

Relative ε -Approximations in Geometry*

Sariel Har-Peled[†]

Micha Sharir[‡]

November 20, 2006

Full version: <http://www.uiuc.edu/~sariel/papers/06/relative>

Abstract

We re-examine relative ε -approximations, previously studied in [Pol86, Hau92, LLS01], [CKMS06], and their relation to certain geometric problems. We give a simple constructive proof of their existence in general range spaces with finite VC-dimension, and of a sharp bound on their size, close to the best known one. We then give a construction of smaller-size relative ε -approximations for range spaces that involve points and halfspaces in two and higher dimensions. The planar construction is based on a new structure—spanning trees with small *relative crossing number*, which we believe to be of independent interest. We also consider applications of the new structures for approximate range counting and related problems.

*Work on this paper by Sariel Har-Peled was partially supported by an NSF CAREER award CCR-0132901. Work by Micha Sharir was supported by a grant from the U.S.-Israel Binational Science Foundation, by NSF Grant CCF-05-14079, by Grant 155/05 from the Israel Science Fund, and by the Hermann Minkowski–MINERVA Center for Geometry at Tel Aviv University.

[†]Department of Computer Science, University of Illinois, 201 N. Goodwin Avenue, Urbana, IL, 61801, USA; sariel@uiuc.edu.

[‡]School of Computer Science, Tel Aviv University, Tel Aviv 69978 Israel, and Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA; michas@post.tau.ac.il.

1 Introduction

The main problem that has motivated the study in this paper is *approximate range counting*. In abstract terms, we are given a *range space* (X, \mathcal{R}) , where X is a set of n objects and \mathcal{R} is a collection of subsets of X , called *ranges*. In a typical geometric setting, X is a subset of some infinite ground set U (e.g., \mathbb{R}^d), and $\mathcal{R} = \{R \cap X \mid R \in \mathcal{R}_U\}$, where \mathcal{R}_U is a collection of subsets (*ranges*) of U of some simple shape (such as halfspaces). (To simplify the notation, we will use \mathcal{R} and \mathcal{R}_U interchangeably.) The goal is to preprocess X into a data structure that supports efficient queries of the form: Given $R \in \mathcal{R}_U$, compute a number t such that

$$(1 - \varepsilon)|X \cap R| \leq t \leq (1 + \varepsilon)|X \cap R|.$$

We refer to such an estimate t as an ε -*approximate count* of $X \cap R$.

The motivation for seeking approximate range counting techniques is that exact range counting is expensive. For instance, consider the classical *halfspace range counting* problem [Mat92], which is the main specific problem studied in this paper. Here, for a point set of size n in \mathbb{R}^d , for $d \geq 2$, the best known algorithm for exact halfspace range counting with near-linear storage takes $O(n^{1-1/d})$ time [Mat92]. As shown in several recent papers, faster solutions exist, for the approximate case, in which the query time is close to $O(n^{1-1/\lfloor d/2 \rfloor})$ [AH05, AH06, AS06, KS06, KRS06]. In particular, the paper [AS06] uses constructs of the kind studied in this paper.

Notice that the problem becomes more challenging as the size of $X \cap R$ decreases. At the extreme, when $|X \cap R| < 1/\varepsilon$, we must produce the count *exactly*. In particular, we need to be able to detect (without any error) *empty ranges*, i.e., those satisfying $X \cap R = \emptyset$. Thus approximate range counting (in the sense defined above) is at least as hard as range emptiness detection.

We make the standard assumption that the range space (X, \mathcal{R}) (or, in fact, (U, \mathcal{R}_U)) has *finite* (i.e., independent of n) *VC-dimension* δ which is indeed the case in many geometric applications; see [Cha01, HW87, Mat99, PA95] for definitions and more details.

Epsilon-approximations. A standard and general technique for tackling the approximate range counting problem is to use ε -approximations. An (*absolute-error*) ε -*approximation* for (X, \mathcal{R}) is a subset $A \subset X$ such that, for each $R \in \mathcal{R}$,

$$\left| \frac{|A \cap R|}{|A|} - \frac{|X \cap R|}{|X|} \right| < \varepsilon. \tag{1}$$

As shown by Vapnik and Chervonenkis [VC71] (see also [Cha01, Mat99, PA95]), there always exist absolute-error ε -approximations of size $\frac{c\delta}{\varepsilon^2} \log \frac{\delta}{\varepsilon}$, where c is an absolute constant. As a matter of fact, any random sample of these many elements of X is an ε -approximation with constant probability. Moreover, such a sample of size $\frac{c\delta}{\varepsilon^2} \log \frac{\delta}{\varepsilon} + \frac{c}{\varepsilon^2} \log \frac{1}{q}$ is an ε -approximation with probability at least $1 - q$, for a sufficiently large constant c . Therefore, to guarantee success with *high probability*, i.e., with probability of failure at most $1/n^{O(1)}$, one needs to choose a sample of size $\frac{c\delta}{\varepsilon^2} \log \frac{\delta}{\varepsilon} + \frac{c'}{\varepsilon^2} \log n = O\left(\frac{1}{\varepsilon^2} \log n\right)$. Approximations of size $O\left(\frac{\delta}{\varepsilon^2} \log \frac{\delta}{\varepsilon}\right)$ can be constructed in deterministic time $O(\delta)^{3\delta} (\varepsilon^{-2} \log \frac{\delta}{\varepsilon})^\delta n$ [Cha04]. In fact, there always exist smaller (absolute-error) ε -approximations, of size

$$O\left(\frac{1}{\varepsilon^{2-2/(\delta'+1)}} \log^{c-1/(\delta'+1)} \frac{1}{\varepsilon}\right),$$

where δ' is the exponent of either the *primal* shatter function (and then $c = 2$) or the *dual* shatter function of the range space (X, \mathcal{R}) (and then $c = 1$); see [Cha01, Cha04, MWW93]. The time to construct these improved nets is roughly the same as above for the dual case. For the primal case the proof is only existential.

