CGL

Computational Geometric Learning

Smallest enclosing ball for probabilistic data

Dan Feldman

Alexander Munteanu

Christian Sohler

CGL Technical Report No.: 70

Part of deliverable:	WP-I/RTD
Site:	TUD
Month:	36

Project co-funded by the European Commission within FP7 (2010–2013)
under contract nr. IST-25582

Abstract

This technical report deals with the smallest enclosing ball problem subject to probabilistic data. In our setting, any of n points may not or may occur at one of finitely many locations, following its own discrete probability distribution. The objective is therefore considered to be a random variable and we aim at finding a center minimizing the expected maximum distance to the points according to their distributions. Our main contribution presented in this report is to develop the first $(1 + \varepsilon)$-approximation algorithm for the probabilistic smallest enclosing ball problem with extensions to the streaming setting.

1 Introduction

In many real-world applications we are faced with missing or uncertain data. For instance, suppose that our knowledge base consists of data on a set of customers. Some customers may not have filled out all information about themselves. Also, some data might be inferred in an uncertain way. An example would be to consider a customer a vegetarian if he or she never buys meat. We cannot be sure this is correct, but from previous statistics on the subject we may estimate a probability of, say, 0.8 for the customer being vegetarian, 0.15 for being vegan and 0.05 for being neither. Another typical scenario in which we must deal with uncertainty is when the data comes from sensor networks, where different features are measured at different sites and must be linked in order to form the complete data set. In such a situation there might be missing data from some of the sensors for each of the records. In a geometric interpretation the points might be completely missing or occur at different locations following some specified probability distribution. As traditional algorithms for data analysis are not suited to deal probabilistic data, we need to generalize these or develop completely new methods which enable us to efficiently analyze data under presence of uncertainty. One important tool and building block for data analysis is the smallest enclosing ball problem. The problem consists of finding a center such that the maximum distance of that center to any point from a set of given points $P \subset \mathbb{R}^d$ is minimized. The traditional problem has been discussed extensively in the literature dating back to J.J. Sylvester in 1857 [26]. Apart from the fact that the problem itself is interesting from a theoretical point of view in several settings, it can be used as a building block for the development and analysis of many geometrical problems. Examples include proximity problems as, for instance, furthest neighbor search, computing the diameter and clustering problems like 1-center or, more generally k-center and facility location [8]. Also support vector machine training can be seen as smallest enclosing ball problems [28]. Recent work shows the necessity of developing efficient algorithms for the smallest enclosing ball problem [25]. In addition to uncertainty in the gathered data, in recent years we have to cope with huge amounts of data referred to as *Big Data* [4, 19]. In typical scenarios the data is only given in a data stream or on external storage where it is prohibitive or even impossible to read the data twice. Therefore we also want our algorithms to work efficiently in the streaming setting [23] where we are only allowed a constant number of passes over the data and our storage is limited to polylogarithmic size. We investigate the smallest enclosing ball problem for probabilistic data. In this setting, our input consists of n finite discrete distributions of points modeling situations in which a point can appear at different locations including the possibility that a point does not appear at all. Our aim is to find a center c such that the *expected* maximum distance of c to points drawn from the input distributions is minimized. We develop the first $(1 + \varepsilon)$-approximation algorithm for the probabilistic smallest enclosing ball problem and extend it to work in the streaming setting. This is done by reduction to near-metric 1-median problems of large size and applying sampling techniques to approximate the resulting problems using only small space.

1.1 Related work

Several recent works have dealt with clustering problems on probabilistic data. One approach was to generalize well-known heuristic algorithms to the uncertain setting. For example Ngai et al. [24] and Chau et al. [5] extended Lloyd's algorithm for the k-means objective [21]. Another clustering algorithm called DBSCAN [7] was also modified to handle probabilistic data by Kriegel and Pfeifle [17, 18] and Xu and Li [30]. The interested reader is referred to [2] for a survey on data mining of

uncertain data.

From the theoretical perspective there is rather few work on data analysis methods for probabilistic data. Cormode and McGregor [6] introduced the study of *probabilistic clustering problems*. They developed approximation algorithms for the probabilistic settings of k-means, k-median as well as k-center clustering. The smallest enclosing ball problem is also referred to in the literature as the 1-center clustering problem. For this problem their results are $O(1)$-approximation algorithms with a blow-up on the number of centers k which apply to arbitrary metrics. If the probability distributions are restricted to the cases that a point exists or not then they achieve a $(1+\varepsilon)$-approximation but again with a larger number of centers. Guha and Muhagala [9] improved upon the previous work. They achieved $O(1)$-approximations while preserving the number of centers. Lammersen, Schmidt and Sohler [20] developed the first clustering algorithms for uncertain datasets in the streaming setting by giving the first probabilistic coreset constructions.

Furthermore, Löffler and Phillips [22] also with Jørgensen [15] studied geometric shape fitting problems in the probabilistic setting. Their approach is to sample instantiations from the input point distributions and compute ε-kernels on them to define distributions on possible solutions to the shape fitting problems. Note that their work also covers the smallest enclosing ball problem but it differs from our setting in the sense that we are looking for an approximation to the expected optimal solution whereas their aim is to define a distribution on deterministic solutions. Still, from a technical point of view we have in common, that both of our works use sampling of instantiations from the input distributions for the approximations.

As our results are mainly achieved by reductions to deterministic 1-median problems and sampling, we would like to mention several results that cover the useful sampling techniques for metric 1-median. The first contribution in this area is due to Indyk [13] who gave the first sampling based approximate comparison for the 1-median objective. Combining it with a randomized tournament tree construction of [16] he achieved a $(1+\varepsilon)$-approximation by taking $O(n)$ samples of size $O(\frac{1}{\varepsilon^5})$. Later Indyk and Thorup [14] showed by a Chernoff type argument, that one single constant size sample is enough to approximate the discrete metric 1-median. The result was published considerably later by Thorup in [27]. Using a little modification based on their contribution, Ackermann, Blömer and Sohler [1] showed that one can get a $(1+\varepsilon)$-approximation for the non-descrete setting for so called *doubling spaces* by solving the problem exactly on a constant size sample.

