Transfer learning with benchmarks and causal invariance

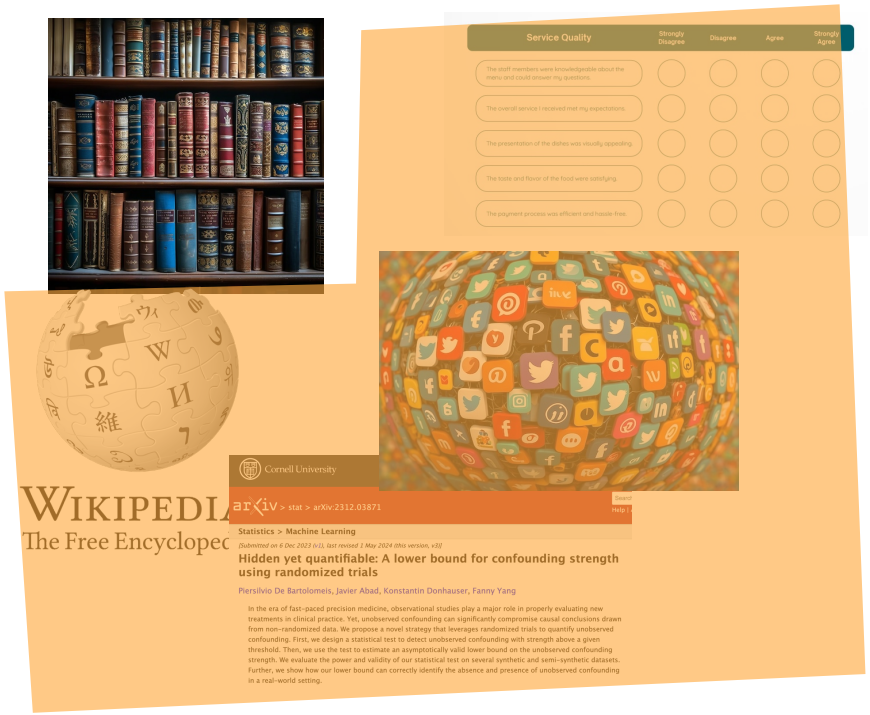
Fanny Yang, Department of Computer Science, ETH Zurich

Statistical Machine Learning group

Distributions shift – a challenge and blessing?

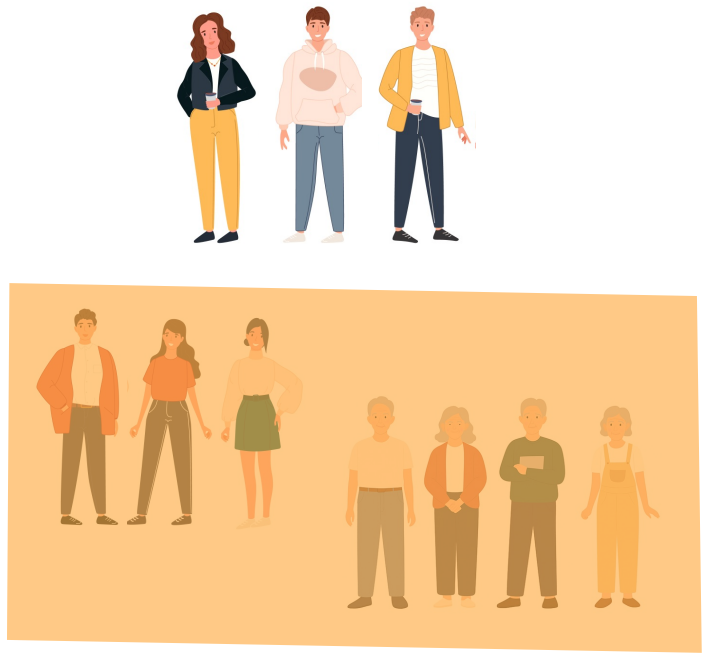
Transfer learning (supervised domain adaptation): Test on **new domain** given only little data

Different text corpuses



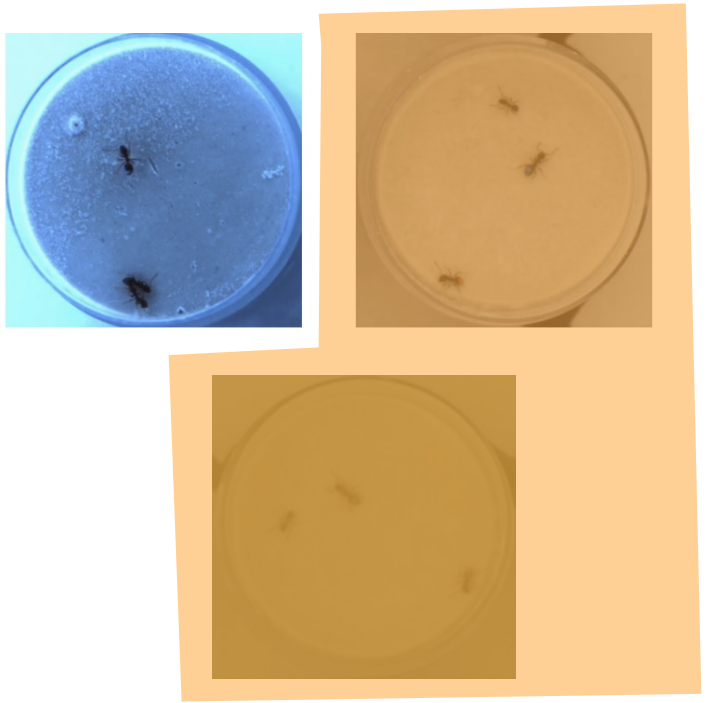
Text → Next token

Different populations



Blood measurements → Disease

Different experimental conditions



Video → Grooming behavior

Heterogeneous data – a challenge and blessing?

Transfer learning (supervised domain adaptation): Test on **new domain** given only little data

Challenge!

Goal: Inference / prediction on new *target task/risk* R_Q (not equal to any R_i)

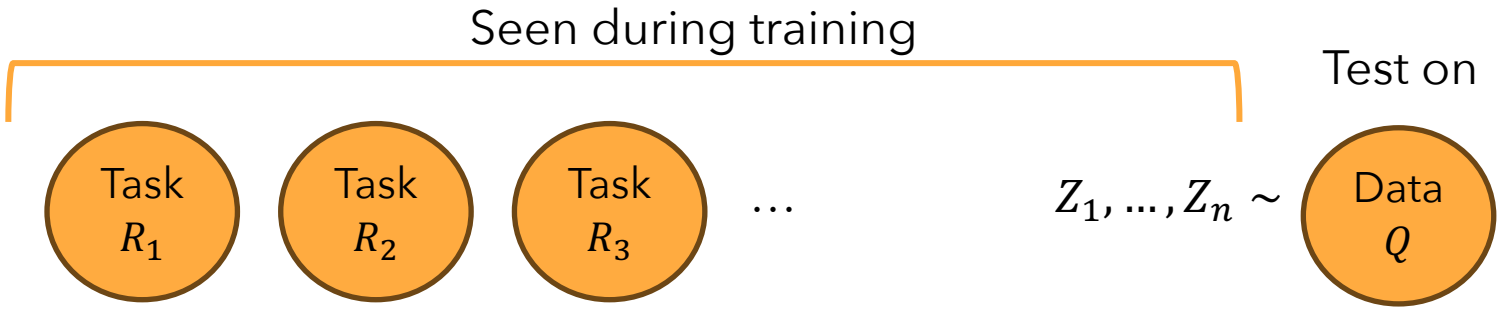
Blessing!

Available data: Lots of labeled data from multiple *sources* R_i

Question today:

How much can data from R_i help?

Setting and relation to practical transfer learning paradigms



Goal: Estimators that use structural knowledge, access to R_i and n samples from Q

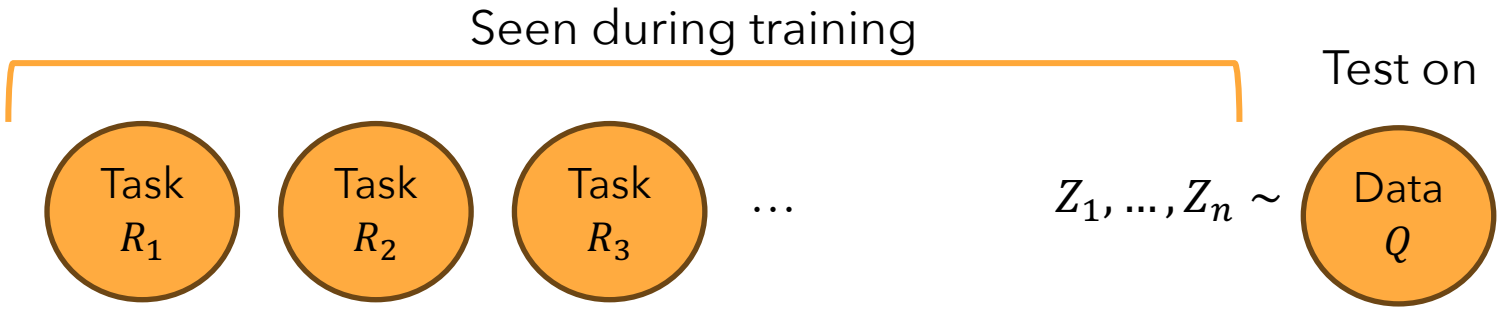
to return $\hat{h} \in H$ that has risk close to $\inf_{h \in H} R_Q(h)$

When H is finite (e.g. trained foundation models):

