Transfer learning with benchmarks

CLS Retreat 2026

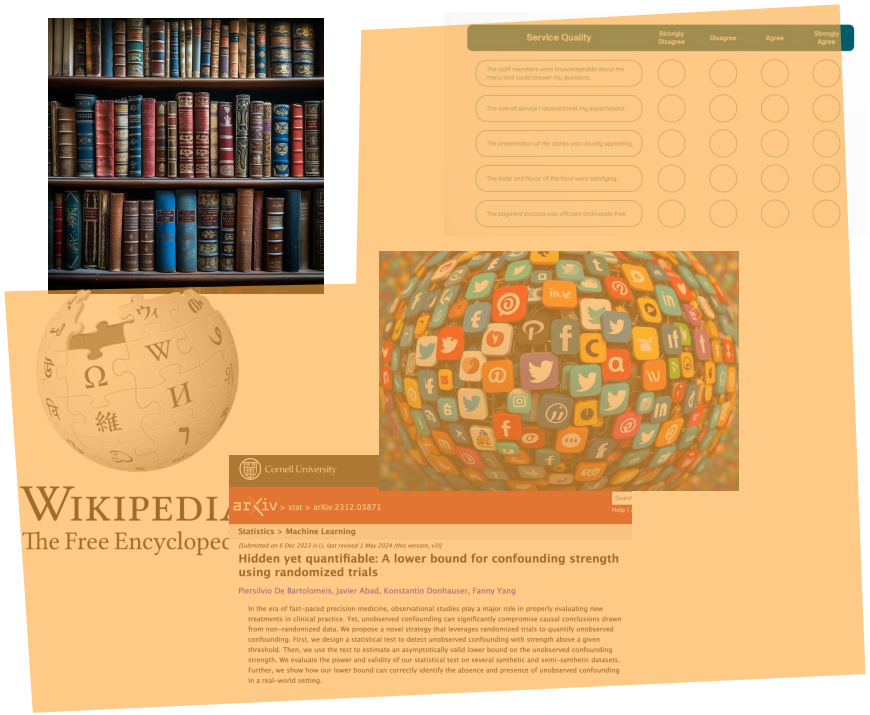
Fanny Yang, Department of Computer Science, ETH Zurich

Statistical Machine Learning group

Distributions shift – a challenge and blessing?

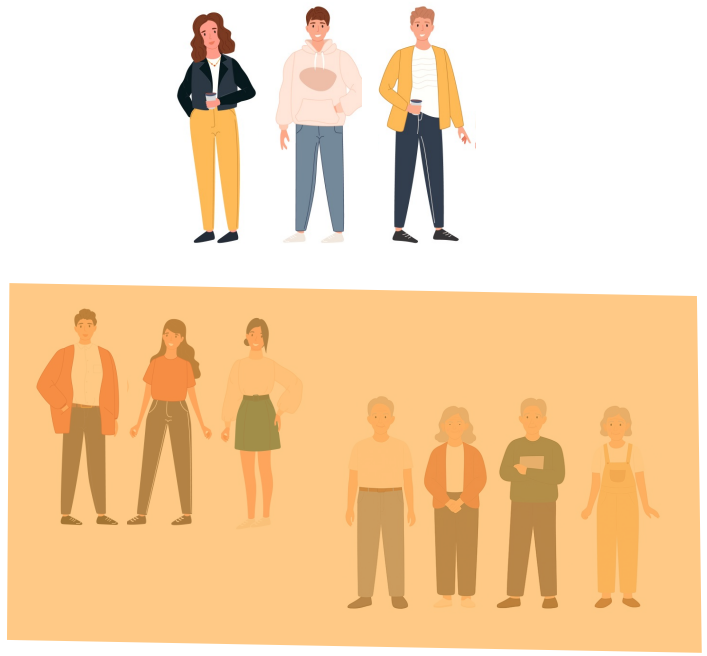
Transfer learning (supervised domain adaptation): Test on **new domain** given only little data

