Algorithms for ε-approximations of Terrains *

Jeff M. Phillips

Department of Computer Science, Duke University, Durham, NC 27708:
jeffp@cs.duke.edu

Abstract. Consider a point set $\mathcal{D}$ with a measure function $\mu : \mathcal{D} \to \mathbb{R}$. Let $\mathcal{A}$ be the set of subsets of $\mathcal{D}$ induced by containment in a shape from some geometric family (e.g. axis-aligned rectangles, half planes, balls, k-oriented polygons). We say a range space $(\mathcal{D}, \mathcal{A})$ has an ε-approximation P if

$$\max_{R \in \mathcal{A}} \left| \frac{\mu(R \cap P)}{\mu(P)} - \frac{\mu(R \cap \mathcal{D})}{\mu(\mathcal{D})} \right| \leq \varepsilon.$$

We describe algorithms for deterministically constructing discrete ε-approximations for continuous point sets such as distributions or terrains. Furthermore, for certain families of subsets $\mathcal{A}$, such as those described by axis-aligned rectangles, we reduce the size of the ε-approximations by almost a square root from $O(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon})$ to $O(\frac{1}{\varepsilon} \operatorname{polylog} \frac{1}{\varepsilon})$. This is often the first step in transforming a continuous problem into a discrete one for which combinatorial techniques can be applied. We describe applications of this result in geo-spatial analysis, biosurveillance, and sensor networks.

1 Introduction

Representing complex objects by point sets may require less storage and may make computation on them faster and easier. When properties of the point set approximate those of the original object, then problems over continuous or piecewise-linear domains are now simple combinatorial problems over point sets. For instance, when studying terrains, representing the volume by the cardinality of a discrete point set transforms calculating the difference between two terrains in a region to just counting the number of points in that region. Alternatively, if the data is already a discrete point set, approximating it with a much smaller point set has applications in selecting sentinel nodes in sensor networks. This paper studies algorithms for creating small samples with guarantees in the form of discrepancy and ε-approximations, in particular we construct ε-approximations of size $O(\frac{1}{\varepsilon} \operatorname{polylog} \frac{1}{\varepsilon})$.

* Work on this paper is supported by a James B. Duke Fellowship, by NSF under a Graduate Research Fellowship and grants CNS-05-40347, CFF-06-35000, and DEB-04-25465, by ARO grants W911NF-04-1-0278 and W911NF-07-1-0376, by an NIH grant 1P50-GM-08183-01, by a DOE grant OEGP200A070505, and by a grant from the U.S. Israel Binational Science Foundation.

ε-approximations. In this paper we study point sets, which we call domains and we label as $\mathcal{D}$, which are either finite sets or are Lebesgue-measureable sets. For a given domain $\mathcal{D}$ let $\mathcal{A}$ be a set of subsets of $\mathcal{D}$ induced by containment in some geometric shape (such as balls or axis-aligned rectangles). The pair $(\mathcal{D}, \mathcal{A})$ is called a *range space*. We say that P is an *ε-approximation* of $(\mathcal{D}, \mathcal{A})$ if

$$\max_{R \in \mathcal{A}} \left| \frac{|R \cap P|}{|P|} - \frac{|R \cap \mathcal{D}|}{|\mathcal{D}|} \right| \leq \varepsilon,$$

where $|\cdot|$ represents the cardinality of a discrete set or the Lebesgue measure for a Lebesgue-measurable set. $\mathcal{A}$ is said to *shatter* a discrete set $X \subseteq \mathcal{D}$ if each subset of X is equal to $R \cap X$ for some $R \in \mathcal{A}$. The cardinality of the largest discrete set X that $\mathcal{A}$ can shatter is known as the *VC-dimension*. A classic result of Vapnik and Chervonenkis [28] states that for any range space $(\mathcal{D}, \mathcal{A})$ with constant VC-dimension v there exists a subset $P \subset \mathcal{D}$ consisting of $O(\frac{v}{\varepsilon^2} \log \frac{v}{\varepsilon})$ points that is an ε-approximation for $(\mathcal{D}, \mathcal{A})$. Furthermore, if each element of P is drawn uniformly at random from $\mathcal{D}$ such that $|P| = O(\frac{v}{\varepsilon^2} \log \frac{v}{\varepsilon\delta})$, then P is an ε-approximation with probability at least $1 - \delta$. Thus, for a large class of range spaces random sampling produces an ε-approximation of size $O(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon})$.

Deterministic construction of ε-approximations. There exist deterministic constructions for ε-approximations. When $\mathcal{D}$ is the unit cube $[0,1]^d$ there are constructions which can be interpreted as ε-approximations of size $O(\frac{1}{\varepsilon^{2d/(d+1)}})$ for half spaces [16] and $O(\frac{1}{\varepsilon^{2d/(d+1)}} \log^{d/(d+1)} \frac{1}{\varepsilon} \text{polylog}(\log \frac{1}{\varepsilon}))$ for balls in d-dimensions [5]. Both have lower bounds of $\Omega(\frac{1}{\varepsilon^{2d/(d+1)}})$ [2]. See Matoušek [17] for more similar results or Chazelle's book [9] for applications. For a domain $\mathcal{D}$, let $\mathcal{R}_d$ describe the subsets induced by axis-parallel rectangles in d dimensions, and let $\mathcal{Q}_k$ describe the subsets induced by k-oriented polygons (or more generally polytopes) with faces described by k predefined normal directions. More precisely, for $\beta = \{\beta_1, \ldots, \beta_k\} \subset \mathbb{S}^{d-1}$, let $\mathcal{Q}_\beta$ describe the set of convex polytopes such that each face has an outward normal $\pm\beta_i$ for $\beta_i \in \beta$. If β is fixed, we will use $\mathcal{Q}_k$ to denote $\mathcal{Q}_\beta$ since it is the size k and not the actual set β that is important. When $\mathcal{D} = [0,1]^d$, then the range space $(\mathcal{D}, \mathcal{R}_d)$ has an ε-approximation of size $O(\frac{1}{\varepsilon} \log^{d-1} \frac{1}{\varepsilon} \text{polylog}(\log \frac{1}{\varepsilon}))$ [12]. Also, for all homothets (translations and uniform scalings) of any particular $Q \in \mathcal{Q}_k$, Skriganov constructs an ε-approximation of size $O(\frac{1}{\varepsilon} \log^{d-1} \frac{1}{\varepsilon} \text{polylog}(\log \frac{1}{\varepsilon}))$. When $\mathcal{D}$ is a discrete point set of size n, ε-approximations of size $O((\frac{1}{\varepsilon} \log \frac{1}{\varepsilon})^{2 - \frac{2}{v+1}})$ exist for bounded VC-dimension v [19], and can be constructed in time $O(n \cdot \frac{1}{\varepsilon^{2v}} \log^v \frac{1}{\varepsilon})$. In this spirit, for $\mathcal{R}_2$ and a discrete point set of size n, Suri, Toth, and Zhou [26] construct an ε-approximation of size $O(\frac{1}{\varepsilon} \log(\varepsilon n) \log^4(\frac{1}{\varepsilon} \log(\varepsilon n)))$ in the context of a streaming algorithm which can be analyzed to run in time $O(n(\frac{1}{\varepsilon} \log^4 \frac{1}{\varepsilon})^3)$.

