EurIPS
Copenhagen

Robust Generalization Under Misspecified Robust Risk and the Role of Limited Target Data

Fanny Yang, CS @ ETH Zurich
Statistical Machine Learning group

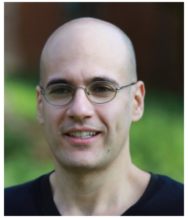
joint work with

Julia Kostin
ETH Zurich

Nicola Gnecco
Imperial College

Kasra Jalaldoust
Columbia

Samory Kpotufe
Columbia

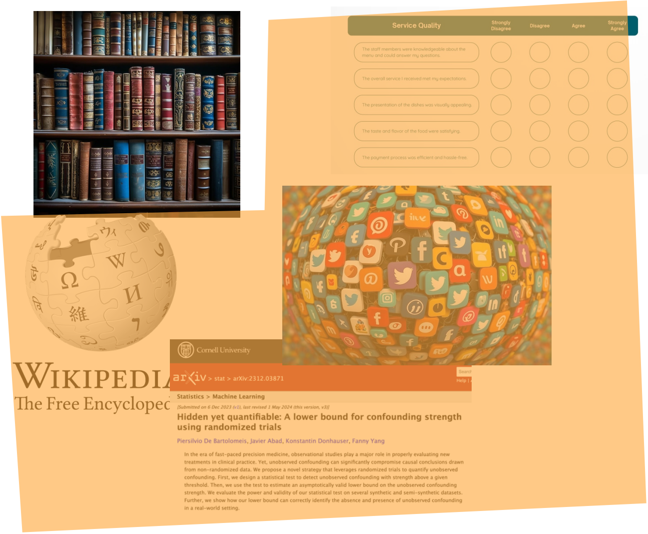


Elias Barenboim
Columbia

Distributions shift – a challenge and blessing?

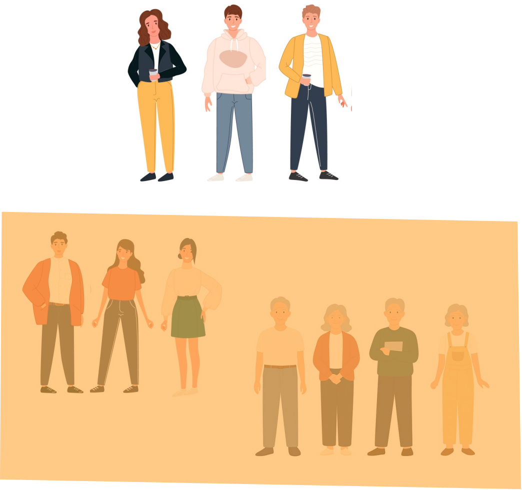
Out-of-distribution generalization: Test on **domains** unseen during training

Different text corpuses



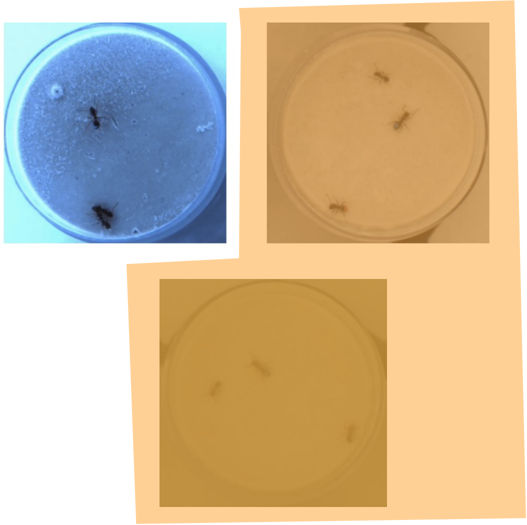
Text → Next token

Different populations



Blood measurements → Disease

Different experimental conditions



Video → Grooming behavior

Heterogeneous data – a challenge and blessing?

Out-of-distribution generalization: Test on **domains** unseen during training

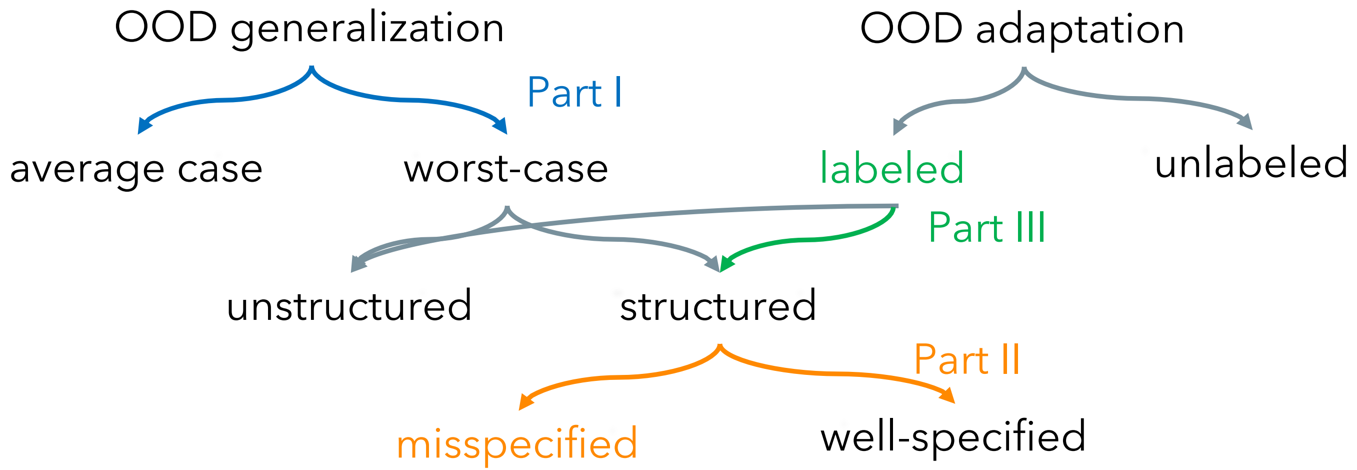
Challenge!

Goal: Inference / prediction on new *target* $\mathbb{P}_{test}$ (not equal to any $\mathbb{P}_i$)

Blessing!

