Range Counting Coresets for Uncertain Data*

Amirali Abdullah, Samira Daruki, and Jeff M. Phillips
School of Computing, University of Utah
Salt Lake City, Utah, USA
amirali@cs.utah.edu, daruki@cs.utah.edu, jeffp@cs.utah.edu

ABSTRACT

We study coresets for various types of range counting queries on uncertain data. In our model each uncertain point has a probability density describing its location, sometimes defined as k distinct locations. Our goal is to construct a subset of the uncertain points, including their locational uncertainty, so that range counting queries can be answered by just examining this subset. We study three distinct types of queries. RE queries return the expected number of points in a query range. RC queries return the number of points in the range with probability at least a threshold. RQ queries returns the probability that fewer than some threshold fraction of the points are in the range. In both RC and RQ coresets the threshold is provided as part of the query. And for each type of query we provide coreset constructions with approximation-size tradeoffs. We show that random sampling can be used to construct each type of coreset, and we also provide significantly improved bounds using discrepancy-based approaches on axis-aligned range queries.

Categories and Subject Descriptors

E.2 [**Data**]: Data Storage Representations;
G.3 [**Mathematics of Computing**]: Probability and Statistics.

Keywords

uncertain data, coresets, discrepancy

1. INTRODUCTION

A powerful notion in computational geometry is the *coreset* [3, 2, 10, 45]. Given a large data set P and a family of queries $\mathcal{A}$, then an η-*coreset* is a subset $S \subset P$ such that for all $r \in \mathcal{A}$ that $\|r(P) - r(S)\| \leq \eta$ (note the notion of distance $\|\cdot\|$ between query results is problem specific and is intentionally left ambiguous for now). Initially used for smallest enclosing ball queries [10] and perhaps most famous in geometry for extent queries as η-kernels [3, 2], the coreset is now employed in many other problems such as clustering [6] and density estimation [45]. Techniques for constructing coresets are becoming more relevant in the era of big data; they summarize a large data set P with a proxy set S of potentially much smaller size that can guarantee error for certain classes of queries. They also shed light onto the limits of how much information can possibly be represented in a small set of data.

In this paper we focus on a specific type of coreset called an η-sample [45, 21, 13] that can be thought of as preserving density queries and that has deep ties to the basis of learning theory [5]. Given a set of objects X (often $X \subset \mathbb{R}^d$ is a point set) and a family of subsets $\mathcal{A}$ of X, then the pair $(X, \mathcal{A})$ is called an *range space*. Often $\mathcal{A}$ are specified by containment in geometric shapes, for instance as all subsets of X defined by inclusion in any ball, any half space, or any axis-aligned rectangle. Now an η-sample of $(X, \mathcal{A})$ is a single subset $S \subset X$ such that

$$\max_{r \in \mathcal{A}} \left| \frac{|X \cap r|}{|X|} - \frac{|S \cap r|}{|S|} \right| \leq \eta.$$

For any query range $r \in \mathcal{A}$, subset S approximates the relative density of X in r with error at most η.

Uncertain points.

Another emerging notion in data analysis is modeling uncertainty in points. There are several formulations of these problems where each point $p \in P$ has an *independent* probability distribution μ_p describing its location and such a point is said to have *locational uncertainty*. *Imprecise points* (also called *deterministic uncertainty*) model where a data point $p \in P$ could be anywhere within a fixed continuous range and were originally used for analyzing precision errors. The worst case properties of a point set P under the imprecise model have been well-studied [19, 20, 7, 22, 33, 36, 37, 44, 30]. *Indecisive points* (or *attribute uncertainty* in database literature [40]) model each $p_i \in P$ as being able to take one of k distinct locations $\{p_{i,1}, p_{i,2}, \ldots, p_{i,k}\}$ with possibly different probabilities, modeling when multiple readings of the same object have been made [26, 43, 17, 16, 1, 4].

We also note another common model of *existential uncertainty* (similar to *tuple uncertainty* in database literature [40] but a bit less general) where the location or value of each $p \in P$ is fixed, but the point may not exist with some probability, modeling false readings [28, 27, 40, 17].

We will focus mainly on the indecisive model of locational

*Support by NSF CCF 0953066, CCF 1115677, and CPS 1035565.

SoCG'13, June 17-20, 2013, Rio de Janeiro, Brazil.

uncertainty since it comes up frequently in real-world applications [43, 4] (when multiple readings of the same object are made, and typically k is small) and can be used to approximately represent more general continuous representations [25, 38].

1.1 Problem Statement

Combining these two notions leads to the question: can we create a coreset (specifically for η-samples) of uncertain input data? A few more definitions are required to rigorously state this question. In fact, we develop three distinct notions of how to define the coreset error in uncertain points. One corresponds to range counting queries, another to querying the mean, and the third to querying the median (actually it approximates the rank for all quantiles).

For an uncertain point set $P = \{p_1, p_2, \ldots, p_n\}$ with each $p_i = \{p_{i,1}, p_{i,2}, \ldots p_{i,k}\} \subset \mathbb{R}^d$ we say that $Q \Subset P$ is a *transversal* if $Q \in p_1 \times p_2 \times \ldots \times p_n$. I.e., $Q = (q_1, q_2, \ldots, q_n)$ is an instantiation of the uncertain data P and can be treated as a "certain" point set, where each q_i corresponds to the location of p_i. $\mathbf{Pr}_{Q \Subset P}[\zeta(Q)]$, (resp. $\mathbf{E}_{Q \Subset P}[\zeta(Q)]$) represents the probability (resp. expected value) of an event $\zeta(Q)$ where Q is instantiated from P according to the probability distribution on the uncertainty in P.

As stated, our goal is to construct a subset of uncertain points $T \subset P$ (including the distribution of each point p's location, μ_p) that preserves specific properties over a family of subsets $(P, \mathcal{A})$. For completeness, the first variation we list cannot be accomplished purely with a coreset as it requires $\Omega(n)$ space.

- *Range Reporting (RR) Queries* support queries of a range $r \in \mathcal{A}$ and a threshold τ, and return all $p_i \in P$ such that $\mathbf{Pr}_{Q \Subset P}[q_i \in r] \geq \tau$. Note that the fate of each $p_i \in P$ depends on no other $p_j \in P$ where $i \neq j$, so they can be considered independently. Building indexes for this model have been studied [15, 18, 42, 47] and effectively solved in $\mathbb{R}^1$ [1].
- *Range Expectation (RE) Queries* consider a range $r \in \mathcal{A}$ and report the expected number of uncertain points in r, $\mathbf{E}_{Q \Subset P}[|r \cap Q|]$. The linearity of expectation allows summing the individual expectations each point $p \in P$ is in r. Single queries in this model have also been studied [23, 11, 24].
- *Range Counting (RC) Queries* support queries of a range $r \in \mathcal{A}$ and a threshold τ, but only return the number of $p_i \in P$ which satisfy $\mathbf{Pr}_{Q \Subset P}[q_i \in r] \geq \tau$. The effect of each $p_i \in P$ on the query is separate from that of any other $p_j \in P$ where $i \neq j$. A random sampling heuristic [46] has been suggested without proof of accuracy.
- *Range Quantile (RQ) Queries* take a query range $r \in \mathcal{A}$, and report the full cumulative density function on the number of points in the range $\mathbf{Pr}_{Q \Subset P}[|r \cap Q|]$. Thus for a query range r, this returned structure can produce for any value $\tau \in [0, 1]$ the probability that τn or fewer points are in r. Since this is no longer an expectation, the linearity of expectation cannot be used to decompose this query along individual uncertain points.

Across all queries we consider, there are two main ways we can approximate the answers. The first and most standard way is to allow an ε-error (for $0 \leq \varepsilon \leq 1$) in the returned answer for RQ, RE, and RC. The second way is to allow an α-error in the *threshold* associated with the query itself. As will be shown, this is not necessary for RR, RE, or RC, but is required to get useful bounds for RQ. Finally, we will also consider probabilistic error δ, demarcating the probability of failure in a randomized algorithm (such as random sampling). We strive to achieve these approximation factors with a small size coreset $T \subset P$ as follows:

RE: For a given range r, let $r(Q) = |Q \cap r|/|Q|$, and let $E_{r(P)} = \mathbf{E}_{Q \Subset P}[r(Q)]$. $T \subset P$ is an *ε-RE coreset* of $(P, \mathcal{A})$ if for all queries $r \in \mathcal{A}$ we have $\left|E_{r(P)} - E_{r(T)}\right| \leq \varepsilon$.

RC: For a range $r \in \mathcal{A}$, let $G_{P,r}(\tau) = \frac{1}{|P|}\left|\{p_i \in P \mid \mathbf{Pr}_{Q \Subset P}[q_i \in r] \geq \tau\}\right|$ be the fraction of points in P that are in r with probability at least some threshold τ. Then $T \subset P$ is an *ε-RC coreset* of $(P, \mathcal{A})$ if for all queries $r \in \mathcal{A}$ and all $\tau \in [0, 1]$ we have $|G_{P,r}(\tau) - G_{T,r}(\tau)| \leq \varepsilon$.

