Coreset Construction and Estimation over Stochastic Data

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Computer Science and Technology

by

Lingxiao Huang

Dissertation Supervisor: Assistant Professor Jian Li

June 2017

Coreset Construction and Estimation over Stochastic Data

by

Lingxiao Huang

Submitted to the Institute for Interdisciplinary Information Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at

TSINGHUA UNIVERSITY

June 2017

Author .....
Institute for Interdisciplinary Information Sciences
April 15, 2015

Certified by .....
Jian Li
Assistant Professor
Thesis Supervisor

Coreset Construction and Estimation over Stochastic Data

by

Lingxiao Huang

Submitted to the Tsinghua University
on April 15, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In recent years, designing algorithms for geometric or combinatorial optimization problems over stochastic data have attracted more and more research interest. In this dissertation, we consider two well-known stochastic geometry models. One is the existential model where each point's location is fixed but only occurs with a certain probability. The other one is the locational model where each point has a probability distribution describing its location. Both stochastic geometry models have been widely studied in recent years. In this dissertation, we mainly focus on the following problems, *coreset construction for shape fitting problems and estimation for combinatorial optimization problems* on these stochastic geometry models.

The first problem concerns with a useful technique for handling large deterministic datasets, called coreset. Roughly speaking, a coreset is a small summary of the original large dataset while guaranteeing that answers for certain queries are provably close to the exact answer of corresponding queries. In this dissertation, we study how to construct coresets on stochastic models. We first extend the concept *ε-kernel coresets* to stochastic data. We consider approximating the expected width (an ε-EXP-KERNEL), as well as the probability distribution on the width (an (ε, τ)-QUANT-KERNEL) for any direction and show how to construct such coresets efficiently. Then we consider two stochastic shape fitting problems, stochastic k-center and stochastic j-flat-center. We propose a new notion called *generalized coresets*, which is a generalization of coresets. We also provide a framework for constructing generalized coresets of constant size for both the stochastic k-center problem and the stochastic j-flat-center problem. Using these generalized coresets, we give the first PTASs (polynomial time approximation schemes) for both stochastic shape fitting problems.

Secondly, we study the problems of computing the expected lengths of several combinatorial or geometric optimization problems in stochastic geometry models, including closest pair, minimum spanning tree, k-clustering, minimum perfect matching, and minimum cycle cover. Most of the above problems are known to be #P-hard.

In this dissertation, we propose two new techniques, called *finding stoch-core* and *Hierachical Partition Family (HPF)*. Combining our new techniques and Monte Carlo method, we obtain the first FPRAS (Fully Polynomial Randomized Approximation Scheme) for most of these problems in stochastic geometry models.

Dissertation Supervisor: Assistant Professor Jian Li

Contents

List of Figures

List of Tables

Chapter 1 Introduction

In recent years, stochastic data are pervasive in applications. Managing, analyzing and optimizing over such stochastic data have become an increasingly important issue and have attracted significant attentions from several research communities including theoretical computer science, databases, machine learning and sensor networks [29, 35, 100]. Theoretically, any deterministic combinatorial or geometry optimization problem has many uncertain counterparts (corresponding to different uncertainty models). A variety of classic problems in deterministic setting have been well studied, while systematic studies of them under uncertainty have only been initiated. For example, suppose we want to build k facilities to serve a set of uncertain demand points, and our goal is to minimize the expectation of the maximum distance from any realized demand point to its closest facility. This problems is called stochastic k-center which is first considered in this dissertation. Estimation and solving optimization problems over stochastic models and data have recently attracted significant attentions in several research communities (see e.g., [96, 100, 102]).

In the following, we list some application examples to exemplify where the stochastic data comes from, and what types of stochastic data we may encounter, see [79] for more examples.

1. **Stochastic shortest path.** Consider a traffic problem where we want to arrive at the airport before a specific time. There are several paths leading to the airport. In deterministic setting, each road costs a certain time. Our goal is to find a shortest path which minimizes the travel time. However in reality, we often know the distribution of the travel time of each road rather than the exact time. In this stochastic setting, we want to pick a path that maximizes the probability we can arrive at the airport on time. This problem has been considered a lot since the 1980s [18, 78, 81, 86, 90, 91].

2. **Fixed set stochastic knapsack.** The knapsack problem is a classic scheduling problem. One natural motivation is as follows. Suppose we are given a series of jobs and a single machine. Each job has a processing time and a profit. Our goal is to pick a subset of jobs which can be finished by a single machine one by one before the deadline and to maximize the total profit of these jobs. However, the processing time of each job is often random and follows from an individual distribution. So a stochastic variant of this problem is called the fixed set stochastic knapsack problem. We still want to choose a set of jobs to maximize the total profit. Since the processing time is random, we have an additional constraint that the probability that finishing all jobs before the deadline is at least some fixed constant $\gamma > 0$. In previous work, many groups of researchers have studied different distributions, including Bernoulli [50, 75], exponential [50] and Gaussian [52].

3. **Adaptive stochastic process.** Note that the solution is chosen in advance for the above two examples. The adaptive variants of these problems have also been studied extensively. For the adaptive stochastic shortest path problem, we know the exact travel time immediately when we pass through a road. Thus, adaptively choosing the following path may increase the probability that we arrive the airport before the deadline.

   Similarly we can consider adaptive policies for stochastic knapsack. In fact, researchers studied a more complicated version where the processing time is also random, and the precise values of the processing time and the profit are revealed when the job is completed. The goal is to gain as much profit as possible. Dean et al. [33] initially studied this problem and proposed a greedy algorithm. Later on, [24, 81] considered the same problem and improved their results.

   Many other adaptive stochastic problems have been well studied, such as adaptive stochastic matching [17] and adaptive stochastic probing [107].

4. **Stochastic geometry.** Theoretically, all deterministic computational geom-

etry problems have natural stochastic counterparts. We take the following stochastic variant of the facility location problem as an example. Suppose we want to build k facilities to serve a set of uncertain demand points (i.e., their locations are random), and our goal is to minimize the expectation of the maximum distance from any realized demand point to its closest facility. A variety of classic geometry problems in deterministic setting have been well studied, while systematic studies of them in different stochastic models have been initiated in recent years. Munteanu et al. [88] studied the stochastic minimum enclosing ball problem in fixed dimensional Euclidean space and gave a PTAS. In this dissertation, we also studied some fundamental problems in this area.

In this dissertation, we focus on two well-known stochastic geometry models: the locational uncertainty model and the existential uncertainty model. Both models have been studied extensively for a variety of computational geometry optimization problems or combinatorial optimization problems, such as closest pairs [70], nearest neighbors [5, 70], minimum spanning trees [65, 71], convex hulls [101], maxima [2], perfect matchings [65], clustering [30, 53] , minimum enclosing balls [88] and range queries [1, 4, 80]. We give the formal definitions of these two stochastic geometry models as follows.

1. Locational uncertainty model: We are given a metric space $\mathcal{P}$. The location of each node $v \in \mathcal{V}$ is a random point in the metric space $\mathcal{P}$ and the probability distribution is given as the input. Formally, we use the term *nodes* to refer to the vertices of the graph, *points* to describe the locations of the nodes in the metric space. We denote the set of nodes as $\mathcal{V} = \{v_1, \ldots, v_m\}$ and the set of points as $\mathcal{P} = \{s_1, \ldots, s_n\}$, where $m = |\mathcal{V}|$ and $n = |\mathcal{P}|$. A realization $\mathbf{r}$ can be represented by an m-dimensional vector $(r_1, \ldots, r_m) \in \mathcal{P}^m$ where point r_i is the location of node v_i for $1 \leq i \leq m$. Let $\mathbf{R}$ denote the set of all possible realizations. We assume that the distributions of the locations of nodes in the metric space $\mathcal{P}$ are independent, thus $\mathbf{r}$ occurs with probability $\Pr[\mathbf{r}] = \prod_{i \in [m]} p_{v_i r_i}$, where p_{vs} represents the probability that the location of

node v is point $s \in \mathcal{P}$. The model is also termed as the *locational uncertainty model* in [71].

2. Existential uncertainty model: A closely related model is the *existential uncertainty model* where the location of a node is a fixed point in the given metric space, but the existence of the node is probabilistic. In this model, we use p_{s_i} to denote the probability that node v_i exists (if exists, its location is s_i). For simplicity, we also use p_i to represent p_{s_i}. A realization $\mathbf{r}$ can be represented by a subset $P \subset \mathcal{P}$ and $\Pr[\mathbf{r}] = \prod_{s_i \in P} p_i \prod_{s_i \notin P}(1 - p_i)$.

Now, we introduce and motivate two classes of problems in this dissertation: coreset construction and estimating the expected values of combinatorial optimization problems over stochastic data. We also briefly state our contributions one by one.

1.1 ε-Kernel Coresets over Stochastic Data

Given a large dataset P and a class C of queries, a coreset S is a dataset of much smaller size such that for every query $r \in C$, the answer $r(S)$ for the small dataset S is close to the answer $r(P)$ for the original large dataset P. Coresets have become more relevant in the era of big data as they summarize large datasets by datasets with potentially much smaller size, and at the same time guarantee the answer to certain classes of queries to be close to the true answer. The notion of a coreset was studied in the directional width problem (in which a coreset is called an ε-kernel) and several other geometric shape fitting problems in the seminal paper [7].

We introduce some notation and briefly review the definition of ε-kernel. For a set P of deterministic points, the *support function* $\mathbf{f}(P, \vec{u})$ is defined to be $\mathbf{f}(P, \vec{u}) = \max_{s \in P}\langle \vec{u}, s\rangle$ for $\vec{u} \in \mathbb{R}^d$, where $\langle ., . \rangle$ is the inner product. The *directional width* of P in direction $u \in \mathbb{R}^d$, denoted by $\omega(P, \vec{u})$, is defined by $\omega(P, \vec{u}) = \mathbf{f}(P, \vec{u}) + \mathbf{f}(P, -\vec{u})$. It is easy to see that the support function and the directional width only depend on the convex hull of P. A subset $Q \subseteq P$ is called an *ε-kernel of P* if for each direction $\vec{u} \in \mathbb{R}^d$, $(1 - \varepsilon)\omega(P, \vec{u}) \leq \omega(Q, \vec{u}) \leq \omega(P, \vec{u})$. For any set of n points, there

is an ε-kernel of size $O(\varepsilon^{-(d-1)/2})$ [7, 8], which can be constructed in $O(n+\varepsilon^{-(d-3/2)})$ time [26, 88].

Our contribution. (Chapter 4) Our main results can be summarized as follows:

1. Suppose $\mathcal{P}$ is a set of stochastic points (in either the existential or locational uncertainty model). Define the *expected* directional width of $\mathcal{P}$ in direction u to be $\omega(\mathcal{P},\vec{u}) = \mathbb{E}_{P\sim\mathcal{P}}[\omega(P,\vec{u})]$, where $P \sim \mathcal{P}$ means that P is a (random) realization of $\mathcal{P}$. We first consider how to construct ε-EXP-KERNEL which is defined as follows:

   Definition 1. *For a constant $\varepsilon > 0$, a set S of (deterministic or stochastic) points in $\mathbb{R}^d$ is called an ε-EXP-KERNEL of $\mathcal{P}$, if for all directions $\vec{u} \in \mathbb{R}^d$,*

   $$(1-\varepsilon)\omega(\mathcal{P},\vec{u}) \leq \omega(S,\vec{u}) \leq \omega(\mathcal{P},\vec{u}).$$

   Our first main result is that an ε-EXP-KERNEL of size $O(\varepsilon^{-(d-1)/2})$ exists for both existential and locational uncertainty model and can be constructed in nearly linear time.

   Theorem 2. *$\mathcal{P}$ is a set of n uncertain points in $\mathbb{R}^d$ (in either locational uncertainty model or existential uncertainty model). There exists an ε-EXP-KERNEL of size $O(\varepsilon^{-(d-1)/2})$ for $\mathcal{P}$. For existential uncertainty model (locational uncertainty model resp.), such an ε-EXP-KERNEL can be constructed in $O(\varepsilon^{-(d-1)}n\log n)$ time, where n is the number of points (possible locations).*

   The existential result is a simple Minkowski sum argument. We first show that there exists a convex polytope M such that for any direction, the directional width of M is exactly the same as the expected directional width of $\mathcal{P}$ (Lemma 16). This immediately implies the existence of a ε-EXP-KERNEL consisting $O(\varepsilon^{-(d-1)/2})$ deterministic points (using the result in [7]), but without the subset constraint. The Minkowski sum argument seems to suggest that the complexity of M is exponential. However, we show that the complexity of M is in fact

polynomial $O(n^{2d-2})$ and we can construct it explicitly in $O(n^{2d-1}\log n)$ time (Theorem 20).

Although the complexity of M is polynomial, we cannot afford to construct it explicitly if we are to construct an ε-EXP-KERNEL in nearly linear time. Thus we construct the ε-EXP-KERNEL without explicitly constructing M. In particular, we show that it is possible to find the extreme vertex of M in a given direction in nearly linear time, by computing the gradient of the support function of M. We also provide quadratic-size data structures that can calculate the exact width $\omega(\mathcal{P},\cdot)$ in logarithmic time under both models in $\mathbb{R}^2$ (Section 4.9).

We also show that under subset constraint (i.e., the ε-EXP-KERNEL is required to be a subset of the original point set, with the same probability distribution for each chosen point), there is no ε-EXP-KERNEL of sublinear size (Lemma 25). However, if there is a constant lower bound $\beta > 0$ on the existential probabilities (called β-assumption), we can construct an ε-EXP-KERNEL of constant size (Theorem 26 and Section 4.6).

2. Sometimes it is useful to obtain more than just the expected value (say of the width) on a query; rather one may want to return (an approximation of) a representation of the full probability distribution that the query can take. So we also consider the construction of the following (ε,τ)-QUANT-KERNEL.

Definition 3. *For a constant $\varepsilon,\tau > 0$, a set $\mathcal{S}$ of stochastic points in $\mathbb{R}^d$ is called an (ε,τ)-*QUANT-KERNEL *of $\mathcal{P}$, if for all directions $\vec{u}$ and all $x \geq 0$,*

$$\Pr_{P\sim\mathcal{P}}\Big[\omega(P,\vec{u}) \leq (1-\varepsilon)x\Big]-\tau \leq \Pr_{S\sim\mathcal{S}}\Big[\omega(S,\vec{u}) \leq x\Big] \leq \Pr_{P\sim\mathcal{P}}\Big[\omega(P,\vec{u}) \leq (1+\varepsilon)x\Big]+\tau.$$

Now, we describe our main results for (ε,τ)-QUANT-KERNELs. We first propose a quite simple but general algorithm for constructing (ε,τ)-QUANT-KERNELs, which achieves the following guarantee.

Theorem 4. *An (ε,τ)-*QUANT-KERNEL *of size $\widetilde{O}\left(\tau^{-2}\varepsilon^{-3(d-1)/2}\right)$ can be constructed in $\widetilde{O}\left(n\tau^{-2}\varepsilon^{-(d-1)}\right)$ time, under both existential and locational uncer-*

tainty models.

The algorithm is surprisingly simple. Take a certain number of i.i.d. realizations, compute an ε-kernel for each realization, and then associate each kernel with probability $1/N$ (so the points are not independent). The analysis requires the VC uniform convergence bound for unions of halfspaces. The details can be found in Section 4.3.1.

For existential uncertainty model, we can improve the size bound as follows.

Theorem 5. *$\mathcal{P}$ is a set of uncertain points in $\mathbb{R}^d$ with existential uncertainty. There exists an (ε, τ)-*QUANT-KERNEL *for $\mathcal{P}$, which consists of a set of independent uncertain points of cardinality $\widetilde{O}(\varepsilon^{-(d-1)}\tau^{-2})\}$. The algorithm for constructing such a coreset runs in $\widetilde{O}(n \log^{O(d)} n)$ time.*

We note that another advantage of the improved construction is that the (ε, τ)-QUANT-KERNEL is a set of independent stochastic points (rather than correlated points as in Theorem 4). We achieve the improvement by two algorithms. The first algorithm transforms the Bernoulli distributed variables into Poisson distributed random variables and creates a probability distribution using the parameters of the Poissons, from which we take a number of i.i.d. samples as the coreset. Our analysis leverages the additivity of Poisson distributions and the VC uniform convergence bound (for halfspaces). However, the number of samples required depends on $\lambda(\mathcal{P})$, so the first algorithm only works when $\lambda(\mathcal{P})$ is small. The second algorithm complements the first one by identifying a convex set K that lies in the convex hull of $\mathcal{P}$ with high probability (K exists when $\lambda(\mathcal{P})$ is large) and uses a small size deterministic ε-kernel to approximate K. The points in $\overline{K} = \mathcal{P} \setminus K$ can be approximated using the same sampling algorithm as in the first algorithm and we can show that $\lambda(\overline{K})$ is small, thus requiring only a small number of samples. Our algorithm can be easily extended to $\mathbb{R}^d$ for any constant d and the size of the coreset is $\widetilde{O}(\tau^{-2}\varepsilon^{-(d-1)})$. In Section 4.3.2, we show such an (ε, τ)-QUANT-KERNEL can be computed in $O(n\text{polylog}n)$ time using an

iterative sampling algorithm. Our technique has some interesting connections to other important geometric problems (such as the Tukey depth problem) [87], may be interesting in its own right.

3. The notion (ε, τ)-QUANT-KERNEL is also not powerful enough for certain shape fitting problems (e.g., the minimum enclosing cylinder problem and the minimum spherical shell problem) in the stochastic setting. The main reason is the appearance of the l_2-norm in the objective function. So we need to be able to handle the fractional powers in the objective function. For a set P of points in $\mathbb{R}^d$, the polar set of P is defined to be $P^\star = \{\vec{u} \in \mathbb{R}^d \mid \langle \vec{u}, s\rangle \geq 0, \forall s \in P\}$. Let r be a positive integer. Given a set P of points in $\mathbb{R}^d$ and $\vec{u} \in P^\star$, we define a function

$$T_r(P, \vec{u}) = \max_{s\in P} \langle \vec{u}, s\rangle^{1/r} - \min_{s\in P} \langle \vec{u}, s\rangle^{1/r}.$$

We only care about the directions in $\mathcal{P}^\star$ (i.e., the polar of the points in $\mathcal{P}$) for which $T_r(P, \vec{u}), \forall P \sim \mathcal{P}$ is well defined.

Definition 6. *For a constant $\varepsilon > 0$, a positive integer r, a set $\mathcal{S}$ of stochastic points in $\mathbb{R}^d$ is called an (ε, r)-FPOW-KERNEL of $\mathcal{P}$, if for all directions $\vec{u} \in \mathcal{P}^\star$,*

$$(1-\varepsilon)\mathbb{E}_{P\sim\mathcal{P}}[T_r(P, \vec{u})] \leq \mathbb{E}_{P\sim\mathcal{S}}[T_r(P, \vec{u})] \leq (1+\varepsilon)\mathbb{E}_{P\sim\mathcal{P}}[T_r(P, \vec{u})].$$

For (ε, r)-FPOW-KERNELS, we provide a linear time algorithm for constructing an (ε, r)-FPOW-KERNEL of size $\widetilde{O}(\varepsilon^{-(rd-r+2)})$ in the existential uncertainty model under the β-assumption where each point is present with probability above β. The algorithm is almost the same as the construction in Section 4.3.1 except that some parameters are different.

Theorem 7. *(Section 4.4) An (ε, r)-FPOW-KERNEL of size $\widetilde{O}(\varepsilon^{-(rd-r+2)})$ can be constructed in $\widetilde{O}\left(n\varepsilon^{-(rd-r+4)/2}\right)$ time in the existential uncertainty model under the β-assumption.*

4. Finally, we show that the above results, combined with the duality and lineariza-

tion arguments [7], can be used to obtain constant size coresets for the function extent problem in the stochastic setting, and to maintain extent measures for stochastic moving points.

Using the above results, we also obtain efficient approximation schemes for various shape-fitting problems in the stochastic setting, such as minimum enclosing ball, minimum spherical shell, minimum enclosing cylinder and minimum cylindrical shell in different stochastic settings. We summarize our application results in the following theorems. The details can be found in Section 4.5.

Theorem 8. *Suppose $\mathcal{P}$ is a set of n independent stochastic points in $\mathbb{R}^d$ under either existential or locational uncertainty model. There are linear time approximation schemes for the following problems: (1) finding a center point c to minimize $\mathbb{E}[\max_{s\in\mathcal{P}} \|s-c\|^2]$; (2) finding a center point c to minimize $\mathbb{E}[\mathsf{obj}(c)] = \mathbb{E}[\max_{s\in P} \|s-c\|^2 - \min_{s\in P} \|s-c\|^2]$. Note that when $d=2$ the above two problems correspond to minimizing the expected areas of the enclosing ball and the enclosing annulus, respectively.*

Under β-assumption, we can obtain efficient approximation schemes for the following shape fitting problems.

Theorem 9. *Suppose $\mathcal{P}$ is a set of n independent stochastic points in $\mathbb{R}^d$, each appearing with probability at least β, for some fixed constant $\beta > 0$. There are linear time approximation schemes for minimizing the expected radius (or width) for the minimum spherical shell, minimum enclosing cylinder, minimum cylindrical shell problems over $\mathcal{P}$.*

1.2 Coreset Construction for Shape Fitting Problems over Stochastic Data

We study two classic geometric optimization problems, the k-center problem and the j-flat center problem in Euclidean spaces. Both problems are important in geometric data analysis. We generalize both problems to the stochastic settings. For the

stochastic k-center problem, we would like to find k points in a fixed dimensional Euclidean space, such that the expected value of the k-center objective is minimized. For the stochastic j-flat-center problem, we seek a j-flat (i.e., a j-dimensional affine subspace) such that the expected value of the maximum distance from any point to the j-flat is minimized. One of the motivations for this stochastic version comes from the stochastic variant of the ℓ_∞ regression problem. We still want to construct coresets for these two shape fitting problems.

In the following, we first define the two stochastic shape fitting problems. Then we briefly introduce our contributions and techniques.

Stochastic k-Center. The deterministic Euclidean k-center problem is a central problem in geometric optimization [11, 8]. It asks for a k-point set F in $\mathbb{R}^d$ such that the maximium distance from any of the n given points to its closest point in F is minimized.

Definition 10. *For a set of points $P \in \mathbb{R}^d$, and a k-point set $F = \{(f_1, \ldots, f_k) \mid f_i \in \mathbb{R}^d, 1 \leq i \leq k\}$, we define $\mathrm{K}(P, F) = \max_{s \in P} \min_{1 \leq i \leq k} \mathrm{d}(s, f_i)$ as the k-center value of F w.r.t. P. We use $\mathcal{F}$ to denote the family of all k-point sets in $\mathbb{R}^d$. Given a set $\mathcal{P}$ of n stochastic points (in either the existential or locational uncertainty model) in $\mathbb{R}^d$, and a k-point set $F \in \mathcal{F}$, we define the expected k-center value of F w.r.t $\mathcal{P}$ as*

$$\mathrm{K}(\mathcal{P}, F) = \mathbb{E}_{P \sim \mathcal{P}}[\mathrm{K}(P, F)].$$

In the stochastic minimum k-center problem, our goal is to find a k-point set $F \in \mathcal{F}$ which minimizes $\mathrm{K}(\mathcal{P}, F)$. In this dissertation, we assume that both the dimensionality d and k are fixed constants.

Stochastic j-Flat-Center. The deterministic j-flat-center problem is defined as follows: given n points in $\mathbb{R}^d$, we would like to find a j-flat F (i.e., a j-dimensional affine subspace) such that the maximum distance from any given point to F is minimized. It is a common generalization of the minimum enclosing ball ($j = 0$), minimum enclosing cylinder ($j = 1$), and minimum width problems ($j = d - 1$), and has been

well studied in computational geometry [8, 45, 106]. Its stochastic version is also naturally motivated by the stochastic variant of the ℓ_∞ regression problem: Suppose we would like to fit a set of points by an affine subspace. However, those points may be produced by some machine learning algorithm, which associates some confidence level to each point (i.e., each point has an existential probability). This naturally gives rise to the stochastic j-flat-center problem. Formally, it is defined as follows.

Definition 11. *Given a set P of n points in $\mathbb{R}^d$, and a j-flat $F \in \mathcal{F}$ ($0 \le j \le d-1$), where $\mathcal{F}$ is the family of all j-flats in $\mathbb{R}^d$, we define the j-flat-center value of F w.r.t. P to be $\mathrm{J}(P,F) = \max_{s\in P} \mathrm{d}(s,F)$, where $\mathrm{d}(s,F) = \min_{f\in F} \mathrm{d}(s,f)$ is the distance between point s and j-flat F. Given a set $\mathcal{P}$ of n stochastic points (in either the existential or locational model) in $\mathbb{R}^d$, and a j-flat $F \in \mathcal{F}$ ($0 \le j \le d-1$), we define the expected j-flat-center value of F w.r.t. $\mathcal{P}$ to be*

$$\mathrm{J}(\mathcal{P},F) = \mathbb{E}_{P\sim\mathcal{P}}[\mathrm{J}(P,F)].$$

In the stochastic minimum j-flat-center problem, our goal is to find a j-flat F which minimizes $\mathrm{J}(\mathcal{P},F)$.

Previous Results and Our contributions. Recall that a *polynomial time approximation scheme (PTAS)* for a minimization problem is an algorithm A that produces a solution whose cost is at most $1+\varepsilon$ times the optimal cost in polynomial time, for any fixed constant $\varepsilon > 0$.

Stochastic k-Center. Cormode and McGregor [30] first studied the stochastic k-center problem in a finite metric graph under the locational uncertainty model, and obtained a bi-criterion constant approximation. Guha and Munagala [53] improved their result to a single-criterion constant factor approximation. Recently, Wang and Zhang [108] studied the stochastic k-center problem on a line, and proposed an efficient exact algorithm. No result better than a constant approximation is known for the Euclidean space $\mathbb{R}^d$ ($d \ge 2$). We obtain the first PTAS for the stochastic k-center problem in $\mathbb{R}^d$.

Theorem 12. *Assume that both k and d are fixed constants. There exists a PTAS for the stochastic minimum k-center problem in $\mathbb{R}^d$, under either the existential or the locational uncertainty model.*

Our result generalizes the PTAS for stochastic minimum enclosing ball by Munteanu et al. [88]. We remark that the assumption that k is a constant is necessary for getting a PTAS, since even the deterministic Euclidean k-center problem is APX-hard for arbitrary k even in $\mathbb{R}^2$ [42].

Stochastic j-Flat-Center. Our main result for the stochastic j-flat-center is as follows.

Theorem 13. *Assume that the dimensionality d is a constant. There exists a PTAS for the stochastic minimum j-flat-center problem, under either the existential or the locational uncertainty model.*

This result also generalizes the PTAS for stochastic minimum enclosing ball (i.e., 0-flat-center) by Munteanu et al. [88]. It also generalizes a previous PTAS for the stochastic minimum enclosing cylinder (i.e., 1-flat-center) problem in the existential model where the existential probability of each point is assumed to be lower bounded by a small fixed constant in Chapter 4.

Our techniques. Our techniques for both problems heavily rely on the powerful notion of coresets. In a typical deterministic geometric optimization problem, an instance P is a set of deterministic (weighted) points. A coreset S of P is a set of (weighted) points, such that the solution for the optimization problem over S is a good approximate solution for P. [1] In Chapter 4, we generalize the notion of ε-kernel coreset (for directional width) to stochastic points. However, our techniques can only handle directional width, and extending it to problems such as stochastic minimum enclosing cylinder requires certain technical assumption.

In this dissertation, we introduce a new framework for solving geometric optimization problems over stochastic points. For a stochastic instance $\mathcal{P}$, we consider

[1] It is possible to define coresets for other classes of optimization problems.

$\mathcal{P}$ as a collection of realizations $\mathcal{P} = \{P \mid P \sim \mathcal{P}\}$. Each realization P has a weight $\Pr[\models P]$, which is its realized probability. Now, we can think the stochastic problem as a certain deterministic problem over (exponential many) all realizations (each being a point set). Our framework constructs an object $\mathcal{S}$ satisfying the following properties.

1. Basically, $\mathcal{S}$ has a constant size description (the constants may depend on d, ε, and k).

2. The objective value for a certain deterministic optimization problem over $\mathcal{S}$ can approximate the objective for the original stochastic problem well. Moreover, the solution to the deterministic optimization over $\mathcal{S}$ is a good approximation for the original problem as well.

At a high level, $\mathcal{S}$ serves very similar roles as the coresets in the deterministic setting. Note that the form of $\mathcal{S}$ may vary for different problems: in stochastic k-center, it is a collection of weighted point sets (we call $\mathcal{S}$ an SKC-Coreset); in stochastic j-flat-center, it is a combination of two collections of weighted point sets for two intermediate problems (we call $\mathcal{S}$ an SJFC-Coreset).

For stochastic k-center under the existential model, we construct an SKC-Coreset $\mathcal{S}$ in two steps. First, we map all realizations to their additive ε-coresets (for deterministic k-centers) [11]. Since there are only a polynomial number of possible additive ε-coresets, the above mapping can partition the space of all realizations into a polynomial number of parts, such that the realizations in each part have very similar objective functions. Moreover, for each additive ε-coresets, it is possible to compute the total probability of the realizations that are mapped to the coreset. In fact, this requires a subtle modification of the construction in [11] so that we can compute the aforementioned probability efficiently. This step has reduced the exponential number of realizations to a polynomial size representation. Next, we define a generalized shape fitting problem, call the *generalized k-median* problem, over the collection of above additive ε-coresets. Then, we need to properly generalize the previous definition of coreset and the total sensitivity (a notion proposed in the deterministic coreset context by Langberg and Schulman [77]), and prove a constant upper bound for the

generalized total sensitivity by relating it to the total sensitivity of the ordinary k-median problem. The SKC-Coreset $\mathcal{S}$ is a generalized coreset for the generalized k-median problem, which consists of a constant number of weighted point sets.

For stochastic k-center under the locational model, computing the weight for each set in the SKC-Coreset $\mathcal{S}$ is somewhat more complicated. We need to reduce the computational problem to a family of bipartite holant problems, and apply the celebrated result by Jerrum, Sinclair, and Vigoda [67].

For the stochastic minimum j-flat-center problem, we proposed an efficient algorithm for constructing an SJFC-Coreset. We utilize several ideas in Chapter 4, as well as prior results on the shape fitting problem. We first partition the realizations $P \sim \mathcal{P}$ into two parts through a construction similar to the (ε, τ)-quant-kernel construction in Chapter 4. Roughly speaking, after linearization, we need to find a convex set $\mathcal{K}$ in a higher dimensional space such that the total probability of any point falling outside $\mathcal{K}$ is small, but not so small such that in each direction the expected directional width of $\mathcal{P}$ is comparable to that of $\mathcal{K}$. Then, for those points inside $\mathcal{K}$, it is possible to use a slight modification of the construction in Chapter 4 to construct a collection of weighted point sets. For the points outside $\mathcal{K}$, since the total probability is small, we reduce the problem to a weighted j-flat-median problem, and use the coreset in [106] (this step is similar to that in [88]). By combining the two collections, we obtain the SJFC-Coreset $\mathcal{S}$ for the problem, which is of constant size. Then, we can easily obtain a PTAS by solving a constant size polynomial system defined by $\mathcal{S}$.

We remark that our overall approach is very different from that in Munteanu et al. [88] (except one aforementioned step and that they also crucially used some machinery from the coreset literature). Munteanu et al. [88] defined a near-metric distance measure $m(A, B) = \max_{a \in A, b \in B} \mathrm{d}(a, b)$ for two non-empty point sets A, B. This near-metric measure satisfies many metric properties, like non-negativity, symmetry and the triangle inequality. By lifting the problem to the space defined by such metric and utilizing a previous coreset result for clustering, they obtained a PTAS for the problem. However, in the more general stochastic minimum k-center problem and

stochastic minimum j-flat-center problem, it is unclear how to translate the distance function between point sets and k-centers or point sets and j-flat sets to a near-metric distance (and still satisfies symmetry and triangle inequality).

1.3 Estimation for Combinatorial Optimization Problems over Stochastic Data

We are interested in the following natural problem over both existential and locational uncertainty models: estimating the expected values of certain statistics of combinatorial objects. In this dissertation, we study several combinatorial or geometry problems in these two models: the closest pair problem, minimum spanning tree, minimum perfect matching (assuming an even number of nodes), k-clustering and minimum cycle cover. We take the minimum spanning tree problem for example. Let MST be the length of the minimum spanning tree (which is a random variable) and $\mathsf{MST}(\mathbf{r})$ be the length of the minimum spanning tree spanning all points in the realization $\mathbf{r}$. We would like to estimate the following quantity:

$$\mathbb{E}[\mathsf{MST}] = \sum_{\mathbf{r}\in\mathbf{R}} \Pr[\mathbf{r}] \cdot \mathsf{MST}(\mathbf{r}).$$

However, the above formula does not give us an efficient way to estimate the expectation since it involves an exponential number of terms. In fact, computing the exact expected value are either NP-hard or #P-hard. Following many of the theoretical computer science literatures on approximate counting and estimation, our goal is to obtain fully polynomial randomized approximation schemes for computing the expected values.

Our contribution. (Chapter 6) We recall that a *fully polynomial randomized approximation scheme (FPRAS)* for a problem f is a randomized algorithm A that takes an input instance x, a real number $\varepsilon > 0$, returns $A(x)$ such that $\Pr[(1-\varepsilon)f(x) \leq A(x) \leq (1+\varepsilon)f(x)] \geq \frac{3}{4}$ and its running time is polynomial in both the size of the

input n and $1/\varepsilon$. Our main contributions can be summarized in Table 1.1.

Problems		Existential	Locational
Closest Pair (§6.1)	$\mathbb{E}[\mathsf{C}]$	FPRAS	FPRAS
	$\Pr[\mathsf{C} \leq 1]$	FPRAS	FPRAS
	$\Pr[\mathsf{C} \geq 1]$	Inapprox	Inapprox
Diameter (§6.1)	$\mathbb{E}[\mathsf{D}]$	FPRAS	FPRAS
	$\Pr[\mathsf{D} \leq 1]$	Inapprox	Inapprox
	$\Pr[\mathsf{D} \geq 1]$	FPRAS	FPRAS
Minimum Spanning Tree (§6.3)	$\mathbb{E}[\mathsf{MST}]$	FPRAS[71]	FPRAS
k-Clustering (§6.2)	$\mathbb{E}[\mathsf{kCL}]$	FPRAS	Open
Perfect Matching (§6.4)	$\mathbb{E}[\mathsf{PM}]$	N.A.	FPRAS
kth Closest Pair (§6.8.1)	$\mathbb{E}[\mathsf{kC}]$	FPRAS	Open
Cycle Cover (§6.5)	$\mathbb{E}[\mathsf{CC}]$	FPRAS	FPRAS
kth Longest m-Nearest Neighbor (§6.6)	$\mathbb{E}[\mathsf{kmNN}]$	FPRAS	Open

Table 1.1: Our results for some problems in different stochastic models.

1. Closest Pair: We use C to denote the minimum distance of any pair of two nodes. If a realization has less than two nodes, C is zero. Computing $\Pr[\mathsf{C} \leq 1]$ exactly in the existential model is known to be #P-hard even in an Euclidean plane [72], but no nontrivial algorithmic result is known before. So is computing $\Pr[\mathsf{C} \geq 1]$. In fact, it is not hard to show that computing $\Pr[\mathsf{C} \geq 1]$ is inapproximable within any factor in a metric space (Section 6.8.2).

   We also consider the problem of computing expected distance $\mathbb{E}[\mathsf{C}]$ between the closest pair in the same model. We prove that the problem is #P-hard in Section 6.8.2 and give the first known FPRAS in Section 6.1. Note that an FPRAS for computing $\Pr[\mathsf{C} \leq 1]$ does not imply an FPRAS for computing $\mathbb{E}[\mathsf{C}]$ [2].

2. Diameter: The problem of computing the expected length of the diameter can be reduced to the closest pair problem as follows. Assume that the longest distance between two points in $\mathcal{P}$ is W. We construct the new instance $\mathcal{P}'$ as follows: for any two points $s, t \in \mathcal{P}$, let their distance be $2W - \mathrm{d}(s,t)$ in $\mathcal{P}'$.

[2] To the contrary, an FPRAS for computing $\Pr[\mathsf{C} \geq 1]$ or $\Pr[\mathsf{C} = 1]$ would imply an FPRAS for computing $\mathbb{E}[\mathsf{C}]$ since $\mathbb{E}[\mathsf{C}] = \sum_{(s_i,s_j)} \Pr[\mathsf{C} = \mathrm{d}(s_i,s_j)]\mathrm{d}(s_i,s_j) = \int \Pr[\mathsf{C} \geq t]\mathrm{d}t = \sum_{(s_i,s_j)} \Pr[\mathsf{C} \geq \mathrm{d}(s_i,s_j)](\mathrm{d}(s_i,s_j) - \mathrm{d}(s_i',s_j'))$.

The new instance is still a metric. The sum of the distance of closest pair in $\mathcal{P}$ and the diameter in $\mathcal{P}'$ is exactly $2W$ (if there are at least two realized points). Hence, the answer for the diameter can be easily derived from the answer for closest pair in $\mathcal{P}'$.

3. Minimum Spanning Tree: Computing $\mathbb{E}[\mathsf{MST}]$ exactly in both uncertainty models is known to be #P-hard [71]. Kamousi, Chan, and Suri [71] developed an FPRAS for estimating $\mathbb{E}[\mathsf{MST}]$ in the existential uncertainty model and a constant factor approximation algorithm in the locational uncertainty model.

   Estimating $\mathbb{E}[\mathsf{MST}]$ is amendable to several techniques. We obtain an FPRAS for estimating $\mathbb{E}[\mathsf{MST}]$ in the locational uncertainty model using the stoch-core techinque in Section 6.3. In fact, the idea in [71] can also be extended to give an alternative FPRAS (Section 6.9). It is not clear how to extend their idea to other problems.

4. Clustering (k-clustering): In the deterministic k-clustering problem, we want to partition all points into k disjoint subsets such that the spacing of the partition is maximized, where the spacing is defined to be the minimum of any $\mathrm{d}(u, v)$ with u, v in different subsets [74]. In fact, the optimal cost of the problem is the length of the $(k-1)$th most expensive edge in the minimum spanning tree [74]. We show how to estimate $\mathbb{E}[\mathsf{kCL}]$ using the HPF (hierarchical partition family) technique in Section 6.2.

5. Perfect Matching: We assume that there are even number of nodes to ensure that a perfect matching always exists. Therefore, only the locational uncertainty model is relevant here. We give the first FPRAS for approximating the expected length of minimum perfect matching in Section 6.4 using a more complicated stoch-core technique.

All of our algorithms run in polynomial time. However, we have not attempted to optimize the exact running time.

Our techniques. Perhaps the simplest and the most commonly used technique for estimating the expectation of a random variable is the Monte Carlo method, that is to use the sample average as the estimate. However, the method is only efficient (i.e., runs in polynomial time) if the variance of the random variable is small (See Lemma 14). To circumvent the difficulty caused by the high variance, a general methodology is to decompose the expectation of the random variable into a convex combination of conditional expectations using the law of total expectation: $\mathbb{E}[X] = \mathbb{E}_Y\big[\,\mathbb{E}[X \mid Y]\,\big] = \sum_y \Pr[Y = y]\,\mathbb{E}[X \mid Y = y]$. Hopefully, $\Pr[Y = y]$ can be estimated (or calculated exactly) efficiently, and the random variable X conditioning on each event y has a low variance. However, choosing the events Y to condition on can be tricky.

We develop two new techniques for choosing such events, each being capable of solving a subset of aforementioned problems. In the first technique, we first identify a set $\mathbb{H}$ of points, called the *stoch-core* of the problem, such that (1): with high probability, all nodes realize in $\mathbb{H}$ and (2): conditioning on event (1), the variance is small. Then, we choose Y to be the number of nodes realized to points not in $\mathbb{H}$. We compute the $(1 \pm \varepsilon)$-estimates for $Y = 0, 1$ using Monte Carlo by (1) and (2). The problematic part is when Y is large, i.e., many nodes realize to points outside $\mathbb{H}$. Even though the probability of such events is very small, the value of X under such events may be considerably large, thus contributing nontrivially. However, we can show that the contribution of such events is dominated by the first few events and thus can be safely ignored. Choosing appropriate stoch-core is easy for some problems, such as closest pair and minimum spanning tree, while it may require additional idea for other problems such as minimum perfect matching.

Our second technique utilizes a notion called *Hierarchical Partition Family (HPF)*. The HPF has n levels, each representing a clustering of all points. For a combinatorial problem, for which the solution is a set of edges, we define Y to be the highest level such that some edge in the solution is an inter-cluster edge. Informally, conditioning on the information of Y, we can essentially bound the variance of X (hence use the Monte Carlo method). To implement Monte Carlo, we need to be able to take samples

efficiently conditioning on Y. We show that such sampling problems can be reduced to, or have connections to, classical approximate counting and sampling problems, such as approximating permanent, counting knapsack.

Chapter 2 Preliminaries

In this chapter, we recall a useful tool, the Chernoff bound. Suppose we want to estimate $\mathbb{E}[X]$. In each Monte Carlo iteration, we take a sample (a realization of all nodes), and compute the value of X for the sample. At the end, we output the average over all samples. The number of samples required by this algorithm is suggested by the following standard Chernoff bound.

Lemma 14. (Chernoff Bound) *Let random variables $X_1, X_2, \ldots, X_N$ be independent random variables taking on values between 0 and U. Let $X = \frac{1}{N}\sum_{i=1}^{N} X_i$ and μ be the expectation of X, for any $\varepsilon > 0$,*

$$\Pr\left[X \in [(1-\varepsilon)\mu, (1+\varepsilon)\mu]\right] \geq 1 - 2e^{-N\frac{\mu}{U}\varepsilon^2/4}.$$

Therefore, for any $\varepsilon > 0$, in order to get an $(1\pm\varepsilon)$-approximation with probability $1 - \frac{1}{\text{poly}(n)}$, the number of samples needs to be $O(\frac{U}{\mu\varepsilon^2}\log n)$. If $\frac{U}{\mu}$, the ratio between the maximum possible value of X and the expected value $\mathbb{E}[X]$, is bounded by a polynomial size of the input, we can use the above Monte Carlo method to estimate $\mathbb{E}[X]$ with a polynomial number of samples.

Chapter 3 Related Work

In this chapter, we will show some prior work about our research problems. We start with introducing some related work about shape fitting problems in both deterministic and stochastic settings. Next, we discuss related work for coreset construction. Then we review some prior work about computing the expected value of combinatorial optimization problems in different stochastic models. We also briefly mention some other stochastic models that is conceptually or technically related to the dissertation.

Shape fitting problem. A number of theoretical results for the k-center problem have been obtained in the past. In deterministic settings, Agarwal and Procopiuc [11] considered the k-center problem and showed that there exists an additive coreset of a constant size which can represent the whole input point set if k and d are both constants. Har-Peled and Varadarajan [61] improved their result in high dimensions and gave an PTAS if k is a constant. Cormode and McGregor [30] considered the k-center problem for the locational model in a finite metric graph, and obtained a bi-criterion constant approximation. Guha and Munagala [53] improved the result to a true constant factor approximation. Munteanu et al. [88] studied the minimum enclosing ball problem (a.k.a. 1-center problem) for stochastic points in fixed dimensional Euclidean space and gave a PTAS. Coresets were also constructed for imprecise points [85] to help derive results for approximating convex hulls and a variety of other shape-fitting problems. Note that their model is different from the existential or locational models.

Other projective clustering problems have also been studied extensively. If k is a constant, the existence of an additive coreset of a constant size for k-line-center, i.e., for the problem of covering P by k congruent cylinders of the minimum radius, was first proved by Agarwal et al. [11]. Har-Peled and Varadarajan [61] obtained an PTAS for k j-flat-center problem while j and k are both constants. Langberg

and Schulman [77] showed that for the weighted k-median/k-means problem, [1] there exists an ε-coreset of size depending polynomially on d and k by bounding the total sensitivity. Varadarajan and Xiao [106] studied the k-line clustering problem and the (j, k) integer projective clustering problem, and showed that there exists an ε-coreset of a poly-logarithmic size.

Coreset construction. There is a large body of literature [93] on constructing coresets for various problems, such as shape fitting [7, 8], shape fitting with outliers [62], clustering [27, 45, 47, 60, 77], integrals [77], matrix approximation and regression [34, 45] and in different settings, such as geometric data streaming [8, 26] and privacy setting [43]. We have introduced some results for shape fitting problems. In this part, we review other applications of coreset construction. For example, Har-Peled and Wang [62] provided a coreset construction approach for handling outliers. From the dual (function extent) perspective, they want to approximate the distance between two level sets in an arrangement of hyperplanes. In the locational model, coresets are created for range counting queries [1] under the subset constraint, but these techniques do not translate to this setting because ε-kernel coresets in general cannot be constructed from a density-preserving subset of the data, as is preserved for the range counting coresets.

Estimation for stochastic combinatorial optimization. Several geometric properties of a set of stochastic points have been studied extensively in the literature under the term *stochastic geometry*. For instance, Bearwood et al. [21] showed that if there are n points uniformly and independently distributed in $[0, 1]^2$, the minimal traveling salesman tour visiting them has an expected length $\Omega(\sqrt{n})$. Asymptotic results for minimum spanning trees and minimum matchings on n points uniformly distributed in unit balls are established by Bertsimas and van Ryzin [23]. Similar results can be found in e.g., [22, 73, 98]. Compared with results in stochastic geometry, we focus on the efficient computation of the statistics, instead of giving explicit mathematical formulas.

[1] The k-median/k-means problem in the existential uncertainty model can be considered as a weighted k-median/k-means problem.

Recently, a number of researchers have begun to explore geometric computing under uncertainty and many classical computational geometry problems have been studied in different stochastic/uncertainty models. Agarwal, Cheng, Tao and Yi [3] studied the problem of indexing probabilistic points with continuous distributions for range queries on a line. Agarwal, Efrat, Sankararaman, and Zhang [5] also studied the same problem in the locational uncertainty model under Euclidean metric. The most probable k-nearest neighbor problem and its variants have attracted a lot of attentions in the database community (See e.g., [28]). Several other problems have also been considered recently, such as computing the expected volume of a set of probabilistic rectangles in a Euclidean space [109], convex hulls [6], skylines (Pareto curves) over probabilistic points [2, 15], and shape fitting [83].

Kamousi, Chan and Suri [71] initiated the study of estimating the expected length of combinatorial objects in this model. They showed that computing the expected length of the nearest neighbor (NN) graph, the Gabriel graph (GG), the relative neighborhood graph (RNG), and the Delaunay triangulation (DT) can be solved exactly in polynomial time, while computing $\mathbb{E}[\mathsf{MST}]$ is #P-hard and there exists a simple FPRAS for approximating $\mathbb{E}[\mathsf{MST}]$ in the existential model. They also gave a deterministic PTAS for approximating $\mathbb{E}[\mathsf{MST}]$ in an Euclidean plane. In another paper [72], they studied the closest pair and (approximate) nearest neighbor problems (i.e., finding the point with the smallest expected distance from the query point) in the same model.

The computational/algorithmic aspects of stochastic geometry have also gained a lot of attention in recent years from the area of wireless networking. In many application scenarios, it is common to assume that the nodes (e.g., sensors) are deployed randomly across a certain area, thereby forming a stochastic network. It is of central importance to study various properties in this network, such as connectivity [55], transmission capacity [56]. We refer the interested reader to a recent survey [57] for more references.

Other stochastic models. Besides the stochastic geometry models, geometric uncertain data has also been studied in the *imprecise* model [16, 64, 76, 84, 89, 92, 104].

In this model, each point is provided with a region where it might be. This originated with the study of imprecision in data representation [54, 94], and can be used to provide upper and lower bounds on several geometric constructs such as the diameter, convex hull, and flow on terrains [36, 104].

Convex hulls have been studied for uncertain points: upper and lower bounds are provided under the imprecise model [41, 85, 89, 104], distributions of circumference and volume are calculated in the locational model [69, 83], the most likely convex hull is found in the existential model in $\mathbb{R}^2$ and shown NP-hard for $\mathbb{R}^d$ for $d > 2$ and in the locational model [101], and the probability a query point is inside the convex hull [6]. As far as we know, the expected complexity of the convex hull under uncertain points has not been studied, although it has been studied [59] under other random data models.

The *randomly weighted graph* model where the edge weights are independent non-negative variables has also been studied extensively. Frieze [48] and Steele [99] showed that the expected value of the minimum spanning tree on such a graph with identically and independently distributed edges is $\zeta(3)/D$ where $\zeta(3) = \sum_{j=1}^{\infty} 1/j^3$ and D is the derivative of the distribution at 0. Alexopoulos and Jacobson [13] developed algorithms that compute the distribution of MST and the probability that a particular edge belongs to MST when edge lengths follow discrete distributions. However, the running time of their algorithms may be exponential in the worst cases. Recently, Emek, Korman and Shavitt [40] showed that computing the kth moment of a class of properties, including the diameter, radius and minimum spanning tree, admits an FPRAS for each fixed k.

Chapter 4 ε-Kernel Coresets over Stochastic Data

In this chapter, we initiate the study of constructing ε-kernel coresets for uncertain points. An ε-kernel coreset approximates the width of a point set in any direction. We consider approximating the expected width (an ε-EXP-KERNEL), as well as the probability distribution on the width (an (ε, τ)-QUANT-KERNEL) for any direction. Then combining with known techniques, we show a few applications to approximating the extent of uncertain functions, maintaining extent measures for stochastic moving points and some stochastic shape fitting problems. We first briefly introduce how to construct an ε-kernel coreset over deterministic point sets.

4.1 ε-Kernel Coresets over Deterministic Data

For a set P of n deterministic points, recall that we define $\mathbf{f}(P, \vec{u})$ to be $\mathbf{f}(P, \vec{u}) = \max_{s \in P} \langle \vec{u}, s \rangle$ for any direction $\vec{u} \in \mathbb{R}^d$ and $\omega(P, \vec{u})$ to be $\omega(P, \vec{u}) = \mathbf{f}(P, \vec{u}) + \mathbf{f}(P, -\vec{u})$. Also recall that a subset $Q \subseteq P$ is called an ε-kernel of P if for any direction $\vec{u} \in \mathbb{R}^d$, $(1-\varepsilon)\omega(P, \vec{u}) \leq \omega(Q, \vec{u}) \leq \omega(P, \vec{u})$. In fact, there exists an ε-kernel of size $O(\varepsilon^{-(d-1)/2})$ with construction time $O(n + \varepsilon^{-(d-3/2)})$, see [26, 110]. In the following, we briefly review the construction.

By Barequet and Har-Peled [19], we first compute a bounding box B in linear time, which has volume at most constant times of the minimum one. By their construction, we also have the property that $\alpha B \subseteq \mathsf{ConvH}(B) \subseteq B$. Here, $\mathsf{ConvH}(B)$ is the convex hull of B and $\alpha > 0$ is some constant depending on d. By applying an affine transformation, we assume that $B = [-1, 1]^d$ without loss of generality. This is because if $M(S)$ is an ε-kernel of $M(P)$ for a non-singular matrix M, then S must be an ε-kernel of P. Since $\alpha B \subseteq \mathsf{ConvH}(B)$, we have $\omega(P, \vec{u}) \geq 2\alpha$ for any direction $\vec{u}$.

Let $\varepsilon' = \sqrt{\varepsilon\alpha}$. We first construct a set $\mathcal{J}$ of size $O(\varepsilon'^{-d+1}) = O(\varepsilon^{-(d-1)/2})$ belonging to the sphere of radius $\sqrt{d}+1$ centered at the origin, satisfying that for any point x on this sphere, there exists a point $y \in \mathcal{J}$ such that $\|x-y\| \leq \varepsilon'$. Here, $\|x-y\|$ is the Euclidean distance between x and y. Next, for each point $y \in \mathcal{J}$, we compute a point $\phi(y) \in P$ which minimizes the distance $\|y - \phi(y)\|$. [1] The output is the collection S of all such $\phi(y)$, which is an ε-kernel of P. Note that the size of S is at most $O(\varepsilon^{-(d-1)/2})$.

We then briefly prove the output is an ε-kernel. Fix a direction $\vec{u}$ and let $s \in P$ be the point maximizing $\langle \vec{u}, s\rangle$. Suppose the ray starting from s in direction $\vec{u}$ intersects the sphere at point x. By the construction of $\mathcal{J}$, there exists a point $y \in \mathcal{J}$ with $\|x-y\| \leq \varepsilon'$. We then discuss the following two cases.

1. If $\phi(y) = s$, then $s \in S$ and $\mathbf{f}(P, \vec{u}) = \mathbf{f}(S, \vec{u})$.

2. If $\phi(y) \neq s$, we construct a ball of radius $\|y-s\|$ centered at y. Since $\|y-\phi(y)\| \leq \|y-s\|$ by the above algorithm, we have that $\phi(y)$ must locate inside this ball. Assume z is the point minimizing $\langle \vec{u}, z'\rangle$ over all points z' in this ball. By previous work [26, 88], it can be shown that $\langle \vec{u}, s\rangle - \langle \vec{u}, z\rangle \leq \alpha\varepsilon$. Thus, we have

$$\mathbf{f}(P, \vec{u}) - \mathbf{f}(S, \vec{u}) \leq \langle \vec{u}, s\rangle - \langle \vec{u}, \phi(y)\rangle \leq \langle \vec{u}, s\rangle - \langle \vec{u}, z\rangle \leq \alpha\varepsilon.$$

The above two cases imply that $\omega(P, \vec{u}) - \omega(S, \vec{u}) \leq 2\alpha\varepsilon \leq \varepsilon\omega(P, \vec{u})$. So the output S is indeed an ε-kernel of P.

In the above definition, we do not require the points in $\mathcal{S}$ are independent. So when they are correlated, we will specify the distribution of $\mathcal{S}$. If all points in $\mathcal{P}$ are deterministic and $\tau < 0.5$, the above definition essentially boils down to requiring $(1-\varepsilon)\omega(\mathcal{P}, \vec{u}) \leq \omega(\mathcal{S}, \vec{u}) \leq (1+\varepsilon)\omega(\mathcal{P}, \vec{u})$. Assuming the coordinates of the input points are bounded, an (ε, τ)-QUANT-KERNEL ensures that for any choice of $\vec{u}$, the cumulative distribution function of $\omega(\mathcal{S}, \vec{u})$ is within a distance ε under the Lévy

[1] In fact, we only need to compute an ε-approximate nearest-neighbor $\phi(y) \in P$ of y, which improves the running time.

metric, to that of $\omega(\mathcal{P}, \vec{u})$. [2].

4.2 ε-Kernels for Expectations of Width

First recall the definition of ε-EXP-KERNEL. Suppose $\mathcal{P}$ is a set of stochastic points (in either the existential or locational uncertainty model). Define the *expected* directional width of $\mathcal{P}$ in direction $\vec{u}$ to be $\omega(\mathcal{P}, \vec{u}) = \mathbb{E}_{P\sim\mathcal{P}}[\omega(P, \vec{u})]$, where $P \sim \mathcal{P}$ means that P is a (random) realization of $\mathcal{P}$.

Definition 15. *For a constant $\varepsilon > 0$, a set S of (deterministic or stochastic) points in $\mathbb{R}^d$ is called an* ε-EXP-KERNEL *of $\mathcal{P}$, if for all directions $\vec{u} \in \mathbb{R}^d$,*

$$(1 - \varepsilon)\omega(\mathcal{P}, \vec{u}) \leq \omega(S, \vec{u}) \leq \omega(\mathcal{P}, \vec{u}).$$

We first state our results in this section for the existential uncertainty model. All results can be extended to the locational uncertainty model, with slightly different bounds (essentially replacing the number of points n with the number of locations m) or assumptions. We describe the difference for locational model in the appendix.

For simplicity of exposition, we assume in this section that all points in $\mathcal{P}$ are in general positions and all p_is are strictly between 0 and 1. For any $s, s' \in \mathbb{R}^d$, we use $\langle s, s'\rangle$ to denote the usual inner product $\sum_{i=1}^{d} s_i s'_i$. For ease of notation, we write $s \succ_{\hat{s}} s'$ as a shorthand notation for $\langle s, \hat{s}\rangle > \langle s', \hat{s}\rangle$. For any $s' \in \mathbb{R}^d$, the binary relation $\succ_{s'}$ defines a total order of all vertices in $\mathcal{P}$. (Ties should be broken in an arbitrary but consistent manner.) We call this order the *canonical order of $\mathcal{P}$ with respect to s*. For any two points s and s', we use $\mathrm{d}(s, s')$ or $\|s - s'\|$ to denote their Euclidean distance. For any two sets of points, A and B, the Minkowski sum of A and B is defined as $A \oplus B := \{a + b \mid a \in A, b \in B\}$. Recall the definitions for a set P of deterministic points and a direction $\vec{u} \in \mathbb{R}^d$, the support function is $\mathbf{f}(P, \vec{u}) = \max_{s\in P}\langle \vec{u}, s\rangle$ and the *directional width* is $\omega(P, \vec{u}) = \mathbf{f}(P, \vec{u}) - \mathbf{f}(P, -\vec{u})$. The

[2] Assuming the coordinates of the input points are bounded, the requirement for an (ε, τ)-QUANT-KERNEL is in fact stronger than that of Lévy distance being no larger than ε as the former requires a multiplicative error on length, which gives better guarantee when the length is small.

support function and the directional width only depend on the convex hull of P.

Lemma 16. *Consider a set $\mathcal{P}$ of uncertain points in $\mathbb{R}^d$ (in either locational uncertainty model or existential uncertainty model). There exists a set S of deterministic points in $\mathbb{R}^d$ (which may not be a subset of $\mathcal{P}$) such that $\omega(\mathcal{P}, \vec{u}) = \omega(S, \vec{u})$ for all $\vec{u} \in \mathbb{R}^d$.*

Proof. By the definition of the expected directional width of $\mathcal{P}$, we have that

$$\omega(\mathcal{P}, \vec{u}) = \mathbb{E}_{P\sim\mathcal{P}}[\omega(P, \vec{u})] = \sum_{P\sim\mathcal{P}} \Pr[P]\Big(\mathbf{f}(P, \vec{u}) + \mathbf{f}(P, -\vec{u})\Big).$$

Consider the Minkowski sum $M = M(\mathcal{P}) := \sum_{P\sim\mathcal{P}} \Pr[P]\mathsf{ConvH}(P)$, where $\mathsf{ConvH}(P)$ is the convex hull of P (including the interior). It is well known that the Minkowski sum of a set of convex sets is also convex. Moreover, it also holds that for all $\vec{u} \in \mathbb{R}^d$ (see e.g., [95]) $\mathbf{f}(M, \vec{u}) = \sum_{P\sim\mathcal{P}} \Pr[P]\mathbf{f}(P, \vec{u})$. Hence, $\omega(\mathcal{P}, \vec{u}) = \omega(M, \vec{u})$ for all $\vec{u} \in \mathbb{R}^d$. □

By the result in [7], we know that for any convex body in $\mathbb{R}^d$, there exists an ε-kernel of size $O(\varepsilon^{-(d-1)/2})$. Combining with Lemma 16, we can immediately obtain the following corollary, which is the first half of Theorem 2.

Corollary 17. *For any $\varepsilon > 0$, there exists an ε-*EXP-KERNEL *of size $O(\varepsilon^{-(d-1)/2})$.*

Recall that in Lemma 16, the Minkowski sum $M = \sum_{P\sim\mathcal{P}} \Pr[P]\mathsf{ConvH}(P)$. Since M is the Minkowski sum of exponential many convex polytopes, so M is also a convex polytope. At first sight, the complexity of M (i.e., number of vertices) could be exponential. However, as we will show shortly, the complexity of M is in fact polynomial.

We need some notations first. For each pair (s, s') of points in $\mathcal{P}$ consider the hyperplane $H_{s,s'}$ that passes through the origin and is orthogonal to the line connecting s and s'. We call these $\binom{n}{2}$ hyperplanes *the separating hyperplanes induced by $\mathcal{P}$* and use Γ to denote the set. Each such hyperplane divides $\mathbb{R}^d$ into 2 halfspaces. For all directions $\vec{u} \in \mathbb{R}^d$ in each halfspace, the order of $\langle s, \vec{u}\rangle$ and $\langle s', \vec{u}\rangle$ is the same (i.e., we

have $s \succ_u s'$ in one halfspace and $s' \succ_u s$ in the other). Those hyperplanes in Γ pass through the origin and thus partition $\mathbb{R}^d$ into d-dimensional polyhedral cones. [3] We denote this *arrangement* as $\mathbb{A}(\Gamma)$.

Consider an arbitrary cone $C \in \mathbb{A}(\Gamma)$. Let $\operatorname{int} C$ denote the interior of C. We can see that for all directions $\vec{u} \in \operatorname{int} C$, the canonical order of $\mathcal{P}$ with respect to $\vec{u}$ is the same (since all directions $\vec{u} \in \operatorname{int} C$ lie in the same set of halfspaces). We use $|M|$ to denote the complexity of M, i.e., the number of vertices in $\mathsf{ConvH}(M)$.

Lemma 18. *Assuming the existential model and $p_i \in (0,1)$ for all $s \in \mathcal{P}$, the complexity of M is the same as the cardinality of $\mathbb{A}(\Gamma)$, i.e., $|M| = |\mathbb{A}(\Gamma)|$. Moreover, each cone $C \in \mathbb{A}(\Gamma)$ corresponds to exactly one vertex s of $\mathsf{ConvH}(M)$ in the following sense: the gradient $\nabla \mathbf{f}(M, \vec{u}) = s$ for all $\vec{u} \in \operatorname{int} C$ (note that here s should be understood as a vector).*

Proof. We have shown that M is a convex polytope. We first note that the support function uniquely defines a convex body (see e.g., [95]). We need the following well known fact in convex geometry (see e.g., [49]): For any convex polytope M, $\mathbb{R}^d$ can be divided into exactly $|M|$ polyhedral cones (of dimension d, ignoring the boundaries), such that each such cone C_s corresponds to a vertex s of M, and for each vector $\vec{u} \in C_s$, it holds $\mathbf{f}(M, \vec{u}) = \langle \vec{u}, s \rangle$ (i.e., the maximum of $\mathbf{f}(M, \vec{u}) = \max_{s' \in M} \langle \vec{u}, s' \rangle$ is achieved by s for all $\vec{u} \in C_s$).[4] See Figure 4-1 for an example in $\mathbb{R}^2$. Hence, for each $\vec{u} \in \operatorname{int} C_s$ the gradient of of the support function (as a function of $\vec{u}$) is exactly s:

$$\nabla \mathbf{f}(M, \vec{u}) = \left\{ \frac{\partial \mathbf{f}(M, \vec{u})}{\partial \vec{u}_j} \right\}_{j \in [d]} = \left\{ \frac{\partial \langle \vec{u}, s \rangle}{\partial \vec{u}_j} \right\}_{j \in [d]} = \left\{ \frac{\partial \sum_{j \in [d]} s_j \vec{u}_j}{\partial \vec{u}_j} \right\}_{j \in [d]} = s, \quad (4.1)$$

where $\vec{u}_j$ is the jth coordinate of $\vec{u}$. With a bit abuse of notation, we denote the set of cones defined above by $\mathbb{A}(M)$.

