Approximation Algorithms for Clustering Uncertain Data

Graham Cormode
AT&T Labs–Research
graham@research.att.com

Andrew McGregor
UC San Diego
andrewm@ucsd.edu

ABSTRACT

There is an increasing quantity of data with uncertainty arising from applications such as sensor network measurements, record linkage, and as output of mining algorithms. This uncertainty is typically formalized as probability density functions over tuple values. Beyond storing and processing such data in a DBMS, it is necessary to perform other data analysis tasks such as data mining. We study the core mining problem of *clustering* on uncertain data, and define appropriate natural generalizations of standard clustering optimization criteria. Two variations arise, depending on whether a point is automatically associated with its optimal center, or whether it must be assigned to a fixed cluster no matter where it is actually located.

For uncertain versions of k-means and k-median, we show reductions to their corresponding weighted versions on data with no uncertainties. These are simple in the unassigned case, but require some care for the assigned version. Our most interesting results are for uncertain k-center, which generalizes both traditional k-center and k-median objectives. We show a variety of bicriteria approximation algorithms. One picks $O(k\epsilon^{-1}\log^2 n)$ centers and achieves a $(1+\epsilon)$ approximation to the best uncertain k-centers. Another picks $2k$ centers and achieves a constant factor approximation. Collectively, these results are the first known guaranteed approximation algorithms for the problems of clustering uncertain data.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Clustering*

General Terms

Algorithms

Keywords

Clustering, Probabilistic Data

PODS'08, June 9–12, 2008, Vancouver, BC, Canada.

1. INTRODUCTION

There is a growing awareness of the need for database systems to be able to handle and correctly process data with uncertainty. Conventional systems and query processing tools are built on the assumption of precise values being known for every attribute of every tuple. But any real dataset has missing values, data quality issues, rounded values and items which do not quite fit any of the intended options. Any real world measurements, such as those arising from sensor networks, have inherent uncertainty around the reported value. Other uncertainty can arise from combination of data values, such as record linkage across multiple data sources (e.g., how sure is it that these two addresses refer to the same location?) and from intermediate analysis of the data (e.g., how much confidence is there in a particular derived rule?). Given such motivations, research is ongoing into how to represent, manage, and process data with uncertainty.

Thus far, most focus from the database community has been on problems of understanding the impact of uncertain data within the DBMS, such as how to answer SQL-style queries over uncertain data. Because of interactions between tuples, evaluating even relatively simple queries is $\#P$-hard, and care is needed to analyze which queries can be "safely" evaluated, avoiding such complexity [9]. However, the equally relevant question of *mining* uncertain data has received less attention. Recent work has studied the cost of computing simple aggregates over uncertain data with limited (sublinear) resources, such as average, count distinct, and median [8, 22]. But beyond this, there has been little principled work on the challenge of mining uncertain data, despite the fact that most data to be mined is inherently uncertain.

We focus on the core problem of *clustering*. Adopting the tuple level semantics, the input is a set of points, each of which is described by a compact probability distribution function (pdf). The pdf describes the possible locations of the point; thus traditional clustering can be modeled as an instance of clustering of uncertain data where each input point is at a fixed location with probability 1. The goal of the clustering is to find a set of *cluster centers*, which minimize the expected cost of the clustering. The cost will vary depending on which formulation of the cost function we adopt, described more formally below. The expectation is taken over all "possible-worlds", that is, all possible configurations of the point locations. The probability of a particular configuration is determined from the individual pdfs, under the usual assumption of tuple independence. Note that even if each pdf only has a constant number of discrete possible locations for a point, explicitly evaluating all possible configurations will be exponential in the number of points, and hence highly impractical.

Given a cost metric, any clustering of uncertain data has a well-defined (expected) cost. Thus, even in this probabilistic setting,

Objective	Metric	Assignment	α	β
k-center (point probability)	Any metric	Unassigned	$1+\epsilon$	$O(\epsilon^{-1}\log^2 n)$
	Any metric	Unassigned	$12+\epsilon$	2
k-center (discrete pdf)	Any metric	Unassigned	$1.582+\epsilon$	$O(\epsilon^{-1}\log^2 n)$
	Any metric	Unassigned	$18.99+\epsilon$	2
k-means	Euclidean	Unassigned	$1+\epsilon$	1
	Euclidean	Assigned	$1+\epsilon$	1
k-median	Any metric	Unassigned	$3+\epsilon$	1
	Euclidean	Unassigned	$1+\epsilon$	1
	Any metric	Assigned	$7+\epsilon$	1
	Euclidean	Assigned	$3+\epsilon$	1

Table 1: Our Results for (α, β)-bicriteria approximations

there is a clear notion of the optimal cost, i.e., the minimum cost clustering attainable. Typically, finding such an optimal clustering is NP-hard, even in the non-probabilistic case. Hence we focus on finding *α-approximation algorithms* which promise to find a clustering of size k whose cost is at most α times the cost of the optimal clustering of size k. We also consider (α, β)-bicriteria approximation algorithms, which find a clustering of size βk whose cost is at most α times the cost of the optimal k-clustering. These give strong guarantees on the quality of the clustering relative to the desired cost-objective. It might be hoped that such approximations would follow immediately from simple generalizations of approximation algorithms for corresponding cost metrics in the non-probabilistic regime. Unfortunately, naive approaches fail, and instead more involved solutions are necessary, generating new approximation schemes for uncertain data.

Clustering Uncertain Data and Soft Clustering. 'Soft clustering' (sometimes also known as probabilistic clustering) is a relaxation of clustering which asks for a set of cluster centers and a fractional assignment for each point to the set of centers. The fractional assignments can be interpreted probabilistically as the probability of a point belonging to that cluster. This is especially relevant in model-based clustering, where the output clusters are themselves distributions (e.g., multivariate Gaussians with particular means and variances): here, the assignments are implicit from the descriptions of the clusters as the ratio of the probability density of each cluster at that point. Although both soft clustering and clustering uncertain data can be thought of notions of "probabilistic clustering", they are quite different: soft clustering takes fixed points as input, and outputs appropriate distributions; our problem has probability distributions as inputs but requires fixed points as output. There is no obvious way to use solutions for one problem to solve the other.

Clearly, one can define an appropriate hybrid generalization, where both input and output can be probabilistic. In this setting, methods such as expectation maximization (EM) [10] can naturally be applied to generate model-based clusterings. However, for clarity, we focus on the formulation of the problem where a 'hard' assignment of clusters is required, so do not discuss soft clustering further.

