


Approximating the Distribution of the Median and other Robust Estimators on Uncertain Data

Kevin Buchin

Department of Mathematics and Computer Science, TU Eindhoven
Eindhoven, The Netherlands

k.a.buchin@tue.nl

 <https://orcid.org/0000-0002-3022-7877>

Jeff M. Phillips

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

jeffp@cs.utah.edu

Pingfan Tang

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

tang1984@cs.utah.edu

Abstract

Robust estimators, like the median of a point set, are important for data analysis in the presence of outliers. We study robust estimators for locationally uncertain points with discrete distributions. That is, each point in a data set has a discrete probability distribution describing its location. The probabilistic nature of uncertain data makes it challenging to compute such estimators, since the true value of the estimator is now described by a distribution rather than a single point. We show how to construct and estimate the distribution of the median of a point set. Building the approximate support of the distribution takes near-linear time, and assigning probability to that support takes quadratic time. We also develop a general approximation technique for distributions of robust estimators with respect to ranges with bounded VC dimension. This includes the geometric median for high dimensions and the Siegel estimator for linear regression.

2012 ACM Subject Classification Theory of computation → Computational geometry

Keywords and phrases Uncertain Data, Robust Estimators, Geometric Median, Tukey Median

Digital Object Identifier 10.4230/LIPIcs.SoCG.2018.16

Related Version A full version of this paper is available at <https://arxiv.org/abs/1601.00630>

Funding NSF CCF-1115677, CCF-1350888, IIS-1251019, ACI-1443046, CNS-1514520, CNS-1564287, and NWO project no. 612.001.207

1 Introduction

Most statistical or machine learning models of noisy data start with the assumption that a data set is drawn iid (independent and identically distributed) from a single distribution. Such distributions often represent some true phenomenon under some noisy observation. Therefore, approaches that mitigate the influence of 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



© Kevin Buchin, Jeff M. Phillips, and Pingfan Tang;
licensed under Creative Commons License CC-BY

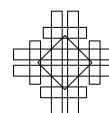
34th International Symposium on Computational Geometry (SoCG 2018).

Editors: Bettina Speckmann and Csaba D. Tóth; Article No. 16; pp. 16:1–16:14

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



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 [11, 18] 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 being equally likely.

Although algorithms for locationally uncertain points have been studied in quite a few contexts over the last decade [10, 16, 13, 4, 12, 3, 1, 2, 23] (see more through discussion in full version [8]), 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 , as well as its generalization the geometric median and the Tukey median for data in \mathbb{R}^d , defined in Section 1.1. Being robust refers to the fact that the median and geometric medians have a *breakdown points* of 0.5, that is, 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 [17]. The Tukey median has a breakdown point between $\frac{1}{d+1}$ and $\frac{1}{3}$ [5].

In this paper, we generalize the median (and other robust estimators) 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 $\text{cost}(x, Q)$; 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 robust. As such we build a distribution over the possible locations of these cost-minimizers. In \mathbb{R}^1 (by defining boundary cases carefully) this distribution is of size at most $O(nk)$, the size of the input, but already in \mathbb{R}^2 it may be as large as k^n .

Our results. We design algorithms to create an approximate support of these median distributions. We create small sets T (called an ε -support) such that each possible median m_Q from a possible point set Q is within a distance $\varepsilon \cdot \text{cost}(m_Q, Q)$ of some $x \in T$. In \mathbb{R} we can create a support set T of size $O(k/\varepsilon)$ in $O(nk \log(nk))$ time. We show that the bound $O(k/\varepsilon)$ is tight since there may be k large enough modes of these distributions, each requiring $\Omega(1/\varepsilon)$ points to represent. In \mathbb{R}^d our bound on $|T|$ is $O(k^d/\varepsilon^d)$, for the Tukey median and the geometric median. If we do not need to cover sets of medians m_Q which occur with probability less than ε , we can get a bound $O(d/\varepsilon^2)$ in \mathbb{R}^d . In fact, this general approach in \mathbb{R}^d extends to other estimators, including the Siegel estimator [19] for linear regression. We then need to map weights onto this support set T . We can do so exactly in $O(n^2k)$ time in \mathbb{R}^1 or approximately in $O(1/\varepsilon^2)$ time in \mathbb{R}^d .

Another goal may be to then construct a 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 $\text{cost}(m_Q, Q)$ where m_Q is the resulting single-point estimate. However, we also show

that already in \mathbb{R}^1 such estimators are not stable with respect to the weights in the median distribution, and 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.

1.1 Formalization of model and notation

We consider a set of n locationally uncertain points $\mathcal{P} = \{P_1, \dots, P_n\}$ so that each P_i has k possible locations $\{p_{i,1}, \dots, p_{i,k}\} \subset \mathbb{R}^d$. Here, $P_i = \{p_{i,1}, \dots, p_{i,k}\}$ is a multiset, which means a point in P_i may appear more than once. Let $P_{\text{flat}} = \cup_i \{p_{i,1}, \dots, p_{i,k}\}$ represent all positions of all points in \mathcal{P} , which implies P_{flat} is also a multiset. 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 $Q \in \mathcal{P}$ is a *traversal* of \mathcal{P} if $Q = \{q_1, \dots, q_n\}$ has each q_i in the domain of P_i (e.g., $q_i = p_{i,j}$ for some j). We denote by $\Pr_{Q \in \mathcal{P}}[\gamma(Q)]$ the probability of the event $\gamma(Q)$, given that Q is a randomly selected traversal from \mathcal{P} , where the selection of each q_i from P_i is independent of $q_{i'}$ from $P_{i'}$.

We are particularly interested in the case where n is large and k is small. For technical simplicity we assume an extended RAM model where k^n (the number of possible traversals of point sets) can be computed in $O(1)$ time and fits in $O(1)$ words of space.

We consider three definitions of medians. In one dimension, given a set $Q = \{q_1, q_2, \dots, q_n\}$ that w.l.o.g. satisfies $q_1 \leq q_2 \leq \dots \leq q_n$, 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 [5], herein we focus on the geometric median and Tukey median. Define $\text{cost}(x, Q) = \frac{1}{n} \sum_{i=1}^n \|x - q_i\|$ where $\|\cdot\|$ is the Euclidian norm. Given a set $Q = \{q_1, q_2, \dots, q_n\} \subset \mathbb{R}^d$, the *geometric median* is defined as $m_Q = \arg \min_{x \in \mathbb{R}^d} \text{cost}(x, Q)$. The Tukey depth [20] of a point p with respect to a set $Q \subset \mathbb{R}^d$ is defined $\text{depth}_Q(p) := \min_{H \in \mathcal{H}_p} |H \cap Q|$ where $\mathcal{H}_p := \{H \text{ is a closed halfspace in } \mathbb{R}^d \mid p \in H\}$. Then a *Tukey median* of a set Q is a point p that can maximize the Tukey depth.

2 Constructing a single point estimate

We begin by exploring the construction of a single point estimator of set of n locationally uncertain points \mathcal{P} . We demonstrate that while the estimator is stable with respect to the value of cost , the actual minimum of that function is not stable and provides an incomplete picture for multimodal uncertainties.

It is easiest to explore this through a weighted point set $X \subset \mathbb{R}^1$. Given a probability distribution defined by $\omega : X \rightarrow [0, 1]$, we can compute its weighted median 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. The first domain is the set T of possible locations of medians under different instantiations of the uncertain points with weights \hat{w} as the probability of those medians being realized; see constructions in Section 3.2 and Section 3.5. Let the resulting weighted median of (T, \hat{w}) be m_T . The second domain is simply the set P_{flat} of all possible locations of \mathcal{P} , and its weight w where $w(p_{i,j})$ is the fraction of $Q \in \mathcal{P}$ which take $p_{i,j}$ as their median (possibly 0). Let the weighted median of (P_{flat}, w) be $m_{\mathcal{P}}$.