In this paper, we consider a variant of this classical structure, which provides *relative-error approximations*. Ideally, we want a subset $A \subset X$ such that, for each $R \in \mathcal{R}$,

$$(1 - \varepsilon) \frac{|X \cap R|}{|X|} \leq \frac{|A \cap R|}{|A|} \leq (1 + \varepsilon) \frac{|X \cap R|}{|X|}. \quad (2)$$

This “definition” suffers however from the same syndrome as the definition of approximate range counting; that is, as $|X \cap R|$ shrinks, the absolute precision of the approximation has to increase. At the extreme, when $A \cap R = \emptyset$, $X \cap R$ must also be empty; in general, we cannot guarantee this property, unless we take $A = X$, which defeats the entire purpose of using small-size ε -approximations to speed up approximate counting.

For this reason, we refine the definition as follows: a *relative (p, ε) -approximation* is a subset $A \subset X$ that satisfies Eq. (2) for each $R \in \mathcal{R}$ with $|R| \geq pn$, where $0 < p < 1$ is another fixed parameter. It is known (see [LLS01]) that there exist subsets with this property of size $\frac{c\delta}{\varepsilon^2 p} \log \frac{1}{p}$, where c is an absolute constant. As a matter of fact, any random sample of these many elements of X is a relative (p, ε) -approximation with constant probability. To guarantee success with probability at least $1 - q$, one needs to sample $\frac{c}{\varepsilon^2 p} \left(\delta \log \frac{1}{p} + \log \frac{1}{q} \right)$ elements of X , for a sufficiently large constant c [LLS01].

To appreciate the above bound on the size of relative (p, ε) -approximations, it is instructive to observe that, for a given parameter p , any absolute error (εp) -approximation A will approximate “large” ranges (of size at least pn) to within *relative* error ε , as is easily checked, so it is a relative (p, ε) -approximation. However, the Vapnik-Chervonenkis bound on the size of A , namely, $\frac{c\delta}{\varepsilon^2 p^2} \log \frac{\delta}{\varepsilon p}$, is larger by roughly a factor of $1/p$ than the bound of [CKMS06, LLS01] stated above.

The existence of a relative (p, ε) -approximation A provides a simple mechanism for approximate range counting: For a range R , count $A \cap R$ exactly, say, by brute force in $O(|A|)$ time, and output $|A \cap R| \cdot |X| / |A|$ as an ε -approximate count of $X \cap R$. However, this will work only for ranges of size at least pn . Aronov and Sharir [AS06] show that an appropriate incorporation of relative (p, ε) -approximations into standard range searching data structures yields a procedure for approximate range counting that works, quite efficiently, for ranges of any size.

Our results. In this paper, we present several constructions and bounds involving relative (p, ε) -approximations. We first give an alternative construction of general relative (p, ε) -approximations, which follows the standard discrepancy-based construction of absolute-error (p, ε) -approximations [Cha01], but uses a more careful analysis that shows that the resulting set is indeed a relative (p, ε) -approximation. That is, for a given threshold $0 < p < 1$, the resulting set A is of size $O\left(\frac{\delta}{\varepsilon^2 p} \log \frac{\delta}{\varepsilon p}\right)$, and, for any range R of size at least pn , gives an absolute approximation error of $\varepsilon |R \cap X| / |X|$. That is,

$$\left| \frac{|R \cap A|}{|A|} - \frac{|R \cap X|}{|X|} \right| < \varepsilon \frac{|R \cap X|}{|X|}.$$

Thus, for such ranges R , the approximate count $\frac{|R \cap A|}{|A|} \cdot |X|$ is an ε -approximate count of $R \cap X$. The construction is randomized, but can easily be derandomized using standard techniques, similar

to those used to obtain the deterministic constructions cited above.

Note that the size of our approximation is slightly worse than the bound of [LLS01], when $p \gg \varepsilon$. However, our construction is useful because it can be enhanced, in certain geometric situations, to yield relative (p, ε) -approximations of smaller size. We study two cases in detail, one involving points in \mathbb{R}^2 and halfplane ranges, and the other involving points in \mathbb{R}^d , $d \geq 3$, and halfspace ranges. Given a threshold parameter $0 < p < 1$, the size of the approximation set is $O\left(\frac{1}{\varepsilon^{4/3}p} \log \frac{1}{\varepsilon p}\right)$ in the plane, and $O\left(\frac{1}{\varepsilon^{3/2}p} \log \frac{1}{\varepsilon p}\right)$ in 3-space; the bounds for higher dimensions are spelled out in Section 4.2.

In the planar case, the construction is based on an interesting generalization of spanning trees with small crossing number, a result that we believe to be of independent interest. Specifically, we show that any finite point set P in the plane has a spanning tree with the following property: For any $k \leq |P|$, any k -shallow line (a line that has at most k points of P in one of the halfplanes that it bounds) crosses at most $O(\sqrt{k} \log(n/k))$ edges of the tree. (The classical construction of Welzl [Wel92] guarantees this property only for $k = n$; i.e., it guarantees the uniform crossing number $O(\sqrt{n})$.) We refer to such a tree as a *spanning tree with low relative crossing number*, and show how to use it in the construction of small-size relative (p, ε) -approximations.

