u) \le 4R^{opt}$. Thus

$$\mathrm{d}_0(u') = \mathrm{dist}(\ell_0, u') \le \mathrm{dist}(\ell_0, u) + \mathrm{dist}(u, u') \le 5R^{opt}.$$

By a symmetric arguement, we have $\mathrm{d}_0(v') \le 5R^{opt}$. From Lemma 2.4 (i), we conclude that $\mathrm{d}_0(x) \le 5R^{opt}$ for any $x \in \mathcal{I}$. Using (1), it then follows that for $i = 1, \ldots, \mathbf{I}$ and any $x \in \mathcal{I}$, we have $\mathrm{d}_i(x) \le \mathrm{d}_0(x) + i\delta R^{opt} \le 6R^{opt}$.

Spread $\lceil 160/\varepsilon^2 \rceil + 1$ equally spaced points on the segment $\mathcal{I}$, and let $S$ denote this set. We say that a point $z \in S$ is *hit* in the $i$-th iteration, if $\mathrm{d}_{i-1}(z) \ge (\varepsilon/2)R^{opt}$ and $\mathrm{d}_i(z) \le (1-\varepsilon/5)\mathrm{d}_{i-1}(z)$. (Intuitively, every time a point $z$ in $S$ is being hit, the value associated with it $\mathrm{d}_i(z)$ goes down "considerably".) Suppose $z$ has being hit $m$ times till the $j$-th iteration; from (1) we have

$$\mathrm{d}_j(z) \le (1-\varepsilon/5)^m \mathrm{d}_0(z) + \mathbf{I}\delta R^{opt} \le (1-\varepsilon/5)^m 5R^{opt} + \mathbf{I}\delta R^{opt}.$$

Thus, for $m = O((1/\varepsilon)\log\frac{1}{\varepsilon})$, we have $5(1-\varepsilon/5)^m \le \varepsilon/4$ and

$$\mathrm{d}_j(z) \le (\varepsilon/4)R^{opt} + \mathbf{I}\cdot(\varepsilon/(4\mathbf{I}))\cdot R^{opt} \le (\varepsilon/2)R^{opt}.$$

Thus, after $z$ is being hit $m = O((1/\varepsilon)\log\frac{1}{\varepsilon})$ times, it can never be hit again. Note that the bound on $m$ is independent of $\mathbf{I}$.

We claim that in every iteration of the algorithm, at least one point of $S$ is being hit. Indeed, consider $p_i$ and $p_i'$, the projection of $p_i$ into $\ell_{opt}$. Let $c_i$ be the projection of $p_i'$ onto $\ell_i$. We know that $\|p_i'p_i\| \le R^{opt}$, and $\|p_ic_i\| > (1+\varepsilon)R^{opt}$ (since $\mathcal{RD}(\ell_{i-1}, P) > (1+\varepsilon)R^{opt}$). From Lemma 2.5, we conclude that $\mathrm{dist}(p_i', p_ic_i) \le (1-\varepsilon/2)\|p_i'c_i\|$. Thus,

$$\mathrm{d}_i(p_i') \le \mathrm{dist}(p_i', p_ic_i) + \delta R^{opt} \le (1-\varepsilon/2)\left\|p_i'c_i\right\| + \delta R^{opt} = (1-\varepsilon/2)\mathrm{d}_{i-1}(p_i') + \delta R^{opt}.$$

Namely, $\mathrm{d}_i(p'_i)$ is significantly smaller than $\mathrm{d}_{i-1}(p'_i)$. Also, we note that $\mathrm{d}_{i-1}(p'_i) \geq \varepsilon R^{opt}$, as otherwise

$$\mathcal{RD}(\ell_{i-1}, P) = \mathrm{dist}(\ell_{i-1}, p_i) \leq \mathrm{dist}(p'_i, p_i) + \mathrm{dist}(\ell_{i-1}, p'_i) \leq R^{opt} + \varepsilon R^{opt} = (1+\varepsilon) R^{opt}.$$

Thus, we have

$$\mathrm{d}_i(p'_i) \leq (1 - \varepsilon/2)\mathrm{d}_{i-1}(p'_i) + \delta R^{opt} \leq (1 - \varepsilon/2)\mathrm{d}_{i-1}(p'_i) + \delta \frac{\mathrm{d}_{i-1}(p')}{\varepsilon} \leq (1 - \varepsilon/3)\mathrm{d}_{i-1}(p'_i),$$

since $\delta < \varepsilon^2/6$ for $\varepsilon \leq 1$ and sufficiently large $c$.

Let $y \in S$ be a point such that $\|p'_i y\| \leq \varepsilon^2 L/320$. We argue that $y$ is hit in the $i$-th iteration. From Lemma 2.4 (ii), we have

$$|\mathrm{d}_{i-1}(y) - \mathrm{d}_{i-1}(p'_i)| \leq \frac{\|p'_i y\|}{u'v'} \cdot 2\max(\mathrm{d}_{i-1}(u'), \mathrm{d}_{i-1}(v')) \leq \frac{\varepsilon^2 L/320}{L} \cdot 12 R^{opt} \leq (\varepsilon^2/16) R^{opt}.$$

Consequently, we have $\mathrm{d}_{i-1}(y) \geq \mathrm{d}_{i-1}(p'_i) - \frac{\varepsilon^2}{16} R^{opt} \geq \varepsilon R^{opt} - \frac{\varepsilon^2}{16} \geq \frac{15\varepsilon}{16} R^{opt}$.

By applying Lemma 2.4 again, this time to $\mathrm{d}_i(\cdot)$,

$$\begin{aligned}
\mathrm{d}_i(y) &\leq \mathrm{d}_i(p'_i) + \frac{\varepsilon^2}{16} R^{opt} \\
&\leq (1 - \varepsilon/3)\mathrm{d}_{i-1}(p'_i) + \frac{\varepsilon^2}{16} R^{opt} \\
&\leq (1 - \varepsilon/3)(\mathrm{d}_{i-1}(y) + \frac{\varepsilon^2}{16} R^{opt}) + \frac{\varepsilon^2}{16} R^{opt} \\
&\leq (1 - \varepsilon/3)\mathrm{d}_{i-1}(y) + \frac{2\varepsilon^2}{16} R^{opt} \\
&\leq (1 - \varepsilon/3)\mathrm{d}_{i-1}(y) + \frac{2\varepsilon}{15}\frac{15\varepsilon}{16} R^{opt} \\
&\leq (1 - \varepsilon/3)\mathrm{d}_{i-1}(y) + \frac{2\varepsilon}{15}\mathrm{d}_{i-1}(y) \\
&\leq (1 - \varepsilon/5)\mathrm{d}_{i-1}(y).
\end{aligned}$$

Thus, the point $y$ is hit in the $i$-th iteration.

We choose the constant $c$ large enough so that the number of iterations $\mathbf{I} = c\frac{1}{\varepsilon^3}\log\frac{1}{\varepsilon}$ is larger than $m \cdot |S|$. Since a point from $S$ is hit in each of the $\mathbf{I}$ iterations, but each point in $S$ is hit at most $m$ times, we have a contradiction.