► **Theorem 1.** $|m_T - m_{\mathcal{P}}| \leq \varepsilon \text{cost}(m_{\mathcal{P}}) \leq \varepsilon \text{cost}(m_Q, Q)$, $Q \in \mathcal{P}$ is any traversal with $m_{\mathcal{P}}$ 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 $|z - y| \leq \varepsilon \text{cost}(z)$ but $|z - x| > \text{cost}(z)$. 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 the sum of all weights of all points greater than $p_{i,j}$ is at most 0.5, and the sum of 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 $|x - p_{i,j}| \leq \varepsilon \text{cost}(m_{\mathcal{P}})$ as desired. ◀

Non-robustness of single point estimates. The geometric median of the set $\{m_Q$ is a geometric median of $Q \mid Q \in \mathcal{P}\}$ 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 $\hat{\text{cost}}(0) = \hat{\text{cost}}(\Delta) = 0$. 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 study in the remainder is more appropriate for noisy data.

3 Approximating the median distribution

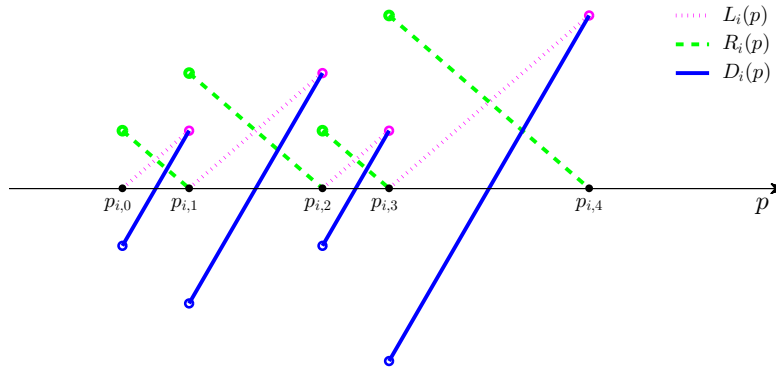
The big challenge in constructing an ε -support T is finding the points $x \in P_{\text{flat}}$ which have small values of $\text{cost}(x, Q)$ (recall $\text{cost}(x, Q) = \frac{1}{n} \sum_{i=1}^n \|x - q_i\|$) for some $Q \in \mathcal{P}$. But this requires determining the smallest cost $Q \in \mathcal{P}$ that has $x \in Q$ and x is the median of Q .

One may think (as the authors initially did) that one could simply use a proxy function $\hat{\text{cost}}(x) = \frac{1}{n} \sum_{i=1}^n \min_{1 \leq j \leq k} \|x - p_{i,j}\|$, which is relatively simple to compute as the lower envelope of cost functions for each P_i . Clearly $\hat{\text{cost}}(x) \leq \text{cost}(x, Q)$ for all $Q \in \mathcal{P}$, so a set \hat{T} satisfying a similar approximation for $\hat{\text{cost}}$ will satisfy our goals for cost . However, there exist (rather adversarial) data sets \mathcal{P} where \hat{T} would require $\Omega(nk)$ points; see full version [8]. On the other hand, we show this is not true for cost . The key difference between cost and $\hat{\text{cost}}$ is that $\hat{\text{cost}}$ does not enforce the use of some $Q \in \mathcal{P}$ of which x is a median. That is, that (roughly) half the points are to the left and half to the right for this Q .

Proxy functions L , R , and D . We handle this problem by first introducing two families of functions, defined precisely shortly. We let $L_i(x)$ (resp. $R_i(x)$) represent the contribution to cost at x from the closest possible location $p_{i,j}$ of an uncertain point P_i to the left (resp. right) of x . This allows us to decompose the elements of this cost. However, it does not help us to enforce this balance. Hence we introduce a third proxy function

$$D_i(x) = L_i(x) - R_i(x)$$

capturing the difference between L_i and R_i . We will show that the choice of which points are used on the left or right of x is completely determined by the D_i values. In particular, we maintain the D_i values (for all $i \in [n]$) in sorted order, and use the i with larger D_i values on the right, and smaller D_i values on the left for the min cost $Q \in \mathcal{P}$.



■ **Figure 1** The plot of $L_i(p)$, $R_i(p)$ and $D_i(p)$.

To define L_i , R_i , and D_i , we first assume that P_{flat} and P_i for all $i \in [n]$ are sorted (this would take $O(nk \log(nk))$ time). Then to simplify definitions we add two dummy points to each P_i , and introduce the notation $\tilde{P}_i = P_i \cup \{p_{i,0}, p_{i,k+1}\}$ and $\tilde{\mathcal{P}} = \{\tilde{P}_1, \tilde{P}_2, \dots, \tilde{P}_n\}$, where $p_{i,0} = \min P_{\text{flat}} - n\Delta$, $p_{i,k+1} = \max P_{\text{flat}} + n\Delta$, and $\Delta = \max P_{\text{flat}} - \min P_{\text{flat}}$. Thus, every point $p \in P_{\text{flat}}$ can be viewed as the median of some traversal of $\tilde{\mathcal{P}}$. Moreover, since we put the $p_{i,0}$ and $p_{i,k+1}$ points far enough out, they will essentially act as points at infinity and not affect the rest of our analysis.

Next, for $p \in P_{\text{flat}}$ we define $\text{cost}(p) = \min\{\text{cost}(p, Q) \mid p \text{ is the median of } Q \text{ and } Q \in \tilde{\mathcal{P}}\}$. Thus, if there exists $Q \in \mathcal{P}$ such that p is the median of Q , then $\text{cost}(p) \leq \text{cost}(p, Q)$.

Now to compute cost and expedite our analysis, for $p \in [\min P_{\text{flat}} - n\Delta, \max P_{\text{flat}} + n\Delta]$, we define $L_i(p) = \min\{|p_i - p| \mid p_i \in \tilde{P}_i \cap (-\infty, p]\}$ and $R_i(p) = \min\{|p_i - p| \mid p_i \in \tilde{P}_i \cap [p, \infty)\}$, and recall $D_i(p) = L_i(p) - R_i(p)$. Obviously, if $p \in \tilde{P}_i$, then $D_i(p) = L_i(p) = R_i(p) = 0$. For example, if $\tilde{P}_i = \{p_{i,0}, p_{i,1}, p_{i,2}, p_{i,3}, p_{i,4}\}$ and $p_{i,0} < p_{i,1} < p_{i,2} < p_{i,3} < p_{i,4}$, then the plot of $L_i(p)$, $R_i(p)$ and $D_i(p)$, is shown in Figure 1.

For the sake of brevity, we now assume n is odd; adjusting a few arguments by $+1$ will adjust for the n is even case.