RQ: For a range $r \in \mathcal{A}$, let $F_{P,r}(\tau) = \mathbf{Pr}_{Q \Subset P}[r(Q) \leq \tau] = \mathbf{Pr}_{Q \Subset P}\left[\frac{|Q \cap r|}{|Q|} \leq \tau\right]$ be the probability that at most a τ fraction of P is in r. Now $T \subset P$ is an *(ε, α)-RQ coreset* of $(P, \mathcal{A})$ if for all $r \in \mathcal{A}$ and $\tau \in [0, 1]$ there exists a $\gamma \in [\tau - \alpha, \tau + \alpha]$ such that $|F_{P,r}(\tau) - F_{T,r}(\gamma)| \leq \varepsilon$. In such a situation, we also say that $F_{T,r}$ is an (ε, α)-*quantization* of $F_{P,r}$.

A natural question is whether we can construct a $(\varepsilon, 0)$-RQ coreset where there is not a secondary α-error term on τ. We demonstrate that there are no useful non-trivial bounds on the size of such a coreset.

When the (ε, α)-quantization $F_{T,r}$ need not be explicitly represented by a coreset T, then Löffler and Phillips [32, 26] show a different small space representation that can replace it in the above definition of an (ε, α)-RQ coreset with probability at least $1 - \delta$. First randomly create $m = O((1/\varepsilon^2)\log(1/\delta))$ transversals $Q_1, Q_2, \ldots, Q_m$, and for each transversal Q_i create an α-sample S_i of $(Q_i, \mathcal{A})$. Then to satisfy the requirements of $F_{T,r}(\tau)$, there exists some $\gamma \in [\tau - \alpha, \tau + \alpha]$ such that we can return $(1/m)|\{S_i \mid r(S_i) \leq \gamma\}|$, and it will be within ε of $F_{P,r}(\tau)$. However, this is subverting the attempt to construct and understand a coreset to answer these questions. A coreset T (our goal) can be used as proxy for P as opposed to querying m distinct point sets. This alternate approach also does not shed light into how much information can be captured by a small size point set, which is provided by bounds on the size of a coreset.

Simple example.

We illustrate a simple example with $k = 2$ and $d = 1$, where $n = 10$ and the $nk = 20$ possible locations of the 10 uncertain points are laid out in order:

$$p_{1,1} < p_{2,1} < p_{3,1} < p_{4,1} < p_{5,1} < p_{6,1} < p_{3,2} < p_{7,1} < p_{8,1} < p_{8,2} < p_{9,1} < p_{10,1} < p_{5,2} < p_{10,2} < p_{2,2} < p_{9,2} < p_{7,2} < p_{4,2} < p_{6,2} < p_{1,2}.$$

We consider a coreset $T \subset P$ that consists of the uncertain points $T = \{p_1, p_3, p_5, p_7, p_9\}$. Now consider a specific range $r \in \mathcal{I}_+$, a one-sided interval that contains $p_{5,2}$ and smaller points, but not $p_{10,2}$ and larger points. We can now see that $F_{T,r}$ is an $(\varepsilon' = 0.1016, \alpha = 0.1)$-quantization of $F_{P,r}$ in Figure 1; this follows since at $F_{P,r}(0.75) = 0.7734$ either $F_{T,r}(x)$ is at most 0.5 for $x \in [0.65, 0.8)$ and is at least 0.875

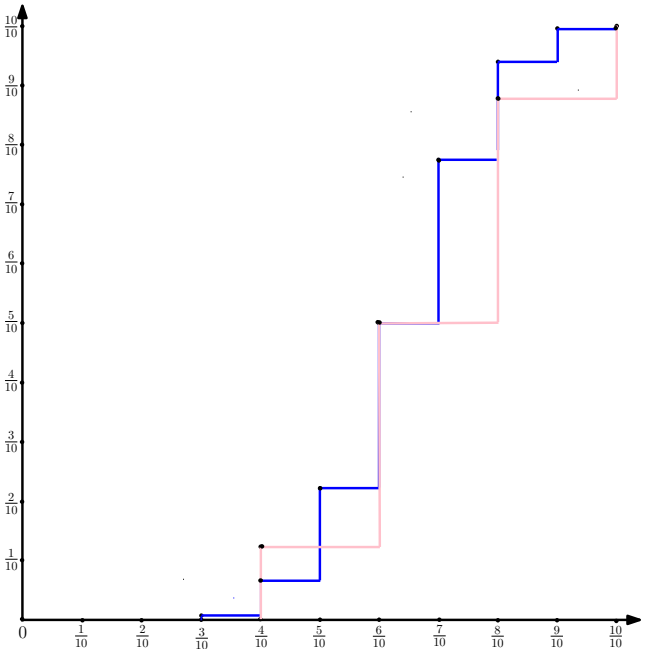

Figure 1: Example cumulative density functions ($F_{T,r}$, in red with fewer steps, and $F_{P,r}$, in blue with more steps) on uncertain point set P and a coreset T for a specific range.

for $x \in [0.8, 0.85]$. Also observe that

$$\left|E_{r(P)} - E_{r(T)}\right| = \left|\frac{13}{20} - \frac{7}{10}\right| = \frac{1}{20} = \varepsilon.$$

When these errors (the (ε', α)-quantization and ε-error) hold for *all* ranges in some range space, then T is an (ε', α)-RQ coreset or ε-RC coreset, respectively.

To understand the error associated with an RC coreset, also consider the threshold $\tau = 2/3$ with respect to the range r. Then in range r, $2/10$ of the uncertain points from P are in r with probability at least $\tau = 2/3$ (points p_3 and p_8). Also $1/5$ of the uncertain points from T are in r with probability at least $\tau = 2/3$ (only point p_3). So there is 0 RC error for this range and threshold.

1.2 Our Results

We provide the first results for RE-, RC-, and RQ-coresets with guarantees. In particular we show that a random sample T of size $O((1/\varepsilon^2)(\nu + \log(1/\delta))$ with probability $1-\delta$ is an ε-RC coreset for any family of ranges $\mathcal{A}$ whose associated range space has VC-dimension ν. Otherwise we enforce that each uncertain point has k possible locations, then a sample T of size $O((1/\varepsilon^2)(\nu + \log(k/\delta))$ suffices for an ε-RE coreset.

Then we leverage discrepancy-based techniques [35, 12] to improve these bounds to $O((1/\varepsilon)\mathsf{poly}(k, \log(1/\varepsilon)))$, for some specific families of ranges $\mathcal{A}$. This is an important improvement since $1/\varepsilon$ can be quite large (say 100 or more), while k, interpreted as the number of readings of a data point, is small for many applications (say 5). In $\mathbb{R}^1$, for one-sided ranges we construct ε-RE and ε-RC coresets of size $O((\sqrt{k}/\varepsilon)\log(k/\varepsilon))$. For axis-aligned rectangles in $\mathbb{R}^d$ we construct ε-RE coresets of size $O((\sqrt{k}/\varepsilon)\log^{\frac{3d-1}{2}}(k/\varepsilon))$ and ε-RC coresets of size $O((k^{3d+\frac{1}{2}}/\varepsilon)\log^{6d-\frac{1}{2}}(k/\varepsilon)))$. Finally, we show that any ε-RE coreset of size t is also an $(\varepsilon, \alpha_{\varepsilon,t})$-RQ coreset with value $\alpha_{\varepsilon,t} = \varepsilon + \sqrt{(1/2t)\ln(2/\varepsilon)}$.

These results leverage new connections between uncertain points and both discrepancy of permutations and colored range searching that may be of independent interest.

2. DISCREPANCY AND PERMUTATIONS

The key tools we will use to construct small coresets for uncertain data is discrepancy of range spaces, and specifically those defined on permutations. Consider a set X, a range space $(X, \mathcal{A})$, and a coloring $\chi : X \to \{-1, +1\}$. Then for some range $A \in \mathcal{A}$, the discrepancy is defined $\mathsf{disc}_\chi(X, A) = |\sum_{x \in X \cap A} \chi(x)|$. We can then extend this to be over all ranges $\mathsf{disc}_\chi(X, \mathcal{A}) = \max_{A \in \mathcal{A}} \mathsf{disc}_\chi(X, A)$ and over all colorings $\mathsf{disc}(X, \mathcal{A}) = \min_\chi \mathsf{disc}_\chi(X, \mathcal{A})$.

Consider a ground set (P, Σ_k) where P is a set of n objects, and $\Sigma_k = \{\sigma_1, \sigma_2, \ldots, \sigma_k\}$ is a set of k permutations over P so each $\sigma_j : P \to [n]$. We can also consider a family of ranges $\mathcal{I}_k$ as a set of intervals defined on *one* of the k permutations so $I_{x,y,j} \in \mathcal{I}_k$ is defined so $P \cap I_{x,y,j} = \{p \in P \mid x < \sigma_j(p) \leq y\}$ for $x < y \in [0, n]$ and $j \in [k]$. The pair $((P, \Sigma_k), \mathcal{I}_k)$ is then a range space, defining a set of subsets of P.

A canonical way to obtain k permutations from an uncertain point set $P = \{p_1, p_2, \ldots, p_n\}$ is as follows. Define the *jth canonical traversal* of P as the set $P_j = \cup_{i=1}^n p_{i,j}$. When each $p_{i,j} \in \mathbb{R}^1$, the sorted order of each canonical traversal P_j defines a permutation on P as $\sigma_j(p_i) = |\{p_{i',j} \in P_j \mid p_{i',j} \leq p_{i,j}\}|$, that is $\sigma_j(p_i)$ describes how many locations (including $p_{i,j}$) in the traversal P_j have value less than or equal to $p_{i,j}$. In other words, σ_j describes the sorted order of the jth point among all uncertain points. Then, given an uncertain point set, let the canonical traversals define the *canonical k-permutation* as (P, Σ_k).

