A Learning Based Hypothesis Test for Harmful Covariate Shift

Tom Ginsberg, Zhongyuan Liang, Rahul G. Krishnan

{tomginsberg, zhongyuan, rahulgk}@cs.toronto.edu

VECTOR INSTITUTE

UNIVERSITY OF TORONTO

Motivation

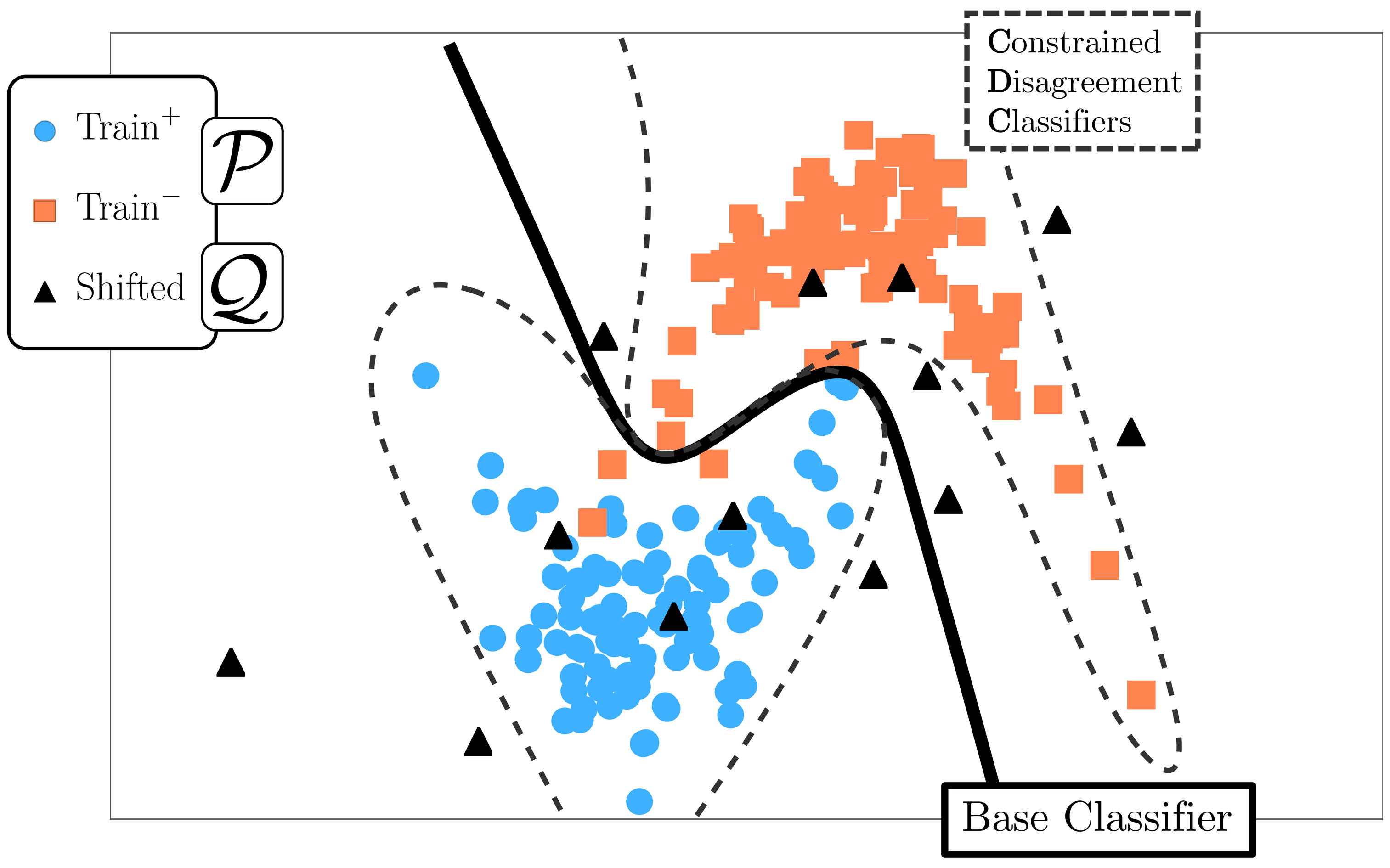
- The ability to **quickly and accurately** identify covariate shifts at test time is a critical component of safe machine learning systems deployed in high-risk domains
- We show how to leverage any deployed classifier as a domain-aware shift detector well-suited to small sample sizes

