

# How to Get Close to the Median Shape\*

Sariel Har-Peled<sup>†</sup>

March 10, 2006

*They sought it with thimbles, they sought it with care;  
They pursued it with forks and hope;  
They threatened its life with a railway-share;  
They charmed it with smiles and soap.  
– The Hunting of the Snark, Lewis Carol*

## Abstract

In this paper, we study the problem of  $L_1$ -fitting a shape to a set of  $n$  points in  $\mathbb{R}^d$  (where  $d$  is a fixed constant), where the target is to minimize the sum of distances of the points to the shape, or alternatively the sum of squared distances. We present a general technique for computing a  $(1 + \varepsilon)$ -approximation for such a problem, with running time  $O(n + \text{poly}(\log n, 1/\varepsilon))$ , where  $\text{poly}(\log n, 1/\varepsilon)$  is a polynomial of constant degree of  $\log n$  and  $1/\varepsilon$  (the power of the polynomial is a function of  $d$ ). This is a linear time algorithm for a fixed  $\varepsilon > 0$ , and is the first subquadratic algorithm for this problem.

Applications of the algorithm include best fitting either a circle, a sphere or a cylinder to a set of points when minimizing the sum of distances (or squared distances) to the respective shape.

## 1 Introduction

Consider the problem of fitting a parameterized shape to given data. This is a natural problem that arises in statistics, learning, data-mining and many other fields. What measure is being used for the quality of fitting has considerable impact of the hardness of the problem of finding the best fitting shape. As a concrete example, let  $P$  be a set of  $n$  points in  $\mathbb{R}^d$ . A typical criterion for measuring how well a shape  $\gamma$  fits  $P$ , denoted as  $\mu(P, \gamma)$ , is the maximum distance between a point of  $P$  and its nearest point on  $\gamma$ , i.e.,  $\mu(P, \gamma) = \max_{p \in P} d(p, \gamma)$ , where  $d(p, \gamma) = \min_{q \in \gamma} \|p - q\|$ . The extent measure of  $P$  is  $\mu(P) = \min_{\gamma \in \mathcal{F}} \mu(P, \gamma)$ , where  $\mathcal{F}$  is a family of shapes (such as points, lines, hyperplanes, spheres, etc.). For example, the problem of finding the minimum radius sphere (resp. cylinder) enclosing  $P$  is the same as finding the

---

\*Alternative titles for this paper include: “How to stay connected with your inner circle” and “How to compute one circle to rule them all”. The latest version of this paper is available from [Har06].

<sup>†</sup>Department of Computer Science; University of Illinois; 201 N. Goodwin Avenue; Urbana, IL, 61801, USA; sariel@uiuc.edu; <http://www.uiuc.edu/~sariel/>. Work on this paper was partially supported by an NSF CAREER award CCR-0132901.

point (resp. line) that fits  $P$  best, and the problem of finding the smallest width slab (resp. spherical shell, cylindrical shell) is the same as finding the hyperplane (resp. sphere, cylinder) that fits  $P$  best.

A natural way of encoding the fitting information for a given a shape  $\gamma$  for the the points of  $P$ , is by creating a point  $\mathbf{d}(P, \gamma) \in \mathbb{R}^n$ , where the  $i$ th coordinate is the distance of the  $i$ th point of  $P$  from  $\gamma$ . Thus, the shape fitting problem mentioned above (of minimizing the distance to the furthest point to the shape), is to find the shape  $\gamma$  that realizes  $\min_{\gamma \in \mathcal{F}} \|\mathbf{d}(P, \gamma)\|_\infty$ . We will refer to this as the  $L_\infty$ -*shape fitting problem*.

The exact algorithms for best shape fitting are generally expensive, e.g., the best known algorithms for computing the smallest volume bounding box containing  $P$  in  $\mathbb{R}^3$  require  $O(n^3)$  time [O’R85]. Consequently, attention has shifted to developing approximation algorithms [BH01, ZS02]. A general approximation technique was recently developed for such problems by Agarwal *et al.* [AHV04]. This implies among other things that one can approximate the circle that best fit a set of points in the plane in  $O(n + 1/\varepsilon^{O(1)})$  time, where the fitting measure is the maximum distance of the point to the circle (in fact, this special case was handled before by Agarwal *et al.* [AAHS00] and by Chan [Cha02]).

The main problem with the  $L_\infty$ -fitting is its sensitivity to noise and outliers. There are two natural remedies.

The first is to change the target function to be less sensitive to outliers. For example, instead of considering the maximum distance, one can consider the average distance. This is the  $L_1$ -*fitting problem*, and here we would like to compute the shape realizing  $\ell_1(\mathcal{F}, P) = \min_{\gamma \in \mathcal{F}} \|\mathbf{d}(P, \gamma)\|_1 = \min_{\gamma \in \mathcal{F}} \sum_{p \in P} d(p, \gamma)$ . Similarly, in the  $L_2$ -*fitting problem*, one would like to minimize the average squared distances of the points to the shape; namely,  $\ell_2(\mathcal{F}, P) = \min_{\gamma \in \mathcal{F}} \|\mathbf{d}(P, \gamma)\|_2^2 = \min_{\gamma \in \mathcal{F}} \sum_{p \in P} (d(p, \gamma))^2$ . The  $L_2$  fitting problem in the case of a single linear subspace is well understood, and is computed via singular value decomposition (SVD). Fast approximation algorithms are known for this problem; see [FKV04, DRVW06] and references therein. As for the  $L_1$ -fitting of a linear subspace, this problem can be solved using linear programming techniques, in polynomial time in high dimensions, and linear time in constant dimension [YKII88]. Recently, Clarkson gave a faster approximation algorithm for this problem [Cla05] which works via sampling.

The problem seems to be harder once the shape we consider is not a linear subspace. There is considerable work on nonlinear regressions [SW89] (i.e., extension of the  $L_2$  least squares technique) for various shapes, but there seems to be no efficient guaranteed approximation algorithm even for the “easy” problem of  $L_1$ -fitting a circle to the data. The hardness seems to arise from the target function being a sum of terms, each term being an absolute value of a difference of a square root of a polynomial and a radius (see Section 2.1.2). In fact, this is an extension of the Fermat-Weber problem and it seems doubtful that an efficient exact solution would exist for such a problem.

The second approach is to specify a number  $k$  of outliers in advance and find the best shape  $L_\infty$ -fitting all but  $k$  of the input points. Har-Peled and Wang showed that there is a coresset for this problem [HW04], and as such it can be solved in  $O(n + \text{poly}(k, \log n, 1/\varepsilon))$  time, for a large family of shapes. The work of Har-Peled and Wang was motivated by the aforementioned problem of  $L_1$ -fitting a circle to a set of points. (The results of Har-Peled

and Wang were recently improved by Agarwal *et al.* [AHY06], but since the improvement is not significant for our purposes we will stick with the older reference.)

**Our Results.** In this paper, we describe a general technique for computing a  $(1 + \varepsilon)$ -approximate solution to the  $L_1$  and  $L_2$ -fitting problems, for a family of shapes which is well behaved (roughly speaking, those are all the shapes that the technique of Agarwal *et al.* [AHV04] can handle). Our algorithm achieves a running time of  $O(n + \text{poly}(\log n, 1/\varepsilon))$ . As such, this work can be viewed as the counterpart to Agarwal *et al.* [AHV04] work on the approximate  $L_\infty$ -fitting problem. This is the first linear time algorithm for this problem.

The only previous algorithm directly relevant for this result, we are aware of, is due to Har-Peled and Koltun [HK05a] that, in  $O(n^2\varepsilon^{-2} \log^2 n)$  time, approximates the best circle  $L_1$ -fitting a set of points in the plane.

**Comment on running time.** The running time of our algorithms is  $O(n + \text{poly}(\log n, 1/\varepsilon)) = O(n + \text{poly}(1/\varepsilon))$ . However, throughout the paper we use the former (and more explicit) bound to emphasize that the running time of the second stage of our algorithms depends on  $n$ , unlike other geometric approximation algorithms.

**Paper organization.** In Section 2 we introduce some necessary preliminaries. In Section 2.1 the problem is stated formally. In Section 3, we provide a (somewhat bizarre) solution for the one-point  $L_1$ -fitting problem in one dimension (i.e., the one-median problem in one dimension). In Section 4, we show how the problem size can be dramatically reduced. In Section 5, a slow approximation algorithm is described for the problem (similar in nature to the algorithm of [HK05a]). In Section 6, we state our main result and some applications. Conclusions are provided in Section 7.

## 2 Preliminaries

Throughout the paper, we refer to the  $x_d$ -parallel direction in  $\mathbb{R}^d$  as *vertical*. Given a point  $x = (x_1, \dots, x_{d-1})$  in  $\mathbb{R}^{d-1}$ , let  $(x, x_d)$  denote the point  $(x_1, \dots, x_{d-1}, x_d)$  in  $\mathbb{R}^d$ . Each point  $x \in \mathbb{R}^d$  is also a vector in  $\mathbb{R}^d$ . Given a geometric object  $A$ ,  $A + x$  represents the object obtained by translating each point in  $A$  by  $x$ .