A geometric view of the permutation range space embeds P as n fixed points in $\mathbb{R}^k$ and considers ranges which are defined by inclusion in $(k-1)$-dimensional slabs, defined by two parallel half spaces with normals aligned along one of the coordinate axes. Specifically, the jth coordinate of the ith point is $\sigma_j(p_i)$, and if the range is on the jth permutation, then the slab is orthogonal to the jth coordinate axis.

Another useful construction from an uncertain point set P is the set P_{cert} of all locations any point in P might occur. Specifically, for every uncertain point set P we can define the corresponding certain point set $P_{\mathsf{cert}} = \bigcup_{i \in [n]} p_i = \bigcup_{j \in [k]} P_j = \bigcup_{i \in [n], j \in [k]} p_{i,j}$. We can also extend any coloring χ on P to a coloring in P_{cert} by letting $\chi_{\mathsf{cert}}(p_{ij}) = \chi(p_i)$, for $i \in [n]$ and $j \in [k]$. Now we can naturally define the discrepancy induced on P_{cert} by any coloring χ of P as $\mathsf{disc}_{\chi_{\mathsf{cert}}}(P_{\mathsf{cert}}, \mathcal{A}) = \max_{r \in \mathcal{A}} \sum_{p_{i,j} \in P_{\mathsf{cert}} \cap r} \chi(p_{i,j})$.

From low-discrepancy to ε-samples.

There is a well-studied relationship between range spaces that admit low-discrepancy colorings, and creating ε-samples of those range spaces [13, 35, 12, 8]. The key relationship is if $\mathsf{disc}(X, \mathcal{A}) = \gamma \log^\omega(n)$, then there exists an ε-sample of $(P, \mathcal{A})$ of size $O((\gamma/\varepsilon)\log^\omega(\gamma/\varepsilon))$ [38], for values γ, ω independent of n or ε. Construct the coloring, and with equal probability discard either all points colored either -1 or those colored $+1$. This roughly halves the point set size, and also implies zero over-count in expectation for any fixed

range. Repeat this coloring and reduction of points until the desired size is achieved. This can be done efficiently in a distributed manner through a merge-reduce framework [13]; The take-away is that a method for a low-discrepancy coloring directly implies a method to create an ε-sample, where the counting error is in expectation zero for any fixed range. We describe and extend these results in much more detail in Appendix A.

3. RE CORESETS

First we will analyze ε-RE coresets through the P_{cert} interpretation of uncertain point set P. The canonical transversals P_j of P will also be useful. In Section 3.2 we will relate these results to a form of discrepancy.

Lemma 3.1. *$T \subset P$ is an ε-RE coreset for $(P, \mathcal{A})$ if and only if $T_{cert} \subset P_{cert}$ is an ε-sample for $(P_{cert}, \mathcal{A})$.*

Proof. First note that since $\mathbf{Pr}[p_i = p_{ij}] = \frac{1}{k}\ \forall i, j$, hence by linearity of expectations we have that $\mathbf{E}_{Q \in P}[|Q \cap r|] = \sum_{i=1}^{n} E[|p_i \cap r|] = \frac{1}{k}|P_{\mathsf{cert}} \cap r|$. Now, direct computation gives us:

$$\left|\frac{|P_{\mathsf{cert}} \cap r|}{|P_{\mathsf{cert}}|} - \frac{|T_{\mathsf{cert}} \cap r|}{|T_{\mathsf{cert}}|}\right| = \left|\frac{|P_{\mathsf{cert}} \cap r|}{k|P|} - \frac{|T_{\mathsf{cert}} \cap r|}{k|T|}\right| = \left|E_{r(P)} - E_{r(T)}\right| < \varepsilon. \qquad \square$$

The next implication enables us to determine an ε-RE coreset on P from ε-samples on each $P_j \Subset P$. Recall P_j is the jth canonical transversal of P for $j \in [k]$, and is defined similarly for a subset $T \subset P$ as T_j.

Lemma 3.2. *Given a range space $(P_{\mathsf{cert}}, \mathcal{A})$, if we have $T \subset P$ such that T_j is an ε-sample for $(P_j, \mathcal{A})$ for all $j \in [k]$, then T is an ε-RE coreset for $(P, \mathcal{A})$.*

Proof. Consider an arbitrary range $r \in \mathcal{R}$, and compute directly $\left|E_{r(P)} - E_{r(T)}\right|$. Recalling that $E_{r(P)} = \frac{|P_{\mathsf{cert}} \cap r|}{|P_{\mathsf{cert}}|}$ and observing that $|P_{\mathsf{cert}}| = k|P|$, we get that:

$$\left|E_{r(P)} - E_{r(T)}\right| = \left|\frac{\sum_{j=1}^{k} |P_j \cap r|}{k|P|} - \frac{\sum_{j=1}^{k} |T_j \cap r|}{k|T|}\right| \leq \frac{1}{k}\sum_{j=1}^{k}\left|\frac{|P_j \cap r|}{|P|} - \frac{|T_j \cap r|}{|T|}\right| \leq \frac{1}{k}(k\varepsilon) = \varepsilon. \ \square$$

3.1 Random Sampling

We show that a simple random sampling gives us an ε-RE coreset of P.

Theorem 3.1. *For an uncertain points set P and range space $(P_{\mathsf{cert}}, \mathcal{A})$ with VC-dimension ν, a random sample $T \subset P$ of size $O((1/\varepsilon^2)(\nu + \log(k/\delta)))$ is an ε-RE coreset of $(P, \mathcal{I})$ with probability at least $1 - \delta$.*

Proof. A random sample T_j of size $O((1/\varepsilon^2)(\nu + \log(1/\delta')))$ is an ε-sample of any $(P_j, \mathcal{A})$ with probability at least $1 - \delta'$ [31]. Now assuming $T \subset P$ resulted from a random sample on P, it induces the k disjoint canonical transversals T_j on T, such that $T_j \subset P_j$ and $|T_j| = O((1/\varepsilon^2)(\nu + \log(1/\delta')))$ for $j \in [k]$. Each T_j is an ε-sample of $(P_j, \mathcal{A})$ for any single $j \in [k]$ with probability at least $1 - \delta'$. Following Lemma 3.2 and using union bound, we conclude that $T \subset P$ is an ε-RE coreset for uncertain point set P with probability at least $1 - k\delta'$. Setting $\delta' = \delta/k$ proves the theorem. $\square$

3.2 RE-Discrepancy and its Properties

Next we extend the well-studied relationship between geometric discrepancy and ε-samples on certain data towards ε-RE coresets on uncertain data.

We first require precise and slightly non-standard definitions.

We introduce a new type of discrepancy based on the expected value of uncertain points called *RE-discrepancy*. Let P_χ^+ and P_χ^- denote the sets of uncertain points from P colored $+1$ or -1, respectively, by χ. Then $\mathsf{RE\text{-}disc}_\chi(P, r) = |P| \cdot |E_{r(P_\chi^+)} - E_{r(P)}|$ for any $r \in \mathcal{A}$. The usual extensions then follow: $\mathsf{RE\text{-}disc}_\chi(P, \mathcal{A}) = \max_{r \in \mathcal{A}} \mathsf{RE\text{-}disc}(P, r)$ and $\mathsf{RE\text{-}disc}(P, \mathcal{A}) = \min_\chi \mathsf{RE\text{-}disc}_\chi(P, \mathcal{A})$. Note that $(P, \mathcal{A})$ is technically not a range space, since $\mathcal{A}$ defines subsets of P_{cert} in this case, not of P.

Lemma 3.3. *Consider a coloring $\chi : P \to \{-1, +1\}$ such that $\mathsf{RE\text{-}disc}_\chi(P, \mathcal{A}) = \gamma \log^\omega(n)$ and $|P_\chi^+| = n/2$. Then the set P_χ^+ is an ε-RE coreset of $(P, \mathcal{A})$ with $\varepsilon = \frac{\gamma}{n}\log(n)$.*

Furthermore, if a subset $T \subset P$ has size $n/2$ and is an $(\frac{\gamma}{n}\log^\omega(n))$-RE coreset, then it defines a coloring χ (where $\chi(p_i) = +1$ for $p_i \in T$) that has $\mathsf{RE\text{-}disc}_\chi(P, \mathcal{A}) = \gamma \log^\omega(n)$.

Proof. We prove the second statement, the first follows symmetrically. We refer to the subset T as P_χ^+. Let $r = \arg\max_{r' \in \mathcal{A}} |E_{r'(P)} - E_{r'(P_\chi^+)}|$. This implies $\frac{\gamma}{n}\log^\omega n \geq |E_{r(P)} - E_{r(P_\chi^+)}| = \frac{1}{n}\mathsf{RE\text{-}disc}_\chi(P, r)$. $\square$

We can now recast RE-discrepancy to discrepancy on P_{cert}. From Lemma 3.1 $\left|\frac{|P_{\mathsf{cert}} \cap r|}{k|P|} - \frac{|T_{\mathsf{cert}} \cap r|}{k|T|}\right| = \left|E_{r(P)} - E_{r(T)}\right|$ and after some basic substitutions we obtain the following.

