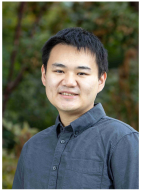
Optimal multi-distribution learning

Yuxin Chen

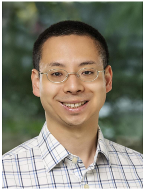
Statistics & Data Science, Wharton, UPenn

Zihan Zhang
Princeton

Wenhao Zhan
Princeton



Simon Du
UWashington



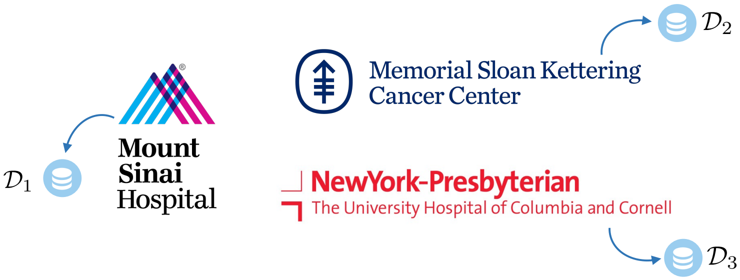
Jason Lee
Princeton

"Optimal multi-distribution learning," Z. Zhang, W. Zhan, Y. Chen, S. Du, J. Lee, arXiv:2312.05134, 2023

In multi-distribution learning, an agent aims to learn a *shared model* to fit multiple (unknown) data distributions

- diverse data sources (e.g., localities, communities, populations)
- heterogeneous objectives $\longrightarrow$ need a balance

- k unknown data distributions $\mathcal{D}_1, \ldots, \mathcal{D}_k$ (e.g., localities, communities, populations)
- hypothesis class $\mathcal{H}$: VC dimension d
- known loss function ℓ (e.g., misclassification error)

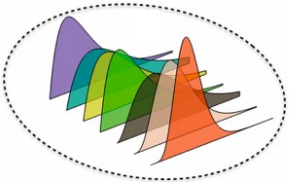


- k unknown data distributions $\mathcal{D}_1, \ldots, \mathcal{D}_k$ (e.g., localities, communities, populations)
- hypothesis class $\mathcal{H}$: VC dimension d
- known loss function ℓ (e.g., misclassification error)

goal: learn an ε-optimal $\underbrace{\text{hypothesis } \widehat{h}}_{\text{possibly random}}$ (in **min-max** sense)

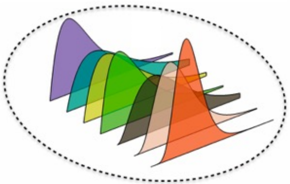
$$\max_{1\le i\le k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i,\widehat{h}} \left[\ell\big(\widehat{h},(x,y)\big)\right] \le \min_{h\in\mathcal{H}} \max_{1\le i\le k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i} \left[\ell\big(h,(x,y)\big)\right] + \varepsilon$$

Mohri et al. '19, Sagawa et al. '19, Blum et al. '17, Buhlmann et al. '15, Guo '23 . . .



distributionally robust learning

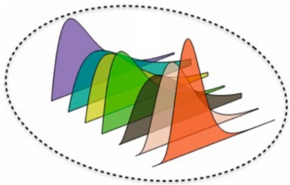
Mohri et al. '19, Sagawa et al. '19, Blum et al. '17, Buhlmann et al. '15, Guo '23 . . .



distributionally robust learning

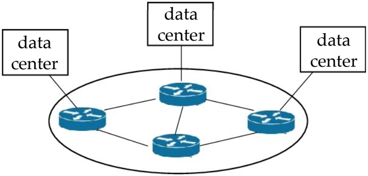
min-max fairness

Mohri et al. '19, Sagawa et al. '19, Blum et al. '17, Buhlmann et al. '15, Guo '23 . . .



distributionally robust learning

min-max fairness



collaborative learning

Adaptive vs. non-adaptive sampling

- **non-adaptive sampling**: pre-determine sample-size budgets for each distribution beforehand

  $\longrightarrow$ loss of data efficiency

Adaptive vs. non-adaptive sampling

- **non-adaptive sampling**: pre-determine sample-size budgets for each distribution beforehand

  $\longrightarrow$ loss of data efficiency

- **adaptive sampling**: sample on demand during learning process

  $\longrightarrow$ this talk

Adaptive vs. non-adaptive sampling

- **non-adaptive sampling**: pre-determine sample-size budgets for each distribution beforehand

