Convex Hulls under Uncertainty

Pankaj K. Agarwal[1], Sariel Har-Peled[2], Subhash Suri[3],
Hakan Yıldız[3], and Wuzhou Zhang[1]

[1] Duke University
[2] University of Illinois, Urbana-Champaign
[3] University of California, Santa Barbara

Abstract. We study the convex-hull problem in a probabilistic setting, motivated by the need to handle data uncertainty inherent in many applications, including sensor databases, location-based services and computer vision. In our framework, the uncertainty of each input point is described by a probability distribution over a finite number of possible locations including a *null* location to account for non-existence of the point. Our results include both exact and approximation algorithms for computing the probability of a query point lying inside the convex hull of the input, time-space tradeoffs for the membership queries, a connection between Tukey depth and membership queries, as well as a new notion of β-hull that may be a useful representation of uncertain hulls.

1 Introduction

The convex hull of a set of points is a fundamental structure in mathematics and computational geometry, with wide-ranging applications in computer graphics, image processing, pattern recognition, robotics, combinatorics, and statistics. Worst-case optimal as well as output-sensitive algorithms are known for computing the convex hull; see the survey [15] for an overview of known results.

In many applications, such as sensor databases, location-based services or computer vision, the location and sometimes even the existence of the data is uncertain, but statistical information can be used as a probability distribution guide for data. This raises the natural computational question: what is a robust and useful convex hull representation for such an uncertain input, and how well can we compute it? We explore this problem under two simple models in which both the location and the existence (presence) of each point is described probabilistically, and study basic questions such as what is the probability of a query point lying inside the convex hull, or what does the probability distribution of the convex hull over the space look like.

Uncertainty models. We focus on two models of uncertainty: unipoint and multipoint. In the *unipoint model*, each input point has a fixed location but it only exists probabilistically. Specifically, the input $\mathcal{P}$ is a set of pairs $\{(p_1, \gamma_1), \ldots, (p_n, \gamma_n)\}$ where each p_i is a point in $\mathbb{R}^d$ and each γ_i is a real number in the range $(0, 1]$ denoting the probability of p_i's existence. The existence probabilities of different points are independent; $P = \{p_1, \ldots, p_n\}$ denotes the set of sites in $\mathcal{P}$.

In the *multipoint model*, each point probabilistically exists at one of multiple possible sites. Specifically, $\mathcal{P}$ is a set of pairs $\{(P_1, \varGamma_1), \ldots, (P_m, \varGamma_m)\}$ where each P_i is a set of n_i points and each $\varGamma_i$ is a set of n_i real values in the range $(0, 1]$. The set $P_i = \left\{p_i^1, \ldots, p_i^{n_i}\right\}$ describes the possible sites for the ith point of $\mathcal{P}$ and the set $\varGamma_i = \left\{\gamma_i^1, \ldots, \gamma_i^{n_i}\right\}$ describes the associated probability distribution. The probabilities γ_i^j correspond to disjoint events and therefore sum to at most 1. By allowing the sum to be less than one, this model also accounts for the possibility of the point not existing (i.e. the *null* location)—thus, the multipoint model generalizes the unipoint model. In the multipoint model, $P = \bigcup_{i=1}^m P_i$ refers to the set of all sites and $n = |P|$.

Our results. The main results of our paper can be summarized as follows.

(A) We show (in Section 2) that the membership probability of a query point $q \in \mathbb{R}^d$, namely, the probability of q being inside the convex hull of $\mathcal{P}$, can be computed in $O(n \log n)$ time for $d = 2$. For $d \geq 3$, assuming the input and the query point are in general position, the membership probability can be computed in $O(n^d)$ time. The results hold for both unipoint and multipoint models.

(B) Next we describe two algorithms (in Section 3) to preprocess $\mathcal{P}$ into a data structure so that for a query point its membership probability in $\mathcal{P}$ can be answered quickly. The first algorithm constructs a *probability map* $\mathbb{M}(\mathcal{P})$, a partition of $\mathbb{R}^d$ into convex cells, so that all points in a single cell have the same membership probability. We show that $\mathbb{M}(\mathcal{P})$ has size $\varTheta(n^{d^2})$, and for $d = 2$ it can be computed in optimal $O(n^4)$ time. The second one is a sampling-based Monte Carlo algorithm for constructing a near-linear-size data structure that can approximate the membership probability with high likelihood in sublinear time for any fixed dimension.

(C) We show (in Section 4) a connection between the membership probability and the Tukey depth, which can be used to approximate cells of high membership probabilities. For $d = 2$, this relationship also leads to an efficient data structure.

(D) Finally, we introduce the notion of *β-hull* (in Section 5) as another approximate representation for uncertain convex hulls in the multipoint model: a convex set C is called *β-dense* for $\mathcal{P}$, for $\beta \in [0, 1]$, if C contains at least β fraction of each uncertain point. The β-hull of $\mathcal{P}$ is the intersection of all β-dense sets for $\mathcal{P}$. We show that for $d = 2$, the β-hull of $\mathcal{P}$ can be computed in $O(n \log^3 n)$ time.

Because of lack of space, many technical details and proofs are omitted from this version and can be found in the full version [3].

Related work. There is extensive and ongoing research in the database community on uncertain data; see [7] for a survey. In the computational geometry community, the early work relied on deterministic models for uncertainty (see e.g. [11]), but more recently probabilistic models of uncertainty, which are closer to the models used in statistics and machine learning, have been explored [1, 2, 9, 10, 14, 16]. The convex-hull problem over uncertain data has received some attention very recently. Suri *et al.* [16] showed that the problem of computing the most likely convex hull of a point set in the multipoint model is NP-hard. Even in the unipoint model, the problem is NP-hard for $d \geq 3$. They also presented an $O(n^3)$-time algorithm for computing the most likely convex hull under the unipoint model in $\mathbb{R}^2$. Zhao *et al.* [17] investigated the problem of computing the probability of each uncertain point lying on the convex hull, where they aimed to return the set of (uncertain) input points whose probabilities of being on the convex hull are at least some threshold. Jørgensen *et al.* [8] showed that the distribution of properties, such as areas or perimeters, of the convex hull of $\mathcal{P}$ may have $\Omega(\Pi_{i=1}^{m} n_i)$ complexity.