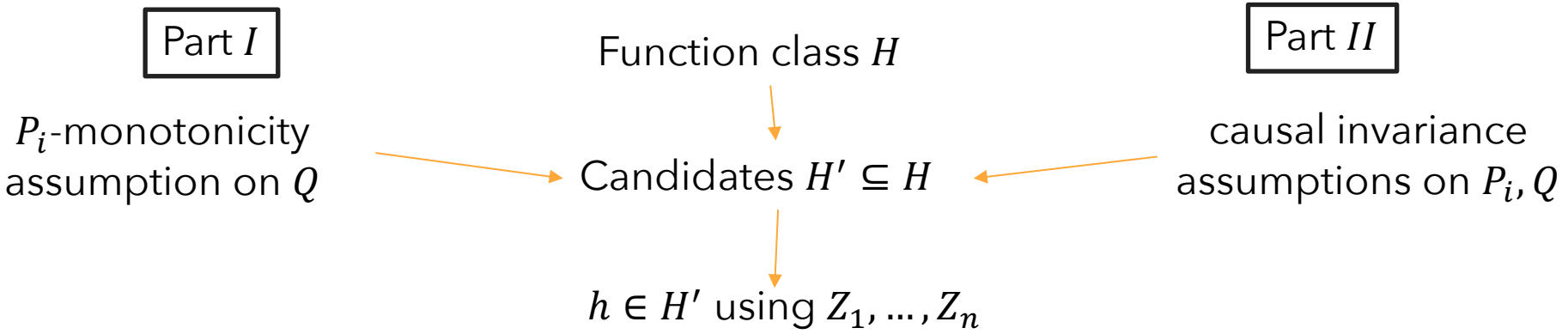
- akin to selecting one foundation model among a set H that you can query but not fine-tune
- **As opposed to related paradigms** : Do not assume access to intermediate layers

  as used in i) representation-based selection, ii) model merging in param. space, iii) logit fusion

Benefits of structure for Transfer Learning



- When can we leverage benchmarks R_i to significantly improve learning Q on function class H?
- Only when assumption on structural relationship between R_i and Q is reasonable and useful!



Question for both: How much can you reduce from H to H'?

Part I: Transfer with benchmarks

"Hedging on the Frontier: Learning New Tasks with Few Samples"
presented at ICML 2026

Tobias
Wegel (ETH)

Federico
Di Gennaro (ETH)

Geelon
So (UCSD)

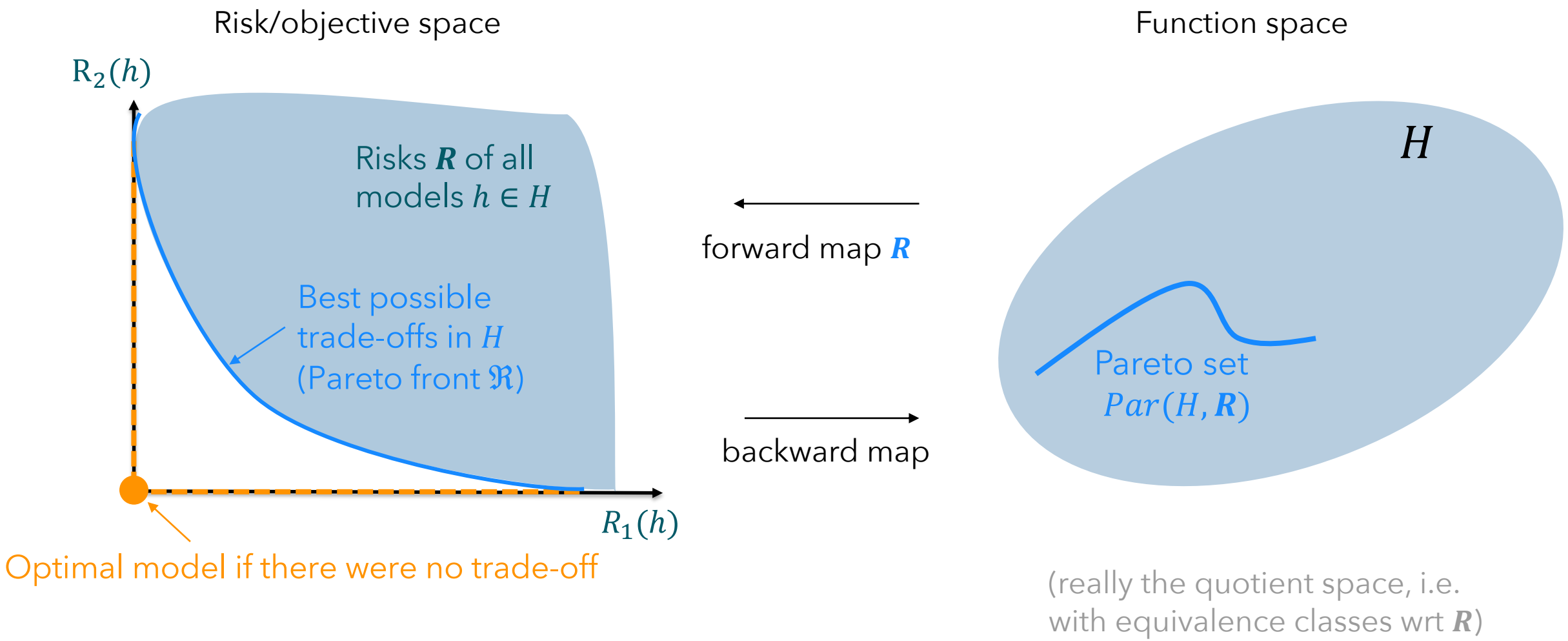
Plan for now

- Short primer on Pareto fronts and sets
- Monotonicity and Moduli of monotonicity
- A pareto cover that depends on Pareto front geometry
- Possible sample complexity gains for Pareto-aware algorithms

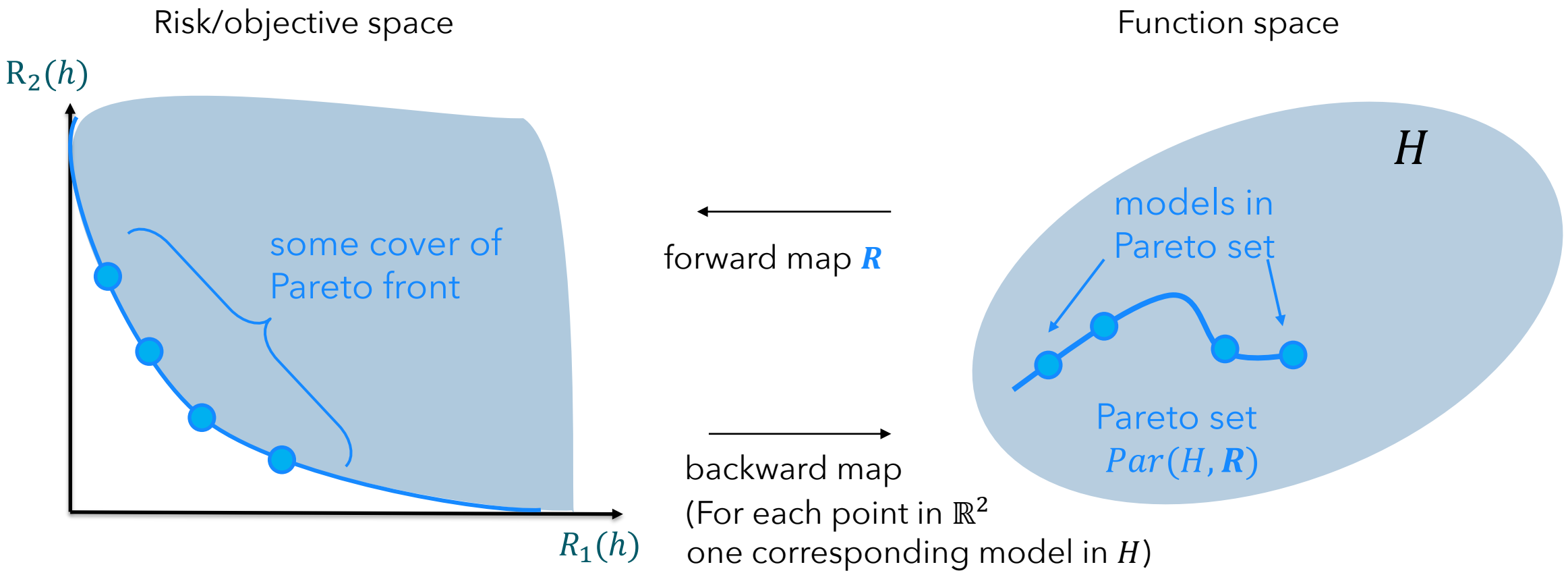
What we won't discuss:

- Fast rates for model aggregation for strongly concave Pareto fronts
  (w/o requiring strongly convex losses) when dictionary includes individual minimizers

Short primer on Pareto fronts and Pareto sets

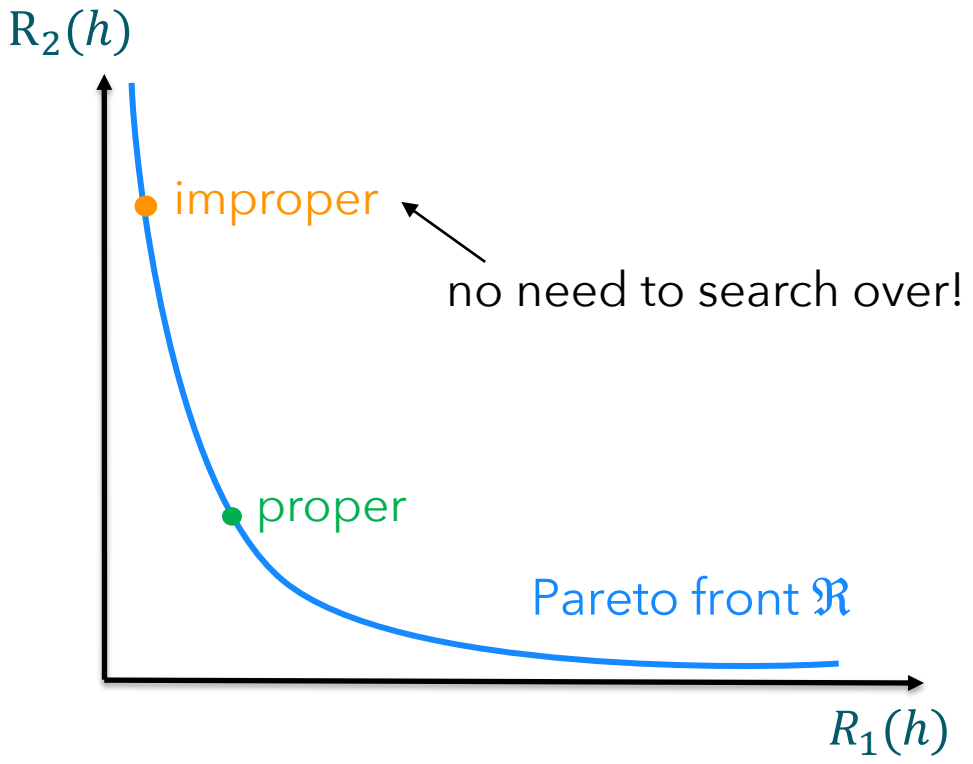


Short primer on Pareto fronts and Pareto sets



In what follows we often draw points in risk space but mean corresponding models in H

Short primer on good (proper) and bad (improper) trade-offs



All Pareto-optimal points on Pareto front are not strictly dominated by any other $h \in H$ (and which to pick depends on preference)

However, you can further distinguish between

- proper trade-offs (small gain-to-loss ratio) vs.
- improper ones (large gain-to-loss ratio)

A new reasonable (?) relatedness assumption

Monotonicity: A reasonable assumption in the era of benchmarks…?