A *surface* is a subset of  $\mathbb{R}^d$  that intersects any vertical line in a single point. A *surface patch* is a portion of a surface such that its vertical projection into  $\mathbb{R}^{d-1}$  is a semi-algebraic set of constant complexity, usually a simplex. Let  $A$  and  $B$  be either a point, a hyperplane, or a surface in  $\mathbb{R}^d$ . We say that  $A$  lies *above* (resp. *below*)  $B$ , denoted by  $A \succeq B$  (resp.  $A \preceq B$ ), if for any vertical line  $\ell$  intersecting both  $A$  and  $B$ , we have that  $x_d \geq y_d$  (resp.  $x_d \leq y_d$ ), where  $(x_1, \dots, x_{d-1}, x_d) = A \cap \ell$  and  $(x_1, \dots, x_{d-1}, y_d) = B \cap \ell$ . (In particular, if both  $A$  and  $B$  are hyperplanes then  $A \succeq B$  implies that  $A$  and  $B$  are parallel hyperplanes.)

Two non-negative numbers  $x$  and  $y$  are  $(1 \pm \varepsilon)$ -*approximation* of each other if  $(1 - \varepsilon)x \leq y \leq (1 + \varepsilon)x$  and  $(1 - \varepsilon)y \leq x \leq (1 + \varepsilon)y$ . We denote this fact by  $x \approx_\varepsilon y$ . Two non-negative functions  $f(\cdot)$  and  $g(\cdot)$  (defined over the same domain) are  $(1 \pm \varepsilon)$ -*approximation* of each other, denoted by  $f \approx_\varepsilon g$ , if  $f(x) \approx_\varepsilon g(x)$ , for all  $x$ .

**Observation 2.1** *Let  $x$  and  $y$  be two positive numbers and  $\varepsilon < 1/4$ . We have: (i) If  $x \approx_{\varepsilon} y$  and  $y \approx_{\varepsilon} z$  then  $x \approx_{3\varepsilon} z$ . (ii) If  $|x - y| \leq \varepsilon x$  then  $x \approx_{2\varepsilon} y$ . (iii) If  $x \leq (1 + \varepsilon)y$  and  $y \leq (1 + \varepsilon)x$  then  $x \approx_{\varepsilon} y$ .*

## 2.1 Problem Statement

### 2.1.1 The Circle Fitting Case

To motivate our exposition we will first consider the problem of  $L_1$ -fitting a circle to a set of points in the plane.

Let  $P = \{p_1, \dots, p_n\}$  be a set of  $n$  points in the plane, and consider the price  $\nu_P(C)$  of  $L_1$ -fitting the circle  $C$  to  $P$ . Formally, for a point  $p_i \in P$  let  $f_i(C) = \left| \|p_i - c\| - r \right|$ , where  $c$  is the center of  $C$ , and  $r$  is the radius of  $C$ . Thus, the overall price, for a circle  $C$  centered at  $(x, y)$  with radius  $r$ , is

$$\nu_P(C) = \nu_P(x, y, r) = \sum_{i=1}^n f_i(C) = \sum_{i=1}^n \left| \|p_i - c\| - r \right| = \sum_{i=1}^n \left| \sqrt{(x_i - x)^2 + (y_i - y)^2} - r \right|,$$

where  $p_i = (x_i, y_i)$ , for  $i = 1, \dots, n$ . We are looking for the circle  $C$  minimizing  $\nu_P(C)$ . This is the circle that best fits the point set under the  $L_1$  metric. Let  $\nu_{\text{opt}}(P)$  denote the price of the optimal circle  $C_{\text{opt}}$ .

Geometrically, each function  $f_i$  induces a surface  $\gamma_i = \left\{ (x_p, y_p, \|p - p_i\|) \mid p \in \mathbb{R}^2 \right\}$  in 3D, which is a cone. A circle is encoded by a point  $C = (x, y, r)$ . The value of  $f_i(C)$  is the vertical distance between the point  $C$  and surface  $\gamma_i$ . Thus, we have a set  $\mathcal{G}$  of  $n$  surfaces in 3D, and we are interested in finding the point that minimizes the sum of vertical distances of this point to the  $n$  surfaces.

### 2.1.2 The General Problem

Formally, for a weighted set of surfaces  $\mathcal{G}$  in  $\mathbb{R}^d$  and  $p$  any point in  $\mathbb{R}^d$  let

$$\nu_{\mathcal{G}}(p) = \sum_{\gamma \in \mathcal{G}} w_{\gamma} \cdot \mathbf{d}_{\lvert}(p, \gamma)$$

denote the  $L_1$  distance of  $p$  from  $\mathcal{G}$ , where  $\mathbf{d}_{\lvert}(p, \gamma)$  is the vertical distance between  $p$  and the surface  $\gamma$  and  $w_{\gamma}$  is the weight associated with  $\gamma$ . Throughout our discussion weights are positive integer numbers. If  $\mathcal{G}$  is unweighted then any surface  $\gamma \in \mathcal{G}$  is assigned weight  $w_{\gamma} = 1$ . We would be interested in finding the point that minimizes  $\nu_{\mathcal{G}}(p)$  when  $p$  is restricted to a domain  $\mathcal{D}_d$ , which is a semi-algebraic set of constant complexity in  $\mathbb{R}^d$ . This is the  $L_1$ -fitting problem. The  $L_2$ -fitting problem is computing the point  $p \in \mathcal{D}_d$  realizing the minimum of  $\mu_{\mathcal{G}}(p) = \sum_{\gamma \in \mathcal{G}} w_{\gamma} \cdot (\mathbf{d}_{\lvert}(p, \gamma))^2$ .

It would be sometime conceptually easier (e.g., see Section 6.1.1) to think about the problem algebraically, where the  $i$ th surface  $\gamma_i$  is an image of a (non-negative)  $(d-1)$ -dimensional function  $f_i(x_1, \dots, x_{d-1}) = \sqrt{p_i(x_1, \dots, x_{d-1})}$ , where  $p_i(\cdot)$  is a constant degree polynomial, for  $i = 1, \dots, n$ . We are interested in approximating one of the following quantities:

$$\begin{aligned}
\text{(i)} \quad & \min_{(x_1, \dots, x_{d-1}) \in \mathcal{D}} \sum_{i=1}^n w_i \cdot f_i(x_1, \dots, x_{d-1}), \\
\text{(ii)} \quad & \nu_{\text{opt}}(\mathcal{G}) = \min_{x \in \mathcal{D}_d} \nu_{\mathcal{G}}(x) = \min_{(x_1, \dots, x_d) \in \mathcal{D}_d} \sum_i w_i \cdot |f_i(x_1, \dots, x_{d-1}) - x_d|, \\
\text{or (iii)} \quad & \mu_{\text{opt}}(\mathcal{G}) = \min_{x \in \mathcal{D}_d} \mu_{\mathcal{G}}(x) = \min_{(x_1, \dots, x_d) \in \mathcal{D}_d} \sum_i w_i \cdot (f_i(x_1, \dots, x_{d-1}) - x_d)^2,
\end{aligned}$$

where  $\mathcal{D} \subseteq \mathbb{R}^{d-1}$  and  $\mathcal{D}_d \subseteq \mathbb{R}^d$  are semi-algebraic sets of constant complexity, and the weights  $w_1, \dots, w_n$  are positive integers. Note that (i) is a special case of (ii), by setting  $\mathcal{D}_d = \mathcal{D} \times \{0\}$ .

To simplify the exposition, we will assume that  $\mathcal{D}_d = \mathbb{R}^d$ . It is easy to verify that our algorithm works also for the more general case with a few minor modifications.

**The linearization dimension.** In the following, a significant parameter in the exposition is the *linearization dimension*  $\mathfrak{d}$ , which is the target dimension we need to map the polynomials  $p_1, \dots, p_n$  so that they all become linear functions. For example, if the polynomials are of the form  $\psi_i(x, y, z) = x^2 + y^2 + z^2 + a_i x + b_i y + c_i z$ , for  $i = 1, \dots, n$ , then they can be linearized by a mapping  $\mathbf{L}(x, y, z) = (x^2 + y^2 + z^2, x, y, z)$ , such that  $h_i(x, y, z, w) = w + a_i x + b_i y + c_i z$  is a linear function and  $\psi_i(x, y, z) = h_i(\mathbf{L}(x, y, z))$ . Thus, in this specific example the linearization dimension is 4. The linearization dimension is always bounded by the number of different monomials appearing in the polynomials  $p_1, \dots, p_n$ . Agarwal and Matoušek [AM94] described an algorithm that computes a linearization of the smallest dimension for a family of such polynomials.