Our results. We answer the question, "*for which ranges spaces can we construct ε-approximations of size $O(\frac{1}{\varepsilon}$ polylog$\frac{1}{\varepsilon})$?*" by describing how to deterministically construct an ε-approximation of size $O(\frac{1}{\varepsilon}$ polylog$\frac{1}{\varepsilon})$ for any domain which can

be decomposed into or approximated by a finite set of constant-size polytopes for families $\mathcal{R}_d$ and $\mathcal{Q}_k$. In particular:

- For a discrete point set $\mathcal{D}$ of cardinality n, we give an algorithm for generating an ε-approximation for $(\mathcal{D}, \mathcal{Q}_k)$ of size $O(\frac{1}{\varepsilon}\log^{2k}\frac{1}{\varepsilon}\text{polylog}(\log\frac{1}{\varepsilon}))$ in $O(n\frac{1}{\varepsilon^3}\text{ polylog}\frac{1}{\varepsilon})$ time. This requires a generalization of the iterative point set thinning algorithm by Chazelle and Matoušek [10] that does not rely on VC-dimension. This implies similar results for $\mathcal{R}_d$ as well.
- For any d-dimensional domain $\mathcal{D}$ that can be decomposed into n k'-oriented polytopes, we give an algorithm for generating an ε-approximation of size $O((k+k')\frac{1}{\varepsilon}\log^{2k}\frac{1}{\varepsilon}\text{polylog}(\log\frac{1}{\varepsilon}))$ for $(\mathcal{D}, \mathcal{Q}_k)$ in time $O((k+k')n\frac{1}{\varepsilon^4}\text{ polylog}\frac{1}{\varepsilon})$.

We are interested in terrain domains $\mathcal{D}$ defined to have a base B (which may, for instance, be a subset of $\mathbb{R}^2$) and a height function $h : B \to \mathbb{R}$. Any point (p, z) such that $p \in B$ and $0 \leq z \leq h(p)$ (or $0 \geq z \geq h(p)$ when $h(p) < 0$) is in the domain $\mathcal{D}$ of the terrain.

- For a terrain domain $\mathcal{D}$ where B and h are piecewise-linear with n linear pieces, our result implies that there exists an ε-approximation of size $O(k\frac{1}{\varepsilon}\log^4\frac{1}{\varepsilon}\text{polylog}(\log\frac{1}{\varepsilon}))$ for $(\mathcal{D}, \mathcal{Q}_k)$, and it can be constructed in $O(n \cdot \frac{1}{\varepsilon^4}\text{ polylog}\frac{1}{\varepsilon})$ time.
- For a terrain domain $\mathcal{D}$ where $B \subset \mathbb{R}^2$ is a rectangle with diameter d and h is smooth (C^2-continuous) with minimum height z^- and largest eigenvalue of its Hessian λ, we give an algorithm for creating an ε-approximation for $(\mathcal{D}, \mathcal{R}_2 \times \mathbb{R})$ of size $O(\frac{1}{\varepsilon}\log^4\frac{1}{\varepsilon}\text{polylog}(\log\frac{1}{\varepsilon}))$ in time $O(\frac{\lambda d^2}{z^-}\frac{1}{\varepsilon^5}\text{ polylog}\frac{1}{\varepsilon})$.

These results improve the running time for a spatial anomaly detection problem in biosurveillance [1], and can more efficiently place or choose sentinel nodes in a sensor network, addressing an open problem [21].

Roadmap. We introduce a variety of new techniques, rooted in discrepancy theory, to create ε-approximations of size $O(\frac{1}{\varepsilon}\text{ polylog}\frac{1}{\varepsilon})$ for increasingly difficult domains. First, Section 2 discusses Lebesgue and combinatorial discrepancy. Section 3 generalizes and improves a classic technique to create an ε-approximation for a discrete point set. Section 4 describes how to generate an ε-approximation for a polygonal domain. When a domain can be decomposed into a finite, disjoint set of polygons, then each can be given an ε-approximation and the union of all these point sets can be given a smaller ε-approximation using the techniques in Section 3. Section 5 then handles domains of continuous, non-polygonal point sets by first approximating them by a disjoint set of polygons and then using the earlier described techniques. Section 6 shows some applications of these results.

2 Lebesgue and Combinatorial Discrepancy

Lebesgue discrepancy. The Lebesgue discrepancy is defined for an n-point set $P \subset [0,1]^d$ relative to the volume of a unit cube $[0,1]^d$. [1] Given a range space

[1] Although not common in the literature, this definition can replace $[0,1]^d$ with an hyper-rectangle $[0, w_1] \times [0, w_2] \times \ldots \times [0, w_d]$.

