ETH zürich

UCL

Achievable distributional robustness when the robust risk is only partially identified

Julia Kostin[1], Nicola Gnecco[2], Fanny Yang[1]
[1]Department of Computer Science, ETH Zurich [2]Gatsby Computational Neuroscience Unit, UCL

Given: $P_1^{X,Y}$ $P_2^{X,Y}$ $P_3^{X,Y}$ P_{test}^{X}

- **Multi-environment training data** $\{P_e^{X,Y}\}_e$
- Some **knowledge of the test distribution shift**

Questions:

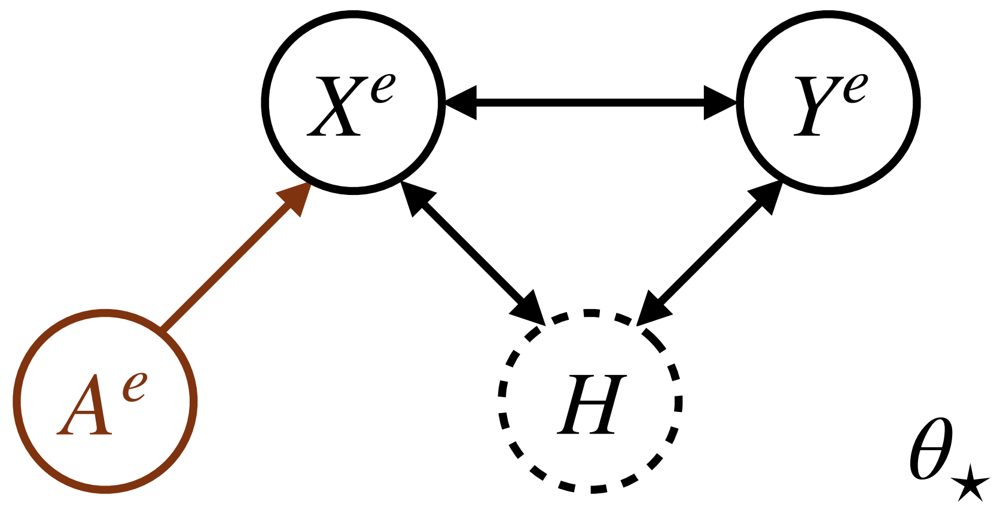
1. How well can any algorithm generalize to $P_{\text{test}}^{X,Y}$ given a collection of different training distributions?
2. What can we do **if there is not enough data heterogeneity** for generalization on test data?

Our work:

- **Introduces a framework** that allows well-defined performance quantification in this more realistic scenario
- Quantifies minimal **identifiable robust risk (i.r.r.) achievable** by *any algorithm* (introduced for linear setting, applicable more generally)
- **Evaluates existing robustness methods** in the harder scenario of insufficient heterogeneity / non-identifiability

Linear setting

Training distribution $P_e^{X,Y}$ for environment e defined by



$$X^e = A^e + \eta;$$
$$Y^e = \beta_\star^\top X^e + \xi,$$

where $(\eta, \xi) \sim \mathcal{N}(0, \Sigma_\star)$ and $\theta_\star = (\Sigma_\star, \beta_\star) \in \Theta$ are *invariant*.

At **test time**, we observe **test shift** $A^e = A^{\text{test}}$ with

$$\mathbb{E}[A^{\text{test}} A^{\text{test}^\top}] \preceq M_{\text{test}} = \gamma \Pi_{\mathcal{M}}.$$

Shift strength **Shift directions**

Allows to incorporate *different granularities of knowledge*:

- Know $\mathbb{E}[A^{\text{test}} A^{\text{test}^\top}]$ ~ have P_{test}^X (domain adaptation)
- Use $\mathcal{M} \subsetneq \mathbb{R}^d$ ~ *some knowledge* of distribution shift
- Use $\mathcal{M} = \mathbb{R}^d$ ~ *no knowledge* (most conservative)