1.1 Our results

In traditional clustering, the three most popular clustering objectives are k-center (to find a set of k cluster centers that minimize the radius of the clusters), k-median (to minimize the sum of distances between points and their closest center) and k-means (to minimize the sum of squares of distances). Each has an uncertain counterparts, where the aim is to find a set of k cluster centers which minimize the *expected* cost of the clustering for k-means, k-median, or k-center objectives. In traditional clustering, the closest center for an input point is well-defined even in the event of ties. But when a single 'point' has multiple possible locations, the meaning is less clear. We consider two variations. In the assigned case, the output additionally assigns a cluster center to each discrete input point. Wherever that point happens to be located, it is always assigned to that center, and the (expected) cost of the clustering is computed accordingly. In the unassigned case, the output is solely the set of cluster centers. Given a particular possible world, each point is assigned to whichever cluster minimizes the distance from its realized location in order to evaluate the cost. Both versions are meaningful: one can imagine clustering data of distributions of consumers' locations to determine where to place k facilities. If the facilities are bank branches, then we may want to assign each customer to a particular branch (so that they can meet personally with their account manager), meaning an assigned solution is needed for branch placement. But customers can also use any ATM, so the unassigned case would apply for ATM placement. We provide results for both assigned and unassigned versions.

- **k-median and k-means.** Due to their linear nature, the unassigned case for k-median and k-means can be efficiently reduced to their non-probabilistic counterparts. But in the assigned version, some more work is needed in order to give required guarantees. In Section 5, we show that uncertain k-means can be directly reduced to weighted deterministic k-means, in both the assigned and unassigned case. We go on to show that uncertain k-median can also be reduced to its deterministic version, but with a constant increase in the approximation factor.

- **k-center.** The uncertain k-center problem is considerably more challenging. Due to the nature of the optimization function, several seemingly intuitive approaches turn out not to be valid. In Section 4, we describe a pair of bicriteria approximation algorithms, for inputs of a particular form one of which achieves a $1+\epsilon$ approximation with a large blow-up in the number of centers, and the other which achieves a constant factor approximation with only $2k$ centers. These apply to general inputs in the unassigned case with a further constant increase in the approximation factor.

- We consider a variety of different models for optimizing the cluster cost in Section 6. Some of these turn out to be provably hard even to approximate, while others yield clusterings which do not seem useful; thus the formulation based on expectation is the most natural of those considered.

Our (α, β)-approximation bounds are summarized in Table 1.

2. RELATED WORK

Clustering data is the topic of much study, and has generated many books devoted solely to the subject. Within database research, various methods have proven popular, such as DBSCAN [13], CURE [16], and BIRCH [28]. Algorithms have been proposed for a variety of clustering problems, such as the k-means, k-median and k-center objectives discussed above. The term 'k-means' is often used casually to refer to Lloyd's algorithm, which provides a heuristic for the k-means objective [25]. It has been shown recently that careful seeding of this heuristic provides a $O(\log n)$ approximation [2]. Other recent work has resulted in algorithms for k-means which guarantee a $1+\epsilon$ approximation, although these are exponential in k [24]. Clustering relative to the k-center or k-median objective is known to be NP-hard [19]. Some of the earliest approximation algorithms were for the k-center problem, and for points in a metric space, the achieved ratio of 2 is the best possible guarantee assuming $P \neq NP$. For k-median, the best known approximation algorithm guarantees a $3+\epsilon$ approximate solution [3], with time cost exponential in ϵ. Approximation algorithms for clustering and its variations continues to be an active area of research in the algorithms community [18, 4, 21]. See the lecture notes by Har-Peled [17, Chapter 4] for a clear introduction.

Uncertain data has recently attracted much interest in the data management community due to increasing awareness of the prevalence of uncertain data. Typically, uncertainty is formalized by providing some compact representation of the possible values of each tuple, in the form of a pdf. Such tuple-level uncertainty models assume that the pdf of each tuple is independent of the others. Prior work has studied the complexity of query evaluation on such data [9], and how to explicitly track the lineage of each tuple over a query to ensure correct results [5]. Query answering over uncertain data remains an active area of study, but our focus is on the complementary area of mining uncertain data, and in particular clustering uncertain data.

The problem of clustering uncertain data has been previously proposed, motivated by the uncertainty of moving objects in spatio-temporal databases [7]. A heuristic provided there was to run Lloyd's algorithm for k-means on the data, using tuple-level probabilities to compute an expected distance from cluster centers (similar to the "fuzzy c-means" algorithm [11]). Subsequent work has studied how to more rapidly compute this distance given pdfs represented by uniform uncertainty within a geometric region (e.g., bounding rectangle or other polygon) [26]. We formalize this concept, and show a precise reduction to weighted k-means in Section 5.1.

Most recently, Aggarwal and Yu [1] proposed an extension of their micro-clustering technique to uncertain data. This tracks the mean and variance of each dimension within each cluster (and so assumes geometric points rather than an arbitrary metric), and uses a heuristic to determine when to create new clusters and remove old ones. This heuristic method is highly dependent on the input order; moreover, this approach has no proven guarantee relative to any of our clustering objectives.

3. PROBLEM DEFINITIONS

In this section, we formalize the definitions of the input and the cost objectives. The input comes in the form of a set of n probability distribution functions (pdfs) describing the locations of n points in a metric space $(\mathcal{X}, d)$. Our results address two cases: when the points are arbitrary locations in d-dimensional Euclidean space, and when the metric is any arbitrary metric. For the latter case, we consider only discrete clusterings, i.e., when the cluster centers are restricted to be among the points identified in the input.

The pdfs specify a set of n random variables $X = \{X_1, \ldots, X_n\}$. The pdf for an individual point, X_i, describes the probability that the i-th point is at any given location $x \in \mathcal{X}$, i.e., $\Pr[X_i = x]$. We mostly consider the case of discrete pdfs, that is, $\Pr[X_i = x]$ is non-zero only for a small number of x's. We assume that

$$\gamma = \min_{i \in [n], x \in \mathcal{X}} \Pr[X_i = x] | \Pr[X_i = x] > 0$$

is only polynomially small. There can be a probability that a point does not appear, which is denoted by $\perp \notin \mathcal{X}$, so that $0 \leq \Pr[X_i = \perp] < 1$. For any point x_i and its associated pdf X_i, define P_i, the probability that the point occurs, as $1 - \Pr[X_i = \perp]$. All our methods apply to the cases where $P_i = 1$, and more generally to $P_i < 1$. The *aspect ratio*, Δ, is defined as the ratio between the greatest distance and smallest distance between pairs of points identified in the input, i.e.,

$$\Delta = \frac{\max_{x,y \in \mathcal{X}, \exists i,j \in [n]: \Pr[X_i = x] \Pr[X_j = y] > 0} d(x,y)}{\min_{x,y \in \mathcal{X}, \exists i,j \in [n]: \Pr[X_i = x] \Pr[X_j = y] > 0} d(x,y)} .$$

A special case of this model is when X_i is of the form

$$\Pr[X_i = x_i] = p_i \quad \text{and} \quad \Pr[X_i = \perp] = 1 - p_i$$

which we refer to as the *point probability* case.