Intuition: If a model h is better than h' *on all* benchmarks/risks $\boldsymbol{R} = (R_1, \dots R_K)$, then it's also better on the new task R_Q

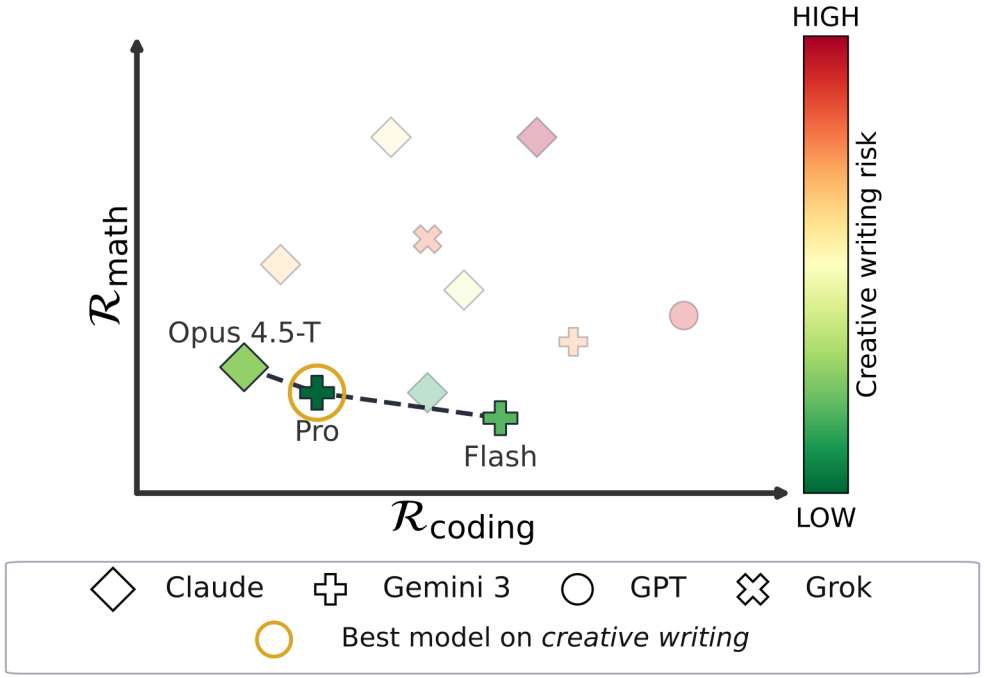
the larger K, the weaker the assumption

Formalization (MON): R_Q is monotone wrt $\boldsymbol{R}$ if

$\boldsymbol{R}(h) \preccurlyeq \boldsymbol{R}(h') \Rightarrow R_Q(h) \leq R_Q(h')$ for all $h, h' \in H$

(domination in all elements)

⇨ Minimizer of R_Q in H is in Pareto set $Par(H, \boldsymbol{R}) \subseteq H$ s.t. it suffices to search on the Pareto set



Data taken from Huggingface LMArena

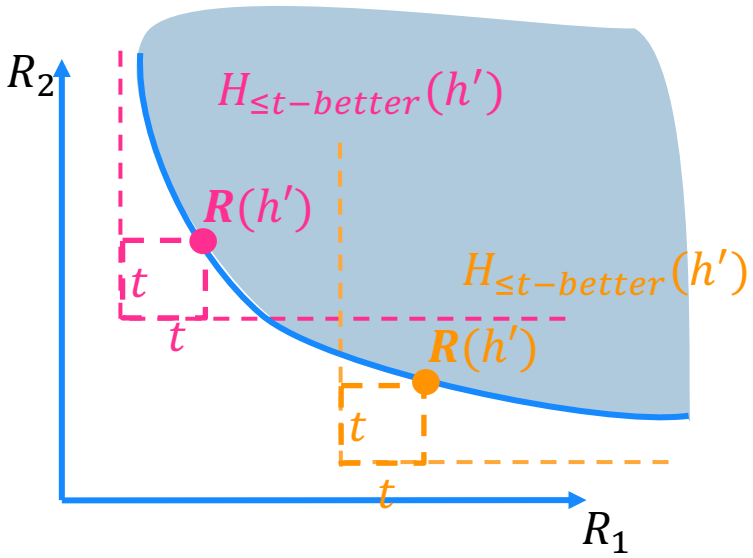
Quantifying monotonicity via the modulus

We can prove bounds depending on "how monotone" R_Q is

Definition (Upper modulus of monotonicity)

For any $h' \in Par(H, \boldsymbol{R})$, how much $R_Q(h)$ can be better than $R_Q(h')$ for all h at most t-better (i.e. $\boldsymbol{R}(h') - t\mathbf{1} \preccurlyeq \boldsymbol{R}(h)$):

$$\bar{\omega}_Q(t) = \sup_{h' \in Par(H,\boldsymbol{R})} \{R_Q(h') - R_Q(h) : h \in H_{\leq t-better}(h')\}$$

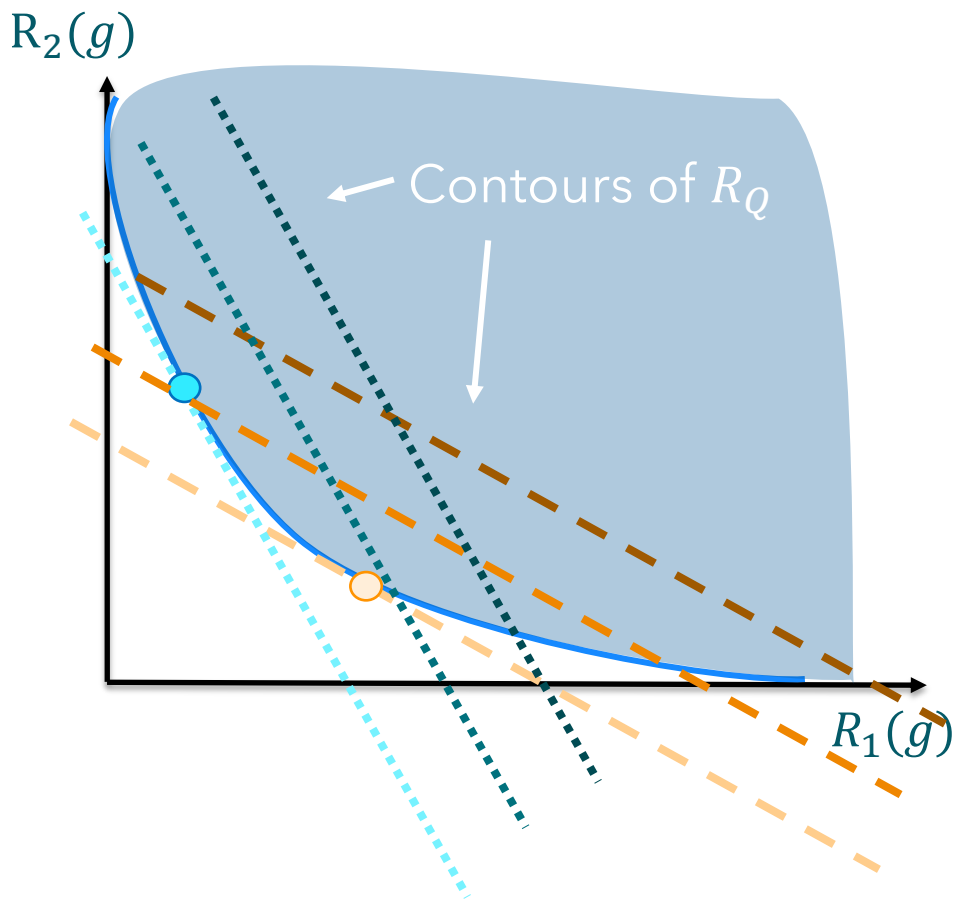


- R_Q is monotone w.r.t. $\boldsymbol{R} \Rightarrow \bar{\omega}_Q(0) = 0$ and minimizer in Pareto set
- $\bar{\omega}_Q(0) > 0$: inherent/irreducible "approximation error" of Pareto set
- how fast $\bar{\omega}_Q(t)$ grows with t: how fast R_Q can improve with "distance"