$([0,1]^d, \mathcal{A})$ and a point set P, the *Lebesgue* discrepancy is defined

$$D(P, \mathcal{A}) = \sup_{R \in \mathcal{A}} |D(P, R)|, \quad \text{where} \quad D(P, R) = n \cdot |R \cap [0,1]^d| - |R \cap P|.$$

Optimized over all n-point sets, define the *Lebesgue discrepancy of* $([0,1]^d, \mathcal{A})$ as

$$D(n, \mathcal{A}) = \inf_{P \subset [0,1]^d, |P|=n} D(P, \mathcal{A}).$$

The study of Lebesgue discrepancy arguably began with the Van der Corput set C_n [27], which satisfies $D(C_n, \mathcal{R}_2) = O(\log n)$. This was generalized to higher dimensions by Hammersley [13] and Halton [12] so that $D(C_n, \mathcal{R}_d) = O(\log^{d-1} n)$. However, it was shown that many lattices also provide $O(\log n)$ discrepancy in the plane [17]. This is generalized to $O(\log^{d-1} n \log^{1+\tau} \log n)$ for $\tau > 0$ over $\mathcal{R}^d$ [22, 23, 6]. For a more in-depth history of the progression of these results we refer to the notes in Matoušek's book [17]. For application of these results in numerical integration see Niederreiter's book [20]. The results on lattices extend to homothets of any $Q_k \in \mathcal{Q}_k$ for $O(\log n)$ discrepancy in the plane [22] and $O(\log^{d-1} n \log^{1+\tau} \log n)$ discrepancy, for $\tau > 0$, in $\mathbb{R}^d$ [24], for some constant k. A wider set of geometric families which include half planes, right triangles, rectangles under all rotations, circles, and predefined convex shapes produce $\Omega(n^{1/4})$ discrepancy and are not as interesting from our perspective.

Lebesgue discrepancy describes an ε-approximation of $([0,1]^d, \mathcal{A})$, where $\varepsilon = f(n) = D(n, \mathcal{A})/n$. Thus we can construct an ε-approximation for $([0,1]^d, \mathcal{A})$ of size $g_{\mathcal{D}}(\varepsilon, \mathcal{A})$ as defined below. (Solve for n in $\varepsilon = D(n, \mathcal{A})/n$.)

$$g_D(\varepsilon, \mathcal{A}) = \begin{cases} O(\frac{1}{\varepsilon} \log^{\tau} \frac{1}{\varepsilon} \operatorname{polylog}(\log \frac{1}{\varepsilon})) & \text{for } D(n, \mathcal{A}) = O(\log^{\tau} n) \\ O((1/\varepsilon)^{1/(1-\tau)}) & \text{for } D(n, \mathcal{A}) = O(n^{\tau}) \end{cases} \tag{1}$$

Combinatorial discrepancy. Given a range space $(X, \mathcal{A})$ where X is a finite point set and a coloring function $\chi : X \to \{-1, +1\}$ we say the *combinatorial discrepancy* of $(X, \mathcal{A})$ colored by χ is

$$\operatorname{disc}_\chi(X, \mathcal{A}) = \max_{R \in \mathcal{A}} \operatorname{disc}_\chi(X \cap R) \quad \text{where}$$

$$\operatorname{disc}_\chi(X) = \sum_{x \in X} \chi(x) = |\{x \in X \ : \ \chi(x) = +1\}| - |\{x \in X \ : \ \chi(x) = -1\}|.$$

Taking this over all colorings and all point sets of size n we say

$$\operatorname{disc}(n, \mathcal{A}) = \max_{\{X : |X|=n\}} \min_{\chi : X \to \{-1,+1\}} \operatorname{disc}_\chi(X, \mathcal{A}).$$

Results about combinatorial discrepancy are usually proved using the partial coloring method [4] or the Beck-Fiala theorem [8]. The partial coloring method usually yields lower discrepancy by some logarithmic factors, but is nonconstructive. Alternatively, the Beck-Fiala theorem actually constructs a low discrepancy coloring, but with a slightly weaker bound. The Beck-Fiala

theorem states that for a family of ranges $\mathcal{A}$ and a point set X such that $\max_{x \in X} |\{A \in \mathcal{A} \ : \ x \in A\}| \leq t$, $\text{disc}(X, \mathcal{A}) \leq 2t - 1$. So the discrepancy is only a constant factor larger than the largest number of sets any point is in.

Srinivasan [25] shows that $\text{disc}(n, \mathcal{R}_2) = O(\log^{2.5} n)$, using the partial coloring method. An earlier result of Beck [3] showed $\text{disc}(n, \mathcal{R}_2) = O(\log^4 n)$ using the Beck-Fiala theorem [8]. The construction in this approach reduces to $O(n)$ Gaussian eliminations on a matrix of constraints that is $O(n) \times O(n)$. Each Gaussian elimination step requires $O(n^3)$ time. Thus the coloring χ in the construction for $\text{disc}(n, \mathcal{R}_2) = O(\log^4 n)$ can be found in $O(n^4)$ time.We now generalize this result.

Lemma 1. $\text{disc}(n, \mathcal{Q}_k) = O(\log^{2k} n)$ *for points in* $\mathbb{R}^d$ *and the coloring that generates this discrepancy can be constructed in* $O(n^4)$ *time, for* k *constant.*

The proof combines techniques from Beck [3] and Matoušek [18].

Proof. Given a class $\mathcal{Q}_k$, each potential face is defined by a normal vector from $\{\beta_1, \ldots, \beta_k\}$. For $j \in [1, k]$ project all points along β_j. Let a *canonical interval* be of the form $\left[\frac{t}{2^q}, \frac{t+1}{2^q}\right)$ for integers $q \in [1, \log n]$ and $t \in [0, 2^q)$. For each direction β_j choose a value $q \in [1, \log n]$ creating 2^q canonical intervals induced by the ordering along β_j. Let the intersection of any k of these canonical intervals along a fixed β_j be a *canonical subset*. Since there are $\log n$ choices for the values of q for each of the k directions, it follows that each point is in at most $(\log n)^k$ canonical subsets. Using the Beck-Fiala theorem, we can create a coloring for X so that no canonical subset has discrepancy more than $O(\log^k n)$.

Each range $R \in \mathcal{Q}_k$ is formed by at most $O(\log^k n)$ canonical subsets. For each ordering by β_i, the interval in this ordering induced by R can be described by $O(\log n)$ canonical intervals. Thus the entire range R can be decomposed into $O(\log^k n)$ canonical subsets, each with at most $O(\log^k n)$ discrepancy.

