On Margin-Based Cluster Recovery with Oracle Queries

Marco Bressan
Dept. of CS, Univ. of Milan, Italy
marco.bressan@unimi.it

Nicolò Cesa-Bianchi
DSRC & Dept. of CS, Univ. of Milan, Italy
nicolo.cesa-bianchi@unimi.it

Silvio Lattanzi
Google
silviol@google.com

Andrea Paudice
Dept. of CS, Univ. of Milan, Italy &
Istituto Italiano di Tecnologia, Italy
andrea.paudice@unimi.it

Abstract

We study an active cluster recovery problem where, given a set of n points and an oracle answering queries like "are these two points in the same cluster?", the task is to recover exactly all clusters using as few queries as possible. We begin by introducing a simple but general notion of margin between clusters that captures, as special cases, the margins used in previous works, the classic SVM margin, and standard notions of stability for center-based clusterings. Under our margin assumptions we design algorithms that, in a variety of settings, recover all clusters exactly using only $\mathcal{O}(\log n)$ queries. For $\mathbb{R}^m$, we give an algorithm that recovers *arbitrary* convex clusters, in polynomial time, and with a number of queries that is lower than the best existing algorithm by $\Theta(m^m)$ factors. For general pseudometric spaces, where clusters might not be convex or might not have any notion of shape, we give an algorithm that achieves the $\mathcal{O}(\log n)$ query bound, and is provably near-optimal as a function of the packing number of the space. Finally, for clusterings realized by binary concept classes, we give a combinatorial characterization of recoverability with $\mathcal{O}(\log n)$ queries, and we show that, for many concept classes in $\mathbb{R}^m$, this characterization is equivalent to our margin condition. Our results show a deep connection between cluster margins and active cluster recoverability.

1 Introduction

This work investigates the problem of exact cluster recovery using oracle queries, in the well-known framework introduced by Ashtiani et al. [2016]. We are given a set X of n points from some domain $\mathcal{X}$ (e.g., from the Euclidean m-dimensional space $\mathbb{R}^m$) and an oracle answering to same-cluster queries of the form "are these two points in the same cluster?" or label queries of the form "which cluster does this point belong to?". The oracle answers are consistent with some clustering $\mathcal{C} = (C_1, \ldots, C_k)$ of X unknown to the algorithm, where k is a fixed constant. The goal is to design an algorithm that recovers $\mathcal{C}$ by using as few queries as possible.

Clearly, if there are no restrictions on $\mathcal{C}$, then any algorithm needs n queries in the worst case. Thus, our goal is to understand when $\mathcal{C}$ can be recovered efficiently; ideally, in polynomial time, and by making $\mathcal{O}(\log n)$ queries. A natural attempt is to consider clusterings with well-separated clusters, since this is what is often considered a "good clustering". The existing work goes precisely in this direction, starting with the following result of Ashtiani et al. [2016] for the case $X \subseteq \mathbb{R}^m$. If every cluster C_i is separated from $X \setminus C_i$ by a ball centered in the center of mass of C_i, and that ball does not intersect other clusters even if expanded by a factor of $1 + \gamma$, then with high probability

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

$\mathrm{poly}(k, m, 1/\gamma) \log n$ queries are sufficient to recover $\mathcal{C}$ in polynomial time. This is called "spherical margin" condition, and γ is called the margin.

Unfortunately, the spherical margin condition is not very realistic, since it imposes a very precise shape upon the clusters. In a generalization attempt, Bressan et al. [2020] showed that one can actually recover clusters with ellipsoidal separators with arbitrary centers, by increasing the number of queries to $\mathrm{poly}(k, m, 1/\gamma)\,(m/\gamma)^m \log n$. This result is achieved via boosted one-sided error learning, which works as follows. Suppose that, by making $f(k, m, \gamma)$ queries, we could identify correctly (with zero mistakes) a constant fraction of the points in some cluster C_i. Then, we could label those points as i, remove them from the dataset, and repeat. It is not hard to show that, after $\mathcal{O}(k \log n)$ rounds, we will have correctly labeled all the input points with high probability. The difficult task is, of course, to learn a constant fraction of some cluster C_i with one-sided error (one-sided error means that any point of X predicted to be in C_i must be in C_i). The key insight in [Bressan et al., 2020] is that, if the clusters have margin γ with respect to their ellipsoidal separators, then roughly $(m/\gamma)^m$ queries are sufficient. This leads to the following question: how much can this approach be extended?

In this work we provide several answers, revealing an interesting connection between margin-based cluster recovery and one-sided error learning. Our main contributions are as follows.

1. We introduce a new notion of margin in $\mathbb{R}^m$, that we call "convex hull margin" (Definition 2). This is a strict generalization of the margins of Bressan et al. [2020], Ashtiani et al. [2016] and of the usual SVM margin, and allows the clusters to have *any shape whatsoever* as long as they are convex. Under the convex hull margin, we develop a novel technique for learning with one-sided error that we call *convex hull expansion trick*. It essentially amounts to sampling many points from a single cluster and "inflate" their convex hull by a factor of $(1 + \gamma)$. This technique yields a polynomial-time exact cluster recovery algorithm that uses $\mathrm{poly}(k, m, 1/\gamma)\,(1 + 1/\gamma)^m \log n$ queries (Theorem 1). The $(1 + 1/\gamma)^m$ dependence on γ and m is significantly better than that of Bressan et al. [2020], and closer to their lower bound of order $(1 + 1/\gamma)^{m/2}$.
2. We introduce a notion of cluster margin for general pseudometric spaces called *one-versus-all margin* (Definition 3). This notion of margin is strictly more general than convex hull margin, and captures, as special cases, standard notions of stability for clustering problems such as k-means or k-centers. We show that, if a clustering has one-versus-all margin, then it can be recovered with $M(\gamma)\,\mathrm{poly}(k) \log n$ queries via a pure learning-theoretic approach (Theorem 3), where $M(\gamma)$ is a quantity related to the packing numbers of the pseudometric space. We show that the dependence on $M(\gamma)$ is essentially optimal, thus characterizing the recoverability of clusterings in this setting.
3. Finally, we show a connection between margin-based learning and exact active cluster recoverability, when clusters are realized by some concept class $\mathcal{H}$ (that is, when for each cluster C_i there is a concept $h_i \in \mathcal{H}$ such that $X \cap h_i = C_i$). We show that if a certain combinatorial parameter, the *coslicing dimension* $\mathrm{cosl}(\mathcal{H})$, is bounded, then one can learn clusterings with $\mathrm{cosl}(\mathcal{H})\,\mathrm{poly}(k) \log n$ label queries; otherwise, $\Omega(n)$ queries are needed in the worst case (Theorem 4). Moreover we show that, for all concept classes in $\mathbb{R}^m$ that are closed under affine transformations and well-behaved in a natural sense, finite coslicing dimension and positive one-versus-all margin are equivalent (Theorem 5).

Note that actively learning a clustering is equivalent, up to a relabeling of the classes, to actively learning a multiclass classifier in the transductive realizable case. Hence our results apply to that case, too. In particular, our $\mathcal{O}(\log n)$ query bounds imply $\widetilde{\mathcal{O}}(\log 1/\varepsilon)$ query bounds for pool-based active learning [McCallum and Nigam, 1998] of multiclass classifiers, where ε is the generalization error. To see this, draw a set X of $\Theta\big(\varepsilon^{-1}(K \log 1/\varepsilon + \log 1/\delta)\big)$ unlabeled samples from the underlying distribution, where K is the relevant measure of capacity (e.g., the Natarajan dimension), run our algorithms over X, and compute a hypothesis consistent with the recovered labeling $\mathcal{C}$. These kinds of reductions are standard in active learning, see for instance [Kane et al., 2017].

On the other hand, our results *do not* apply to actively learning subsets $h \subseteq \mathcal{X}$ (that is, to active learning in the standard sense) if our margin conditions are only enjoyed by the set of positives $X_+ = X \cap h$. To see this, let $X \subset \mathbb{R}^2$ and suppose X_+ contains a single point. Then X_+ satisfies our conditions with unbounded margin w.r.t. the Euclidean distance, but any algorithm needs $\Omega(n)$ label queries to recover it. This does not happen with clustering because, in that case, every class enjoys the margin property, and in particular both the "positives" and the "negatives" for $k = 2$.

Table 1 compares the bounds resulting from the different notions of margins known.

margin	query bound	reference
spherical	$\mathcal{O}\Big(k \log n + k^2 \frac{\log k + \log \frac{1}{\delta}}{\gamma^4}\Big)$	[Ashtiani et al., 2016]
ellipsoidal	$\mathcal{O}\Big(k \log n \left(k^2 m^2 \log k + \max\Big\{2^m, \mathcal{O}\Big(\frac{m}{\gamma} \log \frac{m}{\gamma}\Big)^m\Big\}\right)\Big)$	[Bressan et al., 2020]
convex hull	$\mathcal{O}\Big(k^3 m^5 \left(1 + \frac{1}{\gamma}\right)^m \log\Big(1 + \frac{1}{\gamma}\Big) \log n\Big)$	this work
one-versus-all	$\mathcal{O}(M^*(\gamma)\, k^2 \log k \log n)$	this work

Table 1: Summary of existing margin notions and corresponding known query bounds for same-cluster queries. The first three bounds assume $\mathcal{X} = \mathbb{R}^m$. The fourth bound is for general pseudometric spaces and $M^*(\gamma)$ is roughly a packing number—see Section 4. Note that m can often be replaced by the maximum rank (i.e., the rank of the subspace spanned by the points) of any cluster, see [Bressan et al., 2020]. The spherical margin assumes, for every cluster C_i, that $\forall x \in C_i, y \in X \setminus C_i$, $d(y, \mu_i) > (1+\gamma) d(x, \mu_i)$, where $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$. The ellipsoidal margin assumes that for some PSD matrix $W_i \in \mathbb{R}^{m \times m}$ and some $c_i \in R^m$, $\forall x \in C_i, y \in X \setminus C_i$, $d_W(y, c_i) > \sqrt{1+\gamma}\, d_W(x, c_i)$, where $d_W(a, b) = \sqrt{\langle a - b, W(a-b)\rangle}$; note that for $\gamma \ll 1$ this can be thought of as $d_W(y, c_i) > (1+\gamma) d_W(x, c_i)$.

Related work. Same-cluster queries were introduced formally in [Ashtiani et al., 2016] together with the active cluster recovery problem. Those queries are natural to implement in crowd-sourcing systems, and for this reason they have been extensively studied both in theory [Ailon et al., 2018a,b, Gamlath et al., 2018, Huleihel et al., 2019, Mazumdar and Pal, 2017, Mazumdar and Saha, 2017b,a, Saha and Subramanian, 2019, Vitale et al., 2019] and in practice [Firmani et al., 2018, Gruenheid et al., 2015, Verroios and Garcia-Molina, 2015, Verroios et al., 2017].

Various notions of margin are central in both active learning and cluster recovery [Xu et al., 2004, Balcan et al., 2007, Balcan and Long, 2013, Kane et al., 2017, Bressan et al., 2021]. Our coslicing dimension is similar to the slicing dimension of Kivinen [1995] and the star number of Hanneke and Yang [2015]. Our lower bounds, like many others, are inspired from a construction by Dasgupta [2004]. Our arguments based on packing numbers are similar to those based on the inference dimension of Kane et al. [2017] or the lossless sample compression of Hopkins et al. [2021], as we cannot infer the label of a point only when it is far from already-labeled points. A query bound similar to the one given by our convex hull expansion trick, but worse by a factor roughly 2^m, can be inferred by adapting arguments of Hopkins et al. [2020b]. Combinatorial characterizations of multiclass learning have been proposed in the passive case by Ben-David et al. [1995], Rubinstein et al. [2009], Daniely and Shalev-Shwartz [2014]. Other learning settings related to one-sided and active learning are RPU learning [Rivest and Sloan, 1988] and perfect selective classification [El-Yaniv and Wiener, 2012] — see [Hopkins et al., 2020a] for a discussion.

2 Preliminaries and notation

All missing proofs can be found in the supplementary material. The input is a pair (X, O), where X is a set of n points from some domain $\mathcal{X}$, and O is a label oracle that, when queried on any $x \in X$, returns the cluster id $\mathcal{C}(x)$ of x. The oracle O is consistent with a *latent clustering* $\mathcal{C} = (C_1, \ldots, C_k)$ of X, i.e., a k-tuple of pairwise disjoint sets whose union is X.[1] We allow clusters to be empty. Our goal is to recover $\mathcal{C}$ by making as few queries as possible to O. Queries can be made adaptively, that is, the j-th point to be queried can be chosen as a function of the answers to the first $j-1$ queries. We express the number of queries as a function of k, n, and other parameters to be introduced later. This setting is essentially equivalent to the semi-supervised active clustering (SSAC) framework of Ashtiani et al. [2016], where the oracle answers same-cluster queries $\textsc{scq}(x, y)$ that, for any two points $x, y \in X$, return TRUE iff $\mathcal{C}(x) = \mathcal{C}(y)$. We use label queries instead of SCQ queries for simplicity; any SCQ query can be emulated with two label queries. Conversely, the label of any point can be learned with k SCQ queries, up to a relabeling of the clusters, so our bounds hold for an SCQ oracle as well if multiplied by k. For conciseness, we state query bounds in the form $f(k, m, n, \gamma)$,

[1] In line with previous works, we assume k is fixed and known.

for instance $\mathcal{O}\big(k^2\,(1+{}^1\!/\!_\gamma)^m \log n\big)$. These bounds should be thought of as $\min\big\{n, f(k,m,n,\gamma)\big\}$, since obviously we never need to query the same point twice.