Different text corpuses



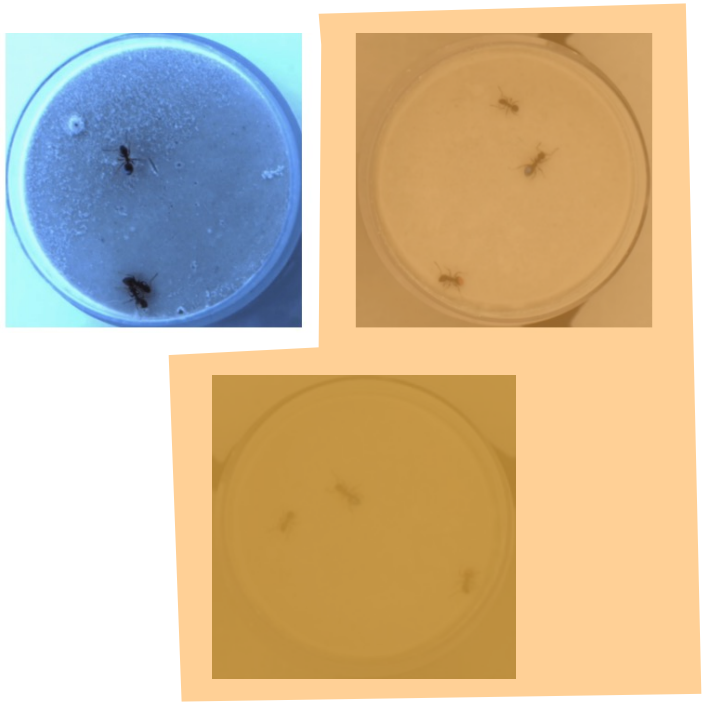
Text → Next token

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)

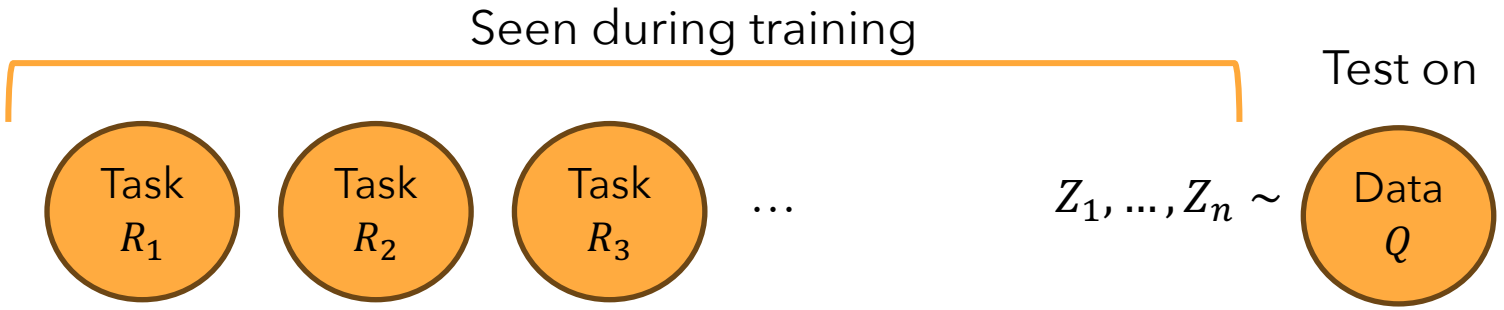
Blessing!

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

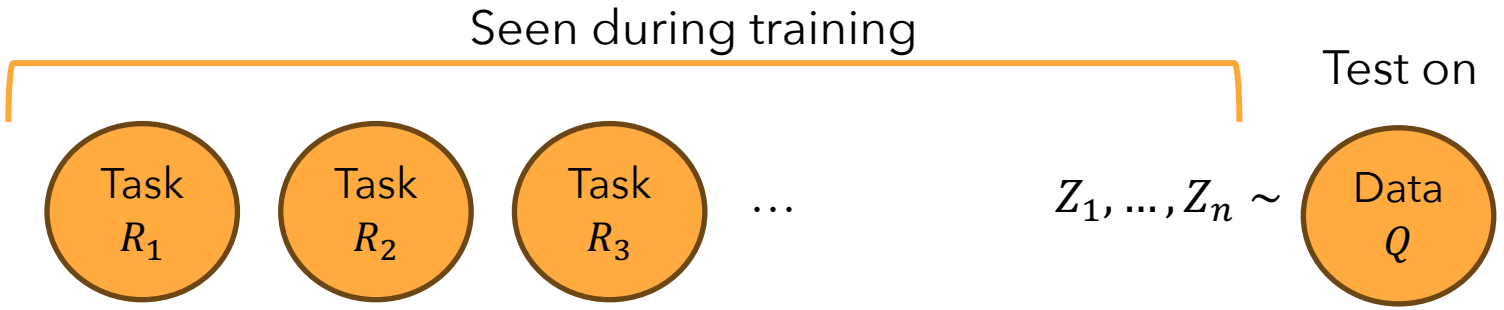
to return $\hat{h} \in H$ that has risk close to $\inf_{h \in H} R_Q(h)$

When H is finite, this could be applied to H being foundation models:

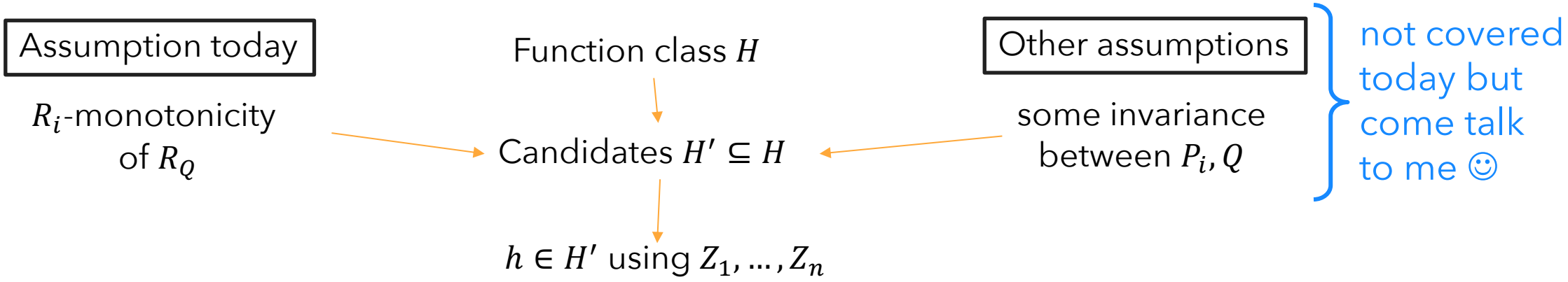
- goal is akin to selecting one foundation model among a set H that you can query but *not fine-tune*
- **As opposed to related paradigms** : Do not assume access to intermediate layers

  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: How much can you reduce from H to H'?

Transfer with benchmarks

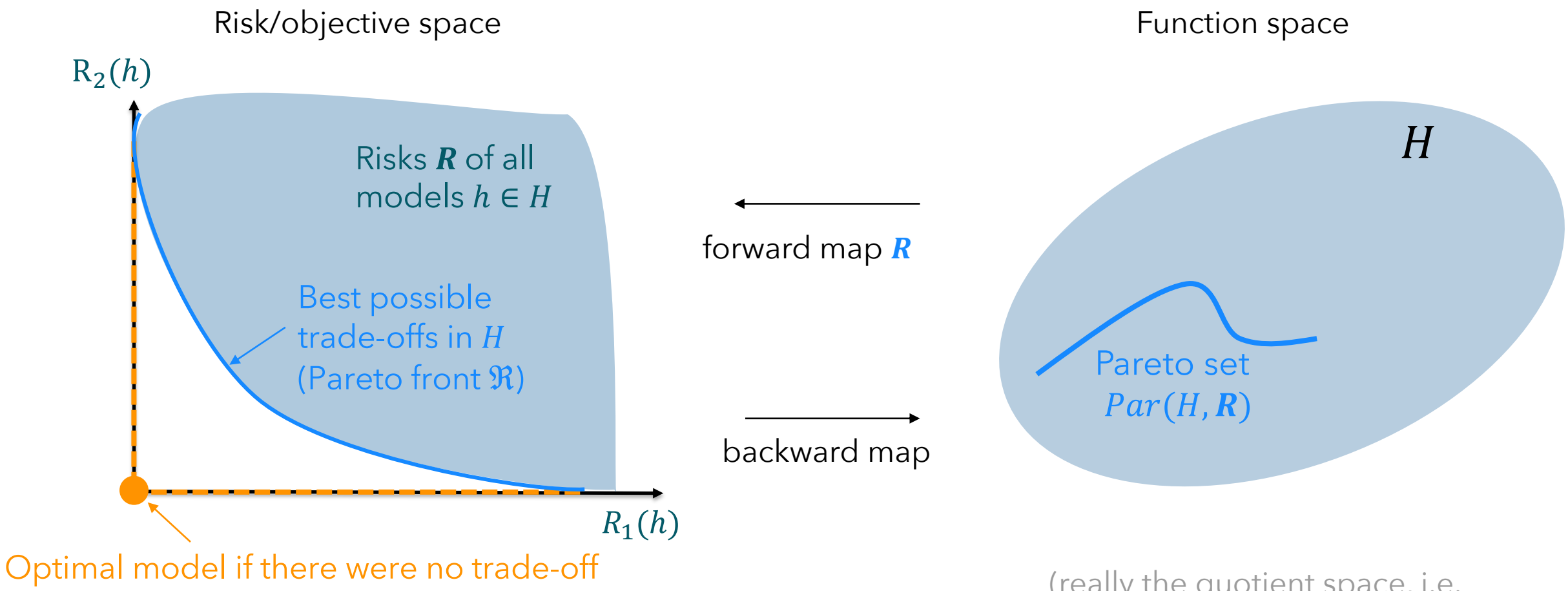
"Hedging on the Frontier: Learning New Tasks with Few Samples"
presented at ICML 2026

Tobias
Wegel (ETH)

Federico
Di Gennaro (ETH)

Geelon
So (UCSD)

Short primer on Pareto fronts and Pareto sets



(really the quotient space, i.e. with equivalence classes wrt $\boldsymbol{R}$)