### 3 Approximate $L_1$ -Fitting in One Dimension

In this section, we consider the one dimensional problem of approximating the distance function of a point  $z$  to a set of points  $\mathbf{Z} = \langle z_1, z_2, \dots, z_n \rangle$ , where  $z_1 \leq z_2 \leq \dots \leq z_n$ . Formally, we want to approximate the function  $\nu_{\mathbf{Z}}(\bar{z}) = \sum_{z_i \in \mathbf{Z}} |z_i - \bar{z}|$ . This is the one-median function for  $\mathbf{Z}$  on the real line. This corresponds to a vertical line in  $\mathbb{R}^d$ , where each  $z_i$  represents the intersection of the vertical line with the surface  $\gamma_i$ . The one dimensional problem is well understood and there exists a coresets for it; see [HM04, HK05b]. Unfortunately, it is unclear how to extend these constructions to the higher dimensional case; specifically, how to perform the operations required in a global fashion on the surfaces so that the construction would hold for all vertical lines. See Remark 3.5 below for more details on this “hardness”. Thus, we present here an alternative construction.

**Definition 3.1** For a set of weighted surfaces  $\mathcal{G}$  in  $\mathbb{R}^d$ , a weighted subset  $\mathcal{S} \subseteq \mathcal{G}$  is an  $\varepsilon$ -coreset for  $\mathcal{G}$  if for any point  $p \in \mathbb{R}^d$  we have  $\nu_{\mathcal{G}}(p) \approx_{\varepsilon} \nu_{\mathcal{S}}(p)$ .

For the sake of simplicity of exposition, in the following we assume that  $\mathcal{G}$  is unweighted. The weighted case can be handled in a similar fashion.

The first step is to partition the points. Formally, we partition  $\mathbf{Z}$  symmetrically into subsets, such that the sizes of the subsets increase as one comes toward the middle of the set. Formally, the set  $L_i = \{z_i\}$  contains the  $i$ th point on the line, for  $i = 1, \dots, m$ , where

$m \geq 10/\varepsilon$  is a parameter to be determined shortly. Similarly,  $R_i = \{z_{n-i+1}\}$ , for  $i = 1, \dots, m$ . Set  $\alpha_m = m$ , and let  $\alpha_{i+1} = \min(\lceil(1 + \varepsilon/10)\alpha_i\rceil, n/2)$ , for  $i = m, \dots, M$ , where  $\alpha_M$  is the first number in this sequence equal to  $n/2$ . Now, let  $L_i = \{z_{\alpha_{i-1}+1}, \dots, z_{\alpha_i}\}$  and  $R_i = \{z_{n-\alpha_{i-1}}, \dots, z_{n-\alpha_i+1}\}$ , for  $i = m+1, \dots, M$ . We will refer to a set  $L_i$  or  $R_i$  as a *chunk*. Consider the partition of  $\mathbf{Z}$  formed by the chunks  $L_1, L_2, \dots, L_M, R_M, \dots, R_2, R_1$ . This is a partition of  $\mathbf{Z}$  into “exponential sets”. The first/last  $m$  sets on the boundary are singletons, and all the other sets grow exponentially in cardinality, till they cover the whole set  $\mathbf{Z}$ .

Next, we pick arbitrary points  $l_i \in L_i$  and  $r_i \in R_i$  and assign them weight  $w_i = |R_i| = |L_i|$ , for  $i = 1, \dots, M$ . Let  $\mathcal{S}$  be the resulting weighted set of points. We claim that this is a coreset for the 1-median function.

But before delving into this, we need the following technical lemma.

**Lemma 3.2** *Let  $A$  be a set of  $n$  real numbers, and let  $\psi$  and  $\bar{z}$  be any two real numbers. We have that  $\left| \nu_A(\bar{z}) - |A| \cdot |\psi - \bar{z}| \right| \leq \nu_A(\psi)$ .*

*Proof:*  $\left| \nu_A(\bar{z}) - |A| \cdot |\psi - \bar{z}| \right| = \left| \sum_{p \in A} |p - \bar{z}| - |A| \cdot |\psi - \bar{z}| \right| \leq \sum_{p \in A} \left| |\bar{z} - p| - |\psi - \bar{z}| \right| \leq \sum_{p \in A} |p - \psi| = \nu_A(\psi)$ , by the triangle inequality.  $\blacksquare$

**Lemma 3.3** *It holds  $\nu_{\mathbf{Z}}(\bar{z}) \approx_{\varepsilon/5} \nu_{\mathcal{S}}(\bar{z})$ , for any  $\bar{z} \in \mathbb{R}$ .*

*Proof:* We claim that

$$\left| \nu_{\mathbf{Z}}(\bar{z}) - \nu_{\mathcal{S}}(\bar{z}) \right| \leq (\varepsilon/10) \nu_{\mathbf{Z}}(\bar{z}),$$

for all  $\bar{z} \in \mathbb{R}$ . Indeed, let  $\tau$  be a median point of  $\mathbf{Z}$  and observe that  $\nu_{\mathbf{Z}}(\tau)$  is a global minimum of this function. We have that

$$\begin{aligned} \mathcal{E} = \left| \nu_{\mathbf{Z}}(\bar{z}) - \nu_{\mathcal{S}}(\bar{z}) \right| &\leq \sum_{i=1}^M \left| \nu_{L_i}(\bar{z}) - |L_i| \cdot |l_i - \bar{z}| \right| + \sum_{i=1}^M \left| \nu_{R_i}(\bar{z}) - |R_i| \cdot |r_i - \bar{z}| \right| \\ &= \sum_{i=m+1}^M \left| \nu_{L_i}(\bar{z}) - |L_i| \cdot |l_i - \bar{z}| \right| + \sum_{i=m+1}^M \left| \nu_{R_i}(\bar{z}) - |R_i| \cdot |r_i - \bar{z}| \right| \\ &\leq \sum_{i=m+1}^M \nu_{L_i}(l_i) + \sum_{i=m+1}^M \nu_{R_i}(r_i), \end{aligned}$$

by Lemma 3.2.

Observe that by construction  $|R_i| \leq (\varepsilon/10) |R_1 \cup \dots \cup R_{i-1}|$ , for  $i > m$ . We claim that  $\sum_{i=m+1}^M \nu_{L_i}(l_i) + \sum_{i=m+1}^M \nu_{R_i}(r_i) \leq (\varepsilon/10) \nu_{\mathbf{Z}}(\tau)$ . To see this, for each point of  $z_i \in \mathbf{Z}$ , let  $I_i$  be the interval with  $z_i$  in one endpoint and the median  $\tau$  in the other endpoint. The total length of those intervals is  $\nu_{\mathbf{Z}}(\tau)$ . Let  $\mathcal{R} = \{I_1, \dots, I_n\}$ .

Consider the interval  $\mathcal{I}_i = \mathcal{I}(R_i)$  which is the shortest interval containing the points of  $R_i$ , for  $i = m+1, \dots, M$ . Clearly, we have  $\nu_{R_i}(r_i) \leq |R_i| \cdot \|\mathcal{I}_i\|$ .

On the other hand, the number of intervals of  $\mathcal{R}$  completely covering  $\mathcal{I}_i$  is at least  $(10/\varepsilon) |R_i|$ , for  $i = m+1, \dots, M$ . As such, we can charge the total length of  $\nu_{R_i}(r_i)$  to the portions of those intervals of  $\mathcal{R}$  covering  $\mathcal{I}_i$ . Thus, every unit of length of the intervals of  $\mathcal{R}$  gets charged at most  $\varepsilon/10$  units.

This implies that the error  $\mathcal{E} \leq (\varepsilon/10)\nu_{\mathbf{Z}}(\tau) \leq (\varepsilon/10)\nu_{\mathbf{Z}}(\bar{\mathbf{z}})$ , which establishes the lemma, by Observation 2.1.  $\blacksquare$

Next, we “slightly” perturb the points of the coresets  $\mathcal{S}$ . Formally, assume that we have points  $l'_1, \dots, l'_M, r'_1, \dots, r'_M$  such that  $|l'_i - l_i|, |r'_i - r_i| \leq (\varepsilon/20)|l_i - r_i|$ , for  $i = 1, \dots, N$ . Let  $\mathcal{R} = \{l'_1, \dots, l'_M, r'_1, \dots, r'_M\}$  be the resulting weighted set. We claim that  $\mathcal{R}$  is still a good coresets.

**Lemma 3.4** *It holds that  $\nu_{\mathbf{Z}}(\bar{\mathbf{z}}) \approx_{\varepsilon} \nu_{\mathcal{R}}(\bar{\mathbf{z}})$ , for any  $\bar{\mathbf{z}} \in \mathbb{R}$ . Namely,  $\mathcal{R}$  is an  $\varepsilon$ -coresets for  $\mathbf{Z}$ .*

*Proof:* By Lemma 3.3 and by the triangle inequality, we have

$$\nu_{\mathcal{S}}(\bar{\mathbf{z}}) = \sum_i (|L_i| \cdot |l_i - \bar{\mathbf{z}}| + |R_i| \cdot |r_i - \bar{\mathbf{z}}|) \geq \sum_i |L_i| \cdot |l_i - r_i|,$$

since for all  $i$  we have  $|L_i| = |R_i|$ . Also, by the triangle inequality  $||l_i - \bar{\mathbf{z}}| - |l'_i - \bar{\mathbf{z}}|| \leq |l_i - l'_i|$ . Thus

$$\left| \nu_{\mathcal{S}}(\bar{\mathbf{z}}) - \nu_{\mathcal{R}}(\bar{\mathbf{z}}) \right| \leq \sum_i |L_i| \cdot |l_i - l'_i| + \sum_i |R_i| |r_i - r'_i| \leq 2 \sum_i |L_i| \frac{\varepsilon}{20} \cdot |l_i - r_i| \leq \frac{\varepsilon}{10} \nu_{\mathcal{S}}(\bar{\mathbf{z}}).$$

Thus  $\mathcal{R}$  is an  $\varepsilon/5$ -coresets of  $\mathcal{S}$ , which is in turn an  $\varepsilon$ -coresets for  $\mathbf{Z}$ , by Observation 2.1.  $\blacksquare$

**Remark 3.5** The advantage of the scheme used in Lemma 3.4 over the constructions of [HM04, HK05b] is that the new framework is more combinatorial and therefore it is more flexible. In particular, the construction can be done in an oblivious way without knowing (even approximately) the optimal value of the 1-median clustering. This is in contrast to the previous constructions that are based on partition of the line into intervals of prespecified length that depends on the value of the optimal solution. As such, they can not be easily extended to handle noise and approximation. The flexibility of the new construction is demonstrated in the following section.

