Shape Fitting on Point Sets with Probability Distributions

Maarten Löffler[1] and Jeff M. Phillips[2]

[1] Department of Computer Science, Utrecht University
[2] Department of Computer Science, Duke University

Abstract. We consider problems on data sets where each data point has uncertainty described by an individual probability distribution. We develop several frameworks and algorithms for calculating statistics on these uncertain data sets. Our examples focus on geometric shape fitting problems. We prove approximation guarantees for the algorithms with respect to the full probability distributions. We then empirically demonstrate that our algorithms are simple and practical, solving for a constant hidden by asymptotic analysis so that a user can reliably trade speed and size for accuracy.

1 Introduction

In gathering data there is a trade-off between quantity and accuracy. The drop in the price of hard drives and other storage costs has shifted this balance towards gathering enormous quantities of data, yet with noticeable and sometimes intentional imprecision. However, often as a benefit from the large data sets, models are developed to describe the pattern of the data error.

For instance, in the gathering of LIDAR data for GIS applications [17], each data point of a terrain can have error in its x- (longitude), y- (latitude) and z-coordinates (height). Greatly simplifying, we could model the uncertainty as a 3-variate normal distribution centered at its recorded value. Similarly, large data sets are gathered with uncertainty in robotic mapping [12], anonymized medical data [1], spatial databases [23], sensor networks [17], and many other areas.

However, much raw data is not immediately given as a set of probability distributions, rather as a set of points. Approximate algorithms may treat this data as exact, construct an approximate answer, and then postulate that since the raw data is not exact, the approximation errors made by the algorithm may be similar to the errors of the imprecise input data. This is a very dangerous postulation.

An algorithm can only provide answers as good as the raw data *and* the models for error on that data. This paper is not about how to construct error models, but how to take error models into account. While many existing algorithms produce approximations with respect only to the raw input data, algorithms in this paper approximate with respect to the raw input data *and* the error models associated with them.

Geometric error models. An early model for imprecise geometric data, motivated by finite precision of coordinates, is ε-*geometry* [14], where each data point is known to lie within a ball of radius ε. This models has been used to study the robustness of problems such as the Delaunay triangulation [6, 18]. This model has been extended to allow different uncertainty regions around each point for object intersection [21] and shape-fitting problems [24]. These approaches give worst case bounds on error, for instance upper and lower bounds on the radius of the minimum enclosing ball. But when uncertainty is given as a probability distribution, then these approaches must use a threshold to truncate the distribution. Furthermore, the answers in this model are quite dependent on the boundary of the uncertainty region, while the true location is likely to be in the interior. This paper thus describes how to use the full probability distribution describing the uncertainty, and to only discretize, as desired, the probability distribution of the final solution.

The database community has focused on similar problems for usually one-dimensional data such as indexing [2], ranking [11], and creating histograms [10].

1.1 Problem Statement

Let $\mu_p : \mathbb{R}^d \to \mathbb{R}^+$ describe the probability distribution of a point p where the integral $\int_{q \in \mathbb{R}^d} \mu_p(q)\ dq = 1$. Let $\mu_P : \mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d \to \mathbb{R}^+$ describe the distribution of a point set P by the joint probability over each $p \in P$. For brevity we write the space $\mathbb{R}^d \times \ldots \times \mathbb{R}^d$ as $\mathbb{R}^{dn}$. For this paper we will assume $\mu_P(q_1, q_2, \ldots, q_n) = \prod_{i=1}^n \mu_{p_i}(q_i)$, so the distribution for each point is independent, although this restriction can be easily circumvented.

Given a distribution μ_P we ask a variety of shape fitting questions about the uncertain point set. For instance, what is the radius of the smallest enclosing ball or what is the smallest axis-aligned bounding box of an uncertain point set. In the presence of imprecision, the answer to such a question is not a single value or structure, but also a *distribution* of answers. The focus of this paper is not just how to answer such shape fitting questions about these distributions, but how to concisely represent them. As a result, we introduce two types of approximate distributions as answers, and a technique to construct coresets for these answers.

ε-**Quantizations.** Let $f : \mathbb{R}^{dn} \to \mathbb{R}^k$ be a function on a fixed point set. Examples include the radius of the minimum enclosing ball where $k = 1$ and the width of the minimum enclosing axis-aligned rectangle along the x-axis and y-axis where $k = 2$. Define the "dominates" binary operator $\preceq$ so that $(p_1, \ldots, p_k) \preceq (v_1, \ldots, v_k)$ is true if for every coordinate $p_i \leq v_i$. Let $\mathbb{X}_f(v) = \{Q \in \mathbb{R}^{dn} \mid f(Q) \preceq v\}$. For a query value v define, $F_{\mu_P}(v) = \int_{Q \in \mathbb{X}_f(v)} \mu_P(Q)\, dQ$. Then F_{μ_P} is the cumulative density function of the distribution of possible values that f can take[1]. Ideally, we would return the function F_{μ_P} so we could quickly answer any query exactly, however, it is not clear how to calculate $F_{\mu_P}(v)$ exactly for

[1] For a function f and a distribution of point sets μ_P, we will always represent the cumulative density function of f over μ_P by F_{μ_P}.

