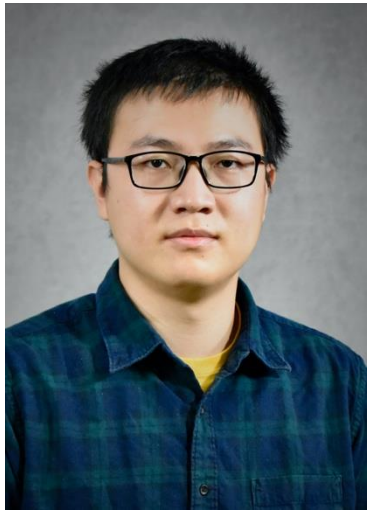
Towards Fundamental Limits for Active Multi-distribution Learning



Chicheng Zhang

THE UNIVERSITY OF ARIZONA

Yihan (Joey) Zhou

TEXAS
The University of Texas at Austin

Motivation

- Classic PAC learning models only concerns one underlying distribution, which can be viewed as the preference of one single group.

$L(h) = 0.7$

$L(h) = 0.3$

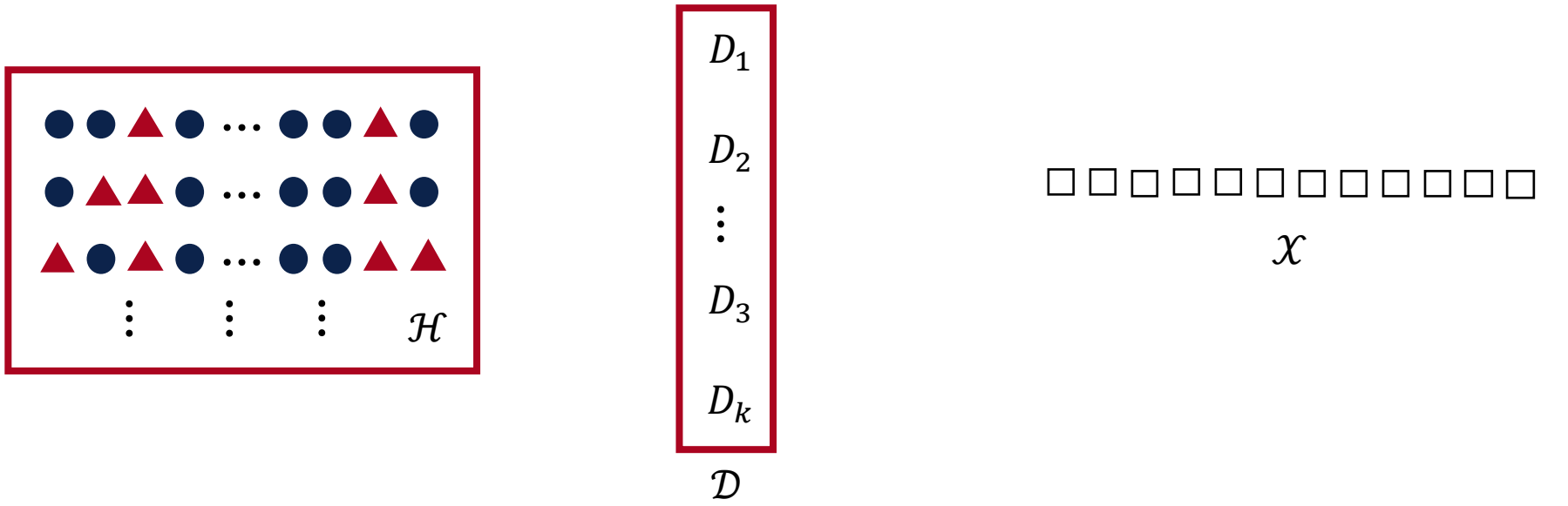
$L(h) = 0.9$

h

- Blum, Haghtalab, Procaccia and Qiao (2017) proposed multi-distribution learning, which measure the performance of any hypothesis on the worst distribution.
- This learning paradigm models fairness, robustness, multi-agent collaboration etc..

