Algorithms for multi-group learning

Daniel Hsu

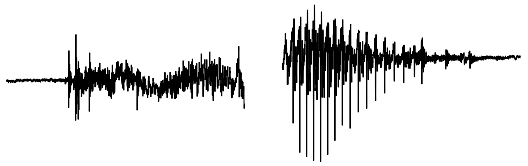
Columbia University

Joint works with: Navid Ardeshir, Sivaraman Balakrishnan, Noah Bergam, Samuel Deng, Jingwen Liu, Christopher Tosh, Lujing Zhang

Yale Statistics and Data Science Seminar

March 30, 2026

Statistical learning



/k/ /a/ …

Training data (labeled examples)
$(x_1, y_1), \ldots, (x_n, y_n)$ from $\mathcal{X} \times \mathcal{Y}$

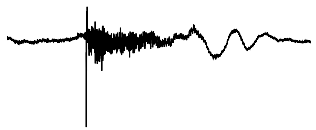
(IID from P)

$w \leftarrow w - \eta \nabla \hat{\mathcal{R}}(w)$

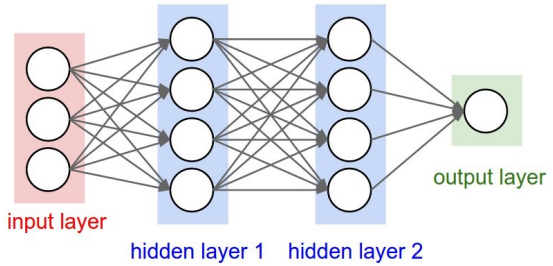
Learning algorithm

$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim P}[\ell(f(x), y)]$

Test point
$x \in \mathcal{X}$



Prediction function
$\hat{f}: \mathcal{X} \to \mathcal{Y}$



Predicted label
$\hat{f}(x) \in \mathcal{Y}$

/t/

Why is statistical learning not enough?

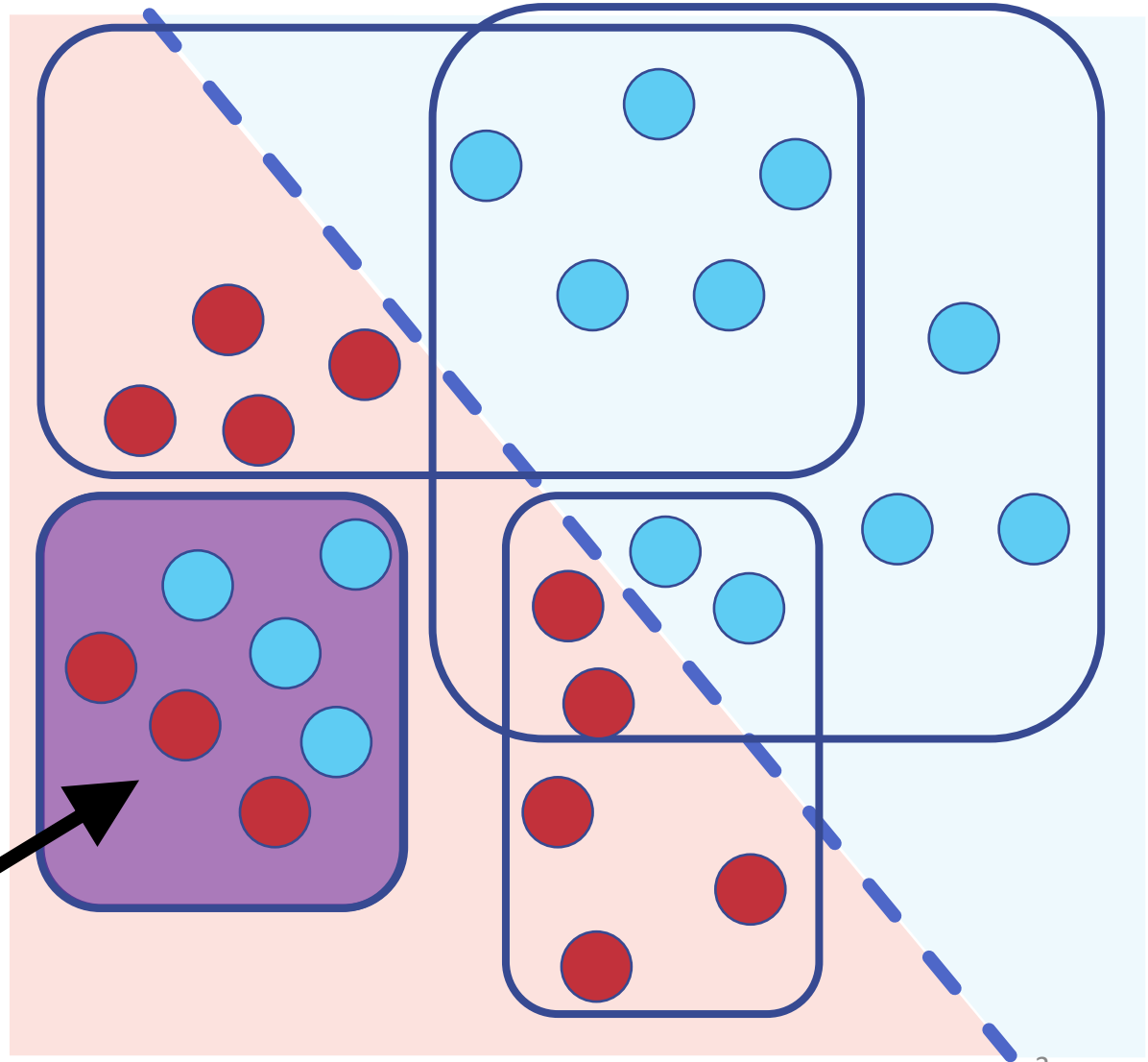
- Aggregate performance over a population P

$$\mathbb{E}_{(x,y)\sim P}[\ell(f(x), y)]$$

- No assurance about any particular instance

$$\ell(f(x), y)$$

- No assurances even for subpopulations/subgroups



Trustworthy AI/ML

Many high-profile failures of ML of 2010s [e.g., Buolamwini & Gebru, 2018] were concerned with individuals & subgroups