Lemma 3.4. *$\mathsf{RE\text{-}disc}_\chi(P, \mathcal{A}) = \frac{1}{k}\mathsf{disc}_{\chi_{cert}}(P_{cert}, \mathcal{A})$.*

This does not immediately solve ε-RE coresets by standard discrepancy techniques on P_{cert} because we need to find a coloring χ on P. A coloring χ_{cert} on P_{cert} may not be consistent across all $p_{i,j} \in p_i$. The following lemma allows us to reduce this to a problem of coloring each canonical transversal P_j.

Lemma 3.5. *$\mathsf{RE\text{-}disc}_\chi(P, \mathcal{A}) \leq \max_j \mathsf{disc}_{\chi_{cert}}(P_j, \mathcal{A})$.*

Proof. For any $r \in \mathcal{A}$ and any coloring χ (and the corresponding χ_{cert}), we can write P as a union of disjoint transversals P_j to obtain

$$\mathsf{disc}_{\chi_{\mathsf{cert}}}(P_{\mathsf{cert}}, r) = \left|\sum_{j=1}^{k}\sum_{p_{ij} \in P_j \cap r} \chi_{\mathsf{cert}}(p_{ij})\right| \leq \sum_{j=1}^{k}\left|\sum_{p_{ij} \in P_j \cap r} \chi_{\mathsf{cert}}(p_{ij})\right| \leq \sum_{j=1}^{k} \mathsf{disc}_{\chi_{\mathsf{cert}}}(P_j, r) \leq k \max_j \mathsf{disc}_{\chi_{\mathsf{cert}}}(P_j, r).$$

Since this holds for every $r \in \mathcal{A}$, hence (using Lemma 3.4)

$$\mathsf{RE\text{-}disc}_\chi(P, \mathcal{A}) = \frac{1}{k}\mathsf{disc}_{\chi_{\mathsf{cert}}}(P_{\mathsf{cert}}, \mathcal{A}) \leq \max_j \mathsf{disc}_{\chi_{\mathsf{cert}}}(P_j, \mathcal{A}).$$

$\square$

3.3 ε-RE Coresets in $\mathbb{R}^1$

Lemma 3.6. *Consider uncertain point set P with $P_{cert} \subset \mathbb{R}^1$ and the range space $(P_{cert}, \mathcal{I}_+)$ with ranges defined by one-sided intervals of the form $(-\infty, x]$, then $\mathsf{RE\text{-}disc}(P, \mathcal{I}) = O(\sqrt{k}\log n)$.*

Proof. Spencer et. al. [41] show that $\mathsf{disc}((P,\Sigma_k),\mathcal{I}_k)$ is $O(\sqrt{k}\log n)$. Since we obtain the Σ_k from the canonical transversals P_1 through P_k, by definition this results in upper bounds on the the discrepancy over all P_j (it bounds the max). Lemma 3.5 then gives us the bound on $\mathsf{RE\text{-}disc}(P,\mathcal{I})$. □

As we discussed in Appendix A the low RE-discrepancy coloring can be iterated in a merge-reduce framework as developed by Chazelle and Matousek [13]. With Theorem A.2 we can prove the following theorem.

Theorem 3.2. *Consider uncertain point set P and range space $(P_{\mathsf{cert}},\mathcal{I}_+)$ with ranges defined by one-sided intervals of the form $(-\infty,x]$, then an ε-RE coreset can be constructed of size $O((\sqrt{k}/\varepsilon)\log(k/\varepsilon))$.*

Since expected value is linear, $\mathsf{RE\text{-}disc}_\chi(P,(-\infty,x]) - \mathsf{RE\text{-}disc}_\chi(P,(-\infty,y)) = \mathsf{RE\text{-}disc}_\chi(P,[y,x])$ for $y < x$ and the above result also holds for the family of two-sided ranges $\mathcal{I}$.

3.4 ε-RE Coresets for Rectangles in $\mathbb{R}^d$

Here let P be a set of n uncertain points where each possible location of a point $p_{i,j} \in \mathbb{R}^d$. We consider a range space $(P_{\mathsf{cert}},\mathcal{R}_d)$ defined by d-dimensional axis-aligned rectangles.

Each canonical transversal P_j for $j \in [k]$ no longer implies a unique permutation on the points (for $d > 1$). But, for any rectangle $r \in \mathcal{R}$, we can represent any $r \cap P_j$ as the disjoint union of points P_j contained in intervals on a predefined set of $(1+\log n)^{d-1}$ permutations [9]. Spencer *et al.* [41] showed there exists a coloring χ such that

$$\max_j \mathsf{disc}_\chi(P_j,\mathcal{R}) = O(D_\ell(n)\log^{d-1} n),$$

where $\ell = (1+\log n)^{d-1}$ is the number of defined permutations and $D_\ell(n)$ is the discrepancy of ℓ permutations over n points and ranges defined as intervals on each permutation. Furthermore, they showed $D_\ell(n) = O(\sqrt{\ell}\log n)$.

To get the RE-discrepancy bound for $P_{\mathsf{cert}} = \cup_{j=1}^k P_j$, we first decompose P_{cert} into the k point sets P_j of size n. We then obtain $(1+\log n)^{d-1}$ permutations over points in each P_j, and hence obtain a family Σ_ℓ of $\ell = k(1+\log n)^{d-1}$ permutations over all P_j. $D_\ell(n) = O(\sqrt{\ell}\log n)$ yields

$$\mathsf{disc}((P,\Sigma_\ell),\mathcal{I}_\ell) = O(\sqrt{k}\log^{\frac{d+1}{2}} n).$$

Now each set $P_j \cap r$ for $r \in \mathcal{R}_d$, can be written as the disjoint union of $O(\log^{d-1} n)$ intervals of Σ_ℓ. Summing up over each interval, we get that $\mathsf{disc}(P_j,\mathcal{R}) = O(\sqrt{k}\log^{\frac{3d-1}{2}} n)$ for each j. By Lemma 3.5 this bounds the RE-discrepancy as well. Finally, we can again apply the merge-reduce framework of Chazelle and Matousek [13] (via Theorem A.2) to achieve an ε-RE coreset.

Theorem 3.3. *Consider uncertain point set P and range space $(P_{\mathsf{cert}},\mathcal{R}_d)$ (for $d > 1$) with ranges defined by axis-aligned rectangles in $\mathbb{R}^d$. Then an ε-RE coreset can be constructed of size $O((\sqrt{k}/\varepsilon)\log^{\frac{3d-1}{2}}(k/\varepsilon))$.*

4. RC CORESETS

Recall that an ε-RC coreset T of a set P of n uncertain points satisfies that for all queries $r \in \mathcal{A}$ and all thresholds $\tau \in [0,1]$ we have $|G_{P,r}(\tau) - G_{T,r}(\tau)| \leq \varepsilon$, where $G_{P,r}(\tau)$ represents the fraction of points from P that are in range r with probability at least τ.

In this setting, given a range $r \in \mathcal{A}$ and a threshold $\tau \in [0,1]$ we can let the pair $(r,\tau) \in \mathcal{A} \times [0,1]$ define a range $R_{r,\tau}$ such that each $p_i \in P$ is either in or not in $R_{r,\tau}$. Let $(P,\mathcal{A}\times[0,1])$ denote this range space. If $(P_{\mathsf{cert}},\mathcal{A})$ has VC-dimension ν, then $(P,\mathcal{A}\times[0,1])$ has VC-dimension $O(\nu+1)$; see Corollary 5.23 in [21]. This implies that random sampling works to construct ε-RC coresets.

Theorem 4.1. *For uncertain point set P and range space $(P_{\mathsf{cert}},\mathcal{A})$ with VC-dimension ν, a random sample $T \subset P$ of size $O((1/\varepsilon^2)(\nu+\log(1/\delta)))$ is an ε-RC coreset of $(P,\mathcal{A})$ with probability at least $1-\delta$.*

Yang *et al.* propose a similar result [46] as above, without proof.

4.1 RC Coresets in $\mathbb{R}^1$

Constructing ε-RC coresets when the family of ranges $\mathcal{I}_+$ represents one-sided, one-dimensional intervals is much easier than other cases. It relies heavily on the ordered structure of the canonical permutations, and thus discrepancy results do not need to decompose and then re-compose the ranges.

Lemma 4.1. *A point $p_i \in P$ is in range $r \in \mathcal{I}_+$ with probability at least $\tau = t/k$ if and only if $p_{i,t} \in r \cap P_t$.*

Proof. By the canonical permutations, since for all $i \in [n]$, we require $p_{i,j} < p_{i,j+1}$, then if $p_{i,t} \in r$, it follows that $p_{i,j} \in r$ for $j \leq t$. Similarly if $p_{i,t} \notin r$, then all $p_{i,j} \notin r$ for $j \geq t$. □

Thus when each canonical permutation is represented upto an error ε by a coreset T, then each threshold τ is represented within ε. Hence, as with ε-RE coresets, we invoke the low-discrepancy coloring of Bohus [9] and Spencer *et al.* [41], and then iterate them (invoking Theorem A.1) to achieve a small size ε-RC coreset.