**Removing the Oracle.** The algorithm as we described it uses $O(\mathbf{I} \log 1/\delta) = O(1/\varepsilon^3 \log^2 1/\varepsilon)$ bits from the oracle. To remove the dependence on the oracle, we simply try all possible strings of size $O(\mathbf{I} \log 1/delta)$, and execute the algorithm on each of these strings. The overall running time of the resulting algorithm is $n \cdot d \cdot C_\varepsilon$, where $C_\varepsilon = \exp\bigl(O(1/\varepsilon^3 \log^2 1/\varepsilon)\bigr)$. We therefore conclude:

**Theorem 3.2** *Let $P$ be a set of $n$ points in $\mathbb{R}^d$ and $0 < \varepsilon \leq 1$ be a parameter. We can compute a line $\ell$, such that $\mathcal{RD}(\ell, P) \leq (1+\varepsilon) R_1^{opt}(P)$, in $n \cdot d \cdot C_\varepsilon$ time, where $C_\varepsilon = \exp\bigl(O(\frac{1}{\varepsilon^3}\log^2\frac{1}{\varepsilon})\bigr)$.*

**Proof of Lemma 3.1:** We generate a set of $O(1/\beta^2)$ pairs of points $(p_1, p_2)$ as follows. Set $\gamma = \beta/8$. Let $p$ be any point in $P$, and let $p_1$ be any point from a $(5r, \gamma r)$-net around $\mathrm{proj}(h, p)$ on $h$. For each choice of $p_1$, we find the point $q \in P$ whose projection $\mathrm{proj}(h, q)$ is furthest from $p_1$, and we choose $p_2$ from a $(5r, \gamma r)$-net around $\mathrm{proj}(h, q)$ on $h$. For each pair $(p_1, p_2)$ that is generated, we include the line through $p_1$ and $p_2$ in our family.

We now argue that our family has the required line in it. Since projection does not expand distances, we have $\mathrm{dist}(\mathrm{proj}(h, \ell_{opt}), \mathrm{proj}(h, p)) \le r$, so there is a choice $p_1^*$ of $p_1$ such that $\mathrm{dist}(\mathrm{proj}(h, \ell_{opt}), p_1^*) \le \gamma r$. Let $\ell'$ be the translation of $\mathrm{proj}(h, \ell_{opt})$ through $p_1^*$. Let $q^* \in P$ be the point whose projection is furthest from $p_1^*$. Since $\mathrm{dist}(\ell', \mathrm{proj}(h, q^*)) \le r + \gamma r \le 2r$, there exists a choice $p_2^*$ for $p_2$ from the $(5r, \gamma r)$-net around $\mathrm{proj}(h, q^*)$ such that (a) $\mathrm{dist}(\ell', p_2^*) \le \gamma r$, and (b) $||p_1^* p_2^*|| \ge ||p_1^* \,\mathrm{proj}(h, q^*)|| + r$. Let $\ell$ be the line through $p_1^*$ and $p_2^*$. Notice that $\ell = \mathrm{Rot}(\ell', p_1^*, p_2^*)$.

For any $s \in P$, let $s'$ denote $\mathrm{proj}(\ell_{opt}, s)$. Since projection does not expand distances, we have

$$||\,\mathrm{proj}(h, s')p_1^*|| \le ||\,\mathrm{proj}(h, s)p_1^*|| + r \le ||\,\mathrm{proj}(h, q^*)p_1^*|| + r \le ||p_1^* p_2^*||.$$

From Lemma 2.6, we conclude that for any $s \in P$,

$$\begin{aligned} \mathrm{dist}(\ell, \mathrm{proj}(h, s')) &\le \mathrm{dist}(\ell', \mathrm{proj}(h, s')) + \frac{||\,\mathrm{proj}(h, s')p_1^*||}{||p_1^* p_2^*||}\mathrm{dist}(\ell', p_2^*) \\ &\le \mathrm{dist}(\ell', \mathrm{proj}(h, s')) + \mathrm{dist}(\ell', p_2^*) \le 2\gamma r. \end{aligned}$$

Let $x \in \mathcal{I}$ be any point. Since $\mathrm{proj}(h, x)$ lies in the convex hull of the set $\{\mathrm{proj}(h, s') | s \in P\}$, we conclude using Lemma 2.4 (i) that $\mathrm{dist}(\ell, \mathrm{proj}(h, x)) \le 2\gamma r \le \beta R_1^{opt}(P)$, which implies that $\ell$ has the properties claimed in the lemma. ■

# 4 A Lower Bound

Megiddo [Meg90] shows that the problem of determining whether there is a line that intersects a given set of balls is NP-hard. The balls in his construction have the same radius, which implies that the problem of computing the radius of the min-enclosing cylinder ($R_1^{opt}$) of a set of points is also NP-hard. We show below that his construction yields the following hardness of approximation result: Unless P = NP, there is no algorithm that, given a set of $n$ points in $\mathbb{R}^d$ and an $0 < \varepsilon \le 1$, runs in time polynomial in $n$, $d$, and $1/\varepsilon$, and returns a number $r$ such that $R_1^{opt}(P) \le r \le (1+\varepsilon)R_1^{opt}(P)$.

Megiddo gives a reduction from 3CNF-satisfiability. Let $\phi$ be a 3CNF formula with $n$ variables and $m$ clauses in which, without loss of generality, we assume that each clause consists of three distinct variables. Let $x_1, \ldots, x_n$ denote the literals and $E_1, \ldots, E_m$ denote the clauses in $\phi$. Let $d = n + 1$. Let $e_i$ denote the point in $\mathbb{R}^d$ with 1 in the $i$'th co-ordinate and 0 elsewhere. Let $r_d = \sqrt{1 - (1/d)}$, and $\alpha$ be chosen so that

$$(12 - 4/d)\alpha^2 = r_d^2.$$

Let $P_1$ be the set of $2d$ points $\{\pm e_i : 1 \le i \le d\}$. Let $Q$ denote the set $\{(x_1, \ldots, x_d) : x_i = \pm 1/\sqrt{d}\}$ of $2^d$ points. Let $L$ denote the set of lines obtained by considering each point $q \in Q$ and taking the line passing through the origin and $q$. Though $Q$ consists of $2^d$ points, $L$ has only $2^{d-1}$ lines. As shown by Megiddo, $R_1^{opt}(P_1)$ equals $r_d$, and this is attained by exactly the lines in $L$. Megiddo constructs a set of $m$ points $P_2 = \{p^1, \ldots, p^m\}$, one for each clause $E_j$ of $\phi$. The point $p^j = (p_1^j, \ldots, p_{n+1}^j)$ is constructed as follows. The last co-ordinate $p_{n+1}^j$ is set to $3\alpha$. For $1 \le i \le n$, if the variable $x_i$ does not occur in $E_j$, then $p_i^j = 0$; if the literal $x_i$ occurs in $E_j$,then $p_i^j = \alpha$; if the literal $\overline{x}_i$ occurs in $E_j$, then $p_i^j = -\alpha$.[1] Megiddo shows that the following properties hold:

[1]Since $\alpha$ may be irrational, we have to instead use a good rational approximaton that can be represented with a small number of bits. We ignore this technicality here since it can be easily accomadated.