Available data: Lots of labeled data from multiple *sources* $\mathbb{P}_i$

Plan today



OOD generalization
Part I
average case
worst-case
OOD adaptation
labeled
unlabeled
Part III
unstructured
structured
Part II
misspecified
well-specified

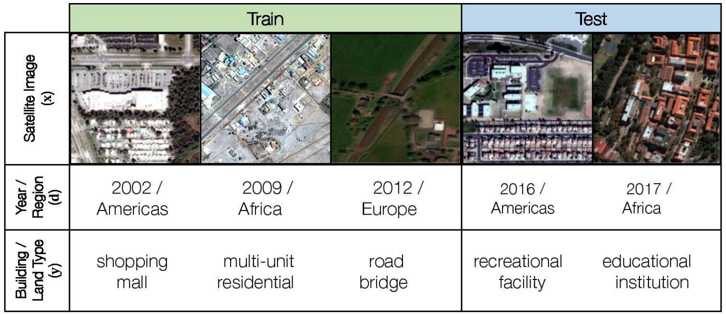
SPEEDUP

I. Out-of-distribution generalization – average case

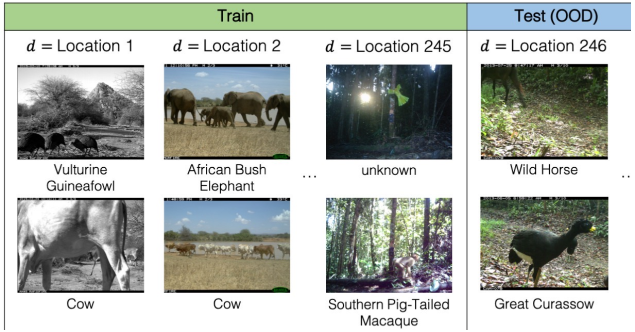
Goal: good performance on a shifted distribution you might "naturally" encounter

(*random* domain shifts). Application where this matters (data from WILDS [Koh, Sagawa et al. '21]):

- Predictions eventually informing policies



- Animal recognition for scientists/rangers



Real-world evaluation: easy - given many environments, pick one to test on, train on rest

Method-wise: ill-defined - need many environments or distributional assumptions on shift

I. Out-of-distribution generalization – worst-case

Goal: require good performance on worst-case domains by attacks/shifts

- Safety–critical applications, e.g. cars/planes (crashes), medical diagnosis (death)

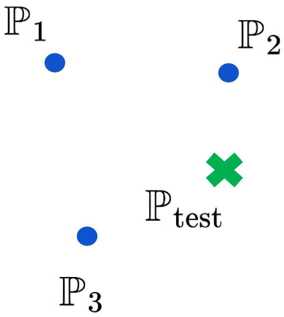
- Forbidden content detection against adversaries, e.g. hate speech, explicit/violent videos

Real-world evaluation: hard - adversarial environments depend on the model → artificial

Method-wise: better defined – “only” need to know the set of shifts

I. "Average-case" vs. worst-case OOD generalization

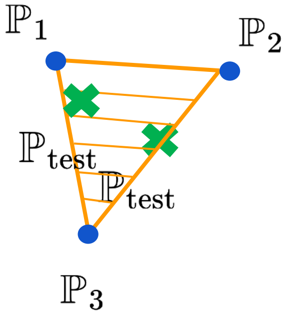
Goal: Do well on a target domain that's naturally shifted "like" source data
e.g. meta-learning, hierarchical Bayesian methods



no gain from robust methods

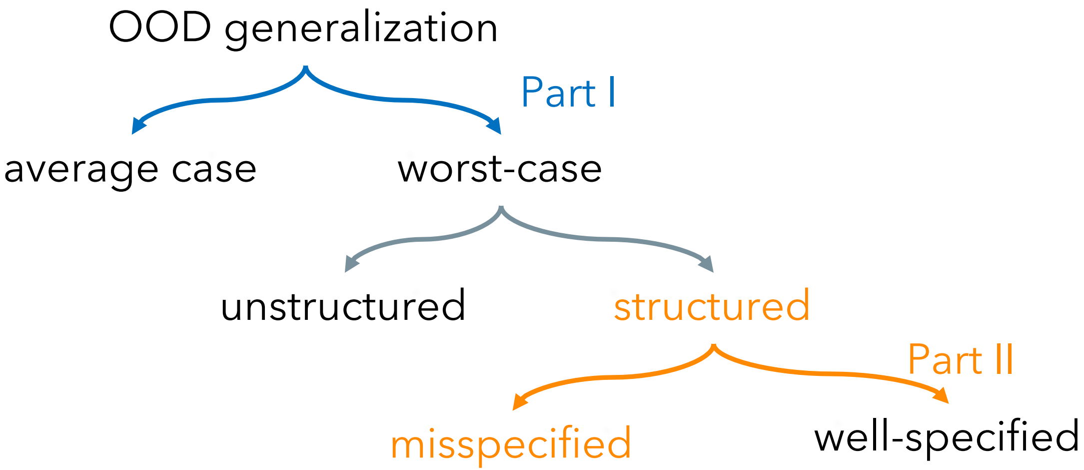
Evaluation on real data:
Randomly held-out domains
e.g. [Guljarani & Lopez-Paz'21, Miller et al. '22, Nastl & Hardt '24]

Goal: Do well on worst possible target domain within a possible set of shifts
e.g. robust generalization



Evaluation on real data:
Hardest held-out domain
rare, see e.g. [Salaudeen et al '25]

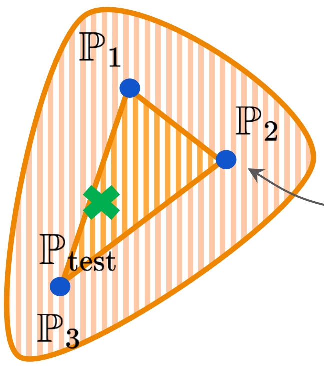
Plan today


OOD generalization
Part I
average case
worst-case
unstructured
structured
Part II
misspecified
well-specified

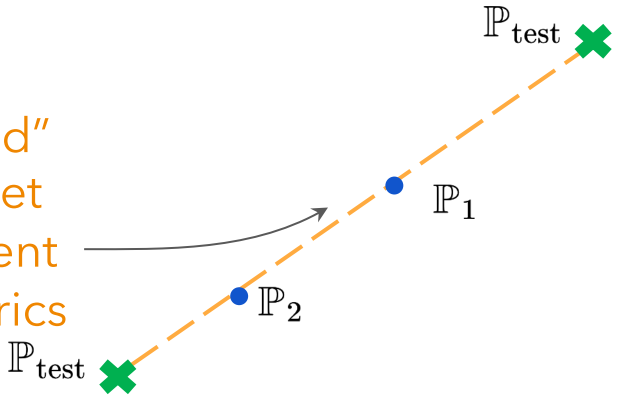
From unstructured to structured shift assumptions

Risk of predictor h on loss on distribution $\mathbb{P}$

Existing robust methods find minimizer of the robust risk $\mathcal{R}_{rob}(h) = \sup_{\mathbb{P} \in \mathcal{P}_{rob}} R(h; \mathbb{P})$ that has guaranteed "small" risk for all $\mathbb{P} \in$ shift-model-specific **robustness set** $\mathcal{P}_{rob}$ of "close" domains.