Problem Definition

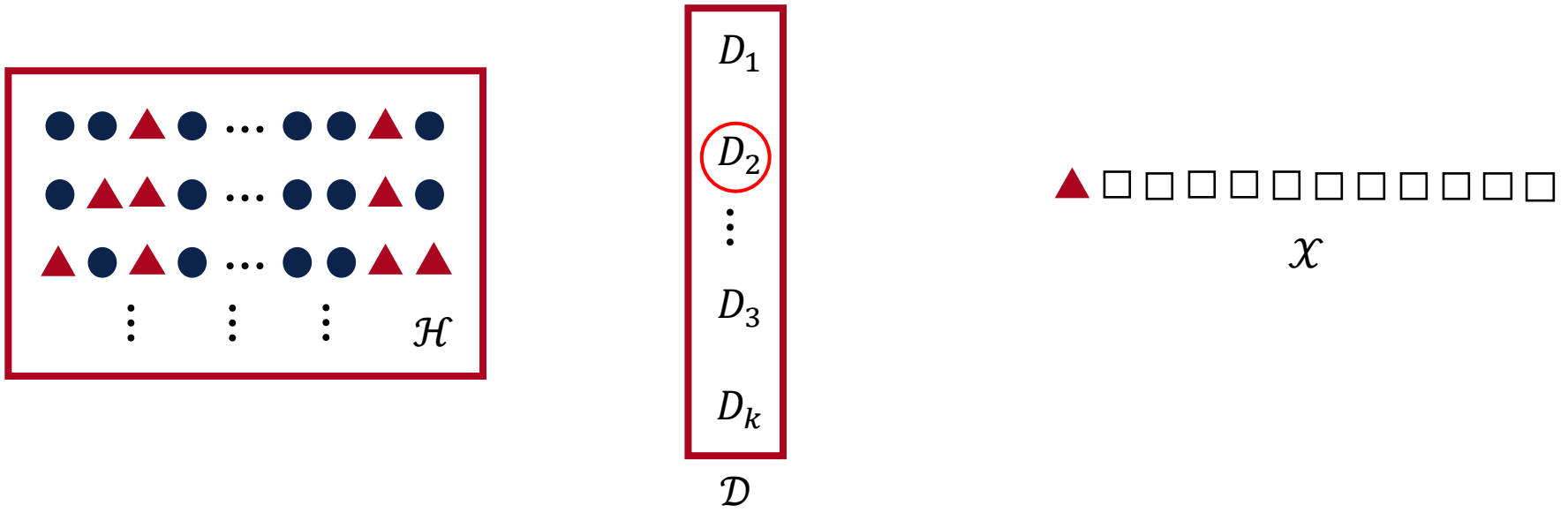
- Given a hypothesis class $\mathcal{H}$ of VC dimension d, a set $\mathcal{D}$ of k distributions $\{D_1, \dots, D_k\}$ over a dataset $\mathcal{X} \times \mathcal{Y}$, an error tolerance parameter ε and a failure parameter δ. Let $L(h) = \max_{i \in [k]} L(h, D_i)$ be the loss and assume $\min_{h \in \mathcal{H}} L(h) = \nu$.



- A learner can pick a distribution D_i and $x \in \mathcal{X}$ and query the label $y \sim D_i(\cdot \mid x)$. The goal is to return a $\hat{h}$ such that $L(\hat{h}) \leq \nu + \varepsilon$ with probability at least $1 - \delta$ with as few queries as possible.

