On Approximate Range Counting and Depth[*]

Peyman Afshani and Timothy M. Chan
School of Computer Science
University of Waterloo
Waterloo, Ontario, N2L 3G1, Canada
{ pafshani, tmchan }@uwaterloo.ca

Abstract

We improve the previous results by Aronov and Har-Peled (SODA'05) and Kaplan and Sharir (SODA'06) and present a randomized data structure of $O(n)$ expected size which can answer 3D *approximate* halfspace range counting queries in $O(\log \frac{n}{k})$ expected time, where k is the actual value of the count. This is the first optimal method for the problem in the standard decision tree model; moreover, unlike previous methods, the new method is Las Vegas instead of Monte Carlo. In addition, we describe new results for several related problems, including approximate Tukey depth queries in 3D, approximate regression depth queries in 2D, and approximate linear programming with violations in low dimensions.

1 Introduction

Halfspace range counting arguably ranks as one of the all-time classic problems in computational geometry: how fast can one count the number of points in a query halfspace q if one is allowed to preprocess the point set in a data structure? The existing space/query trade-offs for this problem are not very satisfactory, even in low dimensions. In 3D (the main case of interest in this paper), for data structures with linear space, one is forced to be content with query time near $O(n^{2/3})$ [27] or, in terms of the actual count k, near $O(k^{2/3})$ [12]. On the other hand, logarithmic query time is attained only by data structures with near-cubic space. This is in contrast with the related problem of halfspace range *reporting* (finding all points inside q), which admits efficient algorithms in low dimension: linear space and $O(\log n + k)$ query in 2D [14] and 3D [1]. (See [2] for further background.)

Since it is generally believed that one cannot beat the query time of $\Omega(n^{2/3})$ with linear space for counting in 3D, researchers have turned to the *approximate* version of the problem. Here, for any fixed constant ε, if the actual count is k, any answer between $(1-\varepsilon)k$ and $(1+\varepsilon)k$ is considered correct. This 3D approximate halfspace range counting problem is the main subject of the paper.

Note that approximation is with respect to the count and not with respect to proximity—the latter option was, for example, investigated by Arya and Mount [4], where ranges are treated as "fuzzy" objects, and points too close to the boundary can either be counted or ignored. These two

[*]A preliminary version of this paper appeared in *Proc. 23rd Sympos. Comput. Geom.*, pages 337–343, 2007. Work of the second author was supported by NSERC.

forms of approximation are vastly different in terms of the techniques involved. Our interest in approximating counts are motivated by applications in computational statistics (see below).

Variants of approximate halfspace range counting were actually considered early on, for example, in a 1986 paper by Edelsbrunner and Welzl [18], who studied the 2D problem with *additive* instead of relative error (and called the problem “halfplanar range estimation”). If an additive error of εn is tolerable, then the problem can be solved with constant space and query time in any fixed dimension by simply working with a sample (so-called *ε-approximation*) of constant size [21, 30]. In particular, when the count k is close to n, we can get low relative error easily. So, the main challenge is in getting low relative error when k is small. In particular, for $k = 0$, we do not tolerate any error; thus, the problem should be at least as hard as range *emptiness* (deciding whether h contains any point). The existence of efficient halfspace range emptiness data structures in 2D and 3D (with $O(n)$ space and $O(\log n)$ query time [17, 31]) suggests that efficient approximate halfspace range counting structures might be possible in the same dimensions.

Indeed, that is the case, as was shown in two recent SODA papers. (Both papers are of particular relevance to us here, as some of the techniques used are related to ours.) The first, by Aronov and Har-Peled [3], described a black-box reduction from approximate halfspace range counting to range emptiness, with polylogarithmic increase in space and query time. In 3D, their resulting data structure needs $O(n \log n)$ space and $O(\log^2 n \log\log n)$ query time and gives a correct approximate answer with high probability, but is Monte Carlo in the sense that the query algorithm does not know if the returned answer is a correct approximation or not. The second paper, by Kaplan and Sharir [22] improved the query time to $O(\log^2 n)$ with the same $O(n \log n)$ space bound, by using a different strategy that combined an approximation technique of Cohen [16] with a new combinatorial lemma about overlaying lower envelopes over all prefixes of a randomly permuted sequence of planes. This query algorithm is also Monte Carlo. Subsequently, in an updated version of the first paper [3], Aronov and Har-Peled showed that the same improved query time of $O(\log^2 n)$ can be obtained directly by their original method, making the overlay lemma unnecessary. On the other hand, Kaplan and Sharir improved their data structure with Ramos in the journal version of their paper and obtained an $O(n \log\log n)$-space data structure with $O(\log^2 n)$ expected query time [23, 24]. Finally, a third paper by Har-Peled and Sharir [19], among other results, describes a data structure with the query time of $O(\log n \log\log n)$ but with a larger space bound of $O(n \log^{O(1)} n)$ for the 3D problem.

All these methods are suboptimal by a logarithmic factor either in space, query time or both. Furthermore, they are all Monte Carlo. Thus, the question of obtaining an $O(n)$-size structure with $O(\log n)$ query time in 3D is left open—this situation is somewhat unsettling, considering the fundamental nature and simplicity of the problem, and the desirability of linear space and logarithmic time in practice.

Our main result. We resolve the remaining open problem in this paper, by presenting a data structure using $O(n)$ expected space which can answer approximate halfspace range counting queries in 3D in $O(\log n)$ expected time or, better still, in $O(\log \frac{n}{k})$ expected time. This is optimal in terms of n and k, because a matching $\Omega(\log \frac{n}{k})$ lower bound in the algebraic decision tree model is not hard to show, even in 2D.[1]

The expected preprocessing time for our data structure is $O(n \log n)$, which is also optimal in

[1]Proof: Consider $\frac{n}{2k}$ points in convex position, each duplicated $2k$ times. Deciding whether a halfplane has approximately less than k points is the same as answering emptiness queries in an input of size $\frac{n}{2k}$. (It is possible to modify the construction for nondegenerate input.)

terms of n. Moreover, unlike the previous algorithms, our query algorithm is *Las Vegas*, meaning that it always produces a correct approximate answer. Our result is obtained by bringing in *shallow cuttings* (see Section 3.1) and, at the same time, applying Kaplan and Sharir's overlay technique in a new way (see Section 3.2), thus reclaiming the usefulness of the overlay lemma.

Other related results. We can apply some of our ideas to obtain new results on approximating the depth of a query point with respect to two definitions of depth that have been popularly studied at the intersection of robust statistics and computational geometry. The *Tukey depth* (also called halfspace depth) of a query point q with respect to a point set P is defined as the minimum number of points of P in any halfspace that contains q. The *regression depth* of a query line q with respect to a 2D point set P is defined as the minimum number of points intersected by q in any continuous motion which turns q into a vertical line [33]. Previous work in computational geometry has mostly focused on estimating the maximum depth rather than building data structures to estimate the depth of a query point/line. Points of maximum Tukey depth and lines of maximum regression depth in 2D can be computed in optimal $O(n \log n)$ expected time [13, 25], but it appears difficult to find data structures with nontrivial worst-case performance that can compute the exact depth of an arbitrary query point/line. The maximum depth in either definition is $\Theta(n)$ for all point sets, so an approximate maximum depth can be found easily by again working with a sample of constant size. The challenge is in approximating the depth of a query point/line, with low relative error, when the point/line is at small depth.

We show how one can approximate the regression depth of any query line in 2D with an $O(n)$-space[2] data structure in $O(\log n)$ time (see Section 4.1). We show how one can approximate the Tukey depth of a query point in 3D with an $O(n)$-space data structure in $O(\log n \log\log n)$ time; in 2D the query time can be improved to $O(\log n)$ (see Section 4.2).

