On Margin-Based Cluster Recovery with Oracle Queries

Anonymous Author(s)
Affiliation
Address
email

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 the Euclidean case, $\mathbb{R}^m$, 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 Euclidean spaces, 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, equivalently, to 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 $\mathcal{C}$ deterministically 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, the question is what assumptions on $\mathcal{C}$ yield query-efficient algorithms. Since a good clustering is often thought of as having well-separated clusters, a natural assumption is that $\mathcal{C}$ satisfies some margin property; and previous work shows precisely that some margin properties yield cluster recovery algorithms that achieve the "gold standard" bound of $\mathcal{O}(\log n)$ queries. The first such algorithm appeared in Ashtiani et al. [2016] for the Euclidean case (i.e., when $X \subseteq \mathbb{R}^m$), with the following result. If every cluster C_i is separated from $X \setminus C_i$ by a ball that is centered in the center of mass of C_i, then $\mathcal{O}(\log n)$ queries are sufficient to recover $\mathcal{C}$ with high probability, provided that, for some fixed $\gamma > 0$, every ball does not include points of other clusters, even if expanded by a factor of $1 + \gamma$. The parameter γ is called *margin*, and the number of queries needed to recover $\mathcal{C}$ grows with $1/\gamma$.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

In a first attempt at generalizing this result, Bressan et al. [2020] showed that with $\mathcal{O}(\log n)$ queries one can actually recover clusters with *ellipsoidal* separators, at the price of a dependence on γ and m of approximately $(m/\gamma)^m$. Interestingly, this result was achieved via boosting of one-sided error learning, which works as follows. Suppose that, by making $\mathcal{O}(1)$ 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 $\mathcal{O}(\log n)$ rounds, we have correctly labeled all the input points with high probability. The difficult task is, of course, using only $\mathcal{O}(1)$ queries to identify correctly a constant fraction of some cluster C_i (this is called learning with one-sided error, since all points predicted to be in C_i are indeed in C_i, while points predicted *not* to be in C_i can be anywhere). The key insight of 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 margin-based approach be extended?

In this work we provide several answers, revealing how margin-based cluster recovery and one-sided error learning are intimately connected. Our main contributions are as follows.

1. We introduce a new notion of margin in Euclidean spaces, 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 simple technique (whose proof, however, is quite technical) yields an algorithm for exact cluster recovery in $\mathbb{R}^m$ which runs in polynomial time and makes $\mathcal{O}(\log n)$ queries (Theorem 1). The dependence of the query bound on γ and m is of order $(1+1/\gamma)^m$, which is significantly better than [Bressan et al., 2020] and closer to the query complexity 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 some standard notions of stability for hard 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 $\mathcal{O}(\log n)$ queries by an algorithm that uses a purely learning-based approach (Theorem 3). The $\mathcal{O}(\log n)$ query bound hides a dependence on the complexity of the pseudometric space expressed in terms of its packing number. We show that such a dependence is essentially optimal, thus characterizing the recoverability of clusterings in this setting.
3. Finally, we show a new 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 of $\mathcal{H}$, called *coslicing dimension*, is bounded, then one can learn clusterings with $\mathcal{O}(\log n)$ label queries; otherwise, $\Omega(n)$ queries are needed in the worst case (Theorem 4). Moreover, we show that if $\mathcal{H}$ is *any* concept class in $\mathbb{R}^m$ that is closed under affine transformations and well-behaved in a natural sense, then the clusters realized by $\mathcal{H}$ are recoverable with $\mathcal{O}(\log n)$ label queries if and only if they have positive one-versus-all margin (Theorem 5).

Our contributions have implications for active learning of binary and multiclass classifiers, too. More precisely, 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], 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 (the VC-dimension or the Natarajan dimension), run our algorithms over X, and compute a hypothesis consistent with the recovered labeling $\mathcal{C}$. These types of reductions are standard in active learning, see for instance [Kane et al., 2017].