Theorem 4.2. *For uncertain point set P and range space $(P_{\mathsf{cert}},\mathcal{I}_+)$ with ranges defined by one-sided intervals of the form $(-\infty,a]$. An ε-RC coreset of $(P,\mathcal{I}_+)$ can be constructed of size $O((\sqrt{k}/\varepsilon)\log(k/\varepsilon))$.*

Extending Lemma 4.1 from one-sided intervals of the form $[-\infty,a] \in \mathcal{I}_+$ to intervals of the form $[a,b] \in \mathcal{I}$ turns out to be non-trivial. It is *not* true that $G_{P,[a,b]}(\tau) = G_{P,[-\infty,b]}(\tau) - G_{P,[-\infty,a]}(\tau)$, hence the two queries cannot simply be subtracted. Also, while the set of points corresponding to the query $G_{P,[-\infty,a]}(\frac{t}{k})$ are a contiguous interval in the tth permutation we construct in Lemma 4.1, the same need not be true of points corresponding to $G_{P,[a,b]}(\frac{t}{k})$. This is a similar difficulty in spirit as noted by Kaplan *et al.* [29] in the problem of counting the number of points of distinct colors in a box where one cannot take a naive decomposition and add up the numbers returned by each subproblem.

We give now a construction to solve this two-sided problem for uncertain points in $\mathbb{R}^1$ inspired by that of Kaplan *et al.* [29], but we require specifying a fixed value of $t \in [k]$. Given an uncertain point $p_i \in P$ assume w.l.o.g that $p_{i,j} < p_{i,j+1}$. Also pretend there is a point $p_{i,k+1} = \eta$ where η is larger than any $b \in \mathbb{R}^1$ from a query range $[a,b]$ (essentially $\eta = \infty$). Given a range $[a,b]$, we consider the right-most set

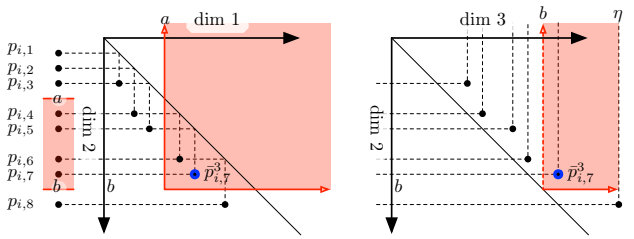

Figure 2: Uncertain point p_i queried by range $[a, b]$. Lifting shown to $\bar{p}_i^3$ along dimensions 1 and 2 (left) and along dimensions 2 and 3 (right).

of t locations of p_i (here $\{p_{i,j-t}, \ldots, p_{i,j}\}$) that are in the range. This satisfies (i) $p_{i,j-t} \geq a$, (ii) $p_{i,j} \leq b$, and (iii) to ensure that it is the right-most such set, $p_{i,j+1} > b$.

To satisfy these three constraints we re-pose the problem in $\mathbb{R}^3$ to designate each contiguous set of t possible locations of p_i as a single point. So for $t < j \leq k$, we map $p_{i,j}$ to $\bar{p}_{i,j}^t = (p_{i,j-t}, p_{i,j}, p_{i,j+1})$. Correspondingly, a range $r = [a, b]$ is mapped to a range $\bar{r} = [a, \infty) \times (-\infty, b] \times (b, \infty)$; see Figure 2. Let $\bar{p}_i^t$ denote the set of all $\bar{p}_{i,j}^t$, and let $\bar{P}^t$ represent $\bigcup_i \bar{p}_i^t$.

Lemma 4.2. *p_i is in interval $r = [a, b]$ with threshold at least t/k if and only if $\bar{p}_i^t \cap \bar{r}^t \geq 1$. Furthermore, no two points $p_{i,j}, p_{i,j'} \in p_i$ can map to points $\bar{p}_{i,j}^t, \bar{p}_{i,j'}^t$ such that both are in a range $\bar{r}^t$.*

Proof. Since $p_{i,j} < p_{i,j+1}$, then if $p_{i,j-t} \geq a$ it implies all $p_{i,\ell} \geq a$ for $\ell \geq j-t$, and similarly, if $p_{i,j} \leq b$ then all $p_{i,\ell} \leq b$ for all $\ell \leq j$. Hence if $\bar{p}_{i,j}^t$ satisfies the first two dimensional constraints of the range $\bar{r}^t$, it implies t points $p_{i,j-t} \ldots, p_{i,j}$ are in the range $[a, b]$. Satisfying the constraint of $\bar{r}^t$ in the third coordinate indicates that $p_{i,j+1} \notin [a, b]$. There can only be one point $p_{i,j}$ which satisfies the constraint of the last two coordinates that $p_{i,j} \leq b < p_{i,j+1}$. And for any range which contains at least t possible locations, there must be at least one such set (and only one) of t consecutive points which has this satisfying $p_{i,j}$. □

Corollary 4.1. *Any uncertain point set $P \in \mathbb{R}^1$ of size n and range $r = [a, b]$ has $G_{P,r}(\frac{t}{k}) = |\bar{P}^t \cap \bar{r}^t|/n$.*

This presents an alternative view of each uncertain point in $\mathbb{R}^1$ with k possible locations as an uncertain point in $\mathbb{R}^3$ with $k-t$ possible locations (since for now we only consider a threshold $\tau = t/k$). Where $\mathcal{I}$ represents the family of ranges defined by two-sided intervals, let $\bar{\mathcal{I}}$ be the corresponding family of ranges in $\mathbb{R}^3$ of the form $[a, \infty) \times (-\infty, b] \times (b, \infty)$ corresponding to an interval $[a, b] \in \mathcal{I}$. Under the assumption (valid under the lifting defined above) that each uncertain point can have at most one location fall in each range, we can now decompose the ranges and count the number of points that fall in each sub-range and add them together. Using the techniques (described in detail in Section 3.4) of Bohus [9] and Spencer *et al.* [41] we can consider $\ell = (k-t)(1+\lceil \log n \rceil)^2$ permutations of $\bar{P}_{\mathsf{cert}}^t$ such that each range $\bar{r} \in \bar{\mathcal{I}}$ can be written as the points in a disjoint union of intervals from these permutations. To extend low discrepancy to *each* of the k distinct values of threshold t, there are k such liftings and $h = k \cdot \ell = O(k^2 \log^2 n)$ such permutations we need to consider. We can construct a coloring

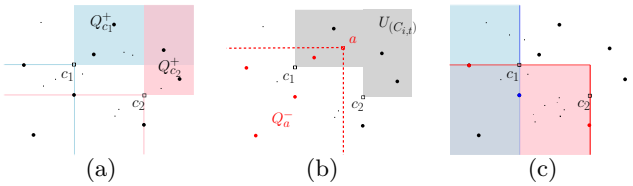

Figure 3: Illustration of uncertain point $p_i \in \mathbb{R}^2$ with $k = 8$ and $t = 3$. (a) All tight negative orthants containing exactly $t = 3$ locations of p_i, their apexes are $C_{i,t} = \{c_1, c_2\}$. (b): $U(C_{i,t})$ is shaded and query Q_a^-. (c): $M(C_{i,t})$, the maximal negative orthants of $C_{i,t}$ that are also bounded in the x-direction.

$\chi : P \to \{-1, +1\}$ such that intervals on each permutation has discrepancy $O(\sqrt{h} \log n) = O(k \log^2 n)$. Recall that for any fixed threshold t we only need to consider the corresponding ℓ permutations, hence the total discrepancy for any such range is at most the sum of discrepancy from all corresponding $\ell = O(k \log^2 n)$ permutations or $O(k^2 \log^4 n)$. Finally, this low-discrepancy coloring can be iterated (via Theorem A.1) to achieve the following theorem.

Theorem 4.3. *Consider an uncertain point set P along with ranges $\mathcal{I}$ of two-sided intervals. We can construct an ε-RC coreset T for $(P, \mathcal{I})$ of size $O((k^2/\varepsilon) \log^4(k/\varepsilon))$.*

4.2 RC Coresets for Rectangles in $\mathbb{R}^d$

The approach for $\mathcal{I}$ can be further extended to $\mathcal{R}_d$, axis-aligned rectangles in $\mathbb{R}^d$. Again the key idea is to define a proxy point set $\bar{P}$ such that $|\bar{r} \cap \bar{P}|$ equals the number of uncertain points in r with at least threshold t. This requires a suitable lifting map and decomposition of space to prevent over or under counting; we employ techniques from Kaplan *et al.* [29].

First we transform queries on axis-aligned rectangles in $\mathbb{R}^d$ to the semi-bounded case in $\mathbb{R}^{2d}$. Denote the x_i-coordinate of a point q as $x_i(q)$, we double all the coordinates of each point $q = (x_1(q), ..., x_\ell(q), ..., x_d(q))$ to obtain point

$$\tilde{q} = (-x_1(q), x_1(q) ..., -x_\ell(q), x_\ell(q), ..., -x_d(q), x_d(q))$$

in $\mathbb{R}^{2d}$. Now answering range counting query $\prod_{i=1}^d [a_i, b_i]$ is equivalent to solving the query $\prod_{i=1}^d [(-\infty, -a_i] \times (-\infty, b_i]]$ on the lifted point set.