Now, consider a cone $C \in \mathbb{A}(\Gamma)$. We show that for all $\vec{u} \in \operatorname{int} C$, $\nabla \mathbf{f}(M, \vec{u})$ is a distinct constant vector independent of $\vec{u}$. In fact, we know that $\mathbf{f}(M, \vec{u}) = \mathbf{f}(\mathcal{P}, \vec{u}) =$

[3] We ignore the lower dimensional cells in the arrangement.

[4] One intuitive way to see this is as follows: The support function for a polytope is just the upper envelope of a finite set of linear functions, thus a piecewise linear function, and the domain of each piece is a polyhedral cone. In fact, we call such a cone C_s the outer normal cones.

$\sum_{s\in\mathcal{P}} \Pr^R(s,\vec{u})\langle s,\vec{u}\rangle$, where $\Pr^R(s,\vec{u}) = \prod_{s'\succ_{\vec{u}} s}(1-p_{s'})p_s$. For all $\vec{u} \in \operatorname{int} C$, the $\Pr^R(s,\vec{u})$ value is the same since the value only depends on the canonical order with respect to $\vec{u}$, which is the same for all $\vec{u} \in C$. Hence, we can get that for all $\vec{u} \in \operatorname{int} C$,

$$\nabla \mathbf{f}(M,\vec{u}) = \sum_{s\in\mathcal{P}} \Pr^R(s,\vec{u})s, \tag{4.2}$$

which is a constant independent of $\vec{u}$. We prove the lemma by showing that the gradient $\nabla\mathbf{f}(M,\vec{u})$ must be different for two adjacent cones C_1, C_2 (separated by some hyperplane in Γ) in $\mathbb{A}(\Gamma)$. Suppose $\vec{u}_1 \in \operatorname{int} C_1$ and $\vec{u}_2 \in \operatorname{int} C_2$. Consider the canonical orders O_1 and O_2 of $\mathcal{P}$ with respect to $\vec{u}_1$ and $\vec{u}_2$ respectively. Since C_1 and C_2 are adjacently, O_1 and O_2 only differ by one swap of adjacent vertices. W.l.o.g., assume that $O_1 = \{s_1,\ldots,s_i,s_{i+1},\ldots,s_n\}$ and $O_1 = \{s_1,\ldots,s_{i+1},s_i,\ldots,s_n\}$. Using (4.2), we can get that

$$\begin{aligned}
\nabla\mathbf{f}(M,\vec{u}_1) - \nabla\mathbf{f}(M,\vec{u}_2) &= \Pr^R(s_i,\vec{u}_1)s_i + \Pr^R(s_{i+1},\vec{u}_1)s_{i+1} - \Pr^R(s_i,\vec{u}_2)s_i - \Pr^R(s_{i+1},\vec{u}_2)s_{i+1}\\
&= D\cdot(p_is_i + (1-p_i)p_{i+1}s_{i+1} - p_{i+1}s_{i+1} - (1-p_{i+1})p_is_i)\\
&= D\cdot p_ip_{i+1}(s_i - s_{i+1}) \neq 0
\end{aligned}$$

where $D = \prod_{j=1}^{i-1}(1-p_j) \neq 0$.

In summary, we have shown in the first paragraph that $\nabla\mathbf{f}(M,\vec{u})$ is piecewise constant, with a distinct constant in each cone in $\mathbb{A}(M)$. The same also holds for $\mathbb{A}(\Gamma)$. This is only possible if $\mathbb{A}(\Gamma)$ (thinking as a partition of $\mathbb{R}^d$) partitions $\mathbb{R}^d$ exactly the same way as $\mathbb{A}(M)$ does. Hence, we have $\mathbb{A}(\Gamma) = \mathbb{A}(M)$ and the lemma follows immediately. □

Since $O(n^2)$ hyperplanes passing through the origin can divide $\mathbb{R}^d$ into at most $O(\binom{n^2}{d-1})$ d-dimensional polyhedral cones (see e.g., [12]), we immediately obtain the following corollary.

Corollary 19. *It holds that* $|M| \leq O(\binom{n^2}{d-1}) = O(n^{2d-2})$.

The proof of Lemma 18 can be easily made constructive. We only need to compute

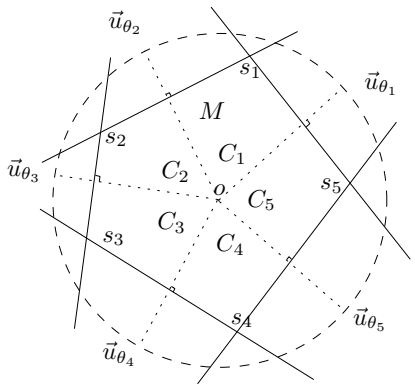

Figure 4-1: The figure depicts a pentagon M in $\mathbb{R}^2$ to illustrate some intuitive facts in convex geometry. (1) The plane can be divided into 5 cones $C_1, \ldots, C_5$, by 5 angles $\theta_1, \ldots, \theta_5$. $\vec{u}_{\theta_i}$ is the unit direction corresponding to angle θ_i. Each cone C_i corresponds to a vertex s_i and for any direction $\vec{u} \in C_i$, $\mathbf{f}(M, \vec{u}) = \langle \vec{u}, s_i \rangle$ and the vector $\nabla \mathbf{f}(M, \vec{u})$ is s_i. (2) Each direction θ_i is perpendicular to an edge of M. $M = \cap_{i=1}^{5} H_i$ where H_i is the supporting halfplane with normal vector $\vec{u}_{\theta_i}$.

the set Γ of all $O(n^2)$ hyperplanes and the arrangement $\mathbb{A}(\Gamma)$ in $O(n^{2d-2})$ time (see e.g., [12, 39]). Given each cone $C \in \mathbb{A}(\Gamma)$, we can calculate $\nabla \mathbf{f}(M, \vec{u})$ for any $\vec{u} \in C$, which gives exactly one vertex of M by (4.1), in $O(n \log n)$ time using the algorithm described in Lemma 49.

Theorem 20. *In $\mathbb{R}^d$ for constant d, the polytope M which defines $\mathbf{f}(\mathcal{P}, \vec{u})$ for any direction $\vec{u}$ can be described with $O(n^{2d-2})$ vertices in $\mathbb{R}^d$, and can be computed in $O(n^{2d-1} \log n)$ time. In $\mathbb{R}^2$, the runtime can be improved to $O(n^2 \log n)$.*

The improved running time in $\mathbb{R}^2$ is derived in Lemma 50 by carefully constructing each vertex of M in $O(1)$ time using its neighboring vertex. The extra $O(\log n)$ is needed to sort the vertices of M to determine neighbors.

4.2.1 A Nearly Linear Time Algorithm for Constructing ε-exp-kernels

Now, we prove the main algorithmic result Theorem 2 of this section: we can find an ε-EXP-KERNEL in nearly linear time. If we already have the Minkowski sum M, we can directly use the algorithm in [7] to find an ε-kernel for M. However, constructing M explicitly takes $O(n^{2d-1} \log n)$ time according to Theorem 20 and this cannot be improved in general as the complexity of M is $O(n^{2d-2})$. Therefore, in order to

achieve a nearly linear time coreset construction, we can not compute M explicitly. For ease of description, we first consider existential uncertainty model. The details for locational uncertainty model can be found in Section 4.6.

Theorem 21. *(second half of Theorem 2, for existential model) $\mathcal{P}$ is a set of n uncertain points in $\mathbb{R}^d$ with existential uncertainty. An ε-*EXP-KERNEL *of size $O(\varepsilon^{-(d-1)/2})$ for $\mathcal{P}$ can be constructed in $O(\varepsilon^{-(d-1)} n \log n)$ time.*

The following simple lemma provides an efficient procedure for finding the extreme vertex in M along any give direction, and is useful in several places later as well.

Lemma 22. *Given any direction $\vec{u} \in \mathbb{R}^d$, we can find in $O(n \log n)$ time a vertex $s^\star \in M$, at which $\langle s, \vec{u} \rangle$ is maximized, over all $s \in M$.*

Proof. Fix an arbitrary direction $\vec{u} \in \mathbb{R}^d$. From the proof of Lemma 18 (in particular (4.1)), we know that the vertex $s^\star \in M$ that maximizes $\langle s, \vec{u} \rangle$ can be computed by $s^\star = \nabla \mathbf{f}(M, \vec{u})$. Using (4.2), $\nabla \mathbf{f}(M, \vec{u})$ can be easily computed in $O(n \log n)$ time (see Lemma 49 for the details). □

Next, we need to find an affine transform T such that the convex polytope $M' = T(M)$ is α-fat for some constant α. We recall that a set P of points is *α-fat*, for some constant $\alpha \leq 1$, if there exists a point $x \in \mathbb{R}^d$, and a unit hypercube $\overline{\mathbb{C}}$ centered at x such that $\alpha\overline{\mathbb{C}} \subset \mathsf{ConvH}(P) \subset \overline{\mathbb{C}}$. According to Chapter 22 in [58], in order to construct such T, it suffices to identify two points in M such that their distance is a constant approximation of the diameter of M. The following lemma (proven in Section 4.6) shows this can be done without computing M explicitly.

Lemma 23. *We find an affine transform T in $O(2^{O(d)} n \log n)$ time, such that the convex polytope $M' = T(M)$ is α-fat for some constant α (α may depend on d).*

After obtaining T, we apply T to $\mathcal{P}$ in linear time. Notice that $M' = T(M(\mathcal{P})) = M(T(\mathcal{P}))$. Therefore, Lemma 22 also holds for M' (i.e., we can search over M' the maximum vertex in any given direction in $O(n \log n)$ time).

Let $\delta = O(\varepsilon\alpha/d)$. We compute a set $\mathcal{I}$ of $O(\delta^{-(d-1)}) = O(\varepsilon^{-(d-1)})$ points on the unit sphere $\mathbb{S}^{d-1}$ such that for any point $\hat{s} \in \mathbb{S}^{d-1}$, there is a point $s \in \mathcal{I}$ such that

$\|s - s'\| \leq \delta$ (see e.g., [10, 25]). For each s in $\mathcal{I}$, we include $-s$ in $\mathcal{I}$ as well. For each vector $s \in \mathcal{I}$, we compute $x(t) = \arg\max_{x \in M'} \langle x, s \rangle$. Based on the previous discussion, all $\{x(s)\}_{s \in \mathcal{I}}$ can be computed in $O(\delta^{-(d-1)} n \log n) = O(\varepsilon^{-(d-1)} n \log n)$ time.

Lemma 24. *$S = \{x(s)\}_{s \in \mathcal{I}}$ is an ε-kernel for M'.* [5]

Finally, we can then run existing ε-kernel algorithms [26, 88] in $O(|S|)$ time to further reduce the size of S to $O(\varepsilon^{-(d-1)/2})$, which finishes the proof of Theorem 21. Lemma 18 and Theorem 21 hold for locational uncertainty models as well. The details can be found in Section 4.6.

4.2.2 ε-exp-kernel Under the Subset Constraint

First, we show that under the subset constraint (i.e., the ε-EXP-KERNEL is required to be a subset of the original point set, with the same probability distribution for each chosen point), there exists no ε-EXP-KERNEL with small size in general. [6]

Lemma 25. *For some constant $\varepsilon > 0$, there exist a set $\mathcal{P}$ of stochastic points such that no $o(n)$ size ε-EXP-KERNEL exists for $\mathcal{P}$ under the subset constraint (for both locational model and existential model).*

Proof. To see this in the existential uncertainty model, simply consider n points, each with existence probability $1/n$. $n/2$ of them co-locate at the origin and the other $n/2$ of them co-locate at $x = 1$. It is not hard to see that the expected length of the diameter is $\Omega(1)$ but the expected length of the diameter of any $o(n)$ size subset is only $o(1)$ (with high probability, no point would even appear).

The case for the locational uncertainty model is as simple. Again, consider n points. For each point, with probability $1/n$, it appears at $x = 1$. Otherwise, its position is the origin (with probability $1 - 1/n$). It is not hard to see that the expected length of the diameter of the original point set is $\Omega(1)$, while that of any $o(n)$ size subset is only $o(1)$ (with high probability, no point would realize at $x = 1$). □

[5] This is a folklore result. A proof of the 2D case can be found in [31]. The general case is a straightforward extension and we provide a proof in Section 4.6 for completeness.

[6] If we require the ε-EXP-KERNEL to be a subset of the original point set, but with possibly different probability for each chosen point, we do not know whether there always exists an ε-EXP-KERNEL with small size.

In light of the above negative result, we make the following β-*assumption*: we assume each possible location realizes a point with probability at least β, for a constant $\beta > 0$. The proof of the following theorem can be found in Section 4.6.

Theorem 26. *Under the β-assumption, in the existential uncertainty model, there is an ε-*EXP-KERNEL *in $\mathbb{R}^d$ of size $O(\beta^{-(d-1)}\varepsilon^{-(d-1)/2}\log(1/\varepsilon))$ that satisfies the subset constraint.*

4.3 ε-Kernels for Probability Distributions of Width

Recall $\mathcal{S}$ is an (ε,τ)-QUANT-KERNEL if for all $x \geq 0$, $\Pr_{P\sim\mathcal{P}}\Big[\omega(P,\vec{u}) \leq (1-\varepsilon)x\Big] - \tau \leq \Pr_{P\sim\mathcal{S}}\Big[\omega(S,\vec{u}) \leq x\Big] \leq \Pr_{P\sim\mathcal{P}}\Big[\omega(P,\vec{u}) \leq (1+\varepsilon)x\Big] + \tau$. For ease of notation, we sometimes write $\Pr\big[\omega(\mathcal{P},\vec{u}) \leq t\big]$ to denote $\Pr_{P\sim\mathcal{P}}\big[\omega(P,\vec{u}) \leq t\big]$, and abbreviate the above as $\Pr\Big[\omega(S,\vec{u}) \leq x\Big] \in \Pr\Big[\omega(\mathcal{P},\vec{u}) \leq (1\pm\varepsilon)x\Big] \pm \tau$. We first provide a simple linear time algorithm for constructing an (ε,τ)-QUANT-KERNEL for both existential and locational models, in Section 4.3.1. The points in the constructed kernel are not independent. Then, for existential models, we provide a nearly linear time (ε,τ)-QUANT-KERNEL construction where all stochastic points in the kernel are independent in Section 4.3.2.

4.3.1 A Simple (ε,τ)-quant-kernel Construction

In this section, we show a linear time algorithm for constructing an (ε,τ)-QUANT-KERNEL for any stochastic model if we can sample a realization from the model in linear time (which is true for both locational and existential uncertainty models).

Algorithm:coreset. Let $N = O\big(\tau^{-2}\varepsilon^{-(d-1)}\log\frac{1}{\varepsilon}\big)$. We sample N independent realizations from the stochastic model. Let $\mathcal{H}_i$ be the convex hull of the present points in the ith realization. For $\mathcal{H}_i$, we use the algorithm in [7] to find a deterministic ε-kernel $\mathcal{E}_i$ of size $O(\varepsilon^{-(d-1)/2})$. Our (ε,τ)-QUANT-KERNEL $\mathcal{S}$ is the following simple stochastic model: with probability $1/N$, all points in $\mathcal{E}_i$ are present. Hence, $\mathcal{S}$ consists of $O\big(\tau^{-2}\varepsilon^{-3(d-1)/2}\log\frac{1}{\varepsilon}\big)$ points (two such points either co-exist or are mutually ex-

clusive). Hence, for any direction $\vec{u}$, $\Pr[\omega(\mathcal{S}, \vec{u}) \leq t] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\omega(\mathcal{E}_i, \vec{u}) \leq t)$, where $\mathbb{I}(\cdot)$ is the indicator function.

For a realization $P \sim \mathcal{P}$, we use $\mathcal{E}(P)$ to denote the deterministic ε-kernel for P. So, $\mathcal{E}(P)$ is a random set of points, and we can think of $\mathcal{E}_1, \ldots, \mathcal{E}_N$ as samples from the random set. Now, we show $\mathcal{S}$ is indeed an (ε, τ)-QUANT-KERNEL. We start with the following simple observation.

Observation 27. *For any $t \geq 0$ and any direction $\vec{u}$, we have that*

$$\Pr[\omega(\mathcal{P}, \vec{u}) \leq t] \leq \Pr_{P\sim\mathcal{P}}[\omega(\mathcal{E}(P), \vec{u}) \leq t] \leq \Pr[\omega(\mathcal{P}, \vec{u}) \leq (1+\varepsilon)t].$$

Proof. For any realization P of $\mathcal{P}$, we have $\frac{1}{1+\varepsilon}\omega(P, \vec{u}) \leq \omega(\mathcal{E}(P), \vec{u}) \leq \omega(P, \vec{u})$. The observation follows by combining all realizations. □

We only need to show that $\mathcal{S}$ is an (ε, τ)-QUANT-KERNEL for $\mathcal{E}(P)$. We need the following two theorems.

Theorem 28. *(Theorem 5.22 in [58]) (VC-dimension) Let $S_1 = (X, \mathcal{R}^1), \ldots, S_k = (X, \mathcal{R}^k)$ be range spaces with VC-dimension $\delta_1, \ldots, \delta_k$, respectively. Next, let $\mathbf{f}(r_1, \ldots, r_k)$ be a function that maps any k-tuple of sets $r_1 \in \mathcal{R}^1, \ldots, r_k \in \mathcal{R}^k$ into a subset of X. Consider the range set*

$$\mathcal{R}' = \{\mathbf{f}(r_1, \ldots, r_k) \mid r_1 \in \mathcal{R}^1, \ldots, r_k \in \mathcal{R}^k\}$$

and the associated range space $(X, \mathcal{R}')$. Then, the VC-dimension of $(X, \mathcal{R}')$ is bounded by $O(k\delta \log k)$, where $\delta = \max_i \delta_i$.

Suppose $(X, \mathcal{R})$ is a range space and μ is a probability measure over X. We say a subset $C \subset X$ is an *ε-approximation* of the range space if for any range $R \in \mathcal{R}$, we have $\Big|\mu_C(R) - \mu(R)\Big| \leq \varepsilon$, where $\mu_C(R) = |C \cap R|/|C|$. We need the following celebrated uniform convergence result, first established by Vapnik and Chervonenkis [105].

Theorem 29. *(See Theorem 4.9 in [14]) Suppose $(X, \mathcal{R})$ is any range space with VC-dimension at most V, where $|X|$ is finite and μ is a probability measure defined*

over X. For any $\varepsilon, \delta > 0$, a random subset $C \subseteq X$ (according to μ) of cardinality $s = O\left(\varepsilon^{-2}\left(V + \log(1/\delta)\right)\right)$ is an ε-approximation for X with probability $1 - \delta$.

Now, we are ready to prove the main lemma in this section.

Lemma 30. *Let $N = O(\tau^{-2}\varepsilon^{-(d-1)}\log(1/\varepsilon))$. For any $t \geq 0$ and any direction $\vec{u}$,*

$$\Pr[\omega(\mathcal{S}, \vec{u}) \leq t] \in \Pr_{P\sim\mathcal{P}}[\omega(\mathcal{E}(P), \vec{u}) \leq t] \pm \tau.$$

Proof. Let $L = O(\varepsilon^{-(d-1)/2})$. We first note that $\mathcal{E}(P)$ has at most n^L possible realizations since each ε-kernel is of size at most L, We first build a mapping g that maps each realization $\mathcal{E}(P)$ to a point in $\mathbb{R}^{dL}$, as follows: Consider a realization P of $\mathcal{P}$. Suppose $\mathcal{E}(P) = \{(x_1^1, \ldots, x_d^1), \ldots, (x_1^L, \ldots, x_d^L)\}$ (if $|\mathcal{E}(P)| < L$, we pad it with $(0, \ldots, 0)$). We let

$$g(\mathcal{E}(P)) = (x_1^1, \ldots, x_d^1, \ldots, x_1^L, \ldots, x_d^L) \in \mathbb{R}^{dL}.$$

For any $t \geq 0$ and any direction $\vec{u} \in \mathbb{R}^d$, note that $\omega(\mathcal{E}(P), \vec{u}) \geq t$ holds if and only if there exists some $1 \leq i, j \leq |\mathcal{E}(P)|, i \neq j$ satisfies that $\sum_{k=1}^{d}(x_k^i - x_k^j)\vec{u}_k \geq t$, which is equivalent to saying that point $g(\mathcal{E}(P))$ is in the union of those $O(|\mathcal{E}(P)|^2)$ halfspaces (for each i, j, we have one such halfspace).

Let X be the image set of g. Let $(X, \mathcal{R}^{i,j})$ $(1 \leq i, j \leq L, i \neq j)$ be a range space, where $\mathcal{R}^{i,j}$ is the set of halfspaces $\{\vec{u} = (\vec{u}_1, \ldots, \vec{u}_d) \in \mathbb{R}^d \mid \sum_{k=1}^{d}(x_k^i - x_k^j)\vec{u}_k \geq t\}$. Let $\mathcal{R}' = \{\cup r_{i,j} \mid r_{i,j} \in \mathcal{R}^{i,j}, i, j \in [L]\}$. Note that each $(X, \mathcal{R}^{i,j})$ has VC-dimension $d + 1$. By Theorem 28, we can see that the VC-dimension of $(X, \mathcal{R}')$ is bounded by $O((d+1)L^2 \lg L^2) = O(\varepsilon^{-(d-1)}\log(1/\varepsilon))$. Notice that $\mathcal{S} = \{\mathcal{E}_1, \ldots, \mathcal{E}_N\}$ is a collection of samples from $\mathcal{E}(P)$. Hence, by Theorem 29, for any t and any direction $\vec{u}$, we have that $\Pr[\omega(\mathcal{S}, \vec{u}) \leq t] \in \Pr_{P\sim\mathcal{P}}[\omega(\mathcal{E}(P), \vec{u}) \leq t] \pm \tau$. □

Combining Observation 27 and Lemma 30, we obtain the following theorem.

Theorem 31. *Let $N = O(\tau^{-2}\varepsilon^{-(d-1)}\log(1/\varepsilon))$. For any $t \geq 0$ and any direction $\vec{u}$, we have that*

$$\Pr[\omega(\mathcal{S}, \vec{u}) \leq t] \in \Pr[\omega(\mathcal{P}, \vec{u}) \leq (1 \pm \varepsilon)t)] \pm \tau.$$

Running time. In each sample, the size of an ε-kernel $\mathcal{K}_i$ is at most $O\big(\varepsilon^{-(d-1)/2}\big)$. Note that we can compute $\mathcal{K}_i$ in $O(n+\varepsilon^{-(d-3/2)})$ time [26, 88]. We take $O\big(\tau^{-2}\varepsilon^{-(d-1)}\log(1/\varepsilon)\big)$ samples in total. So the overall running time is $O\big(n\tau^{-2}\varepsilon^{-(d-1)}\log(1/\varepsilon)+\mathrm{poly}(\frac{1}{\varepsilon\tau})\big) = \widetilde{O}\left(n\tau^{-2}\varepsilon^{-(d-1)}\right)$. In summary, we obtain our main result for (ε,τ)-QUANT-KERNEL in this subsection.

Theorem 4. *(restated) An (ε,τ)-QUANT-KERNEL of size $\widetilde{O}\left(\tau^{-2}\varepsilon^{-3(d-1)/2}\right)$ can be constructed in $\widetilde{O}\left(n\tau^{-2}\varepsilon^{-(d-1)}\right)$ time, under both existential and locational uncertainty models.*

4.3.2 Improved (ε,τ)-quant-kernel for Existential Models

In this section, we show an (ε,τ)-QUANT-KERNEL $\mathcal{S}$ can be constructed in nearly linear time for the existential model, and all points in $\mathcal{S}$ are independent of each other. The size bound $\widetilde{O}(\tau^{-2}\varepsilon^{-(d-1)})$ (see Theorem 5) is better than that in Theorem 4 for the general case, and the independence property may be useful in certain applications. Moreover, some of the insights developed in this section may be of independent interest (e.g., the connection to Tukey depth). Due to the independence requirement, the construction is somewhat more involved. For ease of the description, we assume the Euclidean plane first. All results can be easily extended to $\mathbb{R}^d$. We also assume that all probability values are strictly between 0 and 1 and $0 < \varepsilon, \tau \leq 1/2$ is a fixed constant.

Let $\lambda(\mathcal{P}) = \sum_{s_i\in\mathcal{P}}(-\ln(1-p_i))$. In the following, we present two algorithms. The first algorithm works for any $\lambda(\mathcal{P})$ and produces an (ε,τ)-QUANT-KERNEL $\mathcal{S}$ whose size depends on $\lambda(\mathcal{P})$. In Section 4.3.2, we present the second algorithm that only works for $\lambda(\mathcal{P}) \geq 3\ln(2/\tau)$ but produces an (ε,τ)-QUANT-KERNEL $\mathcal{S}$ with a constant size (the constant only depends on ε, τ and δ). Thus, we can get a constant size (ε,τ)-QUANT-KERNEL by running the first algorithm when $\lambda(\mathcal{P}) \leq 3\ln(2/\tau)$ and running the second algorithm otherwise.

Algorithm 1: For Any $\lambda(\mathcal{P})$

In this section, we present the first algorithm which works for any $\lambda(\mathcal{P})$. We can think of each point s associated with a Bernoulli random variable X_s that takes value 1 with probability p_s and 0 otherwise. Now, we replace the Bernoulli random variable X_s by a Poisson distributed random variable $\widetilde{X}_s$ with parameter $\lambda_s = -\ln(1-p_s)$ (denoted by $\text{Pois}(\lambda_s)$), i.e., $\Pr[\widetilde{X}_s = k] = \frac{1}{k!}\ \lambda_s^k\ e^{-\lambda_s},$ for $k = 0, 1, 2, \ldots$. Here, $\widetilde{X}_s = k$ means that there are k realized points located at the position of s. We call the new instance *the Poissonized instance corresponding to* $\mathcal{P}$. We can check that $\Pr[\widetilde{X}_s = 0] = e^{-\lambda_s} = 1 - p_s = \Pr[X_s = 0]$. Also note that co-locating points do not affect any directional width, so the Poissonized instance is essentially equivalent to the original instance for our problem.

The construction of the (ε, τ)-QUANT-KERNEL $\mathcal{S}$ is as follows: Let $\mathfrak{A}$ be the probability measure over all points in $\mathcal{P}$ defined by $\mathfrak{A}(\{s\}) = \lambda_s/\lambda$ for every $s \in \mathcal{P}$, where $\lambda := \lambda(\mathcal{P}) = \sum_{s\in\mathcal{P}} \lambda_s$. Let τ_1 be a small positive constant to be fixed later. We take $N = \tilde{O}(\tau_1^{-2})$ independent samples from $\mathfrak{A}$ (we allow more than one point to be co-located at the same position), and let $\mathfrak{B}$ be the empirical measure, i.e., each sample point having probability $1/N$. The coreset $\mathcal{S}$ consists of the N sample points in $\mathfrak{B}$, each with the same existential probability $1 - \exp(-\lambda/N)$. A useful alternative view of $\mathcal{S}$ is to think of each point associated with a random variable Y_v following distribution $\text{Pois}(\lambda/N)$ (i.e., the Poissonized instance corresponding to $\mathcal{S}$). This finishes the description of the construction.

Now, we start the analysis. Our goal is to show that $\mathcal{S}$ is indeed an (ε, τ)-QUANT-KERNEL. The following theorem is a special case of Theorem 29 (specialized to the range space consisting of all halfplanes), which shows that the empirical measure $\mathfrak{B}$ is close to the original measure $\mathfrak{A}$ with respect to all half spaces.

Theorem 32. [14, 82] *We denote the set of all halfplanes by* $\mathbb{H}$. *With probability* $1-\delta$, *the empirical measure* $\mathfrak{B}$ *(defined by* $N = O(\tau_1^{-2}\log(1/\delta))$ *independent samples) satisfies the following:*

$$\sup_{H\in\mathbb{H}} |\mathfrak{A}(H) - \mathfrak{B}(H)| \le \tau_1.$$

From now on, we assume that $\mathfrak{B}$ satisfies the statement of Theorem 32. We first observe a simple but useful lemma, which is a consequence of Theorem 32. For a halfplane H, we use $H \models 0$ to denote the event that no point is realized in H.

Lemma 33. *With probability $1-\delta$, for any halfplane $H \in \mathbb{H}$, we have that*

$$\Pr_{\mathcal{S}}\big[H \models 0\big] \in (1 \pm O(\lambda\tau_1))\Pr_{\mathcal{P}}\big[H \models 0\big].$$

Proof. Fix an arbitrary halfplane $H \in \mathbb{H}$. Consider the Poissonized instance corresponding to $\mathcal{P}$. We first observe that $\Pr_{\mathcal{P}}\big[H \models 0\big] = \Pr_{\mathcal{P}}\big[\sum_{s\in\mathcal{P}\cap H} X_s = 0\big]$. Since X_s follows distribution $\mathrm{Pois}(\lambda_s)$, $\sum_{s\in\mathcal{P}\cap H} X_s$ follows Poisson distribution $\mathrm{Pois}\big(\sum_{s\in\mathcal{P}\cap H}\lambda_s\big)$. Similarly, we have that $\Pr_{\mathcal{S}}\big[H \models 0\big] = \Pr_{\mathcal{S}}\big[\sum_{s\in\mathcal{S}\cap H} Y_v = 0\big]$ since $\sum_{s\in\mathcal{S}\cap H} Y_v$ follows $\mathrm{Pois}\big(\sum_{s\in\mathcal{S}\cap H}\lambda/N\big)$. Hence, we can see the following:

$$\begin{aligned}
\Pr_{\mathcal{P}}\big[H \models 0\big] &= \exp\big(-\sum_{s\in\mathcal{P}\cap H}\lambda_s\big) = \exp\big(-\lambda\mathfrak{A}(H)\big) \\
&\in \exp\big(-\lambda(\mathfrak{B}(H)\pm\tau_1)\big) = \exp\big(-\sum_{s\in\mathcal{S}\cap H}\frac{\lambda}{N} \pm \tau_1\lambda\big) \\
&\in (1\pm O(\lambda\tau_1))\exp\big(-\sum_{s\in\mathcal{S}\cap H}\frac{\lambda}{N}\big) = (1\pm O(\lambda\tau_1))\Pr_{\mathcal{S}}\big[H \models 0\big].
\end{aligned}$$

The first inequality follows from Theorem 32 and the second is due to the fact that $e^{-\varepsilon} \geq 1-\varepsilon$ and $e^{\varepsilon} \leq 1+(e-1)\varepsilon$ for any $0<\varepsilon<1$. □

For two real-valued random variables X, Y, we define the Kolmogorov distance $\mathrm{d}_K(X,Y)$ between X and Y to be $\mathrm{d}_K(X,Y) = \sup_{t\in\mathbb{R}}|\Pr[X\leq t] - \Pr[Y\leq t]|$. We also need the following simple lemma.

Lemma 34. *Suppose we have four independent random variables X, X', Y and Y' such that $\mathrm{d}_K(X,X') \leq \varepsilon$ and $\mathrm{d}_K(Y,Y') \leq \varepsilon$ for some $\varepsilon \geq 0$. Then, $\mathrm{d}_K(X+Y, X'+Y') \leq 2\varepsilon$.*

Proof. We need the following useful elementary fact about Kolmogorov distance: Let X, Y, Z be real-valued random variables such that X is independent of Y and independent of Z. Then we have that $\mathrm{d}_K(X+Y, X+Z) \leq \mathrm{d}_K(Y,Z)$. The rest of the proof is

straightforward: $\mathrm{d}_K(X+Y, X'+Y') \le \mathrm{d}_K(X+Y, X+Y') + \mathrm{d}_K(X+Y', X'+Y') \le 2\varepsilon$. The first inequality is the triangle inequality. □

Now, we are ready to show that $\mathcal{S}$ is really an (ε, τ)-QUANT-KERNEL. We note that in this subsection our bound is stronger than (**??**) in that we do not need to relax the length threshold. We first prove the theorem under a simplified assumption: we assume that there is a point $s^\star \in \mathbb{R}^2$ (not necessarily an input point), which we call the special point, that lies in the convex hull of $\mathcal{P}$ with probability at least $1 - \delta/2$. With the assumption, the proof is much simpler but still instructive as the analysis in Section 4.3.2 is an extension of this proof. The general case is proved in Theorem 36 and the proof is more technical and the size bound is slightly worse.

Theorem 35. *Assume that there is a special point $s^\star \in \mathbb{R}^2$ that lies in the convex hull of $\mathcal{P}$ with probability at least $1 - \delta/2$. The parameters of the algorithm are set as*

$$\tau_1 = O\Big(\frac{\tau}{\lambda}\Big) \quad \text{and} \quad N = O\Big(\frac{1}{\tau_1^2}\log\frac{1}{\delta}\Big) = O\Big(\frac{\lambda^2}{\tau^2}\log\frac{1}{\delta}\Big).$$

With probability at least $1 - \delta$, for any $t \ge 0$ and any direction $\vec{u}$, we have that

$$\Pr\Big[\omega(\mathcal{S}, \vec{u}) \le t\Big] \in \Pr\Big[\omega(\mathcal{P}, \vec{u}) \le t\Big] \pm \tau. \tag{4.3}$$

Proof. We first condition on the event that $s^\star$ is in the convex hull of all realized points (which happens with probability at least $1 - \delta/2$). The remainder needs to hold with probability at least $1 - \delta/2$. Under the condition, we can pretend that $s^\star$ is a deterministic point in the original point set (this does not affect any directional width as $s^\star$ is in the convex hull).

Fix an arbitrary direction $\vec{u}$ (w.l.o.g., say it is the x-axis). Rename all points as $s_1, s_2, \ldots, s_n$ according to the increasing order of their projections to $\vec{u}$. Suppose $s^\star$ is renamed as s_k. Let the random variable L be the directional width of $\{s_1, \ldots, s_k\}$ with respect to $\vec{u}$ and R be the directional width of $\{s_k, \ldots, s_n\}$ with respect to $\vec{u}$. Since $s^\star$ is assumed to be within the left and right extents, we can easily see that $\omega(\mathcal{P}, \vec{u}) = L + R$. Similarly, we define L' (R' resp.) to be the directional width of all

points in $\mathcal{S}$ to the left (right resp.) of $s^\star$. Since the convex hull of $\mathcal{S}$ contains $s^\star$, we can also see that $\omega(\mathcal{S}, \vec{u}) = L' + R'$. By Lemma 33, we know that $\mathrm{d}_K(L, L') \leq O(\lambda\tau_1)$ and $\mathrm{d}_K(R, R') \leq O(\lambda\tau_1)$. By Lemma 34, we have that $\mathrm{d}_K(\omega(\mathcal{S}, \vec{u}), \omega(\mathcal{P}, \vec{u})) \leq O(\lambda\tau_1)$. Let $\tau_1 = O(\tau/\lambda)$, the theorem follows. □

Now, we prove the theorem in the general case, where the main difficulty comes from the fact that we can not separate the width into two independent parts L and R. The proof is somewhat technical and can be found in Section 4.7.

Theorem 36. *Let* $\tau_1 = O(\frac{\tau}{\max\{\lambda,\lambda^2\}})$ *and* $N = O(\frac{1}{\tau_1^2}\log\frac{1}{\delta}) = O(\frac{\max\{\lambda^2,\lambda^4\}}{\tau^2}\log\frac{1}{\delta})$. *With probability at least* $1-\delta$, *for any* $t \geq 0$ *and any direction* $\vec{u}$, *we have that* $\Pr\Big[\omega(\mathcal{S}, \vec{u}) \leq t\Big] \in \Pr\Big[\omega(\mathcal{P}, \vec{u}) \leq t\Big] \pm \tau$.

Algorithm 2: For $\lambda(\mathcal{P}) > 3\ln(2/\tau)$

In the second algorithm, we assume that $\lambda(\mathcal{P}) = \sum_{s\in\mathcal{P}} \lambda_s > 3\ln(2/\tau)$. When $\lambda(\mathcal{P})$ is large, we cannot directly use the sampling technique in the previous section since it requires a large number of samples. However, the condition $\lambda(\mathcal{P}) \geq 3\ln(2/\tau)$ implies there is a nonempty convex region $\mathcal{K}$ inside the convex hull of $\mathcal{P}$ with high probability. Moreover, we can show the sum of λ_s values in $\overline{\mathcal{K}} = \mathbb{R}^2 \setminus \mathcal{K}$ is small. Hence, we can use the sampling technique just for $\overline{\mathcal{K}}$ and use the deterministic ε-kernel construction for $\mathcal{K}$.

Now, we describe the details of our algorithm. Again consider the Poissonized instance of $\mathcal{P}$. Imagine the following process. Fix a direction $\vec{u} \in \mathbb{S}^1$. [7] We move a sweep line $\ell_{\vec{u}}$ orthogonal to $\vec{u}$, along the direction $\vec{u}$, to sweep through the points in $\mathcal{P}$. We use $H_{\vec{u}}$ to denote the halfplane defined by $\ell_{\vec{u}}$ (with normal vector $\vec{u}$) and $\overline{H}_{\vec{u}}$ denote its complement. So $\mathcal{P}(\overline{H}_{\vec{u}}) = \mathcal{P} \cap \overline{H}_{\vec{u}}$ is the set of points that have been swept so far. We stop the movement of $\ell_{\vec{u}}$ at the first point such that $\sum_{s\in\overline{H}_{\vec{u}}} \lambda_s \geq \ln(2/\tau)$ (ties should be broken in an arbitrary but consistent manner). One important property about $\overline{H}_{\vec{u}}$ is that $\Pr[\overline{H}_{\vec{u}} \models 0] \leq \tau/2$. We repeat the above process for all directions $\vec{u} \in \mathbb{S}^1$ and let $\mathcal{H} = \cap_{\vec{u}} H_{\vec{u}}$. Since $\lambda(\mathcal{P}) > 3\ln(2/\tau)$, by

[7] Here, $\mathbb{S}^1$ is the surface of the unit ball in $\mathbb{R}^d$.

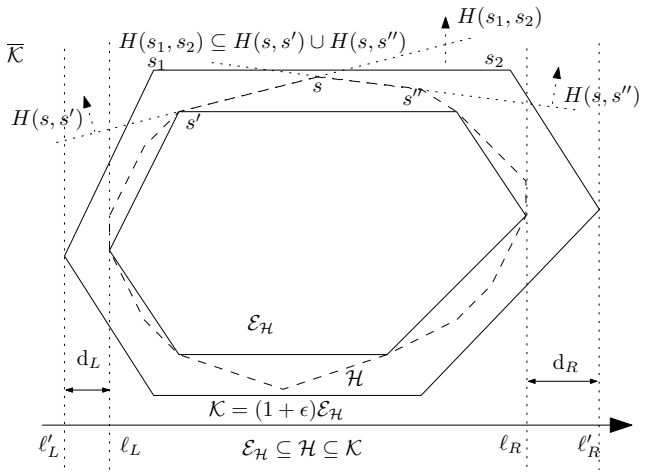

Figure 4-2: The construction of the (ε, τ)-QUANT-KERNEL $\mathcal{S}$. The dashed polygon is $\mathcal{H}$. The inner solid polygon is $\mathsf{ConvH}(\mathcal{E}_{\mathcal{H}})$ and the outer one is $K = (1+\varepsilon)\mathsf{ConvH}(\mathcal{E}_{\mathcal{H}})$. $\overline{\mathcal{K}}$ is the set of points outside $\mathcal{K}$.

Helly's theorem, $\mathcal{H}$ is nonempty. A careful examination of the above process reveals that $\mathcal{H}$ is in fact a convex polytope and each edge of the polytope is defined by two points in $\mathcal{P}$. [8] Moreover, $\mathcal{H}$ is the region of points with Tukey depth at least $\ln(2/\tau)$. [9]

The construction of the (ε, τ)-QUANT-KERNEL $\mathcal{S}$ is as follows. First, we use the algorithm in [7] to find a deterministic ε-kernel $\mathcal{E}_{\mathcal{H}}$ of size $O(\varepsilon^{-1/2})$ for $\mathcal{H}$. One useful property of the algorithm in [7] is that $\mathcal{E}_{\mathcal{H}}$ is a subset of the vertices of $\mathcal{H}$. Hence the convex polytope $\mathsf{ConvH}(\mathcal{E}_{\mathcal{H}})$ is contained in $\mathcal{H}$. Since $\mathcal{E}_{\mathcal{H}}$ is an ε-kernel, $(1+\varepsilon)\mathsf{ConvH}(\mathcal{E}_{\mathcal{H}}))$ (properly shifted) contains $\mathcal{H}$. [10] Let $\mathcal{K} = (1+\varepsilon)\mathsf{ConvH}(\mathcal{E}_{\mathcal{H}})$ and $\overline{\mathcal{K}} = \mathcal{P} \setminus \mathcal{K}$. See Figure 4-2.

Now, we apply the random sampling construction over $\overline{\mathcal{K}}$. More specifically, let $\lambda := \lambda(\overline{\mathcal{K}}) = \sum_{s \in \overline{\mathcal{K}} \cap \mathcal{P}} \lambda_s$. Let $\mathfrak{A}$ be the probability measure over $\mathcal{P} \cap \overline{\mathcal{K}}$ defined by $\mathfrak{A}(\{s\}) = \lambda_s/\lambda$ for every $s \in \mathcal{P} \cap \overline{\mathcal{K}}$. Let $\tau_1 = O(\tau/\lambda)$. We take $N = O(\tau_1^{-2} \log(1/\delta))$ independent samples from $\mathfrak{A}$ and let $\mathfrak{B}$ be the empirical distribution with each sample

[8] This also implies that we only need to do the sweep for $\binom{n}{2}$ directions. In fact, by a careful rotational sweep, we only need $O(n)$ directional sweeps.

[9] The Tukey depth of a point $x \in \mathcal{P}$ is defined as the minimum total weight of points of $\mathcal{P}$ contained in a closed halfspace whose bounding hyperplane passes through x.

[10] In fact, most existing algorithms (e.g., [7]) identify a point in the interior of $\mathcal{H}$ as origin, and compute an ε-kernel $\mathcal{E}_{\mathcal{H}}$ such that $\mathbf{f}(\mathcal{E}_{\mathcal{H}}, \vec{u}) \geq \frac{1}{1+\varepsilon}\mathbf{f}(\mathcal{H}, \vec{u})$ for all directions $\vec{u}$. So, $\mathcal{H} \subseteq (1+\varepsilon)\mathsf{ConvH}(\mathcal{E}_{\mathcal{H}})$ since $\mathbf{f}(\mathcal{H}, \vec{u}) \leq \mathbf{f}((1+\varepsilon)\mathsf{ConvH}(\mathcal{E}_{\mathcal{H}}), \vec{u})$ for all directions $\vec{u}$.

point having probability $1/N$. The (ε, τ)-QUANT-KERNEL $\mathcal{S}$ consists of the N points in $\mathfrak{B}$, each with the same existential probability $1-\exp(-\lambda/N)$, as well as all vertices of $\mathcal{K}$, each with probability 1. This finishes the construction of $\mathcal{S}$.

Now, we show that the size of $\mathcal{S}$ is constant (only depending on ε and δ), which is an immediate corollary of the following lemma.

Lemma 37. $\lambda = \lambda(\overline{\mathcal{K}}) = \sum_{s\in\overline{\mathcal{K}}} \lambda_s = O\Big(\frac{\ln 1/\tau}{\sqrt{\varepsilon}}\Big)$.

Proof. We can see that $\overline{\mathcal{K}}$ is the union of $O(\varepsilon^{-1/2})$ half-planes, each defined by a segment of $\mathcal{K}$. It suffices to show the sum of λ_s values in each half-plane is $O(\ln(1/\tau))$. Consider the half-plane $H(s_1, s_2)$ defined by segment (s_1, s_2) of $\mathcal{K}$. Suppose s is the vertex of $\mathcal{H}$ that is closest to the line (s_1, s_2). Let (s', s) and (s, s'') be the two edges of $\mathcal{H}$ incident on s. Clearly, $H(s', s) \cup H(s, s'')$, the union of the two half-planes defined by (s', s) and (s, s''), strictly contains $H(s_1, s_2)$. See Figure 4-2 for an illustration. Hence, $\sum_{s\in H(s_1,s_2)} \lambda_s$ is at most $2\ln(1/\tau)$. □

Now, we prove the main theorem in this section. The proof is an extension of Theorem 35. Here, the set $\mathcal{K}$ plays a similar role as the special point $s^\star$ in Theorem 35. Unlike Theorem 35, we also need to relax the length threshold here, which is necessary even for deterministic points.

Theorem 38. *In $\mathbb{R}^2$, let $\lambda = \lambda(\overline{\mathcal{K}})$ and $\tau_1 = O(\tau/\lambda)$, and $N = O\Big(\frac{1}{\tau_1^2}\log\frac{1}{\delta}\Big) = O\Big(\frac{\ln^2 1/\tau}{\varepsilon\tau^2}\log\frac{1}{\delta}\Big)$. With probability at least $1-\delta$, for any $t \geq 0$ and any direction $\vec{u}$, we have that*

$$\Pr\Big[\omega(\mathcal{S}, \vec{u}) \leq t\Big] \in \Pr\Big[\omega(\mathcal{P}, \vec{u}) \leq (1 \pm \varepsilon)t\Big] \pm \tau. \tag{4.4}$$

Proof. The proof is similar to Theorem 35. Fix an arbitrary direction $\vec{u}$ (w.l.o.g., say it is the x-axis). Rename all points in $\mathcal{P}$ as $s_1, s_2, \ldots, s_n$ according to the increasing order of their x-coordinates. We use $x(s_i)$ to denote the x-coordinate of s_i. Let ℓ_L (or ℓ_R) be the vertical line that passes the leftmost endpoint of $\mathcal{E}_\mathcal{H}$ (or the rightmost endpoint of $\mathcal{E}_\mathcal{H}$). We use $x(\ell_L)$ (or $x(\ell_R)$) to denote the x coordinate of ℓ_L (or ℓ_R) and let $\mathrm{d}(\ell_L, \ell_R) = |x(\ell_L) - x(\ell_R)|$. Suppose that $s_1, \ldots, s_k$ lie to the left of ℓ_L and $s_r, \ldots, s_n$

lie to the right of ℓ_R. Let the random variable $L = x(\ell_L) - \mathbf{f}(\{s_1, \ldots, s_k\}, -\vec{u})$ and $R = \mathbf{f}(\{s_r, \ldots, s_n\}, \vec{u}) - x(\ell_R)$. Let $W = L + R + \mathrm{d}(\ell_R, \ell_L)$. We can see that W is close to $\omega(\mathcal{P}, \vec{u})$ in the following sense. Let E denote the event that at least one point in $\{s_1, \ldots, s_k\}$ is present and at least one point in $\{s_r, \ldots, s_n\}$ is present. Conditioning on E, W is exactly $\omega(\mathcal{P}, \vec{u})$. Moreover, we can easily see $\Pr[E] \geq (1 - \tau/2)^2 \geq 1 - \tau$. Hence, we have

$$\begin{aligned}\Pr[W \leq t] - \tau \leq (1-\tau)\Pr[W \leq t] &\leq \Pr[\omega(\mathcal{P}, \vec{u}) \leq t \mid E]\Pr[E] \\ &\leq \Pr[\omega(\mathcal{P}, \vec{u}) \leq t] = \Pr[\omega(\mathcal{P}, \vec{u}) \leq t \mid E]\Pr[E] + \Pr[\omega(\mathcal{P}, \vec{u}) \leq t \mid \neg E]\Pr[\neg E] \\ &\leq \Pr[W \leq t] + \tau.\end{aligned}$$

Similarly, we let ℓ'_L (or ℓ'_R) be the vertical line that passes the leftmost endpoint of $\mathcal{K}$ (or the rightmost endpoint of $\mathcal{K}$). Suppose that $s'_1, \ldots, s'_j$ (points in $\mathcal{S}$) lie to the left of ℓ_L and $s'_s, \ldots, s'_N$ lie to the right of ℓ_R. We define $L' = x(\ell'_L) - \mathbf{f}(\{s'_1, \ldots, s'_j\}, -\vec{u})$ and $R' = \mathbf{f}(\{s'_s, \ldots, s'_N\}, \vec{u}) - x(\ell'_R)$. We can also see that $\omega(\mathcal{S}, \vec{u}) = L' + R' + \mathrm{d}(\ell_R, \ell_L)$.

Let $\mathrm{d}_L = x(\ell_L) - x(\ell'_L)$ and $\mathrm{d}_R = x(\ell'_R) - x(\ell_R)$. Let $H_{s'}$ be the half-plane $\{(x, y) \mid x \leq x(\ell'_L) - t\}$. We can see that for any $t \geq 0$,

$$\begin{aligned}\Pr[L \leq t + \mathrm{d}_L] - O(\lambda\tau_1) &= \Pr[X(H_{s'}) = 0] - O(\lambda\tau_1) \\ &\leq \Pr[L' \leq t] = \Pr[Y(H_{s'}) = 0] \leq \Pr[X(H_{s'}) = 0] + O(\lambda\tau_1) \\ &= \Pr[L \leq t + \mathrm{d}_L] + O(\lambda\tau_1),\end{aligned}$$

where the inequalities hold due to Lemma 33. Similarly, we can see that for any $t \geq 0$,

$$\Pr[R \leq t + \mathrm{d}_R] - O(\lambda\tau_1) \leq \Pr[R' \leq t] \leq \Pr[R \leq t + \mathrm{d}_L] + O(\lambda\tau_1).$$

Therefore, by Lemma 34, we have that for any $t > 0$,

$$\Pr[L + R \leq t + \mathrm{d}_L + \mathrm{d}_R] - O(\lambda\tau_1) \leq \Pr[L' + R' \leq t] \leq \Pr[L + R \leq t + \mathrm{d}_L + \mathrm{d}_R] + O(\lambda\tau_1).$$

Therefore, we can conclude that for any $t \geq \mathrm{d}(\ell'_L, \ell'_R)$,

$$\begin{aligned}
\omega(\Pr[\mathcal{S}, \vec{u}) \leq t] &= \Pr[L' + R' + \mathrm{d}(\ell'_L, \ell'_R) \leq t] \\
&\in \Pr[L + R + \mathrm{d}(\ell'_L, \ell'_R) \leq t + \mathrm{d}_L + \mathrm{d}_R] \pm O(\lambda\tau_1) \\
&= \Pr[L + R + \mathrm{d}(\ell_L, \ell_R) \leq t] \pm O(\lambda\tau_1) \\
&= \Pr[W \leq t] \pm O(\lambda\tau_1) \\
&= \Pr[\omega(\mathcal{P}, \vec{u}) \leq t] \pm O(\lambda\tau_1 + \tau).
\end{aligned}$$

Noticing that $\tau \geq \Pr[\omega(\mathcal{P}, \vec{u}) \leq \mathrm{d}(\ell_L, \ell_R)] \geq \Pr[\omega(\mathcal{P}, \vec{u}) \leq (1-\varepsilon)\mathrm{d}(\ell'_L, \ell'_R)]$, we can obtain that, for any $t < \mathrm{d}(\ell'_L, \ell'_R)$, $\Pr[\omega(\mathcal{S}, \vec{u}) \leq t] = 0 \geq \Pr[\omega(\mathcal{P}, \vec{u}) \leq (1-\varepsilon)t] - \tau$. Moreover, it is trivially true that $\Pr[\omega(\mathcal{S}, \vec{u}) \leq t] = 0 \leq \Pr[\omega(\mathcal{P}, \vec{u}) \leq (1-\varepsilon)t] + \tau$. The proof is completed. □

Higher Dimensions. Our constructions can be easily extended to $\mathbb{R}^d$ for any constant $d > 2$. The sampling bound (Theorem 32) still holds if the number of samples is $O(d\tau_1^{-2}\log(1/\delta)) = O(\tau_1^{-2}\log(1/\delta))$. Hence, Theorem 35 and Theorem 36 hold with the same parameters (d is hidden in the constant). In order for Algorithm 2 to work, we need $\lambda(\mathcal{P}) > (d+1)\ln(2/\tau)$ to ensure $\mathcal{H}$ is nonempty. Instead of constructing an ε-kernel $\mathcal{E}_{\mathcal{H}}$ with $O(\varepsilon^{-(d-1)/2})$ vertices, we construct a convex set $\mathcal{K}$ which is the intersection of $O(\varepsilon^{-(d-1)/2})$ halfspaces and satisfies $(1-\varepsilon)\mathcal{K} \subseteq \mathcal{H} \subseteq \mathcal{K}$ (this can be done by either working with the dual, or directly using the construction implicit in [37]).

Now, we briefly sketch how to compute such $\mathcal{K}$ using the dual approach. We first compute the dual $\mathcal{H}^\star$ of $\mathcal{H}$ in $\mathbb{R}^d$. Recall the dual (also called the polar body) $\mathcal{H}^\star$ of $\mathcal{H}$ is defined as the set $\{x \in \mathbb{R}^d \mid \langle x, y\rangle \leq 1, y \in \mathcal{H}\}$. $\mathcal{H}^\star$ has $O(n^d)$ vertices (each corresponding to a face of $\mathcal{H}$). Then, compute an ε-kernel $\mathcal{E}^\star_{\mathcal{H}^\star}$ with $O(\varepsilon^{-(d-1)/2})$ vertices for $\mathcal{H}^\star$. Taking the dual of $\mathcal{E}^\star_{\mathcal{H}^\star}$ gives the desired $\mathcal{K}$, which is an intersection of $O(\varepsilon^{-(d-1)/2})$ halfspaces (each corresponding to a point in $\mathcal{E}^\star_{\mathcal{H}^\star}$). The correctness can be easily seen by an argument through the gauge function $g(\mathcal{E}^\star_{\mathcal{H}^\star}, x) = \min\{\lambda \geq 0 \mid x \in \lambda\mathcal{E}^\star_{\mathcal{H}^\star}\}$. Since $\mathcal{E}^\star_{\mathcal{H}^\star} \subseteq \mathcal{H}^\star \subseteq (1+\varepsilon)\mathcal{E}^\star_{\mathcal{H}^\star}$, we can see that $\frac{1}{1+\varepsilon}g(\mathcal{E}^\star_{\mathcal{H}^\star}, x) =$

$g((1+\varepsilon)\mathcal{E}^{\star}_{\mathcal{H}^{\star}}, x) \leq g(\mathcal{H}^{\star}, x) \leq g(\mathcal{E}^{\star}_{\mathcal{H}^{\star}}, x)$. The correctness follows from the duality between the gauge function and the support function, which says $g(\mathcal{E}^{\star}_{\mathcal{H}^{\star}}, x) = \mathbf{f}(\mathcal{K}, x)$ and $g(\mathcal{H}^{\star}, x) = \mathbf{f}(\mathcal{H}, x)$ for all $x \in \mathbb{S}^{d-1}$ (see e.g., [95]).

We generalize Lemma 37 for $\mathbb{R}^d$ by the following lemma.

Lemma 39. *There is a convex set $\mathcal{K}$, which is an intersection of $O(\varepsilon^{-(d-1)/2})$ halfspaces and satisfies $(1-\varepsilon)\mathcal{K} \subseteq \mathcal{H} \subseteq \mathcal{K}$. Moreover, we have that $\lambda(\overline{\mathcal{K}}) = O(\varepsilon^{-(d-1)/2} \ln(1/\tau))$.*

Plugging the new bound of $\lambda(\overline{\mathcal{K}})$, we can see that it is enough to set

$$N = O(\tau^{-2}\varepsilon^{-(d-1)} \log \frac{1}{\delta} \mathrm{polylog} \frac{1}{\tau}) = \widetilde{O}(\varepsilon^{-(d-1)}\tau^{-2}).$$

Running Time. The algorithm in Section 4.3.2 takes only $O(Nn)$ time (where N is the size of the kernel, which is constant if ε, τ and λ are constant). The algorithm in Section 4.3.2 is substantially slower. The most time consuming part is the construction of $\mathcal{H}$, which is the intersection of all halfspaces. In $\mathbb{R}^d$, we need to sweep $O(n^d)$ directions (each determined by d points). So the polytope $\mathcal{H}$ may have $O(n^d)$ faces. Using the dual approach, we can compute $\mathcal{K}$ in $O(n^d)$ time (linear in the number of points in the dual space) as well. Overall, the running time is $O(n^d)$.

A Nearly Linear Time Algorithm for Constructing (ε, τ)-quant-kernels

We describe a nearly linear time algorithm for constructing an (ε, τ)-QUANT-KERNEL in the existential uncertainty model. As mentioned before, the algorithm in Section 4.3.2 takes linear time. So we only need a nearly linear time algorithm for constructing $\mathcal{H}$ (and $\mathcal{K}$). Note that $\mathcal{H}$ is the set of points in $\mathbb{R}^d$ with Tukey depth at least $\ln(2/\tau)$. One tempting idea is to utilize the notion of ε-approximation (which can be obtained by sampling) to compute the approximate Tukey depth for the points, as done in [87]. However, a careful examination of this approach shows that the sample size needs to be as large as $O(\lambda(\mathcal{P}))$ (to ensure that for every halfspace, the difference between the real weight and the sample weight is less than, say $0.1 \ln(2/\tau)$). Another useful observation is that only points with small (around $\ln(2/\tau)$) Turkey depth are

relevant in constructing $\mathcal{H}$. Hence, we can first sample an ε-approximation of very small size (say $k = O(\log n)$), and use it to quickly identify the region $\mathcal{H}_1$ in which all points have large (i.e., $\lambda(\mathcal{P})/k$) Turkey depth (so $\mathcal{H}_1 \subseteq \mathcal{H}$). Then, we can delete all points inside $\mathcal{H}_1$ and focus on the remaining points. Ideally, the total weight of the remaining points can be reduced significantly and a random sample of the same size k would give an ε'-approximation of the remaining points for some $\varepsilon' < \varepsilon$. We repeat the above until the total weight of the remaining points reduces to a constant, and then a constant size sample suffices. However, it is possible that all points have fairly small Tukey depth (consider the case where all points are in convex position), and no point can be removed. To resolve the issue, we use the idea in Lemma 37: there is a convex set $\mathcal{K}_1$ slightly larger than $\mathcal{H}_1$ such that the weight of points outside $\mathcal{K}_1$ is much smaller. Hence, we can make progress by deleting all points inside $\mathcal{K}_1$. Since $\mathcal{K}_1$ is only slightly larger than $\mathcal{H}_1$, we do not lose too much in terms of the distance. Our algorithm carefully implements the above iterative sampling idea.

For ease of exposition, we first focus on $\mathbb{R}^2$. Consider the Poissonized instance of $\mathcal{P}$. We would like to find two convex sets $\mathcal{H}$ and $\mathcal{K}$ satisfying the following properties.

P1. Assume without loss of generality that the origin is in $\mathcal{H}$. We require that $\frac{1}{1+\varepsilon}\mathcal{K} \subseteq \mathcal{H} \subseteq \mathcal{K}$.

P2. For a direction $\vec{u} \in \mathbb{S}^1$, we use $H(\mathcal{H}, \vec{u})$ to denote the halfplane which does not contain $\mathcal{H}$ and whose boundary is the supporting line of $\mathcal{H}$ with normal direction $\vec{u}$. We require that $\lambda(H(\mathcal{H}, \vec{u})) = \sum_{s \in H(\mathcal{H}, \vec{u})} \lambda_s \geq \ln(2/\tau)$ for all directions $\vec{u} \in \mathbb{S}^1$.

P3. $\lambda(\overline{\mathcal{K}}) = \widetilde{O}(1/\sqrt{\varepsilon})$.

By a careful examination of our analysis in Section 4.3.2, we can see the above properties are all we need for the analysis.

Let $\mathcal{H}^\star$ denote the $\mathcal{H}$ found using the exact algorithm in Section 4.3.2. We use the following set of parameters:

$$z = O(\log n), \quad \varepsilon_1 = O\left(\frac{\varepsilon}{\log n}\right), \quad \varepsilon_2 = O\left(\sqrt{\frac{\varepsilon}{\log n}}\right).$$

Our algorithm proceeds in rounds. Initially, let $\mathcal{H}_0 = \mathsf{ConvH}(\{s \in \mathcal{P} \mid \lambda_s \geq \ln(2/\tau)\})$. In round i (for $1 \leq i \leq z$), we construct two convex sets $\mathcal{H}_i$ and $\mathcal{K}_i$ such that

1. $\mathcal{H}_0 \subseteq \mathcal{K}_0 \subseteq \mathcal{H}_1 \subseteq \mathcal{K}_1 \subseteq \ldots \subseteq \mathcal{H}_z \subseteq \mathcal{K}_z$;
2. $\frac{1}{1+\varepsilon_1}\mathcal{K}_i \subseteq \mathcal{H}_i \subseteq \mathcal{K}_i$ ($\mathcal{K}_i$ and $\mathcal{H}_i$ are very close to each other);
3. $\frac{1}{(1+\varepsilon_1)^i}\mathcal{H}_i \subseteq \mathcal{H}^\star$ ($\mathcal{H}_i$ is almost contained in $\mathcal{H}^\star$);
4. $\lambda(\mathcal{P} \cap \overline{\mathcal{K}}_i) \leq \frac{1}{2}\lambda(\mathcal{P} \cap \overline{\mathcal{K}}_{i-1})$ (the total weight outside $\mathcal{K}_i$ reduces by a factor of at least one half).

We repeat the above process until $\lambda(\overline{\mathcal{K}}_i) \leq \widetilde{O}(1/\sqrt{\varepsilon})$.

Before spelling out the details of our algorithm, we need a few definitions.