"Well-specified" Robustness Set $\mathcal{P}_{rob}$ for different closeness metrics



Unstructured (e.g. Group-DRO)

Confined to convex hull of seen distributions (and small e.g. Wasserstein balls around each)

Structured (invariance-based)

Can be robust very far away as long as it is a "familiar" direction of shift

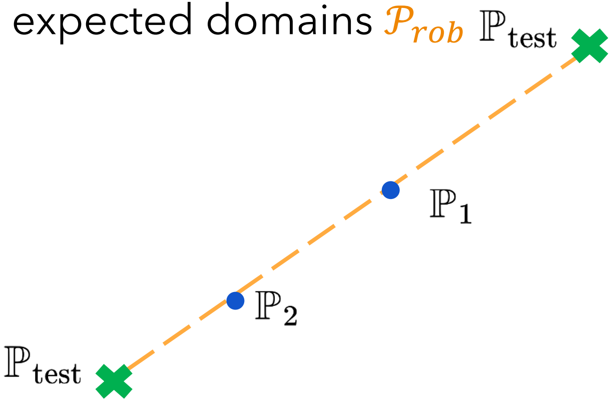
Well-specified vs. misspecified on a high level

Let's say your source domains have shifts in age distribution and salary

Example I (well-specified)

- You expect the target to have shifts in age distribution

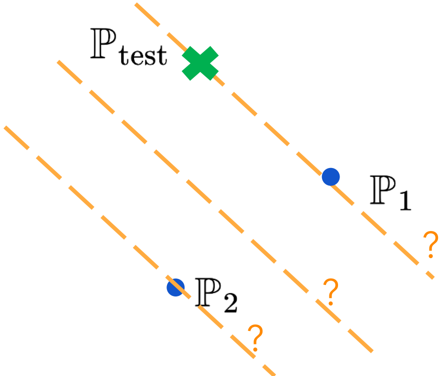
➡ you can "fully imagine" the set of expected domains $\mathcal{P}_{rob}$



Example II (misspecified)

- You expect the target to have shifts in terms of geographical region

➡ you can only "partially imagine" the set of expected domains $\mathcal{P}_{rob}$

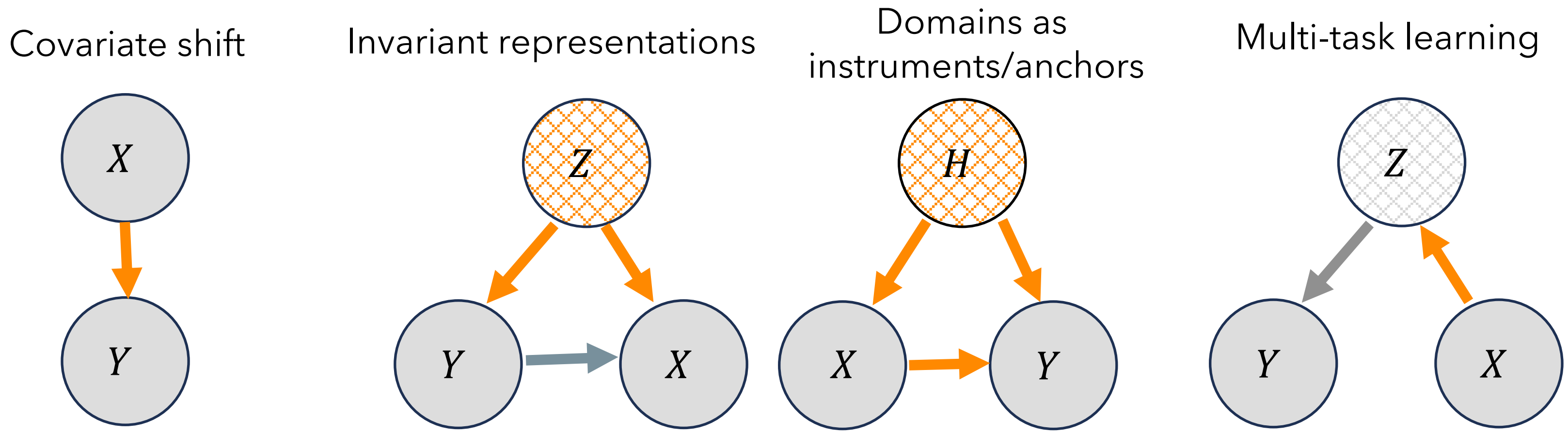


Unified view on many structured shift assumptions

…via invariances in (causal) DAGs/Bayesian networks/graphical models where:

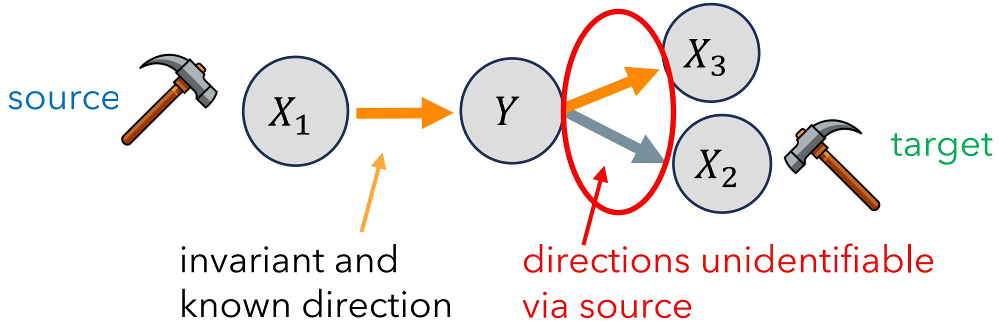
- arrows model (noisy) relationships between child & parent (such as structural equations)
- shifts are modeled as interventions on the variables or varying conditionals

Orange: Invariant across domains, Gray: May shift across domains; Checkered: Unobserved



Misspecification via invariance: partial identifiability

- Misspecified worst-case setting: when we can't identify the invariant mechanisms from sources, and the target shift are different than source shifts!



- Nonidentifiable stable set $\{X_1, X_3\}$
- Hence, unclear how the target shifts affect joint of X, Y!