Examples for monotone and approximately monotone functions

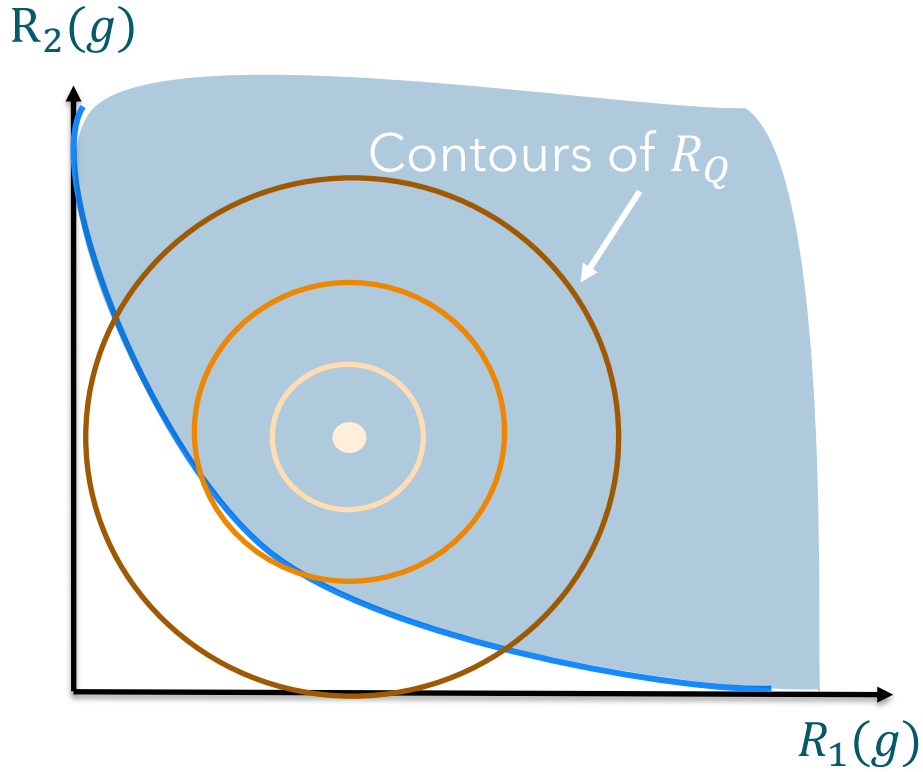
Mixtures $R_Q(\cdot) = \sum_{k=1}^{K} \lambda_k R_k(\cdot)$ with straight contour lines

→ for any λ, minimizer on Pareto front & $\bar{\omega}_Q(t) \leq t$

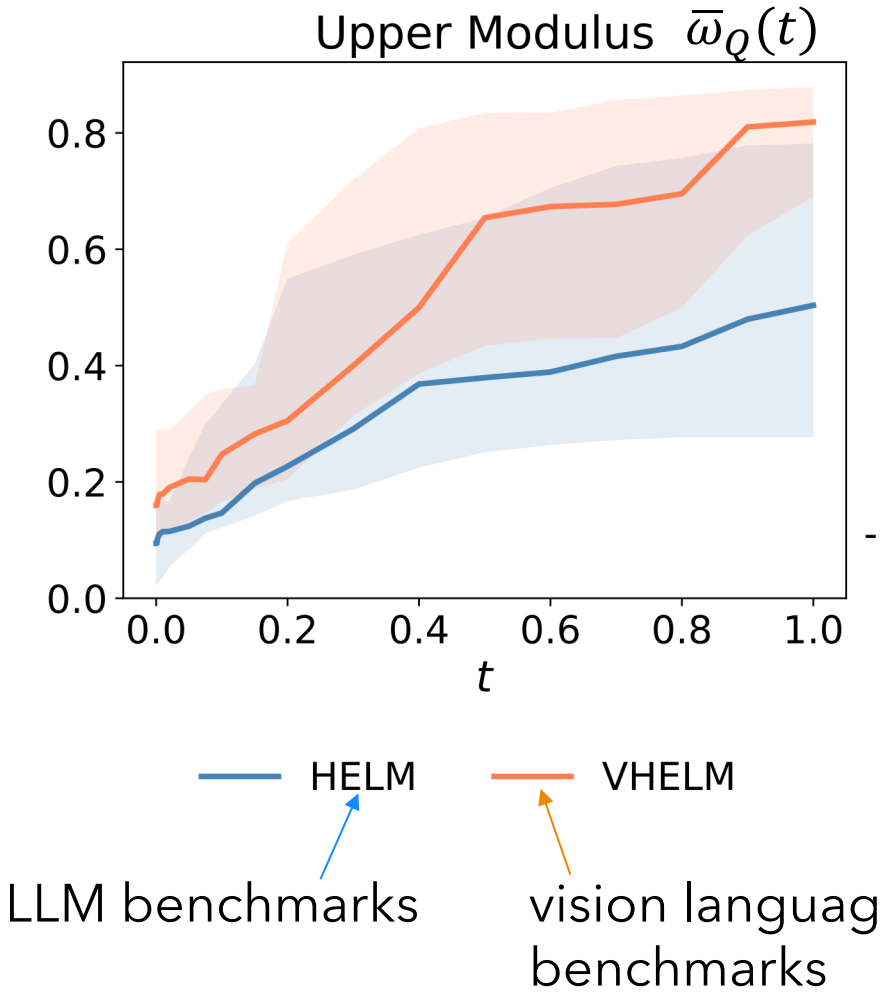


Minimum is close to Pareto set

(approximately monotone with $\bar{\omega}_Q(0) > 0$)

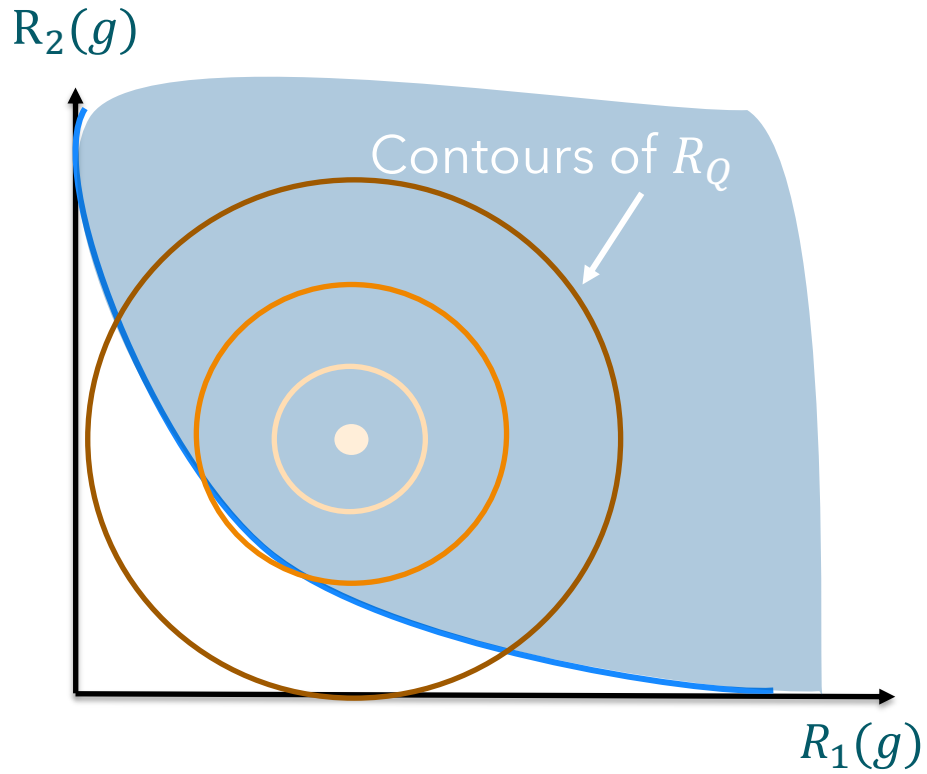


How realistic is it that monotonicity holds (or $\bar{\omega}_Q(0)$ small)



Minimum is close to Pareto set

(approximately monotone with $\bar{\omega}_Q(0) > 0$)



Relation to assumptions in prior work

Similar notions of relatedness were used in the (learning theory) literature:

Modulus of monotonicity extends beyond modulus of transfer (K=1) [Hanneke-Kpotufe '24]

Monotonicity extends beyond mixture distributions considered in [Mansour-Mohri-Suresh-Wu '21, Xiong-Guo-Cai '23]

