Efficiently Certifiable Guarantees for Learning with Distribution Shift

Konstantinos Stavropoulos (UT Austin) based on joint works with: *Gautam Chandrasekaran, Surbhi Goel, Adam Klivans, Vasilis Kontonis, Lin Lin Lee, Abhishek Shetty, Arsen Vasilyan*

Learning in the presence of distribution shift

Hospital A: *Labeled data* (treatment outcomes)

Hospital B: *Only unlabeled data* (patient medical records)

Goal: Decide treatments for B using A

Critical Setting: Guarantees Needed

Attempt #1: Estimate Statistical Distance between Training and Test distributions

Information-theoretic impossibility
e.g. uniformity testing in high dimensions requires exponential number of samples

Attempt #2: Estimate the Discrepancy (Domain Adaptation)
[BBCP'06] [BCKPW'08] [MMR'09]

Discrepancy distance between distributions $\mathscr{D}, \mathscr{D}'$, *with respect to* class of boolean functions $\mathscr{C}$:

$$\mathrm{disc}_{\mathscr{C}}(\mathscr{D}, \mathscr{D}') = \max_{f_1, f_2 \in \mathscr{C}} \left| \Pr_{\mathscr{D}}[f_1(x) \neq f_2(x)] - \Pr_{\mathscr{D}'}[f_1(x) \neq f_2(x)] \right|$$

Out of distribution generalization for classification:

$$\Pr_{\mathscr{D}'}[h(x) \neq f^*(x)] \leq \Pr_{\mathscr{D}}[h(x) \neq f^*(x)] + \mathrm{disc}_{\mathscr{C}}(\mathscr{D}, \mathscr{D}')$$

✓ Small discrepancy between training and test distributions $\Rightarrow$ low test error
(assuming labels always generated by unknown $f^* \in \mathscr{C}$)

✓ Sample complexity of estimating the discrepancy is bounded by the VC-dimension of $\mathscr{C}$

✗ Testing if the discrepancy is small is NP-hard, even when $\mathscr{C}$ is the class of linear classifiers [CKKSV'24] [BGS'18]

Our Approach: Testable Learning with Distribution Shift

Goal: Certify *at least one* of the following

1. The test distribution is different from the training (detect distribution shift)
2. The error of the output classifier h on the test distribution is low

Input: $(x \sim \mathscr{D}, f^*(x))$, $(x \sim \mathscr{D}')$ → $\mathscr{A}$ → **Output**: Either (***REJECT***, $\perp$) or (***ACCEPT***, h)

- **Soundness:** Upon acceptance $\Pr_{x \sim \mathscr{D}'}[h(x) \neq f^*(x)] \leq \epsilon$ (w.h.p.)
- **Completeness:** If $\mathscr{D} = \mathscr{D}'$, then accept (w.h.p.)

✓ Certify low test error with **no assumptions** on test marginal

✓ Results from classical learning theory for **most concept classes** can be enhanced with **efficient** testers for TDS learning

Concept Class	Training Marginal	TDS Runtime	PAC Runtime
Halfspaces	Standard d-dim Gaussian	$d^{\tilde{\Theta}(\log 1/\epsilon)}$	$\mathrm{poly}(d, 1/\epsilon)$
Intersections of k Halfspaces	Standard d-dim Gaussian	$d^{\tilde{O}(\log 1/\epsilon)}(k/\epsilon)^{O(k^3)}$	$\mathrm{poly}(d)(k/\epsilon)^{O(k^2)}$
Degree-2 PTFs	Standard Gaussian or Uniform	$d^{\tilde{O}(1/\epsilon^9)}$	$\mathrm{poly}(d, 1/\epsilon)$
Circuits of size s, depth t	Uniform on d-dim hypercube	$d^{O(\log(s))^{O(t)}\log(1/\epsilon)}$	$d^{O(\log(s))^{t-1}\log(1/\epsilon)}$

[KSV'24a,b] [CKKSV'24] ++

Broader Context: Testable Learning

[Rubinfeld & Vasilyan '23]