Applying the Beck-Fiala construction of size n, this coloring requires $O(n^4)$ time to construct.

Corollary 1. $\mathit{disc}(n, \mathcal{R}_d) = O(\log^{2d} n)$ *and the coloring that generates this discrepancy can be constructed in* $O(n^4)$ *time, for* d *constant.*

A better nonconstructive bound exists due to Matoušek [18], using the partial coloring method. For polygons in $\mathbb{R}^2$ $\text{disc}(n, Q_k) = O(k \log^{2.5} n \sqrt{\log(k + \log n)})$, and for polytopes in $\mathbb{R}^d$ $\text{disc}(n, Q_k) = O(k^{1.5\lfloor d/2 \rfloor} \log^{d+1/2} n \sqrt{\log(k + \log n)})$. For more results on discrepancy see Beck and Chen's book [7].

Similar to Lebesgue discrepancy, the set $P = \{p \in X \mid \chi(p) = +1\}$ generated from the coloring χ for combinatorial discrepancy $\text{disc}(n, \mathcal{A})$ describes an ε-approximation of $(X, \mathcal{A})$ where $\varepsilon = f(n) = \text{disc}(n, \mathcal{A})/n$. Thus, given this value of ε, we can say that P is an ε-approximation for $(X, \mathcal{A})$ of size

$$g(\varepsilon, \mathcal{A}) = \begin{cases} O(\frac{1}{\varepsilon} \log^{\tau} \frac{1}{\varepsilon} \ \text{polylog}(\log \frac{1}{\varepsilon})) & \text{for } \text{disc}(n, \mathcal{A}) = O(\log^{\tau} n) \\ O((1/\varepsilon)^{1/(1-\tau)}) & \text{for } \text{disc}(n, \mathcal{A}) = O(n^{\tau}). \end{cases} \tag{2}$$

The next section will describe how to iteratively apply this process efficiently to achieve these bounds for any value of ε.

3 Deterministic Construction of ε-approximations for Discrete Point Sets

We generalize the framework of Chazelle and Matoušek [10] describing an algorithm for creating an ε-approximation of a range space $(X, \mathcal{A})$. Consider any range space $(X, \mathcal{A})$, with $|X| = n$, for which there is an algorithm to generate a coloring χ that yields the combinatorial discrepancy $\text{disc}_\chi(X, \mathcal{A})$ and can be constructed in time $O(n^w \cdot l(n))$ where $l(n) = o(n)$. For simplicity, we refer to the combinatorial discrepancy we can construct $\text{disc}_\chi(X, \mathcal{A})$ as $\text{disc}(n, \mathcal{A})$ to emphasize the size of the domain, and we use equation (2) to describe $g(\varepsilon, \mathcal{A})$, the size of the ε-approximation it corresponds to. The values $\text{disc}(n, \mathcal{A})$, w, and $l(n)$ are dependent on the family $\mathcal{A}$ (e.g. see Lemma 1), but not necessarily its VC-dimension as in [10]. As used above, let $f(n) = \text{disc}(n, \mathcal{A})/n$ be the value of ε in the ε-approximation generated by a single coloring of a set of size n — the relative error. We require that, $f(2n) \leq (1 - \delta)f(n)$, for constant $0 < \delta \leq 1$; thus it is a geometrically decreasing function.

The algorithm will compress a set X of size n to a set P of size $O(g(\varepsilon, \mathcal{A}))$ such that P is an ε-approximation of $(X, \mathcal{A})$ by recursively creating a low discrepancy coloring. We note that an ε-approximation of an ε'-approximation is an $(\varepsilon + \varepsilon')$-approximation of the original set.

We start by dividing X into sets of size $O(g(\varepsilon, \mathcal{A}))$,[2] here ε is a parameter. The algorithm proceeds in two stages. The first stage alternates between merging pairs of sets and halving sets by discarding points colored $\chi(p) = -1$ by the combinatorial discrepancy method described above. The exception is after every $w + 2$ halving steps, we then skip one halving step. The second stage takes the one remaining set and repeatedly halves it until the error $f(|P|)$ incurred in the remaining set P exceeds $\frac{\varepsilon}{2+2\delta}$. This results in a set of size $O(g(\varepsilon, \mathcal{A}))$.

Algorithm 3.1 Creates an ε-approximation for $(X, \mathcal{A})$ of size $O(g(\varepsilon, \mathcal{A}))$.

1: Divide X into sets $\{X_0, X_1, X_2, \ldots\}$ each of size $4(w + 2)g(\varepsilon, \mathcal{A})$. [2]
2: **repeat** {*Stage 1*}
3: **for** $w + 2$ steps **do** {or stop if only one set is left}
4: MERGE: Pair sets arbitrarily (i.e. X_i and X_j) and merge them into a single set (i.e. $X_i := X_i \cup X_j$).
5: HALVE: Halve each set X_i using the coloring χ from $\text{disc}(X_i, \mathcal{A})$ (i.e. $X_i = \{x \in X_i \mid \chi(x) = +1\}$).
6: MERGE: Pair sets arbitrarily and merge each pair into a single set.
7: **until** only one set, P, is left
8: **repeat** {*Stage 2*}
9: HALVE: Halve P using the coloring χ from $\text{disc}(P, \mathcal{A})$.
10: **until** $f(|P|) \geq \varepsilon/(2 + 2\delta)$

[2] If the sets do not divide equally, artificially increase the size of the sets when necessary. These points can be removed later.

Theorem 1. *For a finite range space $(X, \mathcal{A})$ with $|X| = n$ and an algorithm to construct a coloring $\chi : X \to \{-1, +1\}$ such that*

- *the set $\{x \in X \ : \ \chi(x) = +1\}$ is an α-approximation of $(X, \mathcal{A})$ of size $g(\alpha, \mathcal{A})$ with $\alpha = disc_\chi(X, \mathcal{A})/n$ (see equation (2)).*
- *χ can be constructed in $O(n^w \cdot l(n))$ time where $l(n) = o(n)$.*

then Algorithm 3.1 constructs an ε-approximation for $(X, \mathcal{A})$ of size $O(g(\varepsilon, \mathcal{A}))$ in time $O(w^{w-1} n \cdot g(\varepsilon, \mathcal{A})^{w-1} \cdot l(g(\varepsilon, \mathcal{A})) + g(\varepsilon, \mathcal{A}))$.