  $\longrightarrow$ loss of data efficiency

- **adaptive sampling**: sample on demand during learning process

  $\longrightarrow$ this talk

learning 1 distribution

$$\Downarrow$$

$$\frac{d}{\varepsilon^2}$$

Adaptive vs. non-adaptive sampling

- **non-adaptive sampling**: pre-determine sample-size budgets for each distribution beforehand

  $\longrightarrow$ loss of data efficiency

- **adaptive sampling**: sample on demand during learning process

  $\longrightarrow$ this talk

learning 1 distribution

$\Downarrow$

$$\frac{d}{\varepsilon^2}$$

learning k distributions

non-adaptive

$\Downarrow$

$$\frac{dk}{\varepsilon^2}$$

Adaptive vs. non-adaptive sampling

- **non-adaptive sampling**: pre-determine sample-size budgets for each distribution beforehand

  $\longrightarrow$ loss of data efficiency

- **adaptive sampling**: sample on demand during learning process

  $\longrightarrow$ this talk

learning 1 distribution

$\Downarrow$

$\frac{d}{\varepsilon^2}$

learning k distributions

non-adaptive

$\Downarrow$

$\frac{dk}{\varepsilon^2}$

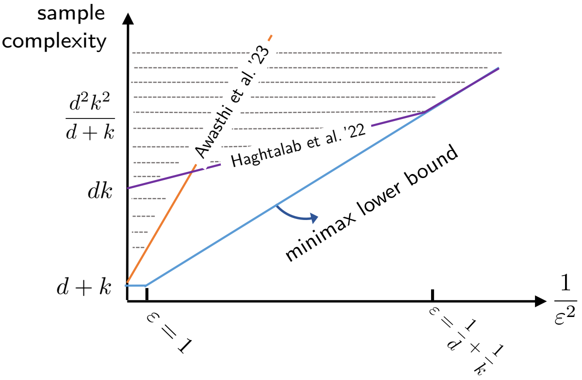
adaptive

$\Downarrow$

Prior works: VC classes

paper	sample complexity
Haghtalab et al. '22	$\frac{d+k}{\varepsilon^2} + \frac{dk}{\varepsilon}$
Awasthi et al. '23	$\frac{d}{\varepsilon^4} + \frac{k}{\varepsilon^2}$
(lower bound) Haghtalab et al. '22	$\frac{d+k}{\varepsilon^2}$

Prior works: VC classes



paper	sample complexity
Haghtalab et al. '22	$\frac{d+k}{\varepsilon^2} + \frac{dk}{\varepsilon}$
Awasthi et al. '23	$\frac{d}{\varepsilon^4} + \frac{k}{\varepsilon^2}$
(lower bound) Haghtalab et al. '22	$\frac{d+k}{\varepsilon^2}$

Can we close the gap between achievability and lower bound?

Proceedings of Machine Learning Research vol 195:1–11, 2023 36th Annual Conference on Learning Theory

Open Problem: The Sample Complexity of Multi-Distribution Learning for VC Classes

Pranjal Awasthi PRANJALAWASTHI@GOOGLE.COM
Google Research, Mountain View, CA, USA

Nika Haghtalab NIKA@BERKELEY.EDU
University of California, Berkeley, CA, USA

Eric Zhao ERIC.ZH@BERKELEY.EDU
University of California, Berkeley, CA, USA

Main results

Theorem 1 (Zhang, Zhan, Chen, Du, Lee '23)

We can design an algorithm that returns randomized hypothesis $\widehat{h}$ s.t.

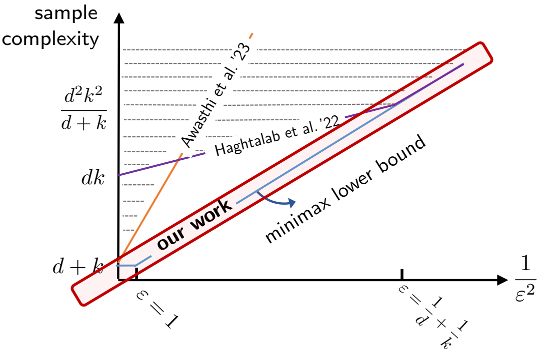
$$\max_{1\le i\le k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i,\widehat{h}} \left[\ell\big(\widehat{h},(x,y)\big)\right] \le \min_{h\in\mathcal{H}} \max_{1\le i\le k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i} \left[\ell\big(h,(x,y)\big)\right] + \varepsilon,$$

with sample complexity

$$\widetilde{O}\left(\frac{d+k}{\varepsilon^2}\right)$$

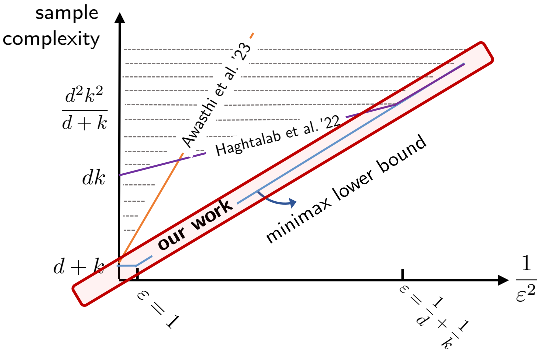
- matches the minimax lower bound (up to log factors)