1. For each point $p \in P_2$, $||p|| \leq 1$.

2. If $\phi$ is satisfiable, then there is a line $\ell \in L$ such that $\mathrm{dist}(\ell, p)^2 \leq (12 - 4/d)\alpha^2 = r_d^2$ for each $p \in P_2$. Thus $R_1^{opt}(P_1 \cup P_2) = r_d$ in this case.

3. If $\phi$ is not satisfiable, then for every line $\ell \in L$, there is a point $p \in P_2$ such that $\mathrm{dist}(\ell, p)^2 = 12\alpha^2 > r_d^2$.

We add, for each point $p$ in $P_2$, the point $-p$ into the set $P_2$. It is easy to check that the properties above hold for the new $P_2$. (We need the fact that for a line $\ell$ through the origin, $\mathrm{dist}(\ell, p) = \mathrm{dist}(\ell, -p)$.) Our goal here is to show that in the case where $\phi$ is not satisfiable, $R_1^{opt}(P_1 \cup P_2)$ is significantly larger than $r_d$. We need the following lemma.

**Lemma 4.1** *Let $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$ be unit vectors such that $|x_i - y_i| \leq 1/14d^2$, for each $1 \leq i \leq d$. Let $\ell_x$ (resp. $\ell_y$) denote the line through the origin and the point $x$ (resp. $y$). Let $p$ be a point such that $||p|| \leq 1$. Then $|\mathrm{dist}(\ell_x, p)^2 - \mathrm{dist}(\ell_y, p)^2| \leq 1/7d$.*

*Proof:* Notice that $||x - y|| \leq 1/14d^{3/2} \leq 1/14d$.

$$
\begin{aligned}
|\mathrm{dist}(\ell_x, p)^2 - \mathrm{dist}(\ell_y, p)^2| &= |\langle p, x\rangle^2 - \langle p, y\rangle^2| \\
&= |(\langle p, x\rangle - \langle p, y\rangle)(\langle p, x\rangle + \langle p, y\rangle)| \\
&\leq 2|(\langle p, x\rangle - \langle p, y\rangle)| \\
&= 2|\langle p, x - y\rangle| \\
&\leq 2||p|| * ||x - y|| \\
&\leq 1/7d.
\end{aligned}
$$

■

Suppose that $\phi$ is unsatisfiable. Let $\ell$ denote the line that achieves $R_1^{opt}(P_1 \cup P_2)$. Since $P_1 \cup P_2$ is symmetric, we may assume that $\ell$ passes through the origin. Suppose $\ell$ also passes through the point $z = (z_1, \ldots, z_d)$, where $z$ is a unit vector.

- **Case 1.** Suppose that $||z_i| - 1/\sqrt{d}| \leq 1/14d^2$ for each $1 \leq i \leq d$. Then there is a point $q = (q_1, \ldots, q_d) \in Q$ such that $|q_i - z_i| \leq 1/14d^2$ for each $1 \leq i \leq d$. Let $\ell_q \in L$ be the line through the origin and $q$. Since $\phi$ is unsatisfiable, there exists a point $p \in P_2$ such that $\mathrm{dist}(\ell_q, p)^2 = 12\alpha^2$. From Lemma 4.1,

$$
\begin{aligned}
\mathrm{dist}(\ell, p)^2 &\geq \mathrm{dist}(\ell_q, p)^2 - 1/7d \\
&= 12\alpha^2 - 1/7d \\
&= 12/(12 - 4/d)r_d^2 - 1/7d \\
&\geq (1 + 1/3d)r_d^2 - 1/7d \\
&\geq r_d^2 + 1/6d - 1/7d \\
&= r_d^2 + 1/42d.
\end{aligned}
$$

- **Case 2.** Suppose that $||z_i| - 1/\sqrt{d}| \geq 1/14d^2$. It is easy to see that there must be a $j$ such that such that $z_j^2 \leq 1/d - 1/196d^3$. Now the square of the distance $\mathrm{dist}(\ell, e_j)^2$ of the point $e_j$ from $\ell$ equals $||e_j||^2 - \langle e_j, z\rangle^2 = 1 - z_j^2$, which is at least $r_d^2 + 1/196d^3$.

Thus, $R_1^{opt}(P_1 \cup P_2)^2 \geq r_d^2 + \min(1/42d, 1/196d^3) = r_d^2 + 1/196d^3 \geq r_d^2(1+1/196d^3)$ for sufficiently large $d$. We conclude that $R_1^{opt}(P_1 \cup P_2) \geq r_d(1 + 1/792d^3)$ for sufficiently large $d$.

Let $k = |P_1 \cup P_2|$. Since $k > d$, we have that $R_1^{opt}(P_1 \cup P_2) \geq r_d(1 + 1/792k^3)$ when $\phi$ is not satisfiable, and $R_1^{opt}(P_1 \cup P_2) = r_d$ when $\phi$ is satisfiable.

**Theorem 4.2** *Unless $P = NP$, there is no polynomial time algorithm that, given a point set $P$ in $\mathbb{R}^d$ with $k$ points, returns a number between $R_1^{opt}(P)$ and $(1 + 1/792k^3)R_1^{opt}(P)$.*

**Corollary 4.3** *Unless $P = NP$, there is no algorithm that, given a set of $n$ points in $\mathbb{R}^d$ and an $0 < \varepsilon \leq 1$, runs in time polynomial in $n$, $d$, and $1/\varepsilon$, and returns a number $r$ such that $R_1^{opt}(P) \leq r \leq (1 + \varepsilon)R_1^{opt}(P)$.*

# 5 Minimum Radius $k$-flat

In this section, we describe an efficient algorithm that, given a set $P$ of $n$ points in $\mathbb{R}^d$ and an $0 < \varepsilon \leq 1$, returns a $k$-flat $\mathcal{F}$ such that $\mathcal{RD}(\mathcal{F}, P) \leq (1+\varepsilon)R_k^{opt}(P)$. We first extend, in Section 5.1, the basic machinery used for the case of a line to the more general case of a $k$-flat. In Section 5.2, we use this machinery in a straightforward fashion to achieve an approximation algorithm for the optimal $k$-flat.

## 5.1 Preliminaries

Let $\mathcal{F}^{opt} = \mathcal{F}_k^{opt}(P)$ denote a $k$-flat that realizes $R_k^{opt}(P)$; there can be more than one such flat but we fix an arbitarary one for the rest of this section. For any $p \in P$, let $p'$ denote $\mathrm{proj}(\mathcal{F}^{opt}, p)$; let $P' = \mathrm{proj}(\mathcal{F}^{opt}, P)$. Let $\mathcal{E}_k^{\mathrm{opt}}(P)$ denote the minimum-volume ellipsoid on affine$(P')$ that encloses $P'$; if affine$(P')$ is $j$-dimensional, we are speaking here of the $j$-dimensional volume [GLS88]. Let $\mathcal{B}_k^{\mathrm{opt}}(P)$ denote the minimum bounding box (on affine$(P')$), with axes aligned with those of $\mathcal{E}_k^{\mathrm{opt}}(P)$, that encloses $\mathcal{E}_k^{\mathrm{opt}}(P)$. For the case $k = 1$, the set $\mathcal{I} = \mathcal{CH}(P')$ was a line segment on the optimal line. For $k > 1$, the convex hull $\mathcal{CH}(P')$ does not necessarily look so simple, so the enclosing box $\mathcal{B}_k^{\mathrm{opt}}(P)$ will play the role of $\mathcal{I}$ in the proofs here.

