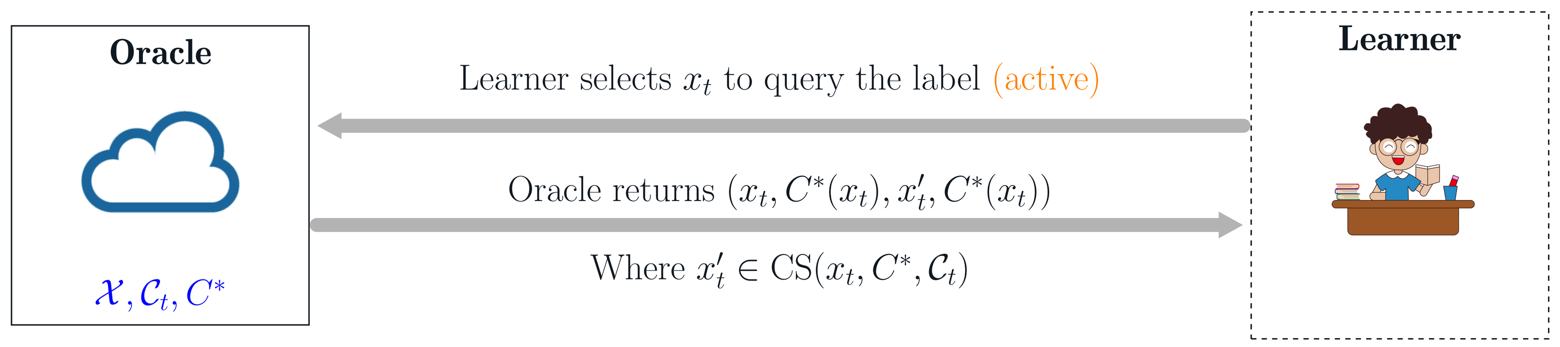
Formal Models of Active Learning from Contrastive Examples

Farnam Mansouri[1] **Hans U. Simon**[2] **Adish Singla**[3] **Yuxin Chen**[4] **Sandra Zilles**[5]

[1]University of Waterloo [2]Ruhr-University Bochum [3]Max Planck Institute for Software Systems [4]University of Chicago [5]University of Regina

Setup



Goal of learning

- **Identification:** $\mathcal{S}_{\mathrm{CS}}(\mathcal{C})$ is the minimum number of queries needed for finding any target labeling rule $C^* \in \mathcal{C}$.
- **$\varepsilon-$approximation:** $\mathcal{S}_{\mathrm{CS}}(\mathcal{C}, \varepsilon)$ is the minimum number of queries needed for finding a C with symmetric difference less than ε (according to uniform distribution over $\mathcal{X}$) to any $C^* \in \mathcal{C}$.

Contrast Set Mapping

- **Minimum Distance Model** $\mathrm{CS}^d_{\mathrm{min}}$**:** Given instance x, provides closest point to x with opposite label.
- **Proximity Model** $\mathrm{CS}^d_{\mathrm{prox}}$**:** Given instance x and distance $r > 0$, returns any x' with opposite label to x such that $d(x, x') < r$, and ω if no such x' exists.

Sample Complexity of Some Classes

Table 1: Complexity of $\varepsilon-$approximation with ℓ_1-metric

	without CS	$\mathrm{CS}^d_{\mathrm{prox}}$	$\mathrm{CS}^d_{\mathrm{min}}$ (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 2: Complexity of identification with Hamming distance

	without CS	$\mathrm{CS}^d_{\mathrm{prox}}$	$\mathrm{CS}^d_{\mathrm{min}}$
(a) Monotonic Monomials	m	$\Theta(\log m)$	1
(b) Monomials	m	$\Theta(\log(m))$	2
(c) Clauses	m	$\Theta(\log(m))$	2
(d) Parity functions	m	$[m-1, m]$	$[m-1, m]$

Monotonic monomials, monomials, clauses and parity are functions over Boolean variables $\{v_1, \ldots, v_m\}$ of the form respectively (a) $v_{i_1} \wedge \ldots \wedge v_{i_k}$; (b) $\ell_{i_1} \wedge \ldots \wedge \ell_{i_k}$; (c) $\ell_{i_1} \vee \ldots \vee \ell_{i_k}$; and (d) $v_{i_1} \oplus \ldots \oplus v_{i_k}$ for some pairwise distinct $i_1, \ldots, i_k$, where $\ell_j \in \{v_j, \bar{v}_j\}$.

Relationship with Self Directed Learning Complexity

Definition 1 ([GS94]) *(informal) Let $\mathcal{C}$ be a concept class over a finite domain $\mathcal{X}$, $n = |\mathcal{X}|$. Assume an unknown $C^* \in \mathcal{C}$. In each of n rounds, a self-directed learner selects a new instance $x \in \mathcal{C}$, predicts a label b, and receives $C^*(x)$ as feedback. A mistake occurs when $C^*(x) \neq b$. The self-directed learning complexity of $\mathcal{C}$, denoted by $\mathrm{SD}(\mathcal{C})$, is the minimum number m such that every self-directed learner makes at least m mistakes for some $C^* \in \mathcal{C}$.*

Theorem 1 *If $\mathcal{C}$ is defined over a finite $\mathcal{X}$, and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{\geq 0}$, then $\mathcal{S}_{\mathrm{CS}^d_{\mathrm{min}}}(\mathcal{C}) \geq \lceil \mathrm{SD}(\mathcal{C})/2 \rceil$.*

Proof sketch:

- We can construct a self-directed learner from a minimum-distance learner L.
- Whenever L poses a query x, execute the following subroutine: (i) choose x and predict 0; (ii) choose x' with minimal $d(x, x')$ and predict the label $C^*(x)$ until a mistake occurs.
- The total number of subroutines is bounded by $2\mathcal{S}_{\mathrm{CS}^d_{\mathrm{min}}}(\mathcal{C})$.
- In fact, the theorem still holds even if d_t changes each time t.