Based on this reduction we can focus on queries of *negative orthants* of the form $\prod_{i=1}^d (-\infty, a_i]$ and represent each orthant by its apex $a = (a_1, ..., a_d) \in \mathbb{R}^d$ as Q_a^-. Similarly, we can define Q_a^+ as *positive orthants* in the form $\prod_{i=1}^d [a_i, \infty) \subseteq \mathbb{R}^d$. For any point set $A \subset \mathbb{R}^d$ define $U(A) = \cup_{a \in A} Q_a^+$.

A *tight* orthant has a location of $p_i \in P$ incident to every bounding facet. Let $C_{i,t}$ be the set of all apexes representing tight negative orthants that contain exactly t locations of p_i; see Figure 3(a). An important observation is that query orthant Q_a^- contains p_i with threshold at least t if and only if it contains at least one point from $C_{i,t}$.

Let $Q_{i,t}^+ = \cup_{c \in C_{i,t}} Q_c^+$ be the locus of all negative orthant query apexes that contain at least t locations of p_i; see Figure 3(b). Notice that $Q_{i,t}^+ = U(C_{i,t})$.

Lemma 4.3. *For any point set $p_i \subset \mathbb{R}^d$ of k points and some threshold $1 \leq t \leq k$, we can decompose $U(C_{i,t})$ into $f(k) = O(k^d)$ pairwise disjoint boxes, $B(C_{i,t})$.*

Proof. Let $M(A)$ be the set of maximal empty negative orthants for a point set A, such that any $m \in M(A)$ is

also bounded in the positive direction along the 1st coordinate axis. Kaplan *et al.* [29] show (within Lemma 3.1) that $|M(A)| = |B(A)|$ and provide a specific construction of the boxes B. Thus we only need to bound $|M(C_{i,t})|$ to complete the proof; see $M(C_{i,t})$ in Figure 3(c). We note that each coordinate of each $c \in C_{i,t}$ must be the same as some $p_{i,j} \in p_i$. Thus for each coordinate, among all $c \in C_{i,t}$ there are at most k values. And each maximal empty tight orthant $m \in M(C_{i,t})$ is uniquely defined by the d coordinates along the axis direction each facet is orthogonal to. Thus $|M(C_{i,t})| \leq k^d$, completing the proof. □

Note that as we are working in a lifted space $\mathbb{R}^{2d}$, this corresponds to $U(C_{i,t})$ being decomposed into $f(k) = O(k^{2d})$ pairwise *disjoint* boxes in which d is the dimensionality of our original point set.

Lemma 4.4. *For negative orthant queries Q_a^- with apex a on uncertain point set P, a point $p_i \in P$ is in Q_a^- with probability at least t/k if a is in some box in $B(C_{i,t})$, and a will lie in at most one box from $B(C_{i,t})$.*

Proof. The query orthant Q_a^- contains point p_i with threshold at least t if and only if Q_a^- contains at least one point from $C_{i,t}$ and this happens only when $a \in U(C_{i,t})$. Since the union of constructed boxes in $B(C_{i,t})$ is equivalent to $U(C_{i,t})$ and they are disjoint, the result follows. □

Corollary 4.2. *The number of uncertain points from P in query range Q_a^- with probability at least t/k is exactly the number of boxes in $\cup_i^n B(C_{i,t})$ that contain a.*

Thus for a set of boxes representing P, we need to perform count stabbing queries with apex a and show a low-discrepancy coloring of boxes.

We do a second lifting by transforming each point $a \in \mathbb{R}^d$ to a semi-bounded box $\bar{a} = \prod_{i=1}^d ((-\infty, a_i] \times [a_i, \infty))$ and each box $b \in \mathbb{R}^d$ of the form $\prod_i^d [x_i, y_i]$ to a point $\bar{b} = (x_1, y_1, ..., x_\ell, y_\ell, ..., x_d, y_d)$ in $\mathbb{R}^{2d}$. It is easy to verify that $a \in b$ if and only if $\bar{b} \in \bar{a}$.

Since this is our second doubling of dimension, we are now dealing with points in $\mathbb{R}^{4d}$. Lifting P to $\bar{P}$ in $\mathbb{R}^{4d}$ now presents an alternative view of each uncertain point $p_i \in P$ as an uncertain point $\bar{p_i}$ in $\mathbb{R}^{4d}$ with $g_k = O(k^{2d})$ possible locations with the query boxes represented as $\bar{\mathcal{R}}$ in $\mathbb{R}^{4d}$.

We now proceed similarly to the proof of Theorem 4.3. For a fixed threshold t, obtain $\ell = g_k \cdot (1 + \lceil \log n \rceil)^{4d-1}$ disjoint permutations of $\bar{P}^t_{\mathsf{cert}}$ such that each range $\bar{r} \in \bar{\mathcal{R}}$ can be written as the points in a disjoint union of intervals from these permutations. For the k distinct values of t, there are k such liftings and $h = O\left(k \cdot g_k \cdot \log^{4d-1} n\right)$ such permutations we need to consider, and we can construct a coloring $\chi : P \to \{-1, +1\}$ so that intervals on each permutation have discrepancy

$$O(\sqrt{h} \log n) = O(k^{d+\frac{1}{2}} \log^{\frac{4d+1}{2}} n).$$

Hence for any such range and specific threshold t, the total discrepancy is the sum of discrepancy from all corresponding $\ell = O\left(g_k \cdot \log^{4d-1} n\right)$ permutations, or $O\left(k^{3d+\frac{1}{2}} \log^{6d-\frac{1}{2}} n\right)$. By applying the iterated low-discrepancy coloring (Theorem A.1), we achieve the following result.

Theorem 4.4. *Consider an uncertain point set P and range space $(P_{\mathsf{cert}}, \mathcal{R}_d)$ with ranges defined by axis-aligned rectangles in $\mathbb{R}^d$. Then an ε-RC coreset can be constructed of size $O\left((k^{3d+\frac{1}{2}}/\varepsilon) \log^{6d-\frac{1}{2}}(k/\varepsilon)\right)$.*

5. RQ CORESETS

In this section, given an uncertain point set P and its ε-RE coreset T, we want to determine values ε' and α so T is an (ε', α)-RQ coreset. That is for any $r \in \mathcal{A}$ and threshold $\tau \in [0, 1]$ there exists a $\gamma \in [\tau - \alpha, \tau + \alpha]$ such that

$$\left| \mathbf{Pr}_{Q \in P}\left[\frac{|Q \cap r|}{|Q|} \leq \tau\right] - \mathbf{Pr}_{S \in T}\left[\frac{|S \cap r|}{|S|} \leq \gamma\right] \right| \leq \varepsilon'.$$

At a high level, our tack will be to realize that both $|Q \cap r|$ and $|S \cap r|$ behave like Binomial random variables. By T being an ε-RE coreset of P, then after normalizing, its mean is at most ε-far from that of P. Furthermore, Binomial random variables tend to concentrate around their mean– and more so for those with more trials. This allows us to say $|S \cap r|/|S|$ is either α-close to the expected value of $|Q \cap r|/|Q|$ or is ε'-close to 0 or 1. Since $|Q \cap r|/|Q|$ has the same behavior, but with more concentration, we can bound their distance by the α and ε' bounds noted before. We now work out the details.

Theorem 5.1. *If T is an ε-RE coreset of P for $\varepsilon \in (0, 1/2)$, then T is an (ε', α)-RQ coreset for P for $\varepsilon', \alpha \in (0, 1/2)$ and satisfying $\alpha \geq \varepsilon + \sqrt{(1/2|T|) \ln(2/\varepsilon')}$.*

Proof. We start by examining a Chernoff-Hoeffding bound on a set of independent random variables X_i so that each $X_i \in [a_i, b_i]$ with $\Delta_i = b_i - a_i$. Then for some parameter $\beta \in (0, \sum_i \Delta_i/2)$

$$\mathbf{Pr}\left[\left|\sum_i X_i - \mathbf{E}\left[\sum_i X_i\right]\right| \geq \beta\right] \leq 2\exp\left(\frac{-2\beta^2}{\sum_i \Delta_i^2}\right).$$

Consider any $r \in \mathcal{A}$. We now identify each random variable $X_i = \mathbb{1}(q_i \in r)$ (that is, 1 if $q_i \in r$ and 0 otherwise) where q_i is the random instantiation of some $p_i \in T$. So $X_i \in \{0, 1\}$ and $\Delta_i = 1$. Thus by equating $|S \cap r| = \sum X_i$

$$\begin{aligned}&\mathbf{Pr}_{S \in T}\left[\left||S \cap r| - \mathbf{E}\left[|S \cap r|\right]\right| \geq \beta|S|\right] \\ &\leq 2\exp\left(\frac{-2\beta^2|S|^2}{\sum_i \Delta_i^2}\right) \\ &= 2\exp(-2\beta^2|S|) \leq \varepsilon'.\end{aligned}$$

Thus by solving for β (and equating $|S| = |T|$)

$$\mathbf{Pr}_{S \in T}\left[\left|\frac{|S \cap r|}{|S|} - \mathbf{E}\left[\frac{|S \cap r|}{|S|}\right]\right| \geq \sqrt{\frac{1}{2|T|}\ln(\frac{2}{\varepsilon'})}\right] \leq \varepsilon'.$$

Now by T being an ε-RE coreset of P then

$$\left|\mathbf{E}_{S \in T}\left[\frac{|S \cap r|}{|S|}\right] - \mathbf{E}_{Q \in P}\left[\frac{|Q \cap r|}{|Q|}\right]\right| \leq \varepsilon.$$

Combining these two we have

$$\mathbf{Pr}_{S \in T}\left[\left|\frac{|S \cap r|}{|S|} - \mathbf{E}_{Q \in P}\left[\frac{|Q \cap r|}{|Q|}\right]\right| \geq \alpha\right] \leq \varepsilon'$$

for $\alpha = \varepsilon + \sqrt{\frac{1}{2|T|}\ln(\frac{2}{\varepsilon'})}$.

