
Formal Models of Active Learning from Contrastive Examples

Farnam Mansouri
University of Waterloo
f5mansou@uwaterloo.ca

Hans U. Simon
Ruhr-University Bochum
hans.simon@rub.de

Adish Singla
Max Planck Institute for Software Systems
adishs@mpi-sws.org

Yuxin Chen
University of Chicago
chenyuxin@uchicago.edu

Sandra Zilles
University of Regina and Amii
sandra.zilles@uregina.ca

Abstract

Machine learning can greatly benefit from providing learning algorithms with pairs of contrastive training examples—typically pairs of instances that differ only slightly, yet have different class labels. Intuitively, the difference in the instances helps explain the difference in the class labels. This paper proposes a theoretical framework in which the effect of various types of contrastive examples on active learners is studied formally. The focus is on the sample complexity of learning concept classes and how it is influenced by the choice of contrastive examples. We illustrate our results with geometric concept classes and classes of Boolean functions. Interestingly, we reveal a connection between learning from contrastive examples and the classical model of self-directed learning.

1 Introduction

In machine learning, contrastive data has been used for various purposes, most notably in order to reduce the number of training samples needed to make high-quality predictions, or to explain the predictions of black-box models. A contrastive example for a labeled training data point (x, b) with label $b \in \{0, 1\}$ could be, e.g., a labeled point $(x', 1 - b)$, such that x' fulfills some additional predefined property. This additional property could be defined in a way so as to convey helpful information for learning. For instance, a constraint on x' could be that it is the *closest* point to x (in the given pool of data points, and with respect to a fixed underlying metric) that has a different label from b . Such examples are often called counterfactuals and have proven useful in various learning settings, including explanation-based supervised learning [8], but also reinforcement learning [16, 22], and learning recommender systems [24].

Intuitively, counterfactuals present a learning algorithm with an important feature or property of a data object that crucially affects its classification. For example, a training data point x might be an image of a cat, together with its class label (the subspecies of cat). A counterfactual piece of information might be an image x' of a cat that looks similar to the cat in x , yet belongs to a different subspecies. The main feature distinguishing x' from x could be highlighted in the image, as a means of explaining the classification of x . Similarly, contrastive information can be used in natural language processing and many other applications, and was applied successfully in representation learning, see,

for example, [5, 7, 19, 20]. Interestingly, also human learners exhibit improved classification abilities when they are taught with the help of such visual explanations [21].

Theoretical studies have analyzed the sample complexity in specific models of learning with contrastive examples, such as PAC-learning of distance functions [3], or active learning with a greedy teacher [26], a randomized teacher [23] or an adversarial teacher [9]. Other formal studies focus on specific aspects of learning with contrastive information, for instance, how negative examples influence the learning process [5] or how individual loss functions affect contrastive learning [15].

One example of the use of contrastive information, particularly relevant to our study, is counterexample-guided program synthesis (CEGIS) in formal methods research [2]. Consider the task of synthesizing a program that satisfies a given specification S . In CEGIS, a synthesizer generates a candidate program P , which is then checked against the specification S by a verifier. If P violates S , the verifier returns a counterexample witnessing the violation. This process is iterated until a program satisfying S is obtained. We can see each candidate program P as a query to the verifier, which acts as an oracle. Importantly, in this setting, the oracle follows a *known procedure*. In particular, one can implement the oracle to return a specific counterexample (e.g., a smallest one) that acts as a contrastive example, and the synthesizer can be *tailored to the knowledge of how the counterexample is generated*. For example, if the synthesizer knows that a smallest counterexample is generated, it can rule out all programs for which there would have been smaller counterexamples.

Motivated by such applications in program synthesis, the main goal of this paper is to introduce and analyze formal settings of learning from contrastive examples, specifically modeling situations in which the learner has *knowledge of the way in which the oracle selects contrastive examples*.

We propose a generic formal framework, encompassing various ways in which contrastive examples are chosen (i.e., not just as counterfactuals); this is the first contribution of our paper. We then analyze several settings in this framework w.r.t. the resulting sample complexity, i.e., the number of examples required for identifying the underlying target concept, resulting in various non-trivial upper and lower bounds on learning classes of geometric objects or Boolean functions (most notably 1-decision lists and monotone DNFs); this constitutes our second main contribution. Finally, as a third contribution, we reveal a surprising connection between the sample complexity of contrastive learning and the mistake bound in the classical notion of self-directed (online) learning [13].

2 Models of Contrastive Learning

Let $\mathcal{X}$ be any (finite or infinite) domain that is bounded under some metric. A concept over $\mathcal{X}$ is a subset of $\mathcal{X}$; we will identify a concept $C \subseteq \mathcal{X}$ with its indicator function on $\mathcal{X}$, given by $C(x) = 1$ if $x \in C$, and $C(x) = 0$ if $x \in \mathcal{X} \setminus C$. A concept class over $\mathcal{X}$ is a set of concepts over $\mathcal{X}$. If C and C' are two concepts over $\mathcal{X}$, then $\Delta(C, C')$ refers to the symmetric difference of C and C' . Moreover, $\delta(C, C')$ is the relative size of $\Delta(C, C')$ over $\mathcal{X}$ measured under the standard uniform distribution. In case $\mathcal{X}$ is finite, this is just $|\Delta(C, C')|/|\mathcal{X}|$.

$S[\text{MQ}](C)$ denotes the minimum worst-case number of membership queries [4] for learning concepts in $\mathcal{C}$, where the worst case is taken over all concepts in $\mathcal{C}$ and the minimum is taken over all membership query learners. Moreover, $\text{VCD}(\mathcal{C})$ denotes the VC dimension of $\mathcal{C}$.

Suppose a concept class $\mathcal{C}$ is fixed and an active learner asks membership queries about an unknown target concept $c \in \mathcal{C}$. In addition to the label of the queried example, the oracle might provide a contrastive example, e.g., a similar example of opposite label. We propose a more generic setting in which the active learner is aided by an oracle that provides labeled examples complementing the actively selected ones in *some* pre-defined form (using a so-called contrast set). The learner tries to identify an unknown target concept $C^* \in \mathcal{C}$ in sequential rounds. In round 1, the version space (i.e., the set of possible target concepts under consideration) is $\mathcal{C}_1 := \mathcal{C}$. Round i , $i \geq 1$, is as follows:

1. The learner poses a membership query consisting of an instance $x_i \in \mathcal{X}$.
2. The oracle's response is a pair $[y_i, (x'_i, y'_i)] \in \{0, 1\} \times (\mathcal{X} \times \{0, 1\})$ such that:
 - $y_i = C^*(x_i)$ and $y'_i = C^*(x'_i)$, i.e., y_i and y'_i are the correct labels of x_i and x'_i , resp., under the target concept,
 - $x'_i \in \text{CS}(x_i, C^*, \mathcal{C}_i)$, where $\text{CS}(x_i, C^*, \mathcal{C}_i) \subseteq \mathcal{X}$ is a set of instances called the *contrast set* of x_i wrt C^* and $\mathcal{C}_i$. The oracle is adversarial, in that it can pick any

$x'_i \in \text{CS}(x_i, C^*, \mathcal{C}_i)$ as part of the response to query x_i . If $\text{CS}(x_i, C^*, \mathcal{C}_i)$ is empty, the oracle returns a dummy response ω , to convey that no contrastive example exists.

3. The version space is updated; in particular, the new version space $\mathcal{C}_{i+1}$ consists of all concepts C in $\mathcal{C}_i$ for which $C(x_i) = y_i$ and:
 - $\text{CS}(x_i, C, \mathcal{C}_i) = \emptyset$ in case the dummy response ω is received,
 - $C(x'_i) = y'_i$ and $x'_i \in \text{CS}(x_i, C, \mathcal{C}_i)$ otherwise.

This definition assumes that the learner knows the function mapping any tuple $(x_i, C, \mathcal{C}_i)$ to the set $\text{CS}(x_i, C, \mathcal{C}_i)$. The core difference to the traditional notion of version space is that the learner can reduce the version space by knowing the contrast set mapping CS: a concept C , even if consistent with all seen labeled examples, is excluded from the version space if $x'_i \notin \text{CS}(x_i, C, \mathcal{C}_i)$ for some i .

The labeled examples collected in the first n rounds, together with the corresponding version spaces, then form an *interaction sequence* of length n wrt CS, given by

$$[(x_1, y_1), (x'_1, y'_1), \mathcal{C}_2], \dots, [(x_n, y_n), (x'_n, y'_n), \mathcal{C}_{n+1}].$$

This interaction sequence ε -approximates the target concept $C^* \in \mathcal{C}$ (for some $\varepsilon \geq 0$) iff $\mathcal{C}_{n+1} \subseteq \{C \in \mathcal{C} \mid \delta(C^*, C) \leq \varepsilon\}$. The sequence identifies the target concept $C^* \in \mathcal{C}$ iff $\mathcal{C}_{n+1} = \{C^*\}$.

Accordingly, we define the following notion of sample complexity.

Definition 1 Assume $\mathcal{X}$, $\mathcal{C}$, CS, and ε are fixed. The ε -approximate contrast sample complexity of a learner L on C^* wrt $\mathcal{C}$, denoted by $\mathcal{S}_{\text{CS}}(L, C^*, \mathcal{C}, \varepsilon)$, is the largest length of any valid interaction sequence $\sigma = [(x_1, y_1), (x'_1, y'_1), \mathcal{C}_2], \dots, [(x_n, y_n), (x'_n, y'_n), \mathcal{C}_{n+1}]$ that ε -approximates $C^* \in \mathcal{C}$ and in which L chooses x_{i+1} on input of the history $[(x_1, y_1), (x'_1, y'_1)], \dots, [(x_i, y_i), (x'_i, y'_i)]$, for all i . (Here the worst case length is taken over the sequence of possible choices of x'_i made by the oracle.) Finally, $\mathcal{S}_{\text{CS}}(L, \mathcal{C}, \varepsilon) = \sup_{C^* \in \mathcal{C}} \mathcal{S}_{\text{CS}}(L, C^*, \mathcal{C}, \varepsilon)$, and $\mathcal{S}_{\text{CS}}(\mathcal{C}, \varepsilon) = \inf_L \mathcal{S}_{\text{CS}}(L, \mathcal{C}, \varepsilon)$. For exact learning, we consider only interaction sequences that identify C^* , and write $\mathcal{S}_{\text{CS}}(\mathcal{C})$ instead of $\mathcal{S}_{\text{CS}}(\mathcal{C}, \varepsilon)$.

Our model assumes the learner has perfect knowledge of the set CS of candidates from which contrastive examples are chosen. This assumption, while often too strong in practice, is indeed realistic in program synthesis settings, where the learning algorithm knows the rules by which the oracle selects counterexamples, as discussed in Section 1 for CEGIS [2]. Here the designer has full control over both the oracle and the learner. In addition, our model is useful for further reasons:

- It can be easily modified to address additional real-world settings, e.g., one can assume that the learner's notion of CS is not identical, but similar to that used by the oracle.
- Lower bounds from our strong model (e.g., our bound in terms of the self-directed learning complexity, Theorem 17) immediately transfer to such weakened versions of the model.

