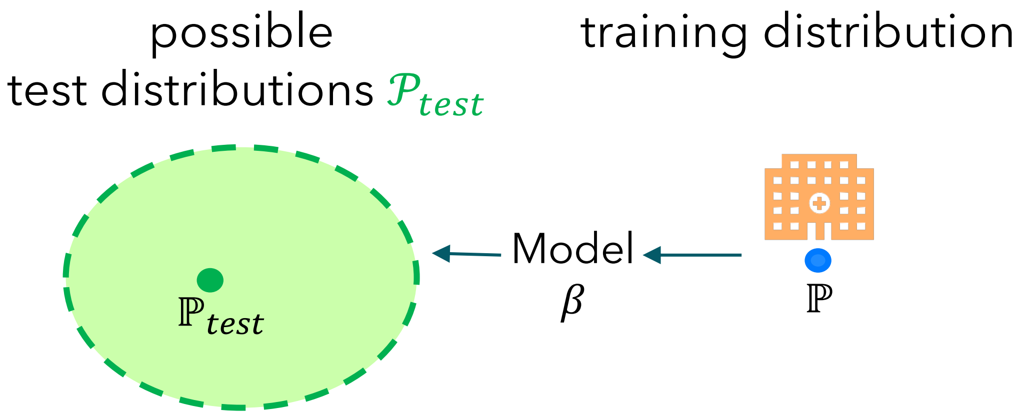
Robust prediction beyond the identifiable case

Oberwolfach Workshop January 2025
on Overparameterization, **Regularization, Identifiability** and Uncertainty

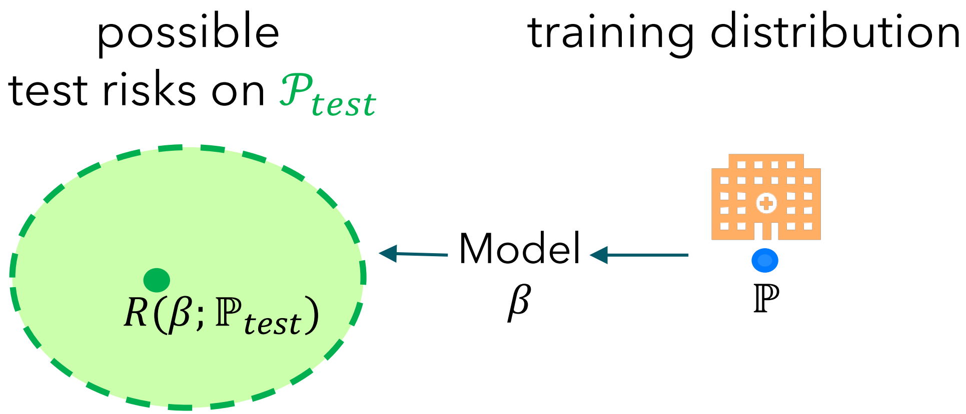
Fanny Yang, joint with Julia Kostin, Nicola Gnecco