Additional related work. Same-cluster 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]. Note that same-cluster queries are essentially equivalent to label queries, the basic mechanism of active learning [Hanneke, 2014].

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 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. Combinatorial characterizations of multiclass learning have been also 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., 2019] for a discussion.

2 Preliminaries and notation

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-uple of pairwise disjoint sets whose union is X.[1] Note that we allow clusters to be empty. Our goal is to recover $\mathcal{C}$ exactly, 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 only for simplicity. Indeed, 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. This implies that our bounds on the number of queries for cluster recovery hold for an SCQ oracle as well, up to a multiplicative factor of k.

We often assume a metric or a pseudometric d over $\mathcal{X}$ (a pseudometric allows two distinct points to have distance 0). 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 $\operatorname{conv}(X)$ for the convex hull of X. The unit Euclidean 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 Euclidean spaces

In this section, we consider the Euclidean setting $\mathcal{X} = \mathbb{R}^m$. We show that the ellipsoidal margin assumption 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*

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

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 new 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 classic SVM margin, which prescribes a distance of $\gamma\, \phi(X)$ between the clusters where $\phi(X)$ is the Euclidean diameter of X. Indeed, since every cluster has Euclidean diameter at most $\phi(X)$, a SVM margin of γ implies a convex hull margin of γ. On the other hand, there are cases with arbitrarily small SVM margin but arbitrarily large convex hull margin, as we see in Section 4. Third, the use of pseudometrics allows us to capture the Euclidean distance between the points after projection on a subspace $V \subset \mathbb{R}^m$, modelling 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) *deterministically outputs $\mathcal{C}$, runs in time $\mathrm{poly}(k, n, m)$, and with high probability makes a number of 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 their ellipsoidal margin of γ_{EL} implies[2] convex hull margin $\gamma \geq \frac{\gamma_{EL}}{3}$ for all $\gamma_{EL} \leq 1$. Hence, in this range, our dependence on γ is better by $\Theta(m^m)$ factors.

As we said, CHEATREC is based on boosting learners with one-sided error. What makes CHEATREC different, however, is the technique used to learn with one-sided error. To explain this, let us recall how the algorithm of Bressan et al. [2020] works. Their idea is to learn an approximate ellipsoidal separator that contains a constant fraction of some cluster C_i, and then remove from it all points that do not belong to C_i. Crucially, to learn the approximate separator, the algorithm needs a number of queries proportional to the VC-dimension of the corresponding class, which for m-dimensional ellipsoids is $\mathcal{O}(m^2)$. Unfortunately, this approach does not work with the convex hull margin, since the class of allowed separators has unbounded VC-dimension (it is the class of all polytopes in $\mathbb{R}^m$).

To bypass this obstacle we develop a novel technique for learning with one-sided error that we call *convex hull expansion trick*. This technique essentially amounts to the following procedure: draw a large labeled sample from X, take all the sampled points that belong to the same cluster C, and inflate their convex hull by a factor $1 + \gamma$. This can be seen as a way to exploit the convex hull margin directly, without going through the VC-dimension of the separators. Such a trick may appear natural, but proving it to work is not trivial and requires a combination of results from probability, convex geometry, and PAC learning. In the rest of the section, we describe the 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 s be a parameter to be set later. In each round, CHEATREC draws from X a uniform random sample S of size $|S| = ks$. Then, using the oracle, it determines the partition $S_1, \ldots, S_k$ induced by $\mathcal{C}$. As there are at most k clusters, for at least one of them (say cluster C) we have $|S_C| \geq s$, a fact that allow us to use PAC bounds later on. Now, let $K = \mathrm{conv}(S_C)$, and let $d \in D$ be the pseudometric that gives the margin of C. Since d is homogeneous and invariant under translation (recall that D contains pseudometrics induced by seminorms), we know that any point y such that $d(y, K) \leq \gamma\phi_d(K)$ must belong to C as well. Therefore, if we pick any point $z \in K$ and compute the scaling $Q = (1 + \gamma)K$ with respect to z, Q will not intersect clusters other than C. Hence, we can safely label all points in $X \cap Q$ as belonging to C. As long as we stick with this, we will not make mistakes — in other words, we learn with one-sided error. However, this is not sufficient: in order to make actual progress, we must guarantee