Consider next the following property of the D_i functions with respect to computing $\text{cost}(p)$ for a point $p \in P_{i_0}$. Let $\{i_1, i_2, \dots, i_{n-1}\} = [n] \setminus \{i_0\}$ be a permutation of uncertain points, except for i_0 , so that $D_{i_1}(p) \leq D_{i_2}(p) \leq \dots \leq D_{i_{n-1}}(p)$. Then to minimize $\text{cost}(p, Q)$, we count the uncertain points P_{i_l} using L_{i_l} if in the permutation $i_l \leq (n-1)/2$ and otherwise count it on the right with R_{i_l} . This holds since for any other permutation $\{j_1, j_2, \dots, j_{n-1}\} = [n] \setminus \{i_0\}$ we have $\sum_{l=\frac{n+1}{2}}^{n-1} D_{i_l}(p) \geq \sum_{l=\frac{n+1}{2}}^{n-1} D_{j_l}(p)$ and thus

$$\begin{aligned} \sum_{l=1}^{\frac{n-1}{2}} L_{i_l}(p) + \sum_{l=\frac{n+1}{2}}^{n-1} R_{i_l}(p) &= \sum_{l=1}^{n-1} L_{i_l}(p) - \sum_{l=\frac{n+1}{2}}^{n-1} D_{i_l}(p) \\ &\leq \sum_{l=1}^{n-1} L_{j_l}(p) - \sum_{l=\frac{n+1}{2}}^{n-1} D_{j_l}(p) = \sum_{l=1}^{\frac{n-1}{2}} L_{j_l}(p) + \sum_{l=\frac{n+1}{2}}^{n-1} R_{j_l}(p). \end{aligned}$$

For $p \in P_{i_0}$, $\text{cost}(p) = \frac{1}{n} \left(\sum_{l=1}^{\frac{n-1}{2}} L_{i_l}(p) + \sum_{l=\frac{n+1}{2}}^{n-1} R_{i_l}(p) \right)$ under this D_i -sorted permutation.

3.1 Computing cost

Now to compute cost for all points $p \in P_{\text{flat}}$, we simply need to maintain the D_i in sorted order, and then sum the appropriate terms from L_i and R_i . Let us first examine a few facts about the complexity of these functions.

The function L_i (resp. R_i) is piecewise-linear, where the slope is always 1 (resp. -1). The breakpoints only occur at $x = p_{i,j}$ for each $p_{i,j} \in P_i$. Hence, they each have complexity $\Theta(k)$ for all $i \in [n]$. The structure of L_i and R_i implies that D_i is also piecewise-linear, where the slope is always 2 and has breakpoints for each $p_{i,j} \in P_i$. Each linear component attains a value $D_i(x) = 0$ when x is the midpoint between two $p_{i,j}, p_{i,j'} \in P_i$ which are consecutive in the sorted order of P_i .

The fact that all D_i have slope 2 at all non-discontinuous points, and these discontinuous points only occur at P_i , implies that the sorted order of the D_i functions does not change in between points of P_{flat} . Moreover, at one of these points of discontinuity $x \in P_{\text{flat}}$, the ordering between D_i s only changes for uncertain points $D_{i'}$ such that there exists a possible location $p_{i',j} \in P_{i'}$ such that $x = p_{i',j}$. This implies that to maintain the sorted order of D_i for any x , as we increase the value of x , we only need to update this order at the nk points in P_{flat} with respect to $D_{i'}$ for which there exists $p_{i',j} \in P_{i'}$ with $p_{i',j} = x$. This takes $O(\log(nk))$ time per update using a balanced BST, and thus $O(nk \log(nk))$ time to define $\text{cost}(x)$ for all values $x \in \mathbb{R}^1$. To compute $\text{cost}(x)$, we also require the values of L_i (or R_i); these can be constructed independently for each $i \in [n]$ in $O(k)$ time after sorting, and in $O(nk \log k)$ time overall.¹ Ultimately, we arrive at the following theorem.

► **Theorem 2.** *Consider a set of n uncertain points \mathcal{P} with k possible locations each. We can compute $\text{cost}(x)$ for all $x \in \mathbb{R}$ such that $x = p_{i,j}$ for some $p_{i,j} \in P_{\text{flat}}$ in $O(nk \log(nk))$ time.*

3.2 Building the ε -support T and bounding its size

We next show that there always exists an ε -support T and it has a size $|T| = O(\frac{k}{\varepsilon})$.

► **Theorem 3.** *Given a set of n uncertain points $\mathcal{P} = \{P_1, \dots, P_n\}$, where $P_i = \{p_{i,1}, \dots, p_{i,k}\} \subset \mathbb{R}$, and $\varepsilon \in (0, 1]$ we can construct an ε -support T that has a size $|T| = O(\frac{k}{\varepsilon})$.*

Proof. We first sort P_{flat} in ascending order, scan $P_{\text{flat}} = \{p_1, \dots, p_{nk}\}$ from left to right and choose one point from P_{flat} every $\lfloor \frac{n}{3} \rfloor$ points, and then put the chosen point into T . Now, suppose p is the median of some traversal $Q \subseteq \mathcal{P}$ and $\text{cost}(p) = \text{cost}(p, Q)$. If $p \notin T$, then there are two consecutive points t, t' in T such that $t < p < t'$. On either side of p there are at least $\lfloor \frac{n}{2} \rfloor$ points in Q , so without loss of generality, we assume $|p - t'| \geq \frac{1}{2}|t - t'|$. Since $|(p, \infty) \cap Q| \geq \frac{n}{2}$ and there are at most $\lfloor \frac{n}{3} \rfloor$ points in $[p, t']$, we have $|(t', \infty) \cap Q| \geq \frac{n}{2} - \lfloor \frac{n}{3} \rfloor \geq \frac{n}{6}$, which implies

$$\begin{aligned} \text{cost}(p) = \text{cost}(p, Q) &\geq \frac{1}{n} \sum_{q \in (t', \infty) \cap Q} |q - p| \geq \frac{1}{n} \sum_{q \in (t', \infty) \cap Q} |t' - p| \\ &\geq \frac{1}{n} \frac{n}{6} |t' - p| = \frac{1}{6} |t' - p| \geq \frac{1}{12} |t - t'|. \end{aligned} \quad (1)$$

For any fixed $\varepsilon \in (0, 1]$, and two consecutive points t, t' ($t < t'$) in T , we put $x_1, \dots, x_{\lceil \frac{12}{\varepsilon} \rceil - 1}$ into T where $x_i = t + \frac{|t-t'|i}{\lceil \frac{12}{\varepsilon} \rceil}$ for $1 \leq i \leq \lceil \frac{12}{\varepsilon} \rceil - 1$. So, for the median $p \in (t, t')$, there exists $x_i \in T$ s.t. $|p - x_i| \leq \frac{\varepsilon}{12} |t - t'|$, and from (1), we know $|p - x_i| \leq \varepsilon \text{cost}(p)$. In total we put $O(\frac{k}{\varepsilon})$ points into T ; thus the proof is completed. ◀

¹ When multiple distinct $p_{i,j}$ coincide at a point x , then more care may be required to compute $\text{cost}(x)$ (depending on the specifics of how the median is defined in these boundary cases). Specifically, we may not want to set $L_i(x) = 0$, instead it may be better to use the value $R_i(x)$ even if $R_i(x) = \alpha > 0$. This is the case when $\alpha < R_{i'}(x) - L_{i'}(x)$ for some other uncertain point $P_{i'}$ (then we say P_i is on the right, and P_i is on the left). This can be resolved by either tweaking the definition of median for these cases, or sorting all $D_i(x)$ for uncertain points P_i with some $p_{i,j} = x$, and some bookkeeping.

► **Remark.** The above construction results in an ε -support T of size $O(k/\varepsilon)$, but does not restrict that $T \subset P_{\text{flat}}$. We can enforce this restriction by for each x placed in T to choose the single nearest point $p \in P_{\text{flat}}$ to replace it in T . This results in an (2ε) -support, which can be made an ε -support by instead adding $\lceil \frac{24}{\varepsilon} \rceil - 1$ points between each pair (t, t') , without affecting the asymptotic time bound.