The goal is to produce a (hard) clustering of the points. We consider two variants of clustering, *assigned clustering* and *unassigned clustering*. In both we must specify a set of k points $C = \{c_1, \ldots, c_k\}$ and in the case of assigned clustering we also specify an assignment from each X_i to a point $c \in C$:

Definition 1. Assigned Clustering: Wherever the i-th point falls, it is always assigned to the same cluster. The output of the clustering is a set of k points $C = \{c_1, \ldots, c_k\}$ and a function $\sigma : [n] \to C$ mapping points to clusters.

Unassigned Clustering: Wherever it happens to fall, the i-th point is assigned to the closest cluster center. In this case, the assignment function τ is implicitly defined by the clusters C, as the function that maps from locations to clusters based on Voronoi cells: $\tau : \mathcal{X} \to C$ such that $\tau(x) = \arg\min_{c \in C} d(x, c)$.

We consider three standard objective functions generalized appropriately to the uncertain data setting. The *cost* of a clustering is formalized as follows. To allow a simple statement of the costs for both the assigned and unassigned cases, define a function $\rho : [n] \times \mathcal{X} \to C$ so that $\rho(i, x) = \sigma(i)$ in the assigned case, and $\rho(i, x) = \tau(x)$ in the unassigned case. These are defined based on the indicator function, $I[A]$, which is 1 when the event A occurs, and 0 otherwise.

Definition 2. The k-median cost, $\mathrm{cost}_{\mathrm{med}}$, is the sum of the distances of points to their center:

$$\mathrm{cost}_{\mathrm{med}}(X, C, \rho) = \sum_{i \in [n], x \in \mathcal{X}} I[X_i = x] d(x, \rho(i, x))$$

The k-means cost, $\mathrm{cost}_{\mathrm{mea}}$, is the sum of the squares of distances:

$$\mathrm{cost}_{\mathrm{mea}}(X, C, \rho) = \sum_{i \in [n], x \in \mathcal{X}} I[X_i = x] d^2(x, \rho(i, x))$$

The k-center cost, $\mathrm{cost}_{\mathrm{cen}}$, is the maximum distance from any point to its associated center:

$$\mathrm{cost}_{\mathrm{cen}}(X, C, \rho) = \max_{i \in [n], x \in \mathcal{X}} \{I[X_i = x] d(x, \rho(i, x))\}$$

These costs are random variables, and so we can consider natural statistical properties such as the expectation and variance of the cost, or the probability that the cost exceeds a certain value. Here we set out to minimize the (expected) cost of the clustering. This generates corresponding optimization problems.

Definition 3. The uncertain k-median problem is to find a set of centers C which minimize the expected k-median cost, i.e.,

$$\min_{C:|C|=k} \mathsf{E}[\mathrm{cost}_{\mathrm{med}}(X, C, \rho)] .$$

The uncertain k-means problem is to find k centers C which minimize the expected k-means cost,

$$\min_{C:|C|=k} \mathsf{E}[\mathrm{cost}_{\mathrm{mea}}(X, C, \rho)] .$$

The uncertain k-center problem is to find k centers C which minimize:

$$\min_{C:|C|=k} \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(X, C, \rho)] .$$

These costs implicitly range over all possible assignments of points to locations (possible worlds). Even in the point probability case, naively computing the cost of a particular clustering by enumerating all possible worlds would take time exponential in the input size. However, each cost can be computed efficiently given C and ρ. In the point-probability case, ρ is implicit from C, we may drop it from our notation.

A special case is when all variables X are of the form $\Pr[X_i = x_i] = 1$, i.e., there is no uncertainty, since the i-th point is always at location x_i. Then we have precisely the "traditional" clustering problem on deterministic data, and the above optimization problems correspond precisely to the standard definitions of k-center, k-median and k-means. We refer to these problems as "deterministic k-center" etc., in order to clearly distinguish them from their uncertain counterparts.

In prior work on computing with uncertain data, a general technique is to sample repeatedly from possible worlds, and compute the expected value of the desired function [8, 22]. Such approaches do not work in the case of clustering, however, since the desired output is a single set of clusters. While it is possible to sample multiple possible worlds and compute clusterings for each, it is unclear how to combine these into a single clustering with some provable guarantees, so more tailored methods are required.

4. UNCERTAIN K-CENTER

In this section, we give results for uncertain k-center clustering. This optimization problem turns out to be richer than its certain counterpart, since it encapsulates aspects of both deterministic k-center and deterministic k-median.

4.1 Characteristics of Uncertain k-center

Clearly, uncertain k-center is NP-hard, since it contains deterministic k-center as a special case when $\Pr[X_i = x_i] = 1$ for all i. Further, this shows that it is hard to approximate uncertain k-center over arbitrary metrics to better than a factor of 2. There exist simple greedy approximation algorithms for deterministic k-center which achieve this factor of 2 in the unweighted case or 3 in the weighted case [12]. We show that such natural greedy heuristics from the deterministic case do not carry over to the probabilistic case for k-center.

Example 1. Consider n points distributed as follows:

$$\begin{array}{llll} \Pr[X_1 = y] & = p & \Pr[X_{i>1} = x] & = p/2 \\ \Pr[X_1 = \perp] & = 1 - p & \Pr[X_{i>1} = \perp] & = 1 - p/2 \end{array}$$

where the two locations x and y satisfy $d(x, y) = 1$. Placing 1 center at x has expected cost p. As n grows larger, placing 1 center at y has expected cost tending to 1. Greedy algorithms for unweighted deterministic k-center pick the first center as an arbitrary point from the input, and so could pick y [15]. Greedy algorithms for weighted deterministic k-center consider each point individually, and pick the point which has the highest weight as the first center [19]: in this case, y has the highest individual weight (p instead of $p/2$) and so would be picked as the center. Thus applying algorithms for the deterministic version of the problems can do arbitrarily badly: they fail to achieve an approximation ratio of $1/p$ for any chosen p. The reason is that the metric of expected cost is quite different from the unweighted and weighted versions of deterministic k-center, and so approximations for the latter do not translate into approximations for the former. □

Our next example gives some further insight.