Proof. Let $2^j = 4(w+2)g(\varepsilon, \mathcal{A})$, for an integer j, be the size of each set in the initial dividing stage (adjusting by a constant if $\delta \leq \frac{1}{4}$). Each round of Stage 1 performs $w + 3$ MERGE steps and $w + 2$ HALVE steps on sets of the same size and each subsequent round deals with sets twice as large. The union of all the sets is an α-approximation of $(X, \mathcal{A})$ (to start $\alpha = 0$) and α only increases in the HALVE steps. The ith round increases α by $f(2^{j-1+i})$ per HALVE step. Since $f(n)$ decrease geometrically as n increases, the size of α at the end of the first stage is asymptotically bounded by the increase in the first round. Hence, after Stage 1 $\alpha \leq 2(w+2)f(4(w+2)g(\varepsilon, \mathcal{A})) \leq \frac{\varepsilon}{2}$. Stage 2 culminates the step before $f(|P|) \geq \frac{\varepsilon}{2+2\delta}$. Thus the final HALVE step creates an $\frac{\varepsilon\delta}{2+2\delta}$-approximation and the entire second stage creates an $\frac{\varepsilon}{2}$-approximation, hence overall Algorithm 3.1 creates an ε-approximation. The relative error caused by each HALVE step in stage 2 is equivalent to a HALVE step in a single round of stage 1.

The running time is also dominated by Stage 1. Each HALVE step of a set of size 2^j takes $O((2^j)^w l(2^j))$ time and runs on $n/2^j$ sets. In between each HALVE step within a round, the number of sets is divided by two, so the running time is asymptotically dominated by the first HALVE step of each round. The next round has sets of size 2^{j+1}, but only $n/2^{j+w+2}$ of them, so the runtime is at most $\frac{1}{2}$ that of the first HALVE step. Thus the running time of a round is less than half of that of the previous one. Since $2^j = O(wg(\varepsilon, \mathcal{A}))$ the running time of the HALVE step, and hence the first stage is bounded by $O(n \cdot (w \cdot g(\varepsilon, \mathcal{A}))^{w-1} \cdot l(g(\varepsilon, \mathcal{A})) + g(\varepsilon, \mathcal{A}))$. Each HALVE step in the second stage corresponds to a single HALVE step per round in the first stage, and does not affect the asymptotics.

We can invoke Theorem 1 along with Lemma 1 and Corollary 1 to compute χ in $O(n^4)$ time (notice that $w = 4$ and $l(\cdot)$ is constant), so $g(\varepsilon, \Omega_k) = O(\frac{1}{\varepsilon} \log^{2k} \frac{1}{\varepsilon} \text{polylog}(\log \frac{1}{\varepsilon}))$ and $g(\varepsilon, \mathcal{R}_d) = O(\frac{1}{\varepsilon} \log^{2d} \frac{1}{\varepsilon} \text{polylog}(\log \frac{1}{\varepsilon}))$. We obtain the following important corollaries.

Corollary 2. *For a set of size n and over the ranges Ω_k an ε-approximation of size $O(\frac{1}{\varepsilon} \log^{2k} \frac{1}{\varepsilon} \text{polylog}(\log \frac{1}{\varepsilon}))$ can be constructed in time $O(n \frac{1}{\varepsilon^3} \text{ polylog} \frac{1}{\varepsilon})$.*

Corollary 3. *For a set of size n and over the ranges $\mathcal{R}_d$ an ε-approximation of size $O(\frac{1}{\varepsilon} \log^{2d} \frac{1}{\varepsilon} \text{polylog}(\log \frac{1}{\varepsilon}))$ can be constructed in time $O(n \frac{1}{\varepsilon^3} \text{ polylog} \frac{1}{\varepsilon})$.*

Weighted case. These results can be extended to the case where each point $x \in X$ is given a weight $\mu(x)$. Now an ε-approximation is a set $P \subset X$ and a weighting $\mu : X \to \mathbb{R}$ such that

$$\max_{R \in \mathcal{A}} \left| \frac{\mu(P \cap R)}{\mu(P)} - \frac{\mu(X \cap R)}{\mu(X)} \right| \leq \varepsilon,$$

where $\mu(P) = \sum_{p \in P} \mu(p)$. The weights on P may differ from those on X. A result from Matoušek [15], invoking the unweighted algorithm several times at a geometrically decreasing cost, creates a weighted ε-approximation of the same asymptotic size and with the same asymptotic runtime as for an unweighted algorithm. This extension is important when we combine ε-approximations representing regions of different total measure. For this case we weight each point relative to the measure it represents.

4 Sampling from Polygonal Domains

We will prove a general theorem for deterministically constructing small ε-approximations for polygonal domains which will have direct consequences on polygonal terrains. A key observation of Matoušek [15] is that the union of ε-approximations of disjoint domains forms an ε-approximation of the union of the domains. Thus for any geometric domain $\mathcal{D}$ we first divide it into pieces for which we can create ε-approximations. Then we merge all of these point sets into an ε-approximation for the entire domain. Finally, we use Theorem 1 to reduce the sample size.

Instead of restricting ourselves to domains which we can divide into cubes of the form $[0,1]^d$, thus allowing the use of Lebesgue discrepancy results, we first expand on a result about lattices and polygons.

Lattices and polygons. For $x \in \mathbb{R}$, let $\lfloor x \rfloor$ represent the fractional part of x, and for $\alpha \in \mathbb{R}^{d-1}$ let $\alpha = (\alpha_1, \ldots, \alpha_{d-1})$. Now given α and m let $P_{\alpha,m} = \{p_0, \ldots, p_{m-1}\}$ be a set of m lattice points in $[0,1]^d$ defined $p_i = (\frac{i}{m}, \lfloor \alpha_1 i \rfloor, \ldots, \lfloor \alpha_{d-1} i \rfloor)$. $P_{\alpha,m}$ is *irrational* with respect to any polytope in $\mathcal{Q}_\beta$ if for all $\beta_i \in \beta$, for all $j \leq d$, and for all $h \leq d-1$, the fraction $\beta_{i,j}/\alpha_h$ is irrational. (Note that $\beta_{i,j}$ represents the jth element of the vector β_i.) Lattices with α irrational (relative to the face normals) generate low discrepancy sets.