Things are more complicated in three (and higher) dimensions. We were unable to extend the planar construction of spanning trees with low relative crossing number to \mathbb{R}^3 (nor to higher dimensions), and this remains an interesting open problem. (We give, in the full version, a counterexample that shows why the planar construction cannot be extended “as is” to 3-space.) Instead, we base our construction on the shallow partition theorem of Matoušek [Mat91b], and construct a set A of size $O\left(\frac{1}{\varepsilon^{3/2}p} \log \frac{1}{\varepsilon p}\right)$, which yields an absolute approximation error of at most εp for halfspaces that contain *at most* pn points. Note that this is the “wrong” inequality—to guarantee small relative error we need this to hold for all ranges with *at least* pn points. To overcome this difficulty, we construct a *sequence* of approximation sets, each capable of producing a relative ε -approximate count for ranges that have roughly a fixed size, where these sizes grow geometrically, starting at pn and ending at roughly n . The sizes of these sets decrease geometrically, so that the size of the first set (that caters to ranges with about pn points), which is $O\left(\frac{1}{\varepsilon^{3/2}p} \log \frac{1}{\varepsilon p}\right)$, dominates asymptotically the overall size of all of them. We output this sequence of sets, and show how to use them to obtain an ε -approximate count of any range with at least pn points.

The situation is somewhat more complicated in higher dimensions. The basic approach used in the three-dimensional case can be extended to higher dimensions, using the appropriate version of the shallow partition theorem. However, the bounds get somewhat more complicated, and apply only under certain restrictions on the relationships between ε , p , and n . We refer the reader to Section 4.2, where these bounds and restrictions are spelled out in detail.

Due to space limitations, we omit many details in this abstract. They are available in the full version of this paper [HS06].

2 Relative (p, ε) -Approximations in General Range Spaces

Our construction is based on the following well known technical ingredient.

Theorem 2.1 ([Cha01]) *Let (X, \mathcal{R}) be a set system defined over $n = |X|$ elements, where $\mathcal{R} = \{R_1, \dots, R_m\}$. One can construct, in $O(nm)$ deterministic time, a coloring $\chi : X \rightarrow \{-1, 1\}$, where each color class has exactly $n/2$ elements, such that, for any $j = 1, \dots, m$, the discrepancy of R_j is $|\chi(R_j)| \leq \sqrt{2|R_j| \ln(2m)}$, where $\chi(R_j) = \sum_{x \in R_j} \chi(x)$.*

Remarks: (1) Clearly, a more precise statement is that $\chi(R_j) \leq \sqrt{2\xi(R_j)\ln(2m)}$, for each j , where $\xi(R)$ is the number of pairs of the matching that the range R separates (or “crosses”). Later, we will use special matchings with small values of the quantities $\xi(R)$, and consequently obtain improved discrepancy bounds, which in turn will lead to improved bounds on the size of the relative approximation sets.

(2) We refer to the process of extracting a subset of X of half the size, by using low-discrepancy coloring, as *halving*. One can show that if the range space (X, \mathcal{R}) has VC dimension δ , then a halving can be constructed in $O(|X|^{\delta+1})$ deterministic time.

Setting $W = \chi^{-1}(1)$ and $B = \chi^{-1}(-1)$, Theorem 2.1 states that, for each R_j ,

$$||W \cap R_j| - |B \cap R_j|| \leq \sqrt{2|R_j|\ln(2m)},$$

and since $|W \cap R_j| = |X \cap R_j| - |B \cap R_j|$, we have, for each R_j ,

$$||X \cap R_j| - 2|B \cap R_j|| \leq \sqrt{2|R_j|\ln(2m)}. \quad (3)$$

Assume now that the set system (X, \mathcal{R}) has finite VC dimension δ . This implies (by Sauer’s lemma [Cha01]) that $|\mathcal{R}| \leq (ne/\delta)^\delta$. To simplify the exposition, we will assume that $\delta > 2$, so that $m = |\mathcal{R}| \leq n^\delta$. Moreover, for any subset $X' \subseteq X$ of size n' , the number of distinct ranges in $\mathcal{R}' = \{R_j \cap X' \mid j = 1, \dots, m\}$ is $m' \leq (n')^\delta$.

We construct a sequence of subsets $P_k \subseteq P_{k-1} \subseteq \dots \subseteq P_0 = X$, with k to be determined shortly, as follows. Set $P_0 = X$ and, for each $i = 1, 2, \dots$, after constructing P_{i-1} , we take a coloring χ_i of P_{i-1} by two colors, as provided in Theorem 2.1, and let P_i be the subset of points of P_{i-1} colored -1 . Put $n_i = |P_i| = n/2^i$, for $i = 1, \dots, k$.

Lemma 2.2 *Let $0 < p < 1$ be a given parameter, and assume that k satisfies $n_k \geq \frac{4\delta}{p} \ln \frac{4\delta}{p}$. Then, for any range R in \mathcal{R} that contains at least pn points of X , and for each $i = 0, \dots, k$, we have*

$$|P_i \cap R| \leq c|P_0 \cap R|/2^i,$$

for some absolute constant c .

Proof: Omitted; see the full version [HS06]. ■

Lemma 2.3 *There exists an index k such that $|P_k| = n_k = \Theta\left(\frac{\delta}{\varepsilon^2 p} \ln \frac{\delta}{\varepsilon p}\right)$, and such that, for any range $R \in \mathcal{R}$ that contains at least pn points, we have $||R \cap P| - 2^k|R \cap P_k|| \leq \varepsilon|R \cap P|$.*