Example 2. Consider n points distributed as follows:

$$\Pr[X_i = x_i] = p_i \qquad \Pr[X_i = \perp] = 1 - p_i$$

where x_i's are some arbitrary set of locations. We can show that if all p_i's are close to 1, the cost is dominated by the point with the greatest distance from its nearest center, i.e., this instance of uncertain k-center is essentially equivalent to deterministic k-center. On the other hand, if all p_i are sufficiently small then the problem is almost equivalent to deterministic k-median. Both statements follow from the next lemma which applies to the point probability case:

LEMMA 1. *For a given set of centers C, let $d_i = d(x_i, C)$. Assume that $d_1 \geq d_2 \geq \ldots \geq d_n$. Then,*

$$\mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(X, C, \rho)] = \sum_i p_i d_i \prod_{j<i} (1 - p_j) . \qquad (1)$$

If $\sum_j p_j \leq \epsilon$ then for all i, $1 \geq \prod_{j<i}(1 - p_j) \geq 1 - \epsilon$ and hence

$$1 - \epsilon \leq \frac{\mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(X, C, \rho)]}{\mathsf{E}[\mathrm{cost}_{\mathrm{med}}(X, C, \rho)]} \leq 1 .$$

Note that from equation (1) it is also clear that if we round all probabilities up to 1 we alter the value of the optimization criterion by at most a $1/\gamma = 1/\min_{i\in[n]} p_i$ factor. But this gives precisely an instance of the deterministic k-center problem, which can be approximated up to a factor 2. So we can easily give a $2/\gamma$ approximation for probabilistic k-center. □

Thus the same probabilistic problem encompasses two quite distinct deterministic problems, both of which are NP-Hard, even to approximate to better than constant factors. More strongly, the same hardness holds even in the case where $P_i = 1$ (so $\Pr[X_i = \perp] = 0$): in the above examples, replace $\perp$ with some point far away from all other points, and allocate $k + 1$ centers instead of 1. Now any near-optimal algorithm in the unassigned case must allocate a center for this far point. This leaves k centers for the remaining problem, whose cost is the same as that in the examples with $\perp$.

Lastly, we show that an intuitive divide-and-conquer approach fails. We might imagine that partitioning the input into ℓ subsets, and finding an α_j approximation on each subset of points, would result in ℓk centers which provide an overall $\max_{j\in[\ell]} \alpha_j$ guarantee. This example shows that this is not the case:

Example 3. Consider the metric space over 4 locations $\{x_1, x_2, c, o\}$ so that:

$$\begin{array}{ll} d(x_1, c) = 4 & d(x_1, o) = 3 \\ d(x_2, c) = 8 & d(x_2, o) = 3 \end{array}$$

The input consists of

$$\begin{array}{ll}\Pr[X_1 = x_1] = 1 & \Pr[X_2 = x_1] = 1\\ \Pr[X_3 = x_2] = \frac{1}{2} & \Pr[X_3 = \perp] = \frac{1}{2}\\ \Pr[X_4 = x_2] = \frac{1}{2} & \Pr[X_4 = \perp] = \frac{1}{2}\end{array}$$

Suppose we partition the input into $\{X_1, X_3\}$ and $\{X_2, X_4\}$. The optimal solution to the induced uncertain 1-center problem is to place a center at o, with cost 3. Our approximation algorithm may decide to place a center at c, which relative to $\{X_1, X_3\}$ is a 2-approximation (and also for $\{X_2, X_4\}$). But on the whole input, placing a center (or rather, two centers) at c has cost 7, and so is no longer a 2-approximation to the optimal cost (placing a center at o still has cost 3 over the full input). Thus approximations on subsets of the input do not translate to approximations on the whole input. □

Instead, one can show the following:

LEMMA 2. *Let X be partitioned into $Y_1 \dots Y_\ell$. For $i \in [\ell]$, let C_i be a set of k points that satisfy,*

$$\mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(Y_i, C_i)] \le \alpha_i \min_{C:|C|=k} \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(X, C)] \ .$$

Then for $\alpha = \sum_{i=1}^{\ell} \alpha_i$,

$$\mathsf{E}\big[\mathrm{cost}_{\mathrm{cen}}(X, \cup_{i\in[\ell]} C_i)\big] \le \alpha \min_{C:|C|=k} \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(X, C)] \ .$$

PROOF. Consider splitting X into just two subsets, Y_1 and Y_2, and finding α_1 approximate centers C_1 for Y_1, and α_2 approximate centers C_2 for Y_2. We can write the cost of using $C_1 \cup C_2$ as

$$\begin{aligned}\mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(X, C_1\cup C_2)] &= \sum_{j\in[t]} \Pr[\mathrm{cost}_{\mathrm{cen}}(Y_1\cup Y_2, C_1\cup C_2) = r_j]\, r_j\\ &\le \sum_{j\in[t]} \Pr[\mathrm{cost}_{\mathrm{cen}}(Y_1, C_1) = r_j]\, r_j\\ &\quad + \sum_{j\in[t]} \Pr[\mathrm{cost}_{\mathrm{cen}}(Y_2, C_2) = r_j]\, r_j\\ &= \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(Y_1, C_1)] + \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(Y_2, C_2)]\\ &\le (\alpha_1 + \alpha_2) \min_{C:|C|=k} \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(X, C)]\end{aligned}$$

This implies the full result by induction. □

Efficient Computation of $\mathrm{cost}_{\mathrm{cen}}$. Lemma 1 implies an efficient way to compute the cost of a given uncertain k-center clustering C against input X in the point probability model: using the same indexing of points, define $X^i = \{X_1 \dots X_i\}$, and so recursively

$$\mathsf{E}\Big[\mathrm{cost}_{\mathrm{cen}}(X^i, C, \rho)\Big] = p_i d_i + (1 - p_i)\mathsf{E}\Big[\mathrm{cost}_{\mathrm{cen}}(X^{i-1}, C, \rho)\Big]$$

$$\text{and } \mathsf{E}\big[\mathrm{cost}_{\mathrm{cen}}(X^0, C, \rho)\big] = 0.$$

In the more general discrete pdf case, for both the assigned and unassigned cases we can form a similar expression, although a more complex form is needed to handle the interactions between points belonging to the same pdf. We omit the straightforward details for brevity. The consequence is that the cost of any proposed k-center clustering can be found in time linear in the input size.

4.2 Cost Rewriting

In subsequent sections, we present bicriteria approximation algorithms for uncertain k-center over point probability distributions, i.e., when all X_i are of the form

$$\Pr[X_i = x_i] = p_i, \qquad \Pr[X_i = \perp] = 1 - p_i.$$

We assume that $\gamma = \min_i \min(p_i, 1 - p_i)$ is only polynomially small. In Section 4.5, we extend our results from the point probability case to arbitrary discrete probability density functions.

Our algorithms begin by rewriting the objective function. Given input X, we consider the set of distances between pairs of points $d(x_i, x_\ell)$. Denote the set of these $t = n(n-1)/2$ distances as $\{r_j\}_{0\le j\le t}$ where $0 = r_0 \le r_1 \le \dots \le r_t$. Then the cost of the clustering with C is

$$\begin{aligned}\mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(X, C)] &= \sum_{j\in[t]} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) = r_j]\, r_j\\ &= \sum_{j\in[t]} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) \ge r_j]\, (r_j - r_{j-1})\end{aligned}$$

Note that $\Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) \ge r]$ is non-increasing as r increases.

4.3 $(1+\epsilon)$ Factor Approximation Algorithm

The following lemma exploits the natural connection between k-center and set-cover, a connection exploited in the optimal asymmetric k-center clustering algorithms [27].