- Standard ML objectives fail to address prereqs. for trustworthy AI/ML

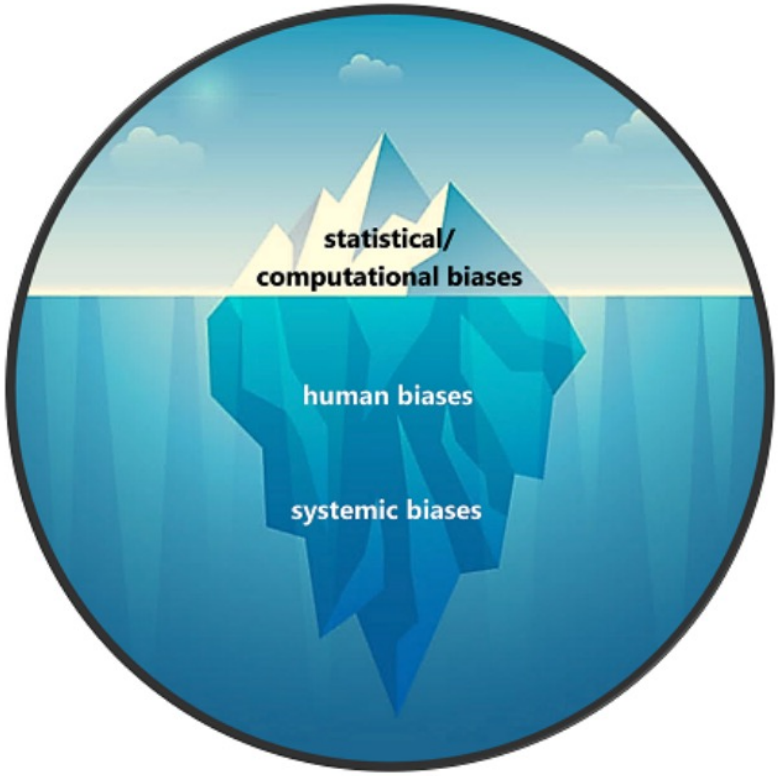


Image credit: https://www.nist.gov/artificial-intelligence/ai-research-identifying-managing-harmful-bias-ai

"Hidden stratification" of training data

[Oakden-Rayner, Dummon, Carneiro, and Ré, 2020]

- Training data is often a data set of convenience, typically stratified
- Downstream applications may require good accuracy just on **specific strata**
- Would like to do as well as if you knew which were the relevant strata

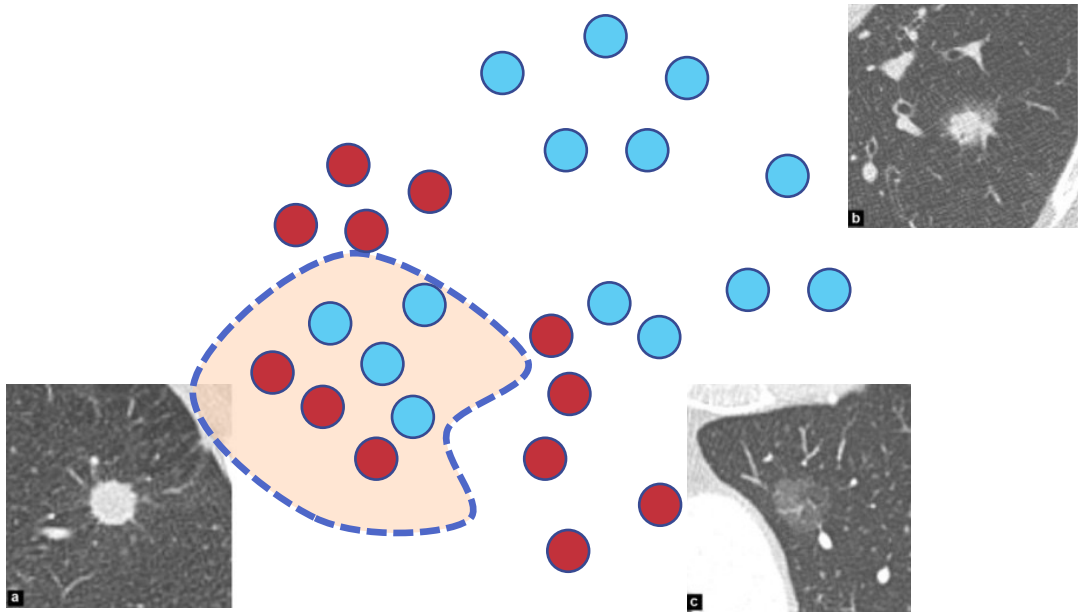
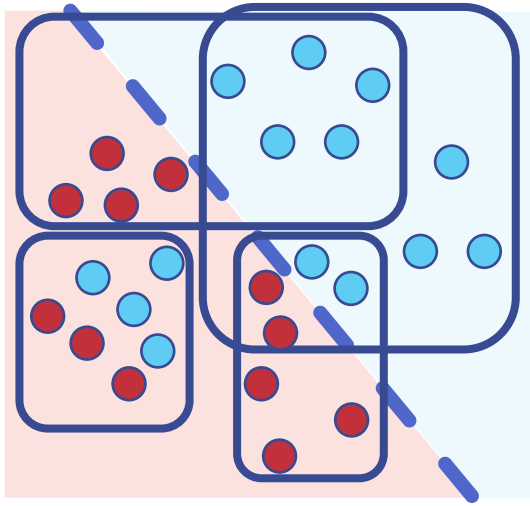


Image credit: Lauren Oakden-Rayner's blog

High-level summary

- **Multi-group learning**: natural generalization of the "classical" setup for supervised learning from statistical learning theory
- Basic statistical convergence rates from "classical" setup can be extended to multi-group setup …
- But requires new algorithms
  - "Classical" learning setup: ERM suffices
  - Multi-group learning setup: have to do more



1. Background and setup

Multi-group learning: brief history

- Formalized by Rothblum & Yona (2021); Blum & Lykouris (2020)
  - Main motivation: fairness in ML
  - This talk: we focus on **binary classification + error rate objective**
    - Many results in this talk extends to other **bounded loss functions**
  - [RY'21] and [BL'20] also consider other objectives (e.g., calibration)
- Related (but non-equivalent) concepts:
  - "Multi-accuracy" [Kim, Ghorbani, Zou, 2019]
  - "Omniprediction" [Gopalan, Kalai, Reingold, Sharan, Wieder, 2022]
- Generalizations:
  - "Multi-calibration" [Hébert-Johnson, Kim, Reingold, Rothblum, 2018]
  - "Multi-objective learning" [Haghtalab, Jordan, Zhao, 2023]
  - "Panprediction" [Balakrishnan, Haghtalab, H., Lee, Zhao, 2026]