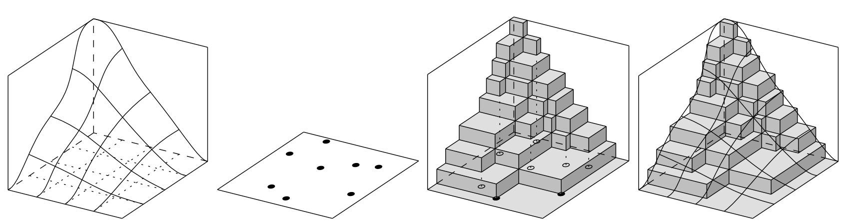
(a) (b) (c)

Fig. 1. (a) The true form of a function from $\mathbb{R} \to \mathbb{R}$. (b) The ε-quantization R as a point set in $\mathbb{R}$. (c) The inferred curve h_R in $\mathbb{R}^2$.

even a single query value v. Rather, we introduce a data structure, which we call an ε-quantization, to answer any such query approximately and efficiently, illustrated in Figure 1 for $k = 1$. An *ε-quantization* is a point set $R \subset \mathbb{R}^k$ which induces a function h_R where $h_R(v)$ describes the fraction of points in R that v dominates. Let $R_v = \{r \in R \mid r \preceq v\}$. Then $h_R(v) = |R_v|/|R|$. For an isotonic (monotonically increasing in each coordinate) function F_{μ_P} and any value v, an ε-quantization, R, guarantees that $|h_R(v) - F_{\mu_P}(v)| \le \varepsilon$. More generally (and, for brevity, usually only when $k > 1$), we say R is a k-variate ε-quantization. A 2-variate ε-quantization is illustrated in Figure 2. The space required to store the data structure for R is dependent only on ε and k, not on $|P|$ or μ_P.

$(\varepsilon, \delta, \alpha)$-Kernels. Rather than compute a new data structure for each measure we are interested in, we can also compute a single data structure (a coreset) that allows us to answer many types of questions. For an isotonic function $F_{\mu_P} : \mathbb{R}^+ \to [0,1]$, an (ε, α)-*quantization* data structure M describes a function $h_M : \mathbb{R}^+ \to [0,1]$ so for any $x \in \mathbb{R}^+$, there is an $x' \in \mathbb{R}^+$ such that (1) $|x - x'| \le \alpha x$ and (2) $|h_M(x) - F_{\mu_P}(x')| \le \varepsilon$. An $(\varepsilon, \delta, \alpha)$-*kernel* is a data structure that can produce an (ε, α)-quantization, with probability at least $1 - \delta$, for F_{μ_P} where f measures the width in any direction and whose size depends only on ε, α, and δ. The notion of (ε, α)-quantizations is generalized to a k-variate version, as are $(\varepsilon, \delta, \alpha)$-kernels.

Shape inclusion probabilities. A *summarizing shape* of a point set $P \subset \mathbb{R}^d$ is a Lebesgue-measureable subset of $\mathbb{R}^d$ that is determined by P. I.e. given a class of shapes $\mathcal{S}$, the summarizing shape $S(P) \in \mathcal{S}$ is the shape that optimizes some aspect with respect to P. Examples include the smallest enclosing ball and the minimum-area axis-aligned bounding rectangle. For a family $\mathcal{S}$ we can study



(a) (b) (c) (d)

Fig. 2. (a) The true form of a 2-variate function. (b) The ε-quantization R as a point set in $\mathbb{R}^2$. (c) The inferred surface h_R in $\mathbb{R}^3$. (d) Overlay of the two images.

the *shape inclusion probability function* $s_{\mu_P} : \mathbb{R}^d \to [0,1]$ (or sip function), where $s_{\mu_P}(q)$ describes the probability that a query point $q \in \mathbb{R}^d$ is included in the summarizing shape[2]. There does not seem to be a closed form for many of these functions. Rather we calculate an ε-sip function $\hat{s} : \mathbb{R}^d \to [0,1]$ such that $\forall_{q\in\mathbb{R}^d} |s_{\mu_P}(q) - \hat{s}(q)| \le \varepsilon$. The space required to store an ε-sip function depends only on ε and the complexity of the summarizing shape.

1.2 Contributions

We describe simple and practical randomized algorithms for the computation of ε-quantizations, $(\varepsilon, \delta, \alpha)$-kernels, and ε-sip functions. Let $T_f(n)$ be the time it takes to calculate a summarizing shape of a set of n points $Q \subset \mathbb{R}^d$, which generates a statistic $f(Q)$ (e.g., radius of smallest enclosing ball). We can calculate an ε-quantization of F_{μ_P}, with probability at least $1-\delta$, in $O(T_f(n)(1/\varepsilon^2)\log(1/\delta))$ time. For univariate ε-quantizations the size is $O(1/\varepsilon)$, and for k-variate ε-quantizations the size is $O(k^2(1/\varepsilon)\log^{2k}(1/\varepsilon))$. We can calculate an $(\varepsilon, \delta, \alpha)$-kernel of size $O((1/\alpha^{(d-1)/2}) \cdot (1/\varepsilon^2)\log(1/\delta))$ in time $O((n + (1/\alpha^{d-3/2}))(1/\varepsilon^2) \cdot \log(1/\delta))$. With probability at least $1-\delta$, we can calculate an ε-sip function of size $O((1/\varepsilon^2)\log(1/\delta))$ in time $O(T_f(n)(1/\varepsilon^2)\log(1/\delta))$.

All of these randomized algorithms are simple and practical, as demonstrated by a series of experimental results. In particular, we show that the constant hidden by the big-O notation is in practice at most 0.5 for all algorithms.