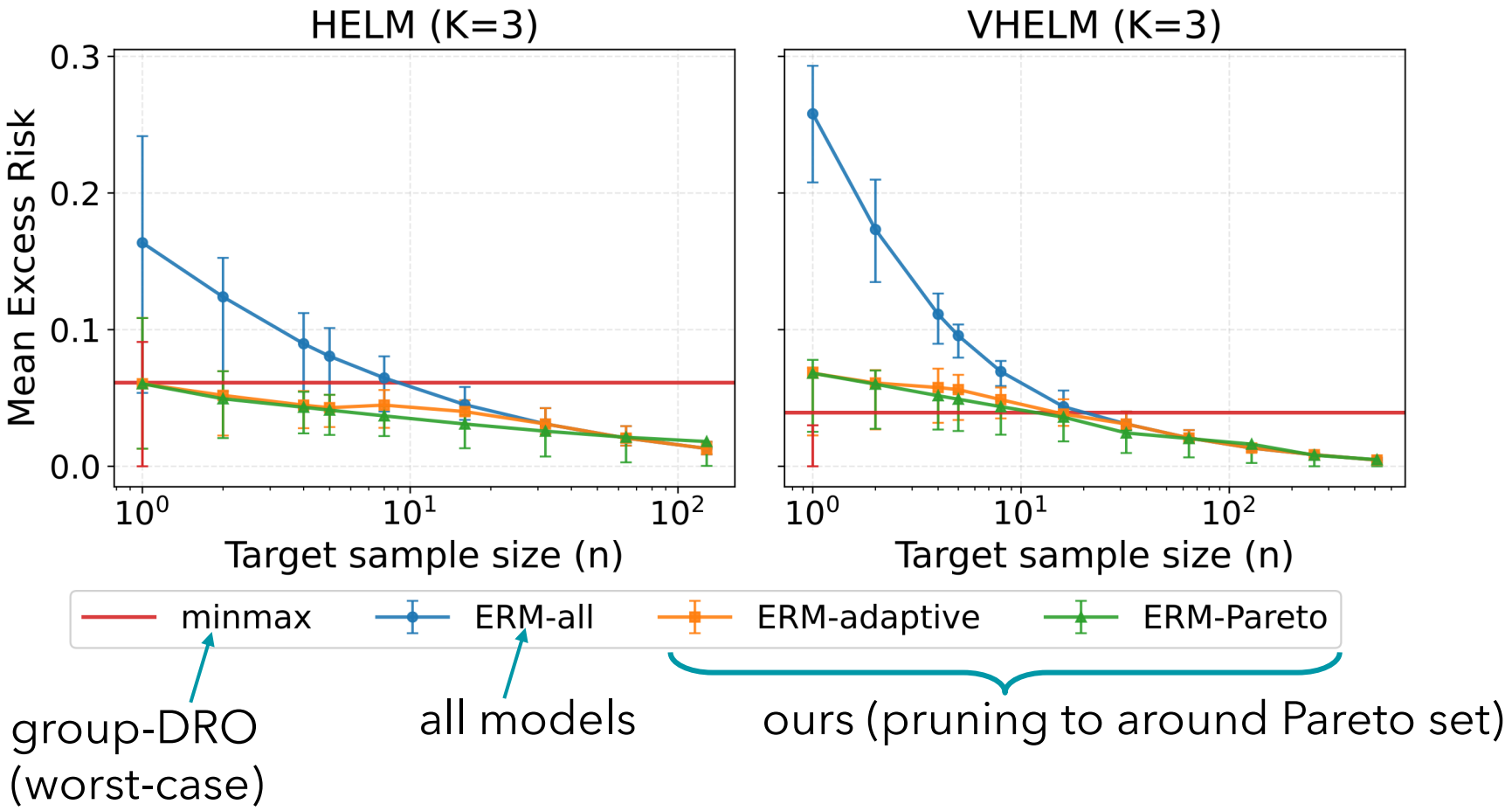
Question: If (modulus of) monotonicity is all we know, how do we make use of it to reduce the excess risk $\mathcal{E}_Q(\hat{h}) := R_Q(\hat{h}) - \inf_{h \in H} R_Q(h^\star)$?

Takeaway: (Approximate) monotonicity allows sample complexity reduction

1. By searching only on Pareto set: Complexity of H → complexity of Pareto set
2. Using a Pareto cover whose size adapts to Pareto geometry → geometry-aware Pareto set complexity

Disclaimer: focus is statistical, not proposing a new method (except for finite H) → left for future work

Finite H: Efficiency gain from only searching in Pareto set



A new measure reflecting Pareto front geometry

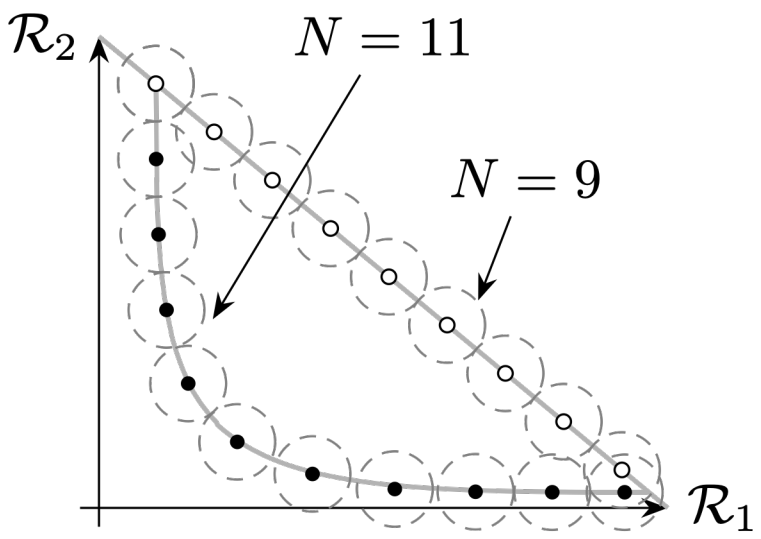
In search of a complexity adapting to Pareto front geometry

Rationale: Under monotonicity we can constrain search to Pareto set

→ complexity should scale with covering number of the Pareto front – but how to cover?

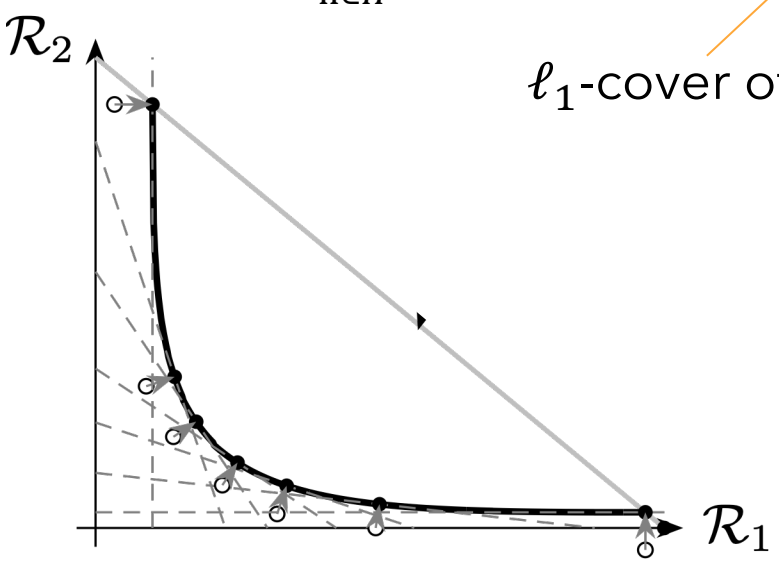
"Naïve" covers:

ℓ_2 −cover of the front in $\mathbb{R}^K$



X unnecessary crowding on points that don't differ much in R_Q (by monotonicity)

simplex cover { $\in \Lambda \subset \Delta^{K-1}$ }



ℓ_1-cover of $K-1$ simplex

X size independent of Pareto front geometry

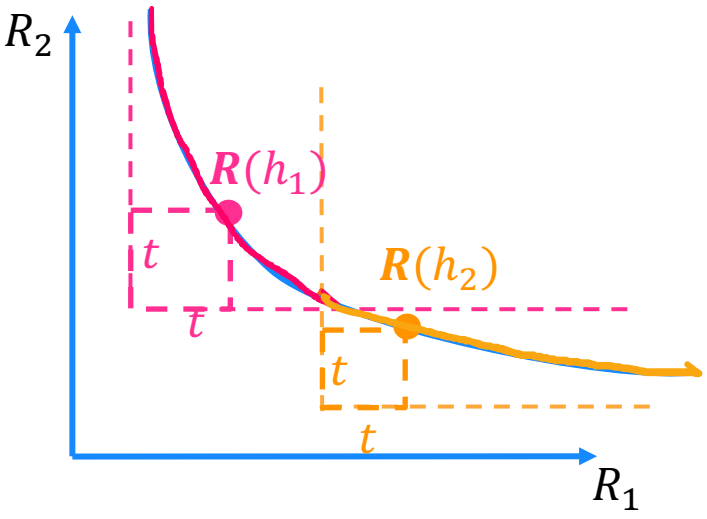
X not enough cover for harder case

A complexity measure that adapts to Pareto front geometry

Desiderata for our cover under (approx.) monotonicity

- not include many models with improper tradeoffs since by monotonicity small potential R_Q gain
- smaller when there are many bad tradeoffs, larger when mostly proper trade-offs

satisfied by



Definition (t-Pareto set and covering number)

$h_1, \dots, h_N$ is t-Pareto set if $\forall h \in Par(H, \boldsymbol{R})$, $\exists h_j$ s.t. $\boldsymbol{R}(h_j) \preccurlyeq \boldsymbol{R}(h) + t1$

$N_{par}(t, H, \mathbf{R})$ denotes the smallest t −Pareto set size

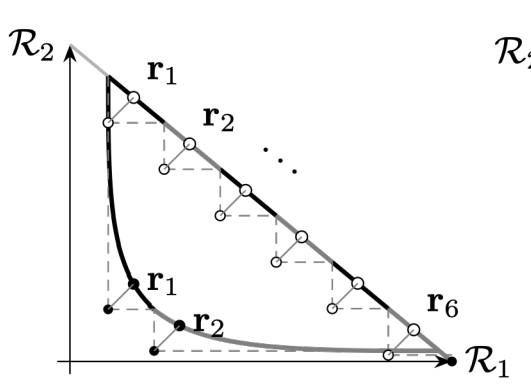
For $h^\star \in argmin_{h \in H} R_Q(h)$ we have for some j: $\boldsymbol{R}(h_j) \preccurlyeq \boldsymbol{R}(h^\star) + t1$ and excess risk $\mathcal{E}_Q(h_j) \coloneqq R_Q(h_j) - R_Q(h^\star) \leq \bar{\omega}_Q(t)$ by definition of $\bar{\omega}_Q(t)$

Comparison with other coverings

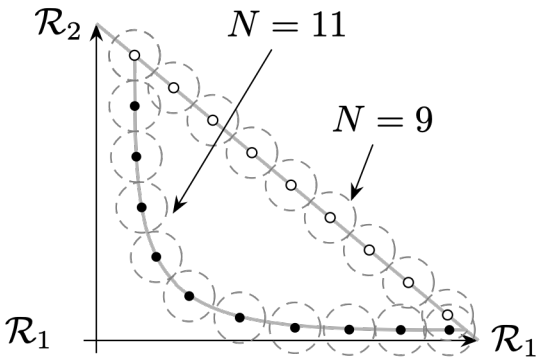
Desiderata for our cover under (approx.) monotonicity

- not include many models with improper tradeoffs since by monotonicity small potential R_Q gain
- smaller when there are many bad tradeoffs, larger when mostly proper trade-offs