Theorem 2. *Let $Q \in \mathcal{Q}_{\beta'}$ be a fixed convex polytope. Let $\beta, \beta' \subset \mathbb{S}^{d-1}$ be sets of k and k' directions, respectively. There is an ε-approximation of $(Q, \mathcal{Q}_\beta)$ of size $O((k+k')\frac{1}{\varepsilon}\log^{d-1}\frac{1}{\varepsilon}\text{polylog}(\log\frac{1}{\varepsilon}))$.*

This ε-approximation is realized by a set of lattice points $P_{\alpha,m} \cap Q$ such that $P_{\alpha,m}$ is irrational with respect to any polytope in $\mathcal{Q}_{\beta \cup \beta'}$.

Proof. Consider polytope tQ_h and lattice $P_{\alpha,m}$, where the uniform scaling factor t is treated as an asymptotic quantity. Skriganov's Theorem 6.1 in [24] claims

$$\max_{v \in \mathbb{R}^d} D(P_{\alpha,m}, tQ_h + v) = O\left(t^{d-1}\rho^{-\theta} + \sum_f S_f(P_{\alpha,m}, \rho) \right)$$

where

$$S_f(P_{\alpha,m}, \rho) = O(\log^{d-1} \rho \log^{1+\tau} \log \rho)$$

for $\tau > 0$, as long as $P_{\alpha,m}$ is irrational with respect to the normal of the face f of Q_h and infinite otherwise, where $\theta \in (0,1)$ and ρ can be arbitrarily large. Note that this is a simplified form yielded by invoking Theorem 3.2 and Theorem 4.5 from [24]. By setting $\rho^\theta = t^{d-1}$,

$$\max_{v \in \mathbb{R}^d} D(P_{\alpha,m}, tQ_h + v) = O(h \log^{d-1} t \log^{1+\tau} \log t). \tag{3}$$

Now by noting that as t grows, the number of lattice points in tQ_h grows by a factor of t^d, and we can set $t = n^{1/d}$ so (3) implies that $D(P_{\alpha,m}, tQ_h) = O(h \log^{d-1} n \log^{1+\tau} \log n)$ for $|P_{\alpha,m}| = m = n$ and $tQ_h \subset [0,1]^d$.

The discrepancy is a sum over the set of h terms, one for each face f, each of which is small as long as $P_{\alpha,m}$ is irrational with respect to f's normal β_f. Hence this lattice gives low discrepancy for any polytope in the analogous family $\mathcal{Q}_\beta$ such that $P_{\alpha,m}$ is irrational with respect to $\mathcal{Q}_\beta$. Finally we realize that any subset $Q \cap Q_k$ for $Q \in \mathcal{Q}_{\beta'}$ and $Q_k \in \mathcal{Q}_\beta$ is a polytope defined by normals from $\beta' \cup \beta$ and we then refer to $g_D(\varepsilon, \mathcal{Q}_{\beta \cup \beta'})$ in (1) to bound the size of the ε-approximation from the given Lebesgue discrepancy.

Remark 1. Skriganov's result [24] is proved under the *whole space* model where the lattice is infinite (tQ_h is not confined to $[0,1]^d$), and the relevant error is the difference between the measure of tQ_h versus the cardinality $|tQ_h \cap P_{\alpha,m}|$, where each $p \in P_{\alpha,m}$ represents 1 unit of measure. Skriganov's main results in this model is summarized in equation (3) and only pertains to a fixed polytope Q_h instead of, more generally, a family of polytopes $\mathcal{Q}_\beta$, as shown in Theorem 2.

Samples for polygonal terrains. Combining the above results and weighted extension of Theorem 1 implies the following results.

Theorem 3. *We can create a weighted ε-approximation of size $O((k+k')\frac{1}{\varepsilon} \cdot \log^{2k} \frac{1}{\varepsilon} \text{polylog}(\log \frac{1}{\varepsilon}))$ of $(\mathcal{D}, \mathcal{Q}_k)$ in time $O((k+k')n\frac{1}{\varepsilon^4} \text{ polylog}\frac{1}{\varepsilon})$ for any d-dimensional domain $\mathcal{D}$ which can be decomposed into n d-dimensional convex k'-oriented polytopes.*

Proof. We divide the domain into n k'-oriented polytopes and then approximate each polytope $Q_{k'}$ with a point set $P_{\alpha,m} \cap Q_{k'}$ using Theorem 2. We observe that the union of these point sets is a weighted ε-approximation of $(\mathcal{D}, \mathcal{Q}_k)$, but is quite large. Using the weighted extension of Theorem 1 we can reduce the point sets to the size and in the time stated.

This has applications to terrain domains $\mathcal{D}$ defined with a piecewise-linear base B and height function $h : B \to \mathbb{R}$. We decompose the terrain so that each linear piece of h describes one 3-dimensional polytope, then apply Theorem 3 to get the following result.

Corollary 4. *For terrain domain $\mathcal{D}$ with piecewise-linear base B and height function $h : B \to \mathbb{R}$ with n linear pieces, we construct a weighted ε-approximation of $(\mathcal{D}, \mathcal{Q}_k)$ of size $O(k\frac{1}{\varepsilon} \log^4 \frac{1}{\varepsilon} \text{polylog}(\log \frac{1}{\varepsilon}))$ in time $O(kn\frac{1}{\varepsilon^4} \text{ polylog}\frac{1}{\varepsilon})$.*

5 Sampling from Smooth Terrains

We can create an ε-approximation for a smooth domain (one which cannot be decomposed into polytopes) in a three stage process. The first stage approximates any domain with a set of polytopes. The second approximates each polytope with a point set. The third merges all point sets and uses Theorem 1 to reduce their size.

This section mainly focuses on the first stage. More formally, we can approximate a non-polygonal domain $\mathcal{D}$ with a set of disjoint polygons P such that P has properties of an ε-approximation.