### 3.1 Variants

Let  $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  be a monotone strictly increasing function (e.g.,  $f(x) = x^2$ ). Consider the function

$$U_{\mathbf{Z}}(\bar{\mathbf{z}}) = \sum_{x \in \mathbf{Z}} f(|x - \bar{\mathbf{z}}|).$$

We claim that the set  $\mathcal{S}$  constructed in Lemma 3.3 is also a coresets for  $U_{\mathbf{Z}}(\cdot)$ . Namely,  $U_{\mathbf{Z}}(\bar{\mathbf{z}}) \approx_{\varepsilon/5} U_{\mathcal{S}}(\bar{\mathbf{z}}) = \sum_{x \in \mathcal{S}} w_x f(|x - \bar{\mathbf{z}}|)$ . To this end, map each point  $x$  of  $\mathbf{Z}$ , to a point of distance  $f(|x - \bar{\mathbf{z}}|)$  from  $\bar{\mathbf{z}}$  (preserving the side of  $\bar{\mathbf{z}}$  on which the point  $x$  lies), and let  $g_{\bar{\mathbf{z}}} : \mathbf{Z} \rightarrow \mathbb{R}$  denote this mapping. Let the resulting set be  $Q = f(\mathbf{Z})$ . Clearly,  $U_{\mathbf{Z}}(\bar{\mathbf{z}}) = \nu_Q(\bar{\mathbf{z}})$ , and let  $\mathcal{T}$  be the coresets constructed for  $Q$  by Lemma 3.3. Observe that  $\mathcal{T} = g_{\bar{\mathbf{z}}}(\mathcal{S})$ , since the construction of the coresets cares only about the ordering of the points, and the ordering is preserved when mapping between  $\mathbf{Z}$  and  $Q$ . Thus, we have that  $U_{\mathbf{Z}}(\bar{\mathbf{z}}) = \nu_Q(\bar{\mathbf{z}}) \approx_{\varepsilon/5} \nu_{\mathcal{T}}(\bar{\mathbf{z}}) = U_{\mathcal{S}}(\bar{\mathbf{z}})$ , as required.

This in particular implies that  $\mu_{\mathbf{Z}}(\bar{\mathbf{z}}) \approx_{\varepsilon/5} \mu_{\mathcal{S}}(\bar{\mathbf{z}})$ , for any  $\bar{\mathbf{z}} \in \mathbb{R}$ , where  $\mu_{\mathbf{Z}}(\bar{\mathbf{z}}) = \sum_{x \in \mathbf{Z}} |\bar{\mathbf{z}} - x|^2$ . In this case, even the modified coresets  $\mathcal{R}$  is still a coresets.

**Lemma 3.6** *It holds that  $\mu_{\mathbf{Z}}(\bar{\mathbf{z}}) \approx_{\varepsilon} \mu_{\mathcal{R}}(\bar{\mathbf{z}})$ , for any  $\bar{\mathbf{z}} \in \mathbb{R}$ . Namely,  $\mathcal{R}$  is an  $\varepsilon$ -coreset of  $\mathbf{Z}$  for the  $\mu(\cdot)$  function.*

*Proof:* Observe that, by the above discussion,  $\mu_{\mathbf{Z}}(\bar{\mathbf{z}}) \approx_{\varepsilon/5} \mu_{\mathcal{S}}(\bar{\mathbf{z}})$ . On the other hand, fix  $\bar{\mathbf{z}} \in \mathbb{R}$ , and assume that  $|l_i - \bar{\mathbf{z}}| < |r_i - \bar{\mathbf{z}}|$ . This implies that  $|r_i - \bar{\mathbf{z}}| \geq |l_i - r_i|/2$ , and we have

$$\begin{aligned} |l'_i - \bar{\mathbf{z}}|^2 + |r'_i - \bar{\mathbf{z}}|^2 &\leq (|l'_i - l_i| + |l_i - \bar{\mathbf{z}}|)^2 + (|r'_i - r_i| + |r_i - \bar{\mathbf{z}}|)^2 \\ &\leq ((\varepsilon/10)|r_i - \bar{\mathbf{z}}| + |l_i - \bar{\mathbf{z}}|)^2 + ((\varepsilon/10)|r_i - \bar{\mathbf{z}}| + |r_i - \bar{\mathbf{z}}|)^2 \\ &\leq (\varepsilon^2/100)|r_i - \bar{\mathbf{z}}|^2 + (\varepsilon/5)|r_i - \bar{\mathbf{z}}||l_i - \bar{\mathbf{z}}| \\ &\quad + |l_i - \bar{\mathbf{z}}|^2 + (1 + \varepsilon/10)^2|r_i - \bar{\mathbf{z}}|^2 \leq (1 + \varepsilon/3)(|l_i - \bar{\mathbf{z}}|^2 + |r_i - \bar{\mathbf{z}}|^2), \end{aligned}$$

since  $|l'_i - l_i|, |r'_i - r_i| \leq (\varepsilon/20)|l_i - r_i|$ . This implies that  $\mu_{\mathcal{S}}(\bar{\mathbf{z}}) \leq (1 + \varepsilon/3)\mu_{\mathcal{R}}(\bar{\mathbf{z}})$ . By applying the same argument in the other direction, we have that  $\mu_{\mathcal{S}}(\bar{\mathbf{z}}) \approx_{\varepsilon/3} \mu_{\mathcal{R}}(\bar{\mathbf{z}})$ , by Observation 2.1 (iii). This in turn implies that  $\mu_{\mathcal{R}}(\bar{\mathbf{z}}) \approx_{\varepsilon} \mu_{\mathbf{Z}}(\bar{\mathbf{z}})$ , as required.  $\blacksquare$

## 4 The Reduction

In this section, we show how to reduce the problem of approximating the  $\nu_{\mathcal{G}}(\cdot)$  function, for a set  $\mathcal{G}$  of  $n$  (unweighted) surfaces in  $\mathbb{R}^d$ , to the problem of approximating the same function for a considerably smaller set of surface patches.

Section 3 provides us with a general framework for how to get a small approximation. Indeed, pick any vertical line  $\ell$ , and consider its intersection points with the surfaces of  $\mathcal{G}$ . Clearly, the function  $\nu_{\mathcal{G}}(\cdot)$  restricted to  $\ell$  can be approximated using the construction of Section 3. To this end, we need to pick levels in the way specified and assign them the appropriate weights. This would guarantee that the resulting function would approximate  $\nu_{\mathcal{G}}(\cdot)$  everywhere.

A major difficulty in pursuing this direction is that the levels we pick have high descriptive complexity. We circumnavigate this difficulty in two stages. In the first stage, we replace those levels by shallow levels, by using random sampling. In the second stage, we approximate these shallow levels such that this introduces small relative error.

**Definition 4.1** For a set  $\mathcal{G}$  of  $n$  surfaces in  $\mathbb{R}^d$ , the *level* of a point  $x \in \mathbb{R}^d$  in the arrangement  $\mathcal{A}(\mathcal{G})$  is the number of surfaces of  $\mathcal{G}$  lying vertically below  $x$ . For  $k = 0, \dots, n-1$ , let  $\mathbf{L}_{\mathcal{G},k}$  represent the surface which is closure of all points on the surfaces of  $\mathcal{G}$  whose level is  $k$ . We will refer to  $\mathbf{L}_{\mathcal{G},k}$  as the *bottom  $k$ -level* or just the  *$k$ -level* of  $\mathcal{G}$ . We define the *top  $k$ -level* of  $\mathcal{G}$  to be  $\mathbf{U}_{\mathcal{G},k} = \mathbf{L}_{\mathcal{G},n-k-1}$ , for  $k = 0, \dots, n-1$ . Note that  $\mathbf{L}_{\mathcal{G},k}$  is a subset of the arrangement of  $\mathcal{G}$ . For  $x \in \mathbb{R}^{d-1}$ , we slightly abuse notation and define  $\mathbf{L}_{\mathcal{G},k}(x)$  to be the value  $x_d$  such that  $(x, x_d) \in \mathbf{L}_{\mathcal{G},k}$ .