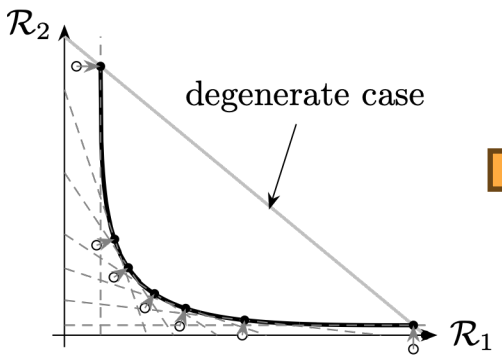
Visualizing different covers for two simple examples of competing risks with $K = 2$:



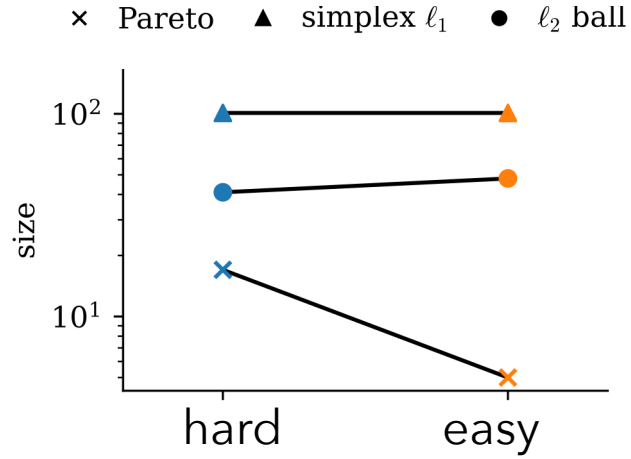
Pareto cover in 2d (ours)



ℓ_2-cover in 2d



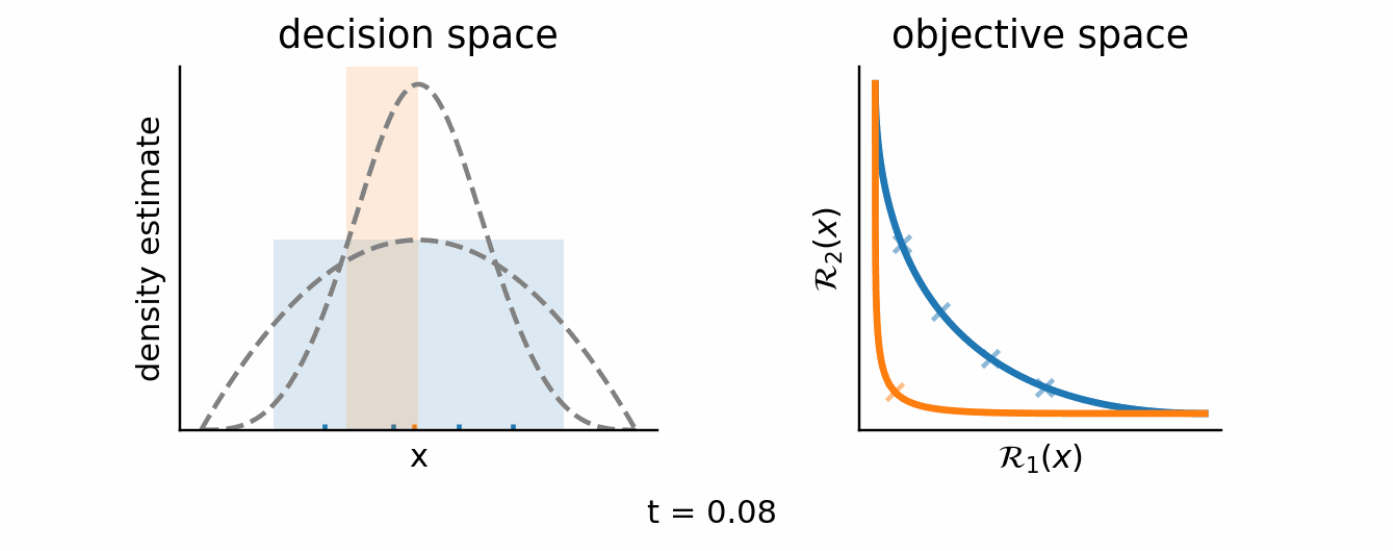
simplex cover



A limiting distribution for the Pareto cover on the Pareto front

Theorem (informal) (WGSY '25)

For smooth Pareto fronts, the distribution of points in Pareto cover as $t \to 0$ follows a distribution μ that puts more mass on models in Pareto set with "good" trade-offs (normal vector close to $(1, \ldots, 1)$)



Interesting independent technical contributions in the proof regarding simplex coverings…

Benefits of being geometry-aware

Pareto-geometry aware transfer algorithms

This notion of cover and its limiting distribution μ naturally give rise to two algorithms that both *"hedge"* on points with good trade-offs

Algorithm (Pareto ERM)

1. Build minimal t −Pareto set

$$H' := \{h_1, \dots, h_{N_{par}(t,H,\boldsymbol{R})}\}$$

2. Compute ERM on

$$\hat{h} \in \operatorname{argmin}_{h \in H'} \hat{R}_Q(h)$$

Algorithm (Pareto Exp. Weights)

1. Let π be pull-back distribution on $Par(H, \boldsymbol{R})$ induced by μ (on $\mathbb{R}^2$)
2. Compute $\hat{\rho}(h) \propto \pi(h) \exp(-\lambda \hat{R}_Q(h))$

$$\hat{h}(\cdot) = \mathbb{E}_{h \sim \hat{\rho}}[h(\cdot)]$$

Recall $\mathcal{E}_Q(\hat{h}, H) = R_Q(\hat{h}) - \inf_{h \in H} R_Q(h)$

$\bar{\omega}_Q(t) = \sup_{h \in Par(H,\boldsymbol{R})} \{R_Q(h) - R_Q(h') : h' \in H_{\leq t-better}(h)\}$

Theoretical guarantees and comparisons

Theorem (excess risk bound for Pareto-ERM/EW) (WGSY '25)

For both estimators $\hat{h}$, for bounded losses, we have* with prob. at least $1 - \delta$

$$\mathcal{E}_Q(\hat{h}, H) \leq \underbrace{\bar{\omega}_Q(t)}_{\text{scale-dependent approximation error}} + O\left(\sqrt{\frac{\log\left(N_{par}(t, H, \boldsymbol{R})/\delta\right)}{n}}\right)$$

- consistent if MON $\Leftrightarrow \bar{\omega}_Q(0) = 0$ (in paper, have consistent algorithm even if $\bar{\omega}_Q(0) \neq 0$)

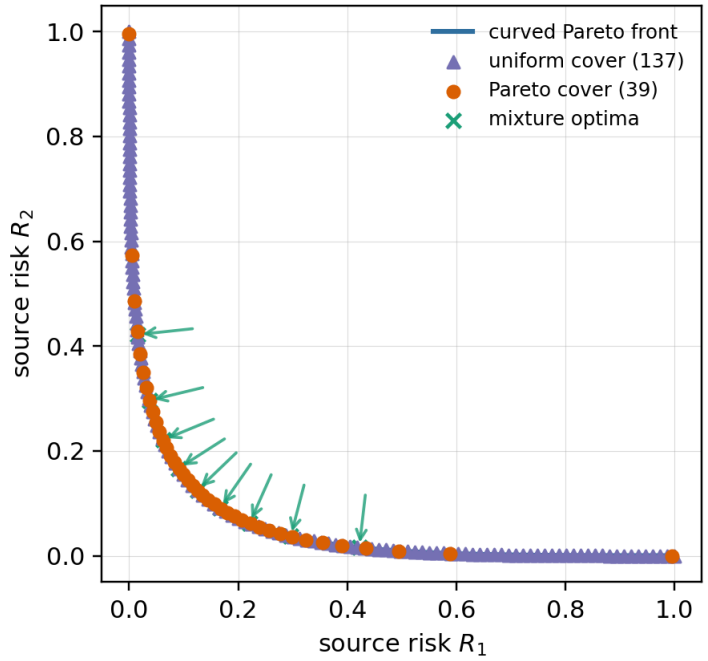
Compare with agnostic algorithms:

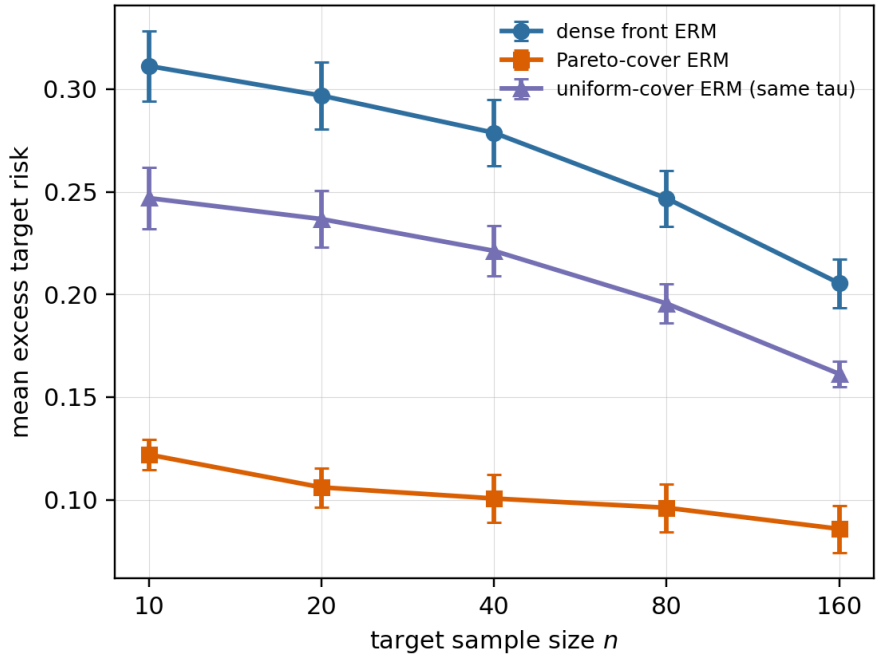
- ERM over uniform ℓ_2-cover: potentially $N_{par}(t, H, \boldsymbol{R}) \ll \left(\frac{1}{t}\right)^{K-1}$
- Hanneke/Kpotufe for $K = 1$ ($N_{par} = 1$): same bound of $\bar{\omega}_Q(0)$ (but ours extends to $K > 1$)
- Mansour for mixtures: algorithm with $O\left(\sqrt{K/n}\right)$ bound is geometry independent, hence non-adaptive