Problem Definition

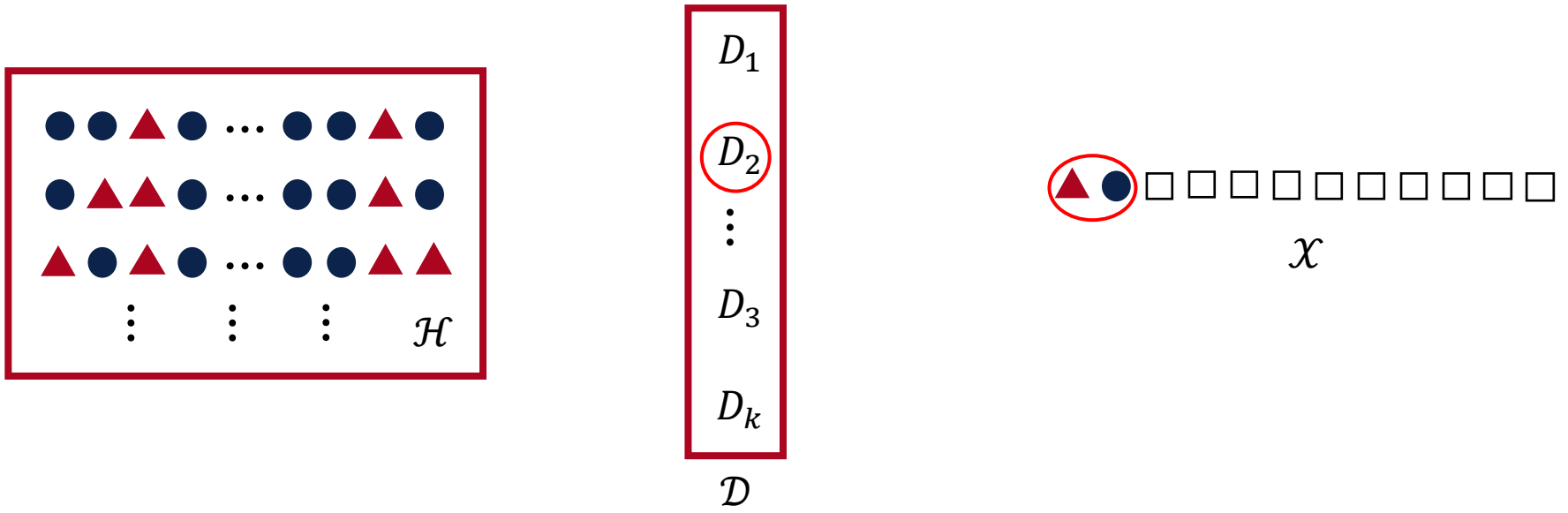
- Given a hypothesis class $\mathcal{H}$ of VC dimension d, a set $\mathcal{D}$ of k distributions $\{D_1, \dots, D_k\}$ over a dataset $\mathcal{X} \times \mathcal{Y}$, an error tolerance parameter ε and a failure parameter δ. Let $L(h) = \max_{i \in [k]} L(h, D_i)$ be the loss and assume $\min_{h \in \mathcal{H}} L(h) = \nu$.



- A learner can pick a distribution D_i and $x \in \mathcal{X}$ and query the label $y \sim D_i(\cdot \mid x)$. The goal is to return a $\hat{h}$ such that $L(\hat{h}) \leq \nu + \varepsilon$ with probability at least $1 - \delta$ with as few queries as possible.