We also realize that if Monte Carlo algorithms are allowed, almost optimal results for many approximate counting problems can be obtained through a simple general reduction using random sampling. In fact, the main idea (see Section 4.3) is already contained in Aronov and Har-Peled's paper [3].

Another related problem of fundamental importance is *linear programming (LP) with violations*. In one version, we are given n halfspaces and we want to find the smallest number k of halfspaces to delete so that the remaining halfspaces have a nonempty intersection. (Equivalently, we want the minimum depth, in another sense of the word, in an arrangement of halfspaces.) This problem was studied by Matoušek [28] and Chan [10]. In 2D and 3D, the current best running time, by Chan, is $O(n \log k + k^2 \log n)$ and $O(n \log k + k^{11/4} n^{1/4} \log^{O(1)} n)$ respectively; in higher dimensions d, the time bound is slightly less than $O(nk^{d+1})$. Aronov and Har-Peled [3] showed that a $(1+\varepsilon)$-factor approximation of the minimum k can be found by significantly faster Monte Carlo algorithms with $O(n \log\log n)$ running time in 2D and $O(n \log^{d+1} n)$ running time in any constant dimension d. We observe that with appropriate data structures, their method actually runs in $O(n \log\log n)$ time for any fixed dimension and, in fact, $O(n)$ time if $k \gg \log n \log\log n$ (see Section 4.4). We also show how to obtain an efficient $O(n \log n)$ *Las Vegas* algorithm in 3D. Similar ideas also lead to an $O(n \log n)$ Las Vegas algorithm for approximating the maximum depth in an arrangement of disks in 2D (for this problem, Aronov and Har-Peled gave an $O(n \log n)$ Monte Carlo algorithm).

[2]This improves over the $O(n \log\log n)$ space bound claimed in the preliminary version of this paper.

2 Preliminaries

Halfspace range searching problems are often easier to study in the dual space where the input is represented by a set, H, of n planes in $\mathbb{R}^3$. Here, the halfspace range reporting and (approximate) counting problems correspond to reporting and (approximately) counting the planes that pass below a query point q. We need the following result, first obtained by Chan [11]:

Lemma A. *With $O(n \log n)$ expected preprocessing time, a 3D halfspace range reporting query with output size k can be answered in $O(\log n + k)$ expected time.*

Subsequent papers on halfspace range reporting [32, 1] have concentrated on improving the space usage of the above data structure. These improvements will not be required here.

We need the following definitions. The *level* of a point $q \in \mathbb{R}^3$ is the number of planes of H that pass below q. Call the locus of all points of level at most k the *$(\le k)$-level* and the boundary of this locus the *k-level*.

The main tool that we use is Matoušek's shallow cutting lemma [26], one version of which is stated below. (The lemma has often been used in previous work on halfspace range reporting [32, 1].) An *ε-cutting* of H is a collection of nonoverlapping cells (tetrahedra) such that each cell intersects at most εn planes of H. The *conflict list* of a cell refers to the list of planes intersecting the cell. We say that the cutting *covers* a region if the union of the simplices contains the region.

Lemma B. *For any set of n planes in $\mathbb{R}^3$ and a parameter k, there exists an $O(\frac{k}{n})$-cutting of size $O(\frac{n}{k})$ that covers the $(\le k)$-level. The cells in the cutting are all vertical prisms unbounded from below.*

Furthermore, we can construct these cuttings for all k of the form $\lfloor (1+\varepsilon)^i \rfloor$ simultaneously in $O_\varepsilon(n \log n)$ expected time.

Proof. The first part is due to Matoušek [26]. The construction time for the cuttings follows from an algorithm by Ramos [32] and Chan [10, Lemma 3.1] observed that the cells can be turned into vertical prisms. $\square$

Throughout this paper, we say that an event X happens *with high probability (w.h.p.)* if $\Pr[X] \ge 1 - 1/n^{c_0}$ for a fixed large constant c_0. We say a number *x is ε-approximately y* iff $y(1-\varepsilon) \le x \le y(1+\varepsilon)$; similarly, *$x$ is ε-approximately less (resp. greater) than y* iff $x \le y(1+\varepsilon)$ (resp. $x \ge y(1-\varepsilon)$). For the sake of simplicity we will not work out the precise dependences of the time/space bounds on ε, since we have not tried to optimize such dependences; we use the O_ε notation to hide factors in terms of ε, which are in all cases $1/\varepsilon^{O(1)}$ or smaller.

3 Approximate Halfspace Range Counting Queries in 3D

In this section, we first present an $O(n)$-space, $O(\log n \log\log n)$-time result for approximate halfspace range counting in 3D. We then refine this method to obtain the optimal $O(n)$-space, $O(\log \frac{n}{k})$-time result.

3.1 Approximate levels by shallow cuttings

We first describe how Lemma B alone can lead to an almost optimal result that beats all previous methods. This method has the additional advantage of being Las Vegas.

As we discussed, approximate halfspace range counting in dual space corresponds to preprocessing a set H of n planes in $\mathbb{R}^3$ in a data structure that can answer the following queries: given any query point q, approximate the number k^* of planes below q. Notice that k^* is also the level of q in H. Thus, for a parameter k, $k \leq k^*$ (resp. $k \geq k^*$) is equivalent to q being above (resp. below) the k-level of H.

Our idea is to use this simple observation but replace exact levels with approximate levels. In 3D, the best upper bound on the complexity of the exact k-level currently is $O(nk^{3/2})$ [34]. However, the total complexity of the k'-level for all $k' = 0, \ldots, k$ is $O(nk^2)$ [15], which means that the average complexity of a k'-level with $(1-\varepsilon)k \leq k' \leq (1+\varepsilon)k$ is $O_\varepsilon(nk)$. This is still too large for our purposes. We show that a form of approximate k-level exists with complexity $O_\varepsilon(\frac{n}{k})$ only, which surprisingly is sublinear for nonconstant k.

We formally define an *ε-approximate $(\leq k)$-level* to be a region that contains the $(\leq (1-\varepsilon)k)$-level and is contained in the $(\leq (1+\varepsilon)k)$-level. (For our purposes, we do not require the region to be a terrain.) Using the shallow cutting lemma, we get:

Lemma 3.1. *For any set H of n planes in $\mathbb{R}^3$ and a parameter k, there exists an $O(\varepsilon)$-approximate $(\leq k)$-level of size $O_\varepsilon(\frac{n}{k})$.*

Furthermore, we can construct such approximate levels for all k of the form $\lfloor (1+\varepsilon)^i \rfloor$ simultaneously in $O_\varepsilon(n \log n)$ expected time; in the same time, we can also build a linear-size data structure that can decide whether a query point lies inside such an approximate $(\leq k)$-level in $O(\log \frac{n}{k})$ time.

Proof. Using Lemma A, we construct an $O(\frac{k}{n})$-cutting of size $O(\frac{n}{k})$, covering the $(\leq k)$-level of H for every k of the form $\lfloor (1+\varepsilon)^i \rfloor$ in $O_\varepsilon(n \log n)$ time. For each vertical prism Δ in each cutting, we find the set of $O(k)$ planes in the conflict list of Δ, denoted by H_Δ; this can be done in $O(\log n + k)$ time each, after $O(n \log n)$ preprocessing time, by applying Lemma A with the vertices of Δ as query points. Next, we construct an ε-cutting (covering Δ) of H_Δ. As is well known, for a constant ε, there exists an ε-cutting (covering $\mathbb{R}^3$) of constant size [17, 30] (more precisely, of size $O(\varepsilon^{-3})$) and the cutting can be constructed in time linear in the number of planes, i.e., $O(k)$ time. (Constant-size cuttings can in fact be constructed by "elementary" methods, such as prune-and-search.) Summing over all $O(n/k)$ prisms for a given k, this process produces $O_\varepsilon(n/k)$ subcells and takes $O_\varepsilon(\frac{n}{k}(\log n + k))$ time. Over all values of k, this amounts to $O_\varepsilon(n \log n)$ preprocessing time.

