(ALMOST) PROVABLE ERROR BOUNDS UNDER DISTRIBUTION SHIFT VIA DISAGREEMENT DISCREPANCY

Carnegie Mellon University
ML MACHINE LEARNING DEPARTMENT

ELAN ROSENFELD SAURABH GARG

PAPER: CODE:

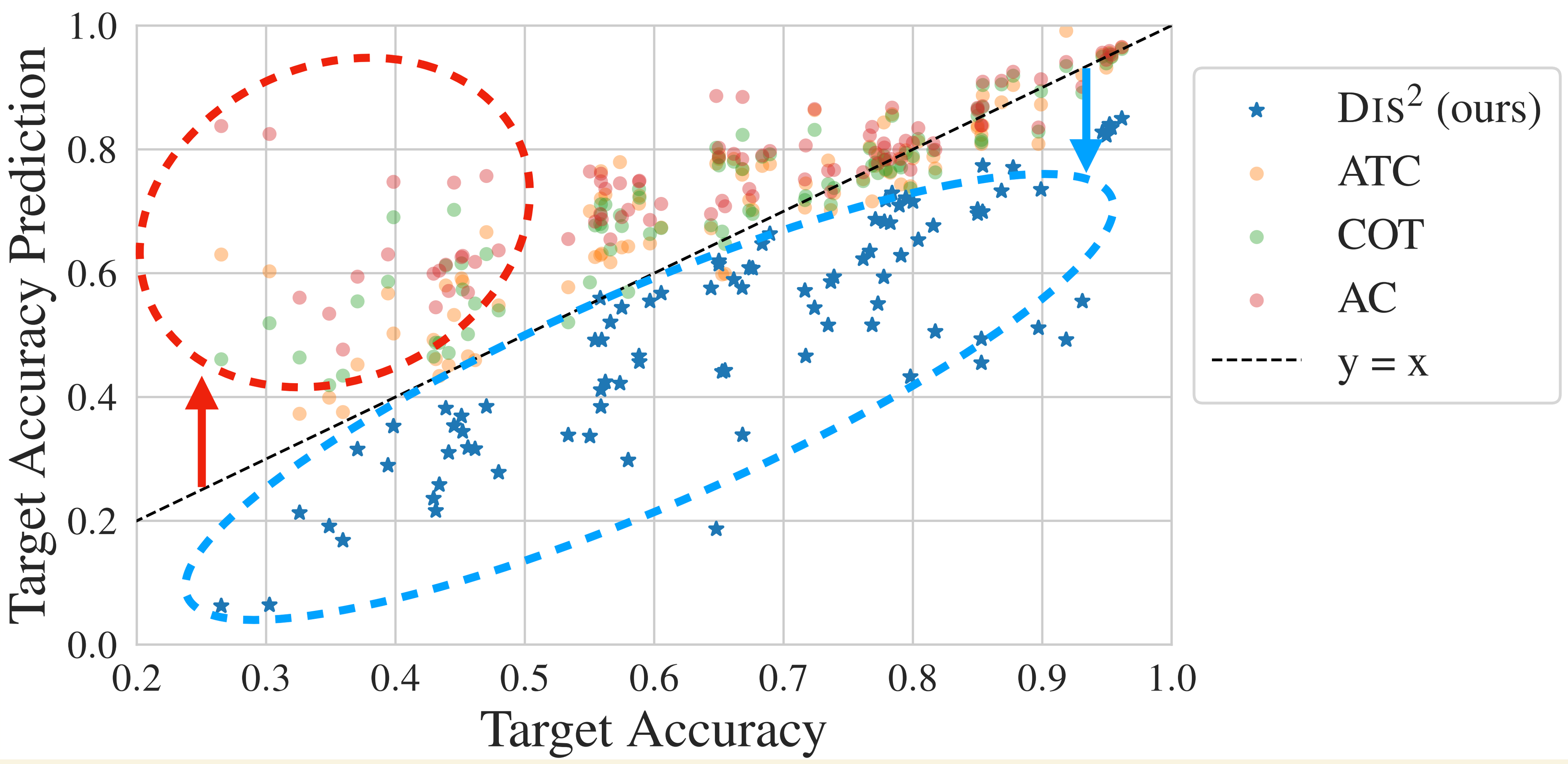
Several methods use unlabeled test data to estimate error under distribution shift.

With one minor assumption, we give a guaranteed error upper bound.

Gives valid, non-vacuous error bounds effectively 100% of the time on real data.

Previous methods consistently ***underestimate* error*** under shift, **especially when the shift is large.**

* (i.e., overestimate accuracy)



Our method gives a (probabilistic) ***upper bound*** and maintains competitive average prediction accuracy.

The bound is extremely simple.

For classifiers h, h' and source/target distributions S, T, define the ***Disagreement Discrepancy*** as their disagreement on T minus their disagreement on S:

$$\Delta(h, h') := \epsilon_T(h, h') - \epsilon_S(h, h')$$

For the true labeling function y^*, $\epsilon_T(h, y^*) = \epsilon_S(h, y^*) + \Delta(h, y^*)$.