Problem Definition

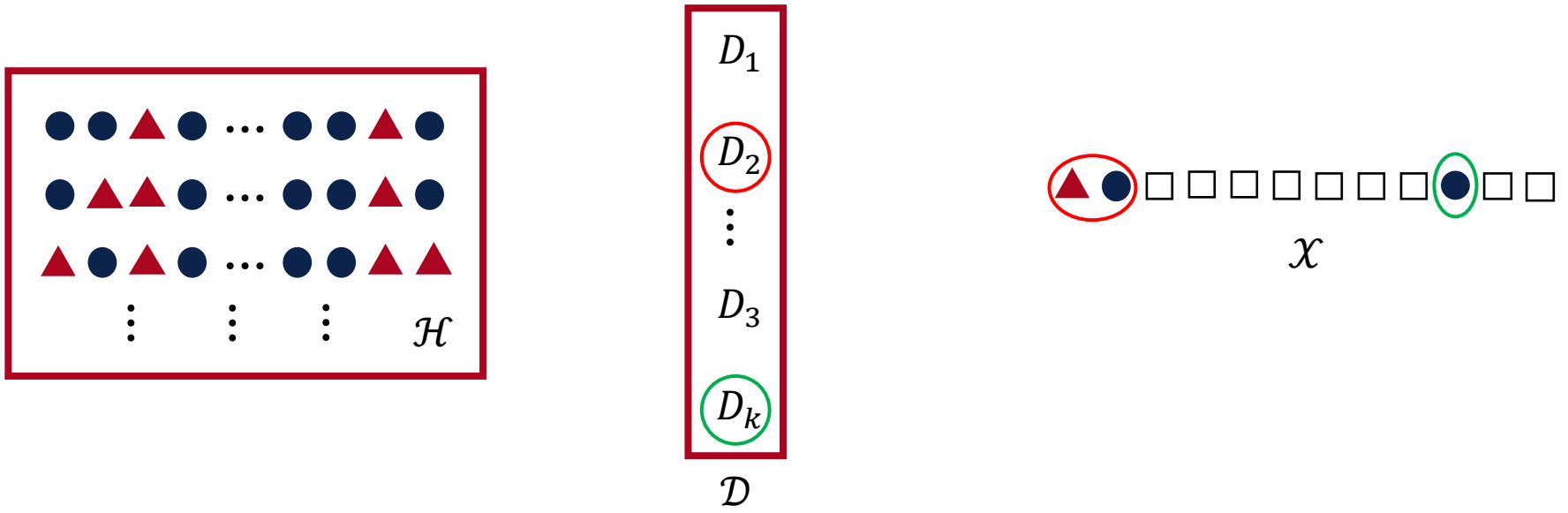
- Given a hypothesis class $\mathcal{H}$ of VC dimension d, a set $\mathcal{D}$ of k distributions $\{D_1, \dots, D_k\}$ over a dataset $\mathcal{X} \times \mathcal{Y}$, an error tolerance parameter ε and a failure parameter δ. Let $L(h) = \max_{i \in [k]} L(h, D_i)$ be the loss and assume $\min_{h \in \mathcal{H}} L(h) = \nu$.



- A learner can pick a distribution D_i and $x \in \mathcal{X}$ and query the label $y \sim D_i(\cdot \mid x)$. The goal is to return a $\hat{h}$ such that $L(\hat{h}) \leq \nu + \varepsilon$ with probability at least $1 - \delta$ with as few queries as possible.

Problem Definition

- Given a hypothesis class $\mathcal{H}$ of VC dimension d, a set $\mathcal{D}$ of k distributions $\{D_1, \dots, D_k\}$ over a dataset $\mathcal{X} \times \mathcal{Y}$, an error tolerance parameter ε and a failure parameter δ. Let $L(h) = \max_{i \in [k]} L(h, D_i)$ be the loss and assume $\min_{h \in \mathcal{H}} L(h) = \nu$.



- A learner can pick a distribution D_i and $x \in \mathcal{X}$ and query the label $y \sim D_i(\cdot \mid x)$. The goal is to return a $\hat{h}$ such that $L(\hat{h}) \leq \nu + \varepsilon$ with probability at least $1 - \delta$ with as few queries as possible.

Main results: realizable setting ($\nu = 0$)

Reference	Model	Label complexity
Uniform sampling + error minimization (naive)	Passive	$\tilde{O}(kd/\varepsilon)$
Blum, Haghtalab, Procaccia and Qiao (2017); Chen, Zhang and Zhou (2018); Nguyen and Zakynthinou (2018)	Passive (sampling from D_i's on demand)	$\tilde{O}((k+d)/\varepsilon)$
Rittler and Chaudhuri (2023)	Active	$\tilde{O}\left(k\, d\, \theta_{\max} \ln\frac{1}{\varepsilon}\right)$

$\theta_{\max}$: maximum disagreement coefficient among the k distributions (Hanneke, 2014)

Can we enjoy the benefit of $O(k+d)$ for active multidistribution learning?