Goal: Get rid of strong assumptions

Instead of assumptions: Provide certifiable guarantees when

1. Direct validation of guarantees
2. Direct verification of assumptions

Are intractable or impossible

Testable Agnostic Learning [RV'23]: No distribution shift, but unknown error benchmark ("optimum error"). *Assumption to be removed:* marginal is well-behaved

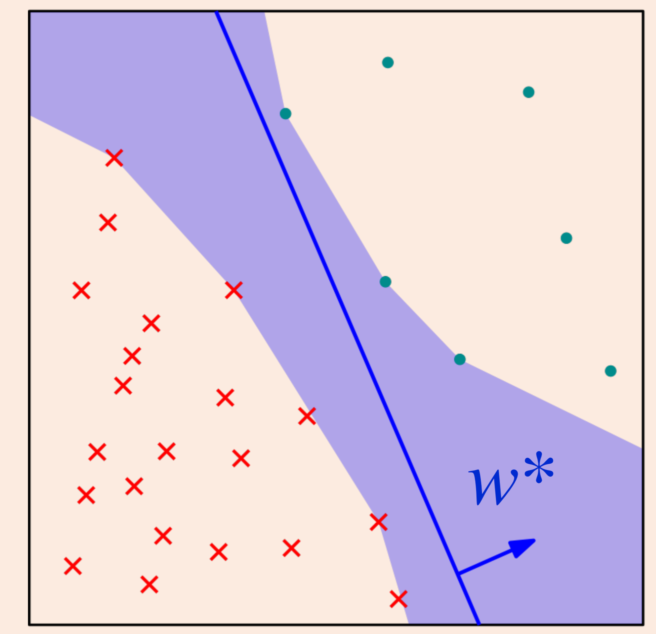
Testable Noise Assumptions [GKSV'25]: Again, no distribution shift. *Assumption to be removed:* The label noise is structured

Case Study: Halfspaces

Suppose: $f^*(x) = \mathrm{sign}(w^* \cdot x - \tau^*)\,,\, w^* \in \mathbb{S}^{d-1}, \tau^* \in \mathbb{R}$

Simpler result: There is an ϵ-TDS learner with runtime $\mathrm{poly}(d) \cdot 2^{O(1/\epsilon)}$

Case I: Low bias $(|\tau^*| \leq 1/\epsilon^{1/2})$



Parameter Recovery

Lemma [DKS'18]: $\mathbb{E}_{x \sim \mathscr{N}}[f^*(x)\, x] = \frac{\exp(-\tau^{*2}/2)}{\sqrt{2\pi}} w^*$

Find:
$\hat{w} : \|\hat{w} - w^*\|_2 \leq O(\epsilon/d)$
$\hat{\tau} : \|\hat{\tau} - \tau^*\|_2 \leq O(\epsilon/d)$
using $\mathrm{poly}(d/\epsilon) \cdot 2^{O(\tau^{*2})}$ samples

Concern: $f^*, \hat{f}$ disagree often under $\mathscr{D}'$

Check the following condition:

$$\mathbb{P}_{x \sim \mathscr{D}'}\left[\exists w', \tau' : \begin{matrix} \|\hat{w} - w'\| \leq O(\epsilon/d) \\ \|\hat{\tau} - \tau'\| \leq O(\epsilon/d) \\ \hat{f}(x) \neq f'(x) \end{matrix} \right] \leq O(\epsilon)$$