1.2 Our contribution

Our main technical result presented in this report is a reduction from the probabilistic smallest enclosing ball problem to related 1-median problems. Although these reductions involve exponential blow-ups in the problem sizes and non-metric spaces, we are able to develop a polynomial time $(1+\varepsilon)$-approximation algorithm for the smallest enclosing ball problem subject to probabilistic data by leveraging the aforementioned sampling results from [1]. Furthermore we show how to extend our algorithm to the streaming setting in only one pass over the data having space complexity and update time per item independent on the number of input distributions. Our result improves upon the polynomial time $O(1)$-approximation algorithms for metric k-center problems from [6, 9] for the special case of the Euclidean space and $k = 1$.

2 Preliminaries

The input to the uncertain smallest enclosing ball problem is a set of n discrete probability distributions. The i-th distribution is defined over m possible locations $q_{i,1}, \ldots, q_{i,m} \in \mathbb{R}^d$. We use the symbol $\perp$ to indicate that a location is not present. With each location a probability $p_{i,j}$ is associated such that $\sum_{j,q_{i,j}\neq\perp} p_{i,j} \leq 1$. The locations together with the probabilities specify the distribution $\mathbf{Pr}[x_i = q_{i,j}] = p_{i,j}$ for any $q_{i,j} \neq \perp$ and $\mathbf{Pr}[x_i = \perp] = 1 - \sum_{j,q_{i,j}\neq\perp} p_{i,j}$. Our only assumption concerning the probabilities $p_{i,j}$ is that they are rational numbers, i.e., $\forall i, j : p_{i,j} \in \mathbb{Q}$. Clearly, this is no unnatural restriction when dealing with finite machine models of computation.

As we will often deal with 1-median problems in our investigations of the uncertain smallest enclosing ball problem, we next define the cost functions for both problems.

Definition 1. *The 1-median cost* $\mathrm{cost}_{\mathrm{MED}}$ *is the sum of the distances for any point to a given center:*

$$\mathrm{cost}_{\mathrm{MED}}(\mathrm{P}, \mathrm{c}) = \sum_{p \in P} d(c, p).$$

The smallest enclosing ball cost $\mathrm{cost}_{\mathrm{SEB}}$ *is the maximum distance of any point to a given center:*

$$\mathrm{cost}_{\mathrm{SEB}}(\mathrm{P}, \mathrm{c}) = \max_{p \in P} d(c, p).$$

Using these definitions we can define the uncertain versions of the 1-median and smallest enclosing ball problems respectively. Note that the expectations are taken over the random choice of the point set P drawn from the n input distributions.

Definition 2. *The uncertain 1-median problem is to find a center c which minimizes the expected 1-median cost, i.e.,*

$$\min_c \mathbf{E}[\mathrm{cost}_{\mathrm{MED}}(P, c)].$$

The uncertain smallest enclosing ball problem is to find a center c which minimizes the expected smallest enclosing ball cost, i.e.,

$$\min_c \mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P, c)].$$

3 The algorithm and intuition for the analysis

The main idea of the algorithm is to distinguish between the case that the overall probability to generate an input point is small and the case that with a reasonable probability at least one input point is generated. In both cases we are able to reduce the uncertain smallest enclosing ball problem to closely related 1-median problems, in which a small sample is sufficient to approximate the best solution.

In the former case, a near-optimal solution is obtained by focussing on the event that a sample from the n distributions contains exactly one point. This, in turn, means that the expected cost can be approximated by a weighted 1-median problem, where each possible location is assigned a weight proportional to its probability.

In the remaining case we can rewrite the cost function to a weighted sum in terms of a near-metric distance measure on instantiations of our n distributions. For any instantiation R, the cost with respect to a fixed center will be the maximum distance of a point of R to the center. We will discretize R to approximate this distance measure using a few points.

Finally, we will argue that in both cases, a near-optimal solution to a sample of constant size will be a good approximation to the best solution. We remark that our algorithm cannot approximate the *cost* of the solution efficiently.

This argumentation will yield our main result which is the following theorem concerning Algorithm 1.

Theorem 3. *Given $0 < \varepsilon, \delta \leq \frac{1}{2}$ and a set $\mathcal{D}$ consisting of n discrete distributions on points from $\mathbb{R}^d$, Algorithm 1 returns a solution to $c \in \mathbb{R}^d$ such that*

$$\mathbf{Pr}\left[\mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P, c)] \leq (1+\varepsilon) \min_{c' \in \mathbb{R}^d} \mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P, c')]\right] \geq 1 - \delta$$

provided it has access to a subroutine that computes a $(1+\varepsilon')$-approximate solution to any near-metric 1-median problem with probability at least $1 - \frac{\delta}{4}$ where near-metric means that the distance measure satisfies non-negativity, symmetry and the triangle inequality.

Algorithm 1: UncertainSmallestEnclosingBall($\mathcal{D}$, ε, δ)

Input : set of n point distributions $\mathcal{D}$,
approximation parameter ε,
failure probability δ

Output: $(1+\varepsilon)$-approximate uncertain smallest enclosing ball of $\mathcal{D}$

$\varepsilon' = \varepsilon\delta/40$;
if $\sum_i \mathbf{Pr}[p_i \neq \perp] \leq \varepsilon$ **then**
  Sample s locations $R_0 = \{q_1, \ldots, q_s\}$ from Q, the set of all $q_{i,j} \neq \perp$ according to the probabilities $p_{i,j} / \sum_{q_{i,j} \neq \perp} p_{i,j}$;
  return a $(1+\varepsilon')$-approximate solution to $\min_x \sum_{k=1}^{s} d(x, q_i)$;
else
  for $k = 1$ **to** s_2 **do**
    $R_k = \emptyset$;
    for $i = 1$ **to** n **do**
      Draw point r_i according to distribution i;
      $R_k = R_k \cup \{r_i\}$;
    Compute D the diameter of R_k;
    Put a grid G with side length $\varepsilon D/\sqrt{d}$ over R_k;
    $S_k = \emptyset$;
    For each nonempty cell in G put a point in S_k that is located in the middle of the cell ;
    Define $m(x, S_k) = \max_{q \in S_k} d(x, q)$;
  return a $(1+\varepsilon')$-approximate solution to $\min_x \sum_{k=1}^{s} m(x, S_k)$;