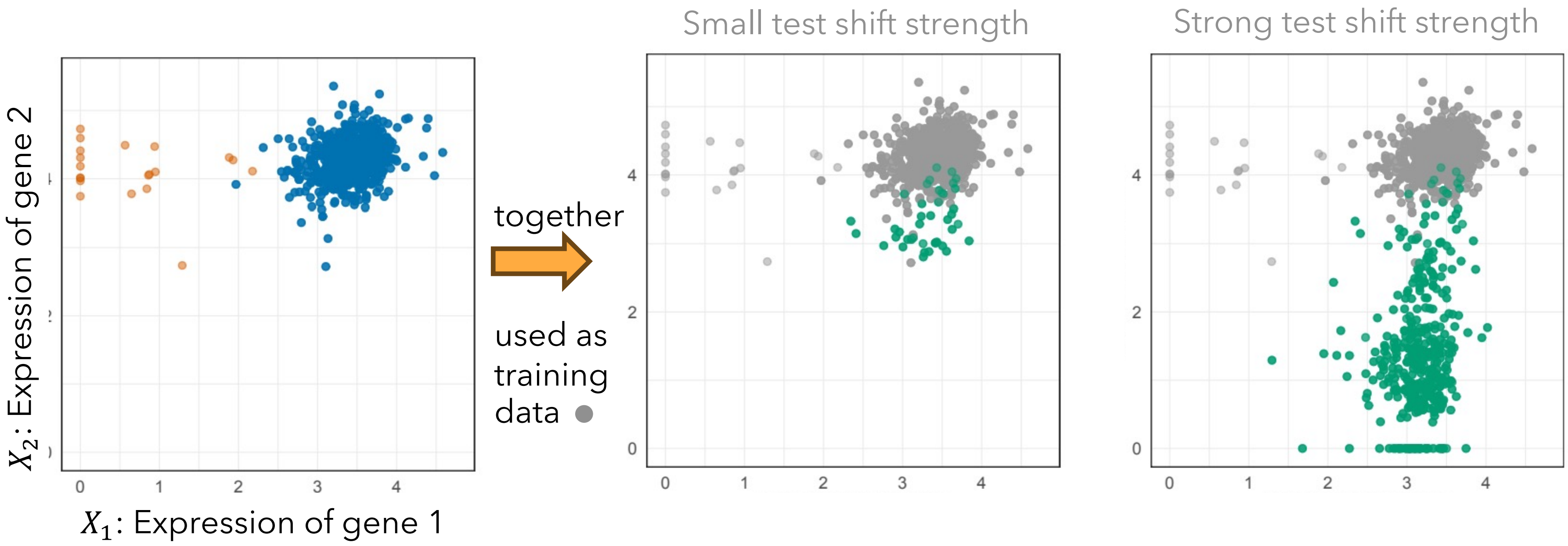
⇨ Then instead of one $\mathcal{P}_{rob}$, we can only identify a set of sets $\mathcal{P}_{rob}$ to be robust against (here two, one per possible direction ←, →)

Important to do so!

⇨ W/o more assumptions for robustness you can only evaluate $\mathfrak{R}_{\mathrm{rob}}(h) = \sup_{\mathcal{P}_{rob}} \sup_{\mathbb{P} \in \mathcal{P}_{rob}} R(h; \mathbb{P})$

Misspecified setting for gene expression data [KG**Y** '24]

Task: Predict the expression of gene 0 using gene 1,2,3 (from [Replogle, et al., 2022])

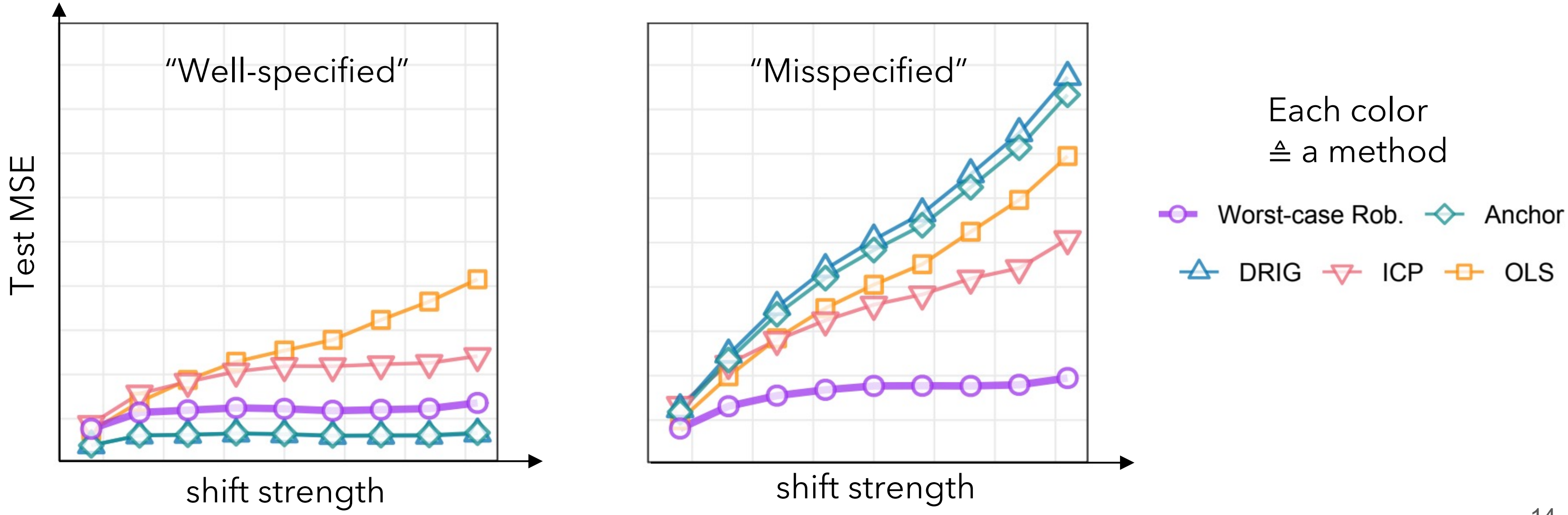


● Data with knocked out gene 1 ● Unintervened data ● Test data (knockout of gene 2)

Misspecified worst-case settings matter [KGY '24]

In such "misspecified" worst-case settings, performance rankings may change!

Experiments on gene expression data, mimicking "worst-case" evaluation



Each color
≙ a method

II. One (unified) table of different OOD paradigms

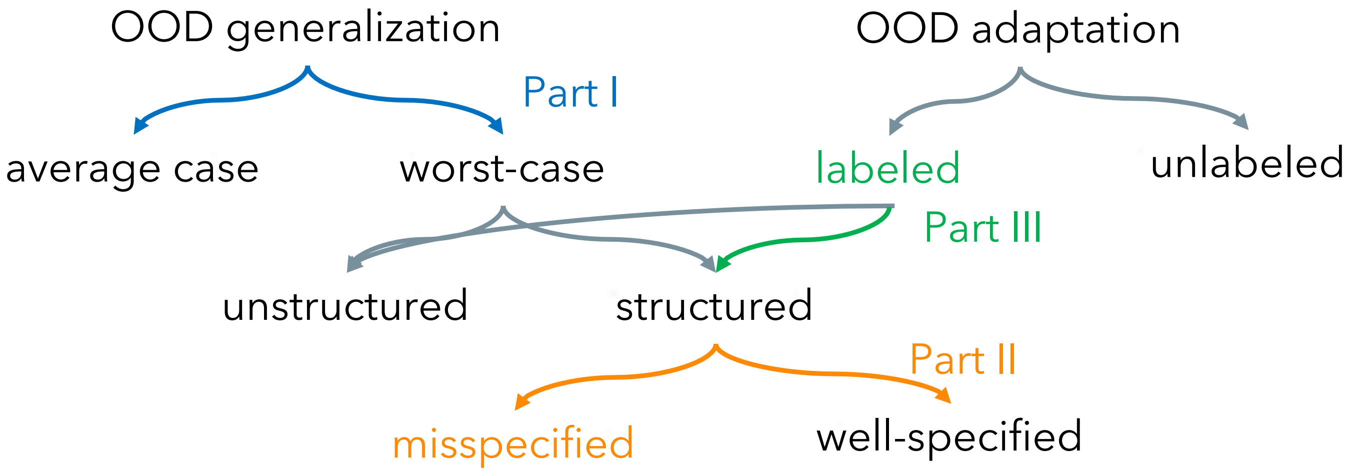
Easier and often more realistic

Harder but more "realistic"