Lemma 2. *If* $|\mathcal{D} \setminus P| \leq \frac{\varepsilon}{2}|\mathcal{D}|$ *and* $P \subseteq \mathcal{D}$ *then* $\max_{R \in \mathcal{A}} \left| \frac{|R \cap P|}{|P|} - \frac{|R \cap \mathcal{D}|}{|\mathcal{D}|} \right| \leq \varepsilon.$

Proof. No range $R \in \mathcal{A}$ can have $\left| \frac{|R \cap P|}{|P|} - \frac{|R \cap \mathcal{D}|}{|\mathcal{D}|} \right| > \varepsilon$ because if $|\mathcal{D}| \geq |P|$ (w.l.o.g.), then $|R \cap \mathcal{D}| - \frac{|\mathcal{D}|}{|P|}|R \cap P| \leq \varepsilon|\mathcal{D}|$ and $|R \cap P|\frac{|\mathcal{D}|}{|P|} - |R \cap \mathcal{D}| \leq \varepsilon|\mathcal{D}|$. The first part follows from $\frac{|\mathcal{D}|}{|P|} \geq 1$ and is loose by a factor of 2. For the second part we can argue

$$|R \cap P|\frac{|\mathcal{D}|}{|P|} - |R \cap \mathcal{D}| \leq |R \cap P|\frac{1}{1-\frac{\varepsilon}{2}} - |R \cap \mathcal{D}| \leq |R \cap \mathcal{D}|\frac{1}{1-\frac{\varepsilon}{2}} - |R \cap \mathcal{D}|$$
$$= \frac{\frac{\varepsilon}{2}}{1-\frac{\varepsilon}{2}}|R \cap \mathcal{D}| \leq \varepsilon|R \cap \mathcal{D}| \leq \varepsilon|\mathcal{D}|.$$

For terrain domains $\mathcal{D}$ defined with a base B and a height function $h : B \to \mathbb{R}$, if B is polygonal we can decompose it into polygonal pieces, otherwise we can approximate it with constant-size polygonal pieces according to Lemma 2. Then, similarly, if h is polygonal we can approximate the components invoking Corollary 4; however, if it is smooth, then we can approximate each piece according to Lemma 2.

We can improve further upon this approach using a stretched version of the Van der Corput Set and dependent on specific properties of the terrain. Consider the case where B is a rectangle with diameter $d_{\mathcal{D}}$ and h is C^2 continuous with minimum value $z_{\mathcal{D}}^-$ and where the largest eigenvalue of its Hessian is $\lambda_{\mathcal{D}}$. For such a terrain $\mathcal{D}$, interesting ranges $\mathcal{R}_2 \times \mathbb{R}$ are generalized cylinders where the first 2 dimensions are an axis-parallel rectangle and the third dimension is unbounded. We can state the following result (proved in the full version).

Theorem 4. *For a domain* $\mathcal{D}$ *with rectangular base* $B \subset \mathbb{R}^2$ *and with a* C^2*-continuous height function* $h : B \to \mathbb{R}$ *we can deterministically create a weighted* ε*-approximation of* $(\mathcal{D}, \mathcal{R}_2 \times \mathbb{R})$ *of size* $O\left(\left(\frac{\lambda_{\mathcal{D}} d_{\mathcal{D}}^2}{z_{\mathcal{D}}^- \varepsilon}\right)\left(\frac{1}{\varepsilon}\log^4\frac{1}{\varepsilon}\text{polylog}(\log\frac{1}{\varepsilon})\right)\right)$*. We reduce the size to* $O(\frac{1}{\varepsilon}\log^4\frac{1}{\varepsilon}\text{polylog}(\log\frac{1}{\varepsilon}))$ *in time* $O\left(\left(\frac{\lambda_{\mathcal{D}} d_{\mathcal{D}}^2}{z_{\mathcal{D}}^-}\right)\frac{1}{\varepsilon^5}\text{ polylog}\frac{1}{\varepsilon}\right)$.

This generalizes in a straightforward way for $B \in \mathbb{R}^d$. Similar results are possible when B is not rectangular or when B is not even piecewise-linear. The techniques of Section 4 are necessary if Q_k is used instead of $\mathcal{R}_2$, and are slower by a factor $O(\frac{1}{\varepsilon})$.

6 Applications

Creating smaller ε-approximations improves several existing algorithms.

Biosurveillance. Let M and B be two points sets in $\mathbb{R}^2$. An important anomaly detection problem for biosurveillance [14, 1] reduces to finding a range (from some family of ranges such as $\mathcal{R}_2$) that maximizes a statistical discrepancy function on M and B, such as $d_P(m_R, b_R) = m_R \ln \frac{m_R}{b_R} + (1 - m_R) \ln \frac{1-m_R}{1-b_R}$, where $m_R = |R \cap M|/|M|$ and $b_R = |R \cap B|/|B|$. Using the results in this paper we can prove the following:

Theorem 5. *Let $|M \cup B| = n$. A range $R \in \mathcal{R}^2$ such that $|d_P(m_R, b_R) - \max_{r \in \mathcal{R}_2} d_P(m_r, b_r)| \leq \varepsilon$ can be deterministically found in $O(n\frac{1}{\varepsilon^3}\text{polylog}(\log \frac{1}{\varepsilon}) + \frac{1}{\varepsilon^4}\text{polylog}(\log \frac{1}{\varepsilon}))$ time.*

A range $R \in \mathcal{R}^2$ such that $|d_P(m_R, b_R) - \max_{r \in \mathcal{R}_2} d_P(m_r, b_r)| \leq \varepsilon + \delta$ can be deterministically found in $O(n\frac{1}{\varepsilon^3}\text{polylog}(\log \frac{1}{\varepsilon}) + \frac{1}{\delta}\frac{1}{\varepsilon^2}\text{polylog}(\log \frac{1}{\varepsilon}))$ time.

This can be generalized to when M and B are terrain domains. This case arises, for example, when each point is replaced with a probability distribution.