Definition 40. *For a set P of weighted points in $\mathbb{R}^d$, we use $\mathsf{TK}(P, \gamma)$ to denote the set of points $x \in \mathbb{R}^d$ with Tukey depth at least γ. It is known that $\mathsf{TK}(P, \gamma)$ is convex (see e.g., [87]). By this definition, $\mathcal{H}^\star = \mathsf{TK}(\mathcal{P}, \ln(2/\tau))$.*

Recall the definition of ε-approximation from Theorem 29. By Theorem 29 (or Theorem 32), we can see that a set of $O(\varepsilon^{-2}\log(1/\delta))$ sampled points is an ε-approximation with probability $1 - \delta$.

We are ready to describe the details of our algorithm. Initially $\mathcal{H}_0 = \mathsf{ConvH}(\{s \in \mathcal{P} \mid \lambda_s \geq \ln(2/\tau)\})$ (obviously $\mathcal{H}_0 \subseteq \mathcal{H}^\star$). Compute a deterministic ε_1-kernel $C_{\mathcal{H}_0}$ of $\mathcal{H}_0$ and let $\mathcal{K}_0 = (1+\varepsilon_1)\mathsf{ConvH}(C_{\mathcal{H}_0})$. Delete all point in $\mathcal{P} \cap \mathcal{K}_0$ and let $\mathcal{P}_1$ be the remaining points in $\mathcal{P}$ (i.e., $\mathcal{P}_1 = \mathcal{P} \cap \overline{\mathcal{K}}_0$). Let $\mathrm{Ver}(\mathcal{K}_0)$ denote all vertices of $\mathcal{K}_0$ (notice that some of them may not be original points in $\mathcal{P}$).

Now, suppose we describe the ith round for general $i > 1$. We have the remaining vertices in $\mathcal{P}_{i-1}$ and $\mathrm{Ver}(\mathcal{K}_{i-1})$. Let each point $s \in \mathcal{P}_i$ has the same old weight λ_s and each point in $\mathrm{Ver}(\mathcal{K}_{i-1})$ has weight $+\infty$ (to make sure every point in $\mathcal{K}_{i-1}$ has Turkey depth $+\infty$). Using random sampling on $\mathcal{P}_i$, obtain an ε_2 -approximation $\mathcal{E}_i$ (of size $L = O(\varepsilon_2^{-2})$) for $\mathcal{P}_i$. Then compute (using the brute-force algorithm described in Section 4.3.2)

$$\mathcal{H}_i = \mathsf{TK}(\mathcal{E}_i \cup \mathrm{Ver}(\mathcal{K}_{i-1}), \max\{4\varepsilon_2\lambda(\mathcal{P}_i), 2\ln(2/\tau)\}).$$

Note that $\mathcal{K}_{i-1} \subseteq \mathcal{H}_i$. Compute a deterministic ε_2-kernel $C_{\mathcal{H}_i}$ of $\mathcal{H}_i$ and let $\mathcal{K}_i = (1+\varepsilon_1)\mathsf{ConvH}(C_{\mathcal{H}_i})$ (hence $\mathsf{ConvH}(C_{\mathcal{H}_i}) \subseteq \mathcal{H}_i \subseteq \mathcal{K}_i$). Then, we delete all points in $\mathcal{P} \cap \mathcal{K}_i$ and add all vertices of $\mathcal{K}_i$ (denoted as $\mathrm{Ver}(\mathcal{K}_i)$). Let $\mathcal{P}_{i+1}$ be the remaining points in $\mathcal{P} \cap \overline{\mathcal{K}}_i$.

Our algorithm terminates when $\lambda(\mathcal{P}_i) \leq \widetilde{O}(1/\sqrt{\varepsilon})$. Suppose the last round is z. Finally, we let $\mathcal{H} = \frac{1}{(1+\varepsilon_1)^z}\mathcal{H}_z$ and $\mathcal{K} = \mathcal{K}_z$.

First we show the algorithm terminates after at most a logarithmic number of rounds.

Lemma 41. $z = O(\log n)$.

Proof. If $\mathcal{H}_i = \mathsf{TK}(\mathcal{E}_i \cup \mathrm{Ver}(\mathcal{K}_{i-1}), 2\ln(2/\tau))$, then we stop since $\lambda(\mathcal{P}_{i+1}) \leq \tilde{O}(1/\sqrt{\varepsilon})$ by Lemma 37. Thus, we only need to bound the number of iterations where $\mathcal{H}_i = \mathsf{TK}(\mathcal{E}_i \cup \mathrm{Ver}(\mathcal{K}_{i-1}), 4\varepsilon_2\lambda(\mathcal{P}_i))$.

Initially, it is not hard to see that $\lambda(\mathcal{P}_1) \leq n\ln(2/\tau)$. Using Lemma 37, we can see that $\lambda(\mathcal{P}_{i+1}) = \lambda(\mathcal{P} \cap \overline{\mathcal{K}}_i) \leq O(5\varepsilon_2\lambda(\mathcal{P}_i)/\sqrt{\varepsilon_1}) \leq \lambda(\mathcal{P}_i)/2$ for the constant defining ε_1 sufficiently large. Hence, $\lambda(\mathcal{P}_i) \leq \lambda(\mathcal{P})/2^i$. □

We need to show $\mathcal{H}$ and $\mathcal{K}$ satisfy P1, P2 and P3. P3 is quite obvious by our algorithm. It is also not hard to see P1 since $(1+\varepsilon_1)^{z+1} \leq 1+\varepsilon$ and

$$\frac{1}{(1+\varepsilon)}\mathcal{K}_z \subseteq \frac{1}{(1+\varepsilon_1)^{z+1}}\mathcal{K}_z \subseteq \frac{1}{(1+\varepsilon_1)^z}\mathcal{H}_z = \mathcal{H} \subseteq \mathcal{H}_z \subseteq \mathcal{K}_z = \mathcal{K}.$$

The most difficult part is to show P2 holds: For every direction $\vec{u} \in \mathbb{S}^1$, $\lambda(H(\mathcal{H}, \vec{u})) = \sum_{s \in H(\mathcal{H},\vec{u})} \lambda_s \geq \ln(2/\tau)$. In fact, we show that $\mathcal{H} \subseteq \mathcal{H}^\star$, from which P2 follows trivially, which suffices to prove the following lemma.

Lemma 42. $\frac{1}{(1+\varepsilon_1)^i}\mathcal{H}_i \subseteq \mathcal{H}^\star$ *for all* $0 \leq i \leq z$. *In particular,* $\mathcal{H} \subseteq \mathcal{H}^\star$.

Proof. We prove the lemma by induction. $\mathcal{H}_0 \subseteq \mathcal{H}^\star$ clearly satisfies the lemma. For ease of notation, we let $\eta = 1/(1+\varepsilon_1)$. Suppose the lemma is true for $\mathcal{H}_{i-1}$, from which we can see that

$$\eta^i\mathcal{K}_{i-1} \subseteq \eta^{i-1}\mathcal{H}_{i-1} \subseteq \mathcal{H}^\star.$$

Now we show the lemma holds for $\mathcal{H}_i$. Consider the ith round. Let $\mathcal{E}_i$ be an $\varepsilon_2\lambda(\mathcal{P}_i)$ -approximation for $\mathcal{P}_i = \mathcal{P} \cap \overline{\mathcal{K}}_{i-1}$ and $\mathcal{H}_i = \mathsf{TK}(\mathcal{E}_i \cup \mathrm{Ver}(\mathcal{K}_{i-1}), 4\varepsilon_2\lambda(\mathcal{P}_i))$. Fix an arbitrary direction $\vec{u} \in \mathbb{S}^1$ (w.l.o.g., assume that $u = (0, -1)$, i.e., the downward direction), let $H(\eta^i\mathcal{H}_i, \vec{u})$ be a halfplane whose boundary is tangent to $\eta^i\mathcal{H}_i$. It suffices to show that $\lambda(H(\eta^i\mathcal{H}_i, \vec{u})) = \sum_{s\in H(\eta^i\mathcal{H}_i,\vec{u})} \lambda_s \geq \ln(2/\tau)$. We move a sweep line $\ell_{\vec{u}}$ orthogonal to $\vec{u}$, along the direction $\vec{u}$ (i.e., from top to bottom), to sweep through the points in $\mathcal{P}_i \cap \mathrm{Ver}(\mathcal{K}_{i-1})$ until the total weight we have swept is at least $\ln(2/\tau)$. We distinguish two cases:

1. $\ell_{\vec{u}}$ hits a point s in $\mathrm{Ver}(\mathcal{K}_{i-1})$ (recall the weight for such point is $+\infty$). We can see that s is the topmost point of $\mathcal{K}_{i-1}$ and $\mathcal{H}_i$ (or equivalently, $\ell_{\vec{u}}$ is also a supporting line for $\mathcal{H}_i$). Since $\eta^i\mathcal{K}_{i-1} \subseteq \mathcal{H}^\star$ by the induction hypothesis, the topmost point of $\eta^i\mathcal{K}_{i-1}$ is lower than that for $\mathcal{H}^\star$. The topmost point of $\eta^i\mathcal{K}_{i-1}$ is also the highest point of $\eta^i\mathcal{H}_i$, from which we can see $H(\eta^i\mathcal{H}_i, \vec{u})$ is lower than $H(\mathcal{H}^\star, \vec{u})$, which implies that $\lambda(H(\eta^i\mathcal{H}_i, \vec{u})) \geq \ln(2/\tau)$.

2. $\ell_{\vec{u}}$ stops moving when it hits an original point in $\mathcal{P}_i$. Since $\max\{3\varepsilon_2\lambda(\mathcal{P}_i), 2\ln(2/\tau) - \varepsilon_2\lambda(\mathcal{P}_i)\} > \ln(2/\tau)$, by definition of $\mathcal{H}_i$, $H(\mathcal{H}_i, \vec{u})$ can not be higher than $\ell_{\vec{u}}$. The boundary of $H(\eta^i\mathcal{H}_i, \vec{u})$ is even lower, from which we can see $\lambda(H(\eta^i\mathcal{H}_i, \vec{u})) \geq \ln(2/\tau)$.

Hence, every point in $\eta^i\mathcal{H}_i$ has Tukey depth at least $\ln(2/\tau)$, which implies the lemma. □

Running time. In each round, we compute in linear time an ε_2-approximation $\mathcal{E}_i$ of size $O(\varepsilon_2^{-2}\log(1/\delta)) = \mathrm{polylog}(n)$ (with $\delta = \mathrm{poly}(n)$ to ensure each probabilistic event succeeds with high probability). $\mathcal{K}_i$ is a dilation of an ε_1-kernel. So the size of $\mathrm{Ver}(\mathcal{K}_i)$ is at most $1/\sqrt{\varepsilon_1} = O(\log^{1/2} n)$. Deciding whether a point is inside $\mathcal{K}_i$ can be solved in $\mathrm{polylog}(n)$ time, by a linear program with $|\mathrm{Ver}(\mathcal{K}_i)|$ variables. To compute $\mathcal{H}_i$, we can use the brute-force algorithm described in Section 4.3.2, which takes $\mathrm{poly}(|\mathcal{E}_i \cap \mathrm{Ver}(\mathcal{K}_{i-1})|) = \mathrm{polylog}(n)$ time. There are logarithmic number of rounds. So the overall running time is $O(n\mathrm{polylog}n)$.

Higher Dimension. Our algorithm can be easily extended to $\mathbb{R}^d$ for any constant $d > 2$. In $\mathbb{R}^d$, we let $\varepsilon_1 = O(\varepsilon/\log n)$ and $\varepsilon_2 = O(\varepsilon/\log n)^{(d-1)/2}$. With the new parameters, we can easily check that Lemma 41 still holds. We can construct an (ε, τ)-QUANT-KERNEL of size $\min\{O\Big(\tau^{-2}\max\{\lambda^2, \lambda^4\}\log(1/\delta)\Big), O(\tau^{-2}\varepsilon^{-(d-1)}\log(1/\delta)\text{polylog}(1/\tau)\}$. The first term is from Theorem 36 and the second from the higher-dimensional extension to Theorem 38. Now, let us examine the running time. In $\mathbb{R}^d$, $|\text{Ver}(\mathcal{K}_i)|$ is at most $\varepsilon_1^{-(d-1)/2} = O(\log^{(d-1)/2} n)$. So deciding whether a point is inside $\mathcal{K}_i$ can be solved in $\log^{O(d)}(n)$ time. Computing $\mathcal{H}_i$ takes $\log^{O(d)}(n)$ time using the brute-force algorithm. So the overall running time is $O(n\log^{O(d)} n)$.

In summary, we obtain the following theorem for (ε, τ)-QUANT-KERNEL.

Theorem 5. *(restated) $\mathcal{P}$ is a set of uncertain points in $\mathbb{R}^d$ with existential uncertainty. Let $\lambda = \sum_{s_i \in \mathcal{P}}(-\ln(1-p_i))$. There exists an (ε, τ)-QUANT-KERNEL for $\mathcal{P}$, which consists of a set of independent uncertain points of cardinality $\min\{\widetilde{O}(\tau^{-2}\max\{\lambda^2, \lambda^4\}), \widetilde{O}(\varepsilon^{-(d-1)}\tau^{-2})\}$. The algorithm for constructing such a coreset runs in $\widetilde{O}(n\log^{O(d)} n)$ time.*

4.3.3 (ε, τ)-quant-kernel Under the Subset Constraint

We show it is possible to construct an (ε, τ)-QUANT-KERNEL in the existential model under the *β-assumption*: each possible location realizes a point with a probability at least β, where $\beta > 0$ is some fixed constant.

Theorem 43. *Under the β-assumption, there is an (ε, τ)-QUANT-KERNEL in $\mathbb{R}^d$, which is of size $O(\mu^{-(d-1)/2}\log(1/\mu))$ and satisfies the subset constraint, in the existential uncertainty model, where $\mu = \min\{\varepsilon, \tau\}$.*

In fact, the algorithm is exactly the same as constructing an ε-EXP-KERNEL and the proof of the above theorem is implicit in the proof of Theorem 26.

4.4 (ε, r)-fpow-kernel Under the β-Assumption

In this section, we show an (ε, r)-FPOW-KERNEL exists in the existential uncertainty model under the β-assumption. Recall that the function $T_r(P, \vec{u}) = \max_{s\in P}\langle \vec{u}, s\rangle^{1/r} - \min_{s\in P}\langle \vec{u}, s\rangle^{1/r}$. For ease of notation, we write $\mathbb{E}[T_r(\mathcal{P}, \vec{u})]$ to denote $\mathbb{E}_{P\sim\mathcal{P}}[T_r(P, \vec{u})]$. Our goal is to find a set $\mathcal{S}$ of stochastic points such that for all directions $\vec{u} \in \mathcal{P}^\star$, we have that $\mathbb{E}[T_r(\mathcal{S}, \vec{u})] \in (1 \pm \varepsilon)\mathbb{E}[T_r(\mathcal{P}, \vec{u})]$.

Our construction of $\mathcal{S}$ is almost the same as that in Section 4.3.1. Suppose we sample N (fixed later) independent realizations and take the ε_0-kernel for each of them. Suppose they are $\{\mathcal{E}_1, \ldots, \mathcal{E}_N\}$ and we associate each a probability $1/N$. We denote the resulting (ε, r)-FPOW-KERNEL by $\mathcal{S}$. Hence, for any direction $\vec{u} \in \mathcal{P}^\star$, $\mathbb{E}[T_r(\mathcal{S}, \vec{u})] = \frac{1}{N}\sum_{i=1}^{N} T_r(\mathcal{E}_i, \vec{u})$ and we use this value as the estimation of $\mathbb{E}[T_r(\mathcal{P}, \vec{u})]$. Now, we show $\mathcal{S}$ is indeed an (ε, r)-FPOW-KERNEL.

Recall that we use $\mathcal{E}(P)$ to denote the deterministic ε-kernel for any realization $P \sim \mathcal{P}$. We first compare $\mathcal{P}$ with the random set $\mathcal{E}(P)$.

Lemma 44. *For any $t \geq 0$ and any direction $\vec{u} \in \mathcal{P}^\star$, we have that*

$$(1 - \varepsilon/2)\mathbb{E}[T_r(\mathcal{P}, \vec{u})] \leq \mathbb{E}_{P\sim\mathcal{P}}[T_r(\mathcal{E}(P), \vec{u}))] \leq \mathbb{E}[T_r(\mathcal{P}, \vec{u})].$$

Proof. By Lemma 4.6 in [7], we have that $(1-\varepsilon/2)T_r(P, \vec{u}) \leq T_r(\mathcal{E}(P), \vec{u}) \leq T_r(P, \vec{u})$. The lemma follows by combining all realizations. □

Now we show that $\mathcal{S}$ is an (ε, r)-FPOW-KERNEL of $\mathcal{E}(P)$. We first prove the following lemma. The proof is almost the same as that of Lemma 30, and can be found in Section 4.8.

Lemma 45. *Let $N = O\big(\varepsilon_1^{-2}\varepsilon_0^{-(d-1)/2}\log(1/\varepsilon_0)\big)$, where $\varepsilon_0 = (\varepsilon/4(r-1))^r$, $\varepsilon_1 = \varepsilon\beta^2$. For any $t \geq 0$ and any direction $\vec{u} \in \mathcal{P}^\star$, we have that*

$$\Pr_{P\sim\mathcal{S}}[\max_{s\in P}\langle \vec{u}, s\rangle^{1/r} \geq t] \in \Pr_{P\sim\mathcal{P}}[\max_{s\in\mathcal{E}(P)}\langle \vec{u}, s\rangle^{1/r} \geq t)] \pm \varepsilon_1/4, \text{ and}$$

$$\Pr_{P\sim\mathcal{S}}[\min_{s\in P}\langle \vec{u}, s\rangle^{1/r} \geq t] \in \Pr_{P\sim\mathcal{P}}[\min_{s\in\mathcal{E}(P)}\langle \vec{u}, s\rangle^{1/r} \geq t)] \pm \varepsilon_1/4.$$

Lemma 46. *Let* $N = O\big(\beta^{-4}\varepsilon^{-(rd-r+4)/2}\log(1/\varepsilon)\big)$ *and* $\varepsilon_0 = (\varepsilon/4(r-1))^r$. $\mathcal{S}$ *constructed above is an* (ε, r)-FPOW-KERNEL *in* $\mathbb{R}^d$.

Proof. Fix a direction $\vec{u} \in \mathcal{P}^\star$. Let $A = \max_{s\in\mathcal{P}}\langle\vec{u}, s\rangle^{1/r}$, $B = \min_{s\in\mathcal{P}}\langle\vec{u}, s\rangle^{1/r}$. We observe that $B \le \max_{s\in P}\langle\vec{u}, s\rangle^{1/r} \le A$ for any realization $P \sim \mathcal{P}$. We also need the following basic fact about the expectation: For a random variable X, if $\Pr[X \ge a] = 1$, then $\mathbb{E}[X] = \int_b^\infty \Pr[X \ge x]\mathrm{d}x + b$ for any $b \le a$. Thus, we have that

$$\begin{aligned}
\mathbb{E}_{P\sim\mathcal{P}}[\max_{s\in\mathcal{E}(P)}\langle\vec{u}, s\rangle^{1/r}] &= \int_B^A \Pr_{P\sim\mathcal{P}}[\max_{s\in\mathcal{E}(P)}\langle\vec{u}, s\rangle^{1/r} \ge x]\mathrm{d}x + B \\
&\le \int_B^A \Pr_{P\sim\mathcal{S}}[\max_{s\in P}\langle\vec{u}, s\rangle^{1/r} \ge x]\mathrm{d}x + B + \varepsilon_1(A-B)/4 \\
&= \mathbb{E}_{P\sim\mathcal{S}}[\max_{s\in P}\langle\vec{u}, s\rangle^{1/r}] + \varepsilon_1(A-B)/4,
\end{aligned}$$

where the first inequality is due to Lemma 45. Similarly, we can show the following two inequalities:

$$\mathbb{E}_{P\sim\mathcal{S}}[\max_{s\in P}\langle\vec{u}, s\rangle^{1/r}] \in \mathbb{E}_{P\sim\mathcal{P}}[\max_{s\in\mathcal{E}(P)}\langle\vec{u}, s\rangle^{1/r}] \pm \varepsilon_1(A-B)/4,$$

$$\mathbb{E}_{P\sim\mathcal{S}}[\min_{s\in P}\langle\vec{u}, s\rangle^{1/r}] \in \mathbb{E}_{P\sim\mathcal{P}}[\min_{s\in\mathcal{E}(P)}\langle\vec{u}, s\rangle^{1/r}] \pm \varepsilon_1(A-B)/4.$$

Recall that $T_r(P, \vec{u}) = \max_{s\in P}\langle\vec{u}, s\rangle^{1/r} - \min_{s\in P}\langle\vec{u}, s\rangle^{1/r}$. By the linearity of expectation, we conclude that

$$\mathbb{E}[T_r(\mathcal{P}, \vec{u})] \in \mathbb{E}_{P\sim\mathcal{P}}[T_r(\mathcal{E}(P), \vec{u})] \pm \varepsilon_1(A-B)/2.$$

Combining Lemma 44, we have that $\mathbb{E}[T_r(\mathcal{S}, \vec{u})] \in (1\pm\varepsilon/2)\mathbb{E}[T_r(\mathcal{P}, \vec{u})] \pm \varepsilon_1(A-B)/2$. By the β-assumption, we know that $\mathbb{E}[T_r(\mathcal{P}, \vec{u})] \ge \beta^2(A-B)$. Thus, $\varepsilon_1(A-B)/2 \le \frac{\varepsilon}{2}\mathbb{E}[T_r(\mathcal{P}, \vec{u})]$, and $\mathbb{E}[T_r(\mathcal{S}, \vec{u})] \in (1\pm\varepsilon)\mathbb{E}[T_r(\mathcal{P}, \vec{u})]$. □

Running time. In each sample, the size of a deterministic ε_0-kernel $\mathcal{E}_i$ is at most $O\big(\varepsilon_0^{-(d-1)/2}\big)$. Note that constructing an ε_0-kernel can be solved in linear time. We take $O\big(\varepsilon_1^{-2}\varepsilon_0^{-(d-1)/2}\log(1/\varepsilon_0)\big)$ samples in total. So the overall running time is $O\big(n\beta^{-4}\varepsilon^{-(rd-r+4)/2}\log(1/\varepsilon) + \mathrm{poly}(1/\varepsilon)\big) = \widetilde{O}\left(n\varepsilon^{-(rd-r+4)/2}\right)$.

Note that each ε_0-kernel contains $O(\varepsilon^{-r(d-1)/2})$ points. We take $N = O\big(\varepsilon_1^{-2}\varepsilon_0^{-(d-1)/2}\log(1/\varepsilon_0)\big)$ independent samples. So the total size of (ε, r)-FPOW-KERNEL is $O(\beta^{-4}\varepsilon^{-(rd-r+2)}\log(1/\varepsilon))$. In summary, we obtain the following theorem.

Theorem 7. *(restated) An (ε, r)-FPOW-KERNEL of size $\widetilde{O}(\varepsilon^{-(rd-r+2)})$ can be constructed in $\widetilde{O}\left(n\varepsilon^{-(rd-r+4)/2}\right)$ time in the existential uncertainty model under the β-assumption. In particular, the (ε, r)-FPOW-KERNEL consists of $N = \widetilde{O}(\varepsilon^{-(rd-r+4)/2})$ point sets, each occuring with probability $1/N$ and containing $O(\varepsilon^{-r(d-1)/2})$ deterministic points.*

4.5 Applications

In this section, we show that our coreset results for the directional width problem readily imply several coreset results for other stochastic problems, just as in the deterministic setting. We introduce these stochastic problems and briefly summarize our results below.

4.5.1 Approximating the Extent of Uncertain Functions

We first consider the problem of approximating the extent of a set $\mathcal{H}$ of uncertain functions. As before, we consider both the existential model and the locational model of uncertain functions.

1. In the existential model, each uncertain function h is a function in $\mathbb{R}^d$ associated with a existential probability p_f, which indicates the probability that h presents in a random realization.

2. In the locational model, each uncertain function h is associated with a finite set $\{h_1, h_2, \ldots\}$ of deterministic functions in $\mathbb{R}^d$. Each h_i is associated with a probability value $p(h_i)$, such that $\sum_i p(h_i) = 1$. In a random realization, h is independently realized to some h_i, with probability $p(h_i)$.

We use $\mathcal{H}$ to denote the random instance, that is a random set of functions. We use $h \in \mathcal{H}$ to denote the event that the deterministic function h is present in the

instance. For each point $x \in \mathbb{R}^d$, we let the random variable $\mathfrak{E}_{\mathcal{H}}(x) = \max_{h\in\mathcal{H}} h(x) - \min_{h\in\mathcal{H}} h(x)$ be the extent of $\mathcal{H}$ at point x. Suppose $\mathcal{S}$ is another set of uncertain functions. We say $\mathcal{S}$ is the ε-EXP-KERNEL for $\mathcal{H}$ if $(1-\varepsilon)\mathfrak{E}_{\mathcal{H}}(x) \le \mathfrak{E}_{\mathcal{S}}(x) \le \mathfrak{E}_{\mathcal{H}}(x)$ for any $x \in \mathbb{R}^d$. We say $\mathcal{S}$ is the (ε,τ)-QUANT-KERNEL for $\mathcal{H}$ if $\Pr_{S\sim\mathcal{S}}\Big[\mathfrak{E}_S(x) \le t\Big] \in \Pr_{H\sim\mathcal{H}}\Big[\mathfrak{E}_H(x) \le (1\pm\varepsilon)t\Big] \pm \phi.$ for any $t \ge 0$ and any $x \in \mathbb{R}^d$.

Let us first focus on linear functions in $\mathbb{R}^d$. Using the *duality transformation* that maps linear function $y = a_1x_1 + \ldots + a_dx_d + a_{d+1}$ to the point $(a_1,\ldots,a_{d+1}) \in \mathbb{R}^{d+1}$, we can reduce the extent problem to the directional width problem in $\mathbb{R}^{d+1}$. Let $\mathcal{H}$ be a set of uncertain linear functions (under either existential or locational model) in $\mathbb{R}^d$ for constant d. From Theorem 20 and Corollary 17, we can construct a set S of $O(n^{2d})$ deterministic linear functions in $\mathbb{R}^d$, such that $\mathfrak{E}_S(x) = \mathbb{E}[\mathfrak{E}_{\mathcal{H}}(x)]$ for any $x \in \mathbb{R}^d$. Moreover, for any $\varepsilon > 0$, there exists an ε-EXP-KERNEL of size $O(\varepsilon^{-d/2})$ and an (ε,τ)-QUANT-KERNEL of size $\widetilde{O}(\tau^{-2}\varepsilon^{-d})$. Using the standard linearization technique [7], we can obtain the following generalization for uncertain polynomials.

Theorem 47. *Let $\mathcal{H}$ be a family of uncertain polynomials in $\mathbb{R}^d$ (under either existential or locational model) that admits linearization of dimension k. We can construct a set M of $O(n^{2k})$ deterministic polynomials, such that $\mathfrak{E}_M(x) = \mathbb{E}[\mathfrak{E}_{\mathcal{H}}(x)]$ for any $x \in \mathbb{R}^d$. Moreover, for any $\varepsilon > 0$, there exists an ε-EXP-KERNEL of size $O(\varepsilon^{-k/2})$ and an (ε,τ)-QUANT-KERNEL of size $\min\{\widetilde{O}(\tau^{-2}\max\{\lambda^2,\lambda^4\}), \widetilde{O}(\varepsilon^{-k}\tau^{-2})\}$. Here $\lambda = \sum_{h\in\mathcal{H}}(-\ln(1-p_h))$.*

Now, we consider functions of the form $u(x) = p(x)^{1/r}$ where $p(x)$ is a polynomial and r is a positive integer. We call such a function a *fractional polynomial*. We still use $\mathcal{H}$ to denote the random set of fractional polynomials. Let $\mathcal{H}^\star \subseteq \mathbb{R}^d$ be the set of points such that for any points $x \in \mathcal{H}^\star$ and any function $\vec{u} \in \mathcal{H}$, we have $u(x) \ge 0$. For each point $x \in \mathcal{H}^\star$, we let the random variable $\mathfrak{E}_{r,\mathcal{H}}(x) = \max_{h\in\mathcal{H}} h(x)^{1/r} - \min_{h\in\mathcal{H}} h(x)^{1/r}$. We say another random set $\mathcal{S}$ of functions is the (ε,r)-FPOW-KERNEL for $\mathcal{H}$ if $(1-\varepsilon)\mathfrak{E}_{r,\mathcal{H}}(x) \le \mathfrak{E}_{r,\mathcal{S}}(x) \le \mathfrak{E}_{r,\mathcal{H}}(x)$ for any $x \in \mathcal{H}^\star$. By the duality transformation and Theorem 7, we can obtain the following result.

Theorem 48. *Let $\mathcal{H}$ be a family of uncertain fractional polynomials in $\mathbb{R}^d$ in the*

*existential uncertainty model under the β-assumption. Further assume that each polynomial admits a linearization of dimension k. For any $\varepsilon > 0$, there exists an (ε, r)-*FPOW-KERNEL *of size $\widetilde{O}(\varepsilon^{-(rk-r+2)})$. Furthermore, the (ε, r)-*FPOW-KERNEL *consists of $N = O\big(\varepsilon^{-(rk-r+4)/2}\big)$ sets, each occurring with probability $1/N$ and containing $O\big(\varepsilon^{-r(k-1)/2}\big)$ deterministic fractional polynomials.*

4.5.2 Stochastic Moving Points

We can extend our stochastic models to moving points. In the existential model, each point s is present with probability p_s and follows a trajectory $s(t)$ in $\mathbb{R}^d$ when present ($s(t)$ is the position of s at time t). In the locational model, each point s is associated with a distribution of trajectories (the support size is finite) and the actual trajectory of s is a random sample for the distribution. Such uncertain trajectory models have been used in several applications in spatial databases [111]. For ease of exposition, we assume the existential model in the following. Suppose each trajectory is a polynomial of t with degree at most r. For each point s, any direction $\vec{u}$ and time t, define the polynomial $f_s(\vec{u}, t) = \langle s(t), \vec{u}\rangle$ and let $\mathcal{H}$ include f_s with probability p_s. For a set $\mathcal{P}$ of points, the directional width at time t is $\mathfrak{E}_{\mathcal{H}}(\vec{u}, t) = \max_{s\in\mathcal{P}} f_s(\vec{u}, t) - \min_{s\in\mathcal{P}} f_s(\vec{u}, t)$. Each polynomial f_s admits a linearization of dimension $k = (r+1)d - 1$. Using Theorem 47, we can see that there is a set M of $O(n^{2k})$ deterministic moving points, such that the directional width of M in any direction $\vec{u}$ is the same as the expected directional width of $\mathcal{P}$ in direction $\vec{u}$. Moreover, for any $\varepsilon > 0$, there exists an ε-EXP-KERNEL (which consists of only deterministic moving points) of size $O(\varepsilon^{-(k-1)/2})$ and an (ε, τ)-QUANT-KERNEL (which consists of both deterministic and stochastic moving points) of size $\widetilde{O}(\varepsilon^{-k}\tau^{-2})$.

4.5.3 Shape Fitting Problems

Theorem 47 can be also applied to some stochastic variants of certain shape fitting problems. We first consider the following variant of the minimum enclosing ball problem over stochastic points. We are given a set $\mathcal{P}$ of stochastic points (under either

existential or locational model), find the center point c such that $\mathbb{E}[\max_{s\in\mathcal{P}} \|s-c\|^2]$ is minimized. It is not hard to see that the problem is equivalent to minimizing the expected area of the enclosing ball in $\mathbb{R}^2$. For ease of exposition, we assume the existential model where s is present with probability p_s. For each point $s \in P$, define the polynomial $h_s(x) = \|x\|^2 - 2\langle x, s\rangle + \|s\|^2$, which admits a linearization of dimension $d+1$ [7]. Let $\mathcal{H}$ be the family of uncertain polynomials $\{h_s\}_{s\in\mathcal{P}}$ (h_s exists with probability p_s). We can see that for any $x \in \mathbb{R}^d$, $\max_{s\in\mathcal{P}} \|x-s\|^2 = \max_{h_s\in\mathcal{H}} h_s(x)$. Using Theorem 47, [11] we can see that there is a set M of $O(n^{2d+2})$ deterministic polynomials such that $\max_{h\in M} h(x) = \mathbb{E}[\max_{s\in\mathcal{P}} \|x-s\|^2]$ for any $x \in \mathbb{R}^d$ and a set S of $O(\varepsilon^{-(d+1)/2})$ deterministic polynomials such that $(1-\varepsilon)\mathbb{E}[\max_{s\in\mathcal{P}} \|x-s\|^2] \le \max_{h\in S} h(x) \le \mathbb{E}[\max_{s\in\mathcal{P}} \|x-s\|^2]$ for any $x \in \mathbb{R}^d$. We can store the set S instead of the original point set in order to answer the following queries: given a point s, return the expected length of the furthest point from s. The problem of finding the optimal center c can be also carried out over S, which can be done in $O(\varepsilon^{-O(d^2)})$ time: We can decompose the arrangement of n semialgebraic surfaces in $\mathbb{R}^d$ into $O(n^{O(d+k)})$ cells of constant description complexity, where k is the linearization dimension (see e.g., [12]). By enumerating all those cells in the arrangement of S, we know which polynomials lie in the upper envelopes, and we can compute the minimum value in each such cell in constant time when d is constant.

The above argument can also be applied to the following variant of the spherical shell for stochastic points. We are given a set $\mathcal{P}$ of stochastic points (under either existential or locational model). Our objective is to find the center point c such that $\mathbb{E}[\mathsf{obj}(c)] = \mathbb{E}[\max_{s\in P} \|s-c\|^2 - \min_{s\in P} \|s-c\|^2]$ is minimized. The problem is equivalent to minimizing the expected area of the enclosing annulus in $\mathbb{R}^2$. The objective can be represented as a polynomial of linearization dimension $k = d+1$. Proceeding as for the enclosing balls, we can show there is a set S of $O(\varepsilon^{-(k-1)/2})$ deterministic polynomials such that $(1-\varepsilon)\mathbb{E}[\mathsf{obj}(c)] \le \mathfrak{E}_S(x) \le \mathbb{E}[\mathsf{obj}(c)]$ for any $x \in \mathbb{R}^d$. We would like to make a few remarks here.

[11] We can see from the proof that all results that hold for width/extent also hold for support function/maximum.

1. Let us take the minimum enclosing ball for example. If we examine the construction of set S, each polynomial $h \in S$ may *not* be of the form $h(x) = \|x\|^2 - 2\langle x, s\rangle + \|s\|^2$, therefore does not translate back to a minimum enclosing ball problem over deterministic points.

2. Another natural objective function for the minimum enclosing ball and the spherical shell problem would be the expected radius $\mathbb{E}[\max_{s\in P} \mathrm{d}(s, c)]$ and the expected shell width $\mathbb{E}[\max_{s\in P} \mathrm{d}(s, c) - \min_{s\in P} \mathrm{d}(s, c)]$. However, due to the fractional powers (square roots) in the objectives, simply using an ε-EXP-KERNEL does not work. This is unlike the deterministic setting. [12] We leave the problem of finding small coresets for the spherical shell problem as an interesting open problem. However, under the β-assumption, we can use (ε, r)-FPOW-KERNELs to handle such fractional powers, as in the next subsection.

Theorem 8. *(restated) Suppose $\mathcal{P}$ is a set of n independent stochastic points in $\mathbb{R}^d$ under either existential or locational uncertainty model. There are linear time approximation schemes for the following problems: (1) finding a center point c to minimize $\mathbb{E}[\max_{s\in\mathcal{P}} \|s - c\|^2]$; (2) finding a center point c to minimize $\mathbb{E}[\mathsf{obj}(c)] = \mathbb{E}[\max_{s\in P} \|s - c\|^2 - \min_{s\in P} \|s - c\|^2]$. Note that when $d = 2$ the above two problems correspond to minimizing the expected areas of the enclosing ball and the enclosing annulus, respectively.*

4.5.4 Shape Fitting Problems (Under the β-assumption)

In this subsection, we consider several shape fitting problems in the existential model *under the β-assumption*. We show how to use Theorem 48 to obtain linear time approximation schemes for those problems.

1. (Minimum spherical shell) We first consider the minimum spherical shell problem. Given a set $\mathcal{P}$ of stochastic points (under the β-assumption), our goal is to find the center point c such that $\mathbb{E}[\max_{s\in P} \|s - c\| - \min_{s\in P} \|s - c\|]$ is minimized. For each point $s \in P$, let $h_s(x) = \|x\|^2 - 2\langle x, s\rangle + \|s\|^2$, which admits a

[12] In particular, there is no stochastic analogue of Lemma 4.6 in [7].

linearization of dimension $d+1$. It is not hard to see that $\mathbb{E}[\max_{s\in P}\|s-c\|] = \mathbb{E}[\max_{s\in P}\sqrt{h_s(c)}]$ and $\mathbb{E}[\min_{s\in P}\|s-c\|] = \mathbb{E}[\min_{s\in P}\sqrt{h_s(c)}]$. Using Theorem 48, we can see that there are $N = \widetilde{O}\big(\varepsilon^{-(d+3)}\big)$ sets S_i, each containing $O\big(\varepsilon^{-(d+1)}\big)$ fractional polynomial $\sqrt{h_s}$s such that for all $x \in \mathbb{R}^d$,

$$\frac{1}{N}\sum_{i\in[N]}(\max_{S_i}\sqrt{h_s(x)} - \min_{S_i}\sqrt{h_s(x)}) \in (1\pm\varepsilon)(\mathbb{E}[\max_{s\in P}\|s-x\|] - \mathbb{E}[\min_{s\in P}\|s-x\|]). \tag{4.5}$$

Note that our (ε, r)-FPOW-KERNEL satisfies the subset constraint. Hence, each function $\sqrt{h_s}$ corresponds to an original point in $\mathcal{P}$. So, we can store N point sets $P_i \subseteq \mathcal{P}$, with $|P_i| = O\big(\varepsilon^{-d}\big)$ as the coreset for the original point set. By (4.5), an optimal solution for the coreset is an $(1+\varepsilon)$-approximation for the original problem.

Now, we briefly sketch how to compute the optimal solution for the coreset. Consider all points in $\cup_i P_i$. Consider the arrangement of $O\big(\varepsilon^{-O(d)}\big)$ hyperplanes, each bisecting a pair of points in $\cup_i P_i$. For each cell C of the arrangement, for any point $s \in C$, the ordering of all points in $\cup_i P_i$ is fixed. We then enumerate all those cells in the arrangement and try to find the optimal center in each cell. Fix a cell C. For any point set P_i, we know which point is the furthest one and which point is the closest one from points in C_0. Say they are $s_i = \arg\max_{s\in P_i}\|s-x\|$ and $s_i' = \arg\min_{s\in P_i}\|s-x\|$. Hence, our problem can be formulated as the following optimization problem:

$$\min_x \frac{1}{N}\sum_i (d_i - d_i'), \text{ s.t. } \quad d_i^2 = \|s_i - x\|^2, d_i'^2 = \|s_i' - x\|^2, d_i, d_i' \geq 0, \forall i \in [N]; x \in C_0.$$

The polynomial system has a constant number of variables and constraints, hence can be solved in constant time. More specifically, we can introduce a new variable t and let $t = \frac{1}{N}\sum_i (d_i - d_i')$. All polynomial constraints define a semi-algebraic set. By using constructive version of Tarski-Seidenberg theorem, we can project out all variables except t and the resulting set is still a semi-

algebraic set (which would be a finite collection of points and intervals in $\mathbb{R}^1$) (See e.g.,[20]).

2. (Minimum enclosing cylinder, Minimum cylindrical shell) Let $\mathcal{P}$ be a set of stochastic points in the existential uncertainty model under the β-assumption. Let $\mathrm{d}(\ell, s)$ denote the distance between a point $s \in \mathbb{R}^d$ and a line $\ell \subset \mathbb{R}^d$. The goal for the minimum enclosing cylinder problem is to find a line ℓ such that $\mathbb{E}[\max_{s\in\mathcal{P}} \mathrm{d}(\ell, s)]$ is minimized, while that for the minimum cylindrical shell problem is to minimize $\mathbb{E}[\max_{s\in\mathcal{P}} \mathrm{d}(\ell, s) - \min_{s\in\mathcal{P}} \mathrm{d}(\ell, s)]$. The algorithms for both problems are almost the same and we only sketch the one for the minimum enclosing cylinder problem.

   We follow the approach in [7]. We represent a line $\ell \in \mathbb{R}^d$ by a $(2d-1)$-tuple $(x_1, \ldots, x_{2d-1}) \in \mathbb{R}^{2d-1}$: $\ell = \{p + tq \mid t \in \mathbb{R}\}$, where $p = (x_1, \cdots, x_{d-1}, 0)$ is the intersection point of ℓ with the hyperplane $x_d = 0$ and $q = (x_d, \ldots, x_{2d-1})$, $\|q\|^2 = 1$ is the orientation of ℓ. Then for any point $s \in \mathbb{R}^d$, we have that

$$\mathrm{d}(\ell, s) = \|(p - s) - \langle p - s, q\rangle q\|,$$

   where the polynomial $\mathrm{d}^2(\ell, s)$ admits a linearization of dimension $O(d^2)$. Now, proceeding as for the minimum enclosing ball problem and using Theorem 48, we can obtain a coreset $\mathcal{S}$ consisting $N = O\big(\varepsilon^{-O(d^2)}\big)$ deterministic point sets $P_i \subseteq \mathcal{P}$.

   We briefly sketch how to obtain the optimal solution for the coreset. We can also decompose $\mathbb{R}^{2d-1}$ (a point x in the space with $\|(x_d, \ldots, x_{2d-1})\| = 1$ represents a line in $\mathbb{R}^d$) into $O\big(\varepsilon^{-O(d^2)}\big)$ semi-algebraic cells such that for each cell, the ordering of the points in $\mathcal{S}$ (by their distances to a line in the cell) is fixed. Note that such a cell is a semi-algebraic cell. For a cell C, assume that $s_i = \arg\max_{s\in P_i} \mathrm{d}(\ell, s_i)$ for all $i \in [N]$, where ℓ is an arbitrary line in C. We can

formulate the problem as the following polynomial system:

$$\min_{l} \frac{1}{N} \sum_{i} d_i, \quad \text{s.t.} \quad d_i^2 = \mathrm{d}^2(\ell, s_i), d_i \geq 0, \forall i \in [N]; \ell = (p, q) \in C_0, \|q\|^2 = 1.$$

Again the polynomial system has a constant number of variables and constraints. Thus, we can compute the optimum in constant time.

Theorem 9. *(restated) Suppose $\mathcal{P}$ is a set of n independent stochastic points in $\mathbb{R}^d$, each appearing with probability at least β, for some fixed constant $\beta > 0$. There are linear time approximation schemes for minimizing the expected radius (or width) for the minimum spherical shell, minimum enclosing cylinder, minimum cylindrical shell problems over $\mathcal{P}$.*

4.6 Missing Details in Section 4.2

4.6.1 Details for Section 4.2.1

Lemma 23. We find an affine transform T in $O(2^{O(d)} n \log n)$ time, such that the convex polytope $M' = T(M)$ is α-fat for some constant α.

Proof. By the results in [19], we only need to construct an approximate bounding box, which can be done as follows: We first identify two points y_1 and y_2 in M such that their distance is a constant approximation of the diameter of M. Then we project the points in M to a hyperplane $H \in \mathbb{R}^{d-1}$ perpendicular to the line through y_1 and y_2, and recursively identify two points among the projected points as the approximate diameter. Hence, it suffices to show how to identify such two points y_1 and y_2. Let $\delta = \arccos(1/2)$. Suppose we are working on $\mathbb{R}^d$. We compute a set $\mathcal{I}$ of $O(\delta^{-(d-1)})$ points on the unit sphere $\mathbb{S}^{d-1}$ such that for any point $s \in \mathbb{S}^{d-1}$, there is a point $\vec{u} \in \mathcal{I}$ such that $\angle(\vec{u}, s) \leq \delta$ (see e.g., [10, 25]). From Lemma 22, we know that we can compute for each direction $\vec{u} \in \mathbb{S}^{d-1}$, the point $x(\vec{u}) \in M$ that maximizes $\langle \vec{u}, x(\vec{u}) \rangle$ in $O(n \log n)$ time. For each $\vec{u} \in \mathcal{I}$, compute both $x(\vec{u})$ and $x(-\vec{u})$, and pick the pair that maximizes $\|x(\vec{u}) - x(-\vec{u})\|$. Now, we argue this

is a constant approximation of the diameter. Suppose the diameter of M is (y_1, y_2) where $y_1, y_2 \in M$. Consider the direction $\vec{u}' = (y_1 - y_2)/\|y_1 - y_2\|$. Without loss of generality, assume $y_1 = \arg\max_y \langle y, \vec{u}' \rangle$ and $y_2 = \arg\max_y \langle y, -\vec{u}' \rangle$. Moreover, there is a direction $\vec{u} \in \mathcal{I}$ such that $\angle(\vec{u}, s) \leq \delta$. Therefore, we can get that

$$\begin{aligned}\omega(M, \vec{u}) &= \mathbf{f}(M, \vec{u}) + \mathbf{f}(M, -\vec{u}) \geq \langle y_1, \vec{u} \rangle + \langle y_2, -\vec{u} \rangle \\ &= \langle \vec{u}, y_1 - y_2 \rangle = \|y_1 - y_2\| \cos\angle(\vec{u}, \vec{u}') \geq \|y_1 - y_2\|/2.\end{aligned}$$

In the third equation, we use the simple fact that $\cos\angle(\vec{u}, \vec{u}') = \langle \vec{u}, \vec{u}' \rangle / \|\vec{u}\|\|\vec{u}'\|$. □

Lemma 24. $S = \{x(\vec{u})\}_{\vec{u} \in \mathcal{I}}$ is an ε-kernel for M'.

Proof. Consider an arbitrary vector $s \in \mathbb{S}^{d-1}$ with $\|s\| = 1$. Suppose the point $a \in M'$ maximizes $\langle s, a \rangle$ and $b \in M'$ maximizes $\langle -s, b \rangle$. Hence, $\omega(M', s) = \langle s, a \rangle - \langle s, b \rangle = \langle s, a - b \rangle$. By the construction of $\mathcal{I}$, there is a direction $\vec{u} \in \mathcal{I}$ (with $\|\vec{u}\| = 1$) such that $\|\vec{u} - s\| \leq \delta$. Then, we can see that

$$\begin{aligned}\omega(S, s) &\geq \langle s, x(\vec{u}) \rangle - \langle s, x(-\vec{u}) \rangle = \langle s, x(\vec{u}) - x(-\vec{u}) \rangle \\ &= \langle \vec{u}, x(\vec{u}) - x(-\vec{u}) \rangle + \langle s - \vec{u}, x(\vec{u}) - x(-\vec{u}) \rangle \\ &\geq \langle \vec{u}, a - b \rangle - \|s - \vec{u}\| \|x(\vec{u}) - x(-\vec{u})\| \\ &= \langle s, a - b \rangle + \langle \vec{u} - s, a - b \rangle - \|s - \vec{u}\| \|x(\vec{u}) - x(-\vec{u})\| \\ &\geq \langle s, a - b \rangle - \|s - \vec{u}\| \|x(\vec{u}) - x(-\vec{u})\| - \|\vec{u} - s\| \|a - b\| \\ &\geq \omega(M', s) - O(\delta d) \geq (1 - \varepsilon)\omega(M', s)\end{aligned}$$

In the last and 2nd to last inequalities, we use the fact that M' is α-fat (i.e., $\alpha\overline{\mathbb{C}} \subset M' \subset \overline{\mathbb{C}}$). □

4.6.2 Details for Section 4.2.2

Theorem 26. Under the β-assumption, there is an ε-EXP-KERNEL in $\mathbb{R}^d$ (for $d = O(1)$), which is of size $O(\beta^{-(d-1)}\varepsilon^{-(d-1)/2}\log(1/\varepsilon))$ and satisfies the subset constraint, in the existential uncertainty model.

Proof. Our algorithm is inspired by the peeling idea in [9]. Let $\varepsilon_1 = \varepsilon\alpha\beta^2/4\sqrt{d}$, where α is a constant defined later. We repeat the following for $L = O(\log_{1-\beta}\varepsilon_1) = O(\log(1/\varepsilon))$ rounds: In round i, we first compute an $(\varepsilon_1/\sqrt{d})$-kernel $\mathcal{S}_i$ (of size $O((\sqrt{d}/\varepsilon_1)^{(d-1)/2}) = O(\beta^{-(d-1)}\varepsilon^{-(d-1)/2})$) for the remaining points (in the deterministic sense) and then delete all points of $\mathcal{S}_i$. Let $\mathcal{S} = \cup_i \mathcal{S}_i$. Now, we show that $\mathcal{S}$ is an ε-EXP-KERNEL for $\mathcal{P}$.

We first establish a lower bound of $\omega(\mathcal{P}, \vec{u})$ for any unit vector $\vec{u} \in \mathbb{S}^{d-1}$. Assume without loss of generality that $\alpha\overline{\mathbb{C}} \subset \mathsf{ConvH}(\mathcal{P}) \subset \overline{\mathbb{C}}$ where $\overline{\mathbb{C}} = [-1, 1]^d$ and α is a constant only depending on d. Since $\alpha\overline{\mathbb{C}} \subset \mathsf{ConvH}(P)$, we know there is a point $s \in \mathsf{ConvH}(P)$ such that $\langle \vec{u}, s\rangle \geq \alpha$ and a different point $s' \in \mathsf{ConvH}(P)$ such that $\langle \vec{u}, s'\rangle \leq -\alpha$. Hence, we have that

$$\omega(\mathcal{P}, \vec{u}) \geq \beta^2(\langle \vec{u}, s\rangle - \langle \vec{u}, s'\rangle) \geq 2\alpha\beta^2.$$

Fix an arbitrary direction $\vec{u} \in \mathbb{S}^{d-1}$. Now, we bound the difference between $\mathbf{f}(\mathcal{P}, \vec{u})$ and $\mathbf{f}(\mathcal{S}, \vec{u})$. We show that for any real value $x \in [-\sqrt{d}, \sqrt{d}]$,

$$\Pr_{P\sim\mathcal{P}}[\mathbf{f}(P, \vec{u}) \geq x] \leq \Pr_{S\sim\mathcal{S}}[\mathbf{f}(S, \vec{u}) \geq x - \varepsilon_1] + \varepsilon_1. \tag{4.6}$$

In fact, a proof of the above statement provides a proof for Theorem 43 (i.e., $\mathcal{S}$ is an (ε, τ)-QUANT-KERNEL as well).

Let $\mathcal{L}_{\mathcal{P}} = \{s_1, s_2, \ldots, s_L\}$ be the set of L points $s \in \mathcal{P}$ that maximize $\langle s, \vec{u}\rangle$ (i.e., the first L vertices in the canonical order w.r.t. $\vec{u}$). Similarly, let $\mathcal{L}_{\mathcal{S}} = \{s'_1, s'_2, \ldots, s'_L\}$ be the set of L points $s' \in \mathcal{S}$ that maximize $\langle s', \vec{u}\rangle$. We distinguish two cases:

1. $\mathcal{L}_{\mathcal{P}} = \mathcal{L}_{\mathcal{S}}$: If $x \geq \langle \vec{u}, s_L\rangle$, we can see that $\Pr_{P\sim\mathcal{P}}[\mathbf{f}(P, \vec{u}) \geq x] = \Pr_{S\sim\mathcal{S}}[\mathbf{f}(S, \vec{u}) \geq x]$. If $x < \langle \vec{u}, s_L\rangle$, both $\Pr_{P\sim\mathcal{P}}[\mathbf{f}(P, \vec{u}) \geq x]$ and $\Pr_{S\sim\mathcal{S}}[\mathbf{f}(S, \vec{u}) \geq x]$ are at least $1 - \prod_{s\in\mathcal{L}_{\mathcal{P}}}(1 - p_s) \geq 1 - (1 - \beta)^L \geq 1 - \varepsilon_1$.

2. Suppose j is the smallest index such that $s_j \neq s'_j$. For $x > \langle \vec{u}, s_j\rangle$, we can see that $\Pr_{P\sim\mathcal{P}}[\mathbf{f}(P, \vec{u}) \geq x] = \Pr_{S\sim\mathcal{S}}[\mathbf{f}(S, \vec{u}) \geq x]$. Now, we focus on the case where $x \leq \langle \vec{u}, s_j\rangle$. From the construction of $\mathcal{S}$, we can see that $\langle s'_{j'}, \vec{u}\rangle \geq \langle s_j, \vec{u}\rangle - \varepsilon_1$

for all $j' \geq j$. [13] Hence, for $x \leq \langle \vec{u}, s_j \rangle$, we can see that

$$\Pr_{S\sim\mathcal{S}}[\mathbf{f}(S,\vec{u}) \geq x - \varepsilon_1] \geq 1 - \prod_{s\in\mathcal{L}_\mathcal{S}} (1-p_s) \geq 1 - \varepsilon_1.$$

So, in either case, (4.6) is satisfied. We also need the following basic fact about the expectation: For a random variable X, if $\Pr[X \geq a] = 1$, then $\mathbb{E}[X] = \int_b^\infty \Pr[X \geq x]\mathrm{d}x + b$ for any $b \leq a$. Since $-\sqrt{d} \leq \mathbf{f}(P,\vec{u}) \leq \sqrt{d}$ for any realization P, we have

$$\begin{aligned}
\mathbf{f}(\mathcal{P},\vec{u}) &= \int_{-\sqrt{d}}^{\infty} \Pr_{P\sim\mathcal{P}}[\mathbf{f}(P,\vec{u}) \geq x]\mathrm{d}x - \sqrt{d} \\
&\leq \int_{-\sqrt{d}}^{\infty} \Pr_{S\sim\mathcal{S}}[\mathbf{f}(S,\vec{u}) \geq x - \varepsilon_1]\mathrm{d}x + 2\sqrt{d}\varepsilon_1 - \sqrt{d} \\
&\leq \int_{-\sqrt{d}-\varepsilon_1}^{\infty} \Pr_{S\sim\mathcal{S}}[\mathbf{f}(S,\vec{u}) \geq x]\mathrm{d}x - \sqrt{d} - \varepsilon_1 + 3\sqrt{d}\varepsilon_1 \\
&= \mathbf{f}(\mathcal{S},\vec{u}) + 3\sqrt{d}\varepsilon_1,
\end{aligned}$$

where the only inequality is due to (4.6) and the fact that $\Pr_{P\sim\mathcal{P}}[\mathbf{f}(P,\vec{u}) \geq x] = \Pr_{S\sim\mathcal{S}}[\mathbf{f}(S,\vec{u}) \geq x] = 0$ for $x > 1$. Similarly, we can get that $\mathbf{f}(\mathcal{S},-\vec{u}) \geq \mathbf{f}(\mathcal{P},-\vec{u}) - 3\varepsilon_1\sqrt{d}$. By the choice of ε_1, we have that $6\sqrt{d}\varepsilon_1 \leq \varepsilon \cdot 2\alpha\beta^2 \leq \varepsilon\omega(\mathcal{P},\vec{u})$. Hence, $\omega(\mathcal{S},\vec{u}) \geq \omega(\mathcal{P},\vec{u}) - 6\sqrt{d}\varepsilon_1 \geq (1-\varepsilon)\omega(\mathcal{P},\vec{u})$. □

4.6.3 Locational uncertainty

Similar results are possible for uncertain points with locational uncertainty. Let $\mathcal{V} = \{v_1,\ldots,v_m\}$ be the set of nodes and $\mathcal{P} = \{s_1,\ldots,s_n\}$ be the collection of possible locations. Now there are n possible locations, and thus $\binom{n}{2}$ hyperplanes Γ that partition $\mathbb{R}^d$. We can replicate all bounds in this setting, except that m replaces n in each bound. The main difficulty is in replicating Lemma 49 that given a direction $\vec{u}$ calculates the vertex of M; for locational uncertain points this is described in Lemma 51. Moreover, the $O(n^2 \log n)$ bound for $\mathbb{R}^2$ is also carefully described in Lemma 52.

In the locational uncertainty model, Lemma 18 also holds with a stronger general

[13] To see this, consider the round in which $s'_{j'}$ is chosen. Let $\hat{s}$ be the vertex minimizing $\langle \hat{s},\vec{u}\rangle$. As s_j is not chosen, we must have $\langle s'_{j'},\vec{u}\rangle - \langle \hat{s},\vec{u}\rangle \geq (1-\varepsilon_1/\sqrt{d})(\langle s_j,\vec{u}\rangle - \langle \hat{s},\vec{u}\rangle)$.

position assumption. With the new general position assumption, it is straightforward to show that the gradient vector is different for two adjacent cones in $\mathbb{A}(\Gamma)$. Other parts of the proof is essentially the same as Lemma 18. The details can be found below. Theorem 2 also holds for the locational model without any change in the proof (the running time becomes $O(n \log^2 n)$).

Now, we prove that Lemma 18 also holds for the locational model. For this purpose, we need a stronger general position assumption: (1) For any $v \in \mathcal{V}$, $\sum_{s \in \mathcal{P}} p_{vs} \in (0, 1)$. This suggests that we need to consider the model with both existential and locational uncertainty. We can make this assumption hold by subtracting an infinitesimal value from each probability value without affecting the directional width in any essential way. (2) For any two nodes $v_1, v_2 \in \mathcal{V}$, two locations $s_1, s_2 \in \mathcal{P}$ and two subsets of locations $S_1, S_2 \subseteq \mathcal{P}$, $p_{v_1 s_1}(\sum_{s \in S_2} p_{v_2 s}) s_1 \neq p_{v_2 s_2}(\sum_{s \in S_1} p_{v_1 s}) s_2$ (this is indeed a general position assumption since we only have a finite number of equations to exclude but uncountable number of choices of the positions).

Lemma 18. (for the locational model)**.** Assuming the locational model and the above general position assumption, the complexity of M is the same as the cardinality of $\mathbb{A}(\Gamma)$, i.e., $|M| = |\mathbb{A}(\mathcal{P})|$. Moreover, each cone $C \in \mathbb{A}(\mathcal{P})$ corresponds to exactly one vertex s of $\mathsf{ConvH}(M)$ in the following sense: $\nabla \mathbf{f}(M, \vec{u}) = s$ for all $\vec{u} \in \operatorname{int} C$.

Proof. The proof is almost the same as that for Lemma 18 except that we need to show $\mathbf{f}(M, \vec{u})$ is different for two adjacent cones in $\mathbb{A}(\Gamma)$. Again, let C_1, C_2 be two adjacent cones separated by some hyperplane in Γ. Suppose $\vec{u}_1 \in \operatorname{int} C_1$ and $\vec{u}_2 \in \operatorname{int} C_2$. Consider the canonical orders O_1 and O_2 of $\mathcal{P}$ with respect to $\vec{u}_1$ and $\vec{u}_2$ respectively. W.l.o.g., assume that $O_1 = \{s_1, \ldots, s_i, s_{i+1}, \ldots, s_n\}$ and $O_1 = \{s_1, \ldots, s_{i+1}, s_i, \ldots, s_n\}$.

Let $\Pr^R(v, s, \vec{u})$ be the probability that the largest point along $\vec{u}$ is uncertain node $v \in \mathcal{V}$ at location $s \in \mathcal{P}$. Using the notations from Lemma 51, $\mathbf{f}(\mathcal{V}, \vec{u})$ can computed by $\sum_{v \in \mathcal{V}, s \in \mathcal{P}} \Pr^R(v, s, \vec{u}) \langle s, \vec{u} \rangle$. Hence, $\nabla \mathbf{f}(\mathcal{V}, \vec{u}) = \sum_{v \in \mathcal{V}, s \in \mathcal{P}} \Pr^R(v, s, \vec{u}) s$.

Suppose s_i is a possible location for v_1 and s_{i+1} is a possible location for v_2. Denote by $\mathcal{P}^R(s, \vec{u})$ the subset of $s' \in \mathcal{P}$ such that $\langle s', \vec{u} \rangle > \langle s, \vec{u} \rangle$ and denote by $\Pr^R_\emptyset(s, \vec{u})$ as the probability that no node $v \in \mathcal{V}$ appears at a larger location than

$s \in \mathcal{P}$ along direction $\vec{u}$. If $v_1 \neq v_2$, we have $\nabla \mathbf{f}(\mathcal{V}, \vec{u}_1) - \nabla \mathbf{f}(\mathcal{V}, \vec{u}_2) = \Pr^R_\emptyset(s_i, \vec{u}_1) \cdot \left(s_i p_{v_1 s_i} \sum_{s' \in \mathcal{P}^R(s, \vec{u}_1)} p_{v_2 s'} - s_{i+1} p_{v_2 s_{i+1}} \sum_{s' \in \mathcal{P}^R(s, \vec{u}_2)} p_{v_1 s'} \right) \neq 0$. If $v_1 = v_2 = v$, we have $\nabla \mathbf{f}(\mathcal{V}, \vec{u}_1) - \nabla \mathbf{f}(\mathcal{V}, \vec{u}_2) = \Pr^R_\emptyset(s_i, \vec{u}_1) \cdot \left(s_i p_{v s_i} - s_{i+1} p_{v s_{i+1}} \right) \neq 0$. □

4.7 Missing Details in Section 4.3

Theorem 36. Let $\tau_1 = O(\frac{\tau}{\max\{\lambda, \lambda^2\}})$ and $N = O(\frac{1}{\tau_1^2} \log \frac{1}{\delta}) = O(\frac{\max\{\lambda^2, \lambda^4\}}{\tau^2} \log \frac{1}{\delta})$. With probability at least $1 - \delta$, for any $t \geq 0$ and any direction $\vec{u}$, we have that $\Pr\left[\omega(\mathcal{S}, \vec{u}) \leq t\right] \in \Pr\left[\omega(\mathcal{P}, \vec{u}) \leq t\right] \pm \tau$.