1.3 Preliminaries: ε-Samples and α-Kernels

ε-Samples. For a set P let $\mathcal{A}$ be a set of subsets of P. In our context usually P will be a point set and the subsets in $\mathcal{A}$ could be induced by containment in a shape from some family of geometric shapes. For example of $\mathcal{A}$, $\mathcal{I}_+$ describes one-sided intervals of the form $(-\infty, t)$. The pair $(P, \mathcal{A})$ is called a *range space*. We say that $Q \subset P$ is an *ε-sample* of $(P, \mathcal{A})$ if

$$\forall_{R\in\mathcal{A}} \left| \frac{\phi(R\cap Q)}{\phi(Q)} - \frac{\phi(R\cap P)}{\phi(P)} \right| \le \varepsilon,$$

where $|\cdot|$ takes the absolute value and $\phi(\cdot)$ returns the measure of a point set. In the discrete case $\phi(Q)$ returns the cardinality of Q. We say $\mathcal{A}$ *shatters* a set S if every subset of S is equal to $R \cap S$ for some $R \in \mathcal{A}$. The cardinality of the largest discrete set $S \subseteq P$ that $\mathcal{A}$ can shatter is the *VC-dimension* of $(P, \mathcal{A})$.

When $(P, \mathcal{A})$ has constant VC-dimension ν, we can create an ε-sample Q of $(P, \mathcal{A})$, with probability $1-\delta$, by uniformly sampling $O((1/\varepsilon^2)(\nu + \log(1/\delta)))$ points from P [25, 16]. There exist deterministic techniques to create ε-samples [19, 9] of size $O(\nu(1/\varepsilon^2)\log(1/\varepsilon))$ in time $O(\nu^{3\nu}n((1/\varepsilon^2)\log(\nu/\varepsilon))^\nu)$. When P is a

[2] For technical reasons, if there are (degenerately) multiple optimal summarizing shapes, we say each is equally likely to be the summarizing shape of the point set.

point set in $\mathbb{R}^d$ and the family of ranges $\mathcal{R}_d$ is determined by inclusion in axis-aligned boxes, then an ε-sample for $(P, \mathcal{R}_d)$ of size $O((d/\varepsilon)\log^{2d}(1/\varepsilon))$ can be constructed in $O((n/\varepsilon^3)\log^{6d}(1/\varepsilon))$ time [22].

For a range space $(P, \mathcal{A})$ the *dual range space* is defined $(\mathcal{A}, P^*)$ where P^* is all subsets $\mathcal{A}_p \subseteq \mathcal{A}$ defined for an element $p \in P$ such that $\mathcal{A}_p = \{A \in \mathcal{A} \mid p \in A\}$. If $(P, \mathcal{A})$ has VC-dimension ν, then $(\mathcal{A}, P^*)$ has VC-dimension $\leq 2^{\nu+1}$. Thus, if the VC-dimension of $(\mathcal{A}, P^*)$ is constant, then the VC-dimension of $(P, \mathcal{A})$ is also constant [20].

When we have a distribution $\mu : \mathbb{R}^d \to \mathbb{R}^+$, such that $\int_{x \in \mathbb{R}} \mu(x)\, dx = 1$, we can think of this as the set P of all points in $\mathbb{R}^d$, where the weight w of a point $p \in \mathbb{R}^d$ is $\mu(p)$. To simplify notation, we write $(\mu, \mathcal{A})$ as a range space where the ground set is this set $P = \mathbb{R}^d$ weighted by the distribution μ.

α-Kernels. Given a point set $P \in \mathbb{R}^d$ of size n and a direction $u \in \mathbb{S}^{d-1}$, let $P[u] = \arg\max_{p \in P} \langle p, u\rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product operator. Let $\omega(P, u) = \langle P[u] - P[-u], u\rangle$ describe the width of P in direction u. We say that $K \subseteq P$ is an *α-kernel* of P if for all $u \in \mathbb{S}^{d-1}$

$$\omega(P, u) - \omega(K, u) \leq \alpha \cdot \omega(P, u).$$

α-kernels of size $O(1/\alpha^{(d-1)/2})$ [4] can be calculated in time $O(n + 1/\alpha^{d-3/2})$ [8, 26]. Computing many extent related problems such as diameter and smallest enclosing ball on K approximates the problem on P [4, 3, 8].

2 Randomized Algorithm for ε-Quantizations

We develop several algorithms with the following basic structure: (1) sample one point from each distribution to get a random point set; (2) construct the summarizing shape of the random point set; (3) repeat the first two steps $O((1/\varepsilon)(\nu + \log(1/\delta)))$ times and calculate a summary data structure. This algorithm only assumes that we can draw a random point from μ_p for each $p \in P$.

2.1 Algorithm for ε-Quantizations

For a function f on a point set P of size n, it takes $T_f(n)$ time to evaluate $f(P)$. We construct an approximation to F_{μ_P} as follows. First draw a sample point q_j from each μ_{p_j} for $p_j \in P$, then evaluate $V_i = f(\{q_1, \ldots, q_n\})$. The fraction of trials of this process that produces a value dominated by v is the estimate of $F_{\mu_P}(v)$. In the univariate case we can reduce the size of $\mathcal{V}$ by returning $2/\varepsilon$ evenly spaced points according to the sorted order.