Cast of characters

- $(X,Y)\sim P$ for data distribution P over $\mathcal{X}\times\{0,1\}$
- $\mathcal{G}$ is (known) family of subsets of $\mathcal{X}$ ("groups")
  - E.g., demographic groups, metric balls, half-spaces
- $\mathcal{H}$ is (known) benchmark class of functions $\mathcal{X}\to\{0,1\}$ ("hypothesis class")
  - E.g., linear classifiers, fixed-size neural nets
- Error rate of f:

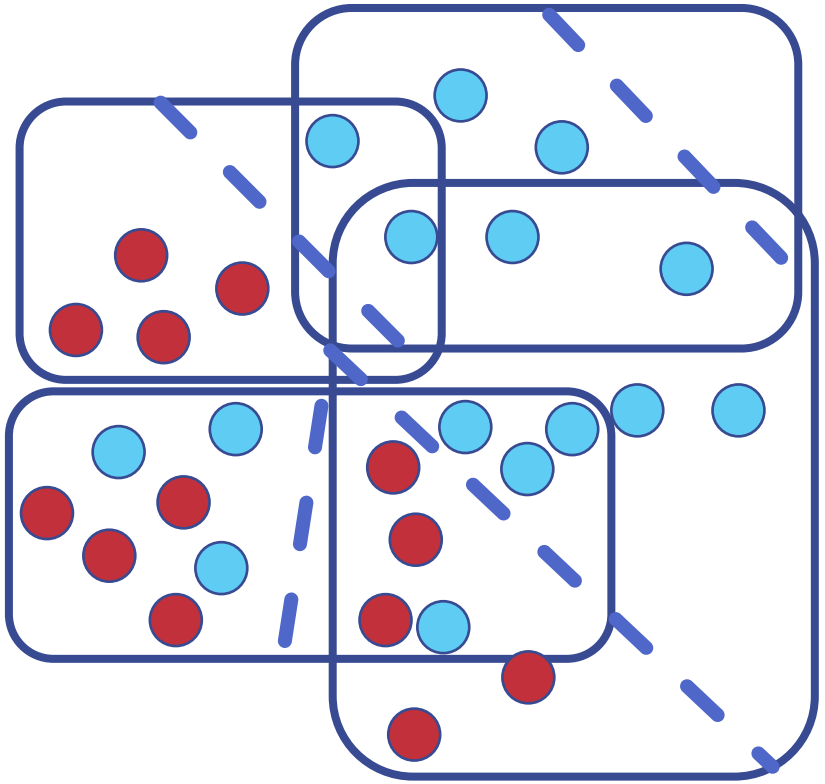
$$\mathrm{err}(f) \coloneqq P(f(X)\neq Y)$$

  - Excess error rate: $\mathrm{err}(f) - \min_{h\in\mathcal{H}} \mathrm{err}(h)$
- Conditional error rate of f on group g:

$$\mathrm{err}(f|g) \coloneqq P(f(X)\neq Y|X\in g)$$

  - Excess conditional error rate on g:

$$\mathrm{err}(f|g) - \min_{h\in\mathcal{H}} \mathrm{err}(h|g)$$



Standard statistical learning

- Given n IID copies of (X, Y), find $\hat{f}: \mathcal{X} \to \{0,1\}$ such that (w.h.p.)

$$\mathrm{err}(\hat{f}) - \min_{h \in \mathcal{H}} \mathrm{err}(h) \leq \epsilon$$

$$\Theta\left(\sqrt{\frac{\mathrm{vc}(\mathcal{H})}{n}}\right)$$

- Suffices to let $\hat{f}$ = empirical risk minimizer (ERM) over $\mathcal{H}$

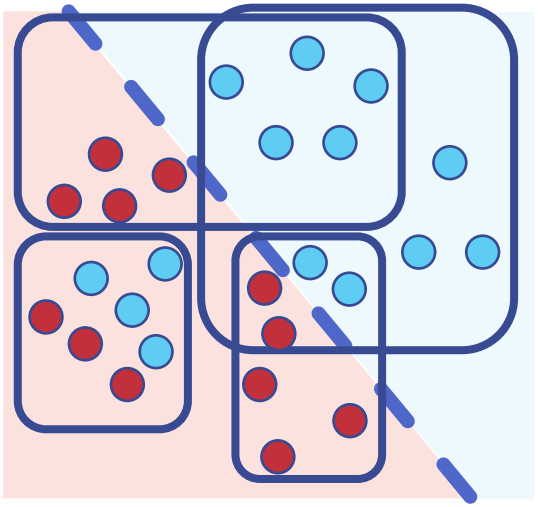
Multi-group statistical learning

- Given n IID copies of (X, Y), find $\hat{f}: \mathcal{X} \to \{0,1\}$ such that (w.h.p.) for each $g \in \mathcal{G}$,

$$\mathrm{err}(\hat{f}|g) - \min_{h \in \mathcal{H}} \mathrm{err}(h|g) \leq \epsilon$$