Lower Bounds

Example 1 *Let $\mathcal{C}$ be a class of functions over variables $\{v_1, \ldots, v_m\}$ and d be the Hamming distance. Denote, $\mathcal{C}' = \{C \oplus v_{m+1} \mid C \in \mathcal{C}\}$. Then*

$$\mathcal{S}[\mathrm{MQ}](\mathcal{C}') = \mathcal{S}_{\mathrm{CS}^d_{\mathrm{min}}}(\mathcal{C}') = \mathcal{S}_{\mathrm{CS}^d_{\mathrm{prox}}}(\mathcal{C}') = \mathcal{S}[\mathrm{MQ}](\mathcal{C}).$$

Definition 2 *Define the class of 1-decision lists DL_m over the Boolean variables $\{v_1, \ldots, v_m\}$ such that*

- *Each function $f_{\mathcal{L}} \in \mathrm{DL}_m$ is represented by a list of shape*
  $$\mathcal{L} = [(\ell_1, b_1), \ldots, (\ell_z, b_z), b_{z+1}] \ ,$$
  where $z \geq 0$, $\ell_i \in \{v_1, \bar{v}_1, \ldots, v_m, \bar{v}_m\}$ and $b_i \in \{0, 1\}$ (each variable occurs at most once).
- *For a point $\mathbf{a} \in \{0,1\}^m$, let $i^* \in [z]$ be minimal number such that ℓ_i applied to $\mathbf{a}$ is 1. If such i^* exists define $f_{\mathcal{L}}(\mathbf{a}) = b_{i^*}$, otherwise define $f_{\mathcal{L}}(\mathbf{a}) = b_{z+1}$.*

Let DL^k_m be the subclass of DL_m which contains all decision lists with at most k label alternations.

Theorem 2 *Let d be the Hamming distance. Then, $\mathrm{SD}(\mathrm{DL}^k_m) \leq km$, while $\mathcal{S}_{\mathrm{CS}^d_{\mathrm{min}}}[\mathrm{DL}^2_m] \geq 2^{m-2} - 1$.*

Definition 3 *An s-term z-MDNF of size m is a function $f : \{0,1\}^m \to \{0,1\}$ of the form $f(v_1, v_2, \cdots, v_m) = M_1 \vee M_2 \vee \cdots \vee M_s$, where each M_i is a monotone monomial with at most z literals. We use $\mathcal{C}^{m,s,z}_{\mathrm{MDNF}}$ to refer to the class of all s-term z-MDNF of size m.*

Theorem 3 (largely due to [GS94; ABM14]) *Let d be the Hamming distance. Then $\mathrm{SD}(\mathcal{C}^{m,s,z}_{\mathrm{MDNF}}) = s$, while $\mathcal{S}_{\mathrm{CS}^d_{\mathrm{min}}}(\mathcal{C}^{m,s,z}_{\mathrm{MDNF}}) \geq \mathcal{S}[\mathrm{MQ}](\mathcal{C}^{m-2,s-1,z-1}_{\mathrm{MDNF}})$ which is at least*

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

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

An Upper Bound for a Dynamic Metric

Theorem 4 *For any for any concept class $\mathcal{C}'$ define $d_{\mathcal{C}'}(x, x') := |\{C \in \mathcal{C}' : C(x) \neq C(x')\}|/|\mathcal{C}'|$. Then $\mathcal{S}_{CS^{(d_{\mathcal{C}_t})}_{\mathrm{min}}}(\mathcal{C}^{m,s,z}_{\mathrm{MDNF}}) \leq s = \mathrm{SD}(\mathcal{C}^{m,s,z}_{\mathrm{MDNF}})$, where $\mathcal{C}_t$ is the given version space at time t.*

Proof idea:

- Let $C^* = M_1 \vee \cdots \vee M_s$. Suppose the learner queries $\mathbf{0}$ exactly s times. For each round $t \in [s]$, the corresponding contrastive example will be associated with some M_{i_t}.
- After the t-th query, the version space will contain only concepts that contain M_{i_t}. This ensures that $i_1, \ldots, i_s$ are pairwise distinct. Consequently, the only concept remaining in the version space at time s will be C^*.

Discussion

- Our bounds for self-directed learning ensure that the complexity of the minimum-distance and proximity models always lies between the complexity of self-directed learning and the sample complexity of membership queries.
- Our framework considers learners and oracles with perfect knowledge. An application in formal methods: the contrast oracle can be viewed as a computer program simulated by the learner to refute a hypothesis.
- Future directions: extension to setups with imperfect knowledge or agnostic learning.

References

[GS94] Sally A. Goldman and Robert H. Sloan. "The Power of Self-Directed Learning". In: *Mach. Learn.* 14.1 (1994), pp. 271–294.

[ABM14] 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)*. 2014, pp. 111–124.