Separability and Convexity of Probabilistic Point Sets*

Martin Fink † John Hershberger ‡ Nirman Kumar † Subhash Suri †

Abstract

We describe an $O(n^d)$ time algorithm for computing the exact probability that two probabilistic point sets are linearly separable in dimension $d \geq 2$, and prove its hardness via reduction from the k-SUM problem. We also show that d-dimensional separability is computationally equivalent to a $(d+1)$-dimensional convex hull membership problem.

1 Introduction

We consider the problems of linear separability and convex hull membership for probabilistic point sets, where a *probabilistic point* is a tuple (p, π) consisting of a point $p \in \mathbb{R}^d$ and its associated probability of existence π. This abstract representation is a convenient way to model data uncertainty in a number of applications including uncertain databases, sensor networks, data cleansing, scientific computing, and machine learning [4, 5]. We present algorithms and hardness results for computing the exact probability that two such probabilistic sets in $\mathbb{R}^d$ are linearly separable (*separability problem*) or that a point lies inside the convex hull of a probabilistic set (*convex hull membership problem*). Specifically, our results include the following.

1. An $O(n^d)$ time and $O(n)$ space algorithm for computing the probability of separation of two probabilistic point sets with a total of n points in d dimensions, for $d \geq 2$.
2. A reduction from the k-SUM problem to the d-dimensional separability problem, for $k = d+1$, as evidence that our $O(n^2)$ bound for $d = 2$ may be almost tight. We also prove $\#P$-hardness of the problem when $d = \Omega(n)$.
3. A linear-time reduction between the convex hull membership problem in d-space and the separability problem in dimension $(d-1)$.
4. Finally, related problems such as probability of non-empty intersection among n probabilistic halfspaces can also be solved in $O(n^d)$ time. We also show how to extend our result to point sets containing degeneracies.

*An expanded version of this work appears in [7].

†University of California, Santa Barbara, CA, USA

‡Mentor Graphics Corp., Wilsonville, OR, USA

Related work. The topic of algorithms for probabilistic (uncertain) data is a subject of extensive and ongoing research in many areas of computer science including databases, data mining, machine learning, combinatorial optimization, theory, and computational geometry. Within computational geometry and databases, a number of papers address nearest neighbor searching, minimum spanning trees, Voronoi diagrams, indexing and skyline queries under the probabilistic model of our paper as well as the locational uncertainty model [1, 2, 10, 11, 13, 12]. Our convex hull membership bound improves upon a recent result of [3], both in time complexity and elimination of the non-degeneracy assumption.

2 Separability of Probabilistic Point Sets

2.1 Preliminaries

Let $\mathcal{A}$ and $\mathcal{B}$ be two probabilistic point sets in $\mathbb{R}^d$ with a total of n points. For notational convenience, we denote a generic probabilistic point as p with the implicit understanding that $\pi(p)$ is the probability associated with p and that all the point probabilities are independent. By the independence of probabilities, a subset A occurs as a random sample of $\mathcal{A}$ with probability

$$\mathbf{Pr}[A] = \prod_{p \in A} \pi(p) \cdot \prod_{p \in \mathcal{A} \setminus A} (1 - \pi(p)).$$

We say that the subsets $A \subseteq \mathcal{A}$ and $B \subseteq \mathcal{B}$ are linearly separable if there is a hyperplane H containing A and B in opposite (open) halfspaces. (The *open* halfspace separation means that no point of $A \cup B$ lies on H, thus enforcing a strict separation.) Define an indicator function $\sigma(\mathcal{A}, \mathcal{B})$ for linear separability

$$\sigma(A, B) = \begin{cases} 1 & \text{if } A, B \text{ are linearly separable} \\ 0 & \text{otherwise,} \end{cases}$$

with $\sigma(\emptyset, \emptyset) = 1$ to handle the trivial case. Then the *separation probability* of $\mathcal{A}$ and $\mathcal{B}$ is the joint sum over all possible samples:

$$\mathbf{Pr}[\sigma(\mathcal{A}, \mathcal{B})] = \sum_{A \subseteq \mathcal{A}, B \subseteq \mathcal{B}} \mathbf{Pr}[A] \cdot \mathbf{Pr}[B] \cdot \sigma(A, B)$$

This is also the *expectation* of the random variable $\sigma(A, B)$. We are interested in the complexity of computing this quantity *exactly*.