We often assume a metric or pseudometric d over $\mathcal{X}$ (a pseudometric allows two distinct points to have distance 0). For $x \in \mathcal{X}$ and $r \geq 0$, the closed ball of radius r centered at x is $B(x,r) = \{y \in \mathcal{X} : d(x,y) \leq r\}$. For any $X \subset \mathcal{X}$, we denote by $\phi_d(X) = \sup_{x,x' \in X} d(x,x')$ the diameter of X measured by d, and we define $\phi_d(\emptyset) = 0$. For any two sets $U, S \subset \mathcal{X}$, we denote by $d(U,S) = \inf_{x \in U, y \in S} d(U,S)$ their distance according to d, and we define $d(U,\emptyset) = \infty$. For any $X \subset \mathbb{R}^m$, we write $\mathrm{conv}(X)$ for the convex hull of X. The unit sphere in $\mathbb{R}^m$ is $S^{m-1} = \{x \in \mathbb{R}^m : \|x\|_2 = 1\}$. We recall some learning-theoretic facts. Let $\mathcal{H}$ be an arbitrary collection of subsets of $\mathcal{X}$ (i.e., a concept class). The intersection class of $\mathcal{H}$ is $I(\mathcal{H}) = \bigcup_{i \in \mathbb{N}} \{h_1 \cap \ldots \cap h_i : h_1, \ldots, h_i \in \mathcal{H}\}$. Given any $S \subset \mathcal{X}$ and any $S' \subseteq S$ realized by some $h^\star \in \mathcal{H}$, the smallest concept in $I(\mathcal{H})$ consistent with S' is defined as $h^\circ = \bigcap \{h \in \mathcal{H} : h \cap S = S'\}$. Note that $h^\circ \subseteq h^\star$. Finally, we recall the definition of learning with one-sided error:

Definition 1 (Kivinen [1995], Definition 4.4)**.** *An algorithm $\mathcal{A}$ learns $\mathcal{H}$ with one-sided error ε and confidence δ with r examples if, for any target concept $h^\star \in \mathcal{H}$ and any probability measure $\mathcal{P}$ over $\mathcal{X}$, by drawing r independent labeled examples from $\mathcal{P}$, the algorithm outputs a concept $h \subseteq h^\star$ such that $\mathcal{P}(h^\star \setminus h) \leq \varepsilon$ with probability at least $1-\delta$.*

3 Margin-based exact recovery of clusters in $\mathbb{R}^m$

In this section we consider the case $\mathcal{X} = \mathbb{R}^m$. We show that the ellipsoidal margin of Bressan et al. [2020] can be significantly generalized, while retaining the $\mathcal{O}(\log n)$ query complexity, by introducing what we call the *convex hull margin*. In a nutshell the convex hull margin says that, given any cluster C, any point not in C is separated by the convex hull of C by a distance at least γ times the diameter of C. Instead of using the Euclidean metric, however, we allow distances to be measured by *any* pseudometric over $\mathbb{R}^m$, which we do not need to know, and which may even differ from cluster to cluster. The only requirement is that the pseudometric be homogeneous and invariant under translation (i.e., induced by a seminorm).

Definition 2 (Convex hull margin)**.** *Let D be the family of all pseudometrics induced by the seminorms over $\mathbb{R}^m$, and let $X \subset \mathbb{R}^m$ be a finite set. A clustering $\mathcal{C} = (C_1, \ldots, C_k)$ of X has convex hull margin γ if for every $i \in [k]$ there exists $d_i \in D$ such that:*

$$d_i\big(X \setminus C_i, \mathrm{conv}(C_i)\big) > \gamma\, \phi_{d_i}(C_i) \tag{1}$$

This definition has a few interesting properties. First, it strictly generalizes the ellipsoidal margin of Bressan et al. [2020] and the spherical margin of Ashtiani et al. [2016]. To see this, let D be the class of all pseudometrics over $\mathbb{R}^m$ that can be written as $d_W(x,y) = \langle x-y, W(x-y)\rangle$ for some positive semidefinite matrix $W \in \mathbb{R}^{m \times m}$ (for the spherical margin, take $W = rI$ where I is the identity matrix). Second, it strictly generalizes (the multiplicative version of) the classic SVM margin. Indeed, under the Euclidean metric, if X has diameter $R = \phi(X)$ and every cluster C can be separated from $X \setminus C$ by a linear separator whose boundary has distance ρ from X, then C has convex hull margin at least $\frac{\rho}{R}$; and there are cases with arbitrarily small (multiplicative) SVM margin but arbitrarily large convex hull margin, see Section 4. Finally, pseudometrics are versatile and can express, for instance, the Euclidean distance between points after a projection on a subspace. This models scenarios where each cluster only "cares" about a certain subset of the features.

Under the convex hull margin, we give a polynomial-time algorithm, named CHEATREC (for Convex Hull ExpAnsion Trick Recovery) that recovers $\mathcal{C}$ using $\mathcal{O}(\log n)$ queries.

Theorem 1. *Let (X,O) be an instance whose latent clustering $\mathcal{C}$ has convex hull margin $\gamma > 0$. Then* CHEATREC(X,O,γ) *outputs $\mathcal{C}$, runs in time* $\mathrm{poly}(k,n,m)$*, and with high probability makes a number of label queries to O bounded by $\mathcal{O}\big(k^2 m^5\,(1+{}^1\!/\!_\gamma)^m \log(1+{}^1\!/\!_\gamma)\log n\big)$.*

To put this result in perspective, consider the algorithm of Bressan et al. [2020]. Under an ellipsoidal margin of γ_{EL}, that algorithm achieves a query bound of roughly $\big(\frac{m}{\gamma_{EL}}\big)^m \log n$. One can check that an ellipsoidal margin of γ_{EL} implies $\gamma \geq \frac{\gamma_{EL}}{3}$ for all $\gamma_{EL} \leq 1$[2]. Hence, in this range, our dependence on γ is better by $\Theta(m)^m$ factors.

[2]Their definition uses squared distances, so the relationship with our margin is $1+\gamma \geq \sqrt{1+\gamma_{EL}}$.

Both CHEATREC and the algorithm of [Bressan et al., 2020] are based on boosting learners with one-sided error. What makes CHEATREC different is, thus, how it learns with one-sided error. To explain this, let us recall how algorithm of Bressan et al. [2020] works. The algorithm starts by drawing a uniform random sample of $\Theta(m^2)$ points S_C from some cluster C. (As there are at most k clusters, to draw s points u.a.r. from some cluster we can just draw ks points u.a.r. from X, query their labels, and take the subset of the majority label). Then, it computes an outer approximation of $\mathrm{conv}(S_C)$ by fitting a minimum-volume ellipsoid $E(S_C)$ to S_C. Since the VC dimension of m-dimensional ellipsoids is $O(m^2)$, by standard PAC bounds $E(S_C)$ contains half of C with good probability. The problem is that $E(S_C)$ approximates $\mathrm{conv}(S_C)$ very roughly, and thus it may intersect a large portion of $X \setminus C$. Indeed, the algorithm makes roughly $(m/\gamma_{EL})^m$ additional queries to filter out all points in $E(S_C) \cap (X \setminus C)$.

This gives us the following idea: instead of spending that many queries on "cleaning" $E(S_C)$, we spend them to build a much larger sample S_C, so that we can approximate $\mathrm{conv}(S_C)$ using a class of convex bodies of VC dimension much higher than $\Theta(m^2)$ — for instance, a polytope on $(1 + 1/\gamma)^m$ vertices. That body should enclose $\mathrm{conv}(S_C)$ quite tightly; perhaps so tightly to contain no point outside C. This intuition leads to our technique, called *convex hull expansion trick*, which essentially amounts to drawing a large sample from some cluster C and inflate its convex hull by a factor $1 + \gamma$. This technique yields bounds that, compared to [Bressan et al., 2020], are strictly better and hold under more general assumptions. In the rest of the section, we describe this trick and sketch the proof of Theorem 1. For the complete proof, see the supplementary material.

3.1 CHEATREC and the convex hull expansion trick

The starting point of CHEATREC is the following idea. Let C be any cluster of $\mathcal{C}$, and let S_C be any subset of C. Let $K = \mathrm{conv}(S_C)$, choose any $z \in K$, and let $Q = (1+\gamma)K$ be the scaling of K with respect to z. Finally, let $d \in D$ be the pseudometric under which C has convex hull margin γ. Since d is homogeneous and invariant under translation (recall that every $d \in D$ is induced by a seminorm), any $y \in X$ such that $d(y, K) \leq \gamma\phi_d(K)$ must belong to C as well. Therefore, $X \cap Q \subseteq C$. That is, Q contains only points of C.

Clearly, knowing that $X \cap Q \subseteq C$ is not enough; in order to make progress, we must guarantee that $X \cap Q$ forms a good fraction of C (in principle $X \cap Q$ could just coincide with S_C, in which case we learn nothing). To this end, we draw a uniform random sample S from X of size $|S| = ks$, where s will be set later, and query all labels of S. For every cluster C we let $S_C = S \cap C$; as there are at most k clusters, at least one of them satisfies $|S_C| \geq s$. The convex hull expansion trick says that, if (1) s is roughly $(1 + 1/\gamma)^m$, and (2) $z = \mu_K$, where μ_K is the center of mass of K, then Q contains a good fraction of C with good probability. This gives the desired query bounds for learning C with one-sided error; to implement the trick in polynomial time, however, we have to replace μ_K with an approximation, since computing it exactly is hard [Rademacher, 2007]. The convex hull expansion trick in fact states that the claim about Q holds even with such an approximation.

We give below a formal statement of the trick, and a sketch of its proof. The uniform probability measure $\mathcal{U}$ over K is defined by $\mathcal{U}(K') = \frac{\mathrm{vol}(K')}{\mathrm{vol}(K)}$ for all measurable $K' \subseteq K$. A probability measure $\mathcal{P}$ is ε-uniform if $|\mathcal{P}(K') - \mathcal{U}(K')| \leq \varepsilon$ for all measurable $K' \subseteq K$.

Lemma 1 (Convex hull expansion trick)**.** *Fix $\gamma > 0$, and let $s = \Theta\big(m^5\big(1 + 1/\gamma\big)^m \log\big(1 + 1/\gamma\big)\big)$ large enough. Let S_C be a sample of s independent uniform random points from some cluster C, and let $K = \mathrm{conv}(S_C)$. Let $X_1, \ldots, X_N$ be independent random points sampled ε-uniformly from K, with $\varepsilon \in \Theta(m^{-1})$ small enough and $N \in \Theta(m^2)$ large enough, and let $z = \frac{1}{N}\sum_{i=1}^{N} X_i$. Finally, let $Q = (1 + \gamma)K$ where the center of the scaling is z. Then $\mathbb{P}\big(|Q \cap C| \geq |C|/2\big) \geq 1/2$.*

Sketch of the proof. For simplicity we ignore factors $\mathrm{poly}(m)$ and $\log(1/\gamma)$. To begin, suppose Q was obtained by scaling K about its own center of mass, μ_K. By a probabilistic argument adapted from [Naszódi, 2018], there exists a polytope P on roughly $(1 + 1/\gamma)^m$ vertices such that $K \subseteq P \subseteq Q$. Now, the class $\mathcal{P}_{t,m}$ of all polytopes on at most t vertices in $\mathbb{R}^m$ has VC-dimension roughly t, see [Kupavskii, 2020]. Let then $t = (1 + 1/\gamma)^m$. Since $P \in \mathcal{P}_{t,m}$, and since P is consistent with S_C (that is, $S_C \subseteq P$), by standard PAC bounds $\mathbb{P}(|P \cap X| \geq |C|/2) \geq 1/2$ as long as $|S_C| = \Omega(t) = \Omega\big((1 + 1/\gamma)^m\big)$. But $P \subseteq Q$, so $\mathbb{P}(|Q \cap X| \geq |C|/2) \geq 1/2$ as well.

Now suppose that, in place of μ_K, we use $z = \frac{1}{N}\sum_{i=1}^{N} X_i$ where each X_i is an independent random points sampled ε-uniformly from K. We consider K is in isotropic position. The radius of K is then at most m, so $\|X_i\|_2 \leq m$. By standard calculations this implies that, if $N = \Theta(m^2)$ and each X_i comes from a distribution that is $\Theta(1/m)$-uniform over K, with good probability z is at distance $\eta = O(1)$ from μ_K. In particular, by increasing N by constant factors, we can make arbitrarily small the probability that $\eta \leq 1/3$. This is sufficient to adapt a result of Naszódi [2018], through an extension of Grunbaum's inequality for convex bodies due to Bertsimas and Vempala [2004], and show that a polytope P such as the one described above still exists, even if we expand K about z. $\square$

We conclude with a note on implementing CHEATREC in polynomial time. First, we should sample the points X_i efficiently. To this end, we transform S_C so that $\mathrm{conv}(S_C)$ is in near-isotropic position. This amounts to computing John's ellipsoid for S_C, and applying the map that turns that ellipsoid into a ball, which takes polynomial time. Afterwards, we can draw every X_i via the "hit-and-run from a corner" algorithm of Lovász and Vempala [2006], in time $\mathrm{poly}(m, |S_C|) \log \frac{m}{\varepsilon}$ — this includes the time to solve a linear program to determine when the random walk of [Lovász and Vempala, 2006] hits the boundary of K. Once we have $z = \frac{1}{N}\sum_{i=1}^{N} X_i$, the only remaining issue is to avoid computing $K = \mathrm{conv}(S_C)$ and $Q = (1+\gamma)K$ explicitly as intersections of halfspaces, as this could take time $|S_C|^{\Theta(m)}$. Instead, we just rescale the sample S_C about z, obtaining a representation of Q as a set of vertices. Then, computing $X \cap Q$ amounts to solving for every $x \in X$ a feasibility problem, which takes polynomial time.