2 Computing the Membership Probability

For simplicity, we describe our algorithms under the unipoint model, and then discuss their extension to the multipoint model. We begin with the 2D case.

2.1 The two-dimensional case

Let $\mathcal{P} = \{(p_1, \gamma_1), \ldots, (p_n, \gamma_n)\}$ be a set of n uncertain points in $\mathbb{R}^2$ under the unipoint model. Recall that $P = \{p_1, \ldots, p_n\}$ is the set of all sites of $\mathcal{P}$. For simplicity, we make the general position assumption on the input, namely, that all coordinates are distinct and no three sites are collinear. A subset $B \subseteq P$ is the outcome of a probabilistic experiment with probability $\gamma(B) = \prod_{p_i \in B} \gamma_i \times \prod_{p_i \notin B} \overline{\gamma_i}$, where $\overline{\gamma_i}$ is the complementary probability $1 - \gamma_i$. By definition, for a point q, the *probability* of q to lie in the convex-hull of B is $\mu(q) = \sum_{B \subseteq P \mid q \in \textsc{ch}(B)} \gamma(B)$, where $\textsc{ch}(B)$ is the convex hull of B. This unfortunately involves an exponential number of terms. However, observe that for a subset $B \subseteq P$, the point q is *outside* $\textsc{ch}(B)$, if and only if q is a vertex of the convex hull $\textsc{ch}(B \cup \{q\})$. So, let $C = \textsc{ch}(B \cup \{q\})$, and V be the set of vertices of C. Then $\mu(q) = 1 - \Pr[\, q \in V \,]$.

If $B = \emptyset$, then clearly $C = \{q\}$ and $q \in V$. Otherwise, $|V| \geq 2$ and $q \in V$ implies that q is an endpoint of exactly two edges on the boundary of C.[4] In this case, the first edge following q in the counter-clockwise order of C is called the *witness edge* of q being in V. Thus, $q \in V$ if and only if $B = \emptyset$ or (exclusively) B has a witness edge, i.e.,

$$\Pr\Big[\, q \in V \,\Big] = \Pr\Big[\, B = \emptyset \,\Big] + \sum_{i=1}^{n} \Pr\Big[\, qp_i \text{ is the witness edge of } q \in V \,\Big] .$$

The first term can be computed in linear time. To compute the ith term in the summation, we observe that qp_i is the witness edge of B if and only if $p_i \in B$ and B contains no sites to the right of the oriented line spanned by the vector

[4] If B consists of a single site p_i, then C is the line segment qp_i. In this case, we consider the boundary of C to be a cycle formed by two edges: one going from q to p_i, and one going from p_i back to q.

$\overrightarrow{qp_i}$, which occurs with probability $\gamma_i \cdot \prod_{p_j \in G_i} \overline{\gamma_j}$, where G_i is the set of sites to the right of $\overrightarrow{qp_i}$. This expression can be computed in $O(n)$ time. It follows that $1 - \mu(q)$, and therefore $\mu(q)$, can be computed in $O(n^2)$ time. The computation time can be improved to $O(n \log n)$ as described in the following paragraph.

Improving the running time. The main idea is to compute the witness edge probabilities in radial order around q. We sort all sites in counter-clockwise order around q. Without loss of generality, assume that the circular sequence $p_1, \ldots, p_n$ is the resulting order. We first compute, in $O(n)$ time, the probability that qp_1 is the witness edge. Then, for increasing values of i from 2 to n, we compute, in $O(1)$ amortized time, the probability that qp_i is the witness edge by updating the probability for qp_{i-1}. In particular, let W_i denote the set of sites in the open wedge bounded by the vectors $\overrightarrow{qp_{i-1}}$ and $\overrightarrow{qp_i}$. Notice that $G_i = G_{i-1} \cup \{p_{i-1}\} \setminus W_i$. It follows that the probability for qp_i can be computed by multiplying the probability for qp_{i-1} with $\frac{\gamma_i}{\gamma_{i-1}} \times \frac{\overline{\gamma_{i-1}}}{\prod_{p_j \in W_i} \overline{\gamma_j}}$. The amortized cost of a single update is $O(1)$ because the total number of multiplications in all the updates is at most $4n$. (Each site affects at most 4 updates.) Finally, notice that we can easily keep track of the set W_i during our radial sweep, as changes to this set follow the same radial order.

Theorem 1. *Given a set of n uncertain points in $\mathbb{R}^2$ under the unipoint model, the membership probability of a query point q can be computed in $O(n \log n)$ time.*

2.2 The *d*-dimensional case

The difficulty in extending the above to higher dimensions is an appropriate generalization of witness edges, which allow us to implicitly sum over exponentially many outcomes without over-counting. Our algorithm requires that all sites, including the query point q, are in general position in the following sense: for $2 \leq k \leq d$, the projection of no $k+1$ points of $P \cup \{q\}$ on a subspace spanned by any subset of k coordinates lies on a $(k-1)$-hyperplane.

