Approximate Distribution of L_1 Median on Uncertain Data ⋆

Jeff M. Phillips and Pingfan Tang
{jeffp,tang1984}@cs.utah.edu

School of Computing, University of Utah
Salt Lake City, UT, USA

Abstract. We study the L_1 median for locationally uncertain points with discrete distributions. That is, each point in a data set has a discrete probability distribution describing its location. The L_1 median is a robust estimator, useful when there are outliers in the point set. However given the probabilistic nature of this data, there is a distribution describing the L_1 median, not a single location. We show how to construct and estimate this median distribution in near-linear or quadratic time in 1 and 2 dimensions.

1 Introduction

Most statistical or machine learning models of noisy data start with the assumption that a data set X is drawn iid (independent and identically distributed) from a single distribution Φ. Since such distributions often represent some true phenomenon plus some noisy observation step, approaches that mitigate the noise involving robust statistics or regularization have become commonplace.

However, many modern data sets are clearly not generated iid, rather each data element represents a separate object or a region of a more complex phenomenon. For instance, each data element may represent a distinct person in a population or an hourly temperature reading. Yet, this data can still be noisy; for instance, multiple GPS locational estimates of a person, or multiple temperature sensors in a city. The set of data elements may be noisy *and* there may be multiple inconsistent readings of each element. To model this noise, the inconsistent readings can naturally be interpreted as a probability distribution.

Given such locationally noisy, non-iid data sets, there are many unresolved and important analysis tasks ranging from classification to regression to summarization. In this paper, we initiate the study of robust estimators [18, 26] on locationally uncertain data. More precisely, we consider an input data set of size n, where each data point's location is described by a discrete probability distribution. We will assume these discrete distributions have a support of at most k points in $\mathbb{R}^d$; and for concreteness and simplicity we will focus on cases where each point has support described by exactly k points, each are equally likely.

⋆ Thanks to supported by NSF CCF-1350888, IIS-1251019, ACI-1443046, and CNS-1514520.

Although algorithms for locationally uncertain points have been studied in quite a few contexts over the last decade [15, 24, 21, 6, 19, 5, 3, 4, 32] (see Section 1.1), few have directly addressed the problem of noise in the data. As the uncertainty is often the direct consequence of noise in the data collection process, this is a pressing concern. As such we initiate this study focusing on the most basic robust estimators: the median for data in $\mathbb{R}^1$, and its generalization the L_1 median for data in $\mathbb{R}^2$. Both estimators can be defined as the point x^* which minimizes x over $\mathsf{cost}(x, Q) = \frac{1}{|Q|}\sum_{q\in Q} \|q - x\|$ for a data set Q. Being robust refers to the fact that if less than 50% of the data points (the outliers) are moved from the true distribution to some location infinitely far away, the estimator remains within the extent of the true distribution [25].

In this paper, we generalize the L_1 median to locationally uncertain data, where the outliers can occur not just among the n data points, but also as part of the discrete distributions representing their possible locations.

The main challenge is in modeling these robust estimators. As we do not have precise locations of the data, there is not a single minimizer of $\mathsf{cost}(x, Q)$; rather there may be as many as k^n possible input point sets Q (the combination of all possible locations of the data). And the expected value of such a minimizer is not robust in the same way that the mean is not. As such we build a distribution over the possible locations of these cost-minimizers. In $\mathbb{R}^1$ this distribution is of size at most $O(nk)$, the size of the input, but in $\mathbb{R}^2$ it may be as large as k^n.

Thus, we design algorithms to create an approximate support of these median distributions. We create small sets T such that each possible median m_Q from a possible point set Q is within a distance $\varepsilon \cdot \mathsf{cost}(m_Q, Q)$ of some $x \in T$. Under reasonable assumptions we can create a set T of size $O(k/\varepsilon)$ in $\mathbb{R}$ in $O(nk \log(nk))$ time. The size $O(k/\varepsilon)$ is essentially tight since there may be k large enough modes of these distributions, each requiring $\Omega(1/\varepsilon)$ points to represent. In $\mathbb{R}^2$ our bound on $|T|$ is $O(k^2/\varepsilon^2)$ under similar assumptions, or $O(d/\varepsilon^2)$ in $\mathbb{R}^d$ when we don't need to cover sets of medians m_Q which occur with probability less than ε. Then we can map weights onto this support set T exactly in $O(n^2k)$ time in $\mathbb{R}^1$ or approximately in either case in $O(1/\varepsilon^2)$ time.

Another goal may be to then construct another single-point estimator of these distributions: the median of these median distributions. In $\mathbb{R}^1$ we can show that this process is stable up to $\mathsf{cost}(m_Q, Q)$ where m_Q is the resulting single-point estimate. However, in either case, we also show that such single point estimates are not stable with respect to the weights in the median distribution, and then hence not stable with respect to the probability of any possible location of an uncertain point. That is, infinitesimal changes to such probabilities can greatly change the location of the single-point estimator. As such, we argue the approximate median distribution (which is stable with respect to these changes) is the best robust representation of such data.

Formalization of model and notation. We consider a set of n locationally uncertain points $\mathcal{P} = \{P_1, \ldots, P_n\}$ so that each P_i has k possible locations $\{p_{i,1}, \ldots, p_{i,k}\} \subset \mathbb{R}^d$. Let $P_{\mathsf{flat}} = \cup_i \{p_{i,1}, \ldots, p_{i,k}\}$ represent all positions of all points in $\mathcal{P}$. We consider each $p_{i,j}$ to be an equally likely (with probability $1/k$)

location of P_i, and can extend our techniques to non-uniform probabilities and uncertain points with fewer than k possible locations. For an uncertain point set $\mathcal{P}$ we say $Q \Subset \mathcal{P}$ is a *traversal* of $\mathcal{P}$ if $Q = \{q_1, \ldots q_n\}$ has each q_i in the domain of P_i (e.g., $q_i = p_{i,j}$ for some j).

We are particularly interested in the case where n can be quite large, but k could be small. For technical simplicity we assume here that the number k^n (the number of possible traversals of point sets) can be computed in $O(1)$ time and fit in $O(1)$ words of space under an extended version of the RAM model.

Given a set $Q = \{q_1, q_2, \ldots, q_n\} \subset \mathbb{R}$ that w.l.o.g. satisfies $q_1 \leq q_2 \leq \ldots \leq q_n$, we define the *median* m_Q as $q_{\frac{n+1}{2}}$ when n is odd and $q_{\frac{n}{2}}$ when n is even. There are several ways to generalize the median to higher dimensions [8], herein we focus on the L_1 median. Define $\mathsf{cost}(p, Q) = \frac{1}{n}\sum_{i=1}^{n} \|p - q_i\|$ where $\|\cdot\|$ is the Euclidian norm. Given a set $Q = \{q_1, q_2, \ldots, q_n\} \subset \mathbb{R}^d$, the L_1 *median* is defined as $m_Q = \arg\min_{p \in \mathbb{R}^d} \mathsf{cost}(p, Q)$. It is typically computed approximately using iterative [31] or other discrete approaches [11, 10]; its true solution may not have a closed form [9].

Main ideas. Since there are k^n possible traversals $Q \Subset \mathcal{P}$, we want to avoid enumerating all of them. Moreover, given a point $x \in T$, we need to determine which other possible m_Q for $Q \Subset \mathcal{P}$ are ε-approximated by x. To do this we introduce a function $\hat{\mathsf{cost}}(x) \leq \mathsf{cost}(x, Q)$ for any Q; it is defined

$$\hat{\mathsf{cost}}(x) = \frac{1}{n}\sum_{i=1}^{n} \min_{1 \leq j \leq k} \|x - p_{i,j}\|.$$

This function can be computed efficiently for all $p_{i,j} \in P_{\mathsf{flat}}$, and then be used as a conservative proxy for cost. We then create a small set T using greedy approaches which are within a constant factor of optimal. The Lipschitz property of cost and $\hat{\mathsf{cost}}$ are essential for the analysis; this property is imperative for robust loss functions (e.g., L_1 and Huber), but not present in non-robust ones like L_2.

Calculating the weights $\hat{w} : T \to [0, 1]$, is easy once we know the probability each point $p_{i,j} \in P_{\mathsf{flat}}$ is the median. We devise a dynamic program to calculate these weights in $\mathbb{R}^1$, that works by carefully tracking the expansion of a polynomial. In $\mathbb{R}^2$ we can no longer use the fact that each m_Q is some $p_{i,j} \in P_{\mathsf{flat}}$. Instead our high probability solution randomly instantiates traversals $Q \Subset \mathcal{P}$, computes their L_1 medians m_Q, and builds an approximate probability distribution from the result.

1.1 Related Work on Uncertain Data

The algorithms and computational geometry communities have recently generated a large amount of research in trying to understand how to efficiently process and represent uncertain data [15, 24, 21, 6, 19, 22, 5, 3, 4, 32, 1], not to mention some motivating systems and other progress from the database community [7, 27, 17, 16, 14, 13]. Some work in this area considers other models, with either worst-case representations of the data uncertainty [29] which do not naturally

allow probabilistic models, or when the data may not exist with some probability [19, 22, 6]. The second model can often be handled as a special case of the locationally uncertain model we study. Among locationally uncertain data, most work focuses on data structures for easy data access [12, 16, 28, 32, 4] but not the direct analysis of data. Among the work on analysis and summarization, such as for histograms [13], convex hulls [6], or clustering [15] it usually focuses on quantities like the expected or most likely value, which may not be stable with respect to noise. This includes estimation of the expected median in a stream of uncertain data [20] or the expected L_1 median as part of k-median clustering of uncertain data [15]. We are are not aware of any work on modeling the probabilistic nature of locationally uncertain data to construct robust estimators of that data, robust to outliers in both the set of uncertain points as well as probability distribution of each uncertain point.

2 Approximating the Median Distribution Support

In this section we describe how to construct T an approximate support of the median distribution. Recall that given a set of uncertain points $\mathcal{P}$, the set T should have the property that for every median m_Q of every traversal $Q \Subset \mathcal{P}$, there exists some $x \in T$ such that $\|x - m_Q\| \leq \varepsilon \mathsf{cost}(m_Q, Q)$, for a chosen error parameter $\varepsilon > 0$.

We first observe in $\mathbb{R}^1$ that $T \subset \mathcal{P}_{\mathsf{flat}}$ since we have defined the median so it must be one of the data points; hence $|T| \leq nk$. We then show how to reduce $|T|$ to $O(k/\varepsilon)$ under reasonable assumptions on how $\mathcal{P}$ is generated. In $\mathbb{R}^2$ we can construct T which is within a constant factor of the optimal size and at most $O(k^2/\varepsilon^2)$ under similar assumptions. Later, in Section 3.1, in $\mathbb{R}^d$ we show a randomized construction of size $O(d/\varepsilon^2)$ with weaker covering guarantees.

cost approximation. The key to these constructions is the function $\hat{\mathsf{cost}}(x) = \frac{1}{n}\sum_{i=1}^{n} \min_{1 \leq j \leq k} \|x - p_{i,j}\|$, which clearly satisfies $\hat{\mathsf{cost}}(m_Q) \leq \mathsf{cost}(m_Q, Q)$ for any $Q \Subset \mathcal{P}$. The following important lemma relates $\hat{\mathsf{cost}}(x)$ to $\mathsf{cost}(m_Q, Q)$.