LEMMA 3. *In polynomial time, for any r we can find C' of size at most $ck\log(n)$ (for some constant c) such that*

$$\Pr\big[\mathrm{cost}_{\mathrm{cen}}(X, C') \ge r\big] \le \min_{C:|C|=k} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) \ge r]$$

PROOF. Let

$$C = \mathrm{argmin}_{C:|C|=k} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) \ge r]$$

and $A = \{x_i : d(x_i, C) < r\}$. For each x_i we define a positive weight, $w_i = -\ln(1-p_i)$. It will be convenient to assume the each $p_i < 1$ so that these weights are not infinite. However, because following argument applies if $p_i \le 1 - \epsilon$ for any $\epsilon > 0$, it can be shown that the argument holds in the limit when $\epsilon = 0$. Note that

$$\Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) \ge r] = 1 - \prod_{i:x_i\notin A} (1-p_i) = 1 - \exp(-\sum_{i\notin A} w_i)\,.$$

We will greedily construct a set C' of size at most $k\log(n\gamma^{-1})$ such that $B = \{x_i : d(x_i, C') < r\}$ satisfies,

$$\sum_{i\notin A} w_i \ge \sum_{i\notin B} w_i$$

and therefore, as required,

$$\Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) \ge r] \ge \Pr\big[\mathrm{cost}_{\mathrm{cen}}(X, C') \ge r\big]\,.$$

We construct C' incrementally: at the j-th step let

$$C'_j = \{c_1, \dots, c_j\}, \quad B_j = \{x_i : d(x_i, C'_j) < r\}\,,$$

and define $t_j = \sum_{i:x_i\in B_j} w_i$. We choose c_{j+1} such that t_{j+1} is maximized. Let $w = \sum_{i\in A} w_i$ and let $s_i = w - t_i$. At each step there exists a choice for c_{j+1} such that $t_{j+1} - t_j \ge s_i/k$. This follows because $\sum_{i\in A\setminus B_j} w_i \ge s_i$ and $|C| = k$. Hence $s_i \le w(1 - 1/k)^i$. Therefore, for $i = k\ln(w/w_{\min})$ we have $s_i < w_{\min}$ and hence $s_i \le 0$. Note that $\ln(w/w_{\min}) \le c\ln n$ for some c because $1/(1-p_i)$ and p_i are $\mathrm{poly}(n)$. □

THEOREM 4. *There exists a polynomial time $(1+\epsilon, c\epsilon^{-1}\log n\log\Delta)$ bi-criteria approximation algorithm for uncertain k-center where Δ is the aspect ratio.*

PROOF. We round up all distances to the nearest power of $(1+\epsilon)$ so that there are $t = \lceil \lg_{1+\epsilon}(\Delta)\rceil$ different distances $r_1 < r_2 <$

$\ldots < r_t$. This will cost us the $(1+\epsilon)$ factor in the objective function. Then, for $j \in [t]$, using Lemma 3, we may find a set C_j of $O(k \log n)$ centers such that

$$\Pr[\mathrm{cost}_{\mathrm{cen}}(X, C_j) \geq r_j] \leq \min_{C:|C|=k} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) \geq r_j] .$$

Taking the union $C_0 \cup \ldots \cup C_t$ as the centers gives the required result. □

4.4 Constant Factor Approximation

The following lemma is based on a result for k-center clustering with outliers by Charikar *et al.* [6]. The outlier problem is defined as follows: given a set of n points and an integer o, find a set of centers such that for the smallest possible value of r, all but at most o points are within distance r of a center. Charikar *et al.* present a 3-approximation algorithm for this problem. In fact, we use their approach for a dual problem, where r is fixed and the number of outliers o is allowed to vary. The proof of the following lemma follows by considering the weighted case of the outlier problem where each x_i receives weight $w_i = \ln \frac{1}{1-p_i}$.

LEMMA 5. *In polynomial time, we can find C' of size k such that*

$$\Pr\big[\mathrm{cost}_{\mathrm{cen}}(X, C') \geq 3r\big] \leq \min_{C:|C|=k} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) \geq r] .$$

The above lemma allows us to construct a bi-criteria approximation for k-center with uncertainty such that α and β are both constant.

THEOREM 6. *There exists a polynomial time $(12+\epsilon, 2)$ bi-criteria approximation for uncertain k-center.*

PROOF. Denote the set of all $t = n(n-1)/2$ distances as $\{r_j\}_{1 \leq j \leq t}$ where $r_1 \leq \ldots \leq r_t$ and let $r_0 = 0$ and $r_{t+1} = \infty$. Let C be the optimum size k set of cluster centers. For $j \in [t]$, using Lemma 5 we find a set C_j of k centers such that

$$\Pr[\mathrm{cost}_{\mathrm{cen}}(X, C_j) \geq r_j] \leq \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C) \geq r_j/3]$$

Note that we may assume,

$$\begin{aligned} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C_j) \geq r_j] \geq\ & \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C_j) \geq r_{j+1}] \\ \geq\ & \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C_{j+1}) \geq r_{j+1}] . \end{aligned}$$

Let ℓ be the smallest j such that

$$1 - \kappa \geq \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C_j) \geq r_j]$$

where $\kappa \in (0,1)$ to be determined.

Let $N = \{X_i : d(x_i, C_\ell) \leq r_\ell\}$ be the set of points "near" to C_ℓ and let $F = X \setminus N$ be the set of points "far" from C_ℓ. We consider clustering the near and far points separately.

We first consider the near points. Note,

$$\begin{aligned} & E[\mathrm{cost}_{\mathrm{cen}}(X, C)] \\ & \quad \geq \sum_{j \in [t]} \Pr\left[\frac{1}{3} r_{j+1} > \mathrm{cost}_{\mathrm{cen}}(X, C) \geq \frac{1}{3} r_j\right] \frac{1}{3} r_j \\ & \quad = \sum_{j \in [t]} \Pr\left[\mathrm{cost}_{\mathrm{cen}}(X, C) \geq \frac{1}{3} r_j\right] (\frac{1}{3} r_j - \frac{1}{3} r_{j-1}) \\ & \quad \geq \frac{1}{3} \sum_{j \in [t]} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C_j) \geq r_j] \, (r_j - r_{j-1}) \\ & \quad \geq \frac{1-\kappa}{3} \sum_{j \in [\ell]} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C_\ell) \geq r_j] \, (r_j - r_{j-1}) \\ & \quad = \frac{1-\kappa}{3} \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(N, C_\ell)] \end{aligned}$$

where the second inequality follows because

$$\frac{\Pr[\mathrm{cost}_{\mathrm{cen}}(N, C_j) \geq r_j]}{\Pr[\mathrm{cost}_{\mathrm{cen}}(N, C_\ell) \geq r_j]} \geq 1 - \kappa$$

for $j \leq \ell$. The last inequality follows because

$$\Pr[\mathrm{cost}_{\mathrm{cen}}(N, C_\ell) \geq r_j] = 0$$

for $j \geq \ell + 1$.

