Geometric Computations on Indecisive Points*

Allan Jørgensen
MADALGO

Maarten Löffler[2]
University of California, Irvine

Jeff M. Phillips[3]
University of Utah

May 3, 2011

Abstract

We study computing with indecisive point sets. Such points have spatial uncertainty where the true location is one of a finite number of possible locations. This data arises from probing distributions a few times or when the location is one of a few locations from a known database. In particular, we study computing distributions of geometric functions such as the radius of the smallest enclosing ball and the diameter. Surprisingly, we can compute the distribution of the radius of the smallest enclosing ball exactly in polynomial time, but computing the same distribution for the diameter is #P-hard. We generalize our polynomial-time algorithm to all LP-type problems. We also utilize our indecisive framework to deterministically and approximately compute on a more general class of uncertain data where the location of each point is given by a probability distribution.

*The second author is funded by NWO under the GOGO project and the Office of Naval Research under grant N00014-08-1-1015; the third author is supported by a subaward to the University of Utah under NSF award 0937060 to CRA.

1 Introduction

We consider uncertain data point sets where each element of the set is not known exactly, but rather is represented by a finite set of candidate elements, possibly weighted, describing the finite set of possible true locations of the data point. The weight of a candidate location governs the probability that the data point is at that particular location. We call a point under this representation an *indecisive point*. Given indecisive input points we study computing full probability distributions (paramount for downstream analysis) over the value of some geometric query function, such as the radius of the smallest enclosing ball.

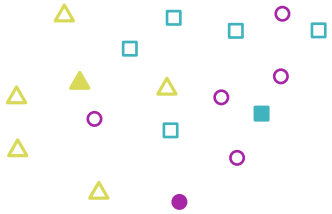


Figure 1: One of the 6^3 possible supports of points from $n = 3$ sets of $k = 6$ points each.

Indecisive points appear naturally in many applications. They play an important role in databases [9, 1, 8, 7, 21], machine learning [5], and sensor networks [25] where a limited number of probes from a certain data set are gathered, each potentially representing the true location of a data point. Alternatively, data points may be obtained using imprecise measurements or are the result of inexact earlier computations.

Consider the problem of tracking the spatial nature of a recently spreading disease. However, due to privacy considerations we do not know exactly who became sick, only their postal codes. But, we do have all coordinates and postal codes of people who live in the area under consideration. In this situation we are given a set of "points," but we do not know the location of each point, only that it is at one of a certain finite number of possible locations. How can we estimate, say, the size of the smallest circle enclosing the patients?

We can generalize the classification of indecisive points to when the true location of each data point is described by a probability distribution. We call these points *uncertain points*. In addition to indecisive points, this general class also includes for instance multivariate normal distributions and all points within a unit disk. More broadly, these data points could represent any (uncertain) geometric object, such as a hyperplane or disc; but since these objects can usually be dualized to points, our exposition will focus on points.

Related work on uncertain points. Many specific ways to model an uncertain point have been introduced throughout the past decades. One of the earliest models for uncertain points was a simple circular region [12], where each point has an infinite set of possible locations: all locations inside a given unit disk. This model has received considerable attention since [4, 15] and can be extended to more complicated regions [23]. Most work in these models focuses on computing the minimum or maximum possible values of a query. Another related model studies *outliers*, where each point is given one location and a probability that the location is correct [14, 2, 3].

The full model of uncertain points, where each point's location is represented by a probability distribution, has received much less attention in the algorithms community, mainly because its generality is difficult to handle and exact solutions seem impossible for all but the most simple questions. Löffler and Phillips [17] study simple randomized algorithms in this model.

There has also been a recent flurry of activity in the database community [9] on problems such as indexing [21], clustering [8], and histogram building [7]. However, the results with detailed algorithmic analysis generally focus on one-dimensional data; furthermore, they often only return the expected value or the most likely answer instead of calculating a full distribution.

1.1 Contribution

We study several geometric measures on sets of indecisive points in Section 2. We compute an exact distribution over the values that these measures can take, not just an expected or most-likely value. Surprisingly,

(a) (b)

Figure 2: (a) An example input consisting of $n = 3$ sets of $k = 6$ points each. (b) One of the 6^3 possible samples of $n = 3$ points.

while for some measures we present polynomial time algorithms (e.g. for the radius of the smallest enclosing ball), other seemingly very similar measures either are #P-Hard (e.g. the diameter) or have a solution size exponential in the number of indecisive input points (e.g. the area of the convex hull). In particular, we show that the family of problems which admit polynomial-time solutions includes *all* LP-type problems [20] with constant combinatorial dimension. #P-hardness results for indecisive data have been shown before [9], but the separation has not been understood as precisely, nor from a geometric perspective.

In Section 3 we extend the above polynomial-time algorithms to uncertain points, a much broader model for uncertainty. We describe detailed results for data points endowed with multivariate normal distributions representing their location. We deterministically reduce uncertain points to indecisive points by creating ε-samples (aka ε-approximations) of their distributions; however, this is not as straightforward as it may seem. It requires a structural theorem (Theorem 3.1) describing a special range space which can account for the dependence among the distributions, dependence that is created by the measure being evaluated. This is required even when the distributions themselves are independent because once an ε-sample has been fixed for one uncertain point, the other ε-samples need to account for the interaction of those fixed points with the measure. All together, these results build important structure required to transition between sets of continuous and discrete distributions with bounded error, and may be of independent interest.

These results provide a deterministic alternative to some of the results in [17]. The determinism in these algorithms is important when there can be no probability of failure, the answer needs to be identical each time it is computed, or when each indecisive point has a small constant number of possible locations.

2 Exact Computations on Indecisive Point Sets

In this section, we assume that we are given a set of n indecisive points, that is, a collection $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_n\}$ of n sets containing k points each, so $Q_i = \{q_{i1}, q_{i2}, \ldots, q_{ik}\}$. We say that a set P of n points is a *support* from $\mathcal{Q}$ if it contains exactly one point from each set Q_i, that is, if $P = \{p_1, p_2, \ldots, p_n\}$ with $p_i \in Q_i$. In this case we also write $P \Subset \mathcal{Q}$. Figure 2(b) shows an example of a set of indecisive points (identified by shading), highlighting a possible support.

When given a set of indecisive points, we assume that each indecisive point is at each of its locations with a fixed probability. Often these probabilities are equal, but for completeness, we describe our algorithms for when each point has a distinct probability of being the true position. For each $Q_i \in \mathcal{Q}$ and each $q_{ij} \in Q_i$, let $w(q_{ij})$ be a positive weight. We enforce $\sum_{j=1}^{k} w(q_{ij}) = k$, and $w(q_{ij})/k$ is the probability that q_{ij} is the true position of Q_i.

We let f denote the function we are interested in computing on $\mathcal{Q}$, that is, f takes a set of n points as input and computes a single real number as output. Since $\mathcal{Q}$ is indecisive, the value of f is not fixed, rather we are interested in the distribution of the possible values. We show that for some measures, we can compute this

distribution in polynomial time, while for others it is #P-hard or the solution itself has size exponential in n.

2.1 Polynomial Time Algorithms

We are interested in the distribution of the value $f(P)$ for each support $P \Subset \mathcal{Q}$. Since there are k^n possible supports, in general we cannot hope to do anything faster than that without making additional assumptions about f. In fact, we will show that some problems (e.g. area of convex hull) the distribution can require that many values, and for other problems (e.g. diameter) computing the distribution is $\#P$-hard, even though the distribution is guaranteed to be much smaller. Define $\tilde{f}(\mathcal{Q}, r)$ as the fraction (measured by weight) of supports of $\mathcal{Q}$ for which f gives a value smaller than or equal to r. We start with a simple example and then generalize. In this version, for simplicity, we assume general position and that k^n can be described by $O(1)$ words. (handled otherwise in Appendix A). First, we will let $f(P)$ denote the radius of the smallest enclosing disk of P in the plane, and show how to solve the decision problem in polynomial time in that case. We then show how to generalize the ideas to other classes of measures.

Smallest enclosing disk. Consider the problem where f measures the radius of the smallest enclosing disk of a support and let all weights be uniform so $w(q_{i,j}) = 1$ for all i and j.