Theorem 1. *For a distribution μ_P of n points, with success probability at least $1 - \delta$, there exists an ε-quantization of size $O(1/\varepsilon)$ for F_{μ_P}, and it can be constructed in $O(T_f(n)(1/\varepsilon^2)\log(1/\delta))$ time.*

Proof. Because $F_{\mu_P} : \mathbb{R} \to [0,1]$ is an isotonic function, there exists another function $g : \mathbb{R} \to \mathbb{R}^+$ such that $F_{\mu_P}(t) = \int_{x=-\infty}^{t} g(x)\,dx$ where $\int_{x\in\mathbb{R}} g(x)\,dx = 1$. Thus g is a probability distribution of the values of f given inputs drawn from μ_P. This implies that an ε-sample of $(g, \mathcal{I}_+)$ is an ε-quantization of F_{μ_P}, since both estimate within ε the fraction of points in any range of the form $(-\infty, x)$.

By drawing a random sample q_i from each μ_{p_i} for $p_i \in P$, we are drawing a random point set Q from μ_P. Thus $f(Q)$ is a random sample from g. Hence, using the standard randomized construction for ε-samples, $O((1/\varepsilon^2)\log(1/\delta))$ such samples will generate an $(\varepsilon/2)$-sample for g, and hence an $(\varepsilon/2)$-quantization for F_{μ_P}, with probability at least $1-\delta$.

Since in an $(\varepsilon/2)$-quantization R every value $h_R(v)$ is different from $F_{\mu_P}(v)$ by at most $\varepsilon/2$, then we can take an $(\varepsilon/2)$-quantization of the function described by $h_R(\cdot)$ and still have an ε-quantization of F_{μ_P}. Thus, we can reduce this to an ε-quantization of size $O(1/\varepsilon)$ by taking a subset of $2/\varepsilon$ points spaced evenly according to their sorted order.

We can construct k-variate ε-quantizations similarly. The output V_i of f is now k-variate and thus results in a k-dimensional point.

Theorem 2. *Given a distribution μ_P of n points, with success probability at least $1-\delta$, we can construct a k-variate ε-quantization for F_{μ_P} of size $O((k/\varepsilon^2)(k+\log(1/\delta)))$ and in time $O(T_f(n)(1/\varepsilon^2)(k+\log(1/\delta)))$.*

Proof. Let $\mathcal{R}_+$ describe the family of ranges where a range $A_p = \{q \in \mathbb{R}^k \mid q \preceq p\}$. In the k-variate case there exists a function $g : \mathbb{R}^k \to \mathbb{R}^+$ such that $F_{\mu_P}(v) = \int_{x\preceq v} g(x)\,dx$ where $\int_{x\in\mathbb{R}^k} g(x)\,dx = 1$. Thus g describes the probability distribution of the values of f, given inputs drawn randomly from μ_P. Hence a random point set Q from μ_P, evaluated as $f(Q)$, is still a random sample from the k-variate distribution described by g. Thus, with probability at least $1-\delta$, a set of $O((1/\varepsilon^2)(k+\log(1/\delta)))$ such samples is an ε-sample of $(g, \mathcal{R}_+)$, which has VC-dimension k, and the samples are also a k-variate ε-quantization of F_{μ_P}.

We can then reduce the size of the ε-quantization R to $O((k^2/\varepsilon)\log^{2k}(1/\varepsilon))$ in $O(|R|(k/\varepsilon^3)\log^{6k}(1/\varepsilon))$ time [22] or to $O((k^2/\varepsilon^2)\log(1/\varepsilon))$ in $O(|R|(k^{3k}/\varepsilon^{2k})\cdot \log^k(k/\varepsilon))$ time [9], since the VC-dimension is k and each data point requires $O(k)$ storage.

2.2 $(\varepsilon, \delta, \alpha)$-Kernels

The above construction works for a fixed family of summarizing shapes. This section builds a single data structure, an $(\varepsilon, \delta, \alpha)$-kernel, for a distribution μ_P in $\mathbb{R}^d$ that can be used to construct (ε, α)-quantizations for several families of summarizing shapes. In particular, an $(\varepsilon, \delta, \alpha)$-kernel of μ_P is a data structure such that in any query direction $u \in \mathbb{S}^{d-1}$, with probability at least $1-\delta$, we can create an (ε, α)-quantization for the cumulative density function of $\omega(\cdot, u)$, the width in direction u.

We follow the randomized framework described above as follows. The desired $(\varepsilon, \delta, \alpha)$-kernel $\mathcal{K}$ consists of a set of $m = O((1/\varepsilon^2)\log(1/\delta))$ $(\alpha/2)$-kernels, $\{K_1, K_2, \ldots, K_m\}$, where each K_j is an $(\alpha/2)$-kernel of a point set Q_j drawn randomly from μ_P. Given $\mathcal{K}$, with probability at least $1-\delta$, we can create an (ε, α)-quantization for the cumulative density function of width over μ_P in any direction $u \in \mathbb{S}^{d-1}$. Specifically, let $M = \{\omega(K_j, u)\}_{j=1}^m$.