[2] Their definition of margin uses squared distances, so if a cluster has margin γ_{EL} by their definition, then it has convex hull margin $\sqrt{1 + \gamma_{EL}} - 1$. This is why we need to make this apparently misplaced observation.

that $X \cap Q$ contains a good fraction of C. It is not obvious why this should be true: in the worst case, $X \cap Q$ could simply coincide with S_C and we would have learned nothing.

It is here that the convex hull expansion trick enters the game. The key idea is to choose $z = \mu_K$, the center of mass of K. Since however computing μ_K is hard [Rademacher, 2007], we use an approximation. We give here the technical statement and a sketch of the proof (full proof in the supplementary material). 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, with probability arbitrarily close to 1, we have $|Q \cap C| \geq \frac{1}{2}|C|$.*

Sketch of the proof. To begin, suppose that Q was obtained by scaling K about its own center of mass μ_K. Then, by a probabilistic argument, a result of Naszódi [2018] implies the existence of a polytope P on roughly $(1 + 1/\gamma)^m$ vertices such that $K \subseteq P \subseteq Q$. Thus, if we can show that $X \cap P$ contains a good fraction of C, then $X \cap Q$ will contain a good fraction of C, too. To ensure that $X \cap P$ contains a good fraction of C, we observe that the class $\mathcal{P}_t$ of all polytopes on at most t vertices has VC-dimension at most $8m^2 t \log_2 t$, by a recent result [Kupavskii, 2020]. Hence, if we let $|S_C| = s \simeq (1 + 1/\gamma)^m$, then by standard PAC bounds $X \cap P$ contains half of C with probability arbitrarily close to 1. This holds because, as $K \subseteq P \subseteq Q$, then P is consistent with S_C.

Now suppose that, in place of μ_K, we use $z = \frac{1}{N}\sum_{i=1}^N X_i$. If K is in isotropic position, then its radius is bounded by m, and therefore $\|X_i\|_2 \leq m$. This implies that, if $N = \Theta(m^2)$ and each X_i comes from a distribution that is $\Theta(1/m)$-uniform over K, then with high 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 $\eta \leq 1/3$. Then, by adapting the result of Naszódi [2018] with an extension of Grunbaum's inequality for convex bodies due to Bertsimas and Vempala [2004], we can show that $|Q \cap C| \geq \frac{1}{2}|C|$ with probability arbitrarily large. □

We conclude with a note on how to implement CHEATREC in polynomial time. The central point is to avoid computing $K = \mathrm{conv}(S_C)$ explicitly as an intersection of halfspaces, since this could take time $|S_C|^{\Theta(m)}$. Thus, we proceed as follows. First, we transform S_C so that $\mathrm{conv}(S_C)$ is in what is called a near-isotropic position. To this end, we compute John's ellipsoid for S_C, and apply the map that turns that ellipsoid into a ball, which takes polynomial time. Afterwards, we can draw $X_1, \ldots, X_N$ efficiently via the "hit-and-run from a corner" algorithm of Lovász and Vempala [2006], in time $\mathcal{O}\big(\mathrm{poly}(m, |S_C|) \log \frac{m}{\varepsilon}\big)$ per sample, including the time to solve a linear program to determine when the walk hits the boundary of K. Once we have z, we rescale S_C about z itself, and label all points in $(1+\gamma)\,\mathrm{conv}(S_C)$ as belonging to C. This amounts to solving for every $x \in X$ a feasibility problem, which takes polynomial time. Finally, observe that the polytope P appearing in the proof is only for the analysis, and is never computed. For a complete discussion, see the full proof.