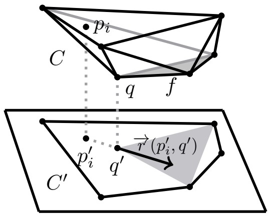
Let B be an outcome, $C = \textsc{ch}(B \cup \{q\})$ its convex hull, and V the vertices of C. Let $\lambda(B \cup \{q\})$ denote the point with the lowest x_d-coordinate in $B \cup \{q\}$. Clearly, if q is $\lambda(B \cup \{q\})$ then $q \in V$; otherwise, we condition the probability based on which point among B is $\lambda(B \cup \{q\})$. Therefore, we can write

$$\Pr\Big[q \in V\Big] = \Pr\Big[q = \lambda(B \cup \{q\})\Big] + \sum_{1 \leq i \leq n} \Pr\Big[p_i = \lambda(B \cup \{q\}) \ \wedge \ q \in V\Big].$$

It is easy to compute the first term. We show below how to compute each term of the summation in $O(n^{d-1})$ time, which gives the desired bound of $O(n^d)$.

Consider an outcome B. Let p_i be an arbitrary point in B. We use p_i as a reference point known to be contained in the hull $C = \textsc{ch}(B \cup \{q\})$. Let B', p'_i and q' denote the projections of B, p_i and q respectively on the hyperplane $x_d = 0$, which we identify with $\mathbb{R}^{d-1}$. Let us define $C' = \textsc{ch}(B' \cup \{q'\}) \subset \mathbb{R}^{d-1}$, and let V' be the vertices of C'.

Let $\overrightarrow{r}(p'_i, q')$ denote the open ray emanating from q' in the direction of the vector $\overrightarrow{p'_i q'}$ (that is, this ray is moving "away" from p'_i). A facet f of C is a *p_i-escaping* facet for q, if q is a vertex of f and the projection of f on $\mathbb{R}^{d-1}$ intersects $\overrightarrow{r}(p'_i, q')$. See the figure on the right. The following lemma is key to our algorithm. The points of C projected into $\partial C'$ form the *silhouette* of C.



Lemma 1. *(A) If $q' \in V'$ then q is a silhouette vertex of C and vice versa.*

(B) q has at most one p_i-escaping facet on C.

(C) The point q is a non-silhouette vertex of the convex-hull C if and only if q has a (single) p_i-escaping facet on C.

Given a subset of sites $P_\alpha \subseteq P \setminus \{p_i\}$ of size $(d-1)$, define $f(P_\alpha)$ to be the $(d-1)$-dimensional simplex $\textsc{ch}(P_\alpha \cup \{q\})$. Since $p_i = \lambda(B \cup \{q\})$ implies $p_i \in B$, we can use Lemma 1 to decompose the ith term as follows:

$$\Pr\Big[p_i = \lambda(B \cup \{q\}) \ \wedge \ q \in V\Big] = \Pr\Big[p_i = \lambda(B \cup \{q\}) \ \wedge \ q' \in V'\Big]$$
$$+ \sum_{\substack{P_\alpha \subseteq P \setminus \{p_i\} \\ |P_\alpha| = (d-1) \\ f(P_\alpha) \text{ is } p_i\text{-escaping for } q}} \Pr\Big[p_i = \lambda(B \cup \{q\}) \ \wedge \ f(P_\alpha) \text{ is a facet of } C\Big].$$

The first term is an instance of the same problem in $(d-1)$ dimensions (for the point q' and the projection of P), and thus is computed recursively. For the second term, we compute the probability that $f(P_\alpha)$ is a facet of C as follows. Let $G_1 \subseteq P$ be the subset of sites which are on the other side of the hyperplane supporting $f(P_\alpha)$ with respect to p_i. Let $G_2 \subseteq P$ be the subset of sites that are below p_i along the x_d-axis. Clearly, $f(P_\alpha)$ is a facet of C (and $p_i = \lambda(B \cup \{q\})$) if and only if all points in P_α and p_i exist in B, and all points in $G_1 \cup G_2$ are absent from B. The corresponding probability can be written as $\gamma_i \times \prod_{p_j \in P_\alpha} \gamma_j \times \prod_{p_j \in G_1 \cup G_2} \overline{\gamma_j}$. This formula is valid only if $P_\alpha \cap G_2 = \emptyset$ and p_i has a lower x_d-coordinate than q; otherwise we set the probability to zero. This expression can be computed in linear time, and the whole summation term can be computed in $O(n^d)$ time. Then, by induction, the computation of the ith term takes $O(n^d)$ time. Notice that the base case of our induction requires computing the probability $\Pr\big[p_i = \lambda(B \cup \{q\}) \ \wedge \ q^{(d-2)} \in V^{(d-2)}\big]$ (where ${}^{(d-2)}$ indicates a projection to $\mathbb{R}^2$). Computing this probability is essentially a two-dimensional membership probability problem on q and P, but is conditioned on the existence of p_i and the non-existence of all sites below p_i along dth axis. Our two dimensional algorithm can be easily adapted to solve this variation in $O(n \log n)$ time as well. Finally, we can improve the computation time for the ith term to $O(n^{d-1})$ by considering the facets $f(P_\alpha)$ in radial order. See the full version of the paper [3] for details.

Remark. The degeneracy of the input is easy to handle in two dimensions, but creates some technical difficulties in higher dimensions that we are currently investigating.

Theorem 2. *Let $\mathcal{P}$ be an uncertain set of n points in the unipoint model in $\mathbb{R}^d$ and q be a point. If the input sites and q are in general position, then one can compute the membership probability of q in $O(n^d)$ time, using linear space.*

Extension to the multipoint model. The algorithm extends to the multipoint model easily by modifying the computation of the probability for an edge or facet. See the full version of the paper [3] for details.