Compare to best $h \in \mathcal{H}$ for group g

- It's possible that no $h \in \mathcal{H}$ can satisfy this requirement

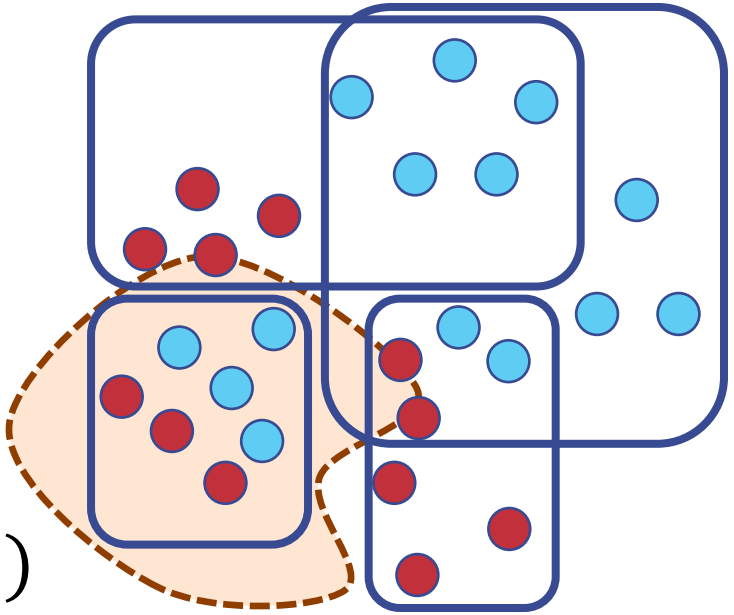


Application: "hidden stratification" of training data

- **Multi-group learning ⇒ guarantees under "hidden stratification"**
  Suppose $\hat{f}$ satisfies multi-group learning guarantee.
  For every $S \subset \mathcal{X}$ that is ϵ-multiplicatively-approx.* by some $g \in \mathcal{G}$,

$$\mathrm{err}\left(\hat{f} \mid S\right) - \min_{h \in \mathcal{H}} \mathrm{err}(h \mid S) \leq O(\epsilon)$$

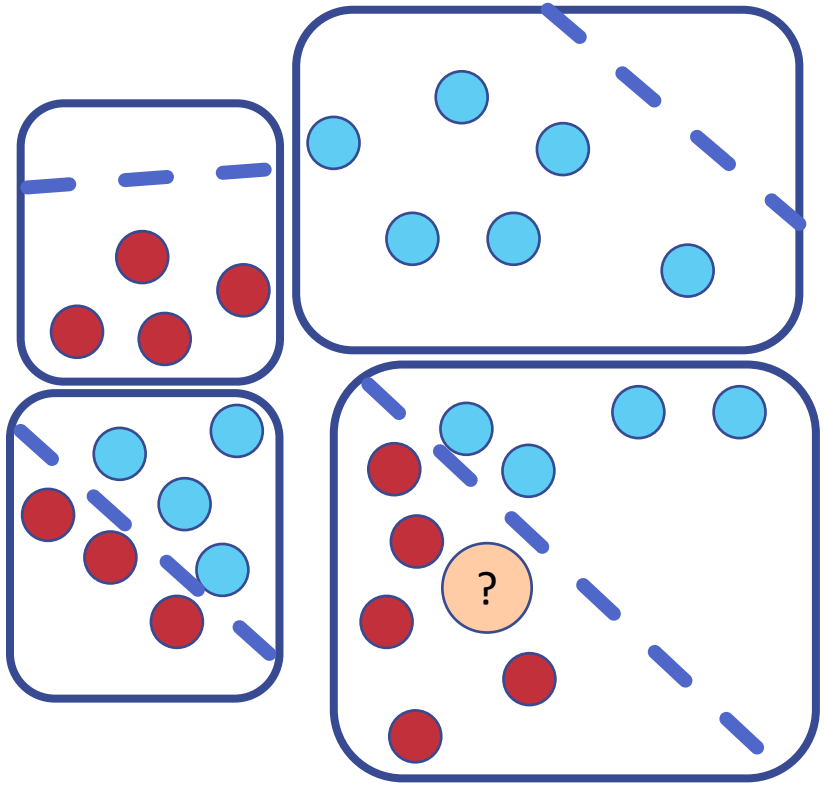
So we'd like $\mathcal{G}$ as "rich" as possible,
and only "pay" $\log|\mathcal{G}|$ or $\mathrm{vc}(\mathcal{G})$



$$*P(g \Delta S) \leq \epsilon \cdot P(g)$$

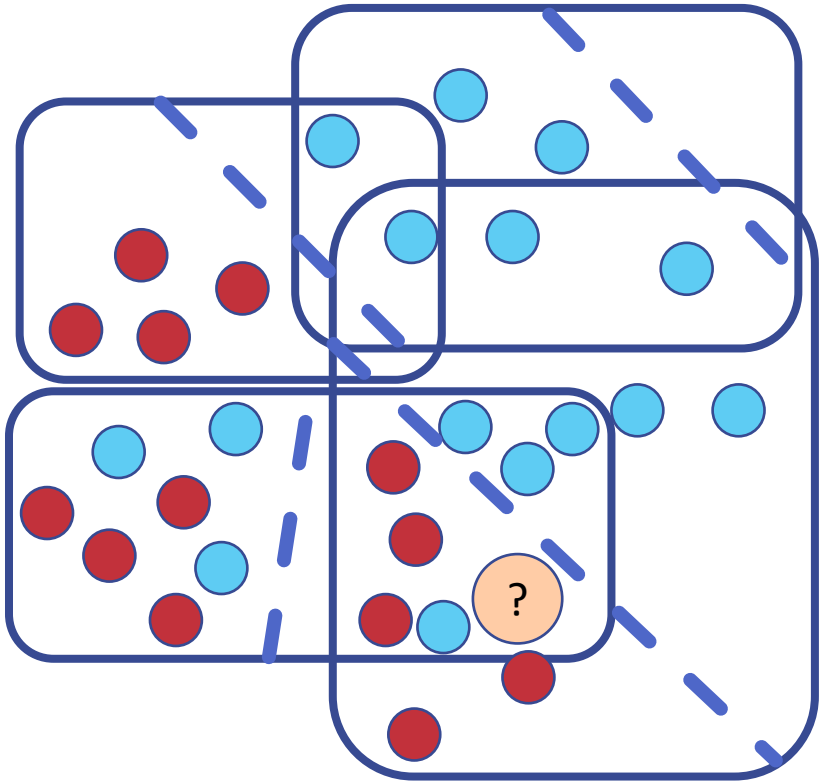
Challenges for multi-group agnostic learning

Easy case: disjoint groups



Fit a predictor to each group

Harder case: overlapping



How do we resolve disagreements among predictors?