4 A more abstract view: the one-versus-all margin

In this section, we consider the case where $\mathcal{X}$ is a generic space equipped with a set of pseudometrics. Like in Section 3, we want to formulate a notion of margin between clusters. However, now we cannot express the margin in terms of the diameter of convex hulls (since $\mathcal{X}$ need not be a vector space and may have no notion of convexity at all), and we cannot apply any "expansion trick" (since the pseudometrics may not be homogeneous and invariant under translation). Instead, we adopt what we call the *one-versus-all margin*. We prove that clusterings with positive one-versus-all margin can be recovered again with $\mathcal{O}(\log n)$ oracle queries. However, we do not provide running time bounds like those of Section 3: our algorithm is essentially based on computing an ERM, which in general may take time superpolynomial in n.

Next, we introduce the one-versus-all margin.

Definition 3 (One-versus-all margin)**.** *Choose 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:*

$$d_i(X \setminus C_i, C_i) > \gamma\, \phi_{d_i}(C_i) \tag{2}$$

Remark. Unlike the convex hull margin (Definition 2), here the pseudometrics $d_1, \ldots, d_k$ are fixed in advance. This is not a weakness of the above definition, but rather a strength of the convex hull margin, which allows one to recover clusters without even knowing $d_1, \ldots, d_k$.

Note also that, just like the convex hull margin, the one-versus-all margin in $\mathbb{R}^m$ is strictly more general than the SVM margin. The fact that SVM margin implies one-versus-all margin follows by the observation in Section 3. The fact that one-versus-all margin does not imply SVM margin is shown in Lemma 2 below.

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 any $u \in \mathbb{R}^2 \setminus \mathbf{0}$.*

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. It is easy to see that $d_1(C_1, C_2) > 0 = \phi_{d_1}(C_1)$, and that $d_2(C_1, C_2) > 0 = \phi_{d_2}(C_2)$. Hence, the one-versus-all margin of $\mathcal{C}$ with respect to d_1, d_2 is unbounded. Yet, for any $\eta > 0$ we can make $d_u(C_1, C_2) \leq \eta\, \phi_{d_u}(X)$ for any nonzero vector $u \in \mathbb{R}^2$, by placing the endpoints of the clusters arbitrarily near the origin. □

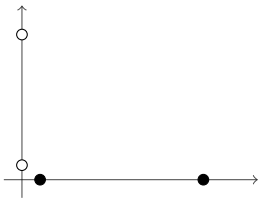

Figure 1: An instance with arbitrarily small SVM margin but unbounded one-versus-all margin.

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

The proof of Theorem 2 is found in the supplementary material.

4.2 Cluster recovery with one-versus-all margin

We conclude by showing that, if $\mathcal{C}$ has one-versus-all margin $\gamma > 0$, then one can recover $\mathcal{C}$ with $\mathcal{O}(\log n)$ queries. Our algorithm, MREC, considers the set of all possible clusters that satisfy the

margin, and for each label selects the smallest hypotheses consistent with the sampled points. To prove that this approach works, we establish a formal connection between one-versus-all margin and one-sided-error learnability of the concept classes induced by all possible clusters with margin.

We need some further notation. As usual, let d be any pseudometric over $\mathcal{X}$. For any $X \subset \mathcal{X}$ and any $r > 0$, we denote by $\mathcal{M}(X, r, d)$ 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, and for any $\gamma > 0$, we define $M(\gamma, d) = \max\{\mathcal{M}\big(B(x, r), \gamma r, d\big) : x \in \mathcal{X}, r > 0\}$. Hence, for any $\gamma > 0$, any ball B in $\mathcal{X}$ contains at most $M(\gamma, d)$ points at pairwise distance greater than γ times the radius of B, and some B attains this bound.[3] 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:*