Statistical Machine Learning group,
Department of Computer Science, ETH Zurich

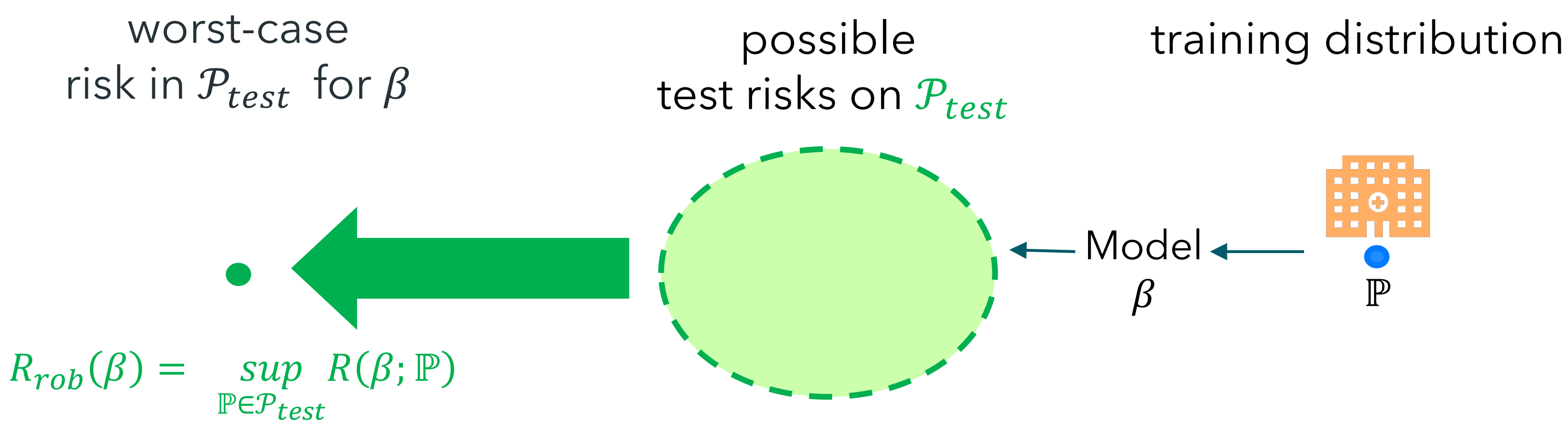

Robust prediction for safety purposes



Robust prediction for safety purposes

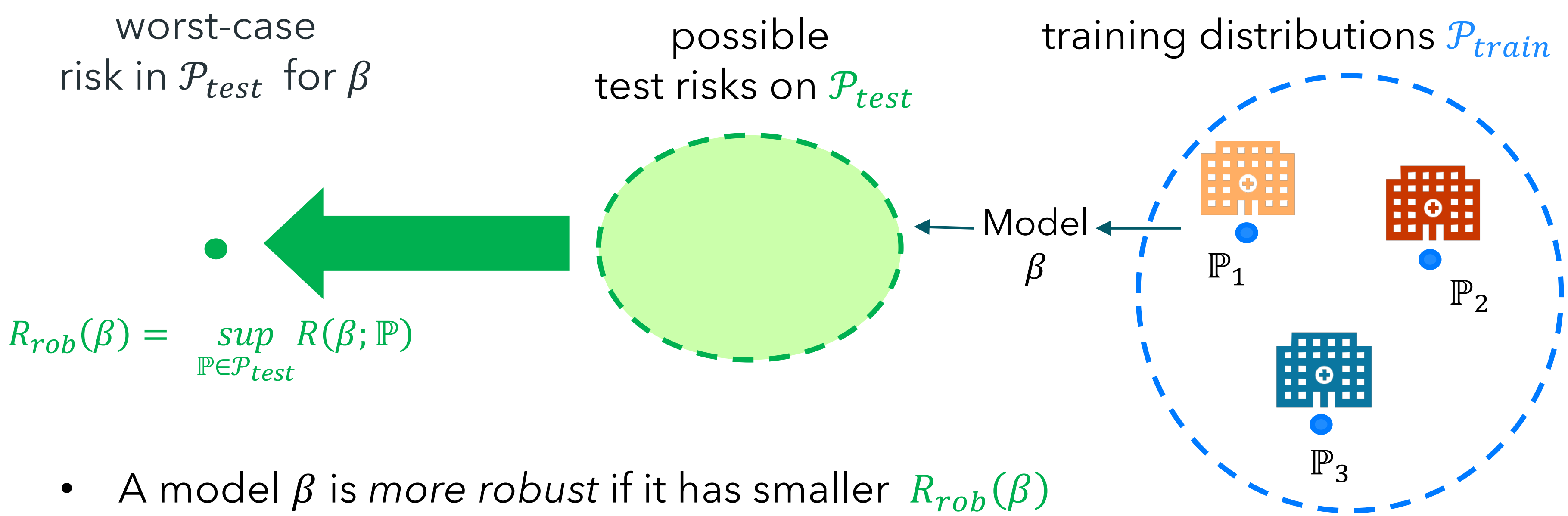


Robust prediction for safety purposes



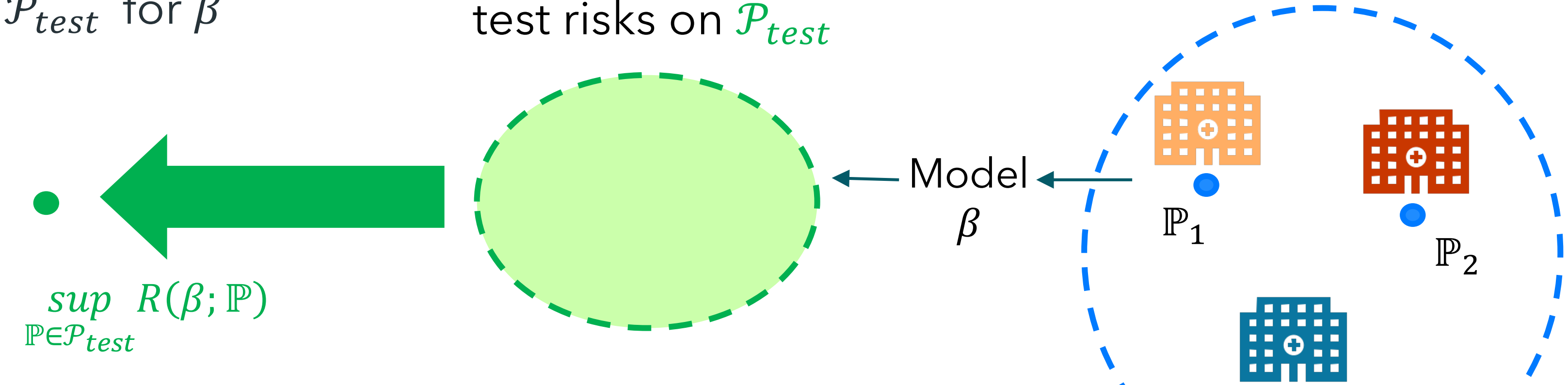
- A model β is *more robust* if it has smaller $R_{rob}(\beta)$

Robust prediction for safety purposes



- A model β is *more robust* if it has smaller $R_{rob}(\beta)$
- Any robustness gains from observing multiple heterogeneous training distributions?

Robustness analysis of methods – what's missing?



$$R_{rob}(\beta) = \sup_{\mathbb{P} \in \mathcal{P}_{test}} R(\beta; \mathbb{P})$$

So far: Given a method returning β, for known $\mathcal{P}_{test}$, how large $R_{rob}(\beta)$?

Problem: $\mathcal{P}_{test}$, hence $R_{rob}(\cdot)$, usually not fully known (*partially identifiable*)

Neglected question: Given partial knowledge about $\mathcal{P}_{test}$,

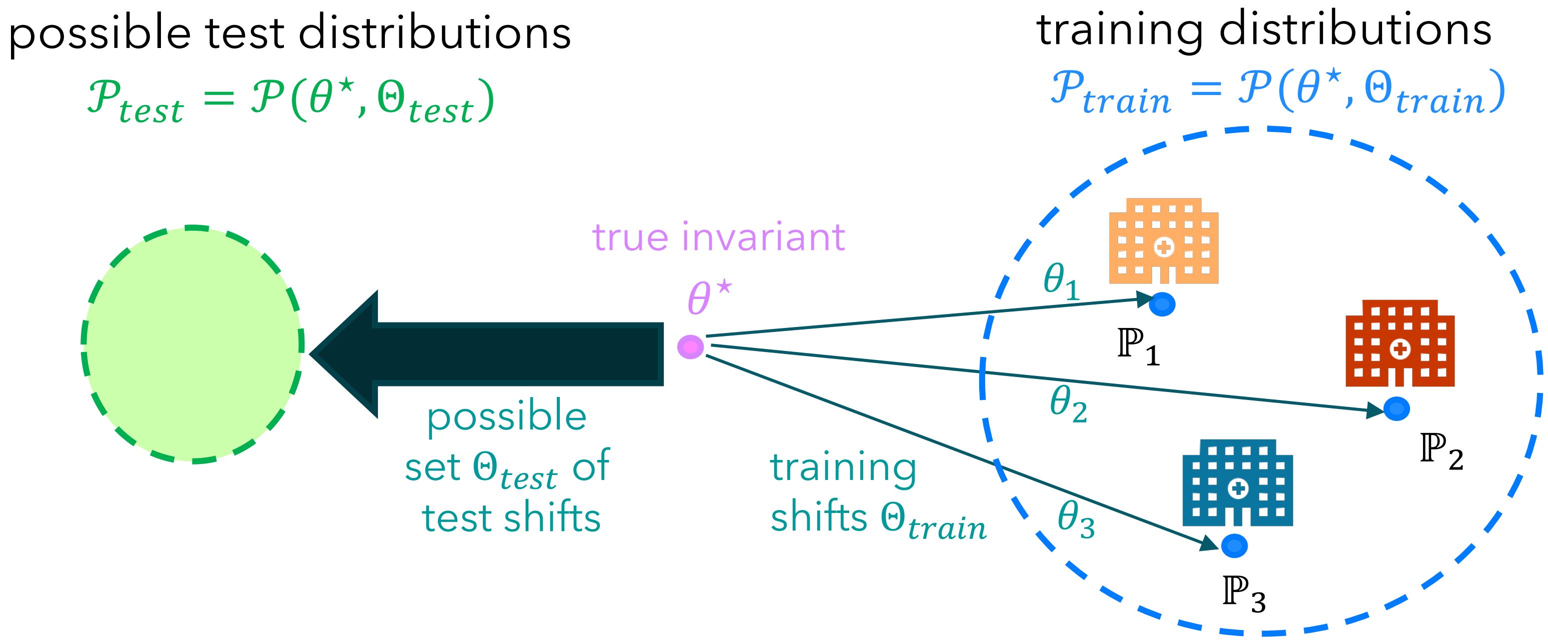
- how robust can *any algorithm* be, i.e. what is the "information-theoretic" (population) limit?
- how well do *existing algorithms perform*, and how close to optimal/adaptive are they?

our work

How can we model partial knowledge of $\mathcal{P}_{test}$/ R_{rob} via its relationship to $\mathcal{P}_{train}$?