Evaluating $\tilde{f}(\mathcal{Q}, r)$ in time polynomial in n and k is not completely trivial since there are k^n possible supports. However, we can make use of the fact that each smallest enclosing disk is in fact defined by a set of at most 3 points that lie on the boundary of the disk. For each support $P \Subset \mathcal{Q}$ we define $B_P \subseteq P$ to be this set of at most 3 points, which we call the *basis* for P. Bases have the property that $f(P) = f(B_P)$.

Now, to avoid having to test an exponential number of supports, we define a *potential basis* to be a set of at most 3 points in $\mathcal{Q}$ such that each point is from a different Q_i. Clearly, there are less than $(nk)^3$ possible potential bases, and each support $P \Subset \mathcal{Q}$ has one as its basis. Now, we only need to count for each potential basis the number of supports it represents. Counting the number of samples that have a certain basis is easy for the smallest enclosing circle. Given a basis B, we count for each indecisive point Q that does not contribute a point to B itself how many of its members lie inside the smallest enclosing circle of B, and then we multiply these numbers.

Now, for each potential basis B we have two values: the number of supports that have B as their basis, and the value $f(B)$. We can sort these $O((nk)^3)$ pairs on the value of f, and the result provides us with the required distribution. We spend $O(nk)$ time per potential basis for counting the points inside and $O(n)$ time for multiplying these values, so combined with $O((nk)^3)$ potential bases this gives $O((nk)^4)$ total time.

Theorem 2.1. *Let $\mathcal{Q}$ be a set of n sets of k points. In $O((nk)^4)$ time, we can compute a data structure of $O((nk)^3)$ size that can tell us in $O(\log(nk))$ time for any value r how many supports of $P \Subset \mathcal{Q}$ satisfy $f(P) \leq r$.*

LP-type problems. The approach described above also works for measures $f : \mathcal{Q} \to \mathbb{R}$ other than the smallest enclosing disk. In particular, it works for LP-type problems [20] that have constant combinatorial dimension. An *LP-type problem* provides a set of constraints H and a function $\omega : 2^H \to \mathbb{R}$ with the following two properties:

MONOTONICITY: For any $F \subseteq G \subseteq H$, $\omega(F) \leq \omega(G)$.

LOCALITY: For any $F \subseteq G \subseteq H$ with $\omega(F) = \omega(G)$ and an $h \in H$ such that $\omega(G \cup h) > \omega(G)$ implies that $\omega(F \cup h) > \omega(F)$.

A *basis* for an LP-type problem is a subset $B \subset H$ such that $\omega(B') < \omega(B)$ for all proper subsets B' of B. And we say that B is a basis for a subset $G \subseteq H$ if $B \subseteq G$, $\omega(B) = \omega(G)$ and B is a basis. A constraint $h \in H$ *violates* a basis B if $w(B \cup h) > w(B)$. The radius of the smallest enclosing ball is an LP-type problem (where the points are the constraints and $\omega(\cdot) = f(\cdot)$) as are linear programming and

many other geometric problems. Let the maximum cardinality of any basis be the *combinatorial dimension* of a problem.

For our algorithm to run efficiently, we assume that our LP-type problem has available the following algorithmic primitive, which is often assumed for LP-type problems with constant combinatorial dimension [20]. For a subset $G \subset H$ where B is known to be the basis of G and a constraint $h \in H$, a *violation test* determines in $O(1)$ time if $\omega(B \cup h) > \omega(B)$; i.e., if h *violates* B. More specifically, given an efficient violation test, we can ensure a stronger algorithmic primitive. A *full violation test* is given a subset $G \subset H$ with known basis B and a constraint $h \in H$ and determines in $O(1)$ time if $\omega(B) < \omega(G \cup h)$. This follows because we can test in $O(1)$ time if $\omega(B) < \omega(B \cup h)$; MONOTONICITY implies that $\omega(B) < \omega(B \cup h)$ only if $\omega(B) < \omega(B \cup h) \leq \omega(G \cup h)$, and LOCALITY implies that $\omega(B) = \omega(B \cup h)$ only if $\omega(B) = \omega(G) = \omega(G \cup h)$. Thus we can test if h violates G by considering just B and h, but if either MONOTONICITY or LOCALITY fail for our problem we cannot.

We now adapt our algorithm to LP-type problems where elements of each Q_i are potential constraints and the ranking function is f. When the combinatorial dimension is a constant β, we need to consider only $O((nk)^\beta)$ bases, which will describe all possible supports.

The full violation test implies that given a basis B, we can measure the sum of probabilities of all supports of $\mathcal{Q}$ that have B as their basis in $O(nk)$ time. For each indecisive point Q such that $B \cap Q = \emptyset$, we sum the probabilities of all elements of Q that do not violate B. The product of these probabilities times the product of the probabilities of the elements in the basis, gives the probability of B being the true basis. See Algorithm 2.1 where the indicator function applied $1(f(B \cup \{q_j\}) = f(B))$ returns 1 if q_j does not violate B and 0 otherwise. It runs in $O((nk)^{\beta+1})$ time.

Algorithm 2.1 Construct Probability Distribution for $f(\mathcal{Q})$.

```
for all potential bases B ⊂ P ∈ Q do
    for i = 1 to n do
        if there is a j such that q_ij ∈ B then
            Set w_i = w(q_ij).
        else
            Set w_i = Σ_{j=1}^{k} w(q_ij) 1(f(B ∪ {q_j}) = f(B)).
    Store a point with value f(B) and weight (1/k^n) Π_i w_i.
```

As with the special case of smallest enclosing disk, we can create a distribution over the values of f given an indecisive point set $\mathcal{Q}$. For each basis B we calculate $\mu(B)$, the summed probability of all supports that have basis B, and $f(B)$. We can then sort these pairs according to the value as f again. For any query value r, we can retrieve $\tilde{f}(\mathcal{Q}, r)$ in $O(\log(nk))$ time and it takes $O(n)$ time to describe (because of its long length).

Theorem 2.2. *Given a set $\mathcal{Q}$ of n indecisive point sets of size k each, and given an LP-type problem $f : \mathcal{Q} \to \mathbb{R}$ with combinatorial dimension β, we can create the distribution of f over $\mathcal{Q}$ in $O((nk)^{\beta+1})$ time. The size of the distribution is $O(n(nk)^\beta)$.*

If we assume general position of $\mathcal{Q}$ relative to f, then we can often slightly improve the runtime needed to calculate $\mu(B)$ using range searching data structures. However, to deal with degeneracies, we may need to spend $O(nk)$ time per basis, regardless.

If we are content with an approximation of the distribution rather than an exact representation, then it is often possible to drastically reduce the storage and runtime. This requires the definition of *ε-quantizations* [17], which is delayed until Section 3 where it is discussed in more detail.

Measures that fit in this framework for points in $\mathbb{R}^d$ include smallest enclosing axis-aligned rectangle (measured either by area or perimeter) ($\beta = 2d$), smallest enclosing ball in the L_1, L_2, or L_∞ metric

($\beta = d + 1$), directional width of a set of points ($\beta = 2$), and, after dualizing, linear programming ($\beta = d$).

2.2 Hardness Results

In this section, we examine some extent measures that do not fit in the above framework. First, diameter does not satisfy the LOCALITY property, and hence we cannot efficiently perform the full violation test. We show that a decision variant of diameter is #P-Hard, even in the plane, and thus (under the assumption that #P $\neq$ P), there is no polynomial time solution. Second, the area of the convex hull does not have a constant combinatorial dimension, thus we can show the resulting distribution may have exponential size.

Diameter. The *diameter* of a set of points in the plane is the largest distance between any two points. We will show that the counting problem of computing $\tilde{f}(\mathcal{Q}, r)$ is #P-hard when f denotes the diameter.

Problem 2.1. *PLANAR-DIAM: Given a parameter d and a set $\mathcal{Q} = \{Q_1, \ldots, Q_n\}$ of n sets, each consisting of k points in the plane, how many supports $P \Subset \mathcal{Q}$ have $f(P) \leq d$?*

We will now prove that Problem 2.1 is #P-hard. Our proof has three steps. We first show a special version of #2SAT has a polynomial reduction from Monotone #2SAT, which is #P-complete [22]. Then, given an instance of this special version of #2SAT, we construct a graph with weighted edges on which the diameter problem is equivalent to this #2SAT instance. Finally, we show the graph can be embedded as a straight-line graph in the plane as an instance of PLANAR-DIAM.