► **Remark.** We can construct a sequence of uncertain data $\{\mathcal{P}(n, k)\}$ such that, for each uncertain data $\mathcal{P}(n, k)$, the optimal ε -support T has a size $\Omega(\frac{k}{\varepsilon})$. For example, for $\varepsilon = \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \dots$, we define $n = \frac{1}{\varepsilon}$, and $p_{i,j} = (j - 1)n + i$ for $i \in [n]$ and $j \in [k]$. Then, for any median $p \in P_{\text{flat}}$, we have $\varepsilon \text{cost}(p) = \frac{2}{n^2} \sum_{i=1}^{\frac{n-1}{2}} i = \frac{n^2-1}{4n^2} < \frac{1}{4}$, hence covering no other points, which implies $|T| = \Omega(nk) = \Omega(\frac{k}{\varepsilon})$.

We can construct the minimal size ε -support T in $O(nk \log(nk))$ time by sorting, and greedily adding the smallest point not yet covered each step. This yields the slightly stronger corollary of Theorem 3.

► **Corollary 4.** Consider a set of n uncertain points $\mathcal{P} = \{P_1, \dots, P_n\}$, where $P_i = \{p_{i,1}, \dots, p_{i,k}\} \subset \mathbb{R}$, and $\varepsilon \in (0, 1]$. We can construct an ε -support T in $O(nk \log(nk))$ time which has the minimal size for any ε -support, and $|T| = O(\frac{k}{\varepsilon})$.

There are multiple ways to generalize the notion of a median to higher dimensions [5]. We focus on two variants: the Tukey median and the geometric median. We start with generalizing the notion of an ε -support to a Tukey median since it more directly follows from the techniques in Theorem 3, and then address the geometric median.

3.3 An ε -Support for the Tukey median

A closely related concept to the Tukey median is a *centerpoint*, which is a point p such that $\text{depth}_Q(p) \geq \frac{1}{d+1}|Q|$. Since for any finite set $Q \in \mathbb{R}^d$ its centerpoint always exists, a Tukey median must be a centerpoint. This means if p is the Tukey median of Q , then for any closed half space containing p , it contains at least $\frac{1}{d+1}|Q|$ points of Q . Using this property, we can prove the following theorem.

► **Theorem 5.** Given a set of n uncertain points $\mathcal{P} = \{P_1, \dots, P_n\}$, where $P_i = \{p_{i,1}, \dots, p_{i,k}\} \subset \mathbb{R}^2$, and $\varepsilon \in (0, 1]$, we can construct an ε -support T for the Tukey median on \mathcal{P} that has a size $|T| = O(\frac{k^2}{\varepsilon^2})$.

Proof. Suppose the projections of P_{flat} on x -axis and y -axis are X and Y respectively. We sort all points in X and choose one point from X every $\lfloor \frac{n}{4} \rfloor$ points, and then put the chosen points into a set X_T . For each point $x \in X_T$ we draw a line through $(x, 0)$ parallel to y -axis. Similarly, we sort all points in Y and choose one point every $\lfloor \frac{n}{4} \rfloor$ points, and put the chosen points into Y_T . For each point $y \in Y_T$ we draw a line through $(0, y)$ parallel to x -axis.

Now, suppose p with coordinates (x_p, y_p) is the Tukey median of some traversal $Q \in \mathcal{P}$ and $\text{cost}(p, Q) = \frac{1}{n} \sum_{q \in Q} \|q - p\|$. If $x_p \notin X_T$ and $y_p \notin Y_T$, then there are $x, x' \in X_T$ and $y, y' \in Y_T$ such that $x < x_p < x'$ and $y < y_p < y'$, as shown in Figure 2(a).

Without loss of generality, we assume $|x_p - x| \geq \frac{1}{2}|x' - x|$ and $|y_p - y| \geq \frac{1}{2}|y - y'|$. Since p is the Tukey median of Q , we have $|Q \cap (-\infty, \infty) \times (-\infty, y_p]| \geq \frac{n}{3}$ where $(-\infty, \infty) \times (-\infty, y_p] = \{(x, y) \in \mathbb{R}^2 \mid y \leq y_p\}$. Recall there are at most $\lfloor \frac{n}{4} \rfloor$ points of P_{flat} in $(-\infty, \infty) \times [y_p, y]$, which implies $|Q \cap (-\infty, \infty) \times (-\infty, y)| \geq \frac{n}{3} - \lfloor \frac{n}{4} \rfloor \geq \frac{n}{12}$. So, we have

$$\text{cost}(p, Q) \geq \frac{1}{n} \sum_{q \in Q \cap (-\infty, \infty) \times (-\infty, y)} \|q - p\| \geq \frac{1}{n} \frac{n}{12} |y - y_p| \geq \frac{1}{24} |y - y'|.$$

16:8 Approximating the Distribution of the Median on Uncertain Data

Using a symmetric argument, we can obtain $\text{cost}(p, Q) \geq \frac{1}{24}|x - x'|$.

For any fixed $\varepsilon \in (0, 1]$, and any two consecutive points x, x' in X_T we put $x_1, \dots, x_{\lceil \frac{48}{\varepsilon} \rceil - 1}$ into X_T where $x_i = x + \frac{|x-x'|i}{\lceil \frac{48}{\varepsilon} \rceil}$. Also, for any two consecutive point y, y' in Y_T , we put $y_1, \dots, y_{\lceil \frac{48}{\varepsilon} \rceil - 1}$ into Y_T where $y_i = y + \frac{|y-y'|i}{\lceil \frac{48}{\varepsilon} \rceil}$. So, for the Tukey median $p \in (x, x') \times (y, y')$, there exist $x_i \in X_T$ and $y_j \in Y_T$ such that $|x_p - x_i| \leq \frac{\varepsilon}{48}|x - x'|$ and $|y_p - y_j| \leq \frac{\varepsilon}{48}|y - y'|$. Since we have shown that $\frac{1}{24}|x - x'|$ and $\frac{1}{24}|y - y'|$ are lower bounds for $\text{cost}(p, Q)$, we obtain

$$\begin{aligned} \|(x_p, y_p) - (x_i, y_j)\| &\leq |x_p - x_i| + |y_p - y_j| \leq \frac{\varepsilon}{48}(|x - x'| + |y - y'|) \\ &\leq \frac{\varepsilon}{48}(24\text{cost}(p, Q) + 24\text{cost}(p, Q)) = \varepsilon\text{cost}(p, Q). \end{aligned}$$

Finally, we define T as $T := X_T \times Y_T$. Then for any $Q \in \mathcal{P}$, if p is the Tukey median of Q , there exists $t \in T$ such that $\|t - p\| \leq \varepsilon\text{cost}(p, Q)$. Thus, T is an ε -support for the Tukey median on \mathcal{P} . Moreover, since $|X_T| = O(\frac{k}{\varepsilon})$ and $|Y_T| = O(\frac{k}{\varepsilon})$, we have $|T| = O(\frac{k^2}{\varepsilon^2})$. ◀

In a straight-forward extension, we can generalize the result of Theorem 5 to d dimensions.