	No data from target (robust generalization)	
Unspecified	impossible	Single-source transfer learning [BCKPV '07, BBCKPV '10, KH'24], continual learning
Misspecified	Partial identifiability [KG**Y**'24] / transportability [JBB'24]	in preparation [KJBK**Y** '26]
Well-specified	Invariance-based learning [PBM '15, ABGL '19, KCJZBZPC '20] GroupDRO and variants [MMR '08, SKHL '19, KSWJPEZ '25]	Multi-source transfer / DA [MMRSW '21, XGC '23, CSEC '24], Using multi-task learning [TJJ '20, DHKLL '20]

A small sample of papers with theoretical guarantees, decipher using LLMs…

Plan today


OOD generalization
Part I
average case
worst-case
OOD adaptation
labeled
unlabeled
Part III
unstructured
structured
Part II
misspecified
well-specified

A simple example for illustration

Assume infinite samples from the source distributions for simplicity

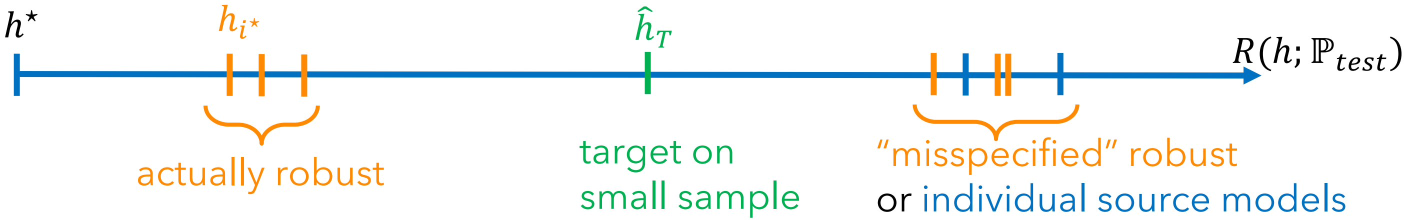
- Recall that in misspecified case: many robustness sets $\mathcal{P}_{rob,1}, \dots, \mathcal{P}_{rob,m}$

$\rightarrow$ multiple robust risks $\rightarrow$ set of robust risk minimizers $h_1, \dots, h_m$ from sources

- Good news: at least for some index $i^\star$, one of h_i has small robust risk!

$\sup_{\mathbb{P} \in \mathcal{P}_{rob}} R(h; \mathbb{P})$

$\rightarrow$ if shift model was correct, for any possible target set, $R(h_{i^\star}; \mathbb{P}_{test}) \leq \mathcal{R}_{rob}(h_{i^\star})$



- Bad news: But which one of those m to pick? $\rightarrow$ Target samples can help us pick!

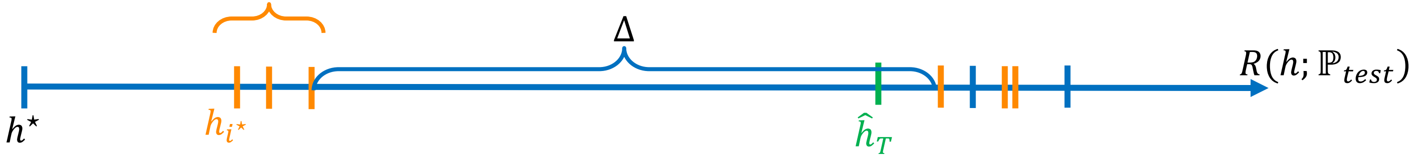
Generally applicable procedure [KJBK**Y** '26]

Two-stage procedure

1. Obtain the set of candidate models $h_1, \dots, h_m$ using invariance-based methods in misspecified / partially identifiable settings; let $h_{(1)}$ best on target risk
2. Using target data, adaptively pick one of the models or the target model $\hat{h}_T$

Informal guarantees we can achieve with stage 2 (currently for linear regression)

- never worse than rate of target estimator $\hat{h}_T$ (no negative transfer!)
- guaranteed to pick w.h.p., if $n_T \geq \frac{\log M}{\Delta}$ (fast adaptation for large shifts!)



Comparison with prior transfer algorithms [KJBK**Y** '26]

Structured transfer (e.g. multi-task learning):

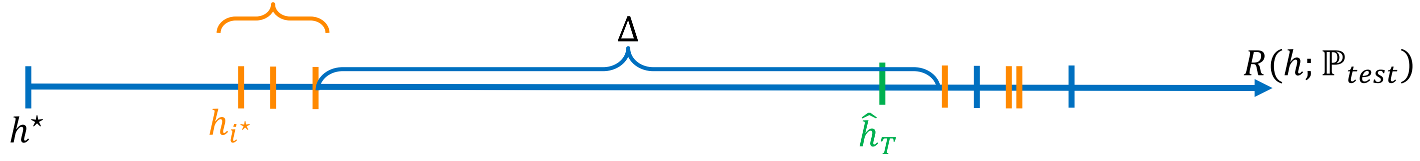
- identifiable case: still rely on few target samples; if invariant, we can gain via source
- unidentifiable case: may pick "wrong", varying representation → large target risks

Unstructured transfer (e.g. only using source models):

- no robustness guarantees against large structured shifts → large target risks

Informal guarantees we can achieve with stage 2

- never worse than rate of target estimator $\hat{h}_T$ (no negative transfer)
- guaranteed to pick w.h.p., if $n_T \geq \frac{\log M}{\Delta}$ (fast adaptation for large shifts!)



Summary

- Worst-case / robust generalization methods should be evaluated on worst cases before concluding that they're futile on randomly held-out datasets

- In "misspecified" worst-case settings, performance rankings may change
→ albeit harder, they're more realistic when you have only few source domains
it's important to at least evaluate and analyze that case!