Lemma 1. *With probability at least $1-\delta$, M is an (ε, α)-quantization for the cumulative density function of the width of μ_P in direction u.*

Proof. The width $\omega(Q_j, u)$ of a random point set Q_j drawn from μ_P is a random sample from the distribution over widths of μ_P in direction u. Thus, with probability at least $1-\delta$, m such random samples would create an ε-quantization. Using the width of the α-kernels K_j instead of Q_j induces an error on each random sample of at most $2\alpha \cdot \omega(Q_j, u)$. Then for a query width w, say there are γm point sets Q_j that have width at most w and $\gamma' m$ α-kernels K_j with width at most w. Note that $\gamma' > \gamma$. Let $\hat{w} = w - 2\alpha w$. For each point set Q_j that has width greater than w it follows that K_j has width greater than $\hat{w}$. Thus the number of α-kernels K_j that have width at most $\hat{w}$ is at most γm, and thus there is a width w' between w and $\hat{w}$ such that the number of α-kernels at most w' is exactly γm.

Since each K_j can be computed in $O(n + 1/\alpha^{d-3/2})$ time, we obtain:

Theorem 3. *We can construct an $(\varepsilon, \delta, \alpha)$-kernel for μ_P on n points in $\mathbb{R}^d$ of size $O((1/\alpha^{(d-1)/2})(1/\varepsilon^2)\cdot\log(1/\delta))$ and in time $O((n+1/\alpha^{d-3/2})\cdot(1/\varepsilon^2)\log(1/\delta))$.*

The notion of (ε, α)-quantizations and $(\varepsilon, \delta, \alpha)$-kernels can be extended to k-dimensional queries or for a series of up to k queries which all have approximation guarantees with probability $1-\delta$.

In a similar fashion, coresets of a point set distribution μ_P can be formed using coresets for other problems on discrete point sets. For instance, sample $m = O((1/\varepsilon^2)\log(1/\delta))$ points sets $\{P_1, \ldots, P_m\}$ each from μ_P and then store α-samples $\{Q_1 \subseteq P_1, \ldots, Q_m \subseteq P_m\}$ of each. This results in an $(\varepsilon, \delta, \alpha)$-sample of μ_P, and can, for example, be used to construct (with probability $1-\delta$) an (ε, α)-quantization for the fraction of points expected to fall in a query disk. Similar constructions can be done for other coresets, such as ε-nets [15], k-center [5], or smallest enclosing ball [7].

2.3 Shape Inclusion Probabilities

For a point set $Q \subset \mathbb{R}^d$, let the summarizing shape $S_Q = \mathcal{S}(Q)$ be from some geometric family $\mathcal{S}$ so $(\mathbb{R}^d, \mathcal{S})$ has bounded VC-dimension ν. We randomly sample m point sets $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ each from μ_P and then find the summarizing shape $S_{Q_j} = \mathcal{S}(Q_j)$ (e.g. minimum enclosing ball) of each Q_j. Let this set of shapes be $S^{\mathcal{Q}}$. If there are multiple shapes from $\mathcal{S}$ which are equally optimal choose one of these shapes at random. For a set of shapes $S' \subseteq \mathcal{S}$, let $S'_p \subseteq S'$

be the subset of shapes that contain $p \in \mathbb{R}^d$. We store S^Ω and evaluate a query point $p \in \mathbb{R}^d$ by counting what fraction of the shapes the point is contained in, specifically returning $|S^\Omega_p|/|S^\Omega|$ in $O(\nu|S^\Omega|)$ time. In some cases, this evaluation can be sped up with point location data structures.

Theorem 4. *Consider a family of summarizing shapes $\mathcal{S}$ where $(\mathbb{R}^d, \mathcal{S})$ has VC-dimension ν and where it takes $T_\mathcal{S}(n)$ time to determine the summarizing shape $\mathcal{S}(Q)$ for any point set $Q \subset \mathbb{R}^d$ of size n. For a distribution μ_P of a point set of size n, with probability at least $1-\delta$, we can construct an ε-sip function of size $O((\nu/\varepsilon^2)(2^{\nu+1} + \log(1/\delta)))$ and in time $O(T_\mathcal{S}(n)(1/\varepsilon^2)\log(1/\delta))$.*

Proof. If $(\mathbb{R}^d, \mathcal{S})$ has VC-dimension ν, then the dual range space $(\mathcal{S}, P^*)$ has VC-dimension $\nu' \leq 2^{\nu+1}$, where P^* is all subsets $\mathcal{S}_p \subseteq \mathcal{S}$, for any $p \in \mathbb{R}^d$, such that $\mathcal{S}_p = \{S \in \mathcal{S} \mid p \in S\}$. Using the above algorithm, sample $m = O((1/\varepsilon^2)(\nu' + \log(1/\delta)))$ point sets Q from μ_P and generate the m summarizing shapes S_Q. Each shape is a random sample from $\mathcal{S}$ according to μ_P, and thus S^Ω is an ε-sample of $(\mathcal{S}, P^*)$.

Let $w_{\mu_P}(S)$, for $S \in \mathcal{S}$, be the probability that S is the summarizing shape of a point set Q drawn randomly from μ_P. For any $\mathcal{S}' \subseteq P^*$, let $W_{\mu_P}(\mathcal{S}') = \int_{S \in \mathcal{S}'} w_{\mu_P}(S)\, dS$ be the probability that some shape from the subset $\mathcal{S}'$ is the summarizing shape of Q drawn from μ_P.