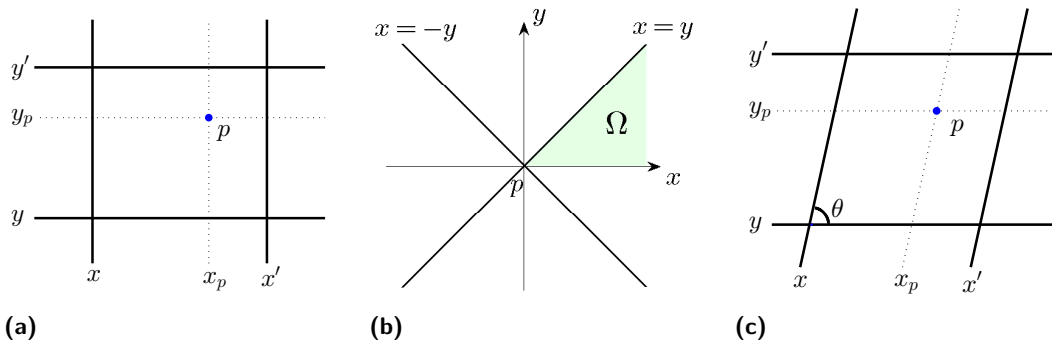
► **Theorem 6.** *Given a set of n uncertain points $\mathcal{P} = \{P_1, \dots, P_n\}$, where $P_i = \{p_{i,1}, \dots, p_{i,k}\} \subset \mathbb{R}^d$, and $\varepsilon \in (0, 1]$, we can construct an ε -support T for the Tukey median on \mathcal{P} that has a size $|T| = O((2d(d+1)(d+2)^2 \frac{k}{\varepsilon})^d)$.*

3.4 An ε -support for the geometric median

Unlike the Tukey median, there does not exist a constant $C > 0$ such that: for any geometric median p of point set $Q \subset \mathbb{R}^d$, any closed halfspace containing p contains at least $\frac{1}{C}|Q|$ points of Q . For example, suppose in \mathbb{R}^2 there are $2n + 1$ points on x -axis with the median point at the origin; this point is also the geometric median. If we move this point upward along the y direction, then the geometric median also moves upwards. However, for the line through the new geometric median and parallel to the x -axis, all $2n$ other points are under this line.

Hence, we need a new idea to adapt the method in Theorem 6 for the geometric median in \mathbb{R}^d . We first consider the geometric median in \mathbb{R}^2 . We show we can find *some* line through it, such that on both sides of this line there are at least $\frac{n}{8}$ points.

► **Lemma 7.** *Suppose p is the geometric median of $Q \subset \mathbb{R}^2$ with size $|Q| = n$. There is a line ℓ through p so both closed half planes with ℓ as boundary contain at least $\frac{n}{8}$ points of Q .*



■ **Figure 2** (a) Tukey median p is in a grid cell formed by x, x' and y, y' . (b) The plane is decomposed into 8 regions with the same shape. (c) Geometric median p is in an oblique grid cell formed by x, x' and y, y' .

Proof. We first build a rectangular coordinate system at the point p , which means p is the origin with coordinates $(x_p, y_p) = (0, 0)$. Then we use the x -axis, y -axis and lines $x = y$, $x = -y$ to decompose the plane into eight regions, as shown in Figure 2(b). Since all these eight regions have the same shape, without loss of generality, we can assume $\Omega = \{(x, y) \in \mathbb{R}^2 \mid x \geq y \geq 0\}$ contains the most points of Q . Then $|\Omega \cap Q| \geq \frac{n}{8}$, otherwise $n = |Q| = |\mathbb{R}^2 \cap Q| \leq 8|\Omega \cap Q| < n$, which is a contradiction.

If $|Q \cap \{p\}| \geq \frac{n}{8}$, i.e., the multiset Q contains p at least $\frac{n}{8}$ times, then obviously this proposition is correct. So, we only need to consider the case $|Q \cap \{p\}| < \frac{n}{8}$. We introduce notations $\tilde{\Omega} = \Omega \setminus \{p\}$ and $\Omega^o = \Omega \setminus \partial\Omega$, and denote the coordinates of any $q \in Q$ as $q = (x_q, y_q)$. From a property of the geometric median (proven in the full version [8]) we know $\sum_{q \in Q \setminus \{p\}} \frac{x_q - x_p}{\|q - p\|} \leq |Q \cap \{p\}|$. Since p is the origin and $Q \setminus \{p\} = (Q \cap \tilde{\Omega}) \cup (Q \setminus \Omega)$, $|Q \cap \{p\}| < \frac{n}{8}$ implies $\sum_{q \in Q \cap \tilde{\Omega}} \frac{x_q}{\|q\|} + \sum_{q \in Q \setminus \Omega} \frac{x_q}{\|q\|} < \frac{n}{8}$. From $\frac{x_q}{\|q\|} = \frac{x_q}{\sqrt{x_q^2 + y_q^2}} \geq \frac{1}{\sqrt{2}}$, $\forall q \in \tilde{\Omega}$ we obtain

$$|Q \cap \tilde{\Omega}| \frac{1}{\sqrt{2}} \leq \sum_{q \in Q \cap \tilde{\Omega}} \frac{x_q}{\|q\|} < \frac{n}{8} - \sum_{q \in Q \setminus \Omega} \frac{x_q}{\|q\|} \leq \frac{n}{8} + |Q \setminus \Omega| \leq \frac{n}{8} + (n - |Q \cap \tilde{\Omega}|)$$

which implies there are not too many points in $\tilde{\Omega}$,

$$|Q \cap \tilde{\Omega}| < \frac{\sqrt{2}n}{(1 + \sqrt{2})} \cdot \frac{9}{8} < 0.66n.$$

Now, we define the two pairs of halfspaces which share a boundary with $\tilde{\Omega}$: $H_1^+ = \{(x, y) \in \mathbb{R}^2 \mid y \geq 0\}$, $H_1^- = \{(x, y) \in \mathbb{R}^2 \mid y \leq 0\}$ and $H_2^+ = \{(x, y) \in \mathbb{R}^2 \mid x - y \geq 0\}$, $H_2^- = \{(x, y) \in \mathbb{R}^2 \mid x - y \leq 0\}$. We assert either $|H_1^+ \cap Q| \geq \frac{n}{8}$ and $|H_1^- \cap Q| \geq \frac{n}{8}$, or $|H_2^+ \cap Q| \geq \frac{n}{8}$ and $|H_2^- \cap Q| \geq \frac{n}{8}$. Otherwise, since $|Q \cap \Omega| \geq \frac{n}{8}$ and $\Omega \subset H_1^+ \cap H_2^+$, we have $|H_1^- \cap Q| < \frac{n}{8}$ and $|H_2^- \cap Q| < \frac{n}{8}$. From $H_1^- \cup H_2^- \cup \Omega^o = \mathbb{R}^2$ we have

$$\begin{aligned} n = |Q| &= |\mathbb{R}^2 \cap Q| = |(H_1^- \cup H_2^- \cup \Omega^o) \cap Q| \leq |H_1^- \cap Q| + |H_2^- \cap Q| + |\Omega^o \cap Q| \\ &\leq |H_1^- \cap Q| + |H_2^- \cap Q| + |\tilde{\Omega} \cap Q| \leq \frac{n}{8} + \frac{n}{8} + 0.66n < n, \end{aligned}$$

which is a contradiction. Therefore, among lines $\ell_1 : y = 0$ and $\ell_2 : x - y = 0$, which both go through p , one of them has at least $n/8$ points from Q on both sides. ◀

► **Theorem 8.** *Given a set of n uncertain points $\mathcal{P} = \{P_1, \dots, P_n\}$, where $P_i = \{p_{i,1}, \dots, p_{i,k}\} \subset \mathbb{R}^2$, and $\varepsilon \in (0, 1]$, we can construct an ε -support T for the geometric median on \mathcal{P} that has a size $|T| = O(\frac{k^2}{\varepsilon^2})$.*

Proof. The idea to prove this theorem is to use several oblique coordinate systems. We consider an oblique coordinate system, the angle between x -axis and y -axis is $\theta \in (0, \frac{\pi}{2}]$, and use the technique in Theorem 5 to generate a grid. More precisely, we project P_{flat} onto the x -axis along the y -axis of the oblique coordinate system to obtain a set X , sort all points in X , and choose one point from X every $\lfloor \frac{n}{9} \rfloor$ points to form a set X_T . Then we use the same method to generate Y and Y_T projecting along the x -axis in the oblique coordinate system. For each point $x \in X_T$ we draw a line through $(x, 0)$ parallel to the (oblique) y -axis, and for each point $y \in Y_T$ we draw a line through $(0, y)$ parallel to the (oblique) x -axis.