Proof. Fix an arbitrary direction $\vec{u}$ (w.l.o.g., say it is the x-axis) and rename all points in $\mathcal{P}$ as $s_1, s_2, \ldots, s_n$ as before. Consider the Poissonized instance of $\mathcal{P}$. Let $s'_1, \ldots, s'_N$ be the N points in $\mathcal{S}$ (also sorted in nondecreasing order of their x-coordinates). Now, we create a coupling between all mass in $\mathfrak{A}$ and that in $\mathfrak{B}$, as follows. We process all points in $\mathfrak{A}$ from left to right, starting with s_1. The process has N rounds. In each round, we assign exactly $1/N$ units of mass in $\mathfrak{A}$ to a point in $\mathfrak{B}$. In the first round, if s_1 contains less than $1/N$ units of mass, we proceed to $s_2, s_3, \ldots s_i$ until we reach $1/N$ units collectively. We split the last point s_i into two points s_{i1} and s_{i2} so that the mass contained in $s_1, \ldots, s_{i-1}, s_{i1}$ is exactly $1/N$, and we assign those points to s'_1. We start the next round with s_{i2}. If s_1 contains more than $1/N$ units of mass, we split s_1 into s_{11} (s_{11} contains $1/N$ units) and s_{12} and we start the second round with s_{12}. We repeat this process until all mass in $\mathfrak{A}$ is assigned.

The above coupling can be viewed as a mass transportation from $\mathfrak{A}$ to $\mathfrak{B}$. We will need one simple but useful property about this transportation: for any vertical line $x = t$, at most τ_1 units of mass are transported across the vertical line (by Theorem 32).

In the construction of the coupling, many points in $\mathfrak{A}$ may be split. We rename them to be $s_1, \ldots, s_L$ (according to the order in which they are processed). The sequence $s_1, \ldots, s_L$ can be divided into N segments, each assigned to a point in $\mathcal{S}$. For a point s'_i in $\mathcal{S}$, let $\mathsf{seg}(i)$ be the segment (the set of points) assigned to s'_i. For any node s and real $t > 0$, we use $H(s, t)$ to denote the right open halfplane defined

by the vertical line $x = x(s) + t$, where $x(s)$ is the x-coordinate of s (see Figure 4-3).

Let X_i (Y_i resp.) be the Poisson distributed random variable corresponding to s_i (s'_i resp.) (i.e., $X_i \sim \text{Pois}(\lambda_{s_i})$ and $Y_i \sim \text{Pois}(\lambda/N)$) for all i. For any $H \subset \mathbb{R}^2$, we write $X(H) = \sum_{s_i \in H \cap \mathcal{P}} X_i$ and $Y(H) = \sum_{s'_i \in H \cap \mathcal{S}} Y_i$. We can rewrite $\Pr[\omega(\mathcal{S}, \vec{u}) \le t]$ as follows:

$$\begin{aligned}\Pr[\omega(\mathcal{S}, \vec{u}) \le t] &= \sum_{i=1}^{N} \Pr[s'_i \text{ is the leftmost point and } \omega(\mathcal{S}, \vec{u}) \le t] + \Pr[\text{no point in } \mathcal{S} \text{ appears}] \\ &= \sum_{i=1}^{N} \Pr[Y_i \neq 0] \Pr\Big[\sum_{j=1}^{i-1} Y_j = 0\Big] \Pr[Y(H(s'_i, t)) = 0] + \Pr[\sum_{s'_i \in \mathcal{S}} Y_i = 0]\end{aligned} \tag{4.7}$$

Similarly, we can write that [14]

$$\begin{aligned}\Pr[\omega(\mathcal{P}, \vec{u}) \le t] &= \sum_{i=1}^{m} \Pr[X_i \neq 0] \Pr\Big[\sum_{j=1}^{i-1} X_j = 0\Big] \Pr[X(H(s_i, t)) = 0] + \Pr[\sum_{s_i \in \mathcal{P}} X_i = 0] \\ &= \sum_{i=1}^{N} \sum_{k \in \mathsf{seg}(i)} \Pr[X_k \neq 0] \Pr\Big[\sum_{j=1}^{k-1} X_j = 0\Big] \Pr[X(H(s_k, t)) = 0] + \Pr[\sum_{s_i \in \mathcal{P}} X_i = 0]\end{aligned} \tag{4.8}$$

We proceed by attempting to show each each summand of (4.8) is close to the corresponding one in (4.7). First, we can see that $\Pr[\sum_{s'_i \in \mathcal{S}} Y_i = 0] = \Pr[\sum_{s_i \in \mathcal{P}} X_i = 0]$ since both $\sum_{s'_i \in \mathcal{S}} Y_i$ and $\sum_{s_i \in \mathcal{P}} X_i$ follow the Poisson distribution $\text{Pois}(\lambda)$.

For any segment i, we can see that $\sum_{k \in \mathsf{seg}(i)} \lambda_{s_k} = \lambda/N$. Moreover, we have $\lambda_{s_k} \le \lambda/N \le \tau/32$, thus $\exp(-\lambda_{s_k}) \in (1 - \lambda_{s_k}, (1 + \tau/16)(1 - \lambda_{s_k}))$.

$$\begin{aligned}\sum_{k \in \mathsf{seg}(i)} \Pr[X_k \neq 0] &= \sum_{k \in \mathsf{seg}(i)} (1 - \exp(-\lambda_{s_k})) \in (1 \pm \frac{\tau}{16}) \sum_{k \in \mathsf{seg}(i)} \lambda_{s_k} \\ &\subset (1 \pm \frac{\tau}{8})(1 - \exp(\frac{\lambda}{N})) = (1 \pm \frac{\tau}{8}) \Pr[Y_i \neq 0].\end{aligned} \tag{4.9}$$

[14] Note that splitting nodes does not change the distribution of $\omega(\mathcal{P}, \vec{u})$: Suppose a node s (corresponding to r.v. X) was spit to two nodes s_1 and s_2 (corresponding to X_1 and X_2 resp.). We can see that $\Pr[X \neq 0] = \Pr[X_1 \neq 0 \text{ and } X_2 \neq 0] = \Pr[X_1 + X_2 \neq 0]$.

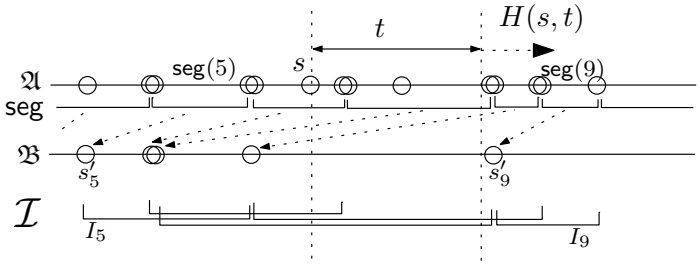

Figure 4-3: Illustration of the interval graph $\mathcal{I}$. For illustration purpose, co-located points (e.g., points that are split in $\mathfrak{A}$) are shown as overlapping points. The arrows indicate the assignment of the segments to the points in $\mathfrak{B}$. Theorem 32 ensures that any vertical line can not stab many intervals.

Then, we notice that for any $k \in \mathsf{seg}(i)$ (i.e., s_k is in the segment assigned to s'_i), it holds that

$$\Pr\Big[\sum_{j=1}^{k} X_j = 0\Big] \in [e^{-i\lambda/N}, e^{-\lambda(i-1)/N}] \subset (1 \pm \frac{\tau}{8})e^{-\lambda(i-1)/N} = (1 \pm \frac{\tau}{8})\Pr\Big[\sum_{j=1}^{i-1} Y_j = 0\Big]. \tag{4.10}$$

The first inequality holds because $\sum_{j=1}^{k} X_j \sim \mathrm{Pois}\big(\sum_{j=1}^{k} \lambda_{s_j}\big)$ and $\lambda(i-1)/N \leq \sum_{j=1}^{k} \lambda_{s_j} \leq \lambda i/N$.

If we can show that $\Pr[X(H(s_k,t)) = 0]$ is close to $\Pr[Y(H(s'_i,t)) = 0]$ for $k \in \mathsf{seg}(i)$, we can finish the proof easily since each summand of (4.8) would be close to the corresponding one in (4.7). However, this is in general not true and we have to be more careful.

Recall that the sequence $s_1, \ldots, s_L$ is divided into N segments. Let $K = \lambda/\tau$. We say that the ith segment (say $\mathsf{seg}(i) = \{s_j, s_{j+1}, \ldots, s_k\}$) is a *good segment* if

$$E_i = \max\Big\{\big|\mathfrak{B}(H(s'_i,t)) - \mathfrak{A}(H(s_j,t))\big|, \big|\mathfrak{B}(H(s'_i,t)) - \mathfrak{A}(H(s_k,t))\big|\Big\} \leq \frac{1}{K}.$$

Otherwise, the segment is *bad*. For a good segment $\mathsf{seg}(i)$ and any $k \in \mathsf{seg}(i)$,

$$\Pr[X(H(s_k,t)) = 0] = \exp\big(-\lambda\mathfrak{A}(H(s_k,t))\big) \in \exp\big(-\lambda\mathfrak{B}(H(s'_i,t)) \pm \lambda/K\big)$$

$$\subset \Pr[Y(H(s'_i,t))=0]e^{\pm\lambda/K} \subset \Pr[Y(H(s'_i,t))=0](1\pm\tau/8). \tag{4.11}$$

We use Gs to denote the set of good segments and Bs the set of bad segments. Now, we consider the summations in both (4.7) and (4.8) with only good segments. We have that

$$\begin{aligned}
&\sum_{i\in\mathsf{Gs}}\sum_{k\in\mathsf{seg}(i)} \Pr[X_k\neq 0]\Pr\Big[\sum_{j=1}^{k-1}X_j=0\Big]\Pr[X(H(s_k,t))=0]\\
\in\ &\sum_{i\in\mathsf{Gs}}\Pr\Big[\sum_{j=1}^{i-1}Y_j=0\Big](1\pm\tau/8)\Pr[Y(H(s'_i,t))=0](1\pm\tau/8)\Pr[Y_i\neq 0](1\pm\tau/8)\\
\subset\ &\sum_{i\in\mathsf{Gs}}\Pr[Y_i\neq 0]\Pr\Big[\sum_{j=1}^{i-1}Y_j=0\Big]\Pr[Y(H(s'_i,t))=0]\pm\tau/2,
\end{aligned}$$

where the first inequality is due to (4.10) and (4.11) and the second holds because (4.9)

Now, we show the total contributions of bad segments to both (4.7) and (4.8) are small. We partition all of the bad segments into $\log(1/\tau)+1$ different sets $B_0,\ldots,B_{\log\frac{1}{\tau}-1}$. Let $B_i=\{i \mid \frac{2^i}{K}<E_i\leq\frac{2^{i+1}}{K}\}$ for $0\leq i\leq\log(1/\tau)-1$. Let $B_{\log\frac{1}{\tau}}=\{i\mid E_i>\frac{1}{\lambda}\}$.

With the above notations, we prove the following crucial inequality:

$$\sum_{i=0}^{\log\frac{1}{\tau}}|B_i|\cdot 2^i = O(\tau_1 NK). \tag{4.12}$$

Now, we prove (4.12). Consider all points $s_1,\ldots,s_L$ and $s'_1,\ldots,s'_L$ lying on the same x-axis. For each i (with $\mathsf{seg}(i)=\{s_j,s_{j+1},\ldots,s_k\}$), we draw the minimal interval I_i that contains s'_i,s_j and s_k. If the ith segment is bad and belongs to B_j, we also say I_i is a *bad interval* of label j. All intervals $\{I_i\}_i$ define an interval graph $\mathcal{I}$. We can see that any vertical line can stab at most $\tau_1 N+1$ intervals, because at most τ_1 unit of mass can be transported across the vertical line, and each interval is responsible for a transportation of exactly $1/N$ units of mass (except the one that

intersects the vertical line). Hence, the interval graph $\mathcal{I}$ can be colored with at most $\tau_1 N + 1$ colors (this is because the clique number of $\mathcal{I}$ is at most $\tau_1 N + 1$ and the chromatic number of an interval graph is the same as its clique number). Consider a color class C (which consists of a set of non-overlapping intervals). Imagine we move an interval I of length t along the x-axis from left to right. When the left endpoint of I passes through an bad interval of label j in C, by the definition of bad segments, the right endpoint of I passes through $\Omega(2^j N/K)$ segments. Suppose the color class C contains b_j bad segments in B_j. Since the right endpoint of I can pass through at most N segments, we have the following inequalities by summing over all labels:

$$\sum_{j=0}^{\log \frac{1}{\tau}-1} b_j \cdot 2^j N/K \leq N.$$

Summing up all color classes, we obtain (4.12).

For B_j $(0 \leq j \leq \log \frac{1}{\tau})$, we can bound the total contribution as follows. By the definition of B_j, we can see that

$$\sum_{i \in B_j} \sum_{k \in \mathsf{seg}(i)} \Pr[X_k \neq 0] \Pr\Big[\sum_{j=1}^{k-1} X_j = 0\Big] \Pr[X(H(s_k, t)) = 0]$$
$$\subset (1 \pm 5 \cdot 2^j \tau) \sum_{i \in B_j} \Pr[Y_i \neq 0] \Pr\Big[\sum_{j=1}^{i-1} Y_j = 0\Big] \Pr[Y(H(s'_i, t)) = 0],$$

Thus, the total contribution of bad segments in B_j $(0 \leq j \leq \log \frac{1}{\tau} - 1)$ to the corresponding summands in ((4.7)-(4.8)) is at most

$$5 \cdot 2^j \tau \sum_{i \in B_j} \Pr[Y_i \neq 0] = 5|B_j| \cdot 2^j \tau \times (1 - \exp(-\frac{\lambda}{N})) = O(|B_i| 2^j \tau \lambda / N),$$

where $\Pr[Y_i \neq 0] = 1 - \exp(-\frac{\lambda}{N})$ (since $Y_i \sim \mathrm{Pois}(\frac{\lambda}{N})$).

For $B_{\log \frac{1}{\tau}}$, the total contribution is bounded by the following.

$$\Big| \sum_{i \in B_{\log \frac{1}{\tau}}} \Big(\sum_{k \in \mathsf{seg}(i)} \Pr[X_k \neq 0] \Pr\Big[\sum_{j=1}^{k-1} X_j = 0\Big] \Pr[X(H(s_k, t)) = 0]$$

$$
\begin{aligned}
&\quad - \Pr[Y_i \neq 0] \Pr\Big[\sum_{j=1}^{i-1} Y_j = 0\Big] \Pr[Y(H(s_i', t)) = 0]\Big)\Big| \\
&\leq \sum_{i \in B_{\log \frac{1}{\tau}}} \sum_{k \in \mathsf{seg}(i)} \Pr[X_k \neq 0] + \sum_{i \in B_{\log \frac{1}{\tau}}} \Pr[Y_i \neq 0] \leq 3 \sum_{i \in B_{\log \frac{1}{\tau}}} \Pr[Y_i \neq 0] \\
&\leq 3|B_{\log \frac{1}{\tau}}| \times (1 - \exp(-\frac{\lambda}{N}) = O(|B_{\log \frac{1}{\tau}}|\lambda/N)
\end{aligned}
$$

Summing up all j and using (4.12), we obtain the following inequality .

$$\sum_{j=0}^{\log \frac{1}{\tau}} O(|B_j| 2^j \tau \lambda / N) = O(\tau_1 \tau \lambda K) \leq \frac{\tau}{4}.$$

This finishes the proof. □

4.8 Missing Details in Section 4.4

Lemma 45. Let $N = O\big(\varepsilon_1^{-2} \varepsilon_0^{-(d-1)/2} \log(1/\varepsilon_0)\big)$, where $\varepsilon_0 = (\varepsilon/4(r-1))^r$, $\varepsilon_1 = \varepsilon\beta^2$. For any $t \geq 0$ and any direction $\vec{u} \in \mathcal{P}^\star$, we have that

$$\Pr_{P \sim \mathcal{S}}\Big[\max_{s \in P} \langle \vec{u}, s\rangle^{1/r} \geq t\Big] \in \Pr_{P \sim \mathcal{P}}\Big[\max_{s \in \mathcal{E}(P)} \langle \vec{u}, s\rangle^{1/r} \geq t)\Big] \pm \varepsilon_1/4, \text{ and}$$

$$\Pr_{P \sim \mathcal{S}}\Big[\min_{s \in P} \langle \vec{u}, s\rangle^{1/r} \geq t\Big] \in \Pr_{P \sim \mathcal{P}}\Big[\min_{s \in \mathcal{E}(P)} \langle \vec{u}, s\rangle^{1/r} \geq t)\Big] \pm \varepsilon_1/4.$$

Proof. The argument is almost the same as that in Lemma 30. Let $L = O(\varepsilon_0^{-(d-1)/2})$. We still build a mapping g that maps each realization $\mathcal{E}(P)$ to a point in $\mathbb{R}^{dL}$, as follows: Consider a realization P of $\mathcal{P}$. Suppose $\mathcal{E}(P) = \{(x_1^1, \ldots, x_d^1), \ldots, (x_1^L, \ldots, x_d^L)\}$ (if $|\mathcal{E}(P)| < L$, we pad it with $(0, \ldots, 0)$). We let $g(\mathcal{E}(P)) = (x_1^1, \ldots, x_d^1, \ldots, x_1^L, \ldots, x_d^L) \in \mathbb{R}^{dL}$. For any $t \geq 0$ and any direction $\vec{u} \in \mathcal{P}^\star$, note that $\max_{s \in \mathcal{E}(P)} \langle \vec{u}, s\rangle^{1/r} \geq t$ holds if and only if there exists some $1 \leq i \leq |\mathcal{E}(P)|$ satisfies that $\sum_{j=1}^d x_j^i \vec{u}_j \geq t^r$, which is equivalent to saying that point $g(\mathcal{E}(P))$ is in the union of the those $|\mathcal{E}(P)|$ half-spaces.

Let X be the image set of g. Let $(X, \mathcal{R}^i)$ $(1 \leq i \leq L))$ be a range space, where $\mathcal{R}^i$ is the set of half spaces $\{\sum_{j=1}^d x_j^i \vec{u}_j \geq t \mid u = (\vec{u}_1, \ldots, \vec{u}_d) \in \mathbb{R}^d, t \geq 0\}$. Let

$\mathcal{R}' = \{\cup r_i \mid r_i \in \mathcal{R}^i, i \in [L]\}$. Note that each $(X, \mathcal{R}^i)$ has VC-dimension $d+1$. By Theorem 28, we have that the VC-dimension of $(X, \mathcal{R}')$ is bounded by $O((d+1)L \lg L) = O(\varepsilon_0^{-(d-1)/2} \log(1/\varepsilon_0))$. Then by Theorem 29, for any t and any direction $\vec{u}$, we have that $\Pr_{P\sim\mathcal{S}}[\max_{s\in P}\langle \vec{u}, s\rangle^{1/r} \geq t] \in \Pr_{P\sim\mathcal{P}}[\max_{s\in\mathcal{E}(P)}\langle \vec{u}, s\rangle^{1/r} \geq t)] \pm \varepsilon_1/4$. The proof for the second statement is the same. □

4.9 Computing the Expected Direction Width

We handle both the existential and location model of uncertain points in this section. For any direction $\vec{u}$, denote by $\omega(\mathcal{P}, \vec{u})$ the expected width of $\mathcal{P}$ along the direction $\vec{u}$, and $\mathbf{f}(\mathcal{P}, \vec{u}) = \mathbb{E}_{P\sim\mathcal{P}}[\max_{p\in P}\langle \vec{u}, p\rangle]$ is the support function. Recall $\omega(\mathcal{P}, \vec{u}) = \mathbf{f}(\mathcal{P}, \vec{u}) - \mathbf{f}(\mathcal{P}, -\vec{u})$ by linearity of expectation.

4.9.1 Computing Expected Width for Existential Uncertainty

The existential model is a bit simpler and we handle that first. Recall in this model we let $\mathcal{P}$ be a set of n uncertain points, and each point $s \in \mathcal{P}$ has a probability p_s. We have the following two lemmas.

Lemma 49. *For any direction $\vec{u}$, we can compute $\omega(\mathcal{P}, \vec{u})$, $\mathbf{f}(\mathcal{P}, \vec{u})$, and $\nabla\mathbf{f}(\mathcal{P}, \vec{u})$ in $O(n \log n)$ time; if the points of $\mathcal{P}$ are already sorted along the direction $\vec{u}$, then we can compute them in $O(n)$ time.*

Proof. Consider any direction $\vec{u}$. Without loss of generality, assume $\|\vec{u}\| = 1$. In the following, we first show how to compute $\mathbf{f}(\mathcal{P}, \vec{u})$. The value $\mathbf{f}(\mathcal{P}, -\vec{u})$ can be computed in a similar manner and we ignore the discussion. After having $\mathbf{f}(\mathcal{P}, \vec{u})$ and $\mathbf{f}(\mathcal{P}, -\vec{u})$, $\omega(\mathcal{P}, \vec{u})$ can be computed immediately by $\omega(\mathcal{P}, \vec{u}) = \mathbf{f}(\mathcal{P}, \vec{u}) - \mathbf{f}(\mathcal{P}, -\vec{u})$. Finally, we will discuss how to compute $\nabla\mathbf{f}(\mathcal{P}, \vec{u})$. Let $\rho(\vec{u})$ be the ray of direction $\vec{u}$ in the plane passing through the origin.

Consider a point $s \in \mathcal{P}$. Note that $\langle s, \vec{u}\rangle$ is the coordinate of the perpendicular projection of s on $\rho(\vec{u})$. Denote by $\mathcal{P}^R(s, \vec{u})$ the subset of points $s' \in \mathcal{P}$ such that $s' >_{\vec{u}} s$ (i.e., $\langle s', \vec{u}\rangle > \langle s, \vec{u}\rangle$). Denote by $\Pr^R(s, \vec{u})$ the probability that s appears in a

realization but all points of $\mathcal{P}^R(s, \vec{u})$ do not appear (i.e., $\langle s, \vec{u}\rangle$ is the largest among all points of $\mathcal{P}$ that appear in the realization). Hence, we have

$$\Pr^R(s, \vec{u}) = p_s \cdot \prod_{s' >_{\vec{u}} s} (1 - p_{s'}). \tag{4.13}$$

Now $\mathbf{f}(\mathcal{P}, \vec{u})$ can be seen as the expected largest coordinate of the projections of the points in $\mathcal{P}$ on $\rho(\vec{u})$. According to the definition of $\Pr^R(s, \vec{u})$, we have $\mathbf{f}(\mathcal{P}, \vec{u}) = \sum_{s\in\mathcal{P}} \Pr^R(s, \vec{u})\langle s, \vec{u}\rangle$.

Based on the above discussion, we can compute $\mathbf{f}(\mathcal{P}, \vec{u})$ in the following way. First, we project all points of $\mathcal{P}$ on $\rho(\vec{u})$ and obtain the coordinate $\langle \vec{u}, s\rangle$ for each $s \in \mathcal{P}$. Second, we sort all points of $\mathcal{P}$ by the coordinates of their projections on $\rho(\vec{u})$. Then, the values $\Pr^R(s, \vec{u})$ for all points $s \in \mathcal{P}$ can be obtained in $O(n)$ time by considering the projection points on $\rho(\vec{u})$ from right to left. Finally, $\mathbf{f}(\mathcal{P}, \vec{u})$ can be computed in additional $O(n)$ time. Therefore, the total time for computing $\mathbf{f}(\mathcal{P}, \vec{u})$ is $O(n \log n)$, which is dominated by the sorting. If the points of $\mathcal{P}$ are given sorted along the direction $\vec{u}$, then we can avoid the sorting step and compute $\mathbf{f}(\mathcal{P}, \vec{u})$ in overall $O(n)$ time.

It remains to compute $\nabla\mathbf{f}(\mathcal{P}, \vec{u})$. Recall that $\nabla\mathbf{f}(\mathcal{P}, \vec{u}) = \sum_{s\in\mathcal{P}} \Pr^R(s, \vec{u})s$ by the proof of Lemma 18. Note that the above has already computed $\Pr^R(s, \vec{u})$ for all points $s \in \mathcal{P}$. Therefore, $\nabla\mathbf{f}(\mathcal{P}, \vec{u})$ can be computed in additional $O(n)$ time. The lemma thus follows. □

Lemma 50. *We can build a data structure of $O(n^2)$ size in $O(n^2 \log n)$ time that can compute $\omega(\mathcal{P}, \vec{u})$, $\mathbf{f}(\mathcal{P}, \vec{u})$, and $\nabla\mathbf{f}(\mathcal{P}, \vec{u})$ in $O(\log n)$ time for any query direction $\vec{u}$. Further, we can construct M explicitly in $O(n^2 \log n)$ time.*

Proof. Consider any direction $\vec{u}$ with $\|\vec{u}\| = 1$. We follow the definitions and notations in the proof of Lemma 49. We first show how to build a data structure to compute $\mathbf{f}(\mathcal{P}, \vec{u})$. Computing $\mathbf{f}(\mathcal{P}, -\vec{u})$ can be done similarly. Again, after having $\mathbf{f}(\mathcal{P}, \vec{u})$ and $\mathbf{f}(\mathcal{P}, -\vec{u})$, $\omega(\mathcal{P}, \vec{u})$ can be computed immediately by $\omega(\mathcal{P}, \vec{u}) = \mathbf{f}(\mathcal{P}, \vec{u}) - \mathbf{f}(\mathcal{P}, -\vec{u})$.

Denote by o the origin. For any *ray* ρ through o in the plane, we refer to the *angle* of ρ as the angle α in $[0, 2\pi)$ such that after we rotate the x-axis around o

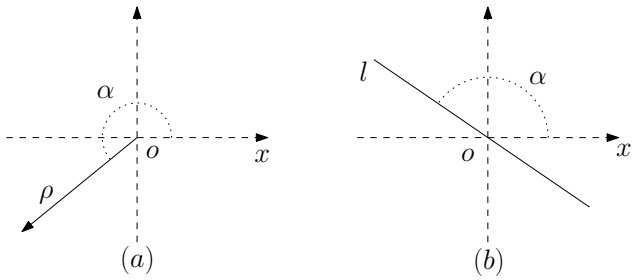

Figure 4-4: Illustrating the definition of the angle α of: (a) a ray ρ and (b) a line l.

counterclockwise by α the x-axis has the same direction as ρ (see Fig. 4-4(a)). For any (undirected) line l through o, we refer to the *angle* of l as the angle α in $[0, \pi)$ such that after we rotate the x-axis around o counterclockwise by α the x-axis is collinear with l.

Recall $\rho(\vec{u})$ is the ray through o with direction $\vec{u}$. We define the *angle* of $\vec{u}$ as the angle of the ray $\rho(\vec{u})$, denoted by $\theta_{\vec{u}}$. For ease of discussion, we assume $\theta_{\vec{u}}$ is in $[0, \pi)$ since the the case $\theta_{\vec{u}} \in [\pi, 2\pi)$ can be handled similarly.

We call the order of the points of $\mathcal{P}$ sorted by the coordinates of their projections on the ray $\rho(\vec{u})$ the *canonical order* of $\mathcal{P}$ with respect to $\vec{u}$. An easy observation is that when we increase the angle $\vec{u}$, the canonical order of $\mathcal{P}$ does not change until $\vec{u}$ is perpendicular to a line containing two points of $\mathcal{P}$. There are $O(n^2)$ lines in the plane each of which contains two points of $\mathcal{P}$ and the directions of these lines partition $[0, \pi)$ into $O(n^2)$ intervals such that if $\theta_{\vec{u}}$ changes in each interval the canonical order of $\mathcal{P}$ does not change. In the following, we show that for each of the above intervals, the value of $\mathbf{f}(\mathcal{P}, \vec{u})$ is a function of the angle $\theta_{\vec{u}}$, and more specifically $\mathbf{f}(\mathcal{P}, \vec{u}) = a \cdot \cos(\theta_{\vec{u}}) + b \cdot \sin(\theta_{\vec{u}})$ where a and b are constants when $\theta_{\vec{u}}$ changes in the interval. As preprocessing for the lemma, we will compute the function $\mathbf{f}(\mathcal{P}, \vec{u})$ for each interval; for each query direction $\vec{u}$, we first find the interval that contains $\theta_{\vec{u}}$ by binary search in $O(\log n)$ time and then obtain the value $\mathbf{f}(\mathcal{P}, \vec{u})$ in constant time using the function for the interval. The details are given below.

For simplicity of discussion, we make a general position assumption that no three points of $\mathcal{P}$ are collinear. For any two points s and s' in $\mathcal{P}$, let $\beta(s, s')$ denote the angle of the line perpendicular to the line containing s and s', and we also say $\beta(s, s')$

is *defined* by s and s'. We sort all $O(n^2)$ angles $\beta(s, s')$ for $s, s' \in \mathcal{P}$ in increasing order, and let $\beta_1, \beta_2, \ldots, \beta_h$ be the sorted list with $h = O(n^2)$. For simplicity, let $\beta_0 = 0$ and $\beta_{h+1} = \pi$. These angles partition $[0, \pi)$ into $h + 1$ intervals. Consider an interval $I_i = (\beta_i, \beta_{i+1})$ for any $0 \leq i \leq h$. Below we compute the function $\mathbf{f}(\mathcal{P}, \vec{u}) = a \cdot \cos(\theta_{\vec{u}}) + b \cdot \sin(\theta_{\vec{u}})$ for $\theta_{\vec{u}} \in (\beta_i, \beta_{i+1})$. Again, note that when $\theta_{\vec{u}}$ changes in I_i, the canonical order of $\mathcal{P}$ does not change.

According to the proof of Lemma 49, $\mathbf{f}(\mathcal{P}, \vec{u}) = \sum_{s \in \mathcal{P}} \Pr^R(s, \vec{u}) \langle s, \vec{u} \rangle$. Since the canonical order of $\mathcal{P}$ does not change for any $\theta_{\vec{u}} \in I_i$, for any $s \in \mathcal{P}$, $\Pr^R(s, \vec{u})$ is a constant when $\theta_{\vec{u}}$ changes in I_i. Next, we consider the coordinate $\langle s, \vec{u} \rangle$ on $\rho(\vec{u})$.

For each point $s \in \mathcal{P}$, let α_s be the angle of the ray originating from o and containing s (i.e., directed from o to s), and let d_s be the length of the line segment $\overline{vo}$. Note that α_s and d_s are fixed for the input. Then, we have (see Fig. 4-5)

$$\langle s, \vec{u} \rangle = d_s \cdot \cos(\alpha_s - \theta_{\vec{u}}) = d_s \cdot \cos(\alpha_v) \cdot \cos(\theta_{\vec{u}}) + d_s \cdot \sin(\alpha_s) \cdot \sin(\theta_{\vec{u}}).$$

Hence, we have the following

$$\begin{aligned}
\mathbf{f}(\mathcal{P}, \vec{u}) &= \sum_{s \in \mathcal{P}} \Pr^R(s, \vec{u}) \langle s, \vec{u} \rangle \\
&= \sum_{s \in \mathcal{P}} \Pr^R(s, \vec{u}) \cdot d_v \cdot [\cos(\alpha_v) \cdot \cos(\theta_{\vec{u}}) + \sin(\alpha_v) \cdot \sin(\theta_{\vec{u}})] \\
&= \cos(\theta_{\vec{u}}) \cdot \Big[\sum_{s \in \mathcal{P}} \Pr^R(s, \vec{u}) \cdot d_v \cdot \cos(\alpha_v) \Big] + \sin(\theta_{\vec{u}}) \cdot \Big[\sum_{s \in \mathcal{P}} \Pr^R(s, \vec{u}) \cdot d_v \cdot \sin(\alpha_v) \Big].
\end{aligned}$$

Let $a = \sum_{s \in \mathcal{P}} \Pr^R(s, \vec{u}) \cdot d_s \cdot \cos(\alpha_s)$ and $b = \sum_{s \in \mathcal{P}} \Pr^R(s, \vec{u}) \cdot d_s \cdot \sin(\alpha_s)$. Hence, a and b are constants when $\theta_{\vec{u}}$ changes in I_i. Then, we have $\mathbf{f}(\mathcal{P}, \vec{u}) = a \cdot \cos(\theta_{\vec{u}}) + b \cdot \sin(\theta_{\vec{u}})$, for any $\theta_{\vec{u}} \in I_i$. Therefore, if we know the two values a and b, we can compute $\mathbf{f}(\mathcal{P}, \vec{u})$ in constant time for any direction $\theta_{\vec{u}} \in I_i$.

In the sequel, we show that we can compute a and b for all intervals $I_i = (\beta_i, \beta_{i+1})$ with $i = 0, 1, \ldots, h$ in $O(n^2)$ time. For each interval I_i, we use $a(I_i)$ and $b(I_i)$ to denote the corresponding a and b respectively for the interval I_i.

Suppose we have computed $a(I_i)$ and $b(I_i)$ for the interval I_i, and also suppose

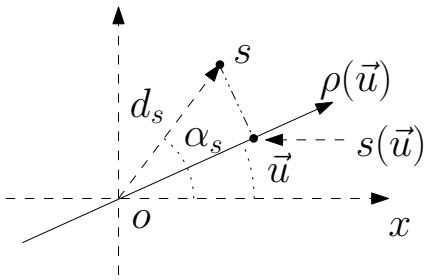

Figure 4-5: Illustrating the computation of the coordinate $x(s, \vec{u})$ on $l(\vec{u})$: $v(\vec{u})$ is the perpendicular projection of s on $l(\vec{u})$. The length of $\overline{ov}$ is d_s.

have computed the value $\Pr^R(s, \vec{u})$ for each point $s \in \mathcal{P}$ when $\theta_{\vec{u}} \in I_i$ (note that $\Pr^R(s, \vec{u})$ is a constant for any $\theta_{\vec{u}} \in I_i$). Initially, we can compute these values for the interval I_0 in $O(n \log n)$ time by Lemma 49. Below, we show that we can obtain $a(I_{i+1})$ and $b(I_{i+1})$ in constant time, based on the above values maintained for I_i.

Recall that $I_i = (\beta_i, \beta_{i+1})$ and $I_{i+1} = (\beta_{i+1}, \beta_{i+2})$. Suppose the angle β_{i+1} is defined by the two points s_1 and s_2 of $\mathcal{P}$. In other words, β_{i+1} is the angle of the line perpendicular to the line through s_1 and s_2. If we increase the angle $\theta_{\vec{u}}$ in (β_i, β_{i+2}), the canonical order of $\mathcal{P}$ does not change except that s_1 and s_2 exchange their order when $\theta_{\vec{u}}$ passes the value β_{i+1}. Therefore, for each point $s \in \mathcal{P} \setminus \{s_1, s_2\}$, the value $\Pr^R(s, \vec{u})$ is a constant for any $\theta_{\vec{u}} \in (\beta_i, \beta_{i+2})$. Based on this observation, we can compute $a(I_{i+1})$ in the following way.

We first analyze the change of the values $\Pr^R(s_1, \vec{u})$ and $\Pr^R(s_2, \vec{u})$ when $\theta_{\vec{u}}$ changes from I_i to I_{i+1}. Let $\vec{u}$ and $\vec{u}'$ be any two directions such that $\theta_{\vec{u}}$ is in I_i and $\theta_{\vec{u}'}$ is in I_{i+1}. Without loss of generality, we assume $\langle s_1, \vec{u} \rangle < \langle s_2, \vec{u} \rangle$, and thus, $\langle s_1, \vec{u}' \rangle > \langle s_2, \vec{u}' \rangle$ since s_1 and s_2 exchange their order. Observe that $\Pr^R(s_1, \vec{u}') = \Pr^R(s_1, \vec{u})/(1 - p_{s_2})$ and $\Pr^R(s_2, \vec{u}') = \Pr^R(s_2, \vec{u}) \cdot (1 - p_1)$. Thus, we can obtain $\Pr^R(s_1, \vec{u}')$ and $\Pr^R(s_2, \vec{u}')$ in constant time since we already maintain $\Pr^R(s_1, \vec{u})$ and $\Pr^R(s_2, \vec{u})$. Consequently, we have

$$\begin{aligned} a(I_{i+1}) &= a(I_i) - [\Pr^R(s_1, \vec{u}) d_{s_1} \cos(\alpha_{s_1}) + \Pr^R(s_1, \vec{u}) d_{s_2} \cos(\alpha_{s_2})] \\ &\quad + [\Pr^R(s_1, \vec{u}') d_{s_1} \cos(\alpha_{s_1}) + \Pr^R(s_1, \vec{u}') d_{s_2} \cos(\alpha_{s_2})] \\ &= a(I_i) + d_{s_1} \cos(\alpha_{s_1}) \cdot [\Pr^R(s_1, \vec{u}') - \Pr^R(s_1, \vec{u})] + d_{s_2} \cos(\alpha_{s_2}) \cdot [\Pr^R(s_2, \vec{u}') - \Pr^R(s_2, \vec{u})]. \end{aligned} \tag{4.14}$$

Hence, after we compute $\Pr^R(s_1, \vec{u}')$ and $\Pr^R(s_2, \vec{u}')$, we can obtain $a(I_{i+1})$ in constant time.

Similarly, we can obtain $b(I_{i+1})$ in constant time. Also, the values $\Pr^R(s_1, \vec{u})$ and $\Pr^R(s_2, \vec{u})$ are updated for $\theta_{\vec{u}} \in I_{i+1}$.

In summary, after the $O(n^2)$ angles $\beta(s, s')$ are sorted in $O(n^2 \log n)$ time, the above computes the functions $\mathbf{f}(\mathcal{P}, \vec{u}) = a(I_i) \cdot \cos(\theta_{\vec{u}}) + b(I_i) \cdot \sin(\theta_{\vec{u}})$ for all intervals I_i with $i = 0, 1, \ldots, h$, in additional $O(n^2)$ time. This finishes our preprocessing.

Consider any query direction $\vec{u}$. By binary search, we first find the two angles β_i and β_{i+1} such that $\beta_i \leq \theta_{\vec{u}} < \beta_{i+1}$. If $\beta_i \neq \theta_{\vec{u}}$, then $\theta_{\vec{u}}$ is in I_i and we can use the function $\mathbf{f}(\mathcal{P}, \vec{u}) = a(I_i)\cos(\theta_{\vec{u}}) + b(I_i)\sin(\theta_{\vec{u}})$ to compute $\mathbf{f}(\mathcal{P}, \vec{u})$ in constant time. If $\beta_i = \theta_{\vec{u}}$, then the function $\mathbf{f}(\mathcal{P}, \vec{u}) = a(I_i)\cos(\theta_{\vec{u}}) + b(I_i)\sin(\theta_{\vec{u}})$ still gives the correct value of $\mathbf{f}(\mathcal{P}, \vec{u})$ since when $\theta_{\vec{u}} = \beta_i$ the projections of the two points of $\mathcal{P}$ defining β_i on $\rho(\vec{u})$ overlap and we can still consider the canonical order of $\mathcal{P}$ for $\theta_{\vec{u}} = \beta_i$ the same as that for $\theta_{\vec{u}} \in I_i$. Hence, the query time is $O(\log n)$.

Next, we show how to compute $\nabla\mathbf{f}(\mathcal{P}, \vec{u})$. Recall that $\nabla\mathbf{f}(\mathcal{P}, \vec{u}) = \sum_{s\in\mathcal{P}} \Pr^R(s, \vec{u})s$ by the proof of Lemma 18. As preprocessing, we compute the value $\sum_{s\in\mathcal{P}} \Pr^R(s, \vec{u})s$ for each interval $\theta_{\vec{u}} \in (\beta_i, \beta_{i+1})$ for $i = 0, 1, \ldots, h$. This can be done in $O(n^2)$ time (after we sort all angles), by using the similar idea as above. Specifically, suppose we already have $\nabla\mathbf{f}(\mathcal{P}, \vec{u}) = \sum_{s\in\mathcal{P}} \Pr^R(s, \vec{u})s$ for $\theta_{\vec{u}} \in (\beta_i, \beta_{i+1})$; then we can compute $\nabla\mathbf{f}(\mathcal{P}, \vec{u}) = \sum_{s\in\mathcal{P}} \Pr^R(s, \vec{u})s$ for $\theta_{\vec{u}} \in (\beta_{i+1}, \beta_{i+2})$ in constant time. This is because when $\theta_{\vec{u}}$ changes from (β_i, β_{i+1}) to $(\beta_{i+1}, \beta_{i+2})$, $\Pr^R(s, \vec{u})$ does not change for any $s \in \mathcal{P} \setminus \{s_1, s_2\}$ and $\Pr^R(s, \vec{u})$ for $s \in \{s_1, s_2\}$ can be updated in constant time, as shown above. Due to the above preprocessing, given any direction $\vec{u}$, we can compute $\nabla\mathbf{f}(\mathcal{P}, \vec{u})$ in $O(\log n)$ time by binary search, similar to that of computing $\mathbf{f}(\mathcal{P}, \vec{u})$. Further, according to Lemma 18, the above preprocessing essentially computes M, in totally $O(n^2 \log n)$ time. The lemma thus follows. □

4.9.2 Computing Expected Width for Locational Uncertainty

In this setting let $\mathcal{V}$ be a set of m uncertain points each taking one of several locations from a set of n locations in $\mathcal{P}$. The probability that a node $v \in \mathcal{P}$ is in location $s \in \mathcal{P}$

is denoted p_{vs}. To simplify analysis and discussion, we assume each location $s \in \mathcal{P}$ only has the potential to be realized by any one uncertain point $v \in \mathcal{V}$.

We now replicate the lemmas in the previous section for this setting. We use the same notation and structure when possible.

Lemma 51. *For any direction $\vec{u}$, we can compute $\omega(\mathcal{V}, \vec{u})$, $\mathbf{f}(\mathcal{V}, \vec{u})$, and $\nabla\mathbf{f}(\mathcal{V}, \vec{u})$ in $O(n \log n)$ time; if the locations of $\mathcal{P}$ are already sorted along the direction $\vec{u}$, then we can compute them in $O(n)$ time.*

Proof. Again, we first compute $\mathbf{f}(\mathcal{V}, \vec{u})$ since $\omega(\mathcal{V}, \vec{u}) = \mathbf{f}(\mathcal{V}, \vec{u}) - \mathbf{f}(\mathcal{V}, -\vec{u})$ and $\mathbf{f}(\mathcal{V}, -\vec{u})$ can be computed similarly.

We follow the structure and proof of Lemma 49 and just note the changes. The first change is that we need to keep a bit more structure since there is now dependence between the different locations of each uncertain node v. Recall that $\mathcal{P}^R(s, \vec{u})$ is the subset of $s' \in \mathcal{P}$ such that $\langle s', \vec{u}\rangle > \langle s, \vec{u}\rangle$ and $\Pr^R_\emptyset(s, \vec{u})$ is the probability that no node $v \in \mathcal{V}$ appears at a larger location than $s \in \mathcal{P}$ along direction $\vec{u}$. To describe this probability we first define a vector A_v indexed by s as $A_v[s] = 1 - \sum_{s' \in \mathcal{P}^R(s,\vec{u})} p_{vs'}$ as the probability that uncertain node v does not appear in any of its possible locations which are after s along direction $\vec{u}$. Now we can define

$$\Pr^R_\emptyset(s, u) = \prod_{v \in \mathcal{P}} A_v[s].$$

Also recall that $\Pr^R(v, s, \vec{u})$ is the probability that the largest point along $\vec{u}$ is uncertain point $v \in \mathcal{V}$ at location $s \in \mathcal{P}$. This updates equation (4.13) to be

$$\Pr^R(v, s, \vec{u}) = p_{vs} \cdot \Pr^R_\emptyset(s, \vec{u}) / A_v[s].$$

Note the two key differences. First we need to sum the probabilities for each location of v since they are mutually exclusive. Second, value $A_v[s]$ needs to be factored out of $\Pr^R_\emptyset(s, u)$ because it is already accounted for in p_{vs} locating s at s, again since they are mutually exclusive.

It follows that $\mathbf{f}(\mathcal{V}, \vec{u}) = \sum_{v \in \mathcal{V}, s \in \mathcal{P}} \Pr^R(v, s, \vec{u}) \langle s, \vec{u}\rangle$.

To compute $\mathbf{f}(\mathcal{V}, \vec{u})$ we again start by projecting all points P onto $\rho(\vec{u})$ obtaining coordinates $\langle \vec{u}, s \rangle$ and sorting, if needed. This takes $O(n \log n)$ time. Given these coordinates, sorted, it now takes a bit more work in the locational setting to show that $\mathbf{f}(\mathcal{V}, \vec{u})$ can be computed in $O(n)$ additional time. We focus on computing all n values $\Pr^R(v, s, \vec{u})$; from there is straightforward to compute $\mathbf{f}(\mathcal{V}, \vec{u})$ in $O(n)$ time.

We sweep over the locations $s \in \mathcal{P}$ from largest $\langle s, \vec{u} \rangle$ value to smallest, and we maintain each $A_v[s]$ and $\Pr^R_\emptyset(s, \vec{u})$ along the way. Given these, it is not hard to calculate $\Pr^R(v, s', \vec{u})$ in constant time with $p_{vs'}$ for $s' \in \mathcal{P}$ as the next smallest value $\langle s', \vec{u} \rangle$. The important observation is that we only need to update $A_v[s]^{\text{new}} = A_v[s]^{\text{old}} - p_{vs}$ if $p_{vs} > 0$ (which by assumption holds for only one $v \in \mathcal{P}$). Then $\Pr^R_\emptyset(s, u)$ is updated by multiplying by $A_v[s]^{\text{new}}/A_v[s]^{\text{old}}$. Both operations can be done in constant time as needed to complete the proof.

It remains to compute $\nabla\mathbf{f}(\mathcal{V}, \vec{u})$. It is not hard to see that in the locational model

$$\nabla\mathbf{f}(\mathcal{V}, \vec{u}) = \sum_{v \in \mathcal{V}, s \in \mathcal{P}} \Pr^R(v, s, \vec{u}) s.$$

Note that the above has already computed the n values $\Pr^R(v, s, \vec{u})$ for all $v \in \mathcal{V}$ and $s \in \mathcal{P}$. Therefore, $\nabla\mathbf{f}(\mathcal{V}, \vec{u})$ can be computed in additional $O(n)$ time. □

Lemma 52. *We can build a data structure of $O(n^2)$ size in $O(n^2 \log n)$ time that can compute $\omega(\mathcal{V}, \vec{u})$, $\mathbf{f}(\mathcal{V}, \vec{u})$, and $\nabla\mathbf{f}(\mathcal{V}, \vec{u})$ in $O(\log n)$ time for any query direction $\vec{u}$. Further, we can construct M explicitly in $O(n^2 \log n)$ time.*

Proof. Again we first discuss the case for computing $\mathbf{f}(\mathcal{V}, \vec{u})$. For ease of discussion, we assume the angle $\theta_{\vec{u}}$ is in $[0, \pi)$. We again follow the structure of Lemma 50. The geometry is largely the same, except that there are $h = O(n^2)$ angles $\beta_1, \beta_2, \ldots, \beta_h$ since each pair $s, s' \in \mathcal{P}$ now defines an angle $\beta(s, s')$. But it remains to compute $\mathbf{f}(\mathcal{V}, \vec{u}) = a \cdot \cos(\theta_{\vec{u}}) + b \cdot \sin(\theta_{\vec{u}})$ for some constants a and b for any $\theta_{\vec{u}}$ in each (β_i, β_{i+1}). The argument is virtually the same, replacing $\Pr^R(s, \vec{u})$ with $\Pr^R(v, s, u)$.

It remains to show that we can calculate the constants $a(I_{i+1})$ and $b(I_{i+1})$ for an interval $I_{i+1} = (\beta_{i+1}, \beta_{i+2})$ efficiently, given the values for interval $I_i = (\beta_i, \beta_{i+1})$.

Assume β_{i+1} is defined for two points $s_1, s_2 \in \mathcal{P}$, where $\langle s_1, \vec{u}\rangle < \langle s_2, \vec{u}\rangle$ for $\theta_{\vec{u}} \in I_i$ and $\langle s_1, \vec{u}'\rangle > \langle s_2, \vec{u}'\rangle$ for $\theta_{\vec{u}'} \in I_{i+1}$. By definition, the ordering among all other pairs of points is unchanged within (β_i, β_{i+2}). Let only v_1 take location s_1 with positive probability and only v_2 take location s_2 with positive probability. We focus on the more general case where $v_1 \neq v_2$; when $v_1 = v_2$, it is easier to update.

We again focus on updating a and the algorithm for b is symmetric. By the $v_1 \neq v_2$ assumption $A_{v_1}[s_1]$ and $A_{v_2}[s_2]$ are unchanged in interval (β_i, β_{i+2}). However from directions $\vec{u}$ to $\vec{u}'$ $A_{v_2}[s_1]$ increases by $p_{v_2 s_2}$ and $A_{v_1}[s_2]$ decreases by $p_{v_1 s_1}$; all other such values are unchanged. Let $A_v[s]^{\vec{u}}$ denote the value in direction $\vec{u}$. Hence

$$\Pr^R_\emptyset(s_1, \vec{u}') = \Pr^R_\emptyset(s_1, \vec{u}) \frac{A^{\vec{u}'}_{v_2}[s_1]}{A^{\vec{u}}_{v_2}[s_1]},$$

and so if we can update $A_{v_2}[s_1]$, we can update $\Pr^R_\emptyset(s_1, \vec{u}')$ in constant time. Only $A_{v_1}[s_2]$ and $\Pr^R_\emptyset(s_2, \vec{u}')$ also need to be updated. And then

$$\Pr^R(v_1, s_1, \vec{u}') = p_{v_1 s_1} \cdot \Pr^R_\emptyset(s_1, \vec{u}')/A_{v_1}[s_1] = \Pr^R(v_1, s_1, \vec{u}) \frac{\Pr^R_\emptyset(s_1, \vec{u}')}{\Pr^R_\emptyset(s_1, \vec{u})}$$

can also be updated in constant time, and similar for $\Pr^R(v_2, s_2, \vec{u}')$. Thus the only remaining difficulty is accessing and updating $A_{v_1}[s_2]$ and $A_{v_2}[s_1]$. We can easily do this for if we store the full size n array $A_v[\cdot]$ for each $v \in \mathcal{V}$. Note that this takes $O(n \cdot m)$ space, but since the output is a structure of size $O(n^2)$ and $m \leq n$, this is not prohibitive. (We note that these full arrays are not explicitly required in Lemma 51, which only requires $O(n)$ space.)

Finally, we can update $a(I_{i+1})$ from $a(I_i)$ similarly to equation (4.14) using $\Pr^R(v, s, \vec{u})$ in place of $\Pr^R(s, \vec{u})$. Thus in $O(m^2)$ time, after sorting all interval breakpoints in $O(n^2 \log n)$ time, we can build a data structure that allows calculation of $\mathbf{f}(\mathcal{V}, \vec{u})$ for any $\vec{u}$ in $O(\log n)$ time.

Next, to compute $\nabla\mathbf{f}(\mathcal{V}, \vec{u})$, recall that in the locational model $\nabla\mathbf{f}(\mathcal{V}, \vec{u}) = \sum_{v\in\mathcal{V}, s\in\mathcal{P}} \Pr^R(v, s, \vec{u})s$ by the proof of Lemma 18. As preprocessing, we compute the value $\sum_{v\in\mathcal{V}, s\in\mathcal{P}} \Pr^R(s, \vec{u})s$ for each interval $\theta_{\vec{u}} \in (\beta_i, \beta_{i+1})$ for $i = 0, 1, \ldots, h$. This can be done in $O(n^2)$ time

(after we sort all angles), by using the similar idea as above. The argument is similar to that in Lemma 50 and we ignore the details. Due to the above preprocessing, given any direction $\vec{u}$, we can compute $\nabla\mathbf{f}(\mathcal{V}, \vec{u})$ in $O(\log n)$ time by binary search. Further, the above preprocessing essentially computes M, in totally $O(n^2 \log n)$ time. The lemma thus follows. □

Chapter 5 Coreset Construction for Stochastic Shape Fitting Problems

Solving geometric optimization problems over uncertain data has become increasingly important in many applications and has attracted a lot of attentions in recent years. In this chapter, we study two important geometric optimization problems, the k-center problem and the j-flat-center problem, over stochastic/uncertain data points in Euclidean spaces. We consider both problems under two popular stochastic geometric models, the existential uncertainty model and the locational uncertainty model. We provide the first PTAS (Polynomial Time Approximation Scheme) for both problems under the two models. Our results generalize the previous results for stochastic minimum enclosing ball and stochastic enclosing cylinder.

5.1 Coreset Construction for Deterministic Shape Fitting Problems

In this section, we introduce an important class of problems in computational geometry, called the shape fitting problems (see e.g., [106]). We also review two useful concepts to bound the size of coresets, called total sensitivity and VC-dimension. Finally, we briefly introduce a framework of constructing coresets for deterministic shape fitting problems.

Definition 53. *(Shape fitting problems) A shape fitting problem is specified by a triple* $(\mathbb{R}^d, \mathcal{F}, \mathrm{d})$. *Here the set* $\mathcal{F}$ *of shapes is a family of subsets of* $\mathbb{R}^d$ *(e.g., all* k*-point sets, or all* j*-flats), and* $\mathrm{d} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{\geq 0}$ *is a symmetric distance function. Define the distance of a point* $s \in \mathbb{R}^d$ *to a shape* $F \in \mathcal{F}$ *to be* $\mathrm{d}(s, F) = \min_{s' \in F} \mathrm{d}(s, s')$. *An instance* P *of the shape fitting problem is a (weighted) point set* $\{s_1, \ldots, s_n\}$ *(*$s_i \in \mathbb{R}^d$*),*

and each s_i has a positive weight $w_i \in \mathbb{R}^+$. The goal is find a shape which best fits P, that is, a shape minimizing $\sum_{s_i \in P} w_i \cdot \mathrm{d}(s_i, F)$ over all shapes $F \in \mathcal{F}$.

In this dissertation, if we consider the Euclidean space $\mathbb{R}^d$, we let the function $\mathrm{d}(\cdot, \cdot)$ be the Euclidean distance function. Next, we introduce a powerful technique for deterministic shape fitting problems.

Definition 54. *(Coreset for shape fitting problems) Given a (weighted) instance P of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathrm{d})$ with a weight function $w : P \to \mathbb{R}^+$, an ε-coreset of S is a (weighted) point set together with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$, such that for any shape $F \in \mathcal{F}$, we have that*

$$\sum_{s_i \in S} w'_i \cdot \mathrm{d}(s_i, F) \in (1 \pm \varepsilon) \sum_{s_i \in P} w_i \cdot \mathrm{d}(s_i, F).^1$$

Total sensitivity and dimension. Feldman and Langberg [45] proposed a framework to construct coresets for a variety of deterministic shape fitting problems. We briefly introduce their framework in the following. To bound the size of coresets for deterministic shape fitting problems, a useful notion is called *total sensitivity*, originally introduced in [77].

Definition 55. *(Total sensitivity of a shape fitting instance). Given an instance $P = \{s_1, \ldots, s_n\}$ of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathrm{d})$, with a weight function $w : P \to \mathbb{R}^+$, the sensitivity of $s_i \in P$ is $\sigma_P(s_i) := \inf\{\beta \geq 0 \mid w_i \cdot \mathrm{d}(s_i, F) \leq \beta \sum_{j \in [n]} w_j \cdot \mathrm{d}(s_j, F), \forall F \in \mathcal{F}\}$. The total sensitivity of P is defined by $\mathfrak{G}_P = \sum_{s_i \in P} \sigma_P(s_i)$.*

We also recall the definition of *VC-dimension* and *shattering dimension* and show their relations.

Definition 56. *(VC-dimension and shattering dimension) Let $P = \{s_1, \ldots, s_n\}$ be an instance of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathrm{d})$. Suppose w_i is the weight of s_i. We consider the range space $(P, \mathcal{R})$, where $\mathcal{R}$ is a family of subsets $R_{F,r}$ of P defined as follows: given an $F \in \mathcal{F}$ and $r \geq 0$, let $R_{F,r} = \{s_i \in P \mid w_i \cdot \mathrm{d}(s_i, F) \geq r\} \in \mathcal{R}$ consist*

[1] The notation $(1 \pm \varepsilon)B$ means the interval $[(1 - \varepsilon)B, (1 + \varepsilon)B]$.

of those points s_i whose weighted distance to the shape F is at least r. We denote the VC-dimension of the instance of the instance P by $\dim_{VC}(P)$, to be the largest integer m, such that there exists a weight function w and a subset $A \subseteq P$ satisfying the property $|\{A \cap R_{F,r} \mid F \in \mathcal{F}, r \geq 0\}| = 2^{|A|}$. We also denote the shattering dimension of the instance P by $\dim(P)$, to be the smallest integer m, such that for any weight function w and $A \subseteq P$ of size $|A| = a \geq 2$, we have $|\{A \cap R_{F,r} \mid F \in \mathcal{F}, r \geq 0\}| \leq a^m$.

The next lemma offers a natural way to bound the VC-dimension of a range space by bounding the shattering dimension.

Lemma 57. *([58, Lemma 5.14]) If $(X, \mathcal{R})$ is a range space with shattering dimension d, then its VC-dimension is bounded by $O(d \log d)$.*

A framework for constructing coresets of shape fitting instances. The main idea of the framework in [45] is based on the following lemma.

Lemma 58. *Given any instance $P = \{s_1, \ldots, s_n\}$ of a shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathrm{d})$, any weight function $w : P \to \mathbb{R}^+$, and any $\varepsilon \in (0, 1]$, there exists an ε-coreset for P of size $O((\frac{\mathfrak{G}_P}{\varepsilon})^2 \dim_{VC}(P))$.*

We take 1-median problem as an example. In this example, $\mathcal{F} = \mathbb{R}^d$ and d is the Euclidean distance function. The framework mainly consists of the following steps.

1. Compute $F \subseteq \mathcal{F}$ which is a constant-factor approximate shape in $\mathcal{F}$. For 1-median, we only need to choose $F = s^* = \arg\min_{s \in P} \sum_{s_i \in P} w_i \cdot \mathrm{d}(s_i, s)$. It is not hard to verify that $F = s^*$ is a 2-approximate shape.

2. For each $s_i \in P$, compute its projection $s_i' = \arg\min_{s \in F} \mathrm{d}(s_i, s)$. Let $P' = \{s_i' : s_i \in P\}$ be the collection of all projection points. For 1-median, we observe that $s_i' = s^*$ for all $i \in [n]$.

3. For each $s_i \in P$, compute an upper bound $\sigma_P(s_i)$ of the sensitivity (see [106, Theorem 7] for details). Compute $\mathfrak{G}_P = \sum_{i \in [n]} \sigma_P(s_i)$ as an upper bound of the total sensitivity. For 1-median, we let $\sigma_P(s_i) = \frac{w_i \cdot \mathrm{d}(s_i, s)}{\sum_{j \in [n]} w_j \cdot \mathrm{d}(s_j, s)} + \frac{3}{n}$ as an upper bound [106]. By these values, we have $\mathfrak{G}_P = \sum_{i \in [n]} \sigma_P(s_i) = 4$.

4. Sample points from P with probabilities proportional to $\sigma_P(s_i)$. Formally speaking, we modify the weight w_i to be $\frac{w_i \cdot \mathfrak{G}_P}{\sigma_P(s_i)}$ and sample the point s_i with probability $\frac{\sigma_P(s_i)}{\mathfrak{G}_P}$. This technique is called importance sampling, see [45, Section 4.1] for more details. By Lemma 58, we need to take $O((\frac{\mathfrak{G}_P}{\varepsilon})^2 \dim_{VC}(P))$ samples. For 1-median, the VC-dimension is $d+1$. Thus, there exists an ε-coreset for P of size $O(d\varepsilon^{-2})$.

5.2 Generalized Shape Fitting Problems and Generalized Coresets

We recall the definitions of the two stochastic shape fitting problems in this chapter.

Definition 59. *For a set of points $P \in \mathbb{R}^d$, and a k-point set $F = \{f_1, \ldots, f_k) \mid f_i \in \mathbb{R}^d, 1 \leq i \leq k\}$, we define $\mathrm{K}(P,F) = \max_{s\in P} \min_{1\leq i\leq k} \mathrm{d}(s, f_i)$ as the k-center value of F w.r.t. P. We use $\mathcal{F}$ to denote the family of all k-point sets in $\mathbb{R}^d$. Given a set $\mathcal{P}$ of n stochastic points (in either the existential or locational uncertainty model) in $\mathbb{R}^d$, and a k-point set $F \in \mathcal{F}$, we define the expected k-center value of F w.r.t $\mathcal{P}$ as*

$$\mathrm{K}(\mathcal{P}, F) = \mathbb{E}_{P\sim\mathcal{P}}[\mathrm{K}(P,F)].$$

In the stochastic minimum k-center problem, our goal is to find a k-point set $F \in \mathcal{F}$ which minimizes $\mathrm{K}(\mathcal{P}, F)$. In this dissertation, we assume that both the dimensionality d and k are fixed constants.

Definition 60. *Given a set P of n points in $\mathbb{R}^d$, and a j-flat $F \in \mathcal{F}$ ($0 \leq j \leq d-1$), where $\mathcal{F}$ is the family of all j-flats in $\mathbb{R}^d$, we define the j-flat-center value of F w.r.t. P to be $\mathrm{J}(P,F) = \max_{s\in P} \mathrm{d}(s,F)$, where $\mathrm{d}(s,F) = \min_{f\in F} \mathrm{d}(s,f)$ is the distance between point s and j-flat F. Given a set $\mathcal{P}$ of n stochastic points (in either the existential or locational model) in $\mathbb{R}^d$, and a j-flat $F \in \mathcal{F}$ ($0 \leq j \leq d-1$), we define the expected j-flat-center value of F w.r.t. $\mathcal{P}$ to be*

$$\mathrm{J}(\mathcal{P}, F) = \mathbb{E}_{P\sim\mathcal{P}}[\mathrm{J}(P,F)].$$

In the stochastic minimum j-flat-center problem, our goal is to find a j-flat F which minimizes $\mathrm{J}(\mathcal{P}, F)$.

In the following, we define generalized shape fitting problems. In fact, we can consider stochastic k-center and stochastic j-flat-center as generalized shape fitting problems. We also give the definition of generalized coresets and introduce a framework for constructing generalized coresets.

Generalized Shape Fitting Problems and Generalized Coresets. In this section, we define the generalized shape fitting problems, which are defined over a collection of (weighted) point sets, (recall the traditional shape fitting problems are defined over a set of (weighted) points). We use $\mathbb{R}^d$ to denote the d-dimensional Euclidean space. Let $\mathrm{d}(s, s')$ denote the Euclidean distance between point s and s' and $\mathrm{d}(s, F) = \min_{s' \in F} \mathrm{d}(s, s')$ for any $F \subset \mathbb{R}^d$. Let $\mathbb{U}^d = \{P \mid P \subset \mathbb{R}^d, |P| \text{ is finite}\}$ be the collection of all finite discrete point sets in $\mathbb{R}^d$.