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

Intuition I: If a model h is better than h' *on all* benchmarks

$\boldsymbol{R} = (R_1, \ldots R_K)$, then it's also better on the new task R_Q

the larger K, the weaker the assumption

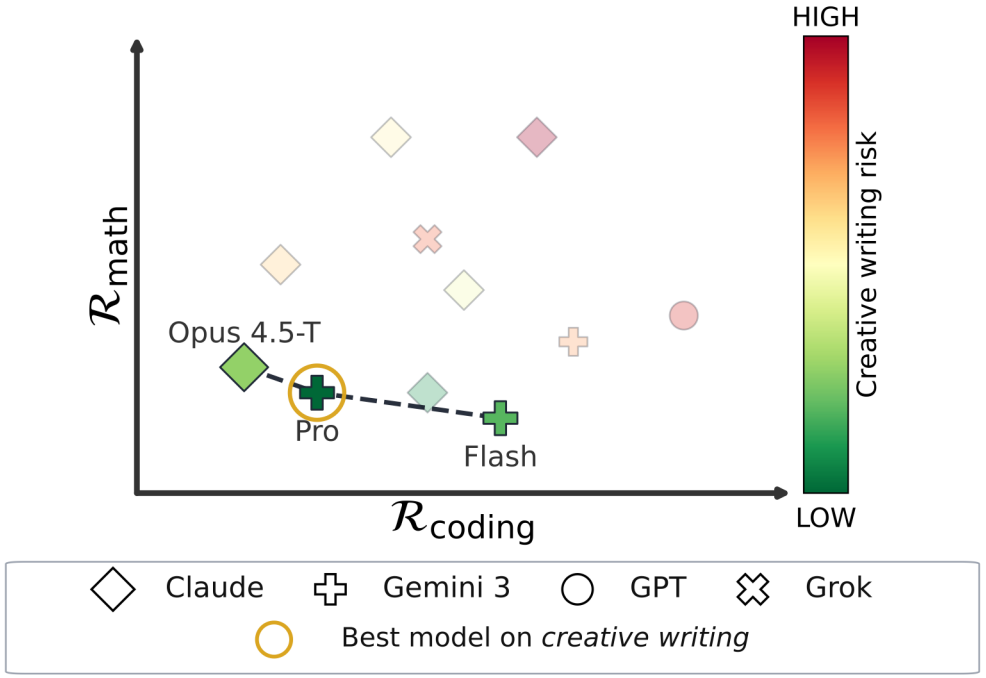
Formalization (MON): R_Q is monotone wrt $\boldsymbol{R}$ if

$\boldsymbol{R}(h) \preccurlyeq \boldsymbol{R}(h') \Rightarrow R_Q(h) \leq R_Q(h')$ 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

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

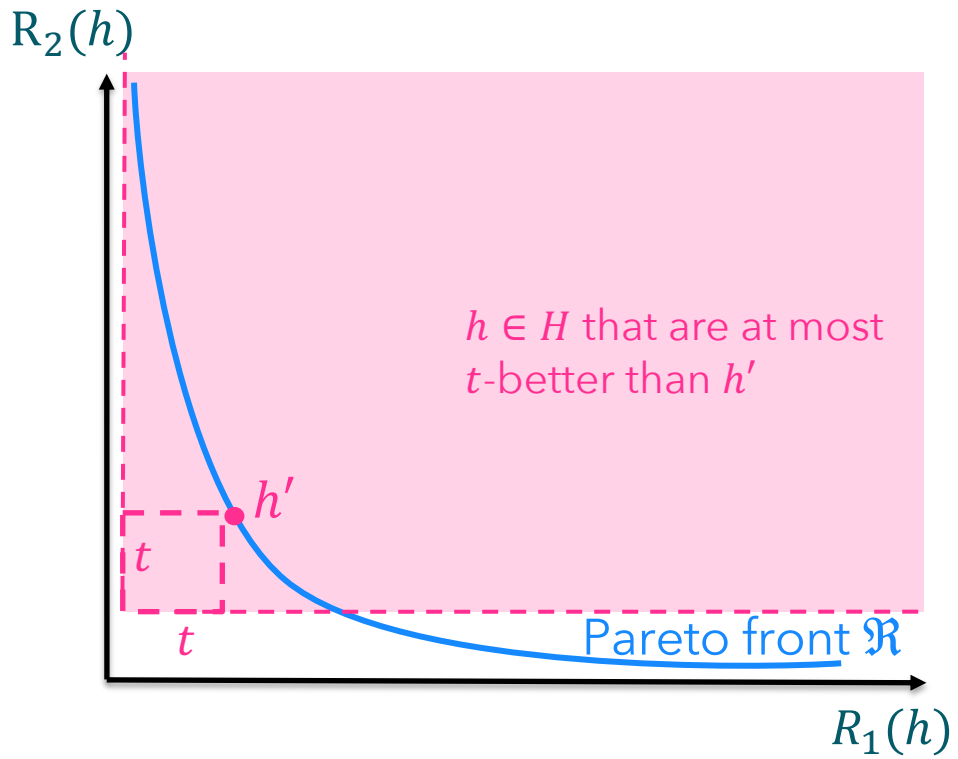
Modulus of monotonicity: Quantifying degree of monotonicity