Let 3CLAUSE-#2SAT be the problem of counting the number of solutions to a 2SAT formula, where each variable occurs in at most three clauses, and each variable is either in exactly one clause or is negated in exactly one clause. Thus, each distinct literal appears in at most two clauses.

Lemma 2.1. *Monotone #2SAT has a polynomial reduction to 3CLAUSE-#2SAT.*

Proof. The Monotone #2SAT problem counts the number satisfying assignments to a #2SAT instance where each clause has at most two variables and no variables are negated. Let $X = \{(x, y_1), (x, y_2), \ldots, (x, y_u)\}$ be the set of u clauses which contain variable x in a Monotone #2SAT instance. We replace x with u variables $\{z_1, z_2, \ldots, z_u\}$ and we replace X with the following $2u$ clauses $\{(z_1, y_1), (z_2, y_2), \ldots, (z_u, y_u)\}$ and $\{(z_1, \neg z_2), (z_2, \neg z_3), \ldots, (z_{u-1}, \neg z_u), (z_u, \neg z_1)\}$. The first set of clauses preserves the relation with other original variables and the second set of clauses ensures that all of the new variables have the same value (i.e. TRUE or FALSE). This procedure is repeated for each original variable that is in more than 1 clause. □

We convert this problem into a graph problem by, for each variable x_i, creating a set $Q_i = \{p_i^+, p_i^-\}$ of two points. Let $\mathcal{Q} = \bigcup_i Q_i$. Truth assignments of variables correspond to a support as follows. If x_i is set TRUE, then the support includes p_i^+, otherwise the support includes p_i^-. We define a distance function f between points, so that the distance is greater than d (long) if the corresponding literals are in a clause, and less than d (short) otherwise. If we consider the graph formed by only long edges, we make two observations. First, the maximum degree is 2, since each literal is in at most two clauses. Second, there are no cycles since a literal is only in two clauses if in one clause the other variable is negated, and negated variables are in only one clause. These two properties imply we can use the following construction to show that the PLANAR-DIAM problem is as hard as counting Monotone #2SAT solutions, which is #P-complete.

Lemma 2.2. *An instance of PLANAR-DIAM reduced from 3CLAUSE-#2SAT can be embedded so $\mathcal{Q} \subset \mathbb{R}^2$.*

Proof. Consider an instance of 3CLAUSE-#2SAT where there are n variables, and thus the corresponding graph has n sets $\{Q_i\}_{i=1}^n$. We construct a sequence Γ of $n' \in [2n, 4n]$ points. It contains all points from $\mathcal{Q}$ and a set of at most as many dummy points. First organize a sequence Γ' so if two points q and p have a long edge, then they are consecutive. Now for any pair of consecutive points in Γ' which do not have a long edge, insert a dummy point between them to form the sequence Γ. Also place a dummy point at the end of Γ.

We place all points on a circle C of diameter $d/\cos(\pi/n')$, see Figure 3. We first place all points on a semicircle of C according to the order of Γ, so each consecutive points are π/n' radians apart. Then for every other point (i.e. the points with an even index in the ordering Γ) we replace it with its antipodal point on C, so no two points are within $2\pi/n'$ radians of each other. Finally we remove all dummy points. This completes the embedding of $\mathcal{Q}$, we now need to show that only points with long edges are further than d apart.

We can now argue that only vertices which were consecutive in Γ are further than d apart, the remainder are closer than d. Consider a vertex p and a circle C_p of radius d centered at p. Let p' be the antipodal point of p on C. C_p intersects C at two points, at $2\pi/n'$ radians in either direction from p'. Thus only points within $2\pi/n'$ radians of p' are further than a distance d from p. This set includes only those points which are adjacent to p in Γ, which can only include points which should have a long edge, by construction.

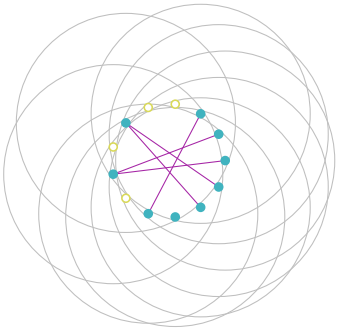


Figure 3: Embedded points are solid, at center of circles of radius d. Dummy points are hollow. Long edges are drawn between points at distance greater than d.

□

Combining Lemmas 2.1 and 2.2:

Theorem 2.3. *PLANAR-DIAM is #P-hard.*

Convex hull. Our LP-type framework also does not work for any properties of the convex hull (e.g. area or perimeter) because it does not have constant combinatorial dimension; a basis could have size n. In fact, the complexity of the distribution describing the convex hull may be $\Omega(k^n)$, since if all points in $\mathcal{Q}$ lie on or near a circle, then every support $P \Subset \mathcal{Q}$ may be its own basis of size n, and have a different value $f(P)$. Thus each support would create a separate data point in the distribution.

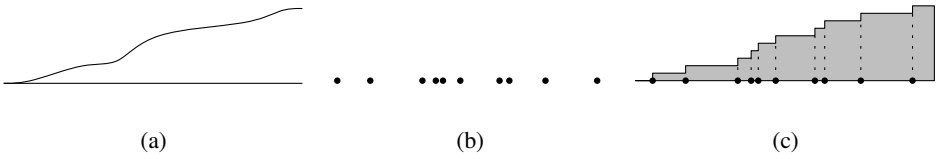
3 Approximate Computations on Uncertain Points

Perhaps the most general way to model an imprecise point is by providing a full probability distribution over $\mathbb{R}^d$; all popular simpler models can be seen as special cases of this. However, probability distributions can be difficult to handle, specifically, it is often impossible to do exact computations on them. In this section we show how the results from Section 2 can be used to approximately answer questions about uncertain points by representing each distribution by a discrete point set, resulting in a set of indecisive points.

In this section, we are given a set $\mathcal{X} = \{X_1, X_2, X_3, \ldots, X_n\}$ of n independent random variables over the universe $\mathbb{R}^d$, together with a set $\mathcal{M} = \{\mu_1, \mu_2, \mu_3, \ldots, \mu_n\}$ of n probability distributions that govern the variables, that is, $X_i \sim \mu_i$. We call a set of points $P = \{p_1, p_2, p_3, \ldots, p_n\}$ a *support* of $\mathcal{X}$, and because of the independence we have probability $Pr[\mathcal{X} = P] = \prod_i Pr[X_i = p_i]$.

As before, we are interested in functions f that measure something about a point set. We now define $\hat{f}(\mathcal{X}, r)$ as the probability that a support P drawn from $\mathcal{M}$ satisfies $f(P) \leq r$. In most cases, we cannot evaluate this function exactly, but previous work [17] describes a Monte Carlo algorithm for approximating $\hat{f}(\mathcal{X}, r)$. Here we show how to make this approximate construction deterministic.

To approximate $\hat{f}(\mathcal{X}, r)$, we construct an *ε-quantization* [17]. Let $g : \mathbb{R} \to \mathbb{R}^+$ be a distribution so $\int_{\mathbb{R}} g(x)\ dx = 1$, and let G be its integral so $G(t) = \int_{-\infty}^{t} g(x)\ dx$. Then $G : \mathbb{R} \to [0, 1]$ is a cumulative density function. Also, let R be a set of points in $\mathbb{R}$, that is, a set of values. Then R induces a function H_R where $H_R(v) = \frac{|\{r \in R | r \leq v\}|}{|R|}$, that is, $H_R(v)$ describes the fraction of points in R with value at most v. We say that R is an ε-quantization of g if H_R approximates G within ε, that is, for all values v we have $|H_R(v) - G(v)| \leq \varepsilon$. Figure 4 shows an example of an ε-quantization. Note that we can apply techniques



(a) (b) (c)

Figure 4: (a) The true form of the cumulative function G. (b) The ε-quantization R as a point set in $\mathbb{R}$. (c) The inferred curve H_R in $\mathbb{R}^2$.

from [17] to deterministically reduce the exact distributions created in Section 2 to ε-quantizations of size $O(1/\varepsilon)$.

The main strategy will be to replace each distribution μ_i by a discrete point set Q_i, such that the uniform distribution over Q_i is "not too far" from μ_i (Q_i is not the most obvious ε-sample of μ_i). Then we apply the algorithms from Section 2 to the resulting set of point sets. Finally, we argue that the result is in fact an ε-quantization of the distribution we are interested in, and we show how to simplify the output in order to decrease the space complexity for the data structure, without increasing the approximation factor too much.