The proof of Theorem 3 will be carried out in the course of the report dealing with the single steps. Namely, the remainder of this report is organized as follows: In Section 4 we present our reductions of the smallest enclosing ball problem with probabilistic data to related 1-median problems on different distance measures and with a significant blow-up in the problem size. Thereafter, in Section 5 we will discuss how to cope with this blow-up using a constant size sample and how to approximate the involved distance measure using space sublinear in the input size using a simple grid construction. Section 6 is dedicated to the proof of our main Theorem which will follow from the preceding results. In Section 7 we discuss how to modify the algorithm to extend our result to the streaming setting by gathering

all the needed samples for both cases in one single pass and using a slightly more complicated grid construction for which we do not need to know the diameter of the various instantiations of point sets arising in a run of the algorithm. Eventually, we summarize and conclude our report in Section 8.

4 The reductions to 1-median

We first deal with the case in which the overall probability to draw an input point is small, i.e., we assume that $\sum_i \mathbf{Pr}[p_i \neq \perp] \leq \varepsilon$ holds. The first step of the reduction is to relate the uncertain smallest enclosing ball objective to the uncertain version of 1-median. The following lemma establishing the relationship is a slight modification of Lemma 1 in [6].

Lemma 4. *If* $\sum_i \mathbf{Pr}[p_i \neq \perp] \leq \varepsilon$ *then*

$$1 - \varepsilon \leq \frac{\mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P, c)]}{\mathbf{E}[\mathrm{cost}_{\mathrm{MED}}(P, c)]} \leq 1.$$

Proof. For a given center c and every $q_{i,j} \neq \perp$ let $A[c, q_{i,j}]$ be the probabilistic event that there exists no point p with $d(c, p) > d(c, q_{i,j})$. Now by assumption clearly

$$1 \geq \mathbf{Pr}[A[c, q_{i,j}]] \geq \prod_{k \neq i} (1 - \mathbf{Pr}[p_k \neq \perp]) \geq 1 - \sum_{k \neq i} \mathbf{Pr}[p_k \neq \perp] > 1 - \varepsilon$$

holds. Therefore we can conclude

$$\begin{aligned} & \sum_{q_{i,j} \neq \perp} p_{i,j} \cdot d(c, q_{i,j}) \\ \geq & \sum_{q_{i,j} \neq \perp} p_{i,j} \cdot d(c, q_{i,j}) \cdot \mathbf{Pr}[A[c, q_{i,j}]] \\ \geq & (1 - \varepsilon) \sum_{q_{i,j} \neq \perp} p_{i,j} \cdot d(c, q_{i,j}) \end{aligned}$$

which proves the proposition by the definitions of the objectives. □

Using the Lemma 4 we can argue that any near-optimal solution to the uncertain 1-median problem is a near-optimal solution to the uncertain smallest enclosing ball problem.

Corollary 5. *If* $\sum_i \mathbf{Pr}[p_i \neq \perp] \leq \varepsilon$ *then any* $(1 + \varepsilon)$*-approximate solution* $\tilde{c}$ *to* $\mathbf{E}[\mathrm{cost}_{\mathrm{MED}}(P, c_{\mathrm{MED}})]$ *is a* $(1 + 4\varepsilon)$*-approximate solution to* $\mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P, c_{\mathrm{SEB}})]$.

Proof. First note that $1/(1 - \varepsilon) \leq (1 + 2\varepsilon)$. By Lemma 4 we have

$$\begin{aligned} \mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P, \tilde{c})] &\leq \mathbf{E}[\mathrm{cost}_{\mathrm{MED}}(P, \tilde{c})] \\ &\leq (1 + \varepsilon)\mathbf{E}[\mathrm{cost}_{\mathrm{MED}}(P, c_{\mathrm{MED}})] \\ &\leq (1 + \varepsilon)\mathbf{E}[\mathrm{cost}_{\mathrm{MED}}(P, c_{\mathrm{SEB}})] \\ &\leq (1 + \varepsilon)\mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P, c_{\mathrm{SEB}})]/(1 - \varepsilon) \\ &\leq (1 + \varepsilon)(1 + 2\varepsilon)\mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P, c_{\mathrm{SEB}})] \\ &\leq (1 + 4\varepsilon)\mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P, c_{\mathrm{SEB}})] \end{aligned}$$

□

By linearity of expectation it can be seen that any uncertain 1-median instance is equivalent to a weighted instance of the non probabilistic version. That is

$$\mathbf{E}[\mathrm{cost}_{\mathrm{MED}}(P,c)] = \mathbf{E}[\sum_{p\in P} d(c,p_i)] = \sum_i \mathbf{E}[d(c,p_i)] = \sum_{i,j} p_{i,j}\, d(c,q_{i,j}).$$

Finally, the last step of our reduction for the first case is to treat the weighted version as an unweighted non-probabilistic version with repeated elements by multiplying the objective by a properly chosen factor. More formally, remember our assumption $p_{i,j} \in \mathbb{Q}$ and let N be the least common multiple of the denominators of the weights $p_{i,j}$ then multiplying $\sum_{i,j} p_{i,j} d(c,q_{i,j})$ by N results in the 1-median objective $\sum_{i,j} n_{i,j} d(c,q_{i,j})$ where $\forall i,j : n_{i,j} \in \mathbb{N}$. Now, by our derivation above, finding a $(1+\varepsilon)$-approximate solution to the latter objective yields a $(1+\varepsilon)$-approximate solution to the uncertain 1-median problem which, in turn, by Corollary 5, is a $(1+4\varepsilon)$-approximate solution to the uncertain smallest enclosing ball problem that we want to solve. Note that the final modification means a blow-up in the number of terms, which, depending on the probabilities, may be even exponential. We will deal with this issue by applying results on metric 1-median problems which state that solving the problem on a sample of constant size yields a good approximation for the original problem.