A first observation is that, without limiting the choice of the contrast set CS, the contrast oracle is extremely powerful; any encoding of concepts in $\mathcal{C}$ as subsets of $\mathcal{X}$ of size at most k can be used to define contrast sets witnessing a sample complexity of at most k :

Proposition 2 Let $\mathcal{C}$ be a countable concept class over a countable $\mathcal{X}$. Let $T : \mathcal{C} \rightarrow 2^{\mathcal{X}}$ be any injective function that maps every concept in $\mathcal{C}$ to a finite set of instances. Then:

1. There is some CS with $\mathcal{S}_{\text{CS}}(\mathcal{C}) \leq \sup_{C \in \mathcal{C}} |T(C)| + 1$.
2. If $T(C) \not\subseteq T(C')$ for $C \neq C'$, then there is some CS such that $\mathcal{S}_{\text{CS}}(\mathcal{C}) \leq \sup_{C \in \mathcal{C}} |T(C)|$.

In particular, if $\mathcal{X}$ is finite, then there is some CS such that $\mathcal{S}_{\text{CS}}(\mathcal{C}) \leq 1 + \min\{k \mid \sum_{i=0}^k \binom{|\mathcal{X}|}{i} \geq |\mathcal{C}|\}$. If $\mathcal{X}$ is countably infinite, then there is some CS such that $\mathcal{S}_{\text{CS}}(\mathcal{C}) = 1$.

Proof. (Sketch.) Let $x_1, x_2, \dots$ be a repetition-free enumeration of $\mathcal{X}$, and $\text{CS}(x_i, C) = \{x_j\}$ for $j' = \min\{j \mid j \geq i, x_j \in T(C)\}$. A learner sets $n_1 = 1$ and starts with iteration 1. In iteration i , it asks a query for x_{n_i} . If it receives $x'_i = x_j$ as a contrastive example, then $j \geq n_i$. The learner will then set $n_{i+1} = j + 1$ and proceed to iteration $i + 1$. See the appendix for details. $\square$

Thus, unlimited choice of the mapping CS reduces learning from a contrast oracle to decoding a smallest encoding of concepts as sample sets. Therefore, the remainder of our study focuses on more natural choices of contrast sets. In particular, when fixing a function $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$ (interpreted as a notion of distance), two natural choices of contrast set mappings are the following:

Minimum distance model. Our first contrast set mapping, dubbed the *minimum distance* model, makes the oracle provide an example closest to x , among those that yield a different label under C than x . For discrete $\mathcal{X}$, this translates to $\text{CS}_{\min}^d(x, C) = \arg \min\{d(x, x') \mid x' \in \mathcal{X}, C(x') \neq C(x)\}$. For continuous $\mathcal{X}$, it can happen that no point $x' \in \mathcal{X}$ with $C(x') \neq C(x)$ attains the infimum of the values $d(x, x')$ for such x' .¹ To accommodate this case, we define $\text{CS}_{\min}^d(x, C) = \arg \min\{d(x, x') \mid x' \in \mathcal{X}, x' = \lim_{i \rightarrow \infty} x'_i \text{ for a Cauchy sequence } (x'_i)_i \text{ with } C(x'_i) \neq C(x) \text{ for all } i\}$. (This definition may in some cases allow the oracle to return a contrastive example with the same label as that of x . This is helpful when no point $x' \in \mathcal{X}$ whose label differs from $C(x)$ attains the infimum of the values $d(x, x')$ for such x' , but mostly for complete metric spaces, where the limit of a Cauchy sequence is guaranteed to exist.²)

Proximity model. For our second contrast set mapping, the learner also selects a radius r independently at each round, allowing it to ask for a contrastive example within a certain vicinity of the query point x . For discrete $\mathcal{X}$, we then define $\text{CS}_{\text{prox}}^d(x, r, C) = \{x' \in \mathcal{X} \mid C(x') \neq C(x), d(x, x') \leq r\}$, i.e., the contrast set contains all x' r -close to x that have the opposite label to x . For continuous $\mathcal{X}$, a similar adaptation as for the minimum distance model yields $\text{CS}_{\text{prox}}^d(x, r, C) = \{x' \in \mathcal{X} \mid d(x, x') \leq r, x' = \lim_{i \rightarrow \infty} x'_i \text{ for a Cauchy sequence } (x'_i)_i \text{ with } C(x'_i) \neq C(x) \text{ for all } i\}$. We refer to this choice of contrast set as the *proximity* model. Note that the oracle here need not choose a contrastive example at minimum distance; any example in $\text{CS}_{\text{prox}}^d(x, r, C)$ is allowed.

Since these mappings do not depend on the version space, we dropped the argument $\mathcal{C}_i$ in $\text{CS}(x, \mathcal{C}_i)$. The following useful observations are easy to prove; the proof of Remark 5 is given in the appendix.

Remark 3 Let $\mathcal{C}$ be a concept class over a complete space $\mathcal{X}$ with metric d , $x \in \mathcal{X}$, and $C \in \mathcal{C}$. Suppose there is some $x' \in \mathcal{X}$ with $C(x) \neq C(x')$ and $d(x, x') < \infty$. Then (i) $\text{CS}_{\min}^d(x, C) \neq \emptyset$, and (ii) if $r \geq d(x, x')$ then $\text{CS}_{\text{prox}}^d(x, r, C) \neq \emptyset$.

Remark 4 Suppose that CS and CS' are mappings that assign to every pair $(x, C) \in \mathcal{X} \times \mathcal{C}$ a subset of $\mathcal{X}$. Suppose, for every $(x, C) \in \mathcal{X} \times \mathcal{C}$, we have (i) $\text{CS}(x, C) \subseteq \text{CS}'(x, C)$, and (ii) $\text{CS}(x, C) = \emptyset$ implies $\text{CS}'(x, C) = \emptyset$. Then $\mathcal{S}_{\text{CS}}(\mathcal{C}) \leq \mathcal{S}_{\text{CS}'}(\mathcal{C})$. Moreover, for any ε , $\mathcal{S}_{\text{CS}}(\mathcal{C}, \varepsilon) \leq \mathcal{S}_{\text{CS}'}(\mathcal{C}, \varepsilon)$.

Remark 5 Let $\mathcal{C}$ be a concept class over a complete space $\mathcal{X}$ with metric d . Then $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \leq \mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}) \leq \mathcal{S}[\text{MQ}](\mathcal{C})$.

When $\mathcal{X}$ is finite, we can extend Remark 5 as follows.

Proposition 6 Fix $\mathcal{X}$ and $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$. Let $S_d(x) = \{d(x, x') \mid x' \in \mathcal{X} \setminus x\}$ and $s_d = \max_{x \in \mathcal{X}} |S_d(x)|$. (Note that $s_d \leq |\mathcal{X}| - 1$ for finite $\mathcal{X}$). Then $\mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}) \leq \lceil \log(s_d) \rceil \cdot \mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C})$.

Proof. Let L learn any $C \in \mathcal{C}$ with $\leq \mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C})$ queries to the minimum distance oracle w.r.t. d . We construct a learner L' with access to the proximity oracle, using L as a subroutine. Suppose L poses a query x . Let $r_{\min} = \min_{x' \in \text{CS}_{\min}^d(x, C)} d(x, x')$. First, L' sorts all numbers in $S_d(x)$ in increasing order and determines the median r of that list. Then L' poses the query (x, r) . The contrastive example returned will be ω iff $d(x, x') < r_{\min}$. Hence, with a binary search using $\lceil \log(s_d) \rceil$ queries, L' can determine an x' with opposite label to x and $d(x, x') = r_{\min}$. This x' is among the admissible answers to the query by L . Thus, L' makes $\leq \lceil \log(s_d) \rceil$ queries for every query that L makes. $\square$

¹For instance, if C is a closed interval in $\mathbb{R}$ and the query point x lies inside this interval, then there is no point outside the interval that has distance closer to x than any other point outside the interval.

²The definition of Cauchy sequence relies on a notion of metric. In our examples in continuous domains, we will use the ℓ_1 -metric both as this metric and as the function d in the superscript of $\text{CS}_{\min}^d$.

	without CS	$\text{CS}_{\text{prox}}^d$	CS_{min}^d (even for exact identification)
(a) thresholds	$\Theta(\log \frac{1}{\varepsilon})$	$\Theta(\log \frac{1}{\varepsilon})$	1
(b) rectangles	$\Theta(k \log(\frac{1}{\varepsilon}))$	$\Omega(\log(\frac{1}{\varepsilon})), O(\log(\frac{k}{\varepsilon}))$	2

Table 1: Asymptotic sample complexity when ε -approximating (a) one-sided threshold functions, and (b) axis-aligned rectangles in k dimensions, for various learning models, with the ℓ_1 -metric.

3 Sample Complexity Under Various Metrics

This section illustrates our notions of learning from contrast oracles under various metrics d . We begin with two examples on the ℓ_1 -metric.

Example 1 Consider the class of one-sided threshold functions $\mathbb{1}\{x \leq \theta^*\}$ for $\theta^* \in \mathcal{X} = [0, 1]$. Table 1(a) displays the asymptotic sample complexity of various models of ε -approximate learning. The result on learning without contrastive examples is known [25]. By a straightforward adversary argument, the sample complexity of the proximity model is in $\Omega(\log 1/\varepsilon)$. As learning without contrast oracle is no stronger than learning in the proximity model, both models here have the same asymptotic complexity. In the minimum-distance model, irrespective of the first queried instance, the contrastive example will be θ^* . Thus, the target can be identified (even exactly!) with one query.

Example 2 Consider the class of all k -dimensional axis-aligned rectangles over $\mathcal{X} = [0, 1]^k$, with the ℓ_1 -distance as metric. Table 1(b) displays the asymptotic sample complexity of various models of ε -approximate learning. The result on learning without contrastive examples is known, see, e.g., [14]. Let C^* be a target rectangle with the corner of lowest ℓ_1 -distance from $\mathbf{0} = (0, \dots, 0)$ being $\mathbf{x}$, and the diagonally opposed corner being $\mathbf{y}$.

In the minimum-distance model, let a learner first query $\mathbf{0}$. If $C^*(\mathbf{0}) = 1$ then clearly $\mathbf{x} = \mathbf{0}$. If $C^*(\mathbf{0}) = 0$ then the contrastive example x' will be $\mathbf{x}$. Similarly, by querying $\mathbf{1}$, the learner can determine $\mathbf{y}$. Thus, $\mathcal{S}_{\text{CS}_{\text{min}}^d} \leq 2$. Also, for determining C^* , at least two distinct positively labelled instances are required, which is impossible to acquire with one query.

To verify the upper bound in the proximity model, note that $d(\mathbf{x}, \mathbf{0}) \leq k$ for any $\mathbf{x}$. The learner starts by querying $x_1 = \mathbf{0}$ and $r_1 = \frac{k}{2}$. The contrastive example will be ω iff $r_1 < d(\mathbf{x}, \mathbf{0})$. Thus, the learner determines whether $d(\mathbf{x}, \mathbf{0}) \leq r_1$ or not. Learning continues with a binary search until finding a radius r_t such that $d(\mathbf{x}, \mathbf{0}) \leq r_t \leq d(\mathbf{x}, \mathbf{0}) + \varepsilon/2$ for $t \geq \log(2k/\varepsilon) + 1$. Let x'_t be the contrastive example. We derive $d(x'_t, \mathbf{x}) \leq d(x'_t, \mathbf{0}) - d(\mathbf{x}, \mathbf{0}) \leq r_t - d(\mathbf{x}, \mathbf{0}) \leq \frac{\varepsilon}{2}$ (noting that $d(x'_t, \mathbf{0}) - d(\mathbf{x}, \mathbf{0}) > 0$ at step t). Similarly with $\log(2k/\varepsilon) + 1$ samples the learner can find a $\mathbf{z} \in [0, 1]^k$ such that $d(\mathbf{z}, \mathbf{y}) \leq \varepsilon/2$. Therefore, the rectangle with bottom left corner x'_t and top right corner $\mathbf{z}$ has error less than ε .