Range spaces and ε-samples. Before we describe the algorithms, we need to formally define range spaces and ε-samples. Given a set of elements Y let $\mathcal{A} \subset 2^Y$ be a family of subsets of Y. For instance, if Y is a point set, $\mathcal{A}$ could be all subsets defined by intersection with a ball. A pair $T = (Y, \mathcal{A})$ is called a *range space*.

We say a set of ranges $\mathcal{A}$ *shatters* a set Y if for every subset $Y' \subseteq Y$ there exists some $A \in \mathcal{A}$ such that $Y' = Y \cap A$. The size of the largest subset $Y' \subseteq Y$ that $\mathcal{A}$ shatters is the *VC-dimension* [24] of $T = (Y, \mathcal{A})$, denoted ν_T.

Let $\mu : Y \to \mathbb{R}^+$ be a measure on Y. For discrete sets Y μ is cardinality, for continuous sets Y μ is a Lebesgue measure. An *ε-sample* (often referred to by the generic term ε-approximation) of a range space $T = (Y, \mathcal{A})$ is a subset $S \subset Y$ such that $\forall A \in \mathcal{A} : \left| \frac{\mu(A \cap S)}{\mu(S)} - \frac{\mu(A \cap Y)}{\mu(Y)} \right| \leq \varepsilon$. A random subset $S \subset Y$ of size $O((1/\varepsilon^2)(\nu_T + \log(1/\delta)))$ is an ε-sample with probability at least $1 - \delta$ [24, 16]. For a range space $T = (Y, \mathcal{A})$ with Y discrete and $\mu(Y) = n$, there are also deterministic algorithms to generate ε-samples of size $O((\nu_T/\varepsilon^2)\log(1/\varepsilon))$ in time $O(\nu_T^{3\nu_T} n((1/\varepsilon^2)\log(\nu_T/\varepsilon))^{\nu_T})$ [6]. Or when the ranges $\mathcal{A}$ are defined by the intersection of k oriented slabs (i.e. axis-aligned rectangles with $k = d$), then an ε-sample of size $O((k/\varepsilon)\log^{2k}(1/\varepsilon))$ can be deterministically constructed in $O((n/\varepsilon^3)\log^{6k}(1/\varepsilon))$ time [19].

In the continuous setting, we can think of each point $y \in Y$ as representing $\mu(y)$ points, and for simplicity represent a weighted range space as $(\mu, \mathcal{A})$ when the domain of the function μ is implicitly known to be Y (often $\mathbb{R}^2$ or $\mathbb{R}^d$). Phillips [19] studies ε-samples for such weighted range spaces with ranges defined as intersection of k intervals; Appendix B extends this. The appendix also supplies constructions that ensure that the resulting ε-samples are in general position, which is helpful if these are to be used in conjunction with the machinery in Section 2. These results, as well as those following, may be of independent interest in further understanding the structure required to transition from continuous to discrete representations of distributions.

General approach (KEY CONCEPTS). Given a distribution $\mu_i : \mathbb{R}^2 \to \mathbb{R}^+$ describing uncertain point X_i and a function f of bounded combinatorial dimension β defined on a support of $\mathcal{X}$, we can describe a straightforward range space $T_i = (\mu_i, \mathcal{A}_f)$, where $\mathcal{A}_f$ is the set of ranges corresponding to the bases of f (e.g., when f measures the radius of the smallest enclosing ball, $\mathcal{A}_f$ would be the set of all balls). More formally, $\mathcal{A}_f$ is the set of subsets of $\mathbb{R}^d$ defined as follows: for every set of β points which define a basis B for f, $\mathcal{A}_f$ contains a range A that contains all points p such that $f(B) = f(B \cup \{p\})$. However, taking

ε-samples from each T_i is *not* sufficient to create sets Q_i such that $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_n\}$ so for all r we have $|\tilde{f}(\mathcal{Q}, r) - \hat{f}(\mathcal{X}, r)| \leq \varepsilon$.

$\hat{f}(\mathcal{X}, r)$ is a complicated joint probability depending on the n distributions and f, and the n straightforward ε-samples do not contain enough information to decompose this joint probability. The required ε-sample of each μ_i should model μ_i in relation to f and any instantiated point q_i representing μ_j for $i \neq j$. The following crucial definition allows for the range space to depend on any $n-1$ points, including the possible locations of each uncertain point.

Let $\mathcal{A}_{f,n}$ describe a family of Lebesgue-measurable sets defined by $n-1$ points $Z \subset \mathbb{R}^d$ and a value w. Specifically, $A(Z, w) \in \mathcal{A}_{f,n}$ is the set of points $\{p \in \mathbb{R}^d \mid f(Z \cup p) \leq w\}$. We describe examples of $\mathcal{A}_{f,n}$ in detail shortly, but first we state the key theorem using this definition. Its proof, delayed until after examples of $\mathcal{A}_{f,n}$, will make clear how $(\mu_i, \mathcal{A}_{f,n})$ exactly encapsulates the right guarantees to approximate $\hat{f}(\mathcal{X}, r)$, and thus why $(\mu_i, \mathcal{A}_f)$ does not.

Theorem 3.1. *Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be a set of uncertain points where each $X_i \sim \mu_i$. For a function f, let Q_i be an ε'-sample of $(\mu_i, \mathcal{A}_{f,n})$ and let $\mathcal{Q} = \{Q_1, \ldots, Q_n\}$. Then for any r, $\left|\hat{f}(\mathcal{X}, r) - \tilde{f}(\mathcal{Q}, r)\right| \leq \varepsilon' n$.*

3.1 Smallest Axis-Aligned Bounding Box by Perimeter

Given a set of points $P \subset \mathbb{R}^2$, let $f(P)$ represent the perimeter of the smallest axis-aligned box that contains P. Let each μ_i be a bivariate normal distribution with constant variance. Solving $f(P)$ is an LP-type problem with combinatorial dimension $\beta = 4$, and as such, we can describe the basis B of a set P as the points with minimum and maximum x- and y-coordinates. Given any additional point p, the perimeter of size ρ can only be increased to a value w by expanding the range of x-coordinates, y-coordinates, or both. As such, the region of $\mathbb{R}^2$ described by a range $A(P, w) \in \mathcal{A}_{f,n}$ is defined with respect to the bounding box of P from an edge increasing the x-width or y-width by $(w - \rho)/2$, or from a corner extending so the sum of the x and y deviation is $(w - \rho)/2$. See Figure 5(Left).

Since any such shape defining a range $A(P, w) \in \mathcal{A}_{f,n}$ can be described as the intersection of $k = 4$ slabs along fixed axis (at $0°$, $45°$, $90°$, and $135°$), we can construct an (ε/n)-sample Q_i of $(\mu_i, \mathcal{A}_{f,n})$ of size $k = O((n/\varepsilon)\log^8(n/\varepsilon))$ in $O((n^6/\varepsilon^6)\log^{27}(n/\varepsilon))$ time [19]. From Theorem 3.1, it follows that for $\mathcal{Q} = \{Q_1, \ldots, Q_n\}$ and any r we have $\left|\hat{f}(\mathcal{X}, r) - \tilde{f}(\mathcal{Q}, r)\right| \leq \varepsilon$.

We can then apply Theorem 2.2 to build an ε-quantization of $f(\mathcal{X})$ in $O((nk)^5) = O(((n^2/\varepsilon)\log^8(n/\varepsilon))^5) = O((n^{10}/\varepsilon^5)\log^{40}(n/\varepsilon))$ time. The size can be reduced to $O(1/\varepsilon)$ within that time bound. Assuming $1/\varepsilon < n$, we can state the following corollary of Theorem 3.1.

Corollary 3.1. *Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be a set of indecisive points where each $X_i \sim \mu_i$ is bivariate normal with constant variance. Let f measure the perimeter of the smallest enclosing axis-aligned bounding box. We can create an ε-quantization of $f(\mathcal{X})$ in $O((n^{10}/\varepsilon^5)\log^{40}(n/\varepsilon))$ time of size $O(1/\varepsilon)$.*

3.2 Smallest Enclosing Disk

Given a set of points $P \subset \mathbb{R}^2$, let $f(P)$ represent the radius of the smallest enclosing disk of P. Let each μ_i be a bivariate normal distribution with constant variance. Solving $f(P)$ is an LP-type problem with combinatorial dimension $\beta = 3$, and the basis B of P generically consists of either 3 points which lie on the boundary of the smallest enclosing disk, or 2 points which are antipodal on the smallest enclosing disk. However, given an additional point $p \in \mathbb{R}^2$, the new basis B_p is either B or it is p along with 1 or 2 points which lie on the convex hull of P.