Now we turn our attention to the remaining case where we have a reasonable probability that a realization of our uncertain point set, i.e., a sample from our n distributions is non-empty and therefore contains at least one actual location. Formally, for the remainder of the section we assume that $\sum_i \mathbf{Pr}[p_i \neq \perp] > \varepsilon$ holds and, under this assumption, aim to reduce the uncertain smallest enclosing ball problem to solving an unweighted 1-median problem with repeated elements defined on a near-metric space.

The idea behind our reduction becomes clear by rewriting the cost function in the following way:

$$\mathbf{E}[\mathrm{cost}_{\mathrm{SEB}}(P,c)] = \sum_{R\neq\emptyset} \mathbf{Pr}[R] \cdot \max_{p\in R} d(c,p)$$

The summation reaches over all non-empty realizations $R \neq \emptyset$ possible to draw from the n input distributions and $\mathbf{Pr}[R]$ is the probability of drawing realization R. As in the previous case, we are faced with a weighted 1-median problem which can be further modified to a corresponding unweighted problem instance. Note that by our reduction, the number of terms underlies an exponential blow-up. At this point we would like to apply the aforementioned sampling results on metric problem instances to get rid of the dependence on the number of terms. But, in this case, it is not a priori clear whether we still deal with a metric space.

For this sake that for non-empty point sets A, B the distance measure $m(A,B) = \max_{a\in A, b\in B} d(a,b)$ is near-metric, proving a more general result than we will need. Note that m cannot be a proper metric since $m(X,X) > 0$ holds for any non-singleton set X. Still, in our next lemma, we are able to prove all remaining properties of metric spaces to hold.

Lemma 6. *Let (X,d) be a metric space and let $P(X)$ be the family of all non-empty subsets of X. For $A, B \in P(X)$ let $m(A,B) = \max_{a\in A, b\in B} d(a,b)$. Then $(P(X), m)$ satisfies the following properties:*

1. $m(A,B) \geq 0$ *(non-negativity)*
2. $A \neq B \Rightarrow m(A,B) > 0$

3. $m(A, B) = m(B, A)$ *(symmetry)*

4. $m(A, C) \leq m(A, B) + m(B, C)$ *(triangle inequality)*

Proof. The non-negativity and symmetry of m follow from the corresponding metric properties of d.

To see that property 2 holds, let a be an element which is contained in one of the sets but not in the other. W.l.o.g. suppose $a \in A \setminus B$. Let $b \in B$ be arbitrary. It follows that $m(A, B) \geq d(a, b) > 0$ since $a \neq b$ and by definition of m.

It remains to prove the triangle inequality. For this sake, let $a \in A, c \in C$ be two elements attaining the maximum distance $m(A, C)$. Let $b \in B$ be chosen arbitrarily. By the standard triangle inequality and by definition of m it follows that $m(A, C) = d(a, c) \leq d(a, b) + d(b, c) \leq m(A, B) + m(B, C)$. □

As we will also need the reverse triangle inequality to hold, this will be proved in our next corollary:

Corollary 7. *Let $(P(x), m)$ be defined as in Lemma 6. Then m satisfies the reverse triangle inequality, i.e.*

$$\forall A, B, C \in P(X) : |m(A, B) - m(B, C)| \leq m(A, C).$$

Proof. Choose arbitrary $A, B, C \in P(X)$. By the standard triangle inequality and symmetry it holds that

$$\begin{aligned} & m(A, B) && \leq m(A, C) + m(B, C) \\ \wedge \quad & m(B, C) && \leq m(A, C) + m(A, B). \end{aligned}$$

Now subtracting the rightmost terms in both lines yields

$$\begin{aligned} & m(A, B) - m(B, C) && \leq m(A, C) \\ \wedge \quad & m(B, C) - m(A, B) && \leq m(A, C) \end{aligned}$$

or equivalently

$$|m(A, B) - m(B, C)| \leq m(A, C).$$

□

The above results enable us to see our problem as a near-metric 1-median problem equipped with the properties of non-negativity, symmetry and the triangle inequality in its simple and also in its reverse form. Note that many results from the literature concerning metric spaces rely only on these properties and do not necessarily need all metric properties to hold. Also, in our case, we will see that the proven properties are all we need to ensure our sampling result holds. Therefore, at this point, our reduction is completed.

5 Achieving sublinear complexity

In both of the above reductions to (near)-metric 1-median problems we have argued that an exponential blow-up in the number of elements might occur. We will deal with that issue in the next subsection by showing that a constant size sample of the elements is enough to get a solution which is a good approximation to the optimal solution.

In the second case of our reduction, the elements themselves consist of up to linearly many points drawn from the distributions. Therefore reducing the number of elements to a constant amount does not necessarily yield sublinear space complexity. We will deal with this issue in the remainder of the section.

5.1 The sampling result

In the previous section we have performed a reduction of the uncertain smallest enclosing ball problem to two different types of the 1-median problem. In both cases we have shown that the underlying space equipped with the distance measures involved satisfy certain metric properties.

In the following we want to leverage these properties to derive a general sampling result which will establish that we can $(1+\varepsilon)$-approximate the uncertain smallest enclosing ball problem with space and time complexity independent of n, the number of input distributions.

For this we will need a known result from [1] which itself is based on a result from [14] published in the appendix of [27]. Lemma 8 formalizes that with high probability any point that is far from being optimal to the original problem is also far from being optimal in a subsampled problem. We will leverage the result in its contrapositive form to conclude that a near optimal solution to the subsampled problem is also near-optimal for the original problem.

In the original paper by Ackermann et al. [1] X is any finite metric space equipped with some metric distance measure d and P is an arbitrary multi-subset of X. Furthermore for the proofs only the properties of symmetry and the triangle inequality are actually needed. Therefore the lemma carries over to our near-metric m, defined in Lemma 6.