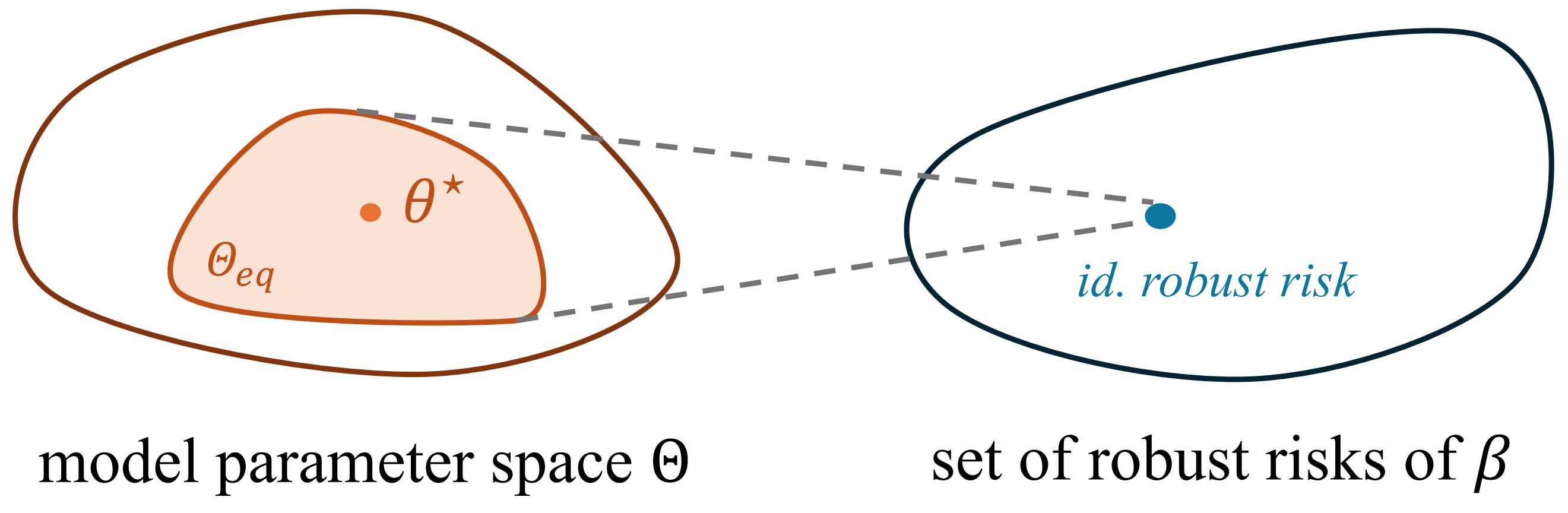
Partially identifiable robustness

Span of shifts seen in training: $\mathcal{S} = \text{range}\left(\sum_{e \in \mathcal{E}_{\text{train}}} \mathbb{E}[A^e A^{e\top}]\right)$

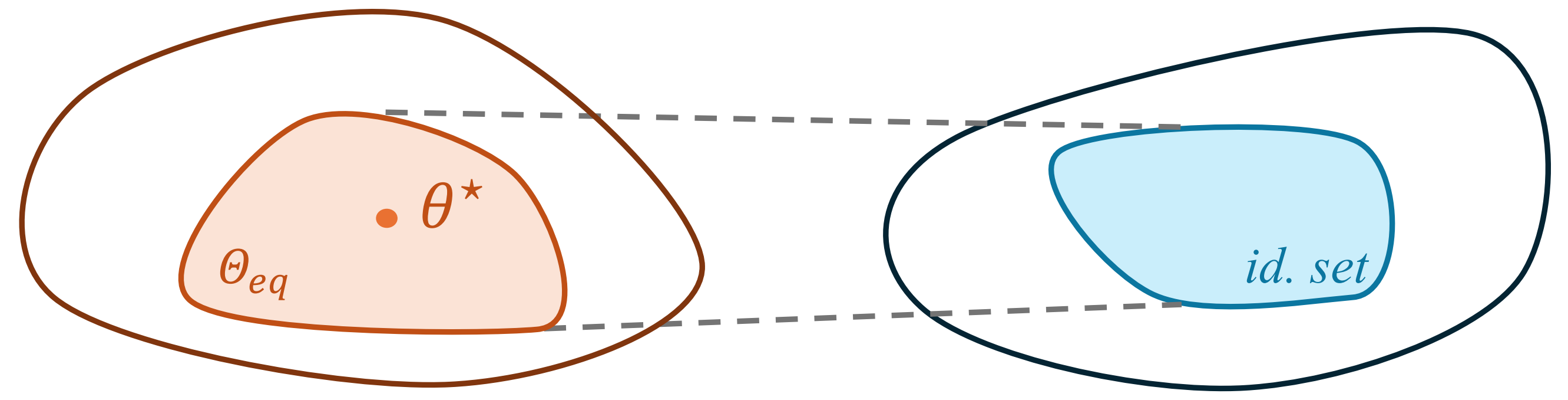
Def.:

Robust risk (**r.r.**) $\mathcal{R}_{\text{rob}}(\beta; \theta, M_{\text{test}})$ - worst-case error w.r.t. distribution shift.

- **Case 1**: $\text{range}\, M_{\text{test}} \subseteq \mathcal{S}$, **robust risk is identifiable**:



- **Case 2**: $\text{range}\, M_{\text{test}} \not\subseteq \mathcal{S}$, **r.r. only partially identifiable**:



Our notion of **identifiable robust risk (i.r.r.)**:

$$\mathcal{R}_{\text{rob,ID}}(\beta; \Theta_{\text{eq}}, M_{\text{test}}) := \sup_{\theta \in \Theta_{\text{eq}}} \mathcal{R}_{\text{rob}}(\beta; \theta, M_{\text{test}}).$$

Then, **minimax identifiable robust risk** reveals achievable performance by any algorithm:

$$\mathfrak{M}(\Theta_{\text{eq}}, M_{\text{test}}) = \inf_{\beta \in \mathbb{R}^d} \mathcal{R}_{\text{rob,ID}}(\beta; \Theta_{\text{eq}}, M_{\text{test}}).$$

Results for the linear setting

Theoretical result 1: Lower bound for minimax i.r.r.

$$\mathfrak{M}(\Theta_{\text{eq}}, \gamma\Pi_{\mathcal{M}}) = \gamma C_{\text{ker}}^2 + \min_{R^\top \beta = 0} \mathcal{R}_{\text{rob}}(\beta; \theta_\star, \gamma S S^\top), \text{ if } \gamma \geq \gamma_{\text{th}}$$

where S, R: orthogonal decomposition of M_{test} such that $\text{range}\, S \subset \mathcal{S}$ and $\text{range}\, R \subset \mathcal{S}^\perp$.

$\Longrightarrow$ For large $\gamma \geq \gamma_{\text{th}}$, optimal predictors refrain in $\text{span}(R)$;

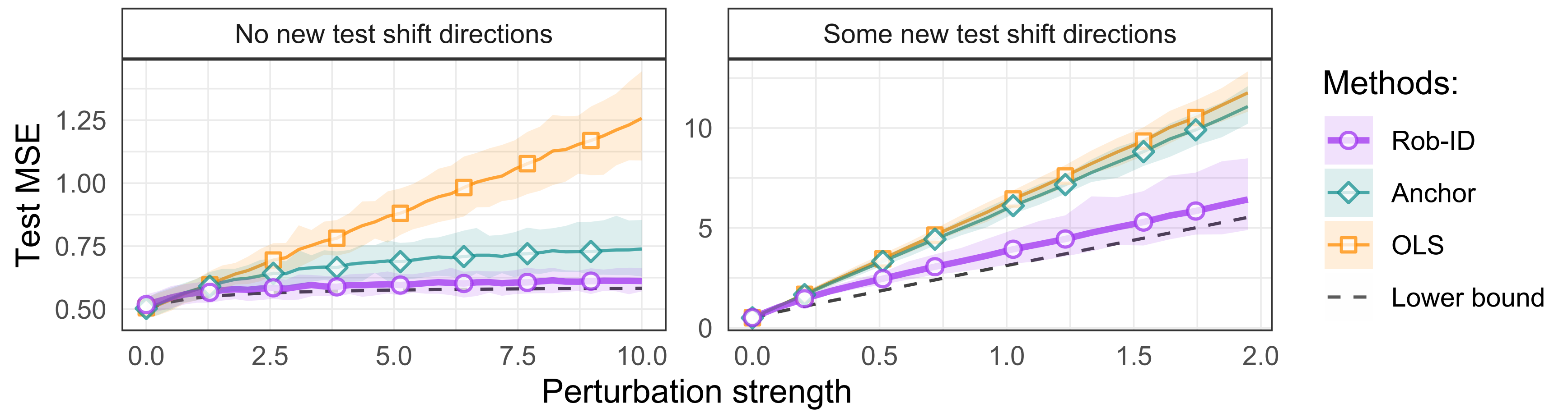
$\Longrightarrow$ Risk **grows linearly w.r.t.** unobserved shift strength γ.