We can start by examining all pairs of points $p_i, p_j \in P$ and the two disks of radius w whose boundary circles pass through them. If one such disk $D_{i,j}$ contains P, then $D_{i,j} \subset A(P, w) \in \mathcal{A}_{f,|P|+1}$. For this to

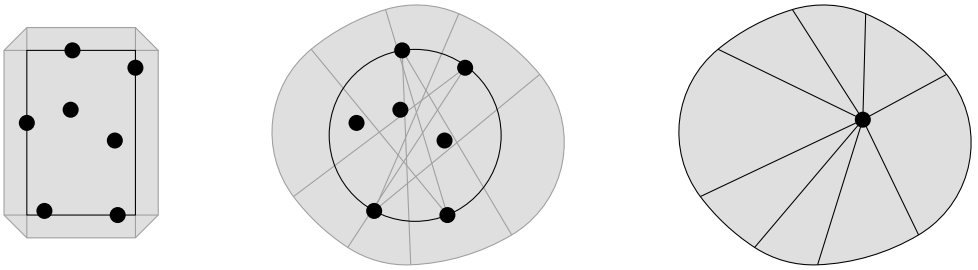


Figure 5: Left: A shape from $\mathcal{A}_{f,n}$ for axis-aligned bounding box, measured by perimeter. Middle: A shape from $\mathcal{A}_{f,n}$ for smallest enclosing ball using the L_2 metric. The curves are circular arcs of two different radii. Right: The same shape divided into wedges from $\mathcal{W}_{f,n}$.

hold, p_i and p_j must lie on the convex hull of P and no point that lies between them on the convex hull can contribute to such a disk. Thus there are $O(n)$ such disks. We also need to examine the disks created where p and one other point $p_i \in P$ are antipodal. The boundary of the union of all such disks which contain P is described by part of a circle of radius $2w$ centered at some $p_i \in P$. Again, for such a disk B_i to describe a part of the boundary of $A(P, w)$, the point p_i must lie on the convex hull of P. The circular arc defining this boundary will only connect two disks $D_{i,j}$ and $D_{k,i}$ because it will intersect with the boundary of B_j and B_k within these disks, respectively. An example of $A(P, w)$ is shown in Figure 5(Middle).

Unfortunately, the range space $(\mathbb{R}^2, \mathcal{A}_{f,n})$ has VC-dimension $O(n)$; it has $O(n)$ circular boundary arcs. So, creating an ε-sample of $T_i = (\mu_i, \mathcal{A}_{f,n})$ would take time exponential in n. However, we can decompose any range $A(P, w) \in \mathcal{A}_{f,n}$ into at most $2n$ "wedges." We choose one point y inside the convex hull of P. For each circular arc on the boundary of $A(P, w)$ we create a wedge by coning that boundary arc to y. Let $\mathcal{W}_f$ describe all wedge shaped ranges. Then $S = (\mathbb{R}^2, \mathcal{W}_f)$ has VC-dimension ν_S at most 9 since it is the intersection of 3 ranges (two halfspaces and one disk) that can each have VC-dimension 3. We can then create Q_i, an $(\varepsilon/2n^2)$-sample of $S_i = (\mu_i, \mathcal{W}_f)$, of size $k = O((n^4/\varepsilon^2)\log(n/\varepsilon))$ in $O((n^2/\varepsilon)^{5+2\cdot 9}\log^{2+9}(n/\varepsilon)) = O((n^{46}/\varepsilon^{23})\log^{11}(n/\varepsilon))$ time, via Corollary B.1 (Appendix B). It follows that Q_i is an (ε/n)-sample of $T_i = (\mu_i, \mathcal{A}_{f,n})$, since any range $A(Z, w) \in \mathcal{A}_{f,n}$ can be decomposed into at most $2n$ wedges, each of which has counting error at most $\varepsilon/2n$, thus the total counting error is at most ε.

Invoking Theorem 3.1, it follows that $\mathcal{Q} = \{Q_1, \ldots, Q_n\}$, for any r we have $\left|\hat{f}(\mathcal{X}, r) - \tilde{f}(\mathcal{Q}, r)\right| \leq \varepsilon$. We can then apply Theorem 2.2 to build an ε-quantization of $f(\mathcal{X})$ in $O((nk)^4) = O((n^{20}/\varepsilon^8)\log^4(n/\varepsilon))$ time. This is dominated by the time for creating the n (ε/n^2)-samples, even though we only need to build one and then translate and scale to the rest. Again, the size can be reduced to $O(1/\varepsilon)$ within that time bound.

Corollary 3.2. *Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be a set of indecisive points where each $X_i \sim \mu_i$ is bivariate normal with constant variance. Let f measure the radius of the smallest enclosing disk. We can create an ε-quantization of $f(\mathcal{X})$ in $O((n^{46}/\varepsilon^{23})\log^{11}(n/\varepsilon))$ time of size $O(1/\varepsilon)$.*

Now that we have seen two concrete examples, we prove Theorem 3.1. More examples can be found in Appendix C.

3.3 Proof of Theorem 3.1

When each X_i is drawn from a distribution μ_i, then we can write $\hat{f}(\mathcal{X}, r)$ as the probability that $f(\mathcal{X}) \leq r$ as follows. Let $1(\cdot)$ be the indicator function, i.e., it is 1 when the condition is true and 0 otherwise.

$$\hat{f}(\mathcal{X}, r) = \int_{p_1} \mu_1(p_1) \ldots \int_{p_n} \mu_n(p_n) 1(f(\{p_1, p_2, \ldots, p_n\}) \leq r)\, dp_n dp_{n-1} \ldots dp_1$$

Consider the inner most integral

$$\int_{p_n} \mu_n(p_n) 1(f(\{p_1, p_2, \ldots, p_n\}) \leq r)\, dp_n,$$

where $\{p_1, p_2 \ldots, p_{n-1}\}$ are fixed. The indicator function is true when

$$f(\{p_1, p_2, \ldots, p_{n-1}, p_n\}) \leq r,$$

and hence p_n is contained in a shape $A(\{p_1, \ldots, p_{n-1}\}, r) \in \mathcal{A}_{f,n}$. Thus if we have an ε'-sample Q_n for $(\mu_n, \mathcal{A}_{f,n})$, then we can guarantee that

$$\int_{p_n} \mu_n(p_n) 1(f(\{p_1, p_2, \ldots, p_n\}) \leq r)\, dp_n \leq \frac{1}{|Q_n|} \sum_{p_n \in Q_n} 1(f(\{p_1, p_s, \ldots, p_n\}) \leq r) + \varepsilon'.$$

We can then move the ε' outside and change the order of the integrals to write:

$$\hat{f}(\mathcal{X}, r) \leq \frac{1}{|Q_n|} \sum_{p_n \in Q_n} \left(\int_{p_1} \mu_1(p_1) \ldots \int_{p_{n-1}} \mu_{n-1}(p_{n-1}) 1(f(\{p_1, \ldots, p_n\}) \leq r) dp_{n-1} \ldots dp_1 \right) + \varepsilon'.$$

Repeating this procedure n times, we get:

$$\begin{aligned} \hat{f}(\mathcal{X}, r) &\leq \left(\prod_{i=1}^{n} \frac{1}{|Q_i|} \right) \sum_{p_1 \in Q_1} \cdots \sum_{p_n \in Q_n} 1(f(\{p_1, p_2, \ldots, p_n\}) \leq r) + \varepsilon' n. \\ &= \tilde{f}(\mathcal{Q}, r) + \varepsilon' n, \end{aligned}$$

where $\mathcal{Q} = \bigcup_i Q_i$.

Similarly we can achieve a symmetric lower bound for $\hat{f}(\mathcal{X}, r)$. □

Acknowledgements

We thank Joachim Gudmundsson and Pankaj Agarwal for helpful discussions in early phases of this work, Sariel Har-Peled for discussions about wedges, and Suresh Venkatasubramanian for organizational tips.