Lemma 1. *For any $Q \Subset \mathcal{P}$, $x \in \mathbb{R}$, $\varepsilon > 0$, if $|x - m_Q| \leq \frac{\varepsilon}{1+\varepsilon}\hat{\mathsf{cost}}(x)$, then we have $|x - m_Q| \leq \varepsilon\hat{\mathsf{cost}}(m_Q) \leq \varepsilon\mathsf{cost}(m_Q, Q)$ where m_Q is the median of Q.*

Proof. We can use the Lipschitz property

$$|\hat{\mathsf{cost}}(x) - \hat{\mathsf{cost}}(y)| \leq |x - y|, \qquad \forall\, x, y \in \mathbb{R},$$

to show

$$\hat{\mathsf{cost}}(x) - \hat{\mathsf{cost}}(m_Q) \leq |\hat{\mathsf{cost}}(x) - \hat{\mathsf{cost}}(m_Q)| \leq |x - m_Q| \leq \frac{\varepsilon}{1+\varepsilon}\hat{\mathsf{cost}}(x),$$

which implies

$$\frac{1}{1+\varepsilon}\hat{\mathsf{cost}}(x) = (1 - \frac{\varepsilon}{1+\varepsilon})\hat{\mathsf{cost}}(x) \leq \hat{\mathsf{cost}}(m_Q).$$

Rearranging these expressions we can show the following, as desired,

$$|x - m_Q| \leq \frac{\varepsilon}{1+\varepsilon}\hat{\mathsf{cost}}(x) \leq \varepsilon\hat{\mathsf{cost}}(m_Q).\square$$

2.1 Computing T in $\mathbb{R}^1$

To compute T, we first observe that we can compute $\hat{\mathsf{cost}}(p_{i,j})$ for all $p_{i,j} \in P_{\mathsf{flat}}$ in $O(nk\log(nk))$ time. $\hat{\mathsf{cost}}$ has at most $n(2k-1)$ critical points where it is not differentiable: It is the sum of n functions $\hat{\mathsf{cost}}_i(x) = \min_{1\leq j\leq k}\|x - p_{i,j}\|$. Each $\hat{\mathsf{cost}}_i(x)$ is the lower envelope of k functions each with a single critical point at $x = p_{i,j}$, and when the lower envelope transitions between two consecutive functions at $(p_{i,j} + p_{i,j+1})/2$ where $p_{i,j}$ and $p_{i,j+1}$ are adjacent in sorted order. We can compute all of these critical points $\widetilde{P}_{\mathsf{flat}} = \cup_{i=1}^{n}\widetilde{P}_i$ in $O(nk)$ time after sorting in $O(nk\log(nk))$ time. Furthermore, we can calculate the value $\hat{\mathsf{cost}}(\tilde{p})$ for all $\tilde{p} \in \widetilde{P}_{\mathsf{flat}}$ in another $O(nk)$ time by scanning the points from smallest to largest, and maintaining $\hat{\mathsf{cost}}_i(x)$ for each i.

Now on the basis of Lemma 1, we can use a greedy algorithm to construct an ε-approximation of the support of $\mathcal{P}$ with T. After taking the smallest valued point ($p_1 \in P_{\mathsf{flat}}$), and setting it to x, it recursively takes the next smallest point $p_i \in P_{\mathsf{flat}}$ such that $p_i > x + \frac{\varepsilon}{1+\varepsilon}\hat{\mathsf{cost}}(x)$, and sets $p_i = x$. Sorting P_{flat} takes $O(nk\log(nk))$ time, and we have used $O(nk\log(nk))$ time to compute and store $\hat{\mathsf{cost}}(p_i)$ for all $p_i \in P_{\mathsf{flat}}$, so in all it takes $O(nk\log(nk))$ time.

Size of T. We now analyze the size of T as a function of n, k, ε. Ideally, we would like it to show that $|T|$ depends only on complexity of the distributions k and the error ε; we show this holds under some reasonable assumptions. In fact, if there exists a constant $\alpha > 0$ such that $\min_{x\in[0,L]}\hat{\mathsf{cost}}(x) \geq \frac{L}{\alpha k}$, then the distances between points in T is at least $\frac{\varepsilon L}{(1+\varepsilon)\alpha k}$. From our construction of T we immediately have the following theorem.

Theorem 1. *Suppose $P_{\mathsf{flat}} \subset [0, L]$, T is as described above and there exists a constant $\alpha > 0$ such that $\min_{x\in[0,L]}\hat{\mathsf{cost}}(x) \geq \frac{L}{\alpha k}$, then we have $|T| \leq \alpha k\frac{1+\varepsilon}{\varepsilon} = O(\alpha k/\varepsilon)$.*

Intuitively, this condition on α says that we cannot have some traversal $Q \Subset \mathcal{P}$ such that all points in Q are very close together relative to L, the diameter of P_{flat}. Moreover, the use of L is for convenience of formal proof statements; a set of κ outliers beyond the range $[0, L]$ can clearly be covered by κ additional points to T (or fewer since if $\kappa < k$, then $\hat{\mathsf{cost}}$ will be very high).

Moreover, we observe two common situations where the (α, L)-assumption holds. First, if some uncertain points P_i are disjoint and well-separated from each other (e.g. for most pairs $i \neq i'$ the convex hull of $\{p_{i,1}, p_{i,2}, \ldots, p_{i,k}\}$ is disjoint with a sufficient separation from the convex hull of $\{p_{i',1}, p_{i',2}, \ldots, p_{i',k}\}$), then α will be sufficiently small, since $\hat{\mathsf{cost}}$ will be at least that gap over k. Second, if each discrete set of locations for P_i is drawn iid from the some (reasonably

bounded) distribution, which could be different from each other $P_{i'}$, then again α will be sufficiently small with high probability.

We say a random variable X is C_0-*bounded* if its the cumulative distribution function (cdf)

$$F(t) = \begin{cases} 0 & \text{if } t < 0 \\ \varphi(t) & \text{if } 0 \le t \le L \\ 1 & \text{if } t > L \end{cases},$$

satisfies[1] $\varphi \in C([0,L]) \cap W^{1,\infty}([0,L])$ such that $\varphi(0) = 0$, $\varphi(L) = 1$ and $\|\varphi'\|_{L_\infty([0,L])} \le C_0$ (i.e., the slope if its cdf F is at most C_0). We now provide the following technical lemma, proved in Appendix A.

Lemma 2. *If each P_i is drawn iid from a distribution represented by C_0-bounded random variable X_i (each X_i could be different) with domain $[0,L]$, $\alpha = LC_0$, and $n > 8\alpha^2(k+1)^2\ln(2/\delta)$, then with probability at least $1-\delta$ we have $\min_x \hat{\mathsf{cost}}(x) \ge 1/(4C_0(k+1))$ and $|T| = O(LC_0k/\varepsilon) = O(\alpha k/\varepsilon)$.*

Now we argue that the value LC_0 is typically constant since it is reasonable for random variables to be C_0-bounded over the domain $[0,L]$. For example a uniform random variable over $[0,L]$ has $C_0 = 1/L$ and $LC_0 = 1$. More generally, distributions with all values at most a constant η times as likely as they would be under a uniform random variable have $LC_0 = \eta$. Basically, a distribution is only not C_0-bounded for a constant C_0 if it has a non-zero probability of instantiating at a specific point.

2.2 The Construction of T in $\mathbb{R}^2$

Given n uncertain points $\mathcal{P}$ in $\mathbb{R}^2$, two sets $A_1, A_2 \subset \mathbb{R}^2$, and $a_1 \in A_1$, $a_2 \in A_2$, for $\varepsilon > 0$ we say a_1 can ε-*cover* a_2 if $\|a_1 - a_2\| \le \varepsilon\hat{\mathsf{cost}}(a_1)$; A_1 can ε-*cover* A_2 if for any $a_2 \in A_2$ there exists $a_1 \in A_1$ such that $\|a_1 - a_2\| \le \varepsilon\hat{\mathsf{cost}}(a_1)$.

For $\hat{\mathsf{cost}}(\cdot)$ defined on $\mathbb{R}^2$, we can use the method in the proof Lemma 1 to obtain a similar result: for any $Q \Subset \mathcal{P}$, $x \in \mathbb{R}^2$, for $\varepsilon > 0$ if $\|x - m_Q\| \le \frac{\varepsilon}{1+\varepsilon}\hat{\mathsf{cost}}(x)$ then $\|x - m_Q\| \le \varepsilon\mathsf{cost}(m_Q, Q)$, where m_Q is the L_1 median of Q. Therefore, if we find a set T which can $(\frac{\varepsilon}{1+\varepsilon})$-cover $CH(P_{\mathsf{flat}})$, the convex hull of P_{flat}, then T should be an $(\frac{\varepsilon}{1+\varepsilon})$-cover of the set of all possible medians, as desired. However, $CH(P_{\mathsf{flat}})$ is an infinite set, and we only want to cover finite points. To solve this problem, we assume $\min_{x\in\mathbb{R}^2} \hat{\mathsf{cost}}(x) \ge \varrho(\mathcal{P}) > 0$, and define the lattice

$$S(\mathcal{P}) = CH(P_{\mathsf{flat}}) \cap \{(\beta i, \beta j) \mid i, j \in \{0, \pm 1, \pm 2, \cdots\}\} \tag{1}$$

where $\beta = \frac{\varepsilon}{2\sqrt{2}(1+\varepsilon)}\varrho(\mathcal{P})$. From the definition of $S(\mathcal{P})$ we know, for any $x \in CH(P_{\mathsf{flat}})$ there exists $s \in S(\mathcal{P})$ such that $\|s - x\| \le \sqrt{2}\beta = \frac{\varepsilon}{2(1+\varepsilon)}\varrho(\mathcal{P})$. If T can

[1] $C([0,L]) = \{f : [0,L] \mapsto \mathbb{R} \mid f \text{ is continuous}\}$, and Sobolev space $W^{1,\infty}([0,L]) = \{g : [0,L] \mapsto \mathbb{R} \mid g \text{ is weakly differentiable and } g' \in L_\infty([0,L])\}$ (cf. [2]).

$\frac{\varepsilon}{2(1+\varepsilon)}$-cover $S(\mathcal{P})$, then there exists $z \in T$ such that $\|z-s\| \leq \frac{\varepsilon}{2(1+\varepsilon)}\hat{\mathsf{cost}}(z)$. So, we have

$$\|z-x\| \leq \|z-s\|+\|s-x\| \leq \frac{\varepsilon}{2(1+\varepsilon)}\hat{\mathsf{cost}}(z)+\frac{\varepsilon}{2(1+\varepsilon)}\varrho(\mathcal{P}) \leq \frac{\varepsilon}{(1+\varepsilon)}\hat{\mathsf{cost}}(z) \quad (2)$$

which implies T can $(\frac{\varepsilon}{1+\varepsilon})$-cover $CH(P_{\mathsf{flat}})$, so we only need to find a set T to $\frac{\varepsilon}{2(1+\varepsilon)}$-cover $S(\mathcal{P})$.