Proof: Fix a range R , and use the notation in the proof of Lemma 2.2. This lemma implies that

$$\begin{aligned} \left| \widehat{\lambda} - 2^k \lambda_k \right| &\leq \sum_{i=1}^k 2^{i-1} \sqrt{2\delta \lambda_{i-1} \ln(2n_i)} \\ &\leq \sum_{i=1}^k 2^{i-1} \sqrt{2\delta \left(c \frac{\widehat{\lambda}}{2^{i-1}} \right) \ln(2n_i)} \leq c_1 2^{k/2} \sqrt{\widehat{\lambda} \ln n_k}, \end{aligned}$$

for some constant c_1 which is proportional to $\sqrt{\delta}$. We want the right-hand side to be smaller than $\varepsilon \widehat{\lambda}$; that is, $c_1 2^{k/2} \sqrt{\widehat{\lambda} \ln n_k} \leq \varepsilon \widehat{\lambda}$, which is equivalent to $c_1^2 2^k \ln n_k \leq \varepsilon^2 \widehat{\lambda}$. Since we assume $\widehat{\lambda} \geq pn$, this will hold if we require that $c_1^2 2^k \ln n_k \leq \varepsilon^2 pn$. Since $n_k = n/2^k$, this amounts to requiring that

$$\frac{c_1^2}{\varepsilon^2 p} \leq \frac{n_k}{\ln n_k},$$

which holds for $n_k \geq 2 \frac{c_1^2}{\varepsilon^2 p} \ln \frac{c_1^2}{\varepsilon^2 p}$. Note that this bound meets the lower bound requirement on n_k , given in Lemma 2.2, provided that c_1 is a sufficiently large multiple of $\sqrt{\delta}$. This completes the proof of the lemma. \blacksquare

We have thus shown the following main result of this section.

Theorem 2.4 *Let (X, \mathcal{R}) be a range space with finite VC dimension δ , where $|X| = n$, and let $0 < \varepsilon < 1$ and $0 < p < 1$ be given parameters. Then one can construct a set $A \subseteq X$ of size $O(\frac{\delta}{\varepsilon^2 p} \log \frac{\delta}{\varepsilon p})$, such that, for each range $R \in \mathcal{R}$ of at least pn points, we have*

$$\left| \frac{|R \cap A|}{|A|} - \frac{|R \cap X|}{|X|} \right| \leq \varepsilon \frac{|R \cap X|}{|X|}.$$

In other words, A is a relative (p, ε) -approximation for (X, \mathcal{R}) . The set A can be computed either in $O(\delta)^{3\delta} \left(\frac{1}{p^2 \varepsilon^2} \log \frac{\delta}{\varepsilon}\right)^\delta n$, or in $O(n^{\delta+1})$ deterministic time.

Proof: Take A to be the set P_k from Lemma 2.3. We only need to verify that it provides the required approximation. Fix a range $R \in \mathcal{R}$, and let $\hat{\lambda} = |R \cap X|$ and $\lambda_k = |R \cap P_k|$. We have that $|\hat{\lambda} - 2^k \lambda_k| \leq \varepsilon \hat{\lambda}$, or, equivalently,

$$\left| \frac{|R \cap X|}{|X|} - \frac{|R \cap A|}{|A|} \right| = \left| \frac{|R \cap X|}{n} - \frac{|R \cap A|}{n/2^k} \right| = \left| \frac{\hat{\lambda}}{n} - \frac{2^k \lambda_k}{n} \right| \leq \frac{\varepsilon \hat{\lambda}}{n},$$

as required. The bound on the construction time is given in the full version [HS06]. \blacksquare

3 Relative (p, ε) -Approximations in the Plane

In this section we present a construction of smaller-size relative (p, ε) -approximations for the range space involving a set of points in the plane and the set of halfplanes as ranges. The key ingredient of the construction is the result of the following subsection.

3.1 Spanning trees with small relative crossing number

We derive a refined “weight-sensitive” version of the classical construct of *spanning trees with small crossing number*, as obtained by Chazelle and Welzl [CW89], with a simplified construction given in [Wel92]. We believe that this refined version is of independent interest, and expect it to have additional applications in the future.

In accordance with standard notation used in the literature, we denote from now on the underlying point set by P .

We first recall the standard result:

Theorem 3.1 ([Wel92]) *Let P be a set of n points in \mathbb{R}^d . One can compute a spanning tree \mathcal{T} of P such that each hyperplane in \mathbb{R}^d crosses at most $O(n^{1-1/d})$ edges of \mathcal{T} .*

Definition 3.2 Let P be a set of n points in the plane. For a line ℓ , let w_ℓ^+ (resp., w_ℓ^-) be the number of points of P lying above (resp., below or on) ℓ , and define the *weight* of ℓ , denoted by w_ℓ , to be $\min(w_\ell^+, w_\ell^-)$.

Let $\mathcal{D}(P, k)$ be the intersection of all closed half-spaces that contain at least $n - k$ points of P . Note that, by the centerpoint theorem (see [Mat03]), the set $\mathcal{D}(P, k)$ is not empty for $k < n/3$. Moreover, the set $\mathcal{D}(P, k)$ is a convex polygon, since it is equal to the intersection of a finite number of halfplanes.

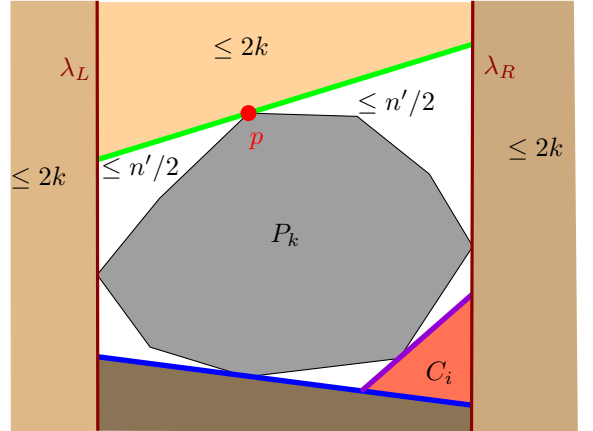
Lemma 3.3 *Let P be a set of n points in the plane. (i) Any line ℓ that avoids the interior of $Q_k = \mathcal{D}(P, k)$ has weight $w_\ell \leq 2k$. (ii) Any line ℓ that intersects the interior of Q_k has weight $w_\ell > k$.*