**Lemma 4.2** *Let  $\mathcal{G}$  be a set of  $n$  surfaces in  $\mathbb{R}^d$ ,  $0 < \delta < 1/4$ , and let  $k$  be a number between 0 and  $n/2$ . Let  $\zeta = \min(ck^{-1}\delta^{-2} \log n, 1)$ , and pick each surface of  $\mathcal{G}$  into a random sample  $\Psi$  with probability  $\zeta$ , where  $c$  is an appropriate constant. Then, with high probability, the*

$\tilde{\kappa}$ -level of  $\mathcal{A}(\Psi)$  lies between the  $(1 - \delta)k$ -level to the  $(1 + \delta)k$ -level of  $\mathcal{A}(\mathcal{G})$ , where  $\tilde{\kappa} = \zeta k = O(\delta^{-2} \log n)$ .

In other words, we have  $\mathbf{L}_{\mathcal{G},(1-\delta)k} \preceq \mathbf{L}_{\Psi,\tilde{\kappa}} \preceq \mathbf{L}_{\mathcal{G},(1+\delta)k}$  and  $\mathbf{U}_{\mathcal{G},(1+\delta)k} \preceq \mathbf{U}_{\Psi,\tilde{\kappa}} \preceq \mathbf{U}_{\mathcal{G},(1-\delta)k}$ .

*Proof:* This follows readily from the Chernoff inequality, and the proof is provided only for the sake of completeness.

Consider a vertical line  $\ell$  passing through a point  $p \in \mathbb{R}^{d-1}$ , and let  $X_i$  be an indicator variable which is 1 if the  $i$ th surface intersecting  $\ell$  (from the bottom) was chosen to the random sample  $\Psi$ . Let  $Y = \sum_{i=1}^{(1+\delta)k} X_i$  be the random variable which is the level of the point  $(p, \mathbf{L}_{\mathcal{G},(1+\delta)k}(p))$  in the resulting arrangement  $\mathcal{A}(\Psi)$ .

Let  $\mu = \mathbf{E}[Y] = \zeta(1 + \delta)k$ . We have by the Chernoff inequality that

$$\Pr[Y < \zeta k] \leq \Pr[Y < (1 - \delta/2)\mu] \leq \exp(-\mu\delta^2/8) = \exp\left(-\frac{(1 + \delta)c}{8} \log n\right) = \frac{1}{n^{O(1)}},$$

by choosing  $c$  to be large enough. There are only  $n^{O(1)}$  combinatorially different orderings of the surfaces of  $\mathcal{G}$  along a vertical line. As such, we can make sure that, with high probability, the  $\tilde{\kappa}$  level in  $\Psi$  (which is just a surface) lies below the  $(1 + \delta)k$  level of  $\mathcal{G}$ .

Similar argument shows that, with high probability, the  $(1 - \delta)k$  level of  $\mathcal{G}$  lies below the  $\tilde{\kappa}$  level of  $\Psi$ .  $\blacksquare$

Lemma 4.2 suggests that instead of picking a specific level in a chunk of levels, as done in Section 3, we can instead pick a level, which is a shallow level of the appropriate random sample, and with high probability this level lies inside the allowable range. The only problem is that even this shallow level might (and will) have unreasonable complexity. We rectify this by doing direct approximation of the shallow levels.

**Definition 4.3** Let  $\mathcal{G}$  be a set of surfaces in  $\mathbb{R}^d$ . The  $(k, r)$ -extent  $\mathcal{G}|_r^k : \mathbb{R}^{d-1} \rightarrow \mathbb{R}$  is defined as the vertical distance between the bottom  $r$ -level and the top  $k$ -level of  $\mathcal{A}(\mathcal{G})$ , i.e., for any  $x \in \mathbb{R}^{d-1}$ , we have

$$\mathcal{G}|_r^k(x) = \mathbf{U}_{\mathcal{G},k}(x) - \mathbf{L}_{\mathcal{G},r}(x).$$

**Definition 4.4** ([HW04]) Let  $\mathcal{F}$  be a set of non-negative functions defined over  $\mathbb{R}^{d-1}$ . A subset  $\mathcal{F}' \subseteq \mathcal{F}$  is  $(k, \varepsilon)$ -sensitive if for any  $r \leq k$  and  $x \in \mathbb{R}^{d-1}$ , we have

$$\mathbf{L}_{\mathcal{F},r}(x) \leq \mathbf{L}_{\mathcal{F}',r}(x) \leq \mathbf{L}_{\mathcal{F},r}(x) + \frac{\varepsilon}{2} \mathcal{F}|_r^k(x); \quad \text{and}$$

$$\mathbf{U}_{\mathcal{F},r}(x) - \frac{\varepsilon}{2} \mathcal{F}|_k^r(x) \leq \mathbf{U}_{\mathcal{F}',r}(x) \leq \mathbf{U}_{\mathcal{F},r}(x).$$

We need the following result of Har-Peled and Wang [HW04]. It states that for well behaved set of functions, one can find a small subset of the functions such that the vertical extent of the subset approximates the extents of the whole set. This holds only for “shallow” levels  $\leq k$ . In our application  $k$  is going to be about  $O(\varepsilon^{-2} \log n)$ . Here is the legalese:

**Theorem 4.5** ([HW04]) Let  $\mathcal{F} = \{p_1^{1/2}, \dots, p_n^{1/2}\}$  be a family of  $d$ -variate functions defined over  $\mathbb{R}^d$ , where  $p_i$  is a  $d$ -variate polynomial, for  $i = 1, \dots, n$ . Given  $k$  and  $0 < \varepsilon < 1$ ,

one can compute, in  $O(n + k/\varepsilon^{2\mathfrak{d}})$  time, a subset  $\mathcal{F}' \subseteq \mathcal{F}$ , such that, with high probability,  $\mathcal{F}'$  is  $(k, \varepsilon)$ -sensitive for  $\mathcal{F}$ , and  $|\mathcal{F}'| = O(k/\varepsilon^{2\mathfrak{d}})$ , where  $\mathfrak{d}$  is the linearization dimension of the polynomials of  $\mathcal{F}$ .

Intuitively, Theorem 4.5 states that shallow levels of depth at most  $k$ , has approximation of size polynomial in  $k$  and  $1/\varepsilon$ , and matching bottom/top  $k$  levels have their mutual distances preserved up to a small multiplicative factor.

**The construction.** We partition the levels of  $\mathcal{A}(\mathcal{G})$  into chunks, according to the algorithm of Section 3, setting  $\mathfrak{m} = O((\log n)/\varepsilon^2)$ . The first top/bottom  $\mathfrak{m}$  levels of  $\mathcal{A}(\mathcal{G})$  we approximate directly by computing a set  $\mathcal{S}_0$  which is  $(\mathfrak{m}, \varepsilon/20)$ -sensitive for  $\mathcal{G}$ , using Theorem 4.5. Next, compute the  $i$ th bottom (resp., top) level of  $\mathcal{A}(\mathcal{S}_0)$ , for  $i = 0, \dots, \mathfrak{m}$ , and let  $\gamma_i$  (resp.,  $\eta_i$ ) denote those levels. We assign weight one for each such surface.

For every pair of chunks of levels  $L_i$  and  $R_i$  from Section 3, for  $i = \mathfrak{m} + 1, \dots, \mathfrak{M}$ , we compute an appropriate random sample  $\Psi_i$ . We remind the reader that  $L_i$  spans the range of levels from  $\alpha_{i-1} + 1$  to  $(1 + \varepsilon/10)\alpha_{i-1}$ ; see Section 3. As such, if we want to find a random level that falls inside this range, we need to set  $\delta = \varepsilon/40$  and  $k = (1 + \varepsilon/20)\alpha_{i-1}$ , and now apply Lemma 4.2, which results in a random set  $\Psi_i$ , such that level  $l_i = O(\varepsilon^{-2} \log n)$  of  $\mathcal{A}(\Psi_i)$  lies between level  $\alpha_{i-1} + 1$  and  $(1 + \varepsilon/10)\alpha_{i-1}$  of  $\mathcal{A}(\mathcal{G})$ . We now approximate the top  $l_i$ -level and bottom  $l_i$ -level of  $\mathcal{A}(\Psi_i)$  by applying Theorem 4.5 to  $\Psi_i$ . This results in a set  $\mathcal{S}_i$  of size  $O(\text{poly}(\log n, 1/\varepsilon))$  of surfaces, such that the extent of the top/bottom  $l_i$  levels of  $\mathcal{A}(\mathcal{S}_{i-M})$ , is an  $(1 \pm \varepsilon/40)$ -approximation to the extent of the top/bottom  $l_i$  levels in  $\mathcal{A}(\Psi_i)$ . We extract the bottom  $l_i$  level and top  $l_i$  level of  $\mathcal{A}(\Psi_i)$ . Let the two resulting surfaces be denoted by  $\gamma_i$  and  $\eta_i$ , respectively, and assign them weight  $|R_i|$ , for  $i = \mathfrak{m} + 1, \dots, \mathfrak{M}$ .