Theorem 3. *Given an uncertain set $\mathcal{P}$ of n points in the multipoint model in $\mathbb{R}^d$ and a point $q \in \mathbb{R}^d$, we can compute the membership probability of q in $O(n \log n)$ time for $d = 2$, and in $O(n^d)$ time for $d \geq 3$ if input sites and q are in general position.*

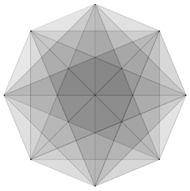
3 Membership Queries

We describe two algorithms – one deterministic and one Monte Carlo – for preprocessing a set of uncertain points for efficient membership-probability queries.

Probability map. The *probability map* $\mathbb{M}(\mathcal{P})$ is the subdivision of $\mathbb{R}^d$ into maximal connected regions so that $\mu(q)$ is the same for all query points q in a region. The following lemma gives a tight bound on the size of $\mathbb{M}(\mathcal{P})$.

Lemma 2. *The worst-case complexity of the probability map of a set of uncertain points in $\mathbb{R}^d$ is $\Theta(n^{d^2})$, under both the unipoint and the multipoint model, where n is the total number of sites in the input.*

Proof. We prove the result for the unipoint model, as the extension to the multipoint model is straightforward. For the upper bound, consider the set H of $O(n^d)$ hyperplanes formed by all d-tuples of points in $\mathcal{P}$. In the arrangement $\mathcal{A}(H)$ formed by these planes, each (open) cell has the same value of $\mu(q)$. This arrangement, which is a refinement of $\mathbb{M}(\mathcal{P})$, has size $O((n^d)^d) = O(n^{d^2})$, establishing the upper bound.



For the lower bound, consider the problem in two dimensions; extension to higher dimensions is straightforward. We choose the sites to be the vertices $p_1, \ldots, p_n$ of a regular n-gon, where each site exists with probability γ, $0 < \gamma < 1$. See the figure on the right. Consider the arrangement $\mathcal{A}$ formed by the line segments $p_i p_j$, $1 \leq i < j \leq n$, and treat each face as relatively open. If $\mu(f)$ denotes the membership probability for a face f of $\mathcal{A}$, then for any two faces f_1 and f_2 of $\mathcal{A}$, where f_1 bounds f_2 (i.e., $f_1 \subset \partial f_2$), we have $\mu(f_1) \geq \mu(f_2)$, and $\mu(f_1) > \mu(f_2)$ if $\gamma < 1$. Thus, the size of the arrangement $\mathcal{A}$ is also a lower bound on the complexity of $\mathbb{M}(\mathcal{P})$. This proves that the worst-case complexity of $\mathbb{M}(\mathcal{P})$ in $\mathbb{R}^d$ is $\Theta(n^{d^2})$. □

We can preprocess this arrangement into a point-location data structure, giving us the following result for $d = 2$.

Theorem 4. *Let $\mathcal{P}$ be a set of uncertain points in $\mathbb{R}^2$, with a total of n sites. $\mathcal{P}$ can be preprocessed in $O(n^4)$ time into a data structure of size $O(n^4)$ so that for any point $q \in \mathbb{R}^d$, $\mu(q)$ can be computed in $O(\log n)$ time.*

See the full version of the paper [3] for details.

Remark. For $d \geq 3$, due to our general position assumption, we can compute the membership probability only for d-faces of $\mathbb{M}(\mathcal{P})$, and not for the lower-dimensional faces. In that case, by utilizing a point-location technique in [5], one can build a structure that can report the membership probability of a query point (inside a d-face) in $O(\log n)$ time, with a preprocessing cost of $O(n^{d^2+d})$.

Monte Carlo algorithm. The size of the probability map may be prohibitive even for $d = 2$, so we describe a simple, space-efficient Monte Carlo approach for quickly approximating the membership probability, within absolute error. Fix a parameter $s > 1$, to be specified later. The preprocessing consists of s rounds, where the algorithm creates an outcome A_j of $\mathcal{P}$ in each round j. Each A_j is preprocessed into a data structure so that for a query point $q \in \mathbb{R}^d$, we can determine whether $q \in \textsc{ch}(A_j)$.

For $d \leq 3$, we can build each $\textsc{ch}(A_j)$ explicitly and use linear-size point-location structures with $O(\log n)$ query time. This leads to total preprocessing time $O(sn \log n)$ and space $O(sn)$. For $d \geq 4$, we use the data structure in [13] for determining whether $q \in A_j$, for all $1 \leq j \leq s$. For a parameter t such that $n \leq t \leq n^{\lfloor d/2 \rfloor}$ and for any constant $\sigma > 0$, using $O(st^{1+\sigma})$ space and preprocessing, it can compute in $O(\frac{sn}{t^{1/\lfloor d/2 \rfloor}} \log^{2d+1} n)$ time whether $q \in \textsc{ch}(A_j)$ for every j.

Given a query point $q \in \mathbb{R}^d$, we check for membership in $\textsc{ch}(A_j)$, for every $j \leq s$. If q lies in k of them, we return $\widehat{\mu}(q) = k/s$ as our estimate of $\mu(q)$. Thus, the query time is $O(\frac{sn}{t^{1/\lfloor d/2 \rfloor}} \log^{2d+1} n)$ for $d \geq 4$, $O(s \log n)$ for $d = 3$, and $O(\log n + s)$ for $d = 2$ (using fractional cascading).

It remains to determine the value of s so that $|\mu(q) - \widehat{\mu}(q)\,| \leq \varepsilon$ for all queries q, with probability at least $1-\delta$. For a fixed q and outcome A_j, let X_i be the random indicator variable, which is 1 if $q \in \textsc{ch}(A_j)$ and 0 otherwise. Since $\mathsf{E}[X_i] = \mu(q)$ and $X_i \in \{0, 1\}$, using a Chernoff-Hoeffding bound on $\widehat{\mu}(q) = k/s = (1/s) \sum_i X_i$, we observe that $\Pr[\,|\widehat{\mu}(q) - \mu(q)\,| \geq \varepsilon\,] \leq 2 \exp(-2\varepsilon^2 s) \leq \delta'$. By Lemma 2, we need to consider $O(n^{d^2})$ distinct queries. If we set $1/\delta' = O(n^{d^2}/\delta)$ and $s = O((1/\varepsilon^2) \log(n/\delta))$, we obtain the following theorem.