Don't know this... 😞

Optimize over *critics* h' in hypothesis class $\mathscr{H}$ to find an *upper bound* $\Delta(h, h')$.

Assumption: Define $h^* := \arg\max_{h' \in \mathscr{H}} \Delta(h, h')$. We assume $\Delta(h, y^*) \le \Delta(h, h^*)$.

Immediately implies population bound: $\epsilon_T(h, y^*) \le \epsilon_S(h, y^*) + \Delta(h, h^*)$.

With a bit more work, we arrive at the probabilistic empirical bound:

This we can estimate! 😁

Theorem: With probability $\ge 1 - \delta$,

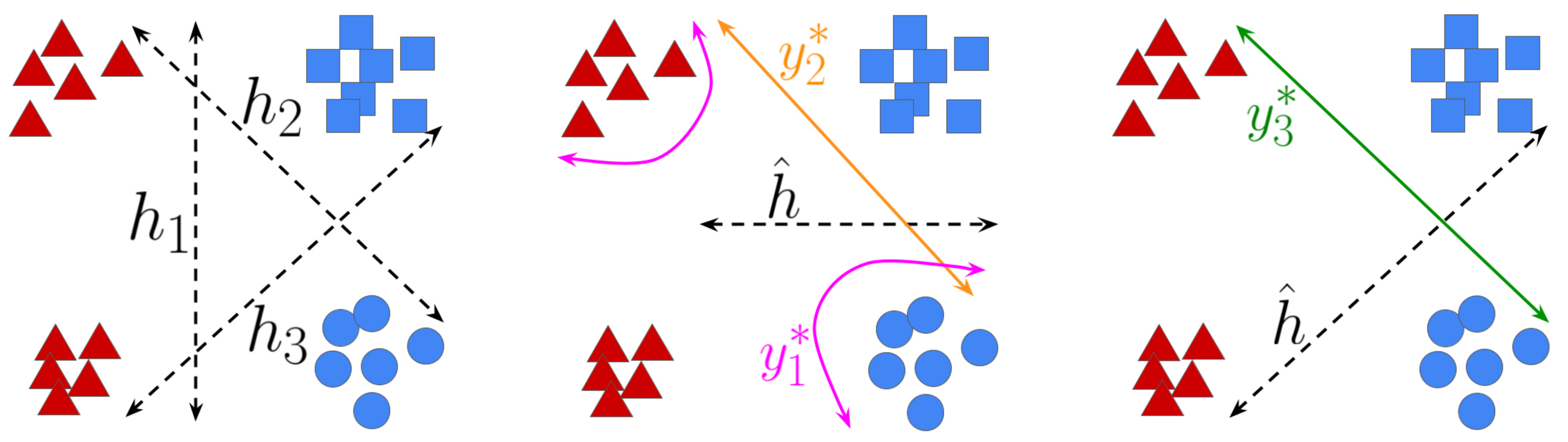
$$\epsilon_T(h, y^*) \le \hat{\epsilon}_S(h, y^*) + \hat{\Delta}(h, h^*) + \sqrt{\frac{(n_S + 4n_T)\ \log 1/\delta}{2 n_S n_T}}$$

[1] Domain-Adjusted Regression or: ERM May Already Learn Features Sufficient for Out-of-Distribution Generalization. Rosenfeld et al. 2022

We choose $\mathscr{H}$ as set of linear classifiers. **Why should we expect the assumption to hold?**

1. Labeling function fixed ahead of time—**it is not *chosen* to maximize $\Delta(h, y^*)$.**
2. Linear classifier can achieve excellent accuracy even under distribution shift.[1] **So y^* is already close to linear.**

ADVANTAGE OF DIS² OVER BOUNDS BASED ON H AND HΔH-DIVERGENCE



Prediction Method (Domain Adversarial?)	Coverage (↑) ✗	Coverage (↑) ✓	Conditional Overestimation (↓) ✗	Conditional Overestimation (↓) ✓	MAE (↓) ✗	MAE (↓) ✓
AC	0.1000 ± .032	0.0333 ± .023	0.1194 ± .012	0.1123 ± .012	0.1091 ± .011	0.1091 ± .012
DoC	0.1667 ± .040	0.0167 ± .017	0.1237 ± .012	0.1096 ± .012	0.1055 ± .011	0.1083 ± .012
ATC NE	0.2889 ± .048	0.1333 ± .044	0.0824 ± .009	0.0969 ± .012	0.0665 ± .007	0.0854 ± .011
COT	0.2554 ± .047	0.1667 ± .049	0.0860 ± .009	0.0948 ± .011	0.0700 ± .007	0.0808 ± .010
DIS²	0.9889 ± .011	0.7500 ± .058	0.0011 ± .000	0.0475 ± .007	0.1489 ± .011	0.0945 ± .010
DIS² (no δ term)	0.7556 ± .048	0.4333 ± .065	0.0771 ± .013	0.0892 ± .011	0.0887 ± .009	0.0637 ± .008