Proof: Omitted; see the full version [HS06]. ■

Lemma 3.4 *The set $P \setminus \mathcal{D}(P, k)$ can be covered by disjoint convex polygons C_1, \dots, C_u , each containing at most $2k$ points of P , such that any line intersects at most $O(\log(n/k))$ of these polygons.*

Proof: We construct the polygons C_i iteratively, as follows. Let λ_L and λ_R be the two vertical lines supporting $\mathcal{D}(P, k)$ on its left and on its right, respectively. C_1 (resp., C_2) is the halfplane to the left (resp., right) of λ_L (resp., λ_R). The construction maintains the invariant that the complement of the union of the polygons C_1, \dots, C_i constructed so far is a convex polygon K_i that contains $\mathcal{D}(P, k)$ and passes through some of its vertices, so that $K_i \setminus \mathcal{D}(P, k)$ consists of pairwise disjoint connected “pockets”. (Initially, after constructing C_1 and C_2 , we have two pockets—the regions lying respectively above and below $\mathcal{D}(P, k)$, between λ_L and λ_R .)

Each step of the construction picks a pocket that contains more than $2k$ points of P , and finds a line ℓ that supports $\mathcal{D}(P, k)$ at a vertex of the pocket, and subdivides the pocket into two sub-pockets and a third piece that lies on the other side of ℓ . The line ℓ is chosen so that the two resulting sub-pockets contain an equal number of points of P . The third piece, which clearly contains at most $2k$ points of P , is taken to be the next polygon C_{i+1} , and the construction continues in this manner until each pocket has at most $2k$ points. We then terminate the construction, adding all the pockets to the output collection of polygons. Note that each “non-final” C_i is a triangle, having a “base” whose relative interior passes through a vertex of $\mathcal{D}(P, k)$, and two other sides, each of which is a portion of a base of an earlier polygon. A similar property holds for final polygons, except that their “base” is a connected portion of the boundary of $\mathcal{D}(P, k)$, possibly consisting of several edges.



We claim that each line ℓ intersects at most $O(\log(n/k))$ polygons. For this, define the *weight* $w(C_i)$ of a non-final polygon C_i , for $i \geq 3$, to be the number of points of P in the pocket that was split when C_i was created; the weight of each final polygon is the number of points of P that it contains, which is at most $2k$. Define the *level* of C_i to be $\lfloor \log_2 w(C_i) \rfloor$. It is easily checked that ℓ crosses at most two final polygons, and it can cross both (non-base) sides of at most one (final or non-final) polygon. Any other polygon C_i crossed by ℓ is such that ℓ enters it through its base, reaching it from another polygon whose level is, by construction, strictly smaller than that of C_i . Since there are only $O(\log(n/k))$ distinct levels, the claim follows. ■

Lemma 3.5 *Let $k > 0$ be a pre-specified parameter. One can construct a spanning tree \mathcal{T} for $P \setminus Q_k$ such that each line intersects at most $O(\sqrt{k} \log(n/k))$ edges of \mathcal{T} .*

Proof: Construct the decomposition of $P \setminus \mathcal{D}(P, k)$ into u covering polygons C_1, \dots, C_u , using Lemma 3.4. Let Σ denote the union of the boundaries of all these polygons. Regard Σ as a straight-edge graph on the vertices of the polygons, and let \mathcal{T}_0 denote any spanning tree of Σ .

For each $i = 1, \dots, u$, construct a spanning tree \mathcal{T}_i of $P \cap C_i$ with crossing number $O(k^{1/2})$, using Theorem 3.1. In addition, connect one point of $P \cap C_i$ to an arbitrary vertex of C_i . Let \mathcal{T}^* denote the union of $\mathcal{T}_0, \mathcal{T}_1, \dots, \mathcal{T}_u$, plus the connecting segments just introduced. Clearly, \mathcal{T}^* is a tree.

Let ℓ be any line in the plane. The proof of the preceding lemma implies that ℓ intersects at most $O(\log(n/k))$ of the polygons C_i . Hence, ℓ crosses at most $O(\log(n/k))$ edges of \mathcal{T}_0 , at most $O(\log(n/k))$ of the connecting segments, and it can cross edges of at most $O(\log(n/k))$ trees \mathcal{T}_i , for $i = 1, \dots, u$. Since ℓ crosses at most $O(k^{1/2})$ edges of each such tree, we conclude that ℓ crosses at most $O(\sqrt{k} \log(n/k))$ edges of \mathcal{T}^* .

Finally, we get rid of the extra ‘‘Steiner vertices’’ of \mathcal{T}^* (those not belonging to $P \setminus Q_k$) in a straightforward manner, by making \mathcal{T}^* a rooted tree, at some point of $P \setminus Q_k$, and by replacing each path connecting a point $u \in P \setminus Q_k$ to an ancestor $v \in P \setminus Q_k$, where all inner vertices of the path are Steiner points, by the straight segment uv . This produces a spanning tree \mathcal{T} of $P \setminus Q_k$, whose crossing number is at most that of \mathcal{T}^* . ■

Theorem 3.6 *Given a point set P of n points in the plane, one can construct a spanning tree \mathcal{T} for P such that any line ℓ crosses at most $O(\sqrt{w_\ell} \log(n/w_\ell))$ edges of \mathcal{T} . The tree \mathcal{T} can be constructed in $O(n^{1+\varepsilon})$ (deterministic) time, for any fixed $\varepsilon > 0$.*

Proof: We construct a sequence of subsets of P , as follows. Put $P_0 = P$. At the i th step, $i \geq 1$, consider the polygon $Q_i = \mathcal{D}(P_{i-1}, 2^i)$, and let $P_i = P_{i-1} \cap Q_i$. We stop when P_i becomes empty.