- Transfer learning can take advantage of invariance-based assumptions (if true) and target data can help significantly in misspecified worst-case settings

References

J. Kostin, N. Gnecco, F. Yang "Achievable distributional robustness when the robust risk is only partially identified", NeurIPS 2024

Transfer learning and (misspecified) robustness:

J. Kostin, K. Jalaldoust, E. Barenboim, S. Kpotufe, F. Yang, In preparation

SML group at ETH Zurich:
sml.inf.ethz.ch

A gene expression experiment (rewrite)

As mentioned, evaluation for worst-case is hard with real-world data

We cannot take the sup over expected distributions -> pick hardest

- Task: Predict X_i expression of gene i as a function of X_{i_1}, X_{i_2} expressions of 3 other genes
- Source domains $e \triangleq$ knock-out of gene X_j (+ observational)
- Misspecified setting: Test domain $\triangleq$ knock-out of genes $X_k \neq X_j$ not in source domain
- Well-specified setting: Test domain $\triangleq$ knock-out of gene X_j
- shift strength $\gamma \triangleq$ distance of covariates to mean in observational environment

Some more details on experiments

Question: In the misspecified case,

What's the best we can do (lower bound for R_rob) and how do existing methods rank

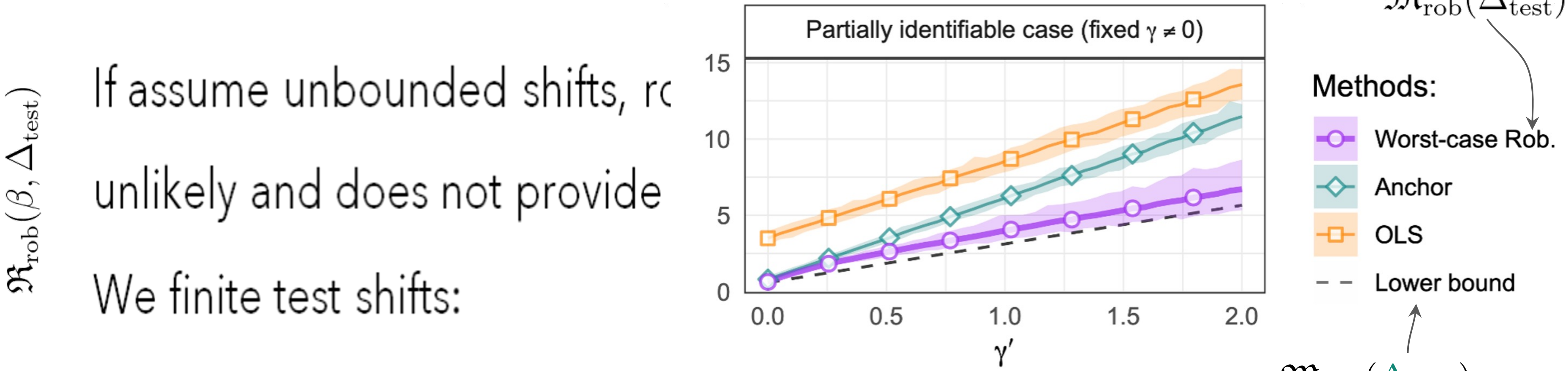
*[Replogle, et al., 2022]

Theoretical lower bounds for R_rob

- If assume unbounded shifts, robust risks would → ∞

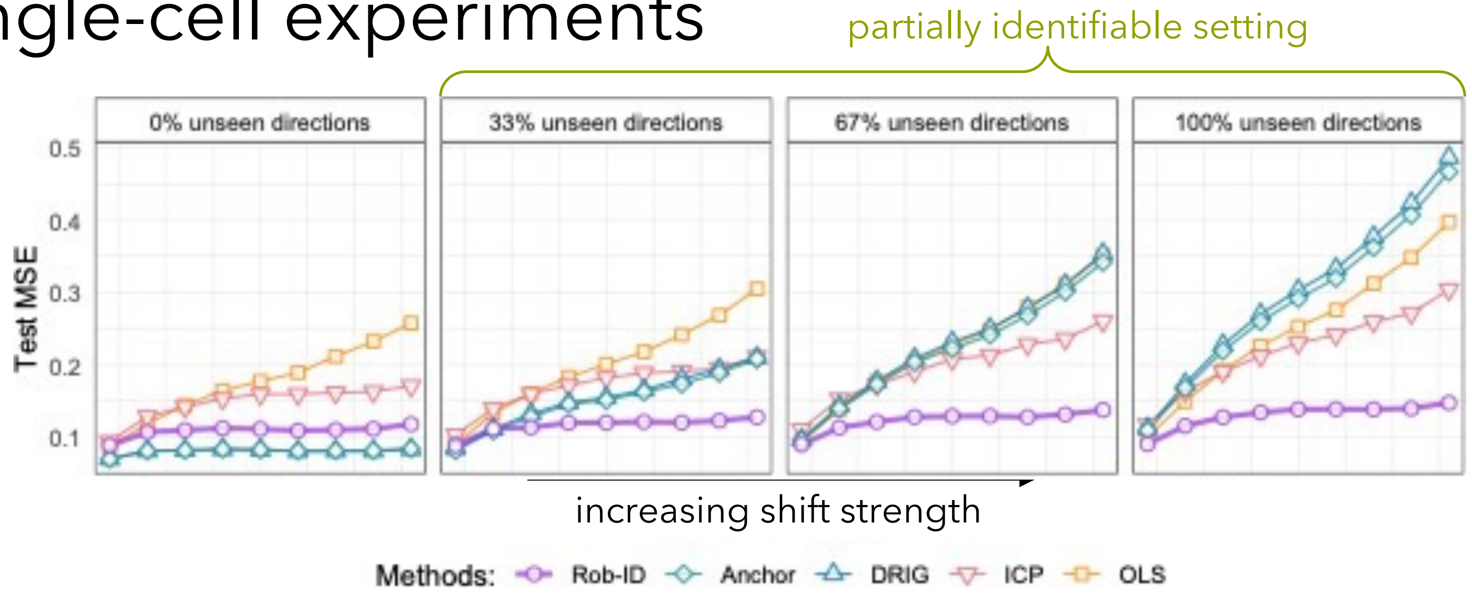
  unlikely and does not provide quantitative comparison of methods
- We finite test shifts: $\Delta_{\text{test}} = \gamma\Delta_{\text{tr}} + \gamma'\Delta_{\text{new}}$ (Rothenhaeusler $\gamma' = 0$)

$\widehat{\mathfrak{M}}_{\text{rob}}(\Delta_{\text{test}})$



$$\mathfrak{M}_{\text{rob}}(\Delta_{\text{test}}) = \inf_{\beta} \mathfrak{R}_{\text{rob}}(\beta, \Delta_{\text{test}})$$

Single-cell experiments



In partially identifiable cases (new test shift directions γ' large) anchor regression and OLS

- are far from optimal (slope in γ' can be smaller for the minimax quantity)
- are similar (term with γ' dominates) unless for very large γ