4 The one-versus-all margin

In this section, we let $\mathcal{X}$ be a generic space equipped with a set of pseudometrics. Like in Section 3, we want to formulate a notion of margin between clusters. However, we cannot express the margin in terms of diameter of convex hulls (since $\mathcal{X}$ need not be a vector space and may have no notion of convexity at all). Thus, we introduce a notion called *one-versus-all margin*. We prove that clusterings with positive one-versus-all margin can be recovered with $\mathcal{O}(\log n)$ oracle queries. However, we do not provide a running time analysis; the implementation of our algorithm depends on $\mathcal{X}$ and on the specific class of clusterings, and in general may take time superpolynomial in the size of the input. Let us introduce the one-versus-all margin.

Definition 3 (One-versus-all margin)**.** *Fix k pseudometrics $d_1, \ldots, d_k$ over $\mathcal{X}$. A clustering $\mathcal{C} = (C_1, \ldots, C_k)$ of a finite set $X \subset \mathcal{X}$ has one-versus-all margin γ with respect to $d_1, \ldots, d_k$ if for all $i \in [k]$ we have $d_i(X \setminus C_i, C_i) > \gamma\, \phi_{d_i}(C_i)$.*

Remark. Unlike the convex hull margin (Definition 2), here $d_1, \ldots, d_k$ are fixed in advance. The reason is that here $\mathcal{X}$ may not be a linear space, $C_1, \ldots, C_k$ may not be convex, or $d_1, \ldots, d_k$ may not be induced by seminorms. Hence, we cannot apply techniques like the convex hull expansion trick, which work *for every pseudometric* including $d_1, \ldots, d_k$. As a consequence, it is unclear how one can recover $\mathcal{C}$ efficiently without knowing $d_1, \ldots, d_k$ in advance.

Before continuing, we shall prove that the one-versus-all margin can be arbitrarily large while the multiplicative SVM margin is arbitrarily small. This also proves an analogous claim for the convex hull margin in Section 3.

Lemma 2. *For any $u \in \mathbb{R}^2$ let $d_u(x, y) = |\langle u, x - y\rangle|$. For any $\eta > 0$ there exists a clustering $\mathcal{C} = (C_1, C_2)$ on a set $X \subset \mathbb{R}^2$ that has arbitrarily large one-versus-all margin with respect to $d_{(0,1)}, d_{(1,0)}$, and yet $d_u(C_1, C_2) \leq \eta\, \phi_{d_u}(X)$ for all $u \in \mathbb{R}^2$.*

Proof sketch. Consider Figure 1. The two points along the x axis belong to C_1, and the two points along the y axis belong to C_2. The pseudometric $d_1 = d_{(0,1)}$ measures the distance along the y axis, and the pseudometric $d_2 = d_{(1,0)}$ measures the distance along the x axis. Therefore $d_1(C_1, C_2) > 0 = \phi_{d_1}(C_1)$, and $d_2(C_1, C_2) > 0 = \phi_{d_2}(C_2)$. Hence, the one-versus-all margin of $\mathcal{C}$ with respect to d_1, d_2 is unbounded. Yet, for any $\eta > 0$ and any $u \in \mathbb{R}^2$ we can make $d_u(C_1, C_2) \leq \eta\, \phi_{d_u}(X)$ by placing the endpoints of the clusters arbitrarily near the origin. $\square$

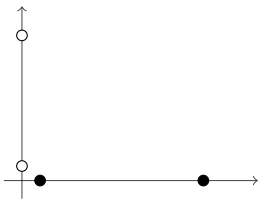

Figure 1: An instance with arbitrarily small SVM margin but unbounded one-versus-all margin.

4.1 The one-versus-all-margin captures the stability of center-based clusterings

Fix any pseudometric d over $\mathcal{X}$. A clustering $\mathcal{C} = (C_1, \ldots, C_k)$ of X is *center-based* if there exist k points $c_1, \ldots, c_k \in \mathcal{X}$, called *centers*, such that for every $i \in [k]$ and every $x \in C_i$ we have $d(x, c_j) > d(x, c_i)$ for all $j \neq i$. In other terms, every point is assigned to the nearest center. It is well known that many popular center-based clustering problems, such as k-means or k-centers, are NP-hard to solve in general. However, those problems become polynomial-time solvable (or approximable) if the solution $\mathcal{C}$ meets certain stability properties. We show that two of these properties, the α-center proximity of Awasthi et al. [2012] and the $(1+\varepsilon)$-perturbation resilience of Bilu and Linial [2012], imply positive one-versus-all margin. Let us recall these properties. We define a $(1+\varepsilon)$-perturbation of d as any function d' (which need not be a pseudometric) such that $d \leq d' \leq (1+\varepsilon)d$.

Definition 4. *Let $\mathcal{C}$ be a center-based clustering.*

- *$\mathcal{C}$ satisfies the α-center proximity property with $\alpha > 1$ if, for all $i \in [k]$, for all $x \in C_i$ and all $j \neq i$ we have $d(x, c_j) > \alpha\, d(x, c_i)$.*
- *$\mathcal{C}$ is $(1+\varepsilon)$-perturbation resilient with $\varepsilon > 0$ if it is induced by the same centers $c_1, \ldots, c_k$ under any $(1+\varepsilon)$-perturbation of d.*

It is known that $(1+\varepsilon)$-perturbation resilience implies α-center proximity with $\alpha = 1+\varepsilon$, see [Awasthi et al., 2012]. Our result is:

Theorem 2. *If $\mathcal{C}$ satisfies α-center proximity, then it has one-versus-all margin $\gamma \geq \frac{(\alpha-1)^2}{2(\alpha+1)}$. Hence, if $\mathcal{C}$ satisfies $(1+\varepsilon)$-perturbation stability, then it has one-versus-all margin $\gamma \geq \frac{\varepsilon^2}{2(\varepsilon+2)}$.*

4.2 Cluster recovery with one-versus-all margin

We conclude by showing how a clustering with one-versus-all margin $\gamma > 0$ can be recovered with $\mathcal{O}(\log n)$ queries. Our algorithm, MREC, draws uniform random points from X and then selects the smallest hypothesis consistent with the points sampled from each cluster (see Section 2). We will show that the VC-dimension of the class from which that hypothesis is taken can be bounded in terms of $\frac{1}{\gamma}$. This implies that, at every round, MREC learns a constant fraction of labels without mistakes.

We need some further notation. Let d be any pseudometric over $\mathcal{X}$. For any $X \subset \mathcal{X}$ and any $r > 0$, let $\mathcal{M}(X, r, d)$ be the maximum cardinality of any $A \subseteq X$ such that $d(x, y) > r$ for all distinct $x, y \in A$. From now on we assume that $\mathcal{M}(X, r, d)$ is bounded[3], and for any $\gamma > 0$, we define $M(\gamma, d) = \max\{\mathcal{M}\big(B(x, r), \gamma r, d\big) : x \in \mathcal{X}, r > 0\}$. Thus any ball B contains at most $M(\gamma, d)$ points at pairwise distance greater than γ times the radius of B, and some B attains this bound. Finally, by $\text{vc-dim}(H, X)$ we denote the VC-dimension of a generic concept class H over a set X.

Lemma 3 (One-versus-all margin implies one-sided-error learnability)**.** *Let d be any pseudometric over $\mathcal{X}$. For any finite $X \subset \mathcal{X}$ and any $\gamma > 0$, define the effective concept class over X:*

$$H = \{C \subseteq X \,:\, d(X \setminus C, C) > \gamma\, \phi_d(C)\} \tag{2}$$

Then $H = I(H)$, and $\text{vc-dim}(H, X) \leq M^*(\gamma, d)$ where $M^*(\gamma, d) = \max(2, M(\gamma, d))$. Therefore, H can be learned with one-sided error ε and confidence δ with $\mathcal{O}\big(\varepsilon^{-2}(M^*(\gamma, d) \log {}^1\!/_\varepsilon + {}^1\!/_\delta)\big)$ examples by choosing the smallest consistent hypothesis in H.

Sketch of the proof. To prove that $H = I(H)$, one can take any two $C_1, C_2 \in H$ and show that $C_1 \cap C_2$ satisfies the margin condition, too. To prove that $\text{vc-dim}(H, X) \leq M^*(\gamma, d)$, we have two steps. First, let $\text{sl}(H, X)$ be the *slicing dimension* of H. This is the size of the largest subset $S \subseteq X$ *sliced* by H, i.e., such that for every $x \in S$ there is $C \in H$ giving $S \setminus x = S \cap C$, see [Kivinen, 1995]. As the same work shows, we have $\text{vc-dim}(I(H), X) \leq \text{sl}(H, X)$; hence, to prove $\text{vc-dim}(H, X) \leq M^*(\gamma, d)$ it suffices to prove that $\text{sl}(H, X) \leq M^*(\gamma, d)$. To this end, we use a packing argument. Suppose that $S \subseteq X$ is sliced by H, choose any $x \in S$, and let $C \in H$ such that $S \setminus x = S \cap C$. By construction of H, we know that $d(C, x) > \gamma\phi_d(C)$. Since $S \setminus x \subseteq C$, this yields:

$$d\big(S \setminus x, x\big) \geq d(C, x) > \gamma\, \phi_d(C) \geq \gamma\, \phi_d(S \setminus x) \tag{3}$$

[3] If $\mathcal{M}(X, r, d)$ is not bounded, then our results can be extended in the natural way, that is, we can prove a lower bound of $\Omega(n)$ queries for instances of n points.

It can be shown that this implies $d(S \setminus x, x) > \gamma\phi_d(S)$ for all $x \in S$, and, in turn, $|S| \leq M^*(\gamma, d)$. The claim on the learnability with one-sided error holds by choice of the smallest consistent hypothesis in $H = I(H)$, combined with standard PAC bounds. $\square$

We present our main result. Let $M(\gamma) = \max_{d \in \{d_1, \ldots, d_k\}} M(\gamma, d)$ with $d_1, \ldots, d_k$ as in Definition 3.

Theorem 3. *Let (X, O) be any instance whose latent clustering $\mathcal{C}$ has one-versus-all margin $\gamma > 0$ with respect to $d_1, \ldots, d_k$. Then* MREC$(X, O, \gamma, d_1, \ldots, d_k)$ *outputs $\mathcal{C}$ while making, with high probability, at most $\mathcal{O}(M^*(\gamma) k \log k \log n)$ label queries to O, where $M^*(\gamma) = \max(2, M(\gamma))$. Moreover, for any algorithm $\mathcal{A}$ and for any $\gamma > 0$, there are instances with one-versus-all margin γ on which $\mathcal{A}$ makes $\Omega(M(2\gamma))$ label queries in expectation.*

Sketch of the proof. For the lower bounds, we take a set X on $M(2\gamma)$ points at pairwise distance larger than 2γ times the radius of X, which is at least γ times the diameter of X, and we draw a random clustering $\mathcal{C}$ in the form $(x, X \setminus x)$. One can see that $\mathcal{C}$ has one-versus-all margin γ, and simple arguments, coupled with Yao's principle for Monte Carlo algorithms, show that any algorithm needs $\Omega(M(2\gamma))$ queries in the worst case to return $\mathcal{C}$.

For the upper bounds, we show how to learn an expected constant fraction of X with one-sided error using $\Theta(M^*(\gamma) k \log k)$ queries; the rest follows by our general boosting argument. To begin, for each $i \in [k]$ we consider the effective concept class:

$$H_i = \{C \subseteq X \,:\, d_i(C, X \setminus C) > \gamma\, \phi_{d_i}(C)\} \tag{4}$$

We then set $\varepsilon = 1/2k$ and $\delta = 1/2$, and draw a labeled sample S of size $\Theta(\varepsilon^{-1}(M^*(\gamma) \log 1/\varepsilon + \log 1/\delta)) = \Theta(M^*(\gamma) k \log k)$. Finally, for each $i \in [k]$ we choose the smallest hypothesis $\widehat{C}_i \in H_i$ consistent with the subset $S_i \subseteq S$ labeled as i, and we assign label i to all points in $\widehat{C}_i$. Clearly, $\widehat{C}_i \subseteq C_i$, therefore we are learning with one-sided error. Moreover, by Lemma 3, with probability at least $1/2$ we have $|\widehat{C}_i| \geq |C_i| - \varepsilon|X| = |C_i| - |X|/2k$. As $|\widehat{C}_i| \geq 0$, this implies $\mathbb{E}|\widehat{C}_i| \geq |C_i|/2 - |X|/4k$. By summing over all i, this shows that we are labelling correctly at least $|X|/4$ points in expectation. $\square$

Remark. By Theorem 3, in $\mathbb{R}^m$ MREC yields a $\mathcal{O}(\log n)$ query bound even when the clusters are *not* convex. Note however that this does not mean that MREC subsumes CHEATREC, for two reasons. First, as noted above, here $d_1, \ldots, d_k$ are known in advance, whereas MREC does not need to know them. Second, MREC works by computing the smallest hypothesis $\widehat{C}_i$ consistent with S_i (see the proof of Theorem 3), which in general may take superpolynomial time. Indeed, CHEATREC runs in polynomial time by *not* computing $\widehat{C}_i$ at all.

5 One-versus-all clusterings

In this section we consider the case where the clusters of $\mathcal{C}$ are realized by some concept class $\mathcal{H}$ over $\mathcal{X}$. This is the clustering equivalent of the one-versus-all multiclass classifiers (see, e.g., [Shalev-Shwartz and Ben-David, 2014]). We show that $\mathcal{C}$ is actively recoverable if and only if $\mathcal{H}$ has finite *coslicing dimension*, a combinatorial quantity similar to the star number of Hanneke and Yang [2015] and the slicing dimension of Kivinen [1995]. We also show that, for a wide family of concept classes in $\mathbb{R}^m$, a finite coslicing dimension is equivalent to a positive one-versus-all margin.