After constructing $CH(P_{\mathsf{flat}})$ in $O(nk\log(nk))$ time, to compute $S(\mathcal{P})$ we need a $\hat{\mathsf{cost}}$ lower bound $\varrho(\mathcal{P})$. It can be obtained in $O(nk^2)$ time according to Lemma 3, by considering only any one uncertain point, or the bound can be improved by a factor n by considering all uncertain points in $O(n^2k^2)$ time.

Lemma 3. *Given a set of n uncertain points $\mathcal{P} = \{P_1, \ldots, P_n\}$, if $\varrho(\mathcal{P}) = \frac{1}{n+1}\min_{1\leq j\leq k}\hat{\mathsf{cost}}(p_{1,j})$, then we have $\min_{x\in\mathbb{R}^2}\hat{\mathsf{cost}}(x) \geq \varrho(\mathcal{P})$.*

Proof. Suppose $x^* = \arg\min_{x\in\mathbb{R}^2}\hat{\mathsf{cost}}(x)$, and

$$\hat{\mathsf{cost}}(x^*) = \frac{1}{n}\sum_{i=1}^{n}\|x^*-p_{i,j_i}\| \quad \text{where} \quad j_i = \arg\min_{1\leq j\leq k}\|x^*-p_{i,j}\|, \forall\, i \in \{1,2,\cdots,n\}.$$

Thus using the Lipschitz property of $\hat{\mathsf{cost}}$

$$\hat{\mathsf{cost}}(p_{1,j_1}) - \hat{\mathsf{cost}}(x^*) \leq |\hat{\mathsf{cost}}(p_{1,j_1}) - \hat{\mathsf{cost}}(x^*)| \leq \|x^* - p_{1,j_1}\| \leq n\hat{\mathsf{cost}}(x^*)$$

which implies $\hat{\mathsf{cost}}(p_{1,j_1}) \leq (n+1)\hat{\mathsf{cost}}(x^*)$. So, we obtain

$$\min_{x\in\mathbb{R}^2}\hat{\mathsf{cost}}(x) = \hat{\mathsf{cost}}(x^*) \geq \frac{1}{n+1}\hat{\mathsf{cost}}(p_{1,j_1}) = \varrho(\mathcal{P}).\square$$

Now, to construct T, we arbitrarily add points from S to T one at a time, among the points in S which are not already $(\frac{\varepsilon}{2(1+\varepsilon)})$-covered by other points in T; details are provided in Algorithm B.1 in Appendix B. We can show (see Appendix B) the set T has size within a constant factor of the size of the optimal such domain T^*. Let R be half the diameter of $\mathcal{P}_{\mathsf{flat}}$. Alternatively, we can show (in Appendix C) the size of T is at most $O(\alpha k^2/\varepsilon^2)$; here $\alpha^2 = (R/k)/\min_{x\in CH(P_{\mathsf{flat}})}\hat{\mathsf{cost}}(x)$. Moreover, we can show under similar C_0-bounded assumption on n distributions from which each P_i is drawn iid, that roughly $\alpha = C_0^2R^2$; again under reasonable assumptions (as in $\mathbb{R}^1$), we can assume α is constant.

3 Assigning a Weight to T

In this section, we show how to assign a weight to T which approximates the probability distribution of medians. We provide an optimal algorithm in $\mathbb{R}^1$, and an unrestricted randomized algorithm.

Define the weight of $p_{i,j} \in P_{\mathsf{flat}}$ as $w(p_{i,j}) = \frac{1}{k^n}|\{Q \Subset \mathcal{P} \mid p_{i,j} \text{ is the median of Q}\}|$. Suppose T is constructed by our greedy algorithm for $\mathbb{R}^1$. For $p_{i,j} \in P_{\mathsf{flat}}$, letting $x = \max_{\tilde{x} \in T}\{\tilde{x} \le p_{i,j}\}$ and $y = \min_{\tilde{y} \in T}\{\tilde{y} > p_{i,j}\}$, we introduce a map $f_T : P_{\mathsf{flat}} \to T$,

$$f_T(p_{i,j}) = \begin{cases} y & \text{if } |p_{i,j} - y| \le \frac{\varepsilon}{1+\varepsilon}\hat{\mathsf{cost}}(y) \text{ and } |p_{i,j} - y| < |p_{i,j} - x| \\ x & \text{otherwise.} \end{cases}$$

Intuitively, this maps each $p_{i,j} \in P_{\mathsf{flat}}$ onto the closest point $x \in T$, unless it violates the ε-approximation property which another further point satisfies.

Now for each $x \in T$, define weight of x as $\hat{w}(x) = \sum_{\{p_{i,j} \in P_{\mathsf{flat}} | f_T(p_{i,j}) = x\}} w(p_{i,j})$. So we first compute the weight of each point in P_{flat} and then obtain the weight of points in T on another linear sweep. Our ability to calculate the weights w for each point in P_{flat} is summarized in the next lemma, with corollary about $\hat{w}$ following. The algorithm, explained in detail within the proof, is a dynamic program that expands a specific polynomial, where in the final state, the coefficients correspond with the probability of each point being the median.

Lemma 4. *We can outputs $w(p_{i,j})$ for all points in P_{flat} in $\mathbb{R}^1$ in $O(n^2k)$ time.*

Proof. For any $p_{i_0} \in P_{i_0}$, we define

$$l_j = \begin{cases} |\{p \in P_j \mid p \le p_{i_0}\}| & \text{if } 1 \le j \le i_0 - 1 \\ |\{p \in P_{j+1} \mid p \le p_{i_0}\}| & \text{if } i_0 \le j \le n-1 \end{cases}, \quad r_j = \begin{cases} |\{p \in P_j \mid p \ge p_{i_0}\}| & \text{if } 1 \le j \le i_0 - 1 \\ |\{p \in P_{j+1} \mid p \ge p_{i_0}\}| & \text{if } i_0 \le j \le n-1 \end{cases}.$$

Then, if n is odd, we have

$$w(p_{i_0}) = \frac{1}{k^n} \sum_{\substack{S_1 \cap S_2 = \emptyset \\ S_1 \cup S_2 = \{1,\cdots,n-1\}}} (l_{i_1} \cdot l_{i_2} \cdot \ldots \cdot l_{i_{\frac{n-1}{2}}} \cdot r_{j_1} \cdot r_{j_2} \cdot \ldots \cdot r_{j_{\frac{n-1}{2}}})$$

where $S_1 = \{i_1, i_2, \cdots, i_{\frac{n-1}{2}}\}$ and $S_2 = \{j_1, j_2, \cdots, j_{\frac{n-1}{2}}\}$, and if n is even, we have

$$w(p_i) = \frac{1}{k^n} \sum_{\substack{S_1 \cap S_2 = \emptyset \\ S_1 \cup S_2 = \{1,\cdots,n-1\}}} (l_{i_1} \cdot l_{i_2} \cdot \ldots \cdot l_{i_{\frac{n}{2}-1}} \cdot r_{j_1} \cdot r_{j_2} \cdot \ldots \cdot r_{j_{\frac{n}{2}}})$$

where $S_1 = \{i_1, i_2, \cdots, i_{\frac{n}{2}-1}\}$ and $S_2 = \{j_1, j_2, \cdots, j_{\frac{n}{2}}\}$.

We next describe the algorithm for n odd; the case for n even is similar. To compute $\sum_{\substack{S_1 \cap S_2 = \emptyset \\ S_1 \cup S_2 = \{1,\cdots,n-1\}}} (l_{i_1} \cdot l_{i_2} \cdot \ldots \cdot l_{i_{\frac{n-1}{2}}} \cdot r_{j_1} \cdot r_{j_2} \cdot \ldots \cdot r_{j_{\frac{n-1}{2}}})$, we construct the following polynomial:

$$(l_1 x + r_1)(l_2 x + r_2) \cdots (l_{n-1} x + r_{n-1}), \tag{3}$$

and $\sum_{\substack{S_1 \cap S_2 = \emptyset \\ S_1 \cup S_2 = \{1,\cdots,n-1\}}} (l_{i_1} \cdot l_{i_2} \cdot \ldots \cdot l_{i_{\frac{n-1}{2}}} \cdot r_{j_1} \cdot r_{j_2} \cdot \ldots \cdot r_{j_{\frac{n-1}{2}}})$ is the coefficient of $x^{\frac{n-1}{2}}$. We define $\rho_{i,j}$ $(1 \le i \le n-1, 0 \le j \le i)$ as the coefficient

of x^j in the polynomial $(l_1x + r_1)\cdots(l_ix + r_i)$ and then it is easy to check $\rho_{i,j} = l_i\rho_{i-1,j-1} + r_i\rho_{i-1,j}$, so we can use dynamic programming to compute $\rho_{n-1,0}, \rho_{n-1,1}, \cdots, \rho_{n-1,n-1}$.

Algorithm 3.1 Compute $\rho_{n-1,0}, \rho_{n-1,1}, \cdots, \rho_{n-1,n-1}$

Let $\rho_{1,0} = r_1, \rho_{1,1} = l_1, \rho_{1,2} = 0$.
for $i = 2$ to $n-1$ **do**
  for $j = 0$ to i **do**
    $\rho_{i,j} = l_i\rho_{i-1,j-1} + r_i\rho_{i-1,j}$
  $\rho_{i,i+1} = 0$
return $\rho_{n-1,0}, \rho_{n-1,1}, \cdots, \rho_{n-1,n-1}$.

In Algorithm 3.1, $\rho_{n-1,\frac{n-1}{2}} = \sum_{\substack{S_1\cap S_2=\emptyset \\ S_1\cup S_2=\{1,\cdots,n-1\}}} (l_{i_1}\cdot l_{i_2}\cdot\ldots\cdot l_{i_{\frac{n-1}{2}}}\cdot r_{j_1}\cdot r_{j_2}\cdot \ldots\cdot r_{j_{\frac{n-1}{2}}})$. Suppose for $p_{i_0} \in P_{i_0}$ we have obtained $\rho_{n-1,0}, \rho_{n-1,1}, \ldots, \rho_{n-1,n-1}$ by Algorithm 3.1, and then we consider $p_{i'_0} = \min\{p \in P_{\mathsf{flat}} - P_{i_0} \mid p \geq p_{i_0}\}$. We assume $p_{i'_0} \in P_{i'_0}$, and if $i'_0 < i_0$, we construct a polynomial

$$(l_1x + r_1)\cdots(l_{i'_0-1}x + r_{i'_0-1})(\tilde{l}_{i'_0}x + \tilde{r}_{i'_0})(l_{i'_0+1}x + r_{i'_0+1})\cdots(l_{n-1}x + r_{n-1}) \quad (4)$$

and if $i'_0 > i_0$, we construct a polynomial

$$(l_1x + r_1)\cdots(l_{i'_0-2}x + r_{i'_0-2})(\tilde{l}_{i'_0-1}x + \tilde{r}_{i'_0-1})(l_{i'_0}x + r_{i'_0})\cdots(l_{n-1}x + r_{n-1}) \quad (5)$$

where $\tilde{l}_{i'_0} = \tilde{l}_{i'_0-1} = |\{p \in P_{i_0} \mid p \leq p_{i'_0}\}|$, $\tilde{r}_{i'_0} = \tilde{r}_{i'_0-1} = |\{p \in P_{i_0} \mid p \geq p_{i'_0}\}|$.