References

[1] P. Agrawal, O. Benjelloun, A. D. Sarma, C. Hayworth, S. Nabar, T. Sugihara, and J. Widom. Trio: A system for data, uncertainty, and lineage. In *PODS*, 2006.

[2] H.-K. Ahn, S. W. Bae, S.-S. Kim, M. Korman, I. Reinbacher, and W. Son. Square and rectangle covering with outliers. In *Proc. 25th European Workshop on Computational Geometry*, pages 273–276, 2009.

[3] R. Atanassov, P. Bose, M. Couture, A. Maheshwari, P. Morin, M. Paquette, M. Smid, and S. Wuhrer. Algorithms for optimal outlier removal. *Journal of Discrete Algorithms*, 7(2):239–248, 2009.

[4] D. Bandyopadhyay and J. Snoeyink. Almost-Delaunay simplices: Nearest neighbor relations for imprecise points. In *SODA*, 2004.

[5] J. Bi and T. Zhang. Support vector classification with input data uncertainty. In *NIPS*, 2004.

[6] B. Chazelle and J. Matousek. On linear-time deterministic algorithms for optimization problems in fixed dimensions. *Journal of Algorithms*, 21:579–597, 1996.

[7] G. Cormode and M. Garafalakis. Histograms and wavelets of probabilitic data. In *ICDE*, 2009.

[8] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In *PODS*, 2008.

[9] N. Dalvi and D. Suciu. Efficient query evaluation on probabilitic databases. *The VLDB Journal*, 16:523–544, 2007.

[10] G. N. Frederickson and D. B. Johnson. Generalized selection and ranking: Sorted matrices. *SIAM Journal on Computing*, 13:14–30, 1984.

[11] M. Fürer. Faster integer multiplication. *SIAM J. Computing*, 39:979–1005, 2009.

[12] L. J. Guibas, D. Salesin, and J. Stolfi. Epsilon geometry: building robust algorithms from imprecise computations. In *SoCG*, 1989.

[13] S. Har-Peled. Chapter 5: On complexity, sampling, and ε-nets and ε-samples. http://valis.cs.uiuc.edu/~sariel/teach/notes/aprx/lec/05_vc_dim.pdf, May 2010.

[14] S. Har-Peled and Y. Wang. Shape fitting with outliers. In *Proc. 19th Symposium on Computational Geometry*, pages 29–38, New York, NY, USA, 2003. ACM.

[15] M. Held and J. S. B. Mitchell. Triangulating input-constrained planar point sets. *Information Processing Letters*, 109(1), 2008.

[16] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the samples complexity of learning. *Journal of Computer and System Sciences*, 62:516–527, 2001.

[17] M. Löffler and J. Phillips. Shape fitting on point sets with probability distributions. In *ESA*, 2009.

[18] J. Matousek. *Geometric Discrepancy*. Springer, 1999.

[19] J. M. Phillips. Algorithms for ε-approximations of terrains. In *ICALP*, 2008.

[20] M. Sharir and E. Welzl. A combinatorial bound for linear programming and related problems. In *STACS*, 1992.

[21] Y. Tao, R. Cheng, X. Xiao, W. K. Ngai, B. Kao, and S. Prabhakar. Indexing multi-dimensional uncertain data with arbitrary probability density functions. In *VLDB*, 2005.

[22] L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8:410–421, 1979.

[23] M. van Kreveld and M. Löffler. Largest bounding box, smallest diameter, and related problems on imprecise points. *Computational Geometry: Theory and Applications*, 43:419–433, 2010.

[24] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[25] Y. Zou and K. Chakrabarty. Uncertainty-aware and coverage-oriented deployment of sensor networks. *Journal of Parallel and Distributed Computing*, 2004.

A Algorithm 2.1 in RAM model

In this section we will analyze Algorithm 2.1 without the assumption that the integer k^n can be stored in $O(1)$ words. To simplify the results, we assume a RAM model where a word size contains $O(\log(nk))$ bits, each weight $w(q_{i,j})$ can be stored in one word, and the weight of any basis $f(B)$, the product of n $O(1)$ word weights, can thus be stored in $O(n)$ words. The following lemma describes the main result we will need relating to large numbers.

Lemma A.1. *In the RAM model where a word size has b bits we can calculate the product of n numbers where each is described by $O(b)$ bits in $(nb\log^2 n)2^{O(\log^* n)}$ time.*

Proof. Using Fürer's recent result [11] we can multiply two m-bit numbers in $m\log m 2^{O(\log^* m)}$ bit operations. The product of n m-bit numbers has $O(\log((2^m)^n) = O(nm)$ bits, and can be accomplished with $n-1$ pairwise multiplications.

We can calculate this product efficiently from the bottom up, starting with $n/2$ multiplications of two $m = O(b)$ bit numbers. Then we perform $n/4$ multiplications of two $2m$ bit numbers, and so on. Since each operation on $O(b)$ bits takes $O(1)$ time in our model, Fürer's result clearly upper bounds the RAM result. The total cost of this can be written as

$$\sum_{i=0}^{\log n} \frac{n}{2^{i-1}}(2^i m)\log(2^i m)2^{O(\log^*(2^i m))} = 2nb\sum_{i=0}^{\log n}(i+\log b)2^{O(\log^*(2^i \log b))} = nb\log n\log(nb)2^{O(\log^*(nb))}.$$

Since we assume $b = O(\log n)$ we can simplify this bound to $(nb\log^2 n)2^{O(\log^* n)}$. □

This implies that Algorithm 2.1 takes $O((nk)^{\beta+1}) + ((nk)^\beta n\log(nk)\log^2 n)2^{O(\log^* n)}$ time where each w_i can be described in $O(b) = O(\log(nk))$ bits. This dominates the single division and all other operations. Now we can rewrite Theorem 2.2 without the restriction that k^n can be stored in $O(1)$ words.

Theorem A.1. *Given a set $\mathcal{Q}$ of n indecisive point sets of size k each, and given an LP-type problem $f : \mathcal{Q} \to \mathbb{R}$ with combinatorial dimension β, we can create the distribution of f over $\mathcal{Q}$ in $O((nk)^{\beta+1}) + ((nk)^\beta n\log(nk)\log^2 n)2^{O(\log^* n)}$ time. The size of the distribution is $O(n(nk)^\beta)$.*

B ε-Samples of Distributions

In this section we explore conditions for continuous distributions such that they can be approximated with bounded error by discrete distributions (point sets). We state specific results for multi-variate normal distributions.

We say a subset $W \subset \mathbb{R}^d$ is *polygonal approximable* if there exists a polygonal shape $S \subset \mathbb{R}^d$ with m facets such that $\phi(W \setminus S) + \phi(S \setminus W) \leq \varepsilon\phi(W)$ for any $\varepsilon > 0$. Usually, m is dependent on ε, for instance for a d-variate normal distribution $m = O((1/\varepsilon^{d+1})\log(1/\varepsilon))$ [19]. In turn, such a polygonal shape S describes a continuous point set where $(S, \mathcal{A})$ can be given an ε-sample Q using $O((1/\varepsilon^2)\log(1/\varepsilon))$ points if $(S, \mathcal{A})$ has bounded VC-dimension [18] or using $O((1/\varepsilon)\log^{2k}(1/\varepsilon))$ points if $\mathcal{A}$ is defined by a constant k number of directions [19]. For instance, where $\mathcal{A} = \mathcal{B}$ is the set of all balls then the first case applies, and when $\mathcal{A} = \mathcal{R}_2$ is the set of all axis-aligned rectangles then either case applies.

A shape $W \subset \mathbb{R}^{d+1}$ may describe a distribution $\mu : \mathbb{R}^d \to [0, 1]$. For instance for a range space $(\mu, \mathcal{B})$, then the range space of the associated shape W_μ is $(W_\mu, \mathcal{B} \times \mathbb{R})$ where $\mathcal{B} \times \mathbb{R}$ describes balls in $\mathbb{R}^d$ for the first d coordinates and any points in the $(d+1)$th coordinate.