Our results

For simplicity, assume each group $g \in \mathcal{G}$ has $P(g) = \Omega(1)$*

Simple/practical algorithms:

- Excess error rate for group $g \in \mathcal{G}$: [Tosh & H., 2022; Zhang, H., Balakrishnan, 2026+]

$$\tilde{O}\left(\left(\frac{\mathrm{vc}(\mathcal{G}) + \mathrm{vc}(\mathcal{H})}{n}\right)^{1/3}\right)$$

- Can improve rate to $\tilde{O}\left(n^{-2/5}\right)$

Near-optimal algorithms:

- Excess error rate for group $g \in \mathcal{G}$: [Ardeshir et al, 2026; Bergam, Deng, H., 2026+]

$$O\left(\left(\frac{\mathrm{vc}(\mathcal{G}) \log n + \mathrm{vc}(\mathcal{H})}{n}\right)^{1/2}\right)$$

- Under "group-realizability", can improve rate to $\tilde{O}(n^{-1})$

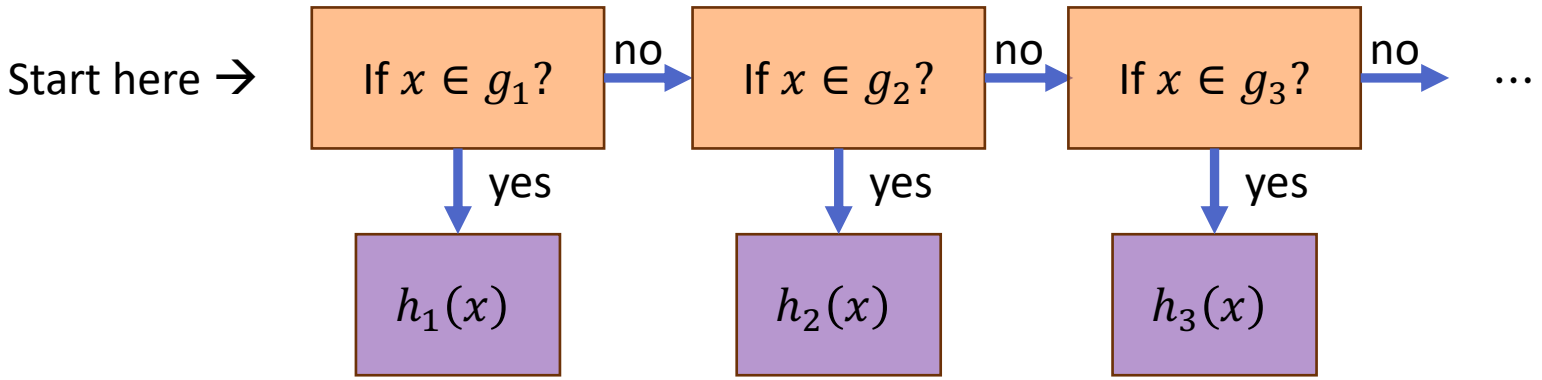
*Can typically replace n with $n \cdot P(g)$

2. Simple/practical algorithms

A simple/practical multi-group learning algorithm

- "PREPEND" algorithm [Tosh & H., 2022]
  - Learns a **decision list**

"if $x \in g_1$ then return $h_1(x)$ else if $x \in g_2$ then return $h_2(x)$ else if …"



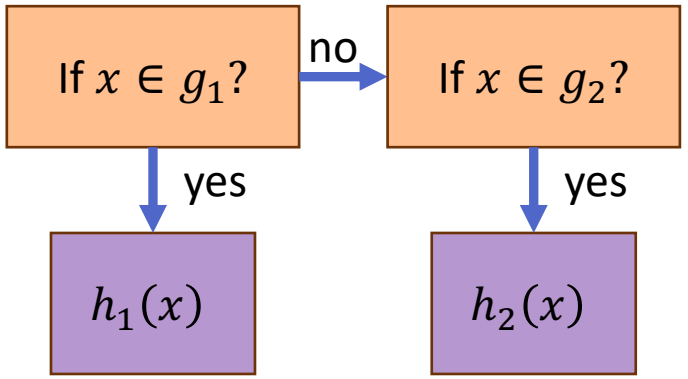
  - Algorithm independently proposed by Globus-Harris, Kearns, Roth (2022) in different but related context

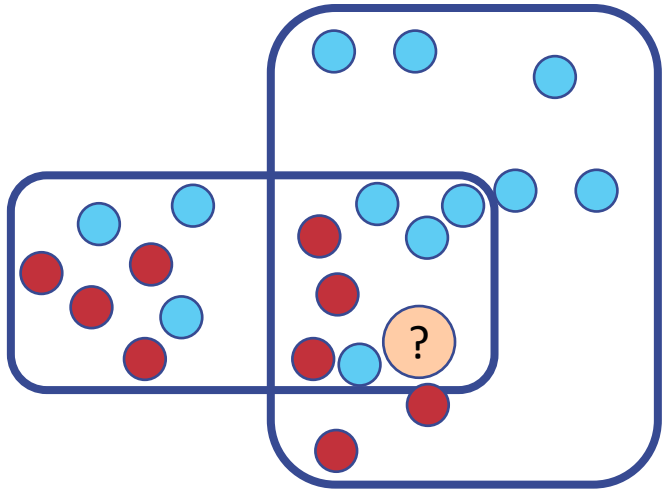
What does not work

- Cannot determine list solely from (estimates of) "low-order" statistics

$$P(g), \qquad \mathrm{err}(h \mid g)$$

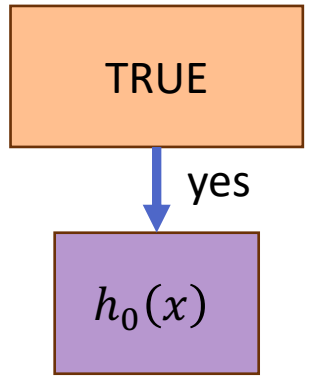
  - Suppose $g \cap g' \neq \emptyset$
    - What should be done for $x \in g \cap g'$?
    - It may depend on $P(g \cap g')$