*for strongly convex can achieve fast rate $O\left(\log\left(N_{par}(t, H, R)/\delta\right)/n\right)$

Infinite H: Efficiency gains wrt uniform (Euclidean) cover of Pareto set

For simplicity consider targets that are mixtures $R_Q = wR_1 + (1-w)R_2$





Pareto cover "maximally" uses structure and concentrates on proper trade-offs

→ achieves smaller risk with the same # samples from target Q

Part II: Transfer with invariances

"How useful is Causal Invariance for Domain Adaptation in Finite Sample Settings?"
arxiv preprint, feedback welcome

Julia Kostin
ETH Zurich

Kasra Jalaldoust
Columbia

Elias Barenboim
Columbia

Samory Kpotufe
Columbia

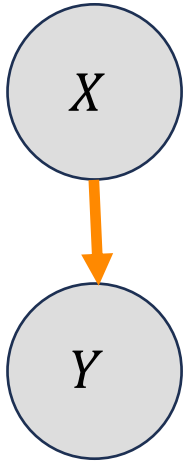
Invariance: Unified view on many structured shift assumptions

It's natural to assume that some aspects of data generating mechanism stay invariant

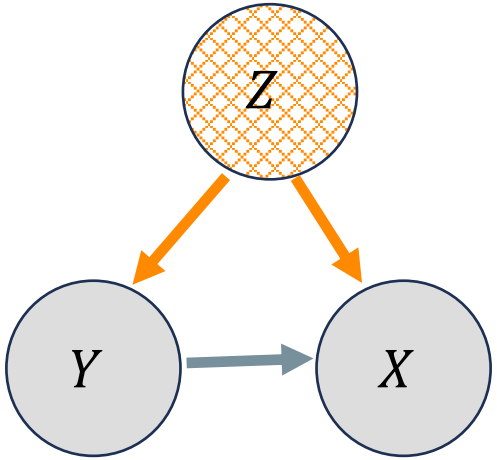
many are visualizable using invariances in DAGs / graphical models

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

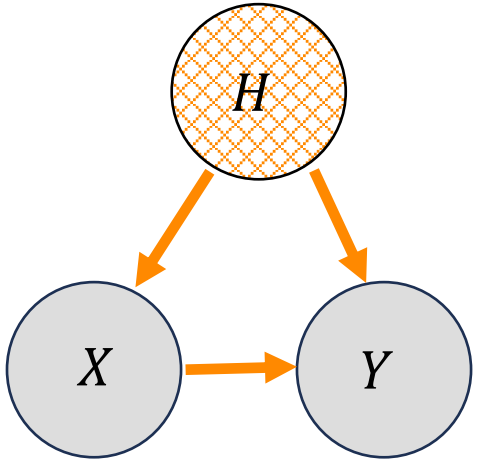
Covariate shift



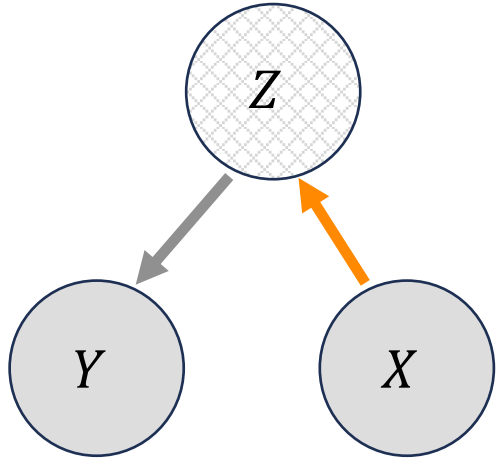
Invariant representations



Domains as instruments/anchors

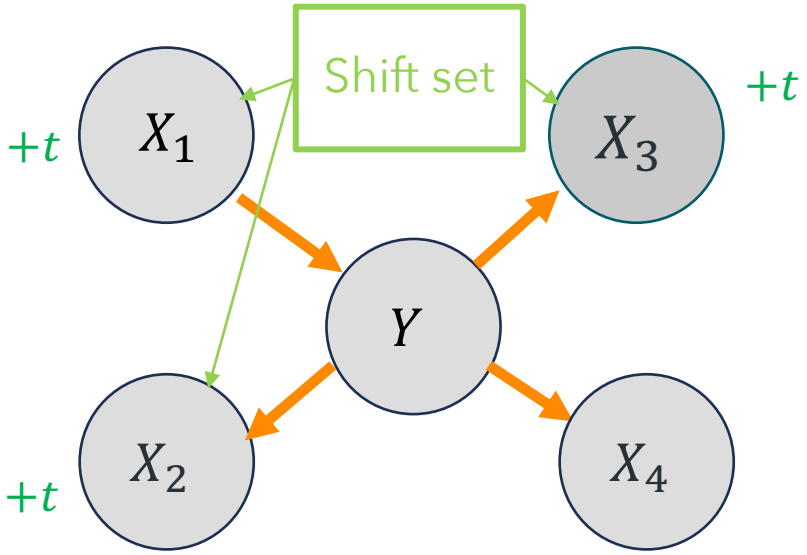


Multi-task learning



Much studied in robust domain generalization (no fixed Q but worst-case over set)

Subset feature invariance under partial identifiability



In this work, for simplicity we focus on (P, Q)-invariant subsets, i.e. $P(Y|X_I) = Q(Y|X_I)$ across P, Q.

- What does invariance buy us? If I is (P, Q)-invariant, then

  $h_{I.P} = \mathbb{E}_P[Y|X_I] = \mathbb{E}_Q[Y|X_I] = h_{I,Q} = h_I \Rightarrow R_Q(h_{I,P}) = R_Q(h_{I,Q})$

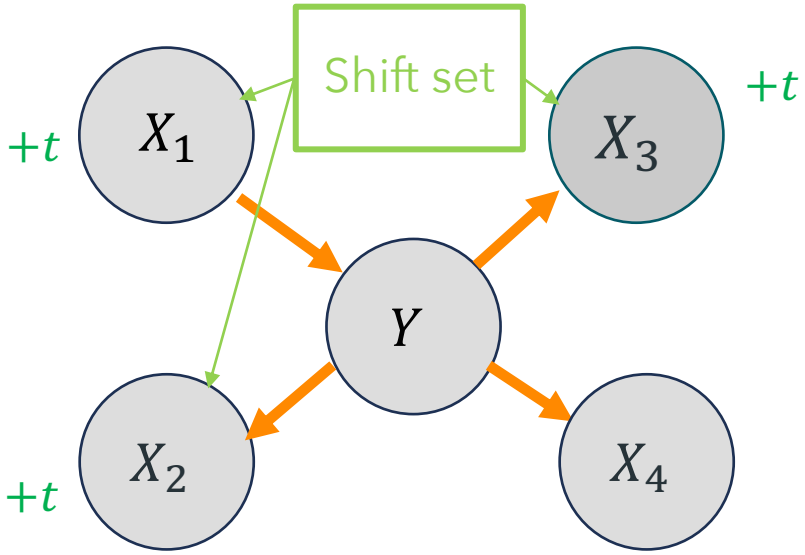
  👍 can obtain $h_{I,P}$ purely by training using P (lots of data) and achieve at least bounded risk on Q

  👎 In general not useful: I may not be predictive in Q.

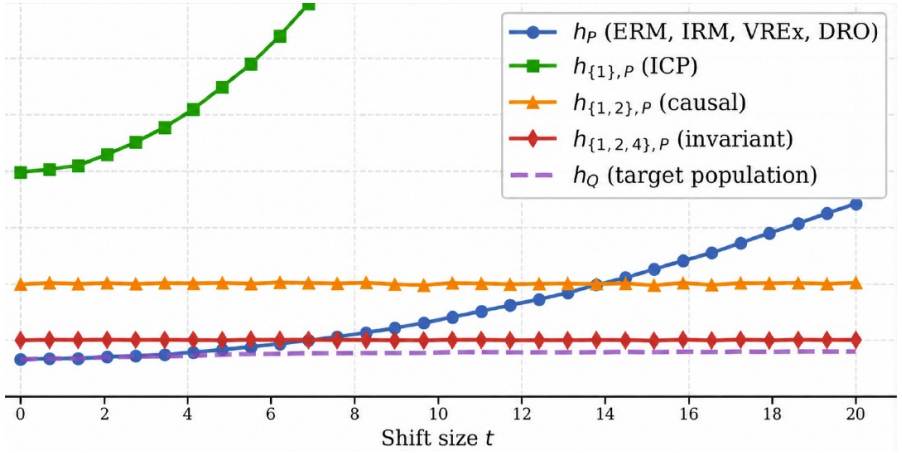
- Even if we assume there is invariant set $I^\star$ with small excess risk $\mathcal{E}_Q(h_{I^\star}) = R_Q(h_{I^\star}) - \inf_{h \in H} R_Q(h) < \epsilon$

  👎 $I^\star$ is usually unknown, might have collection $\mathfrak{I}$ (e.g. confidence set) of subsets containing $I^\star$

Subset feature invariance under partial identifiability



In this work, for simplicity we focus on (P, Q)-invariant subsets, i.e. $P(Y|X_I) = Q(Y|X_I)$ across P, Q.