The lower bound on $\mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}, \varepsilon)$ is inherited from that in Example 1, using the same argument.

The next few examples concern concept classes over the Boolean domain $B_m = \{0, 1\}^m$ and learning from a contrast oracle in the minimum distance model with respect to the **Hamming distance as metric**. For $I \subseteq [m]$, we denote by $\mathbf{1}_I$ the Boolean vector with 1s in positions indexed by I and 0s elsewhere; let $\mathbf{1} := \mathbf{1}_{[m]}$. The notation $\mathbf{0}_I$ and $\mathbf{0}$ is understood analogously.

The next two toy examples will show that a certain metric (here the Hamming distance) may be suitable in the context of a specific concept class, but will become misleading in the minimum distance model when a slight representational change is made to the concept class. These toy examples are merely meant to illustrate the power and limitations of our abstract model in extreme situations.

Example 3 Let $\mathcal{X} = \{0, 1\}^m$. Let $\mathcal{C}_{\text{mmon}}^m$ consist of all monotone monomials, i.e., logical formulas of the form $v_{i_1} \wedge \dots \wedge v_{i_k}$ for some pairwise distinct $i_1, \dots, i_k$ and some $k \in \{0, \dots, m\}$. The concept associated with such a formula contains the boolean vector $(b_1, \dots, b_m)$ iff $b_{i_1} = \dots = b_{i_k} = 1$ (where the empty monomial is the constant 1 function). Let d be the Hamming distance. Then:

- (1) $\mathcal{S}[\text{MQ}](\mathcal{C}_{\text{mmon}}^m) = m$. (2) $\mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}_{\text{mmon}}^m) = 1$. (3) $\mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}_{\text{mmon}}^m) = \Theta(\log m)$.

Proof. (Sketch.) (1) is folklore. For (2), a learner will first query $\mathbf{0}$, which has the label 0 unless the target is the empty monomial. The unique positively labeled example of smallest Hamming distance to the all-zeroes vector has ones exactly in $v_{i_1}, \dots, v_{i_k}$, where the target concept is $v_{i_1} \wedge \dots \wedge v_{i_k}$.

Thus, a single query identifies the target. The upper bound in (3) follows from Proposition 6, while the lower bound is obtained via case distinction. Details are in the appendix. $\square$

Now, we will see how a small syntactic change from $\mathcal{C}_{\text{mon}}^m$ to a class $\mathcal{C}'_{\text{mon}}$ (adding an $(m+1)$ st component to the boolean vectors) can make contrastive examples useless (see appendix for details).

Example 4 Let $\mathcal{X}' = \{0, 1\}^{m+1}$. The concept class $\mathcal{C}'_{\text{mon}}$ over $\mathcal{X}'$ is defined to contain, for each $C \in \mathcal{C}_{\text{mon}}^m$, the concept $C' \in \mathcal{C}'_{\text{mon}}$ constructed via $C'(b_1, \dots, b_m, 0) = C(b_1, \dots, b_m)$ and $C'(b_1, \dots, b_m, 1) = 1 - C(b_1, \dots, b_m)$. No other concepts are contained in $\mathcal{C}'_{\text{mon}}$. Let d be the Hamming distance. Then $\mathcal{S}[\text{MQ}](\mathcal{C}'_{\text{mon}}) = \mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}'_{\text{mon}}) = \mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}'_{\text{mon}}) = m$.

While the Hamming distance is useless for the class $\mathcal{C}'_{\text{mon}}$ in Example 4, other functions d might still allow for efficient learning in this setting. Such functions, however, might be rather “unnatural” as a notion of distance (arguably though, the class $\mathcal{C}'_{\text{mon}}$ in Example 4 is also somewhat “unnatural”). We will hence focus most of our study on intuitively “natural” functions d . When learning Boolean functions, we consider Hamming distance a natural choice.

Let $\mathcal{C}_{\text{mon}}^m$ denote the class of monomials over the Boolean variables $v_1, \dots, v_m$ (including the empty monomial which represents the constant-1 function). We assume that each Boolean variable occurs at most once (negated or not negated) in a monomial. Since monomials like $v_i \wedge \bar{v}_i$ are thus excluded, the constant-0 function does not belong to $\mathcal{C}_{\text{mon}}^m$. A concept class dual to $\mathcal{C}_{\text{mon}}^m$ is the class of clauses over the Boolean variables $\{v_1, \dots, v_m\}$ (including the empty clause which represents the constant-0 function but excluding clauses like $v_i \vee \bar{v}_i$ which represent the constant-1 function). We denote this class by $\mathcal{C}_{\text{claus}}^m$, and obtain the following result, the full proof of which is given in the appendix.

Theorem 7 $\mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}_{\text{mon}}^m \cup \mathcal{C}_{\text{claus}}^m) = 2$, where d is the Hamming distance.

Proof. (Sketch.) One first constructs a learner $L_{\wedge}$ for $\mathcal{C}_{\text{mon}}^m$, as in the proof of Example 3, and then a learner $L_{\vee}$ for $\mathcal{C}_{\text{claus}}^m$ by dualization from $L_{\wedge}$. It suffices to describe a learner L that maintains simulations of $L_{\wedge}$ and $L_{\vee}$ and either comes in both simulations to the same conclusion about the target concept or realizes an inconsistency in one of the simulations, which can then be aborted. $\square$

For $\mathcal{S}_{\text{CS}_{\text{prox}}^d}$, an upper bound is obtained from Proposition 6 and Theorem 7, noting that the Hamming distance can take only m distinct values for pairs (x, x') with $x \neq x'$.

Corollary 8 $\mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}_{\text{mon}}^m) \cup \mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}_{\text{claus}}^m) \leq 2\lceil \log m \rceil$, where d is the Hamming distance.

We conclude this subsection with an example of a somewhat natural concept class for which the minimum distance model under the Hamming distance is not very powerful.

Example 5 Let $\mathcal{X} = \{0, 1\}^m$ and let $\mathcal{C}_{\text{par}}$ be the class of parity functions, i.e., logical formulas of the form $v_{i_1} \oplus \dots \oplus v_{i_k}$ for some pairwise distinct $i_1, \dots, i_k$ and some $k \in \{0, \dots, m\}$. Then $\mathcal{S}[\text{MQ}](\mathcal{C}_{\text{par}}) = m$ and $m-1 \leq \mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}_{\text{par}}) = \mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}_{\text{par}}) \leq m$.

Proof. $\mathcal{S}[\text{MQ}](\mathcal{C}_{\text{par}}) = m$ is trivial, and immediately yields $\mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}_{\text{par}}) \leq m$, $\mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}_{\text{par}}) \leq m$. Suppose the target concept is $c^* = v_{i_1} \oplus \dots \oplus v_{i_k}$. For any instance $\mathbf{x}$, we have $d(\mathbf{x}, \mathbf{x}') = 1$ and $c^*(\mathbf{x}) \neq c^*(\mathbf{x}')$, where $\mathbf{x}'$ results from $\mathbf{x}$ by flipping the value of component i_1 . When always using a query radius $r_t = 1$, the proximity model is thus equivalent to the minimum distance model. The lower bound is witnessed by an oracle that contrasts any $\mathbf{x}_t$ with the example $\mathbf{x}'_t$ resulting from $\mathbf{x}_t$ by flipping the component i_1 . This provides no information about variables other than v_{i_1} . $\square$

We conclude this section with the **discrete metric** d_0 , defined by $d_0(x, x') = 0$ if $x = x'$, and $d_0(x, x') = 1$ if $x \neq x'$. Now $\text{CS}_{\text{min}}^{d_0}(x, C) \subseteq \text{CS}_{\text{min}}^d(x, C)$ for every $(x, C) \in \mathcal{X} \times \mathcal{C}$ and every $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$. Moreover, $\text{CS}_{\text{min}}^d(x, C) = \emptyset$ implies $\text{CS}_{\text{min}}^{d_0}(x, C) = \emptyset$. Remark 4 thus yields:

Corollary 9 For every concept class $\mathcal{C}$, the term $\mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C})$ is maximized, if the underlying function $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$ equals the discrete metric d_0 .

In what follows, if O is a list of oracles, we use $\mathcal{S}[O]$ to denote the sample complexity of learning with access to all oracles in O . Let EX^+ be an oracle returning a positive example x , i.e., an $x \in \mathcal{X}$

with $C^*(x) = 1$ (or a dummy response if no such x exists). When called several times, the oracle may repeatedly present the same example. Let EX^- be the analogous oracle for a negative example. The following result says that, under the metric d_0 , the contrast oracle does not provide significantly more information than the membership oracle (although, for some special concept classes, this small extra information can make a big difference). The proof is given in the appendix.

Theorem 10 $\mathcal{S}[\text{EX}^+, \text{EX}^-, \text{MQ}](\mathcal{C}) - 2 \leq \mathcal{S}_{\text{CS}_{\min}^{d_0}}(\mathcal{C}) \leq \mathcal{S}[\text{EX}^+, \text{EX}^-, \text{MQ}](\mathcal{C})$.

Example 6 Consider $\mathcal{C}$ to be the class of all singleton concepts over a finite domain $\mathcal{X}$ where $n := |\mathcal{X}| \geq 2$. It is known that $\mathcal{S}[\text{MQ}](\mathcal{C}) = n - 1$. But a single call of the oracle EX^+ reveals the identity of the target concept. It follows that $\mathcal{S}[\text{EX}^+](\mathcal{C}) = \mathcal{S}[\text{EX}^+, \text{EX}^-, \text{MQ}](\mathcal{C}) = \mathcal{S}_{\text{CS}_{\min}^{d_0}}(\mathcal{C}) = 1$.

4 Lower Bounds

In this section, we will derive some lower bounds on the sample complexity in our models of learning from contrastive examples. All our results here are for finite instance spaces.

4.1 A Metric-Independent Lower Bound

Self-directed learning [13] is a model of online learning, in which the sequence of data to be labeled is chosen by the learner as opposed to the environment. The learner gets immediate feedback on every single prediction, which allows it to adapt its data selection to the version space sequentially. It is then assessed in terms of the number of incorrect predictions it makes across all data instances in $\mathcal{X}$; as usual, we consider the worst case over all concepts in the underlying class $\mathcal{C}$.

Definition 11 ([13]) Let $\mathcal{C}$ be a concept class over a finite domain $\mathcal{X}$, $n = |\mathcal{X}|$. A self-directed learner L for $\mathcal{C}$ interacts with an oracle in n rounds. In each round, L selects a not previously selected instance $x \in \mathcal{X}$ and predicts a label b for x . The oracle knows the target $C^* \in \mathcal{C}$ and returns the label $C^*(x)$; if $b \neq C^*(x)$, the learner incurs a mistake. L proceeds to the next round until all instances in $\mathcal{X}$ have been labeled. The cost of L on C^* is the number of mistakes made in these n rounds. The cost of L on $\mathcal{C}$ is the maximum cost of L on any $C \in \mathcal{C}$. The self-directed learning complexity of $\mathcal{C}$, denoted by $\text{SD}(\mathcal{C})$, is the smallest number c such that there exists a self-directed learner for $\mathcal{C}$ whose cost on $\mathcal{C}$ is c .