The general scheme to create an ε-sample for $(S, \mathcal{A})$, where $S \in \mathbb{R}^d$ is a polygonal shape, is to use a lattice Λ of points. A *lattice* Λ in $\mathbb{R}^d$ is an infinite set of points defined such that for d vectors $\{v_1, \ldots, v_d\}$ that form a basis, for any point $p \in \Lambda$, $p + v_i$ and $p - v_i$ are also in Λ for any $i \in [1, d]$. We first

create a discrete $(\varepsilon/2)$-sample $M = \Lambda \cap S$ of $(S, \mathcal{A})$ and then create an $(\varepsilon/2)$-sample Q of $(M, \mathcal{A})$ using standard techniques [6, 19]. Then Q is an ε-sample of $(S, \mathcal{A})$. For a shape S with m $(d-1)$-faces on its boundary, any subset $A' \subset \mathbb{R}^d$ that is described by a subset from $(S, \mathcal{A})$ is an intersection $A' = A \cap S$ for some $A \in \mathcal{A}$. Since S has m $(d-1)$-dimensional faces, we can bound the VC-dimension of $(S, \mathcal{A})$ as $\nu = O((m + \nu_{\mathcal{A}}) \log(m + \nu_{\mathcal{A}}))$ where $\nu_{\mathcal{A}}$ is the VC-dimension of $(\mathbb{R}^d, \mathcal{A})$. Finally the set $M = S \cap \Lambda$ is determined by choosing an arbitrary initial origin point in Λ and then uniformly scaling all vectors $\{v_1, \ldots, v_d\}$ until $|M| = \Theta((\nu/\varepsilon^2) \log(\nu/\varepsilon))$ [18]. This construction follows a less general but smaller construction in Phillips [19].

It follows that we can create such an ε-sample K of $(S, \mathcal{A})$ of size $|M|$ in time $O(|M| m \log |M|)$ by starting with a scaling of the lattice so a constant number of points are in S and then doubling the scale until we get to within a factor of d of $|M|$. If there are n points inside S, it takes $O(nm)$ time to count them. We can then take another ε-sample of $(K, \mathcal{A})$ of size $O((\nu_{\mathcal{A}}/\varepsilon^2) \log(\nu_{\mathcal{A}}/\varepsilon))$ in time $O(\nu_{\mathcal{A}}^{3\nu_{\mathcal{A}}} |M| ((1/\varepsilon^2) \log(\nu_{\mathcal{A}}/\varepsilon))^{\nu_{\mathcal{A}}})$.

Theorem B.1. *For a polygonal shape $S \subset \mathbb{R}^d$ with m (constant size) facets, we can construct an ε-sample for $(S, \mathcal{A})$ of size $O((\nu/\varepsilon^2) \log(\nu/\varepsilon))$ in time $O(m(\nu/\varepsilon^2) \log^2(\nu/\varepsilon))$, where $(S, \mathcal{A})$ has VC-dimension $\nu_{\mathcal{A}}$ and $\nu = O((\nu_{\mathcal{A}} + m) \log(\nu_{\mathcal{A}} + m))$.*

This can be reduced to size $O((\nu_{\mathcal{A}}/\varepsilon^2) \log(\nu_{\mathcal{A}}/\varepsilon))$ in time $O((1/\varepsilon^{2\nu_{\mathcal{A}}+2})(m+\nu_{\mathcal{A}}) \log^{\nu_{\mathcal{A}}}((m+\nu_{\mathcal{A}})/\varepsilon) \log(m/\varepsilon)$.

We can consider the specific case of when $W \subset \mathbb{R}^3$ is a d-variate normal distribution $\mu : \mathbb{R}^d \to \mathbb{R}^+$. Then $m = O((1/\varepsilon^d) \log(1/\varepsilon))$ and $|M| = O((m/\varepsilon^2) \log m \log(m/\varepsilon)) = O((1/\varepsilon^{d+2}) \log^3(1/\varepsilon))$.

Corollary B.1. *Let $\mu : \mathbb{R}^d \to \mathbb{R}^+$ be a d-variate normal distribution with constant standard deviation. We can construct an ε-sample of $(\mu, \mathcal{A})$ with VC-dimension $\nu_{\mathcal{A}} \geq 2$ (and where $d \leq \nu_{\mathcal{A}} \leq 1/\varepsilon$) of size $O((\nu_{\mathcal{A}}/\varepsilon^2) \log(\nu_{\mathcal{A}}/\varepsilon))$ in time $O((1/\varepsilon^{2\nu_{\mathcal{A}}+d+2}) \log^{\nu_{\mathcal{A}}+1}(1/\varepsilon))$.*

For convenience, we restate a tighter, but less general theorem from Phillips, here slightly generalized.

Theorem B.2 ([19])**.** *Let $\mu : \mathbb{R}^d \to \mathbb{R}^+$ be a d-variate normal distribution with constant standard deviation. Let $(\mu, \mathcal{Q}_k)$ be a range space where the ranges are defined as the intersection of k slabs with fixed normal directions. We can construct an ε-sample of $(\mu, \mathcal{Q}_k)$ of size $O((1/\varepsilon) \log^{2k}(1/\varepsilon))$ in time $O((1/\varepsilon^{d+4}) \log^{6k+3}(1/\varepsilon))$.*

B.1 Avoiding Degeneracy

An important part of the above construction is the arbitrary choice of the origin points of the lattice Λ. This allows us to arbitrarily shift the lattice defining M and thus the set Q. In Section 2.1 we need to construct n ε-samples $\{Q_1, \ldots, Q_n\}$ for n range spaces $\{(S_1, \mathcal{A}), \ldots, (S_n, \mathcal{A})\}$. In Algorithm 2.1 we examine sets of $\nu_{\mathcal{A}}$ points, each from separate ε-samples that define a minimal shape $A \in \mathcal{A}$. It is important that we do not have two such (possibly not disjoint) sets of $\nu_{\mathcal{A}}$ points that define the same minimal shape $A \in \mathcal{A}$. (Note, this does not include cases where say two points are antipodal on a disk and any other point in the disk added to a set of $\nu_{\mathcal{A}} = 3$ points forms such a set; it refers to cases where say four points lie (degenerately) on the boundary of a disc.) We can guarantee this by enforcing a property on all pairs of origin points p and q for $(S_i, \mathcal{A})$ and $(S_j, \mathcal{A})$. For the purpose of construction, it is easiest to consider only the lth coordinates p_l and q_l for any pair of origin points or lattice vectors (where the same lattice vectors are used for each lattice). We enforce a specific property on every such pair p_l and q_l, for all l and all distributions and lattice vectors.

First, consider the case where $\mathcal{A} = \mathcal{R}_d$ describes axis-aligned bounding boxes. It is easy to see that if for all pairs p_l and q_l that $(p_l - q_l)$ is irrational, then we cannot have $> 2d$ points on the boundary of an axis-aligned bounding box, hence the desired property is satisfied.

Now consider the more complicated case where $\mathcal{A} = \mathcal{B}$ describes smallest enclosing balls. There is a polynomial of degree 2 that describes the boundary of the ball, so we can enforce that for all pairs p_l and q_l

that $(p_l - q_l)$ is of the form $c_1(r_{p_l})^{1/3} + c_2(r_{q_l})^{1/3}$ where c_1 and c_2 are rational coefficients and r_{p_l} and r_{q_l} are distinct integers that are not multiple of cubes. Now if $\nu = d + 1$ such points satisfy (and in fact define) the equation of the boundary of a ball, then no $(d + 2)$th point which has this property with respect to the first $d + 1$ can also satisfy this equation.

More generally, if $\mathcal{A}$ can be described with a polynomial of degree p with ν variables, then enforce that every pair of coordinates are the sum of $(p + 1)$-roots. This ensures that no $\nu + 1$ points can satisfy the equation, and the undesired situation cannot occur.

C Computing Other Measures on Uncertain Points

We generalize this machinery to other LP-type problems f defined on a set of points in $\mathbb{R}^d$ and with constant combinatorial dimension. Although, in some cases (like smallest axis-aligned bounding box by perimeter) we are able to show that $(\mathbb{R}^d, \mathcal{A}_{f,n})$ has constant VC-dimension, for other cases (like radius of the smallest enclosing disk) we cannot and need to first decompose each range $A(Z, w) \in \mathcal{A}_{f,n}$ into a set of disjoint "wedges" from a family of ranges $\mathcal{W}_f$.

To simplify the already large polynomial runtimes below we replace the runtime bound in Theorem 2.2 with $O((nk)^{\beta+1} \log^4(nk))$.