III: Primer on invariance-based methods

parameter varying with e

real-valued, i.e., regression

- Distributions: $(\theta_\star, \theta_e) \mapsto \mathbb{P}_e \sim (X, Y)$

invariant parameter

covariate vector

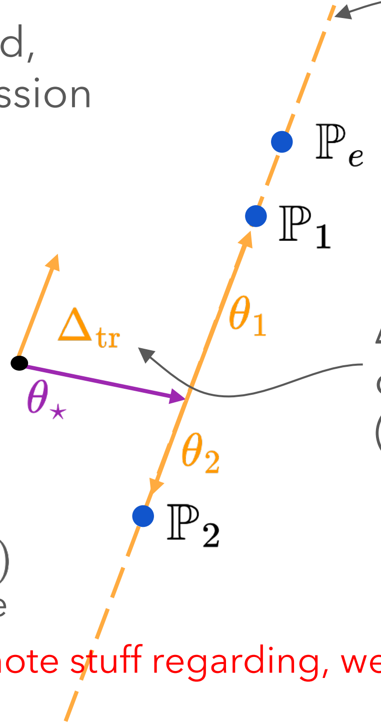
- Concrete example:

$$X = \theta_e + \eta$$
$$Y = X^\top \beta_\star + \xi$$

exogenous noise
$(\eta, \xi) \sim N(0, \Sigma_\star)$
invariant covariance

Invariant parameter $\theta_\star = (\beta_\star, \Sigma_\star)$

"Well-specified" Robust Set $\mathcal{P}_{\text{test}}(\theta_\star, \Delta_{\text{tr}})$



Δ_{tr} represents shift subspace of training environments (here a directional vector)

note stuff regarding, we allow confounding in Sigma!!

Disclaimer: diagram purely illustrative and informal, geometry should not be directly interpreted!

III: Primer on invariance-based methods

parameter varying with e

real-valued,
i.e., regression

- Distributions: $(\theta_\star, \theta_e) \mapsto \mathbb{P}_e \sim (X, Y)$

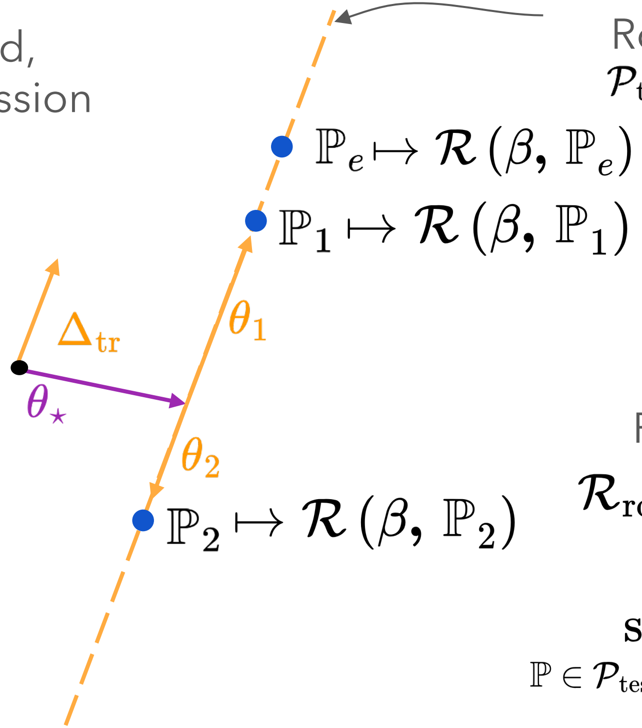
invariant parameter

covariate
vector

- Given a model $\beta : X \mapsto \widehat{Y}$

  the risk of β on possible environments

$$\mathcal{R}(\beta, \mathbb{P}_e) = \mathbb{E}_{\mathbb{P}_e}\left[\ell(\beta(X), Y)\right]$$



"Well-specified"
Robust Set
$\mathcal{P}_{\mathrm{test}}(\theta_\star, \Delta_{\mathrm{tr}})$

Robust risk

$$\mathcal{R}_{\mathrm{rob}}(\beta, \theta_\star, \Delta_{\mathrm{tr}}) = \sup_{\mathbb{P} \in \mathcal{P}_{\mathrm{test}}(\theta_\star, \Delta_{\mathrm{tr}})} \mathcal{R}(\beta, \mathbb{P})$$

Disclaimer: diagram purely illustrative and informal, geometry should not be directly interpreted!

Well-specified case: identifiable robust sets and risks

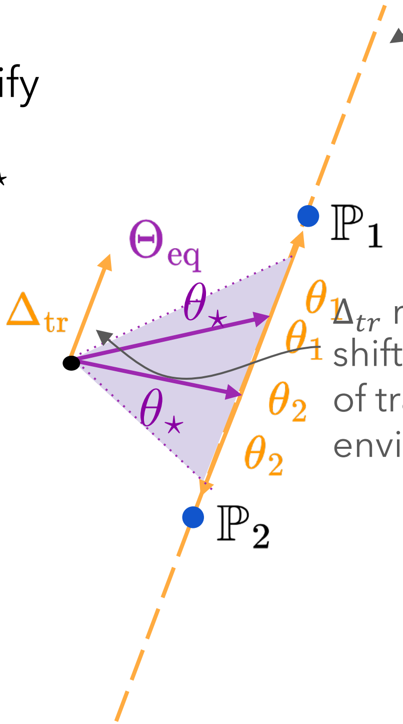
In general, we can only compute/identify

an equivalent set $\boldsymbol{\Theta_{eq}} \supset \boldsymbol{\theta_\star}$ instead of $\theta_\star$

Robust risk identifiable (prior work)

only when expected test shifts $\theta_e \in \Delta_{tr}$

align with train shifts (right)



Δ_{tr} represents shift subspace of training environments

Robust Set **identifiable**

Robust Set **identifiable**
for all $\tilde\theta \in \Theta_{eq}$
In general, we can only compute/identify
an equivalent set $\boldsymbol{\Theta_{eq}} \supset \boldsymbol{\theta_\star}$ instead of $\theta_\star$
Robust risk identifiable (prior work)
only when expected test shifts $\theta_e \in \Delta_{tr}$
align with train shifts (right)

for all $\tilde\theta \in \Theta_{eq}$

Robust Risk **identifiable**

$$\mathcal{R}_{\text{rob}}(\beta, \theta_\star, \Delta_{\text{tr}}) = \sup_{\mathbb{P} \in \mathcal{P}_{\text{test}}(\theta_\star, \Delta_{\text{tr}})} \mathcal{R}(\beta, \mathbb{P}) = \sup_{\mathbb{P} \in \mathcal{P}_{\text{test}}(\tilde\theta, \Delta_{\text{tr}})} \mathcal{R}(\beta, \mathbb{P}) = \mathcal{R}_{\text{rob}}\left(\beta, \tilde\theta, \Delta_{\text{tr}}\right)$$