Lemma 8 (Lemma 3.3 from [1])**.** *Let $\varepsilon \leq 1$ and $b \in \mathbb{R}^d$ be an arbitrary point with $\operatorname{cost}_{\mathrm{MED}}(P,b) > (1+\frac{4}{5}\varepsilon)\operatorname{cost}_{\mathrm{MED}}(P,c)$. A uniform sample multiset $S \subset P$ of size s satisfies*

$$\mathbf{Pr}\left[\operatorname{cost}_{\mathrm{MED}}(S,b) \leq \operatorname{cost}_{\mathrm{MED}}(S,c) + \frac{\varepsilon s}{5n}\operatorname{cost}_{\mathrm{MED}}(P,c)\right] \leq \exp\left(-\frac{\varepsilon^2 s}{144}\right).$$

Based on this result we are able to prove the following lemma which establishes that our algorithm returns a $(1+\varepsilon)$-approximate solution to the uncertain smallest enclosing ball problem given an arbitrary algorithm, which returns a near-optimal solution to the related problem based only a constant size sample. The proof closely follows the argumentation from Lemma 3.4 from [1].

Lemma 9. *Let $c_S = \operatorname{argmin}_{x\in X}\operatorname{cost}_{\mathrm{MED}}(S,x)$ and let $\tilde{c}_S$ be an $(1+\frac{\varepsilon\delta}{40})$-approximate solution, i.e., $\operatorname{cost}_{\mathrm{MED}}(S,\tilde{c}_S) \leq (1+\frac{\varepsilon\delta}{40})\operatorname{cost}_{\mathrm{MED}}(S,c_S)$. Then for every $\delta > 0$ there exists a constant λ_δ such that every uniform sample multiset $S \subseteq P$ of size $s \geq \lambda_\delta \frac{d}{\varepsilon^2}\log\frac{1}{\varepsilon}$ satisfies*

$$\mathbf{Pr}\left[\operatorname{cost}_{\mathrm{MED}}(P,\tilde{c}_S) \leq (1+\varepsilon)\operatorname{cost}_{\mathrm{MED}}(P,c)\right] \geq 1 - \frac{3\delta}{4}.$$

Proof. Let $r = \frac{12s}{\delta n}\operatorname{cost}_{\mathrm{MED}}(P,c)$ and let $B_1(c,r) \subset \mathbb{R}^d$ be the ball with radius r centered at c. Similarly for $r' = \frac{r}{3}$ define $B_2(c,r')$. Since r is chosen to be a multiple of the expected distance of a randomly chosen point, we can use Markov's inequality and the union bound to infer that $S \subseteq B_2$ holds with probability at least $1-\frac{\delta}{4}$. See [1, Lemma 3.2] for details. Conditioned on this event, for every $x \in \mathbb{R}^d$ B we get $\operatorname{cost}_{\mathrm{MED}}(S,x) \geq 2r's$ but we also know that $\operatorname{cost}_{\mathrm{MED}}(S,c) \leq r's$. Therefore any $(1+\varepsilon)$-approximate solution with respect to S is contained in B_1. Particularly this holds for $\tilde{c}_S$ since $\frac{\varepsilon\delta}{40} < \varepsilon$.

It is a well known fact (see e.g. [3, 10] or [12, Ch. 10]) that any ball $B(\sigma,\rho) \subseteq \mathbb{R}^d$ can be covered by at most $2^{O(d)}$ balls of radius $\frac{\rho}{2}$. From this we can deduce that our ball B can be covered by at most $2^{jO(d)}$ balls of radius $\frac{\rho}{2^j}$ for any $j \in \mathbb{N}$ by applying the construction recursively. We choose $j = \lceil \log\frac{120s}{\varepsilon\delta}\rceil$. By our argumentation above, this yields a set of $l \leq (\frac{240m}{\varepsilon\delta})^{O(d)}$ balls covering B_1, each having radius

$\frac{\varepsilon}{10n}\operatorname{cost}_{\mathrm{MED}}(P,c)$. Let $C=\{c_1,\ldots,c_l\}$ be the set of their centers. Furthermore define $C_{\mathrm{bad}}=\{b\in C \mid \operatorname{cost}_{\mathrm{MED}}(P,b)>(1+\frac{4}{5}\varepsilon)\operatorname{cost}_{\mathrm{MED}}(P,c)\}$.

Now, by the union bound and using Lemma 8 it follows that for any δ there exists a constant λ_δ such that for any $s\geq\lambda_\delta\frac{d}{\varepsilon^2}\log\frac{1}{\varepsilon}$ we have

$$\mathbf{Pr}\left[\operatorname{cost}_{\mathrm{MED}}(S,b)\leq\operatorname{cost}_{\mathrm{MED}}(S,c)+\frac{\varepsilon s}{5n}\operatorname{cost}_{\mathrm{MED}}(P,c)\right] \leq\left(\frac{240m}{\varepsilon\delta}\right)^{O(d)}\exp\left(-\frac{\varepsilon^2 s}{144}\right)\leq\frac{\delta}{4}.$$

Thus, we conclude with probability at least $1-\frac{\delta}{4}$, that

$$\operatorname{cost}_{\mathrm{MED}}(S,b)>\operatorname{cost}_{\mathrm{MED}}(S,c)+\frac{\varepsilon s}{5n}\operatorname{cost}_{\mathrm{MED}}(P,c)$$

holds for every $b\in C_{\mathrm{bad}}$.