We approximate the sip function at $p \in \mathbb{R}^d$ by returning the fraction $|S^\Omega_p|/m$. The true answer to the sip function at $p \in \mathbb{R}^d$ is $W_{\mu_P}(\mathcal{S}_p)$. Since S^Ω is an ε-sample of $(\mathcal{S}, P^*)$, then with probability at least $1-\delta$

$$\left| \frac{|S^\Omega_p|}{m} - \frac{W_{\mu_P}(\mathcal{S}_p)}{1} \right| = \left| \frac{|S^\Omega_p|}{|S^\Omega|} - \frac{W_{\mu_P}(\mathcal{S}_p)}{W_{\mu_P}(P^*)} \right| \leq \varepsilon.$$

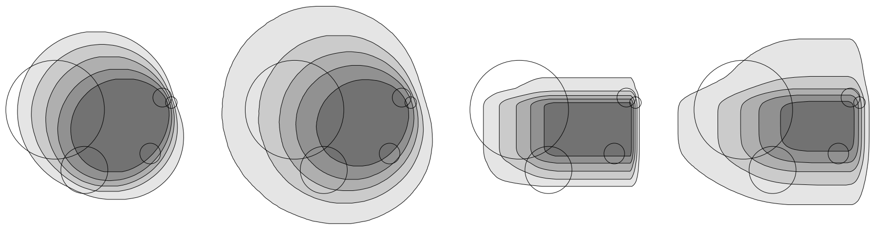
Since for the family of summarizing shapes $\mathcal{S}$ the range space $(\mathbb{R}^d, \mathcal{S})$ has VC-dimension ν, each can be stored using that much space.

Using deterministic techniques [9] the size can then be reduced to $O(2^{\nu+1}(\nu/\varepsilon^2) \cdot \log(1/\varepsilon))$ in time $O((2^{3(\nu+1)} \cdot (\nu/\varepsilon^2)\log(1/\varepsilon))^{2^{\nu+1}} \cdot 2^{3(\nu+1)}(\nu/\varepsilon^2)\log(1/\delta))$.

Representing ε-sip functions by isolines. Shape inclusion probability functions are density functions. A convenient way of visually representing a density function in $\mathbb{R}^2$ is by drawing the isolines. A γ-*isoline* is a collection of closed curves bounding a region of the plane where the density function is greater than γ.

In each part of Figure 3 a set of 5 circles correspond to points with a probability distribution. In part (a,c) the probability distribution is uniform over the inside of the circles. In part (b,d) it is drawn from a normal distribution with standard deviation the radius. We generate ε-sip functions for smallest enclosing ball in Figure 3(a,b) and for smallest axis-aligned rectangle in Figure 3(c,d).

In all figures we draw approximations of $\{.9, .7, .5, .3, .1\}$-isolines. These drawing are generated by randomly selecting $m = 5000$ (Figure 3(a,b)) or $m = 25000$ (Figure 3(c,d)) shapes, counting the number of inclusions at different points in



(a) (b) (c) (d)

Fig. 3. The sip for the smallest enclosing ball (a,b) or smallest enclosing axis-aligned rectangle (c,d), for uniformly (a,c) or normally (b,d) distributed points.

the plane and interpolating to get the isolines. The innermost and darkest region has probability $> 90\%$, the next one probability $> 70\%$, etc., the outermost region has probability $< 10\%$.

3 Measuring the Error

We have established asymptotic bounds of $O((1/\varepsilon^2)(\nu + \log(1/\delta))$ random samples for constructing ε-quantizations and ε-sip functions. In this section we empirically demonstrate that the constant hidden by the big-O notation is approximately 0.5, indicating that these algorithms are indeed quite practical. Additionally, we show that we can reduce the size of ε-quantizations to $2/\varepsilon$ without sacrificing accuracy and with only a factor 4 increase in the runtime. We also briefly compare the (ε, α)-quantizations produced with $(\varepsilon, \delta, \alpha)$-kernels to ε-quantizations. We show that the $(\varepsilon, \delta, \alpha)$-kernels become useful when the number of uncertain points becomes large, i.e. exceeding 1000.

Univariate ε-quantizatons. We consider a set of $n = 50$ points samples in $\mathbb{R}^3$ chosen randomly from the boundary of a cylinder piece of length 10 and radius 1. We let each point represent the center of 3-variate Gaussian distribution with standard deviation 2 to represent the probability distribution of an uncertain point. This set of distributions describes an uncertain point set $\mu_P : \mathbb{R}^{3n} \to \mathbb{R}^+$.

We want to estimate three statistics on μ_P: dwid, the width of the points set in a direction that makes an angle of 75° with the cylinder axis; diam, the diameter of the point set; and seb_2, the radius of the smallest enclosing ball (using code from Bernd Gärtner [13]). We can create ε-quantizations with m samples from μ_P, where the value of m is from the set $\{16, 64, 256, 1024, 4096\}$.

We would like to evaluate the ε-quantizations versus the ground truth function F_{μ_P}; however, it is not clear how to evaluate F_{μ_P}. Instead, we create another ε-quantization Q with $\eta = 100000$ samples from μ_P, and treat this as if it were the ground truth. To evaluate each sample ε-quantization R versus Q we find the maximum deviation (i.e. $d_\infty(R, Q) = \max_{q \in \mathbb{R}} |h_R(q) - h_Q(q)|$) with h defined with respect to diam or dwid. This can be done by for each value $r \in R$ evaluating $|h_R(r) - h_Q(r)|$ and $|(h_R(r) - 1/|R|) - h_Q(r)|$ and returning the maximum of both values over all $r \in R$.