We now consider the far points. Let C^* be an $(3+\epsilon)$-approximation to the (weighted) k-median problem on F (i.e., each point x_i has weight p_i). This can be found in polynomial time using the result of Arya *et al.* [3]. Then, by Lemma 1 (note that $\prod_{x_i \in F}(1-p_i) \geq \kappa$),

$$\begin{aligned} \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(F, C^*)] \leq\ & (3+\epsilon)\kappa^{-1} \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(F, C)] \\ \leq\ & (3+\epsilon)\kappa^{-1} \mathsf{E}[\mathrm{cost}_{\mathrm{cen}}(X, C)] . \end{aligned}$$

Appealing to Lemma 2, we deduce that $C^* \cup C_\ell$ are $2k$ centers that achieve a

$$\frac{3+\epsilon}{\kappa} + \frac{3}{1-\kappa}$$

approximation to the objective function. Setting $\kappa = 1/2$ gives the stated result (rescaling ϵ as appropriate.) □

4.5 Extension to General Density Functions

So far we have been considering the point probability case for uncertain k-center, in which there is no difference between the assigned and unassigned versions. In this section, we show that we can use these solutions in the unassigned case for general pdfs, and only lose a constant factor in the objective function. The following lemma makes this concrete:

LEMMA 7. *Let $0 \leq p_{i,j} \leq 1$ satisfy $P_i = \sum_j p_{i,j} \leq 1$. Then*

$$1 - \prod_{i,j}(1-p_{i,j}) \leq 1 - \prod_i (1-P_i) \leq g(P^*) \left(1 - \prod_{i,j}(1-p_{i,j})\right)$$

where $P^* = \max_i P_i$ and $g(x) = x/(1-\exp(-x))$.

PROOF. The first inequality follows by the union bound. To prove the next inequality we first note that $1 - p_{i,j} \leq \exp(-p_{i,j})$ and hence $\prod_{i,j} 1 - p_{i,j} \leq \exp(-\sum_i P_i)$. We now prove that

$$g(P^*)(1 - \exp(-\sum_{i \in [n]} P_i)) \geq (1 - \prod_{i \in [n]} (1 - P_i)) \qquad (2)$$

by induction on n.

$$\begin{aligned} 1 - \prod_{i \in [n]} (1-P_i) &= (1-P_n)(1 - \prod_{i \in [n-1]} (1-P_i)) + P_n \\ &\leq g(P^*)(1-P_n)(1 - e^{-\sum_{i \in [n-1]} P_i}) + P_n \\ &\leq g(P^*)(e^{-P_n}(1 - e^{-\sum_{i \in [n-1]} P_i}) + P_n / g(P^*)) \\ &\leq g(P^*)(e^{-P_n} + P_n/c(P^*) - e^{-\sum_{i \in [n]} P_i}) \\ &\leq g(P^*)(1 - e^{-\sum_{i \in [n]} P_i}) . \end{aligned}$$

This requires $\exp(-P_n) + P_n/g(P^*) \leq 1$ which is satisfied by the definition of $g(P^*)$. For the base case, we derive the same requirement on $g(P^*)$. □

Hence, in the unassigned clustering case with a discrete pdf, we can treat each possible point as a separate independent point pdf, and be accurate upto a $g(1) = e/(e-1)$ factor in the worst case. We then appeal to the results in the previous sections, where now n denotes the total size of the input (i.e. the sum of description sizes of all pdfs).

5. UNCERTAIN K-MEANS AND K-MEDIAN

The definitions of k-means and k-median are based on the sum of costs over all points. This linearity allows the use of linearity of expectation to efficiently compute and optimize the cost of clustering. We outline how to use reductions to appropriately formulated weighted instances of deterministic clustering. In some cases, solutions to the weighted versions are well-known, but we note that weighted clustering can be easily reduced to unweighted clustering with only a polynomial blow-up in the problem size (on the assumption that the ratio between the largest and smallest non-zero weights is polynomial) by replacing each weighted point with an appropriate number of points of unit weight.

5.1 Uncertain k-means

The k-means objective function is defined as the expectation of a linear combination of terms. Consequently,

$$\mathsf{E}[\mathrm{cost}_{\mathrm{mea}}(X, C, \rho)] = \sum_{i\in[n], x\in\mathcal{X}} \Pr[X_i = x] d^2(x, \rho(i, x)). \quad (3)$$

So given a clustering C and ρ, the *cost* of that clustering can be computed quickly, by computing the contribution of each point independently, and summing.

THEOREM 8. *For $\mathcal{X} = (\mathbb{R}^d, \ell_2)$, there exists a randomized, polynomial time $(1+\epsilon)$-approximation for uncertain k-means.*

PROOF. First observe that the result for the unassigned version of uncertain k-means can be immediately reduced to a weighted version of the non-probabilistic problem: by linearity of expectation the cost to be minimized is exactly the same as that of a weighted instance of k-means, where the weight on each point X_i is $\Pr[X_i = x]$. Applying known results for k-means gives the desired accuracy and time bounds [20, 24, 14].

The assigned clustering version of the problem can be reduced to the point probability case where the assigned and unassigned versions are identical. To show this we first define

$$\mu_i = \frac{1}{P_i} \sum_{x\in\mathcal{X}} x \Pr[X_i = x] \text{ and } \sigma_i^2 = \sum_{x\in\mathcal{X}} d(x, \mu_i)^2 \Pr[X_i = x],$$

i.e., μ_i and σ_i^2 are the weighted (vector) mean and the (scalar) variance of X_i, respectively. Then we can rewrite $\mathsf{E}[\mathrm{cost}_{\mathrm{mea}}(X, C, \rho)]$ using properties of Euclidean distances as:

$$\begin{aligned}
\mathsf{E}[\mathrm{cost}_{\mathrm{mea}}(X, C, \rho)] &= \sum_{i\in[n], x\in\mathcal{X}} \Pr[X_i = x] d(x, \rho(i))^2 \\
&= \sum_{i\in[n], x\in\mathcal{X}} \Pr[X_i = x] \Big(d(x, \mu_i)^2 \\
&\qquad + d(\mu_i, \rho(i))^2 + 2d(x, \mu_i) d(\mu_i, \rho(i)) \Big) \\
&= \sum_{i\in[n]} \sigma_i^2 + \sum_{i\in[n], x\in\mathcal{X}} P_i d^2(\mu_i, \rho(i)) \\
&= \sum_{i\in[n]} \sigma_i^2 + \mathsf{E}\big[\mathrm{cost}_{\mathrm{mea}}(X', C, \rho)\big],
\end{aligned}$$

where X' is the input in the point probability case defined by $\Pr[X'_i = \mu_i] = P_i$. Note that $\sum_{i\in[n]} \sigma_i^2$ is a non-negative scalar that does not depend on C and ρ. Hence any α-approximation algorithm for the minimization of $\mathsf{E}[\mathrm{cost}_{\mathrm{mea}}(X', C, \rho)]$ gives an α-approximation for the minimization of $\mathsf{E}[\mathrm{cost}_{\mathrm{mea}}(X, C, \rho)]$. But now, as in the non-assigned case, known results for k-means can be used to find an $(1+\epsilon)$ approximation for the minimization of the instance of weighted deterministic k-means given by X'. □