Remember our $(1+\frac{\varepsilon\delta}{40})$-approximate solution $\tilde{c}_S$ of S. Let $q\in C$ be the closest point from C to $\tilde{c}_S$. From $\tilde{c}_S\in B_1$ we know that $d(q,\tilde{c}_S)\leq\frac{\varepsilon}{10n}\operatorname{cost}_{\mathrm{MED}}(P,c)$. Moreover, by the triangle inequality and the (near)-optimal properties of $\tilde{c}_S$ and c_S respectively it follows that

$$\begin{aligned}\operatorname{cost}_{\mathrm{MED}}(S,q) &\leq \operatorname{cost}_{\mathrm{MED}}(S,\tilde{c}_S)+\frac{\varepsilon s}{10n}\operatorname{cost}_{\mathrm{MED}}(P,c)\\ &\leq \operatorname{cost}_{\mathrm{MED}}(S,c_S)+\frac{\varepsilon\delta}{40}\operatorname{cost}_{\mathrm{MED}}(S,c_S)+\frac{\varepsilon s}{10n}\operatorname{cost}_{\mathrm{MED}}(P,c)\\ &\leq \operatorname{cost}_{\mathrm{MED}}(S,c)+\frac{\varepsilon\delta}{40}\operatorname{cost}_{\mathrm{MED}}(S,c)+\frac{\varepsilon s}{10n}\operatorname{cost}_{\mathrm{MED}}(P,c)\end{aligned}$$

Note that $\mathbf{E}[\operatorname{cost}_{\mathrm{MED}}(S,c)]=\frac{s}{n}\operatorname{cost}_{\mathrm{MED}}(P,c)$ and therefore, again using Markov's inequality, we see that, with probability at least $1-\frac{\delta}{4}$, we can bound $\operatorname{cost}_{\mathrm{MED}}(S,c)$ by $\frac{4s}{\delta n}\operatorname{cost}_{\mathrm{MED}}(P,c)$ from above. Assuming this bound to hold, we can continue our derivation

$$\begin{aligned}\operatorname{cost}_{\mathrm{MED}}(S,q) &\leq \operatorname{cost}_{\mathrm{MED}}(S,c)+\frac{\varepsilon\delta}{40}\operatorname{cost}_{\mathrm{MED}}(S,c)+\frac{\varepsilon s}{10n}\operatorname{cost}_{\mathrm{MED}}(P,c)\\ &\leq \operatorname{cost}_{\mathrm{MED}}(S,c)+2\cdot\frac{\varepsilon s}{10n}\operatorname{cost}_{\mathrm{MED}}(P,c)\\ &= \operatorname{cost}_{\mathrm{MED}}(S,c)+\frac{\varepsilon s}{5n}\operatorname{cost}_{\mathrm{MED}}(P,c)\end{aligned}$$

to deduce that, again with probability at least $1-\frac{\delta}{4}$, we have $\operatorname{cost}_{\mathrm{MED}}(S,q)<\operatorname{cost}_{\mathrm{MED}}(S,q)$ for every $b\in C_{\mathrm{bad}}$. This means that $q\notin C_{\mathrm{bad}}$ an therefore we have $\operatorname{cost}_{\mathrm{MED}}(P,q)\leq(1+\frac{4}{5}\varepsilon)\operatorname{cost}_{\mathrm{MED}}(P,c)$. Finally, again leveraging the triangle inequality we can conclude

$$\begin{aligned}\operatorname{cost}_{\mathrm{MED}}(P,\tilde{c}_S) &\leq \operatorname{cost}_{\mathrm{MED}}(P,q)+nd(q,\tilde{c}_S)\\ &\leq \left(1+\frac{4}{5}\varepsilon\right)\operatorname{cost}_{\mathrm{MED}}(P,c)+\frac{\varepsilon}{10}\operatorname{cost}_{\mathrm{MED}}(P,c)\\ &\leq (1+\varepsilon)\operatorname{cost}_{\mathrm{MED}}(P,c).\end{aligned}$$

Note, that all assumed events occur with probability $(1-\frac{\delta}{4})^3>1-\frac{3\delta}{4}$ which yields the proposition. □

5.2 Discretizing the instantiations

Note that an instantiation denoted by R may have linear size. In order to get rid of the dependence on n, we discretize R by covering the points from R using a

grid of side length $\varepsilon D/\sqrt{d}$ where D is the diameter of R. In the following we show that evaluating the distance measure m on R and on its discretization S changes the value only by a $(1 \pm \varepsilon)$-factor. This, in turn, implies that an optimal solution regarding the discretized instantiations S is a near-optimal solution to the same objective when evaluating with respect to R.

Lemma 10. *Let $R \subset \mathbb{R}^d$ be some set of points with diameter D. Let G be a grid of side length $\varepsilon D/\sqrt{d}$ which covers all points of R. For every nonempty cell in G, let S contain a point that is located in the middle of the cell. Then*

$$\forall p \in \mathbb{R}^d : (1-\varepsilon)m(p,R) \leq m(p,S) \leq (1+\varepsilon)m(p,R).$$

Proof. First note that by construction, for every point $x \in R$ there exists a point $y \in S$ at distance at most $d(x,y) \leq \varepsilon D/2$ (located in the same cell) and vice versa. Now fix an arbitrary $p \in \mathbb{R}^d$. From the triangle inequality it follows that $m(p,S) = d(p,y) \leq d(p,x) + d(x,y) \leq m(p,R) + \varepsilon D/2 \leq m(p,R) + \varepsilon m(p,R)$. To see that the last inequality holds, choose $x_1, x_2 \in R$ attaining $d(x_1,x_2) = D$. Then from the triangle inequality and the pigeonhole principle we know that

$$D = d(x_1,x_2) \leq 2\max\{d(p,x_1), d(p,x_2)\} \leq 2m(p,R).$$

The lower bound follows from $m(p,R) \leq m(p,S) + \varepsilon m(p,R)$ which can be derived in a similar way. □

Corollary 11. *Let opt minimize $C(p) = \sum_i m(p,R_i)$, and let p^* minimize $S(p) = \sum_i m(p,S_i)$ where $R_i \subset \mathbb{R}^d$ and S_i are their corresponding discretizations as defined in Lemma 10. Then $C(p^*) \leq (1+\varepsilon)C(opt)$.*

Proof. First note that $1/(1-\varepsilon) \leq (1+2\varepsilon)$. Therefore, by applying Lemma 10 termwise it holds that $C(p^*) \leq (1+2\varepsilon)S(p^*)$. Furthermore $S(p^*) \leq S(opt)$ since p^* minimizes S. Another application of Lemma 10 yields $S(opt) \leq (1+\varepsilon)C(opt)$. Now putting all together results in

$$C(p^*) \leq (1+2\varepsilon)(1+\varepsilon)C(opt) \leq (1+4\varepsilon)C(opt)$$

which concludes the proof. □

6 Proof of Theorem 3

In the previous sections we have gathered all the results which we need to combine in order to prove Theorem 3.