Theorem 5. *Let $\mathcal{P}$ be a set of uncertain points in $\mathbb{R}^d$ under the multipoint model with a total of n sites, and let $\varepsilon, \delta \in (0, 1)$ be parameters. For $d \geq 4$, $\mathcal{P}$ can be preprocessed, for any constant $\sigma > 0$, in $O((t^{1+\sigma}/\varepsilon^2) \log \frac{n}{\delta})$ time, into a data structure of size $O((t^{1+\sigma}/\varepsilon^2) \log \frac{n}{\delta})$, so that with probability at least $1-\delta$, for any query point $q \in \mathbb{R}^2$, $\widehat{\mu}(q)$ satisfying $|\mu(q) - \widehat{\mu}(q)\,| \leq \varepsilon$ and $\widehat{\mu}(q) > 0$ can be returned in $O(\frac{n}{t^{1/\lfloor d/2 \rfloor} \varepsilon^2} \log \frac{n}{\delta} \log^{2d+1} n)$ time, where t is a parameter and $n \leq t \leq n^{\lfloor d/2 \rfloor}$. For $d \leq 3$, the preprocessing time and space are $O(\frac{n}{\varepsilon^2} \log \log \frac{n}{\delta} \log n)$ and $O(\frac{n}{\varepsilon^2} \log \frac{n}{\delta})$,* *respectively. The query time is* $O(\frac{1}{\varepsilon^2}\log(\frac{n}{\delta})\log n)$ *(resp.* $O(\frac{1}{\varepsilon^2}\log\frac{n}{\delta})$*) for* $d = 3$ *(resp.* $d = 2$*).*

4 Tukey Depth and Convex Hull

The membership probability is neither a convex nor a continuous function, as suggested by the example in the proof of Lemma 2. In this section, we establish a helpful structural property of this function, intuitively showing that the probability stabilizes once we go deep enough into the "region". Specifically, we show a connection between the Tukey depth of a point q with its membership probability; in two dimensions, this also results in an efficient data structure for approximating $\mu(q)$ quickly within a small absolute error.

Estimating $\mu(q)$**.** Let Q be a set of weighted points in $\mathbb{R}^d$. For a subset $A \subseteq Q$, let $w(A)$ be the total weight of points in A. Then the *Tukey depth* of a point $q \in \mathbb{R}^d$ with respect to Q, denoted by $\tau(q, Q)$, is $\min w(Q \cap H)$ where the minimum is taken over all halfspaces H that contain q.[5] If Q is obvious from the context, we use $\tau(q)$ to denote $\tau(q, Q)$. Before bounding $\mu(q)$ in terms of $\tau(q, Q)$, we prove the following lemma.

Lemma 3. *Let* Q *be a finite set of points in* $\mathbb{R}^d$*. For any* $p \in \mathbb{R}^d$*, there is a set* $\mathcal{S} = \{S_1, \ldots, S_T\}$ *of* d*-simplices formed by* Q *such that (i) each* S_i *contains* p *in its interior; (ii) no pair of them shares a vertex; and (iii)* $T \geq \lceil \tau(p, Q)/d \rceil$*.*

We now use Lemma 3 to bound $\mu(p)$ in terms of $\tau(p, P)$.

Theorem 6. *Let* $\mathcal{P}$ *be a set of* n *uncertain points in the uniform unipoint model, that is, each point is chosen with the same probability* $\gamma > 0$*. Let* P *be the set of sites in* $\mathcal{P}$*. There is a constant* $c > 0$ *such that for any point* $p \in \mathbb{R}^d$ *with* $\tau(p, P) = t$*, we have* $(1-\gamma)^t \leq 1 - \mu(p) \leq d \exp\left(-\frac{\gamma t}{cd^2}\right)$*.*

Proof. For the first inequality, fix a closed halfspace H that contains t points of P. If none of these t points is chosen then p does not appear in the convex hull of the outcome, so $1 - \mu(p) \geq (1-\gamma)^t$.

Next, let $\mathcal{S}$ be the set of simplices of Lemma 3, and let V be its set of vertices, where $T \geq \lceil t/d \rceil$. Let $n' = |V| = (d+1)T$. Set $\varepsilon = \frac{1}{d+1}$. A random subset of V of size $O(\frac{d}{\varepsilon}\log\frac{1}{\varepsilon\delta}) = O(d^2 \log\frac{d}{\delta})$ is an ε-net for halfspaces, with probability at least $1 - \delta$.

In particular, any halfspace passing through p, contains at least T points of V. That is, all these halfspaces are ε-heavy and would be stabbed by an ε-net. Now, if we pick each point of V with probability γ, it is not hard to argue that the resulting sample R is an ε-net[6]. Indeed, the expected size (and

[5] If the points in Q are unweighted, then $\tau(q, Q)$ is simply the minimum number of points that lie in a closed halfspace that contains q.

[6] The standard argument uses slightly different sampling, but this is a minor technicality, and it is not hard to prove the ε-net theorem with this modified sampling model.

in with sufficiently large probability) of $R \cap V$ is $n'' = n'\gamma = (d+1)T\gamma \geq t\gamma$. As such, for some constant c, we need the minimal value of δ such that the inequality $t\gamma \geq cd^2 \ln \frac{d}{\delta}$ holds, which is equivalent to $\exp\left(\frac{t\gamma}{cd^2}\right) \geq \frac{d}{\delta}$. This in turn is equivalent to $\delta \geq d \exp\left(-\frac{t\gamma}{cd^2}\right)$. Thus, we set $\delta = d \exp\left(-\frac{t\gamma}{cd^2}\right)$.