This complexity notion was analyzed relative to other learning-theoretic parameters [6, 11], and extended to infinite concept classes and multi-class settings [10]. Interestingly, SD can be used to derive a metric-independent lower bound on the sample complexity of the minimum-distance model.

Theorem 12 If $\mathcal{C}$ is defined over a finite $\mathcal{X}$, and $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$, then $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \geq \lceil \text{SD}(\mathcal{C})/2 \rceil$.

Proof. Fix $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$, and let L be a learner with access to a minimum-distance oracle w.r.t. d , which makes at most $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C})$ queries to identify any $C \in \mathcal{C}$. We construct a self-directed learner L' for $\mathcal{C}$ which calls L as a subroutine. If L poses a query for x , then L' predicts 0 on x . After receiving feedback, L' passes on the correct label b for x to L . In addition, L' must provide a contrastive example for x , which is obtained as follows: L' sorts all instances in $\mathcal{X}$ in increasing order w.r.t. their distance from x as measured by d . It then predicts the label b for instances in this ordered list until it observes the first mistake, say for x' . Then x' must be a closest instance to x with a label opposite from b . Hence, x' is passed on to L as a contrastive example. If no such x' is found, then L' gives the response ω to L . This proceeds until all instances in $\mathcal{X}$ have been labeled.

Each time L' calls L , it incurs at most two mistakes, namely one for its prediction on the instance x queried by L , and one for its prediction on the contrastive instance x' . In total, thus L' makes at most $2t$ mistakes, where t is the number of queries posed by L . We obtain the desired bound. $\square$

Remark 13 For each $m \in \mathbb{N}$, there is a concept class $\mathcal{C}_m$ with $\text{VCD}(\mathcal{C}_m) = 2$ and $\text{SD}(\mathcal{C}_m) \geq m$ [18]. In particular, there is no relationship between VCD and SD. Thus, no matter which function $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$ is chosen, there exists no upper bound on $\mathcal{S}_{\text{CS}_{\min}^d}$ as a function of VCD.

We will see multiple cases in which the lower bound provided by Theorem 12 is not met. As a small warm-up example, there are finite classes of VC dimension 1 for which $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) = 2$, while every class of VC dimension 1 satisfies $\text{SD} = 1$ (see the appendix for proof details):

Example 7 Let $\mathcal{X} = \{x_1, x_2, x_3\}$. Let $\mathcal{C}$ consist of the concepts $\{x_1, x_3\}$, $\{x_3\}$, $\{x_2, x_3\}$, and $\{x_2\}$. Then $\mathcal{C}$ satisfies $\text{SD}(\mathcal{C}) = \text{VCD}(\mathcal{C}) = 1$, but $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \geq 2$ for any $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$.

However, for classes of VC dimension 1, we cannot increase the gap between SD and $\min_d \mathcal{S}_{\text{CS}_{\min}^d}$:

Theorem 14 For any finite $\mathcal{C}$ with $\text{VCD}(\mathcal{C}) = 1$ there exists a metric d with $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \leq 2$.

The proof (see appendix) uses the fact that every $x \in \mathcal{X}$ has at most one label $b \in \{0, 1\}$ such that (x, b) “teaches” a unique $C \in \mathcal{C}$ in the model of recursive teaching [11]. The instances in $\mathcal{X}$ can be partitioned according to these labels. This partitioning is used to define a metric that places instances in the same part closer to one another than instances in different parts. Now a learner choosing one specific query from each of the two parts will receive enough information to identify the target.

4.2 Lower Bounds Using the Hamming Distance

Let DL_m be the class of 1-decision lists over the Boolean variables $v_1, \dots, v_m$. Each list is of shape

$$\mathcal{L} = [(\ell_1, b_1), \dots, (\ell_z, b_z), b_{z+1}] , \quad (1)$$

where $z \geq 0$, $\ell_i \in \{v_1, \bar{v}_1, \dots, v_m, \bar{v}_m\}$ and $b_i \in \{0, 1\}$. We assume that each variable occurs at most once in $\mathcal{L}$ (negated or not negated). We say that a point $\mathbf{a} \in B_m$ is *absorbed by item k of $\mathcal{L}$* if the literal ℓ_k applied to $\mathbf{a}$ evaluates to 1 while the literals $\ell_1, \dots, \ell_{k-1}$ applied to $\mathbf{a}$ evaluate to 0. The list $\mathcal{L}$ represents the following Boolean function $f_{\mathcal{L}}$: (i) If $\mathbf{a}$ is absorbed by the k 'th item of $\mathcal{L}$, then $f_{\mathcal{L}}(\mathbf{a}) = b_k$. (ii) If $\mathbf{a}$ is absorbed by none of the z items of $\mathcal{L}$, then $f_{\mathcal{L}}(\mathbf{a}) = b_{z+1}$.

The following lemma observes a structural property of a concept class $\mathcal{C}$ which allows us to lower-bound the sample complexity in the minimum-distance model (with the Hamming distance) using the membership query complexity of a certain subclass of $\mathcal{C}$.

Lemma 15 Suppose that $\mathcal{C}$ is a concept class over B_m , and $\mathcal{C}'$ is a subclass of $\mathcal{C}$ with the properties

(P1) For each $\mathbf{a} \in B_m$ with $a_m = y_m$ and each $C \in \mathcal{C}'$, we have that $C(\mathbf{a}) = 1$.

(P2) For each $\mathbf{a}$ with $a_m = y'_m$ and $a_{m-1} = y_{m-1}$ and each $C \in \mathcal{C}'$, we have that $C(\mathbf{a}) = 0$.

For $i \in \{m, m-1\}$, here $y_i \in \{0, 1\}$ is fixed, $y'_i = 1 - y_i$, and $\mathcal{C}''$ is the set of all concepts of shape

$$(a_1, \dots, a_{m-2}) \mapsto C(a_1, \dots, a_{m-2}, y'_{m-1}, y'_m) \text{ for } C \in \mathcal{C}' .$$

Then $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \geq \mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}') \geq \mathcal{S}[\text{MQ}](\mathcal{C}'')$, where d is the Hamming distance.

Proof. (Sketch.) Suppose L learns $\mathcal{C}'$ from q queries to a contrast oracle in the minimum distance model. We show that L can be transformed into L' which learns $\mathcal{C}''$ from at most q queries to a membership oracle. To this end, let C'' be the target concept in $\mathcal{C}''$ and let C' with $C'(\mathbf{a}_{y'_{m-1}y'_m}) = C''(\mathbf{a})$ be the corresponding concept in $\mathcal{C}'$. Note that L has not uniquely identified C' in $\mathcal{C}'$ as long as the subfunction C'' is not uniquely identified in $\mathcal{C}''$. L' can therefore identify C'' in $\mathcal{C}''$ by maintaining a simulation of L until L has identified C' in $\mathcal{C}$. Details are given in the appendix. $\square$

This lemma can be used, for example, to prove $\mathcal{S}_{\text{CS}_{\min}^d}(\text{DL}_m) \geq 2^{m-2} - 1$, where d is the Hamming distance. Our proof of this statement, however, establishes an even stronger result. To formulate this stronger result, we first need to introduce some notation.

Consider a decision list $\mathcal{L}$ of the form (1). We say that $\mathcal{L}$ has k label alternations if the number of distinct indices $i \in [z]$ such that $b_{i+1} \neq b_i$ equals k . Let DL_m^k be the subclass of DL_m which contains all decision lists with at most k label alternations. It is well known that (i) DL_m^0 contains only the constant-1 and the constant-0 function, and (ii) $\text{DL}_m^1 = \mathcal{C}_{\text{mon}}^m \cup \mathcal{C}_{\text{claus}}^m$. Now the proof of the following claim builds on Lemma 15; see the appendix for details.

Corollary 16 $\mathcal{S}_{\text{CS}_{\min}^d}(\text{DL}_m^2) \geq 2^{m-2} - 1$, where d is the Hamming distance.

Interestingly, this bound on $\mathcal{S}_{\text{CS}_{\min}^d}(\text{DL}_m)$ (using Hamming distance) is not an effect of our metric-independent lower bound in terms of SD (Thm. 12). As seen below, $\mathcal{S}_{\text{CS}_{\min}^d}(\text{DL}_m)$ asymptotically exceeds $\text{SD}(\text{DL}_m)$. Thus, the lower bound from Thm. 12 is not always asymptotically tight.

For DL_m , we saw that the sample complexity in the minimum Hamming distance model is exponential in m . By contrast, SD is at most quadratic in m . To prove this, let a *block* in $[(\ell_1, b_1), \dots, (\ell_z, b_z), b_{z+1}]$ be any maximal substring $(\ell_i, b_i), \dots, (\ell_{i+j}, b_{i+j})$ with $b_i = \dots = b_{i+j}$.

Theorem 17 *There exists a self-directed learner for DL_m that makes, on any target list $C^* \in \text{DL}_m$, at most $4km$ mistakes where k is the number of blocks in C^* .*

Proof. (Sketch.) For each of the $2m$ literals, 2 queries suffice to determine whether it occurs in the first block of the target list. Iteratively, one consumes $\leq 4m$ queries per block; see the appendix. $\square$

We now present a second example of a natural class of Boolean functions for which SD is asymptotically smaller than $\mathcal{S}_{\text{CS}_{\min}^d}(C)$, where d is the Hamming distance.

Definition 18 Fix $s, z, m \in \mathbb{N}$. An s -term z -MDNF of size m is a function $f : \{0, 1\}^m \rightarrow \{0, 1\}$ of the form $f(v_1, v_2, \dots, v_m) = M_1 \vee M_2 \vee \dots \vee M_s$, where each M_i is a monotone monomial with at most z literals. We use $\mathcal{C}_{\text{MDNF}}^{m,s,z}$ to refer to the class of all s -term z -MDNF of size m .

Theorem 19 (largely due to [13, 1]) Let d be the Hamming distance. Then $\text{SD}(\mathcal{C}_{\text{MDNF}}^{m,s,z}) = s$, while $\mathcal{S}_{\text{CS}_{\min}^d} \geq \mathcal{S}[\text{MQ}](\mathcal{C}_{\text{MDNF}}^{m-2,s-1,z-1})$, which is at least

$$(z-1)(s-1)\log(m-2) + \alpha,$$

$$\text{where } \alpha = \left(\frac{z-1}{s-1}\right)^{s-2} \text{ if } z > s, \text{ and } \alpha = \left(\frac{2s-2}{z-1}\right)^{(z-1)/2} \text{ if } z \leq s.$$

Proof. See [13] for $\text{SD}(\mathcal{C}_{\text{MDNF}}^{m,s,z}) = s$. [1] showed the lower bound on $\mathcal{S}[\text{MQ}](\mathcal{C}_{\text{MDNF}}^{m-2,s-1,z-1})$.