Intuition II: If a model is a lot better on some benchmarks but not much better/worse on others, it can't be much better on a new task (think of holistic assessments of PhD candidates)

Formalization: Encoded in size of modulus $\bar{\omega}_Q(t)$, in words:

for any h', how much smaller $R_Q(h)$ can be than $R_Q(h')$

for all h that are at most t-better on all benchmarks



- $\bar{\omega}_Q(0) = 0 \Leftrightarrow$ monotonicity holds and
  $\bar{\omega}_Q(0) > 0$ is "approximation error" of Pareto set
- how fast $\bar{\omega}_Q(t)$ grows with $t \approx$ "rate" of R_Q improvement with benchmarks

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

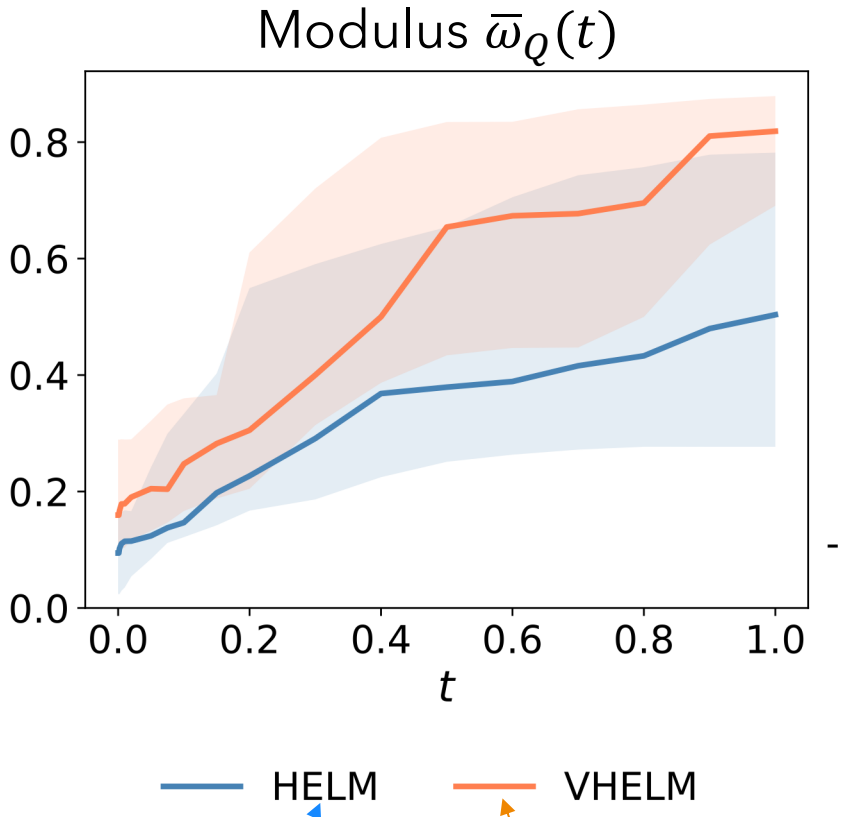
Modulus of monotonicity: a quantification

Intuition II: If a model is a lot better on some benchmarks but not much better/worse on others, it can't be much better on a new task (think of holistic assessments of PhD candidates)

Formalization: Encoded in size of modulus $\bar{\omega}_Q(t)$, in words:

for any h', how much smaller $R_Q(h)$ can be than $R_Q(h')$

for all h that are at most t-better on all benchmarks



- $\bar{\omega}_Q(0) = 0 \Leftrightarrow$ monotonicity holds and
  $\bar{\omega}_Q(0) > 0$ is "approximation error" of Pareto set
- how fast $\bar{\omega}_Q(t)$ grows with $t \approx$ "rate" of R_Q improvement with benchmarks

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

How many samples for model selection under (approx.) monotonicity?

Goal: If we believe the intuitions are true, **quantify** how much we can make use of it to reduce excess risk

Disclaimer: focus is statistical, not computational (method can be implemented for finite H so far)

Ansatz: (Approximate) monotonicity allows sample complexity reduction in two ways