Proof of Theorem 3. Clearly all steps of the algorithm can be performed in polynomial time. So we proceed with proving the correctness of our algorithm.

We begin with the case $\sum_i \mathbf{Pr}[p_i \neq \perp] \leq \varepsilon$: We know by Corollary 5 that we can turn our attention to $(1+\varepsilon)$-approximating the uncertain 1-median solution while loosing only a factor of $1+4\varepsilon$. Our further reduction to a deterministic input version of the 1-median problem brings no additional loss and enables us to apply the sampling result formalized in Lemma 9. Since by scaling the objective, sampling proportional to the probabilities corresponds to uniform sampling from the expanded sum, the assumptins of Lemma 9 are fulfilled. Now using an $(1+\varepsilon')$-approximation to the subsampled problem obtained with probability at least $1 - \frac{\delta}{4}$ as assumed in our proposition, we can conclude that with probability at least $(1 - \frac{\delta}{4})(1 - \frac{3\delta}{4}) > 1 - \delta$ the solution is a $1+\varepsilon$-approximation to the 1-median problem. Thus, by Corollary 5, we can conclude that we have a $(1+4\varepsilon)$-approximation to the original probabilistic smallest enclosing ball problem.

It remains to deal with the case $\sum_i \mathbf{Pr}[p_i \neq \perp] > \varepsilon$: In this case the reduction involves a non-metric distance measure, but in Lemma 6 and Corollary 7 we have shown that all metric properties needed in our subsequent reasoning from Section 5, namely non-negativity, symmetry and the triangle inequality hold. In particular, after the reduction to the near-metric certain version of unweighted 1-median, we can again apply Lemma 9. Note that the sets R_k are uniform samples from the set of all possible non-empty realizations of the uncertain point set. Also, we know from Corollary 11 that by discretizing the sets to keep them small, we loose only a factor of $(1 + 4\varepsilon)$. Therefore we can conclude that by our assumptions we end up with an approximation factor of at most $(1 + \varepsilon)(1 + 4\varepsilon) \leq 1 + 7\varepsilon$ with probability at least $1 - \delta$.

Rescaling ε and δ concludes the proof. □

7 Extensions to the streaming setting

In the streaming setting, our space requirements are limited to be of order at most $O(\text{polylog}\, n)$ and we are allowed only $O(1)$ passes over the input data. In order to translate our algorithm into a single-pass algorithm with a space bound even independent of n (though exponential in d), note that by Lemma 9 in both of the cases in which our algorithm operates, we only need a constant size sample of the elements in order to get a good approximation. In the first case we need to sample $s = O(\frac{d}{\varepsilon^2} \log \frac{1}{\varepsilon})$ of the locations $q_{i,j} \neq \perp$ proportional to their probabilities $p_{i,j}$ with repetition which can be done by running s independent copies of the well-known reservoir sampling approach [29]. At the same time we also sample everything we need for the second case. That is, we sample s times independently from the n distributions one by one as they arrive in the data stream. By our reasoning from Section 5.2, for every independent copy we can store a core set of size $O((\frac{\sqrt{d}}{\varepsilon})^d)$, which allows us to efficiently approximate the distance measure defined on the original instantiations. All in all we can compute and store all information that we need in one single pass over the data using $O(sd + s(\frac{\sqrt{d}}{\varepsilon})^d) = O((\frac{\sqrt{d}}{\varepsilon})^d \frac{d^2}{\varepsilon^2} \log \frac{1}{\varepsilon})$ space.

The drawback in our construction is that, while streaming the input in one pass, the diameter of a realization may grow quite often and therefore may require many re-computations and re-insertions of the discretized point set. We can get around this problem by replacing the grid construction by a concentric exponential grid which we can maintain efficiently in the streaming setting. We note that our construction is quite similar to the one used in [11] for coresets to the k-median problem:

The first point will be the center of the grid. When the second point arrives, we put a d-dimensional cube around the first, i.e., covering both points. Let the side length of the cube be l. Then we subdivide it into equal cells of side length $\frac{\varepsilon l}{4\sqrt{d}}$. Whenever a point is inserted outside the current range of the grid, we double its side length until the new point is covered. This results in concentric *levels* L_i of side length $l2^i$ where we subdivide the space $L_i \setminus L_{i-1}$ into cells of side length $\frac{\varepsilon l 2^i}{4\sqrt{d}}$. Thus, the cells also become coarser and coarser with increasing distance from the center.

Such a grid may have $O(\log D)$ levels since, at the end of the stream, our grid construction is supposed to cover all the sampled input points. To reduce this factor, remember the proof of Lemma 10 and note that for any point x and its counterpart in the grid y we are interested in a bound on $d(x, y)$. Since our cells

become coarser and coarser, after at most $j = i + \lceil \log \frac{4\sqrt{d}}{\varepsilon} \rceil$ levels it holds that

$$\frac{\varepsilon l 2^{i+\left\lceil \log \frac{4\sqrt{d}}{\varepsilon} \right\rceil}}{4\sqrt{d}} \geq l2^i,$$

i.e., the coarsest cells are at least as large as the whole grid at the i-th level. Thus, we can collapse everything up to the i-th level and store only the center point as a representative and, more importantly, we only need to store $O(\log \frac{d}{\varepsilon})$ levels instead of $O(\log D)$.

As a consequence of the alternative construction, our space requirements are little larger than before but the insertion of a point x can be done in time $O(d)$ since deciding at which level we need to insert can be achieved by computing $||x||_\infty$ and deciding into which actual cell the point is inserted can be done by inspecting every coordinate one by one. The actual insertion can be achieved in time $O(1)$ using arrays.