For each i , construct a spanning tree \mathcal{T}_i for $P_{i-1} \setminus Q_i$, using Lemma 3.5 (with $k = 2^i$). Connect the resulting trees by straight segments into a single spanning tree \mathcal{T} of P .

We claim that \mathcal{T} is the desired spanning tree. Indeed, consider an arbitrary line ℓ of weight k . Observe that ℓ cannot cross any of the polygons Q_i , for $i > U = \lceil \log_2 k \rceil$, since any line that crosses such a polygon must be of weight at least $2^{i+1} > k$, by Lemma 3.3(ii).

Thus ℓ crosses only the first $O(\log k)$ layers of our construction. Hence, the number of edges of \mathcal{T} that ℓ crosses is at most

$$\sum_{i=1}^U O\left(\sqrt{2^i} \log(n/2^i)\right) = O(\sqrt{k} \log(n/k)),$$

as asserted. ■

3.2 Relative (p, ε) -approximations for halfplanes

We can turn the above construction of a spanning tree with small relative crossing number into a construction of a relative (p, ε) -approximation for a set of points in the plane and halfplane ranges, as follows.

Let P be a set of n points in the plane, and let \mathcal{T} be a spanning tree of P as provided in Theorem 3.6. We replace \mathcal{T} by a perfect matching M of P , with the same relative crossing number, i.e., the number of pairs of M that are separated by a halfplane of weight k is at most $O(\sqrt{k} \log(n/k))$. This is done in a standard manner—we first convert \mathcal{T} to a spanning path with the same relative crossing number, and then pick every other edge of the path.

We now construct a coloring of P with low discrepancy, by randomly coloring the points in each pair of M , as in Theorem 2.1. The analysis in the proof of that theorem (see the Remark following the theorem) yields the following variant.

Lemma 3.7 *Given a set P of n points in the plane, one can construct a coloring $\chi : P \mapsto \{-1, 1\}$, such that, for any halfplane h that contains k points of P , we have*

$$\chi(h \cap P) = O(k^{1/4} \log n).$$

The coloring is balanced—each color class consists of exactly $n/2$ points of P .

We now continue with the analysis of Section 2, using the improved discrepancy bound of the preceding lemma. This can be shown to lead to the following improved bound.

Theorem 3.8 *Given a set P of points in the plane, and parameters $0 < \varepsilon < 1$ and $0 < p < 1$, one can construct a relative (p, ε) -approximation subset of size $O\left(\frac{1}{\varepsilon^{4/3} p} \log^{4/3} \frac{1}{\varepsilon p}\right)$.*

See the full version [HS06] for details concerning the construction time.

4 Relative (p, ε) -Approximations in Higher Dimensions

4.1 Relative (p, ε) -Approximations in \mathbb{R}^3

The construction in higher dimensions is different from the planar one, because of our present inability to extend the construction of spanning trees with low relative crossing number to three or higher dimension.

The main technical step in the construction is given in the following theorem.

Theorem 4.1 *Let P be a set of n points in \mathbb{R}^3 , and let $0 < \varepsilon < 1$, $0 < p < 1$ be given parameters. Then there exists a set $A \subset P$, of size $O\left(\frac{1}{\varepsilon^{3/2} p} \log \frac{1}{\varepsilon p}\right)$, such that, for any halfspace h of weight at most pn , we have*

$$\left| \frac{|h \cap A|}{|A|} - \frac{|h \cap P|}{|P|} \right| \leq \varepsilon p. \quad (4)$$

Remark: Let us note right away the difference between Eq. (4) and the situation in the preceding sections. That is, up to now we have handled ranges of size *at least* pn , whereas Eq. (4) applies to ranges of size *at most* pn . This issue requires a somewhat less standard construction, that will culminate in a *sequence* of approximation sets, each catering to a different range of halfspace weights. Nevertheless, the overall size of these sets will satisfy the above bound, and the cost of accessing them will be small. See below for details.

Proof: We apply the shallow partition theorem of Matoušek [Mat91b], with $r = 1/p$, to obtain a partition of P into $s \leq 2r$ subsets P_1, \dots, P_s , each of size at most $k = pn$, such that any k -shallow halfspace (namely, a halfspace that contains at most k points of P) separates at most $c \log r = c \log(1/p)$ subsets, for some absolute constant c . (Note that if h meets any P_i , it has to separate it, because h is too shallow to fully contain P_i .) Without loss of generality, we can carry out the construction so that the size of each P_i is even.

We then construct, for each subset P_i , a spanning tree of P_i with crossing number $O(k^{2/3})$ [Wel92], and convert it to a perfect matching of P_i , with the same asymptotic bound on its crossing number, namely, the maximum number of pairs in the matching that a halfspace separates.

We then color each matched pair independently, as above, coloring at random one of its points by -1 and the other by 1 , with equal probabilities. Let R_1 be the set of points colored -1 ; we have $|R_1| = n/2$. With high probability, the discrepancy of any halfspace h is at most $\sqrt{6\xi(h) \ln(2n)}$, where $\xi(h)$ is the crossing number of h . Since h is assumed to be k -shallow, it follows by construction

that $\xi(h) = O\left(\sum_i k_i^{2/3}\right)$, where $k_i = |h \cap P_i|$, and where the sum extends over those $O(\log r)$ subsets for which $k_i > 0$. Using Hölder's inequality, this yields $\xi(h) = O(k^{2/3} \log^{1/3} r)$, so the discrepancy of h is $O(k^{1/3} \log^{2/3} n)$.

We continue recursively in this manner for j steps, producing a sequence of subsets $R_0 = P, R_1, \dots, R_j$, where R_i is obtained from R_{i-1} using the above coloring procedure. When constructing R_i from R_{i-1} , for $i \geq 2$, we use the parameter $k_{i-1} = k \min\{c/2^{i-1}, 1\}$, where c is the constant derived in the following lemma, which is a variant of Lemma 2.2.