It is easy to check, the weight of $p_{i'_0}$ is $w(p_{i'_0}) = \frac{1}{k^n}\tilde{\rho}_{n-1,\frac{n-1}{2}}$ where $\frac{1}{k^n}\tilde{\rho}_{n-1,\frac{n-1}{2}}$ is the coefficient of $x^{\frac{n-1}{2}}$ in (4) if $i'_0 < i_0$ or (5) if $i'_0 > i_0$. Since (3) and (4) have only one different factor, we obtain the coefficients of (4) from the coefficients of (3) in $O(n)$ time. We recover the coefficients of $(l_1x + r_1)\cdots(l_{i'-1}x + r_{i'-1})(l_{i'_0+1}x + r_{i'_0+1})\cdots(l_{n-1}x + r_{n-1})$ from $\rho_{n-1,0}, \rho_{n-1,1}, \cdots, \rho_{n-1,n-1}$, and then use these coefficients to compute the coefficients of (4). Similarly, if $i'_0 > i_0$, we obtain the coefficients of (5) from the coefficients of (3). Therefore, we can use $O(n^2)$ time to compute the weight of the first point in P_{flat} and then use $O(n)$ time to compute the weight of other points. The whole time is $O(n^2)+nkO(n) = O(n^2k)$. □

Corollary 1. *We can assign $\hat{w}(x)$ to each $x \in T$ in $\mathbb{R}^1$ in $O(n^2k)$ time.*

3.1 Simultaneous Randomized Domain T and Weight

Each point $m_Q \in \{m_Q \text{ is an } L_1 \text{ median of } Q \mid Q \Subset \mathcal{P}\}$ may take a distinct value. Thus even calculating that set, let alone their weights in the case of duplicates, would require at least $\Omega(k^n)$ time. Rather, here we show how to randomly endow the set T constructed in the previous section with approximate weights.

First define $M_{\mathcal{P}} = \{q \text{ is the } L_1 \text{ median of } Q | \ Q \Subset \mathcal{P}\}$, and the map $f_T : M_{\mathcal{P}} \mapsto T$:

$$f_T(q) = \arg\min \left\{ \|q - z\| \mid z \in T, z \text{ can } \frac{\varepsilon}{1+\varepsilon}\text{-cover } q \right\}$$

The minimum value point in $\left\{ \|q - z\| \mid z \in T, z \text{ can } \frac{\varepsilon}{1+\varepsilon}\text{-cover } q \right\}$ may be not unique, in this case, we choose $z = \arg\min \left\{ \|q - z\| \mid z \in T, z \text{ can } \frac{\varepsilon}{1+\varepsilon}\text{-cover } q \right\}$ with minimum coordinates as the value of $f_T(q)$, to ensure the uniqueness of $f_T(q)$. Being more careful, in $\mathbb{R}^d$ for $d > 1$ we cannot compute q exactly, but can within any factor $\phi < \varepsilon$ [31, 11, 10]. So we need to relax the above definitions so that q is the ϕ-approximate result, and then we ensure that z can $(\frac{\varepsilon-\phi}{1+\varepsilon-\phi})$-cover q in the map f_T. This will still provide a valid cover for our purposes and does not affect the resulting bounds. For simplicity, we omit the discussion of ϕ from the remaining description.

Now, for $i \in \{1, 2 \cdots, m\}$, we define the weight of z_i

$$\hat{w}_i = \sum_{q \in M_{\mathcal{P}}, f_T(q) = z_i} w(q)$$

where $w(q) = \frac{1}{k^n} |\{Q \Subset \mathcal{P} \mid q \text{ is the } L_1 \text{ median of } Q\}|$.

Our goal will be to approximate these weights, for which we use Algorithm 3.2 to obtain approximate values $\hat{w}_i$. Initially let $c_i = 0$ be the weight for each $z_i \in T$, and proceed in a series of $N = O(\frac{1}{\varepsilon^2}(d + \log \frac{1}{\delta}))$ rounds (in $\mathbb{R}^d$). In each round, we create a random traversal $Q \Subset \mathcal{P}$, compute its (ϕ-approximate) L_1 median m_Q, and assign and increment by $1/N$ the weight of the (appropriately defined) $z_j \in T$ which ε-covers m_Q.

Algorithm 3.2 Approximate the weight of points in T

Input: $\mathcal{P}$, $\frac{\varepsilon}{1+\varepsilon}\hat{\mathsf{cost}}(z_i)$ for each $z_i \in T$, and a positive integer N
Initialize $c_i = 0$ for $i = 1, 2, \cdots, m$
for $j = 1$ to N **do**
  Randomly choose $Q \Subset \mathcal{P}$
  $q =$ the L_1 median of Q
  $z_i = f_T(q)$
  $c_i = c_i + 1$
return $\frac{c_i}{N}$ as the approximate value of $\hat{w}_i$ for $i = 1, 2, \cdots, m$

Below in Theorem 2, we show that the approximated weight of each $z_i \in T$ is within ε of what should be its true weight with probability at least $1 - \delta$ (via straight-forward application of a VC-dimension theory [30, 23]). Alternatively, we can skip the construction of T and simply let the set of medians $\{m_Q\}$ constructed through this iterative process represent T. Note that in both cases, points $z_j \in T$ with $\hat{w}(z_j) \leq \varepsilon$ might be given a weight 0 and not be part of T, even if they are required to cover a median m_Q which may occur, albeit, with a very small probability.

Theorem 2. *Suppose $\varepsilon > 0$, $\delta \in (0,1)$ and $\{c_1, c_2, \cdots, c_m\}$ is obtained from Algorithm 3.2. If in Algorithm 3.2 $N = O((1/\varepsilon^2)(d + \log(\frac{1}{\delta})))$, then we have*

$$\mathbf{Pr}\left[\max_{i\in[m]} |c_i - \hat{w}_i| \leq \varepsilon\right] > 1 - \delta.$$

Proof. We use Vapnik-Chervonenkis theory [30], considering a family $\mathcal{R}$ of queries (with bounded VC-dimension, for instance balls of any size in $\mathbb{R}^d$ have VC-dimension $\nu = O(d)$), for instance let $\mathcal{R}$ be all balls. Now consider any probability distribution μ defined over $\mathbb{R}^d$ and $N = O((1/\varepsilon^2)(\nu + \log(1/\delta)))$ iid samples X from μ. Then we know [23] that, with probability at least $1 - \delta$, the empirical distribution μ_X defined by X satisfies

$$\max_{R\in\mathcal{R}} |R(\mu) - R(\mu_X)| \leq \varepsilon,$$

where $R(\mu)$ is the measure of μ restricted to that range R. For discrete distributions μ defined on some point set Z, we can consider balls $R \in \mathcal{R}$ small enough to distinguish each $z \in Z$.

Now to complete the proof, we simply realize that each step of the for loop in Algorithm 3.2 constructs q which is a iid random sample from the distribution of medians of $\mathcal{P}$. Thus the set of these points constitutes X in the above VC-dimension result, and the claim follows. □

4 Constructing a Single Point Estimate

Given a discrete domain $X \subset \mathbb{R}^1$ (a point set) and a probability distribution defined by function $\omega : X \to [0,1]$, we can compute its weighted median. Assuming X is sorted, this takes $O(|X|)$ time by scanning from smallest to largest until the sum of weights reaches 0.5.

There are two situations whereby we obtain such a discrete weighted domain. First is the set T described by the greedy approximation algorithm from Section 2.1, and the resulting weight $\hat{w}$ from Section 3. Let the resulting single point estimate be m_T. The second domain is the set P_{flat} of all possible locations of $\mathcal{P}$, and its weight w where $w(p_{i,j})$ is the fraction of $Q \Subset \mathcal{P}$ which take $p_{i,j}$ as their median (possibly 0). Let the resulting single point estimate be $m_{\mathcal{P}}$.

Theorem 3. *$|m_T - m_{\mathcal{P}}| \leq \varepsilon \hat{\mathsf{cost}}(m_{\mathcal{P}}) \leq \varepsilon \mathsf{cost}(m_Q, Q)$, $Q \Subset \mathcal{P}$ is any traversal with m_P as its median.*

Proof. We can divide $\mathbb{R}$ into $|T|$ intervals, one associated with each $x \in T$, as follows. Each $z \in \mathbb{R}$ is in an interval associated with $x \in T$ if z is closer to x than any other point $y \in T$, unless $|z - y| \leq \frac{\varepsilon}{1-\varepsilon}\hat{\mathsf{cost}}(y)$ but $|z - x| \geq \frac{\varepsilon}{1-\varepsilon}\hat{\mathsf{cost}}(x)$. Thus a point $p_{i,j}$ whose weight $w(p_{i,j})$ contributes to $\hat{w}(x)$, is in the interval associated with x.

Thus, if $p_{i,j} = m_{\mathcal{P}}$, then all weights of all points greater than $p_{i,j}$ is at most 0.5, and all weights of points less than $p_{i,j}$ is less than 0.5. Hence if $m_{\mathcal{P}}$ is in

an interval associated with $x \in T$, then the sum of all weights of points $p_{i,j}$ in intervals greater than that of x must be at most 0.5 and those less than that of x must be less than 0.5. Hence $m_T = x$, and $|x - p_{i,j}| \leq \frac{\varepsilon}{1+\varepsilon}\hat{\mathsf{cost}}(x) \leq \varepsilon\hat{\mathsf{cost}}(m_{\mathcal{P}})$ as desired. □

4.1 Non-Robustness of Single Point Estimates

Unfortunately, the L_1 median of the set $\{m_Q \text{ is an } L_1 \text{ median of } Q \mid Q \Subset \mathcal{P}\}$ is not stable under small perturbations in weights; it stays within the convex hull of the set, but otherwise not much can be said, even in $\mathbb{R}^1$. Consider the example with $n = 3$ and $k = 2$, where $p_{1,1} = p_{1,2} = p_{2,1} = 0$ and $p_{2,2} = p_{3,1} = p_{3,2} = \Delta$ for some arbitrary Δ. The median will be at 0 or Δ, each with probability 1/2, depending on the location of P_2. We can also create a more intricate example where $\hat{\mathsf{cost}}(0) = \hat{\mathsf{cost}}(\Delta) = 0$. As these examples have m_Q at 0 or Δ equally likely with probability 1/2, then canonically in $\mathbb{R}^1$ we would have the median of this distribution at 0, but a slight change in probability (say from sampling) could put it all the way at Δ. This indicates that a representation of the distribution of medians (as we provide in Sections 2 and 3) is more appropriate for noisy data.