Now, to see that Lemma 10 can be adapted to hold for the new grid construction, remember the definitions from its proof. Let $\ell = l2^i$ (for some i) be the side length of the final cube covering all input points. Then there must exist a point at distance more than $\frac{\ell}{4}$ from the center of our grid. The latter is also an input point by construction. Therefore we can deduce $D > \frac{s}{4}$ and consequently $d(x, y) \leq \frac{\varepsilon \ell}{8} < \frac{\varepsilon D}{2}$ which means that our reasoning from Section 5.2 still applies. Consequently we can conclude the following corollary.

Corollary 12. *For any $0 < \varepsilon, \delta \leq \frac{1}{2}$ there exists a single-pass streaming algorithm using $O((\frac{\sqrt{d}}{\varepsilon})^d \frac{d^2}{\varepsilon^2} \log^2 \frac{d}{\varepsilon})$ space and update time $O(d)$ per item returning a $(1+\varepsilon)$-approximation to the optimal solution for the probabilistic smallest enclosing ball problem with probability at least $1 - \delta$ provided it has access to a subroutine that computes a $(1+\varepsilon')$-approximate solution to any near-metric 1-median problem with probability at least $1 - \frac{\delta}{4}$ where near-metric means that the distance measure satisfies non-negativity, symmetry and the triangle inequality.*

Proof. Using the modifications to Algorithm 1 discussed above, the proposition follows by reasoning similarly as in the proof of Theorem 3 for the non-streaming setting. □

8 Conclusion

In this technical report we have dealt with the smallest enclosing ball problem in the presence of uncertain or, more precisely, probabilistic data, where every input point follows its own probability distribution on several locations where a location may also indicate that the point is not present at all. We have developed a randomized polynomial time $(1+\varepsilon)$-approximation algorithm and have discussed how to extend it to work efficiently in the streaming setting, making it work in a single pass over the data. Our main technical contribution is a reduction to two different 1-median problems, one of which is defined on a more complex near-metric space. A grid construction from the theory of coresets has been applied to approximate the underlying distance measure and known sampling results for metric 1-median have been leveraged to cope with significant blow-ups in the problem sizes due to the reductions. Our result improves upon the polynomial time $O(1)$-approximation algorithms for metric k-center problems from [6, 9] for the special case of the Euclidean space and $k = 1$.

References

[1] M. R. Ackermann, J. Blömer, and C. Sohler. Clustering for metric and non-metric distance measures. *ACM Transactions on Algorithms*, 6(4):59:1–59:26, 2010.

[2] C. C. Aggarwal and P. S. Yu. A survey of uncertain data algorithms and applications. *IEEE Trans. Knowl. Data Eng.*, 21(5):609–623, 2009.

[3] P. Assouad. Plongements lipschitziens dans $\mathbb{R}^n$. *Bull. Soc. Math. France*, 111(4):429–448, 1983.

[4] M. Beyer. Gartner says solving 'big data' challenge involves more than just managing volumes of data. Gartner, Retrieved October 30th, 2013. http://www.gartner.com/newsroom/id/1731916.

[5] M. Chau, R. Cheng, B. Kao, and J. Ng. Uncertain data mining: An example in clustering location data. In *Proceedings of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 199–204, 2006.

[6] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In *Proceedings of the 27th Symposium on Principles of Database Systems (PODS)*, pages 191–200, 2008.

[7] 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 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996.

[8] A. Goel, P. Indyk, and K. R. Varadarajan. Reductions among high dimensional proximity problems. In *Proceedings of the 12th Annual Symposium on Discrete Algorithms (SODA)*, pages 769–778, 2001.

[9] S. Guha and K. Munagala. Exceeding expectations and clustering uncertain data. In *Proceedings of the 28th Symposium on Principles of Database Systems (PODS)*, pages 269–278, 2009.

[10] A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *Proceedings of the 44th Symposium on Foundations of Computer Science (FOCS)*, pages 534–543, 2003.

[11] 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 (STOC)*, pages 291–300, 2004.

[12] J. Heinonen. *Lectures on analysis on metric spaces*. Universitext, Springer, New York, 2001.

[13] P. Indyk. Sublinear time algorithms for metric space problems. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing (STOC)*, pages 428–434, 1999.

[14] P. Indyk and M. Thorup. Approximate 1-medians. Unpublished manuscript, 2000.

[15] A. Jørgensen, M. Löffler, and J. M. Phillips. Geometric computations on indecisive and uncertain points. *CoRR*, abs/1205.0273, 2012.

[16] J. M. Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC)*, pages 599–608, 1997.

[17] H.-P. Kriegel and M. Pfeifle. Density-based clustering of uncertain data. In *Proceedings of the 11th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 672–677, 2005.

[18] H.-P. Kriegel and M. Pfeifle. Hierarchical density-based clustering of uncertain data. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM)*, pages 689–692, 2005.

[19] A. Labrinidis and H. V. Jagadish. Challenges and opportunities with big data. *PVLDB*, 5(12):2032–2033, 2012.

[20] C. Lammersen, M. Schmidt, and C. Sohler. Probabilistic k-median clustering in data streams. In *10th International Workshop on Approximation and Online Algorithms (WAOA)*, pages 70–81, 2012.

[21] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–136, 1982.

[22] M. Löffler and J. M. Phillips. Shape fitting on point sets with probability distributions. In *Proceedings of the 17th Annual European Symposium on Algorithms (ESA)*, pages 313–324, 2009.

[23] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.

[24] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip. Efficient clustering of uncertain data. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)*, pages 436–445, 2006.

[25] P. Rai, H. D. III, and S. Venkatasubramanian. Streamed learning: One-pass svms. *CoRR*, abs/0908.0572, 2009.

[26] J. Sylvester. A question in the geometry of situation. *Quarterly Journal of Mathematics*, 79(1), 1857.

[27] M. Thorup. Quick k-median, k-center, and facility location for sparse graphs. *SIAM J. Comput.*, 34(2):405–432, 2005.

[28] I. W. Tsang, J. T. Kwok, and P.-M. Cheung. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

[29] J. S. Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.

[30] H. Xu and G. Li. Density-based probabilistic clustering of uncertain data. In *International Conference on Computer Science and Software Engineering (CSSE)*, pages 474–477, 2008.