PREPEND algorithm

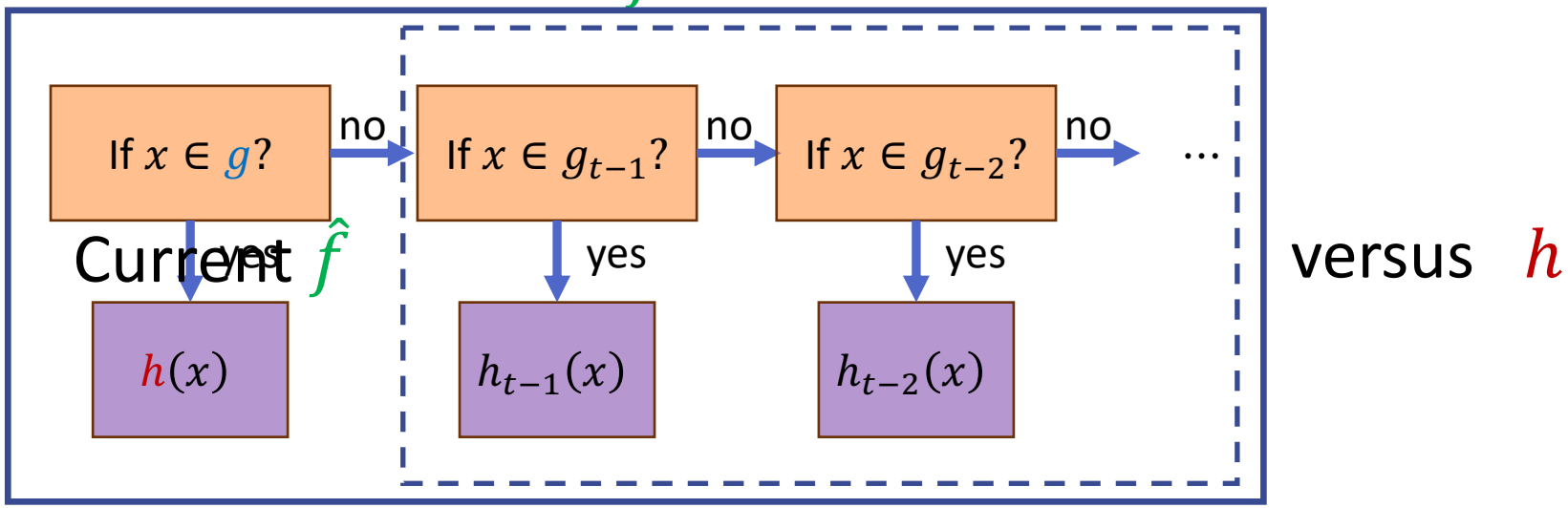
- Pick some (default) $h_0 \in \mathcal{H}$, initialize $\hat{f}$ to be the decision list "if TRUE then return $h_0(x)$"



- While there exists $(g, h) \in \mathcal{G} \times \mathcal{H}$ such that

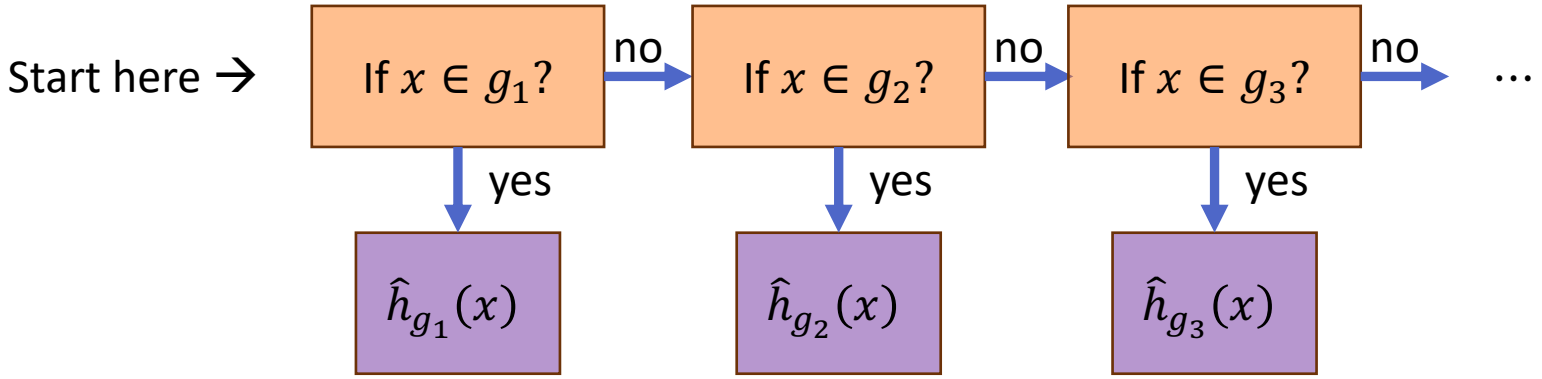
$$\widehat{\mathrm{err}}(\hat{f} \mid g) > \widehat{\mathrm{err}}(h \mid g) + \epsilon$$

  - Prepend "if $x \in g$ then return $h(x)$ else" to current decision list $\hat{f}$



Some comments

1. If $\mathcal{G}$ has small cardinality: can replace $\mathcal{H}$ with $\widehat{\mathcal{H}} := \{\hat{h}_g : g \in \mathcal{G}\}$, where $\hat{h}_g := \min_{h \in \mathcal{H}} \widehat{\mathrm{err}}(h|g)$ (ERM for group g)
   - Then decision list is just an ordering of (a subset of) $\mathcal{G}$

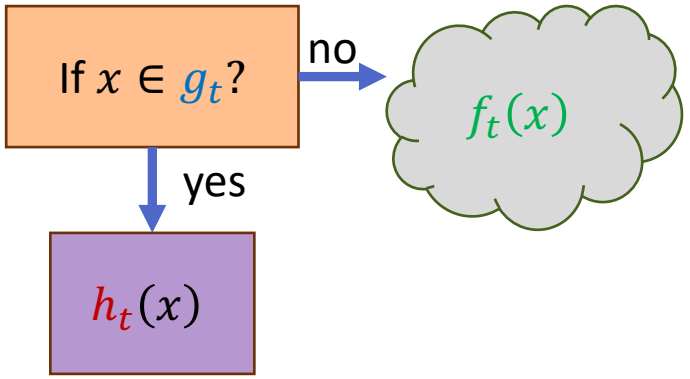


2. PREPEND can choose same $g \in \mathcal{G}$ in multiple loop iterations
   - These iterations simply move chosen g to front of the ordering
   - Length of list never more than $|\mathcal{G}|$

Analysis of PREPEND

- In iteration t, update current f_t to new f_{t+1} by prepending

"if $x \in g_t$ then return $h_t(x)$ else"