For a point $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and a real number $\alpha$, let $\alpha x$ denote the point $(\alpha x_1, \ldots, \alpha x_d)$. For a set $X \subseteq \mathbb{R}^d$, let $\alpha X$ denote the set $\{\alpha x \mid x \in \mathbb{R}^d\}$. We say that a compact, convex set $X \in \mathbb{R}^d$ is *symmetric* if there exists a point $x \in X$ such that for any $y \in \mathbb{R}^d$, $x + y \in X$ if and only if $x - y \in X$. We call such a point $x$ the *center* of $X$. It is easy to see that a compact, convex, symmetric set has a unique center. For a compact, convex, symmetric set $X$ with center $x$ and a given $\alpha \geq 0$, the *concentric scaling* of $X$ by $\alpha$ is the set $\{x + \alpha(y - x) \mid y \in X\}$. The following lemma is the reason we use $\mathcal{B}_k^{\mathrm{opt}}(P)$ and not any bounding box for $P'$.

**Lemma 5.1** *Let $\mathcal{G}$ be any flat, and let $\mathrm{d} : \mathbb{R}^d \to \mathbb{R}$ be the distance function $\mathrm{d}_{\mathcal{G}}(\cdot)$.*

*(i) Let $\mathcal{B}_{in}$ and $\mathcal{B}_{out}$ be compact, convex, symmetric bodies centered at $v \in \mathbb{R}^d$ such that $\mathcal{B}_{out}$ is the concentric scaling of $\mathcal{B}_{in}$ by a factor of $\alpha > 1$. We have*

$$\max_{x \in \mathcal{B}_{out}} \mathrm{d}(x) \leq \alpha \max_{x \in \mathcal{B}_{in}} \mathrm{d}(x).$$

*(ii) Let $P'$ and $\mathcal{B}_k^{\mathrm{opt}}(P)$ be as defined. We have*

$$\max_{x \in \mathcal{B}_k^{\mathrm{opt}}(P)} \mathrm{d}(x) \leq k^3 \max_{p' \in P'} \mathrm{d}(p').$$

*(iii) For any flat $\mathcal{H}$, we have*

$$\max_{x \in \mathrm{proj}(\mathcal{H}, \mathcal{B}_k^{\mathrm{opt}}(P))} \mathrm{d}(x) \le k^3 \max_{p' \in P'} \mathrm{d}(\mathrm{proj}(\mathcal{H}, p')).$$

*Proof:* (i) Let $z \in \mathcal{B}_{out}$ be a point such that $\mathrm{d}(z) = \max_{x \in \mathcal{B}_{out}} \mathrm{d}(x)$. If $z = v$, the claim is immediate. Otherwise, let $\ell$ be the line through $v$ and $z$, and let its intersection with $\mathcal{B}_{in}$ be the segment $yw$. Let us assume that $||wz|| \le ||yz||$. It is easy to check that $||wz||/||yw|| \le (\alpha - 1)/2$. From Lemma 2.4, we see that

$$\mathrm{d}(z) \le \mathrm{d}(w) + (\alpha - 1)\max(\mathrm{d}(y), \mathrm{d}(w)) \le \alpha \max(\mathrm{d}(y), \mathrm{d}(w)) \le \alpha \max_{x \in \mathcal{B}_{in}} \mathrm{d}(x).$$

(ii) Let $\mathcal{E}_k^{\mathrm{opt}}(P)$ denote the min-enclosing ellipsoid of $P'$, and let $\mathcal{E}$ be the concentric scaling of $\mathcal{E}_k^{\mathrm{opt}}(P)$ by a factor of $1/k^2$. It follows from John's theorem [GLS88, Section 4.6] that $\mathcal{E} \subseteq \mathcal{CH}(P') \subseteq \mathcal{E}_k^{\mathrm{opt}}(P)$. Recall that $\mathcal{B}_k^{\mathrm{opt}}(P)$ is the bounding box of $\mathcal{E}_k^{\mathrm{opt}}(P)$ which is aligned with its main axises. Let $\mathcal{B}_{in}$ be the concentric scaling of $\mathcal{B}_k^{\mathrm{opt}}(P)$ by a factor of $1/k^3$. It is easy to argue that that $\mathcal{B}_{in} \subseteq \mathcal{E}$ and so $\mathcal{B}_{in} \subseteq \mathcal{CH}(P')$. By part (i) above, we have

$$\max_{x \in \mathcal{B}_k^{\mathrm{opt}}(P)} \mathrm{d}(x) \le k^3 \max_{x \in \mathcal{B}_{in}} \mathrm{d}(x) \le k^3 \max_{x \in \mathcal{CH}(P')} \mathrm{d}(x) \le k^3 \max_{p' \in P'} \mathrm{d}(p'),$$

where the last inequality follows from the convexity established in Lemma 2.4 (i).

(iii) We note that $\mathrm{proj}(\mathcal{H}, \mathcal{B}_k^{\mathrm{opt}}(P))$ is a concentric scaling by a factor of $k^3$ of $\mathrm{proj}(\mathcal{H}, \mathcal{B}_{in})$ and that $\mathrm{proj}(\mathcal{H}, \mathcal{B}_{in}) \subseteq \mathcal{CH}(\mathrm{proj}(\mathcal{H}, P')) \subseteq \mathrm{proj}(\mathcal{H}, \mathcal{B}_k^{\mathrm{opt}}(P))$. The claim follows by an argument similar to that in (ii). ∎

The statement of the following lemma is folklore but we include a proof here for completeness. The algorithm can be viewed as a version of Gram-Schmidt orthogonalization carried out for $k$ stages.

**Lemma 5.2** *Given a set $P$ of $n$ points in $\mathbb{R}^d$, and a parameter $k$, we can compute in $O(ndk)$ time a $k$-flat $\mathcal{F}$, such that $\mathcal{RD}(\mathcal{F}, P) \le 2^{k+1} R_k^{opt}(P)$.*

*Proof:* We may assume that affine$(P)$ has dimension at least $k+1$; otherwise, we simply compute and return a $k$-flat that contains $P$. Let $p_1$ be an arbitrary point of $P$, and let $p_i \in P$ be the point such that $\mathcal{RD}(\mathcal{G}_{i-1}, P) = \mathrm{dist}(\mathcal{G}_{i-1}, p_i)$, for $i = 2, \ldots, k+1$, where $\mathcal{G}_{i-1} = \mathrm{affine}(p_1, \ldots, p_{i-1})$. We set $\mathcal{G}_{k+1}$ to be the flat $\mathcal{F}$ that we return. Let $\mathcal{F}_i$ be the $k$-flat that minimizes $\mathcal{RD}(\mathcal{F}, P)$ over all $k$-flats $\mathcal{F}$ that contain $\mathcal{G}_i$. We claim that $\mathcal{RD}(\mathcal{F}_i, P) \le 2^i R_k^{opt}(P)$. For $i = 1$, the claim is trivial, as we can just translate the optimal $k$-flat so that it passes through $p_1$. We have, $\mathcal{RD}(\mathcal{F}_1, P) \le 2R_k^{opt}(P)$.