Let $\mathcal{X}$ be any domain, $X \subset \mathcal{X}$ any finite set, and $\mathcal{C} = (C_1, \ldots, C_k)$ a clustering of X. Let $\mathcal{H}$ be any concept class over $\mathcal{X}$. We say that $\mathcal{C}$ is realized by $\mathcal{H}$ if for all $i \in [k]$ there is some $h_i \in \mathcal{H}$ such that $C_i = X \cap h_i$. For example, the ellipsoidal clusters of Bressan et al. [2020] can be formulated by letting $\mathcal{X} = \mathbb{R}^m$ and letting $\mathcal{H}$ to be the family of all ellipsoids in $\mathbb{R}^m$, and the convex clusters of Section 3 can be formulated by letting $\mathcal{X} = \mathbb{R}^m$ and letting $\mathcal{H}$ to be the family of all polytopes in $\mathbb{R}^m$. Clearly, we expect the number of queries needed to recover $\mathcal{C}$ to depend on $\mathcal{H}$. This leads us to the question: what can we say about the recoverability of $\mathcal{C}$ in terms of $\mathcal{H}$?

5.1 The coslicing dimension

We show that the number of queries needed to actively recover $\mathcal{C}$ depends on a combinatorial quantity that we call *coslicing dimension* of $\mathcal{H}$.

Definition 5. *We say that $\mathcal{H}$ coslices $X \subseteq \mathcal{X}$ if for all $x \in X$ there exist two concepts $h_x^+, h_x^- \in \mathcal{H}$ such that $X \cap h_x^+ = \{x\}$ and $X \cap h_x^- = X \setminus \{x\}$. The coslicing dimension of $\mathcal{H}$ is:*

$$\operatorname{cosl}(\mathcal{H}) = \sup\{|X| : X \text{ is cosliced by } \mathcal{H}\}$$

If $\mathcal{H}$ coslices arbitrarily large sets then we let $\operatorname{cosl}(\mathcal{H}) = \infty$.

For instance, let $\mathcal{X} = \mathbb{R}^m$. If $\mathcal{H}$ is the class of linear separators, then $\operatorname{cosl}(\mathcal{H}) = \infty$ (take X to be the set of vertices of an n-vertex polytope and use the hyperplane separator theorem). If instead $\mathcal{H}$ is the class of axis-aligned boxes, it can be shown that $\operatorname{cosl}(\mathcal{H}) = 2m$. Our main result is:

Theorem 4. *If $\operatorname{cosl}(\mathcal{H}) < \infty$ then there is an algorithm that, given any n-point instance whose latent clustering $\mathcal{C}$ is realized by $\mathcal{H}$, recovers $\mathcal{C}$ with $\mathcal{O}(\operatorname{cosl}(\mathcal{H})\, k \log k \log n)$ label queries with high probability. Moreover, for any algorithm $\mathcal{A}$ there are instances on $\operatorname{cosl}(\mathcal{H})$ points whose latent clustering $\mathcal{C}$ is realized by $\mathcal{H}$ where $\mathcal{A}$ makes $\Omega(\operatorname{cosl}(\mathcal{H}))$ label queries in expectation to return $\mathcal{C}$. As a consequence, if $\operatorname{cosl}(\mathcal{H}) = \infty$ then in the worst case any algorithm needs $\Omega(n)$ label queries in expectation to recover an n-point clustering realized by $\mathcal{H}$.*

Sketch of the proof. We follow the same ideas of the proof of Theorem 3. For the lower bound, we take any X cosliced by $\mathcal{H}$ with $|X| = \operatorname{cosl}(\mathcal{H})$, and we draw a random clustering of X in the form $(x, X \setminus x)$.

For the upper bound, for each $i \in [k]$ we define the effective concept class:

$$H_i = \{C \,:\, C = C_i' \wedge (C_1', \ldots, C_k') \in P_k(X)\} \tag{5}$$

where $P_k(X)$ is the set of all clusterings of X realized by $\mathcal{H}$. As observed in the proof of Lemma 3, we have the general relationship $\text{vc-dim}(I(H_i), X) \leq \operatorname{sl}(H_i, X)$ whenever $\operatorname{sl}(H_i, X) < \infty$. Therefore, if we show that $\operatorname{sl}(H_i, X) \leq \operatorname{cosl}(\mathcal{H})$, by drawing a labeled sample of size $\Theta(\operatorname{cosl}(\mathcal{H})\, k \log k)$ we can recover the labels of an expected constant fraction of X, as in the proof of Lemma 3. To prove that $\operatorname{sl}(H_i, X) \leq \operatorname{cosl}(\mathcal{H})$, let $U = \{x_1, \ldots, x_\ell\} \subseteq X$ be sliced by H_i. By construction of H_i, there are ℓ clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_\ell$ realized by $\mathcal{H}$ and such that $\mathcal{C}_j = (x_j, U \setminus x_j)$ for all $j \in [\ell]$. This implies that U is cosliced by $\mathcal{H}$. Hence, $|U| \leq \operatorname{cosl}(\mathcal{H})$ and so $\operatorname{sl}(H_i, X) \leq \operatorname{cosl}(\mathcal{H})$. □

Remark. In the case of convex clusters in $\mathbb{R}^m$, we would let $\mathcal{H}$ be the class of all convex polytopes, obtaining $\operatorname{cosl}(\mathcal{H}) = \infty$ and thus Theorem 4 would not provide any useful bound. This is true even if $\mathcal{C}$ has convex hull margin $\gamma > 0$, although by Theorem 1 we know that $\mathcal{C}$ can be recovered with $\mathcal{O}(\log n)$ queries. The same holds for the one-versus-all margin. This is because we defined the coslicing dimension as a function of $\mathcal{H}$, through which we cannot capture margin properties — indeed, the margin depends on the instance (X, O) rather than on $\mathcal{H}$. We note however that one could fix this by redefining the coslicing dimension in the form $\operatorname{cosl}(\mathcal{H}, \mathcal{I})$, where $\mathcal{I}$ is a class of instances, and adapt Theorem 4 correspondingly. Then, under the assumption that all the instances $(X, O) \in \mathcal{I}$ have margin $\gamma > 0$, one could bound $\operatorname{cosl}(\mathcal{H}, \mathcal{I})$ as a function of γ, recovering the same type of bounds of Theorem 1 and Theorem 3.

5.2 The one-versus-all margin, again!

We look again at the case $\mathcal{X} = \mathbb{R}^m$. Consider any concept class $\mathcal{H}$. Theorem 4 and the remark above say that, if $\operatorname{cosl}(\mathcal{H}, \mathcal{I}) < \infty$ where $\mathcal{I}$ is the class of allowed instances, then the latent clustering of any instance can be recovered with $\mathcal{O}(\log n)$ queries. Note that we are not using any notion of margin here — only the finiteness of $\operatorname{cosl}(\mathcal{H}, \mathcal{I})$.

Here we show that, for a wide family of concept classes in $\mathbb{R}^m$, a finite coslicing dimension and a positive one-versus-all margin are actually equivalent. This equivalence is established by proving that the $\mathcal{O}(\log n)$ query bound is attainable *if and only* if the instances have positive one-versus-all margin. In what follows, we assume that $\mathcal{H}$ satisfies:

Definition 6. *A concept class $\mathcal{H}$ in $\mathbb{R}^m$ is non-fractal if there is $h \in \mathcal{H}$ such that both h and its complement contain a ball of positive radius.*

This assumption avoids pathological cases; for instance, $\mathcal{H}$ containing only hypotheses that are affine transformations of Cantor's set. Our result is:

Theorem 5. *Let $\mathcal{H}$ be a concept class in $\mathbb{R}^m$ that is non-fractal and closed under affine transformations. There is an algorithm that, given any instance whose latent clustering $\mathcal{C}$ has one-versus-all margin γ and is realized by $\mathcal{H}$, returns $\mathcal{C}$ while making $\mathcal{O}(Mk \log k \log n)$ label queries with high probability, where $M = \max\left(2, (1 + 4/\gamma)^m\right)$. Moreover, for any algorithm $\mathcal{A}$, there exist arbitrarily large n-point instances, whose latent clustering $\mathcal{C}$ has arbitrarily small one-versus-all margin and is realized by $\mathcal{H}$, where $\mathcal{A}$ makes $\Omega(n)$ label queries in expectation to recover $\mathcal{C}$.*

Sketch of the proof. The upper bounds follow by Theorem 3 and the packing number of $\mathbb{R}^m$. For the lower bounds, we show that arbitrarily large packings of a sphere are cosliced by $\mathcal{H}$. We use Figure 2 for reference. Let $h \in \mathbb{R}^m$ be such that both h and its complement $\overline{h}$ contain a ball of positive radius. Then, for any $\rho > 0$ there exist a ball $B = B(c, r) \subseteq h$ with $r > 0$, and a point $x \in \overline{h}$ such that $d(B, x) \leq \rho$. Now take a sphere S of radius $r' \ll r$ with center on the segment xc. Let $\eta = \sup_{y \in S \setminus B} d(x, y)$, and let X be an η-packing of S, that is, a subset of points of S such that $d(x', x'') > \eta$ for all distinct $x', x'' \in X$. Note how this implies that every $x' \in X \setminus x$ necessarily lies in $S \cap B$, and therefore, in $S \cap h$. Moreover, by letting $\eta/r' \to 0$, we can take X arbitrarily large. Since $\mathcal{H}$ is closed under affine transformations, by rotating X it follows that for every $x \in X$ there is $h_x \in \mathcal{H}$ such that $X \cap h_x = X \setminus x$. By applying the same argument to $\overline{h}$ and by complementation we can show that $X \cap h'_x = \{x\}$ for some $h'_x \in \mathcal{H}$ for all $x \in X$ as well. Hence $\mathcal{H}$ coslices arbitrarily large sets. To conclude, invoke Theorem 4. □

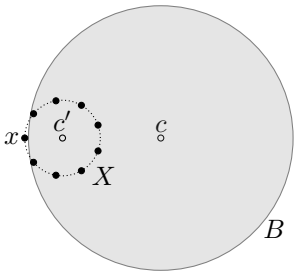

Figure 2: The η-packing X is in B, and thus in h, except for x that lies in $\overline{h}$. By taking B arbitrarily close to x, we can make η arbitrarily small and thus X arbitrarily large.

Note that Theorem 5 applies to several basic concept classes $\mathcal{H}$. For instance, when $\mathcal{H}$ is the class of all linear separators, the class of all ellipsoids, the class of all polytopes, and the class of all convex bodies (bounded or not, and possibly degenerate). It also includes more complex classes with non-convex concepts, for instance, the class of all finite or infinite disjoint unions of balls, polytopes, or convex bodies.

6 Future Work

An interesting direction for future work is to explore the power of different types of queries to avoid the exponential dependency on the dimensionality of the ambient space. Another open problem is to understand if our margin conditions are sufficient to obtain PTASes for k-means and other center-based clustering problems. A third open problem is to close the gap in the exponential dependence on the dimension in the Euclidean case, which is $(1+1/\gamma)^m$ in our upper bound and roughly $(1+1/\gamma)^{m/2}$ in the lower bound of Bressan et al. [2020].

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge partial support by the Google Focused Award "Algorithms and Learning for AI" (ALL4AI). Nicolò Cesa-Bianchi is also supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR) and by the EU Horizon 2020 ICT-48 research and innovation action under grant agreement 951847, project ELISE (European Learning and Intelligent Systems Excellence).

References

N. Ailon, A. Bhattacharya, and R. Jaiswal. Approximate correlation clustering using same-cluster queries. In *Proc. of LATIN*, pages 14–27, 2018a.

N. Ailon, A. Bhattacharya, R. Jaiswal, and A. Kumar. Approximate clustering with same-cluster queries. In *Proc. of ITCS*, volume 94, pages 40:1–40:21, 2018b.

H. Ashtiani, S. Kushagra, and S. Ben-David. Clustering with same-cluster queries. In *Advances in Neural Information Processing Systems*, volume 29, pages 3216–3224, 2016.

P. Awasthi, A. Blum, and O. Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(1):49–54, 2012.

M.-F. Balcan and P. Long. Active and passive learning of linear separators under log-concave distributions. In *Proc. of COLT*, pages 288–316, 2013.

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proc. of COLT*, pages 35–50, 2007.

A. Basu and T. Oertel. Centerpoints: A link between optimization and convex geometry. *SIAM Journal on Optimization*, 27(2):866–889, 2017.

S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. Long. Characterizations of learnability for classes of $\{0, ..., n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

D. Bertsimas and S. Vempala. Solving convex programs by random walks. *J. ACM*, 51(4):540–556, July 2004.

Y. Bilu and N. Linial. Are stable instances easy? *Comb. Probab. Comput.*, 21(5):643–660, Sept. 2012.

M. Bressan, N. Cesa-Bianchi, S. Lattanzi, and A. Paudice. Exact recovery of mangled clusters with same-cluster queries. In *Advances in Neural Information Processing Systems*, volume 33, pages 9324–9334, 2020.

M. Bressan, N. Cesa-Bianchi, S. Lattanzi, and A. Paudice. Exact recovery of clusters in finite metric spaces using oracle queries. In *Proc. of COLT*, volume 134, pages 775–803, 2021.

A. Daniely and S. Shalev-Shwartz. Optimal learners for multiclass problems. In *Proc. of COLT*, volume 35, pages 287–316, 2014.

S. Dasgupta. Analysis of a greedy active learning strategy. In *Proc. of NIPS*, page 337–344, 2004.

R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2), 2012.

D. Firmani, S. Galhotra, B. Saha, and D. Srivastava. Robust entity resolution using a crowd oracle. *IEEE Data Eng. Bull.*, 41(2):91–103, 2018.

B. Gamlath, S. Huang, and O. Svensson. Semi-supervised algorithms for approximately optimal and accurate clustering. In *Proc. of ICALP*, pages 57:1–57:14, 2018.