2.2 Reduction to Anchored Separability

There are $O(n^d)$ combinatorially distinct separating hyperplanes induced by $\mathcal{A} \cup \mathcal{B}$, so a natural idea is to decompose the sum into probabilities over these planes. However, many different hyperplanes may be separating for the same sample pair, so we must avoid over-counting by assigning each pair to a unique *canonical* hyperplane.[1] Our main insight is the following: if we introduce an extra point z into the input, then the canonical hyperplane can be defined uniquely (and computed efficiently) with respect to z. We call this additional point z the *anchor point*.

We initially assume that the input points are in general position, and choose z *above* (in the dth coordinate) all the input points and in general position with respect to $\mathcal{A} \cup \mathcal{B}$. The non-degeneracy assumption can be eliminated, as briefly explained in Section 5. We assign $\pi(z) = 1$ so that the anchor is always included in the sample.

If $A \subseteq \mathcal{A}$ and $B \subseteq \mathcal{B}$ are two random samples and H a hyperplane separating them, then z lies either (i) on the same side as A, (ii) on the same side as B, or (iii) on the hyperplane H. The following lemma shows that case (iii) precisely counts the double-counting between cases (i) and (ii).

Lemma 1 *There exist separating hyperplanes H_1, H_2 with z lying on the same side of H_1 as A but on the same side of H_2 as B if and only if there is another hyperplane H that passes through z and separates A from B.*

Let $\mathcal{P} + z$ be the shorthand for the probabilistic point set $\mathcal{P} \cup \{z\}$, with $\pi(z) = 1$. Let $\mathbf{Pr}\big[\sigma(z, \mathcal{A}, \mathcal{B})\big]$ denote the probability that sets $\mathcal{A}$ and $\mathcal{B}$ are linearly separable by a hyperplane passing through z. By the preceding lemma, we have the following.

Lemma 2 *Given two probabilistic point sets $\mathcal{A}$ and $\mathcal{B}$, we have the following equality:*

$$\mathbf{Pr}\big[\sigma(\mathcal{A}, \mathcal{B})\big] = \mathbf{Pr}\big[\sigma(\mathcal{A}+z, \mathcal{B})\big] + \mathbf{Pr}\big[\sigma(\mathcal{A}, \mathcal{B}+z)\big] - \mathbf{Pr}\big[\sigma(z, \mathcal{A}, \mathcal{B})\big].$$

Computing $\mathbf{Pr}\big[\sigma(\mathcal{A}+z, \mathcal{B})\big]$ and $\mathbf{Pr}\big[\sigma(\mathcal{A}, \mathcal{B}+z)\big]$ requires solving two instances of *anchored separability*, once with z included in $\mathcal{A}$ and once in $\mathcal{B}$, and this is the problem we solve in the following subsection. The calculation of the remaining term $\mathbf{Pr}\big[\sigma(z, \mathcal{A}, \mathcal{B})\big]$ can be reduced to an instance of separability in dimension $d-1$, as shown below.

Consider any sample $A \subseteq \mathcal{A}$ and $B \subseteq \mathcal{B}$. We centrally project all these points onto the hyperplane $x_d = 0$ from the anchor point z: the image of a point $p \in \mathbb{R}^d$ is the point $p' \in \mathbb{R}^{d-1}$ at which the line connecting z to p intersects the hyperplane $x_d = 0$. All points of $\mathcal{A} \cup \mathcal{B}$ have a well-defined projection because z lies above all of them.

Lemma 3 *Let $A \subseteq \mathcal{A}$ and $B \subseteq \mathcal{B}$ be two sample sets, and let A', B' be their projections onto $x_d = 0$ with respect to z. Then A and B are separable by a hyperplane passing through z if and only if A' and B' are linearly separable in $x_d = 0$.*

3 Computing Anchored Separability

We now describe our main technical result, namely, how to compute the probability of anchored separability $\mathbf{Pr}\big[\sigma(\mathcal{A}+z, \mathcal{B})\big]$. Given a hyperplane H, we can easily compute the probability that $\mathcal{A}+z$ lies in H^+ and $\mathcal{B}$ lies in H^-. The separation probabilities for different hyperplanes, however, are not independent, and so our algorithm "assigns" each separable sample to a unique hyperplane, which geometrically is the hyperplane that separates $A+z$ from B and lies at *maximum distance* from the anchor z. We introduce the concept of a *shadow cone* to formalize these canonical hyperplanes (see Fig. 1).