- Setting up unified shift robustness view via invariance (+ one example)
- From fully identifiable (prior work) to partially identifiable (our work) R_{rob}
- Measure of robustness and hardness in partially identifiable case

Unified view of shift robustness using invariance



- Assume that $(\theta^\star, \theta_e)$ parameterize distributions $\mathbb{P}_e$ with $\theta^\star$ invariant and θ_e varying with e
- Viewpoint includes traditional shift concepts (covariate shift, spurious correlations, domain mixtures, neighborhood) & causality-based ones (IRM-related or next slide)

Imagine simple linear example for concreteness…

Assume that joint distributions in each "environment" e in train and test environments are defined by

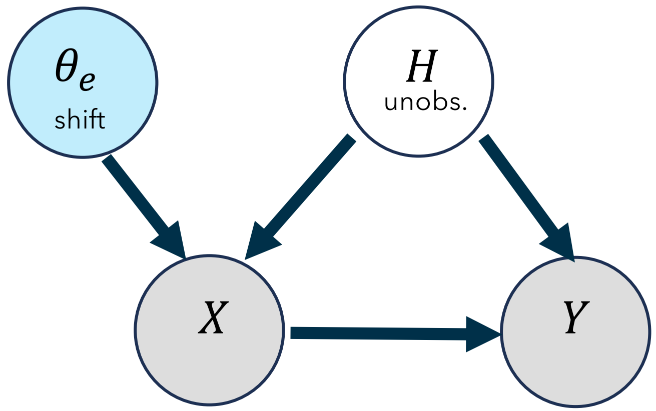
mean shifts varying with e
(assume ref. env has $\theta_e = 0$)

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

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

with invariant $\theta_\star = (\beta_\star, \Sigma_\star)$ same across environments

Possible underlying causal model (most simplified version)

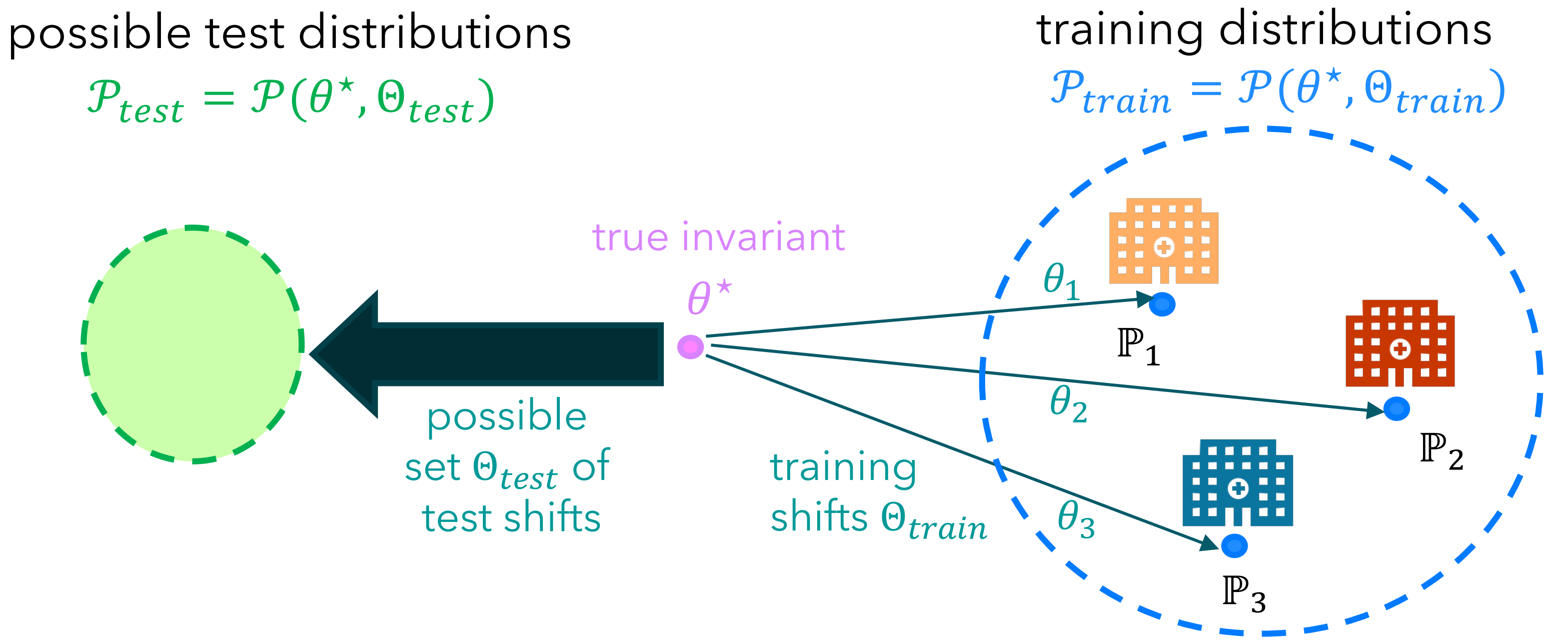


We allow cross-covariance $\Sigma_{\star,\eta\xi} \neq 0$ corresponding to confounding

$\Longrightarrow$ allows not only covariate shift, but also shift in $\mathbb{E}[Y|X]$!

*our calculations also incorporate covariance shifts and beyond Gaussians

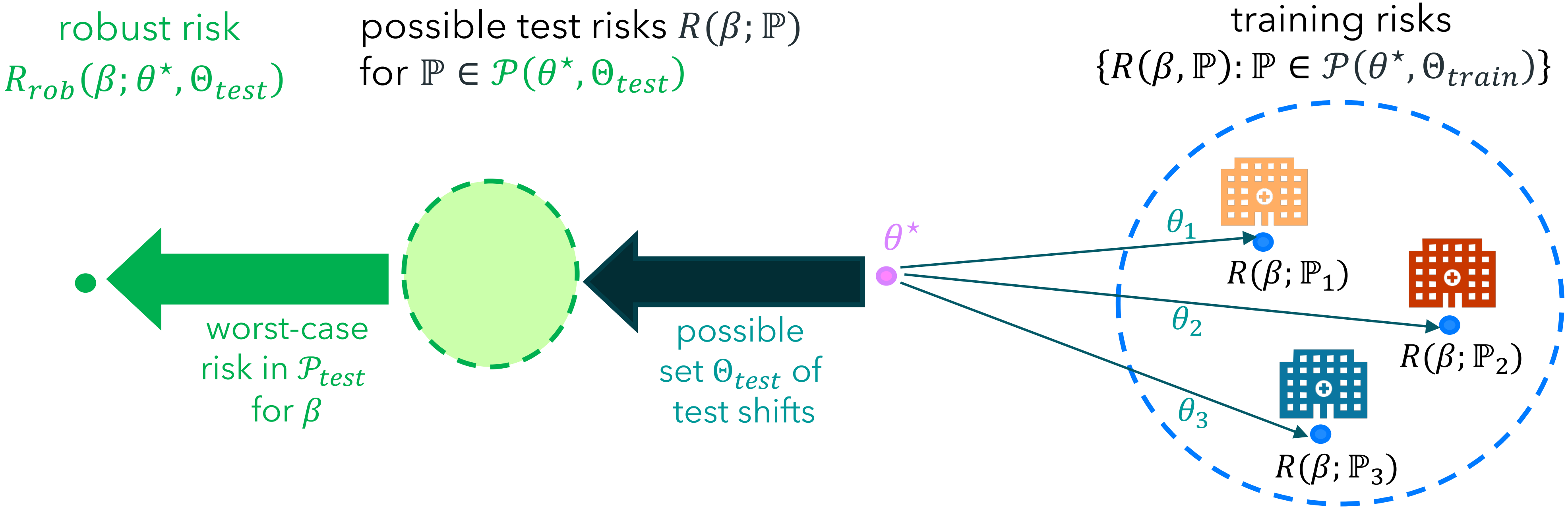
Unified view of shift robustness using invariance