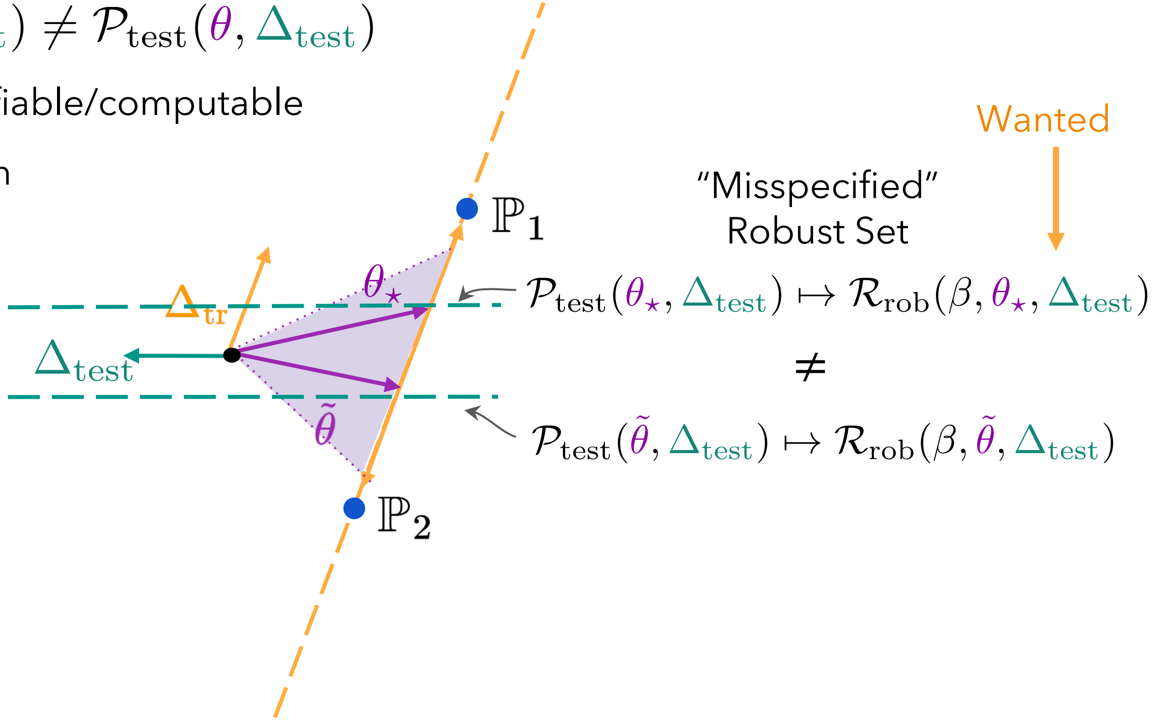
*limited range of Θ_{eq} if additionally assume $||\theta_\star||$ bounded

Our work: When robust risk is partially identifiable

Generally though, $\mathcal{P}_{\text{test}}(\theta_\star, \Delta_{\text{test}}) \neq \mathcal{P}_{\text{test}}(\tilde{\theta}, \Delta_{\text{test}})$

for $\tilde{\theta} \in \Theta_{eq}$ → robust risk non-identifiable/computable

because $\Delta_{new} \neq \Delta_{tr}$ and $\theta_\star$ unknown

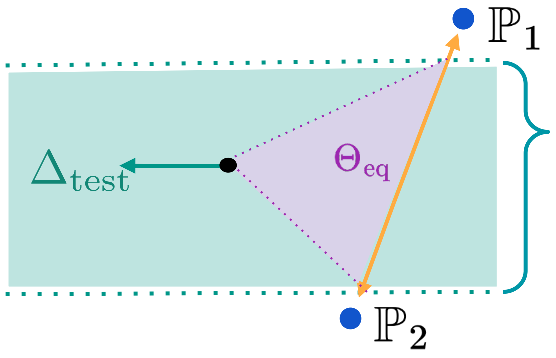


Our work: When robust risk is partially identifiable

Generally though, $\mathcal{P}_{\text{test}}(\theta_\star, \Delta_{\text{test}}) \neq \mathcal{P}_{\text{test}}(\tilde{\theta}, \Delta_{\text{test}})$

for $\tilde{\theta} \in \Theta_{eq}$ → robust risk non-identifiable/computable

because $\Delta_{new} \neq \Delta_{tr}$ and $\theta_\star$ unknown



Set of robust sets

$\{\mathcal{P}_{\text{test}}(\theta_\star, \Delta_{\text{test}})\}_{\theta\in\Theta_{\text{eq}}}$

Robustness of a model β can then be measured using set of all robust risks induced by $\theta \in \Theta_{eq}$:

- Output whole set / interval
- Output the **worst-case robust risk**

$$\mathfrak{R}_{\text{rob}}(\beta, \Delta_{\text{test}}) = \sup_{\theta\in\Theta_{\text{eq}}} \mathcal{R}_{\text{rob}}(\beta, \theta, \Delta_{\text{test}})$$

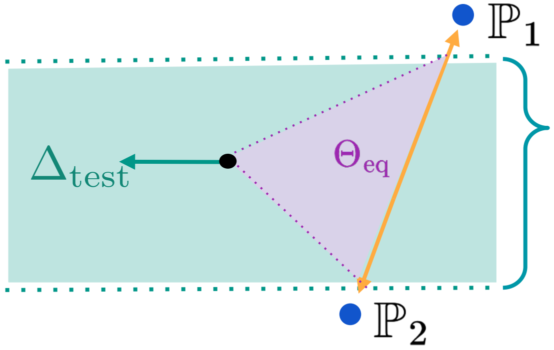
Set of robust risks

$\{\mathcal{R}_{\text{rob}}(\beta, \theta, \Delta_{\text{test}})\}_{\theta\in\Theta_{\text{eq}}}$

What we can do with the worst-case robust risk

- Can define a notion of achievable robustness (by any method/penalty/algorithm)

$$\mathfrak{M}_{\mathrm{rob}}(\Delta_{\mathrm{test}}) = \inf_{\beta} \underbrace{\mathfrak{R}_{\mathrm{rob}}(\beta, \Delta_{\mathrm{test}})}_{\sup_{\theta \in \Theta_{\mathrm{eq}}} \mathcal{R}_{\mathrm{rob}}(\beta, \theta, \Delta_{\mathrm{test}})}$$



Set of robust sets

$$\{\mathcal{P}_{\mathrm{test}}(\theta_{\star}, \Delta_{\mathrm{test}})\}_{\theta \in \Theta_{\mathrm{eq}}}$$

Set of robust risks

$$\{\mathcal{R}_{\mathrm{rob}}(\beta, \theta, \Delta_{\mathrm{test}})\}_{\theta \in \Theta_{\mathrm{eq}}}$$

- Can evaluate it on existing algorithms and how close they are to this lower bound
- Can estimate (minimizer achieving) $\inf_{\beta} \mathfrak{R}_{\mathrm{rob}}(\beta, \Delta_{\mathrm{test}})$

Summary

Framework for partially identifiable distribution shift, beyond identifiable/non-identifiable dichotomy

Defined a measure of achievable robustness, grounded in worst-case performance across compatible models

Computed achievable robust risk in a linear setting and showed existing methods (robust or not) can behave similarly

J. Kostin, N. Gnecco, F. Yang "Achievable distributional robustness when the robust risk is only partially identified", NeurIPS 2024