5 Conclusion

We initiate the study of robust estimators for uncertain data, by studying the L_1 median on locationally uncertain data points. We show how to efficiently create approximate distributions for the location of these medians in $\mathbb{R}^1$, and generalize these approaches to $\mathbb{R}^2$, and also via a simple randomized algorithm to $\mathbb{R}^d$. We also argue that although we can use such distributions to calculate a single-point representation of these distributions, it is not very stable to the input distributions, and serves as a poor representation when the true scenario is multi-modal; hence further motivating our distributional approach.

References

1. A. Abdullah, S. Daruki, and J. M. Phillips. Range counting coresets for uncertain data. In *SOCG*, 2013.
2. R. A. Adams. *Sobolev Spaces 2 edition*. Academic Press, 2003.
3. P. K. Agarwal, B. Aronov, S. Har-Peled, J. M. Phillips, K. Yi, and W. Zhang. Nearest-neighbor searching under uncertainty II. In *PODS*, 2013.
4. P. K. Agarwal, S.-W. Cheng, Y. Tao, and K. Yi. Indexing uncertain data. In *PODS*, 2009.
5. P. K. Agarwal, A. Efrat, S. Sankararaman, and W. Zhang. Nearest-neighbor searching under uncertainty. In *PODS*, 2012.
6. P. K. Agarwal, S. Har-Peled, S. Suri, H. Yildiz, and W. Zhang. Convex hulls under uncertainty. In *ESA*, 2014.
7. P. Agrawal, O. Benjelloun, A. D. Sarma, C. Hayworth, S. Nabar, T. Sugihara, and J. Widom. Trio: A system for data, uncertainty, and lineage. In *PODS*, 2006.

8. G. Aloupis. Geometric measures of data depth. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*. AMS, 2006.
9. C. Bajaj. The algebraic degree of geometric optimization problems. *Discrete and Computational Geometry*, 3:177–191, 1988.
10. P. Bose, A. Maheshwari, and P. Morin. Fast approximations for sums of distances clustering and the Fermet-Weber problem. *CGTA*, 24:135–146, 2003.
11. R. Chandrasekaran and A. Tamir. Algebraic optimization: The Fermet-Weber location problem. *Mathematical Programming*, 46:219–224, 1990.
12. R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. S. Vitter. Efficient indexing methods for probabilistic threshold queries over uncertain data. In *VLDB*, 2004.
13. G. Cormode and M. Garafalakis. Histograms and wavelets of probabilitic data. In *ICDE*, 2009.
14. G. Cormode, F. Li, and K. Yi. Semantics of ranking queries for probabilistic data and expected ranks. In *ICDE*, 2009.
15. G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In *PODS*, 2008.
16. N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. In *VLDB*, 2004.
17. N. N. Dalvi, C. Ré, and D. Suciu. Probabilistic databases: Diamonds in the dirt. *Commun. ACM*, 52(7):86–94, 2009.
18. D. Donoho and P. J. Huber. The notion of a breakdown point. In P. Bickel, K. Doksum, and J. Hodges, editors, *A Festschrift for Erich L. Lehmann*, pages 157–184. Wadsworth International Group, 1983.
19. L. Huang and J. Li. Minimum spanning trees, perfect matchings and cycle covers over stochastic points in metric spaces. Technical report, arXiv:1209.5828, 2012.
20. T. Jayram, A. McGregor, S. Muthukrishnan, and E. Vee. Estimating statistical aggregates on probabilistic data streams. *ACM TODS*, 33:1–30, 2008.
21. A. G. Jørgensen, M. Löffler, and J. M. Phillips. Geometric computation on indecisive points. In *WADS*, 2011.
22. P. Kamousi, T. M. Chan, and S. Suri. Stochastic minimum spanning trees in euclidean spaces. In *SOCG*, 2011.
23. Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the samples complexity of learning. *Journal of Computer and System Science*, 62:516–527, 2001.
24. M. Löffler and J. Phillips. Shape fitting on point sets with probability distributions. In *ESA*, 2009.
25. H. P. Lopuhaa and P. J. Rousseeuw. Breakdown points of affine equivaniant estimators of multivariate location and converiance matrices. *The Annals of Statistics*, 19:229–248, 1991.
26. P. J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, pages 283–297, 1985.
27. A. D. Sarma, O. Benjelloun, A. Halevy, S. Nabar, and J. Widom. Representing uncertain data: models, properties, and algorithms. *VLDBJ*, 18:989–1019, 2009.
28. Y. Tao, R. Cheng, X. Xiao, W. K. Ngai, B. Kao, and S. Prabhakar. Indexing multi-dimensional uncertain data with arbitrary probability density functions. In *VLDB*, 2005.
29. M. van Kreveld and M. Löffler. Largest bounding box, smallest diameter, and related problems on imprecise points. *CGTA*, 43:419–433, 2010.
30. V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

31. E. Weiszfeld. Sur le point pour lequel la somme des distances de n points dennes est minimum. *Tohoku Mathmatics*, 43:355–386, 1937.
32. Y. Zhang, X. Lin, Y. Tao, W. Zhang, and H. Wang. Efficient computation of range aggregates against uncertain location based queries. *IEEE TKDE*, 24:1244–1258, 2012.

APPENDIX

A The proof of Lemma 2

Recall that $C([0,L]) = \{f : [0,L] \mapsto \mathbb{R} \mid f \text{ is continuous}\}$, and Sobolev space $W^{1,\infty}([0,L]) = \big\{g : [0,L] \mapsto \mathbb{R} \mid g \text{ is weakly differentiable and } g' \in L_\infty([0,L])\big\}$ (cf. [2]).

We first show the following lemma which depends on a couple of technical lemmas. We will show that this implies Lemma 2 and then return to prove the technical lemmas.

Lemma 5. *For each $i \in \{1,2,\ldots,n\}$, suppose $X_{i,1}, X_{i,2}, \ldots, X_{i,k}$ are independent and identically distributed C_0-bounded random variables. For any $x \in [0,L]$, $i \in \{1,2,\cdots,n\}$, suppose Y_i is given by $Y_i(x) = \min_{0\le j\le k}\{|x - X_{i,j}|\}$ and $Y_1, Y_2, \cdots, Y_n$ are mutually independent.*

If $\overline{Y}_n(x) = \frac{1}{n}\sum_{i=1}^n Y_i(x)$, then for any $\eta > 0$ and $\delta \in (0,1)$ we have

$$\mathbf{Pr}\Big[\min_{0\le x\le L} \overline{Y}_n(x) \ge \frac{1}{2C_0(k+1)} - \eta\Big] > 1-\delta, \quad \forall\ n > \frac{L^2}{2\eta^2}\ln\frac{2}{\delta}.$$

Proof. By Chernoff-Hoeffding inequality, for any $x \in [0,L]$, we have

$$\mathbf{Pr}\big[|\overline{Y}_n(x) - \mathbf{E}(\overline{Y}_n(x))| > \eta\big] \le 2\exp\big(\frac{-2\eta^2}{\sum_{i=1}^n(\frac{1}{n}L)^2}\big) = 2\exp(\frac{-2n\eta^2}{L^2}) < \delta.$$

where the last line follows from $n > \frac{L^2}{2\eta^2}\ln\frac{2}{\delta}$. For any fixed $x \in [0,L]$, the condition $|\overline{Y}_n(x) - \mathbf{E}(\overline{Y}_n(x))| \le \eta$ implies

$$\overline{Y}_n(x) \ge -\eta + \mathbf{E}(\overline{Y}_n(x)) \ge -\eta + \min_{0\le x'\le L}\mathbf{E}(\overline{Y}_n(x')) \ge -\eta + \frac{1}{2C_0(k+1)},$$

with the last step following from Lemma 7. Thus, we obtain

$$\mathbf{Pr}\big[-\eta + \frac{1}{2C_0(k+1)} \le \overline{Y}_n(x)\big] > 1-\delta, \quad \forall\ n > \frac{L^2}{2\eta^2}\ln\frac{2}{\delta},\ \forall\ x \in [0,L].$$

If the value of $X_{i,j}$ is given for all $1 \le i \le n$ and $1 \le j \le k$, then $\overline{Y}_n(x)$ is a continuous function of x on the closed interval $[0,L]$, so we can assume $\min_{0\le x\le L}\overline{Y}_n(x) = \overline{Y}_n(x_0)$ where $x_0 \in [0,L]$ depends on $X_{i,j}$. Combining these results, we prove the following as desired

$$\mathbf{Pr}\big[-\eta + \frac{1}{2C_0(k+1)} \le \overline{Y}_n(x_0) = \min_{0\le x\le L}\overline{Y}_n(x)\big] > 1-\delta, \quad \forall\ n > \frac{L^2}{2\eta^2}\ln\frac{2}{\delta}.$$

□

If for all $i \in \{1, 2, \ldots, n\}$ the set of possible locations $\{p_{i,1}, p_{i,2}, \ldots, p_{i,k}\}$ of each uncertain point P_i are drawn iid from separate C_0-bounded random variables (namely X_i), then we can use Lemma 5 to provide an upper bound for α, a lower bound for $\hat{\mathsf{cost}}(x)$, and hence upper bound on $|T|$, with probability at least $1-\delta$. In this case, the quantity $\bar{Y}_n(x) = \hat{\mathsf{cost}}(x)$, so we have $\mathbf{Pr}[\min_{0\le x\le L} \hat{\mathsf{cost}}(x) \ge \frac{1}{2C_0(k+1)} - \eta] \ge 1-\delta$ for n sufficiently large as $n > \frac{L^2}{2\eta^2}\ln\frac{2}{\delta}$. Setting $\eta = 4C_0(k+1)$ we obtain

$$\min_{0\le x\le L} \hat{\mathsf{cost}}(x) = \min_{0\le x\le L} \overline{Y}_n(x) \ge \frac{1}{2C_0(k+1)} - \eta > \frac{1}{4C_0(k+1)} \ge \frac{L}{\alpha 4(k+1)}.$$

Hence, letting $\alpha = LC_0$, and considering $n \ge 8\alpha^2(k+1)^2\ln\frac{2}{\delta}$, via Theorem 1 we have $\mathbf{Pr}[|T| = O(\alpha k/\varepsilon) = O(LC_0k/\varepsilon)] > 1-\delta$. Thus showing Lemma 2.

Technical Lemmas. Since in Lemma 2 we assume k points are independently sampled from a distribution and only require the cumulative distribution function of this distribution weakly differentiable, to prove Lemma 2 we need to generalize some integration formula to weakly differentiable functions.

Lemma 6. *Suppose $f \in C(\mathbb{R}) = \{f : \mathbb{R} \mapsto \mathbb{R} |\ f$ is continuous$\}$, and $x = g(t)$ satisfies $g \in C([t_1, t_2]) \cap W^{1,\infty}([t_1, t_2])$ or $g \in C([t_2, t_1]) \cap W^{1,\infty}([t_2, t_1])$, then we have*

$$\int_{g(t_1)}^{g(t_2)} f(x)dx = \int_{t_1}^{t_2} f(g(t))g'(t)dt \tag{6}$$

where g' is the weak derivative of g.