Definition 61. *(Generalized shape fitting problems) A generalized shape fitting problem is specified by a triple* $(\mathbb{R}^d, \mathcal{F}, \mathsf{dist})$. *Here the set $\mathcal{F}$ of shapes is a family of subsets of $\mathbb{R}^d$ (e.g., all k-point sets, or all j-flats), and* $\mathsf{dist} : \mathbb{U}^d \times \mathcal{F} \to \mathbb{R}^{\geq 0}$ *is a generalized distance function, defined as* $\mathsf{dist}(P, F) = \max_{s \in P} \mathrm{d}(s, F)$ *for a point set $P \in \mathbb{U}^d$ and a shape $F \in \mathcal{F}$.* [2] *An instance $\mathbf{S}$ of the generalized shape fitting problem is a (weighted) collection $\{S_1, \ldots, S_m\}$ ($S_i \in \mathbb{U}^d$) of point sets, and each S_i has a positive weight $w_i \in \mathbb{R}^+$. For any shape $F \in \mathcal{F}$, define the total generalized distance from $\mathbf{S}$ to F to be* $\mathsf{dist}(\mathbf{S}, F) = \sum_{S_i \in \mathbf{S}} w_i \cdot \mathsf{dist}(S_i, F)$. *Given an instance $\mathbf{S}$, our goal is to find a shape $F \in \mathcal{F}$, which minimizes the total generalized distance* $\mathsf{dist}(\mathbf{S}, F)$.

If we replace $\mathbb{U}^d$ with $\mathbb{R}^d$, the above definition reduces to the traditional shape fitting problem, see Definition 53. Here, we give an example for Definition 61.

Example. Consider a generalized shape fitting problem where $\mathcal{F}$ is the collection of all 2-point sets in $\mathbb{R}^2$. In this case, for a point $s \in \mathbb{R}^2$ and a 2-point set $F \in \mathcal{F}$, the function $\mathrm{d}(s, F) = \min_{f \in F} \mathrm{d}(s, f)$ is the Euclidean distance between s and its

[2] Note that dist may not be a metric in general.

nearest point $f \in F$. For a point set $P \in \mathbb{U}^2$ and a 2-point set $F \in \mathcal{F}$, the function $\mathsf{dist}(P, F) = \max_{s \in P} \mathrm{d}(s, F)$ is the farthest distance from some point $s \in P$ to F.

Then we construct an instance $\mathbf{S} = \{S_1, S_2, S_3\}$ ($S_i \in \mathbb{U}^2$) of this generalized shape fitting problem as follows. Let $S_1 = \{s_1 = (0,0), s_2 = (0,2)\}$, $S_2 = \{s_3 = (6,0), s_4 = (6,2)\}$, and $S_3 = \{s_5 = (0,1)\} \in$. Each S_i has a positive weight, where $w_1 = w_2 = 1$ and $w_3 = 2$. Then our goal is to find a 2-point set $F \in \mathcal{F}$, which minimizes the following total generalized distance

$$\mathsf{dist}(\mathbf{S}, F) = \sum_{S_i \in \mathbf{S}} w_i \cdot \mathsf{dist}(S_i, F) = \mathsf{dist}(S_1, F) + \mathsf{dist}(S_2, F) + 2\mathsf{dist}(S_3, F).$$

Consider a 2-point set $F^* = \{f_1 = (0,1), f_2 = (6,1)\}$. We can compute that $\mathsf{dist}(S_1, F^*) = \max_{s \in S_1} \mathrm{d}(s, F^*) = \mathrm{d}(s_1, F^*) = \mathrm{d}(s_1, f_1) = 1$. By the same way, we compute that $\mathsf{dist}(S_2, F^*) = \mathrm{d}(s_3, f_2) = 1$ and $\mathsf{dist}(S_3, F^*) = \mathrm{d}(s_5, f_1) = 0$. Thus, we have that $\mathsf{dist}(\mathbf{S}, F^*) = 1 + 1 + 0 = 2$. In fact, we can prove that F^* is the optimal 2-point set which minimizes the total generalized distance $\mathsf{dist}(\mathbf{S}, F)$.

Now, we define what is a coreset for a generalized shape fitting problem.

Definition 62. *(Generalized Coreset) Given a (weighted) instance* $\mathbf{S}$ *of a generalized shape fitting problem* $(\mathbb{R}^d, \mathcal{F}, \mathsf{dist})$ *with a weight function* $w : \mathbf{S} \to \mathbb{R}^+$*, a generalized* ε*-coreset of* $\mathbf{S}$ *is a (weighted) collection* $\mathcal{S} \subseteq \mathbf{S}$ *of point sets, together with a weight function* $w' : \mathcal{S} \to \mathbb{R}^+$*, such that for any shape* $F \in \mathcal{F}$*, we have that*

$$\sum_{S_i \in \mathcal{S}} w_i' \cdot \mathsf{dist}(S_i, F) \in (1 \pm \varepsilon) \sum_{S_i \in \mathbf{S}} w_i \cdot \mathsf{dist}(S_i, F)$$

(or more compactly, $\mathsf{dist}(\mathcal{S}, F) \in (1 \pm \varepsilon)\mathsf{dist}(\mathbf{S}, F)$*. We denote the cardinality of the coreset* $\mathcal{S}$ *as* $|\mathcal{S}|$*.*

Definition 62 also generalizes the prior definition in [106], where each $S_i \in \mathbf{S}$ contains only one point.

Total sensitivity and dimension. To bound the size of the generalized coresets,

we need the notion of *total sensitivity*, originally introduced in [77].

Definition 63. *(Total sensitivity of a generalized shape fitting instance). Let $\mathbb{U}^d$ be the collection of all finite discrete point sets $P \subset \mathbb{R}^d$, and let $\mathsf{dist} : \mathbb{U}^d \times \mathcal{F} \to \mathbb{R}^{\geq 0}$ be a continuous function. Given an instance $\mathbf{S} = \{S_i \mid S_i \subset \mathbb{U}^d, 1 \leq i \leq n\}$ of a generalized shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathsf{dist})$, with a weight function $w : \mathbf{S} \to \mathbb{R}^+$, the sensitivity $S_i \in \mathbf{S}$ is $\sigma_{\mathbf{S}}(S_i) := \inf\{\beta \geq 0 \mid w_i \cdot \mathsf{dist}(S_i, F) \leq \beta \cdot \mathsf{dist}(\mathbf{S}, F), \forall F \in \mathcal{F}\}$. The total sensitivity of $\mathbf{S}$ is defined by $\mathfrak{G}_{\mathbf{S}} = \sum_{S_i \in \mathbf{S}} \sigma_{\mathbf{S}}(S_i)$.*

Note that this definition generalizes the one in [77]. In fact, if each $S_i \in \mathbf{S}$ contains only one point, this definition is equivalent to Definition 55.

We also generalize the definition of *dimension* defined in [45] (it is in fact the primal shattering dimension (See e.g., [45, 58]) of a certain range space. It plays a similar role to VC-dimension).

Definition 64. *(Generalized dimension) Let $\mathbf{S} = \{S_i \mid S_i \in \mathbb{U}^d, 1 \leq i \leq n\}$ be an instance of a generalized shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathsf{dist})$. Suppose w_i is the weight of S_i. We consider the range space $(\mathbf{S}, \mathcal{R})$, where $\mathcal{R}$ is a family of subsets $R_{F,r}$ of $\mathbf{S}$ defined as follows: given an $F \in \mathcal{F}$ and $r \geq 0$, let $R_{F,r} = \{S_i \in \mathbf{S} \mid w_i \cdot \mathsf{dist}(S_i, F) \geq r\} \in \mathcal{R}$ consist of the sets S_i whose weighted distance to the shape F is at least r. Finally, we denote the generalized dimension of the instance $\mathbf{S}$ by $\dim(\mathbf{S})$, to be the smallest integer m, such that for any weight function w and $\mathcal{A} \subseteq \mathbf{S}$ of size $|\mathcal{A}| = a \geq 2$, we have $|\{\mathcal{A} \cap R_{F,r} \mid F \in \mathcal{F}, r \geq 0\}| \leq a^m$.*

The definition [77] is a special case of the above definition when each $S_i \in \mathbf{S}$ contains only one point. On the other hand, the above definition is a special case of Definition 7.2 [45] if thinking each $w_i \cdot \mathsf{dist}(S_i, \cdot) = g_i(\cdot)$ as a function from $\mathcal{F}$ to $\mathbb{R}^{\geq 0}$.

We first have the following lemma for bounding the size of generalized coresets by the generalized total sensitivity and dimension.

Lemma 65. *Given any instance $\mathbf{S} = \{S_i \mid S_i \subset \mathbb{U}^d, 1 \leq i \leq n\}$ of a generalized shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathsf{dist})$, any weight function $w : \mathbf{S} \to \mathbb{R}^+$, and any $\varepsilon \in (0, 1]$, there exists a generalized ε-coreset for $\mathbf{S}$ of cardinality $O((\frac{\mathfrak{G}_{\mathbf{S}}}{\varepsilon})^2 \dim(\mathbf{S}) \log \dim(\mathbf{S}))$.*

The proof is a straightforward extension of the following theorem (a restatement of Theorem 4.1 and its proof in [45]). Lemma 65 is a direct corollary from the following theorem.

Theorem 66. *Let $D = \{g_i \mid 1 \leq i \leq n\}$ be a set of n functions. For each $g \in D$, $g : X \to \mathbb{R}^{\geq 0}$ is a function from a ground set X to $[0, +\infty)$. Let $0 < \varepsilon < 1/4$ be a constant. Let $q : D \to \mathbb{R}^+$ be a function on D such that*

$$q(g) \geq \max_{x \in X} \frac{g(x)}{\sum_{g \in D} g(x)}. \tag{5.1}$$

Then there exists a collection $\mathcal{S} \subseteq D$ of functions, together with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$, such that for every $x \in X$

$$|\sum_{g \in D} g(x) - \sum_{g \in \mathcal{S}} w'(g) \cdot g(x)| \leq \varepsilon \sum_{g \in Y} g(x),$$

Moreover, the size of $\mathcal{S}$ is

$$O\left(\left(\frac{\sum_{g \in D} q(g)}{\varepsilon}\right)^2 \dim(D) \cdot \log(\dim(D))\right),$$

where $\dim(D)$ is the generalized shattering dimension of D (see Definition 7.2 in [45]). [3]

Now we are ready to prove Lemma 65.

Proof. Suppose that we are given a (weighted) instance $\mathbf{S} = \{S_i \mid S_i \subset \mathbb{R}^d, 1 \leq i \leq n\}$ of a generalized shape fitting problem $(\mathbb{R}^d, \mathcal{F}, \mathsf{dist})$, with a weight function $w : \mathbf{S} \to \mathbb{R}^+$. A generalized ε-coreset is a collection $\mathcal{S} \subseteq \mathbf{S}$ of point sets, together with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$ such that, for any shape $F \in \mathcal{F}$, we have

$$\sum_{S_i \in \mathcal{S}} w'_i \cdot \mathsf{dist}(S_i, F) \in (1 \pm \varepsilon) \sum_{S_i \in \mathbf{S}} w_i \cdot \mathsf{dist}(S_i, F). \tag{5.2}$$

[3] Note that there is an additional $\log(\dim(D))$ term comparing to Lemma 58. This is because we consider shattering dimension instead of VC-dimension. By Lemma 57, we should have this additional term.

For every $S_i \in \mathbf{S}$ and $F \in \mathcal{F}$, let $g_i(F) = w_i \cdot \mathsf{dist}(S_i, F)$ and $D = \{g_i \mid S_i \in \mathbf{S}\}$. Define

$$\begin{aligned} q(g_i) &= \sigma_{\mathbf{S}}(S_i) + \frac{1}{n} = \inf\{\beta \geq 0 \mid w_i \cdot \mathsf{dist}(S_i, F) \\ &\leq \beta \cdot \sum_{S_i \in \mathbf{S}} w_i \cdot \mathsf{dist}(S_i, F), \forall F \in \mathcal{F}\} + \frac{1}{n}. \end{aligned}$$

It is not hard to verify that this definition satisfies Inequality (5.1). The additional $1/n$ term will be useful in Section 5.3, where we need a lower bound of $q(g_i)$. Thus, we have $\mathfrak{G}_{\mathbf{S}} + 1 = \sum_{S_i \in \mathbf{S}}(\sigma_{\mathbf{S}}(S_i) + 1/n) = \sum_{g_i \in D} q(g_i)$. Recall that $\dim(\mathbf{S})$ is the generalized shattering dimension of $\mathbf{S}$. By Theorem 66, we conclude that there exists a collection $\mathcal{S}$ of cardinality $O\big((\frac{\mathfrak{G}_{\mathbf{S}}}{\varepsilon})^2 \dim(\mathbf{S}) \cdot \log(\dim(\mathbf{S}))\big)$ with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$ satisfying Inequality (5.2). □

5.3 Stochastic Minimum k-Center

In this section, we consider the stochastic minimum k-center problem in $\mathbb{R}^d$ in the stochastic model. Let $\mathcal{F}$ be the family of all k-point sets of $\mathbb{R}^d$, and let $\mathcal{P}$ be the set of stochastic points. Our main technique is to construct an SKC-Coreset $\mathcal{S}$ of constant size. For any k-point set $F \in \mathcal{F}$, $\mathrm{K}(\mathcal{S}, F)$ should be a $(1 \pm \varepsilon)$-estimation for $\mathrm{K}(\mathcal{P}, F) = \mathbb{E}_{P \sim \mathcal{P}}[\mathrm{K}(P, F)]$. Recall that $\mathrm{K}(P, F) = \max_{s \in P} \min_{f \in F} \mathrm{d}(s, f)$ is the k-center value between two point sets P and F. Constructing $\mathcal{S}$ includes two main steps: 1) Partition all realizations via additive ε-coresets, which reduces an exponential number of realizations to a polynomial number of point sets. 2) Show that there exists a generalized coreset of constant cardinality for the generalized k-median problem defined over the above set of polynomial point sets. Finally, we enumerate polynomially many possible collections $\mathcal{S}_i$ (together with their weights). We show that there is an SKC-Coreset $\mathcal{S}$ among those candidate. By solving a polynomial system for each $\mathcal{S}_i$, and take the minimum solution, we can obtain a PTAS.

We first need the formal definition of an additive ε-coreset [11] as follows.

Definition 67. *(additive ε-coreset) Let $B(f, r)$ denote the ball of radius r centered*

at point f. For a set of points $P \in \mathbb{U}^d$, we call $Q \subseteq P$ an additive ε-coreset *of P if for every k-point set $F = \{f_1, \ldots, f_k\}$, we have*

$$P \subseteq \cup_{i=1}^{k} B(f_i, (1+\varepsilon)\mathrm{K}(Q, F)),$$

i.e., the union of all balls $B(f_i, (1+\varepsilon)\mathrm{K}(Q,F))$ $(1 \leq i \leq k)$ covers P. [4]

5.3.1 Existential uncertainty model

We first consider the existential uncertainty model.

Step 1: Partitioning realizations

We first provide an algorithm $\mathbb{A}$, which can construct an additive ε-coreset for any deterministic point set. We can think $\mathbb{A}$ as a mapping from all realizations of $\mathcal{P}$ to all possible additive ε-coresets. The mapping naturally induces a partition of all realizations. Note that we do not run $\mathbb{A}$ on every realization.

Algorithm $\mathbb{A}$ for constructing additive ε-coresets. Given a realization $P \sim \mathcal{P}$, we first compute an approximation value r_P of the optimal k-center value $\min_{F \in \mathcal{F}} \mathrm{K}(P, F)$. Then we build a Cartesian grid $G(P)$ of side length depending on r_P. Let $\mathcal{C}(P) = \{C \mid C \in G, C \cap P \neq \emptyset\}$ be the collection of those nonempty cells (i.e., cells that contain at least one point in P). In each non-empty cell $C \in \mathcal{C}(P)$, we maintain the point $s^C \in C \cap P$ of smallest index. Let $\mathcal{E}(P) = \{s^C \mid C \in G\}$, which is an additive ε-coreset of P. Finally the output of $\mathbb{A}(P)$ is $\mathcal{E}(P), G(P)$, and $\mathcal{C}(P)$. The details can be found in Section 5.5.

Note that we do not use the construction of additive ε-coresets [11], because it is not easy to recover the set of original realizations with a certain additive ε-coreset. We need the set of additive ε-coresets to have some extra properties (in particular, Lemma 70 below), which allows us to compute certain probability values efficiently.

We first have the following lemma.

Lemma 68. *The running time of $\mathbb{A}$ on any n point set P is $O(kn^{k+1})$. Moreover,*

[4] Our definition is slight weaker than that in [11]. The weaker definition suffices for our purpose.

the output $\mathcal{E}(P)$ is an additive ε-coreset of P of size at most $O(k/\varepsilon^d)$.

Denote $\mathcal{E}(\mathcal{P}) = \{\mathcal{E}(P) \mid P \sim \mathcal{P}\}$ be the collection of all possible additive ε-coresets. By Lemma 68, we know that each $S \in \mathcal{E}(\mathcal{P})$ is of size at most $O(k/\varepsilon^d)$. Thus, the cardinality of $\mathcal{E}(\mathcal{P})$ is at most $n^{O(k/\varepsilon^d)}$. For a point set S, denote $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S] = \sum_{P:P\sim\mathcal{P},\mathcal{E}(P)=S} \Pr[\models P]$ to be the probability that the additive ε-coreset of a realization is S. The following simple lemma states that we can have a polynomial size representation for the objective function $\mathrm{K}(\mathcal{P}, F)$.

Lemma 69. *Given $\mathcal{P}$ of n points in $\mathbb{R}^d$ in the existential uncertainty model, for any k-point set $F \in \mathcal{F}$, we have that*

$$\sum_{S\in\mathcal{E}(\mathcal{P})} \Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S] \cdot \mathrm{K}(S, F) \in (1 \pm \varepsilon)\mathrm{K}(\mathcal{P}, F).$$

Proof. By the definition of $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S]$, we can see that for any k-point set $F \in \mathcal{F}$,

$$\begin{aligned}\sum_{S\in\mathcal{E}(\mathcal{P})} \Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S] \cdot \mathrm{K}(S, F) &= \sum_{S\in\mathcal{E}(\mathcal{P})} \sum_{P:P\sim\mathcal{P},\mathcal{E}(P)=S} \Pr[\models P] \cdot \mathrm{K}(S, F) \\ \in (1 \pm \varepsilon) \sum_{S\in\mathcal{E}(\mathcal{P})} \sum_{P:P\sim\mathcal{P},\mathcal{E}(P)=S} \Pr[\models P] \cdot \mathrm{K}(P, F) &= (1 \pm \varepsilon)\mathrm{K}(\mathcal{P}, F).\end{aligned}$$

The inequality above uses the definition of additive ε-coresets (Definition 67). □

We can think $\mathcal{P} \to \mathcal{E}(\mathcal{P})$ as a mapping, which maps a realization $P \sim \mathcal{P}$ to its additive ε-coreset $\mathcal{E}(P)$. The mapping partitions all realizations $P \sim \mathcal{P}$ into a polynomial number of additive ε-coresets. For each possible additive ε-coreset $S \in \mathcal{E}(\mathcal{P})$, we denote $\mathcal{E}^{-1}(S) = \{P \sim \mathcal{P} \mid \mathcal{E}(P) = S\}$ to be the collection of all realizations mapping to S. By the definition of $\mathcal{E}(\mathcal{P})$, we have that $\cup_{S\in\mathcal{E}(\mathcal{P})}\mathcal{E}^{-1}(S) = \mathcal{P}$.

Now, we need an efficient algorithm to compute $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S]$ for each additive ε-coreset $S \in \mathcal{E}(\mathcal{P})$. The following lemma states that the mapping constructed by algorithm $\mathbb{A}$ has some nice properties that allow us to compute the probabilities. This is also the reason why we cannot directly use the original additive ε-coreset

construction algorithm in [11]. The proof is somewhat subtle and can be found in Section 5.5.

Lemma 70. *Consider a subset S of at most $O(k/\varepsilon^d)$ points. Run algorithm $\mathbb{A}(S)$, which outputs an additive ε-coreset $\mathcal{E}(S)$, a Cartesian grid $G(S)$, and a collection $\mathcal{C}(S)$ of nonempty cells. If $\mathcal{E}(S) \neq S$, then $S \notin \mathcal{E}(\mathcal{P})$ (i.e., S is not the output of $\mathbb{A}$ for any realization $P \sim \mathcal{P}$).* [5] *If $|S| \leq k$, then $\mathcal{E}^{-1}(S) = \{S\}$. Otherwise if $\mathcal{E}(S) = S$ and $|S| \geq k+1$, then a point set $P \sim \mathcal{P}$ satisfies $\mathcal{E}(P) = S$ if and only if*

P1. For any cell $C \notin \mathcal{C}(S)$, $C \cap P = \emptyset$.

P2. For any cell $C \in \mathcal{C}(S)$, assume that point $s^C = C \cap S$. Then $s^C \in P$, and any point $s' \in C \cap \mathcal{P}$ with a smaller index than that of s^C does not appear in the realization P.

Thanks to Lemma 70, now we are ready to show how to compute $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S]$ efficiently for each $S \in \mathcal{E}(\mathcal{P})$. We enumerate every point set of size $O(k/\varepsilon^d)$. For a set S, we first run $\mathbb{A}(S)$ and output a Cartesian grid $G(S)$ and a point set $\mathcal{E}(S)$. We check whether $S \in \mathcal{E}(\mathcal{P})$ by checking whether $\mathcal{E}(S) = S$ or $|S| \leq k$. If $S \in \mathcal{E}(\mathcal{P})$, we can compute $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S]$ using the Cartesian grid $G(S)$. See Algorithm 1 for details. We also give an example to explain Algorithm 1, see Figure 5.3.1.

The following lemma asserting the correctness of Algorithm 1 is a simple consequence of Lemma 70.

Lemma 71. *For any point set S, Algorithm 1 computes exactly the total probability*

$$\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S] = \sum_{P:P\sim\mathcal{P},\mathcal{E}(P)=S} \Pr[\vDash P]$$

in $O(n^{O(k/\varepsilon^d)})$ time.

Proof. Run $\mathbb{A}(S)$, and we obtain a point set $\mathcal{E}(S)$. If $\mathcal{E}(S) \neq S$, we have that $S \notin \mathcal{E}(\mathcal{P})$ by Lemma 70. Thus, $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S] = 0$. If $|S| \leq k$, we have that $\mathcal{E}^{-1}(S) = \{S\}$ by Lemma 70. Thus, $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S] = \Pr[\vDash S]$.

[5] It is possible that some point set S satisfies Definition 67 for some realization P, but is not the output of $\mathbb{A}(S)$.

Algorithm 1 Computing $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]$

1 For each point set $S\sim\mathcal{P}$ of size $|S|=O(k/\varepsilon^d)$, run algorithm $\mathbb{A}(S)$. Assume that the output is a point set $\mathcal{E}(S)$, a Cartesian grid $G(S)$, and a cell collection $\mathcal{C}(S)=\{C\mid C\in G, C\cap S\neq\emptyset\}$.
2 If $\mathcal{E}(S)\neq S$, output $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]=0$. If $|S|\leq k$, output $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]=\Pr[\vDash S]$.
3 For a cell C, suppose $C\cap\mathcal{P}=\{t_i\mid t_i\in\mathcal{P}, 1\leq i\leq m\}$. W.l.o.g., assume that $t_1,\ldots,t_m$ are in increasing order of their indices. For $C\notin\mathcal{C}(S)$, let

$$Q(C)=\Pr_{P\sim\mathcal{P}}\Big[P\cap C=\emptyset\Big]=\prod_{i=1}^{m}(1-p_i)$$

be the probability that no point in C is realized. If $C\in\mathcal{C}(S)$, assume that point $t_j\in C\cap S$, and let

$$Q(C)=\Pr_{P\sim\mathcal{P}}\Big[t_j\in P \text{ and } \{t_1,\ldots,t_{j-1}\}\cap P=\emptyset\Big]=p_j\cdot\prod_{i=1}^{j-1}(1-p_i)$$

be the probability that t_j appears, but $t_1,\ldots,t_{j-1}$ do not appear.
4 Output $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]=\prod_{C\in G(S)}Q(C)$.

Otherwise if $\mathcal{E}(S)=S$ and $|S|\geq k+1$, by Lemma 70, each realization $P\in\mathcal{E}^{-1}(S)$ satisfies P1 and P2. Then combining the definition of $Q(C)$, and the independence of all cells, we can see that $\prod_{C\in\mathcal{C}}Q(C)$ is equal to $\sum_{P\in\mathcal{E}^{-1}(S)}\Pr[\vDash P]=\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]$.

For the running time, note that we only need to consider at most $n^{O(k/\varepsilon^d)}$ point sets $S\sim\mathcal{P}$. For each S, Algorithm 1 needs to run $\mathbb{A}(S)$, which costs $O(kn^{k+1})$ time by Lemma 68. Step 2 and 3 only cost linear time. Thus, we can compute all probabilities $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]$ in $O(n^{O(k/\varepsilon^d)})$ time. □

Step 2: Existence of generalized coreset via generalized total sensitivity

Recall that $\mathcal{E}(\mathcal{P})$ is a collection of polynomially many point sets of size $O(k/\varepsilon^d)$. By Lemma 69, we can focus on a generalized k-median problem: finding a k-point set $F\in\mathcal{F}$ which minimizes $\mathrm{K}(\mathcal{E}(\mathcal{P}),F)=\sum_{S\in\mathcal{E}(\mathcal{P})}\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]\cdot\mathrm{K}(S,F)$. In fact, the generalized k-median problem is a special case of the generalized shape fitting problem we defined in Definition 61. Here, we instantiate the shape family $\mathcal{F}$ to be the collection of all k-point sets. Note that the k-center objective $\mathrm{K}(\mathcal{E}(\mathcal{P}),F)$ is

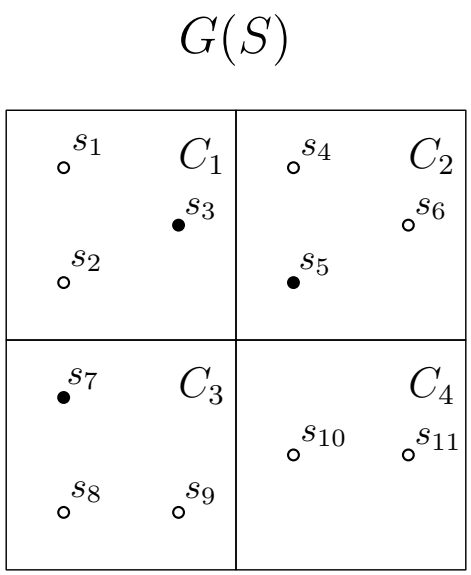

Figure 5-1: An example for Algorithm 1 when $k = 2$. In this figure, $\mathcal{P} = \{s_1, \ldots, s_{11}\}$ consists of all points, and $S = \{s_3, s_5, s_7\}$ consists of black points. Then by Lemma 70, we have that $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S] = p_3p_5p_7(1-p_1)(1-p_2)(1-p_4)(1-p_{10})(1-p_{11})$. Now we run Algorithm 1 on S. In Step 1, we first construct a Cartesian grid $G(S)$ as in the figure, and construct a cell collection $\mathcal{C}(S) = \{C_1, C_2, C_3\}$ since $C_4 \cap S = \emptyset$. Note that $\mathcal{E}(S) = S$ (by Lemma 70) and $|S| = 3 > k$. We directly go to Step 3 and want to compute the value $Q(C_i)$ for each cell C_i. For cell C_1, two rectangle points s_1 and s_2 are of smaller index than $s_3 \in S$. So we compute that $Q(C_1) = p_3(1-p_1)(1-p_2)$. Similarly, we compute $Q(C_2) = p_5(1-p_4)$, $Q(C_3) = p_7$, and $Q(C_4) = (1-p_{10})(1-p_{11})$. Finally in Step 4, we output $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S] = \prod_{C\in G(S)} Q(C) = p_3p_5p_7(1-p_1)(1-p_2)(1-p_4)(1-p_{10})(1-p_{11})$.

indeed a generalized distance function in Definition 61. To make things concrete, we formalize it below. Recall that $\mathbb{U}^d$ is the collection of all finite discrete point sets in $\mathbb{R}^d$.

Definition 72. *A generalized k-median problem is specified by a triple $(\mathbb{R}^d, \mathcal{F}, \mathrm{K})$. Here $\mathcal{F}$ is the family of all k-point sets in $\mathbb{R}^d$, and $\mathrm{K} : \mathbb{U}^d \times \mathcal{F} \to \mathbb{R}^{\geq 0}$ is a generalized distance function defined as follows: for a point set $P \in \mathbb{U}^d$ and a k-point set $F \in \mathcal{F}$, $\mathrm{K}(P, F) = \max_{s\in P} \mathrm{d}(s, F) = \max_{s\in P} \min_{f\in F} \mathrm{d}(s, f)$. An instance $\mathbf{S}$ of the generalized k-median problem is a (weighted) collection $\{S_1, \ldots, S_m\}$ ($S_i \in \mathbb{U}^d$) of point sets, and each S_i has a positive weight $w_i \in \mathbb{R}^+$. For any k-point set $F \in \mathcal{F}$, the total generalized distance from $\mathbf{S}$ to F is $\mathrm{K}(\mathbf{S}, F) = \sum_{S_i\in\mathbf{S}} w_i \cdot \mathrm{K}(S_i, F)$. The goal of the*

generalized k-median problem (GKM) is to find a k-point set F which minimizes the total generalized distance $\mathrm{K}(\mathbf{S}, F)$.

Recall that a generalized ε-coreset is a sub-collection $\mathcal{S} \subseteq \mathbf{S}$ of point sets, together with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$, such that for any k-point set $F \in \mathcal{F}$, we have $\sum_{S\in\mathcal{S}} w'(S) \cdot \mathrm{K}(S, F) \in (1 \pm \varepsilon) \sum_{S\in\mathbf{S}} w(S) \cdot \mathrm{K}(S, F)$ (or $\mathrm{K}(\mathcal{S}, F) \in (1 \pm \varepsilon)\mathrm{K}(\mathbf{S}, F)$). This generalized coreset will serve as the SKC-Coreset for the original stochastic k-center problem.

Our main lemma asserts that a constant sized generalized coreset exists, as follows.

Lemma 73. *(main lemma) Given an instance $\mathcal{P}$ of n stochastic points in $\mathbb{R}^d$, let $\mathcal{E}(\mathcal{P})$ be the collection of all additive ε-coresets. There exists a generalized ε-coreset $\mathcal{S} \subseteq \mathcal{E}(\mathcal{P})$ of cardinality $|\mathcal{S}| = O(\varepsilon^{-(d+2)}dk^4)$, together with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$, which satisfies that for any k-point set $F \in \mathcal{F}$,*

$$\sum_{S\in\mathcal{S}} w'(S) \cdot \mathrm{K}(S, F) \in (1 \pm \varepsilon) \sum_{S\in\mathcal{E}(\mathcal{P})} \Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S] \cdot \mathrm{K}(S, F).$$

Now, we prove Lemma 73 by showing a constant upper bound on the cardinality of a generalized ε-coreset. This is done by applying Lemma 65 and providing constant upper bounds for both the total sensitivity and the generalized dimension of the generalized k-median instance.

Given an instance $\mathbf{S} = \{S_i \mid S_i \in \mathbb{U}^d, 1 \le i \le n\}$ of a generalized k-median problem with a weight function $w : \mathbf{S} \to \mathbb{R}^+$, we denote F^* to be the k-point set which minimizes the total generalized distance $\mathrm{K}(\mathbf{S}, F) = \sum_{S\in\mathbf{S}} w(S) \cdot \mathrm{K}(S, F)$ over all $F \in \mathcal{F}$. W.l.o.g., we assume that $\mathrm{K}(\mathbf{S}, F^*) > 0$. Since if $\mathrm{K}(\mathbf{S}, F^*) = 0$, there are at most k different points in the instance.

We first construct a *projection instance* P^* of a weighted k-median problem for $\mathbf{S}$, and relate the total sensitivity $\mathfrak{G}_{\mathbf{S}}$ to $\mathfrak{G}_{P^*}$. Recall that $\mathfrak{G}_{\mathbf{S}} = \sum_{S\in\mathbf{S}} \sigma_{\mathbf{S}}(S)$ is the total sensitivity of $\mathbf{S}$. Our construction of P^* is as follows. For each point set $S_i \in \mathbf{S}$, assume that $F_i^* \in \mathcal{F}$ is the k-point set satisfying that $F_i^* = \mathsf{argmax}_F \frac{w(S_i)\cdot\mathrm{K}(S_i,F)}{\mathrm{K}(\mathbf{S},F)}$, i.e., the sensitivity $\sigma_{\mathbf{S}}(S_i)$ of S_i is equal to $\frac{w(S_i)\mathrm{K}(S_i,F_i^*)}{\mathrm{K}(\mathbf{S},F_i^*)}$. Let $s_i^* \in S_i$ denote the point

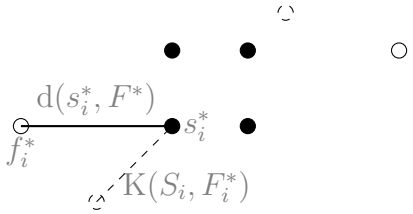

Figure 5-2: In the figure, S_i is the black point set, F^* is the white point set, and F_i^* is the dashed point set. Here, $s_i^* \in S_i$ is the farthest point to F_i^* satisfying $\mathrm{d}(s_i^*, F_i^*) = \mathrm{K}(S_i, F_i^*)$, and $f_i^* \in F^*$ is the closest point to s_i^* satisfying $\mathrm{d}(s_i^*, f_i^*) = \mathrm{d}(s_i^*, F^*)$.

farthest to F_i^* (breaking ties arbitrarily). Let $f_i^* \in F^*$ denote the point closest to s_i^* (breaking ties arbitrarily). Denote P^* to be the multi-set $\{f_i^* \mid S_i \in \mathbf{S}\}$, and denote the weight function $w' : P^* \to \mathbb{R}^+$ to be $w'(f_i^*) = w(S_i)$ for any $i \in [n]$. Thus, P^* is a weighted k-median instance in $\mathbb{R}^d$ with a weight function w'. See Figure 5-2 for an example of the construction of P^*.

Lemma 74. *Given an instance $\mathbf{S} = \{S_i \mid S_i \in \mathbb{U}^d, 1 \le i \le n\}$ of a generalized k-median problem in $\mathbb{R}^d$ with a weight function $w : \mathbf{S} \to \mathbb{R}^+$, let P^* be its projection instance. Then, we have $\mathfrak{G}_{\mathbf{S}} \le 2\mathfrak{G}_{P^*} + 1$.*

Proof. First note that we have the following fact. Given $i, j \in [n]$, recall that $s_j^* \in S_j$ is the farthest point to F_j^*, and $f_j^* \in F^*$ is the closest point to s_j^*. Let $f \in F_i^*$ be the point closest to s_j^*.

$$\begin{aligned}\mathrm{K}(S_j, F_i^*) + \mathrm{K}(S_j, F^*) &\ge \mathrm{d}(s_j^*, F_i^*) + \mathrm{d}(s_j^*, F^*) = \mathrm{d}(s_j^*, F_i^*) + \mathrm{d}(s_j^*, f_j^*) \\ &= \mathrm{d}(s_j^*, f) + \mathrm{d}(s_j^*, f_j^*) \ge \mathrm{d}(f_j^*, f) \ge \mathrm{d}(f_j^*, F_i^*), \end{aligned} \tag{5.3}$$

The first inequality follows from the definitions of $\mathrm{K}(S_j, F_i^*)$ and $\mathrm{K}(S_j, F^*)$. The first equality follows from the definition of f_j^*. The second inequality follows from the triangle inequality, and the last inequality is by the definition of $\mathrm{d}(f_j^*, F_i^*)$.

Then we have the following fact:

$$\begin{aligned}\sum_{f \in P^*} w'(f) \cdot \mathrm{d}(f, F_i^*) = \sum_{f_j^* \in P^*} w'(f_j^*) \cdot \mathrm{d}(f_j^*, F_i^*) &\le \sum_{S_j \in \mathbf{S}} w(S_j) \cdot \big(\mathrm{K}(S_j, F^*) + \mathrm{K}(S_j, F_i^*)\big) \\ &= \mathrm{K}(\mathbf{S}, F^*) + \mathrm{K}(\mathbf{S}, F_i^*) \le 2\mathrm{K}(\mathbf{S}, F_i^*), \end{aligned} \tag{5.4}$$

since $\mathrm{K}(\mathbf{S}, F^*) \leq \mathrm{K}(\mathbf{S}, F_i^*)$ and Inequality (5.3).

Let $f' \in F_i^*$ be the point closest to f_i^*. We also notice the following fact:

$$\begin{aligned}\mathrm{K}(S_i, F^*) + \mathrm{d}(f_i^*, F_i^*) &\geq \mathrm{d}(s_i^*, f_i^*) + \mathrm{d}(f_i^*, F_i^*) = \mathrm{d}(s_i^*, f_i^*) + \mathrm{d}(f_i^*, f') \\ &\geq \mathrm{d}(s_i^*, f') \geq \mathrm{d}(s_i^*, F_i^*) = \mathrm{K}(S_i, F_i^*). \end{aligned} \tag{5.5}$$

The first inequality follows from the definition of f_i^*, the second inequality follows from the triangle inequality, and the last inequality follows from the definition of $\mathrm{d}(s_i^*, F_i^*)$.

Now we are ready to analyze $\sigma_{\mathbf{S}}(S_i)$ for some $S_i \in \mathbf{S}$. We can see that

$$\begin{aligned} w(S_i) \cdot \mathrm{K}(S_i, F_i^*) &\leq w(S_i) \cdot \mathrm{K}(S_i, F^*) + w(S_i) \cdot \mathrm{d}(f_i^*, F_i^*) && \text{[by (5.5)]} \\ &\leq w(S_i) \cdot \mathrm{K}(S_i, F^*) + \sigma_{P^*}(f_i^*) \cdot \left(\sum_{f \in P^*} w'(f) \cdot \mathrm{d}(f, F_i^*) \right) && \text{[by the definition of } \sigma_{P^*}\text{]} \\ &\leq w(S_i) \cdot \mathrm{K}(S_i, F^*) + 2\sigma_{P^*}(f_i^*) \cdot \mathrm{K}(\mathbf{S}, F_i^*) && \text{[by (5.4)]} \\ &= \frac{w(S_i) \cdot \mathrm{K}(S_i, F^*)}{\mathrm{K}(\mathbf{S}, F_i^*)} \cdot \mathrm{K}(\mathbf{S}, F_i^*) + 2\sigma_{P^*}(f_i^*) \cdot \mathrm{K}(\mathbf{S}, F_i^*) \\ &\leq \left(\frac{w(S_i) \cdot \mathrm{K}(S_i, F^*)}{\mathrm{K}(\mathbf{S}, F^*)} + 2\sigma_{P^*}(f_i^*) \right) \mathrm{K}(\mathbf{S}, F_i^*). && \text{[by } \mathrm{K}(\mathbf{S}, F_i^*) \geq \mathrm{K}(\mathbf{S}, F^*)\text{]} \end{aligned}$$

Finally, we bound the total sensitivity as follows:

$$\mathfrak{G}_{\mathbf{S}} = \sum_{S_i \in \mathbf{S}} \sigma_{\mathbf{S}}(S_i) \leq \sum_{S_i \in \mathbf{S}} \left(\frac{w(S_i) \cdot \mathrm{K}(S_i, F^*)}{\mathrm{K}(\mathbf{S}, F^*)} + 2\sigma_{P^*}(f_i^*) \right) = 1 + 2\mathfrak{G}_{P^*}.$$

This finishes the proof of the lemma. □

Since P^* is an instance of a weighted k-median problem, we know that the total sensitivity $\mathfrak{G}_{P^*}$ is at most $2k+1$, by [77, Theorem 9]. [6] Then combining Lemma 74, we have the following lemma which bounds the total sensitivity of $\mathfrak{G}_{\mathbf{S}}$.

Lemma 75. *Consider an instance* $\mathbf{S}$ *of a generalized* k*-median problem* $(\mathbb{R}^d, \mathcal{F}, \mathrm{K})$*. The total sensitivity* $\mathfrak{G}_{\mathbf{S}}$ *is at most* $4k+3$*.*

[6] Theorem 9 in [77] bounds the total sensitivity for the unweighted version. However, the proof can be extended to the weighted version in a straightforward way.

Now the remaining task is to bound the generalized dimension $\dim(\mathbf{S})$. Consider the range space $(\mathbf{S}, \mathcal{R})$, $\mathcal{R}$ is a family of subsets $R_{F,r}$ of $\mathbf{S}$ defined as follows: given an $F \in \mathcal{F}$ and $r \geq 0$, let $R_{F,r} = \{S_i \in \mathbf{S} \mid w_i \cdot \mathrm{K}(S_i, F) \geq r\} \in \mathcal{R}$. Here w_i is the weight of $S_i \in \mathbf{S}$. We have the following lemma.

Lemma 76. *Consider an instance $\mathbf{S}$ of a generalized k-median problem in $\mathbb{R}^d$. If each point set $S \in \mathbf{S}$ is of size at most L, then the generalized dimension $\dim(\mathbf{S})$ is $O(dkL)$.*

Proof. Consider a mapping $g : \mathbf{S} \to \mathbb{R}^{dL}$ constructed as follows: suppose $S_i = \{x^1 = (x_1^1, \ldots, x_d^1), \ldots, x^L = (x_1^L, \ldots, x_d^L)\}$ (if $|S_i| < L$, we pad it with $x^1 = (x_1^1, \ldots, x_d^1)$). We let

$$g(S_i) = (x_1^1, \ldots, x_d^1, \ldots, x_1^L, \ldots, x_d^L) \in \mathbb{R}^{dL}.$$

For any $t \geq 0$ and any k-point set $F \in \mathcal{F}$, we observe that $w_i \cdot \mathrm{K}(S_i, F) \geq r$ holds if and only if there exists some $1 \leq j \leq L$ satisfying that $w_i \cdot \mathrm{d}(x^j, F) \geq r$, which is equivalent to saying that point $g(S_i)$ is in the union of the following L sets $\{(x_1^1, \ldots, x_d^1, \ldots, x_1^L, \ldots, x_d^L) \mid \mathrm{d}(x^j, F) \geq r/w_i\}$ $(j \in [L])$.

Let X be the image set of g. Let $(X, \mathcal{R}^j)$ $(1 \leq j \leq L)$ be L range spaces, where each $\mathcal{R}^j$ consists of all subsets $R_{F,r}^j = \{(x_1^1, \ldots, x_d^1, \ldots, x_1^L, \ldots, x_d^L) \in X \mid \mathrm{d}(x^j, F) \geq r\}$ for all $F \in \mathcal{F}$ and $r \geq 0$. Note that each $(X, \mathcal{R}^j)$ has shattering dimension dk by [45]. Let $\mathcal{R}' = \{\cup R_j \mid R_j \in \mathcal{R}^j, i \in [L]\}$. Using the standard result for bounding the shattering dimension of the union of set systems (e.g.,[58, Theorem 5.22]), we can see that the shattering dimension of $(X, \mathcal{R}')$ (which is the generalized dimension of $\mathbf{S}$) is bounded by $O(dkL)$. □

Note that an additive ε-coreset is of size at most $O(k/\varepsilon^d)$. Then combining Lemma 65, 75 and 76, we directly obtain Lemma 73. Combining Lemma 69 and 73, we have the following theorem.

Theorem 77. *Given an instance $\mathcal{P}$ of n points in $\mathbb{R}^d$ in the existential uncertainty model, there exists an* SKC-Coreset *$\mathcal{S}$ of $\tilde{O}(\varepsilon^{-(d+2)} d^2 k^4)$ [7] point sets with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$, which satisfies that,*

[7] Here, we hide a $\log k + \log(1/\varepsilon)$ term in the notion $\tilde{O}()$.

1. *For each point set* $S \in \mathcal{S}$*, we have* $S \subseteq \mathcal{P}$ *and* $|S| = O(k/\varepsilon^d)$.

2. *For any* k*-point set* $F \in \mathcal{F}$*, we have* $\sum_{S\in\mathcal{S}} w'(S) \cdot \mathrm{K}(S, F) \in (1 \pm \varepsilon)\mathrm{K}(\mathcal{P}, F)$.

PTAS for stochastic minimum k-center. It remains to give a PTAS for the stochastic minimum k-center problem. For an instance $\mathcal{E}(\mathcal{P})$ of a generalized k-median problem, if we can compute the sensitivity $\sigma_{\mathcal{E}(\mathcal{P})}(S)$ efficiently for each point set $S \in \mathcal{E}(\mathcal{P})$, then we can construct an SKC-Coreset by importance sampling (The details of the sampling technique are the same as described in [45, Section 4.1]). However, it is unclear how to compute the sensitivity $\sigma_{\mathcal{E}(\mathcal{P})}(S)$ efficiently. Instead, we enumerate all weighted sub-collections $\mathcal{S}_i \subseteq \mathcal{E}(\mathcal{P})$ of cardinality at most $\tilde{O}(\varepsilon^{-(d+2)}d^2k^4)$. We claim that we only need to enumerate $O(n^{\tilde{O}(\varepsilon^{-(2d+2)}d^2k^5)})$ polynomially many sub-collections $\mathcal{S}_i$ together with their weight functions, such that there exists a generalized ε-coreset of $\mathcal{E}(\mathcal{P})$. [8] We will show the details later.

In the next step, for each weighted sub-collection $\mathcal{S} \subseteq \mathcal{E}(\mathcal{P})$ with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$, we briefly sketch how to compute the optimal k-point set F such that $\mathrm{K}(\mathcal{S}, F)$ is minimized. We cast the optimization problem as a constant size polynomial system.

Denote the space $\mathcal{F} = \{(y^1, \ldots, y^k) \mid y^i \in \mathbb{R}^d, 1 \le i \le k\}$ to be the collection of ordered k-point sets ($(y^1, y^2, \ldots, y^k) \in \mathcal{F}$ and $(y^2, y^1, \ldots, y^k) \in \mathcal{F}$ to be two different k-point sets if $y^1 \neq y^2$). We first divide the space $\mathcal{F}$ into pieces $\{\mathcal{F}^i\}$, as follows: Let $L = O(k/\varepsilon^d)$ and $\mathcal{L} = (l_1, \ldots, l_L)$ $(1 \le l_j \le k, \forall j \in [L])$ be a sequence of integers, and let $b \in [L]$ be an index. Consider a point set $S = \{x^1 = (x^1_1, \ldots, x^1_d), \ldots, x^L = (x^L_1, \ldots, x^L_d)\} \in \mathcal{S}$ and a k-point set $F = \{y^1 = (y^1_1, \ldots, y^1_d), \ldots, y^k = (y^k_1, \ldots, y^k_d)\} \in \mathcal{F}$. We give the following definition.

Definition 78. *The* k*-center value* $\mathrm{K}(S, F)$ *is* decided *by* $\mathcal{L}$ *and* b *if the following two properties hold.*

1. *For any* $i \in [L]$ *and any* $j \in [k]$, $\mathrm{d}(x^i, y^{l_i}) \le \mathrm{d}(x^i, y^j)$*, i.e., the closest point to* x^j *is* $y^{l_j} \in F$.

[8] We remark that even though we enumerate the weight function, computing $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P) = S]$ is still important for our algorithm. See Lemma 81 for the details of the enumeration algorithm.

2. *For any* $i \in [L]$, $\mathrm{d}(x^i, y^{l_i}) \leq \mathrm{d}(x^b, y^{l_b})$, *i.e., the* k*-center value* $\mathrm{K}(S, F) = \mathrm{d}(x^b, y^{l_b})$.

For each point set $S_i \in \mathcal{S}$, we enumerate an integer sequence $\mathcal{L}_i$ and an index b_i. Given a collection $\{\mathcal{L}_i, b_i\}_i$ (index i ranges over all S_i in $\mathcal{S}$), we construct a piece $\mathcal{F}^{\{\mathcal{L}_i,b_i\}_i} \subseteq \mathcal{F}$ as follows: for any point set $S_i \in \mathcal{S}$ and any k-point set $F \in \mathcal{F}^{\{\mathcal{L}_i,b_i\}_i}$, the k-center value $\mathrm{K}(S_i, F)$ is *decided* by $\mathcal{L}_i$ and b_i. According to Definition 78, $\mathcal{F}^{\{\mathcal{L}_i,b_i\}_i}$ is defined by a polynomial system.

Then, we solve our optimization problem in each piece $\mathcal{F}^{\{\mathcal{L}_i,b_i\}_i}$. By definition 78, for any point set $S_i \in \mathcal{S}$ and any k-point set $F \in \mathcal{F}^{\{\mathcal{L}_i,b_i\}_i}$, the k-center value $\mathrm{K}(S_i, F) = \mathrm{d}(x^{b_i}, y^{\mathcal{L}_i(b_i)})$ ($x^{b_i} \in S_i$, $y^{\mathcal{L}_i(b_i)} \in F$). Here, the index $\mathcal{L}_i(b_i)$ is the b_i-th item of $\mathcal{L}_i$. Hence, our problem can be formulated as the following optimization problem:

$$\min_{F} \sum_{S_i \in \mathcal{S}} w'(S_i) \cdot g_i, \quad \text{s.t.,}\ g_i^2 = \|x^{b_i} - y^{\mathcal{L}_i(b_i)}\|^2, g_i \geq 0, \forall i \in [L]; y^{\mathcal{L}_i(b_i)} \in F; F \in \mathcal{F}^{\{\mathcal{L}_i,b_i\}_i}.$$

By Definition 78, there are at most $kL|\mathcal{S}|$ constraints, which is a constant. Thus, the polynomial system has dk variables and $O(kL|\mathcal{S}|)$ constraints, hence can be solved in constant time. Note that there are at most $O(k^{L|\mathcal{S}|})$ different pieces $\mathcal{F}^{\{\mathcal{L}_i,b_i\}_i} \subseteq \mathcal{F}$, which is again a constant. Thus, we can compute the optimal k-point set for the weighted sub-collection $\mathcal{S}$ in constant time.

Now we return to the stochastic minimum k-center problem. Recall that we first enumerate all possible weighted sub-collections $\mathcal{S}_i \subseteq \mathcal{E}(\mathcal{P})$ of cardinality at most $\tilde{O}(\varepsilon^{-(d+2)} d^2 k^4)$. Then we compute the optimal k-point set F^i for each weighted sub-collection $\mathcal{S}_i$ as above, and compute the expected k-center value $\mathrm{K}(\mathcal{P}, F^i)$. [9] Let $F^* \in \mathcal{F}$ be the k-point set which minimizes the expected k-center value $\mathrm{K}(\mathcal{P}, F^i)$ over all F^i. By Lemma 81, there is one sub-collection $\mathcal{S}_i$ with a weight function w' satisfying that $\mathrm{K}(\mathcal{S}_i, F^i) \leq (1+\varepsilon) \min_{F \in \mathcal{F}} \mathrm{K}(\mathcal{P}, F)$. Thus, we conclude that F^* is a $(1+\varepsilon)$-approximation for the stochastic minimum k-center problem. For the running

[9] It is not hard to compute $\mathrm{K}(\mathcal{P}, F^i)$ in $O(n \log n)$ time by sorting all points in $\mathcal{P}$ in non-increasing order according to their distances to F^i.

time, we enumerate at most $O(n^{\tilde{O}(\varepsilon^{-(2d+2)}d^2k^5)})$ weighted sub-collections. Moreover, computing the optimal k-point set for each sub-collection costs constant time. Then the total running time is at most $O(n^{\tilde{O}(\varepsilon^{-(2d+2)}d^2k^5)})$. Thus, we have the following corollary.

Corollary 79. *If both k and d are constants, given an instance $\mathcal{P}$ of n stochastic points in $\mathbb{R}^d$ in the existential uncertainty model, there exists a PTAS for the stochastic minimum k-center problem in $O(n^{\tilde{O}(\varepsilon^{-(2d+2)}d^2k^5)})$ time.*

Enumerating possible generalized ε-coresets. Given an instance $\mathbf{S} = \{S_i \mid S_i \in \mathbb{U}^d, 1 \le i \le N\}$ of a generalized k-median problem in $\mathbb{R}^d$ with a weight function $w : \mathbf{S} \to \mathbb{R}^+$, now we show how to enumerate polynomially many sub-collections $\mathcal{S}_i \subseteq \mathbf{S}$ together with their weight functions, such that there exists a generalized ε-coreset of $\mathbf{S}$. Recall that $\sigma_{\mathbf{S}}(S_i)$ is the sensitivity of S_i, and $\mathfrak{G}_{\mathbf{S}} = \sum_{i\in[N]} \sigma_{\mathbf{S}}(S_i)$ is the total sensitivity. Also recall that $\dim(\mathbf{S})$ is the generalized dimension of $\mathbf{S}$. Define $q(S_i) = \sigma_{\mathbf{S}}(S_i) + 1/N$ for $1 \le i \le M$, and define $q_{\mathbf{S}} = \sum_{1\le i\le N} q(S_i)$. Note that $q_{\mathbf{S}} = \mathfrak{G}_{\mathbf{S}} + 1 \le 4k + 4$ by Lemma 75. Our algorithm is as follows.

1. Let $M = O((\frac{q_{\mathbf{S}}}{\varepsilon})^2 \dim(\mathbf{S}))$. Let $L = \frac{10}{\varepsilon}(\log M + \log N + \log k)$.
2. Enumerate all collections $\mathcal{S}_i \subseteq \mathbf{S}$ of cardinality at most M. Note that we only need to enumerate at most N^M collections.
3. For a collection $\mathcal{S} \subseteq \mathbf{S}$, w.l.o.g., assume that $\mathcal{S} = \{S_1, S_2, \cdots, S_m\}$ $(m \le M)$. Enumerate all sequences $\left((1+\varepsilon)^{a_1}, \ldots, (1+\varepsilon)^{a_m}\right)$ where each $0 \le a_i \le L$ is an integer.
4. Given a collection $\mathcal{S} = \{S_1, S_2, \cdots, S_m\}$ and a sequence $\left((1+\varepsilon)^{a_1}, \ldots, (1+\varepsilon)^{a_m}\right)$, we construct a weight function $w' : \mathcal{S} \to \mathbb{R}^+$ as follows: for a point set $S_i \in \mathcal{S}$, denote $w'(S_i)$ to be $(1+\varepsilon)^{a_i} \cdot w(S_i)/M$. Recall that $w(S_i)$ is the weight of $S_i \in \mathbf{S}$.

Analysis. Recall that given an instance $\mathcal{P}$ of a stochastic minimum k-center problem, we first reduce to an instance $\mathbf{S} = \mathcal{E}(\mathcal{P})$ of a generalized k-median problem. Note

that the cardinality of $\mathbf{S}$ is at most $n^{O(k/\varepsilon^d)}$, and the cardinality of a generalized ε-coreset is at most $M = \tilde{O}(\varepsilon^{-(d+2)}d^2k^4)$ by Theorem 77. Thus, we enumerate at most $N^M = n^{\tilde{O}(\varepsilon^{-(2d+2)}d^2k^5)}$ polynomially many sub-collections $\mathcal{S}_i \subseteq \mathbf{S}$. For each collection $\mathcal{S}_i$, we construct at most $M^{L+1} = O(n^{O(k/\varepsilon^d)})$ polynomially many weight functions. In total, we enumerate $N^M \cdot M^{L+1} = O(n^{\tilde{O}(\varepsilon^{-(2d+2)}d^2k^5)})$ polynomially many weighted sub-collections.

It remains to show that there exists a generalized ε-coreset of $\mathbf{S}$. We first have the following lemma.

Lemma 80. *Given an instance $\mathbf{S} = \{S_i \mid S_i \in \mathbb{U}^d, 1 \leq i \leq N\}$ of a generalized k-median problem in $\mathbb{R}^d$ with a weight function $w : \mathbf{S} \to \mathbb{R}^+$, there exists a generalized ε-coreset $\mathcal{S} \subseteq \mathbf{S}$ with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$, such that*

$$\sum_{S \in \mathcal{S}} w'(S) \cdot \mathrm{K}(S, F) \in (1 \pm \varepsilon) \sum_{S \in \mathbf{S}} w(S) \cdot \mathrm{K}(S, F).$$

The cardinality of $\mathcal{S}$ is at most $M = O((\frac{q_{\mathbf{S}}}{\varepsilon})^2 \dim(\mathbf{S}))$. Moreover, each weight $w'(S)$ $(S \in \mathcal{S})$ has the form that $w'(S) = \frac{c \cdot q_{\mathbf{S}} \cdot w(S)}{q(S) \cdot M}$, where $1 \leq c \leq M$ is an integer.

Proof. For each $S \in \mathbf{S}$, let $g_S : \mathcal{F} \to \mathbb{R}^+$ be defined as $g_S(F) = w(S) \cdot \mathrm{K}(S, F)/q(S)$. Let $D = \{g_S \mid S \in \mathbf{S}\}$ be a collection, together with a weight function $w'' : D \to \mathbb{R}^+$ defined as $w''(g_S) = q(S)$. Note that for any k-point set $F \in \mathcal{F}$, we have that

$$\sum_{g_S \in G} w''(g_S) \cdot g_S(F) = \sum_{S \in \mathbf{S}} w(S) \cdot \mathrm{K}(S, F) = \mathrm{K}(\mathbf{S}, F).$$

By Theorem 4.1 in [45], we can randomly sample (with replacement) a collection $\mathcal{S} \subseteq D$ of cardinality at most $M = O((\frac{q_{\mathbf{S}}}{\varepsilon})^2 \dim(\mathbf{S}))$, together with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$ defined as $w'(g_S) = q_{\mathbf{S}}/M$. Then the multi-set $\mathcal{S}$ satisfies that for every $F \in \mathcal{F}$,

$$\sum_{g_S \in \mathcal{S}} w'(g_S) \cdot g_S(F) \in (1 \pm \varepsilon) \sum_{g_S \in G} w''(g_S) \cdot g_S(F) = (1 \pm \varepsilon)\mathrm{K}(\mathbf{S}, F).$$

By the definition of g_S and w', we prove the lemma. □

We are ready to prove the following lemma.

Lemma 81. *Among all sub-collections $\mathcal{S} \subseteq \mathbf{S}$ of cardinality at most $M = O((\frac{q_{\mathbf{S}}}{\varepsilon})^2 \dim(\mathbf{S}))$, together with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$ of the form $w'(S_i) = (1+\varepsilon)^{a_i} \cdot w(S_i)/M$ ($0 \leq a_i \leq \frac{10(\log M + \log N + \log k)}{\varepsilon}$ is an integer), there exists a generalized ε-coreset of $\mathbf{S}$.*

Proof. By Lemma 80, there exists a generalized ε-coreset $\mathcal{S} \subseteq \mathcal{S}$ of cardinality at most M together with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$ defined as follows: each weight $w'(S)$ ($S \in \mathcal{S}$) has the form that $w'(S) = \frac{c_S \cdot q_{\mathbf{S}} \cdot w(S)}{q(S) \cdot M}$ for some integer $1 \leq c_S \leq M$. W.l.o.g., we assume that $\mathcal{S} = \{S_1, S_2, \ldots, S_m \mid S_i \in \mathbf{S}\}$ ($m \leq M$).

By the definition of $q(S)$, we have that $1/N \leq q(S) \leq q_{\mathbf{S}} = \mathfrak{G}_{\mathbf{S}} + 1 \leq 4k + 4$. Then we conclude that for each $S \in \mathcal{S}$,

$$1 \leq \frac{c_S \cdot q_{\mathbf{S}}}{q(S)} \leq (4k+4)MN.$$

For $1 \leq i \leq m$, let $a_i = \lfloor \log_{1+\varepsilon}(\frac{c_{S_i} \cdot q_{\mathbf{S}}}{q(S_i)}) \rfloor$. Note that each a_i satisfies that $0 \leq a_i \leq \frac{10(\log M + \log N + \log k)}{\varepsilon}$. Thus, we have enumerated the following sub-collection $\mathcal{S} = \{S_1, S_2, \cdot, S_m \mid S_i \in \mathbf{S}\}$ with a weight function $w'' : \mathcal{S} \to \mathbb{R}^+$, such that $w''(S_i) = (1+\varepsilon)^{a_i} \cdot w(S_i)/M$. Moreover, for any k-point set F, we have the following inequality.

$$\begin{aligned}\sum_{1 \leq i \leq m} w''(S_i) \cdot \mathrm{K}(S_i, F) &= \sum_{1 \leq i \leq m} \frac{(1+\varepsilon)^{a_i} \cdot w(S_i)}{M} \cdot \mathrm{K}(S_i, F) \in (1 \pm \varepsilon) \sum_{1 \leq i \leq m} \frac{c_S \cdot q_{\mathbf{S}} \cdot w(S)}{q(S_i) \cdot M} \cdot \mathrm{K}(S_i, F) \\ &= (1 \pm \varepsilon) \sum_{1 \leq i \leq m} w'(S_i) \cdot \mathrm{K}(S_i, F) \in (1 \pm 3\varepsilon) \sum_{S \in \mathbf{S}} w(S) \cdot \mathrm{K}(S, F).\end{aligned}$$

The last inequality is due to the assumption that the sub-collection $\mathcal{S}$ with a weight function w' is a generalized ε-coreset of $\mathbf{S}$. Let $\varepsilon' = \varepsilon/3$, we prove the lemma. □

5.3.2 Locational uncertainty model

Next, we consider the stochastic minimum k-center problem in the locational uncertainty model. Given an instance of m nodes $v_1, \ldots, v_m$ which may locate in the point set $\mathcal{P} = \{s_1, \ldots, s_n \mid s_i \in \mathbb{R}^d, 1 \leq i \leq n\}$, our construction of additive ε-coresets and the method for bounding the total sensitivity is exactly the same as in the existential

uncertainty model. The only difference is that for an additive ε-coreset S, how to compute the probability $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]=\sum_{P:P\sim\mathcal{P},\mathcal{E}(P)=S}\Pr[\models P]$. Here, $P\sim\mathcal{P}$ is a realized point set according to the probability distribution of $\mathcal{P}$. Run $\mathbb{A}(S)$, and construct a Cartesian grid $G(S)$. Denote $T(S)=\left(\cup_{P:P\sim\mathcal{P},\mathcal{E}(P)=S}P\right)\setminus S$ to be the collection of all points s which might be contained in some realization $P\sim\mathcal{P}$ with $\mathcal{E}(P)=S$. Recall that $\mathcal{C}(S)=\{C\in G\mid |C\cap S|=1\}$ is the collection of d-dimensional Cartesian cells C which contains a point $s^C\in S$. By Lemma 70, for any realization P with $\mathcal{E}(P)=S$, we have the following observations.