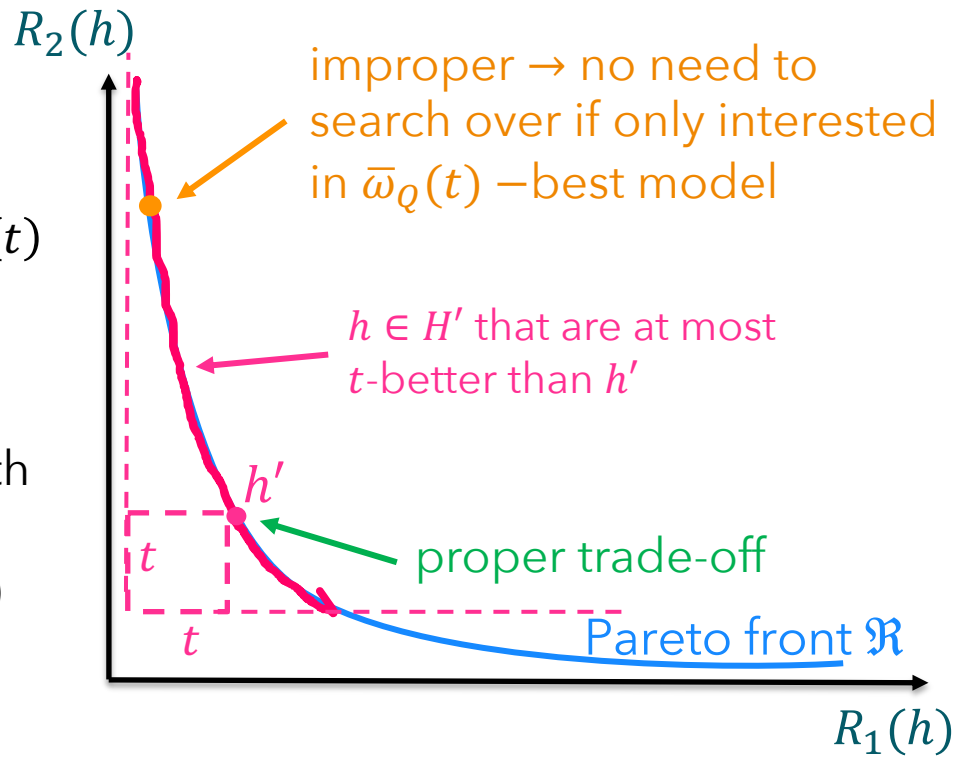
1. Minimizers on Pareto set → reduce search to complexity of Pareto set $Par(H, \boldsymbol{R}) \subseteq H$
2. A Pareto front geometry-aware Pareto cover and complexity measure for transfer

The modulus and proper trade-offs

We can shrink search in H to search in $H' = Par(H, \boldsymbol{R})$.

- In order to find a $\bar{\omega}_Q(t)$ −best model in H' we can reduce search to an *even smaller subset*, using the definition of $\bar{\omega}_Q(t)$
- If we include some models with proper trade-offs (little gain/loss ratio) → no need to include many models with improper trade-offs (at most $\bar{\omega}_Q(t)$ better on R_Q (intuition II))

Next: Formalization using Pareto covers



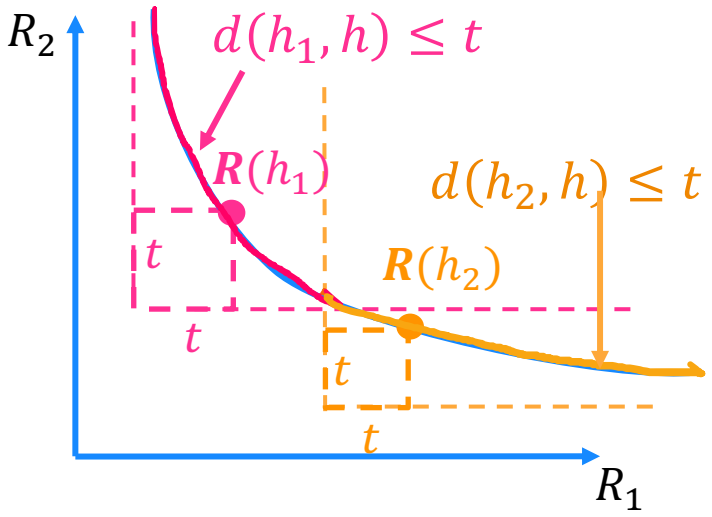
*Modulus of monotonicity extends beyond modulus of transfer (K=1) [Hanneke-Kpotufe '24]

A distance & complexity measure that adapts to Pareto front geometry

- Complexity of learning in a set H' is generally captured by size of a t-cover defined by a distance d:

  How many (finite) points $h_j \in H'$ do I need s.t. that for any $h \in H'$, I have $d(h_j, h) \leq t$?