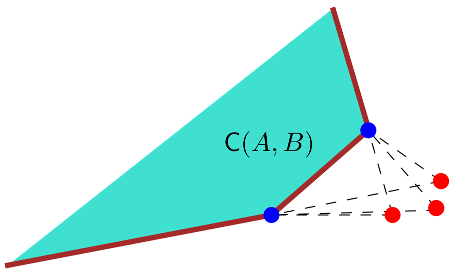

Figure 1: A shadow cone in two dimensions.

Given two points $u, v \in \mathbb{R}^d$, let $shadow(u, v) = \{\lambda v + (1-\lambda)u \mid \lambda \geq 1\}$ be the ray originating at v and directed along the line uv away from u. Given two sets of points A and B, with $A \cap B = \emptyset$, we define their *shadow cone* $\mathsf{C}(A, B)$ as the union of $shadow(u, v)$ for all $u \in CH(A)$ and $v \in CH(B)$, where $CH()$ denotes the convex hull.

$\mathsf{C}(A, B)$ is a (possibly unbounded) convex polytope, each of whose faces is defined by a subset of (at most d) points in $A \cup B$, and the defining set always includes at least one point of B. The following lemma states the important connection between the shadow cone and hyperplane separability.

Lemma 4 *$A+z$ and B can be separated by a hyperplane if and only if $z \notin \mathsf{C}(A, B)$.*

3.1 Canonical Separating Hyperplanes

Since $\mathsf{C}(A, B)$ is a convex set, there is a *unique* nearest point $p = \mathrm{np}(z, \mathsf{C}(A, B))$ on the boundary of $\mathsf{C}(A, B)$

[1] Dualizing the points to hyperplanes can simplify the enumeration of separating planes for the summation but does not address the over-counting problem.

with minimum distance to z. We define our *canonical hyperplane* $H(z, A, B)$ as the one that passes through p and is orthogonal to the vector $p - z$. The following lemma states the definition of canonical separators.

Lemma 5 *Let C be a d-dimensional convex polyhedron, z a point not contained in C, and p the point of C at minimum distance from z. If p lies in the relative interior of the face F of C, then the hyperplane H through p that is orthogonal to $p - z$ contains F. This hyperplane contains C in one of its closed halfspaces, and is the hyperplane farthest from z with this property.*

We turn the separation question around and instead of asking "which hyperplane separates a particular sample pair A, B," we ask "for which pairs of samples A, B is H a canonical separator?" The latter formulation allows us to compute the separation probability $\mathbf{Pr}\big[\sigma(\mathcal{A}+z, \mathcal{B})\big]$ by considering at most $O(n^d)$ possible hyperplanes.

3.2 The Algorithm

Our algorithm enumerates all subsets $I \subseteq \mathcal{A}$ and $J \subseteq \mathcal{B}$, with $|I \cup J| \leq d$ and $|J| \geq 1$, and assigns to the hyperplane $H(z, I, J)$ the separation probability of *all those samples $A \cup B$ that are separable and for which $H(z, I, J)$ is the canonical separator $H(z, A, B)$*. Let $\mathbf{Pr}\big[H(z, I, J)\big]$ denote the probability that the points defining the hyperplane $H(z, I, J)$ are in the sample and none of the remaining points of $\mathcal{A} \cup \mathcal{B}$ lies on its *incorrect side*. Then, it's easy to check that

$$\begin{aligned}\mathbf{Pr}\big[H(z, I, J)\big] &= \prod_{u \in I \cup J} \pi(u) \times \prod_{u \in \mathcal{A} \cap H^-} (1 - \pi(u)) \\ &\times \prod_{u \in \mathcal{B} \cap H^+} (1 - \pi(u)).\end{aligned}$$

The pseudo-code below describes our algorithm.

Algorithm AnchoredSep:

Input: The point sets $\mathcal{A} + z$ and $\mathcal{B}$
Output: Their separation probability
$\alpha = \mathbf{Pr}\big[\sigma(\mathcal{A} + z, \mathcal{B})\big]$
$\alpha = \prod_{u \in \mathcal{B}} (1 - \pi(u))$;
forall the
$I \subseteq \mathcal{A}, J \subseteq \mathcal{B}$ *where* $|I \cup J| \leq d, J \neq \emptyset$ **do**
  let $p = \mathrm{np}(z, \mathsf{C}(I, J))$;
  if *p lies in the relative interior of* $\mathsf{C}(I, J)$ **then**
    $\alpha = \alpha + \mathbf{Pr}\big[H(z, I, J)\big]$;
  end
end
return α;

Theorem 6 **AnchoredSep** *correctly computes the probability* $\mathbf{Pr}\big[\sigma(\mathcal{A} + z, \mathcal{B})\big]$.