Given a fixed "ground truth" quantization Q we repeat this process for $\tau = 500$ trials of R, each returning a $d_\infty(R,Q)$ value. The set of these τ maximum deviations values results in another quantization S for each of diam and dwid, plotted in Figure 4. Intuitively, the maximum deviation quantization S describes the sample probability that $d_\infty(R,Q)$ will be less than some query value.

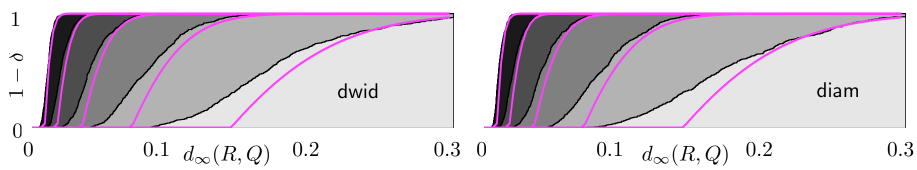


Fig. 4. Shows quantizations of $\tau = 500$ trials for $d_\infty(R,Q)$ where Q and R measure dwid and diam. The size of each R is $m = \{16, 64, 256, 1024, 4096\}$ (from right to left) and the "ground truth" quantization Q has size $\eta = 100000$. Smooth, thick curves are $1-\delta = 1-\exp(-2m\varepsilon^2+1)$ where $\varepsilon = d_\infty(R,Q)$.

Note that the maximum deviation quantizations S are similar for both statistics (and others we tried), and thus we can use these plots to estimate $1-\delta$, the sample probability that $d_\infty(R,Q) \leq \varepsilon$, given a value m. We can fit this function as approximately $1-\delta = 1-\exp(-m\varepsilon^2/C+\nu)$ with $C = 0.5$ and $\nu = 1.0$. Thus solving for m in terms of ε, ν, and δ reveals: $m = C(1/\varepsilon^2)(\nu+\log(1/\delta))$. This indicates the big-O notation for the asymptotic bound of $O((1/\varepsilon^2)(\nu+\log(1/\delta))$ [16] for ε-samples only hides a constant of approximately 0.5.

Maximum error in sip functions. We can perform a similar analysis on sip functions. We use the same input data as is used to generate Figure 3(b) and create sip functions R for the smallest enclosing ball using $m = \{16, 36, 81, 182, 410\}$ samples from μ_P. We compare this to a "ground truth" sip function Q formed using $\eta = 5000$ sampled points. The maximum deviation between R and Q in this context is defined $d_\infty(R,Q) = \max_{q\in\mathbb{R}^2} |R(q) - Q(q)|$ and can be found by calculating $|R(q) - Q(q)|$ for all points $q \in \mathbb{R}^2$ at the intersection of boundaries of discs from R or Q.

We repeat this process for $\tau = 100$ trials, for each value of m. This creates a quantization S (for each value of m) measuring the maximum deviation for the sip functions. These maximum deviation quantizations are plotted in Figure 5. We fit these curves with a function $1-\delta = 1-\exp(-m\varepsilon^2/C+\nu)$ with $C = 0.5$ and $\nu = 2.0$, so $m = C(1/\varepsilon^2)(\nu + \log 1/\delta)$. Note that the dual range space $(\mathcal{B}, \mathbb{R}^{2*})$, defined by disks $\mathcal{B}$ has VC-dimension 2, so this is exactly what we would expect.

Maximum error in k-variate quantizations. We extend these experiments to k-variate quantizations by considering the width in k different directions. As expected, the quantizations for maximum deviation can be fit with an equation $1-\delta = 1-\exp(-m\varepsilon^2/C+k)$ with $C = 0.5$, so $m \leq C(1/\varepsilon^2)(k+\log 1/\delta)$. For $k > 2$, this bound for m becomes too conservative. Figures omitted for space.

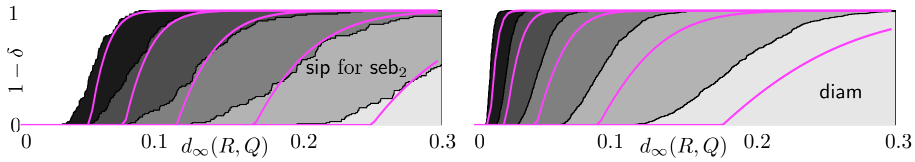


Fig. 5. Left: Quantization of $\tau = 100$ trials of maximum deviation between sip functions for smallest enclosing disc with $m = \{16, 36, 81, 182, 410\}$ (from right to left) sample shapes versus a "ground truth" sip function with $\eta = 5000$ sample shapes. Right: Quantization of $\tau = 500$ trials for $d_\infty(R, Q)$ where Q and R measure diam. Size of each R is $m = \{64, 256, 1024, 4096, 16384\}$, then compressed to size $\{8, 16, 32, 64, 128\}$ (resp., from right to left) and the "ground truth" quantization Q has size $\eta = 100000$.

3.1 Compressing ε-Quantizations

Theorem 1 describes how to compress the size of a univariate ε-quantization to $O(1/\varepsilon)$. We first create an $(\varepsilon/2)$-quantization of size m, then sort the values V_i, and finally take every $(m\varepsilon/2)$th value according to the sorted order. This returns an ε-quantization of size $2/\varepsilon$ and requires creating an initial ε-quantization with 4 times as many samples as we would have without this compression. The results, shown in Figure 5 using the same setup as in Figure 4, confirms that this compression scheme works better than the worst case claims. We only show the plot for diam, but the results for dwid and seb$_2$ are nearly identical. In particular, the error is smaller than the results in Figure 4, but it takes 4 times as long.