5.2 Uncertain k-median

Similarly to k-means, linearity of expectation means that the uncertain k-median cost of any proposed clustering can be quickly computed since

$$\mathsf{E}[\mathrm{cost}_{\mathrm{med}}(X, C, \rho)] = \sum_{i\in[n], x\in\mathcal{X}} \Pr[X_i = x] d(x, \rho(i, x))].$$

Uncertain k-median (Unassigned Clustering). Again due to the linearity of the cost metric, solutions to the weighted instance of the k-median problem can be applied to solve the uncertain version for the unassigned case.

THEOREM 9. *For $\mathcal{X} = (\mathbb{R}^d, \ell_2)$, there exists a randomized, polynomial time $(1+\epsilon)$-approximation for uncertain k-median (unassigned). For arbitrary metrics, the approximation factor is $(3+\epsilon)$-approximation.*

The result follows by the reduction to weighted k-median and appealing to bounds for k-median for arbitrary metrics [3] or in the Euclidean case [23].

Uncertain k-median (Assigned Clustering). For k-means assigned case, it was possible to find the mean of each point distribution, and then cluster these means. This succeeds due to the properties of the k-means objective function in Euclidean space. For k-median assigned case, we adopt an approach that is similar in spirit, but whose analysis is more involved. The approach has two stages: first, it finds a near-optimal 1-clustering of each point distribution, and then it clusters the resulting points.

THEOREM 10. *Given an algorithm for weighted k-median clustering with approximation factor α, we can find a $(2\alpha+1)$ approximation for the uncertain k-median problem.*

PROOF. For each X_i, find the discrete 1-median y_i by simply finding the point from the (discrete) distribution X_i which minimizes the cost.

Let $P_i = 1 - \Pr[X_i = \perp]$, as before, and let

$$T = \sum_{i\in[n], x\in\mathcal{X}} \Pr[X_i = x] d(x, y_i),$$

be the expected cost of assigning each point to its discrete 1-median. Define OPT, the optimal cost of an assigned k-median clustering, as

$$OPT = \min_{C,\sigma:|C|=k} \mathsf{E}[\mathrm{cost}_{\mathrm{med}}(X, C, \sigma))].$$

Finally, set $OPT_Y = \min_{C:|C|=k} \mathsf{E}\big[\sum_i P_i d(\rho(i, x), y_i)\big]$, the optimal cost of weighted k-median clustering of the chosen 1-medians $Y = \cup_{i\in[n]}\{y_i\}$.

Now $T \le OPT$ because it is an optimal assigned n-median clustering, whose optimal cost can be no worse than an optimal assigned k-median clustering, i.e.,

$$\min_{C,\sigma:|C|=n} \mathsf{E}[\mathrm{cost}_{\mathrm{med}}(X, C, \sigma)] \le \min_{C,\sigma:|C|=k} \mathsf{E}[\mathrm{cost}_{\mathrm{med}}(X, C, \sigma)].$$

Let $C = \mathrm{argmin}_{C:|C|=k} \mathsf{E}[\mathrm{cost}_{\mathrm{med}}(X, C, \rho)]$ be a set of k medians which achieves the optimal cost and let $\sigma : [n] \to C$ be the allocation of uncertain points to these centers. Then

$$\begin{aligned}
T + OPT &= \sum_{i\in[n], x\in\mathcal{X}} \Pr[X_i = x](d(x, \sigma(i)) + d(x, y_i)) \\
&\ge \sum_{i\in[n]} d(y_i, \sigma(i)) P_i \ge OPT_Y.
\end{aligned}$$

The α-approximation algorithm is used to find a set of k-medians C' for Y. Define $\sigma'(i)$ by $\sigma'(i) = \arg\min_{c \in C} d(c, y_i)$. This bounds the cost of the clustering with centers C as

$$\begin{aligned} &E[\,\mathrm{cost}_{\mathrm{med}}(X, C', \sigma')] \\ &\quad = \sum_{i \in [n], x \in \mathcal{X}} \Pr[X_i = x] d(x, \sigma'(i)) \\ &\quad \leq \sum_{i \in [n], x \in \mathcal{X}} \Pr[X_i = x](d(x, y_i) + d(y_i, \sigma'(i))) \\ &\quad \leq T + \alpha OPT_Y \leq \alpha(T + OPT) + T \leq (2\alpha + 1)OPT\,. \end{aligned}$$

□

6. EXTENSIONS

We consider a variety of related formulations, and show that of these, optimizing to minize the expected cost is the most natural and feasible.

Bounded probability clustering. One possible alternate formulation of uncertain clustering is to require only that the probability of the cost of the clustering exceeding a certain amount is bounded. However, one can easily show that such a formulation does not admit any approximation.

THEOREM 11. *Given a desired cost τ, it NP-hard to approximate to any constant factor*

$$\min_{C:|C|=k} \Pr[\mathrm{cost}_{\mathrm{med}}(X, C, \rho) > \tau]\,,$$
$$\min_{C:|C|=k} \Pr[\mathrm{cost}_{\mathrm{mea}}(X, C, \rho) > \tau]\,,$$
$$\textit{or} \min_{C:|C|=k} \Pr[\mathrm{cost}_{\mathrm{cen}}(X, C, \rho) > \tau]\,.$$

PROOF. The hardness follows from considering deterministic arrangements of points, i.e., each point occurs in exactly one location with certainty. If there is an optimal clustering with cost τ or less, then the probability of exceeding this is 0; otherwise, the probability that the optimal cost exceeds τ is 1. Hence, any algorithm which approximates this probability correctly decides whether there is a k clustering with cost τ. But this is known to be NP-hard for the three deterministic clustering cost functions. □

Consequently, similar optimization goals such as "minimize the expectation subject to $\Pr[\mathrm{cost}_{\mathrm{mea}}(X, C, \rho) > \tau] < \delta$" are also NP-hard.