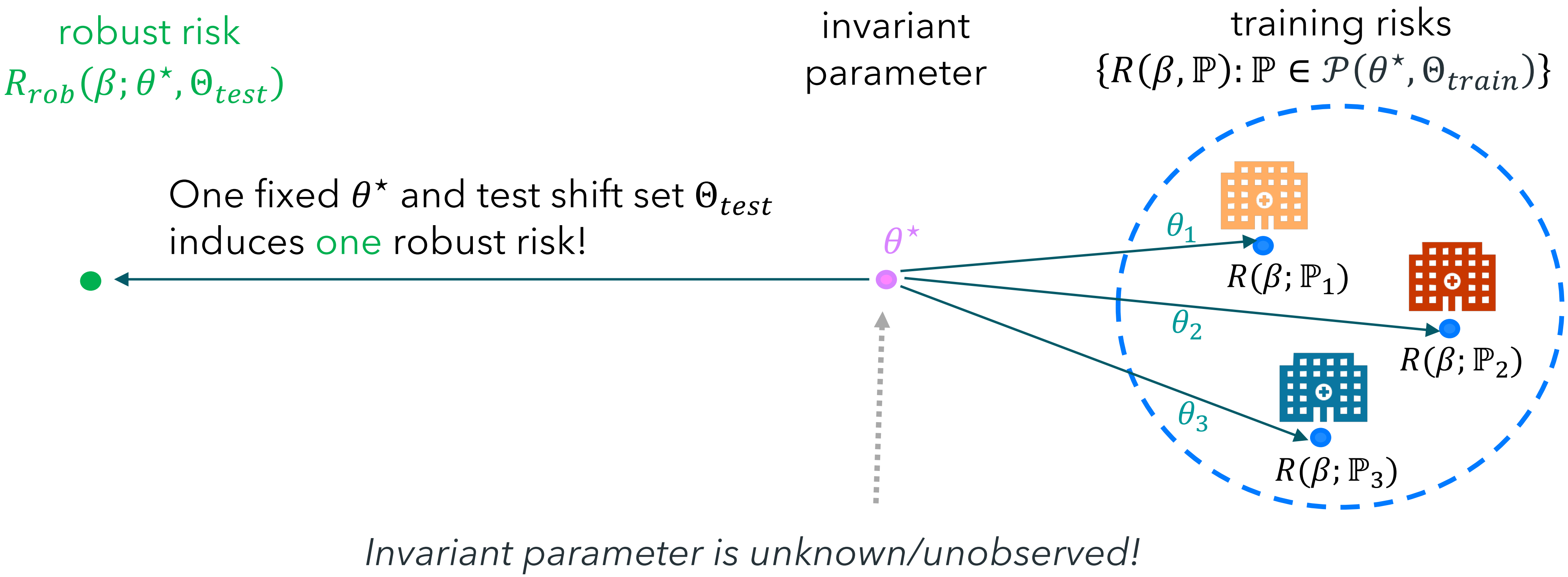
Remember: we're interested in answering, given some invariance assumption & any $\Theta_{\text{test}}, \Theta_{train}$

- how robust can *any algorithm* be, i.e. what is the "information-theoretic" (population) limit?
- how do *existing algorithms perform*, and how close to optimal/adaptive are they?

Measuring robustness via robust risk



Measuring robustness via robust risk



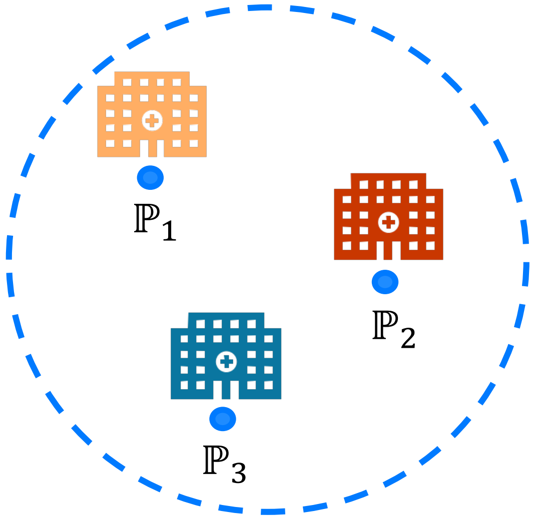
Prior work: Assuming identifiable robust risk

robust risk
$R_{rob}(\beta; \theta^\star, \Theta_{test})$

training distributions
$\mathcal{P}_{train} = \mathcal{P}(\theta^\star, \Theta_{train})$

Robust risk identifiable, i.e. computable

using observed $\mathcal{P}_{train}, \Theta_{train}$ and Θ_{test}?