Assume the claim is correct for $\mathcal{F}_{i-1}$. Let $\mathcal{F}'_i$ be the rotation $\mathrm{Rot}(\mathcal{F}_{i-1}, \mathcal{G}_{i-1}, p_i)$ of $\mathcal{F}_{i-1}$ around $\mathcal{G}_{i-1}$, so it passes through $p_i$. From Lemma 2.6, it follows that for any $q \in P$,

$$\begin{aligned} \mathrm{dist}(\mathcal{F}'_i, q) &\le \mathrm{dist}(\mathcal{F}_{i-1}, q) + \frac{\mathrm{dist}(\mathcal{G}_{i-1}, q)}{\mathrm{dist}(\mathcal{G}_{i-1}, p)} \mathrm{dist}(\mathcal{F}_{i-1}, p) \\ &\le \mathrm{dist}(\mathcal{F}_{i-1}, q) + \mathrm{dist}(\mathcal{F}_{i-1}, p) \le 2\mathcal{RD}(\mathcal{F}_{i-1}, P). \end{aligned}$$

This implies that $\mathcal{RD}(\mathcal{F}'_i, P) \le 2\mathcal{RD}(\mathcal{F}_{i-1}, P)$, which in turn implies that $\mathcal{RD}(\mathcal{F}_i, P) \le 2\mathcal{RD}(\mathcal{F}_{i-1}, P)$. Thus $\mathcal{RD}(\mathcal{F}_{k+1}, P) \le 2^{k+1} R_k^{opt}(P)$. However $\mathcal{F}_{k+1} = \mathcal{G}_{k+1}$, since $\mathcal{G}_{k+1}$ is itself a $k$-flat. It is clear that $\mathcal{G}_{k+1}$ can be computed in $O(ndk)$ time. ∎

The following is a generalization of Lemma 3.1 for $k > 1$.

**Lemma 5.3** *Let $P$, $\mathcal{F}^{opt}$, $\mathcal{B}_k^{\rm opt}(P)$ be as defined above. Given $P$, a $(k+1)$-flat $\mathcal{H}$, a parameter $0 < \beta \leq 1$ and a number $r$ such that $R_k^{opt}(P) \leq r \leq 2^{k+1} R_k^{opt}(P)$, we can compute in $nd\exp(O(k^3\log(1/\beta)))$ time a family of $\exp(O(k^3\log(1/\beta)))$ $k$-flats on $\mathcal{H}$ such that at least one $k$-flat $\widehat{\mathcal{F}}$ in the family has the property that for any $x \in \mathcal{B}_k^{\rm opt}(P)$,*

$$\mathrm{dist}(\widehat{\mathcal{F}}, x) \leq \mathrm{dist}(\mathcal{K}, x) + \beta R_k^{opt}(P),$$

*where $\mathcal{K} = \mathrm{proj}(\mathcal{H}, \mathcal{F}^{opt})$.*

*Proof:* Let $\gamma = \beta/(2^{k+1}k^3(k+1))$. We specify a set of sequences $v_1, \ldots, v_{k+1}$ of points on $\mathcal{H}$ as follows. Let $p_1$ be any point in $P$, and let $v_1$ be chosen from a $(5r, \gamma r)$-net around $\mathrm{proj}(\mathcal{H}, p_1)$ on $\mathcal{H}$. For $2 \leq i \leq k+1$, we choose $v_i$ given the choice of $v_1, \ldots, v_{i-1}$ as follows. Let $\mathcal{G}_{i-1} = \mathrm{affine}(v_1, \ldots, v_{i-1})$. Let $p_i$ be the point in $P$ whose projection onto $\mathcal{H}$ is furthest from $\mathcal{G}_{i-1}$. We choose $v_i$ from a $(5r, \gamma r)$-net around $\mathrm{proj}(\mathcal{H}, p_i)$ on $\mathcal{H}$.

For each choice of $v_1, \ldots, v_{k+1}$, we add the $k$-flat $\mathrm{affine}(v_1, \ldots, v_{k+1})$ to our family. The claims in the lemma about the size of the family and the running time are readily verified.

We now argue that there is a sequence $v_1^*, \ldots, v_{k+1}^*$ such that $\widehat{\mathcal{F}} = \mathrm{affine}(v_1^*, \ldots, v_{k+1}^*)$ has the properties claimed in the lemma. We assume for simplicity that $\mathcal{K} = \mathrm{proj}(\mathcal{H}, \mathcal{F}^{opt})$ is a $k$-flat; it is straightforward to extend this arguement when it is a lower dimensional flat. For any $p \in P$, let $p'$ denote $\mathrm{proj}(\mathcal{F}^{opt}, p)$. Let $v_1^*$ be the point in the $(5r, \gamma r)$-net around $\mathrm{proj}(\mathcal{H}, p_1)$ that is closest to $\mathcal{K}$, and $\mathcal{F}_1$ be the translation of $\mathcal{K}$ through $v_1^*$. Since $\mathrm{dist}(\mathcal{K}, \mathrm{proj}(\mathcal{H}, p_1)) \leq \mathrm{dist}(\mathcal{F}^{opt}, p_1) \leq r$, we hav $\mathrm{dist}(\mathcal{K}, v_1^*) \leq \gamma r$. It follows that for any $p \in P$, $\mathrm{dist}(\mathcal{F}_1, \mathrm{proj}(\mathcal{H}, p')) \leq \gamma r$.