Sensor Networks. Let $\mathcal{D}$ be a set of points describing the location of sensors. If $P \subseteq \mathcal{D}$ is an *ε-sentinel* of $(\mathcal{D}, \mathcal{A})$, then for all $R \in \mathcal{A}$ (1) if $|R \cap \mathcal{D}| \geq \varepsilon|\mathcal{D}|$ then $|R \cap P| \geq \varepsilon\frac{3}{4}|P|$, and (2) if $|R \cap P| \geq \varepsilon\frac{3}{4}|P|$ then $|R \cap \mathcal{D}| \geq \frac{\varepsilon|\mathcal{D}|}{2}$. Previous work constructs ε-sentinels for half spaces [21] of size $O(\frac{1}{\varepsilon})$ and in expected time $O(\frac{n}{\varepsilon}\log n)$ or for any $\mathcal{A}$ with bounded VC-dimension v [11] of size $O(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$ and in time $O(n\frac{1}{\varepsilon^{2v}}\log^v\frac{1}{\varepsilon})$. Noting that an $\frac{\varepsilon}{4}$-approximation can be used as an ε-sentinel, we can state the following.

Theorem 6. *For a discrete point set $\mathcal{D}$ of size n, we can compute ε-sentinels for $(\mathcal{D}, \mathcal{Q}_k)$ of size $O(\frac{1}{\varepsilon}\log^{2k}\frac{1}{\varepsilon}\text{polylog}(\log\frac{1}{\varepsilon}))$ in time $O(n\frac{1}{\varepsilon^3}\textit{polylog}(\log\frac{1}{\varepsilon}))$.*

Furthermore, we can create $O(n\varepsilon/\log^{2k}\frac{1}{\varepsilon})$ disjoint sets of ε-sentinels in $O(n\frac{1}{\varepsilon^3}\log(n\varepsilon)\text{polylog}(\log\frac{1}{\varepsilon}))$ total time.

We can extend this result to place an ε-sentinel to cover a polygonal domain $\mathcal{D}$ as well. Details and further results are in the full version.

Acknowledgments. I would like to thank Pankaj Agarwal for many helpful discussions including finding a bug in an earlier version of the proof of Lemma 1, Shashidhara Ganjugunte, Hai Yu, Yuriy Mileyko, and Esther Ezra for a careful proofreading, Jirka Matoušek for useful pointers, Subhash Suri for posing a related problem, and Don Rose for discussions on improving the Beck-Fiala Theorem.

References

1. D. Agarwal, A. McGregor, J. M. Phillips, S. Venkatasubramanian, and Z. Zhu. Spatial scan statistics: Approximations and performance study. In *Proceedings 12th ACM SIGKDD Knowledge Discovery & Data Mining*, pages 24–33, 2006.

2. R. Alexander. Principles of a new method in the study of irregularities of distribution. *Inventiones Mathematicae*, 103:279–296, 1991.
3. J. Beck. Balanced two-coloring of finite sets in the square I. *Combinatorica*, 1:327–335, 1981.
4. J. Beck. Roth's estimate on the discrepancy of integer sequences is nearly sharp. *Combinatorica*, 1:319–325, 1981.
5. J. Beck. Irregularities of distribution I. *Acta Mathematics*, 159:1–49, 1987.
6. J. Beck. Probabilistic diophantine approximation, I Kronecker sequences. *Annals of Mathematics*, 140:451–502, 1994.
7. J. Beck and W. Chen. *Irregularities of Distribution*. Cambridge University Press, 1987.
8. J. Beck and T. Fiala. "integer-making" theorems. *Disc. App. Math.*, 3:1–8, 1981.
9. B. Chazelle. *The Discrepancy Method*. Cambridge University Press, 2000.
10. B. Chazelle and J. Matousek. On linear-time deterministic algorithms for optimization problems in fixed dimensions. *Journal of Algorithms*, 21:579–597, 1996.
11. S. Gandhi, S. Suri, and E. Welzl. Catching elephants with mice: Sparse sampling for monitoring sensor networks. In *Proceedings 5th Embedded Networked Sensor Systems*, pages 261–274, 2007.
12. J. H. Halton. On the efficiency of certain quasi-random sequences of points in evaluating multidimensional integrals. *Numerical Mathematics*, 2:84–90, 1960.
13. J. M. Hammersly. Monte Carlo methods for solving multivariable problems. *Annals of New York Acadamy of Science*, 86:844–874, 1960.
14. M. Kulldorff. A spatial scan statistic. *Comm. in Stat.: T&M*, 26:1481–1496, 1997.
15. J. Matoušek. Approximations and optimal geometric divide-and-conquer. In *Proceedings 23rd Symposium on Theory of Computing*, pages 505–511, 1991.
16. J. Matoušek. Tight upper bounds for the discrepancy of halfspaces. *Discrete and Computational Geometry*, 13:593–601, 1995.
17. J. Matoušek. *Geometric Discrepancy*. Springer, 1999.
18. J. Matoušek. On the discrepancy for boxes and polytopes. *Monatsh. Math.*, 127:325–336, 1999.
19. J. Matoušek, E. Welzl, and L. Wernisch. Discrepancy and approximations for bounded VC-dimension. *Combinatorica*, 13:455–466, 1993.
20. H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, 1992.
21. N. Shrivastava, S. Suri, and C. D. Tóth. Detecting cuts in sensor networks. *ACM Transactions on Sensor Networks*, 4(10), 2008.
22. M. Skriganov. Lattices in algebraic number fields and uniform distributions modulo 1. *Leningrad Mathematics Journal*, 1:535–558, 1990.
23. M. Skriganov. Constructions of uniform distributions in terms of geometry of numbers. *St. Petersburg Mathematics Journal*, 6:635–664, 1995.
24. M. Skriganov. Ergodic theory on $SL(n)$, diophantine approximations and anomalies in the lattice point problem. *Inventiones Mathematicae*, 132:1–72, 1998.
25. A. Srinivasan. Improving the discrepancy bound for sparse matrices: Better approximations for sparse lattice approximation problems. In *Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 692–701, 1997.
26. S. Suri, C. D. Tóth, and Y. Zhou. Range counting over multidimensional data streams. In *Proceedings 20th Symposium on Computational Geometry*, pages 160–169, 2004.
27. J. G. van der Corput. Verteilungsfunktionen I. *Aka. Wet. Ams.*, 38:813–821, 1935.
28. V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Prob. and its Applic.*, 16:264–280, 1971.