A. Giannopoulos. Notes on isotropic convex bodies, 2003. URL `http://users.uoa.gr/~apgiannop/isotropic-bodies.pdf`.

A. Gruenheid, B. Nushi, T. Kraska, W. Gatterbauer, and D. Kossmann. Fault-tolerant entity resolution with the crowd. *arXiv preprint arXiv:1512.00537*, 2015.

S. Hanneke and L. Yang. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16(12): 3487–3602, 2015.

M. Hopkins, D. Kane, and S. Lovett. The power of comparisons for actively learning linear classifiers. In *Advances in Neural Information Processing Systems*, volume 33, pages 6342–6353, 2020a.

M. Hopkins, D. Kane, S. Lovett, and G. Mahajan. Noise-tolerant, reliable active classification with comparison queries. In *Proc. of COLT*, volume 125, pages 1957–2006, 2020b.

M. Hopkins, D. Kane, S. Lovett, and M. Moshkovitz. Bounded memory active learning through enriched queries. In *Proc. of COLT*, pages 2358–2387, 2021.

W. Huleihel, A. Mazumdar, M. Médard, and S. Pal. Same-cluster querying for overlapping clusters. In *Advances in Neural Information Processing Systems*, volume 32, pages 10485–10495, 2019.

D. M. Kane, S. Lovett, S. Moran, and J. Zhang. Active classification with comparison queries. In *Proc. of IEEE FOCS*, pages 355–366, 2017.

L. G. Khachiyan. Rounding of polytopes in the real number model of computation. *Mathematics of Operations Research*, 21(2):307–320, 1996.

J. Kivinen. Learning reliably and with one-sided error. *Mathematical Systems Theory*, 28(2):141–172, 1995.

A. Kupavskii. The VC-dimension of k-vertex d-polytopes. *Combinatorica*, 40(6):869–874, 2020.

L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006.

A. Mazumdar and S. Pal. Semisupervised Clustering, AND-Queries and Locally Encodable Source Coding. In *Advances in Neural Information Processing Systems*, volume 30, pages 6489–6499, 2017.

A. Mazumdar and B. Saha. Query complexity of clustering with side information. In *Advances in Neural Information Processing Systems*, volume 30, pages 4682–4693. 2017a.

A. Mazumdar and B. Saha. Clustering with noisy queries. In *Advances in Neural Information Processing Systems*, volume 30, pages 5788–5799, 2017b.

A. McCallum and K. Nigam. Employing em and pool-based active learning for text classification. In *Proc. of ICML*, page 350–358, 1998.

M. Naszódi. Approximating a convex body by a polytope using the epsilon-net theorem. *Discrete & Computational Geometry*, 61(3):686–693, Mar. 2018.

L. A. Rademacher. Approximating the centroid is hard. In *Proc. of ACM SoCG*, page 302–305, 2007.

R. L. Rivest and R. Sloan. Learning complicated concepts reliably and usefully. In *Proc. of COLT*, pages 69–79, 1988.

B. I. Rubinstein, P. L. Bartlett, and J. H. Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *Journal of Computer and System Sciences*, 75(1):37–59, 2009.

B. Saha and S. Subramanian. Correlation Clustering with Same-Cluster Queries Bounded by Optimal Cost. In *Proc. of ESA*, pages 81:1–81:17, 2019.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

V. Verroios and H. Garcia-Molina. Entity resolution with crowd errors. In *Proc. of IEEE ICDE*, pages 219–230, 2015.

V. Verroios, H. Garcia-Molina, and Y. Papakonstantinou. Waldo: An adaptive human interface for crowd entity resolution. In *Proc. of ACM SIGMOD*, pages 1133–1148, 2017.

F. Vitale, A. Rajagopalan, and C. Gentile. Flattening a hierarchical clustering through active learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 15263–15273, 2019.

L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. *Advances in Neural Information Processing Systems*, pages 1537–1544, 2004.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The assumptions and conditions under which our results hold are clearly specified in the paper.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a purely theoretical paper and does not imply any foreseeable societal consequence.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

On Margin-Based Cluster Recovery with Oracle Queries (Supplementary Material)

Marco Bressan
Dept. of CS, Univ. of Milan, Italy
marco.bressan@unimi.it

Nicolò Cesa-Bianchi
DSRC & Dept. of CS, Univ. of Milan, Italy
nicolo.cesa-bianchi@unimi.it

Silvio Lattanzi
Google
silviol@google.com

Andrea Paudice
Dept. of CS, Univ. of Milan, Italy &
Istituto Italiano di Tecnologia, Italy
andrea.paudice@unimi.it

1 Proof of Lemma 1

Lemma 1 (Convex hull expansion trick)**.** *Fix $\gamma > 0$, and let $s = \Theta\big(m^5\big(1 + 1/\gamma\big)^m \log\big(1 + 1/\gamma\big)\big)$ large enough. Let S_C be a sample of s independent uniform random points from some cluster C, and let $K = \mathrm{conv}(S_C)$. Let $X_1, \dots, X_N$ be independent random points sampled ε-uniformly from K, with $\varepsilon \in \Theta(m^{-1})$ small enough and $N \in \Theta(m^2)$ large enough, and let $z = \frac{1}{N}\sum_{i=1}^N X_i$. Finally, let $Q = (1+\gamma)K$ where the center of the scaling is z. Then $\mathbb{P}\big(|Q \cap C| \geq |C|/2\big) \geq 1/2$.*

Preliminaries. Without loss of generality, we assume that K has full rank (as one can always work in the subspace spanned by S_C, which can be computed in time $\mathcal{O}(|S_C|m)$). For technical reasons, we let $\vartheta = \frac{1}{1+\gamma} \in (0,1)$ and prove the lemma for $s = \Omega\big(\frac{m^5}{\vartheta(1-\vartheta)^m} \ln \frac{1}{\vartheta(1-\vartheta)}\big)$ large enough. To see that any $s \in \Theta\big(m^5\big(1 + 1/\gamma\big)^m \ln\big(1 + 1/\gamma\big)\big)$ satisfies this assumption, first substitute ϑ to get:

$$\frac{1}{\vartheta(1-\vartheta)^m} \ln \frac{1}{\vartheta(1-\vartheta)} = (1+\gamma)\left(\frac{1+\gamma}{\gamma}\right)^m \ln \frac{(1+\gamma)^2}{\gamma} \tag{6}$$

which is in $\mathcal{O}\big((1 + 1/\gamma)^m (1+\gamma)\ln(1 + 1/\gamma)\big)$. Now note that $(1+\gamma)\ln\big(1 + 1/\gamma\big)$ is bounded by $\mathcal{O}\big(\gamma \cdot 1/\gamma\big) = \mathcal{O}(1)$ for $\gamma > 1$, and by $2\ln\big(1 + 1/\gamma\big) = \mathcal{O}\big(\ln 1/\gamma\big)$ when $\gamma \leq 1$. Hence, in any case the term $(1+\gamma)\ln(1 + 1/\gamma)$ is in $\mathcal{O}(\ln(1 + 1/\gamma))$. Therefore:

$$\big(1 + 1/\gamma\big)^m \ln\big(1 + 1/\gamma\big) = \Omega\left(\frac{1}{\vartheta(1-\vartheta)^m} \ln \frac{1}{\vartheta(1-\vartheta)}\right) \tag{7}$$

as claimed.

Before starting with the actual proof, we introduce some further definitions and notation.

Definition 7. *A convex body $K \subset \mathbb{R}^m$ is in* isotropic position[4] *if it has center of mass in the origin, $\int_K x\,dx = 0$, and moment of inertia 1 in every direction, $\int_K \langle x, u\rangle^2\,dx = 1$ for all $u \in S^{m-1}$.*

We define the norm $\|\cdot\|_K = \|f(\cdot)\|_2$ where $f = f_K$ is the unique affine transformation such that $f(K)$ is in isotropic position. We let $R(K) = \sup_{x\in K}\|x\|_2$ denote the Euclidean radius of K, and we let $R_K(K) = \sup_{x\in K}\|x\|_K$ denote the isotropic radius of K. We also let $d_K(x,y) = \|x-y\|_K$ be the isotropic distance of K.

Now, the proof has two steps. First, we show that the point $z = \frac{1}{N}\sum_{i=1}^N X_i$ is close to the centroid μ_K of K, according to $d_K(\cdot)$, with good probability. Second, we show that if this is the case, then

[4] Not to be confused with the definition of Giannopoulos [2003], where the assumption $\int_K \langle x, u\rangle^2\,dx = 1$ is replaced by $\mathrm{vol}(K) = 1$.

$\frac{K}{\vartheta}$, where the scaling is meant about z, contains a polytope P which contains K and thus S_C, and which belongs to a class with VC dimension $\mathcal{O}\big(\frac{m^5}{\vartheta(1-\vartheta)^m}\big)$. By standard PAC bounds this implies that $|P \cap C| \geq \frac{1}{2}|C|$, with a probability that can be made arbitrarily close to 1 by adjusting the constants.

Step 1: z is close to the centroid of K

Let μ_K be the center of mass of K. We prove:

Lemma 4. *Fix any $\eta, p \geq 0$, and choose any $\varepsilon \leq \frac{\eta}{4(m+1)}$ and $N \geq \frac{16(m+1)^2}{p^2\eta^2}$. If $X_1, \ldots, X_N$ are drawn independently and ε-uniformly at random from K, and $\overline{X} = \frac{1}{N}X_i$, then:*

$$\mathbb{P}\left(d_K\big(\overline{X}, \mu_K\big) \leq \eta\right) \geq 1 - p \tag{8}$$

For the proof of Lemma 4, we need two ancillary results.

Lemma 5. $R_K(K) \leq m + 1$.

Proof. Consider K in isotropic position, and let $K' = \vartheta K$ where $\vartheta = \mathrm{vol}(K)^{-1/m}$, so that $\mathrm{vol}(K') = 1$. Then, K' is in isotropic position according to the definition of Giannopoulos [2003]. In this case, [Giannopoulos, 2003, Theorem 1.2.4] implies $R(K') \leq (m+1)L_K$, where L_K is the *isotropic constant* which, for all $u \in S^{m-1}$, satisfies $\int_{K'} \langle x, u\rangle^2 \, dx = L_K^2$. Since $K' = \vartheta K$ and $\int_K \langle x, u\rangle^2 \, dx = 1$ by the isotropy of K, we have $L_K = \vartheta$. Hence $R(K') \leq (m+1)\vartheta$, that is, $R(K) \leq m + 1$. □

Lemma 6. *If X is drawn from an ε-uniform distribution over K, then $\|\mathbb{E}X\|_K \leq 2\varepsilon(m+1)$.*

Proof. Since X is ε-uniform over K, there exists a coupling (X, Y) with $\mathbb{P}(X \neq Y) \leq \varepsilon$ and Y uniform over K. Since $\|\mathbb{E}Y\|_K = 0$, we have:

$$\|\mathbb{E}X\|_K = \|\mathbb{E}[X - Y]\|_K \leq \mathbb{P}(X \neq Y) \sup_{x,y \in K} d_K(x, y) \leq \varepsilon \, 2R_K(K) \leq 2\varepsilon(m+1) \tag{9}$$

where the last inequality is given by Lemma 5. □

Proof of Lemma 4. For the sake of the analysis we look at K from its isotropic position. Note that the X_i are still ε-uniform over K, since the affine map that places K in isotropic position preserves volume ratios. The claim becomes:

$$\mathbb{P}\big(\|\overline{X}\|_2 \leq \eta\big) \geq 1 - p \tag{10}$$

Now, $\|\overline{X}\|_2 \leq \|\mathbb{E}\overline{X}\|_2 + \|\overline{X} - \mathbb{E}\overline{X}\|_2$. Thus, we show that $\|\mathbb{E}\overline{X}\|_2 \leq \frac{\eta}{2}$, and that $\|\overline{X} - \mathbb{E}\overline{X}\|_2 \leq \frac{\eta}{2}$ with probability at least $1 - p$. For the first part, by Lemma 6, and since $\varepsilon \leq \frac{\eta}{4(m+1)}$, we obtain:

$$\|\mathbb{E}\overline{X}\|_2 = \|\mathbb{E}X_i\|_2 \leq 2\varepsilon(m+1) \leq \frac{\eta}{2} \tag{11}$$

For the second part, by Lemma 5 we have $\|X_i\|_2 \leq m+1$ for all i. Therefore, if we let $Y_i = X_i - \mathbb{E}X_i$ for all i, we have $\|Y_i\|_2 \leq 2(m+1)$ and thus $\|Y_i\|_2^2 \leq 4(m+1)^2$. Now let $\overline{Y} = \frac{1}{N}\sum_{i=1}^N Y_i$. Since the Y_i are independent and with $\mathbb{E}Y_i = 0$, then $\mathbb{E}\langle Y_i, Y_j\rangle = 0$ whenever $i \neq j$ and therefore:

$$\mathbb{E}\|\overline{Y}\|_2^2 = \mathbb{E}\left(\frac{1}{N^2}\sum_{i,j=1}^N \langle Y_i, Y_j\rangle\right) = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\|Y_i\|_2^2 \leq \frac{4(m+1)^2}{N} \tag{12}$$

Plugging in our value $N \geq \frac{16(m+1)^2}{p^2\eta^2}$, and using Jensen's inequality, we obtain:

$$\big(\mathbb{E}\|\overline{Y}\|_2\big)^2 \leq \mathbb{E}\|\overline{Y}\|_2^2 \leq \frac{p^2\eta^2}{4} \tag{13}$$

Therefore $\mathbb{E}\|\overline{Y}\|_2 \leq \frac{p\eta}{2}$, which by Markov's inequality implies that $\mathbb{P}\big(\|\overline{Y}\|_2 > \frac{\eta}{2}\big) < p$. By noting that $\overline{Y} = \overline{X} - \mathbb{E}\overline{X}$, the proof is complete. □

Step 2: showing a tight enclosing polytope

We prove:

Lemma 7. *Let $z \in \mathbb{R}^m$ such that $d_K(z, \mu_K) \leq \frac{1}{e} - \frac{1}{3}$. For any $\vartheta \in (0,1)$ there exists a polytope P on $t = \mathcal{O}\big(\frac{m^2}{\vartheta(1-\vartheta)^m}\big)$ vertices such that $K \subseteq P \subseteq \frac{K}{\vartheta}$, where the scaling is about z.*

For the proof, we need some ancillary results.