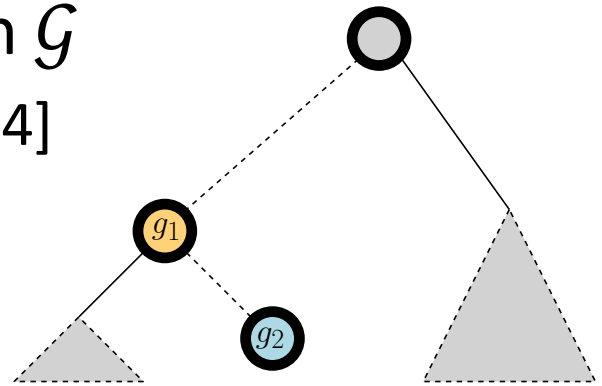


$$\begin{aligned}\mathrm{err}(f_{t+1}) &= P(g_t)\mathrm{err}(\,h_t \mid g_t\,) + P(g_t^{\mathrm{c}})\mathrm{err}(\,f_t \mid g_t^{\mathrm{c}}\,)\\ &\le P(g_t)(\mathrm{err}(\,f_t \mid g_t\,) - \epsilon/2) + P(g_t^{\mathrm{c}})\mathrm{err}(\,f_t \mid g_t^{\mathrm{c}}\,)\\ &= \mathrm{err}(f_t) - P(g_t)\epsilon/2\end{aligned}$$

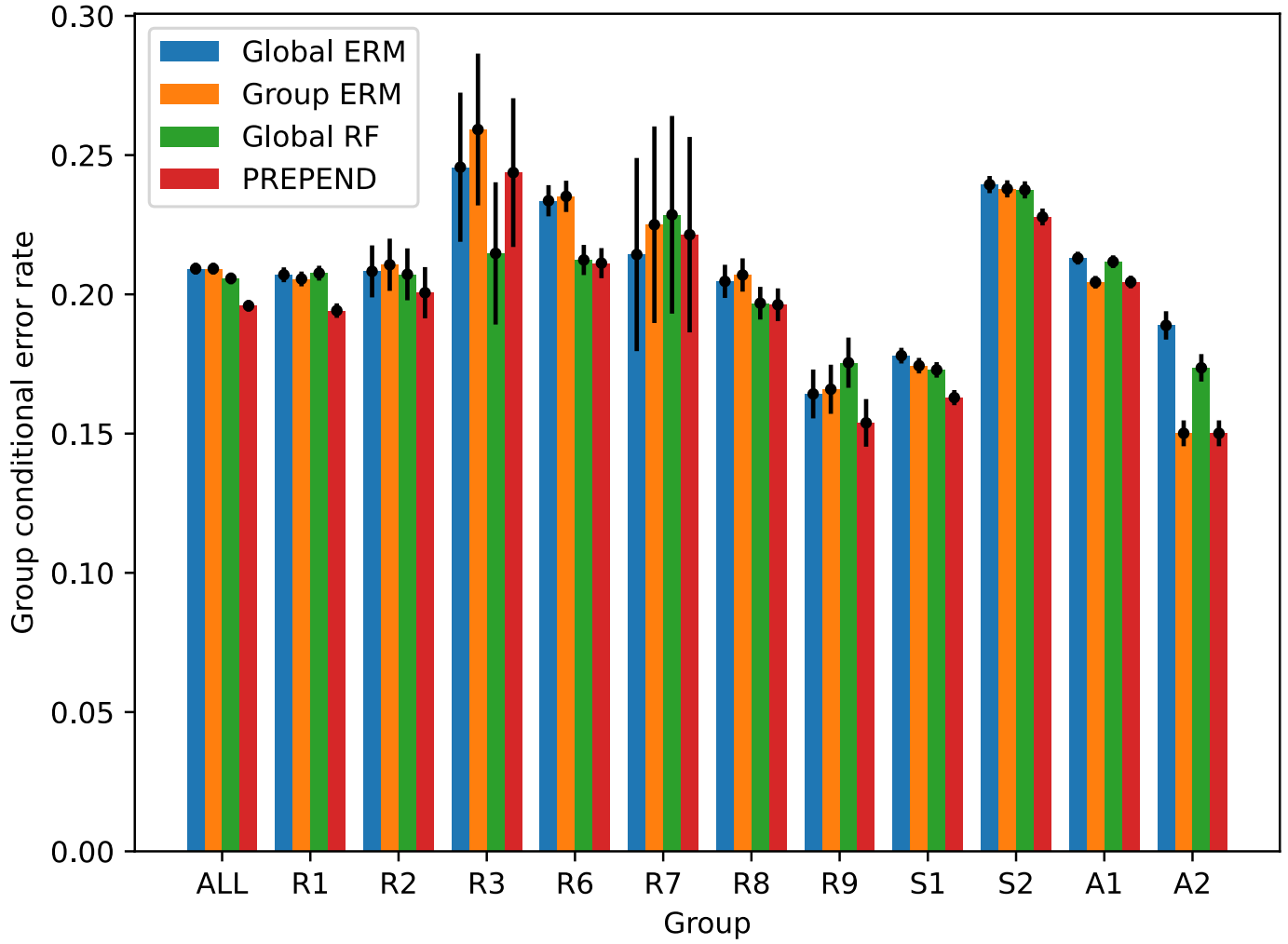
- Must stop within $O(1/\epsilon)$ iterations

Efficiency of decision list learning?