Let $2 \leq i \leq k+1$ and suppose we have constructed a sequence $v_1^*, \ldots, v_{i-1}^*$ and a $k$-flat $\mathcal{F}_{i-1}$ in $\mathcal{H}$ containing $v_1^*, \ldots, v_{i-1}^*$ such that for any $p \in P$, $\mathrm{dist}(\mathcal{F}_{i-1}, \mathrm{proj}(\mathcal{H}, p')) \leq (i-1)\gamma r$. This implies that

$$\begin{aligned} \mathrm{dist}(\mathcal{F}_{i-1}, \mathrm{proj}(\mathcal{H}, p)) &\leq \mathrm{dist}(\mathcal{F}_{i-1}, \mathrm{proj}(\mathcal{H}, p')) \\ &\quad + || \mathrm{proj}(\mathcal{H}, p)\, \mathrm{proj}(\mathcal{H}, p')|| \\ &\leq (i-1)\gamma r + r \leq 2r. \end{aligned}$$

Let $p_i^* \in P$ be the point whose projection $\mathrm{proj}(\mathcal{H}, p_i^*)$ is furthest from $\mathcal{G}_i^* = \mathrm{affine}(v_1^*, \ldots, v_{i-1}^*)$. We choose $v_i^*$ to be any point from the $(5r, \gamma r)$-net around $\mathrm{proj}(\mathcal{H}, p_i^*)$ such that (1) $\mathrm{dist}(\mathcal{F}_{i-1}, v_i^*) \leq \gamma r$ and (2) $\mathrm{dist}(\mathcal{G}_i^*, v_i^*) \geq \mathrm{dist}(\mathcal{G}_i^*, \mathrm{proj}(\mathcal{H}, p_i^*)) + r$. It is not hard to see that such a $v_i^*$ exists since $\mathrm{dist}(\mathcal{F}_{i-1}, \mathrm{proj}(\mathcal{H}, p_i^*)) \leq 2r$. Let $\mathcal{F}_i = \mathrm{Rot}(\mathcal{F}_{i-1}, \mathcal{G}_i^*, v_i^*)$.

For any $p \in P$, we have

$$\begin{aligned} \mathrm{dist}(\mathcal{G}_i^*, \mathrm{proj}(\mathcal{H}, p')) &\leq \mathrm{dist}(\mathcal{G}_i^*, \mathrm{proj}(\mathcal{H}, p)) + r \\ &\leq \mathrm{dist}(\mathcal{G}_i^*, \mathrm{proj}(\mathcal{H}, p_i^*)) + r \\ &\leq \mathrm{dist}(\mathcal{G}_i^*, v_i^*). \end{aligned}$$

From Lemma 2.6, we conclude that for any $p \in P$,

$$\begin{aligned} \mathrm{dist}(\mathcal{F}_i, \mathrm{proj}(\mathcal{H}, p')) &\leq \mathrm{dist}(\mathcal{F}_{i-1}, \mathrm{proj}(\mathcal{H}, p')) + \frac{\mathrm{dist}(\mathcal{G}_i^*, \mathrm{proj}(\mathcal{H}, p'))}{\mathrm{dist}(\mathcal{G}_i^*, v_i^*)} \mathrm{dist}(\mathcal{F}_{i-1}, v_i^*) \\ &\leq \mathrm{dist}(\mathcal{F}_{i-1}, \mathrm{proj}(\mathcal{H}, p')) + \mathrm{dist}(\mathcal{F}_{i-1}, v_i^*) \leq i\gamma r. \end{aligned}$$

We conclude that for any $p \in P$, $\mathrm{dist}(\widehat{\mathcal{F}}, \mathrm{proj}(\mathcal{H}, p')) \leq (k+1)\gamma r$, where $\widehat{\mathcal{F}} = \mathcal{F}_{k+1} = \mathrm{affine}(v_1^*, \ldots, v_{k+1}^*)$. Applying Lemma 5.1 (iii), we have that for any $x \in \mathcal{B}_k^{\rm opt}(P)$,

$$\mathrm{dist}(\widehat{\mathcal{F}}, \mathrm{proj}(\mathcal{H}, x)) \leq k^3 \max_{p \in P} \mathrm{dist}(\widehat{\mathcal{F}}, \mathrm{proj}(\mathcal{H}, p')) \leq k^3(k+1)\gamma r \leq \beta R_k^{opt}(P),$$

which implies the lemma. ■

**Definition 5.4** A sequence of distance functions $\mathrm{d}_0, \ldots, \mathrm{d}_{\mathbf{I}}$, where $\mathrm{d}_i : \mathbb{R}^d \to \mathbb{R}$ is *$\mathcal{U}$-bounded*, if $\forall x \in \mathcal{B}_k^{\mathrm{opt}}(P)$, we have $\mathrm{d}_i(x) \le \mathcal{U}$ for $i = 0, \ldots, \mathbf{I}$.

**Lemma 5.5** *Let $\mathrm{d}_0, \ldots, \mathrm{d}_{\mathbf{I}}$ be a sequence of $\mathcal{U}$-bounded distance functions, and let $\mathcal{C}_\varepsilon = 3\varepsilon^2 R_k^{opt}(P)/(200\mathcal{U})$, where $0 < \varepsilon \le 1$. Let $x \in \mathcal{B}_k^{\mathrm{opt}}(P)$ be a point and $i$ an integer, such that $\mathrm{d}_i(x) \le (1 - \varepsilon/3)\mathrm{d}_{i-1}(x)$ and $\mathrm{d}_{i-1}(x) \ge \varepsilon R_k^{opt}(P)$. Let $\mathcal{B}$ be the translate of $\mathcal{C}_\varepsilon \mathcal{B}_k^{\mathrm{opt}}(P)$ centered at $x$. Then for any $z \in \mathcal{B}$, we have (i) $\mathrm{d}_{i-1}(z) \ge (\varepsilon/2)R_k^{opt}(P)$ and (ii) $\mathrm{d}_i(z) \le (1 - \varepsilon/5)\mathrm{d}_{i-1}(z)$.*

*Proof:* Consider any $z \in \mathcal{B}$. Let us assume that that $z \ne x$, for otherwise the claim is immediate. Let $\ell$ be the line through $x$ and $z$ and let the the segment $wy$ denote the intersection of $\ell$ with $\mathcal{B}'$, the concentric scaling by 2 of $\mathcal{B}_k^{\mathrm{opt}}(P)$. It is easy to check that $||xz||/||wy|| \le \mathcal{C}_\varepsilon$. From Lemma 5.1, we see that $\mathrm{d}_{i-1}(w), \mathrm{d}_{i-1}(y) \le 2\mathcal{U}$. We obtain from Lemma 2.4 that

$$|\mathrm{d}_{i-1}(z) - \mathrm{d}_{i-1}(x)| \le \frac{||xz||}{||wy||} 2\max(\mathrm{d}_{i-1}(w), \mathrm{d}_{i-1}(y)) \le 4\mathcal{C}_\varepsilon \mathcal{U}.$$

Consequently,

$$\mathrm{d}_{i-1}(z) \ge \mathrm{d}_{i-1}(x) - 4\mathcal{C}_\varepsilon\mathcal{U} = \varepsilon R_k^{opt}(P) - (12\varepsilon^2/200)R_k^{opt}(P) \ge (9\varepsilon/10)R_k^{opt}(P).$$

which implies (i). By a similar application of Lemma 2.4 to $\mathrm{d}_i(\cdot)$, we get

$$\begin{aligned}
\mathrm{d}_i(z) &\le \mathrm{d}_i(x) + 4\mathcal{C}_\varepsilon\mathcal{U} \\
&\le (1 - \varepsilon/3)\mathrm{d}_{i-1}(x) + 4\mathcal{C}_\varepsilon\mathcal{U} \\
&\le (1 - \varepsilon/3)(\mathrm{d}_{i-1}(z) + 4\mathcal{C}_\varepsilon\mathcal{U}) + 4C_\varepsilon\mathcal{U} \\
&\le (1 - \varepsilon/3)\mathrm{d}_{i-1}(z) + 8\mathcal{C}_\varepsilon\mathcal{U} \\
&\le (1 - \varepsilon/3)\mathrm{d}_{i-1}(z) + \frac{2\varepsilon}{15} \cdot \frac{9\varepsilon R_k^{opt}(P)}{10} \\
&\le (1 - \varepsilon/3)\mathrm{d}_{i-1}(z) + (2\varepsilon/15)\mathrm{d}_{i-1}(z) \\
&\le (1 - \varepsilon/5)\mathrm{d}_{i-1}(z)
\end{aligned}$$