Theoretical result 2: Performance of existing methods

- For large new shifts, **empirical risk minimization (OLS)** yields error akin to known **invariance-based methods,** e.g.:
  - ***Anchor regression*** [Rothenhäusler et al. 2021] or
  - ***DRIG*** [Shen et al. 2023]).
- They are **provably worse than the minimax predictor**

Experiments confirm theoretical conclusions:

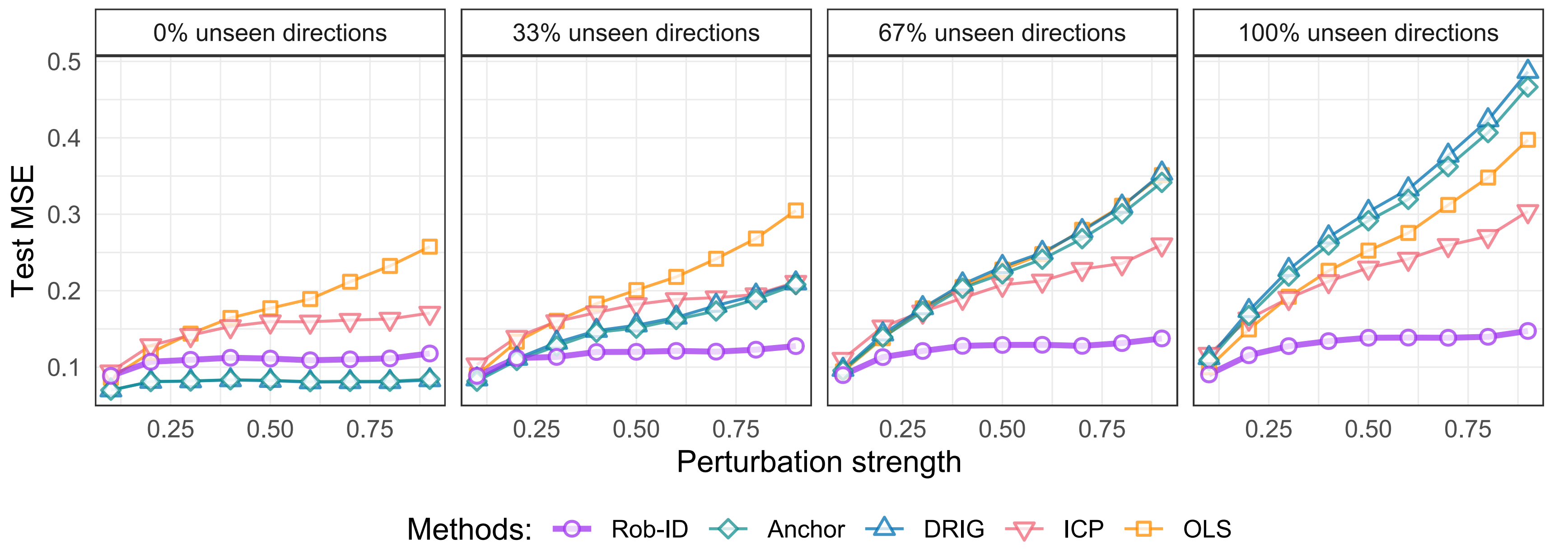
left: case 1, identifiable. **right**: case 2, partially identifiable



where **Rob-ID** is empirical minimizer of the identifiable r.r.

Comparison on real-world dataset

Performance of various invariance-based OOD methods, evaluated on **real-world gene expression dataset** [Replogle et al. 2022] in 1) ***identifiable case*** (left) vs. 2) ***partially identifiable case*** (others)



- **Ranking** of robust prediction methods **changes** in partially identifiable settings!
- Minimizer of the i.r.r. **outperforms existing methods despite possible assumption violations** in real data.

Call to evaluate robustness methods on partially identifiable scenarios theoretically & experimentally!