Main results



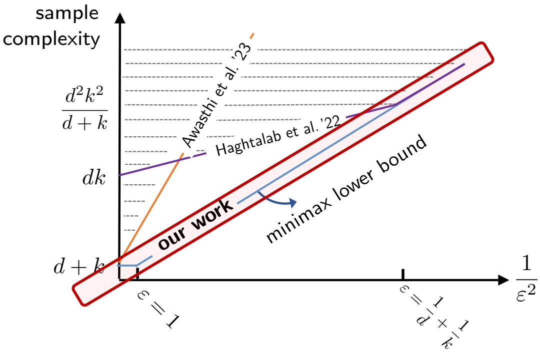
- matches the minimax lower bound (up to log factors)
- solves a COLT open problem (concurrent work: Peng '23)

Main results



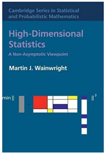
- matches the minimax lower bound (up to log factors)
- solves a COLT open problem (concurrent work: Peng '23)
- can be extended to Rademacher classes

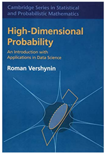
Main results



- matches the minimax lower bound (up to log factors)
- solves a COLT open problem (concurrent work: Peng '23)
- can be extended to Rademacher classes
- algorithm is $\underbrace{\text{oracle-efficient}}_{\text{only needs to call ERM oracle}}$ (solves another COLT open problem)

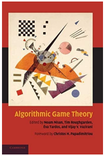
Algorithm design

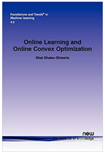




high-dimensional
statistics

game theory





online learning

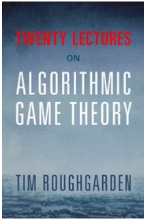
A game-theoretic view

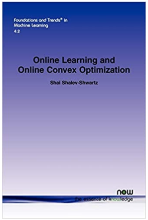
$$\color{blue}{\min_{h \in \mathcal{H}}} \color{red}{\max_{1 \le i \le k}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i} \left[\ell(h,(x,y))\right]$$

max-player:
finding least favorable distribution

min-player:
finding most favorable hypothesis

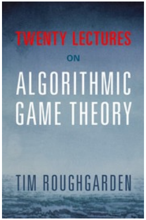
Preliminaries: learning in games

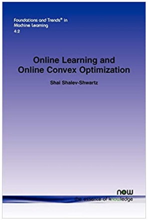




- **no-regret algorithm**: online algorithm w/ $\underbrace{\text{sub-linear regret}}_{\frac{1}{T}\text{Regret}(T)\to 0}$ over any adversary
  - e.g., Hedge algorithm (equivalent to online mirror descent)

Preliminaries: learning in games





- **no-regret algorithm**: online algorithm w/ $\underbrace{\text{sub-linear regret}}_{\frac{1}{T}\mathsf{Regret}(T) \to 0}$ over any adversary
  - e.g., Hedge algorithm (equivalent to online mirror descent)

- **best-response**: play argmin or argmax (not always no-regret)

$$\min_{h\in\mathcal{H}} \max_{1\le i\le k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\left[\ell(h,(x,y))\right]$$

min-player

max-player

- min-player/max-player: no-regret/no-regret (Haghtalab et al. '22)

$$\frac{d+k}{\varepsilon^2} + \frac{dk}{\varepsilon} \qquad \text{(burn-in due to covering of } \mathcal{H}\text{)}$$

- min-player/max-player: best-response/no-regret (Awasthi et al. '23)

$$\frac{d}{\varepsilon^4} + \frac{k}{\varepsilon^2} \qquad \text{(lack of sample reuse)}$$

Our approach: best-response/no-regret

At iteration t:

- min-player computes **empirical best response**

$$h^t \approx \arg\min_{h \in \mathcal{H}} L(h, w^t)$$

  - $L(h, w) := \sum_{i=1}^{k} w_i \underset{(x,y)\sim\mathcal{D}_i}{\mathbb{E}} \left[\ell\big(h, (x, y)\big)\right]$ (loss w.r.t. weighted dist)

Our approach: best-response/no-regret

At iteration t:

- min-player computes **empirical best response**

$$h^t \approx \arg\min_{h \in \mathcal{H}} L(h, w^t)$$

  - $L(h, w) := \sum_{i=1}^{k} w_i \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}_i} \left[\ell\big(h, (x, y)\big) \right]$ (loss w.r.t. weighted dist)

- max-player runs **Hedge** to update $\underbrace{\text{mixed distribution } w^t \in \Delta_k}_{\text{weighted distribution } \sum_i w_i^t \mathcal{D}_i}$

Our approach: best-response/no-regret

At iteration t:

- min-player computes **empirical best response**

$$h^t \approx \arg\min_{h \in \mathcal{H}} L(h, w^t)$$

  - $L(h, w) \coloneqq \sum_{i=1}^{k} w_i \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i} \big[\ell\big(h, (x, y)\big)\big]$ (loss w.r.t. weighted dist)

- max-player runs **Hedge** to update $\underbrace{\text{mixed distribution } w^t \in \Delta_k}_{\text{weighted distribution } \sum_i w_i^t \mathcal{D}_i}$

$$w_i^t \;\propto\; w_i^{t-1} \exp(\eta \widehat{r}_i^t) \qquad \text{with } \widehat{r}_i^t: \text{ empirical risk for } \mathcal{D}_i$$

Our approach: best-response/no-regret

At iteration t:

- min-player computes **empirical best response**

$$h^t \approx \arg\min_{h \in \mathcal{H}} L(h, w^t)$$

  - $L(h, w) \coloneqq \sum_{i=1}^{k} w_i \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i} \left[\ell\big(h, (x, y)\big)\right]$ (loss w.r.t. weighted dist)

- max-player runs **Hedge** to update $\underbrace{\text{mixed distribution } w^t \in \Delta_k}_{\text{weighted distribution } \sum_i w_i^t \mathcal{D}_i}$

$$w_i^t \;\propto\; w_i^{t-1} \exp(\eta \widehat{r}_i^t) \qquad \text{with } \widehat{r}_i^t : \text{ empirical risk for } \mathcal{D}_i$$

Output: randomized hypothesis $\widehat{h} \sim \text{Uniform}\big(\{h^t\}_{1\le t\le T}\big)$

Key algorithmic distinction from prior work

$$\underbrace{\text{adaptive sampling}}_{\#\text{samples from } \mathcal{D}_i \text{ based on } \{w_i^t\}} \quad + \text{ sample reuse}$$

Key algorithmic distinction from prior work

$$\underbrace{\text{adaptive sampling}}_{\#\text{samples from } \mathcal{D}_i \text{ based on } \{w_i^t\}} \quad + \text{ sample reuse}$$

Sampling strategy at iteration t:

- **best-response:** $\underbrace{\text{have } \frac{d+k}{\varepsilon^2} w_i^t \text{ samples available}}_{\text{reuse samples}}$ from $\mathcal{D}_i$

Key algorithmic distinction from prior work

$$\underbrace{\text{adaptive sampling}}_{\#\text{samples from } \mathcal{D}_i \text{ based on } \{w_i^t\}} \quad + \text{ sample reuse}$$

Sampling strategy at iteration t:

- **best-response:** $\underbrace{\text{have } \frac{d+k}{\varepsilon^2} \max\limits_{1 \le \tau \le t} w_i^\tau \text{ samples available}}_{\text{reuse samples}}$ from $\mathcal{D}_i$

Key algorithmic distinction from prior work

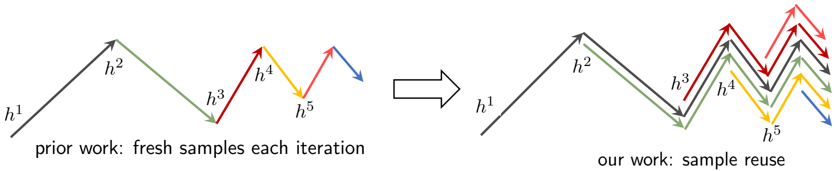
$$\underbrace{\text{adaptive sampling}}_{\#\text{samples from } \mathcal{D}_i \text{ based on } \{w_i^t\}} + \text{ sample reuse}$$

Sampling strategy at iteration t:

- **best-response:** $\underbrace{\text{have } \frac{d+k}{\varepsilon^2} \max_{1 \le \tau \le t} w_i^\tau \text{ samples available from } \mathcal{D}_i}_{\text{reuse samples}}$
- **no-regret:** $\underbrace{\text{draw } k \max_{1 \le \tau \le t} w_i^\tau \text{ samples from } \mathcal{D}_i}_{\text{fresh samples}}$