For each subcell $\delta \subset \Delta$ of the ε-cutting, we use Lemma A to compute the level ℓ_δ of some arbitrary point in δ, and if $\ell_\delta \leq k$, we include δ in the approximate $(\leq k)$-level. This construction satisfies the desired property, because each subcell δ intersects $O(\varepsilon k)$ planes so the levels of any two points in the subcell differ by at most $O(\varepsilon k)$. The time for this step is again $O_\varepsilon(\frac{n}{k}(\log n + k))$, and thus the total preprocessing time is $O_\varepsilon(n \log n)$.

For a given query point q, we can find the vertical prism Δ in the $O(\frac{k}{n})$-cutting containing q in $O(\log \frac{n}{k})$ time by planar point location [17, 31] on the xy-projection of the prisms. Afterwards, we can find the subcell $\delta \subset \Delta$ containing q in constant time and see if δ was included in the approximate $(\leq k)$-level. $\square$

The above lemma immediately suggests a data structure for our problem:

Theorem 3.2. *With $O_\varepsilon(n \log n)$ expected preprocessing time one can build a data structure of size $O_\varepsilon(n)$ which can answer approximate 3D halfspace range counting queries in $O_\varepsilon(\log n \log \log n)$ worst-case time. The query algorithm is always correct.*

Proof. Let $k_i = \lfloor (1+\varepsilon)^i \rfloor$ for $i = 1, \ldots, \lceil \log_{1+\varepsilon} n \rceil$, and construct an approximate $(\le k_i)$-level for each i. The expected preprocessing time is $O_\varepsilon(n \log n)$. The total space is given by a geometric series $O(\sum_i \frac{n}{k_i}) = O_\varepsilon(n)$.

To approximate the number k^* of points below the query point q, we do an approximate binary search. For a given k_i, we can determine whether k^* is $O(\varepsilon)$-approximately less than k_i or $O(\varepsilon)$-approximately greater than k_i, by testing whether q lies inside the approximate $(\le k_i)$-level or not, in $O(\log \frac{n}{k_i})$ time. After $O(\log \log_{1+\varepsilon} n)$ iterations of binary search on the k_i's, we arrive at a value that is $O(\varepsilon)$-approximately k^*. The query time is $O(\log n \log \log_{1+\varepsilon} n)$. □

Remark. The above data structure uses a hierarchy of shallow cuttings and is similar in spirit to Chan's data structure for halfspace range reporting [11], which uses a hierarchy of lower envelopes of random samples. Lower envelopes of samples share some similar characteristics as shallow cuttings and are more practical for implementation but, without additional ideas, do not seem to yield Las Vegas results as good as the above data structure. In the next method, though, we will employ lower envelopes of random subsets, but in conjunction with our shallow-cutting-based method.

3.2 Using randomized incremental construction

For our optimal method, we need an additional technique by Kaplan and Sharir:

Lemma 3.3. *Let $h_1, \ldots, h_n$ be a random permutation of n given planes in $\mathbb{R}^3$. With $O(n \log n)$ expected preprocessing time one can build a data structure of expected size $O(n)$ so that given a query point q, one can find the smallest index j such that h_j lies below q in $O(\log n)$ expected time. In fact, the expected query bound can be reduced to $O(\log j)$.*

Proof. The first part (ignoring the space bound) was originally due to Kaplan and Sharir [22], and is derived from their combinatorial lemma stating that the overlay of all the lower envelopes encountered during a randomized incremental construction has expected complexity $O(n \log n)$. An alternative method was described in subsequent paper of Kaplan, Ramos, and Sharir [23] (which uses linear space).

For the improvement to $O(\log j)$, let $j_i = 2^{2^i}$ for $i = 1, 2, \ldots, \lceil \log \log n \rceil$ and build the above data structure for the prefix $h_1, \ldots, h_{j_i}$, which is itself a random permutation, for each i. The expected preprocessing time and space remain asymptotically unchanged. To answer a query, we query the prefix $h_1, \ldots, h_{j_i}$ for $i = 1$, $i = 2$, and so on, until an answer is found. The total expected query time is $O(\sum_{j_{i-1} \le j} \log j_i) = O(\log j)$. □

The usefulness of the above lemma is explained by the following observation:

Observation 3.4. *Let $h_1, \ldots, h_n$ be a random permutation of a set H. Given any subset $S \subseteq H$ of size k^*, let j be the smallest index with $h_j \in S$. Let $k = \frac{n}{j}$. Then the probability that $k^* < k/b$ or $k^* > bk$ for a parameter $b > 0$ is $O(1/b)$.*

Proof. The event $k^* < k/b$ implies that $j < \frac{n}{bk^*}$ and so at least one of $h_1, \ldots, h_{n/(bk^*)}$ is in S; this happens with probability at most $\frac{n}{bk^*} \cdot \frac{k^*}{n} = 1/b$. On the other hand, $k^* > bk$ implies that $j > bn/k^*$ and so $h_1, \ldots, h_{bn/k^*}$ are all not in S; this happens with probability at most $(1 - k^*/n)^{bn/k^*} = 1/e^{\Omega(b)}$. □

The key new idea is to use Kaplan and Sharir's lemma to obtain an initial estimate $k = \frac{n}{j}$ which approximates the unknown count k^* well with good probability ("well" and "good" in the sense of the above observation). With the availability of this initial estimate, we can speed up the query time of Theorem 3.2: namely, we can replace the approximate binary search (which is the cause of the extra $\log\log n$ factor) with a simple linear search. In the analysis, we bound the overall expected query time by a geometric series.

Theorem 3.5. *With $O_\varepsilon(n \log n)$ expected preprocessing time, one can build a data structure of expected size $O_\varepsilon(n)$ which can answer approximate 3D halfspace range counting queries in $O_\varepsilon(\log \frac{n}{k^*})$ expected time for any fixed query halfspace. Here k^* is the actual value of the count and the query algorithm is always correct.*

Proof. Our data structure consists of the data structure from Theorem 3.2 which includes approximate ($\leq k_i$)-levels for $k_i = \lfloor (1+\varepsilon)^i \rfloor, i = 1, \ldots, \lceil \log_{1+\varepsilon} n \rceil$. We augment this with the data structure in Lemma 3.3 applied to a random permutation of the n planes dual to the input points.

To approximate the number k^* of planes below a query point q, we first compute the smallest index j such that h_j lies below q in $O(\log j)$ time. Let $k = \frac{n}{j}$ and suppose $k_s \leq k < k_{s+1}$. We apply a linear search starting at k_s. Recall that we can determine whether k^* is $O(\varepsilon)$-approximately less than $k_{s\pm i}$ or $O(\varepsilon)$-approximately greater than $k_{s\pm i}$, in $O(\log \frac{n}{k_{s\pm i}})$ time by querying an approximate level. If k^* is $O(\varepsilon)$-approximately less than k_s, we repeatedly $O(\varepsilon)$-approximately compare k^* with $k_{s-1}, k_{s-2}, \ldots$; otherwise, we repeatedly $O(\varepsilon)$-approximately compare k^* with $k_{s+1}, k_{s+2}, \ldots$ With $O(i)$ iterations of the search, we eventually arrive at a value $k_{s\pm i}$ that is $O(\varepsilon)$-approximately k^*.