Let p with coordinates (x_p, y_p) be the geometric median of some traversal $Q \subseteq \mathcal{P}$ and $\text{cost}(p, Q) = \frac{1}{n} \sum_{q \in Q} \|q - p\|$. If $x_p \notin X_T$ and $y_p \notin Y_T$, then there are $x, x' \in X_T$ and $y, y' \in Y_T$ such that $x_p \in (x, x')$ and $y_p \in (y, y')$, as shown in Figure 2(c).

16:10 Approximating the Distribution of the Median on Uncertain Data

If we have the condition:

$$\begin{aligned} |Q \cap (-\infty, \infty) \times (-\infty, y_p]| &\geq \frac{n}{8}, & |Q \cap (-\infty, \infty) \times [y_p, \infty)| &\geq \frac{n}{8}, \\ |Q \cap (-\infty, x_p] \times (-\infty, \infty)| &\geq \frac{n}{8}, & |Q \cap [x_p, \infty) \times (-\infty, \infty)| &\geq \frac{n}{8}, \end{aligned} \quad (2)$$

then we can make the following computation.

Without loss of generality, we assume $|x_p - x| \geq \frac{1}{2}|x' - x|$ and $|y_p - y| \geq \frac{1}{2}|y - y'|$. There are at most $\lfloor \frac{n}{9} \rfloor$ points of P_{flat} in $(-\infty, \infty) \times [y_p, y]$, which implies $|Q \cap (-\infty, \infty) \times (-\infty, y)| \geq \frac{n}{8} - \lfloor \frac{n}{9} \rfloor \geq \frac{n}{72}$. So, we have

$$\text{cost}(p, Q) \geq \frac{1}{n} \sum_{q \in Q \cap (-\infty, \infty) \times (-\infty, y)} \|q - p\| \geq \frac{1}{n} \frac{n}{72} |y - y_p| \geq \frac{\sin(\theta)}{144} |y - y'|.$$

Similarly, we can prove $\text{cost}(p, Q) \geq \frac{\sin(\theta)}{144} |x - x'|$.

For any fixed $\varepsilon \in (0, 1]$, and any two consecutive points x, x' in X_T we put $x_1, \dots, x_{\lfloor \frac{288}{\varepsilon \sin(\theta)} \rfloor - 1}$ into X_T where $x_i = x + \frac{|x - x'|i}{\lfloor \frac{288}{\varepsilon \sin(\theta)} \rfloor}$. Also, for any two consecutive point y, y' in Y_T , we put $y_1, \dots, y_{\lfloor \frac{288}{\varepsilon \sin(\theta)} \rfloor - 1}$ into Y_T where $y_i = y + \frac{|y - y'|i}{\lfloor \frac{288}{\varepsilon \sin(\theta)} \rfloor}$. So, for the L_1 median $p \in (x, x') \times (y, y')$, there exist $x_i \in X_T$ and $y_j \in Y_T$ such that $|x_p - x_i| \leq \frac{\varepsilon \sin(\theta)}{288} |x - x'|$ and $|y_p - y_j| \leq \frac{\varepsilon \sin(\theta)}{288} |y - y'|$. Since we have shown that both $\frac{\sin(\theta)}{144} |x - x'|$ and $\frac{\sin(\theta)}{144} |y - y'|$ are lower bounds for $\text{cost}(p, Q)$, using the distance formula in an oblique coordinate system, we have

$$\begin{aligned} \|(x_p, y_p) - (x_i, y_j)\| &\leq ((x_p - x_i)^2 + (y_p - y_j)^2 + 2(x_p - x_i)(y_i - y_p) \cos(\theta))^{\frac{1}{2}} \\ &\leq ((x_p - x_i)^2 + (y_p - y_j)^2 + 2|x_p - x_i||y_i - y_p|)^{\frac{1}{2}} \\ &= |x_p - x_i| + |y_p - y_j| \leq \frac{\varepsilon \sin(\theta)}{288} (|x - x'| + |y - y'|) \\ &\leq \frac{\varepsilon \sin(\theta)}{288} \left(\frac{144}{\sin(\theta)} \text{cost}(p, Q) + \frac{144}{\sin(\theta)} \text{cost}(p, Q) \right) = \varepsilon \text{cost}(p, Q). \end{aligned}$$

Therefore, if all k^n geometric medians of traversals satisfy (2) and $\theta \in (0, \frac{\pi}{2}]$ is a constant then $T = X_T \times Y_T$ is an ε -support of size $O\left(\frac{k^2}{(\sin(\theta)\varepsilon)^2}\right)$ for the geometric median on \mathcal{P} .

Although we cannot find an oblique coordinate system to make (2) hold for all k^n medians, we can use several oblique coordinate systems. Using the result of Lemma 7, for any geometric median of n points Q , we know there exists a line ℓ through p and parallel to a line in $\{\ell_1 : y = 0, \ell_2 : x - y = 0, \ell_3 : x = 0, \ell_4 : x + y = 0\}$, such that in both sides of this line, there are at least $\frac{n}{8}$ points of Q . Since we did not make any assumption on the distribution of points in Q , if we rotate $\ell_1, \ell_2, \ell_3, \ell_4$ anticlockwise by $\frac{\pi}{8}$ around the origin, we can obtain four lines $\ell'_1, \ell'_2, \ell'_3, \ell'_4$, and there exists a line ℓ' through p and parallel to a line in $\{\ell'_1, \ell'_2, \ell'_3, \ell'_4\}$, such that on both sides of this line, there are at least $\frac{n}{8}$ points of Q . The angle between ℓ and ℓ' is at least $\frac{\pi}{8}$.

Therefore, given $\mathcal{L} = \{\ell_1, \ell_2, \ell_3, \ell_4\}$ and $\mathcal{L}' = \{\ell'_1, \ell'_2, \ell'_3, \ell'_4\}$, for each pair $(\ell, \ell') \in \mathcal{L} \times \mathcal{L}'$, we take ℓ and ℓ' as x -axis and y -axis respectively to build an oblique coordinate system, and then use the above method to compute a set $T(\ell, \ell')$. Since for any geometric median p there must be an oblique coordinate system based on some $(\ell, \ell') \in \mathcal{L} \times \mathcal{L}'$ to make (2) hold for p , we can take $T = \cup_{\ell \in \mathcal{L}, \ell' \in \mathcal{L}'} T(\ell, \ell')$ as an ε -support for geometric median on \mathcal{P} , and the size of T is $|T| = O\left(16 \frac{k^2}{(\sin(\frac{\pi}{8})\varepsilon)^2}\right) = O\left(\frac{k^2}{\varepsilon^2}\right)$. \blacktriangleleft

The result of Theorem 8 can be generalized to \mathbb{R}^d and details are in the full version [8].