- **Excess error rate bound**: $\tilde{O}\left(\left(\frac{\mathrm{VC}(\mathcal{G})+\mathrm{VC}(\mathcal{H})}{n}\right)^{1/3}\right)$ for each group g
- Also: Can get $O\left(n^{-1/2}\right)$ if $n \gtrsim |\mathcal{G}|$ (VC dim. of $|\mathcal{G}|$-length decision lists)
- Q: Can we learn a decision list with better rate without paying $|\mathcal{G}|$?
- Some progress: [Zhang, H., Balakrishnan, 2026+]
  - Randomized update threshold improves rate from $\tilde{O}\left(n^{-1/3}\right)$ to $\tilde{O}\left(n^{-2/5}\right)$
  - Uses ideas from *Adaptive Data Analysis* (Dwork et al, 2015; Hardt, 2017)
- Can also improve rate by assuming structure on $\mathcal{G}$
  - E.g., laminar groups → $\tilde{O}\left(n^{-1/2}\right)$ [Deng & H., 2024]



Employment prediction in California



2016 American Community Survey

Groups:

ALL	overall population
R{1,2,3,6,7,8,9}	group by race
S{1,2}	group by sex
A{1,2}	group by age

Global ERM: `logreg` on all data
Group ERM: `logreg` on group
Global RF: random forest on all data

Data is from "Folkstable" package
[Ding, Hardt, Miller, Schmidt, 2021]

Recap: PREPEND

- Learns a **decision list** structure over benchmark hypotheses
  - Both learning algorithm and final predictor are simple and interpretable
- Excess error rate bound: $\tilde{O}(n^{-2/5})$
  - Prior work of [RY'21]: $\tilde{O}(n^{-1/8})$
- Only gets $O(n^{-1/2})$ excess error rate for $n \gtrsim |\mathcal{G}|$…

3. Getting near-optimal excess error rate

Min-max stochastic optimization

- Multi-group learning as min-max stochastic optimization
  [e.g., Nemirovski, Juditsky, Lan, Shapiro, 2009]

$$\min_{f} \max_{g \in \mathcal{G}} \{\mathrm{err}(f|g) - \min_{h \in \mathcal{H}} \mathrm{err}(h|g)\}$$

- Our (more modest) goal: Find f with value $\leq \epsilon$
  - Satisfied by Bayes predictor and optimal decision list ($\epsilon \leq 0$)
  - But statistical cost to estimate either can be too high

- Equivalent: stochastic constraint satisfaction

$$\text{Find } f \text{ s.t. } \mathrm{err}(f|g) - \mathrm{err}(h|g) \leq \epsilon \quad \text{for all } (g,h) \in \mathcal{G} \times \mathcal{H}$$

  - Seems easier to work in the dual space

Algorithm for stochastic constraint satisfaction

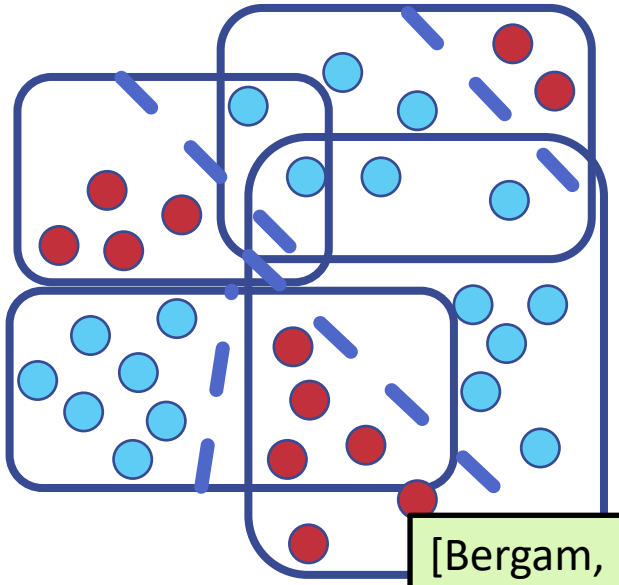
- Algorithm motivated by online learning [Blum & Lykouris, 2020; Tosh & H., 2022]
  - Simulate "online" version of multi-group learning problem (repeated game):
    - 1. Adversary: choose weighting $\vec{w} = \left(w_{g,h}\right)_{(g,h)\in\mathcal{G}\times\mathcal{H}}$ over constraints
    - 2. Learner: choose predictor f (based solely on $\vec{w}, \mathcal{G}, \mathcal{H}$) that ensures

$$\sum_{g,h} w_{g,h}\{\mathrm{err}(f|g) - \mathrm{err}(h|g)\} \leq \epsilon$$

    - (Repeat 1 and 2 above several times)
  - Finally aggregate all predictors chosen by learner ("online-to-batch")
- **Excess error rate bound**: $O\left(\left(\frac{\log|\mathcal{G}|+\log|\mathcal{H}|)}{n}\right)^{1/2}\right)$ for each group g
  - Only "Adversary" uses data, to decide how to choose next $\vec{w}$
  - Can replace $\log|\mathcal{H}|$ with $\tilde{O}\left(\mathrm{vc}(\mathcal{H})\right)$

Special case: group-realizable setting [Ardeshir et al, 2026]

- "Group-realizability" assumption:
  $\min_{h \in \mathcal{H}} \mathrm{err}(h|g) = 0$ for all $g \in \mathcal{G}$
  - Weaker than "realizability"
    (unless $\mathcal{G}$ contains entire domain)



- Let $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ be the class of functions compatible with group-realizability
  - It's possible for $\mathrm{vc}(\mathcal{C}_{\mathcal{G},\mathcal{H}}) = \infty$ even when $\mathrm{vc}(\mathcal{G}) \vee \mathrm{vc}(\mathcal{H}) < \infty$

Theorem: ERM over $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ gives (w.h.p.) predictor $\hat{f}$

$$\mathrm{err}(\hat{f}|g) \leq \tilde{O}\left(\frac{\mathrm{vc}(\mathcal{G}) + \mathrm{vc}(\mathcal{H})}{n}\right) \quad \text{for each group } g$$

[Bergam, Deng, H., 2026+]: Multi-group "One Inclusion Graph" gives predictor $\hat{f}$ with

$$\mathbb{E}[\mathrm{err}(\hat{f}|g)] \leq \frac{\mathrm{vc}(\mathcal{H}|_g)}{n \cdot P(g)}$$

for each group g

(Algorithm extends to non-group-realizable setting with $n^{-1/2}$ rate)

Summary

- **Multi-group learning**: extension of statistical learning that is addresses many practical concerns in trustworthy AI/ML
- Tools from statistical learning theory are useful here, but **need to remix the algorithmic ideas**
- Open: Simpler near-optimal algorithms? Polynomial-time algorithms? Combinatorial characterization of $(\mathcal{G}, \mathcal{H})$-multi-group-learnability?

Thank you!

Support: ONR under grant N00014-24-1-2700; by the NSF under grants CCF-1740833, IIS-1563785, and IIS-2040971; and by a JP Morgan Faculty Award