Proof. Without loss of generality, we assume $t_1 < t_2$ and then define

$$j(x) = \begin{cases} c_0 e^{-\frac{1}{1-x^2}} & \text{if } |x| < 1 \\ 0 & \text{if } |x| \ge 1 \end{cases}, \quad j_\delta(x) = \frac{1}{\delta} j(\frac{x}{\delta}),$$

where $\delta > 0$ and constant c_0 satisfies $\int_{\mathbb{R}} j(x)dx = 1$. We extend g to make it satisfy $g \in C([t_1 - 1, t_2 + 1]) \cap W^{1,\infty}([t_1 - 1, t_2 + 1])$, and for $\delta \in (0, 1)$ define

$$g_\delta(t) = \int_{\mathbb{R}} g(t')j_\delta(t - t')dt', \quad \forall\ t \in [t_1, t_2].$$

From the properties of mollifier j_δ, we know $g_\delta \in C^\infty([t_1, t_2])$ and

$$\lim_{\delta\to 0} \|g_\delta - g\|_{C([t_1,t_2])} = 0, \quad \lim_{\delta\to 0} g'_\delta(t) = g'(t) \quad \text{a.e. } [t_1, t_2]. \tag{7}$$

Since

$$\int_{g_\delta(t_1)}^{g_\delta(t_2)} f(x)dx = \int_{t_1}^{t_2} f(g_\delta(t))g'_\delta(t)dt, \tag{8}$$

letting $\delta \to 0$ in (8), by (7) and Lebesgue's dominated convergence theorem, we obtain (6). □

Now, using Lemma 6, we can give the proof of Lemma 7. Recall, we say a random variable X is C_0-*bounded* if its the cumulative distribution function

$$F(t) = \begin{cases} 0 & \text{if } t < 0 \\ \varphi(t) & \text{if } 0 \le t \le L \\ 1 & \text{if } t > L \end{cases},$$

satisfies $\varphi \in C([0,L]) \cap W^{1,\infty}([0,L])$ such that $\varphi(0) = 0$, $\varphi(L) = 1$ and $\|\varphi'\|_{L_\infty([0,L])} \le C_0$.

Lemma 7. Suppose $X_1, X_2, \ldots, X_k$ are independent and identically distributed C_0-bounded random variables. For $x \in [0,L]$, if $Y(x) = \min_{1\le j\le k}\{|x - X_j|\}$, then we have

$$\min_{0\le x\le L} \mathbf{E}(Y(x)) \ge \frac{1}{2C_0(k+1)}, \tag{9}$$

where $\mathbf{E}(Y(x))$ is the expected value of $Y(x)$.

Proof. If $0 \le x \le \frac{1}{2}L$, then by a direct computation we have

$$\mathbf{E}(Y(x)) = \int_0^x (1 - \varphi(x+y) + \varphi(x-y))^k dy + \int_x^{L-x} (1 - \varphi(x+y))^k dy. \tag{10}$$

For any $y \in [0,x]$, from $0 \le \varphi'(x+y) + \varphi'(x-y) \le 2C_0$, we have

$$(1-\varphi(x+y)+\varphi(x-y))^k(\varphi'(x+y)+\varphi'(x-y)) \le (1-\varphi(x+y)+\varphi(x-y))^k 2C_0,$$

which implies

$$\int_0^x (1-\varphi(x+y)+\varphi(x-y))^k(\varphi'(x+y)+\varphi'(x-y))dy \le \int_0^x (1-\varphi(x+y)+\varphi(x-y))^k 2C_0 dy. \tag{11}$$

Letting $\tau = 1 - \varphi(x+y) + \varphi(x-y)$, by Lemma 6 and (11), we obtain

$$\int_0^x (1-\varphi(x+y)+\varphi(x-y))^k dy \ge \frac{1}{2C_0}\int_{1-\varphi(2x)}^1 \tau^k d\tau = \frac{1}{2C_0(k+1)}\big(1-(1-\varphi(2x))^{k+1}\big). \tag{12}$$

Similarly, we have

$$\int_x^{L-x} (1 - \varphi(x+y))^k \varphi'(x+y) dy \le \int_x^{L-x} (1 - \varphi(x+y))^k C_0 dy$$

which implies

$$\int_x^{L-x} (1-\varphi(x+y))^k dy \ge \frac{1}{C_0}\int_0^{1-\varphi(2x)} \tau^k d\tau = \frac{1}{C_0(k+1)}(1-\varphi(2x))^{k+1}. \tag{13}$$

From (10), (12) and (13), we obtain

$$\mathbf{E}(Y(x)) \ge \frac{1}{2C_0(k+1)}\big(1 - (1-\varphi(2x))^{k+1} + 2(1-\varphi(2x))^{k+1}\big) \ge \frac{1}{2C_0(k+1)}. \tag{14}$$

For $x \in (\frac{1}{2}L, L]$, we can use a similar approach to prove

$$\mathbf{E}(Y(x)) \geq \frac{1}{2C_0(k+1)}. \tag{15}$$

Therefore, from (14) and (15) we have

$$\mathbf{E}(Y(x)) \geq \frac{1}{2C_0(k+1)}, \quad \forall\, x \in [0, L],$$

which implies (9).

□

B Size Bound of T in $\mathbb{R}^2$

Here we analyze the size of T resulting from running Algorithm B.1.

Algorithm B.1 Construct T from $\mathcal{P}$ in $\mathbb{R}^2$

Input: $\mathcal{P}$ and $\varepsilon > 0$
Compute $\varrho(\mathcal{P}) = \frac{1}{2}\min_{1\leq j\leq k} \hat{\mathsf{cost}}(p_{1,j})$
Construct $S = S(\mathcal{P})$ according to (1)
$T = \emptyset$
while $S \neq \emptyset$ **do**
  Choose $z \in S$ arbitrarily
  $T = T \cup \{z\}$
  $S = S \setminus \{s \in S \mid z \text{ can } (\frac{\varepsilon}{2(1+\varepsilon)})\text{-cover } s\}$
return T.

Our main theorem shows that $T(\mathcal{P})$ is within a constant factor of the optimal size such cover.

Theorem 4. *For a set of n uncertain points $\mathcal{P}$ in $\mathbb{R}^2$ and $\varepsilon \in (0, 1]$, suppose $\min_{x\in\mathbb{R}^2} \hat{\mathsf{cost}}(x) > 0$ and $T(\mathcal{P})$ is constructed from Algorithm B.1, then there exists a constant C^* independent of $\mathcal{P}$ and ε such that*

$$|T(\mathcal{P})| \leq C^*|T^*(\mathcal{P})|, \quad \forall\ \mathcal{P} \text{ in } \mathbb{R}^2$$

where

$$T^*(\mathcal{P}) = arg\min\left\{|T| \;\middle|\; T \subset \mathbb{R}^2, T \text{ can } \frac{\varepsilon}{2(1+\varepsilon)}\text{-cover } S(\mathcal{P})\right\},$$

and $S(\mathcal{P})$ is given by (1).

Proof. Suppose $T(\mathcal{P}) = \{z_1, z_2, \cdots, z_m\}$ is constructed as described above, and if $j > i$ then z_i is put into $T(\mathcal{P})$ before z_j. We consider adaptively-sized balls around each z_i defined as

$$B(z_i, r_i) = \{x \in \mathbb{R}^2 \mid \|x - z_i\| \leq r_i\} \quad \text{where} \quad r_i = \tfrac{\varepsilon}{2(1+\varepsilon)}\hat{\mathsf{cost}}(z_i).$$

Recall, it is the union of these balls that must cover $S(\mathcal{P})$. Since $j > i$ implies z_i is put into T before z_j, from Algorithm B.1 we know z_j is not $\frac{\varepsilon}{2(1+\varepsilon)}$-covered by z_i, which implies

$$\|z_j - z_i\| > r_i, \quad \forall\, j > i \geq 1. \tag{16}$$

From the Lipschitz property of $\hat{\mathsf{cost}}$ and $\varepsilon \in (0, 1]$ we have

$$-\frac{1}{4} \leq \frac{r_i - r_j}{\|z_i - z_j\|} \leq \frac{1}{4}, \quad \forall\, i \neq j. \tag{17}$$

Now, we divide the proof of this theorem into several steps. The first two are structural results about pairs of points $z_i, z_j \in T$. The third result shows that no single ball $B(z_i, r_i)$ can intersect too many other balls. The fourth result relates this to the largest independent set I from the result of any run of our algorithm, showing it must have size at least $\frac{1}{118}$ of the size of $T(\mathcal{P})$. The fifth step shows that any ball $B(x, r)$ does not intersect too many balls from the indpendent set. Finally, the sixth step combines these result to bound the size of any run to the optimal run.

Step 1 (pairs are not too close). If $j \neq i$ and $B(z_j, r_j) \cap B(z_i, r_i) \neq \emptyset$, then we have

$$\frac{4}{5} r_i \leq \|z_j - z_i\| \leq \frac{8}{3} r_i \tag{18}$$

and

$$r_j \geq \frac{3}{5} r_i. \tag{19}$$

We prove this with some algebraic manipulation of (17). If $j > i$, from (16) we have $\|z_j - z_i\| > r_i > \frac{4}{5} r_i$. If $j < i$, we assume $\|z_j - z_i\| < \frac{4}{5} r_i$ which implies

$$\|x_j - x_i\| < r_i - \frac{1}{4}\|z_j - z_i\|. \tag{20}$$

From (17) we have

$$r_i - \frac{1}{4}\|z_j - z_i\| \leq r_j. \tag{21}$$

From (20) and (21) we obtain

$$\|z_j - z_i\| < r_j$$

which is contradictory to $j < i$ and (16). So, we have $\|z_j - z_i\| \geq \frac{4}{5} r_i$.

To prove the other inequality in (18), we assume $\|z_j - z_i\| > \frac{8}{3} r_i$ which implies

$$r_i < \frac{3}{8}\|z_j - z_i\| \tag{22}$$

From (17) and (22) we have

$$r_j \le \frac{1}{4}\|z_j - z_i\| + r_i < \frac{1}{4}\|z_j - z_i\| + \frac{3}{8}\|z_j - z_i\| = \frac{5}{8}\|z_j - z_i\|. \tag{23}$$

From (22) and (23) we obtain

$$r_j + r_i < \frac{5}{8}\|z_j - z_i\| + \frac{3}{8}\|z_j - z_i\| = \|z_j - z_i\|$$

which is contradictory to $B(z_i, r_i) \cap B(z_j, r_j) \neq \emptyset$. So we have $\|z_j - z_i\| \le \frac{8}{3} r_i$.