Let C' be the subclass of $\mathcal{C}_{\text{MDNF}}^{m,s,z}$ with concepts of the form $M_1 \vee \dots \vee M_{s-1} \vee v_m$ where $v_{m-1} \in M_i$ for $1 \leq i \leq s-1$. Then C' satisfies properties (P1) and (P2) of Lemma 15 with $y_m = 1$, and $y_{m-1} = 0$. Then $C'' = \text{MDNF}_{m-2,z-1,s-1}$, and the result follows from Lemma 15. $\square$

Theorem 19 reveals a gap between SD and the sample complexity of the minimum distance model under the Hamming distance. While we do not know of any “natural” function $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$ with which to close this gap, it can be closed if we allow the learner and the contrast oracle to share a new function d after every query. By $\mathcal{S}_{\text{CS}_{\min}^{(d^t)}}$, we denote the sample complexity resulting from interaction sequences at which, in step t , the set CS is defined using the distance function d^t . The same proof as for Theorem 12 shows that our lower bound in terms of SD (and thus the lack of an upper bound in terms of VCD, cf. Remark 13) persists despite this strengthening of the model.

Theorem 20 Fix $(d_t)_{t=1}^\infty$ and $\mathcal{C}$ over a finite domain $\mathcal{X}$. Then $\mathcal{S}_{\text{CS}_{\min}^{(d^t)}}(\mathcal{C}) \geq \lceil \text{SD}(\mathcal{C})/2 \rceil$.

However, the dynamic distance function approach helps to overcome obstacles noted in Theorem 19. One natural way to define dynamically changing functions d^t is to make them dependent on the version space. Defining $d_{C'}(x, x') := |\{C \in C' : C(x) \neq C(x')\}|/|C'|$ for any $C' \subset \mathcal{C}$, and recalling that $\mathcal{C}_t$ refers to the version space obtained from $\mathcal{C}$ after t queries, we get (see appendix for a proof):

Theorem 21 $\mathcal{S}_{\text{CS}_{\min}^{(d_{C_t})}}(\mathcal{C}_{\text{MDNF}}^{m,s,z}) \leq s = \text{SD}(\mathcal{C}_{\text{MDNF}}^{m,s,z})$, where $\mathcal{C}_t = (\mathcal{C}_{\text{MDNF}}^{m,s,z})_t$ for any t .

5 Discussion and Conclusions

We proposed and analyzed a generic framework for active learning with contrastive examples. In studying the sample complexity in two of its instantiations (the minimum distance and the proximity model), we observed interesting connections to other models of query learning, and most notably to self-directed learning, which also led to a new result on self-directed learning of decision lists (Theorem 17). Our definition allows modelling contrastive learning with a multitude of contrast set rules, and it can easily be extended to encompass passive learners receiving contrastive examples.

Our framework allows the learner to reason with perfect knowledge about the choice of the contrast set. While this is unrealistic in some practical situations, it has applications in formal methods, where the contrast oracle is a computer program that could be simulated by the learner in order to refute a hypothesis (a concept C can be excluded if the oracle does not provide the same counterexamples when C is the target). Moreover, the learner's knowledge of the mapping CS implicitly connects our framework to *self-supervised learning*. Consider, for example, passive learning under the minimum distance model. Here the learner is passively sampling examples (x, y) , which are supplemented with contrastive examples from $CS_{\min}^d$. Following our model, the learner knows that the contrastive example x' to x is the closest one to x that has a label different from y . In practice, this can be implemented as a self-supervised mechanism in which the learner internally labels with y all data points that are strictly closer to x than x' .

In future work, the condition that the learner has perfect knowledge of the set CS can be softened so as to encompass more learning settings. To this end, note that contrastive examples can still be very helpful in case of a mismatch between the learner's assumption on the oracle and the actual strategy of the oracle. Consider for instance the case of learning halfspaces: (a) First assume the oracle is the perfect minimum-distance oracle, with respect to the ℓ_2 metric. A learner with perfect knowledge of the oracle will succeed after just a single query x_1 , if it conjectures the halfspace perpendicular to the vector $x'_1 - x_1$ at the point x'_1 , and consistent with the observed labels for x_1 and x'_1 . This is because the decision boundary of the target concept is perpendicular to the vector $x'_1 - x_1$ at the point x'_1 . Note that this holds even if the learner picks the first query x_1 at random. (b) Now, suppose the same learner still expects the minimum-distance oracle, but the oracle returns a point x''_1 whose distance from the true minimum-distance counterexample may be up to γ . In this case, it is not hard to show that the classification error of the halfspace perpendicular to $x''_1 - x_1$ at x''_1 is at most

$$\frac{\cos^{-1}\left(\frac{r}{r+\gamma}\right)}{\pi}.$$

Here r refers to the distance from x_1 to the target hyperplane. Thus, even when the initial point x_1 is chosen at random, the expected classification error remains linear in γ , with just a single contrastive example.

There are also scenarios where receiving an instance with the exact minimum distance is not necessarily optimal. For example, when learning a halfspace in the presence of label noise, the perfect minimum-distance oracle would provide the learner with outliers that may be far away from the decision boundary. Here a softer variant of the minimum-distance model may be preferable. One such alternative is an oracle that returns the closest instance x' to the queried instance x , such that the majority of the k nearest neighbors of x' (for some parameter k) have a different label than x .

Acknowledgments and Disclosure of Funding

The authors would like to thank Valentio Iverson and Mohamadsadegh Khosravani for insightful discussions. Moreover, thanks are due to the anonymous reviewers for their helpful feedback on an earlier version of this paper.

Y. Chen acknowledges the support of the National Science Foundation under Grant no. IIS 2313131. S. Zilles was supported through a Canada CIFAR AI Chair at the Alberta Machine Intelligence Institute (Amii), through a Natural Sciences and Engineering Research Council (NSERC) Canada Research Chair, through the New Frontiers in Research Fund (NFRF) under grant no. NFRFE-2023-00109 and through the NSERC Discovery Grants program under grant no. RGPIN-2017-05336.

References

- [1] Hasan Abasi, Nader H. Bshouty, and Hanna Mazzawi. On exact learning monotone dnf from membership queries. In *Proceedings of the 25th International Conference on Algorithmic Learning Theory (ALT)*, pages 111–124, 2014.
- [2] Alessandro Abate, Cristina David, Pascal Kesseli, Daniel Kroening, and Elizabeth Polgreen. Counterexample guided inductive synthesis modulo theories. In *Proceedings of the 30th International Conference on Computer Aided Verification (CAV)*, pages 270–288, 2018.

- [3] Noga Alon, Dmitrii Avdiukhin, Dor Elboim, Orr Fischer, and Grigory Yaroslavtsev. Optimal sample complexity of contrastive learning. *ArXiv*, abs/2312.00379, 2023.
- [4] Dana Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, 1988.
- [5] Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 7187–7209, 2022.
- [6] Shai Ben-David, Nadav Eiron, and Eyal Kushilevitz. On self-directed learning. In *Proceedings of the 8th Annual Conference on Computational Learning Theory (COLT)*, pages 136–143, 1995.
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, (ICML)*, pages 1597–1607, 2020.
- [8] Sanjoy Dasgupta, Akansha Dey, Nicholas Roberts, and Sivan Sabato. Learning from discriminative feature feedback. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3959–3967, 2018.
- [9] Sanjoy Dasgupta and Sivan Sabato. Robust learning from discriminative feature feedback. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 973–982, 2020.
- [10] Pramith Devulapalli and Steve Hanneke. The dimension of self-directed learning. In *Proceedings of the International Conference on Algorithmic Learning Theory, (ALT)*, 2024.
- [11] Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, VC-dimension and sample compression. *J. Mach. Learn. Res.*, 15(1):3107–3131, 2014.
- [12] Sally Floyd. Space-bounded learning and the vapnik-chervonenkis dimension. In Ronald L. Rivest, David Haussler, and Manfred K. Warmuth, editors, *Proceedings of the 2nd Annual Workshop on Computational Learning Theory (COLT)*, pages 349–364, 1989.
- [13] Sally A. Goldman and Robert H. Sloan. The power of self-directed learning. *Mach. Learn.*, 14(1):271–294, 1994.
- [14] Steve Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Carnegie Mellon University, 2009.
- [15] Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 5000–5011, 2021.
- [16] Tobias Huber, Maximilian Demmler, Silvan Mertes, Matthew L. Olson, and Elisabeth André. Ganterfactual-rl: Understanding reinforcement learning agents’ strategies through visual counterfactual explanations. In Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh, editors, *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1097–1106, 2023.
- [17] Hayato Kobayashi and Ayumi Shinohara. Complexity of teaching by a restricted number of examples. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009.
- [18] Christian Kuhlmann. On teaching and learning intersection-closed concept classes. In *Proceedings of the 4th European Conference on Computational Learning Theory, EuroCOLT ’99*, page 168–182, Berlin, Heidelberg, 1999. Springer-Verlag.
- [19] Weixin Liang, James Zou, and Zhou Yu. ALICE: active learning with contrastive natural language explanations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4380–4391, 2020.

- [20] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
- [21] Oisín Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3820–3828, 2018.
- [22] Matthew L. Olson, Roli Khanna, Lawrence Neal, Fuxin Li, and Weng-Keen Wong. Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artif. Intell.*, 295:103455, 2021.
- [23] Stefanos Poulis and Sanjoy Dasgupta. Learning with feature feedback: from theory to practice. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1104–1113, 2017.
- [24] Yuta Saito and Thorsten Joachims. Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances. In *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys)*, pages 828–830, 2021.
- [25] Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.
- [26] Chaoqi Wang, Adish Singla, and Yuxin Chen. Teaching an active learner with contrastive examples. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 17968–17980, 2021.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We provide formal statements and proofs of all contributions claimed in the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We explain the limitations of our work in Sections 2 and 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each formal statement begins with stating the premises under which it is claimed. Proofs not given in full in the main body are included in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This is a purely theoretical paper, without experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This is a purely theoretical paper, without experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This is a purely theoretical paper, without experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This is a purely theoretical paper, without experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: This is a purely theoretical paper, without experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This is a purely theoretical paper, without any anticipated harm or societal impact of any kind.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: We do not anticipate any broader societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our research does not involve data or predictive/generative models, and does not pose any risks of the mentioned kind.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were not used as any non-standard, important, or original component of this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Appendix

This appendix provides the proofs omitted from the main paper.

Proposition 2 Let $\mathcal{C}$ be a countable concept class over a countable $\mathcal{X}$. Let $T : \mathcal{C} \rightarrow 2^{\mathcal{X}}$ be any injective function that maps every concept in $\mathcal{C}$ to a finite set of instances. Then:

1. There is some CS with $\mathcal{S}_{\text{CS}}(\mathcal{C}) \leq \sup_{C \in \mathcal{C}} |T(C)| + 1$.
2. If $T(C) \not\subseteq T(C')$ for $C \neq C'$, then there is some CS such that $\mathcal{S}_{\text{CS}}(\mathcal{C}) \leq \sup_{C \in \mathcal{C}} |T(C)|$.

In particular, if $\mathcal{X}$ is finite, then there is some CS such that $\mathcal{S}_{\text{CS}}(\mathcal{C}) \leq 1 + \min\{k \mid \sum_{i=0}^k \binom{|\mathcal{X}|}{i} \geq |\mathcal{C}|\}$. If $\mathcal{X}$ is countably infinite, then there is some CS such that $\mathcal{S}_{\text{CS}}(\mathcal{C}) = 1$.

Proof. Let $T : \mathcal{C} \rightarrow \mathcal{X}$ be such that $T(C) \neq T(C')$ whenever $C \neq C'$. Let $x_1, x_2, \dots$ be a fixed repetition-free enumeration of all the elements in $\mathcal{X}$.