Lemma 4.2 *For any halfspace h with at most $k = pn$ points of P , we have, for any $i \leq j$,*

$$|h \cap R_i| \leq \frac{ck}{2^i} = \frac{c pn}{2^i},$$

for an appropriate absolute constant c , provided that $n_j \geq \frac{2}{p} \ln \frac{1}{p}$.

We thus have

$$\left| \frac{|h \cap R_{j-1}|}{|R_{j-1}|} - \frac{|h \cap R_j|}{|R_j|} \right| = \frac{\chi(h, R_{j-1})}{|R_{j-1}|} = O\left(\frac{2^{j-1} k_{j-1}^{1/3} \log^{2/3}(n/2^{j-1})}{n}\right),$$

for $j = 0, \dots$. Substituting $k_i = k \min\{c/2^i, 1\}$, for each i , and adding up the inequalities, it is easily checked that the last right-hand side dominates, so we obtain

$$\left| \frac{|h \cap P|}{|P|} - \frac{|h \cap R_j|}{|R_j|} \right| = O\left(\frac{2^{2j/3} k^{1/3} \log^{2/3}(n/2^{j-1})}{n}\right).$$

We choose j so that this bound is at most $\varepsilon p = \varepsilon k/n$. That is, $2^j = O\left(\frac{\varepsilon^{3/2} pn}{\log(n/2^j)}\right)$. Hence, the size of R_j is

$$|R_j| = \frac{n}{2^j} = O\left(\frac{\log(n/2^j)}{\varepsilon^{3/2} p}\right) = O\left(\frac{1}{\varepsilon^{3/2} p} \log \frac{1}{\varepsilon p}\right).$$

Taking $A = R_j$ completes the proof. ■

Theorem 4.3 *Given a set P of n points in \mathbb{R}^3 , and two parameters $0 < \varepsilon < 1$, $0 < p < 1$, we can construct $O\left(\log \frac{1}{p}\right)$ subsets of P , A_1, \dots, A_k , of total size $O\left(\frac{1}{\varepsilon^{3/2} p} \log \frac{1}{\varepsilon p}\right)$, so that, given any halfspace h containing qn points of P , for $q \geq p$, we can find a set A_t that satisfies*

$$\left| \frac{|h \cap A_t|}{|A_t|} - \frac{|h \cap P|}{|P|} \right| \leq \varepsilon \frac{|h \cap P|}{|P|}.$$

The (brute-force) time it takes to search for A_t and obtain the count $|h \cap A_t|$ is $O\left(\frac{1}{\varepsilon^{3/2} q} \log \frac{1}{\varepsilon q}\right)$.

See the full version [HS06] for details concerning the searching procedure.

4.2 Higher dimensions

The preceding construction can be generalized to higher dimensions, with some complications. We first introduce the following parameters:

$$\gamma = 1 + \frac{1 - \frac{1}{d^*}}{d + 1}, \quad \text{where } d^* = \lfloor d/2 \rfloor, \quad \text{and } \mu = \frac{2d}{d + 1}.$$

Note that, for $d \geq 4$, $\gamma > 1$ (and tends to 1 as d increases), and $\mu < 2$ (and tends to 2 as d increases).

The analogous version of Theorem 4.1 is

Theorem 4.4 *Let P be a set of n points in \mathbb{R}^d , $d \geq 4$, and let $0 < \varepsilon < 1$, $0 < p < 1$, be as above. Then there exists a set $A \subset P$, of size $O\left(\frac{d^{\mu/2}}{\varepsilon^\mu p^\gamma} \log \frac{d}{\varepsilon p}\right)$, such that, for any halfspace h that is (at most) pn -shallow, we have $\left|\frac{|h \cap A|}{|A|} - \frac{|h \cap P|}{|P|}\right| \leq \varepsilon p$, provided that $n = \Omega\left(\frac{d^{\mu/2}}{p^\gamma} \log^{\mu/2} \frac{d}{p}\right)$.*

Proof: Omitted; see the full version [HS06]. ■

We thus have the following result.

Theorem 4.5 *Given a set P of n points in \mathbb{R}^d , and two parameters $0 < \varepsilon < 1$, $0 < p < 1$, we can construct $O\left(\log \frac{1}{p}\right)$ subsets of P , A_1, \dots, A_k , of total size $O\left(\frac{d^{\mu/2}}{\varepsilon^\gamma p^\mu} \log^{\mu/2} \frac{1}{\varepsilon p}\right)$, so that, given any halfspace h containing qn points of P , for $q \geq p$, we can find a set A_t that satisfies $\left|\frac{|h \cap A_t|}{|A_t|} - \frac{|h \cap P|}{|P|}\right| \leq \varepsilon \frac{|h \cap P|}{|P|}$. The (brute-force) time it takes to search for A_t and obtain the count $|h \cap A_t|$ is $O\left(\frac{d^{\mu/2}}{\varepsilon^\gamma q^\mu} \log^{\mu/2} \frac{1}{\varepsilon q}\right)$.*

5 Conclusions

In this paper, we showed how to use low-discrepancy halving to construct small-size relative (p, ε) -approximations. We then gave a construction of even smaller-size approximations for halfplane ranges, by revisiting the classical construction of spanning trees with low crossing number, and modifying it to be weight-sensitive. We then gave similar constructions of relative approximation sets for halfspace ranges in higher dimensions, using a different approach, and showed how to use them for approximate halfspace range counting. We also revisited (this is presented only in the full version) the approximate range-counting problem in two and three dimensions and provided better algorithms than those previously known.