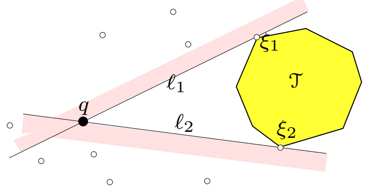
Now, with probability at least $1 - \delta$, for a point p in $\mathbb{R}^d$ with Tukey depth at least t, we have that p is in the convex-hull of the sample. □

Theorem 6 can be extended to the case when each point p_i of $\mathcal{P}$ is chosen with different probability, say, γ_i. In order to apply Theorem 6, we convert $\mathcal{P}$ to a multiset $\mathcal{Q}$, as follows. We choose a parameter $\eta = \frac{\delta}{10n}$. For each point $p_i \in \mathcal{P}$, we make $w_i = \left\lceil \frac{\ln(1-\gamma_i)}{\ln(1-\eta)} \right\rceil$ copies of p_i, each of which is selected with probability η. We can apply Theorem 6 to $\mathcal{Q}$ and show that if $\tau(q, \mathcal{Q}) \geq \frac{d^2}{\eta} \ln(2d/\delta)$, then $\mu(q, \mathcal{Q}) \geq (1 - \delta/2)$. Omitting the further details, we conclude the following.

Corollary 1. *Let $\mathcal{P} = \{(p_1, \gamma_1), \ldots, (p_n, \gamma_n)\}$ be a set of n uncertain points in $\mathbb{R}^d$ under the unipoint model. For $1 \leq i \leq n$, set $w_i = \left\lceil \frac{\ln(1-\gamma_i)}{\ln(1-\delta/10n)} \right\rceil$ to be the weight of point p_i. If the (weighted) Tukey depth of a point $q \in \mathbb{R}^d$ in $\{p_1, \ldots, p_n\}$ is at least $\frac{10d^2n}{\delta} \ln(2d/\delta)$, then $\mu(q, \mathcal{P}) \geq 1 - \delta$.*

Data structure. Let $\mathcal{P}$ be a set of points in the uniform unipoint model in $\mathbb{R}^2$, i.e., each point appears with probability γ. We now describe a data structure to estimate $\mu(q)$ for a query point $q \in \mathbb{R}^2$, within additive error $1/n$. We fix a parameter $t_0 = \frac{c}{\gamma} \ln n$ for some constant $c > 0$. Let $\mathcal{T} = \left\{x \in \mathbb{R}^2 \mid \tau(x, \mathcal{P}) \geq t_0\right\}$ be the set of all points whose Tukey depth in P is at least t_0. $\mathcal{T}$ is a convex polygon with $O(n)$ vertices [12]. By Theorem 6, $\mu(q) \geq 1 - 1/n^2$ for all points $q \in \mathcal{T}$, provided that the constant c is chosen appropriately. We also preprocess P for halfspace range reporting queries [6]. $\mathcal{T}$ can be computed in time $O(n \log^3 n)$ [12], and constructing the half-plane range reporting data structure takes $O(n \log n)$ time [6]. So the total preprocessing time is $O(n \log^3 n)$, and the size of the data structure is linear.

A query is answered as follows. Given a query point $q \in \mathbb{R}^2$, we first test in $O(\log n)$ time whether $q \in \mathcal{T}$. If the answer is yes, we simply return 1 as $\mu(q)$. If not, we compute in $O(\log n)$ time the two tangents ℓ_1, ℓ_2 of $\mathcal{T}$ from q. For $i = 1, 2$, let $\xi_i = \ell_i \cap \mathcal{T}$, and let ℓ_i^- be the half-plane bounded by ℓ_i that does not contain $\mathcal{T}$. Set $\mathcal{P}_q = \mathcal{P} \cap (\ell_1^- \cup \ell_2^-)$ and $n_q = |\mathcal{P}_q|$. Let R_q be the subset of $\mathcal{P}_q$ by choosing each point with probability γ.



By querying the half-plane range reporting data structure with each of these two tangent lines, we compute the set $\mathcal{P}_q$ in time $O(\log n + n_q)$. Let $\omega_q = \Pr[\, q \notin \textsc{ch}(R_q \cup \mathcal{T})\,]$. We compute ω_q, in $(n_q \log n_q)$ time, by adapting the algorithm for computing $\mu(q)$ described in Section 2.

The correctness and efficiency of the algorithm follow from the following lemma, whose proof is omitted from this version.

Lemma 4. *For any point $q \notin \mathcal{T}$, (i) $|\Pr[\, q \in \textsc{ch}(R_q \cup \mathcal{T})\,] - \mu(q)| \le 1/n$; (ii) $n_q \le 4t_0 = O(\gamma^{-1}\log n)$.*

By Lemma 4, $n_q = O(\gamma^{-1}\log n)$, so the query takes $O(\gamma^{-1}\log(n)\log\log n)$ time. We thus obtain the following.