■

## 5.2 The Algorithm

We compute, in $O(ndk)$ time, a $k$-flat $\mathcal{F}_0$ such that $\mathcal{RD}(\mathcal{F}_0, P) \le 2^{k+1}R_k^{opt}(P)$, using the algorithm of Lemma 5.2. We compute a sequence of $k$-flats $\mathcal{F}_0, \ldots, \mathcal{F}_{\mathbf{I}}$, where

$$\mathbf{I} = \frac{\exp(ck^2)}{\varepsilon^{2k+1}} \log\frac{1}{\varepsilon},$$

and $c$ is a sufficiently large constant to be determined below. We describe below how the flat $\mathcal{F}_i$ is computed from $\mathcal{F}_{i-1}$ in the $i$-th iteration.

In the $i$-th iteration, we compute $p_i$, the point of $P$ furthest away from $\mathcal{F}_{i-1}$. Let $\widehat{\mathcal{F}_i}$ denote the projection of $\mathcal{F}^{opt}$ onto the $(k+1)$-flat $\mathcal{H}_i$ containing $\mathcal{F}_{i-1}$ and $p_i$. Using the algorithm of Lemma 5.3 (setting $r = \mathcal{RD}(\mathcal{F}_0, P)$), we compute a family of $k$-flats on $\mathcal{H}_i$ such that at least one flat $\mathcal{F}$ in the family has the property that for any $x \in \mathcal{B}_k^{\mathrm{opt}}(P)$, $\mathrm{dist}(\mathcal{F}, x) \le \mathrm{dist}(\widehat{\mathcal{F}_i}, x) + \delta R_k^{opt}(P)$, where $\delta = \varepsilon/4\mathbf{I}$. We present the oracle with this family, and the oracle picks the flat from the family with the above property. It can do this by specifying $O(k^3 \log 1/\delta)$ bits. Let $\mathcal{F}_i$ be the $k$-flat chosen by the oracle.

At the end of the $\mathbf{I}$-th iteration, we return the best $k$-flat from the sequence $\mathcal{F}_0, \ldots, \mathcal{F}_{\mathbf{I}}$. That is, we return the flat $\mathcal{F}$ from the sequence that minimizes $\mathcal{RD}(\mathcal{F}, P)$. We argue below that $\mathcal{RD}(\mathcal{F}, P) \leq (1+\varepsilon)R_k^{opt}(P)$ for such a flat $\mathcal{F}$. Let us assume the contrary, that is, $\mathcal{RD}(\mathcal{F}_i, P) > (1+\varepsilon)R_k^{opt}(P)$, for each $0 \leq i \leq \mathbf{I}$. We will derive a contradiction.

**Proof of Correctness.** Let $\mathrm{d}_i(x)$ denote the distance $\mathrm{dist}(\mathcal{F}_i, x)$ of a point $x \in \mathbb{R}^d$ from $\mathcal{F}_i$. For each $1 \leq i \leq \mathbf{I}$ and any $x \in \mathcal{B}_k^{\mathrm{opt}}(P)$, we have

$$\mathrm{d}_i(x) \leq \mathrm{dist}(\widehat{\mathcal{F}_i}, x) + \delta R^{opt} \leq \mathrm{d}_{i-1}(x) + \delta R^{opt}.$$

In particular, this implies that $\mathrm{d}_i(x) \leq \mathrm{d}_0(x) + i\delta R^{opt} \leq \mathrm{d}_0(x) + R^{opt}$. On the other hand, we have that for any $x \in \mathcal{B}_k^{\mathrm{opt}}(P)$,

$$\begin{aligned} \mathrm{d}_0(x) &\leq k^3 \max_{p \in P} \mathrm{dist}(\mathcal{F}_0, p') \\ &\leq k^3 \max_{p \in P} \mathrm{dist}(\mathcal{F}_0, p) + ||pp'|| \\ &\leq k^3(2^{k+1}+1)R^{opt}. \end{aligned}$$

It follows that for $i = 1, \ldots, \mathbf{I}$ and $x \in \mathcal{B}_k^{\mathrm{opt}}(P)$, $\mathrm{d}_j(x) \leq (k^3(2^{k+1}+1)+1)R^{opt} \leq k^3 2^{k+2} R^{opt}$. Thus, $\mathrm{d}_1, \ldots, \mathrm{d}_{\mathbf{I}}$ is a $(k^3 2^{k+2} R_k^{opt}(P))$-bounded sequence (see Definition 5.4).

We partition $\mathcal{B}_k^{\mathrm{opt}}(P)$ into a grid $\mathcal{G}$, where each grid cell is a copy of $\mathcal{C}_\varepsilon \mathcal{B}_k^{\mathrm{opt}}(P)$, where $\mathcal{C}_\varepsilon = 3\varepsilon^2/(200 \cdot 2^{k+2}k^3)$. Let $S$ be the set of vertices of $\mathcal{G}$. Clearly, $|S| = O\big((200 \cdot 2^{k+2}k^3)/\varepsilon^2\big)^k = e^{O(k^2)}/\varepsilon^{2k}$. $S$ has the property that any translate of $\mathcal{C}_\varepsilon \mathcal{B}_k^{\mathrm{opt}}(P)$ centered at a point in $\mathcal{B}^{\mathrm{opt}}$ intersects some point in $S$. We say that a point $z \in S$ is *hit* in the $i$-th iteration if $\mathrm{d}_{i-1}(z) \geq (\varepsilon/2)R_k^{opt}(P)$ and $\mathrm{d}_i(z) \leq (1-\varepsilon/5)\mathrm{d}_{i-1}(z)$.

Suppose that a point $z \in S$ has being hit $m$ times until the $j$-th iteration; we have

$$\mathrm{d}_j(z) \leq (1-\varepsilon/5)^m \mathrm{d}_0(z) + \mathbf{I}\delta R^{opt} \leq (1-\varepsilon/5)^m k^3 2^{k+2} R^{opt} + \mathbf{I}\delta R^{opt}.$$

Thus, for $m = O(\frac{k}{\varepsilon}\log\frac{1}{\varepsilon})$, we have

$$\mathrm{d}_j(z) \leq (\varepsilon/4)R^{opt} + \mathbf{I} \cdot (\varepsilon/(4\mathbf{I})) \cdot R^{opt} \leq (\varepsilon/2)R^{opt}.$$

Thus, after $z$ is hit $m = O(\frac{k}{\varepsilon}\log\frac{1}{\varepsilon})$ times, it can never be hit again. Note that the bound on $m$ is independent of $\mathbf{I}$.