Theorem 6 (Bertsimas and Vempala [2004], Theorem 3)**.** *Let K be a convex set in isotropic position and z be a point at distance t from its centroid. Then any halfspace containing z contains at least $\frac{1}{e} - t$ of the volume of K.*

Now we adapt a result by Naszódi [2018], recalled here for convenience. We say that a halfspace F supports a convex body from outside if F intersects the boundary of the body, but not its interior.

Lemma 8 (Lemma 2.1, Naszódi [2018])**.** *Let $0 < \vartheta < 1$, and F be a halfspace that supports ϑK from outside, where the scaling is about μ_K. Then:*

$$\mathrm{vol}(K \cap F) \geq \mathrm{vol}(K) \cdot (1-\vartheta)^m \frac{1}{e} \tag{14}$$

Our adaptation is:

Lemma 9. *Let $z \in \mathbb{R}^m$, let $0 < \vartheta < 1$, and let F be a half-space that supports ϑK from outside, where the scaling is about z. Then:*

$$\mathrm{vol}(K \cap F) \geq \mathrm{vol}(K) \cdot (1-\vartheta)^m \left(\frac{1}{e} - d_K(\mu_K, z)\right) \tag{15}$$

Proof. The proof is similar to the proof of the original lemma. See Figure 3 for reference. Let F_0 be a translate of F whose boundary contains z, and let $K_0 = K \cap F_0$. Let F_1 be a translate of F that supports K from outside, and let $p \in F_1 \cap K$. Now consider $K_0' = \vartheta p + (1-\vartheta)K_0$, that is, the homothetic copy of K_0 with center p and ratio $1-\vartheta$. The crucial observation is that $F = \vartheta p + (1-\vartheta)F_0$, which implies $K_0' \subset K \cap F$. Clearly $\mathrm{vol}(K_0') = (1-\vartheta)^m \,\mathrm{vol}(K_0)$. Moreover, by Theorem 6 we have $\mathrm{vol}(K_0) \geq \mathrm{vol}(K)(1/e - t)$ where $t = d_K(\mu_K, z)$; this holds because mapping K to its isotropic position preserves volume ratios. This concludes the proof. □

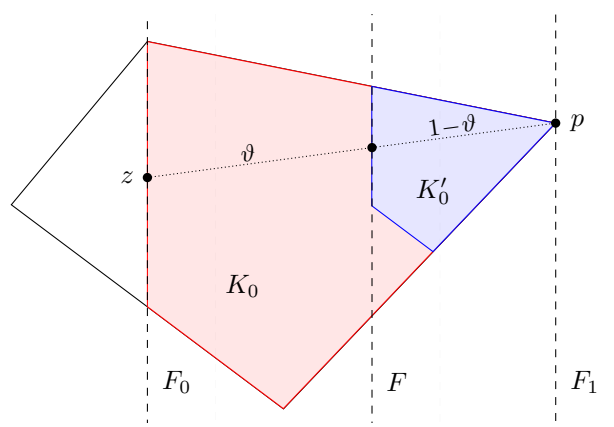

Figure 3: a visual proof of Lemma 9, with $d(z,p) = 1$ for simplicity.

Proof of Lemma 7. We adapt the construction behind [Naszódi, 2018, Theorem 1.2], by replacing $\varepsilon = \frac{(1-\vartheta)^m}{e}$ with $\varepsilon = \frac{(1-\vartheta)^m}{3}$. The theorem then says that, if we set:

$$t = \left\lceil C \frac{(m+1)3}{(1-\vartheta)^m} \ln \frac{3}{(1-\vartheta)^m} \right\rceil \tag{16}$$

and we let $X_1, \dots, X_t$ be t points chosen independently and uniformly at random from K, and let $P = \mathrm{conv}(X_1, \dots, X_t)$, then $\vartheta K \subseteq P \subseteq K$ with probability at least $1 - p$, where

$$p = 4 \left(11C^2 \left(\frac{(1-\vartheta)^m}{3} \right)^{C-2} \right)^{m+1} \tag{17}$$

Now, we choose $C = \Theta(\frac{1}{\vartheta} \ln \frac{1}{\vartheta})$ large enough. On the one hand, we obtain:

$$t = \left\lceil C \frac{(m+1)3}{(1-\vartheta)^m} \ln \frac{3}{(1-\vartheta)^m} \right\rceil = \mathcal{O}\left(\frac{m^2}{\vartheta(1-\vartheta)^m} \ln \frac{1}{\vartheta} \ln \frac{1}{1-\vartheta} \right) \tag{18}$$

Since $\vartheta \in (0,1)$, we have $\ln \frac{1}{\vartheta} \ln \frac{1}{1-\vartheta} = \ln(1 + 1/x) \ln(1+x)$ where $x = \frac{\vartheta}{1-\vartheta} > 0$. However, $\ln(1 + 1/x) \ln(1+x) < 1$ for all $x > 0$. Therefore, $t \in \mathcal{O}\big(\frac{m^2}{\vartheta(1-\vartheta)^m}\big)$. On the other hand, by setting C large enough we can make $C^2 \big(\frac{(1-\vartheta)^m}{3}\big)^{C-2}$ arbitrarily small, and therefore $p < 1$.

Since $p < 1$, we conclude that *there exists* a polytope P on $t \in \mathcal{O}\big(\frac{m^2}{\vartheta(1-\vartheta)^m}\big)$ vertices such that $\vartheta K \subseteq P \subseteq K$. To conclude, instead of P simply take $\frac{P}{\vartheta}$ where the scaling is about z. □

Wrap-up

First, by Lemma 4, by taking $N \in \mathcal{O}(m^2)$ large enough we can make $d_K(\mu_K, z) \leq \frac{1}{e} - \frac{1}{3}$ with probability arbitrarily close to 1. Now let $\mathcal{P}_t$ be the family of all polytopes in $\mathbb{R}^m$ on at most t vertices. For $t \in \mathcal{O}\big(\frac{m^2}{\vartheta(1-\vartheta)^m}\big)$ large enough, Lemma 7 implies that there exists some $P \in \mathcal{P}_t$ such that $K \subseteq P \subseteq \frac{K}{\vartheta}$.

Now we prove that, by choosing s large enough, with probability arbitrarily close to 1 we have $|\frac{K}{\vartheta} \cap C| \geq \frac{1}{2}|C|$. First, by a recent result of Kupavskii [2020], we have $\text{vc-dim}(\mathcal{P}_t) \leq 8m^2 t \log_2 t$. For our t this yields

$$\text{vc-dim}(\mathcal{P}_t) = \mathcal{O}\bigg(m^2 \frac{m^2}{\vartheta(1-\vartheta)^m} \ln \frac{m^2}{\vartheta(1-\vartheta)^m} \bigg) = \mathcal{O}\bigg(\frac{m^5}{\vartheta(1-\vartheta)^m} \ln \frac{1}{\vartheta(1-\vartheta)} \bigg) \tag{19}$$

where in the last equality we used $\ln \frac{m^2}{\vartheta(1-\vartheta)^m} = \ln \frac{m^{(2/m)m}}{\vartheta(1-\vartheta)^m} \leq m \ln \frac{m^{2/m}}{\vartheta(1-\vartheta)} = \mathcal{O}\big(m \ln \frac{1}{\vartheta(1-\vartheta)}\big)$. Hence, by choosing $|S_C| = s = \mathcal{O}\big(\frac{m^5}{\vartheta(1-\vartheta)^m} \ln \frac{1}{\vartheta(1-\vartheta)}\big)$ large enough, for any constant c, ε, δ we can make:

$$|S_C| \geq c \, \frac{\text{vc-dim}(\mathcal{P}_t) \, \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta}}{\varepsilon} \tag{20}$$

Since P is consistent with S_C, that is, $P \supset S_C$, then by standard PAC bounds we have $|P \cap C| \geq (1-\varepsilon)|C|$ with probability at least $1 - \delta$. But $P \subseteq \frac{K}{\vartheta}$, and therefore $|\frac{K}{\vartheta} \cap C| \geq (1-\varepsilon)|C|$ with probability at least $1 - \delta$. By adjusting the constants this yields the thesis of Lemma 1.

2 Proof of Theorem 1

Theorem 1. *Let (X, O) be an instance whose latent clustering $\mathcal{C}$ has convex hull margin $\gamma > 0$. Then* CHEATREC(X, O, γ) *outputs $\mathcal{C}$, runs in time* $\mathrm{poly}(k, n, m)$*, and with high probability makes a number of label queries to O bounded by $\mathcal{O}\big(k^2 m^5 \, (1 + 1/\gamma)^m \log(1 + 1/\gamma) \log n\big)$.*

We give the pseudocode of the algorithm for reference. First, we prove the correctness and the query bound. Then we show the running time bound. Note that, for readability, the pseudocode given here is high-level; the actual implementation is more complex, see below.

Correctness and query bound. We prove that, at each round, for some i we recover at least half of the points in C_i with probability $1 - \delta$, where δ can be made arbitrarily small by adjusting the constants. Let S_i be the subset of the sample S having label i. Since there are at most k clusters and $|S| = ks$, for some i we will have $|S_i| \geq s$. Now we apply Lemma 1 to $K = \mathrm{conv}(S_i)$. Since s satisfies the hypotheses, the lemma says that $\widehat{C_i} = Q \cap X$ has size $|\widehat{C_i}| \geq \frac{|C_i|}{2}$ with probability

Algorithm 1 HULLTRICK$(K, 1+\gamma)$

let $N = \Omega(m^2)$ large enough
let $\varepsilon = \mathcal{O}(m^{-1})$ small enough
draw N i.i.d. random points $X_1, \ldots, X_N$ from any ε-uniform distribution over K
let $z = \frac{1}{N}\sum_{i=1}^{N} X_i$
return $z + (1+\gamma)(K - z)$

Algorithm 2 CHEATREC(X, O, γ)

while $X \neq \emptyset$ **do**
  let $s = \mathcal{O}\big(m^5 \big(1 + {}^1\!/\!_\gamma\big)^m \ln\big(1 + {}^1\!/\!_\gamma\big)\big)$ large enough
  draw a uniform random sample S of size $\min(|X|, ks)$ from X, without repetition
  learn the labels of S with ks queries to O
  let S_i be the points of S having label i
  for $i = 1, \ldots, k$ **do**
    $K = \mathrm{conv}(S_i)$
    $Q =$ HULLTRICK$(K, 1+\gamma)$
    $\widehat{C}_i = Q \cap X$
    label all points of $\widehat{C}_i$ with label i
    $X = X \setminus \widehat{C}_i$

arbitrarily close to 1 (that is, with probability $1 - \delta$ as above). It remains to show that $\widehat{C_i} \subseteq C_i$. Let d be any pseudometric that is homogeneous and invariant under translation. Then, any point $y \in (1+\gamma)K \cap X$ satisfies $d(y, K) \leq \gamma\, \phi_d(K)$. But $K = \mathrm{conv}(S_i) \subseteq \mathrm{conv}(C_i)$. Therefore $\phi_d(K) \leq \phi_d(C_i)$. Hence $d(y, \mathrm{conv}(C_i)) \leq \gamma\, \phi_d(C_i)$. By the convex margin assumption, this implies that $y \in C_i$. This also proves the correctness of the algorithm. The total query bound follows as in Lemma 3 of [Bressan et al., 2020], whose algorithm also recovers an expected fraction $\frac{1}{4k}$ of all points in each round.

Running time bound. First, we analyze the running time of CHEATREC excluding the call to HULLTRICK. Drawing and labeling the samples obviously cost $\mathcal{O}(n)$ time throughout the entire execution. In the main loop, K is actually not computed explicitly — see below. Similarly, Q is simply a set of points obtained by rescaling S_i about some point in space. Thus, computing $\widehat{C}_i = Q \cap X$ and labeling its points boils down to deciding, for all $x \in X$, if x can be written as $\sum_{x_j \in Q_j} \lambda_j x_j$ for a set of coefficients $\lambda_1, \ldots, \lambda_{|Q|} \in [0, 1]$. This can be done with polynomial precision using any polynomial-time solver for linear programs (say, the ellipsoid method).

Let us now turn to HULLTRICK. The computationally expensive part is drawing N points from an ε-uniform distribution over $K = \mathrm{conv}(S_i)$. This can be done with any method of choice. Here, we consider the "hit-and-run from a corner" algorithm of Lovász and Vempala [2006], which implements a fast mixing random walk whose stationary distribution is uniform over any convex body. We remark that other methods for computing approximate centers exist, see for example [Basu and Oertel, 2017].

The implementation is as follows. First, we put K in near-isotropic position by computing the minimum volume enclosing ellipsoid (MVEE) and then applying an affine transformation to make the MVEE into the ball of unit radius. As shown in [Khachiyan, 1996], this operation takes time $|S_i| m^2 \big(\ln m + \ln\ln |S_i|\big)$. After this transformation, let μ be the center of the unit ball that contains K. Observe that μ is at distance at least $\frac{1}{m}$ from the boundary of K: this holds since, by John's theorem, the ball of radius $\frac{1}{m}$ centered at μ is entirely contained in K. Now we execute the hit-and-run from a corner algorithm starting at μ. By the results of Lovász and Vempala [2006], we have the following bound.