- The geometry of "balls" in previous slide corresponds to a distance function d (definition omitted)



Definition (t-Pareto set and covering number)

$h_1, \ldots, h_N$ is t-Pareto set if $\forall h \in Par(H, \boldsymbol{R})$, $\exists h_j$ s.t. $d(h_j, h) \leq t$

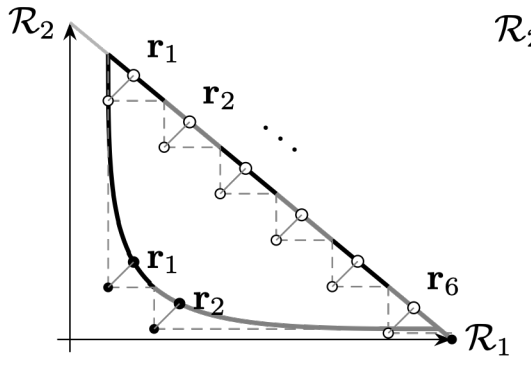
$N_{par}(t, H, \mathbf{R})$ denotes the smallest t −Pareto set size

Comparison with other natural coverings

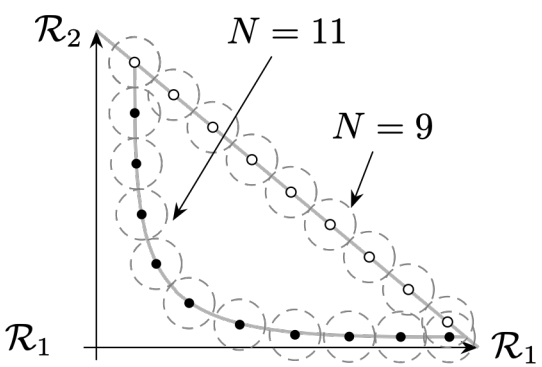
Our cover aligns with intuition II:

- focuses on models with proper tradeoffs instead of improper tradeoffs
- smaller when there are many improper tradeoffs we can ignore, larger when mostly proper trade-offs

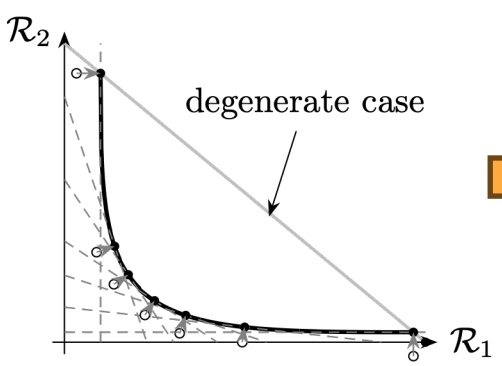
Visualizing different covers for two simple examples of competing risks with $K = 2$:



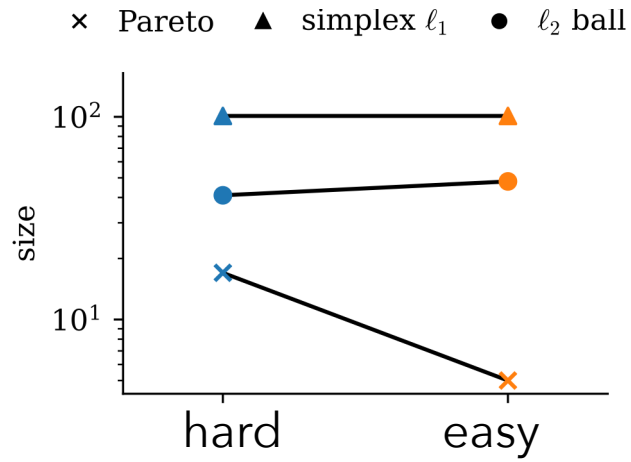
Pareto cover in 2d (ours)



ℓ_2-cover in 2d



simplex cover



👎 includes many improper tradeoffs

👎 size independent of Pareto front

Pareto-geometry aware transfer algorithms

This notion of cover and its limiting distribution μ naturally give rise to algorithms that *"hedge"* on points with good trade-offs (Pareto Exponential Weights in paper)

Algorithm (Pareto ERM)