A naïve implementation of **AnchoredSep** runs in $O(n^{d+1})$ time and $O(n)$ space, but it can be improved to $O(n^d)$ time using duality and topological sweep.

Theorem 7 *Let $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}^d$ be two probabilistic sets of n points in general position, for $d \geq 2$. We can compute their probability of hyperplane separation $\mathbf{Pr}\big[\sigma(\mathcal{A}, \mathcal{B})\big]$ in $O(n^d)$ worst-case time.*

4 Lower Bounds

We now argue that the separability problem is at least as hard as the k-SUM problem for $k = d + 1$, for any fixed d. We also show that the problem is $\#P$-hard when $d = \Omega(n)$.

The k-SUM problem is a generalization of 3-SUM, which is a classical hard problem in computational geometry [8, 9]. We use the following variant: Given k sets containing a total of n real numbers, grouped into a single set Q and $k - 1$ sets $R_1, R_2, \ldots, R_{k-1}$, determine whether there exist $k - 1$ elements $r_i \in R_i$, one per set R_i, and an element $q \in Q$ such that $\sum_{i=1}^{k-1} r_i = q$. We have the following result.

Theorem 8 *The d-dimensional hyperplane separability problem is at least as hard as $(d + 1)$-SUM.*

The problem is $\#P$-hard for $d = \Omega(n)$.

Lemma 9 *Computing $\mathbf{Pr}\big[\sigma(\mathcal{A}, \mathcal{B})\big]$ is $\#P$-hard if the dimension d is not a constant.*

Proof. We reduce the $\#P$-hard problem of counting independent sets in a graph [14] to the separability problem. Consider an undirected graph $G = (V, E)$ on the vertex set $\{1, 2, \ldots, n\}$. For each i, we construct an n-dimensional point $a_i = (0, \ldots, 1, \ldots, 0)$, namely, the unit vector along the ith axis. The collection of points $\{a_1, \ldots, a_i, \ldots, a_n\}$, each with associated probability $\pi_i = 1/2$, is our point set $\mathcal{A}$. Next, for each edge $e = (i, j) \in E$, we construct a point b_{ij} at the midpoint of the line segment connecting a_i and a_j. The set of points b_{ij}, each with associated probability 1, is the set $\mathcal{B}$. It is easy to see that there is a one-to-one correspondence between separable subsets of $\mathcal{A} \cup \mathcal{B}$ and the independent sets of G. Each separable sample occurs precisely with probability $(1/2)^n$, and therefore we can count the number of independent sets using the separation probability $\mathbf{Pr}\big[\sigma(\mathcal{A}, \mathcal{B})\big]$. ☐

5 Handling Input Degeneracies

We deal with degenerate inputs through a problem-specific symbolic perturbation within the framework of Simulation of Simplicity [6]. We convert degenerate non-separable samples into non-degenerate samples that are still non-separable. We first choose the anchor z above all points in $\mathcal{P} = \mathcal{A} \cup \mathcal{B}$ and outside the affine span of every d-tuple of $\mathcal{P}$. For each $a \in \mathcal{A}$, we define a perturbed point $a' = a + \epsilon \cdot (a - z)$, and for each $b \in \mathcal{B}$, define $b' = b + \epsilon \cdot (z - b)$, where $\epsilon > 0$ is infinitesimally small. Let $\mathcal{A}', \mathcal{B}'$ be the sets of perturbed points corresponding to $\mathcal{A}$ and $\mathcal{B}$. We prove that $A + z$ and B are strictly separable by a hyperplane if and only if $A' + z$ and B' are. Furthermore, if some hyperplane H with $z \notin H$ is a non-strict separator of $A' + z$ and B' for some ϵ, then H is a strict separator for any $\epsilon_0 < \epsilon$.