To prove (19), we assume $r_j < \frac{3}{5} r_i$ which implies

$$r_i < \frac{5}{2}(r_i - r_j). \tag{24}$$

From (17) and (24) we have

$$r_i < \frac{5}{2}(r_i - r_j) \le \frac{5}{2} \cdot \frac{1}{4}\|z_j - z_i\| = \frac{5}{8}\|z_j - z_i\|. \tag{25}$$

So, from the assumption $r_j < \frac{3}{5} r_i$ and (25) we obtain

$$r_j + r_i < \frac{3}{5} r_i + r_i = \frac{8}{5} r_i < \frac{8}{5} \cdot \frac{5}{8}\|z_j - z_i\| = \|z_j - z_i\|$$

which is contradictory to $B(z_i, r_i) \cap B(z_j, r_j) \neq \emptyset$. So we have (19).

Step 2 (shrunken balls will not intersect). For any $i \neq j$, we have

$$B(z_i, \frac{4}{9} r_i) \cap B(z_j, \frac{4}{9} r_j) = \emptyset. \tag{26}$$

This follows easily from the results of Step 1. Without loss of generality, we assume $j > i$, so from (16) we have

$$2r_i < 2\|z_j - z_i\|. \tag{27}$$

From (17) we have

$$r_j - r_i \le \frac{1}{4}\|z_j - z_i\|. \tag{28}$$

Adding (27) and (28), we obtain

$$r_j + r_i < \frac{9}{4}\|z_j - z_i\|$$

which is equivalent to $\frac{4}{9}(r_j + r_i) < \|z_j - z_i\|$ implies (26).

Step 3 (no ball intersects too many others). For any $i \in \{1, 2, \cdots, m\}$, if $J_i = \{j \in \{1, 2, \cdots, m\} \mid B(z_j, r_j) \cap B(z_i, r_i) \neq \emptyset, j \neq i\}$, then we have

$$|J_i| \leq 117. \tag{29}$$

To prove (29), we use an area argument derived from the Step 2 fact that if we shrink balls enough they cannot intersect. To start, we define

$$A\big(z_i, \frac{8}{15}r_i, \frac{44}{15}r_i\big) = \big\{x \in \mathbb{R}^2 \big| \ \frac{8}{15}r_i \leq \|x - z_i\| \leq \frac{44}{15}r_i\big\}$$

For $j \in J_i$, from (18) we have

$$B(z_j, \frac{4}{15}r_i) \subset A\big(z_i, \frac{8}{15}r_i, \frac{44}{15}r_i\big) \tag{30}$$

From (19) we have $B(z_j, \frac{4}{15}r_i) \subset B(z_j, \frac{4}{9}r_j)$, so from (30) we obtain

$$\begin{aligned} &\mathbf{m}\left(B(z_j, \frac{4}{9}r_j) \cap A\big(z_i, \frac{8}{15}r_i, \frac{44}{15}r_i\big)\right) \\ \geq &\mathbf{m}\left(B(z_j, \frac{4}{15}r_i) \cap A\big(z_i, \frac{8}{15}r_i, \frac{44}{15}r_i\big)\right) = \mathbf{m}\left(B(z_j, \frac{4}{15}r_i)\right) = \pi\big(\frac{4}{15}r_i\big)^2 \end{aligned} \tag{31}$$

where $\mathbf{m}(\cdot)$ represents the area of a set.

Moreover, from (26) for any $j' \neq j$ we have

$$B(z_j, \frac{4}{9}r_j) \cap B(z_{j'}, \frac{4}{9}r_{j'}) = \emptyset. \tag{32}$$

From (31) and (32) we obtain

$$\begin{aligned} |J_i|\pi(\frac{4}{15}r_i)^2 &\leq \sum_{j \in J_i} \mathbf{m}\left(B(z_j, \frac{4}{9}r_j) \cap A\big(z_i, \frac{8}{15}r_i, \frac{44}{15}r_i\big)\right) \\ &= \mathbf{m}\left(\bigcup_{j \in J_i}\left(B(z_j, \frac{4}{9}r_j) \cap A\big(z_i, \frac{8}{15}r_i, \frac{44}{15}r_i\big)\right)\right) \\ &\leq \mathbf{m}\left(A\big(z_i, \frac{8}{15}r_i, \frac{44}{15}r_i\big)\right) = \pi(\frac{44}{15}r_i)^2 - \pi(\frac{8}{15}r_i)^2 = 117\pi(\frac{4}{15}r_i)^2 \end{aligned}$$

which implies (29).

Step 4 (existence of large independent set). There exists $I \subset \{1, 2, \cdots, m\}$ such that

$$B(z_i, r_i) \cap B(z_j, r_j) = \emptyset, \quad \forall\ i, j \in I, \tag{33}$$

and

$$|I| \geq \frac{1}{118}|T(\mathcal{P})|. \tag{34}$$

To prove the existence of I, we convert $T(\mathcal{P}) = \{z_1, z_2, \cdots, z_m\}$ to a graph $G = (V, E)$ where $V = \{v_1, v_2, \cdots, v_m\}$ and $E = \{(v_i, v_j) |\ v_i, v_j \in V,\ B(z_i, r_i) \cap B(z_j, r_j) \neq \emptyset\}$. From (29) we know

$$\max_{1 \leq i \leq m} \text{degree}(v_i) \leq 117. \tag{35}$$

By (35) and Brook's theorem we know $\chi(G)$, the chromatic number of G, satisfies

$$\chi(G) \leq \max_{1 \leq i \leq m} \text{degree}(v_i) + 1 \leq 118. \tag{36}$$

Suppose V' is the largest independent set of G. We define $I = \{i \in \{1, 2, \cdots, m\} |\ v_i \in V'\}$ and obviously I satisfies (33) and $|I| = |V'|$. From (36) and the relationship between $|V'|$ and $\chi(G)$, we obtain

$$|I| = |V'| \geq \frac{|V|}{\chi(G)} \geq \frac{|V|}{118} = \frac{|T(\mathcal{P})|}{118}. \tag{37}$$

Step 5 (small intersection of $B(x, r)$ with independent set). For any $x \in \mathbb{R}^2$, $r > 0$ satisfying

$$-\frac{1}{4}\|x - z_i\| \leq r - r_i \leq \frac{1}{4}\|x - z_i\| \quad \forall\ i \in I, \tag{38}$$

we define $I_{(x,r)} = \{i \in I \mid B(z_i, r_i) \cap B(x, r) \neq \emptyset\}$ where $I \subset \{1, 2, \cdots, m\}$ satisfies (33), then we have

$$|I_{(x,r)}| \leq 30. \tag{39}$$

Using the method in Step 1, from (38), we can obtain

$$\|z_i - x\| \leq \frac{8}{3}r \text{ and } \quad r_i \geq \frac{3}{5}r, \ \ \forall\ i \in I_{(x,r)}$$

which implies

$$B(z_i, \frac{3}{5}r) \subset B(z_i, r_i) \ \text{ and } \quad B(z_i, \frac{3}{5}r) \subset B(x, \frac{49}{15}r), \quad \forall\ i \in I_{(x,r)}. \tag{40}$$

So, from (40) we have

$$\mathbf{m}\left(B(z_i, r_i) \cap B(x, \frac{49}{15}r)\right) \geq \mathbf{m}\left(B(z_i, \frac{3}{5}r) \cap B(x, \frac{49}{15}r)\right) = \mathbf{m}\left(B(z_i, \frac{3}{5}r))\right) = \pi(\frac{3}{5}r)^2, \ \ \forall\, i \in I_{(x,r)}. \tag{41}$$

Since I satisfies (33), from (41) we obtain

$$\begin{aligned} |I_{(x,r)}|\pi(\frac{3}{5}r)^2 \leq & \sum_{i \in I_{(x,r)}} \mathbf{m}\left(B(z_i, r_i) \cap B(x, \frac{49}{15}r)\right) = \mathbf{m}\left(\bigcup_{i \in I_{(x,r)}} \left(B(z_i, r_i) \cap B(x, \frac{49}{15}r)\right)\right) \\ \leq & \mathbf{m}\left(B(x, \frac{49}{15}r)\right) = \pi(\frac{49}{15}r)^2 \leq 30\pi(\frac{3}{5}r)^2 \end{aligned}$$

which implies (39).

Step 6 (putting it all together). Suppose $T \subset \mathbb{R}^2$ can $\frac{\varepsilon}{2(1+\varepsilon)}$-cover $S(\mathcal{P})$. For any $x \in T$ and $r = \frac{\varepsilon}{2(1+\varepsilon)}\hat{\mathsf{cost}}(x)$, we know x and r satisfy (38), which implies $|I_{(x,r)}| \leq 30$. This means each point in T can $\frac{\varepsilon}{2(1+\varepsilon)}$-cover at most 30 points in $\{z_i \in T(\mathcal{P}) | i \in I\}$. Since T must $\frac{\varepsilon}{2(1+\varepsilon)}$-cover $\{z_i \in T(\mathcal{P}) \mid i \in I\} \subset S(\mathcal{P})$, we have

$$|I| \leq 30|T|. \tag{42}$$

From (37) and (42) we have

$$|T| \geq \frac{|I|}{30} \geq \frac{1}{30} \cdot \frac{1}{118}|T(\mathcal{P})| = \frac{1}{3540}|T(\mathcal{P})|, \quad \forall\, T \subset \mathbb{R}^2 \text{ and } T \text{ can } \frac{\varepsilon}{2(1+\varepsilon)}\text{-cover } S(\mathcal{P}). \tag{43}$$

Setting $C^* = 3540$, from (43) we complete the proof. □

Remark: In the proof of Theorem 4, we set $C^* = 3540$. However, since the bounds in (29) and (42) are not tight, as well as the argument in Step 6, the true value of C^* is likely much smaller than 3540.

C Expected Lower Bound of $\hat{\mathsf{cost}}$ under C_0-bounded iid Assumption in $\mathbb{R}^2$

In this section we provide show that if the uncertain points are each drawn iid from separate C_0-bounded distributions, then the size of T is small (specifically $O(\alpha k^2/\varepsilon^2)$, where α is a constant that depends on C_0) with high probability.

Before discussing the the existence of the constant α in more detail, we establish a lemma which is similar to the $\mathbb{R}^1$ case and is the basis for estimating the lower bound of the expected value of $\hat{\mathsf{cost}}(x)$.