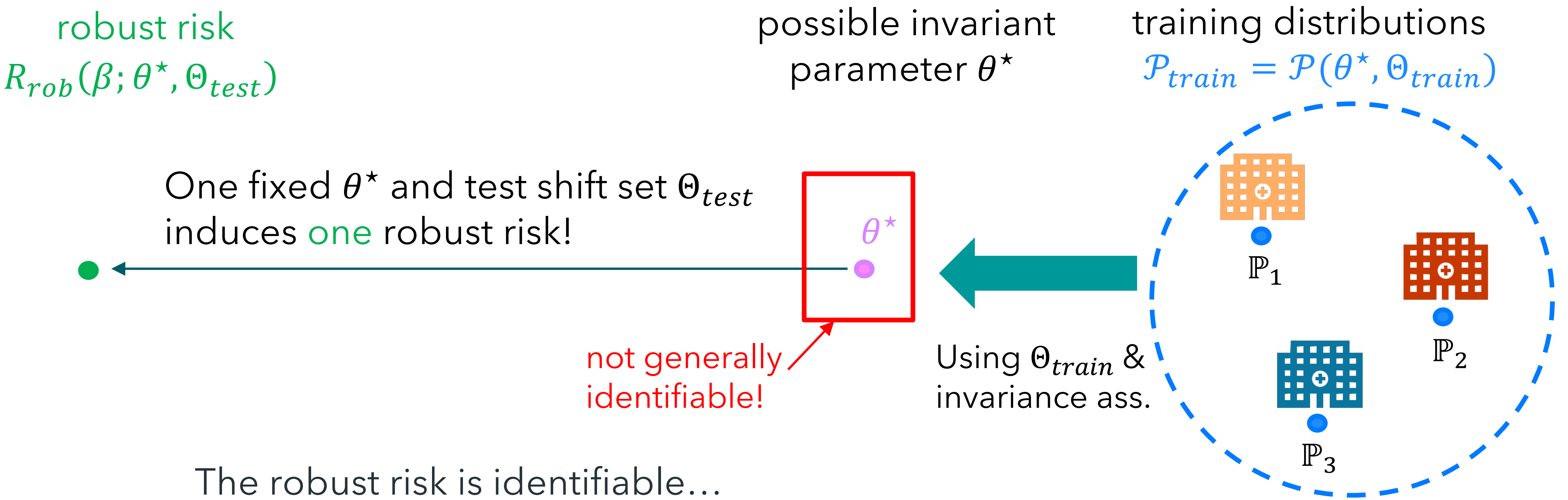


Previous work on robustness only considers identifiable case

- for invariance-based shift models this only holds for specific combinations of Θ_{test}, Θ_{train}
- any other combination naturally corresponds to some kind of partial knowledge of $\mathcal{P}_{test}$

(always true for DRO/mixture of domains with no $\theta^\star$ such as in e.g. Mansour et al. '08, Sagawa et al. '19)

Prior work I: Identifiable invariant mechanism $\theta^\star$

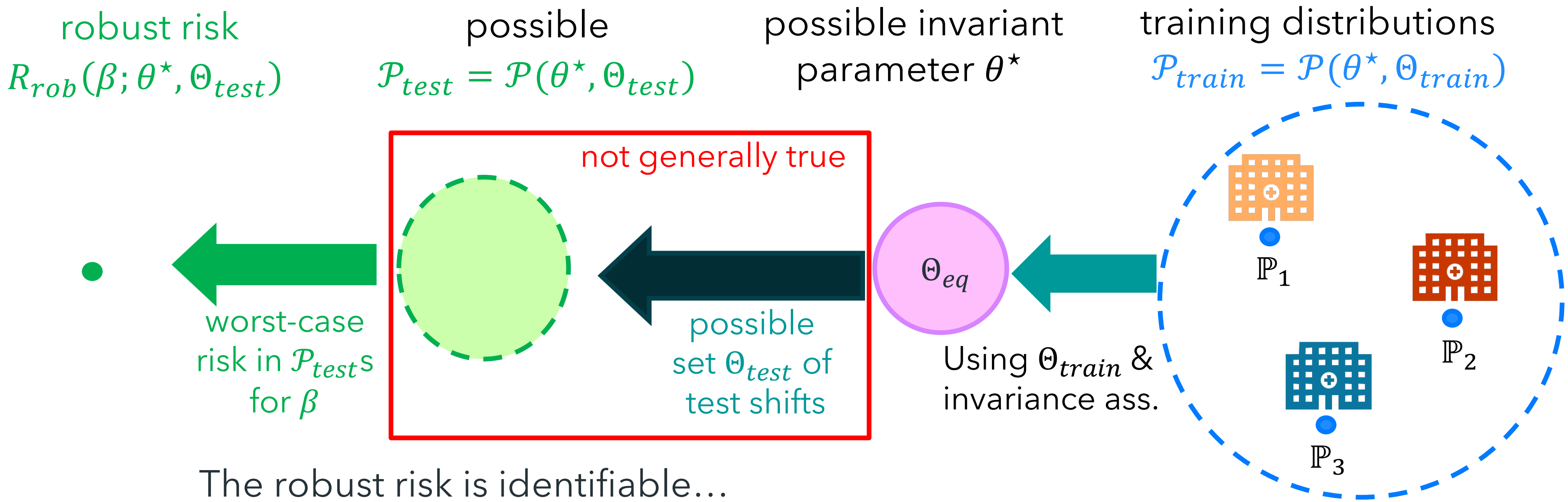


The robust risk is identifiable…

- when Θ_{train} is heterogeneous enough to identify $\theta^\star$

e.g. Peters et al. '16, Rojas-Carulla et al. '18, Arjovsky et al. '19, Krueger '20

Prior work II: Only identifiable robust risk

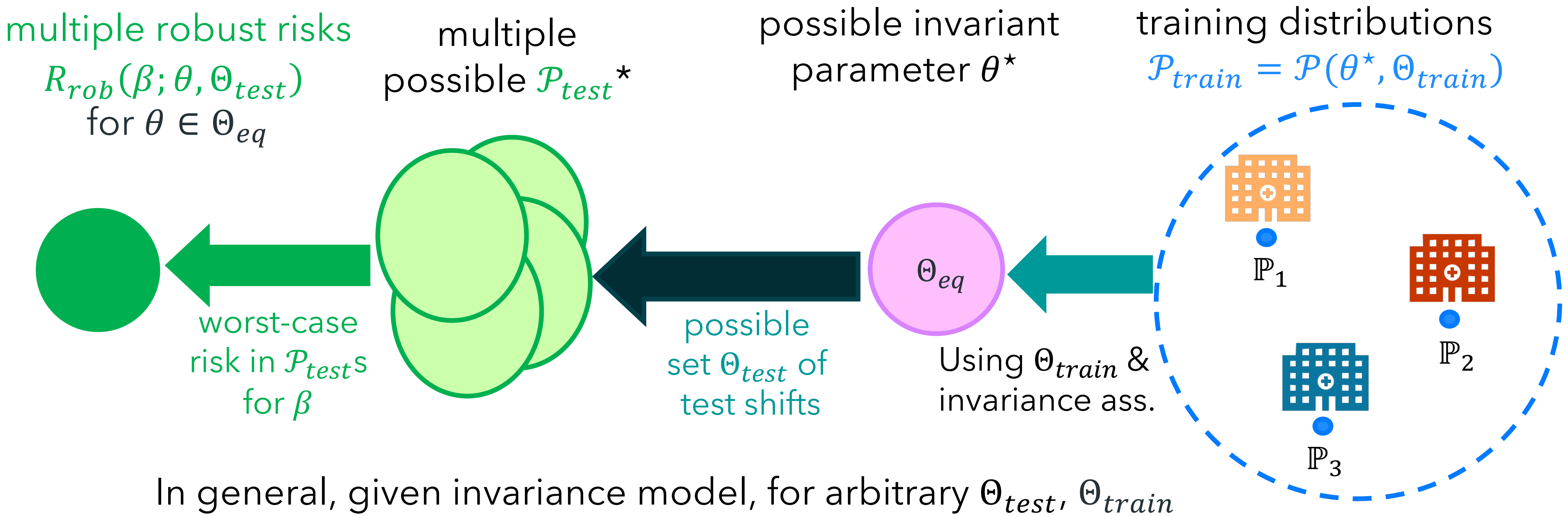


The robust risk is identifiable…

- when Θ_{train} is heterogeneous enough to identify $\theta^\star$
- when Θ_{test} similar to Θ_{train} if one can't identify $\theta^\star$