The probability that k^* is $O(\varepsilon)$-approximately $k_{s\pm i} \approx (1+\varepsilon)^{\pm i} k$ is at most $O(1/(1+\varepsilon)^i)$ by Observation 3.4. Thus, the total expected query time is upper-bounded by

$$\sum_{i=1}^{\infty} \frac{1}{(1+\varepsilon)^i} \cdot O\left(i \log \frac{n(1+\varepsilon)^i}{k^*}\right) \;=\; O_\varepsilon\left(\log \frac{n}{k^*}\right).$$

This completes the proof. $\square$

4 Related Problems

Similar techniques can be applied to solve other related problems.

4.1 Approximate regression depth queries in 2D

The problem of computing the regression depth of a query line in 2D reduces to the following in dual space: Given a set H of n lines in $\mathbb{R}^2$, preprocess them in a data structure so that given any query point q, we can find the minimum number k^* of lines intersected by a ray over all rays originating from q (see Figure 1(a)). Following [6], call k^* the *undirected depth* of q. Call the locus of all points of undirected depth at most k the ($\leq k$)-*envelope*. Call the boundary of this locus the k-*envelope*. In measuring the complexity of a k-envelope or a polygonal chain, we will include all vertices in the arrangement that lie on the chain, including those making angles of π (in other words, every line contributes as many vertices as its number of intersections with the boundary - see Firgure 1(b)).

We use the ideas outlined in Theorem 3.2. We first need an analog of Lemma 3.1. Define an ε-*approximate* ($\leq k$)-*envelope* to be a region that contains the ($\leq (1-\varepsilon)k$)-envelope and is contained

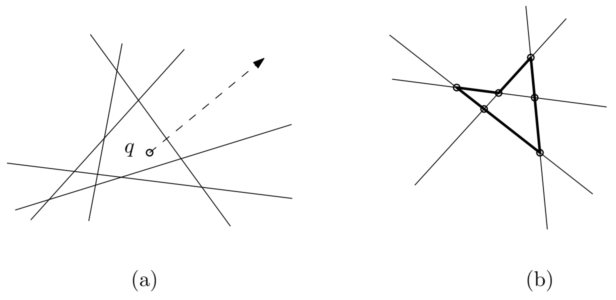

Figure 1: (a) A point of regression depth one. (b) The 1-envelope of a set of four lines containing six vertices. The vertices are marked with small circles.

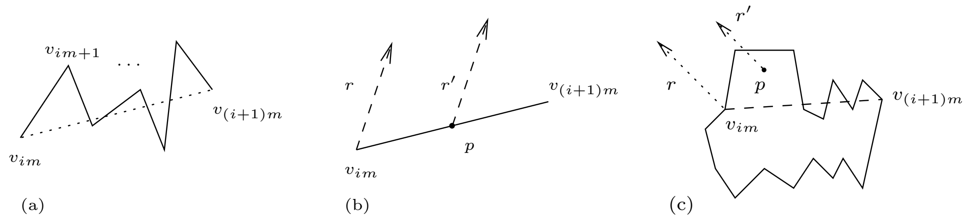

Figure 2: (a) An edge of a simplified polygon (in dotted line). (b) Proof of Lemma 4.2(i). (c) Proof of Lemma 4.2(ii).

in the ($\leq (1+\varepsilon)k$)-envelope. In the preliminary version of this paper, we prove the existence of an approximate ($\leq k$)-envelope of size $O(\min\{n, \frac{n}{k}\log n\})$. Here, we present a different proof of a slightly improved bound of $O(\frac{n}{k}\alpha(k))$, where $\alpha(\cdot)$ is the inverse Ackermann function.

Lemma 4.1. *For any set H of n lines in $\mathbb{R}^2$, the total complexity of the k'-envelope over all $k' = 0, \ldots, k$ is $O(nk)$.*

Proof. First observe that in 2D, the 0-envelope consists of all unbounded cells in the arrangement and has linear complexity by the zone theorem [17], since the unbounded cells intersect the line $x = -M$ or $x = M$ for a sufficiently large M.

We apply Clarkson and Shor's technique [15], by picking a random sample in which each line is chosen with probability $\frac{1}{k}$. The lemma follows from the fact that a vertex v of the arrangement which lies in the ($\leq k$)-envelope of H has $\Omega(\frac{1}{k^2})$ chance of surviving in the sample and being a 0-envelope vertex. $\square$

In 2D, one can obtain an approximate level of size $O(\frac{n}{k})$ by taking an exact level of size $O(n)$ and applying a "simplification" process (e.g., as in [18]). We show that approximate envelopes can be constructed in a similar fashion. The modification is not trivial, as k-envelopes have a more complicated geometry than k-levels. In particular, the k-envelope may consist of multiple polygons; all points inside these polygons have undirected depth at least k and all points outside have undirected depth at most k.

Given a polygon $A = \langle v_0, v_1, \ldots, v_t\rangle$, we define its *$m$-simplification* as the polygon $A' = \langle v_0, v_m, v_{2m}, \ldots, v_{m\lfloor \frac{t}{m}\rfloor}, v_t\rangle$ (see Figure 2(a)). Note that even if A is a simple polygon, the simplified polygon A' may self-intersect. We say that a point p is *outside* a nonsimple polygon C if p

is in the outer connected component of $\mathbb{R}^2 - C$ (i.e., the component containing infinity); otherwise, p is *inside* C. Similarly, p is outside a collection C of polygons (possibly with holes) if p is in the outer component of $\mathbb{R}^2 - \bigcup_{A \in C} A$. The following lemma encompasses all the properties of simplified polygons that we need:

Lemma 4.2. *Let C be the collection of polygons defining the k-envelope and C' be the collection of the m-simplifications of these polygons. Then* (i) *all points inside C' have undirected depth at least $k - m$;* (ii) *all points outside C' have undirected depth at most $k + m$; and* (iii) *C' has at most $O(|C'|m)$ crossings.*

Proof. (i) Consider any ray r' from an arbitrary point inside C' which intersects C', say, at a point p on the segment $v_{im}v_{(i+1)m}$. Draw a ray r parallel to r' from v_{im} (Figure 2(b)). Then r hits at least k lines since v_{im} has undirected depth k. Any line which crosses only one of these two rays intersects the segment $v_{im}v_{(i+1)m}$ and thus creates a vertex on the chain $\langle v_{im}, v_{im+1}, \ldots, v_{(i+1)m}\rangle$, but there are only m such vertices. So, r' hits at least $k - m$ lines.

(ii) Let p be outside C'. Then p must be outside the simplified polygon A' for some $A = \langle v_0, v_1, \ldots, v_t\rangle$ in C. If p is outside A, then p has undirected depth at most k. We may assume that p is inside A.

Since p is outside A', it is possible to connect p to infinity through a curve c not crossing A'. Since p is inside A, the number of intersections between c and A must be odd. This implies that for some v_{im}, the number of intersections between c and the chain $A_i = \langle v_{im}, v_{im+1}, \cdots, v_{(i+1)m}\rangle$ is odd.

Observe that p is inside the shape formed by adding the segment $v_{im}v_{(i+1)m}$ of A' to the chain A_i, since the number of intersections of c with this boundary is odd. Thus, we have a situation similar to Figure 2(c). Consider a ray r from v_{im} which crosses k lines, and draw a ray r' parallel to r from p. Then any line which crosses r' but not r must also cross the chain A_i, but there can be at most m such lines. So, p has undirected depth at most $k + m$.

(iii) Suppose two segments $v_{im}v_{(i+1)m}$ and $w_{jm}w_{(j+1)m}$ of C' cross. Since the original k-envelope C does not have crossings, one of two possibilities must hold: the chain $A_i = \langle v_{im}, v_{im+1}, \ldots, v_{(i+1)m}\rangle$ intersects the segment $w_{jm}w_{(j+1)m}$, in which case we give $w_{jm}w_{(j+1)m}$ one charge; or the chain $B_j = \langle w_{jm}, w_{jm+1}, \ldots, w_{(j+1)m}\rangle$ intersects the segment $v_{im}v_{(i+1)m}$, in which case we give $v_{im}v_{(i+1)m}$ one charge.