There are several interesting open problems for further research. The main one is to extend the construction of spanning trees with small relative crossing number to three and higher dimensions. Another open problem is to improve Theorem 3.6. A minor further improvement of Theorem 3.6 is possible by plugging the construction of Theorem 3.6 into the construction of Lemma 3.5. This still falls short of the desired spanning tree with crossing number $O(\sqrt{w_\ell})$, for a line ℓ of weight w_ℓ . We leave this as an open problem for further research.

Interestingly, the partition of Lemma 3.4 can be interpreted as a strengthening of the shallow partition theorem of Matoušek [Mat91b] in two dimensions. It is quite possible that a similar (but probably weaker) strengthening is possible in three and higher dimensions.

References

- [AdBMS98] P. K. Agarwal, M. de Berg, J. Matoušek, and O. Schwarzkopf, Constructing levels in arrangements and higher order Voronoi diagrams, *SIAM J. Comput.* 27 (1998), 654–667.
- [AH05] B. Aronov and S. Har-Peled, On approximating the depth and related problems, *Proc. 16th Annu. ACM-SIAM Sympos. Discrete Algo.*, 2005, 886–894.

- [AH06] B. Aronov and S. Har-Peled, On approximating the depth and related problems, manuscript, 2006. Full version of [AH05], available from <http://www.uiuc.edu/~sariel/papers/04/depth>.
- [AS06] B. Aronov and M. Sharir, Approximate halfspace range counting, in preparation.
- [CW89] B. Chazelle and E. Welzl, Quasi-optimal range searching in spaces with finite VC dimension, *Discrete Comput. Geom.* 4 (1989), 467–490.
- [Cha01] B. Chazelle, *The Discrepancy Method: Randomness and Complexity*, Cambridge University Press, New York, 2001.
- [Cha04] B. Chazelle, The discrepancy method in computational geometry, chapter 44, in *Handbook of Discrete and Computational Geometry*, 2nd Edition, J.E. Goodman and J. O’Rourke, Eds., CRC Press, Boca Raton, 2004, 983–996.
- [Coh97] E. Cohen, Size-estimation framework with applications to transitive closure and reachability, *J. Comput. Syst. Sci.* 55 (1997), 441–453.
- [CK04] E. Cohen and H. Kaplan, Spatially-decaying aggregation over a network: model and algorithms, *SIGMOD ’04: Proc. 2004 ACM SIGMOD Internat. Conf. on Management of Data*, 2004, 707–718.
- [CKMS06] E. Cohen, H. Kaplan, Y. Mansour and M. Sharir, manuscript, 2006.
- [Har07] S. Har-Peled, How to get close to the median shape, *Comput. Geom. Theory Appl.* 36 (2007), 39–51.
- [Hau92] D. Haussler, Decision theoretic generalizations of the PAC model for neural nets and other learning applications, *Inf. Comput.* 100 (1992), 78–150.
- [HS06] S. Har-Peled and M. Sharir, Relative ε -approximations in geometry, Manuscript, 2006. Available from <http://www.uiuc.edu/~sariel/papers/06/relative>.
- [HW87] D. Haussler and E. Welzl, Epsilon nets and simplex range queries, *Discrete Comput. Geom.* 2 (1987), 127–151.
- [KRS06] H. Kaplan, E. Ramos and M. Sharir, in preparation.
- [KS06] H. Kaplan and M. Sharir, Randomized incremental construction of three-dimensional convex hulls and planar Voronoi diagrams, and approximate range counting, *Proc. 17th ACM-SIAM Sympos. Discrete Algorithms* (2006), 484–493.
- [LLS01] Y. Li, P. M. Long, and A. Srinivasan, Improved bounds on the sample complexity of learning, *J. Comput. Syst. Sci.* 62 (2001), 516–527.
- [Mat90] J. Matoušek, Construction of epsilon-nets, *Discrete Comput. Geom.* 5 (1990), 427–448.
- [Mat91a] J. Matoušek, Computing the center of planar point sets, in *Computational Geometry: Papers from the DIMACS Special Year* (J.E. Goodman, R. Pollack and W. Steiger, eds.), AMS, Providence, RI, 1991, pp. 221–230.
- [Mat91b] J. Matoušek, Reporting points in halfspaces, *Comput. Geom. Theory Appl.* 2 (1991), 169–186.

- [Mat92] J. Matoušek, Efficient partition trees, *Discrete Comput. Geom.* 8 (1992), 315–334.
- [Mat99] J. Matoušek, *Geometric Discrepancy*, Algorithms and Combinatorics, Vol. 18, Springer Verlag, Heidelberg, 1999.
- [Mat03] J. Matoušek, *Using the Borsuk-Ulam Theorem*, Universitext. Springer-Verlag, Berlin, 2003, Lectures on topological methods in combinatorics and geometry, Written in cooperation with Anders Björner and Günter M. Ziegler.
- [MWW93] J. Matoušek, E. Welzl and L. Wernisch, Discrepancy and approximations for bounded VC-dimension, *Combinatorica* 13 (1993), 455–466.
- [PA95] J. Pach and P.K. Agarwal, *Combinatorial Geometry*, Wiley Interscience, New York, 1995.
- [Pol86] D. Pollard, Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions, Manuscript, 1986.
- [SS05] H. Shaul and M. Sharir, Ray shooting amid balls, farthest point from a line, and range emptiness searching, *Proc. 16th ACM-SIAM Sympos. Discrete Algorithms* (2005), 525–534.
- [ST86] N. Sarnak and R. E. Tarjan, Planar point location using persistent search trees, *Commun. ACM* 29 (1986), 669–679.
- [VC71] V.N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and its Applications* 16 (1971), 264–280.
- [Wel92] E. Welzl, On spanning trees with low crossing numbers, In *Data Structures and Efficient Algorithms, Final Report on the DFG Special Joint Initiative*, volume 594 of *Lect. Notes in Comp. Sci.*, pages 233–249, Springer-Verlag, 1992.