e.g. Rothenhaeusler et al. '21, Shen et al. '23

Our work: general partially identifiable robust risk



In general, given invariance model, for arbitrary Θ_{test}, Θ_{train}

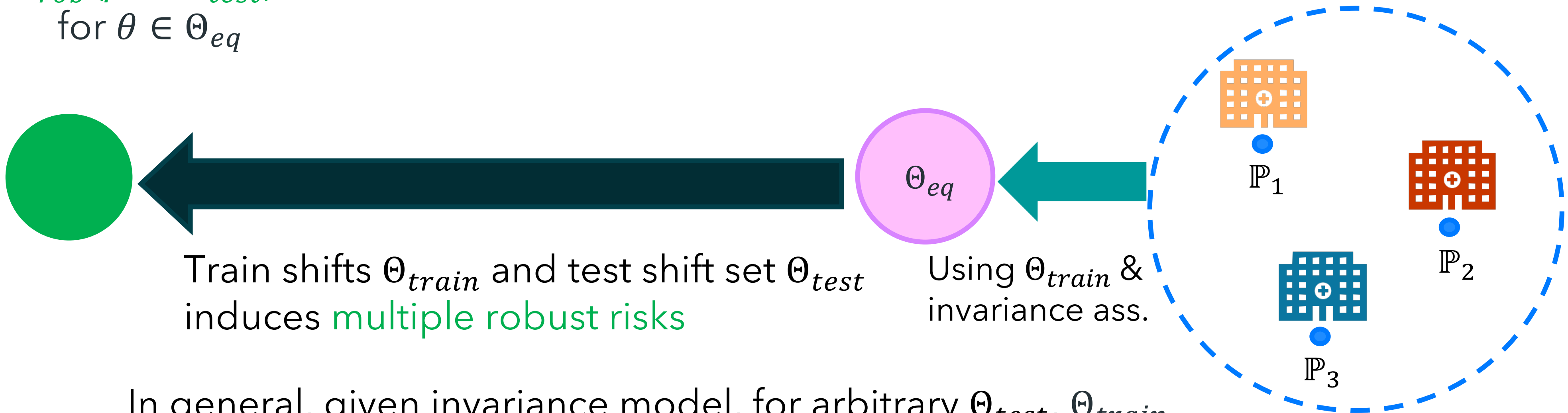
we end up only with partially/set-identifying the robust risk!

Our work: general partially identifiable robust risk

multiple robust risks
$R_{rob}(\beta; \theta, \Theta_{test})$
for $\theta \in \Theta_{eq}$

possible invariant parameter $\theta^\star$

training distributions
$\mathcal{P}_{train} = \mathcal{P}(\theta^\star, \Theta_{train})$



Train shifts Θ_{train} and test shift set Θ_{test} induces multiple robust risks

Using Θ_{train} & invariance ass.

In general, given invariance model, for arbitrary Θ_{test}, Θ_{train} we end up only with partially/set-identifying the robust risk!

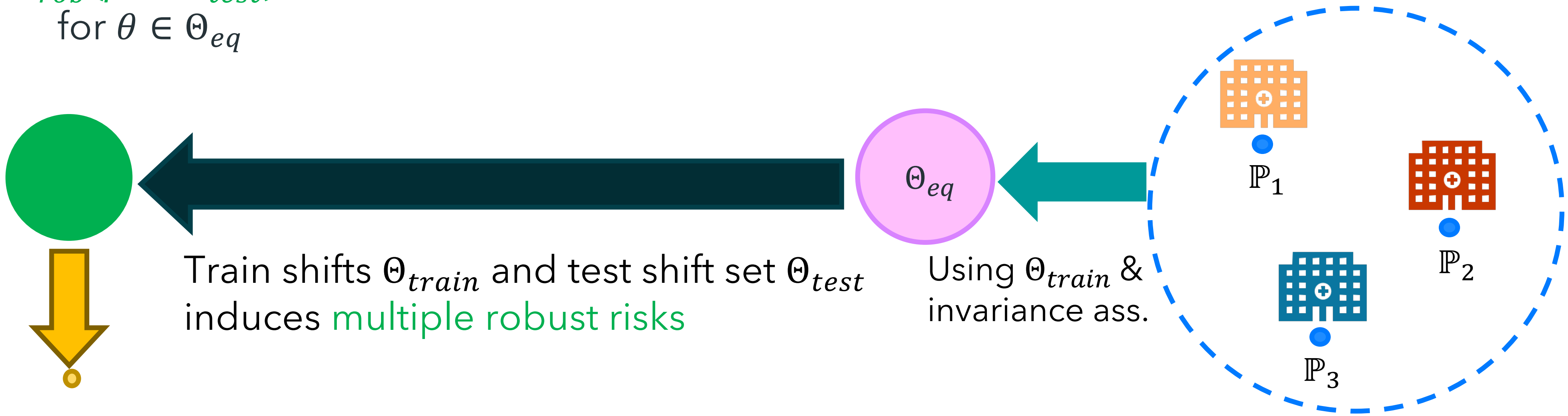
How do we even measure robustness in this case?

Quantifying robustness in partial identifiable setting

multiple robust risks
$R_{rob}(\beta;\theta,\Theta_{test})$
for $\theta \in \Theta_{eq}$

possible invariant parameter $\theta^\star$

training distributions
$\mathcal{P}_{train} = \mathcal{P}(\theta^\star, \Theta_{train})$



Train shifts Θ_{train} and test shift set Θ_{test} induces multiple robust risks

Using Θ_{train} & invariance ass.

worst-case robust risk
$\mathfrak{R}_{rob}(\beta;\Theta_{train},\Theta_{test}) = \max_{\theta\in\Theta_{eq}} R_{rob}(\beta;\theta,\Theta_{test})$

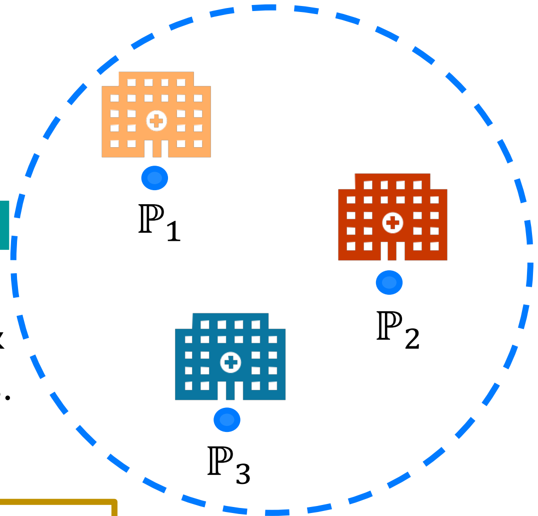
…want small robust risk even for the hardest true $\theta \in \Theta_{eq}$ that could have induced $\mathcal{P}_{train}$

Achievable robustness in partial identifiable setting

multiple robust risks
$R_{rob}(\beta; \theta, \Theta_{test})$
for $\theta \in \Theta_{eq}$

possible invariant parameter $\theta^\star$

training distributions
$\mathcal{P}_{train} = \mathcal{P}(\theta^\star, \Theta_{train})$



Θ_{eq}

Using Θ_{train} & invariance ass.