Approach: Detecting Shift with Model Disagreement

- **We define** *harmful covariate shift* as a shift where a model's behaviour is poorly specified due to lack of learned invariances
- **We identify** harmful shifts by answering the question:

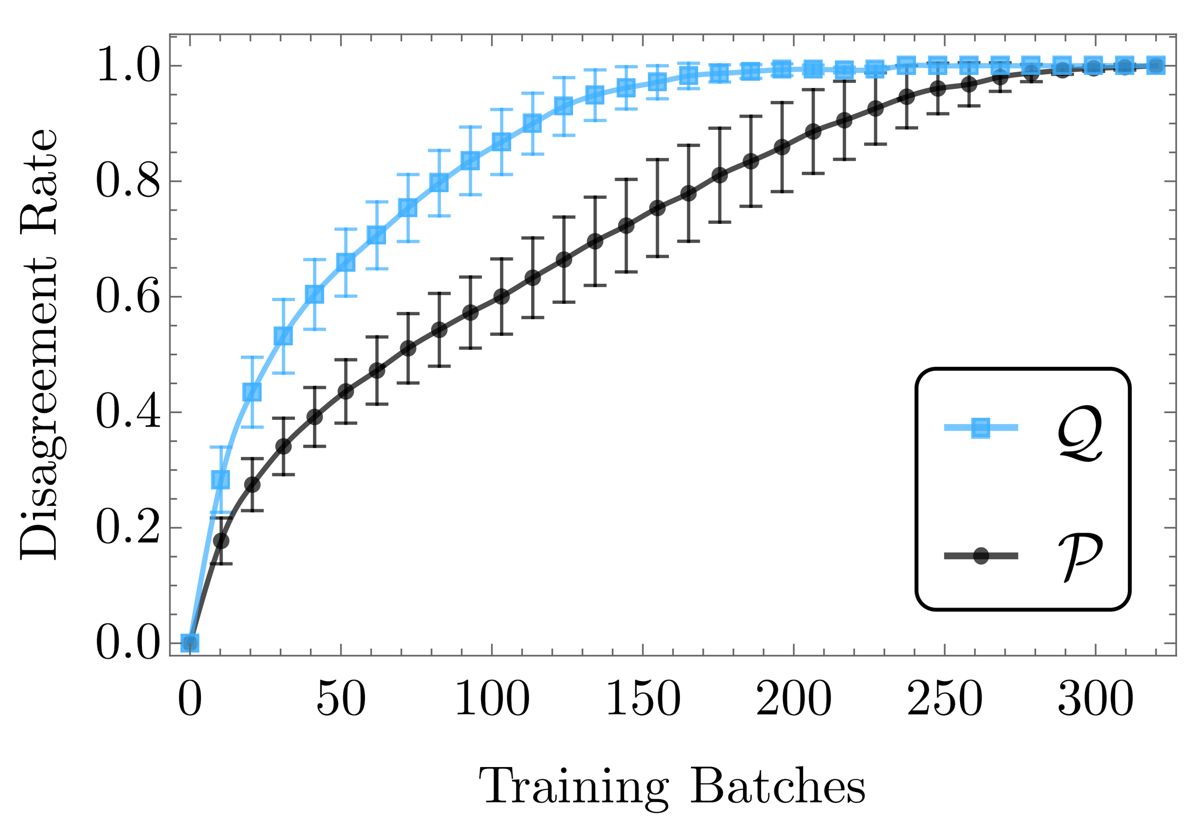
> *can we train multiple models to meet similar performance criteria on the training dataset $\mathcal{P}$ while disagreeing with each other on an unlabeled test set $\mathcal{Q}$?*

- **We give an algorithm** for *C*onstrained *D*isagreement *C*lassifiers (CDCs) which maximize classification disagreement on $\mathcal{Q}$ while constrained to predict consistently on $\mathcal{P}$



- The rate that CDCs disagree is a powerful and sample-efficient statistic for identifying covariate shift $\mathcal{P} \neq \mathcal{Q}$.