Lemma 8. *Suppose $(X_1, Y_1,), (X_2, Y_2), \cdots, (X_k, Y_k)$ are independent and identically distributed two dimensional random variables. The joint density function of (X_1, Y_1) is $f(\tilde{x}, \tilde{y})$ which satisfies*

$$f(\tilde{x}, \tilde{y}) \in L_\infty(\mathbb{R}^2),\ \|f\|_{L_\infty(\mathbb{R}^2)} \leq C_0,\ \text{and } f(\tilde{x}, \tilde{y}) = 0,\ \forall\ (\tilde{x}, \tilde{y}) \notin B((0,0), R)$$

where R and C_0 are positive constants. For any fixed $(x, y) \in B((0,0), R)$, if $Z_j(x, y)$ and $Z(x, y)$ are defined by

$$\begin{aligned} Z_j(x,y) &= \sqrt{(X_j - x)^2 + (Y_j - y)^2},\ j = 1, 2, \cdots, k\ , \\ Z(x,y) &= \min\{Z_1(x,y), Z_2(x,y), \cdots, Z_k(x,y)\}\ , \end{aligned}$$

then we have

$$\min_{(x,y)\in B((0,0),R)} \mathbf{E}(Z(x,y)) \geq \frac{1}{4\pi R C_0(k+1)}. \tag{44}$$

Proof. We first assume $f \in C^\infty(B((0,0),R))$. A direct computation yields the cumulative distribution function of $Z(x,y)$:

$$F_{Z(x,y)}(z) = \begin{cases} 0 & \text{if } z \le 0 \\ 1 - \Big(1 - \iint\limits_{B((x,y),z)} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}\Big)^k & \text{if } 0 < z \le R - \sqrt{x^2+y^2} \\ 1 - \Big(1 - \iint\limits_{B((x,y),z)\cap B((0,0),R)} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}\Big)^k & \text{if } R - \sqrt{x^2+y^2} < z \le R + \sqrt{x^2+y^2} \\ 1 & \text{if } z > R + \sqrt{x^2+y^2} \end{cases}$$

and its expected value $\mathbf{E}(Z(x,y)) = I_1 + I_2$, where

$$I_1 = \int_0^{R-\sqrt{x^2+y^2}} \Big(1 - \iint\limits_{B((x,y),z)} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}\Big)^k dz,$$
$$I_2 = \int_{R-\sqrt{x^2+y^2}}^{R+\sqrt{x^2+y^2}} \Big(1 - \iint\limits_{B((x,y),z)\cap B((0,0),R)} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}\Big)^k dz.$$

Now, we estimate the lower bound of I_1 and I_2 respectively. Using the polar coordinates

$$\begin{cases} \tilde{x} = x + r\cos\theta \\ \tilde{y} = y + r\sin\theta \end{cases}, \tag{45}$$

we have

$$\iint\limits_{B((x,y),z)} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} = \int_0^z \int_0^{2\pi} f(x + r\cos\theta, y + r\sin\theta) r d\theta dr. \tag{46}$$

To estimate I_1, we introduce the transformation

$$1 - \iint\limits_{B((x,y),z)} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} = \tau_1. \tag{47}$$

From (46) and (47) we have

$$d\tau_1 = -\Big(\int_0^{2\pi} f(x + z\cos\theta, y + z\sin\theta) z d\theta\Big) dz$$

which implies

$$
\begin{aligned}
I_1 =& \int_{1-\iint\limits_{B((x,y),R-\sqrt{x^2+y^2})} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}}^{1} \frac{\tau_1^k}{\int_0^{2\pi} f(x+z\cos\theta, y+z\sin\theta)zd\theta}d\tau_1 \\
\geq& \frac{1}{2\pi C_0(R-\sqrt{x^2+y^2})} \int_{1-\iint\limits_{B((x,y),R-\sqrt{x^2+y^2})} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}}^{1} \tau_1^k d\tau_1 \qquad (48) \\
\geq& \frac{1}{2\pi C_0 R(k+1)}\Big(1-\big(1-\iint\limits_{B((x,y),R-\sqrt{x^2+y^2})} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}\big)^{k+1}\Big).
\end{aligned}
$$

To estimate I_2, using the polar coordinates (45), we have

$$
\begin{aligned}
& B((x,y),z)\cap B((0,0),R) \\
=& \big\{(x+r\cos\theta, y+r\sin\theta)|\theta_1(z)\leq\theta\leq\theta_2(z), 0\leq r\leq z\big\} \qquad (49) \\
& \cup\big\{(x+r\cos\theta, y+r\sin\theta)|\theta_2(z)\leq\theta\leq\theta_1(z)+2\pi, 0\leq r\leq r(\theta)\big\}
\end{aligned}
$$

where

$$
r(\theta) = -(x\cos\theta + y\sin\theta) + \sqrt{(x\cos\theta+y\sin\theta)^2+(R^2-x^2-y^2)}\ ,
$$

and $\theta_1(z)$, $\theta_2(z)$ $(\theta_1(z)<\theta_2(z))$ are two roots of

$$
(x+z\cos\theta)^2+(y+z\sin\theta)^2 = R^2
$$

and satisfy

$$
(x+z\cos\theta, y+z\sin\theta)\in B((0,0),R),\ \ \forall\,\theta\in(\theta_1(z),\theta_2(z)), z\in(R-\sqrt{x^2+y^2}, R+\sqrt{x^2+y^2}).
$$

Moreover, it is easy to check

$$
r(\theta_1(z)) = r(\theta_2(z)) = z. \qquad (50)
$$

Thus, from (49) we obtain

$$
\begin{aligned}
& \iint\limits_{B((x,y),z)\cap B((0,0),R)} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} \\
=& \int_{\theta_1(z)}^{\theta_2(z)}\int_0^z f(x+r\cos\theta, y+r\sin\theta)rdrd\theta + \int_{\theta_2(z)}^{\theta_1(z)+2\pi}\int_0^{r(\theta)} f(x+r\cos\theta, y+r\sin\theta)rdrd\theta. \qquad (51)
\end{aligned}
$$

Introducing the transformation

$$
1-\iint\limits_{B((x,y),z)\cap B((0,0),R)} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} = \tau_2\ ,
$$

from (50) and (51) we have

$$-\frac{d\tau_2}{dz} = \int_{\theta_1(z)}^{\theta_2(z)} f(x + z\cos\theta, y + z\sin\theta) z d\theta$$

which implies

$$\begin{aligned} I_2 =& \int_0^{1-\iint\limits_{B((x,y),R-\sqrt{x^2+y^2})} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}} \frac{\tau_2^k}{\int_{\theta_1(z)}^{\theta_2(z)} f(x + z\cos\theta, y + z\sin\theta) z d\theta} d\tau_2 \\ \geq& \frac{1}{2\pi C_0(R + \sqrt{x^2+y^2})} \int_0^{1-\iint\limits_{B((x,y),R-\sqrt{x^2+y^2})} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}} \tau_2^k d\tau_2 \qquad (52) \\ \geq& \frac{1}{4\pi C_0 R(k+1)} \big(1 - \iint\limits_{B((x,y),R-\sqrt{x^2+y^2})} f(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}\big)^{k+1}. \end{aligned}$$

Therefore, from (48) and (52) we have

$$\mathbf{E}(Z(x,y)) = I_1 + I_2 \geq \frac{1}{2\pi C_0(k+1)}\big(\frac{1}{R} - \frac{1}{2R}\big) = \frac{1}{4\pi R C_0(k+1)}.$$

For the case $f \notin C^\infty(B((0,0),R))$, we can use a sequence of smooth functions to approximate f, and obtain (44) for each smooth function and then use Lebesgue's dominated convergence theorem to show f also satisfies (44). Thus, the proof of this lemma is completed.

On the basis of this lemma, using Chernoff-Hoeffding inequality, we can obtain a lower bound of the expected value of $\hat{\mathsf{cost}}(x)$ in Theorem 5, and the proof is similar to that of Lemma 5.

Theorem 5. *Suppose for fixed $i \in \{1, 2, \cdots, n\}$, $(X_{i,1}, Y_{i,1}), (X_{i,2}, Y_{i,2}), \cdots, (X_{i,k}, Y_{i,k})$ are independent and identically distributed two dimensional random variables. The joint density function of $(X_{i,1}, Y_{i,1})$ is $f_i(\tilde{x}, \tilde{y})$ which satisfies*

$$f_i(\tilde{x},\tilde{y}) \in L_\infty(\mathbb{R}^2),\ \|f_i\|_{L_\infty(B((0,0),R))} \leq C_0,\ \text{and } f_i(\tilde{x},\tilde{y}) = 0,\ \forall\ (\tilde{x},\tilde{y}) \notin B((0,0),R)$$

where R and C_0 are positive constants. For any fixed $(x, y) \in B((0,0),R)$, suppose $Z^i(x,y)$ is given by

$$Z^i(x,y) = \min_{1\leq j\leq k} \left\{ \sqrt{(X_{i,j}-x)^2 + (Y_{i,j}-y)^2} \right\},$$

and $Z^1, Z^2, \cdots, Z^n$ are mutually independent.

If $\overline{Z}_n(x,y) \triangleq \frac{1}{n}\sum_{i=1}^n Z^i(x,y)$, then for any $\eta > 0$ and $\delta \in (0,1)$ we have

$$\mathbf{Pr}\left[\min_{(x,y)\in B((0,0),R)} \overline{Z}_n(x,y) \geq \frac{1}{4\pi R C_0(k+1)} - \eta \right] > 1 - \delta, \qquad \forall\ n > \frac{2R^2}{\eta^2} \ln\frac{2}{\delta}. \qquad (53)$$

Remark: Under the condition of Theorem 5, we know there exists a constant $\alpha = (R^2C_0)^2$ such that $\mathbf{Pr}\big[|T(\mathcal{P})| = O(\alpha k^2 \frac{1}{\varepsilon^2})\big] > 1-\delta$ for any $n > 128\alpha^2(k+1)^2 \ln \frac{2}{\delta}$. In fact, if $n > 128R^4C_0^2(k+1)^2 \ln \frac{2}{\delta}$, then we can find $\eta_0 \in (0, \frac{1}{8RC_0(k+1)})$ such that $n > \frac{2R^2}{\eta_0^2} \ln \frac{2}{\delta} > 128R^4C_0^2(k+1)^2 \ln \frac{2}{\delta}$. Moreover, since $\eta_0 \in (0, \frac{1}{8RC_0(k+1)})$, we have

$$\begin{aligned}
&\min_{(x,y)\in B((0,0),R)} \overline{Z}_n(x,y) \geq \frac{1}{4\pi RC_0(k+1)} - \eta_0 \\
\Rightarrow\quad & |T(\mathcal{P})| \leq O\left(\frac{R^2}{\big(\frac{\varepsilon}{1+\varepsilon}\min_{x\in B((0,0),R)} \hat{\mathsf{cost}}(x)\big)^2}\right) = O\left(\frac{R^2(\frac{1}{\varepsilon})^2}{\big(\min_{(x,y)\in B((0,0),R)} \overline{Z}_n(x,y)\big)^2}\right) \\
\leq & O\left(\frac{R^2(\frac{1}{\varepsilon})^2}{\big(-\eta_0 + (4\pi RC_0(k+1))^{-1}\big)^2}\right) = O\left(R^4C_0^2k^2\frac{1}{\varepsilon^2}\right) \\
\Rightarrow\quad & |T(\mathcal{P})| = O\left(\alpha k^2 \frac{1}{\varepsilon^2}\right)
\end{aligned}$$

which implies

$$\mathbf{Pr}\left[|T(\mathcal{P})| = O(\alpha k^2 \frac{1}{\varepsilon^2})\right] \geq \mathbf{Pr}\left[\min_{(x,y)\in B((0,0),R)} \overline{Z}_n(x,y)\right]. \tag{54}$$

Since $n > \frac{2R^2}{\eta_0^2} \ln \frac{2}{\delta}$, from (53) and (54) we obtain

$$\mathbf{Pr}\left[|T| = O(\alpha k^2 \frac{1}{\varepsilon^2})\right] > 1 - \delta.$$