Train shifts Θ_{train} and test shift set Θ_{test} induces multiple robust risks

worst-case robust risk
$\mathfrak{R}_{rob}(\beta; \Theta_{train}, \Theta_{test}) = \max_{\theta \in \Theta_{eq}} R_{rob}(\beta; \theta, \Theta_{test})$

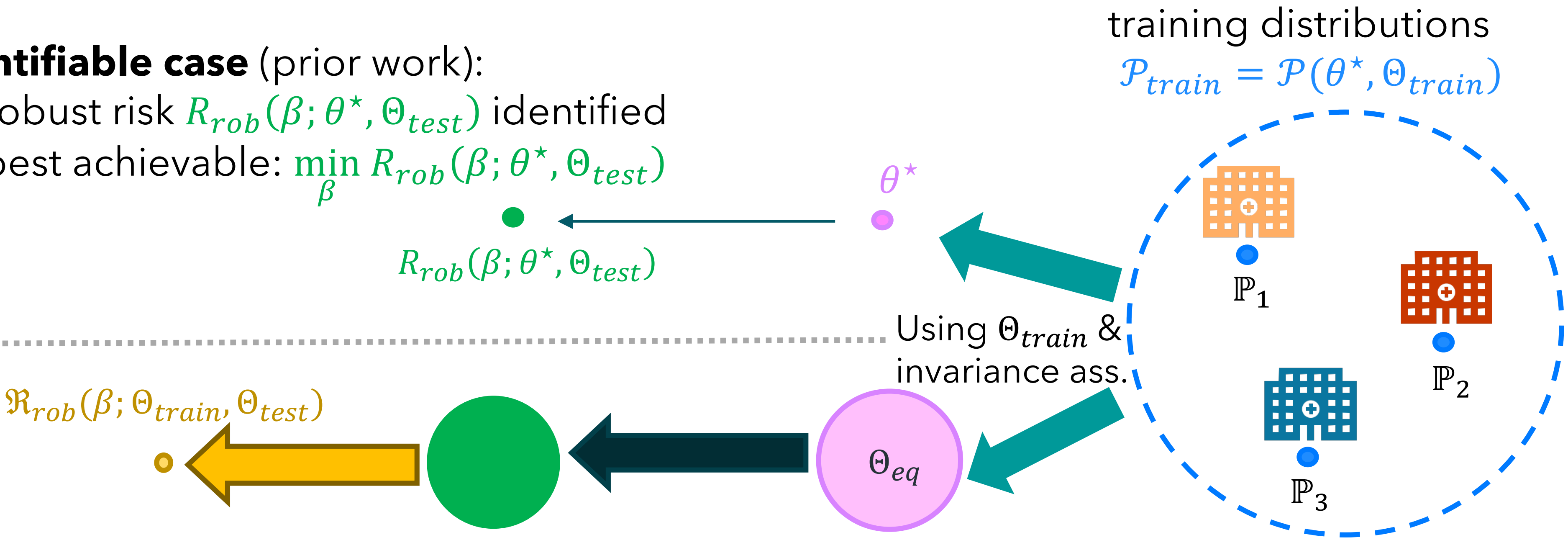
and achievable worst-case robust risk
$\mathfrak{M}(\Theta_{train}, \Theta_{test}) = \min_{\beta} \mathfrak{R}_{rob}(\beta; \Theta_{train}, \Theta_{test})$
$= \min_{\beta} \max_{\theta \in \Theta_{eq}} R_{rob}(\beta; \theta, \Theta_{test})$

allow us to quantify robustness in the non-identifiable case

Summary of differences in identifiability

Identifiable case (prior work):

- robust risk $R_{rob}(\beta; \theta^\star, \Theta_{test})$ identified
- best achievable: $\min_\beta R_{rob}(\beta; \theta^\star, \Theta_{test})$



Partially identifiable case (ours):

- only worst-case robust risk $\mathfrak{R}_{rob}(\beta; \Theta_{train}, \Theta_{test}) = \max_{\theta \in \Theta_{eq}} R_{rob}(\beta; \theta, \Theta_{test})$ identified
- best-achievable: $\mathfrak{M}(\Theta_{train}, \Theta_{test}) = \min_\beta \mathfrak{R}_{rob}(\beta; \Theta_{train}, \Theta_{test})$

Remember: we're interested in answering, given some invariance assumption & any $\Theta_{\text{test}}, \Theta_{train}$

- how robust can *any algorithm* be, i.e. what is the "information-theoretic" (population) limit?
- how do *existing algorithms perform*, and how close to optimal/adaptive are they?

⇩

We quantify for invariance-based methods in this general setting

- the best achievable robustness $\mathfrak{M}(\Theta_{train}, \Theta_{test})$
- how ranking of different methods wrt $\mathfrak{R}_{rob}(\beta; \Theta_{train}, \Theta_{test})$ changes drastically with varying $\Theta_{test}, \Theta_{train}$

theoretically for linear model
empirically for real data

Simple linear example for concreteness

Assume that joint distributions in each "environment" e in train and test environments are defined by

mean shifts varying with e (assume ref. env has $\theta_e = 0$)

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

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

with invariant $\theta_\star = (\beta_\star, \Sigma_\star)$ same across environments

Test time shifts assumptions

M_{seen}: covariance with range in span of seen shift directions

$range(M_{seen}) \subset span\,\{\theta_e\}_{e\in[k]}$

M_{unseen}: projection matrix onto unseen directions with

$range(M_{seen}) \perp span\,\{\theta_e\}_{e\in[k]}$

Mean shifts during test time assumed to lie in $\Theta_{test} = \{\theta_{test} \colon \theta_{test}\theta_{test}^\top \preccurlyeq \gamma M_{seen} + \gamma' M_{unseen}\}$

shift strengths

$X^e = A^e + \eta$
$Y^e = \beta_\star^\top X^e + \xi$
$\Theta_{test} = \{\theta_{test}: \theta_{test}\theta_{test}^\top \preccurlyeq \gamma M_{seen} + \gamma' M_{unseen}\}$

Achievable and achieved robustness

Trends concluded from formal statement

In partially identifiable case & new test shift directions γ' large, anchor regression and OLS

- are far from achievable robustness $\mathfrak{M}(\Theta_{train}, \Theta_{test}) = \min_\beta \mathfrak{R}_{rob}(\beta; \Theta_{eq}, \Theta_{test})$
- have similar linear robustness when term with unseen directions γ' dominates

Corollary [KGY' 24] (informal) – Performance comparison in the partially identifiable setting