Combining these statements, for any $x \leq M - \alpha \leq M - \alpha'$ we have $\varepsilon' > F_{T,r}(x) \geq 0$ and $\varepsilon' > F_{P,r}(x) \geq 0$ (and symmetrically for $x \geq M + \alpha \geq M + \alpha'$). It follows that $F_{T,r}$ is an (ε', α)-quantization of $F_{P,r}$.

Since this holds for any $r \in \mathcal{A}$, by T being an ε-RE coreset of P, it follows that T is also an (ε', α)-RQ coreset of P. □

We can now combine this result with specific results for ε-RE coresets to get size bounds for (ε, α)-RQ coresets. To achieve the below bounds we set $\varepsilon = \varepsilon'$.

Corollary 5.1. *Consider uncertain point set P with range space $(P_{\mathsf{cert}}, \mathcal{A})$. There exists a $(\varepsilon, \varepsilon + \sqrt{(1/2|T|)\ln(2/\varepsilon)})$-RQ coreset of $(P, \mathcal{A})$ of size $|T| =$*

- $O((1/\varepsilon^2)(\nu + \log(k/\delta)))$ *when $\mathcal{A}$ has VC-dimension ν, with probability $1 - \delta$ (Theorem 3.1),*
- $O((\sqrt{k}/\varepsilon)\log(k/\varepsilon))$ *when $\mathcal{A} = \mathcal{I}$ (Theorem 3.2), and*
- $O\left((\sqrt{k}/\varepsilon)\log^{\frac{3d-1}{2}}(\frac{k}{\varepsilon})\right)$ *when $\mathcal{A} = \mathcal{R}_d$ (Theorem 3.3).*

Finally we discuss why the α term in the (ε', α)-RQ coreset T is needed. Recall from Section 3 that approximating the value of $\mathsf{E}_{Q \in P}\left[\frac{|Q \cap r|}{|Q|}\right]$ with $\mathsf{E}_{S \in T}\left[\frac{|S \cap r|}{|S|}\right]$ for all r corresponds to a low-discrepancy sample of P_{cert}. Discrepancy error immediately implies we will have at least the ε horizontal shift between the two distributions and their means, unless we could obtain a zero discrepancy sample of P_{cert}. Note this ε-horizontal error corresponds to the α term in an (ε', α)-RQ coreset. When P is very large, then due to the central limit theorem, $F_{P,r}$ will grow very sharply around $\mathsf{E}_{Q \in P}\left[\frac{|Q \cap r|}{|Q|}\right]$. In the worst case $F_{T,r}$ may be $\Omega(1)$ vertically away from $F_{P,r}$ on either side of $\mathsf{E}_{S \in T}\left[\frac{|S \cap r|}{|S|}\right]$, so no reasonable amount of ε' vertical tolerance will make up for this gap.

On the other hand, the ε' vertical component is necessary since for very small probability events (that is for a fixed range r and small threshold τ) on P, we may need a much smaller value of τ (smaller by $\Omega(1)$) to get the same probability on T, requiring a very large horizontal shift. But since it is a very small probability event, only a small vertical ε' shift is required.

The main result of this section then is showing that there exist pairs (ε', α) which are both small.

6. CONCLUSION AND OPEN QUESTIONS

This paper defines and provides the first results for coresets on uncertain data. These can be essential tools for monitoring a subset of a large noisy data set, as a way to approximately monitor the full uncertainty.

There are many future directions on this topic, in addition to tightening the provided bounds especially for other range spaces. Can we remove the dependence on k without random sampling? Can coresets be constructed over uncertain data for other queries such as minimum enclosing ball, clustering, and extents?

7. REFERENCES

[1] Agarwal, P. K., Cheng, S.-W., Tao, Y., and Yi, K. Indexing uncertain data. In *PODS* (2009).

[2] Agarwal, P. K., Har-Peled, S., and Varadarajan, K. Geometric approximations via coresets. *Current Trends in Combinatorial and Computational Geometry (E. Welzl, ed.)* (2007).

[3] Agarwal, P. K., Har-Peled, S., and Varadarajan, K. R. Approximating extent measure of points. *Journal of ACM 51*, 4 (2004), 2004.

[4] Agrawal, P., Benjelloun, O., Sarma, A. D., Hayworth, C., Nabar, S., Sugihara, T., and Widom, J. Trio: A system for data, uncertainty, and lineage. In *PODS* (2006).

[5] Anthony, M., and Bartlett, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[6] Bādoiu, M., Har-Peled, S., and Indyk, P. Approximate clustering via core-sets. In *STOC* (2002).

[7] Bandyopadhyay, D., and Snoeyink, J. Almost-Delaunay simplices: Nearest neighbor relations for imprecise points. In *SODA* (2004).

[8] Beck, J. Roth's estimate of the discrepancy of integer sequences is nearly sharp. *Combinatorica 1* (1981), 319–325.

[9] Bohus, G. On the discrepancy of 3 permutations. *Random Structures and Algorithms 1*, 2 (1990), 215–220.

[10] Bădoiu, M., and Clarkson, K. Smaller core-sets for balls. In *SODA* (2003).

[11] Burdick, D., Deshpande, P. M., Jayram, T., Ramakrishnan, R., and Vaithyanathan, S. OLAP over uncertain and imprecise data. In *VLDB* (2005).

[12] Chazelle, B. *The Discrepancy Method*. Cambridge, 2000.

[13] Chazelle, B., and Matousek, J. On linear-time deterministic algorithms for optimization problems in fixed dimensions. *Journal of Algorithms 21* (1996), 579–597.

[14] Chazelle, B., and Welzl, E. Quasi-optimal range searching in spaces of finite VC-dimension. *Discrete and Computational Geometry 4* (1989), 467–489.

[15] Cheng, R., Xia, Y., Prabhakar, S., Shah, R., and Vitter, J. S. Efficient indexing methods for probabilistic threshold queries over uncertain data. In *VLDB* (2004).

[16] Cormode, G., and Garafalakis, M. Histograms and wavelets of probabilitic data. In *ICDE* (2009).

[17] Cormode, G., Li, F., and Yi, K. Semantics of ranking queries for probabilistic data and expected ranks. In *ICDE* (2009).

[18] Dalvi, N., and Suciu, D. Efficient query evaluation on probabilistic databases. In *VLDB* (2004).

[19] Guibas, L. J., Salesin, D., and Stolfi, J. Epsilon geometry: building robust algorithms from imprecise computations. In *SoCG* (1989).

[20] Guibas, L. J., Salesin, D., and Stolfi, J. Constructing strongly convex approximate hulls with inaccurate primitives. *Algorithmica 9* (1993), 534–560.

[21] Har-Peled, S. *Geometric Approximation Algorithms*. American Mathematical Society, 2011.

[22] Held, M., and Mitchell, J. S. B. Triangulating input-constrained planar point sets. *Information Processing Letters 109*, 1 (2008).

[23] Jayram, T., Kale, S., and Vee, E. Efficient aggregation algorithms for probabilistic data. In *SODA* (2007).

[24] Jayram, T., McGregor, A., Muthukrishnan, S., and Vee, E. Estimating statistical aggregates on probabilistic data streams. In *PODS* (2007).

[25] Jørgensen, A. G., Löffler, M., and Phillips, J. M. Geometric computation on indecisive and uncertain points. arXiv:1205.0273.

[26] Jørgensen, A. G., Löffler, M., and Phillips, J. M. Geometric computation on indecisive points. In *WADS* (2011).

[27] Kamousi, P., Chan, T. M., and Suri, S. The stochastic closest pair problem and nearest neighbor search. In *WADS* (2011).

[28] Kamousi, P., Chan, T. M., and Suri, S. Stochastic minimum spanning trees in euclidean spaces. In *SOCG* (2011).

[29] Kaplan, H., Rubin, N., Sharir, M., and Verbin, E. Counting colors in boxes. In *SODA* (2007).

[30] Kruger, H. Basic measures for imprecise point sets in $\mathbb{R}^d$. Master's thesis, Utrecht University, 2008.

[31] Li, Y., Long, P. M., and Srinivasan, A. Improved bounds on the samples complexity of learning. *Journal of Computer and System Science 62* (2001), 516–527.

[32] Löffler, M., and Phillips, J. Shape fitting on point sets with probability distributions. In *ESA* (2009), Springer Berlin / Heidelberg.

[33] Löffler, M., and Snoeyink, J. Delaunay triangulations of imprecise points in linear time after preprocessing. In *SOCG* (2008).

[34] Matousek, J. Approximations and optimal geometric divide-and-conquer. *Journal of Computer and System Sciences 50*, 2 (1995), 203 – 208.

[35] Matoušek, J. *Geometric Discrepancy*. Springer, 1999.

[36] Nagai, T., and Tokura, N. Tight error bounds of geometric problems on convex objects with imprecise coordinates. In *Jap. Conf. on Discrete and Comput. Geom.* (2000), LNCS 2098, pp. 252–263.