Learning to Disagree and Two-Sample Testing



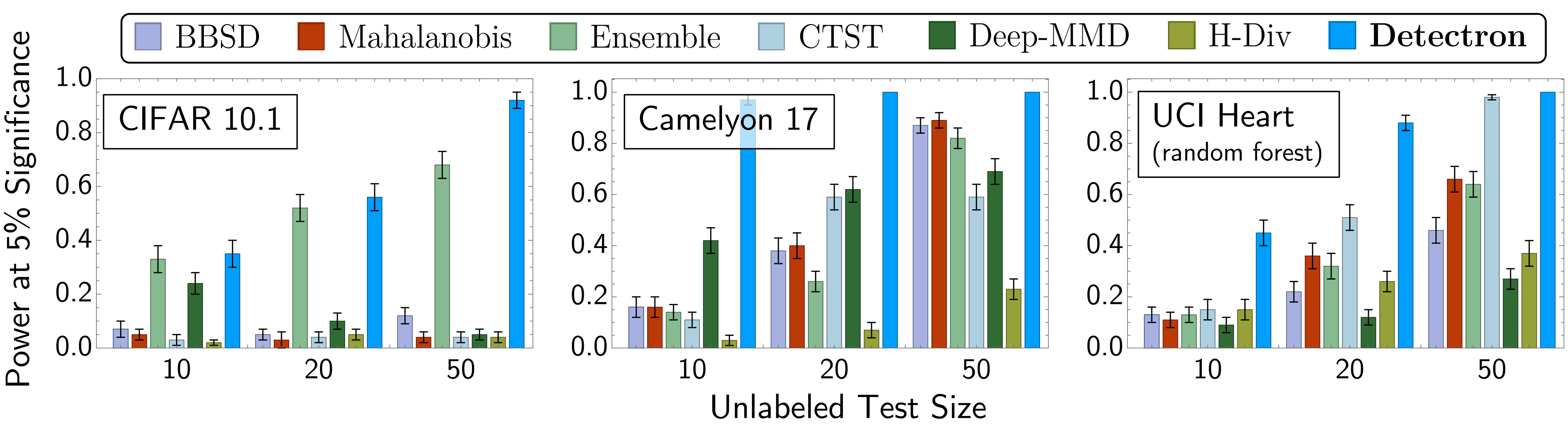
- We train a model g to **agree** with a set of labelled training data, while disagreeing with a baseline model f on unlabeled data

$$\mathcal{L}(\mathcal{P}, \mathcal{Q}) = \frac{1}{|\mathcal{P}| + |\mathcal{Q}|}\left(\overbrace{\sum_{(x,y)\in\mathcal{P}} \mathcal{A}(g(x), y)}^{Agreement} + \lambda \overbrace{\sum_{\tilde{x}\in\mathcal{Q}} \mathcal{D}(g(x), f(x))}^{Disagreement}\right)$$

- A permutation test is used to bound the significance level for rejecting $\mathcal{H}_0$:

> $\mathcal{H}_0$: g will disagree on $\mathcal{P}$ and $\mathcal{Q}$ with the same rate
> $\mathcal{H}_a$: g is more likely disagree on $\mathcal{Q}$ compared to $\mathcal{P}$ $\implies$ harmful shift

Performance on Shift Detection Benchmarks



Conclusion and Future Work

- We present a promising technique to perform two sample testing with high statistical power using a pretrained classifier
- **Future directions:**

> Improving computational runtime | Establishing a theoretical connection between model complexity generalization error and test power | Extending to arbitrary tasks beyond classification | Large-scale experiments (e.g. ImageNet)