For large γ' fixed γ, $\mathfrak{M}(\Theta_{train}, \Theta_{test}) = \min_\beta \mathfrak{R}_{rob}(\beta; \Theta_{eq}, \Theta_{test}) = C^2\gamma' + c_1$

vs. Anchor regression*: $\mathfrak{R}_{rob}(\beta_{anchor}; \Theta_{train}, \Theta_{test}) = \left(C + h(\gamma)\right)^2 \gamma' + c_2$

vs. Ordinary least squares: $\mathfrak{R}_{rob}(\beta_{OLS}; \Theta_{train}, \Theta_{test}) = \left(C + h(1)\right)^2 \gamma' + c_3$

(h is decreasing function, c are constant in γ')

* Rothenhaeusler et al. '21

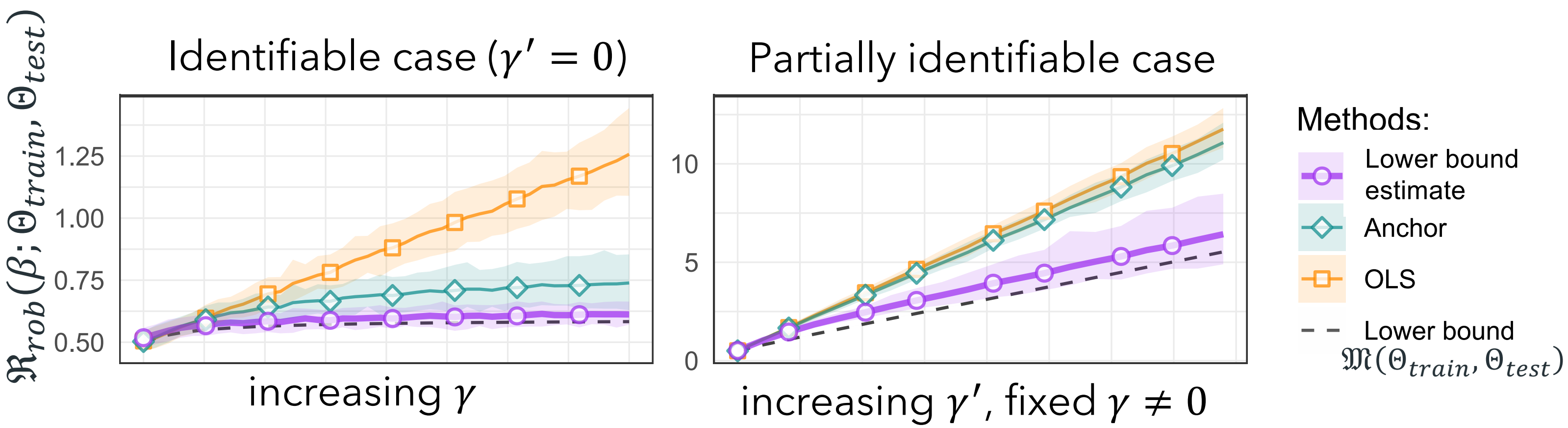
$X^e = A^e + \eta$
$Y^e = \beta_\star^\top X^e + \xi$
$\Theta_{test} = \{\theta_{test} \colon \theta_{test}\theta_{test}^\top \preccurlyeq \gamma M_{seen} + \gamma' M_{unseen}\}$

Experimental comparison in linear setting

Trends concluded from formal statement

In partially identifiable case & new test shift directions γ' large, anchor regression and OLS

- are far from achievable robustness $\mathfrak{M}(\Theta_{train}, \Theta_{test}) = \min_\beta \mathfrak{R}_{rob}(\beta; \Theta_{eq}, \Theta_{test})$
- have similar linear robustness when term with unseen directions γ' dominates



- Using correct Θ_{test}
- Average over random draws of $\theta^\star$

Real-world data setting

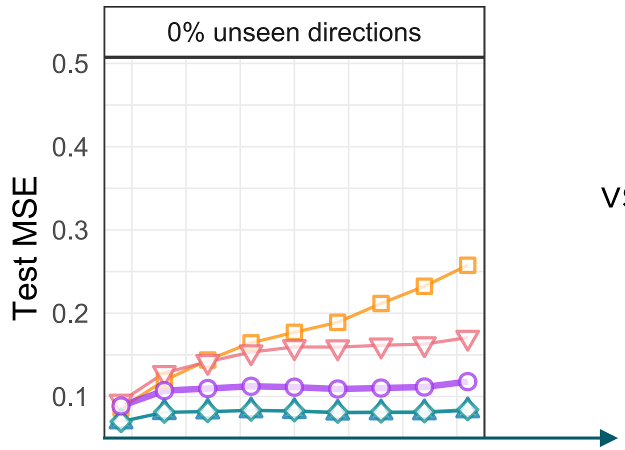
Single–cell gene expression dataset with

- prediction task: expression of one gene (target) as a function of expressions of 3 others
- per target, 3 different environments ($\triangleq$ individual gene knocked out) + observational
  - training environments: observational + 1 knocked-out environment
  - shift strengths γ, γ' $\triangleq$ distance of covariates to mean in observational environment
  - partially identified setting: test data also includes some percentage of samples from knock-out environments not seen during training
- Results are averaged across these scenarios

*[Replogle, et al., 2022]

Experimental comparison in real-world setting

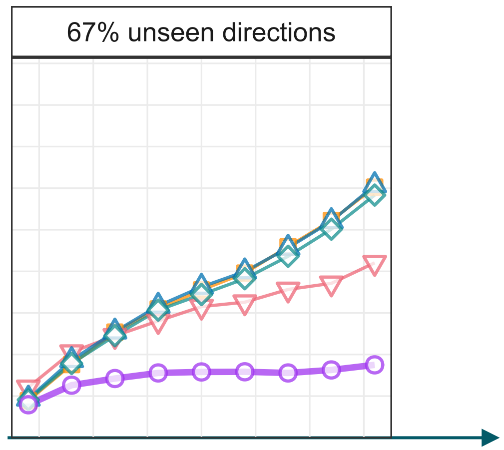
identifiable case ($\gamma' = 0$)



increasing shift strength γ

vs.

partially identifiable case ($\gamma' \neq 0$)



increasing shift strength $\gamma = \gamma'$

Test set here consists of:

- 67% from knock-outs unseen during training
- 33% held-out data from seen knock-out env

Different invariance-based Methods: Rob-ID Anchor DRIG ICP OLS

$\underset{\beta}{\text{argmin}}\ \widehat{\Re}_{rob}(\beta; \Theta_{train}, \Theta_{test})$ assuming previous synthetic setting where M_{unseen} is most conservatively chosen to be entire $\left(span\ \{\theta_e\}_{e\in[k]}\right)^{\perp}$ (no reason to do well!)

Summary

How analyze the more general partially identifiable setting (vs. focusing on identifiable vs. non-identifiable)

- introduced measures of (achievable) robustness
- computed them for a linear example and compared achievable robustness with prior methods

Future work

- Apply on other types of invariant mechanisms (see e.g. Francesco's, Arthur's talk, and beyond causality)
- Use achievable robustness for active selection of training distributions

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