1. For any cell $C\notin\mathcal{C}(S)$, $C\cap P=\emptyset$. It means that for any point $s\in C\cap\mathcal{P}$, we have $s\notin T(S)$.

2. For any cell $C\in\mathcal{C}(S)$ and any point $s'\in C\cap\mathcal{P}$ with a smaller index than that of s^C, we have $s'\notin P$. It means that $s'\notin T(S)$.

By the above observations, we conclude that $T(S)$ is the collection of those points s' belonging to some cell $C\in\mathcal{C}(S)$ and with a larger index than that of s^C.

Then we reduce the counting problem $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]$ to a family of bipartite holant problems. We first give the definition of holant problems.

Definition 82. *An instance of a holant problem is a tuple* $\Lambda=\left(G(V,E),(g_u)_{u\in V}\right),(w_e)_{e\in E}$, *where for every* $u\in V$, $g_u:\{0,1\}^{E_u}\to\mathbb{R}^+$ *is a function, where* E_u *is the set of edges incident to* $\vec{u}$. *For every assignment* $\sigma\in\{0,1\}^E$, *we define the weight of* σ *as*

$$w_\Lambda(\sigma)\triangleq\prod_{u\in V}g_u\left(\sigma\mid_{E_u}\right)\prod_{e\in\sigma}w_e.$$

Here $\sigma\mid_{E_u}$ *is the assignment of* E_u *with respect to the assignment* σ. *We denote the value of the holant problem* $Z(\Lambda)\triangleq\sum_{\sigma\in\{0,1\}^E}w_\Lambda(\sigma)$.

For a counting problem $\Pr_{P\sim\mathcal{P}}[\mathcal{E}(P)=S]$, w.l.o.g., we assume that $S=\{s_1,\ldots,s_{|S|}\}$. Then we construct a family of holant instance $\Lambda_{\mathcal{L}}$ as follows.

1. Enumerate all integer sequences $\mathcal{L}=(l_1,\ldots,l_{|S|},l_t)$ such that $\sum_{1\le i\le|S|}l_i+l_t=n$, $l_i\ge 1$ $(1\le i\le|S|)$, and $l_t\ge 0$. Let $\mathbf{L}$ be the collection of all these integer sequences $\mathcal{L}$.

2. For a sequence $\mathcal{L}$, assume that $\Lambda_{\mathcal{L}} = \left(G(U, V, E), (g_{\vec{u}})_{u \in U \cup V}\right)$ is a holant instance on a bipartite graph, where $V = \{v_1, \ldots, v_n\}$, and $U = S \cup \{t\}$ (we use vertex t to represent the collection $T(S)$).

3. The weight function $w : E \to \mathbb{R}^+$ is defined as follows:

   (a) For a vertex $v_i \in V$ and a vertex $s_j \in S$, $w_{ij} = p_{ij}$.

   (b) For a vertex $v_i \in V$ and $t \in V$, $w_{it} = \sum_{s_j \in T(S)} p_{ij}$.

4. For each vertex $v \in V$, the function $g_v = (= 1)$. [10] For each vertex $s_i \in S$, the function $g_{s_i} = (= l_i)$, and the function $g_t = (= l_t)$.

Since each $S \in \mathcal{E}(\mathcal{P})$ is of constant size, we only need to enumerate at most $O(m^{|S|+1}) = \text{poly}(n)$ integer sequences $\mathcal{L}$. Given an integer sequence $\mathcal{L} = (l_1, \ldots, l_{|S|}, l_t)$, we can see that $Z(\Lambda_{\mathcal{L}})$ is exactly the probability that l_i nodes are realized at point $s_i \in S$ ($\forall 1 \leq i \leq |S|$), and l_t nodes are realized inside the point set $T(S)$. Then by Lemma 70, we have the following equality:

$$\Pr_{P \sim \mathcal{P}}[\mathcal{E}(P) = S] = \sum_{\mathcal{L} \in \mathbf{L}} Z(\Lambda_{\mathcal{L}}).$$

It remains to show that we can compute each $Z(\Lambda_{\mathcal{L}})$ efficiently. Fortunately, we have the following lemma.

Lemma 83. *([67],[103]) For any bipartite graph $\Lambda_{\mathcal{L}}$ with a specified integer sequence $\mathcal{L}$, there exists an FPRAS to compute the holant value $Z(\Lambda_{\mathcal{L}})$.*

Thus, we have the following theorem.

Theorem 84. *If both k and d are constants, given an instance of m stochastic nodes in $\mathbb{R}^d$ in the locational uncertainty model, there exists a PTAS for the stochastic minimum k-center problem.*

Combining Theorem 77 and 84, we obtain the main result Theorem 12.

[10] Here the function $g_v = (= i)$ means that the function value g_v is 1 if exactly i edges incident to v are of value 1 in the assignment. Otherwise, $g_v = 0$

5.4 Stochastic Minimum j-Flat-Center

In this section, we consider a generalized shape fitting problem, the minimum j-flat-center problem in the stochastic models. Let $\mathcal{F}$ be the family of all j-flats in $\mathbb{R}^d$. Our main technique is to construct an SJFC-Coreset of constant size, which satisfies that for any j-flat $F \in \mathcal{F}$, we can use the SJFC-Coreset to obtain a $(1\pm\varepsilon)$-estimation for the expected j-flat-center value $\mathrm{J}(\mathcal{P}, F)$. Then since the SJFC-Coreset is of constant size, we have a polynomial system of constant size to compute the optimum in constant time.

Let $B = \sum_{1\leq i\leq n} p_i$ be the total probability. We discuss two different cases. If $B < \varepsilon$, we reduce the problem to a weighted j-flat-median problem, which has been studied in [106]. If $B \geq \varepsilon$, the construction of an SJFC-Coreset can be divided into two parts. We first construct a convex hull, such that with high probability (say $1-\varepsilon$) that all points are realized inside the convex hull. Then we construct a collection of point sets to estimate the contribution of points insider the convex hull. On the other hand, for the case that some point appears outside the convex hull, we again reduce the problem to a weighted j-flat-median problem. The definition of the weighted j-flat-median problem is as follows.

Definition 85. *For some $0 \leq j \leq d-1$, let $\mathcal{F}$ be the family of all j-flats in $\mathbb{R}^d$. Given a set P of n points in $\mathbb{R}^d$ together with a weight function $w : P \to \mathbb{R}^+$, denote $\mathsf{cost}(P, F) = \sum_{s_i\in P} w_i \cdot \mathrm{d}(s_i, F)$. A weighted j-flat-median problem is to find a shape $F \in \mathcal{F}$ which minimizes the value $\mathsf{cost}(P, F)$.*

5.4.1 Case 1: $B < \varepsilon$

In the first case, we show that the minimum j-flat-center problem can be reduced to a weighted j-flat-median problem. We need the following lemmas.

Lemma 86. *If $B < \varepsilon$, for any j-flat $F \in \mathcal{F}$, we have $\sum_{s_i\in\mathcal{P}} p_i \cdot \mathrm{d}(s_i, F) \in (1 \pm \varepsilon) \cdot \mathrm{J}(\mathcal{P}, F)$.*

Proof. For a j-flat $F \in \mathbb{R}^d$, w.l.o.g., we assume that $\mathrm{d}(s_i, F)$ is non-decreasing in i.

Thus, we have

$$\mathrm{J}(\mathcal{P}, F) = \sum_{i\in[n]} p_i \cdot \mathrm{d}(s_i, F) \cdot \prod_{j>i}(1-p_j)$$

Since $B < \varepsilon$, for any $i \in [n]$, we have that $1-\varepsilon \le 1 - \sum_{j\in[n]} p_i \le \prod_{j>i}(1-p_j) \le 1$. So we prove the lemma. □

By Lemma 86, we reduce the original problem to a weighted j-flat-median problem, where each point $s_i \in \mathcal{P}$ has weight p_i. We then need the following lemma to bound the total sensitivity.

Lemma 87. *(Theorem 18 in [106])* [11] *Consider the weighted j-flat-median problem where $\mathcal{F}$ is the set of all j-flats in $\mathbb{R}^d$. The total sensitivity of any weighted n-point set is $O(j^{1.5})$.*

On the other hand, we know that the dimension of the weighted j-flat-median problem is $O(jd)$ by [45]. Then by Lemma 65, there exists an ε-coreset $\mathcal{S} \subseteq \mathcal{P}$ of cardinality $O(j^4 d\varepsilon^{-2}\log(jd)) = \tilde{O}(j^4 d\varepsilon^{-2})$ to estimate the j-flat-median value $\sum_{s_i\in\mathcal{P}} p_i \cdot \mathrm{d}(s_i, F)$ for any j-flat $F \in \mathcal{F}$. [12] Moreover, we can compute a constant approximation j-flat in $O(ndj^{O(j^2)})$ time by [44]. Then by [106], we can construct an ε-coreseet $\mathcal{S}$ in $O(ndj^{O(j^2)})$ time. Combining Lemma 86, we conclude the main lemma in this subsection.

Lemma 88. *Given an instance $\mathcal{P}$ of n stochastic points in $\mathbb{R}^d$, if the total probability $\sum_i p_i < \varepsilon$, there exists an* SJFC-Coreset *of cardinality $\tilde{O}(j^4 d\varepsilon^{-2})$ for the minimum j-flat-center problem. Moreover, we have an $O(ndj^{O(j^2)})$ time algorithm to compute the* SJFC-Coreset.

5.4.2 Case 2: $B \ge \varepsilon$

Note that if F is a j-flat, the function $\mathrm{d}(x, F)^2$ has a linearization. Here, a linearization is to map the function $\mathrm{d}(x, F)^2$ to a k-variate linear function through variate

[11] Theorem 18 in [106] bounds the total sensitivity for the unweighted version. However, the proof can be extended to the weighted version in a straightforward manner.

[12] We remark that for the j-flat-median problem, Feldman and Langberg [45] showed that there exists a coreset of size $O(jd\varepsilon^{-2})$. However, it is unclear how to generalize their technique to weighted version.

embedding. The number k is called the dimension of the linearization, see [8]. We have the following lemma to bound the dimension of the linearization.

Lemma 89. *([46]) Suppose F is a j-flat in $\mathbb{R}^d$, the function $\mathrm{d}(x,F)^2$ ($x \in \mathbb{R}^d$) has a linerization. Let D be the dimension of the linearization. If $j = 0$, we have $D = d+1$. If $j = 1$, we have $D = O(d^2)$. Otherwise, for $2 \leq j \leq d-1$, we have $D = O(j^2d^3)$.*

Suppose $\mathcal{P}$ is an instance of n stochastic points in $\mathbb{R}^d$. For each j-flat $F \in \mathbb{R}^d$, let $h_F(x) = \mathrm{d}(x,F)^2$ ($x \in \mathbb{R}^d$), which admits a linearization of dimension $O(j^2d^3)$ by Lemma 89. Now, we map each point $s \in \mathcal{P}$ into an $O(j^2d^3)$ dimensional point s' and map each j-flat $F \in \mathbb{R}^d$ into an $O(j^2d^3)$ dimensional direction $\vec{u}$, such that $\mathrm{d}(s,F) = \langle s', \vec{u}\rangle^{1/2}$. For convenience, we still use $\mathcal{P}$ to represent the collection of points after linearization. Recall that $\Pr[\vDash P]$ is the realized probability of the realization $P \sim \mathcal{P}$. By this mapping, we translate our goal into finding a direction $\vec{u} \in \mathbb{R}^{O(j^2d^3)}$, which minimizes the expected value $\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2}] = \sum_{P\sim\mathcal{P}} \Pr[\vDash P] \cdot \max_{x\in P}\langle \vec{u}, x\rangle^{1/2}$. We also denote $\mathcal{P}^\star = \{\vec{u} \in \mathbb{R}^d \mid \langle \vec{u}, s\rangle \geq 0, \forall s \in \mathcal{P}\}$ to be the polar set of $\mathcal{P}$. We only care about the directions in the polar set $\mathcal{P}^\star$ for which $\langle \vec{u}, s\rangle^{1/2}$, $\forall s \in \mathcal{P}$ is well defined.

We first construct a convex hull $\mathcal{H}$ to partition the realizations into two parts. Our construction uses the method of (ε,τ)-QUANT-KERNEL construction in Chapter 4. For any normal vector (direction) $\vec{u}$, we move a sweep line $l_{\vec{u}}$ orthogonal to $\vec{u}$, along the direction $\vec{u}$, to sweep through the points in $\mathcal{P}$. Stop the movement of $\ell_{\vec{u}}$ at the first point such that $\Pr[\mathcal{P} \cap \overline{H}_{\vec{u}})] \geq \varepsilon'$, where $\varepsilon' = \varepsilon^{O(j^2d^3)}$ is a fixed constant. Denote $H_{\vec{u}}$ to be the halfplane defined by the sweep line $\ell_{\vec{u}}$ (orthogonal to the normal vector $\vec{u}$) and $\overline{H}_{\vec{u}}$ to be its complement. Denote $\mathcal{P}(\overline{H}_{\vec{u}}) = \mathcal{P} \cap \overline{H}_{\vec{u}}$ to be the set of points swept by the sweep line $l_{\vec{u}}$. We repeat the above process for all normal vectors (directions) $\vec{u}$, and let $\mathcal{H} = \cap_{\vec{u}} H_{\vec{u}}$. Since the total probability $B \geq \varepsilon$, $\mathcal{H}$ is nonempty by Helly's theorem. We also know that $\mathcal{H}$ is a convex hull by Chapter 4. Moreover, we have the following lemma.

Lemma 90. *(Lemma 39 and Theorem 5) Suppose the dimensionality is d. There is a convex set $\mathcal{K}$, which is an intersection of $O(\varepsilon^{-(d-1)/2})$ halfspaces and satisfies*

$(1-\varepsilon)\mathcal{K} \subseteq \mathcal{H} \subseteq \mathcal{K}$. *Moreover,* $\mathcal{K}$ *can be constructed in* $O(n \log^{O(d)} n)$ *time.*

By the above lemma, we construct a convex set $\mathcal{K} = \cap_{\vec{u}} \mathcal{K}_{\vec{u}}$, which is the intersection of $O(\varepsilon^{-O(j^2d^3)})$ halfspaces $\mathcal{K}_{\vec{u}}$ ($\vec{u}$ is the direction orthogonal to the halfspace $\mathcal{K}_{\vec{u}}$). Let $\overline{\mathcal{K}}_{\vec{u}}$ be the complement of $\mathcal{K}_{\vec{u}}$, and let $\mathcal{P}(\overline{\mathcal{K}}_{\vec{u}}) = \mathcal{P} \cap \overline{\mathcal{K}}_{\vec{u}}$ be the set of points in $\overline{\mathcal{K}}_{\vec{u}}$. Denote $\mathcal{P}(\overline{\mathcal{K}})$ to be the set of points outside the convex set $\mathcal{K}$. Then we have the following lemma, which shows that the total probability outside $\mathcal{K}$ is very small.

Lemma 91. *Let* $\mathcal{K}$ *be a convex set constructed as in Lemma 90. The total probability* $\Pr[\mathcal{P}(\overline{\mathcal{K}})] \leq \varepsilon$.

Proof. Assume that $\mathcal{K} = \cap_{\vec{u}} \mathcal{K}_{\vec{u}}$. Consider a halfspace $\mathcal{K}_{\vec{u}}$. By Lemma 90, the convex set $\mathcal{K}$ satisfies that $\mathcal{H} \subseteq \mathcal{K}$. Thus, we have that $\Pr[\mathcal{P}(\overline{\mathcal{K}}_{\vec{u}})] \leq \Pr[\mathcal{P}(\overline{H}_{\vec{u}})] \leq \varepsilon'$ by the definition of $\overline{H}_{\vec{u}}$.

Note that $\Pr[\mathcal{P}(\overline{\mathcal{K}})]$ is upper bounded by the multiplication of ε' and the number of halfspaces of $\mathcal{K}$. By Lemma 90, there are at most $O(\varepsilon^{-O(j^2d^3)})$ halfspaces $\mathcal{K}_{\vec{u}}$. Thus, we have that $\Pr[\mathcal{P}(\overline{\mathcal{K}})] \leq \varepsilon$. □

Our construction of SJFC-Coreset is consist of two parts. For points inside $\mathcal{K}$, we construct a collection $\mathcal{S}_1$. Our construction is almost the same as (ε, r)-fpow-kernel construction in Chapter 4, except that the cardinality of the collection $\mathcal{S}_1$ is different. For completeness, we provide the details of the construction here. Let $\mathcal{P}(\mathcal{K})$ be the collection of points in $\mathcal{K} \cap \mathcal{P}$, then $\mathcal{P}(\mathcal{K})$ is also an instance of a stochastic minimum j-flat-center problem. We show that we can estimate $\mathbb{E}_{P \sim \mathcal{P}(\mathcal{K})}[\max_{x \in P} \langle \vec{u}, x \rangle^{1/2}]$ by $\mathcal{S}_1$. For the rest points outside $\mathcal{K}$, we show that the contribution for the objective function $\mathbb{E}_{P \sim \mathcal{P}}[\max_{x \in P} \langle \vec{u}, x \rangle^{1/2}]$ is almost linear and can be reduced to a weighted j-flat-median problem as in Case 1.

We first show how to construct $\mathcal{S}_1$ for points inside $\mathcal{K}$ as follows.

1. Sample $N = O((\varepsilon'\varepsilon)^{-2} \varepsilon^{-O(j^2d^3)} \log(1/\varepsilon)) = O(\varepsilon^{-O(j^2d^3)})$ independent realizations restricted to $\mathcal{P}(\mathcal{K})$.

2. For each realization S_i, use the algorithm in [7] to find a deterministic ε-kernel $\mathcal{E}_i$ of size $O(\varepsilon^{-O(j^2d^3)})$. Here, a deterministic ε-kernel $\mathcal{E}_i$ satisfies that $(1 -$

$\varepsilon)CH(S_i) \subseteq CH(\mathcal{E}_i) \subseteq CH(S_i)$, where $CH(\cdot)$ is the convex hull of the point set.

3. Let $\mathcal{S}_1 = \{\mathcal{E}_i \mid 1 \leq i \leq N\}$ be the collection of all ε-kernels, and each ε-kernel $\mathcal{E}_i$ has a weight $1/N$.

Hence, the total size of $\mathcal{S}_1$ is $O(\varepsilon^{-O(j^2d^3)})$. For any direction $\vec{u} \in \mathcal{P}^\star$, we use $\frac{1}{N}\sum_{\mathcal{E}_i \in \mathcal{S}_1} \max_{x \in \mathcal{E}_i} \langle \vec{u}, x\rangle^{1/2}$ as an estimation of $\mathbb{E}_{P\sim\mathcal{P}(\mathcal{K})}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2}]$. By Chapter 4, we have the following lemma.

Lemma 92. *(Lemma 44-46) For any direction $\vec{u} \in \mathcal{P}^\star$, let $M_{\vec{u}} = \max_{x\in\mathcal{P}(\mathcal{K})}\langle \vec{u}, x\rangle^{1/2}$. We have that*

$$\frac{1}{N}\sum_{\mathcal{E}_i\in\mathcal{S}_1} \max_{x\in\mathcal{E}_i}\langle \vec{u}, x\rangle^{1/2} \in (1\pm\varepsilon/2)\mathbb{E}_{P\sim\mathcal{P}(\mathcal{K})}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2}] \pm \varepsilon'\varepsilon(1-\varepsilon)M_{\vec{u}}/4$$

Now we are ready to prove the following lemma.

Lemma 93. *For any direction $\vec{u} \in \mathcal{P}^\star$, we have the following property.*

$$\frac{1}{N}\sum_{\mathcal{E}_i\in\mathcal{S}_1} \max_{x\in\mathcal{E}_i}\langle \vec{u}, x\rangle^{1/2} + \sum_{s_i\in\mathcal{P}(\overline{\mathcal{K}})} p_i\cdot\langle \vec{u}, s_i\rangle^{1/2} \in (1\pm 4\varepsilon)\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2}].$$

Proof. Let E be the event that no point is present in $\overline{\mathcal{K}}$. By the fact $\Pr[\overline{\mathcal{K}}] \leq \varepsilon$, we have that $\Pr[E] = \Pi_{s_i\in\mathcal{P}(\overline{\mathcal{K}})}(1-p_i) \geq 1-\sum_{s_i\in\mathcal{P}(\overline{\mathcal{K}})} p_i \geq 1-\varepsilon$. Thus, we conclude that $1-\varepsilon \leq \Pr[E] \leq 1$ We first rewrite $\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2}]$ as follows:

$$\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2}] = \Pr[E]\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2} \mid E] + \Pr[\overline{E}]\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2} \mid \overline{E}]$$
$$=\Pr[E]\cdot\mathbb{E}_{P\sim\mathcal{P}(\mathcal{K})}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2}] + \Pr[\overline{E}]\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2} \mid \overline{E}]$$

For event E, we bound the term $\Pr[E]\cdot\mathbb{E}_{P\sim\mathcal{P}(\mathcal{K})}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2}]$ via the collection $\mathcal{S}_1$. Let $M_{\vec{u}} = \max_{x\in\mathcal{P}(\mathcal{K})}\langle \vec{u}, x\rangle^{1/2}$. By Lemma 92, for any direction $\vec{u} \in \mathcal{P}^\star$, we have that

$$\frac{1}{N}\sum_{\mathcal{E}_i\in\mathcal{S}_1} \max_{x\in\mathcal{E}_i}\langle \vec{u}, x\rangle^{1/2} \in (1\pm\varepsilon/2)\mathbb{E}_{P\sim\mathcal{P}(\mathcal{K})}[\max_{x\in P}\langle \vec{u}, x\rangle^{1/2}] \pm \varepsilon'\varepsilon(1-\varepsilon)M_{\vec{u}}/4$$

By Lemma 90, we have that $(1-\varepsilon)\mathcal{K} \subseteq \mathcal{H}$. Then by the construction of $\mathcal{H}_{\vec{u}}$, we have that $\Pr[\mathcal{P} \cap (1-\varepsilon)\overline{\mathcal{K}}_{\vec{u}}] \geq \varepsilon'$. Thus, we obtain that

$$\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}] \geq \varepsilon'(1-\varepsilon)\max_{x\in\mathcal{P}(\mathcal{K})}\langle\vec{u},x\rangle^{1/2} = \varepsilon'(1-\varepsilon)M_{\vec{u}}.$$

So we conclude that

$$(1-2\varepsilon)\Pr[E]\cdot\mathbb{E}_{P\sim\mathcal{P}(\mathcal{K})}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}] - \varepsilon\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}] \leq \frac{1}{N}\sum_{\mathcal{E}_i\in\mathcal{S}_1}\max_{x\in\mathcal{E}_i}\langle\vec{u},x\rangle^{1/2}$$
$$\leq(1+2\varepsilon)\Pr[E]\cdot\mathbb{E}_{P\sim\mathcal{P}(\mathcal{K})}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}] + \varepsilon\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}], \tag{5.6}$$

since $1-\varepsilon \leq \Pr[E] \leq 1$.

For event $\overline{E}$, without loss of generality, we assume that the n points $s_1,\ldots,s_n$ in $\mathcal{P}$ are sorted in nondecreasing order according to the inner product $\langle\vec{u},s_i\rangle$. Assume that $s_{i_1},\ldots,s_{i_l}$ $(i_1 < i_2 < \ldots < i_l)$ are points in $\mathcal{P}(\overline{\mathcal{K}})$. Let E_j be the event that point s_{i_j} is present and all points s_{i_k} are not present for $k > j$. We have that

$$\Pr[\overline{E}]\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2} \mid \overline{E}] = \sum_{j\in[l]}\Pr[E_j]\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2} \mid E_j]$$
$$=\sum_{j\in[l]}p_{i_j}\cdot\Big(\prod_{j+1\leq k\leq l}(1-p_{i_k})\Big)\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2} \mid E_j].$$

By the above equality, on one hand, we have that

$$\Pr[\overline{E}]\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2} \mid \overline{E}] \geq (1-\varepsilon)\sum_{j\in[l]}p_{i_j}\cdot\langle\vec{u},s_{i_j}\rangle^{1/2}, \tag{5.7}$$

since $\max_{x\in P}\langle\vec{u},x\rangle^{1/2} \geq \langle\vec{u},s_{i_j}\rangle^{1/2}$ if event E_j happens. On the other hand, the

following inequality also holds.

$$\begin{aligned}
&\Pr[\overline{E}]\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}\mid\overline{E}]=\sum_{j\in[l]}\Pr[E_j]\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}\mid E_j]\\
\leq&\sum_{j\in[l]}\Pr[E_j]\cdot\mathbb{E}_{P\sim\mathcal{P}}[\langle\vec{u},s_{i_j}\rangle^{1/2}+\max_{x\in P\cap\mathcal{P}(\mathcal{K})}\langle\vec{u},x\rangle^{1/2}\mid E_j]\\
\leq&\sum_{j\in[l]}p_{i_j}\cdot\left(\mathbb{E}_{P\sim\mathcal{P}}[\langle\vec{u},s_{i_j}\rangle^{1/2}\mid E_j]+\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P\cap\mathcal{P}(\mathcal{K})}\langle\vec{u},x\rangle^{1/2}\mid E_j]\right)\\
\leq&\sum_{j\in[l]}p_{i_j}\cdot\langle\vec{u},s_{i_j}\rangle^{1/2}+\sum_{j\in[l]}p_{i_j}\cdot\mathbb{E}_{P\sim\mathcal{P}(\mathcal{K})}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}]\leq\sum_{j\in[l]}p_{i_j}\cdot\langle\vec{u},s_{i_j}\rangle^{1/2}+\varepsilon\cdot\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}].
\end{aligned}\tag{5.8}$$

The last inequality holds since that $\sum_{j\in[l]}p_{i_j}=\Pr[\mathcal{P}(\overline{\mathcal{K}})]\leq\varepsilon$ by Lemma 91. Combining Inequalities (5.6), (5.7) and (5.8), we prove the lemma. □

By Lemma 88, we construct a point set $\mathcal{S}_2$ to estimate $\sum_{s_i\in\mathcal{P}(\overline{\mathcal{K}})}p_i\cdot\mathrm{d}(s_i,F)$ with a weight function $w':\mathcal{S}_2\to\mathbb{R}$. We have that the size of $\mathcal{S}_2$ can be bounded by $O(j^4d\varepsilon^{-2})$. Then $\mathcal{S}=\mathcal{S}_1\cup\mathcal{S}_2$ is a collection of constant size, which satisfies the following property:

$$\frac{1}{N}\sum_{\mathcal{E}_i\in\mathcal{S}_1}\max_{x\in\mathcal{E}_i}\langle\vec{u},x\rangle^{1/2}+\sum_{s_i\in\mathcal{S}_2}w_i'\cdot\langle\vec{u},s_i\rangle^{1/2}\in(1+O(\varepsilon))\mathbb{E}_{P\sim\mathcal{P}}[\max_{x\in P}\langle\vec{u},x\rangle^{1/2}].\tag{5.9}$$

Here w_i' is the weight of s_i in $\mathcal{S}_2$. We can think $\mathcal{S}_2=\{\{s_i\}\mid 1\leq s_i\leq|\mathcal{S}_2|\}$ as a collection of singleton point sets $\{s_i\}$. Then by Inequality 5.9, we have that $\mathcal{S}$ is a generalized ε-coreset satisfying Definition 62. We conclude the following lemma.

Lemma 94. *Given an instance $\mathcal{P}$ of n stochastic points of the stochastic minimum j-flat-center problem in the existential model, if the total probability $\sum_i p_i\geq\varepsilon$, there exists an* SJFC-Coreset *$\mathcal{S}$ containing $O(\varepsilon^{-O(j^2d^3)}+j^4d\varepsilon^{-2})$ point sets of size at most $O(\varepsilon^{-O(j^2d^3)})$, together with a weight function $w':\mathcal{S}\to\mathbb{R}^+$, which satisfies that for any j-flat $F\in\mathcal{F}$,*

$$\sum_{S\in\mathcal{S}}w'(S)\cdot\mathrm{J}(S,F)\in(1\pm\varepsilon)\mathrm{J}(\mathcal{P},F).$$

Combining Lemma 88 and Lemma 94, we can obtain the following theorem.

Theorem 95. *Given an instance $\mathcal{P}$ of n stochastic points in the existential model, there is an* SJFC-Coreset *of size $O(\varepsilon^{-O(j^2d^3)} + j^4d\varepsilon^{-2})$ for the minimum j-flat-center problem. Moreover, we have an $O(n\log^{O(d)} n + \varepsilon^{-O(j^2d^3)}n)$ time algorithm to compute the* SJFC-Coreset*.*

Proof. We only need to prove the running time. Recall that the SJFC-Coreset $\mathcal{S}$ can be divided into two parts $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$. For the first part $\mathcal{S}_1$, we construct the convex hull $\mathcal{K}$ in $O(n\log^{O(d)} n)$ by Lemma 90. Then we construct $\mathcal{S}_1$ by taking $O(\varepsilon^{-O(j^2d^3)})$ independent realizations restricted to $\mathcal{P}(\mathcal{K})$. For each sample, we construct a deterministic ε-kernel in $O(n + \varepsilon^{-(d-3/2)})$ by [110]. So the total time for constructing $\mathcal{S}_1$ is $O(n\log^{O(d)} n + \varepsilon^{-O(j^2d^3)}n)$. On the other hand, we can construct $\mathcal{S}_2$ in $O(ndj^{O(j^2)})$ time by Lemma 88. Thus, we prove the theorem. □

PTAS for stochastic minimum j-flat-center. Given an SJFC-Coreset $\mathcal{S}$ together with a weight function $w' : \mathcal{S} \to \mathbb{R}^+$ by Theorem 95, it remains to show how to compute the optimal j-flat for $\mathcal{S}$. Our goal is to find the optimal j-flat F^* such that the total generalized distance $\sum_{S\in\mathcal{S}} w'(S)\cdot \mathrm{J}(S, F^*)$ is minimized. The argument is similar to the stochastic minimum k-center problem.

We first divide the family $\mathcal{F}$ of j-flats into a constant number of sub-families. In each sub-family $\mathcal{F}' \subseteq \mathcal{F}$, we have the following property: for each $S_i \in \mathcal{S}$, and each j-flat $F \in \mathcal{F}'$, the point $s^i = \arg\max_{s\in S_i} \mathrm{d}(s, F)$ is fixed. By Lemma 41, we have that $h_F(x) = \mathrm{d}(x, F)^2$ $(x \in \mathbb{R}^d)$ admits a linearization of dimension $O(j^2d^3)$. For each sub-family $\mathcal{F}'$, we can formulate the optimization problem as a polynomial system of constant degree, a constant number of variables, and a constant number of constraints. Then we can compute the optimal j-flat in constant time for all sub-families $\mathcal{F}' \subseteq \mathcal{F}$. Thus, we can compute the optimal j-flat-center for the SJFC-Coreset $\mathcal{S}$ in constant time. We then have the following corollary.

Corollary 96. *If the dimensionality d is a constant, given an instance of n stochastic points in $\mathbb{R}^d$ in the existential uncertainty model, there exists a PTAS for the stochastic minimum j-flat-center problem in $O(n\log^{O(d)} n + \varepsilon^{-O(j^2d^3)}n)$ time.*

Locational Uncertainty Model. Note that in the locational uncertainty model, we only need to consider Case 2. We use the same construction as in the existential model. Let $p_i = \sum_j p_{v_j s_i}$. Similarly, we make a linearization for the function $\mathrm{d}(x, F)^2$, where $x \in \mathbb{R}^d$ and $F \in \mathcal{F}$ is a j-flat. Using this linearization, we also map $\mathcal{P}$ into $O(j^2d^3)$-dimensional points. For the jth node and a set P of points, we denote $p_j(P) = \sum_{s_i \in P} p_{v_j s_i}$ to be the total probability that the jth node locates inside P.

By the condition $\Pr[\mathcal{P}(\overline{\mathcal{K}})] \leq \varepsilon$, we have that $\Pr[E] = 1 - \prod_{j\in[m]}(1 - p_j(\overline{\mathcal{K}})) \leq 1 - (1 - \varepsilon) = \varepsilon$, where event E represents that there exists a point present in $\overline{\mathcal{K}}$. So we can regard those points outside $\mathcal{K}$ independent. On the other hand, for any direction $\vec{u}$, since $\Pr[\mathcal{P} \cap (1-\varepsilon)\overline{\mathcal{H}}_{\vec{u}}] \geq \varepsilon'$, we have that $\Pr[E_{\vec{u}}] = 1 - \prod_{j\in[m]}(1 - p_j(\mathcal{P} \cap (1-\varepsilon)\overline{H}_{\vec{u}})) \geq 1 - (1 - \frac{\varepsilon'}{m})^m \geq \varepsilon'/2$, where event $E_{\vec{u}}$ represents that there exists a point present in $\mathcal{P} \cap (1-\varepsilon)\overline{H}_{\vec{u}}$. Moreover, we can use the same method to construct a collection $\mathcal{S}_1$ as an estimation for the point set $\mathcal{P}(\mathcal{K})$ in the locational uncertainty model. So Lemma 93 still holds. Then by Lemma 94, we can construct an SJFC-Coreset of constant size.

Theorem 97. *Given an instance $\mathcal{P}$ of n stochastic points in the locational uncertainty model, there is an* SJFC-Coreset *of cardinality $O(\varepsilon^{-O(j^2d^3)} + j^4d\varepsilon^{-2})$ for the minimum j-flat-center problem. Moreover, we have a polynomial time algorithm to compute the gerneralized ε-coreset.*

By a similar argument as in the existential model, we can give a PTAS for the locational uncertainty model. Then combining with Corollary 96, we prove the main result Theorem 13.

5.5 Constructing additive ε-coresets

In this section, we first give the algorithm for constructing an additive ε-coreset. We construct Cartesian grids and maintain one point from each nonempty grid cell, which is similar to [11]. However, our algorithm is more complicated. See Algorithm 2 for details.

Now we analyze the algorithm.

Algorithm 2 Constructing additive ε-coresets ($\mathbb{A}$)

1 Input: a realization $P \sim \mathcal{P}$. W.l.o.g., assume that $P = \{s_1, \ldots, s_m\}$.
2 Let $r_P = \min_{F:F\subseteq\mathcal{P},|F|=k} \mathrm{K}(P, F)$. If $r_P = 0$, output $\mathcal{E}(P) = P$. Otherwise assume that $2^a \leq r_P < 2^{a+1}$ ($a \in \mathcal{Z}$).
3 Draw a d-dimensional Cartesian grid $G_1(P)$ of side length $\varepsilon 2^a/4d$ centered at point 0^d.
4 Let $\mathcal{C}_1(P) = \{C \mid C \in G, C \cap P \neq \emptyset\}$ be the collection of those cells which intersects P.
5 For each cell $C \in \mathcal{C}_1(P)$, let $s^C \in C \cap P$ be the point in C of smallest index. Let $\mathcal{E}_1(P) = \{s^C \mid C \in \mathcal{C}_1(P)\}$.
6 Compute $r_{\mathcal{E}_1(P)} = \min_{F:F\subseteq\mathcal{P},|F|=k} \mathrm{K}(\mathcal{E}_1(P), F)$. If $r_{\mathcal{E}_1(P)} \geq 2^a$, let $\mathcal{E}(P) = \mathcal{E}_1(P)$, $G(P) = G_1(P)$, and $\mathcal{C}(P) = \mathcal{C}_1(P)$.
7 If $r_{\mathcal{E}_1(P)} < 2^a$, draw a d-dimensional Cartesian grid $G_2(P)$ of side length $\varepsilon 2^a/8d$ centered at point 0^d. Repeat Step 4 and 5, construct $\mathcal{C}_2(P)$ and $\mathcal{E}_2(P)$ based on the new Cartesian grid $G_2(P)$. Let $\mathcal{E}(P) = \mathcal{E}_2(P)$, $G(P) = G_2(P)$, and $\mathcal{C}(P) = \mathcal{C}_2(P)$.
8 Output $\mathcal{E}(P)$, $G(P)$, and $\mathcal{C}(P)$.

Lemma 98. *r_P is a 2-approximation for the minimum k-center problem w.r.t. P.*

Proof. By Gonzalez's greedy algorithm [51], there exists a subset $F \subseteq P \subseteq \mathcal{P}$ of size k such that the k-center value $\mathrm{K}(P, F)$ is a 2-approximation for the minimum k-center problem w.r.t. P. Thus, we prove the lemma. □

By the above lemma, we have the following lemma.

Lemma 68. The running time of $\mathbb{A}$ on any n point set P is $O(kn^{k+1})$. Moreover, the output $\mathcal{E}(P)$ is an additive ε-coreset of P of size at most $O(k/\varepsilon^d)$.

Proof. Since r_P is a 2-approximation, $\mathcal{E}(P)$ is an additive ε-coreset of P of size $O(k/\varepsilon^d)$ by Theorem 2.4 in [11]. For the running time, consider computing r_P in Step 2 (also $r_{\mathcal{E}_1(P)}$ in Step 6). There are at most n^k point sets $F \subseteq \mathcal{P}$ such that $|F| = k$. Note that computing $\mathrm{K}(P, F)$ costs at most nk time. Thus, it costs $O(kn^{k+1})$ time to compute r_P (also $r_{\mathcal{E}_1(P)}$) for all k-point sets $F \subseteq \mathcal{P}$. On the other hand, it only costs linear time to construct the Cartesian grid $G(P)$, the cell collection $\mathcal{C}(P)$ and $\mathcal{E}(P)$ after computing r_P and $r_{\mathcal{E}_1(P)}$, which finishes the proof. □

We then give the following lemmas, which is useful for proving Lemma 70.

Lemma 99. *For two point sets P, P', if $P' \subseteq P$, then $r_{P'} \leq r_P$. Moreover, if P' is an additive ε-coreset of P, then $(1-\varepsilon)r_P \leq r_{P'} \leq r_P$.*

Proof. Suppose $F \subseteq \mathcal{P}$ is the k-point set such that the k-center value $\mathrm{K}(P, F) = r_P$. Since $P' \subseteq P$, we have $\mathrm{K}(P', F) \leq r_P$. Thus, we have $r_{P'} \leq \mathrm{K}(P', F) \leq r_P$.

Moreover, assume that P' is an additive ε-coreset of P. Suppose $F' \subseteq \mathcal{P}$ is the k-point set such that the k-center value $\mathrm{K}(P', F') = r_{P'}$. Then by Definition 67, we have $\mathrm{K}(P, F') \leq (1+\varepsilon)r_{P'}$. Thus, we have $(1-\varepsilon)r_P \leq (1-\varepsilon)\mathrm{K}(P, F') < r_{P'} \leq r_P$. □

Lemma 100. *Assume that a point set $P' = \mathcal{E}(P)$ for another point set $P \sim \mathcal{P}'$. Running $\mathbb{A}(P')$ and $\mathbb{A}(P)$, assume that we obtain two Cartesian grids $G(P')$ and $G(P)$ respectively. Then we have $G(P') = G(P)$.*

Proof. If $r_P = 0$, we have that $r_{P'} \leq r_P = 0$ by Lemma 99. Thus we do not construct the Cartesian grid for both P and P'. Otherwise assume that $2^a \leq r_P < 2^{a+1}$ ($a \in \mathcal{Z}$). Run $\mathbb{A}(P)$. In Step 5, we construct a Cartesian grid $G_1(P)$ of side length $\varepsilon 2^a/4d$, a cell collection $\mathcal{C}_1(P)$, and a point set $\mathcal{E}_1(P)$. Since $\mathcal{E}_1(P)$ is an additive ε-coreset of P by [11], we have $2^{a-1} < (1-\varepsilon)r_P \leq r_{\mathcal{E}_1(P)} \leq r_S < 2^{a+1}$. Then we consider the following two cases.

Case 1: $r_{\mathcal{E}_1(P)} \geq 2^a$. Then $P' = \mathcal{E}(P) = \mathcal{E}_1(P)$, and $G(P) = G_1(P)$ in this case. Running $\mathbb{A}(P')$, we have that $2^a \leq r_{\mathcal{E}_1(P)} = r_{P'} \leq r_P < 2^{a+1}$ by Lemma 99. Thus, we construct a Cartesian grid $G_1(P') = G_1(P)$ of side length $\varepsilon 2^a/4d$, and a point set $\mathcal{E}_1(P')$ in Step 5. Since $G_1(P') = G_1(P)$ and $P' = \mathcal{E}_1(P)$, we have that $\mathcal{E}_1(P') = P'$ by the construction of $\mathcal{E}_1(P')$. Thus, $r_{\mathcal{E}_1(P')} = r_{\mathcal{E}_1(P)} \geq 2^a$, and we obtain that $G(P') = G_1(P')$ in Step 6, which proves the lemma.

Case 2: $2^{a-1} \leq r_{\mathcal{E}_1(P)} < 2^a$. Then in Step 7, we construct a Cartesian grid $G_2(P)$ of side length $\varepsilon 2^a/8d$ for P, a cell collection $\mathcal{C}_2(P)$, and a point set $\mathcal{E}_2(P)$. In this case, we have that $\mathcal{E}(P) = \mathcal{E}_2(P)$, $G(P) = G_2(P)$, and $\mathcal{C}(P) = \mathcal{C}_2(P)$. Now run $\mathbb{A}(P')$, and obtain $\mathcal{E}(P')$, $G(P')$, and $\mathcal{C}(P')$. By Lemma 99, we have

$$2^{a+1} > r_P \geq r_{P'} = r_{\mathcal{E}(P)} \geq (1-\varepsilon)r_P > 2^{a-1}.$$

We need to consider two cases. If $2^{a-1} \leq r_{P'} < 2^a$, we construct a Cartesian grid

$G_1(P')$ of side length $\varepsilon 2^a/8d$, and a point set $\mathcal{E}_1(P')$ in Step 5. Since $G_1(P') = G_2(P)$ and $P' = \mathcal{E}_2(P)$, we have that $\mathcal{E}_1(P') = P'$ by the construction of $\mathcal{E}_1(P')$. Then we let $G(P') = G_1(P')$ in Step 6. In this case, both $G(P)$ and $G(P')$ are of side length $\varepsilon 2^a/8d$, which proves the lemma.

Otherwise if $2^a \leq r_{P'} < 2^{a+1}$, we construct the Cartesian grid $G_1(P') = G_1(P)$ of side length $\varepsilon 2^a/4d$, a cell collection $\mathcal{C}_1(P')$, and a point set $\mathcal{E}_1(P')$ in Step 5. We then prove that $\mathcal{E}_1(P') = \mathcal{E}_1(P)$. Since all Cartesian grids are centered at point 0^d, a cell in $G_1(P)$ can be partitioned into 2^d equal cells in $G_2(P)$. Rewrite a cell $C^* \in G_1(P)$ as $C^* = \cup_{1 \leq i \leq 2^d} C_i$ where each $C_i \in G_2(P)$. Assume that point $s^* \in C^* \cap P = \cup_{1 \leq i \leq 2^d}(C_i \cap P)$ has the smallest index, then point s^* is also the point in $C^* \cap \mathcal{E}_2(P)$ of smallest index. Since $\mathcal{E}(P) = \mathcal{E}_2(P)$, we have that s^* is the point in $C^* \cap \mathcal{E}(P)$ of smallest index. Considering $\mathcal{E}_1(P')$, note that for each cell $C^* \in \mathcal{C}_1(P')$, $\mathcal{E}_1(P')$ only contains the point in $C^* \cap P'$ of smallest index. Since $P' = \mathcal{E}(P)$, we have that $\mathcal{E}_1(P') = \mathcal{E}_1(P)$. Thus, we conclude that $r_{\mathcal{E}_1(P')} = r_{\mathcal{E}_1(P)} < 2^a$. Then in Step 7, we construct a Cartesian grid $G_2(P') = G_2(P)$ of side length $\varepsilon 2^a/8d$ for P'. Finally, we output $G(P') = G_2(P') = G(P)$, which proves the lemma. □

Recall that we denote $\mathcal{E}(\mathcal{P}) = \{\mathcal{E}(P) \mid P \sim \mathcal{P}\}$ to be the collection of all possible additive ε-coresets. For any S, we denote $\mathcal{E}^{-1}(S) = \{P \sim \mathcal{P} \mid \mathcal{E}(P) = S\}$ to be the collection of all realizations mapped to S. Now we are ready to prove Lemma 70.

Lemma 70. (restated) Consider a subset S of at most $O(k/\varepsilon^d)$ points. Run algorithm $\mathbb{A}(S)$, which outputs an additive ε-coreset $\mathcal{E}(S)$, a Cartesian grid $G(S)$, and a collection $\mathcal{C}(S)$ of nonempty cells. If $\mathcal{E}(S) \neq S$, then $S \notin \mathcal{E}(\mathcal{P})$ (i.e., S is not the output of $\mathbb{A}$ for any realization $P \sim \mathcal{P}$). If $|S| \leq k$, then $\mathcal{E}^{-1}(S) = \{S\}$. Otherwise if $\mathcal{E}(S) = S$ and $|S| \geq k+1$, then a point set $P \sim \mathcal{P}$ satisfies $\mathcal{E}(P) = S$ if and only if

P1. For any cell $C \notin \mathcal{C}(S)$, $C \cap P = \emptyset$.

P2. For any cell $C \in \mathcal{C}(S)$, assume that point $s^C = C \cap S$. Then $s^C \in P$, and any point $s' \in C \cap \mathcal{P}$ with a smaller index than that of s^C does not appear in the realization P.

Proof. If $\mathcal{E}(S) \neq S$, we have that $r_S > 0$. Assume that $S \in \mathcal{E}(\mathcal{P})$. There must exist some point set $P \sim \mathcal{P}$ such that $\mathcal{E}(P) = S$. By Lemma 100, running $\mathbb{A}(P)$ and $\mathbb{A}(S)$, we obtain the same Cartesian grid $G(P) = G(S)$. Since $\mathcal{E}(S) \neq S$, there must exist a cell $C \in \mathcal{C}(S)$ such that $|C \cap S| \geq 2$ (by the construction of $\mathcal{E}(S)$). Note that $C \in G(P)$. We have $|C \cap \mathcal{E}(P)| = 1$, which is a contradiction with $\mathcal{E}(P) = S$. Thus, we conclude that $S \notin \mathcal{E}(\mathcal{P})$.

If $|S| \leq k$, assume that there exists another point set $P \neq S$, such that $\mathcal{E}(P) = S$. By Lemma 68, we know that S is an additive ε-coreset of P. By Definition 67, we have $S \subseteq P$ and $\mathrm{K}(P, S) \leq (1 + \varepsilon)\mathrm{K}(S, S) = 0$. Thus we conclude that $P = S$. On the other hand, we have $\mathcal{E}(S) = S$ since $r_S = 0$. So we conclude that $\mathcal{E}^{-1}(S) = \{S\}$.

If $|S| \geq k+1$ and $\mathcal{E}(S) = S$, we have that $r_S > 0$. Running $\mathbb{A}(P)$ and $\mathbb{A}(S)$, assume that we obtain two Cartesian grids $G(P)$ and $G(S)$ respectively. By Lemma 100, if $\mathcal{E}(P) = S$, then we have $G(P) = G(S)$. Moreover, by the construction of $\mathcal{E}(P)$, P1 and P2 must be satisfied.

We then prove the 'only if' direction. If P1 and P2 are satisfied, we have that S is an additive ε-coreset of P satisfying Definition 67 by [11]. Then by Lemma 99, we have that $(1 - \varepsilon)r_P \leq r_S \leq r_P$. Assume that $2^a \leq r_S < 2^{a+1}$ ($a \in \mathcal{Z}$), we conclude $2^a \leq r_P < 2^{a+2}$. Now run $\mathbb{A}(S)$. In Step 5, assume that we construct a Cartesian grid $G_1(S)$ of side length $\varepsilon 2^a/4d$, a cell collection $\mathcal{C}_1(S)$, and a point set $\mathcal{E}_1(S)$. Since $\mathcal{E}_1(S)$ is an additive ε-coreset of S by [11], we have $2^{a-1} < (1 - \varepsilon)r_S \leq r_{\mathcal{E}_1(S)} \leq r_S < 2^{a+1}$. Then we consider the following two cases.

Case 1: $2^a \leq r_{\mathcal{E}_1(S)} < 2^{a+1}$. In this case, we have that $G(S) = G_1(S)$, $\mathcal{C}(S) = \mathcal{C}_1(S)$, and $S = \mathcal{E}(S) = \mathcal{E}_1(S)$. Running $\mathbb{A}(P)$, assume that we obtain $G(P)$, $\mathcal{C}(P)$, and $\mathcal{E}(P)$. Consider the following two cases. If $2^a \leq r_P < 2^{a+1}$, we construct a Cartesian grid $G_1(P) = G(S)$ of side length $\varepsilon 2^a/4d$, and a point set $\mathcal{E}_1(P)$ in Step 5. Since P1 and P2 are satisfied, we know that $\mathcal{E}_1(P) = S$. Then since $2^a \leq r_{\mathcal{E}_1(P)} = r_S < 2^{a+1}$, we obtain that $\mathcal{E}(P) = \mathcal{E}_1(P) = S$ in this case. Otherwise if $2^{a+1} \leq r_S < 2^{a+2}$, run $\mathbb{A}(P)$. We construct a Cartesian grid $G_1(P)$ of side length $\varepsilon 2^a/2d$, and a point set $\mathcal{E}_1(P)$ in Step 5. Since P1 and P2 are satisfied, we have that $\mathcal{E}_1(P) \subseteq S$. Thus, we have $r_{\mathcal{E}_1(P)} \leq r_S < 2^{a+1}$ by Lemma 99. Then in Step 7, we construct a Cartesian

grid $G_2(P) = G_1(S)$ of side length $\varepsilon 2^a/4d$, and a point set $\mathcal{E}_2(P)$. In this case, we have that $G(P) = G_2(P) = G_1(S)$, and $\mathcal{E}(P) = \mathcal{E}_2(P)$. By P1 and P2, we have that $\mathcal{E}(P) = \mathcal{E}_2(P) = S$.

Case 2: $2^{a-1} \leq r_{\mathcal{E}_1(S)} < 2^a$. Running $\mathbb{A}(S)$, we construct a Cartesian grid $G_2(S)$ of side length $\varepsilon 2^a/8d$, and a point set $\mathcal{E}_2(S)$ in Step 7. In this case, we have that $G(S) = G_2(S)$, and $S = \mathcal{E}(S) = \mathcal{E}_2(S)$. Since $\mathcal{E}_1(S)$ is an additive ε-coreset of S, we conclude that $\mathcal{E}_1(S)$ is also an additive 3ε-coreset of P satisfying Definition 67. Then we have that $2^a \leq r_P \leq (1+3\varepsilon) r_{\mathcal{E}_1(S)} < 2^{a+1}$ by Lemma 99. Running $\mathbb{A}(P)$, we construct a Cartesian grid $G_1(P) = G_1(S)$ of side length $\varepsilon 2^a/4d$, and a point set $\mathcal{E}_1(P)$ in Step 5. Since P1 and P2 are satisfied, we know that $\mathcal{E}_1(P) = \mathcal{E}_1(S)$. Thus, we have $2^{a-1} \leq r_{\mathcal{E}_1(P)} = r_{\mathcal{E}_1(S)} < 2^a$. Then in Step 7, we construct a Cartesian grid $G_2(P) = G_2(S)$ of side length $\varepsilon 2^a/8d$, and a point set $\mathcal{E}_2(P)$. Again by P1 and P2, we have that $\mathcal{E}_2(P) = \mathcal{E}_2(S)$. Thus, we output $\mathcal{E}(P) = \mathcal{E}_2(P) = S$, which finishes the proof. □

Chapter 6 Estimating the Expected Value of Combinatorial Optimization Problems over Stochastic Data

In this chapter, we consider the stochastic geometry model where the location of each node is a random point in a given metric space, or the existence of each node is uncertain. We study the problems of computing the expected lengths of several combinatorial or geometric optimization problems over stochastic points, including closest pair, minimum spanning tree, k-clustering, minimum perfect matching, and minimum cycle cover. We also consider the problem of estimating the probability that the length of closest pair, or the diameter, is at most, or at least, a given threshold. Most of the above problems are known to be #P-hard. We obtain FPRAS (Fully Polynomial Randomized Approximation Scheme) for most of them in both the existential and locational uncertainty models. Our result for stochastic minimum spanning trees in the locational uncertain model improves upon the previously known constant factor approximation algorithm. Our results for other problems are the first known to the best of our knowledge.

6.1 The Closest Pair Problem

6.1.1 Estimating $\Pr[\mathsf{C} \leq 1]$

As a warmup, we first demonstrate how to use the stoch-core technique for the closest pair problem in the existential uncertainty model. Given a set of points $\mathcal{P} = \{s_1, \ldots, s_n\}$ in the metric space, where each point $s_i \in \mathcal{P}$ is present with probability p_i. We use C to denote the distance between the closest pair of vertices in the realized graph. If the realized graph has less than two points, C is zero. The goal is

to compute the probability $\Pr[\mathsf{C} \leq 1]$.

For a set H of points and a subset $S \subseteq H$, we use $H\langle S\rangle$ to denote the event that among all points in H, all and only points in S are present. For any nonnegative integer i, let $H\langle i\rangle$ denote the event $\bigvee_{S\subseteq H:|S|=i} H\langle S\rangle$, i.e., the event that exactly i points are present in H.

The *stoch-core* of the closest pair problem is simply defined to be

$$\mathbb{H} = \left\{ s_i \mid p_i \geq \frac{\varepsilon}{n^2} \right\}.$$

Let $\mathbb{F} = \mathcal{P} \setminus \mathbb{H}$. We consider the decomposition

$$\Pr[\mathsf{C} \leq 1] = \sum_{i=0}^{|\mathbb{F}|} \Pr[\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1] = \sum_{i=0}^{|\mathbb{F}|} \Pr[\mathbb{F}\langle i\rangle] \cdot \Pr[\mathsf{C} \leq 1 \mid \mathbb{F}\langle i\rangle].$$

Our algorithm is very simple: estimate the first three terms (i.e., $i = 0, 1, 2$) and use their sum as our final answer.

We can see that $\mathbb{H}$ satisfies the two properties of a stoch-core mentioned in the introduction:

1. The probability that all nodes are realized in $\mathbb{H}$, i.e., $\Pr[\mathbb{F}\langle 0\rangle]$, is at least $1 - n \cdot \frac{\varepsilon}{n^2} = 1 - \frac{\varepsilon}{n}$;

2. If there exist two points $s_i, s_j \in \mathbb{H}$ such that $\mathrm{d}(s_i, s_j) \leq 1$, we have $\Pr[\,\mathsf{C} \leq 1 \mid \mathbb{F}\langle 0\rangle\,] \geq \frac{\varepsilon^2}{n^4}$; otherwise, $\Pr[\mathsf{C} \leq 1 \mid \mathbb{F}\langle 0\rangle] = \Pr[\mathbb{H}\langle 0\rangle \mid \mathbb{F}\langle 0\rangle] + \Pr[\mathbb{H}\langle 1\rangle \mid \mathbb{F}\langle 0\rangle]$. Note that we can compute $\Pr[\mathbb{H}\langle 0\rangle \mid \mathbb{F}\langle 0\rangle]$ and $\Pr[\mathbb{H}\langle 1\rangle \mid \mathbb{F}\langle 0\rangle]$ in polynomial time. We do not consider this case in the following analysis.

Both properties guarantee that the random variable $I(\mathsf{C} \leq 1)$, conditioned on $\mathbb{F}\langle 0\rangle$, is poly-bounded [1], hence we can easily get a $(1 \pm \varepsilon)$-estimation for $\Pr[\mathbb{F}\langle 0\rangle \wedge \mathsf{C} \leq 1]$ with polynomial many samples with high probability. Similarly, $\Pr[\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1]$ can also be estimated with polynomial number of samples for $i = 1, 2$. The algorithm can be found in Algorithm 3.

[1] $I()$ is the indicator function. Note that $\mathbb{E}[I(\mathsf{C} \leq 1)] = \Pr[\mathsf{C} \leq 1]$.

Algorithm 3 Estimating $\Pr[\mathsf{C} \leq 1]$

1 Estimate $\Pr[\mathbb{F}\langle 0\rangle \wedge \mathsf{C} \leq 1]$: Take $N_0 = O\big((n/\varepsilon)^4 \ln n\big)$ independent samples. Suppose M_0 is the number of samples satisfying $\mathsf{C} \leq 1$ and $\mathbb{F}\langle 0\rangle$. $T_0 \leftarrow \frac{M_0}{N_0}$.
2 Estimate $\Pr[\mathbb{F}\langle 1\rangle \wedge \mathsf{C} \leq 1]$: For each point $s_i \in \mathbb{F}$, take $N_1 = O((n/\varepsilon)^4 \ln n)$ independent samples conditioning on the event $\mathbb{F}\langle\{s_i\}\rangle$. Suppose there are M_i samples satisfying $\mathsf{C} \leq 1$. $T_1 \leftarrow \sum_{s_i \in \mathbb{F}} p_i M_i / N_1$.
3 Estimate $\Pr[\mathbb{F}\langle 2\rangle \wedge \mathsf{C} \leq 1]$: For each point pair $s_i, s_j \in \mathbb{F}$, take $N_2 = O((n/\varepsilon)^4 \ln n)$ independent samples conditioning on the event $\mathbb{F}\langle\{s_i, s_j\}\rangle$. Suppose there are M_{ij} samples satisfying $\mathsf{C} \leq 1$. $T_2 \leftarrow \sum_{s_i, s_j \in \mathbb{F}} p_i p_j M_{ij} / N_2$.
4 Output: $T_0 + T_1 + T_2$

Lemma 101. *Steps 1,2,3 in Algorithm 3 provide* $(1 \pm \varepsilon)$*-approximations for* $\Pr[\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1]$ *for* $i = 0, 1, 2$ *respectively, with high probability.*

Theorem 102. *There is an FPRAS for estimating the probability of the distance between the closest pair of nodes is at most* 1 *in the existential uncertainty model.*

Proof. We only need to show that the contribution from the rest of terms (where more than two points outside stoch-core $\mathbb{H}$ are present) is negligible compared to the third term. Suppose S is the set of all present points such that $\mathsf{C} \leq 1$ and there are at least 3 points not in $\mathbb{H}$. Suppose s_i, s_j are the closest pair in S. We associate S with a smaller set $S' \subset S$ by making 1 present point in $(S \cap \mathbb{F}) \setminus \{s_i, s_j\}$ absent (if there are several such S', we choose an arbitrary one). We denote it as $S \sim S'$. We use the notation $S \in F_i$ to denote that the realization S satisfies $(\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1)$. Then, we can see that for $i \geq 3$,

$$\Pr[\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1] = \sum_{S : S \in F_i} \Pr[S] \leq \sum_{S' : S' \in F_{i-1}} \sum_{S : S \sim S'} \Pr[S].$$

For a fixed S', there are at most m different sets S such that $S \sim S'$ and $\Pr[S] \leq \frac{2\varepsilon}{n^2}\Pr[S']$ for any such S. Hence, we have that

$$\sum_{S : S \sim S'} \Pr[S] \leq \frac{2\varepsilon}{n}\Pr[S'].$$

Therefore,

$$\Pr[\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1] \leq \frac{2\varepsilon}{n} \cdot \sum_{S':S'\in F_{i-1}} \Pr[S'] = \frac{2\varepsilon}{n} \cdot \Pr[\mathbb{F}\langle i-1\rangle \wedge \mathsf{C} \leq 1].$$

Hence, overall we have $\sum_{i\geq 3} \Pr[\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1] \leq \varepsilon \Pr[\mathbb{F}\langle 2\rangle \wedge \mathsf{C} \leq 1]$. This finishes the analysis. □

Note that the number of samples is dominated by estimating $\Pr[\mathbb{F}\langle 2\rangle \wedge \mathsf{C} \leq 1]$. Since there are $O\big(n^2\big)$ different pairs $s_i, s_j \in \mathbb{F}$. We take N_2 independent samples for each pair. Overall, we take $O\big(\frac{n^6}{\varepsilon^4} \ln n\big)$ independent samples.

Locational Uncertainty Model. The algorithm for the locational uncertainty model is similar to the one for the existential uncertainty model. Here we briefly sketch the algorithm. For ease of exposition, we assume that for each point, there is only one node that may be realized at this point. In principle, if more than one node may be realized at the same point, we can create multiple copies of the point co-located at the same place.

For any node $v \in \mathcal{V}$ and point $s \in \mathcal{P}$, we use the notation $v \vDash s$ to denote the event that node v is realized at point s. Let $p_{vs} = \Pr[v \vDash s]$, i.e., the probability that node v is realized at point s. For each point $s \in \mathcal{P}$, we let $p(s)$ denote the probability that point s is present ($p(s) = p_{vs}$, v is the unique node which may be realized at s). Let $H\langle i\rangle$ denote the event that exactly i nodes are realized to the point set H.

We construct the stoch-core $\mathbb{H} = \{s \mid p(s) \geq \frac{\varepsilon}{(nm)^2}\}$. Let $\mathbb{F} = \mathcal{P} \setminus \mathbb{H}$. Then we rewrite $\Pr[\mathsf{C} \leq 1] = \sum_{0\leq i\leq n} \Pr[\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1]$. We only need to estimate the first three terms.

Estimating $\Pr[\mathbb{F}\langle 0\rangle \wedge \mathsf{C} \leq 1]$.

1. If there exist two points $s, s' \in \mathbb{H}$ with $\mathrm{d}(s, s') \leq 1$ which correspond to different nodes, then $\Pr[\mathbb{F}\langle 0\rangle \wedge \mathsf{C} \leq 1] \geq p(s)p(s') \geq \frac{\varepsilon^2}{(nm)^4}$ by the definition of stoch-core , we can simply estimate $\Pr[\mathbb{F}\langle 0\rangle \wedge \mathsf{C} \leq 1]$ by taking $O(\frac{(nm)^4}{\varepsilon^4} \ln n)$ independent samples using the Monte Carlo method.