Any line which intersects the segment $v_{im}v_{(i+1)m}$ must create a vertex on the chain A_i, but there are only m such vertices. Thus, each of the $|C'|$ segments receives $O(m)$ charges. □

With one additional combinatorial fact stated below, we can derive our lemma on approximate envelopes.

Lemma 4.3. *Given an arrangement of N line segments in $\mathbb{R}^2$ with X intersections, the outer face has complexity $O(N\alpha(\lceil X/N \rceil))$.*

Proof. By a standard result [20], the outer face has complexity $O(N\alpha(N))$. To obtain an X-sensitive bound, we use known results on *intersection-sensitive cuttings* (e.g., [5]): there exists a partition of $\mathbb{R}^2$ into $O(r + X(r/N)^2)$ triangles, such that each triangle intersects $O(N/r)$ segments. For each triangle Δ, let S_Δ be the segments clipped to Δ. Since each vertex in Δ of the outer face of the overall arrangement must be on the outer face of S_Δ, the complexity of the overall outer face is at most $O\left(\sum_\Delta |S_\Delta|\alpha(|S_\Delta|)\right) = O((r + X(r/N)^2) \cdot (N/r)\alpha(N/r))$. Setting $r = \min\{N^2/X, N\}$ yields the $O(N\alpha(\lceil X/N \rceil))$ bound. $\square$

Lemma 4.4. *For any set of n lines in $\mathbb{R}^2$ and a parameter k, there exists an $O(\varepsilon)$-approximate $(\leq k)$-envelope of size $O_\varepsilon(\frac{n}{k}\alpha(k))$.*

Proof. According to Lemma 4.1, the average complexity of a k'-envelope for a random k' between $(1-\varepsilon)k$ and $(1+\varepsilon)k$ is $O_\varepsilon(n)$. Let C be such a k'-envelope. We return the outer face of the m-simplifications of the polygons in C, with the parameter $m = \varepsilon k$. By Lemma 4.2(i,ii), the resulting polygon is an $O(\varepsilon)$-approximate $(\leq k)$-envelope.

Note that any polygon in C with complexity less than m can be simplified to the empty polygon and be discarded. The total number of vertices in the simplified polygons is $N = O_\varepsilon(\frac{n}{k})$ and the number of crossings is $X = O_\varepsilon(Nk)$ by Lemma 4.2(iii). By Lemma 4.3, the complexity of the outer face is $O_\varepsilon(\frac{n}{k}\alpha(k))$. $\square$

We can now use the above lemma to prove the following theorem.

Theorem 4.5. *One can build a data structure that uses $O_\varepsilon(n)$ space and can answer approximate regression depth queries in 2D in $O_\varepsilon(\log n)$ worst-case time.*

Proof. Let $k_i = \lfloor (1+\varepsilon)^i \rfloor$ for $i = 1, \ldots, \lceil \log_{1+\varepsilon} n \rceil$, and construct an $(\varepsilon/3)$-approximate $(\leq k_i)$-envelope E_i for each i by Lemma 4.4. The total size of the E_i's is $O_\varepsilon(\sum_i \frac{n}{k_i}\alpha(k_i)) = O_\varepsilon(n)$. Finally, we store all these approximate envelopes in a point location data structure [17, 31]. Note that there are no intersections between the boundaries of the E_i's.

To approximate k^* for a query point q, we return the smallest k_i such that q lies inside E_i. This can be done in $O_\varepsilon(\log n)$ time by a single planar point location query on the combined subdivision formed by the boundaries of the E_i's. $\square$

4.2 Approximate Tukey depth queries in 3D

The problem of computing the Tukey depth of a query point in 3D reduces to the following in dual space: Given a set H of n planes in $\mathbb{R}^3$, preprocess them in a data structure so that given any query plane q, we can find the smallest value k^* such that q intersects the $(\leq k^*)$-level in the arrangement of H. (Technically, we need to compute also the smallest value k^{**} such that q intersects the $(\geq n - k^{**})$-level, and return the smaller of the two values, but computing k^{**} is similar.)

We adapt the method used in Theorem 3.2 to solve this problem:

Theorem 4.6. *One can preprocess a 3D point set of size n in $O_\varepsilon(n \log n)$ expected time into a data structure of $O_\varepsilon(n)$ size such that the Tukey depth of any query point q can be approximated in $O_\varepsilon(\log n \log\log n)$ worst-case time. The query algorithm is always correct.*

Proof. Let $k_i = \lfloor (1+\varepsilon)^i \rfloor$ for $i = 1, \ldots, \lceil \log_{1+\varepsilon} n \rceil$, and construct an approximate $(\leq k_i)$-level L_i for each i by Lemma 3.1, as in the proof of Theorem 3.2. For each i, we compute the upper hull U_i of the $O(\frac{n}{k_i})$ vertices of L_i. This takes time $O(\sum_i \frac{n}{k_i} \log \frac{n}{k_i}) = O_\varepsilon(n \log n)$.

To approximate k^* for a query plane q, we do an approximate binary search. For a given k_i, we can determine whether k^* is $O(\varepsilon)$-approximately greater than k_i or $O(\varepsilon)$-approximately less than k_i, by testing whether q lies strictly above the upper hull U_i or not, in $O(\log n)$ time (back in primal space, this corresponds testing whether a point lies below a lower envelope of planes, which reduces to planar point location). After $O(\log \log_{1+\varepsilon} n)$ iterations of binary search, we arrive at a value that is $O(\varepsilon)$-approximately k^*. The query time is $O(\log n \log \log_{1+\varepsilon} n)$. $\square$

Remarks. For Tukey depth queries in 2D, we can improve the query time to $O_\varepsilon(\log n)$ by the same idea as in the proof of Theorem 4.5 (replacing binary search with a single planar point location query).

4.3 General approximate counting through reduction to small counts

The concept of approximate levels employed in the previous sections provides almost optimal results where the correctness is guaranteed with probability one. However, it is not always possible to take this approach if such approximate levels are difficult to build. Nonetheless, as Aronov and Har-Peled [3] observed Monte Carlo approximation schemes can be obtained through simpler ideas such as random sampling. They proposed a general technique to reduce a approximate counting problem to a corresponding emptiness problem. In this section, we propose a similar reduction but to a counting problem when the count is small (bounded by $O(\log n)$). In contrast to their technique, our reduction introduces no overhead in the space and only a tiny increase in the query time.

Our method uses an easy application of the Chernoff bound to obtain the following lemma which helps us "approximately compare" the actual value of the count with another fixed parameter.

Observation 4.7. *Let c be an arbitrary constant, c_ε be a sufficiently large constant depending on ε and c, and k be a fixed parameter such that $k \geq c_\varepsilon \log n$. For a set P of n elements consider a random sample, R, where each element of P is included with probability $p = \frac{c_\varepsilon \log n}{k}$. For any subset $S \subseteq P$ of size k^*, if $k^* \leq k$ (resp. $k^* \geq k$) then with probability at least $1 - n^{-c}$ we have $|S \cap R| \leq (1+\varepsilon)pk$ (resp. $|S \cap R| \geq (1-\varepsilon)pk$).*

Proof. Let $X_i = 1$ if the i-th element of S is in R, and 0 otherwise; these are independent, identically distributed, random variables. Let $X = |S \cap R| = \sum_{i=1}^{k^*} X_i$, which has mean $\mu = pk^*$.