Minimizing Variance. Rather than minimizing the expectation, it may be desirable to choose a clustering which has minimal variance, i.e., is more reliable. However, we can reduce this to the original problem of minimizing expectation. For example, in the point probability case when each $X_i = x_i$ with probability p_i and is absent otherwise

$$\mathrm{Var}(\mathrm{cost}_{\mathrm{med}}(X, C, \rho)) = \sum_i (p_i - p_i^2) d^2(x_i, \rho(i, x_i))$$

i.e., $\mathsf{E}[\mathrm{cost}_{\mathrm{mea}}(X, C, \rho)]$ where p_i is replaced by $p_i - p_i^2$. The same idea extends to minimizing the sum of p-th powers of the distances, since these too can be rewritten using cumulant relations. Note that if for all $p_i \leq \epsilon$ then $p_i(1 - \epsilon) \leq p_i - p_i^2 \leq p_i$, and so there is little difference between $\mathsf{E}[\mathrm{cost}_{\mathrm{mea}}]$ and $\mathrm{Var}\,(\mathrm{cost}_{\mathrm{med}})$ when the probabilities are small. But a clustering that minimizes $\mathrm{Var}(\mathrm{cost}_{\mathrm{cen}}(X, C, \rho))$ can be a "very bad" clustering in terms of its expected cost when the probabilities are higher, as shown by this example:

Example 4. Consider picking 1-center on two points such that $\Pr[X_1 = x_1] = 1$ and $\Pr[X_2 = x_2] = 0.5$. Then

$$\mathrm{Var}(\mathrm{cost}_{\mathrm{cen}}(X, \{x_2\})) = 0,$$

while

$$\mathrm{Var}(\mathrm{cost}_{\mathrm{cen}}(X, \{x_1\})) = d^2(x_1, x_2)/4.$$

But

$$\mathrm{cost}_{\mathrm{cen}}(X, \{x_2\}) = d(x_1, x_2)$$

while $\mathrm{cost}_{\mathrm{cen}}(X, \{x_1\}) = \frac{1}{2} d(x_1, x_2)$: the zero-variance solution costs twice as much (in expectation) as the higher-variance solution. □

Expected Clustering Cost. Our problem was to produce a clustering that is good over all possible worlds. An alternative is to instead ask, given uncertain data, what is the expected cost of the optimal clustering in each possible world? That is,

- k-median: $\mathsf{E}\big[\min_{C:|C|=k} \mathrm{cost}_{\mathrm{med}}(X, C, \rho)\big]$
- k-means: $\mathsf{E}\big[\min_{C:|C|=k} \mathrm{cost}_{\mathrm{mea}}(X, C, \rho)\big]$
- k-center: $\mathsf{E}\big[\min_{C:|C|=k} \mathrm{cost}_{\mathrm{cen}}(X, C, \rho)\big]$

This gives a (rather conservative) lower bound on the cost of producing a single clustering. The form of this optimization problem is closer to that of previously studied problems [8, 22], and so may be more amenable to similar approaches, such as sampling a number of possible worlds and estimating the cost in each.

Continuous Distributions. We have, thus far, considered discrete input distributions. In some cases, it is also useful to study continuous distributions, such as a multi-variate Gaussian, or an area within which the point is uniformly likely to appear. Some of our techniques naturally and immediately handle certain continuous distributions: following the analysis of uncertain k-means (Section 5.1), we only have to know the location of the mean in the assigned case, which can typically be easily calculated or is a given parameter of the distribution. It remains open to fully extend these results to continuous distributions. In particular, it sometimes requires a complicated integral just to compute the cost of a proposed clustering C under a metric such as k-center, where we need to evaluate

$$\int_0^\infty \Pr[\mathrm{cost}_{\mathrm{cen}}(C, X, \rho) \geq r] r \mathrm{d}r$$

over a potentially arbitrary collection of continuous pdfs. Correctly evaluating such integrals can require careful arguments based on numerical precision and appropriate rounding.

Facility Location. We have focused on the clustering version of problems. However, it is equally feasible to study related problems, in particular formulations such as Facility Location, where instead of a fixed number of clusters, there is a facility cost associated with opening a new center, and a service cost for assigning a point to a center, and the goal is to minimize the overall cost. Formalizing this as a deterministic facility cost and expected service cost is straightforward, and means that this and other variations are open for further study.

Acknowledgments

We thank S. Muthukrishnan and Aaron Archer for some stimulating discussions. We also thank Chandra Chekuri, Bolin Ding, and Nitish Korula for assistance in clarifying proofs.

7. REFERENCES

[1] C. Aggarwal and P. S. Yu. Framework for clustering uncertain data streams. In *IEEE International Conference on Data Engineering*, 2008.

[2] D. Arthur and S. Vassilvitskii. kmeans++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[3] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004.

[4] M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *ACM Symposium on Theory of Computing*, pages 250–257, 2002.

[5] O. Benjelloun, A. D. Sarma, A. Y. Halevy, and J. Widom. Uldbs: Databases with uncertainty and lineage. In *International Conference on Very Large Data Bases*, 2006.

[6] M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 642–651, 2001.

[7] M. Chau, R. Cheng, B. Kao, and J. Ngai. Uncertain data mining: An example in clustering location data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2006.

[8] G. Cormode and M. N. Garofalakis. Sketching probabilistic data streams. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 281–292, 2007.

[9] N. N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *VLDB J.*, 16(4):523–544, 2007.

[10] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[11] J. C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.

[12] M. Dyer and A. Frieze. A simple heuristic for the p-center problem. *Operations Research Letters*, 3:285–288, 1985.

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226, 1996.

[14] D. Feldman, M. Monemizadeh, and C. Sohler. A PTAS for k-means clustering based on weak coresets. In *Symposium on Computational Geometry*, 2007.

[15] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38(2-3):293–306, 1985.

[16] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient clustering algorithm for large databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 73–84, 1998.

[17] S. Har-Peled. Geometric approximation algorithms. `http://valis.cs.uiuc.edu/~sariel/teach/notes/aprx/book.pdf`, 2007.

[18] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *ACM Symposium on Theory of Computing*, pages 291–300, 2004.

[19] D. Hochbaum and D. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184, May 1985.

[20] M. Inaba, N. Katoh, and H. Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering (extended abstract). In *Symposium on Computational Geometry*, pages 332–339, 1994.

[21] P. Indyk. Algorithms for dynamic geometric problems over data streams. In *ACM Symposium on Theory of Computing*, 2004.

[22] T. S. Jayram, A. McGregor, S. Muthukrishnan, and E. Vee. Estimating statistical aggregates on probabilistic data streams. In *ACM Symposium on Principles of Database Systems*, pages 243–252, 2007.

[23] S. G. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. In *Proceedings of European Symposium on Algorithms*, 1999.

[24] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *IEEE Symposium on Foundations of Computer Science*, 2004.

[25] J. B. MacQueen. Some method for the classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Structures*, pages 281–297, 1967.

[26] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip. Efficient clustering of uncertain data. In *IEEE International Conference on Data Mining*, 2006.

[27] R. Panigrahy and S. Vishwanathan. An $O(\log^* n)$ approximation algorithm for the asymmetric p-center problem. *J. Algorithms*, 27(2):259–268, 1998.

[28] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 103–114, 1996.