2. If no such two points $s, s' \in \mathbb{H}$ exist, $\Pr[\mathbb{F}\langle 0\rangle \wedge \mathsf{C} \leq 1] = 0$.

Estimating $\Pr[\mathbb{F}\langle 1\rangle \wedge \mathsf{C} \leq 1]$**.** We first rewrite this term by $\sum_{v\in\mathcal{V},s\in\mathbb{F}} \Pr[\mathbb{F}\langle 1\rangle \wedge \mathsf{C} \leq 1 \wedge v \vDash s]$. For a node $v \in \mathcal{V}$ and point $s \in \mathbb{F}$, we denote $\mathsf{B}_s = \{s' \in \mathbb{H} : \mathrm{d}(s, s') \leq 1\}$. If B_s contains any point corresponding to a node other than v, we can use Monte Carlo for estimating $\Pr[\mathbb{F}\langle 1\rangle \wedge \mathsf{C} \leq 1 \mid v \vDash s]$ since it is at least $\frac{\varepsilon}{(nm)^2}$. Otherwise, computing $\Pr[\mathbb{F}\langle 1\rangle \wedge \mathsf{C} \leq 1 \mid v \vDash s]$ is equivalent to computing $\Pr[\mathbb{F}\langle 0\rangle \wedge \mathsf{C} \leq 1]$ in the instance without v (since v is at distance more than 1 from any other nodes).

Estimating $\Pr[\mathbb{F}\langle 2\rangle \wedge \mathsf{C} \leq 1]$**.** We rewrite it as $\sum_{v,v'\in\mathcal{V},s,s'\in\mathbb{F}} \Pr[\mathbb{F}\langle 2\rangle \wedge \mathsf{C} \leq 1 \wedge v \vDash s \wedge v' \vDash s']$. We estimate each term in the same way as the former case. We do not repeat the argument here.

Analysis. Similar to the existential uncertainty model, we can show that the contribution of $\sum_{3\leq i\leq n} \Pr[\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1]$ is negligible. The argument is almost the same as before. Suppose S is a realization such that $\mathsf{C} \leq 1$ and there are at least 3 points not in $\mathbb{H}$. Suppose v_i, v_j are the closest pair in S. We associate S with S', where S' is obtained by sending node v in S (except v_i, v_j) located in $\mathbb{F}$ to a point $s \in \mathbb{H}$ such that $p_{vs} \geq \frac{1}{2n}$. We denote it as $S \sim S'$. Then for a fixed S', there are at most nm different sets S such that $S \sim S'$ and $\Pr[S] \leq \frac{2\varepsilon}{m}\Pr[S']$ for any such S. The rest arguments are the same.

Theorem 103. *There is an FPRAS for estimating the probability of the distance between the closest pair of nodes is at most* 1 *in the locational uncertainty model.*

The number of samples is dominated by estimating $\Pr[\mathbb{F}\langle 2\rangle \wedge \mathsf{C} \leq 1]$. Since there are $O(m^2)$ different pairs of nodes $v, v' \in \mathcal{V}$ and $O(n^2)$ different pairs of points $s, s' \in \mathbb{F}$, we separate $\mathbb{F}\langle 2\rangle$ into $O(m^2n^2)$ different terms. For each term, we take $O\big(\frac{(nm)^4}{\varepsilon^4}\ln n\big)$ independent samples. Thus, we take $O\big(\frac{m^6n^6}{\varepsilon^4}\ln n\big)$ independent samples in total.

6.1.2 Estimating $\mathbb{E}[\mathsf{C}]$

In this section, we consider the problem of estimating $\mathbb{E}[\mathsf{C}]$, where C is the distance of the closest pair of present points, in the existential uncertainty model. Now, we introduce our second main technique, the *hierarchical partition family (HPF)*

technique, to solve this problem. An HPF is a family Ψ of partitions of $\mathcal{P}$, formally defined as follows.

Definition 104. *(Hierarchical Partition Family (HPF)) Let T be any minimum s-panning tree spanning all points of $\mathcal{P}$. Suppose that the edges of T are $e_1, \ldots, e_{m-1}$ with $\mathrm{d}(e_1) \geq \mathrm{d}(e_2) \geq \ldots \geq \mathrm{d}(e_{m-1})$. Let $E_i = \{e_i, e_{i+1}, \ldots, e_{m-1}\}$. The HPF $\Psi(\mathcal{P})$ consists of m partitions $\Gamma_1, \ldots, \Gamma_n$. Γ_1 is the entire point set $\mathcal{P}$. Γ_i consists of i disjoint subsets of $\mathcal{P}$, each corresponding to a connected component of $G_i = G(\mathcal{P}, E_i)$. Γ_n consists of all singleton points in $\mathcal{P}$. It is easy to see that Γ_j is a refinement of Γ_i for $j > i$. Consider two consecutive partitions Γ_i and Γ_{i+1}. Note that G_i contains exactly one more edge (i.e., e_i) than G_{i+1}. Let μ'_{i+1} and μ''_{i+1} be the two components (called the* split components*) in Γ_{i+1}, each containing an endpoint of e_i. Let $\nu_i \in \Gamma_i$ be the connected component of G_i that contains e_i. We call ν_i the* special component *in Γ_i. Let $\Gamma'_i = \Gamma_i \setminus \nu_i$.*

We observe two properties of $\Psi(\mathcal{P})$ that are useful later.

P1. Consider a component $C \in \Gamma_i$. Let s_1, s_2 be two arbitrary points in C. Then $\mathrm{d}(s_1, s_2) \leq (n-1)\mathrm{d}(e_i)$ (this is because s_1 and s_2 are connected in G_i, and e_i is the longest edge in G_i).

P2. Consider two different components C_1 and C_2 in Γ_i. Let $s_1 \in C_1$ and $s_2 \in C_2$ be two arbitrary points. Then $\mathrm{d}(s_1, s_2) \geq \mathrm{d}(e_{i-1})$ (this is because the minimum inter-component distance is $\mathrm{d}(e_{i-1})$ in G_i).

Let the random variable Y be smallest integer i such that there is at most one present point in each component of Γ_{i+1}. Note that if $Y = i$ then each component of Γ_i contains at most one point, except that the special component ν_i contains exactly two present points. The following lemma is a simple consequence of P1 and P2.

Lemma 105. *Conditioning on $Y = i$, it holds that $\mathrm{d}(e_i) \leq \mathsf{C} \leq n\mathrm{d}(e_i)$ (hence, C is poly-bounded).*

Consider the following expansion of $\mathbb{E}[\mathsf{C}]$:

$$\mathbb{E}[\mathsf{C}] = \sum_{i=1}^{m-1} \Pr[Y = i]\mathbb{E}[\mathsf{C} \mid Y = i].$$

For a fixed i, $\Pr[Y = i]$ can be estimated as follows: For a component $C \subset \mathcal{P}$, we use $C\langle j\rangle$ to denote the event that exactly j points in C are present, $C\langle s\rangle$ the event that only s is present in C and $C\langle \leq j\rangle$ the event that no more than j points in C are present. Let μ'_i and μ''_i be the two split components in Γ_i. Note that

$$\Pr[Y = i] = \Pr[\mu'_{i+1}\langle 1\rangle] \cdot \Pr[\mu''_{i+1}\langle 1\rangle] \cdot \prod_{C \in \Gamma'_i} \Pr[C\langle \leq 1\rangle].$$

Each term can be easily computed in polynomial time. The remaining is to show how to estimate $\mathbb{E}[\mathsf{C} \mid Y = i]$. Since C is poly-bounded, it suffices to give an efficient algorithm to take samples conditioning on $Y = i$. This is again not difficult: We take exactly one point $s \in \mu'_{i+1}$ with probability $\Pr[\mu'_{i+1}\langle s\rangle]/\Pr[\mu'_{i+1}\langle 1\rangle]$. Same for μ''_{i+1}. For each $C \in \Gamma'_i$, take no point from C with probability $\Pr[C\langle 0\rangle]/\Pr[C\langle \leq 1\rangle]$; otherwise, take exactly one point $s \in C$ with probability $\Pr[C\langle s\rangle]/\Pr[C\langle \leq 1\rangle]$.

By Lemma 105, conditioning on $Y = i$, taking $O(\frac{n}{\varepsilon^2} \ln n)$ independent samples are enough using the Monte Carlo method. Since there are m levels, we take $O\big(\frac{n^2}{\varepsilon^2} \ln n\big)$ independent samples in total. This finishes the description of the FPRAS in the existential uncertainty model.

Locational Uncertainty Model. Our algorithm is almost the same as the existential model. We first construct the HPF $\Psi(\mathcal{P})$. The random variable Y is defined in the same way. The only difference is how to estimate $\Pr[Y = i]$ and how to take samples efficiently conditioning on $Y = i$. First consider estimating $\Pr[Y = i]$. We can consider the problem as the following bins-and-balls problem: we have m balls (corresponding to nodes) and i bins (corresponding to components in Γ_i). Each ball v is thrown to bin C with probability $p_{vC} = \sum_{s \in C} p_{vs}$ (note that $\sum_C p_{vC} = 1$). We want to compute the probability that each of the first and second bins (corresponding to the two split components) contains exactly one ball, and for other bins each

contains at most one ball. Consider the following $i \times i$ ($i \geq m$) matrix M with $M_{vC} = \begin{cases} p_{vC} = \sum_{s \in C} p_{vs}, & \text{for } v \in [m] \text{ and } C \in [i]; \\ 1, & \text{otherwise} \end{cases}$. It is not difficult to see that the permanent

$$\mathrm{Per}(M) = \sum_{\sigma \in \mathbb{S}_i} \prod_{v} M_{v\sigma(v)}$$

is exactly the probability that each bin contains at most one ball. To enforce each of the first two bins contains exactly one ball, simply consider the Laplace expansion of $\mathrm{Per}(M)$, expanded along the first two columns, and retain those relevant terms:

$$\Pr[Y = i] = \sum_{k \in [n]} \sum_{j \in [n], j \neq k} M_{k1} M_{j2} \mathrm{Per}(M^{\star}_{kj})$$

where $M^{\star}_{kj}$ is M with the 1st and 2nd columns and kth and jth rows removed. Then, we can use the celebrated result for approximating permanent by Jerrum, Sinclair, and Vigoda [**?**] to get an FPRAS for approximating $\Pr[Y = i]$. In fact, the algorithm in [**?**] provides a fully polynomial time approximate sampler for perfect matchings [2]. This can be easily translated to an efficient sampler conditioning on $Y = i$ [3]. Finally, we remark that the above algorithm can be easily modified to handle the case with both existential and locational uncertainty model.

Theorem 106. *There is an FPRAS for estimating the expected distance between the closest pair of nodes in both existential and locational uncertainty models.*

kth Closest Pair. In addition, we consider the problem of the expected distance $\mathbb{E}[\mathsf{kC}]$ between the kth closest pair under the existential uncertainty model. We use the HPF technique, and construct an efficient sampler via a dynamic programming. The details can be found in Section 6.8.1.

[2] The approximate sampler can return in poly-time a permutation $\sigma \in \mathbb{S}_i$ with probability $(1 \pm \varepsilon) \prod_s M_{s\sigma(s)} / \mathrm{Per}(M)$.

[3] We can also use the generic reduction by Jerrum, Valiant and Vazirani [68] which can turn an FPRAS into a poly-time approximate sampler for self-reducible relations.

6.2 k-Clustering

In this section, we study the k-clustering problem in the existential uncertainty model. According to [74], the optimal objective value for k-clustering is the $(k-1)$th most expensive edge of the minimum spanning tree. We consider estimating $\mathbb{E}[\mathsf{kCL}]$ under the existential uncertainty model.

Denote the point set $\mathcal{P} = \{s_1, \ldots, s_n\}$, where each point $s_i \in \mathcal{P}$ is present with probability p_i. We construct the HPF $\Psi(\mathcal{P})$. Let the random variable Y be the largest integer i such that at most $k-1$ components in Γ_i contain at least one present point. Let $\Gamma'_i = \Gamma_i \setminus \nu_i$. Note that if $Y = i$ then at most $k-2$ components in Γ'_i contain present points while the special component ν_i contains at least two present points, since both component μ'_{i+1} and μ''_{i+1} contain at least one present point. By the property P1 and P2 of HPF, we have the following lemma.

Lemma 107. *Conditioning on $Y = i$, it holds that $\mathrm{d}(e_i) \le \mathsf{kCL} \le n\mathrm{d}(e_i)$ (hence, kCL is poly-bounded)..*

Proof. Since Γ_{i+1} contains at least k nonempty components, any spanning tree must have at least $k-1$ inter-component edges. Any inter-component edge is of length at least $\mathrm{d}(e_i)$, so is the $(k-1)$th expensive edge. Now we show the other direction. Assume w.l.o.g. that all pairwise distances are distinct. Consider a realization satisfying $Y = i$ and the graphical matroid which consists of all forests of the realization. Suppose $\mathsf{kCL} = \mathrm{d}(e)$ for some edge e. Let E_e be all edges with length no larger than e in this realization. We can see that $\mathsf{rank}(E_e) = n - k + 1$ where rank is the matroid rank function and n the number of present points in the realization. Hence, any spanning tree contains no more than $n-k+1$ edges from E_e. Equivalently, the $(k-1)$th most expensive edge of any spanning tree is no smaller than kCL. Moreover, since Γ_i has no more than $k-1$ nonempty components, there exists a spanning tree such that the $(k-1)$th most expensive edge is an intra-component edge in Γ_i. The lemma follows from P1.

□

Consider the following expansion $\mathbb{E}[\mathsf{kCL}] = \sum_{i=1}^{m-1} \Pr[Y = i]\mathbb{E}[\mathsf{kCL} \mid Y = i]$. Recall that for a component $C \subset \mathcal{P}$, we use $C\langle j\rangle$ to denote the event that exactly j points in C are present, $C\langle s\rangle$ the event that only s is present in C and $C\langle \leq j\rangle$ ($C\langle \geq j\rangle$) the event that at most (at least) than j points in C are present. For a partition Γ on $\mathcal{P}$, we use $\Gamma\langle j, \geq 1\rangle$ to denote the event that exactly j components in Γ contain at least one present point. Note that

$$\Pr[Y = i] = \Pr[\mu'_{i+1}\langle \geq 1\rangle] \cdot \Pr[\mu''_{i+1}\langle \geq 1\rangle] \cdot \Pr[\Gamma'_i\langle k-2, \geq 1\rangle].$$

Note that $\Pr[\mu'_{i+1}\langle \geq 1\rangle]$ and $\Pr[\mu''_{i+1}\langle \geq 1\rangle]$ can be easily computed in polynomial time. The remaining task is to show how to compute $\Pr[\Gamma'_i\langle k-2, \geq 1\rangle]$ and how to estimate $\mathbb{E}[\mathsf{kCL} \mid Y = i]$. We first present a simple lemma which is useful later.

Lemma 108. *For a component C and $j \in \mathbb{Z}$, we can compute $\Pr[C\langle j\rangle]$ (or $\Pr[C\langle \geq j\rangle]$) in polynomial time. Moreover, there exists a poly-time sampler to sample present points from C conditioning on $C\langle j\rangle$ (or $C\langle \geq j\rangle$).*

Proof. The idea is essentially from [38]. W.l.o.g, we assume that the points in C are $s_1, \ldots, s_n$. We denote the event that among the first a points, exactly b points are present by $E[a, b]$ and denote the probability of $E[a, b]$ by $\Pr[a, b]$. Note that our goal is to compute $\Pr[n, j]$, which can be solved by the following dynamic program:

1. If $a < b$, $\Pr[a, b] = 0$. If $a = b$, $\Pr[a, b] = \prod_{1 \leq l \leq a} p_l$. If $b = 0$, $\Pr[a, b] = \prod_{1 \leq l \leq a}(1 - p_l)$.
2. For $a > b$ and $b \geq 1$, $\Pr[a, b] = p_a \Pr[a-1, b-1] + (1 - p_a)\Pr[a-1, b]$.

We can also use this dynamic program to construct an efficient sampler. Consider the point s_n. With probability $p_n \Pr[n-1, j-1]/\Pr[n, j]$, we make it present and then recursively consider the point s_{n-1} conditioning on the event $E[n-1, j-1]$. With probability $(1 - p_n)\Pr[n-1, j]/\Pr[n, j]$, we discard it and then recursively sample conditioning on the event $E[n-1, j]$. $\Pr[C\langle \geq j\rangle]$ can be handled in the same way and we omit the details. □

Computing $\Pr[\Gamma'_i\langle k-2, \geq 1\rangle]$**.** Now, it is ready to show how to compute $\Pr[\Gamma'_i\langle k-2, \geq 1\rangle]$ in polynomial time. Note that for each component $C_j \in \Gamma'_i$, we can easily compute $q_j = \Pr[C_j\langle \geq 1\rangle]$ in polynomial time. Since all components in Γ'_i are disjoint, using Lemma 108 (consider each component C_j in Γ'_i as a point with existential probability q_j), we can compute $\Pr[\Gamma'_i\langle k-2, \geq 1\rangle]$.

To take samples conditioning on $Y = i$, we first sample $k-2$ components in Γ'_i which contain present points. Then for these $k-2$ components and μ'_{i+1}, μ''_{i+1}, we independently sample present points in each component using Lemma 108. By Lemma 107, for estimating $\mathbb{E}[\mathsf{kCL} \mid Y = i]$, we need to take $O\big(\frac{n}{\varepsilon^2}\ln n\big)$ independent samples. So we take $O\big(\frac{n^2}{\varepsilon^2}\ln n\big)$ independent samples in total.

Theorem 109. *There is an FPRAS for estimating the expected length of k-th expensive edge in the minimum spanning tree in the existential uncertainty model.*

6.3 Minimum Spanning Trees

We consider the problem of estimating the expected size of minimum spanning tree in the locational uncertainty model. In this section, we briefly sketch how to solve it using our stoch-core method. Recall that the term nodes refers to the vertices $\mathcal{V}$ of the spanning tree and points describes the locations in $\mathcal{P}$. For ease of exposition, we assume that for each point, there is only one node that may realize at this point.

Recall that we use the notation $v \vDash s$ to denote the event that node v is present at point s. Let $p_{vs} = \Pr[v \vDash s]$. Since node v is realized with certainty, we have $\sum_{s\in\mathcal{P}} p_{vs} = 1$. For each point $s \in \mathcal{P}$, we let $p(s)$ denote the probability that point s is present. For a set H of points, let $p(H) = \sum_{s\in H} p(s)$, i.e., the expected number of points present in H. For a set H of points and a set S of nodes, we use $H\langle S\rangle$ to denote the event that all and only nodes in S are realized to some points in H. If S only contains one node, say v, we use the notation $H\langle v\rangle$ as the shorthand for $H\langle\{v\}\rangle$. Let $H\langle i\rangle$ denote the event $\bigvee_{S:|S|=i} H\langle S\rangle$, i.e., the event that exactly i nodes are in H. We use $\mathrm{diam}(H)$, called the diameter of H, to denote $\max_{s,t\in H} \mathrm{d}(s,t)$. Let $\mathrm{d}(p,H)$ be the closest distance between point p and any point in H.

Finding stoch-core Firstly, we find in poly-time the stoch-core $\mathbb{H}$ as follows:

Algorithm 4 Constructing stoch-core $\mathbb{H}$ for Estimating $\mathbb{E}[MST]$

1 Among all points r with $p(r) \geq \frac{\varepsilon}{16n}$, find the furthest two points s and t.
2 Set $\mathbb{H} \leftarrow \mathsf{B}(s, \mathrm{d}(s,t)) = \{s' \in \mathcal{P} \mid \mathrm{d}(s', s) \leq \mathrm{d}(s,t)\}$.

Lemma 110. *Algorithm 4 finds a stoch-core $\mathbb{H}$ such that*

Q1. $p(\mathbb{H}) \geq m - \frac{\varepsilon}{16} = m - O(\varepsilon)$

Q2. $\mathbb{E}[\,\mathsf{MST} \mid \mathbb{H}\langle m \rangle\,] = \Omega\Big(\mathrm{diam}(\mathbb{H})\frac{\varepsilon^2}{n^2}\Big)$.

Furthermore, the algorithm runs in linear time.

Proof. For each point r that is not in $\mathbb{H}$, we know $p(r) < \frac{\varepsilon}{16n}$. Therefore, we have that and $p(\mathcal{P} \setminus \mathbb{H}) < \frac{\varepsilon}{16}$. and $p(\mathbb{H}) \geq m - \frac{\varepsilon}{16}$. Consider two cases:

1. Points s and t relate to different nodes. In this case, we have that

$$\mathbb{E}[\mathsf{MST} \mid \mathbb{H}\langle m \rangle] \geq \mathrm{d}(s,t)\Pr[\exists (v,u), v \neq u, v \vDash s, u \vDash t] = \mathrm{d}(s,t)p(s)p(t) \geq \mathrm{d}(s,t)\frac{\varepsilon^2}{256n^2}.$$

2. Points s and t relate to the same node v. In this case, conditioning on the event that a different node u is realized to an arbitrary point q, $\mathbb{E}[\mathsf{MST} \mid \mathbb{H}\langle m \rangle] \geq \mathrm{d}(s,q)\Pr[v \vDash s] + \mathrm{d}(t,q)\Pr[v \vDash t] \geq \mathrm{d}(s,t)\frac{\varepsilon}{16n}$.

In either case, $\mathbb{H}$ satisfies both Q1 and Q2.

□

Estimating $\mathbb{E}[\mathsf{MST}]$ Let $\mathbb{F} = \mathcal{P} \setminus \mathbb{H}$. We rewrite $\mathbb{E}[\mathsf{MST}]$ by $\sum_{i \geq 0} \mathbb{E}[\mathsf{MST} \mid \mathbb{F}\langle i \rangle] \cdot \Pr[\mathbb{F}\langle i \rangle]$. We only need to estimate $\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle 0 \rangle\,] \cdot \Pr[\mathbb{F}\langle 0 \rangle]$ and $\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle 1 \rangle\,] \cdot \Pr[\mathbb{F}\langle 1 \rangle]$.

Lemma 111. *Algorithm 5 produces a $(1 \pm \varepsilon)$-estimate for the first term with high probability.*

Algorithm 5 Estimating $\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle 0\rangle\,] \cdot \Pr[\mathbb{F}\langle 0\rangle]$

1 Take $N_0 = O\big(\frac{mn^2}{\varepsilon^4} \ln m\big)$ random samples. Set $A \leftarrow \emptyset$ at the beginning.
2 For each sample G_i, if it satisfies $\mathbb{F}\langle 0\rangle$, $A \leftarrow A \cup \{G_i\}$.
3 $T_0 \leftarrow \frac{1}{N_0} \sum_{G_i \in A} \mathsf{MST}(G_i)$.

Proof. Based on the event $\mathbb{F}\langle 0\rangle$, the length of MST is at most $m \cdot \mathrm{diam}(\mathbb{H})$. Due to (Q2), we have a poly-bounded random variable and can therefore obtain a $(1 \pm \varepsilon)$-estimate for $\mathbb{E}[\,\mathsf{MST} \mid \mathbb{H}\langle m\rangle\,]$ using the Monte Carlo method with $O\big(\frac{mn^2}{\varepsilon^4} \ln m\big)$ samples satisfying $\mathbb{H}\langle m\rangle$ (by Lemma 14). By the first property of $\mathbb{H}$, with probability close to 1, a sample satisfies $\mathbb{H}\langle m\rangle$. So, the expected time to obtain an useful sample is bounded by a constant. Overall, we can obtain a $(1 \pm \varepsilon)$-estimate of the first term with using $N_0 = O\big(\frac{mn^2}{\varepsilon^4} \ln m\big)$ samples with high probability. □

Algorithm 6 Estimating $\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle 1\rangle\,] \cdot \Pr[\mathbb{F}\langle 1\rangle]$

1 Set $B \leftarrow \{s \mid s \in \mathbb{F}, \mathrm{d}(s, \mathbb{H}) < \frac{m}{\varepsilon} \cdot \mathrm{diam}(\mathbb{H})\}$. Let $\mathsf{CI}(v)$ be the event that v is the only node that is realized to some point $s \in B$.
2 Conditioning on $\mathsf{CI}(v)$, take $N_1 = O\big(\frac{mn^2}{\varepsilon^5} \ln m\big)$ independent samples.
Let $A_v \leftarrow \{G_{v,i} \mid 1 \le i \le N_1\}$ be the set of N_1 samples for $\mathsf{CI}(v)$.
3 $T_v \leftarrow \frac{1}{N_1} \sum_{G_{v,i} \in A_v} \mathsf{MST}(G_{v,i})$ (estimating $\mathbb{E}[\,\mathsf{MST} \mid \mathsf{CI}(v)]$)
4 $T_1 \leftarrow \sum_{v \in \mathcal{V}} \Big(\Pr[\mathsf{CI}(v)] T_v + \sum_{s \in \mathbb{F} \setminus B} \Pr[\mathbb{F}\langle v\rangle \wedge v \vDash s] \, \mathrm{d}(s, \mathbb{H}) \Big)$.

Lemma 112. *Algorithm 6 produces a $(1 \pm \varepsilon)$-estimate for the second term with high probability.*

Analysis Note that the number of samples is asymptotically dominated by estimating $\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle 1\rangle\,] \cdot \Pr[\mathbb{F}\langle 1\rangle]$. For each node $v \in \mathcal{V}$, we take N_1 independent samples. Thus, we need to take $O\big(\frac{m^2n^2}{\varepsilon^5} \ln m\big)$ independent samples. Now, we analyze the performance guarantee of our algorithm. We need to show that the total contribution from the scenarios where more than one node are not in the stoch-core is very small. We need some notations first. Suppose S is the set of nodes realized out of stoch-core $\mathbb{H}$. We use $\mathbb{F}_S$ to denote the set of all possible realizations of all nodes in S to points in $\mathbb{F}$ (we can think of each element in $\mathbb{F}_S$ as an $|S|$-dimensional vector where each

coordinate is indexed by a node in S and its value is a point in $\mathbb{F}$). Similarly, we denote the set of realizations of $\bar{S} = V \setminus S$ to points in $\mathbb{H}$ by $\mathbb{H}_{\bar{S}}$. For any $F_S \in \mathbb{F}_S$ and $H_{\bar{S}} \in \mathbb{H}_{\bar{S}}$, we use $(F_S, H_{\bar{S}})$ to denote the event that both F_S and $H_{\bar{S}}$ happen and $\mathsf{MST}(F_S, H_{\bar{S}})$ to denote the length of the minimum spanning tree under the realization $(F_S, H_{\bar{S}})$. We need the following combinatorial fact.

Lemma 113. *Consider a particular realization $(F_S, H_{\bar{S}})$, where S is the set of nodes realized out of $\mathbb{H}$. $|S| \geq 2$. Let $d = \mathrm{d}(v_S, u_S) = \min_{v \in S, u \in \bar{S}}\{\mathrm{d}(u, v)\}$ where $v_S \in F_S$, $u_S \in H_{\bar{S}}$. The realization $(F_{S'}, H_{\bar{S}'})$ is obtained from $(F_S, H_{\bar{S}})$ by sending the node v_S to $\mathbb{H}$, where $S' = S \setminus v_S$. Then $\mathsf{MST}(F_S, H_{\bar{S}}) \leq 4\mathsf{MST}(F_{S'}, H_{\bar{S}'})$.*

Proof. We have

$$4\mathsf{MST}(F_{S'}, H_{\bar{S}'}) \geq 2\mathsf{MST}(F_{S'}, H_{\bar{S}'}) + 2d \geq \mathsf{MST}(F_{S'}, H_{\bar{S}}) + 2d \geq \mathsf{MST}(F_S, H_{\bar{S}})$$

The second inequality holds since the length of the minimum spanning tree is at most two times the length of the minimum Steiner tree (We consider $\mathsf{MST}(F_{S'}, H_{\bar{S}})$ as a Steiner tree connecting all nodes in $F_{S'} \cup H_{\bar{S}}$).

□

The only remaining part for establishing Theorem 115 is to show the following essential lemma.

Lemma 114. *For any $\varepsilon > 0$, if $\mathbb{H}$ satisfies the properties in Lemma 110, we have that*

$$\sum_{i>1} \mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle i\rangle] \cdot \Pr[\mathbb{F}\langle i\rangle] \leq \varepsilon \cdot \mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle 1\rangle] \cdot \Pr[\mathbb{F}\langle 1\rangle].$$

Proof. We claim that for any $i > 1$, $\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle i+1\rangle] \cdot \Pr[\mathbb{F}\langle i+1\rangle] \leq \frac{\varepsilon}{2}\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle i\rangle] \cdot \Pr[\mathbb{F}\langle i\rangle]$. If the claim is true, then we can show the lemma easily by noticing that, for any $m \geq 2$, $\sum_{i>1} \mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle i\rangle]\Pr[\mathbb{F}\langle i\rangle] \leq \sum_{i=1}^{m-1} \left(\frac{\varepsilon}{2}\right)^i \mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle 1\rangle]\Pr[\mathbb{F}\langle 1\rangle] \leq \varepsilon\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle 1\rangle]\Pr[\mathbb{F}\langle 1\rangle]$. Now, we prove the claim. First, we rewrite the LHS as follows:

$$\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle i+1\rangle] \cdot \Pr[\mathbb{F}\langle i+1\rangle] = \sum_{|S|=i+1} \sum_{F_S \in \mathbb{F}_S} \sum_{H_{\bar{S}} \in \mathbb{H}_{\bar{S}}} \big(\Pr[(F_S, H_{\bar{S}})] \cdot \mathsf{MST}(F_S, H_{\bar{S}})\big),$$

Similarly, the RHS can be written as:

$$\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle i\rangle] \cdot \Pr[\mathbb{F}\langle i\rangle] = \sum_{|S'|=i} \sum_{F_{S'}\in\mathbb{F}_{S'}} \sum_{H_{\bar{S}'}\in\mathbb{H}_{\bar{S}'}} \big(\Pr[(F_{S'}, H_{\bar{S}'})] \cdot \mathsf{MST}(F_{S'}, H_{\bar{S}'})\big).$$

For each pair $(F_S, H_{\bar{S}})$, let $C(F_S, H_{\bar{S}}) = \Pr[F_S, H_{\bar{S}}] \cdot \mathsf{MST}(F_S, H_{\bar{S}})$. Consider each pair $(F_S, H_{\bar{S}})$ with $|S| = i+1$ as a seller and each pair $(F_{S'}, H_{\bar{S}'})$ with $|S'| = i$ as a buyer. The seller $(F_S, H_{\bar{S}})$ wants to sell the term $C(F_S, H_{\bar{S}})$ and the buyers want to buy all this term. The buyer $(F_{S'}, H_{\bar{S}'})$ has a budget of $C(F_{S'}, H_{\bar{S}'})$. We show that there is a charging scheme such that each term $C(F_S, H_{\bar{S}})$ is fully paid by the buyers and each buyer spends at most an $\frac{\varepsilon}{2}$ fraction of her budget. Note that the existence of such a charging scheme suffices to prove the claim.

Suppose we are selling the term $C(F_S, H_{\bar{S}})$. Consider the following charging scheme. Suppose $v \in S$ is the node closest to any node in $\bar{S}$. Let $S' = S \setminus \{v\}$ and $F_{S'}$ be the restriction of F_S to all coordinates in S except v. We say $(F_{S'}, H_{\bar{S}'})$ is consistent with $(F_S, H_{\bar{S}})$, denoted as $(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})$, if $H_{\bar{S}'}$ agrees with $H_{\bar{S}}$ for all vertices in $\bar{S}$. and $F_{S'}$ agrees with F_S for all vertices in $S \setminus \{v\}$. Intuitively, $(F_{S'}, H_{\bar{S}'})$ can be obtained from $(F_S, H_{\bar{S}})$ by sending v to an arbitrary point in $\mathbb{H}$. Let

$$Z(F_S, H_{\bar{S}}) = \sum_{(F_{S'}, H_{\bar{S}'})\sim(F_S, H_{\bar{S}})} \Pr[(F_{S'}, H_{\bar{S}'})].$$

We need the following inequality later: For any fixed $(F_{S'}, H_{\bar{S}'})$,

$$\sum_{(F_S, H_{\bar{S}})\sim(F_{S'}, H_{\bar{S}'})} \frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})} \le \sum_{v\in\bar{S}'} \frac{\Pr(v\in\mathbb{F})}{\Pr(v\in\mathbb{H})} \le \frac{\varepsilon}{8}.$$

To see the inequality, for a fixed node v, consider the quantity

$$\sum_{(F_S, H_{\bar{S}})\sim(F_{S'}, H_{\bar{S}'}), \bar{S}=\bar{S}'\setminus\{v\}} \frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})}.$$

A crucial observation here is that the denominators of all terms are in fact the same, by the definition of Z, which is $\sum \Pr[(F'_{S'}, H'_{\bar{S}'})]$, and the summation is over all

$(F'_{S'}, H'_{\bar{S}'})$s which are the same as $(F_{S'}, H_{\bar{S}'})$ except that the location of v is a different point in $\mathbb{H}$. The numerator is the summation over all $(F_S, H_{\bar{S}})$s which are the same as $(F_{S'}, H_{\bar{S}'})$ except that the location of v is a different point in $\mathbb{F}$. Canceling out the same multiplicative terms from the numerators and the denominator, we can see it is at most $\frac{\Pr(v\in\mathbb{F})}{\Pr(v\in\mathbb{H})}$.

Now, we specify how to charge each buyer. For each buyer $(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})$, we charge her the following amount of money

$$\frac{\Pr[(F_{S'}, H_{\bar{S}'})] \cdot C(F_S, H_{\bar{S}})}{Z(F_S, H_{\bar{S}})}$$

We can see that $C(F_S, H_{\bar{S}})$ is fully paid by all buyers consistent with $(F_S, H_{\bar{S}})$. It remains to show that each buyer $(F_{S'}, H_{\bar{S}'})$ has been charged at most $\frac{\varepsilon}{2}C(F_{S'}, H_{\bar{S}'})$. By the above charging scheme, the terms $(F_S, H_{\bar{S}})$s in LHS that charge buyer $(F_{S'}, H_{\bar{S}'})$ are consistent with $(F_{S'}, H_{\bar{S}'})$. Now, we can see that the total amount of money charged to buyer $(F_{S'}, H_{\bar{S}'})$ can be bounded as follows:

$$\begin{aligned}
\sum_{(F_S,H_{\bar{S}})\sim(F_{S'},H_{\bar{S}'})} \frac{\Pr[F_{S'}, H_{\bar{S}'}] \cdot C(F_S, H_{\bar{S}})}{Z(F_S, H_{\bar{S}})} &\le 4\mathsf{MST}(F_{S'}, H_{\bar{S}'}) \cdot \sum_{(F_S,H_{\bar{S}})\sim(F_{S'},H_{\bar{S}'})} \frac{\Pr[F_{S'}, H_{\bar{S}'}] \cdot \Pr[(F_S, H_{\bar{S}})]}{Z(F_S, H_{\bar{S}})} \\
&= 4\mathsf{MST}(F_{S'}, H_{\bar{S}'})\Pr[F_{S'}, H_{\bar{S}'}] \cdot \sum_{(F_S,H_{\bar{S}})\sim(F_{S'},H_{\bar{S}'})} \frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})} \\
&\le \frac{\varepsilon}{2}\mathsf{MST}(F_{S'}, H_{\bar{S}'})\Pr[F_{S'}, H_{\bar{S}'}]
\end{aligned}$$

The first inequality follows from Lemma 113. This completes the proof.

□

Theorem 115. *There is an FPRAS for estimating the expected length of the minimum spanning tree in the locational uncertainty model.*

Finally, we remark that the problem can be solved by a variety of methods. The stoch-core method presented in this section is not the simplest one, but may be still helpful for understanding a very similar but somewhat more technical application of the method to minimum perfect matching (see Section 6.4).

6.4 Minimum Perfect Matchings

In this section, we consider the minimum perfect matching (PM) problem. We use the stoch-core method. The same stoch-core construction for MST can not be directly used here since PM can be much smaller than MST. For example, suppose there are only two points. There are even number of nodes residing at each point. In this case, PM is 0. Now, if we change the location of one particular node to the other point, the value of PM increase dramatically while the value of MST stays the same. In some sense, PM is more sensitive to the location of nodes, hence requires new stoch-core construction. There are two major differences from the algorithm for MST. First, the stoch-core is composed by several clusters of points, instead of a single ball. Second, we need a more careful charging argument.

Finding stoch-core. First, we show how to find in poly-time the stoch-core $\mathbb{H}$. Initially, $\mathbb{H}$ consists of all singleton points, each being a component by itself. Then, we gradually grow the ball from each point, and merge two components if they touch. We stop until certain properties Q1 and Q2 are satisfied. See the Pseudo-code in Algorithm 7 for details. For a node v and a set H of points, we let $p_v(H) = \sum_{s\in H} p_{vs}$. We use $\mathrm{diam}(H)$, called the diameter of H, to denote $\max_{s,s'\in H\cap\mathcal{P}} \mathrm{d}(s,s')$.

Algorithm 7 Constructing stoch-core $\mathbb{H}$ for Estimating $\mathbb{E}[\mathsf{PM}]$

1 Initially, $t \leftarrow 0$ and each point $s \in \mathcal{P}$ is a component $\mathbb{H}_{\{s\}} = \mathsf{B}(s,t)$ by itself.
2 Gradually increase t. If two different components $\mathbb{H}_{S_1}$ and $\mathbb{H}_{S_2}$ intersect (where $\mathbb{H}_S := \cup_{s\in S}\mathsf{B}(s,t)$), merge them into a new component $\mathbb{H}_{S_1\cup S_2}$.
3 Stop increasing t while the first time the following two conditions are satisfied by components at t. [4]

Q1. For each node v, there is a unique component $\mathbb{H}_j$ such that $p_v(\mathbb{H}_j) \geq 1 - O(\frac{\varepsilon}{mn^3})$. We call $\mathbb{H}_j$ the stoch-core of node v, denoted as $\mathbb{H}(v)$.

Q2. For all j, $|\{v \in \mathcal{V} \mid \mathbb{H}(v) = \mathbb{H}_j\}|$ is even.

4 Output the stopping time T and the components $\mathbb{H}_1, \ldots, \mathbb{H}_k$.

We need the following lemma which is useful for bounding $\mathbb{E}[\mathsf{PM}]$ from below.

[4] Note that we only need to consider those $t = \mathrm{d}(s,s')/2$ for some points $s, s' \in \mathcal{P}$. Thus, we compute on at most $O(n^2)$ different time ts.

Lemma 116. *For any two disjoint sets H_1 and H_2 of points, and any node v, we have*

$$\mathbb{E}[\mathsf{PM}] \geq \min\{p_v(H_1), p_v(H_2)\} \cdot \mathrm{d}(H_1, H_2)/n.$$

Here, $\mathrm{d}(H_1, H_2) = \min_{s \in H_1 \cap \mathcal{P}, s' \in H_2 \cap \mathcal{P}} \mathrm{d}(s, s')$.

Proof. Suppose $s = \arg\max_{\hat{s}}\{p_{v\hat{s}} \mid \hat{s} \in H_1\}$, and $s' = \arg\max_{\hat{s}}\{p_{v\hat{s}} \mid \hat{s} \in H_2\}$. Obviously, we have $p_{vs} \geq \frac{p_v(H_1)}{n}$ and $p_{vs'} \geq \frac{p_v(H_2)}{n}$. So it suffices to show $\mathbb{E}[\mathsf{PM}] \geq \min\{p_{vs}, p_{vs'}\} \cdot \mathrm{d}(s, t)$. We first see that

$$\begin{aligned}\mathbb{E}[\mathsf{PM}] &\geq p_{vs}\mathbb{E}[\mathsf{PM} \mid v \vDash s] + p_{vs'}\mathbb{E}[\mathsf{PM} \mid v \vDash s'] \\ &\geq \min\{p_{vs}, p_{vs'}\}\Big(\mathbb{E}[\mathsf{PM} \mid v \vDash s] + \mathbb{E}[\mathsf{PM} \mid v \vDash s']\Big).\end{aligned}$$

Then it is sufficient to prove that $\mathbb{E}[\mathsf{PM} \mid v \vDash s] + \mathbb{E}[\mathsf{PM} \mid v \vDash s'] \geq \mathrm{d}(s, s')$. Fix a realization of all nodes except v. Conditioning on this realization, we consider the following two minimum perfect matchings, one for the case $v \vDash s$, (denoted as PM_1) and the other one for $v \vDash s'$ (denoted as PM_2). Consider the symmetric difference

$$\mathsf{PM}_1 \oplus \mathsf{PM}_2 := (\mathsf{PM}_1 \setminus \mathsf{PM}_2) \cup (\mathsf{PM}_2 \setminus \mathsf{PM}_1).$$

We can see that it is a path $(s, p_1, p_2, \ldots, p_k, s')$, such that $(s, p_1) \in \mathsf{PM}_1$,$(p_1, p_2) \in \mathsf{PM}_2, \ldots, (p_k, s') \in \mathsf{PM}_2$. So $\mathsf{PM}_1 + \mathsf{PM}_2 \geq \mathrm{d}(s, s')$ by the triangle inequality. Therefore, we have $\mathbb{E}[\mathsf{PM} \mid v \vDash s] + \mathbb{E}[\mathsf{PM} \mid v \vDash s'] \geq \mathrm{d}(s, s') \geq \mathrm{d}(H_1, H_2)$. □

By Q1, Q2 and the above lemma, we can show that the following additional property holds.

Lemma 117. *Q3.* $\mathbb{E}[\mathsf{PM}] = \Omega(\frac{\varepsilon D}{mn^5})$ *where* $D = \max_i\{\mathrm{diam}(\mathbb{H}_i)\}$.

Proof. Note that the stopping time T must exist, because the set of all points satisfies the first two properties. Now, we show that Q3 also holds. Firstly, note that $D \leq 2mT$. Secondly, consider $T' = T - \varepsilon$ for some infinitesimal $\varepsilon > 0$. At time T', consider two situations:

1. There exists a node v, such that $\forall j, p_v(\mathbb{H}_j) < 1 - O(\frac{\varepsilon}{mn^3})$. Then there must exist two components C_1 and C_2 such that $p_v(C_1) > \Omega(\frac{\varepsilon}{mn^3})$ and $p_v(C_2) > \Omega(\frac{\varepsilon}{mn^3})$. Moreover, since C_1 and C_2 are two distinct components, $\mathrm{d}(C_1, C_2) \geq 2T'$. Then, by Lemma 116, we have $\mathbb{E}[\mathsf{PM}] \geq \Omega(\frac{\varepsilon}{mn^4}) \cdot 2T \geq \Omega(\frac{\varepsilon D}{mn^5})$.

2. Suppose that Q1 is true but Q2 is still false. Suppose $\mathbb{H}_j$ is a component which homes odd number of nodes. Note that with probability at least $(1 - \frac{1}{mn^3})^m \approx 1$, each node is realized to a point in its stoch-core. When this is the case, there is at least one node in $\mathbb{H}_j$ that needs to be matched with some node outside $\mathbb{H}_j$, which incurs a cost of at least $2T$.

□

Estimating $\mathbb{E}[\mathsf{PM}]$. We use $\mathbb{H}\langle m\rangle$ to denote the event that for each node v, $v \vDash \mathbb{H}(v)$. We denote the event that there are exactly i nodes which are realized out of their stoch-cores by $\mathbb{F}\langle i\rangle$. Again, we only need to estimate two terms: $\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle 0\rangle]] \cdot \Pr[\mathbb{F}\langle 0\rangle]$ and $\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle 1\rangle] \cdot \Pr[\mathbb{F}\langle 1\rangle]$. Using Properties Q1, Q2 and Q3, we can estimate these terms in polynomial time. Our final estimation is simply the sum of the first two terms.

Algorithm 8 Estimating $\mathbb{E}[\,\mathsf{PM} \mid \mathbb{F}\langle 0\rangle\,] \cdot \Pr[\mathbb{F}\langle 0\rangle]$

1 Take $N_1 = O(\frac{m^2 n^5}{\varepsilon^4} \ln m)$ independent samples. Set $A \leftarrow \emptyset$ at the beginning.
2 For each sample G_i, if it satisfies $\mathbb{H}\langle m\rangle$, $A \leftarrow A \cup \{G_i\}$.
$T_0 \leftarrow \frac{1}{N_1} \sum_{G_i \in A} \mathsf{PM}(G_i)$.

Lemma 118. *Algorithm 6.4 produces a $(1 \pm \varepsilon)$-estimate for the first term with high probability.*

Proof. Note that $\Pr[\mathbb{H}\langle m\rangle]$ is close to 1 (by union bound) and can be computed exactly. To estimate $\mathbb{E}[\mathsf{PM} \mid \mathbb{H}\langle m\rangle]]$, the algorithm takes the average of $N_1 = O(\frac{m^2 n^5}{\varepsilon^4} \ln m)$ samples. Note that conditioning on $\mathbb{H}\langle m\rangle$, the minimum perfect matching could be at most mD. We distinguish the following two cases.

1. $\mathbb{E}[\mathsf{PM} \mid \mathbb{H}\langle m\rangle] \geq \frac{\varepsilon}{2}\mathbb{E}[\mathsf{PM}] = \Omega(\frac{\varepsilon^2 D}{mn^5})$. We can get a $(1 \pm \varepsilon)$-approximation using the Monte Carlo method with $O(\frac{m^2n^5}{\varepsilon^4}\ln m)$ samples. Therefore PM is poly-bounded conditioning on $\mathbb{H}\langle n\rangle$.

2. $\mathbb{E}[\mathsf{PM} \mid \mathbb{H}\langle m\rangle] < \frac{\varepsilon}{2}\mathbb{E}[\mathsf{PM}]$. Then the probability that the sample average is larger than $\varepsilon\mathbb{E}[\mathsf{PM}]$ is at most $\text{poly}(\frac{1}{m})$ by Chernoff Bound. We can thus ignore this part safely.

□

Algorithm 9 Estimating $\mathbb{E}[\,\mathsf{PM} \mid \mathbb{F}\langle 1\rangle\,] \cdot \Pr[\mathbb{F}\langle 1\rangle]$

1 For each node v, set $B_v \leftarrow \{s \mid s \in \mathcal{P} \setminus \mathbb{H}(v), \mathrm{d}(s, \mathbb{H}(v)) < \frac{4mD}{\varepsilon}$. Let $\mathsf{CI}(v)$ be the event that v is the only node that is realized to some point $s \in B_v$.
2 Conditioning on $\mathsf{CI}(v)$, take $N_1 = O\big(\frac{m^2n^5}{\varepsilon^4}\ln m\big)$ independent samples. Let $A_v \leftarrow \{G_{v,i} \mid 1 \leq i \leq N_2\}$ be the set of N_1 samples for $\mathsf{CI}(v)$.
3 $T_v \leftarrow \frac{1}{N_1}\sum_{G_{v,i} \in A_v} \mathsf{PM}(G_{v,i})$ (estimating $\mathbb{E}[\,\mathsf{PM} \mid \mathsf{CI}(v)]$)
4 $T_1 \leftarrow \sum_{v\in\mathcal{V}}\Big(\Pr[\mathsf{CI}(v)]T_v + \sum_{s\in\mathbb{F}\setminus B_v}\Pr[\mathbb{F}\langle v\rangle \wedge v \vDash s]\,\mathrm{d}(s, \mathbb{H}(v))\Big)$.

Lemma 119. *Algorithm 6.4 produces a $(1\pm\varepsilon)$-estimate for the second term with high probability.*

Analysis Note that the number of samples is asymptotically dominated by estimating $\mathbb{E}[\,\mathsf{PM} \mid \mathbb{F}\langle 1\rangle\,]\cdot\Pr[\mathbb{F}\langle 1\rangle]$. For each node $v \in \mathcal{V}$, we take N_1 independent samples. Thus, we need to take $O\big(\frac{m^3n^5}{\varepsilon^4}\ln m\big)$ independent samples in total.

We still need to show that for $i > 1$, the contribution from event $\mathbb{F}\langle i\rangle$ is negligible. Suppose S is the set of nodes that are realized out of their stoch-cores. We use $\mathbb{F}_S$ and $\mathbb{H}_{\bar{S}}$ to denote the set of all realizations of the all nodes in S to points out of their stoch-cores, and the set of realizations of $\bar{S} = V \setminus S$ to points in their stoch-cores respectively. We use $\mathsf{PM}(F_S, H_{\bar{S}})$ to denote the length of the minimum perfect matching under the realization $(F_S, H_{\bar{S}})$, where $F_S \in \mathbb{F}_S$ and $H_{\bar{S}} \in \mathbb{H}_{\bar{S}}$. The following combinatorial fact plays the same role in the charging argument as Lemma 113 does in the previous section. Differing from the MST problem, we can not achieve a similar bound as the one in Lemma 113 since $\mathsf{PM}(F_S, H_{\bar{S}})$ may decrease significantly if we

send only one node outside its stoch-core back to its stoch-core. However, we show that in such case, if we send one more node back to its stoch-core, $\mathsf{PM}(F_S, H_{\bar{S}})$ can still be bounded.

We need the following structural result about minimum perfect matchings, which is essential for our charging argument.

Lemma 120. *Fix a realization $(F_S, H_{\bar{S}})$. We use $\ell(v)$ to denote $\mathrm{d}(v, \mathbb{H}(v))$ for all nodes $v \in S$. Suppose $v_1 \in S$ has the smallest ℓ value and v_2 has the second smallest ℓ value. Let $S' = S \backslash \{v_1\}$, $S'' = S' \backslash \{v_2\}$. Further let $(F_{S'}, H_{\bar{S}'})$ be a realization obtained from $(F_S, H_{\bar{S}})$ by sending v_1 to a point in its stoch-core $\mathbb{H}(v_1)$ and $(F_{S''}, H_{\bar{S}''})$ be a realization obtained from $(F_{S'}, H_{\bar{S}'})$ by sending v_2 to a point in its stoch-core $\mathbb{H}(v_2)$. Then we have that $\mathsf{PM}(F_S, H_{\bar{S}}) \leq 2(n+2)\mathsf{PM}(F_{S'}, H_{\bar{S}'}) + 2(n+2)\mathsf{PM}(F_{S''}, H_{\bar{S}''})$*

Proof. Let $\mathrm{d} = \min_v \ell(v)$ and $D = \max_i \operatorname{diam}(\mathbb{H}_i)$. Note that $d \geq \frac{D}{n}$ as $d \geq 2T$ and $D \leq 2nT$. We distinguish the following three cases:

1. $\mathsf{PM}(F_S, H_{\bar{S}}) \leq \frac{\mathrm{d}}{2}$. Using a similar argument to the one in Lemma 116, we have

$$\mathsf{PM}(F_{S'}, H_{\bar{S}'}) + \mathsf{PM}(F_S, H_{\bar{S}}) \geq \ell(v) = \mathrm{d}$$

   So, we have $\mathsf{PM}(F_S, H_{\bar{S}}) \leq \mathsf{PM}(F_{S'}, H_{\bar{S}'})$ in this case.

2. $\mathsf{PM}(F_S, H_{\bar{S}}) \geq (n+2)\mathrm{d}$. By the triangle inequality, we can see that

$$\mathsf{PM}(F_{S'}, H_{\bar{S}'}) + (n+1)d \geq \mathsf{PM}(F_{S'}, H_{\bar{S}'}) + d + D \geq \mathsf{PM}(F_S, H_{\bar{S}})$$

   So, we have $\mathsf{PM}(F_S, H_{\bar{S}}) \leq (n+2)\mathsf{PM}(F_{S'}, H_{\bar{S}'})$.

3. $\frac{\mathrm{d}}{2} \leq \mathsf{PM}(F_S, H_{\bar{S}}) \leq (n+2)\mathrm{d}$.

   (a) $\mathsf{PM}(F_{S'}, H_{\bar{S}'}) \geq \frac{\mathrm{d}}{2}$. We directly have $\mathsf{PM}(F_S, H_{\bar{S}}) \leq 2(n+2)\mathsf{PM}(F_{S'}, H_{\bar{S}'})$.

   (b) $\mathsf{PM}(F_{S'}, H_{\bar{S}'}) \leq \frac{\mathrm{d}}{2}$. By Lemma 116, we have

$$\mathsf{PM}(F_{S'}, H_{\bar{S}'}) + \mathsf{PM}(F_{S''}, H_{\bar{S}''}) \geq \mathrm{d}$$

Then we have $\mathsf{PM}(F_S, H_{\bar{S}}) \leq 2(n+2)\mathsf{PM}(F_{S''}, H_{\bar{S}''})$.

In summary, we prove the lemma. □

The remaining is to establish the following key lemma. The proof is similar to, but more involved than that of Lemma 114.

Lemma 121. *For any $\varepsilon > 0$, if $\mathbb{H}$ satisfies the properties Q1, Q2 in Algorithm 7, we have that*

$$\sum_{i>1} \mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle i\rangle] \cdot \Pr[\mathbb{F}\langle i\rangle] \leq \varepsilon \cdot \mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle 0\rangle] \cdot \Pr[\mathbb{F}\langle 0\rangle] + \varepsilon \cdot \mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle 1\rangle] \cdot \Pr[\mathbb{F}\langle 1\rangle].$$

Proof. We claim that for any $i > 1$,

$$\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle i+1\rangle]\cdot\Pr[\mathbb{F}\langle i+1\rangle] \leq \frac{\varepsilon}{6}\big(\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle i\rangle]\cdot\Pr[\mathbb{F}\langle i\rangle]+\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle i-1\rangle]\cdot\Pr[\mathbb{F}\langle i-1\rangle]\big)$$

If the claim is true, the lemma can be proven easily as follows. For ease of notation, we use $A(i)$ to denote $\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle i\rangle] \cdot \Pr[\mathbb{F}\langle i\rangle]$. First, we can see that

$$A(i+2) + A(i+1) \leq \frac{\varepsilon}{6}A(i+1) + \frac{2\varepsilon}{6}A(i) + \frac{\varepsilon}{6}A(i-1) \leq \frac{\varepsilon}{2}(A(i) + A(i-1)).$$

So if i is odd, $A(i+2) + A(i+1) \leq (\frac{\varepsilon}{2})^{(i+1)/2}(A(1) + A(0))$. Therefore, $\sum_{i>1} A(i) \leq \frac{\varepsilon/2}{1-\varepsilon/2}(A(1) + A(0)) \leq \varepsilon(A(1) + A(0))$. Now, we prove the claim. Again, we rewrite the LHS as

$$\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle i+1\rangle] \cdot \Pr[\mathbb{F}\langle i+1\rangle] = \sum_{|S|=i+1} \sum_{F_S} \sum_{H_{\bar{S}}} \Big(\Pr[F_S, H_{\bar{S}}] \cdot \mathsf{PM}(F_S, H_{\bar{S}}) \Big).$$

Similarly, we have the RHS to be

$$\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle i\rangle] \cdot \Pr[\mathbb{F}\langle i\rangle] = \sum_{|S'|=i} \sum_{F_{S'}} \sum_{H_{\bar{S}'}} \Big(\Pr[F_{S'}, H_{\bar{S}'}] \cdot \mathsf{PM}(F_{S'}, H_{\bar{S}'}) \Big) \text{ and}$$

$$\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle i-1\rangle] \cdot \Pr[\mathbb{F}\langle i-1\rangle] = \sum_{|S''|=i-1} \sum_{F_{S''}} \sum_{H_{\bar{S}''}} \Big(\Pr[F_{S''}, H_{\bar{S}''}] \cdot \mathsf{PM}(F_{S''}, H_{\bar{S}''}) \Big).$$

Let $C(F_S, H_{\bar{S}}) = \Pr[F_S, H_{\bar{S}}] \cdot \mathsf{PM}(F_S, H_{\bar{S}})$. Consider all $(F_{S'}, H_{\bar{S}'})$ with $|S'| = i$ and all $(F_{S''}, H_{\bar{S}''})$ with $|S''| = i-1$ as buyers. The buyers want to buy all terms in LHS. The budget of buyer $(F_{S'}, H_{\bar{S}'})/(F_{S''}, H_{\bar{S}''})$ is $C(F_{S'}, H_{\bar{S}'})/C(F_{S''}, H_{\bar{S}''})$. We show there is a charging scheme such that each term $C(F_S, H_{\bar{S}})$ is fully paid by the buyers and each buyer spends at most an $\frac{\varepsilon}{6}$ fraction of her budget.

Suppose we are selling the term $C(F_S, H_{\bar{S}})$. Consider the following charging scheme. Suppose $v_1 \in S$ the node that is realized to point $s_1 \in \mathcal{P} \setminus \mathbb{H}(v_1)$ which is the closest point to its stoch-core in F_S. Suppose $v_2 \in S$ the node that is realized to point $s_2 \in \mathcal{P} \setminus \mathbb{H}(v_2)$ which is the second closest point to its stoch-core in F_S. Let $S' = S \setminus \{v_1\}$, $S'' = S' \setminus \{v_2\}$. If $(F_{S'}, H_{\bar{S}'})$ is obtained from $(F_S, H_{\bar{S}})$ by sending v_1 to a point in its stoch-core $\mathbb{H}(v_1)$, we say $(F_{S'}, H_{\bar{S}'})$ is consistent with $(F_S, H_{\bar{S}})$, denoted as $(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})$. If $(F_{S''}, H_{\bar{S}''})$ is obtained from $(F_{S'}, H_{\bar{S}'})$ by sending v_2 to a point in its stoch-core $\mathbb{H}(v_2)$, we say $(F_{S''}, H_{\bar{S}''})$ is consistent with $(F_{S'}, H_{\bar{S}'})$, denoted as $(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})$. Let

$$Z(F_S, H_{\bar{S}}) = \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})} \Pr[(F_{S'}, H_{\bar{S}'})], \quad \text{and}$$

$$Z(F_{S'}, H_{\bar{S}'}) = \sum_{(F_{S''}, H_{\bar{S}''}) \sim (F_{S'}, H_{\bar{S}'})} \Pr[F_{S''}, H_{\bar{S}''}]$$

Now, we claim that for any fixed $(F_{S''}, H_{\bar{S}''})$,

$$\sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_{S''}, H_{\bar{S}''})} \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})} \leq \sum_{v \in \bar{S}''} \frac{\Pr[v \notin \mathbb{H}(v)]}{\Pr[v \in \mathbb{H}(v)]}.$$

The proof of the claim is essentially the same as in Lemma 114. We first observe that for a fixed node $v = S' \setminus S''$, the denominators of all terms are in fact the same by the definition of Z. Then, the proof can be completed by canceling out the same multiplicative terms from the numerators and the denominator.

Now, we specify how to charge each buyer. For each buyer $(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})$,

we charge $(F_{S'}, H_{\bar{S}'})$ the following amount of money

$$2(n+2)\Pr[F_S, H_{\bar{S}}]\mathsf{PM}(F_{S'}, H_{\bar{S}'}) \cdot \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_S, H_{\bar{S}})},$$

and we charge each buyer $(F_{S''}, H_{\bar{S}''})$ consistent with $(F_{S'}, H_{\bar{S}'})$ the following amount of money

$$2(n+2)\Pr[F''_S, H_{\bar{S}''}]\mathsf{PM}(F_{S''}, H_{\bar{S}''}) \cdot \frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})} \cdot \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})}.$$

In this case, we call $(F_{S''}, H_{\bar{S}''})$ a *sub-buyer* of the term $C(F_S, H_{\bar{S}})$. By Lemma 120, we can see that $A(F_S, H_{\bar{S}})$ is fully paid. To prove the claim, it suffices to show that each buyer $(F_{S'}, H_{\bar{S}'})$ and each sub-buyer $(F_{S''}, H_{\bar{S}''})$ has been charged at most $\frac{\varepsilon}{6}A(F_{S'}, H_{\bar{S}'})$ dollars. By the above charging scheme, the terms in LHS that are charged to buyer $(F_{S'}, H_{\bar{S}'})$ are consistent with $(F_{S'}, H_{\bar{S}'})$. Using the same argument as in Lemma 114, we can show that the spending of $(F_{S'}, H_{\bar{S}'})$ as a buyer is at most

$$\frac{\varepsilon}{nm} \cdot \mathsf{PM}(F_{S'}, H_{\bar{S}'}) \cdot \Pr[F_{S'}, H_{\bar{S}'}].$$

For notational convenience, we let $B = 2(n+2)\mathsf{PM}(F_{S''}, H_{\bar{S}''})\Pr[F_{S''}, H_{\bar{S}''}]$. The spending of $(F_{S''}, H_{\bar{S}''})$ as a sub-buyer can be bounded as follows:

$$\begin{aligned}
& B \cdot \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_{S''}, H_{\bar{S}''})} \sum_{(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})} \left(\frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})} \cdot \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})} \right) \\
\leq & B \cdot \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_{S''}, H_{\bar{S}''})} \sum_{(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})} \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})} \\
\leq & B \cdot mn \cdot \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_{S''}, H_{\bar{S}''})} \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})} \\
\leq & B \cdot mn \cdot \sum_{v \in \bar{S}''} \frac{\Pr[v \notin \mathbb{H}(v)]}{\Pr[v \in \mathbb{H}(v)]} \\
\leq & \frac{\varepsilon}{6} \cdot \mathsf{PM}(F_{S''}, H_{\bar{S}''}) \cdot \Pr[F_{S''}, H_{\bar{S}''}]
\end{aligned}$$

In the first inequality, we use the fact that $\frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})} \leq 1$. Note that for each $(F_{S'}, H_{\bar{S}'})$,

there are at most mn different $(F_S, H_{\bar{S}})$ such that $(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})$. So we have the second inequality. This completes the proof of the lemma.

□

Theorem 122. *Assuming the locational uncertainty model and that the number of nodes is even, there is an FPRAS for estimating the expected length of the minimum perfect matching.*

Remark. We have also tried to use the HPF method for this problem. The problem can be essentially reduced to the following bins-and-balls problem: Again each ball is thrown to the bins with nonuniform probabilities and we want to estimate the probability that each bin contains even number of balls. To the best of our knowledge, the problem is not studied before. The structure of the problem is somewhat similar to the permanent problem. We attempted to use the MCMC technique developed in [67], but the details become overly messy and we have not been able to provide a complete proof.