$$H = \{C \subseteq X \,:\, d(X \setminus C, C) > \gamma\, \phi_d(C)\} \tag{3}$$

Then $H = I(H)$, and $\text{vc-dim}(H, X) \leq M$ where $M = \max(2, M(\gamma, d))$. As a consequence, H can be learned with one-sided error ε and confidence δ with $\mathcal{O}\big(\varepsilon^{-2}(M \log {}^1\!/_\varepsilon + {}^1\!/_\delta)\big)$ 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 $\text{vc-dim}(H, X) \leq M$, 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 $\text{vc-dim}(H, X) \leq M$ it suffices to prove that $\text{sl}(H, X) \leq M$. 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:

$$d\big(S \setminus x, x\big) \geq d(C, x) > \gamma\, \phi_d(C) \geq \gamma\, \phi_d(S \setminus x) \tag{4}$$

It can be shown that this implies $d(S \setminus x, x) > \gamma\phi_d(S)$ for all $x \in S$, and, in turn, $|S| \leq M$. 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. □

We can now present our main result. We 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, γ) *deterministically outputs $\mathcal{C}$ while making, with high probability, at most $\mathcal{O}(Mk \log k \log n)$ queries to O, where $M = \max(2, M(\gamma))$. 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 $\Omega(M(2\gamma))$ 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 $\Theta(Mk \lg k)$ queries; the rest follows by our general boosting argument. To begin, for each $i \in [k]$ we let $H_i = \{C \subseteq X \,:\, d_i(C, X \setminus C) > \gamma\, \phi_{d_i}(C)\}$. We then set $\varepsilon = {}^1\!/_{2k}$ and $\delta = {}^1\!/_2$, and draw a labeled sample S of size $\Theta(\varepsilon^{-1}(M \log {}^1\!/_\varepsilon + \log {}^1\!/_\delta)) = \Theta(Mk \log k)$. 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 $\widehat{C}_i$. 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. □

[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.

Remark. By Theorem 3, in $\mathbb{R}^m$ MREC yields a $\mathcal{O}(\log n)$ query bound even when the clusters are *not* convex. However, this does not mean that MREC subsumes CHEATREC. First, as noted above, here $d_1, \ldots, d_k$ are known in advance. 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 study active cluster recovery when the clusters can be realized by binary concepts from some concept class $\mathcal{H}$. In other words, this is the clustering equivalent of the classic one-versus-all formulation of multiclass classifiers (see, e.g., [Shalev-Shwartz and Ben-David, 2014]). We show that the active recoverability of clusterings in this setting is driven by the *coslicing dimension* of $\mathcal{H}$, 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 many natural concept classes in $\mathbb{R}^m$, a bounded coslicing dimension is equivalent to a positive one-versus-all margin.

As usual, let $\mathcal{X}$ be any domain, $X \subset \mathcal{X}$ be any finite set, and $\mathcal{C} = (C_1, \ldots, C_k)$ be 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 that the number of active queries needed to recover $\mathcal{C}$ depends on the complexity of $\mathcal{H}$. This leads us to the following question: what can we say about the recoverability of $\mathcal{C}$ in terms of the concept class $\mathcal{H}$?

5.1 A general characterization: the coslicing dimension

In this section, we characterize the active recoverability of $\mathcal{C}$ via the *coslicing dimension* of $\mathcal{H}$, a combinatorial quantity similar to the slicing dimension of Kivinen [1995] and the star number of Hanneke and Yang [2015].

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 $\mathcal{O}(\operatorname{cosl}(\mathcal{H})\, k \log k \log n)$ 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 $\Omega(\operatorname{cosl}(\mathcal{H}))$ queries in expectation to return $\mathcal{C}$. As a consequence, if $\operatorname{cosl}(\mathcal{H}) = \infty$ then any algorithm needs $\Omega(n)$ queries in expectation to recover an n-point clustering realized by $\mathcal{H}$.*

Sketch of the proof. The proof follows the same ideas of the proof of Theorem 3. For the lower bounds, 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 bounds, for each $i \in [k]$ we define:

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

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 $\operatorname{sl}(H_i, X) \leq \operatorname{cosl}(\mathcal{H})$. $\square$

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 $\mathcal{H}$, whereas the margin depends on the instance (X, O). To fix this, one can redefine 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, if every instance $(X, O) \in \mathcal{I}$ has margin $\gamma > 0$, one can bound $\operatorname{cosl}(\mathcal{H}, \mathcal{I})$ as a function of γ, retrieving the same type of bounds of Theorem 1 and Theorem 3.