To prove statement 1, define CS by $\text{CS}(x_i, C) = \{x_{j'}\}$ for $j' = \min\{j \mid j \geq i, x_j \in T(C)\}$. Now an active learner operates as follows. Initially, it sets $n_1 = 1$ and starts with iteration 1. In iteration i , it asks a query for x_{n_i} . If it receives $x'_i = x_j$ as a contrastive example, then note that $j \geq n_i$. The learner will then set $n_{i+1} = j + 1$ and proceed to iteration $i + 1$.

Upon its i th query, the learner will find out the label for x_{n_i} (which is irrelevant) and the i th element of $T(C)$ in the enumeration $x_1, x_2, \dots$. After $|T(C)|$ queries, the learner has seen all elements from $T(C)$, and will receive a dummy response for its next query. From its responses, it has now inferred that its target is a concept that is mapped to $T(C)$ by mapping T . By injectivity of T , this means that the learner has uniquely identified C . This proves statement 1. Since an injective mapping T using all subsets of $\mathcal{X}$ of size at most k^* can encode $\sum_{i=0}^{k^*} \binom{|\mathcal{X}|}{i}$ concepts, such T would yield $\mathcal{S}_{\text{CS}}(\mathcal{C}) \leq 1 + \min\{k \mid \sum_{i=0}^k \binom{|\mathcal{X}|}{i} \geq |\mathcal{C}|\}$.

For statement 2, note that, if $T(C) \not\subseteq T(C')$ for distinct $C, C' \in \mathcal{C}$, then the learner can infer C correctly after having seen all elements in $T(C)$. With this result, for countably infinite $\mathcal{X}$, any mapping T identifying each concept with a different instance in $\mathcal{X}$, results in $\mathcal{S}_{\text{CS}}(\mathcal{C}) = 1$. $\square$

Remark 5 Let $\mathcal{C}$ be a concept class over a complete space $\mathcal{X}$ with metric d . Then $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \leq \mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}) \leq \mathcal{S}[\text{MQ}](\mathcal{C})$.

Proof. The second inequality holds by definition. For the first inequality, consider any $r > 0$ and $(x, C) \in \mathcal{X} \times \mathcal{C}$. Due to Remark 3, $\text{CS}_{\min}^d(x, C) = \emptyset$ iff there is no x' with $C(x') \neq C(x)$. Moreover, $\text{CS}_{\text{prox}}^d(x, r, C) = \emptyset$ iff there are no points with different label converging to a point at distance at most r from x . Therefore, either $\text{CS}_{\text{prox}}^d(x, r, C) = \emptyset$, or $\text{CS}_{\text{prox}}^d(x, r, C) \supseteq \text{CS}_{\min}^d(x, C) \neq \emptyset$. In the former case, the contrastive oracle in the minimum distance model also conveys that there are no points with different label converging to a point at distance at most r from x . Consequently, the contrastive oracle in the proximity model does not provide extra information. In the later case, any contrastive example given by the contrastive oracle in the minimum distance model can be also given by the contrastive oracle in the proximity model. Again, the contrastive oracle in the proximity model does not provide extra information. $\square$

Example 3 Let $\mathcal{X} = \{0, 1\}^m$. Let $\mathcal{C}_{\text{mmon}}^m$ consist of all monotone monomials, i.e., logical formulas of the form $v_{i_1} \wedge \dots \wedge v_{i_k}$ for some pairwise distinct $i_1, \dots, i_k$ and some $k \in \{0, \dots, m\}$. The concept associated with such a formula contains the boolean vector $(b_1, \dots, b_m)$ iff $b_{i_1} = \dots = b_{i_k} = 1$ (where the empty monomial is the constant 1 function). Let d be the Hamming distance. Then:

- (1) $\mathcal{S}[\text{MQ}](\mathcal{C}_{\text{mmon}}^m) = m$.
- (2) $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}_{\text{mmon}}^m) = 1$.
- (3) $\mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}_{\text{mmon}}^m) = \Theta(\log m)$.

Proof. It remains to prove the lower bound from Claim (3).

Let $(\mathbf{x}, r)$ be the first query point of the learner. Let w be the Hamming weight of $\mathbf{x}$. The logarithmic lower bound is now obtained by an adversary argument.

If $w > m/3$, then the oracle returns $(\mathbf{x}, 1)$ and $(\mathbf{x}', 0)$ where $\mathbf{x}'$ is obtained from $\mathbf{x}$ by flipping one of the 1-entries in $\mathbf{x}$ to 0. Note that, if $j_1, \dots, j_{w-1}$ denote the positions of the 1-entries in $\mathbf{x}'$, then the index set associated with the variables in C^* could be any subset of $\{j_1, \dots, j_{w-1}\}$.

If $r \geq m/3$ and $w \leq m/3$, then we may assume without loss of generality that $r \leq m - w$. The oracle returns $(\mathbf{x}, 0)$ and $(\mathbf{x}', 1)$ where $\mathbf{x}'$ is obtained by adding r ones to $\mathbf{x}$. Note that, if $j_1, \dots, j_r$ denote the positions of the additional ones, then the index set associated with the variables in C^* could be any subset of $\{j_1, \dots, j_r\}$.

If $r < m/3$ and $w \leq m/3$, then the oracle returns $(\mathbf{x}, 0)$ and ω . Let J with $|J| = w$ be the set of indices j such that $x_j = 1$. Fix a set $K \subseteq [m]$ such that $|K| = r + 1$ and $K \cap J = \emptyset$. Let $I = [m] \setminus (J \cup K)$ be the set of the remaining indices in $[m]$. Let us make the commitment that the variables x_k with $k \in K$ are contained in C^* . Even if the learner knew about this commitment, it would still have to determine the I -indices of the remaining variables in C^* . These indices could form an arbitrary subset of I .

In any case the learner can reduce the search space for the index set of C^* not more than by a factor of 3. Hence the learner requires $\Omega(\log m)$ queries for the exact identification of C^* . $\square$

Example 4 Let $\mathcal{X}' = \{0, 1\}^{m+1}$. The concept class $\mathcal{C}'_{\text{mmon}}$ over $\mathcal{X}'$ is defined to contain, for each $C \in \mathcal{C}_{\text{mmon}}^m$, the concept $C' \in \mathcal{C}'_{\text{mmon}}$ constructed via $C'(b_1, \dots, b_m, 0) = C(b_1, \dots, b_m)$ and $C'(b_1, \dots, b_m, 1) = 1 - C(b_1, \dots, b_m)$. No other concepts are contained in $\mathcal{C}'_{\text{mmon}}$. Let d be the Hamming distance. Then $\mathcal{S}[\text{MQ}](\mathcal{C}'_{\text{mmon}}) = \mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}'_{\text{mmon}}) = \mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}'_{\text{mmon}}) = m$.

Proof. A membership query learner using m queries will work as described in the proof of Example 3, yet will leave the $(m + 1)$ st component of every queried vector equal to 0. Thus, the learner will identify which vectors of the form $(b_1, \dots, b_m, 0)$ belong to the target concept. The latter immediately implies which vectors of the form $(b_1, \dots, b_m, 1)$ belong to the target concept. Again, by a standard information-theoretic argument, fewer membership queries will not suffice in the worst case. Thus $\mathcal{S}[\text{MQ}](\mathcal{C}'_{\text{mmon}}) = m$.

Upon asking any query $(b_1, \dots, b_m, b_{m+1})$, a learner in the minimum distance model will receive the correct answer, plus the contrastive example $(b_1, \dots, b_m, 1 - b_{m+1})$. This contrastive example does not provide any additional information to the learner. Hence, the learner has to ask as many queries as an MQ-learner, i.e., $\mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}'_{\text{mmon}}) = m$.

Finally, according to Remark 5, $\mathcal{S}_{\text{CS}_{\text{min}}^d} \leq \mathcal{S}_{\text{CS}_{\text{prox}}^d} \leq \mathcal{S}[\text{MQ}]$. Thus, $\mathcal{S}_{\text{CS}_{\text{prox}}^d}(\mathcal{C}'_{\text{mmon}}) = m$. $\square$

Remark 22 Example 4 can be generalized: each concept class $\mathcal{C}$ over domain $\mathcal{X} = \{0, 1\}^n$ can be (redundantly) extended to a concept class $\mathcal{C}'$ over $\mathcal{X}' = \{0, 1\}^{n+1}$ that renders contrastive examples in the minimum distance model useless for learning concepts in $\mathcal{C}'$, under the Hamming distance.

Theorem 7 $\mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}_{\text{mon}}^m \cup \mathcal{C}_{\text{claus}}^m) = 2$, where d is the Hamming distance.

To prove this result, we begin with a straightforward extension of Example 3.

Lemma 23 $\mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}_{\text{mon}}^m) = 2$, where d is the Hamming distance.

Proof. The proof is analogous to that of Example 3, just here the learner asks two queries, one for **0** and one for **1**. The two contrastive examples then determine the target monomial, by a similar reasoning as for Example 3. $\square$

By duality³, Lemma 23 implies the following result:

Corollary 24 $\mathcal{S}_{\text{CS}_{\text{min}}^d}(\mathcal{C}_{\text{claus}}^m) = 2$, where d is the Hamming distance.

We can now proceed with the proof of Theorem 7.

Proof. (of Theorem 7.) Let $L_{\wedge}$ be the learner of $\mathcal{C}_{\text{mon}}^m$ as it is described in the proof of Lemma 23. Let $L_{\vee}$ be the learner of $\mathcal{C}_{\text{claus}}^m$ which is obtained from $L_{\wedge}$ by dualization. It suffices to describe a learner L who maintains simulations of $L_{\wedge}$ and $L_{\vee}$ and either comes in both simulations to the same

³The mapping $f(\mathbf{v}) \mapsto \bar{f}(\bar{\mathbf{v}})$ transforms the monomial $\bigwedge_{i \in I} \bar{v}_i \wedge \bigwedge_{j \in J} v_j$ into the clause $\bigvee_{i \in I} \bar{v}_i \vee \bigvee_{j \in J} v_j$, and vice versa. The same mapping can be used to dualize the proof of Lemma 23 in the sense that, within this proof, every classification label is negated and every point from B_m is componentwise negated. This concerns the query points chosen by the learner as well as the contrastive examples returned by the oracle.

conclusion about the target concept⁴ or realizes an inconsistency in one of the simulations, which can then be aborted. Details follow.

If the target concept is representable as a literal, then L will not run into problems because both simulations lead to the same (correct) result. From now we assume that the target concept cannot be represented by a literal. Remember that both learners, $L_{\wedge}$ and $L_{\vee}$, choose the two query points **1** and **0**. Let us first consider the case that the target function is the 1-constant (= empty monomial). Then the oracle returns the label 1 for both query points along with an error message that there is no point of opposite label in B_m . After having received this error message, L can abort both simulations because it has identified the 1-constant as the target function. The reasoning for the constant-0 function is analogous.