3.5 Assigning a weight to T in \mathbb{R}^1

Here we provide an algorithm to assign a weight to T in \mathbb{R}^1 , which approximates the probability distribution of median. For T in \mathbb{R}^d , we provide a randomized algorithm in Section 4.1.

Define the weight of $p_{i,j} \in P_{\text{flat}}$ as $w(p_{i,j}) = \frac{1}{k^n} |\{Q \in \mathcal{P} \mid p_{i,j} \text{ is the median of } Q\}|$, the probability it is the median. Suppose T is constructed by our greedy algorithm for \mathbb{R}^1 . For $p_{i,j} \in P_{\text{flat}}$, we introduce a map $f_T : P_{\text{flat}} \rightarrow T$,

$$f_T(p_{i,j}) = \arg \min \{ |x - p_{i,j}| \mid x \in T, |x - p_{i,j}| \leq \varepsilon \text{cost}(p_{i,j}) \},$$

where $\text{cost}(p_{i,j}) = \min \{ \text{cost}(p_{i,j}, Q) \mid p_{i,j} \text{ is the median of } Q \text{ and } Q \in \mathcal{P} \}$.

Intuitively, this maps each $p_{i,j} \in P_{\text{flat}}$ 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 $\hat{w}(x) = \sum_{\{p_{i,j} \in P_{\text{flat}} \mid f_T(p_{i,j})=x\}} w(p_{i,j})$. So we first compute the weight of each point in P_{flat} and then obtain the weight of points in T in another linear sweep. Our ability to calculate the weights w for each point in P_{flat} is summarized in the next lemma. The algorithm, explained within the proof, is a dynamic program that expands a specific polynomial similar to Li *et.al.* [14], where in the final state, the coefficients correspond with the probability of each point being the median.

► **Lemma 9.** *We can output $w(p_{i,j})$ 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 the following terms to count the number of points to the left (l_j) or right (r_j) of it in the j th uncertain point (excluding P_{i_0}):

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

Then, if n is odd, we can write the weight of p_{i_0} as

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

where $S_1 = \{i_1, i_2, \dots, i_{\frac{n-1}{2}}\}$ and $S_2 = \{j_1, j_2, \dots, j_{\frac{n-1}{2}}\}$. This sums over all partitions S_1, S_2 of uncertain points on the left or right of p_{i_0} for which it is the median, and each term is the product of ways each uncertain point can be on the appropriate side. We define $w(p_{i_0})$ similarly when n is even, then the last index of S_2 is $j_{\frac{n}{2}}$.

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

$$(l_1x + r_1)(l_2x + r_2) \cdots (l_{n-1}x + r_{n-1}), \tag{3}$$

where $\sum_{\substack{S_1 \cap S_2 = \emptyset \\ S_1 \cup S_2 = \{1, \dots, n-1\}}} (l_{i_1} \cdot l_{i_2} \cdot \dots \cdot l_{i_{\frac{n-1}{2}}} \cdot r_{j_1} \cdot r_{j_2} \cdot \dots \cdot r_{j_{\frac{n-1}{2}}})$ is the coefficient of $x^{\frac{n-1}{2}}$. We define $\rho_{i,j}$ ($1 \leq i \leq n-1, 0 \leq j \leq i$) 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 $\rho_{i,j} = l_i \rho_{i-1,j-1} + r_i \rho_{i-1,j}$. Thus we can use dynamic programming to compute $\rho_{n-1,0}, \rho_{n-1,1}, \dots, \rho_{n-1,n-1}$, as shown in Algorithm 1.

Thus Algorithm 1 computes the weight $\frac{1}{k^n} w(p_{i_0}) = \rho_{n-1, \frac{n-1}{2}}$ for a single $p_{i_0} \in P_{\text{flat}}$. Next we show, we can reuse much of the structure to compute the weight for another point; this will ultimately shave a factor n off of running Algorithm 1 nk times.

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

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 $\rho_{n-1,0}, \rho_{n-1,1}, \dots, \rho_{n-1,n-1}$.

Suppose for $p_{i_0} \in P_{i_0}$ we have obtained $\rho_{n-1,0}, \rho_{n-1,1}, \dots, \rho_{n-1,n-1}$ by Algorithm 1, and then we consider $p_{i'_0} = \min\{p \in P_{\text{flat}} \setminus P_{i_0} \mid p \geq p_{i_0}\}$. 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 $\tilde{l}'_{i'_0} = \tilde{l}'_{i'_0-1} = |\{p \in P_{i_0} \mid p \leq p_{i'_0}\}|$ and $\tilde{r}'_{i'_0} = \tilde{r}'_{i'_0-1} = |\{p \in P_{i_0} \mid p \geq p_{i'_0}\}|$.

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 $(l_1x + r_1) \cdots (l'_{i'_0-1}x + r'_{i'_0-1})(l'_{i'_0+1}x + r'_{i'_0+1}) \cdots (l_{n-1}x + r_{n-1})$ from $\rho_{n-1,0}, \rho_{n-1,1}, \dots, \rho_{n-1,n-1}$, 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 each other point. The whole time is $O(n^2) + nkO(n) = O(n^2k)$. ◀

► **Corollary 10.** We can assign $\hat{w}(x)$ to each $x \in T$ in \mathbb{R}^1 in $O(n^2k)$ time.

4 A randomized algorithm to construct a covering set

In this section we describe a much more general randomized algorithm for robust estimators on uncertain data. It constructs an approximate covering set of the support of the distribution of the estimator, and estimates the weight at the same time. The support of the distribution is not as precise compared to the techniques in the previous section in that the new technique may fail to cover regions with small probability of containing the estimator.

Suppose $\mathcal{P} = \{P_1, P_2, \dots, P_n\}$ is a set of uncertain data, where for $i \in [n]$, $P_i = \{p_{i,1}, p_{i,2}, \dots, p_{i,k}\} \subseteq \mathcal{X}$ for some domain \mathcal{X} . An estimator $E : \{Q \mid Q \subseteq \mathcal{P}\} \mapsto Y$ maps $Q \subseteq \mathcal{P}$ to a metric space (Y, φ) . Let $B(y, r) = \{y' \in Y \mid \varphi(y, y') \leq r\}$ be a ball of radius r in that metric space. We denote ν as the VC-dimension of the range space (Y, \mathcal{R}) induced by these balls, with $\mathcal{R} = \{B(y, r) \mid y \in Y, r \geq 0\}$.

We now analyze the simple algorithm which randomly instantiates traversals $Q \subseteq \mathcal{P}$, and constructs their estimators $z = E(Q)$. Repeating this N times builds a domain $T = \{z_1, z_2, \dots, z_N\}$ each with weight $w(z_i) = 1/N$.

► **Theorem 11.** For $\varepsilon > 0$ and $\delta \in (0, 1)$, set $N = O((1/\varepsilon^2)(\nu + \log(1/\delta)))$. Then, with probability at least $1 - \delta$, for any $B \in \mathcal{R}$ we have $|\sum_{z \in T \cap B} w(z) - \Pr_{Q \subseteq \mathcal{P}}[E(Q) \in B]| \leq \varepsilon$.

Proof. Let T^* be the true support of $E(Q)$ where $Q \subseteq \mathcal{P}$, and let $w^* : T^* \rightarrow \mathbb{R}^+$ be the true probability distribution defined on T^* ; e.g., for discrete T^* , then for any $z' \in T^*$,

$w^*(z') = \Pr_{Q \in \mathcal{P}}[E(Q) = z']$. Then each random z generated is a random draw from w^* . Hence for a range space with bounded VC-dimension [21] ν , we can apply the sampling bound [15] for ε -approximations of these range spaces to prove our claim. \blacktriangleleft

In Theorem 11, for $z_i \in T$, if we choose $B = B(z_i, r) \in \mathcal{R}$ with r small enough such that $T \cap B$ only contains z_i , then we obtain the following.