[37] Ostrovsky-Berman, Y., and Joskowicz, L. Uncertainty envelopes. In *21st European Workshop on Comput. Geom.* (2005), pp. 175–178.

[38] Phillips, J. M. Algorithms for ε-approximations of terrains. In *ICALP* (2008).

[39] Phillips, J. M. *Small and Stable Descriptors of Distributions for Geometric Statistical Problems*. PhD thesis, Duke University, 2009.

[40] Sarma, A. D., Benjelloun, O., Halevy, A., Nabar, S., and Widom, J. Representing uncertain data: models, properties, and algorithms. *The VLDB Journal 18*, 5 (2009), 989–1019.

[41] Spencer, J., Srinivasan, A., and Tetai, P. Discrepancy of permutation families. Unpublished manuscript, 2001.

[42] Tao, Y., Cheng, R., Xiao, X., Ngai, W. K., Kao, B., and Prabhakar, S. Indexing multi-dimensional uncertain data with arbitrary probability density functions. In *VLDB* (2005).

[43] van der Merwe, R., Doucet, A., de Freitas, N., and Wan, E. The unscented particle filter. In *NIPS* (2000), vol. 8, pp. 351–357.

[44] van Kreveld, M., and Löffler, M. Largest bounding box, smallest diameter, and related problems on imprecise points. *Computational Geometry: Theory and Applications 43* (2010), 419–433.

[45] Vapnik, V., and Chervonenkis, A. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications 16* (1971), 264–280.

[46] Yang, S., Zhang, W., Zhang, Y., and Lin, X. Probabilistic threshold range aggregate query processing over uncertain data. *Advances in Data and Web Management* (2009), 51–62.

[47] Zhang, Y., Lin, X., Tao, Y., Zhang, W., and Wang, H. Efficient computation of range aggregates against uncertain location based queries. *IEEE Transactions on Knowledge and Data Engineering 24* (2012), 1244–1258.

APPENDIX

A. LOW DISCREPANCY TO ε-CORESET

Mainly in the 90s Chazelle and Matousek [14, 13, 34, 35, 12] led the development of method to convert from a low-discrepancy coloring to a coreset that allowed for approximate range queries. Here we summarize and generalize these results.

We start by restating a results of Phillips [38, 39] which generalizes these results, here we state it a bit more specifically for our setting.

Theorem A.1 (Phillips [38, 39])**.** *Consider a point set P of size n and a family of subsets $\mathcal{A}$. Assume an $O(n^\beta)$ time algorithm to construct a coloring $\chi : P \to \{-1, +1\}$ so $\mathsf{disc}_\chi(P, \mathcal{A}) = O(\gamma \log^\omega n)$ where β, γ, and ω are constant algorithm parameters dependent on $\mathcal{A}$, but not P (or n). There exists an algorithm to construct an ε-sample of $(P, \mathcal{A})$ of size $g(\varepsilon, \mathcal{A}) = O((\gamma/\varepsilon)\log^\omega(\gamma/\varepsilon))$ in time $O(n \cdot g(\varepsilon, \mathcal{A})^{\beta-1})$.*

Note that we ignored non-exponential dependence on ω and β since in our setting they are data and problem independent constants. But we are more careful with γ terms since they depend on k, the number of locations of each uncertain point.

We restate the algorithm and analysis here for completeness, using $g = g(\varepsilon, \mathcal{A})$ for shorthand. Divide P into n/g parts $\{\bar{P}_1, \bar{P}_2, \ldots, \bar{P}_{n/g}\}$ of size $k = 4(\beta + 2)g$. Assume this divides evenly and n/g is a power of two, otherwise pad P and adjust g by a constant. Until there is a single set, repeat the following two stages. In stage 1, for $\beta+2$ steps, pair up all remaining sets, and for all pairs (e.g. P_i and P_j) construct a low-discrepancy coloring χ on $P_i \cup P_j$ and discard all points colored -1 (or $+1$ at random). In the $(\beta + 3)$rd step pair up all sets, but do not construct a coloring and halve. That is every epoch ($\beta + 3$ steps) the size of remaining sets double, otherwise they remain the same size. When a single set remains, stage 2 begins; it performs the color-halve part of the above procedure until $\mathsf{disc}(P, \mathcal{A}) \leq \varepsilon n$ as desired.

We begin analyzing the error on a single coloring.

Lemma A.1. *The set $P^+ = \{p \in P \mid \chi(p) = +1\}$ is an $(\mathsf{disc}_\chi(P, \mathcal{A})/n)$-sample of $(P, \mathcal{A})$.*

Proof.

$$\max_{R\in\mathcal{A}}\left|\frac{|P\cap R|}{|P|}-\frac{|P^+\cap R|}{|P^+|}\right| = \max_{R\in\mathcal{A}}\left|\frac{|P\cap R|-2|P^+\cap R|}{n}\right| \leq \frac{\mathsf{disc}_\chi(P,\mathcal{A})}{n}. \qquad \square$$

We also note two simple facts [12, 35]:

(S1) If Q_1 is an ε-sample of P_1 and Q_2 is an ε-sample of P_2, then $Q_1 \cup Q_2$ is an ε-sample of $P_1 \cup P_2$.

(S2) If Q is an ε_1-sample of P and S is an ε_2 sample of Q, then S is an $(\varepsilon_1+\varepsilon_2)$-sample of P.

Note that (S1) (along with Lemma A.1) implies the arbitrarily decomposing P into n/g sets and constructing colorings of each achieves the same error bound as doing so on just one. And (S2) implies that chaining together rounds adds the error in each round. It follows that if we ignore the $(\beta+3)$rd step in each epoch, then there is 1 set remaining after $\log(n/g)$ steps. The error caused by each step is $\mathsf{disc}(g,\mathcal{A})/g$ so the total error is $\log(n/g)(\gamma\log^\omega g)/g = \varepsilon$. Solving for g yields $g = O(\frac{\gamma}{\varepsilon}\log(\frac{n\varepsilon}{\gamma})\log^\omega(\frac{\gamma}{\varepsilon}))$.

Thus to achieve the result stated in the theorem the $(\beta+3)$rd step skip of a reduce needs to remove the $\log(n\varepsilon/\gamma)$ term from the error. This works! After $\beta+3$ steps, the size of each set is $2g$ and the discrepancy error is $\gamma\log^\omega(2g)/2g$. This is just more than half of what it was before, so the total error is now:

$$\sum_{i=0}^{\frac{\log(n/g)}{\beta+3}} (\beta+3)\gamma\log^\omega(2^i g)/(2^i g) = \Theta(\beta(\gamma\log^\omega g)/g) = \varepsilon.$$

Solving for g yields $g = O(\frac{\beta\gamma}{\varepsilon}\log^\omega(1/\varepsilon))$ as desired. Stage 2 can be shown not to asymptotically increase the error.

To achieve the runtime we again start with the form of the algorithm without the halve-skip on every $(\beta+3)$rd step. Then the first step takes $O((n/g)\cdot g^\beta)$ time. And each ith step takes $O((n/2^{i-1})g^{\beta-1})$ time. Since each subsequent step takes half as much time, the runtime is dominated by the first $O(ng^{\beta-1})$ time step.

For the full algorithm, the first epoch ($\beta+3$ steps, including a skipped halve) takes $O(ng^{\beta-1})$ time, and the ith epoch takes $O(n/2^{(\beta+2)i}(g2^i)^{\beta-1}) = O(ng^{\beta-1}/2^{3i})$ time. Thus the time is still dominated by the first epoch. Again, stage 2 can be shown not to affect this runtime, and the total runtime bound is achieved as desired, and completes the proof.

Finally, we state a useful corollary about the expected error being 0. This holds specifically when we choose to discard the set P^+ or $P^- = \{p\in P \mid \chi(p) = -1\}$ at random on each halving.

Corollary A.1. *The expected error for any range $R\in\mathcal{A}$ on the ε-sample T created by Theorem A.1 is*

$$\mathsf{E}\left[\frac{|R\cap P|}{|P|}-\frac{|T\cap R|}{|T|}\right] = 0.$$

Note that there is no absolute value taken inside $\mathsf{E}[\cdot]$, so technically this measures the expected undercount.

RE-discrepancy.

We are also interested in achieving these same results for RE-discrepancy. To this end, the algorithms are identical. Lemma 3.3 replaces Lemma A.1. (S1) and (S2) still hold. Nothing else about the analysis depends on properties of disc or RE-disc, so Theorem A.1 can be restated for RE-discrepancy.

Theorem A.2. *Consider an uncertain point set P of size n and a family of subsets $\mathcal{A}$ of P_{cert}. Assume an $O(n^\beta)$ time algorithm to construct a coloring $\chi: P\to\{-1,+1\}$ so $\mathsf{RE\text{-}disc}_\chi(P,\mathcal{A}) = O(\gamma\log^\omega n)$ where β, γ, and ω are constant algorithm parameters dependent on $\mathcal{A}$, but not P (or n). There exists an algorithm to construct an ε-RE coreset of $(P,\mathcal{A})$ of size $g(\varepsilon,\mathcal{A}) = O((\gamma/\varepsilon)\log^\omega(\gamma/\varepsilon))$ in time $O(n\cdot g(\varepsilon,\mathcal{A})^{\beta-1})$.*