Let us now consider the case that the target function can be represented by a monomial M^* of length at least 2 resp. by a corresponding decision list. If M^* contains at least one negated and at least one not negated variable, then the oracle returns label 0 for both query points plus an extra-point of opposite label. If the target function were a clause, then the case of label 0 for both query points can happen only if the target function would be the 0-constant (so that no extra-point of opposite label could be returned by the oracle). Hence L may abort the simulation of $L_{\vee}$ and identify the target function via $L_{\wedge}$. Let us now assume that M^* does not contain negated variables so that M^* is of the form $\bigwedge_{j \in J} v_j$ for some $J \subseteq [m]$ with $|J| \geq 2$. Then, upon query point **0**, the oracle returns label 0 plus the extra-point **1_J**. The learner $L_{\vee}$ concludes from seeing label 0 that the target clause does not contain a negated variable. But $L_{\vee}$ would expect to see an extra-point with only a single 1-component (because a clause with only unnegated variables can be satisfied by a single 1-bit in the right position). Hence L can abort the simulation of $L_{\vee}$ and continue with the simulation of $L_{\wedge}$ until M^* is identified.

All cases that we did not discuss are dual to cases that we actually have discussed. Thus the proof of the corollary is accomplished. $\square$

Theorem 10 $\mathcal{S}[\text{EX}^+, \text{EX}^-, \text{MQ}](\mathcal{C}) - 2 \leq \mathcal{S}_{\text{CS}_{\min}^{d_0}}(\mathcal{C}) \leq \mathcal{S}[\text{EX}^+, \text{EX}^-, \text{MQ}](\mathcal{C})$.

Proof. As for the second inequality, it suffices to show that each call of one of the oracles $\text{EX}^+, \text{EX}^-, \text{MQ}$ can be simulated by a single call of the contrast oracle. This is clearly true for a call of the MQ-oracle (because the contrast oracle is even more informative). A call of EX^+ -oracle can be simulated as follows:

1. Choose an (arbitrary) instance $x \in \mathcal{X}$ as input of the contrast oracle and get back a pair (b, x') such that $b = C^*(x) \neq C^*(x')$.
2. If $b = 1$, then return x (as the desired positive example). Otherwise return x' .

A similar reasoning shows that a call of the EX^- -oracle can be simulated by a call of the contrast oracle. Hence the second inequality is valid.

As for the first inequality, it suffices to show that q queries of the contrast oracle can be simulated by q calls of the membership oracle and two extra-calls which are addressed to the oracles providing a (positive or negative) example. For $i = 1, \dots, q$, let x_i be the i -th query instance that is given as input to the contrast oracle and let (b_i, x'_i) with $b_i = C^*(x_i) \neq C^*(x'_i)$ be the pair returned by the latter. Note that x'_i can be any instance with label opposite to b_i because the discrete metric assigns the same distance value, 1, to any such pair (x, x') . The above q queries can therefore be simulated as follows:

1. Call the oracle EX^- and obtain an instance y_0 such that $C^*(y_0) = 0$.
2. Call the oracle EX^+ and obtain an instance y_1 such that $C^*(y_1) = 1$.
3. For $i = 1, \dots, q$, choose x_i as input of the MQ-oracle and get back the label $b_i = C^*(x_i)$. If $b_i = 0$, then return $(0, y_1)$ as a valid answer of the contrast oracle when the latter is called with query instance x_i . If $b_i = 1$, then return $(1, y_0)$ instead.

This completes the proof. $\square$

Example 7 Let $\mathcal{X} = \{x_1, x_2, x_3\}$. Let $\mathcal{C}$ consist of the concepts $\{x_1, x_3\}$, $\{x_3\}$, $\{x_2, x_3\}$, and $\{x_2\}$. Then $\mathcal{C}$ satisfies $\text{SD}(\mathcal{C}) = \text{VCD}(\mathcal{C}) = 1$, but $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \geq 2$ for any $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0}$.

⁴This happens when the target concept is a literal, i.e., both a monomial of length 1 and a clause of length 1.

Table 2: A concept class $\mathcal{C}$ with $\text{SD}(\mathcal{C}) = \text{VCD}(\mathcal{C}) = 1$, but $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \geq 2$ for any metric d .

	x_1	x_2	x_3
C_1	1	0	1
C_2	0	0	1
C_3	0	1	1
C_4	0	1	0

Proof. The concept class is shown in tabular form in Table 2.

Let d be any distance function and let $i(1), i(2), i(3)$ be any permutation of 1, 2, 3. Let $E_{\{i(1), i(2)\}}$ be the event that (L chooses query point $x_{i(1)}$ and $d(x_{i(2)}, x_{i(1)}) \leq d(x_{i(3)}, x_{i(1)})$) or (L chooses query point $x_{i(2)}$ and $d(x_{i(1)}, x_{i(2)}) \leq d(x_{i(3)}, x_{i(2)})$). Note that at least one of the events $E_{\{1,2\}}$, $E_{\{1,3\}}$ and $E_{\{2,3\}}$ must occur. It suffices therefore to show that, in any case, the oracle has a response resulting in a version space of size 2. This can be achieved as follows:

- In case of $E_{\{1,2\}}$, the response $\{(x_1, 0), (x_2, 1)\}$ would lead to the version space $\{C_3, C_4\}$.⁵
- In case of $E_{\{1,3\}}$, the response $\{(x_1, 0), (x_3, 1)\}$ would lead to the version space $\{C_2, C_3\}$.
- In case of $E_{\{2,3\}}$, the response $\{(x_2, 0), (x_3, 1)\}$ would lead to the version space $\{C_1, C_2\}$.

□

Theorem 14 For any finite $\mathcal{C}$ with $\text{VCD}(\mathcal{C}) = 1$ there exists a metric d with $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \leq 2$.

To prove this theorem, we observe the following useful fact, which was noted by [11].

Lemma 25 ([11]) Let $\mathcal{C}$ be any concept class with $\text{VCD}(\mathcal{C}) = 1$. Then there is at least one $(x, b) \in \mathcal{X} \times \{0, 1\}$ such that there exists exactly one $C \in \mathcal{C}$ with $C(x) = b$.

Proof. (of Theorem 14). It is well-known that every finite concept class of VCD 1 is contained in a concept class of size $|\mathcal{X}| + 1$ whose VCD is still 1, see, e.g., [12]. Thus, we may assume that $\mathcal{C} = |\mathcal{X}| + 1$. Set $n = |\mathcal{X}|$. By an iterative application of Lemma 25, we see that there exists an order $C_1, \dots, C_n, C_{n+1}$ of the concepts in $\mathcal{C}$, an order $x_1, \dots, x_n$ of the instances in $\mathcal{X}$ and a binary sequence $b_1, \dots, b_n$ such that, for $i = 1, \dots, n$, $C = C_i$ is the only concept in $\mathcal{C} \setminus \{C_1, \dots, C_{i-1}\}$ which satisfies the equation $C(x_i) = b_i$. Let $i^+(1) < \dots < i^+(r)$ and $i^-(1) < \dots < i^-(s)$ with $r + s = n$ be given by $b_{i^+(j)} = 1$ for $j = 1, \dots, r$ and $b_{i^-(j)} = 0$ for $j = 1, \dots, s$. In the sequel, we assume that $r, s \geq 1$.⁶ The following implications are easy to verify.

- (I1) If $C(x_j) = 1 - b_j$ for all $j \in [n]$, then $C = C_{n+1}$.
- (I2) If $C(x_{i^+(j)}) = 0$ for all $j \in [r]$, and j^- is the smallest $j \in [s]$ with $C(x_{i^-(j)}) = 0$, then $C = C_{i^-(j^-)}$.
- (I3) If $C(x_{i^-(j)}) = 1$ for all $j \in [s]$, and j^+ is the smallest $j \in [r]$ with $C(x_{i^+(j)}) = 1$, then $C = C_{i^+(j^+)}$.
- (I4) If (i) j^- is the smallest $j \in [s]$ such that $C(x_{i^-(j)}) = 0$, (ii) j^+ is the smallest $j \in [r]$ such that $C(x_{i^+(j)}) = 1$, and (iii) $i^+(j^+) > i^-(j^-)$, then $C = C_{i^-(j^-)}$.
- (I5) If (i) j^- is the smallest $j \in [s]$ such that $C(x_{i^-(j)}) = 0$, (ii) j^+ is the smallest $j \in [r]$ such that $C(x_{i^+(j)}) = 1$, and (iii) $i^-(j^-) > i^+(j^+)$, then $C = C_{i^+(j^+)}$.

It thus suffices to choose a metric d such that the oracle can be forced (with 2 minimum-distance queries) to make one of the above five implications applicable. To this end, we define

$$d(x_i, x_j) = \begin{cases} |i - j| & \text{if } b_i = b_j \\ n & \text{if } b_i \neq b_j \end{cases}.$$

⁵The learner can take no advantage from knowing which of x_1, x_2 is in the role of the query point and which is in the role of the contrastive example.

⁶The proof will become simpler if r or s equals 0.

The two query points, chosen by the learner, are $x_{i^+(1)}$ and $x_{i^-(1)}$.

Suppose first that the oracle returns the label 1 for $x_{i^+(1)}$. If $i^+(1) = 1$, then $C = C_1$ and we are done. We may therefore assume that $i^+(1) > 1$, which implies that $i^-(1) = 1$. If the oracle returns the label 0 for $x_{i^-(1)} = x_1$, then $C = C_1$ and we are done. We may therefore assume that the oracle returns the label 1 for x_1 . If there is no contrastive example for $(x_1, 1)$, then (I3) applies. Otherwise, let $(x_j, 0)$ be the contrastive example for $(x_1, 1)$. If $j < i^+(1)$, then (I4) applies. If $j > i^+(1)$, then either (I3) (in case $d(x_1, x_j) = n$) or (I5) (in case $d(x_1, x_j) < n$) applies.

Due to symmetry, one can reason analogously if the oracle returns the label 0 for $x_{i^-(1)}$.

Finally, consider the case that the oracle returns the label 0 for $x_{i^+(1)}$ and the label 1 for $x_{i^-(1)}$. It follows that the contrastive example for $x_{i^+(1)}$ must be labeled 1 and the contrastive example for $x_{i^-(1)}$ must be labeled 0. The following implications are easy to verify:

1. If the contrastive example for $x_{i^+(1)}$ is an instance x_i with $b_i = 0$ and the contrastive example for $x_{i^-(1)}$ is an instance x_i with $b_i = 1$, then (I1) applies.
2. If the contrastive example for $x_{i^+(1)}$ is an instance x_i with $b_i = 0$ and the contrastive example for $x_{i^-(1)}$ is an instance of the form $x_{i^-(j^-)}$, then (I2) applies.
3. If the contrastive example for $x_{i^-(1)}$ is an instance x_i with $b_i = 1$ and the contrastive example for $x_{i^+(1)}$ is an instance of the form $x_{i^+(j^+)}$, then (I3) applies.
4. If (i) the contrastive example for $x_{i^+(1)}$ is an instance of the form $x_{i^+(j^+)}$, (ii) the contrastive example for $x_{i^-(1)}$ is an instance of the form $x_{i^-(j^-)}$, and (iii) $i^+(j^+) > i^-(j^-)$, then (I4) applies.
5. If (i) the contrastive example for $x_{i^+(1)}$ is an instance of the form $x_{i^+(j^+)}$, (ii) the contrastive example for $x_{i^-(1)}$ is an instance of the form $x_{i^-(j^-)}$, and (iii) $i^-(j^-) > i^+(j^+)$, then (I5) applies.