minimax rate of model aggregation

Recall: we know P, have n samples from Q and access to $\mathfrak{T}$

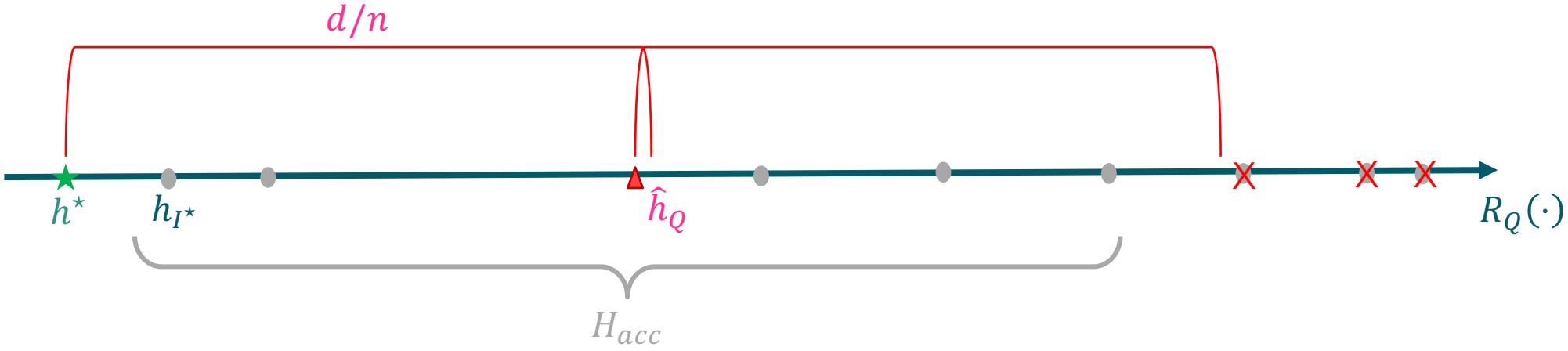
Further the best $I^\star \in \mathfrak{T}$ has small excess $\mathcal{E}_Q(h_{I^\star}) \ll r_{agg}(n, |\mathfrak{T}|)$?

Question: When (in terms of n, shift) is it possible to *benefit* from $\mathfrak{T}$ to achieve excess risk close to $\mathcal{E}_Q(h_{I^\star}) \ll r_{agg}(n, |\mathfrak{T}|)$ & never be worse than simply training on n?

(non-negative transfer)

When can we beat just aggregation and train-on-target?

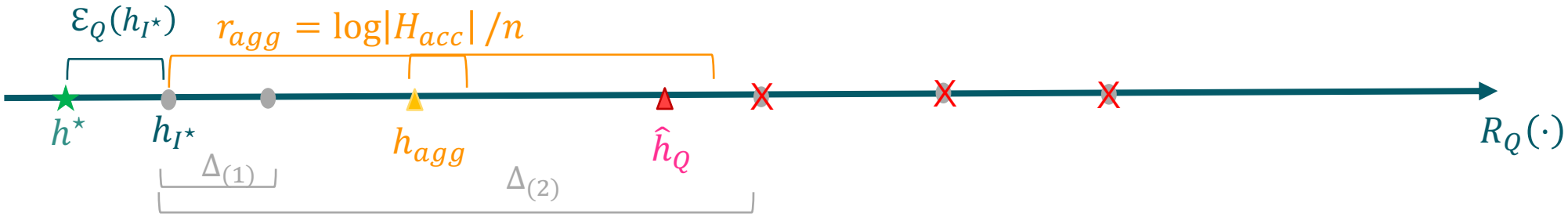
To cleanly quantify, our analysis (first of that kind) considers the linear regression setting and provides an answer that depends on the *margin structure* between the risks i.e. $\Delta_I := R_Q(h_I) - R_Q(h_{I^\star})$



$\mathcal{E}_Q(h) = R_Q(h) - \inf_{h \in H} R_Q(h)$
and $\mathcal{E}_Q(h_{I^\star}) \ll \log|H_{acc}|/n$

When can we beat just aggregation and target?

To cleanly quantify, our analysis (first of that kind) considers the linear regression setting and provides an answer that depends on the *margin structure* between the risks i.e. $\Delta_I := R_Q(h_I) - R_Q(h_{I^\star})$



- A two-stage estimator can achieve non-negative transfer (never worse than target rate $\frac{d}{n}$) and

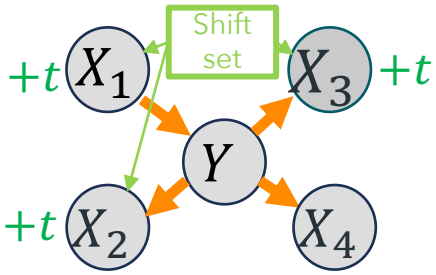
necessary

- If $\Delta_{(1)} > \frac{\log|H_{acc}|}{n}$: can achieve $\mathcal{E}_Q(\hat{h}) = \mathcal{E}_Q(h_{I^\star})$ when $n > \Omega(d)$
- If $\Delta_{(1)} \ll \frac{\log|H_{acc}|}{n}$ and $\Delta_{(2)} - \Delta_{(1)} > \frac{\log|H_{acc}|}{n}$: can get $\mathcal{E}_Q(\hat{h}) \ll \mathcal{E}_Q(h_{I^\star}) + \Delta_{(1)}$ when $n > \Omega(d)$

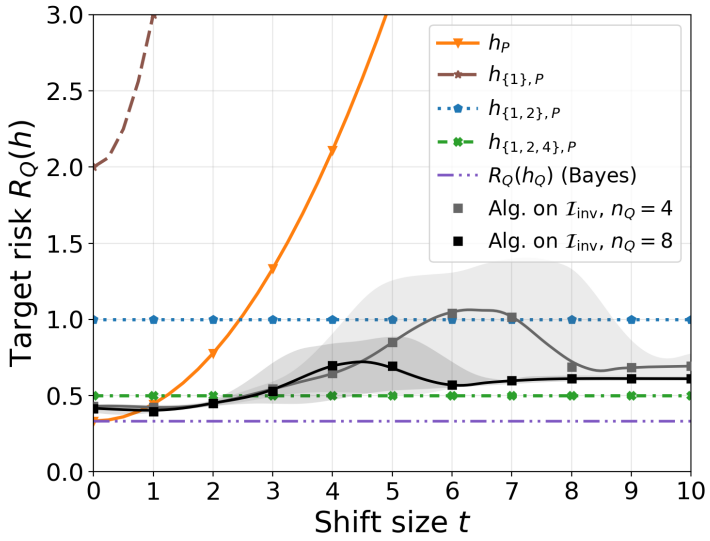
much better than aggregation

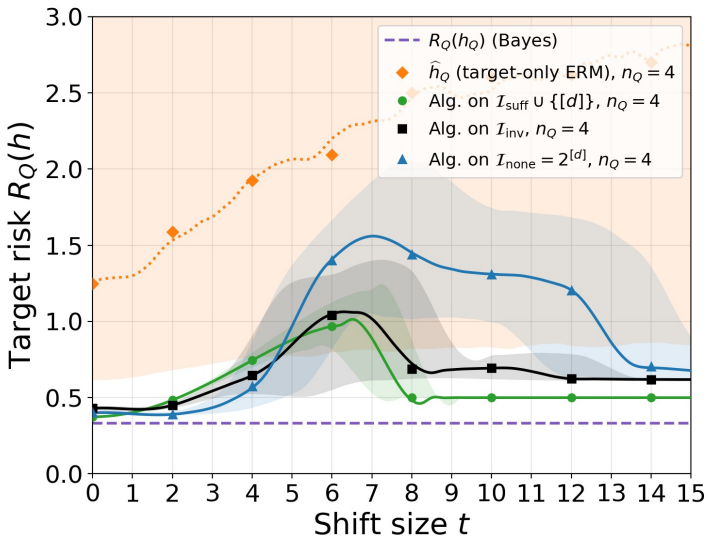
- For causal SCM (like example), $\Delta_{(1)}$ is bounded for truly invariant sets, $\Delta_{(2)} - \Delta_{(1)}$ increases with shift

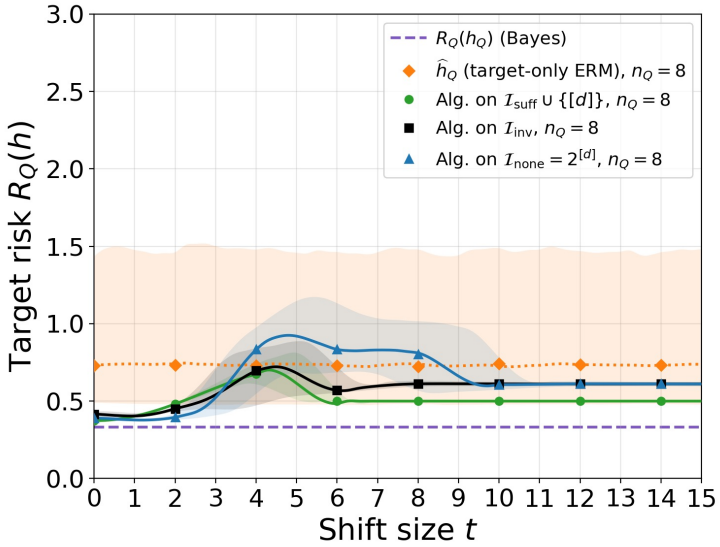
Visualizing the trends on synthetic example



Given different kind of collections (containing different invariant and other subsets)







- non-negative transfer – never worse than the target only model
- as shift increases transfer gets harder; but when it's very large, easier to detect good invariant sets
- as we increase n, the necessary shift at which it detects well drops

Takeaway

- In the case of monotonicity, the more "proper trade-offs" the harder
- In the case of invariances, the larger the shifts, the easier to take advantage of a good model

- T. Wegel, F. di Gennaro, G. So, FY "*Hedging on the Frontier: Learning New Tasks with Few Samples*" ICML 2026

Related: Pareto set learning with finite samples

- T. Wegel, G. So, J. Park, FY "On the sample complexity of semi-supervised multi-objective learning" NeurIPS 2025
- J. Kostin, K. Jalaldoust, E. Barenboim, S. Kpotufe, FY, "*How Useful is Causal Invariance for Domain Adaptation in Finite-Sample Settings*?", arxiv preprint

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