No information about w^*

Concern: $w^* \cdot x$ often large under $\mathscr{D}'$

Concentration Certificates

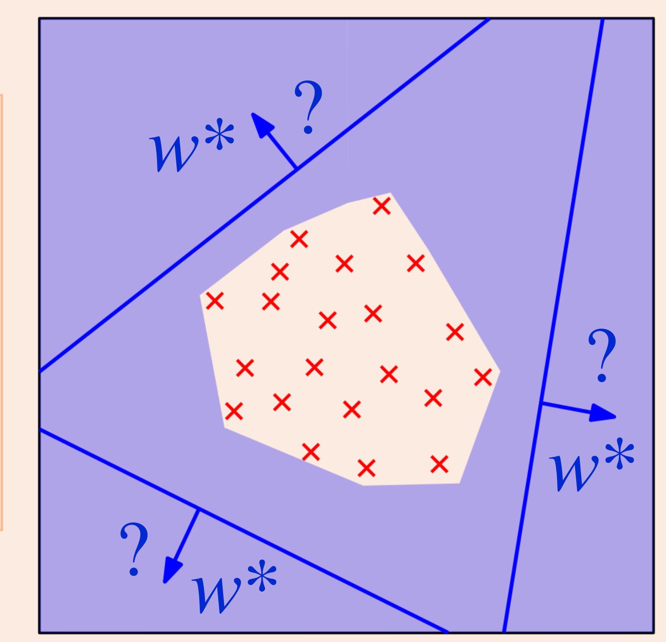
$$\mathbb{P}_{x \sim \mathscr{D}'}[w^* \cdot x > \tau^*] \leq \mathbb{P}_{x \sim \mathscr{D}'}[w^* \cdot x > \epsilon^{-1/2}] \leq \epsilon\, \mathbb{E}_{x \sim \mathscr{D}'}[(w^* \cdot x)^2]$$

Check: $\mathbb{E}_{\mathscr{D}'}[x_i x_j] \approx \mathbb{E}_{\mathscr{N}}[x_i x_j]\,, \forall i, j$

So: $\mathbb{E}_{x \sim \mathscr{D}'}[(w^* \cdot x)^2] \approx \mathbb{E}_{x \sim \mathscr{N}}[(w^* \cdot x)^2] = 1$

For tight results: Check degree $\log(1/\epsilon)$

Case II: High bias $(\tau^* > 1/\epsilon^{1/2})$



Universal TDS Learners
Handle Benign Shifts

Accept whenever $\mathscr{D}'$ such that:

1. $\mathbb{E}_{x \sim \mathscr{D}'}[(v \cdot x)^4] \leq C\,, \forall v \in \mathbb{S}^{d-1}$

2. $\mathbb{P}_{x \sim \mathscr{D}'}[\,|v \cdot x| \leq r] \leq Cr\,, \forall v \in \mathbb{S}^{d-1}$

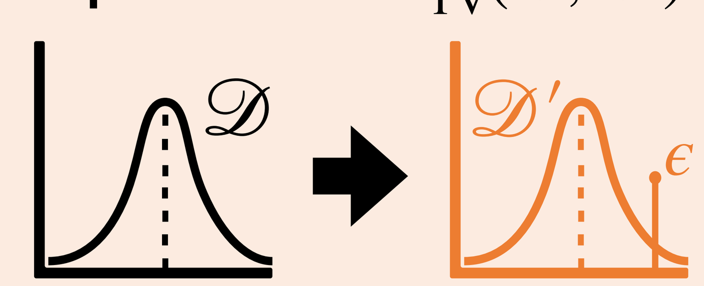
Instead of moment matching:

Check: $\sup_{v \in \mathbb{S}^{d-1}} \mathbb{E}_{\mathscr{D}'}[(v \cdot x)^2] \leq C$ using eigenvalue decomposition

For improved runtime: Certify subgaussianity via SoS [DHPT'24]

Tolerant TDS Learners [GSSV'24]
Handle Moderate Amounts of Shift

Accept whenever $\mathrm{d}_{\mathrm{TV}}(\mathscr{D}, \mathscr{D}') \leq \epsilon$



While: $\sup_{v \in \mathbb{S}^{d-1}} \mathbb{E}_{\mathscr{D}'}[(v \cdot x)^2] > 10$

Find r s.t. $\mathbb{P}_{x \sim \mathscr{D}'}[(v_{\max} \cdot x)^2 > r]$ is at least $2\,\mathbb{P}_{x \sim \mathscr{N}}[(v_{\max} \cdot x)^2 > r]$

Condition $\mathscr{D}'$ on $(v_{\max} \cdot x)^2 \leq r$

Check: Mass of rejected region $O(\epsilon)$