Theorem 7. *Let $\mathcal{P}$ be a set of n uncertain points in $\mathbb{R}^2$ in the unipoint model, where each point appears with probability γ. $\mathcal{P}$ can be preprocessed in $O(n\log^3 n)$ time into a linear-size data structure that, for any point $q \in \mathbb{R}^2$, returns a value $\widetilde{\mu}(q)$ in $O(\gamma^{-1}\log(n)\log\log n)$ time such that $|\widetilde{\mu}(q) - \mu(q)\,| \le 1/n$.*

5 β-Hull

In this section, we consider the multipoint model, i.e., $\mathcal{P}$ is a set of m uncertain point defined by the pairs $\{(P_1, \Gamma_1), \ldots, (P_m, \Gamma_m)\}$. A convex set $C \subseteq \mathbb{R}^2$ is called *β-dense* with respect to $\mathcal{P}$ if it contains β-fraction of each (P_i, Γ_i), i.e., $\sum_{p_i^j} \gamma_i^j \ge \beta$ for all $i \le m$. The *β-hull* of $\mathcal{P}$, denoted by $\textsc{ch}_\beta(\mathcal{P})$, is the intersection of all convex β-dense sets with respect to $\mathcal{P}$. Note that for $m = 1$, $\textsc{ch}_\beta(\mathcal{P})$ is the set of points whose Tukey depth is at least $1-\beta$. We first prove an $O(n)$ upper bound on the complexity of $\textsc{ch}_\beta(\mathcal{P})$ and then describe an algorithm for computing it.

Theorem 8. *Let $\mathcal{P} = \{(P_1, \Gamma_1), \ldots, (P_m, \Gamma_m)\}$ be a set of m uncertain points in $\mathbb{R}^2$ under the multipoint model with $P = \bigcup_{i=1}^m P_i$ and $|P| = n$. For any $\beta \in [0,1]$, $\textsc{ch}_\beta(\mathcal{P})$ has $O(n)$ vertices.*

Proof. We call a convex β-dense set C *minimal* if there is no convex β-dense set C' such that $C' \subset C$. A minimal convex β-dense set C is the convex hull of $P \cap C$. Therefore C is a convex polygon whose vertices are a subset of P. Obviously $\textsc{ch}_\beta(\mathcal{P})$ is the intersection of minimal convex β-dense sets. Therefore each edge of $\textsc{ch}_\beta(\mathcal{P})$ lies on a line passing through a pair of points of P, i.e., $\textsc{ch}_\beta(\mathcal{P})$ is the intersection of a set H of halfplanes, each bounded by a line passing through a pair of points of P. Next we argue that $|H| \le 2n$.

Fix a point $p \in P$. We claim that H contains at most two halfplanes whose bounding lines pass through p. Indeed if $p \in \mathrm{int}(\textsc{ch}_\beta(\mathcal{P}))$, then no bounding line of H passes through p; if $p \in \partial(\textsc{ch}_\beta(\mathcal{P}))$, then at most two bounding lines of H pass through p; and if $p \notin \textsc{ch}_\beta(\mathcal{P})$, then there are two tangents to $\textsc{ch}_\beta(\mathcal{P})$ from p. Hence at most two bounding lines of H pass through p, as claimed. □

Algorithm. We describe the algorithm for computing the upper boundary $\mathcal{U}$ of $\textsc{ch}_\beta(\mathcal{P})$. The lower boundary of $\textsc{ch}_\beta(\mathcal{P})$ can be computed analogously. It will be easier to compute $\mathcal{U}$ in the dual plane. Let $\mathcal{U}^*$ denote the dual of $\mathcal{U}$. We call a line ℓ passing through a point $p \in P_i$ *β-tangent* of P_i at p if one of the open half-planes bounded by ℓ contains less than β-fraction of points of P_i but the corresponding closed half-plane contains at least β-fraction of points.

Recall that the dual of a point $p = (a, b)$ is the line $p^* : y = ax - b$, and the dual of a line $\ell : y = mx + c$ is the point $\ell^* = (m, -c)$. The point p lies above/below/on the line ℓ if and only if the dual point ℓ^* lies above/below/on the dual line p^*. Set $P_i^* = \left\{ p_i^{j*} \mid p_i^j \in P_i \right\}$ and $P^* = \bigcup_{i=1}^m P_i^*$. For a point $q \in \mathbb{R}^2$ and for $i \leq m$, let $\kappa(q, i) = \sum \gamma_i^j$ where the summation is taken over all points $p_i^j \in P_i$ such that q lies below the dual line p_i^{j*}. We define the β-level $\varLambda_i$ of P_i^* to be the upper boundary of the region $\left\{ q \in \mathbb{R}^2 \mid \kappa(q, i) \geq \beta \right\}$. $\varLambda_i$ is an x-monotone polygonal chain composed of the edges of the arrangement $\mathcal{A}(P_i^*)$; the dual line of a point on $\varLambda_i$ is a β-tangent line of P_i. Let $\varLambda$ be the lower envelope of $\varLambda_1, \ldots, \varLambda_m$.

Let ℓ be the line supporting an edge of $\mathcal{U}$. It can be proved that the dual point ℓ^* is a vertex of $\varLambda$. Next, let q be a vertex of $\mathcal{U}$, then q cannot lie above any β-tangent line of any P_i, which implies that the dual line q^* passes through a pair of vertices of $\varLambda$ and does not lie below any vertex of $\varLambda$. Hence, each vertex of $\mathcal{U}$ corresponds to an edge of the upper boundary of the convex hull of $\varLambda$. This observation suggests that $\mathcal{U}^*$ can be computed by adapting an algorithm for computing the convex hull of a level in an arrangement of lines [4, 12]. We begin by describing a simple procedure, which will be used as a subroutine in the overall algorithm.