Suppose $k^* \leq k$. According to the Chernoff bound [29],

$$\Pr[X > (1+\delta)\mu] \;\leq\; \begin{cases} e^{-\delta^2\mu/4} & \text{if } \delta \leq 2e-1 \\ 2^{-(1+\delta)\mu} & \text{if } \delta > 2e-1. \end{cases}$$

We substitute $\delta = \varepsilon k/k^*$ and see that $\Pr[X > (1+\varepsilon)pk] \leq \Pr[X > pk^* + \varepsilon pk] \leq \max\{e^{-\varepsilon^2(k/k^*)^2pk^*/4},\, 2^{-pk^*-\varepsilon pk}\} \leq \max\{e^{-\varepsilon^2 pk/4},\, 2^{-\varepsilon pk}\} \leq n^{-c}$, if c_ε is sufficiently large.

Next, suppose $k^* \geq k$. According to the Chernoff bound,

$$\Pr[X < (1-\delta)\mu] \;\leq\; e^{-\delta^2\mu/2}.$$

Then $\Pr[X < (1-\varepsilon)pk] \leq \Pr[X < (1-\varepsilon)pk^*] \leq e^{-\varepsilon^2 pk^*/2} \leq e^{-\varepsilon^2 pk/2} \leq n^{-c}$, if c_ε is sufficiently large. $\square$

The above observation suggests the following reduction.

Theorem 4.8. *Given any point set of size n and a "threshold" parameter ℓ, suppose there is a data structure that can decide whether the number of points in a query range is at most ℓ, and if so, return this number. Let $P_{\text{small}}(n)$, $S_{\text{small}}(n)$, and $Q_{\text{small}}(n, \ell)$ be the (expected) preprocessing time, space, and query time of this data structure (assuming these functions are well-behaved).*

Then with $O_\varepsilon(P_{\text{small}}(n))$ preprocessing time, $O_\varepsilon(S_{\text{small}}(n))$ space and $O_\varepsilon(Q_{\text{small}}(n, \log n) \log\log n)$ expected query time one can build a data structure that can approximate the number of points in a query range.

The algorithm is correct w.h.p. for any fixed query range.

Proof. For each i, we take a random sample R_i of the given point set P with sampling probability $p_i = \frac{c_\varepsilon \log n}{k_i}$ in which $k_i = \lfloor (1+\varepsilon)^i \rfloor$ for $i = \lceil \log_{1+\varepsilon}(c_\varepsilon \log n) \rceil, \ldots, \lceil \log_{1+\varepsilon} n \rceil$. We then build a small-count data structure for every sample R_i, with threshold $\ell = c_\varepsilon \log n$. Since the expected size of R_i geometrically decreases as i increases, the total preprocessing time and space is asymptotically unchanged.

To approximate the (unknown) number k^* of points inside a query region h, we do an approximate binary search. By Observation 4.7, for a given k_i, we can determine whether k^* is $O(\varepsilon)$-approximately less than k_i or $O(\varepsilon)$-approximately greater than k_i, by testing whether $|h \cap R_i| \leq c_\varepsilon \log n$. This can be done by using the small-count data structure for R_i. After $O(\log\log_{1+\varepsilon} n)$ iterations of binary search on the k_i's, we arrive at a value that $O(\varepsilon)$-approximates k^* w.h.p. (We can readjust ε by a constant factor if necessary.) The query time is $O(Q_{\text{small}}(n, c_\varepsilon \log n) \log\log_{1+\varepsilon} n)$.

One special case remains: what if k^* is less than all the k_i's, i.e., $k^* \leq c_\varepsilon \log n$? In this case, we can directly answer the query using a small-count data structure for P in $Q_{\text{small}}(n, c_\varepsilon \log n)$ time. $\square$

Remarks. For example, for 3D halfspace range counting, we can use known results for 3D halfspace range reporting [1, 11, 32] to get $P_{\text{small}}(n) = O(n \log n)$, $S_{\text{small}}(n) = O(n)$, and $Q_{\text{small}}(n, \ell) = O(\log n + \ell)$. Thus, this implies a Monte Carlo data structure with $O(\log n \log\log n)$ query time using linear space, which is already close to optimal. Of course, for 3D approximate halfspace range counting specifically, the optimal solution presented in Section 3 is better.

Ironically, essentially the same Chernoff-bound observation was used before in Aronov and Har-Peled's paper [3], not for halfspace range counting but for offline problems like linear programming with violations. One reason that Aronov and Har-Peled obtained a worse result for range counting is perhaps their insistence on reducing approximate counting to *emptiness*, although to be fair, their main focus was not in optimizing logarithmic factors.

4.4 Approximate LP with violations

Lastly, we address the problem of LP with violations: given a set H of n halfspaces in $\mathbb{R}^d$, find a point that violates (i.e., lies outside) the smallest number k_{opt} of halfspaces.

We can apply Theorem 4.8 to get the following reduction, which was originally obtained by Aronov and Har-Peled [3]. We include the proof to prepare for a later result:

Theorem 4.9. *Suppose there is an algorithm for LP with violations running in $T_{\text{small}}(n, k_{\text{opt}})$ time (assuming that the function T_{small} is well-behaved). Then we can solve the problem approximately in $O_\varepsilon(T_{\text{small}}(\frac{n}{k_{\text{opt}}} \log n, \log n))$ time. This algorithm is correct w.h.p.*

Proof. We first solve the approximate decision problem: given a threshold k, determine whether k_{opt} is ε-approximately less than k or ε-approximately greater than k.

The solution is simple: take a random sample R with sampling probability $p = \frac{c_\varepsilon \log n}{k}$ and solve the problem for R with threshold $pk = c_\varepsilon \log n$ in $O(T_{\text{small}}(\frac{c_\varepsilon n \log n}{k}, c_\varepsilon \log n))$ time. By Observation 4.7, if a fixed point violates ε-approximately less (resp. greater) than pk halfspaces in R, then it violates $O(\varepsilon)$-approximately less (resp. greater) than k halfspaces in H w.h.p. Thus, the answer is correct w.h.p., since the number of combinatorially "different" points is polynomial in n.

To approximate k_{opt}, we just run the above approximate decision algorithm for $k = n, \lfloor(1-\varepsilon)n\rfloor, \lfloor(1-\varepsilon)^2 n\rfloor, \ldots$, until k is $O(\varepsilon)$-approximately k_{opt}. The total running time forms a geometric series and increases only by a constant factor. $\square$

In d dimensions, Matoušek [28] gave an algorithm for LP with violations with running time $O(nk_{\text{opt}}^{d+1})$. For small $k_{\text{opt}} < n^\alpha$ for a constant $\alpha > 0$ depending on d, Chan [7, 8] showed how to improve the running time to $O(n \log k_{\text{opt}})$ by using data structures for linear programming queries:

Lemma C. *LP with violations in a constant dimension d can be solved in $T_{\text{small}}(n, k_{\text{opt}}) = O(n \log k_{\text{opt}} + k_{\text{opt}}^{O(1)})$ time.*

Combined with Theorem 4.8, the above lemma implies an algorithm for the approximate version of the problem with running time $O_\varepsilon(T_{\text{small}}(\frac{n}{k_{\text{opt}}} \log n, \log n)) = O_\varepsilon(\frac{n}{k_{\text{opt}}} \log n \log\log n + \log^{O(1)} n)$ for all k_{opt}. For $k_{\text{opt}} = \omega(\log n \log\log n)$, this is sublinear. For $k_{\text{opt}} = O(\log n \log\log n)$, we can switch back to the algorithm in Lemma C itself. Thus we have the following result.