We now argue that for each $1 \leq i \leq \mathbf{I}$, some point of $S$ is hit in the $i$-th iteration. Using an arguement identical to the one for lines, we conclude that $\mathrm{d}_i(p'_i) \leq (1-\varepsilon/3)\mathrm{d}_{i-1}(p'_i)$ and $\mathrm{d}_{i-1}(p'_i) \geq \varepsilon R^{opt}$. (Recall that $p'_i = \mathrm{proj}(\mathcal{F}^{opt}, p_i)$.) Let $\mathcal{B}$ denote the translate of $\mathcal{C}_\varepsilon \mathcal{B}_k^{\mathrm{opt}}(P)$ centered at $p'_i$. As noted above, some point $z \in S$ lies in $\mathcal{B}$. Now Lemma 5.5 tells us that $z$ is hit in the $i$-th iteration.

We choose $c$ large enough so that the number of iterations $\mathbf{I}$ is larger than $m \cdot |S|$. Since a point from $S$ is hit in each of the $\mathbf{I}$ iterations, but each point in $S$ is hit at most $m$ times, we have a contradiction.

**Removing the Oracle.** The algorithm as we described it uses $O(k^3 \log(4\mathbf{I}/\varepsilon))$ bits from the oracle in each iteration, and therefore $O(\mathbf{I}k^3 \log(4\mathbf{I}/\varepsilon)) = \frac{e^{O(k^2)}}{\varepsilon^{2k+1}} \log^2 \frac{1}{\varepsilon}$ bits overall. To remove the dependence on the oracle, we simply try all possible strings of size $\frac{e^{O(k^2)}}{\varepsilon^{2k+1}} \log^2 \frac{1}{\varepsilon}$, and execute the algorithm on each of these strings. The overall running time of the resulting algorithm is $nd \cdot D_\varepsilon$, where $D_\varepsilon = \exp\left(\frac{e^{O(k^2)}}{\varepsilon^{2k+1}} \log^2 \frac{1}{\varepsilon}\right)$. We therefore conclude:

**Theorem 5.6** *Given a set $P$ of $n$ points in $\mathbb{R}^d$ and a parameter $0 < \varepsilon \leq 1$, we can compute, in $n \cdot d \cdot \exp\left(\frac{e^{O(k^2)}}{\varepsilon^{2k+1}} \log^2 \frac{1}{\varepsilon}\right)$ time, a $k$-flat $\mathcal{F}$ such that $\mathcal{RD}(\mathcal{F}, P) \leq (1+\varepsilon)R_k^{opt}(P)$.*

**Remark 5.7** In estimating the running time, we have not explicitly kept track of the number of bits needed to represent the numbers involved and have assumed that all simple arithmetic operations can be performed in unit time. We note however that the maximum number of bits needed to represent any such number depends only on $\varepsilon$, $k$, and the maximum number of bits needed to represent any input number. In particular, the number of bits is independent of $n$ and $d$. A careful scrutiny tells us that the bound on the running time claimed in Theorem 5.6 remains valid even if we take into account the bit complexity of the arithmetic operations.

# 6 Conclusions

We have presented a linear-time approximation algorithm for fitting low-dimensional flats in high dimensions. This is a substantial improvement over previous algorithms. Currently, the bottleneck in extending this approach to fitting a small number of flats seems to be that we do not know how to do this in low dimensions. Also, unlike previous algorithms, it can not be used directly for fitting a $k$-flat in the presence of outliers. These are open questions for further research. In particular, there is a huge gap between our result, and the fastest algorithms for those more general problems [HV02]. Is this gap inherent to the problems, or just a byproduct of our techniques? Namely, can outliers or multiple flats be handled in high dimensions in near-linear time?

# Acknowledgments

The authors wish to thank Ken Clarkson, Piotr Indyk, and Piyush Kumar for helpful discussions concerning the problems studied in this paper.

# References

[BC02] M. Bădoiu and K. L. Clarkson. Smaller core-sets for balls. In *Proc. 14th ACM-SIAM Sympos. Discrete Algorithms*, 250–257, 2003.

[BGKvL90] H. L. Bodlaender, P. Gritzmann, V. Klee, and J. van Leeuwen. The computational complexity of norm-maximization. *Combinatorica*, 10:203–225, 1990.

[BH01] G. Barequet and S. Har-Peled. Efficiently approximating the minimum-volume bounding box of a point set in three dimensions. *J. Algorithms*, 38:91–109, 2001.

[BHI02] M. Bădoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *Proc. 34th Annu. ACM Sympos. Theory Comput.*, pages 250–257, 2002.

[Bri02] A. Brieden. On geometric optimization problems likely not contained in APX. To appear, 2002.

[Ede87] H. Edelsbrunner. *Algorithms in Combinatorial Geometry.* EATCS Monographs on Theoretical Computer Science. Springer-Verlag, Heidelberg, 1987.

[FKS96] U. Faigle, W. Kern, and M. Streng. Note on the computational complexity of $j$-radii of polytopes in $\mathbb{R}^n$. *Mathematical Programming*, 73:1–5, 1996.

[GK93] P. Gritzmann and V. Klee. Computational complexity of inner and outer $j$-radii of polytopes in finite-dimensional normed spaces. *Math. Program.*, 59:163–213, 1993.

[GK94] P. Gritzmann and V. Klee. On the complexity of some basic problems in computational convexity: I. containment problems. *Discrete Math.*, 136:129–174, 1994.

[GLS88] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics.* Springer-Verlag, Berlin Heidelberg, 2nd edition, 1988. 2nd edition 1994.

[HV01] S. Har-Peled and K. R. Varadarajan. Approximate shape fitting via linearization. In *Proc. 42nd Annu. IEEE Sympos. Found. Comput. Sci.*, pages 66–73, 2001.

[HV02] S. Har-Peled and K. R. Varadarajan. Projective clustering in high dimensions using core-sets. In *Proc. 18th Annu. ACM Sympos. Comput. Geom.*, pages 312–318, 2002.

[KMY03] P. Kumar, J. S. B. Mitchell, and E. A. Yildirim. Computing core-sets and approximate smallest enclosing hyperspheres in high dimensions. In *Proc. ALENEX*, 45–55, 2003.

[Mat02] J. Matoušek. *Lectures on Discrete Geometry.* Springer, 2002.

[Meg90] N. Megiddo. On the complexity of some geometric problems in unbounded dimension. *J. Symb. Comput.*, 10:327–334, 1990.

[Nes98] Y. Nesterov. Global quadratic optimization via conic relaxation. Technical report, Catholic University of Louvaine, Belgium, 1998.

[NRT99] A. Nemirovski, C. Roos, and T. Terlaky. On maximization of quadratic forms over intersection of ellipsoids with common center. *Mathematical Programming*, 86(3):463–473, 1999.

[VVZ02] K. R. Varadarajan, S. Venkatesh, and J. Zhang. Approximating the radii of point sets in high dimensions. In *Proc. 43th Annu. IEEE Sympos. Found. Comput. Sci.*, 2002.