Lemma 5. *Given a line ℓ, the intersection points of ℓ and $\varLambda$ can be computed in $O(n \log n)$ time.*

Proof. We sort the intersections of the lines of P^* with ℓ. Let $\langle q_1, \ldots, q_u \rangle$, $u \leq n$, be the sequence of these intersection points. For every $i \leq m$, $\kappa(q_1, i)$ can be computed in a total of $O(n)$ time. Given $\{\kappa(q_{j-1}, i) \mid 1 \leq i \leq m\}$, $\{\kappa(q_j, i) \mid 1 \leq i \leq m\}$ can be computed in $O(1)$ time. A point $q_j \in \varLambda$ if $q_j \in \varLambda_i$ for some i and lies below $\varLambda'_i$ for all other i'. This completes the proof of the lemma. □

The following two procedures can be developed by plugging Lemma 5 into the parametric-search technique [4, 12].

(A) Given a point q, determine whether q lies above $\mathcal{U}^*$ or return the tangent lines of $\mathcal{U}^*$ from q. This can be done in $O(n \log^2 n)$ time.

(B) Given a line ℓ, compute the edges of $\mathcal{U}^*$ that intersect ℓ, in $O(n \log^3 n)$ time. (Procedure (B) uses (A) and parametric search.)

Given (B), we can now compute $\mathcal{U}^*$ as follows. We fix a parameter $r > 1$ and compute a $(1/r)$-cutting[7] $\varXi = \{\varDelta_1, \ldots, \varDelta_u\}$, where $u = O(r^2)$. For each $\varDelta_i$, we do the following. Using (B) we compute the edges of $\mathcal{U}^*$ that intersect $\partial\varDelta_i$. We can then deduce whether $\varDelta_i$ contains any vertex of $\mathcal{U}^*$. If the answer is yes, we solve the problem recursively in $\varDelta_i$ with the subset of lines of P^* that cross $\varDelta_i$. We omit the details from here and conclude the following.

Theorem 9. *Given a set $\mathcal{P}$ of uncertain points in $\mathbb{R}^2$ under the multipoint model with a total of n sites, and a parameter $\beta \in [0, 1]$, the β-hull of $\mathcal{P}$ can be computed in $O(n \log^3 n)$ time.*

[7] A $(1/r)$-*cutting* of P^* is a triangulation $\varXi$ of $\mathbb{R}^2$ such that each triangle of $\varXi$ crosses at most n/r lines of P^*.

Acknowledgments. P. Agarwal and W. Zhang are supported by NSF under grants CCF-09-40671, CCF-10-12254, and CCF-11-61359, by ARO grants W911NF-07-1-0376 and W911NF-08-1-0452, and by an ERDC contract W9132V-11-C-0003. S. Har-Peled is supported by NSF grants CCF-09-15984 and CCF-12-17462. S. Suri and H. Yıldız are supported by NSF grants CCF-1161495 and CNS-1035917.

References

1. Agarwal, P.K., Aronov, B., Har-Peled, S., Phillips, J.M., Yi, K., Zhang, W.: Nearest neighbor searching under uncertainty II. In: Proc. 32nd ACM Sympos. Principles Database Syst. pp. 115–126 (2013)
2. Agarwal, P.K., Cheng, S., Tao, Y., Yi, K.: Indexing uncertain data. In: Proc. 28th ACM Sympos. Principles Database Syst. pp. 137–146 (2009)
3. Agarwal, P.K., Har-Peled, S., Suri, S., Yıldız, H., Zhang, W.: Convex hulls under uncertainty. CoRR abs/1406.6599 (2014), http://arxiv.org/abs/1406.6599
4. Agarwal, P.K., Sharir, M., Welzl, E.: Algorithms for center and Tverberg points. ACM Trans. Algo. 5(1), 5:1–5:20 (Dec 2008)
5. Chazelle, B.: Cutting hyperplanes for divide-and-conquer. Discrete Comput. Geom. 9(1), 145–158 (1993)
6. Chazelle, B., Guibas, L.J., Lee, D.T.: The power of geometric duality. BIT 25(1), 76–90 (1985)
7. Dalvi, N.N., Ré, C., Suciu, D.: Probabilistic databases: Diamonds in the dirt. Commun. ACM 52(7), 86–94 (2009)
8. Jørgensen, A., Löffler, M., Phillips, J.: Geometric computations on indecisive points. In: Proc. 12th Workshop Algorithms Data Struct. pp. 536–547 (2011)
9. Kamousi, P., Chan, T.M., Suri, S.: Closest pair and the post office problem for stochastic points. In: Proc. 12th Workshop Algorithms Data Struct. pp. 548–559 (2011)
10. Kamousi, P., Chan, T., Suri, S.: Stochastic minimum spanning trees in euclidean spaces. In: Proc. 27th Annu. Sympos. Comput. Geom. pp. 65–74 (2011)
11. Löffler, M.: Data Imprecision in Computational Geometry. Ph.D. thesis, Dept. Computer Sci. (2009)
12. Matoušek, J.: Computing the center of planar point sets. In: Goodman, J.E., Pollack, R., Steiger, W. (eds.) Computational Geometry: Papers from the DIMACS Special Year, pp. 221–230. Amer. Math. Soc. (1991)
13. Matoušek, J., Schwarzkopf, O.: Linear optimization queries. In: Proc. 8th Annu. Sympos. Comput. Geom. pp. 16–25 (1992)
14. Phillips, J.: Small and Stable Descriptors of Distributions for Geometric Statistical Problems. Ph.D. thesis, Dept. Computer Sci. (2009)
15. Seidel, R.: Convex hull computations. In: Goodman, J.E., O'Rourke, J. (eds.) Handbook of Discrete and Computational Geometry, pp. 495–512. CRC Press (2004)
16. Suri, S., Verbeek, K., Yıldız, H.: On the most likely convex hull of uncertain points. In: Proc. 21st Annu. European Sympos. Algorithms. pp. 791–802 (2013)
17. Zhao, Z., Yan, D., Ng, W.: A probabilistic convex hull query tool. In: Proc. 15th Int. Conf. on Ext. Database Tech. pp. 570–573 (2012)