Note that  $\gamma_i$  and  $\eta_i$  no longer have constant complexity, but their complexity is bounded by  $O(\text{poly}(\log n, 1/\varepsilon))$ . Let  $\mathcal{H} = \{\gamma_1, \eta_1, \dots, \gamma_{\mathfrak{M}}, \eta_{\mathfrak{M}}\}$  be the resulting set of weighted surfaces, and observe that the complexity of the arrangement  $\mathcal{A}(\mathcal{H})$  is  $O(\text{poly}(\log n, 1/\varepsilon))$ . Furthermore, the analysis of Section 3 implies that  $\nu_{\mathcal{G}}(p) \approx_{\varepsilon} \nu_{\mathcal{H}}(p)$ , for any point  $p \in \mathbb{R}^d$ .

**Implementation details.** To get a linear running time, we need to carefully implement the above algorithm. First, observe that we computed  $O(\varepsilon^{-1} \log n)$  random samples  $\Psi_{\mathfrak{m}+1}, \dots, \Psi_{\mathfrak{M}}$ . Observe that if two random samples are generated by sampling every surface with probabilities which are similar (up to a factor of two), then we can just use the same random sample. Thus, we need to generate random samples only for probabilities which are powers of two (implying that only  $O(\log n)$  random samples are needed). In particular, let  $\Upsilon_i$  be a random sample generated by picking each surface of  $\mathcal{G}$  with probability  $1/2^i$ .

To perform this sampling quickly we generate the  $(i + 1)$ th random sample by picking each surface of  $\Upsilon_i$  into  $\Upsilon_{i+1}$  with probability half (the sequence of random samples  $\mathcal{G} = \Upsilon_0, \Upsilon_1, \dots, \Upsilon_{O(\log n)}$  is sometimes referred to as a *gradation*). Namely, each  $\Upsilon_i$  serves as a replacement for a sequence of random samples  $\Psi_{\alpha}, \dots, \Psi_{\beta}$  which were generated using similar probabilities, where  $\alpha$  and  $\beta$  are a function of  $i$ .

Next, we need to approximate the “shallow” levels of  $\Psi_i$  up to level  $\xi = O(\max(l_{j_i}, \dots, l_{j_{i+1}-1})) = O(\varepsilon^{-2} \log n)$ . Again, we are performing the computation of the shallow levels for a batch of samples of  $\Psi$  using a single sample of  $\Upsilon$  (i.e., will approximate the top/bottom  $O(\xi)$ -levels of  $\Upsilon_i$  and this would supply us with the surfaces approximating all the required levels

in  $\Psi_\alpha, \dots, \Psi_\beta$ ). Using Theorem 4.5, this takes  $O(|\Upsilon_i| + \text{poly}(\log n, 1/\varepsilon))$  time. By the Chernoff inequality, with high probability, we have  $|\Upsilon_i| = O(n/2^i)$ . Thus the overall running time, with high probability, is  $\sum_i O(n/2^i + \text{poly}(\log n, 1/\varepsilon)) = O(n + \text{poly}(1/\varepsilon, \log n))$ . Putting everything together, we have:

**Theorem 4.6** *Given a set of  $n$  unweighted surfaces  $\mathcal{G}$  in  $\mathbb{R}^d$ , as in Section 2.1.2, and a parameter  $\varepsilon$ , one can compute a set  $\mathcal{H}$  of surface patches, such that each patch is a portion of a surface of  $\mathcal{G}$  which is defined over a region in  $\mathbb{R}^{d-1}$  (such a region is a semi-algebraic set of constant descriptive complexity). The number of surface patches is  $O(\text{poly}(1/\varepsilon, \log n))$ . Furthermore,  $\nu_{\mathcal{G}}(p) \approx_{\varepsilon} \nu_{\mathcal{H}}(p)$  and  $\mu_{\mathcal{G}}(p) \approx_{\varepsilon} \mu_{\mathcal{H}}(p)$ , for any point  $p \in \mathbb{R}^d$ . The algorithm takes  $O(n + \text{poly}(\log n, 1/\varepsilon))$  time and it succeeds with high probability.*

*The total weight of surfaces interesting any vertical line  $\ell$  is equal to  $|\mathcal{G}|$ .*

The algorithm of Theorem 4.6 is a Monte-Carlo algorithm. In particular, it might fail with low probability. It is not clear if there is an efficient way to detect such a (low probability) failure.

Theorem 4.6 shows that given an instance of any of the problems defined in Section 2.1.2, we can quickly reduce the problem size to a small weighted set of surface patches. This, while beneficial, still leaves us with the mundane task of solving the problem on the reduced instance. Since we no longer have to care too much about efficiency the problem becomes more manageable and we tackle it in the next section.

## 5 A Slow Approximation Algorithm

Let  $\mathcal{G}$  be a set of  $n$  weighted surface patches in  $\mathbb{R}^d$ , such that any vertical line intersects surfaces with total weight  $W$ . In this section, we show how to solve any of the problems of Section 2.1.2. A (roughly) quadratic time algorithm for the special case of a circle was given by Har-Peled and Koltun [HK05a], and the algorithm described here is somewhat similar to their algorithm. We demonstrate our algorithm for the case of approximating  $\nu_{\mathcal{G}}(p) = \sum_{\gamma \in \mathcal{G}} w_\gamma \cdot \mathbf{d}_|(p, \gamma)$ .

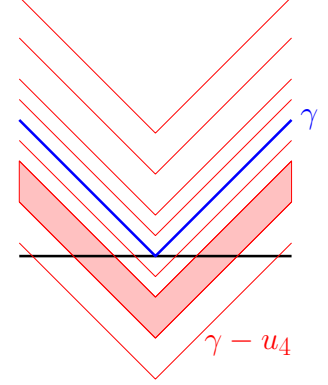
Let  $\Phi$  be the decomposition of  $\mathbb{R}^{d-1}$  into constant complexity cells, such that for each cell  $\Delta \in \Phi$ , we have that any two vertical lines in  $\mathbb{R}^d$  intersecting  $\Delta$  cross the same set of surface patches of  $\mathcal{G}$ . Thus,  $\Phi$  induces a decomposition of  $\mathbb{R}^d$  into vertical prisms, such that we have to solve our problem inside each such prism. The number of different prisms is  $O(n^{2\mathfrak{d}})$ , where  $\mathfrak{d}$  is the linearization dimension of  $\mathcal{G}$ . Now, we need to solve the problem inside each prism, for a weighted set of *surfaces* (instead of surface patches).

So, consider a cell  $\Delta \in \Phi$  and let  $\mathfrak{N}$  denote the vertical prism that has  $\Delta$  for a base. Let  $\mathcal{H}$  be the resulting set of surfaces active in  $\mathfrak{N}$ . We compute, in  $O(n)$  time, a vertical segment  $\sigma \subseteq \mathfrak{N}$  that stabs all the surfaces of  $\mathcal{H}$ , and its length is at most twice the length of the shortest vertical segment that intersect all the surfaces of  $\mathcal{H}$  inside  $\mathfrak{N}$ . This can be done by the algorithm of [AHV04].

The basic idea is to replace the “unfriendly” distance function  $\mathbf{d}_|(p, \gamma)$ , associated with  $\gamma \in \mathcal{H}$ , appearing in  $\nu_{\mathcal{G}}(p)$  by its level-sets. Namely, for each term in the summation of  $\nu_{\mathcal{G}}(p)$  we will generate several level-sets, such that instead of computing  $\mathbf{d}_|(p, \gamma)$ , we will use

the relevant level-set value. Somewhat imprecisely, the level-set of  $d_1(p, \gamma)$  is a surface and the value associated with the region between two consecutive level-sets will be the value  $d_1(\cdot, \gamma)$  on the higher level-set. This process is somewhat similar to height contours used in drawing topographical maps. Since every level-set is a surface, this induces an arrangement of surfaces. For any point  $p \in \mathbb{R}^d$ , we can now compute  $\nu_{\mathcal{G}}(p)$  by just figuring out in between what level-sets  $p$  lies. Therefore, evaluating  $\nu_{\mathcal{G}}(p)$  is reduced to performing a point-location query in arrangement of surfaces and returning the value associated with the face containing  $p$ .

Note that  $\|\sigma\|/2$  is a lower bound for the value of  $\nu_{\mathcal{H}}(\cdot)$  in this prism and  $W \cdot \|\sigma\|$  is an upper bound on the value of  $\nu_{\text{opt}} = \min_{p \in \mathbb{B}} \nu_{\mathcal{H}}(p)$ , where  $W = w_{\mathcal{H}} \geq n$ , where  $w_{\mathcal{H}}$  denotes the total weight of the surfaces of  $\mathcal{H}$ . As such, let  $u = \varepsilon\|\sigma\|/(10W^2)$ . Next, let  $u_i = iu$ , for  $i = 1, \dots, m = O(1/\varepsilon)$ . Let  $u_i = (1 + \varepsilon/20)u_{i-1}$ , for  $i = m + 1, \dots, M = O(\log_{1+\varepsilon/20} W) = O(\varepsilon^{-1} \log W)$ . Note that  $u_M > W^2\|\sigma\| > 2\nu_{\text{opt}}$ . Thus, if a point  $p \in \mathbb{B}$  is in distance more than  $u_M$  away from any surface of  $\mathcal{H}$ , then its distance from this single surface is enough to ensure that  $\nu_{\mathcal{H}}(p) > 2\nu_{\text{opt}}$ . In particular, for every surface  $\gamma \in \mathcal{H}$ , we create  $O(M)$  copies of it as follows:



$\gamma - u_M, \gamma - u_{M-1}, \dots, \gamma - u_1, \gamma, \gamma + u_1, \dots, \gamma + u_M$ , where  $\gamma + x$  is the result of translating the surface  $\gamma$  up by distance  $x$ . Let  $\mathcal{J}(\gamma)$  denote the resulting “stack” (i.e., set) of surfaces.