Lemma 10 (See Lovász and Vempala [2006], Corollary 1.2)**.** *Assume hit-and-run is started from μ. For any $\varepsilon > 0$, the distribution of the random walk after*

$$t = \Theta\left(m^4 \ln \frac{m}{\varepsilon}\right)$$

steps is ε-uniform over K.

It remains to implement the hit-and-run algorithm over K. To this end we need to solve the following problem: given a generic point $x \in K$ and a vector $u \in S^{n-1}$, determine the intersection of the ray $\{x + \alpha u\}_{\alpha \geq 0}$ with the boundary of K. This amounts to solving a linear program that searches for the maximum value $\alpha \geq 0$ such that $x + \alpha u$ can be written as $\sum_{x_j \in S_i} \lambda_j x_j$ for a set of coefficients $\lambda_1, \ldots, \lambda_{|S_i|} \in [0, 1]$. We can solve such an LP in time $t_K = \mathrm{poly}(|S_i|, m)$ with polynomial precision using any polynomial-time solver for linear programs. The total time to draw the N samples is therefore:

$$\mathcal{O}\left(|S_i| m^2 \big(\ln m + \ln\ln |S_i|\big) + N t_K\, m^4 \ln \frac{m}{\varepsilon}\right) \tag{21}$$

As we set $N = \mathcal{O}(m^2)$ and $\varepsilon = \mathcal{O}(m^{-1})$, the total running time is:

$$\mathcal{O}\left(|S_i| m^2 \big(\ln m + \ln\ln |S_i|\big) + t_K\, m^6 \ln m\right) \tag{22}$$

which is in $\mathrm{poly}(|S_i|, m) = \mathrm{poly}(n, m)$.

3 Proof of Lemma 2

Lemma 2. *For any $u \in \mathbb{R}^2$ let $d_u(x, y) = |\langle u, x - y\rangle|$. For any $\eta > 0$ there exists a clustering $\mathcal{C} = (C_1, C_2)$ on a set $X \subset \mathbb{R}^2$ that has arbitrarily large one-versus-all margin with respect to $d_{(0,1)}, d_{(1,0)}$, and yet $d_u(C_1, C_2) \leq \eta\, \phi_{d_u}(X)$ for all $u \in \mathbb{R}^2$.*

Let $u_1 = (1, 0)$ and $u_2 = (0, 1)$, and for some constant a independent of η and to be fixed later, consider the set X consisting of the four points (see Figure 1):

$$p_1 = \eta\, u_1, \quad q_1 = a\, u_1 \tag{23}$$

$$p_2 = \eta\, u_2, \quad q_2 = a\, u_2 \tag{24}$$

Finally, let $\mathcal{C} = (C_1, C_2)$ where $C_1 = \{p_1, q_1\}$ and $C_2 = \{p_2, q_2\}$.

Consider the two pseudometrics $d_1 = d_{(0,1)}$ and $d_2 = d_{(1,0)}$. Then $\phi_{d_1}(C_1) = 0$ and $d_1(C_1, C_2) = \eta$, and vice versa, $\phi_{d_2}(C_2) = 0$ and $d_2(C_1, C_2) = \eta$. Thus, the one-versus-all margin of $\mathcal{C}$ with respect to d_1, d_2 is unbounded.

Now choose any $u \in \mathbb{R}^2 \setminus \mathbf{0}$. Without loss of generality, by rescaling we can assume u to be a unit vector. In this case, we have $d_u(C_1, C_2) \leq d_u(p_1, p_2) \leq \|p_1 - p_2\|_2 = \eta\sqrt{2}$. Yet, the convex hull of X contains a ball of radius $\Omega(1)$, and therefore $\phi_d(X) = \Omega(1)$, where the constants depend on a. Hence, we can make $d(C_1, C_2) \leq \eta\, \phi_d(X)$ by choosing a appropriately.

4 Proof of Theorem 2

Theorem 2. *If $\mathcal{C}$ satisfies α-center proximity, then it has one-versus-all margin $\gamma \geq \frac{(\alpha-1)^2}{2(\alpha+1)}$. Hence, if $\mathcal{C}$ satisfies $(1 + \varepsilon)$-perturbation stability, then it has one-versus-all margin $\gamma \geq \frac{\varepsilon^2}{2(\varepsilon+2)}$.*

Consider any cluster C_i with center c_i. We must show that any $y \in C_j$ with $j \neq i$ satisfies $d(y, x) > \frac{(\alpha-1)^2}{2(\alpha+1)} \phi_d(C_i)$ for all $x \in C_i$. Let $x' = \arg\max_{x \in C_i} d(x, c_i)$. Clearly, if $d(x', c_i) = 0$ then all points of C_i coincide and $\phi_d(C_i) = 0$. If this is the case, then by the α-center proximity, for any $x \in C_i$ and any $y \in C_j$ we have $d(y, x) = d(y, c_i) > \alpha\, d(y, c_j) \geq 0$. Therefore $d(y, x) > a\, \phi_d(C_i)$ for any $a > 0$, which proves our claim.

Suppose instead that $d(x', c_i) > 0$. Choose any $x \in C_i$ and any $y \in C_j$. Now, we have two cases.

Case 1: $x = c_i$**.** In this case we proceed as follows. Bear in mind that $d(x', c_i) = \phi_d(C_i)$, $d(y, c_i) = d(y, x)$, and $d(y, c_j) < \frac{1}{\alpha} d(y, c_i)$.

$$\begin{aligned}
\phi_d(C_i) &= d(x', c_i) && (25)\\
&< \frac{1}{\alpha} d(x', c_j) && (26)\\
&\leq \frac{1}{\alpha}\Big(d(x', c_i) + d(c_i, y) + d(y, c_j)\Big) && (27)\\
&< \frac{1}{\alpha}\left(\phi_d(C_i) + d(y, x) + \frac{1}{\alpha} d(y, c_i)\right) && (28)\\
&= \frac{1}{\alpha}\phi_d(C_i) + \frac{1}{\alpha} d(y, x) + \frac{1}{\alpha^2} d(y, x) && (29)
\end{aligned}$$

from which we infer

$$d(y, x) > \frac{\alpha(\alpha - 1)}{\alpha + 1}\phi_d(C_i) > \frac{(\alpha - 1)^2}{2(\alpha + 1)}\phi_d(C_i) \tag{30}$$

Case 2: $x \neq c_i$**.** In this case, $d(x, c_i) > 0$, and we start by deriving:

$$\begin{aligned}
d(y, x) &\geq d(x, c_j) - d(y, c_j) && (31)\\
&> d(x, c_j) - \frac{1}{\alpha} d(y, c_i) && (32)\\
&\geq d(x, c_j) - \frac{1}{\alpha}\big(d(y, x) + d(x, c_i)\big) && (33)\\
&= -\frac{1}{\alpha} d(y, x) + d(x, c_j) - \frac{1}{\alpha} d(x, c_i) && (34)
\end{aligned}$$

and thus

$$d(y, x)\left(\frac{\alpha + 1}{\alpha}\right) > d(x, c_j) - \frac{1}{\alpha} d(x, c_i) \tag{35}$$

which yields

$$d(y, x) > \frac{\alpha}{\alpha + 1} d(x, c_j) - \frac{1}{\alpha + 1} d(x, c_i) \tag{36}$$

Let $\beta = \frac{d(x', c_i)}{d(x, c_i)}$. Observe that $\phi_d(C_i) \leq 2\beta d(x, c_i)$ and therefore $d(x, c_i) \geq \frac{\phi_d(C_i)}{2\beta}$.

Now we consider two cases. First, suppose that $\beta \leq \frac{\alpha+1}{\alpha-1}$. In this case, we apply $d(x, c_j) > \alpha d(x, c_i)$ to (36) to obtain:

$$\begin{aligned}
d(y, x) &> \frac{\alpha^2}{\alpha + 1} d(x, c_i) - \frac{1}{\alpha + 1} d(x, c_i) && (37)\\
&= d(x, c_i)(\alpha - 1) && (38)\\
&\geq \frac{\phi_d(C_i)}{2\beta}(\alpha - 1) && (39)\\
&\geq \phi_d(C_i)\frac{(\alpha - 1)^2}{2(\alpha + 1)} && (40)
\end{aligned}$$

Suppose instead that $\beta > \frac{\alpha+1}{\alpha-1}$. Since we chose $x' \in C_i$ such that $d(x', c_i) = \beta d(x, c_i)$, we obtain:

$$\begin{aligned}
d(x, c_j) &> d(x', c_j) - d(x, x') && (41)\\
&> \alpha d(x', c_i) - \big(d(x, c_i) + d(x', c_i)\big) && (42)\\
&= \alpha\beta d(x, c_i) - \big(d(x, c_i) + \beta d(x, c_i)\big) && (43)\\
&= d(x, c_i)((\alpha - 1)\beta - 1) && (44)
\end{aligned}$$

Combining this with (36), we obtain:

$$d(y,x) > d(x,c_i)\left(\frac{\alpha}{\alpha+1}((\alpha-1)\beta-1)-\frac{1}{\alpha+1}\right) \tag{45}$$

$$= d(x,c_i)\left(\frac{\alpha(\alpha-1)\beta}{\alpha+1}-1\right) \tag{46}$$

$$\geq \frac{\phi_d(C_i)}{2\beta}\left(\frac{\alpha(\alpha-1)\beta}{\alpha+1}-1\right) \tag{47}$$

$$= \phi_d(C_i)\left(\frac{\alpha(\alpha-1)}{2(\alpha+1)}-\frac{1}{2\beta}\right) \tag{48}$$

$$> \phi_d(C_i)\left(\frac{\alpha(\alpha-1)-(\alpha-1)}{2(\alpha+1)}\right) \tag{49}$$

$$= \phi_d(C_i)\frac{(\alpha-1)^2}{2(\alpha+1)} \tag{50}$$

Hence, in all cases, we obtain $d(y,x) > \phi_d(C_i)\frac{(\alpha-1)^2}{2(\alpha+1)}$. This concludes the proof.

5 Proof of Lemma 3

Lemma 3 (One-versus-all margin implies one-sided-error learnability)**.** *Let d be any pseudometric over $\mathcal{X}$. For any finite $X \subset \mathcal{X}$ and any $\gamma > 0$, define the effective concept class over X:*

$$H = \{C \subseteq X \,:\, d(X \setminus C, C) > \gamma\,\phi_d(C)\} \tag{2}$$

Then $H = I(H)$, and $\text{vc-dim}(H,X) \leq M^*(\gamma,d)$ where $M^*(\gamma,d) = \max(2, M(\gamma,d))$. Therefore, H can be learned with one-sided error ε and confidence δ with $\mathcal{O}\big(\varepsilon^{-2}(M^*(\gamma,d)\log 1/\varepsilon + 1/\delta)\big)$ examples by choosing the smallest consistent hypothesis in H.

For the first claim, we start by showing that $H = I(H)$. Let $C_1, C_2 \in H$. We show that $C := C_1 \cap C_2 \in H$, too. Consider any $y \in X \setminus C$, and without loss of generality assume that $y \notin C_1$. Since $C \subseteq C_1$, we have $\phi_d(C_1) \geq \phi_d(C)$ and $d(y,C) \geq d(y,C_1)$. Moreover, $d(y,C_1) > \gamma\,\phi_d(C_1)$ since $C_1 \in H$. Therefore:

$$d(y,C) \geq d(y,C_1) > \gamma\,\phi_d(C_1) \geq \gamma\,\phi_d(C) \tag{51}$$

proving that $d(y,C) > \gamma\,\phi_d(C)$. Since this holds for all $y \in X \setminus C$, we have $C \in H$ as well. Therefore, $H = I(H)$.

Since $H = I(H)$, then $\text{vc-dim}(H,X) = \text{vc-dim}(I(H),X)$. Now, we use the following results of Kivinen [1995]:

Definition 8. *[Kivinen [1995], Definition 5.11] Let $\mathcal{X}$ be any domain and $\mathcal{H} \subseteq 2^{\mathcal{X}}$ be a concept class. We say that $\mathcal{H}$ slices $X \subset \mathcal{X}$ if, for each $x \in X$, there is $h \in \mathcal{H}$ such that $X \cap h = X \setminus \{x\}$. The slicing dimension of $(\mathcal{H},\mathcal{X})$, denoted by $\text{sl}(\mathcal{H},\mathcal{X})$, is the maximum size of a set sliced by $\mathcal{H}$. If $\mathcal{H}$ slices arbitrarily large sets, then we let $\text{sl}(\mathcal{H},\mathcal{X}) = \infty$.*

Lemma 11 (Kivinen [1995], Lemma 5.19)**.** *If $\text{sl}(\mathcal{H},\mathcal{X}) < \infty$, then $\text{vc-dim}(I(\mathcal{H}),\mathcal{X}) \leq \text{sl}(\mathcal{H},\mathcal{X})$.*

We will now show that $\text{sl}(H,X) \leq M^*(\gamma,d)$, where $M^*(\gamma,d)$ is finite by assumption. By Lemma 11, this will imply $\text{vc-dim}(H,X) \leq \text{sl}(H,X)$.

Let $S \subseteq X$ be any set sliced by H. We show that $|S| \leq M^*(\gamma,d)$. The case $|S| \leq 2$ is trivial since $M^*(\gamma,d) \geq 2$. Suppose then that $|S| \geq 3$. Choose $a,b \in S$ such that $d(a,b) = \phi_d(S)$. Since S is sliced by H, for any $x \in S$ we have $S \setminus x = S \cap C$ for some $C \in H$. Since $S \setminus x \subseteq C$, we have $d\big(S \setminus x, x\big) \geq d(C,x)$ and $\phi_d(C) \geq \phi_d(S \setminus x)$. Moreover, $d(C,x) > \gamma\,\phi_d(C)$, since $x \in X \setminus C$ and $C \in H$. Therefore:

$$d\big(S \setminus x, x\big) \geq d(C,x) \tag{52}$$

$$> \gamma\,\phi_d(C) \tag{53}$$

$$\geq \gamma\,\phi_d(S \setminus x) \tag{54}$$

Hence, $d(S \setminus x, x) > \gamma\, \phi_d(S \setminus x)$ for all $x \in S$.