6 Convexity and Related Problems

Given a probabilistic set of points $\mathcal{P}$, the convex hull membership probability of a query point z is the probability that z lies in the convex hull of $\mathcal{P}$. We write this as $\mathbf{Pr}\big[z \in CH(\mathcal{P})\big] = \sum_{P \subseteq \mathcal{P},\ z \in CH(P)} \mathbf{Pr}\big[P\big]$. Without loss of generality, assume that the query point is $z = (0, 0, \ldots, 1)$, and define the *central projection* of $p \in \mathcal{P}$ as the point p' at which the line pz meets the plane $x_d = 0$. Let the set $\mathcal{A}$ (resp. $\mathcal{B}$) be the central projections of all those points in $\mathcal{P}$ with $x_d > 1$ (resp. with $x_d < 1$), where each point inherits the associated probability of its corresponding point in $\mathcal{P}$. The sets $\mathcal{A}$ and $\mathcal{B}$ are $(d-1)$-dimensional probabilistic points, with $|\mathcal{A}| + |\mathcal{B}| = n$. We show the following equality

$$\mathbf{Pr}\big[z \in CH(\mathcal{P})\big] \;=\; 1 - \mathbf{Pr}\big[\sigma(\mathcal{A}, \mathcal{B})\big],$$

which proves that d-dimensional convex hull membership can be computed in the same time bound as the $(d-1)$-dimensional separability. Similarly, the probability that n probabilistic halfspaces have non-empty intersection can be computed in the same time bound as d-dimensional separability.

7 Concluding Remarks

We considered the problem of hyperplane separability for probabilistic point sets. Our main result is that, given two sets of n probabilistic points in $\mathbb{R}^d$, we can compute in $O(n^d)$ time the exact probability that their random samples are linearly separable. The same technique and result lead to similar bounds for several other problems, including the probability that a query point lies inside the convex hull of n probabilistic points, or the probability that n probabilistic halfspaces have non-empty intersection. We also proved that the d-dimensional separability problem is at least as hard as the $(d+1)$-SUM problem [8, 9], which implies that our $O(n^2)$ algorithms for 2-dimensional separability or 3-dimensional convex hull membership are nearly optimal.

References

[1] P. K. Agarwal, S. W. Cheng, and K. Yi. Range searching on uncertain data. *ACM Trans. on Algorithms*, 8(4):43:1–43:17, 2012.

[2] P. K. Agarwal, A. Efrat, S. Sankararaman, and W. Zhang. Nearest-neighbor searching under uncertainty. In *PODS*, pages 225–236, 2012.

[3] P. K. Agarwal, S. Har-Peled, S. Suri, H. Yildiz, and W. Zhang. Convex hulls under uncertainty. In *Proc. ESA*, pages 37–48, 2014.

[4] C. C. Aggarwal. *Managing and Mining Uncertain Data*. Springer, 2009.

[5] C. C. Aggarwal and P. S. Yu. A survey of uncertain data algorithms and applications. *IEEE TKDE.*, 21(5):609–623, 2009.

[6] H. Edelsbrunner and E. P. Mücke. Simulation of simplicity: A technique to cope with degenerate cases in geometric algorithms. *ACM Trans. on Graphics*, 9(1):66–104, 1990.

[7] M. Fink, J. Hershberger, N. Kumar, and S. Suri. Hyperplane separability and convexity of probabilistic point sets. In *Proc. 32nd SoCG (to appear)*, 2016.

[8] A. Gajentaan and M. H. Overmars. On a class of $O(n^2)$ problems in computational geometry. *CGTA*, 5(3):165–185, 1995.

[9] A. Gronlund and S. Pettie. Threesomes, degenerates, and love triangles. In *Proc. 55th FOCS*, pages 621–630, 2014.

[10] P. Kamousi, T. M. Chan, and S. Suri. Stochastic minimum spanning trees in Euclidean spaces. In *Proc. 27th SoCG*, pages 65–74, 2011.

[11] H. P. Kriegel, P. Kunath, and M. Renz. Probabilistic nearest-neighbor query on uncertain objects. In *Advances in Databases: Concepts, Systems and Applications*, pages 337–348. 2007.

[12] S. Suri and K. Verbeek. On the most likely Voronoi diagram and nearest neighbor searching. In *Proc. 25th ISAAC*, pages 338–350, 2014.

[13] S. Suri, K. Verbeek, and H. Yıldız. On the most likely convex hull of uncertain points. In *Proc. 21st ESA*, pages 791–802, 2013.

[14] L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM J. Comput.*, 8(3):410–421, 1979.