The surfaces of  $\mathcal{J}(\gamma)$  partition the (domain of interest in the) prism into regions where the function  $d_1(p, \gamma)$  is the same up to a multiplicative factor of  $(1 + \varepsilon/20)$ . The only region that fails to comply with this condition is the region in between  $\gamma - u_1$  and  $\gamma + u_1$ . Thus, if we approximate the value of  $d_1(p, \gamma)$  by the value of this function on the surface in this stack just above  $p$ , we get an  $(1 \pm \varepsilon/20)$ -approximation of this function, except for the region between  $\gamma - u_1$  and  $\gamma + u_1$ .

In particular, let  $\mathcal{J} = \cup_{\gamma \in \mathcal{H}} \mathcal{J}(\gamma)$ . Consider any point  $p \in \mathbb{B}$ , and let  $a_i = d_1(p, \gamma_i)$ , where  $\gamma_i \in \mathcal{H}$ , for  $i = 1, \dots, n$ , such that  $|a_i| \leq u_M$ . Also, let  $b_i$  be the maximum of the values associated with the surfaces just above and below  $p$  in the stack  $\mathcal{J}(\gamma_i)$ . Thus, we have that

$$\begin{aligned} \nu_{\mathcal{H}}(p) &\leq \overline{\nu}_{\mathcal{H}}(p) = \sum_i b_i \leq \sum_i \left( u + \left(1 + \frac{\varepsilon}{20}\right) a_i \right) = n \cdot \frac{\varepsilon\|\sigma\|}{10W^2} + \left(1 + \frac{\varepsilon}{20}\right) \sum_i a_i \\ &\leq \left(1 + \frac{\varepsilon}{5}\right) \nu_{\mathcal{H}}(p), \end{aligned}$$

since  $\nu_{\mathcal{H}}(p) \geq \|\sigma\|$ .

Thus, let  $\mathcal{A}$  be the arrangement  $\mathcal{A}(\mathcal{J})$ , and compute for every face  $F$  of  $\mathcal{A}$  the value of  $\overline{\nu}_{\mathcal{H}}(p)$ , where  $p \in F$ . By the above argument, the point  $p_{\text{opt}}$  realizing  $\min_{x \in \mathbb{B}} \nu_{\mathcal{H}}(x)$  is approximated correctly by  $\overline{\nu}_{\mathcal{H}}(p_{\text{opt}})$ . As such, any point inside the face of  $\mathcal{A}$  with the lowest associated value of  $\overline{\nu}_{\mathcal{H}}$  is the required approximation.

It is easy to verify that the same algorithm with minor modifications would also enable us to approximate the minimum of the mean function  $\mu_{\mathcal{G}}(\cdot)$ . We conclude:

**Theorem 5.1** *Let  $\mathcal{G}$  be a set of  $n$  weighted surface patches in  $\mathbb{R}^d$ , with linearization dimension  $\mathfrak{d}$ , such that any vertical line intersects surfaces with total weight  $W$ . Let  $\varepsilon > 0$  be a*

parameter. Then one can compute, in  $O(n^{3d+1}\varepsilon^{-d}\log^d W)$  time, a point  $x \in \mathbb{R}^d$ , such that  $\nu_{\mathcal{G}}(x) \leq (1 + \varepsilon)\nu_{\text{opt}}(\mathcal{G})$ , where  $\nu_{\text{opt}}(\mathcal{G}) = \min_{x \in \mathbb{R}^d} \nu_{\mathcal{G}}(x)$ .

One can also compute, in the same time complexity, a point  $y \in \mathbb{R}^d$ , such that  $\mu_{\mathcal{G}}(y) \leq (1 + \varepsilon)\mu_{\text{opt}}(\mathcal{G})$ , where  $\mu_{\text{opt}}(\mathcal{G}) = \min_{x \in \mathbb{R}^d} \mu_{\mathcal{G}}(x)$ .

*Proof:* The correctness follows from the above discussion. As for the running time, there are  $O(n^{2d})$  prisms. In each prism we have at most  $n$  surfaces, and every surface get replicated  $O(\varepsilon^{-1} \log W)$  times. The complexity of the arrangement inside each prism is  $O((n\varepsilon^{-1} \log W)^{d+1})$ . A careful implementation would require time proportional to the complexity of all those arrangements, which is  $O(n^{3d+1}\varepsilon^{-d}\log^d W)$ , as claimed. ■

## 6 The Main Result and Some Applications

By plugging Theorem 4.6 into Theorem 5.1, we get the main result of this paper:

**Theorem 6.1** *Given a set of  $n$  unweighted surfaces  $\mathcal{G}$  in  $\mathbb{R}^d$ , as defined in Section 2.1.2, and a parameter  $0 < \varepsilon < 1/4$ , then one can compute, in  $O(n + \text{poly}(\log n, 1/\varepsilon))$  time, a point  $x \in \mathbb{R}^d$ , such that  $\nu_{\mathcal{G}}(x) \leq (1 + \varepsilon)\nu_{\text{opt}}(\mathcal{G})$ , where  $\nu_{\text{opt}}(\mathcal{G}) = \min_{x \in \mathbb{R}^d} \nu_{\mathcal{G}}(x)$ .*

*One can also compute, in the same time complexity, a point  $y \in \mathbb{R}^d$ , such that  $\mu_{\mathcal{G}}(y) \leq (1 + \varepsilon)\mu_{\text{opt}}(\mathcal{G})$ , where  $\mu_{\text{opt}}(\mathcal{G}) = \min_{x \in \mathbb{R}^d} \mu_{\mathcal{G}}(x)$ .*

*The algorithm is randomized and succeeds with high probability.*

### 6.1 Applications

The discussion Section 2.1.1 implies that we can readily apply Theorem 6.1 to the problem of  $L_1$ -fitting a circle to a set of points in the plane. Note, that in fact the same reduction would work for the  $L_2$ -fitting problem, and for fitting a sphere to points in higher dimensions. We conclude:

**Theorem 6.2 ( $L_1/L_2$ -fitting points to a circle/sphere)** *Let  $P$  be a set of  $n$  points in  $\mathbb{R}^d$ , and  $\varepsilon > 0$  a parameter. One can  $(1 + \varepsilon)$ -approximate the sphere best fitting the points of  $P$ , where the price is the sum of Euclidean distances of the points of  $P$  to the sphere. The running time of the algorithm is  $O(n + \text{poly}(\log n, 1/\varepsilon))$ , and the algorithm succeeds with high probability.*

*Similarly, one can  $(1 + \varepsilon)$ -approximate the sphere that minimizes the sum of square distances of the points to the sphere.*

To our knowledge, Theorem 6.2 is the first subquadratic algorithm for this problem. A roughly quadratic time algorithm for the problem of  $L_1$ -fitting a circle to points in the plane was provided by Har-Peled and Koltun [HK05a].

#### 6.1.1 $L_1/L_2$ -Fitting a cylinder to a point-set

Let  $P = \{p_1, \dots, p_n\}$  be a set of  $n$  points in  $\mathbb{R}^d$ ,  $\ell$  be a line in  $\mathbb{R}^d$  parameterized by a point  $q \in \ell$ , and a direction  $\vec{v}$  on the unit sphere  $\mathbb{S}^{(n)} \subseteq \mathbb{R}^d$ , and let  $r$  be the radius of the cylinder

having  $\ell$  as its center. We denote by  $\mathfrak{C} = \mathfrak{C}(q, \vec{v}, r)$  the cylinder having  $\ell = \cup_{t \in \mathbb{R}} (q + t\vec{v})$  as its center. For a point  $p_i \in P$ , we have that its distance from  $\mathfrak{C}$  is

$$f_i(q, \vec{v}, r) = d(p_i, \mathfrak{C}) = \left| \|p_i - q - \langle p_i - q, \vec{v} \rangle \vec{v}\| - r \right| = \left| \sqrt{p_i(q, \vec{v}, r)} - r \right|,$$

where  $p_i(q, \vec{v}, r)$  is a polynomial with linearization dimension  $O(d^4)$  (as can be easily verified), for  $i = 1, \dots, n$ . The linearization dimension in this case can be reduced with more care, see [AHV04]. Thus, the overall price of fitting  $\mathfrak{C}$  to the points of  $P$  is  $\sum_i f_i(\mathfrak{C})$ . This falls into our framework, and we get:

**Theorem 6.3** *Let  $P$  be a set of  $n$  points in  $\mathbb{R}^d$ , and  $\varepsilon > 0$  a parameter. One can  $(1 + \varepsilon)$ -approximate the cylinder that best fits the points of  $P$ , where the price is the sum of Euclidean distances of the points of  $P$  to the cylinder. The running time of the algorithm is  $O(n + \text{poly}(\log n, 1/\varepsilon))$ , and the algorithm succeeds with high probability.*

*Similarly, one can  $(1 + \varepsilon)$ -approximate the cylinder that minimizes the sum of square distances of the points of  $P$  to the cylinder.*

Interestingly, in two dimensions, an algorithm similar to the one in Theorem 6.3 solves the problem of finding two parallel lines that minimizes the sum of distances of the points to the lines (i.e., each point contributes its distance to the closer of the two lines).