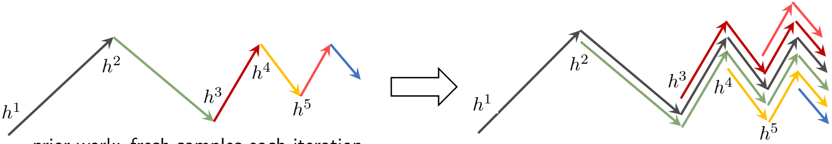
Key technical challenges

1. complicated statistical dependency due to sample reuse



Key technical challenges

1. complicated statistical dependency due to sample reuse



prior work: fresh samples each iteration

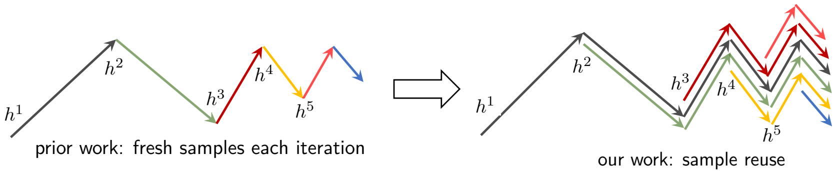
our work: sample reuse

2. need to bound the algorithm trajectory in a fine-grained manner

$$\text{sample complexity} \asymp \frac{d+k}{\varepsilon^2}\underbrace{\sum_{i=1}^{k}\max_{1\le t\le T} w_i^t}_{}$$

Key technical challenges

1. complicated statistical dependency due to sample reuse



2. need to bound the algorithm trajectory in a fine-grained manner

$$\text{sample complexity} \asymp \frac{d+k}{\varepsilon^2} \underbrace{\sum_{i=1}^{k} \max_{1 \le t \le T} w_i^t}_{\widetilde{O}(1)}$$

concentration + doubling trick + combinatorics

Concurrent work: Peng et al. '23

Peng et al. '23 established a sample complexity of

$$\frac{d+k}{\varepsilon^2}\left(\frac{k}{\varepsilon}\right)^{o(1)}$$

which also solved the COLT open problem

Concurrent work: Peng et al. '23

Peng et al. '23 established a sample complexity of

$$\frac{d+k}{\varepsilon^2}\left(\frac{k}{\varepsilon}\right)^{o(1)}$$

which also solved the COLT open problem

- optimal up to some sub-polynomial term
- a very different algorithm

  — recursive structure to eliminate non-optimal hypotheses

Necessesity of randomization

Our alg. returns randomized hypothesis ...

Necessesity of randomization

Our alg. returns randomized hypothesis ...

Question: is it possible to find an ε-optimal <u>deterministic</u> hypothesis w/ the same sample complexity **(another COLT open problem)**?

Necessesity of randomization

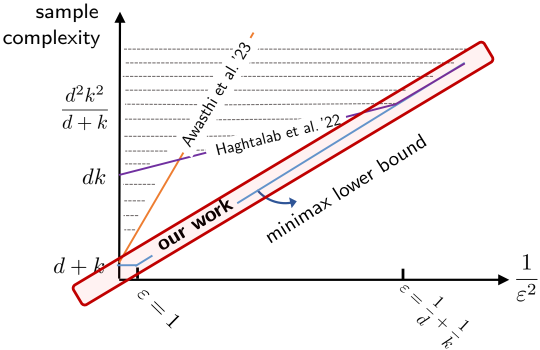
Our alg. returns randomized hypothesis ...

Question: is it possible to find an ε-optimal deterministic hypothesis w/ the same sample complexity **(another COLT open problem)**?

Answer: No!

- finding an ε-optimal deterministic policy needs $\Omega(\frac{dk}{\varepsilon^2})$ samples

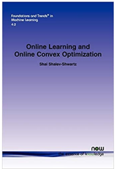
Summary: multi-distribution learning

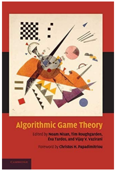


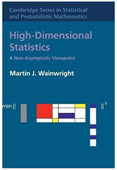
- settles the sample complexity of MDL under on-demand sampling
- solves 3 COLT open problems posed by Awasthi et al. '23

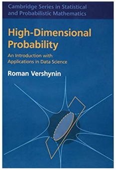
Concluding remarks

Advancing frontier of statistical learning requires integrated thinking of modern statistics, optimization & game theory









online learning & games

(high-dimensional) statistics

"Optimal multi-distribution learning," Z. Zhang, W. Zhan, Y. Chen, S. Du, J. Lee, arXiv:2312.05134, 2023