Hence each possible response of the oracle leads to the exact identification of the target concept. $\square$

Lemma 15 Suppose that $\mathcal{C}$ is a concept class over B_m , and $\mathcal{C}'$ is a subclass of $\mathcal{C}$ with the properties

(P1) For each $\mathbf{a} \in B_m$ with $a_m = y_m$ and each $C \in \mathcal{C}'$, we have that $C(\mathbf{a}) = 1$.

(P2) For each $\mathbf{a}$ with $a_m = y'_m$ and $a_{m-1} = y_{m-1}$ and each $C \in \mathcal{C}'$, we have that $C(\mathbf{a}) = 0$.

For $i \in \{m, m-1\}$, here $y_i \in \{0, 1\}$ is fixed, $y'_i = 1 - y_i$, and $\mathcal{C}''$ is the set of all concepts of shape

$$(a_1, \dots, a_{m-2}) \mapsto C(a_1, \dots, a_{m-2}, y'_{m-1}, y'_m) \text{ for } C \in \mathcal{C}'.$$

Then $\mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}) \geq \mathcal{S}_{\text{CS}_{\min}^d}(\mathcal{C}') \geq \mathcal{S}[\text{MQ}](\mathcal{C}'')$, where d is the Hamming distance.

Proof. Suppose that L learns $\mathcal{C}'$ from a contrast oracle in the minimum distance model at the expense of q queries. It suffices to show that L can be transformed into L' which learns $\mathcal{C}''$ from a membership oracle at the expense of at most q queries. To this end, let C'' be the target concept in $\mathcal{C}''$ and let C' with $C'(\mathbf{a}'_{m-1}y'_m) = C''(\mathbf{a})$ be the corresponding concept in $\mathcal{C}'$. Note that L has not uniquely identified C' in $\mathcal{C}'$ as long as the subfunction C'' is not uniquely identified in $\mathcal{C}''$. L' can therefore uniquely identify C'' in $\mathcal{C}''$ by maintaining a simulation of L until L has uniquely identified C' in $\mathcal{C}$. In order to explain how L' simulates the contrast oracle, we proceed by case distinction:

- If L chooses a query point of the form $\mathbf{a}'_{m-1}y'_m$, then L' chooses the query point $\mathbf{a}$ and receives the label $b := C''(\mathbf{a})$ from its membership oracle. Then L' returns b and the point $\mathbf{a}'$ to L where $\mathbf{a}' = \mathbf{a}'_{m-1}y_m$ if $b = 0$ and $\mathbf{a}' = \mathbf{a}_{m-1}y'_m$ if $b = 1$. Note that $(b, \mathbf{a}')$ is among the admissible answers of the contrast oracle.
- If L chooses a query point of the form $\mathbf{a}_{m-1}y'_m$, then L' returns the label 0 and the additional point $\mathbf{a}_{m-1}y_m$. Again, this is among the admissible answers of the contrast oracle.
- If L chooses a query point of the form $\mathbf{a}_{m-1}y_m$, then L' returns the label 1 and the additional point $\mathbf{a}'_{m-1}y'_m$. Again, this is among the admissible answers of the contrast oracle.
- If L chooses a query-point of the form $\mathbf{a}'_{m-1}y_m$, then L' chooses the query point $\mathbf{a}$ and receives the label $b = C''(\mathbf{a})$ from its membership oracle. Then L' returns the label 1 and

the point $\mathbf{a}'$ to L where $\mathbf{a}' = \mathbf{a}y'_{m-1}y'_m$ if $b = 0$ and $\mathbf{a}' = \mathbf{a}y_{m-1}y'_m$ if $b = 1$. Again, this is among the admissible answers of the contrast oracle. Here note that, in case $b = 1$, there is no admissible answer at distance 1 from the query point, so that a contrastive example at distance 2 can be chosen.

Clearly L'' can maintain this simulation until it reaches exact identification of the target concept. Moreover, the number of query point chosen by L' does not exceed the number of query points chosen by L . $\square$

Corollary 16 $\mathcal{S}_{\text{CS}_{\min}^d}(\text{DL}_m^2) \geq 2^{m-2} - 1$, where d is the Hamming distance.

Proof. Let $\mathcal{C} = \text{DL}_m^2$. Moreover, let $\mathcal{C}' \subseteq \mathcal{C}$ be a subset of DL_m^2 of decision lists of the form

$$\mathcal{C} = [(v_m, 1), (v_{m-1}, 0), \underbrace{(\ell_1, b_1), \dots, (\ell_z, b_z), b_{z+1}}_{=\mathcal{C}' \in \text{DL}_{m-2}^1}],$$

where $z \geq 0$, $\ell_i \in \{v_1, \bar{v}_1, \dots, v_{m-2}, \bar{v}_{m-2}\}$ and $b_i \in \{0, 1\}$. $\mathcal{C}'$ satisfies properties (P1) and (P2) with $y_{m-1} = y_m = 1$ in Lemma 15. Then $\mathcal{C}'' = \text{DL}_{m-2}^1$. $\mathcal{C}_{\text{mon}}^{m-2}$ is a subclass of DL_{m-2}^1 and $2^{m-2} - 1$ membership queries are needed for learning $\mathcal{C}_{\text{mon}}^{m-2}$ [17]. $\square$

Theorem 17 *There exists a self-directed learner for DL_m that makes, on any target list $C^* \in \text{DL}_m$, at most $4km$ mistakes where k is the number of blocks in C^* .*

Proof. Recall $B_m = \{0, 1\}^m$. For any set of literals $\ell_{i_1}, \dots, \ell_{i_j}$, we use $B_m(\ell_{i_1} = b_1, \dots, \ell_{i_j} = b_j)$ to denote the $(m - j)$ -dimensional subcube resulting from fixing the value of ℓ_{i_t} to b_t , for $t \in [j]$. We call a literal ℓ b -pure if the target list C^* assigns label b to every point in $B_m(\ell = 1)$.

The b -purity of $\ell = v_i$ (resp. of $\ell = \bar{v}_i$) can be checked by testing whether C^* restricted to $B_m(v_i = 1)$ (resp. to $B_m(v_i = 0)$) degenerates to the constant- b function. This is done at the expense of at most 2 mistakes per literal: The learner predicts 0 for a first point in $B_m(\ell = 1)$, receives the correct label b (which may or may not count as one mistake), and then predicts b for further points in $B_m(\ell = 1)$ until a mistake is made. If no mistake is made on further points in $B_m(\ell = 1)$, then ℓ is b -pure; otherwise it is not pure. Hence, after having made at most $4m$ mistakes, the learner L knows all pure-literals, say $\ell_1, \dots, \ell_t$, and L also knows the (unique) bit $b \in \{0, 1\}$ with respect to which $\ell_1, \dots, \ell_t$ are pure.

At this point L knows the first block $(\ell_1, b), \dots, (\ell_t, b)$ of the list C^* . L can now proceed iteratively in order to learn the remaining blocks. As for the second block, all query points are taken from the $(m - t)$ -dimensional subcube $B_m(\ell_1 = 1 - b, \dots, \ell_t = 1 - b)$ so that they are not absorbed by the first block. In this way, C^* is identified blockwise at the expense of at most $4m$ queries per block. $\square$

Theorem 21 $\mathcal{S}_{\text{CS}_{\min}^{(d_{C_t})}}(\mathcal{C}_{\text{MDNF}}^{m,s,z}) \leq s = \text{SD}(\mathcal{C}_{\text{MDNF}}^{m,s,z})$, where $\mathcal{C}_t = (\mathcal{C}_{\text{MDNF}}^{m,s,z})_t$ for any t .

The proof of this theorem makes use of the following terminology. Suppose $\mathbf{a} = (a_1, \dots, a_m)$, $\mathbf{b} = (b_1, \dots, b_m) \in B^m$ for some m . We write $\mathbf{a} \leq \mathbf{b}$ if $a_i = 1$ implies $b_i = 1$. Moreover, we implicitly identify the Boolean vector $\mathbf{a}$ with the corresponding monotone monomial.

First we introduce a helpful lemma.

Lemma 26 *Let $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \{0, 1\}^m$ such that $\mathbf{a} \leq \mathbf{b} \leq \mathbf{c}$. Then for any $\mathcal{C}' \subseteq \mathcal{C}_{\text{MDNF}}^{m,s,z}$ we have $d_{\mathcal{C}'}(\mathbf{a}, \mathbf{b}) \leq d_{\mathcal{C}'}(\mathbf{a}, \mathbf{c})$.*

Proof. Consider any $C \in \mathcal{C}_{\text{MDNF}}^{m,s,z}$. Since C is monotone, if $C(\mathbf{a}) = 1$ then we have $C(\mathbf{b}) = 1$, $C(\mathbf{c}) = 1$. Similarly, if $C(\mathbf{a}) = 0$ and $C(\mathbf{b}) = 1$ then $C(\mathbf{c}) = 1$ as well. This completes the proof. $\square$

Proof. (of Theorem 21). The claim $\text{SD}(\mathcal{C}_{\text{MDNF}}^{m,s,z}) = s$ was proven by [13] and already stated in Theorem 19. It suffices to define a learner that witnesses $\mathcal{S}_{\text{CS}_{\min}^{(d_{C_t})}}(\mathcal{C}_{\text{MDNF}}^{m,s,z}) \leq s$.

Define a learner that asks, in each of s iterations, a query for the vector $\mathbf{0}$. Since $\mathbf{0}$ has label 0, all the contrastive examples have label 1. Let $\mathbf{b}^t$ be the contrastive example for the t th query.

First, we argue that $\mathbf{b}^1$ is a monomial in the target concept C^* . Suppose $\mathbf{b}^1$ is not a monomial. Since it has label 1, there is a monomial $\mathbf{b}' \leq \mathbf{b}^1$. Using Lemma 26, we have $d_{C_1}(0, \mathbf{b}') \leq d_{C_1}(0, \mathbf{b}^1)$. Moreover, since $\mathbf{b}^1 \neq \mathbf{b}'$, there exists $C \in \mathcal{C}_{\text{MDNF}}^{m,s,z}$ such that $C(\mathbf{b}') = 0$, but $C(\mathbf{b}^1) = 1$. Thus, $d_{C_1}(0, \mathbf{b}') < d_{C_1}(0, \mathbf{b}^1)$. Therefore $\mathbf{b}'$ should have been the contrastive example, which is a contradiction.

Next, we prove inductively that $\{\mathbf{b}^1, \dots, \mathbf{b}^s\}$ is the set of monomials in C^* . Let $\{\mathbf{b}^1, \dots, \mathbf{b}^{t-1}\}$ be $t - 1$ distinct monomials in C^* . We prove that $\mathbf{b}^t$ is a monomial in C^* distinct from $\mathbf{b}^1, \dots, \mathbf{b}^{t-1}$.

Note that $\mathcal{C}_{t-1}$ is the set of concepts that have label 1 on all of $\mathbf{b}^1, \dots, \mathbf{b}^{t-1}$. We also know that all concepts give label 0 to the all zero vector. Thus, according to Lemma 26 for any $\mathbf{a}$ with $\mathbf{b}^q \leq \mathbf{a}$ for some $q \in [t - 1]$, we have $d_{C_1}(0, \mathbf{a}) \geq d_{C_1}(0, \mathbf{b}^q) = 1$ (maximum distance). Thus, $\mathbf{b}^q \not\leq \mathbf{b}^t$. With an argument similar to one we used for the base case, $\mathbf{b}^t$ is also a monomial. This completes the proof. $\square$