Theorem 4.10. *We can solve the problem of approximate LP with violations in constant dimension in $O_\varepsilon(\min\left\{\frac{n}{k_{\text{opt}}} \log n \log\log n + \log^{O(1)} n,\ n \log k_{\text{opt}} + k_{\text{opt}}^{O(1)}\right\})$ time. The bound is always at most $O(n \log\log n)$. This algorithm is correct w.h.p.*

The above algorithm is quintessentially Monte Carlo. We show how to obtain a Las Vegas algorithm in 3D (and thus in 2D as well), by using shallow cuttings:

Theorem 4.11. *We can solve the problem of approximate LP with violations in 3D in $O_\varepsilon(n \log n)$ expected time. This algorithm is always correct.*

Proof. We follow the algorithm from Theorem 4.9. Consider a random sample R in which each plane is chosen with probability $p = \frac{c_\varepsilon \log n}{k}$. Intuitively, this random sample attempts to compare approximately k_{opt} to k. Instead of relying on a high probability bound we try to certify that the chosen sample is indeed "good". Let H^- (resp. H^+) be the bounding planes of the lower (resp. upper) halfspaces. Let $R^- = R \cap H^-$ (resp. $R^+ = R \cap H^+$). We only describe this process for R^-; R^+ can be certified in a similar manner.

Following the algorithm of Theorem 4.9, we need to certify that $(\leq pk)$-level of R^- is contained within the $(\leq (1+O(\varepsilon))k)$-level of H^- and contains the $(\leq (1-O(\varepsilon))k)$-level of H^-. By Observation 4.7, this certification succeeds w.h.p., since the number of combinatorially different points is polynomial in n; if certification fails at any moment, we can afford to switch to a brute-force algorithm.

To perform the certification, we follow the approximate level construction from the proof of Lemma 3.1. Consider each vertical prism Δ of the $O(\frac{k}{n})$-cutting covering the $(\leq 2k)$-level of H^-. Recall that Δ is further divided into a constant number of subcells δ via an ε-cutting, and we have computed the level ℓ_δ of an arbitrary point in each subcell δ.

Let R^-_Δ be the planes in R^- intersecting Δ. Note that since Δ intersects $O(k)$ planes of H^-, $|R^-_\Delta|$ is binomially distributed with mean $O(pk)$; so $E[g(|R^-_\Delta|)] = O(g(pk))$ for any polynomially bounded function g satisfying $g(2n) = O(g(n))$.

We construct the arrangement of R^-_Δ. If a feature f of this arrangement intersects a subcell δ, then $f \cap \delta$ lies between $(\ell_\delta - O(\varepsilon)k)$-level and $(\ell_\delta + O(\varepsilon)k)$-level of H^-, because the levels of any two points in a subcell differ by at most $O(\varepsilon k)$. We just need to compare ℓ_δ approximately with k. As there are only a constant number of subcells in Δ, the cost per Δ is $O(|R^-_\Delta|^3)$, which has expected value $O(p^3k^3) = O_\varepsilon(\log^3 n)$.

In fact, we can reduce some of the logarithmic factors: instead of examining all features f of the whole arrangement of R^-_Δ, it suffices to construct an approximate $(\leq k')$-level of R^-_Δ, over all k' of the form $\lfloor (1+\varepsilon)^i \rfloor$, of total size $O_\varepsilon(|R^-_\Delta|)$, and examine only features of the boundaries of these approximate levels. By Lemma 3.1, this reduces the cost per Δ to $O_\varepsilon(|R^-_\Delta| \log |R^-_\Delta|)$, which has expected value $O_\varepsilon(\log n \log\log n)$.

The total expected extra cost for the certification is thus $O_\varepsilon(\frac{n}{k} \log n \log\log n)$. We can approximate $k_{\rm opt}$ as before. The total expected time of the algorithm remains $O_\varepsilon(n \log n + \frac{n}{k_{\rm opt}} \log n \log\log n)$. As before, for $k_{\rm opt} < \log\log n$, we can directly use Lemma C. $\square$

Using a similar approach, we can also obtain an $O(n \log n)$ Las Vegas algorithm for the following problem: given n disks in 2D, find a point that has the maximum *depth*, i.e., lies inside the maximum number of disks.

Theorem 4.12. *We can compute a $(1+\varepsilon)$-factor approximation to the maximum depth in an arrangement of n disks in 2D in $O_\varepsilon(n \log n)$ expected time. This algorithm is always correct.*

Proof. First we can get a Monte-Carlo $O_\varepsilon(n \log n)$ algorithm, as described by Aronov and Har-Peled [3]: Let $T_{\rm small}(n, k_{\rm opt})$ denote the time required to solve the problem when the depth is $k_{\rm opt}$; we can naively bound $T_{\rm small}(n, k_{\rm opt})$ by $O(n \log n + nk_{\rm opt})$ by explicitly constructing the entire disk arrangement [3]. By an appropriate analog of Theorem 4.9, we can then solve the problem in time $O_\varepsilon(T_{\rm small}(\frac{n}{k_{\rm opt}} \log n, \log n)) = O_\varepsilon(\frac{n}{k_{\rm opt}} \log^2 n)$, which is $O_\varepsilon(n \log n)$ for $k_{\rm opt} \geq \log n$. For $k_{\rm opt} < \log n$, we can switch to the $O(n \log n + nk_{\rm opt})$ algorithm.

Now, to turn this into a Las Vegas algorithm, we use certification via shallow cuttings again. By applying the standard lifting map to transform the disks into planes in 3D [17], the problem is equivalent to approximating the maximum level for an arrangement of n planes over all points on the unit paraboloid. We can certify whether a sample of planes R chosen during Theorem 4.9 is "good" in exactly the same manner as in the proof of Theorem 4.11. The total expected time becomes $O_\varepsilon(n \log n + \frac{n}{k_{\rm opt}} \log^2 n)$. As before, for $k_{\rm opt} < \log n$, we can switch to the $O(n \log n + nk_{\rm opt})$ algorithm. $\square$

An intriguing question that remains is whether approximate LP with violations can be solved in linear time, Las Vegas or Monte Carlo. We don't know how even in 2D. For constant-factor approximations, however, we can obtain a linear-time Monte Carlo algorithm in 2D, by using the following subroutine:

Lemma D. *Let H^- and H^+ be two sets of lines in $\mathbb{R}^2$ of total size n. Given K, we can find the largest $k \leq K$ such that there is a line separating the k-level of H^- and the $(|H^+| - k)$-level of H^+, in $O(n + K(n/K)^\delta \log n)$ expected time for any fixed $\delta > 0$.*

Proof. The result follows from known techniques by Chan [13, 9]. Specifically, Chan [13] considered the problem of finding a point of maximum Tukey depth for a given point set; in the dual, this problem in 2D is equivalent to the problem stated in the lemma in the case when $H^- = H^+$ and $K = n$. The same algorithm can be applied to the setting where the input consists of two sets H^- and H^+. The expected running time is $O(n \log n)$.

To obtain a time bound that depends on K, we note that the algorithm in [13] uses an "implicit linear programming" technique to reduce the problem to the following decision problem:

> Given $k \le K$ and a line ℓ, decide whether ℓ separates the k-level of H^- and the $(|H^+| - k)$-level of H^+.

Specifically, if the decision problem can be solved in $D(n, K)$ time and $D(n, K)/n^\delta$ is a monotone increasing function in n for some fixed $\delta > 0$, then the optimization problem can be solved in $O(D(n, K))$ expected time. The decision problem amounts to deciding whether a given line intersects a k-level ($k \le K$); as observed in [9, proof of Theorem 5.2], this can actually be done in $O(n + K \log n)$ time (the subproblem reduces to a one-dimensional problem of selecting the K smallest and K largest elements from a list of n elements). We replace $O(n + K \log n)$ with the upper bound $D(n, K) = O(n + K(n/K)^\delta \log n)$ (so that $D(n, K)/n^\delta$ is monotone increasing). □

Theorem 4.13. *We can obtain a factor-$O(1)$ approximation for LP with violations in 2D in $O_\varepsilon(n)$ expected time. This algorithm is correct w.h.p.*

Proof. Let H^- (resp. H^+) be the bounding lines of the lower (resp. upper) halfplanes. Let L_k^- denote the k-level of H^- and L_k^+ denote the $(|H^+| - k)$-level of H^+. Given an upper bound K to k_{opt}, we use Lemma D to find the largest $k \le K$ such that there is a line separating L_k^- and L_k^+.