► **Corollary 12.** *For $\varepsilon > 0$ and $\delta \in (0, 1)$, set $N = O((1/\varepsilon^2)(\nu + \log(1/\delta)))$. Then, with probability at least $1 - \delta$, for any $z \in Y$ we have $|w(z) - \Pr_{Q \in \mathcal{P}}[E(Q) = z]| \leq \varepsilon$.*

► **Remark.** We can typically define a metric space (Y, φ) where $\nu = O(1)$; for instance for point estimators (e.g., the geometric median), define a projection into \mathbb{R}^1 so no z_i s map to the same point, then define distance φ as restricted to the distance along this line, so metric balls are intervals (or slabs in \mathbb{R}^d); these have $\nu = 2$.

4.1 Application to geometric median

For each $Q \in \mathcal{P}$, the geometric median m_Q 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. But it is straightforward to apply this randomized approach. For $P_{\text{flat}} \in \mathbb{R}^d$, the natural metric space (Y, φ) is $Y = \mathbb{R}^d$ and φ as the Euclidian distance.

However, there is no known closed form solution for the geometric median; it can be computed within any additive error ϕ through various methods [22, 9, 7, 6]. As such, we can state a slightly more intricate corollary.

► **Corollary 13.** *Set $\varepsilon > 0$ and $\delta \in (0, 1)$ and $N = O((1/\varepsilon^2)(d + \log(1/\delta)))$. For an uncertain point set \mathcal{P} with $P_{\text{flat}} \subset \mathbb{R}^d$, let the estimator E be the geometric median, and let E_ϕ be an algorithm that finds an approximation to the geometric median within additive error $\phi > 0$. Run the algorithm using E_ϕ . Then for any ball $B = B(x, r) \in \mathcal{R}$, there exists² another ball $B' = B(x, r')$ with $|r - r'| \leq \phi$ such that with probability at least $1 - \delta$, $|\sum_{z \in T \cap B'} w(z) - \Pr_{Q \in \mathcal{P}}[E(Q) \in B]| \leq \varepsilon$.*

4.2 Application to Siegel estimator

The Siegel (repeated median) estimator [19] is a robust estimator S for linear regression in \mathbb{R}^2 with optimal breakdown point 0.5. For a set of points Q , for each $q_i \in Q$ it computes slopes of all lines through q_i and each other $q' \in Q$, and takes their median a_i . Then it takes the median a of the set $\{a_i\}_i$ of all median slopes. The offset b of the estimated line $\ell : y = ax + b$, is the median of $(y_i - ax_i)$ for all points $q_i = (x_i, y_i)$. For uncertain data $P_{\text{flat}} \subset \mathbb{R}^2$, we can directly apply our general technique for this estimator.

We use the following metric space (Y, φ) . Let $Y = \{\ell \mid \ell \text{ is a line in } \mathbb{R}^2 \text{ with form } y = ax + b, \text{ where } a, b \in \mathbb{R}\}$. Then let φ be the Euclidean distance in the standard dual; for two lines $\ell : y = ax + b$ and $\ell' : y = a'x + b'$, define $\varphi(\ell, \ell') = \sqrt{(a - a')^2 + (b - b')^2}$. By examining the dual space, we see that (Y, \mathcal{R}) with $\mathcal{R} = \{B(\ell, r) \mid \ell \in Y, r \geq 0\}$ and $B(\ell, r) = \{\ell' \in Y \mid \varphi(\ell, \ell') \leq r\}$ has a VC-dimension 3.

From the definition of the Siegel estimator [19], there can be at most $O(n^3 k^3)$ distinct lines in $T = \{S(Q) \mid Q \in \mathcal{P}\}$. By Corollary 12, setting $N = O((1/\varepsilon^2) \log(1/\delta))$, then with probability at least $1 - \delta$ for all $z \in T$ we have $|w(z) - \Pr_{Q \in \mathcal{P}}[S(Q) = z]| \leq \varepsilon$.

² To simplify the discussion on degenerate behavior, define ball B' , so any point q on its boundary can be defined inside or outside of B , and this decision can be different for each q , even if they are co-located.

References

- 1 Pankaj K. Agarwal, Boris Aronov, Sariel Har-Peled, Jeff M. Phillips, Ke Yi, and Wuzhou Zhang. Nearest-neighbor searching under uncertainty II. In *PODS*, 2013.
- 2 Pankaj K. Agarwal, Siu-Wing Cheng, Yufei Tao, and Ke Yi. Indexing uncertain data. In *PODS*, 2009.
- 3 Pankaj K. Agarwal, Alon Efrat, Swaminathan Sankararaman, and Wuzhou Zhang. Nearest-neighbor searching under uncertainty. In *PODS*, 2012.
- 4 Pankaj K. Agarwal, Sariel Har-Peled, Subhash Suri, Hakan Yildiz, and Wuzhou Zhang. Convex hulls under uncertainty. In *ESA*, 2014.
- 5 Greg Aloupis. Geometric measures of data depth. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*. AMS, 2006.
- 6 Sanjeev Arora, Prabhakar Raghavan, and Satish Rao. Approximation schemes for Euclidean k -medians and related problems. In *STOC*, 1998.
- 7 Prosenjit Bose, Anil Maheshwari, and Pat Morin. Fast approximations for sums of distances clustering and the Fermat-Weber problem. *CGTA*, 24:135–146, 2003.
- 8 Kevin Buchin, Jeff M. Phillips, and Pingfan Tang. Approximating the distribution of the median and other robust estimators on uncertain data. *ArXiv e-prints*, 2018. [arXiv: 1601.00630](https://arxiv.org/abs/1601.00630).
- 9 R. Chandrasekaran and A. Tamir. Algebraic optimization: The Fermat-Weber location problem. *Mathematical Programming*, 46:219–224, 1990.
- 10 Graham Cormode and Andrew McGregor. Approximation algorithms for clustering uncertain data. In *PODS*, 2008.
- 11 David Donoho and Peter 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. 1983.
- 12 Lingxiao Huang and Jian Li. Approximating the expected values for combinatorial optimization problems over stochastic points. In *ICALP*, 2015.
- 13 Allan G. Jørgensen, Maarten Löffler, and Jeff M. Phillips. Geometric computation on indecisive points. In *WADS*, 2011.
- 14 Jian Li, Barna Saha, and Amol Deshpande. A unified approach to ranking in probabilistic databases. In *VLDB*, 2009.
- 15 Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the samples complexity of learning. *Journal of Computer and System Science*, 62:516–527, 2001.
- 16 Maarten Löffler and Jeff Phillips. Shape fitting on point sets with probability distributions. In *ESA*, 2009.
- 17 Hendrik P. Lopuhaa and Peter J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19:229–248, 1991.
- 18 Peter J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, pages 283–297, 1985.
- 19 Andrew F. Siegel. Robust regression using repeated medians. *Biometrika*, 82:242–244, 1982.
- 20 J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the 1974 International Congress of Mathematics, Vancouver*, volume 2, pages 523–531, 1975.
- 21 Vladimir Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Th. Probability and Applications*, 16:264–280, 1971.
- 22 Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937.
- 23 Ying Zhang, Xuemin Lin, Yufei Tao, and Wenjie Zhang. Uncertain location based range aggregates in a multi-dimensional space. In *Proceedings 25th IEEE International Conference on Data Engineering*, 2009.