Lemma C.1. *If the disjoint union of m shapes from $\mathcal{W}_f$ can form any shape from $\mathcal{A}_{f,n}$, then an (ε/m)-sample of $(M, \mathcal{W}_f)$ is an ε-sample of $(M, \mathcal{A}_{f,n})$.*

Proof. For any shape $A \in \mathcal{A}_{f,n}$ we can create a set of m shapes $\{W_1, \ldots, W_m\} \subset \mathcal{W}_f$ whose disjoint union is A. Since each range of $\mathcal{W}_f$ may have error ε/m, their union has error at most ε. □

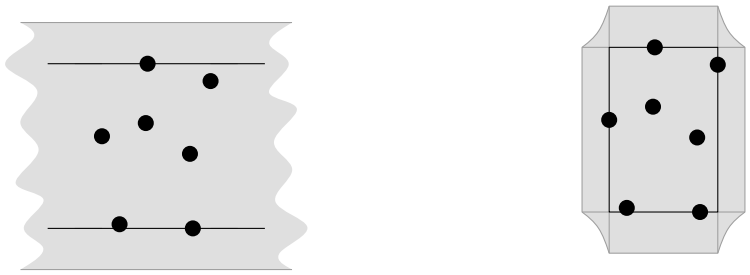
We study several example cases for which we can deterministically compute ε-quantizations. For each case we show an example element of $\mathcal{A}_{f,n}$ on an example of 7 points.

To facilitate the analysis, we define the notion of shatter dimension, which is similar to VC-dimension. The shatter function $\pi_T(m)$ of a range space $T = (Y, \mathcal{A})$ is the maximum number of sets in T where $|Y| = m$. The *shatter dimension* σ_T of a range space $T = (Y, \mathcal{A})$ is the minimum value such that $\pi_T(m) = O(m^{\sigma_T})$. If a range $A \in \mathcal{A}$ is defined by k points, then $k \geq \sigma_T$. It can be shown [13] that $\sigma_T \leq \nu_T$ and $\nu_T = O(\sigma_T \log \sigma_T)$. And, in general, the basis size of the related LP-type problem is bounded $\beta \leq \sigma_T$.

C.1 Directional Width

We first consider the problem of finding the width along a particular direction u (dwid). Given a point set P, $f(P)$ is the width of the minimum slab containing P, as in Figure 6(a). This can be thought of as a one-dimensional problem by projecting all points P using the operation $\langle \cdot, u \rangle$. The directional width is then just the difference between the largest point and the smallest point. As such, the VC-dimension of $(\mathbb{R}^d, \mathcal{A}_f)$ is 2. Furthermore, $\mathcal{A}_{f,n} = \mathcal{A}_f$ in this case, so $(\mathbb{R}^d, \mathcal{A}_{f,n})$ also has VC-dimension 2. For $1 \leq i \leq n$, we can then create an (ε/n)-sample Q_i of $(\mu_i, \mathcal{A}_{f,n})$ of size $k = O(n/\varepsilon)$ in $O((n/\varepsilon)\log(n/\varepsilon))$ time given basic knowledge of the distribution μ_i. We can then apply Theorem 2.2 to build an ε-quantization in time $O((kn)^{2+1} \log^4(n^2/\varepsilon)) = O(n^6/\varepsilon^3 \log^4(n/\varepsilon))$ for the dwid case.

We can actually evaluate $\tilde{f}(\mathcal{Q}, r)$ for all values of r faster using a series of sweep lines. Each of the $O(n^4/\varepsilon^2)$ potential bases are defined by a left and right end point. Each of the $O(n^2/\varepsilon)$ points in $\bigcup_i Q_i$ could be a left or right end point. We only need to find the $O(1/\varepsilon)$ widths (defined by pairs of end points) that wind up in the final ε-quantization. Using a Frederickson and Johnson approach [10], we can search for each width in $O(\log(n/\varepsilon))$ steps. At each step we are given a width ω and need to decide what fraction of supports have width at most ω. We can scan from each of the possible left end points and count the number of supports that have width at most ω. For each ω, this can be performed in $O(n^2/\varepsilon)$ time with a pair of simultaneous sweep lines. The total runtime is $O(1/\varepsilon) \cdot O(\log(n/\varepsilon)) \cdot O(n^2/\varepsilon) = O((n^2/\varepsilon^2)\log(n/\varepsilon))$.



(a) (b)

Figure 6: (a) Directional (vertical) width. (b) Axis-aligned bounding box, measured by area. The curves are hyperbola parts.

Theorem C.1. *We can create an ε-quantization of size $O(1/\varepsilon)$ for the dwid problem in $O((n^2/\varepsilon^2)\log(n/\varepsilon))$ time.*

C.2 Axis-aligned Bounding Box

We now consider the set of problems related to axis-aligned bounding boxes in $\mathbb{R}^d$. For a point set P, we minimize $f(P)$, which either represents the d-dimensional volume of $S(P)$ (the aabbv case — minimizes the area in $\mathbb{R}^2$) or the $(d-1)$-dimensional volume of the boundary of $S(P)$ (the aabbp case — minimizes the perimeter in $\mathbb{R}^2$). Figures 5(Left) and 6(b) show two examples of elements of $\mathcal{A}_{f,n}$ for the aabbp case and the aabbv case in $\mathbb{R}^2$. For both $(\mathbb{R}^2, \mathcal{A}_{f,n})$ has a shatter dimension of 4 because the shape is determined by the x-coordinates of 2 points and the y-coordinates of 2 points. This generalizes to a shatter dimension of $2d$ for $(\mathbb{R}^d, \mathcal{A}_{f,n})$, and hence a VC-dimension of $O(d \log d)$. The smaller VC-dimension in the aabbp case discussed in detail above can be extended to higher dimensions.

Hence, for $1 \leq i \leq n$, for both cases we can create an (ε/n)-sample Q_i of $(\mu_i, \mathcal{A}_{f,n})$, each of size $k = O((n^2/\varepsilon^2)\log(n/\varepsilon))$ in total time $O(((n/\varepsilon)\log(n/\varepsilon))^{O(d\log d)})$ via Corollary B.1. In $\mathbb{R}^d$, we can construct the ε-quantization in $O((kn)^{2d+1}\log^4(nk)) = O((n^{6d+3}/\varepsilon^{4d+2})\log^{2d+1}(n/\varepsilon))$ time via Theorem 2.2.

Theorem C.2. *We can create an ε-quantization of size $O(1/\varepsilon)$ for the aabbp or aabbv problem on n d-variate normal distributions in $O(((n/\varepsilon)\log(n/\varepsilon))^{O(d\log d)})$ time.*

C.3 Smallest Enclosing Ball

Figure 7 shows example elements of $\mathcal{A}_{f,n}$ for smallest enclosing ball, for metrics L_∞ (the seb_∞ case) and L_1 (the seb_1 case) in $\mathbb{R}^2$. An example element of $\mathcal{A}_{f,n}$ for smallest enclosing ball for the L_2 metric (the seb_2 case) was shown in Figure 5(Middle). For seb_∞ and seb_1, $(\mathbb{R}^d, \mathcal{A}_{f,n})$ has VC-dimension $2d$ because the shapes are defined by the intersection of halfspaces from d predefined normal directions. For seb_1 and seb_∞, we can create n (ε/n)-samples Q_i of each $(\mu_i, \mathcal{A}_{f,n})$ of size $k = O((n/\varepsilon)\log^{2d}(n/\varepsilon))$ in total time $O(n(n/\varepsilon)^{d+4}\log^{6k+3}(n/\varepsilon))$ via Theorem B.2. We can then create an ε-quantization in $O((nk)^{2d+1}\log^4(nk)) = O((n^{4d+2}/\varepsilon^{2d+1})\log^{4d^2+2d+4}(n/\varepsilon))$ time via Theorem 2.2.

Theorem C.3. *We can create an ε-quantization of size $O(1/\varepsilon)$ for the seb_1 or seb_∞ problem on n d-variate normal distributions in $O((n^{4d+2}/\varepsilon^{2d+1})\log^{4d^2+2d+4}(n/\varepsilon))$ time.*

We conjecture this technique can be extended to $\mathbb{R}^d$ for seb_2, but we cannot figure how to decompose a shape $A \in \mathcal{A}_{f,n}$ into a polynomial number of wedges with constant VC-dimension.

(a) (b)

Figure 7: (a) Smallest enclosing ball, L_∞ metric. (b) Smallest enclosing ball, L_1 metric.