This work	Active	$\tilde{O}\left((k+d)\, \theta_{\max} \ln\frac{1}{\varepsilon}\right)$
This work	Active	$\Omega\big((k+d)\, \theta_{\max}\big)$

Main results: non-realizable setting ($\nu > 0$)

Reference	Model	Label complexity
Haghtalab, Jordan and Zhao (2022); Zhang, Zhan, Chen, Du and Lee (2024); Peng (2024).	Passive (sampling from D_i's on demand)	$\tilde{O}\left(\frac{k+d}{\varepsilon^2}\right)$
Rittler and Chaudhuri (2023)	Active	$\tilde{O}\left(k\, d\, \theta_{\max}^2 \left(1+\frac{\nu^2}{\varepsilon^2}\right)+\frac{k}{\varepsilon^2}\right)$

Can we enjoy the benefit of $O(k+d)$ for active multidistribution learning?

This work	Active	$\tilde{O}\left((k+d)\theta_{\max}\left(1+\frac{\nu^2}{\varepsilon^2}\right)+\frac{k\nu}{\varepsilon^2}\right)$

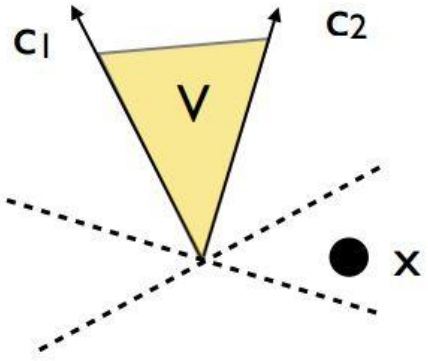
Is the $O(k/\varepsilon^2)$ term fundamental?

does not appear in vanilla active learning ($k = 1$)

This work	Active (proper learning)	$\Omega\left(\frac{k\nu}{\varepsilon^2}\right),\ k \geq 2$

Technical highlights: upper bounds

- Algorithm: reduce to passive multidistribution learning via a disagreement-based approach (Hanneke, 2014)



- Nonrealizable setting ($\nu > 0$):
  - Active learning helps when ν is small (Kääriäinen, 2007)
  - Along the way, we obtain refined sample complexity bounds of passive multidistribution learning $\tilde{O}\left(\frac{(k+d)(\nu+\epsilon)}{\varepsilon^2}\right)$
  - Smoothly interpolates between realizable (Blum, Haghtalab, Procaccia and Qiao, 2017) and nonrealizable (Zhang, Zhan, Chen, Du and Lee, 2024) regimes

Technical highlights: lower bound

- Lower bound $\Omega\left(\frac{kv}{\varepsilon^2}\right)$:
- Unique challenge for active multidistribution learning ($k \geq 2$): needs to balance the errors under different distributions

Distinguishing these requires $\Omega\left(\frac{v}{\varepsilon^2}\right)$ labels

Error	D_1	D_2
h_1	0	$v - 2\epsilon$
h_2	v	$\frac{v}{2} - 2\epsilon$

$L(h_1) = v - 2\epsilon$

$L(h_2) = v$

→ Should choose h_1

Error	D_1	D_2
h_1	0	$v + 2\epsilon$
h_2	v	$\frac{v}{2} + 2\epsilon$

$L(h_1) = v + 2\epsilon$

$L(h_2) = v$

→ Should choose h_2

- Confirms the intuition in Rittler and Chaudhuri (2023)

Conclusion and open problems

- We design new algorithms for active multi-distribution learning with $O(k+d)$ label complexities
- An $\Omega\left(\frac{k\nu}{\varepsilon^2}\right)$ label complexity is necessary for proper learners, highlighting a challenge unique to active multidistribution learning

- Additional results:
  - Distribution-independent analysis (similar to Hanneke and Yang (2015))
  - $\Omega\left(\frac{k\nu}{\varepsilon^2}\right)$ label complexity lower bound for any learner

- Open questions:
  - How to design algorithms without knowing ν?
  - Computationally efficient algorithms?

Thank you!