Now, suppose first that $\gamma \geq 1$. Since $|S| \geq 3$, any $x \in S \setminus \{a, b\}$ yields the absurd:

$$\begin{aligned}
\phi_d(S) &\geq d(S \setminus x, x) && (55)\\
&> \phi_d(S \setminus x) && \text{since } \gamma \geq 1 \quad (56)\\
&= d(a, b) && \text{since } a, b \in S \setminus x \quad (57)\\
&= \phi_d(S) && \text{by the choice of } a, b \quad (58)
\end{aligned}$$

Hence we must have $|S| \leq 2$, which implies $|S| \leq M^*(\gamma, d)$. Suppose instead that $\gamma < 1$. Then, for any two distinct points $x, y \in S$, we have $d(x, y) > \gamma\phi_d(S)$. This is trivially true if $x = a$ and $y = b$; otherwise, assuming $x \notin \{a, b\}$, it follows by the fact that $d(x, y) \geq d(S \setminus x, x) > \gamma\phi_d(S \setminus x) = \gamma\phi_d(S)$, as seen above. Moreover, S is contained in the closed ball $B(x, \phi_d(S))$ for any $x \in S$. Therefore, $\mathcal{M}(B(x, r), \gamma r, d) \geq |S|$ for $r = \phi_d(S)$ and some $x \in \mathcal{X}$. By definition of $M(\gamma, d)$ this implies that $|S| \leq M(\gamma, d)$, and $M(\gamma, d) \leq M^*(\gamma, d)$. This concludes the proof.

6 Proof of Theorem 3

Theorem 3. *Let (X, O) be any instance whose latent clustering $\mathcal{C}$ has one-versus-all margin $\gamma > 0$ with respect to $d_1, \ldots, d_k$. Then* MREC$(X, O, \gamma, d_1, \ldots, d_k)$ *outputs $\mathcal{C}$ while making, with high probability, at most $\mathcal{O}(M^*(\gamma)k \log k \log n)$ label queries to O, where $M^*(\gamma) = \max(2, M(\gamma))$. Moreover, for any algorithm $\mathcal{A}$ and for any $\gamma > 0$, there are instances with one-versus-all margin γ on which $\mathcal{A}$ makes $\Omega(M(2\gamma))$ label queries in expectation.*

For the lower bounds, let $\gamma' = 2\gamma$. By definition of $M(\gamma')$, there exists a set X of $M(\gamma')$ points that, according to some $d \in \{d_1, \ldots, d_k\}$, lies within a ball of radius $r > 0$, and thus has diameter at most $2r$, and such that $d(x, y) > \gamma' r = 2\gamma r$ for all distinct $x, y \in X$. Now choose $x \in X$ uniformly at random, and define $\mathcal{C} = (x, X \setminus x)$. The argument above shows that $d(x, X \setminus x) > \gamma\phi_d(X)$, which implies that $\mathcal{C}$ has one-versus-all margin γ. Clearly, in expectation over the distribution of $\mathcal{C}$, any exact cluster recovery algorithm must make $\Omega(|X|) = \Omega(M(2\gamma))$ queries. By Yao's principle for Monte Carlo algorithms, then, any such algorithm makes $\Omega(M(2\gamma))$ queries on some instance.

Let us turn to the upper bounds. The pseudocode of MREC is given below. As in the proof sketch, for each $i \in [k]$ the class H_i is defined as:

$$H_i = \{C \subseteq X \,:\, d_i(X \setminus C, C) > \gamma\, \phi_{d_i}(C)\} \tag{59}$$

By Lemma 3 and by definition of learning with one-sided error (Definition 1), we have $\widehat{C}_i \subseteq C_i$ and therefore MREC never misclassifies any point, and moreover:

$$\mathbb{P}\left(\left|C_i \setminus \widehat{C}_i\right| \leq \varepsilon|X|\right) \geq 1 - \delta \tag{60}$$

In our case, that is, with $\varepsilon = 1/2k$ and $\delta = 1/2$, and since $|C_i \setminus \widehat{C}_i| = |C_i| - |\widehat{C}_i|$, this yields:

$$\mathbb{P}\left(\left|\widehat{C}_i\right| \geq |C_i| - \frac{|X|}{2k}\right) \geq \frac{1}{2} \tag{61}$$

which, since $\left|\widehat{C}_i\right| \geq 0$, implies:

$$\mathbb{E}\left|\widehat{C}_i\right| \geq \frac{1}{2} \cdot \left(|C_i| - \frac{|X|}{2k}\right) = \frac{|C_i|}{2} - \frac{|X|}{4k} \tag{62}$$

By summing over all $i \in [k]$, at each round MREC correctly labels, and removes from X, an expected number of points equal to:

$$\mathbb{E}\left|\widehat{C}_1 \cup \ldots \cup \widehat{C}_k\right| = \sum_{i=1}^{k} \mathbb{E}\left|\widehat{C}_i\right| \geq \sum_{i=1}^{k} \left(\frac{|C_i|}{2} - \frac{|X|}{4k}\right) = \frac{|X|}{2} - \frac{|X|}{4} = \frac{|X|}{4} \tag{63}$$

Thus, at each round MREC gets rid of an expected fraction $1/4$ of all points still in X. By a standard probabilistic argument this implies that, with high probability, all points are correctly labeled within $\mathcal{O}(\log n)$ rounds, see Lemma 3 of [Bressan et al., 2020]. Therefore, the total number of queries is bounded by $\mathcal{O}(M^*(\gamma)k \log k)$ with high probability. This concludes the proof.

Algorithm 3 MREC$(X, O, \gamma, d_1, \ldots, d_k)$

if $X = \emptyset$ **then return**
for each $i \in [k]$ let H_i as in Equation 59
draw a sample S of $|S| = \Theta\big(M^*(\gamma) k \ln k\big)$ points u.a.r. from X
use O to learn the labels of S
for $i \in [k]$ **do**
  let S_i be the subset of S having label i
  let $\widehat{C}_i$ be the smallest set in H_i consistent with S_i
  give label i to every $x \in \widehat{C}_i$
let $X' = X \setminus \cup_{i \in [k]} \widehat{C}_i$
MREC$(X', O, \gamma, d_1, \ldots, d_k)$

7 Proof of Theorem 4

As said in the sketch, the proof follows the same ideas of the proof of Theorem 3.

For the lower bounds, suppose that $\mathrm{cosl}(\mathcal{H}) < \infty$, and let X be a set cosliced by $\mathcal{H}$ with $|X| = \mathrm{cosl}(\mathcal{H})$. We draw a random uniform element $x \in X$, and we consider the clustering $\mathcal{C} = (x, X \setminus x)$. By definition of sliced set, $\mathcal{C}$ is realised by $\mathcal{H}$, and therefore it satisfies the assumptions. Now the same arguments of the lower bounds of Theorem 3 imply that any algorithm needs $\Omega(|X|) = \Omega(\mathrm{cosl}(\mathcal{H}))$ queries on some instance to return $\mathcal{C}$. Clearly, if $\mathrm{cosl}(\mathcal{H}) = \infty$ this means that we can take $|X| = n$ arbitrarily large, whence the second lower bound.

For the upper bounds, let $P_k(X)$ be the set of all k-clusterings of X that are realised by $\mathcal{H}$. Then, for each $i \in [k]$ we define:

$$H_i = \{C' \,:\, C' = C'_i \,\wedge\, (C'_1, \ldots, C'_k) \in P_k(X)\} \tag{64}$$

As in MREC, we learn each class H_i with one-sided error by choosing the smallest hypothesis in $I(H_i)$ that is consistent with S_i, for a labeled sample S of size $\Theta(\text{vc-dim}(I(H_i), X)\, k \ln k)$. As shown in the proof of Lemma 3, a result of Kivinen [1995] implies that if $\mathrm{sl}(H_i, X) < \infty$ then $\text{vc-dim}(I(H_i), X) \leq \mathrm{sl}(H_i, X)$. Therefore, to prove the theorem we only need to show that $\mathrm{sl}(H_i, X) \leq \mathrm{cosl}(\mathcal{H})$. To this end, suppose that $U = \{x_1, \ldots, x_\ell\} \subseteq X$ is sliced by H_i. By construction of H_i, this means that there are ℓ clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_\ell$, each one realised by $\mathcal{H}$, such that $\mathcal{C}_i = (\{x_i\}, U \setminus \{x_i\})$ for all $i \in [k]$. This implies that U is cosliced by $\mathcal{H}$. Hence, $|U| \leq \mathrm{cosl}(\mathcal{H})$ and so $\mathrm{sl}(H_i, X) \leq \mathrm{cosl}(\mathcal{H})$, as claimed. The rest of the proof is similar to the proof of Theorem 3, and shows that at each round we recover an expected constant fraction of all points, and that therefore all points will be recovered with high probability after $\mathcal{O}(\log n)$ rounds. This shows that we can recover $\mathcal{C}$ by making with high probability at most $\mathcal{O}(\mathrm{cosl}(\mathcal{H}) k \log k \log n)$ queries.

8 Proof of Theorem 5

Theorem 5. *Let $\mathcal{H}$ be a concept class in $\mathbb{R}^m$ that is non-fractal and closed under affine transformations. There is an algorithm that, given any instance whose latent clustering $\mathcal{C}$ has one-versus-all margin γ and is realized by $\mathcal{H}$, returns $\mathcal{C}$ while making $\mathcal{O}(M k \log k \log n)$ label queries with high probability, where $M = \max\big(2, (1 + 4/\gamma)^m\big)$. Moreover, for any algorithm $\mathcal{A}$, there exist arbitrarily large n-point instances, whose latent clustering $\mathcal{C}$ has arbitrarily small one-versus-all margin and is realized by $\mathcal{H}$, where $\mathcal{A}$ makes $\Omega(n)$ label queries in expectation to recover $\mathcal{C}$.*

The upper bound follows from Theorem 3, by using the standard fact that in the Euclidean metric the unit ball has covering number $\mathcal{N}(B(x,1), \varepsilon) \leq (1 + 2/\varepsilon)^m$ for all $\varepsilon > 0$. As it is well-known, we have $\mathcal{M}(B(x,1), \varepsilon) \leq \mathcal{N}(B(x,1), \varepsilon/2)$, which for $\varepsilon = \gamma$ yields $M(\gamma) \leq (1 + 4/\gamma)^m$.

We now prove the lower bound. Having instances with "arbitrarily small one-versus-all margin" means that $\gamma = 0$, see Definition 3. Take any $h \in \mathcal{H}$ such that both h and its complement $\overline{h}$ contain a ball of positive radius. Note that this implies that, for any $\rho > 0$, there exists a closed ball B with

radius $r > 0$ such that:

$$B \subseteq h \tag{65}$$

$$\exists x \in \overline{h} \,:\, d(B, x) \leq \rho \tag{66}$$

Let c be the center of B. Consider a sphere S of radius r' that contains x and whose center c' lies on the affine subspace $x + \alpha(c - x)$. Let $\eta = \sup_{y \in S \setminus B} d(x, y)$ and let X be an η-packing of S. Note that, since $\gamma = 0$, we can choose ρ and r' arbitrarily small, and in particular we can make the ratio $\frac{\eta}{r'}$ arbitrarily small. This implies that we can make X arbitrarily large, see Figure 4.

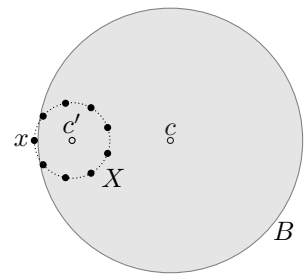

Figure 4: An η-packing X such that $X \cap h = X \setminus \{x\}$ and $X \cap \overline{h} = \{x\}$. The ball B is by construction entirely in h, whereas x is by construction in $\overline{h}$.

Now, consider $x' \in X \setminus \{x\}$. As by construction $d(x', x) > \eta$, and as $r' < r$, we must have $x' \in B$. Therefore, $x' \in h$. Hence the concept $h_x = h$ is such that $X \cap h_x = X \setminus \{x\}$. Now, for any $x' \in X \setminus \{x\}$, there is a rotation R with fixed point c' and such that $R(x') = x$. Hence, $h_{x'} = R^{-1} h_x$ is such that $X \cap h_{x'} = X \setminus \{x'\}$. Since R^{-1} is an affine transformation, $h_{x'} \in \mathcal{H}$ as well. Hence, for every $x \in X$ there exists some concept $h_x \in \mathcal{H}$ such that $X \cap h_x = X \setminus \{x\}$.

Now consider the complement $\overline{h}$ of h. Note that $\overline{h} \in \mathrm{co}(\mathcal{H})$. The first part of the argument above can be applied to $\overline{h}$ as well, showing that for every $x \in X$ there exists $\overline{h_x} \in \mathrm{co}(\mathcal{H})$ such that $X \cap \overline{h_x} = X \setminus \{x\}$. Now consider the complement h_x of $\overline{h_x}$. Clearly $h_x \in \mathcal{H}$, and moreover, $X \cap h_x = \{x\}$.

Hence, for any $x \in X$ there are two concepts $h_x^-, h_x^+ \in \mathcal{H}$ such that $X \cap h_x^- = X \setminus \{x\}$ and $X \cap h_x^+ = \{x\}$. Hence, every 2-clustering $\mathcal{C}$ of X in the form $C_1 = \{x\}, C_2 = X \setminus \{x\}$ is realized by $\mathcal{H}$. Note that this holds with X fixed; we just need to transform the concetps appropriately. It is they immediate to see that any algorithm must perform $\Omega(|X|)$ queries on some instance (X, O). As we can make X arbitrarily large, this completes the proof.