3.2 $(\varepsilon, \delta, \alpha)$-Kernels versus ε-Quantizations

We compare $(\varepsilon, \delta, \alpha)$-kernels to with ε-quantizations for diam, dwid, and seb$_2$, using code from Hai Yu [26] for α-kernels. Using the same setup as in Figure 4 with $n = 5000$ input points, we set $\varepsilon = 0.2$ and $\delta = 0.1$, resulting in $m = 40$ point sets sampled from μ_P. We also generated α-kernels of size at most 40. The $(\varepsilon, \delta, \alpha)$-kernel has a total of 1338 points. We calculated ε-quantizations and (ε, α)-quantizations for diam, dwid, and seb$_2$, each compressed to a size 10 shown in Figure 6. This method starts becoming useful in compressing μ_P when $n \gg 1000$ (otherwise the total size of the $(\varepsilon, \delta, \alpha)$-kernel may be larger than μ_P) or if computing $f_{\mathcal{S}}$ is very expensive.

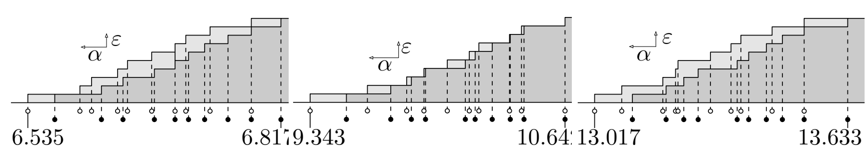


Fig. 6. (ε, α)-quantization (white points) and ε-quantization (black points) for (left) seb$_2$, (center) dwid, and (right) diam.

Acknowledgements. Thanks to Pankaj K. Agarwal for many helpful discussions.

References

1. Charu C. Agarwal and Philip S. Yu, editors. *Privacy Preserving Data Mining: Models and Algorithms*. Springer, 2008.
2. Pankaj K. Agarwal, Siu-Wing Cheng, Yufei Tao, and Ke Yi. Indexing uncertain data. In *PODS*, 2009.
3. Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi Varadarajan. Geometric approximations via coresets. *C. Trends Comb. and Comp. Geom. (E. Welzl)*, 2007.
4. Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Approximating extent measure of points. *J. ACM*, 51(4):2004, 2004.
5. Pankaj K. Agarwal, Cecilia M. Procopiuc, and Kasturi R. Varadarajan. Approximation algorithms for k-line center. In *ESA*, pages 54–63, 2002.
6. Deepak Bandyopadhyay and Jack Snoeyink. Almost-Delaunay simplices: Nearest neighbor relations for imprecise points. In *SODA*, pages 403–412, 2004.
7. Mihai Bădoiu and Ken Clarkson. Smaller core-sets for balls. In *SODA*, 2003.
8. Timothy Chan. Faster core-set constructions and data-stream algorithms in fixed dimensions. *Computational Geometry: Theory and Applications*, 35:20–35, 2006.
9. Bernard Chazelle and Jiri Matousek. On linear-time deterministic algorithms for optimization problems in fixed dimensions. *J. Algorithms*, 21:579–597, 1996.
10. Graham Cormode and Minos Garafalakis. Histograms and wavelets of probabilitic data. In *ICDE*, 2009.
11. Graham Cormode, Feifie Li, and Ke Yi. Semantics of ranking queries for probabilistic data and expected ranks. In *ICDE*, 2009.
12. Austin Eliazar and Ronald Parr. Dp-slam 2.0. In *ICRA*, 2004.
13. Bernd Gärtner. Fast and robust smallest enclosing balls. In *ESA*, 1999.
14. Leonidas J. Guibas, D. Salesin, and J. Stolfi. Epsilon geometry: building robust algorithms from imprecise computations. In *SoCG*, pages 208–217, 1989.
15. David Haussler and Emo Welzl. epsilon-nets and simplex range queries. *Disc. & Comp. Geom.*, 2:127–151, 1987.
16. Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the samples complexity of learning. *J. Comp. and Sys. Sci.*, 62:516–527, 2001.
17. T. M. Lillesand, R. W. Kiefer, and J. W. Chipman. *Remote Sensing and Image Interpretaion*. John Wiley & Sons, 2004.
18. Maarten Löffler and Jack Snoeyink. Delaunay triangulations of imprecise points in linear time after preprocessing. In *SoCG*, pages 298–304, 2008.
19. Jiri Matousek. Approximations and optimal geometric divide-and-conquer. In *SToC*, pages 505–511, 1991.
20. Jiri Matousek. *Geometric Discrepancy; An Illustrated Guide*. Springer, 1999.
21. T. Nagai and N. Tokura. Tight error bounds of geometric problems on convex objects with imprecise coordinates. In *JCDCG*, 2000.
22. Jeff M. Phillips. Algorithms for ε-approximations of terrains. In *ICALP*, 2008.
23. S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Pearsons, 2001.
24. Marc van Kreveld and Maarten Löffler. Largest bounding box, smallest diameter, and related problems on imprecise points. *Comp. Geom.: The. and App.*, 2009.
25. Vladimir Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *The. of Prob. App.*, 16:264–280, 1971.
26. Hai Yu, Pankaj K. Agarwal, Raghunath Poreddy, and Kasturi R. Varadarajan. Practical methods for shape fitting and kinetic data structures using coresets. In *SoCG*, 2004.