5.2 The one-versus-all margin, again!

We look again at the Euclidean case, $\mathcal{X} = \mathbb{R}^m$. Consider any concept class $\mathcal{H}$. Theorem 4 and the above remark say that, if $\operatorname{cosl}(\mathcal{H}, \mathcal{I}) < \infty$, where $\mathcal{I}$ is the class of allowed instances, then any clustering $\mathcal{C}$ realized by $\mathcal{H}$ can be recovered with $\mathcal{O}(\log n)$ queries, with no need for the one-versus-all margin (Definition 3). We show that, for a wide family of concept classes in $\mathbb{R}^m$, both things are actually equivalent: we can achieve the $\mathcal{O}(\log n)$ bound *if and only* if the instances have positive one-versus-all margin. This establishes a connection between one-versus-all margin and active learnability of clusterings realized by concept classes in $\mathbb{R}^m$. 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 $\mathcal{O}(Mk \log k \log n)$ 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 $\Omega(n)$ 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. $\square$

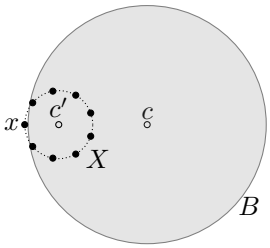

Figure 2: The η-packing X is in B, and thus in h, except for x that lies in $\overline{h}$. By taking B arbitrarily close to x, we can make η arbitrarily small and thus X arbitrarily large.

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 whose hypotheses are not convex: for instance, the class of all finite or infinite disjoint unions of balls, polytopes, or convex bodies.

References

N. Ailon, A. Bhattacharya, and R. Jaiswal. Approximate correlation clustering using same-cluster queries. In *Proc. of LATIN*, pages 14–27, 2018a. doi: 10.1007/978-3-319-77404-6_2.

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. doi: 10.4230/LIPIcs.ITCS.2018.40.

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. doi: https://doi.org/10.1016/j.ipl.2011.10.006.

M.-F. Balcan and P. Long. Active and passive learning of linear separators under log-concave distributions. In *Proc. of COLT*, volume 30, 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. doi: 10.1145/1008731.1008733.

Y. Bilu and N. Linial. Are stable instances easy? *Comb. Probab. Comput.*, 21(5):643–660, Sept. 2012. doi: 10.1017/S0963548312000193.

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 (to appear)*, 2021.

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

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. doi: 10.4230/LIPIcs.ICALP.2018.57.

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. *CoRR*, abs/1512.00537, 2015. URL http://arxiv.org/abs/1512.00537.

S. Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014. doi: 10.1561/2200000037.

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

M. Hopkins, D. M. Kane, and S. Lovett. The power of comparisons for actively learning linear classifiers. *arXiv preprint arXiv:1907.03816*, 2019.

M. Hopkins, D. Kane, S. Lovett, and M. Moshkovitz. Bounded memory active learning through enriched queries. *CoRR*, abs/2102.05047, 2021. URL https://arxiv.org/abs/2102.05047.

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. doi: 10.1109/FOCS.2017.40.

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. doi: 10.1007/BF01191474.

A. Kupavskii. The VC-dimension of k-vertex d-polytopes. *Combinatorica*, 40(6):869–874, 2020. doi: 10.1007/s00493-020-4475-4.

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. doi: 10.1007/s00454-018-9977-0.

L. A. Rademacher. Approximating the centroid is hard. In *Proc. of ACM SoCG*, page 302–305, 2007. doi: 10.1145/1247069.1247123.

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. doi: 10.4230/LIPIcs.ESA.2019.81.

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]