### 6.1.2 1-Median Clustering of Partial Data

Consider an input of  $n$  points  $p_1, \dots, p_n$ , where the points are not explicitly given. Instead, we are provided with a set  $\mathfrak{F} = \{f_1, \dots, f_n\}$  of  $n$  flats, such that  $p_i \in f_i$ , where a *flat* is an affine subspace of  $\mathbb{R}^d$ . This naturally arises when we have partial information about a point and the point must comply with certain linear constraints that define its flat.

It is now natural to want to best fit or cluster the partial data. For example, we might wish to compute the smallest ball that encloses all the partial points. This boils down to computing the smallest ball  $\mathbf{b}$  that intersects all the flats (i.e., we assume the real point  $p_i$  lies somewhere in the intersection of the ball  $\mathbf{b}$  and  $f_i$ ). An approximation algorithm for this problem that has polynomial dependency on the dimension  $d$  (but bad dependency on the dimensions of the flats) was recently published by Gao *et al.* [GLS06].

Here, we are interested in finding the point  $\mathbf{c}$  that minimizes the sum of distances of the point  $\mathbf{c}$  to the flats  $f_1, \dots, f_n$ . Namely, this is the 1-median clustering problem for partial data.

Consider a flat  $f$  which contains the point  $q$ , and is spanned by the unit vectors  $\vec{v}_1, \dots, \vec{v}_k$ . That is  $f = \left\{ q + t_1 \vec{v}_1 + \dots + t_k \vec{v}_k \mid t_1, \dots, t_k \in \mathbb{R} \right\}$ . Then, we have that the distance of  $p \in \mathbb{R}^d$  from the flat  $f$  is

$$d(p, f) = \left\| p - q - \sum_{i=1}^k \langle p - q, \vec{v}_i \rangle \vec{v}_i \right\| = \sqrt{\psi(p)},$$

where  $\psi(\cdot)$  is a polynomial with linearization dimension  $O(d^2)$ . Thus, the problem of 1-median clustering of partial data is no more than finding the point  $p$  that minimizes the

function  $\nu_{\mathfrak{F}}(p) = \sum_i d(p, f_i)$ . Approximating the minimum of this function can be done using Theorem 6.1. We conclude:

**Theorem 6.4** *Let  $\mathfrak{F} = \{f_1, \dots, f_n\}$  be a set of  $n$  flats in  $\mathbb{R}^d$ , and  $\varepsilon > 0$  a parameter. One can compute a point  $p \in \mathbb{R}^d$ , such that  $\nu_{\mathfrak{F}}(p)$  is a  $(1 + \varepsilon)$ -approximation to  $\min_q \nu_{\mathfrak{F}}(q)$ . The running time of the algorithm is  $O(n + \text{poly}(\log n, 1/\varepsilon))$ , and the algorithm succeeds with high probability.*

Note that 1-mean clustering in this case is trivial as it boils down to a minimization of a quadratic polynomial.

## 7 Conclusions

In this paper, we have described in this paper a general approximation technique for problems of  $L_1$ -fitting of a shape to a set of points in low dimensions. The running time of the new algorithm is  $O(n + \text{poly}(\log n, 1/\varepsilon))$ , which is linear running time for a fixed  $\varepsilon$ . The constant powers hiding in the polylogarithmic term are too embarrassing to be explicitly stated, but are probably somewhere between 20 to 60 even just for the problem of  $L_1$ -fitting a circle to a set of points in the plane. Namely, this algorithm is only of theoretical interest. As such, the first open problem raised by this work is to improve these constants. A considerably more interesting problem is to develop a practical algorithm for this family of problems.

A natural tempting question is whether one can use the techniques in this paper for the problem of  $L_1$ -fitting a spline or a Bezier curve to a set of points. Unfortunately, the resulting surfaces in the parametric space are no longer nice functions. Therefore, the algorithmic difficulty here is overshadowed by algebraic considerations. We leave this as an open problem for further research.

Another natural question is whether one can use the techniques of Har-Peled and Wang [HW04] directly, to compute a coresets for this problem, and solve the problem on the coresets directly (our solution did a similar thing, by breaking the parametric space into a small number of prisms, and constructing a small “sketch” inside each such region). This would be potentially a considerable simplification over our current involved and messy approach. There is unfortunately a nasty technicality that requires that a coresets for the  $L_1$ -fitting of linear function is also a coresets if we take the square root of the functions (as holds for the construction of Section 3). It seems doubtful that this claim holds in general, but maybe a more careful construction of a coresets for the case of planes in three dimensions would still work. We leave this as open problem for further research.

The author believes that the algorithm presented in this paper should have other applications. We leave this as an open problem for further research.

## Acknowledgments

The author thanks Pankaj Agarwal, Arash Farzan, John Fischer, Vladlen Koltun, Bardia Sadri, Kasturi Varadarajan, and Yusu Wang for useful and insightful discussions related to the problems studied in this paper.

The author also thanks the anonymous referees for their detailed and patient comments on the manuscript.

## References

- [AAHS00] P. K. Agarwal, B. Aronov, S. Har-Peled, and M. Sharir. Approximation and exact algorithms for minimum-width annuli and shells. *Discrete Comput. Geom.*, 24(4):687–705, 2000.
- [AHV04] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Approximating extent measures of points. *J. Assoc. Comput. Mach.*, 51(4):606–635, 2004.
- [AHY06] P. Agarwal, S. Har-Peled, and H. Yu. Robust shape fitting via peeling and grating coresets. In *Proc. 17th ACM-SIAM Sympos. Discrete Algorithms*, pages 182–191, 2006.
- [AM94] P. K. Agarwal and J. Matoušek. On range searching with semialgebraic sets. *Discrete Comput. Geom.*, 11:393–418, 1994.
- [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.
- [Cha02] T. M. Chan. Approximating the diameter, width, smallest enclosing cylinder and minimum-width annulus. *Internat. J. Comput. Geom. Appl.*, 12(2):67–85, 2002.
- [Cla05] K. L. Clarkson. Subgradient and sampling algorithms for  $\ell_1$  regression. In *Proc. 16th ACM-SIAM Sympos. Discrete Algorithms*, pages 257–266, Philadelphia, PA, USA, 2005. Society for Industrial and Applied Mathematics.
- [DRVW06] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *Proc. 17th ACM-SIAM Sympos. Discrete Algorithms*, pages 1117–1126, New York, NY, USA, 2006. ACM Press.
- [FKV04] A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. Assoc. Comput. Mach.*, 51(6):1025–1041, 2004.
- [GLS06] J. Gao, M. Langberg, and L. Schulman. Analysis of incomplete data and an intrinsic-dimension helly theorem. In *Proc. 17th ACM-SIAM Sympos. Discrete Algorithms*, pages 464–473, 2006.
- [Har06] S. Har-Peled. How to get close to the median shape. Available from [http://www.uiuc.edu/~sariel/papers/05/11\\_fitting/](http://www.uiuc.edu/~sariel/papers/05/11_fitting/), 2006.
- [HK05a] S. Har-Peled and V. Koltun. Separability with outliers. In *Proc. 16th Annu. Internat. Sympos. Algorithms Comput.*, pages 28–39, 2005.

- [HK05b] S. Har-Peled and A. Kushal. Smaller coresets for  $k$ -median and  $k$ -means clustering. In *Proc. 21st Annu. ACM Sympos. Comput. Geom.*, pages 126–134, 2005.
- [HM04] S. Har-Peled and S. Mazumdar. Coresets for  $k$ -means and  $k$ -median clustering and their applications. In *Proc. 36th Annu. ACM Sympos. Theory Comput.*, pages 291–300, 2004.
- [HW04] S. Har-Peled and Y. Wang. Shape fitting with outliers. *SIAM J. Comput.*, 33(2):269–285, 2004.
- [O’R85] J. O’Rourke. Finding minimal enclosing boxes. *Internat. J. Comput. Inform. Sci.*, 14:183–199, 1985.
- [SW89] G.A.F. Seber and C.J. Wild. *Nonlinear Regression*. John Wiley & Sons, 1989.
- [YKII88] P. Yamamoto, K. Kato, K. Imai, and H. Imai. Algorithms for vertical and orthogonal  $l_1$  linear approximation of points. In *Proc. 4th Annu. ACM Sympos. Comput. Geom.*, pages 352–361. ACM Press, 1988.
- [ZS02] Y. Zhou and S. Suri. Algorithms for a minimum volume enclosing simplex in three dimensions. *SIAM J. Comput.*, 31(5):1339–1357, 2002.