We claim that this k approximates k_{opt} to within a factor of 4. Suppose there is a point q that violates k constraints. Then q is below L_k^- and above L_k^+, so no line can separate the L_k^- and L_k^+. Conversely, suppose no line separates L_k^- and L_k^+. Then the upper hull of L_k^- and the lower hull of L_k^+ must intersect, so there is a point q below an upper hull edge u^-v^- of L_k^- and above a lower hull edge u^+v^+ of L_k^+. Since u^- and v^- are on the k-level of H^-, q lies below the $(2k)$-level of H^-; similarly, q is above the $(2k)$-level of H^+. So q violates at most $4k$ constraints.

We set $K = n/\log^2 n$. If $k_{\text{opt}} \le n/\log^2 n$, the algorithm will successfully compute k_{opt} in expected time $O(n+K(n/K)^\delta \log n) = O(n)$. If $k_{\text{opt}} > n/\log^2 n$, we can switch to the Monte Carlo algorithm in Theorem 4.10 running in time $O_\varepsilon(\frac{n}{k_{\text{opt}}} \log n \log\log n + \log^{O(1)} n)$. □

Remark. The approximation factor 4 can be improved, but it is not clear how to bring the constant arbitrarily close to 1.

References

[1] P. Afshani and T. M. Chan. Optimal halfspace range reporting in three dimensions. In *Proceedings of the 20th Annual Symposium on Discrete Algorithms*, pages 180–186, 2009.

[2] P. K. Agarwal and J. Erickson. Geometric range searching and its relatives. In B. Chazelle, J. E. Goodman, and R. Pollack, editors, *Advances in Discrete and Computational Geometry*, volume 223 of *Contemporary Mathematics*, pages 1–56. American Mathematical Society, Providence, RI, 1999.

[3] B. Aronov and S. Har-Peled. On approximating the depth and related problems. In *Proceedings of the 16th Annual ACM–SIAM Symposium on Discrete Algorithms*, pages 886–894, 2005. Updated version at `http://valis.cs.uiuc.edu/~sariel/research/papers/04/depth/`, downloaded in November 2006.

[4] S. Arya and D. M. Mount. Approximate range searching. *Computational Geometry: Theory and Applications*, 17(3-4):135–152, 2000.

[5] M. de Berg and O. Schwarzkopf. Cuttings and applications. *International Journal of Computational Geometry and Applications*, 5:343–355, 1995.

[6] M. Bern and D. Eppstein. Multivariate regression depth. In *Proceedings of the 16th Annual Symposium on Computational Geometry*, pages 315–321, 2000.

[7] T. M. Chan. Fixed-dimensional linear programming queries made easy. In *Proceedings of the 12th Annual ACM Symposium on Computational Geometry*, pages 284–290, 1996.

[8] T. M. Chan. Output-sensitive results on convex hulls, extreme points, and related problems. *Discrete and Computational Geometry*, 16:369–387, 1996.

[9] T. M. Chan. Geometric applications of a randomized optimization technique. *Discrete and Computational Geometry*, 22:547–567, 1999.

[10] T. M. Chan. Low-dimensional linear programming with violations. *SIAM Journal on Computing*, 34:879–893, 2000.

[11] T. M. Chan. Random sampling, halfspace range reporting, and construction of $(\leq k)$-levels in three dimensions. *SIAM Journal on Computing*, 30(2):561–575, 2000.

[12] T. M. Chan. On enumerating and selecting distances. *International Journal of Computational Geometry and Applications*, 11:291–304, 2001.

[13] T. M. Chan. An optimal randomized algorithm for maximum Tukey depth. In *Proceedings of the 15th Annual ACM–SIAM Symposium on Discrete Algorithms*, pages 430–436, 2004.

[14] B. Chazelle, L. J. Guibas, and D. T. Lee. The power of geometric duality. *BIT*, 25(1):76–90, 1985.

[15] K. L. Clarkson and P. W. Shor. Applications of random sampling in computational geometry, II. *Discrete and Computational Geometry*, 4:387–421, 1989.

[16] E. Cohen. Size-estimation framework with applications to transitive closure and reachability. *Journal of Computer and System Sciences*, 55:441–453, 1997.

[17] H. Edelsbrunner. *Algorithms in Combinatorial Geometry*, volume 10 of *EATCS Monographs on Theoretical Computer Science*. Springer-Verlag, Heidelberg, West Germany, 1987.

[18] H. Edelsbrunner and E. Welzl. Constructing belts in two-dimensional arrangements with applications. *SIAM Journal on Computing*, 15:271–284, 1986.

[19] S. Har-Peled and M. Sharir. Relative ε-approximations in geometry. `http://valis.cs.uiuc.edu/~sariel/research/papers/06/relative/`, 2006. Also with B. Aronov, in *Proceedings of the 23rd ACM Symposium on Computational Geometry*, pages 327–336, 2007.

[20] S. Hart and M. Sharir. Nonlinearity of Daveport-Schinzel sequences and of generalized path compression schemes. *Combinatorica*, 6(2):151–177, 1986.

[21] D. Haussler and E. Welzl. Epsilon-nets and simplex range queries. *Discrete and Computational Geometry*, 2:127–151, 1987.

[22] H. Kaplan and M. Sharir. Randomized incremental constructions of three-dimensional convex hulls and planar Voronoi diagrams, and approximate range counting. In *Proceedings of the 17th Annual ACM–SIAM Symposium on Discrete Algorithms*, pages 484–493, 2006.

[23] H. Kaplan, E. Ramos, and M. Sharir. Range minima queries with respect to a random permutation, and approximate range counting. To appear in *Discrete and Computational Geometry*.

[24] H. Kaplan, E. Ramos, and M. Sharir. The overlay of minimization diagrams in a randomized incremental construction. Manuscript, 2007.

[25] S. Langerman and W. Steiger. The complexity of hyperplane depth in the plane. *Discrete and Computational Geometry*, 30(2):299–309, August 2003.

[26] J. Matoušek. Reporting points in halfspaces. *Computational Geometry: Theory and Applications*, 2(3):169–186, 1992.

[27] J. Matoušek. Range searching with efficient hierarchical cuttings. *Discrete and Computational Geometry*, 10(2):157–182, 1993.

[28] J. Matoušek. On geometric optimization with few violated constraints. *Discrete and Computational Geometry*, 14:365–384, 1995.

[29] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, New York, NY, 1995.

[30] K. Mulmuley. *Computational Geometry: An Introduction Through Randomized Algorithms*. Prentice Hall, Englewood Cliffs, NJ, 1994.

[31] F. P. Preparata and M. I. Shamos. *Computational Geometry: An Introduction*. Springer-Verlag, New York, NY, 1985.

[32] E. A. Ramos. On range reporting, ray shooting and k-level construction. In *Proceedings of the 14th Annual Symposium on Computational Geometry*, pages 390–399, 1999.

[33] P. J. Rousseeuw and M. Hubert. Regression depth. *Journal of American Statistical Association*, 94(446):388–402, 1999.

[34] M. Sharir, S. Smorodinsky, and G. Tardos. An improved bound for k-sets in three dimensions. *Discrete and Computational Geometry*, 26:195–204, 2001.