6.5 Minimum Cycle Covers

In this section, we consider the expected length of minimum cycle cover problem. In the deterministic version of the cycle cover problem, we are asked to find a collection of node-disjoint cycles such that each node is in one cycle and the total length is minimized. Here we assume that each cycle contains at least two nodes. If a cycle contains exactly two nodes, the length of the cycle is two times the distance between these two nodes. The problem can be solved in polynomial time by reducing the problem to a minimum bipartite perfect matching problem. [5] W.l.o.g., we assume that no two edges in $\mathcal{P} \times \mathcal{P}$ have the same length. For ease of exposition, we assume that for each point, there is only one node that may realize at this point. In principle,

[5] If we require each cycle consist at least three nodes, the problem is still poly-time solvable by a reduction to minimum perfect matching by Tutte [103]. Hartvigsen [63] obtained a polynomial time algorithm for minimum cycle cover with each cycle having at least 4 nodes Cornuéjols and Pulleyblank [32] have reported that Papadimitriou showed the NP-completeness of minimum cycle cover with each cycle having at least 6 nodes.

if more than one nodes may realize at the same point, we can create multiple copies of the point co-located at the same place, and impose a distinct infinitesimal distance between each pair of copies, to ensure that no two edges have the same distance.

We need the notion of the nearest neighbor graph, denoted by NN. For an undirected graph, an edge $e = (u, v)$ is in the nearest neighbor graph if u is the nearest neighbor of v, or vice versa. We also use NN to denote its length. $\mathbb{E}[\mathsf{NN}]$ can be computed exactly in polynomial time [71]. As a warmup, we first show that $\mathbb{E}[\mathsf{NN}]$ is a 2-approximation of $\mathbb{E}[\mathsf{CC}]$ in the following lemma.

Lemma 123. $\mathbb{E}[\mathsf{NN}] \leq \mathbb{E}[\mathsf{CC}] \leq 2\mathbb{E}[\mathsf{NN}]$.

Proof. We show that $\mathsf{NN} \leq \mathsf{CC} \leq 2\mathsf{NN}$ satisfies for each possible realization. We prove the first inequality. For each node u, there are two edges incident on u. Suppose they are e_{u1} and e_{u2}. We have $\mathsf{CC} = \frac{\sum_u (\mathrm{d}(e_{u1}) + \mathrm{d}(e_{u2}))}{2} \geq \mathsf{NN}$. The second inequality can be seen by doubling all edges in NN and the triangle inequality. □

We denote the longest edge in NN (and also its length) by L. Note that L is also a random variable. By the law of total expectation, we estimate $\mathbb{E}[\mathsf{CC}]$ based on the following formula:

$$\mathbb{E}[\mathsf{CC}] = \sum_{e \in \mathcal{P} \times \mathcal{P}} \Pr[\mathsf{L} = e] \cdot \mathbb{E}[\mathsf{CC} \mid \mathsf{L} = e]$$

It is obvious to see that $\frac{\mathsf{NN}}{m} \leq \mathsf{L} \leq \mathsf{NN}$. Combined with Lemma 123, we have that

$$\mathrm{d}(e) \leq \mathbb{E}[\mathsf{CC} \mid \mathsf{L} = e] \leq 2m\mathrm{d}(e). \tag{6.1}$$

However, it is not clear to us how to estimate $\Pr[\mathsf{L} = e]$ and how to take samples conditioning on event $\mathsf{L} = e$ efficiently. To circumvent the difficulty, we consider some simpler events. Consider a particular edge $e = (s, t) \in \mathcal{P} \times \mathcal{P}$. Denote as $N_s(t)$ the event that the nearest neighbor of s is t. Let L_{st} be the event the longest edge L in NN is $e = (s, t)$. Let $A_s(t) = N_s(t) \wedge L_{st}$. First we rewrite $\mathbb{E}[\mathsf{CC} \mid \mathsf{L} = e] \cdot \Pr[\mathsf{L} = e]$ by

$$\begin{aligned}\mathbb{E}[\mathsf{CC} \mid \mathsf{L} = e] \cdot \Pr[\mathsf{L} = e] =& \mathbb{E}[\mathsf{CC} \mid A_s(t) \vee A_t(s)] \cdot \Pr[A_s(t) \vee A_t(s)] \\ =& \mathbb{E}[\mathsf{CC} \mid A_s(t)] \cdot \Pr[A_s(t)] + \mathbb{E}[\mathsf{CC} \mid A_t(s)] \cdot \Pr[A_t(s)]\end{aligned}$$

$$- \mathbb{E}[\mathsf{CC} \mid A_s(t) \wedge A_t(s)] \cdot \Pr[A_s(t) \wedge A_t(s)]$$

Now, we show how to estimate $\mathbb{E}[\mathsf{CC} \mid A_s(t)] \cdot \Pr[A_s(t)]$ for each edge $e = (s, t)$. The other two terms can be estimated in the same way. Also notice that the third term is less than both the first term and the second term. Therefore, for any points s and t, we have the following fact which is useful later:

$$\mathbb{E}[\mathsf{CC}] \geq \mathbb{E}[\mathsf{CC} \mid \mathsf{L} = e] \cdot \Pr[\mathsf{L} = e] \geq \mathbb{E}[\mathsf{CC} \mid A_s(t)] \cdot \Pr[A_s(t)]. \tag{6.2}$$

By the above inequality, we can see that the total error for estimating the three terms is negligible compared to $\mathbb{E}[\mathsf{CC} \mid \mathsf{L} = e] \cdot \Pr[\mathsf{L} = e]$. Moreover, we have that

$$\begin{aligned}\mathbb{E}[\mathsf{CC} \mid A_s(t)] \cdot \Pr[A_s(t)] &= \mathbb{E}[\mathsf{CC} \mid A_s(t)] \cdot \Pr[L_{st} \wedge N_s(t)] \\ &= \mathbb{E}[\mathsf{CC} \mid A_s(t)] \cdot \Pr[L_{st} \mid N_s(t)] \cdot \Pr[N_s(t)]\end{aligned}$$

Suppose v is the node that may be realized to point s and u is the node that may be realized to point t. We use B as a shorthand notation for $\mathsf{B}(s, \mathrm{d}(s, t))$. We first observe that $\Pr[N_s(t)]$ can be computed exactly in poly-time as follows:

$$\Pr[N_s(t)] = p_{vs} \cdot p_{ut} \cdot \prod_{w \neq v,u} \big(1 - p_w(\mathsf{B})\big)$$

Also note that we can take samples conditioning on the event $N_s(t)$ (the corresponding probability distribution for node v is: $\Pr[v \vDash r \mid N_s(t)] = \frac{p_{vr}}{1-p_w(\mathsf{B})}$).

Estimating $\mathbb{E}[\,\mathsf{CC} \mid A_s(t)\,] \cdot \Pr[L_{st} \mid N_s(t)]$. Next, we show how to estimate $\mathbb{E}[\mathsf{CC} \mid A_s(t)] \cdot \Pr[L_{st} \mid N_s(t)]$. The high level idea is the following. We take samples conditioning on $N_s(t)$. If $\Pr[L_{st} \mid N_s(t)]$ is large (i.e., at least $1/\mathrm{poly}(nm)$), we can get enough samples satisfying L_{st}, thus $A_s(t)$. Therefore, we can get $(1 \pm \varepsilon)$-approximation for both $\Pr[L_{st} \mid N_s(t)]$ and $\mathbb{E}[\mathsf{CC} \mid A_s(t)]$ in poly-time (we also use the fact that if $A_s(t)$ is true, CC is at least $\mathrm{d}(s, t)$ and at most $2m\mathrm{d}(s, t)$). However, if $\Pr[L_{st} \mid N_s(t)]$ is small, it is not clear how to obtain a reasonable estimate of this value. In this case,

we show the contribution of the term to our final answer is extremely small and even an inaccurate estimation of the term will not affect our answer in any significant way with high probability.

Now, we elaborate the details. We iterate the following steps for N times ($N = O(\frac{m^2n^4}{\varepsilon^3}(\ln m + \ln n))$ suffices). Since there are $O(n^2)$ different edges between points, we totally need $O(\frac{m^2n^6}{\varepsilon^3}(\ln m + \ln n))$ iterations.

- Suppose we are in the ith iteration. We take a sample G_i of the stochastic graph conditioning on the event $N_s(t)$. We compute the nearest neighbor graph $\mathsf{NN}(G_i)$ and the minimum length cycle cover $\mathsf{CC}(G_i)$. If $e = (s, t)$ is the longest edge in $\mathsf{NN}(G_i)$, let $I_i = 1$. Otherwise $I_i = 0$.

Our estimate of $\mathbb{E}[\,\mathsf{CC} \mid A_s(t)\,] \cdot \Pr[L_{st} \mid N_s(t)]$ is the following:

$$\left(\frac{\sum_{i=1}^{N} I_i \cdot \mathsf{CC}(G_i)}{\sum_{i=1}^{N} I_i}\right)\left(\frac{\sum_{i=1}^{N} I_i}{N}\right) = \frac{\sum_{i=1}^{N} I_i \cdot \mathsf{CC}(G_i)}{N}$$

It is not hard to see that the expectation of $\frac{\sum_{i=1}^{N} I_i \cdot \mathsf{CC}(G_i)}{N}$ is exactly $\mathbb{E}[\,\mathsf{CC} \mid A_s(t)\,] \cdot \Pr[L_{st} \mid N_s(t)]$.

We distinguish the following two cases:

1. $\Pr[L_{st} \mid N_s(t)] \geq \frac{\varepsilon}{2mn^4}$. By Lemma 14, $\frac{\sum_{i=1}^{N} I_i}{N} \in (1 \pm \varepsilon)\Pr[L_{st} \mid N_s(t)]$ with high probability. In this case, we have enough successful samples (samples with $I_i = 1$) to guarantee that $\frac{\sum_{i=1}^{N} I_i \mathsf{CC}(G_i)}{\sum_{i=1}^{N} I_i}$ is a $(1\pm\varepsilon)$-approximation of $\mathbb{E}[\,\mathsf{CC} \mid A_s(t)\,]$ with high probability, again by Lemma 14. We note that under the condition $A_s(t)$, we can get a $(1 \pm \varepsilon)$-approximation since CC is at least $\mathrm{d}(s, t)$ and at most $2n\mathrm{d}(s, t)$.

2. $\Pr[L_{st} \mid N_s(t)] < \frac{\varepsilon}{2mn^4}$. We note that $I_i = 0$ means that while $N_s(t)$ happens, the longest edge L in NN is longer than $e = (s, t)$. Suppose $e' = (s', t')$ is the edge with the maximum $\Pr[L_{s't'}|N_s(t)]$. Since $\Pr[L_{st} \mid N_s(t)] \leq \frac{\varepsilon}{2mn^4}$, $e' = (s', t')$ must be different from $e = (s, t)$ and $\Pr[L_{s't'} \mid N_s(t)] \geq \frac{4mn^2}{\varepsilon}\Pr[L_{st} \mid N_s(t)]$.

Hence, we have that

$$\begin{aligned}
\mathbb{E}[\mathsf{CC} \mid A_s(t)] \cdot \Pr[A_s(t)] &= \mathbb{E}[\mathsf{CC} \mid A_s(t)] \cdot \Pr[L_{st} \mid N_s(t)] \cdot \Pr[N_s(t)] \\
&\leq 2m \cdot \mathrm{d}(s,t) \cdot \frac{\varepsilon}{4mn^2} \cdot \Pr[L_{s't'} \mid N_s(t)] \cdot \Pr[N_s(t)] \\
&\leq \frac{\varepsilon}{2n^2} \cdot \mathrm{d}(s',t') \cdot \Pr[L_{s't'} \mid N_s(t)] \cdot \Pr[N_s(t)] \\
&\leq \frac{\varepsilon}{2n^2} \cdot \mathbb{E}[\,\mathsf{CC} \mid A_{s'}(t')] \cdot \Pr[L_{s't'}] \\
&\leq \frac{\varepsilon}{2n^2} \cdot \mathbb{E}[\mathsf{CC}]
\end{aligned}$$

The first and third inequalities are due to (6.1) and the fourth are due to (6.2). By Chernoff Bound, we have that

$$\Pr\left[\frac{\sum_{i=1}^{N} I_i \cdot \mathsf{CC}(G_i)}{N} \geq \frac{\varepsilon}{n^2} \cdot \mathbb{E}[\mathsf{CC}]\right] \leq \frac{e^{-m}}{n^2}$$

Then, with probability at least $1 - \mathrm{poly}(\frac{1}{m})$, the contribution from all such edges is less than $\varepsilon\mathbb{E}[\mathsf{CC}]$.

Summing up, we have obtained the following theorem.

Theorem 124. *There is an FPRAS for estimating the expected length of the minimum length cycle cover in both the locational uncertainty model and the existential uncertainty model.*

Finally, we remark that our algorithm also works in presence of both locational uncertainty and node uncertainty, i.e., the existence of each node is a Bernoulli random variable. It is not hard to extend our technique to handle the case where each cycle is required to contain at least three nodes. This is done by considering the longest edge in the $2\mathsf{NN}$ graph (each node connects to the nearest and the second nearest neighbors). The extension is fairly straightforward and we omit the details here.

6.6 kth Longest m-Nearest Neighbor

We consider the problem of computing the expected length of the kth longest m-nearest neighbor (i.e., for each point, find the distance to its m-nearest neighbor, then compute the kth longest one among these distances) in the existential uncertainty model. We use kmNN to denote the length of the kth longest m-nearest neighbor.

Similar to k-clustering, we use the HPF $\Psi(\mathcal{P})$ for estimating $\mathbb{E}[\mathsf{kmNN}]$. We call a component a small component if it contains at most m present points. Let the random variable Y be the largest integer i such that there are at most $k-1$ present points among those small components in Γ_i. We can see that if $Y = i$ then the special component ν_i is not a small component, while both μ'_{i+1} and μ''_{i+1} should not be empty, and one of μ'_{i+1} and μ''_{i+1} must be a small component. Moreover, Γ'_i contains at most $k-1$ present points among those small components.

We can rewrite $\mathbb{E}[\mathsf{kmNN}]$ by $\mathbb{E}[\mathsf{kmNN}] = \sum_{i=1}^{n} \Pr[Y = i]\mathbb{E}[\mathsf{kmNN} \mid Y = i]$. By the Property P1 and P2 of $\Psi(\mathcal{P})$, we directly have the following lemma.

Lemma 125. *Conditioning on $Y = i$, it holds that $\mathrm{d}(e_i) \leq \mathsf{kmNN} \leq n\mathrm{d}(e_i)$.*

For a partition Γ on $\mathcal{P}$, we use $\Gamma\langle \#j, \leq m\rangle$ to denote the event that there are exactly j present points among those small components in Γ. The remaining task is to show how to compute $\Pr[Y = i]$ and how to estimate $\mathbb{E}[\mathsf{kmNN} \mid Y = i]$. We first prove the following lemma.

Lemma 126. *For a partition Γ on $\mathcal{P}$, we can compute $\Pr[\Gamma\langle \#j, \leq m\rangle]$ in polynomial time. Moreover, there exists a polynomial time sampler for sampling present points in Γ conditioning on $\Gamma\langle \#j, \leq m\rangle$.*

Proof. W.l.o.g, we assume that the components in Γ are $C_1, \ldots, C_n$. We denote $E[a, b]$ the event that among the first a components, exactly b points are present in those small components. We denote the probability of $E[a, b]$ by $\Pr[a, b]$. Note that our goal is to compute $\Pr[n, j]$. We have the following dynamic program:

1. If $\sum_{1\leq l\leq a} \min\{m, |C_l|\} < b$, $\Pr[a, b] = 0$. If $b = 0$, $\Pr[a, b] = \prod_{1\leq l\leq a}(\Pr[C_l\langle 0\rangle] + \Pr[C_l\langle \geq m + 1\rangle])$.

2. For $1 \leq b \leq \sum_{1\leq l\leq a} \min\{m, |C_l|\}$, $\Pr[a,b] = \sum_{0\leq l\leq n} \Pr[C_a\langle l\rangle] \cdot \Pr[a-1, b-l] + \Pr[C_a\langle \geq m+1\rangle] \cdot \Pr[a-1,b]$.

Thus we can compute $\Pr[n,j]$ in polynomial time. Similar to Lemma 108, we can also construct a polynomial uniform sampler.

□

To prove Theorem 128, we only need the following lemma.

Lemma 127. *We can compute* $\Pr[Y=i]$ *in polynomial time. Moreover, there exists a polynomial time sampler conditioning on* $Y=i$.

Proof. By the definition of $Y=i$, we can rewrite $\Pr[Y=i]$ as follows:

$$\begin{aligned}\Pr[Y=i] = &\sum_{1\leq n_1\leq m, m+1-n_1\leq n_2\leq m} \Pr[\mu'_{i+1}\langle n_1\rangle] \cdot \Pr[\mu''_{i+1}\langle n_2\rangle] \cdot \left(\sum_{k-n_1-n_2\leq l\leq k-1} \Pr[\Gamma'_i\langle \#l, \leq m\rangle]\right) \\ &+ \sum_{m+1\leq n_1\leq |\mu'_{i+1}|, 1\leq n_2\leq m} \Pr[\mu'_{i+1}\langle n_1\rangle] \cdot \Pr[\mu''_{i+1}\langle n_2\rangle] \cdot \left(\sum_{k-n_2\leq l\leq k-1} \Pr[\Gamma'_i\langle \#l, \leq m\rangle]\right) \\ &+ \sum_{1\leq n_1\leq m, m+1\leq n_2\leq |\mu''_{i+1}|} \Pr[\mu'_{i+1}\langle n_1\rangle] \cdot \Pr[\mu''_{i+1}\langle n_2\rangle] \cdot \left(\sum_{k-n_1\leq l\leq k-1} \Pr[\Gamma'_i\langle \#l, \leq m\rangle]\right)\end{aligned}$$

Note that we can compute $\Pr[Y=i]$ in polynomial time by Lemma 126. Using the same argument as in Lemma 129, we can construct a polynomial uniform sampler conditioning on $Y=i$. By Lemma 125, we only need to take $O(\frac{n}{\varepsilon^2}\ln n)$ independent samples for estimating $\mathbb{E}[\mathsf{kmNN} \mid Y=i]$. So we take $O(\frac{n^2}{\varepsilon^2}\ln n)$ independent samples in total.

□

Theorem 128. *There is an FPRAS for estimating the expected length of the kth longest m-nearest neighbor in the existential uncertainty model.*

6.7 Missing Proofs

6.7.1 Closest Pair

Lemma 101. *Steps 1,2,3 in Algorithm 3 provide* $(1\pm\varepsilon)$*-approximations for* $\Pr[\mathbb{F}\langle i\rangle \wedge \mathsf{C} \leq 1]$ *for* $i = 0, 1, 2$ *respectively, with high probability.*

Proof. As we just argued, $\Pr[\mathbb{F}\langle 1\rangle \wedge \mathsf{C} \leq 1]$ can be estimated since $I(\mathsf{C} \leq 1)$, conditioned on $\mathbb{F}\langle 0\rangle$, is poly-bounded. For estimating $\Pr[\mathbb{F}\langle 1\rangle \wedge \mathsf{C} \leq 1]$, we first rewrite this term by $\sum_{s_i\in\mathbb{F}} \Pr[\mathbb{F}\langle\{s_i\}\rangle \wedge \mathsf{C} \leq 1]$. For a point $s_i \in \mathbb{F}$, note that $\Pr[\mathbb{F}\langle\{s_i\}\rangle \wedge \mathsf{C} \leq 1] = \Pr[\mathbb{F}\langle\{s_i\}\rangle] \cdot \Pr[\mathsf{C} \leq 1 \mid \mathbb{F}\langle\{s_i\}\rangle]$. Since we have that $p_i(1 - \frac{\varepsilon}{n}) \leq \Pr[\mathbb{F}\langle\{s_i\}\rangle] \leq p_i$ by the first property of the stoch-core $\mathbb{H}$, we can use p_i to estimate $\Pr[\mathbb{F}\langle\{s_i\}\rangle]$. For estimating $\Pr[\mathsf{C} \leq 1 \mid \mathbb{F}\langle\{s_i\}\rangle]$, we denote $\mathsf{B}_{s_i} = \{t \in \mathbb{H} : \mathrm{d}(s_i, t) \leq 1\}$. If B_{s_i} is not empty, we can use Monte Carlo for estimating $\Pr[\mathsf{C} \leq 1 \mid \mathbb{F}\langle\{s_i\}\rangle]$ since its value is at least $\frac{\varepsilon}{n^2}$. Otherwise, computing $\Pr[\mathsf{C} \leq 1 \mid \mathbb{F}\langle\{s_i\}\rangle]$ is equivalent to computing $\Pr[\mathsf{C} \leq 1 \mid \mathbb{F}\langle 0\rangle]$ in the instance without s_i (since s_i is at distance more than 1 from any other point). The proof for $\Pr[\mathbb{F}\langle 2\rangle \wedge \mathsf{C} \leq 1]$ is almost the same and we do not repeat it.

□

6.7.2 Minimum Spanning Tree

Lemma 112. *Algorithm 6 produces a* $(1\pm\varepsilon)$*-estimate for the second term with high probability.*

Proof. To compute the second term, we first rewrite it as follows:

$$\mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle 1\rangle\,] \cdot \Pr[\mathbb{F}\langle 1\rangle] = \sum_{v\in\mathcal{V}} \Big(\sum_{s\in F} \Pr[\mathbb{F}\langle v\rangle \wedge v \vDash s]\, \mathbb{E}[\,\mathsf{MST} \mid \mathbb{F}\langle v\rangle, v \vDash s] \Big)$$

Fix a node v. To estimate $\sum_{s\in F} \Pr[\mathbb{F}\langle v\rangle \wedge v \vDash s]\, \mathbb{E}[\mathsf{MST} \mid \mathbb{F}\langle v\rangle, v \vDash s]$, we consider the following two situations:

1. Point $s \in B$, i,e, $\mathrm{d}(s, \mathbb{H}) < \frac{m}{\varepsilon} \cdot \mathrm{diam}(\mathbb{H})$.

We estimate the sum for all $s \in B$. Notice that the sum is in fact $\Pr[\mathsf{CI}(v)] \cdot \mathbb{E}[\,\mathsf{MST} \mid \mathsf{CI}(v)]$. We can see that $\Pr[\mathsf{CI}(v)]$ can be computed exactly in linear time. We argue that the quality of the estimation taken on $N_1 = O\big(\frac{mn^2}{\varepsilon^5} \ln m\big)$ samples is sufficient by considering the following two cases:

(a) Assume that $\mathbb{E}[\,\mathsf{MST} \mid \mathsf{CI}(v)] \geq \frac{1}{2}\mathbb{E}[\,\mathsf{MST} \mid \mathbb{H}\langle m\rangle] \geq \Omega\Big(\frac{\varepsilon^2}{n^2}\Big)\mathrm{diam}(\mathbb{H})$. In this case, we have a poly-bounded random variable. This is because under the condition $\mathsf{CI}(v)$, the maximum possible length of any minimum spanning tree is $O(\frac{m}{\varepsilon}\mathrm{diam}(\mathbb{H}))$. Hence we can use Monte Carlo to get a $(1 \pm \varepsilon)$-approximation of $\mathbb{E}[\,\mathsf{MST} \mid \mathsf{CI}(v)]$ with $O\big(\frac{mn^2}{\varepsilon^5} \ln m\big)$ samples.

(b) Otherwise, we assume that $\mathbb{E}[\,\mathsf{MST} \mid \mathsf{CI}(v)] \leq \frac{1}{2}\mathbb{E}[\,\mathsf{MST} \mid \mathbb{H}\langle m\rangle]]$. Let V_0 be the collection of these nodes. The probability that the sample average is larger than $\mathbb{E}[\mathsf{MST} \mid \mathbb{H}\langle m\rangle]]$ is at most $\mathrm{poly}(\frac{1}{m})$ by Chernoff Bound. The probability that for all nodes $v \in V_0$, the sample average are at most $\mathbb{E}[\mathsf{MST} \mid \mathbb{H}\langle m\rangle]]$ is at least $1 - \mathrm{poly}(\frac{1}{m})$ by union bound. If this is the case, we can see their total contribution to the final estimation of $\mathbb{E}[\mathsf{MST}]$ is less than $\varepsilon\mathbb{E}[\,\mathsf{MST} \mid \mathbb{H}\langle m\rangle]]\Pr[\mathbb{H}\langle m\rangle]$. In fact, this is because

$$\sum_{v \in V_0} \Pr[\mathsf{CI}(v)] \cdot T_v \leq \sum_{v \in V_0} \Pr[\mathsf{CI}(v)] \cdot \mathbb{E}[\,\mathsf{MST} \mid \mathbb{H}\langle m\rangle] < \varepsilon\mathbb{E}[\,\mathsf{MST} \mid \mathbb{H}\langle m\rangle]\Pr[\mathbb{H}\langle m\rangle].$$

The second inequality is due to the fact that $\sum_{v \in V_0} \Pr[\mathsf{CI}(v)] \leq m - p(\mathbb{H}) < \varepsilon/16 < \varepsilon\Pr[\mathbb{H}\langle m\rangle]$.

2. Point $s \in \mathbb{F} \setminus B$, each term has $\mathrm{d}(s, \mathbb{H}) > \frac{m}{\varepsilon} \cdot \mathrm{diam}(\mathbb{H})$.
   We just use $\mathrm{d}(s, \mathbb{H})$ as the estimation of $\mathbb{E}[\mathsf{MST} \mid \mathbb{F}\langle v\rangle, v \vDash s]$. This is because the length of MST is always at least $\mathrm{d}(s, \mathbb{H})$ and at most $\mathrm{d}(s, \mathbb{H}) + m \cdot \mathrm{diam}(\mathbb{H}) \leq (1 + \varepsilon)\mathrm{d}(s, \mathbb{H})$.

□

6.7.3 Minimum Perfect Matching

Lemma 119. *Algorithm 6.4 produces a $(1 \pm \varepsilon)$-estimate for the second term with high probability.*

Proof. To compute the second term, we first rewrite it as follows:

$$\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle 1\rangle] \cdot \Pr[\mathbb{H}\langle 1\rangle] = \sum_{v\in\mathcal{V}}\Big(\sum_{s\notin\mathbb{H}(v)} \Pr[\mathbb{F}\langle v\rangle \wedge v \vDash s]\ \mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle v\rangle, v \vDash s]\Big).$$

Fix a particular node v. To estimate $\sum_{s\in\mathbb{F}} \Pr[\mathbb{F}\langle v\rangle \wedge v \vDash s]\,\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle v\rangle, v \vDash s]$, we consider the following two situations:

1. Point $s \in B_v$, i,e, $\mathrm{d}(s, \mathbb{H}(v)) < \frac{4mD}{\varepsilon}$.

   We estimate the sum for all $s \in B^v$. Notice that the sum is in fact $\Pr[\mathsf{CI}(v)] \cdot \mathbb{E}[\,\mathsf{PM} \mid \mathsf{CI}(v)]$. We can see that $\Pr[\mathsf{CI}(v)]$ can be computed exactly in linear time. We argue that the quality of the estimation taken on $N_2 = O\big(\frac{m^2n^5}{\varepsilon^4}\ln m\big)$ samples is poly-bounded by considering the following two cases:

   (a) Assume that $\mathbb{E}[\,\mathsf{PM} \mid \mathsf{CI}(v)] \geq \frac{1}{2}\mathbb{E}[\,\mathsf{PM} \mid \mathbb{H}\langle m\rangle] = \Omega\Big(\frac{\varepsilon D}{mn^5}\Big)$. In this case, our estimation is poly-bounded. This is because under the condition $\mathsf{CI}(v)$, the maximum possible length of any minimum perfect matching is $O(\frac{mD}{\varepsilon})$. Hence we can use Monte Carlo to get a $(1 \pm \varepsilon)$-approximation of $\mathbb{E}[\,\mathsf{PM} \mid \mathsf{CI}(v)]$ with $O\big(\frac{m^2n^5}{\varepsilon^4}\ln m\big)$ samples.

   (b) Otherwise, we assume that $\mathbb{E}[\,\mathsf{PM} \mid \mathsf{CI}(v)] \leq \frac{1}{2}\mathbb{E}[\,\mathsf{PM} \mid \mathbb{H}\langle m\rangle]]$. Let V_0 be the collection of these nodes. The probability that the sample average is larger than $\mathbb{E}[\mathsf{PM} \mid \mathbb{H}\langle m\rangle]]$ is at most $\mathrm{poly}(\frac{1}{m})$ by Chernoff Bound. The probability that for each node $v \in V_0$, the sample average is at most $\mathbb{E}[\mathsf{PM} \mid \mathbb{H}\langle m\rangle]]$ is at least $1 - \mathrm{poly}(\frac{1}{m})$ by union bound. If this is the case, we can see their total contribution to the final estimation of $\mathbb{E}[\mathsf{PM}]$ is less than $\varepsilon\mathbb{E}[\,\mathsf{PM} \mid \mathbb{H}\langle m\rangle]]\Pr[\mathbb{H}\langle m\rangle]$. In fact, this is because

   $$\sum_{v\in V_0} \Pr[\mathsf{CI}(v)]\cdot T_v \leq \sum_{v\in V_0} \Pr[\mathsf{CI}(v)]\cdot\mathbb{E}[\,\mathsf{PM} \mid \mathbb{H}\langle m\rangle]] < \varepsilon\mathbb{E}[\,\mathsf{PM} \mid \mathbb{H}\langle m\rangle]]\Pr[\mathbb{H}\langle m\rangle].$$

The second inequality is due to the fact that $\sum_{v\in V}\Pr[\mathsf{CI}(v)] \leq m - \sum_{v\in V_0} p_v(\mathbb{H}(v)) \leq \frac{\varepsilon}{n^3} < \varepsilon\Pr[\mathbb{H}\langle m\rangle]$.

2. Point $s \in \mathcal{P}\setminus(B_v \cup \mathbb{H}(v))$, each term has $\mathrm{d}(s,\mathbb{H}(v)) > \frac{4mD}{\varepsilon}$. The algorithm uses $\mathrm{d}(s,\mathbb{H}(v))$ as the estimation of $\mathbb{E}[\mathsf{PM} \mid \mathbb{F}\langle v\rangle, v \vDash s]$. Note that the length of PM is always at least $\mathrm{d}(s,\mathbb{H}(v)) - mD \geq (1-\frac{\varepsilon}{4})\mathrm{d}(s,\mathbb{H}(v))$. This is because such an instance PM contains a path from s to some point $t \in \mathbb{H}(v)$ deleting no more than m segments of length at most D (each segment is in some $\mathbb{H}_j$). On the other hand, the length of PM is at most $\mathrm{d}(s,\mathbb{H}(v)) + mD \leq (1+\frac{\varepsilon}{4})\mathrm{d}(s,\mathbb{H}(v))$. So it is a $(1\pm\varepsilon)$-estimation.

□

6.8 The Closest Pair Problem

6.8.1 Estimating kth Closest Pair in the Existential Uncertainty Model

Again, we construct the HPF $\Psi(\mathcal{P})$. Let the random variable Y be the largest integer i such that there are at least k point collisions in Γ_i. Here we use a point collision to denote that a pair of points are present in the same component. Note that if there are exactly i points in a component, the amount of point collisions in this component is $\binom{i}{2}$. We denote as $\Gamma\langle\#j\rangle$ the event that there are exactly j point collisions among the partition Γ on $\mathcal{P}$. Similarly, we can rewrite $\mathbb{E}[\mathsf{kC}]$ by $\mathbb{E}[\mathsf{kC}] = \sum_{i=1}^{m-1}\Pr[Y=i]\mathbb{E}[\mathsf{kC} \mid Y=i]$.

We use dynamic programming technique to achieve an FPRAS for computing $\mathbb{E}[\mathsf{kC}]$. Note that conditioning on $Y=i$, the value of kC is between $\mathrm{d}(e_i)$ and $m\cdot\mathrm{d}(e_i)$. So we only need to show the following lemma.

Lemma 129. *We can compute* $\Pr[Y=i]$ *in polynomial time. Moreover, there exists a polynomial time sampler conditioning on* $Y=i$.

Proof. We denote $E[a,b]$ $(1\leq a\leq i-1,\ b\leq k)$ the event that among the first a components in Γ'_i, there are exactly $b\leq k$ point collisions. We denote the probability

of $E[a,b]$ by $\Pr[a,b]$. We give the dynamic programming as follows.

1. If $\sum_{1\le j\le a}\binom{|C_j|}{2} < b$, $\Pr[a,b]=0$. If $b=0$, $\Pr[a,b]=\prod_{1\le j\le a}\Pr[C_j\langle\le 1\rangle]$. If $b<0$, $\Pr[a,b]=0$.

2. If $\sum_{1\le j\le a}\binom{|C_j|}{2}\ge b$, $1\le b\le k$, $\Pr[a,b]=\sum_{0\le l\le n_a}\Pr[C_a\langle l\rangle]\cdot\Pr[a-1,b-\binom{l}{2}]$.

By the above dynamic programming, we can compute $\Pr[i-1,l]$ for $0\le l\le k-1$ in polynomial time.

By the definition of $Y=i$, it is no hard to see that we can rewrite $\Pr[Y=i]$ as follows:

$$\Pr[Y=i]=\sum_{1\le n_1\le|\mu'_{i+1}|,1\le n_2\le|\mu''_{i+1}|}\Pr[\mu'_{i+1}\langle n_1\rangle]\cdot\Pr[\mu''_{i+1}\langle n_2\rangle]\cdot\left(\sum_{k-\binom{n_1+n_2}{2}\le l\le k-1-\binom{n_1}{2}-\binom{n_2}{2}}\Pr[\Gamma'_i\langle\#l\rangle]\right)$$

Note that we can compute $\Pr[Y=i]$ in polynomial time. We need to describe our sampler conditioning on $Y=i$. We first sample the event $\mu'_{i+1}\langle n_1\rangle\wedge\mu''_{i+1}\langle n_2\rangle$ with probability $\Pr[\mu'_{i+1}\langle n_1\rangle\wedge\mu''_{i+1}\langle n_2\rangle\mid Y=i]$. Then conditioning on $k-\binom{n_1+n_2}{2}\le l\le k-1-\binom{n_1}{2}-\binom{n_2}{2}$, we sample the total number of point collisions in Γ'_i. Then we sample the number of present points in each component in Γ'_i using the dynamic programming. Finally, based on the number of present points in each component, we sample the present points by Lemma 108.

Using the Monte Carlo method, we only need to take $O(\frac{n}{\varepsilon^2}\ln n)$ independent samples for estimating $\mathbb{E}[\mathsf{kC}\mid Y=i]$. Thus, we totally take $O(\frac{n^2}{\varepsilon^2}\ln n)$ independent samples.

□

Theorem 130. *There is an FPRAS for estimating the expected distance between the kth closest pair in the existential uncertainty model.*

6.8.2 Hardness for Closest Pair

Theorem 131. *Computing* $\Pr[\mathsf{C}\ge 1]$ *is #P-hard to approximate within any factor in a metric space in both the existential and locational uncertainty models.*

Proof. First consider the existential uncertainty model. Consider a metric graph G with edge weights being either 0.9 or 1.8. Each vertex in this graph exists with probability 1/2. Let G' be the unweighted graph with the same number of vertices. G' contains only those edges corresponding to edges with weight 0.9 in G. It is not hard to see that

$$\Pr[\mathsf{C} \geq 1] = \#\text{independent sets of size at least two in } G' \cdot \frac{1}{2^n}.$$

The right hand side is well known to be inapproximable for arbitrary graphs [97].

For the locational model, let the instance be G (with n vertices $s_1, \ldots, s_n$) with n additional vertices $t_1, \ldots, t_n$ which are far away from each other and any vertex in G. Let the probability distribution of node v_i be $p_{v_i s_i} = 1/2$, and $p_{v_i t_i} = 1/2$. We can see that in this locational uncertainty model, the value $\Pr[\mathsf{C} \geq 1]$ is the same as that in the corresponding existential model G.

□

Theorem 132. *Computing* $\mathbb{E}[\mathsf{C}]$ *exactly in both the existential and locational uncertainty models is #P-hard in a metric space.*

Proof. Consider a metric graph G with edge weights being either 1 or 2. Each vertex in this graph exists with probability 1/2. Note that

$$\mathbb{E}[\mathsf{C}] = \Pr[\mathsf{C} = 1] + 2\Pr[\mathsf{C} = 2] = (\Pr[\mathsf{C} \leq 1] - \Pr[\mathsf{C} = 0]) + 2(1 - \Pr[\mathsf{C} \leq 1])$$

Computing $\Pr[\mathsf{C} = 0]$ can be easily done in polynomial time. Computing $\Pr[\mathsf{C} \leq 1]$ in such a graph is as hard as counting independent sets in general graphs, hence is also #P-hard (as in Theorem 131). So, computing $\mathbb{E}[\mathsf{C}]$ is #P-hard as well.

For the locational model, let the instance be G (with n vertices $s_1, \ldots, s_n$) with n additional vertices $t_1, \ldots, t_n$ which satisfies $\mathrm{d}(s_i, t_j) = \mathrm{d}(t_i, t_j) = 5$ $(1 \leq i, j \leq m, i \neq j)$. Let the probability distribution of node v_i be $p_{v_i s_i} = 1/2$, and $p_{v_i t_i} = 1/2$. It is not hard to see that in this locational uncertainty model, the value $\mathbb{E}[\mathsf{C}]$ is linearly related to the value $\mathbb{E}[\mathsf{C}]$ in the existential model G. Therefore, computing $\mathbb{E}[\mathsf{C}]$ is

also #P-hard in the locational uncertainty model.

□

6.9 Another FPRAS for MST

W.l.o.g., we assume that for each point, there is only one node that may be realized to this point. Our algorithm is a slight generalization of the one proposed in [71]. Let $\mathbb{E}[i]$ be the expected MST length conditioned on the event that all nodes $\{v_1, \ldots, v_m\}$ are realized to points in $\{s_i, \ldots, s_n\}$ (denote the event by $\mathsf{In}(i, n)$). Let $\mathbb{E}'[i]$ be the expected MST length conditioned on the event that all nodes $\{v_1, \ldots, v_n\}$ are realized to $\{s_i, \ldots, s_n\}$ and at least one node is realized to s_i. We use $s \vDash s$ to denote the event that node v is realized to point s. Note that

$$\mathbb{E}[i] = \mathbb{E}'[i]\Pr[\exists v, v \vDash s_i \mid \mathsf{In}(i, m)] + \mathbb{E}[i+1]\Pr[\nexists v, v \vDash s_i \mid \mathsf{In}(i, m)]$$

For a particular point s_i, we reorder the points $\{s_i, \ldots, s_n\}$ as $\{s_i = r_i, \ldots, r_n\}$ in increasing order of distance from s_i. Let $\mathbb{E}'[i, j]$ be the expected MST length for all nodes conditioned on the event that all nodes are realized to $\{r_i, \ldots, r_j\}$ (denoted as $\mathsf{In}'(i, j)$) and $\exists v, v \vDash s_i$. Let $\mathbb{E}''[i, j]$ be the expected MST length for all nodes conditioned on the event $\mathsf{In}'(i, j) \wedge (\exists v, v \vDash r_i) \wedge (\exists s', s' \vDash r_j)$. We can see that

$$\begin{aligned}\mathbb{E}'[i, j] =& \mathbb{E}''[i, j]\Pr[\exists v', v' \vDash r_j \mid \mathsf{In}'(i, j), \exists v, v \vDash r_i] \\ &+ \mathbb{E}'[i, j-1]\Pr[\nexists v, v \vDash r_i \mid \mathsf{In}'(i, j), \exists v, v \vDash r_i]\end{aligned}$$

It is not difficult to see the probability $\Pr[\exists v', v' \vDash r_j \mid \mathsf{In}'(i, j), \exists v, v \vDash r_i]$ can be computed in polynomial time. Here we use the assumption that for each point, only one node that may realize to it. Moreover, we can also take samples conditioning on event $\mathsf{In}'(i, j) \wedge (\exists v, v \vDash r_i) \wedge (\exists v', v' \vDash r_j)$. Therefore $\mathbb{E}''[i, j]$ can be approximated within a factor of $(1 \pm \varepsilon)$ using the Monte Carlo method in polynomial time since it is poly-bounded. The number of samples needed can be bounded by $O\big(\frac{mn^2}{\varepsilon^2} \ln m\big)$.

We can easily generalize the above algorithm to the case where $\sum_{j=1}^{n} p_{ij} \leq 1$,

i.e., node i may not be present with some certainty. Indeed, this can be done by generalizing the definition of $\mathsf{In}(i,j)$ (and similarly $\mathsf{In}'(i,j)$) to be the event that each node is either absent or realized to some point in $\{r_i,\ldots,r_j\}$.

Chapter 7 Concluding Remarks

In this dissertation, we study two famous stochastic geometry models, the locational uncertainty model and the existential uncertainty model.

In the first part of the dissertation, we study how to construct coresets for different stochastic problems. We initiate the study of constructing ε-kernel coresets in stochastic geometry models. We consider approximating the expected width (an ε-EXP-KERNEL), as well as the probability distribution on the width (an (ε, τ)-QUANT-KERNEL) for any direction. We provide efficient algorithms for constructing such ε-kernel coresets in nearly linear time. Our ε-kernel coresets have a few applications, including approximating the extent of uncertain functions, maintaining extent measures for stochastic moving points and giving PTAS for some stochastic shape fitting problems.

We also study another two important stochastic geometric optimization problems, the k-center problem and the j-flat-center problem in stochastic geometry models in Euclidean spaces. We first think each of the stochastic problems as a certain deterministic problem over (exponential many) all possible realizations (each being a point set). By this view, we introduce a new notion called *generalized coreset*, which is a collection of realizations (instead of points for coresets). We also propose a new framework for generalized coreset construction. By the framework, we provide the first PTAS (Polynomial Time Approximation Scheme) for both stochastic geometry optimization problems, which generalize the previous results for stochastic minimum enclosing ball [88] and stochastic enclosing cylinder [66].

The second part of the dissertation is to estimate the expected value of a variety combinatorial objects over stochastic data. Several geometric properties of a set of stochastic points have been studied extensively in the literature under the term *stochastic geometry*. For instance, it is well known that if there are n points uniformly and

independently distributed in $[0,1]^2$, the minimal traveling salesman tour/minimum spanning tree/minimum matching visiting them has an expected length $\Theta(\sqrt{n})$ [21, 28]. Compared with results in stochastic geometry, we focus on the efficient computation of the statistics, instead of giving explicit mathematical formulas.

We study the problems of computing the expected lengths of several combinatorial or geometric optimization problems in both models, including closest pair, minimum spanning tree, k-clustering, minimum perfect matching, and minimum cycle cover. We also consider the problem of estimating the probability that the length of closest pair, or the diameter, is at most, or at least, a given threshold. Most of the above problems are known to be #P-hard. Caused by the high variance, we can not directly use Monte Carlo method to estimate the expected value. Thus, we develop two new techniques: *stoch-core* and *Hierarchical Partition Family (HPF)*. Both techniques are used to decompose the expectation of certain random variable into a convex combination of conditional expectations, such that each conditional expectation has a low variance. Combining our new techniques and Monte Carlo method, we obtain FPRAS (Fully Polynomial Randomized Approximation Scheme) for most of these problems in stochastic geometry models.

There are still many open optimization problems over different stochastic models. In this dissertation, we study the discrete stochastic models. In practice, the distribution of point locations often follows some continuous distribution in practice, such as GPS system, robot control, and so on. Many fundamental issues in this domain, such as many classic geometry computation and optimization problems are still not well understood by researchers. We believe it is a fruitful direction for further research.

Bibliography

[1] A. Abdullah, S. Daruki, and J.M. Phillips. Range counting coresets for uncertain data. In *Proceedings 29th ACM Syposium on Computational Geometry*, pages 223–232, 2013.

[2] P. Afshani, P.K. Agarwal, L. Arge, K.G. Larsen, and J.M. Phillips. (Approximate) uncertain skylines. In *Proceedings of the 14th International Conference on Database Theory*, pages 186–196, 2011.

[3] P.K. Agarwal, S.W. Cheng, Y. Tao, and K. Yi. Indexing uncertain data. In *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 137–146. ACM, 2009.

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

[5] P.K. Agarwal, A. Efrat, S. Sankararaman, and W. Zhang. Nearest-neighbor searching under uncertainty. In *Proceedings of the 31st Symposium on Principles of Database Systems*, pages 225–236, 2012.

[6] P.K. Agarwal, S. Har-Peled, S. Suri, H. Yıldız, and W. Zhang. Convex hulls under uncertainty. In *Proceedings of the 22nd Annual European Symposium on Algorithms*, pages 37–48, 2014.

[7] P.K. Agarwal, S. Har-Peled, and K. Varadarajan. Approximating extent measures of points. *Journal of the ACM*, 51(4):606–635, 2004.

[8] P.K. Agarwal, S. Har-Peled, and K. Varadarajan. Geometric approximation via coresets. *Combinatorial and Computational Geometry*, 52:1–30, 2005.

[9] P.K. Agarwal, S. Har-Peled, and H. Yu. Robust shape fitting via peeling and grating coresets. *Discrete & Computational Geometry*, 39(1-3):38–58, 2008.

[10] P.K. Agarwal, J. Matoušek, and S. Suri. Farthest neighbors, maximum spanning trees and related problems in higher dimensions. *Computational Geometry - Theory and Applications*, 1(4):189–201, 1992.

[11] P.K. Agarwal and C.M. Procopiuc. Exact and approximation algorithms for clustering. *Algorithmica*, 33(2):201–226, 2002.

[12] P.K. Agarwal and M. Sharir. Arrangements and their applications. *Handbook of Computational Geometry,* J. Sack and J. Urrutia (eds.), pages 49–119. Elsevier, Amsterdam, The Netherlands, 2000.

[13] C. Alexopoulos and J.A. Jacobson. State space partition algorithms for stochastic systems with applications to minimum spanning trees. *Networks*, 35(2):118–138, 2000.

[14] M. Anthony and P.L. Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[15] M.J. Atallah, Y. Qi, and H. Yuan. Asymptotically efficient algorithms for skyline probabilities of uncertain data. *ACM Trans. Datab. Syst*, 32(2):12, 2011.

[16] D. Bandyopadhyay and J. Snoeyink. Almost-Delaunay simplices: Nearest neighbor relations for imprecise points. In *Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms*, pages 410–419, 2004.

[17] N. Bansal, A. Gupta, J. Li, J. Mestre, V. Nagarajan, and A. Rudra. When *lp* is the cure for your matching woes: Improved bounds for stochastic matchings. In *European Symposium on Algorithms*, pages 218–229. Springer, 2010.

[18] J.F. Bard and J.E. Bennett. Arc reduction and path preference in stochastic acyclic networks. *Management Science*, 37(2):198–215, 1991.

[19] G. Barequet and S. Har-Peled. Efficiently approximating the minimum-volume bounding box of a point set in three dimensions. *Journal of Algorithms*, 38(1):91–109, 2001.

[20] S. Basu, R. Pollack, and M. Roy. Algorithms in real algebraic geometry. *AMC*, 10:12, 2011.

[21] J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. In *Proc. Cambridge Philos. Soc*, pages 55:299–327, 1959.

[22] M. W. Bern and D. Eppstein. Worst-case bounds for suadditive geometric graphs. In *Symposium on Computational Geometry*, pages 183–188, 1993.

[23] D.J. Bertsimas and G. van Ryzin. An asymptotic determination of the minimum spanning tree and minimum matching constants in geometrical probability. *Operations Research Letters*, 9(4):223–231, 1990.

[24] A. Bhalgat. A $(2+\varepsilon)$-approximation algorithm for the stochastic knapsack problem. *Unpublished manuscript*, 2011.

[25] T.M. Chan. Approximating the diameter, width, smallest enclosing cylinder, and minimum-width annulus. In *Proceedings of the 16th Annual Son Computational Geometry*, pages 300–309, 2000.

[26] T.M. Chan. Faster core-set constructions and data-stream algorithms in fixed dimensions. *Computational Geometry: Theory and Applications*, 35:20–35, 2006.

[27] K. Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

[28] R. Cheng, J. Chen, M. Mokbel, and C. Chow. Probabilistic verifiers: Evaluating constrained nearest-neighbor queries over uncertain data. In *ICDE*, 2008.

[29] R. Cheng, J. Chen, and X. Xie. Cleaning uncertain data with quality guarantees. *Proceedings of the VLDB Endowment*, 1(1):722–735, 2008.

[30] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In *Proceedings of the 27th Symposium on Principles of Database Systems*, pages 191–200, 2008.

[31] G. Cormode and S. Muthukrishnan. Radial histograms for spatial streams. Technical Report 2003-11, Center for Discrete Mathematics and Computer Science (DIMACS), 2003.

[32] G. Cornuejols and W. Pulleyblank. A matching problem with side constraints. *Discrete Math.*, 29, 1980.

[33] B.C. Dean, M.X. Goemans, and J. Vondrdk. Approximating the stochastic knapsack problem: The benefit of adaptivity. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 208–217. IEEE, 2004.

[34] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the 17th ACM-SIAM symposium on Discrete algorithm*, pages 1117–1126, 2006.

[35] X. Dong, A.Y. Halevy, and C. Yu. Data integration with uncertainty. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pages 687–698, 2007.

[36] A. Driemel, H. HAverkort, M. Löffler, and R.I. Silveira. Flow computations on imprecise terrains. *Journal of Computational Geometry*, 4:38–78, 2013.

[37] R.M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227–236, 1974.

[38] M. Dyer. Approximate counting by dynamic programming. In *ACM Symposium on Theory of Computing*, pages 693–699, 2003.

[39] H. Edelsbrunner, J. O'Rourke, and R. Seidel. Constructing arrangements of lines and hyperplanes with applications. *SIAM Journal on Computing*, 15(2):341–363, 1986.

[40] Y. Emek, A. Korman, and Y. Shavitt. Approximating the statistics of various properties in randomly weighted graphs. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1455–1467. SIAM, 2011.

[41] W. Evans and J. Sember. The possible hull of imprecise points. In *Proceedings of the 23rd Canadian Conference on Computational Geometry*, 2011.

[42] T. Feder and D. Greene. Optimal algorithms for approximate clustering. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 434–444. ACM, 1988.

[43] D. Feldman, A. Fiat, H. Kaplan, and K. Nissim. Private coresets. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pages 361–370, 2009.

[44] D. Feldman, A. Fiat, and M. Sharir. Coresets forweighted facilities and their applications. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 315–324. IEEE, 2006.

[45] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing*, pages 569–578, 2011.

[46] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1434–1453. SIAM, 2013.

[47] D. Feldman and L.J. Schulman. Data reduction for weighted and outlier-resistant clustering. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1343–1354. SIAM, 2012.

[48] A.M. Frieze. On the value of a random minimum spanning tree problem. *Discrete Applied Mathematics*, 10(1):47–56, 1985.

[49] P.K. Ghosh and K.V. Kumar. Support function representation of convex bodies, its application in geometric computing, and some related representations. *Computer Vision and Image Understanding*, 72(3):379–403, 1998.

[50] A. Goel and P. Indyk. Stochastic load balancing and related problems. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 579–586. IEEE, 1999.

[51] T.F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[52] V. Goyal and R. Ravi. A ptas for the chance-constrained knapsack problem with random item sizes. *Operations Research Letters*, 38(3):161–164, 2010.

[53] S. Guha and K. Munagala. Exceeding expectations and clustering uncertain data. In *Proceedings of the 28th Symposium on Principles of Database Systems*, pages 269–278, 2009.

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

[55] P. Gupta and P.R. Kumar. Critical power for asymptotic connectivity. In *Proceedings of the 37th IEEE Conference on Decision and Control*, volume 1, pages 1106–1110. IEEE, 1998.

[56] P. Gupta and P.R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, 2000.

[57] M. Haenggi, J.G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE Journal on Selected Areas in Communications*, 27(7):1029–1046, 2009.

[58] S. Har-Peled. *Geometric approximation algorithms*, volume 173. American mathematical society Providence, 2011.

[59] S. Har-Peled. On the expected complexity of random convex hulls. *arXiv:1111.5340*, 2011.

[60] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 291–300, 2004.

[61] S. Har-Peled and K. Varadarajan. Projective clustering in high dimensions using core-sets. In *Proceedings of the eighteenth annual symposium on Computational geometry*, pages 312–318. ACM, 2002.

[62] S. Har-Peled and Y. Wang. Shape fitting with outliers. *SIAM Journal on Computing*, 33(2):269–285, 2004.

[63] D. Hartvigsen. An extension of matching theory. phd thesis, carnegie-mellon university. 1984.

[64] M. Held and J.S.B. Mitchell. Triangulating input-constrained planar point sets. *Information Processing Letters*, 109(1):54–56, 2008.

[65] L. Huang and J. Li. Approximating the expected values for combinatorial optimization problems over stochastic points. In *Automata, Languages, and Programming*, pages 910–921. Springer, 2015.

[66] Lingxiao Huang, Jian Li, Jeff M Phillips, and Haitao Wang. ε-kernel coresets for stochastic points. In *European Symposium on Algorithms*. Springer, 2016.

[67] M. Jerrum, A. Sinclair, and E. Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004.

[68] M. Jerrum, L.G. Valiant, and V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[69] A.G. Jørgensen, M. Löffler, and J.M. Phillips. Geometric computation on indecisive points. In *Proceedings of the 12th Algorithms and Data Structure Symposium*, pages 536–547, 2011.

[70] P. Kamousi, T.M. Chan, and S. Suri. The stochastic closest pair problem and nearest neighbor search. In *Proceedings of the 12th Algorithms and Data Structure Symposium*, pages 548–559, 2011.

[71] P. Kamousi, T.M. Chan, and S. Suri. Stochastic minimum spanning trees in Euclidean spaces. In *Proceedings of the 27th annual ACM symposium on Computational Geometry*, pages 65–74. ACM, 2011.

[72] P. Kamousi, T.M. Chan, and S. Suri. Closest pair and the post office problem for stochastic points. *Computational Geometry*, 47(2):214–223, 2014.

[73] H.J. Karloff. How long can a Euclidean traveling salesman tour be? In *J. Discrete Math*, page 2(1). SIAM, 1989.

[74] J. Kleinberg and T. Eva. *Algorithm design*. Pearson Education India, 2006.

[75] J. Kleinberg, Y. Rabani, and É. Tardos. Allocating bandwidth for bursty connections. *SIAM Journal on Computing*, 30(1):191–217, 2000.

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

[77] M. Langberg and L.J. Schulman. Universal ε-approximators for integrals. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, 2010.

[78] J. Li and A. Deshpande. Maximizing expected utility for stochastic combinatorial optimization problems. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 797–806. IEEE, 2011.

[79] J. Li and Y. Liu. Approximation algorithms for stochastic combinatorial optimization problems. *Journal of the Operations Research Society of China*, 4(1):1–47, 2016.

[80] J. Li and H. Wang. Range queries on uncertain data. *Theoretical Computer Science*, 609:32–48, 2016.

[81] J. Li and W. Yuan. Stochastic combinatorial optimization via poisson approximation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 971–980. ACM, 2013.

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

[83] M. Löffler and J. Phillips. Shape fitting on point sets with probability distributions. In *Proceedings of the 17th European Symposium on Algorithms*, pages 313–324, 2009.

[84] M. Löffler and J. Snoeyink. Delaunay triangulations of imprecise points in linear time after preprocessing. In *Proceedings of the 24th Sympoium on Computational Geometry*, pages 298–304, 2008.

[85] M. Löffler and M. van Kreveld. Approximating largest convex hulls for imprecise points. *Journal of Discrete Algorithms*, 6:583–594, 2008.

[86] R.P. Loui. Optimal paths in graphs with stochastic or multidimensional weights. *Communications of the ACM*, 26(9):670–676, 1983.

[87] J. Matoušek. Computing the center of planar point sets. *Discrete and Computational Geometry*, 6:221, 1991.

[88] A. Munteanu, C. Sohler, and D. Feldman. Smallest enclosing ball for probabilistic data. In *Proceedings of the thirtieth annual symposium on Computational geometry*, page 214. ACM, 2014.

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

[90] E. Nikolova. Approximation algorithms for reliable stochastic combinatorial optimization. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 338–351. Springer, 2010.

[91] E. Nikolova, J.A. Kelner, M. Brand, and M. Mitzenmacher. Stochastic shortest paths via quasi-convex maximization. In *European Symposium on Algorithms*, pages 552–563. Springer, 2006.

[92] Y. Ostrovsky-Berman and L. Joskowicz. Uncertainty envelopes. In *Abstracts of the 21st European Workshop on Comput. Geom.*, pages 175–178, 2005.

[93] J.M. Phillips. Coresets and sketches. In *Handbook of Discrete and Computational Geometry*, number Chapter 49. CRC Press, 3rd edition, 2016.

[94] D. Salesin, J. Stolfi, and L.J. Guibas. Epsilon geometry: building robust algorithms from imprecise computations. In *Proceedings of the 5th Symposium on Computational Geometry*, pages 208–217, 1989.

[95] R. Schneider. *Convex bodies: the Brunn-Minkowski theory*, volume 44. Cambridge University Press, 1993.

[96] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*, volume 16. SIAM, 2014.

[97] A. Sly. Computational transition at the uniqueness threshold. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 287–296. IEEE, 2010.

[98] T.L. Snyder and J.M. Steele. A priori bounds on the Euclidean traveling salesman. In *J. Comput*, page 24(3). SIAM, 1995.

[99] J.M. Steele. On frieze's $\zeta(3)$ limit for lengths of minimal spanning trees. *Discrete Applied Mathematics*, 18(1):99–103, 1987.

[100] D. Suciu, D. Olteanu, C. Ré, and C. Koch. Probabilistic databases. *Synthesis Lectures on Data Management*, 3(2):1–180, 2011.

[101] S. Suri, K. Verbeek, and H. Yıldız. On the most likely convex hull of uncertain points. In *Proceedings of the 21st European Symposium on Algorithms*, pages 791–802, 2013.

[102] C. Swamy and D. B. Shmoys. Approximation algorithms for 2-stage stochastic optimization problems. pages 37(1):33–46, 2006.

[103] W.T. Tutte. A short proof of the factor theorem for finite graphs. *Canad. J. Math.*, 6, 1954.

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

[105] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[106] K. Varadarajan and X. Xiao. On the sensitivity of shape fitting problems. In *32nd International Conference on Foundations of Software Technology and Theoretical Computer Science*, page 486, 2012.

[107] J. Vondrak, C. Chekuri, and R. Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 783–792. ACM, 2011.

[108] H. Wang and J. Zhang. One-dimensional k-center on uncertain data. *Theoretical Computer Science*, 602:114–124, 2015.

[109] H. Yıldız, L. Foschini, J. Hershberger, and S. Suri. The union of probabilistic boxes: Maintaining the volume. *European Symposia on Algorithms*, pages 591–602, 2011.

[110] H. Yu, P.K. Agarwal, R. Poreddy, and K. Varadarajan. Practical methods for shape fitting and kinetic data structures using coresets. *Algorithmica*, 52(3):378–402, 2008.

[111] K. Zheng, G. Trajcevski, X. Zhou, and P. Scheuermann. Probabilistic range queries for uncertain trajectories on road networks. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 283–294, 2011.

Acknowledgements

First, I would like to thank my advisor, Jian Li, for his invaluable help and patient instructions. He always offers me advices of great value and inspiration of new ideas. I enjoyed discussing with him and have learnt a lot from him. It is my honor to be his academic student.

I would like to take this opportunity to thank all of my collaborators: Pingzhong Tang, Yicheng Liu, Lingqing Ai, Xian Wu, Longbo Huang, Qicai Shi, Jeff M. Phillips, Haitao Wang, Pinyan Lu, Chihao Zhang, Hu Ding, Yu Liu, Yifei Jin and Lunjia Hu. I have benefited a lot and improved my research skills while working with them. I am indebted to Professor Pinyan Lu, who was my mentor during my visit to MSRA in Shanghai in the spring of 2015, and Chihao Zhang, who helped me a lot during my internship at MSRA. My PhD life in IIIS was so enjoyable with my dear friends. I would like to thank my friends, Yu, Jianan, Ye, Ruichuang, Linyun, Xian, Lingqing, Yifei, Wei, Mengwen, Dong, Hao, Zhize and Qicai (I have been moving a lot). We have so many colorful shared memory of learning and playing together. I would also like to thank my Dota-mates who won the champion of 9cg in Tsinghua with me: Lingqing, Xian, Linyun, Jianan and Ruichuang.

Finally, I must give my most special thanks to my mother Xifeng and my girl friend Zhongjun, who always accompanied me. Whenever I have met difficulties, they always support me and encourage me. Their selfless love has made this all possible.

Declaration

I solemnly declare that the submissions of the dissertation are the results of my independent research work under the guidance of the instructor. As far as I know, the research results of this dissertation do not contain any content that is copyrighted by others, except as already quoted in the text. Other individuals and collectives who have contributed to the research work involved in this paper have been identified in a clear manner.

Signature：_____Date：_____

Personal Resume, Academic Papers Published During the Study and Research Results

Personal Resume

2008-2012 Bachelor Degree, IIIS, Tsinghua University

2012-2017 PhD Degree, IIIS, Tsinghua University; Advisor: Prof. Jian Li

Research interest: algorithm design (including approximation algorithm, computational geometry and random algorithm), machine learning and game theory

Published Academic Papers

1. AAMAS 2014. Egalitarian Pairwise Kidney Exchange: Fast Algorithms via Linear Programming and Parametric Flow. Authors: Jian Li, Yicheng Liu, Lingxiao Huang, Pingzhong Tang.
2. SIGMETRICS 2014. The Multi-shop Ski Rental Problem. Authors: Lingqing Ai, Xian Wu, Lingxiao Huang, Longbo Huang, Pingzhong Tang, Jian Li
3. COCOON 2015. Approximation Algorithms for the Connected Sensor Cover Problem. Authors: Lingxiao Huang, Jian Li, Qicai Shi.
4. ICALP 2015. Approximating the Expected Values for Combinatorial Optimization Problems over Stochastic Points Authors: Lingxiao Huang, Jian Li.
5. SODA 2016. Canonical Paths for MCMC: from Art to Science Authors: Lingxiao Huang, Pinyan Lu, Chihao Zhang.
6. ICML 2016. K-Means Clustering with Distributed Dimensions Authors: Hu Ding, Yu Liu, Lingxiao Huang, Jian Li.
7. ESA 2016. "-kernal Coresets for Stochastic Points Authors: Lingxiao Huang, Jian Li, Je_ M. Phillips, Haitao Wang.
8. SODA 2017. Stochastic k-Center and j-Flat-Center Problems Authors: Lingxiao Huang, Jian Li.