1. Build minimal t −Pareto set

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

2. Compute ERM on

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

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

For estimator $\hat{h}$, for bounded losses, we have w.p. at least $1 - \delta$

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

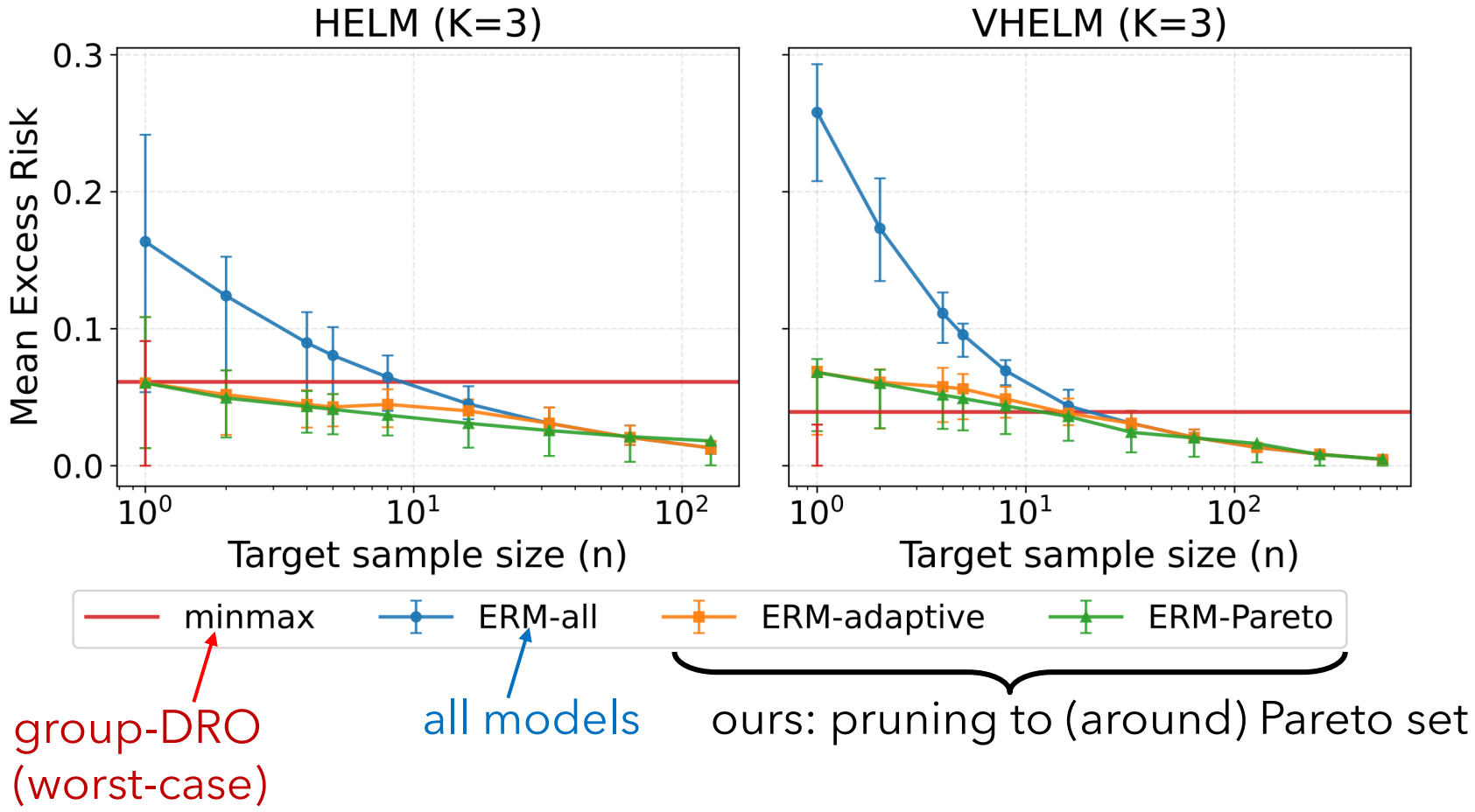
$$\leq \underbrace{\bar{\omega}_Q(t)}_{\text{scale-dependent approximation error}} + O\left(\sqrt{\frac{\log(N_{par}(t,H,\boldsymbol{R})/\delta)}{n}}\right)$$

Comparison with other works:

Without reducing to Pareto set: $\sqrt{\log(|H|/\delta)}$

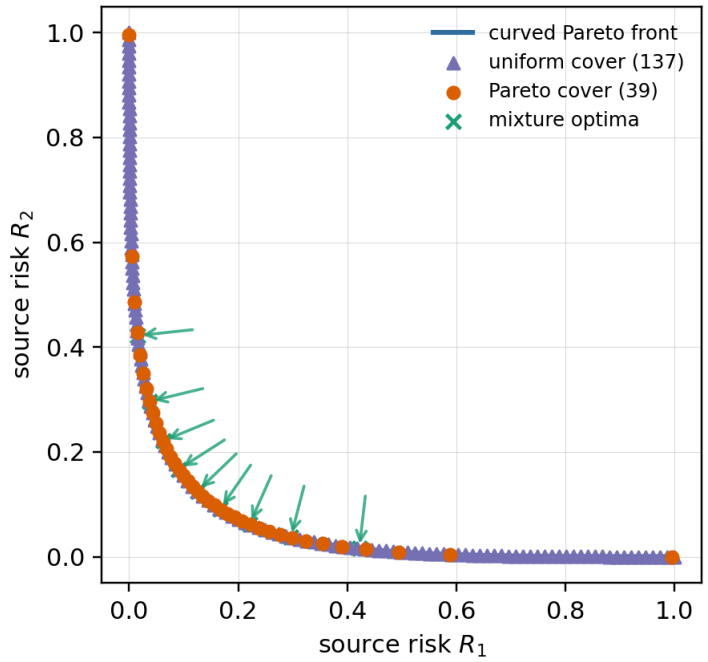
Generic ℓ_p-cover $\sim(1/t)^{K-1} \gg N_{par}(t, H, \boldsymbol{R})$

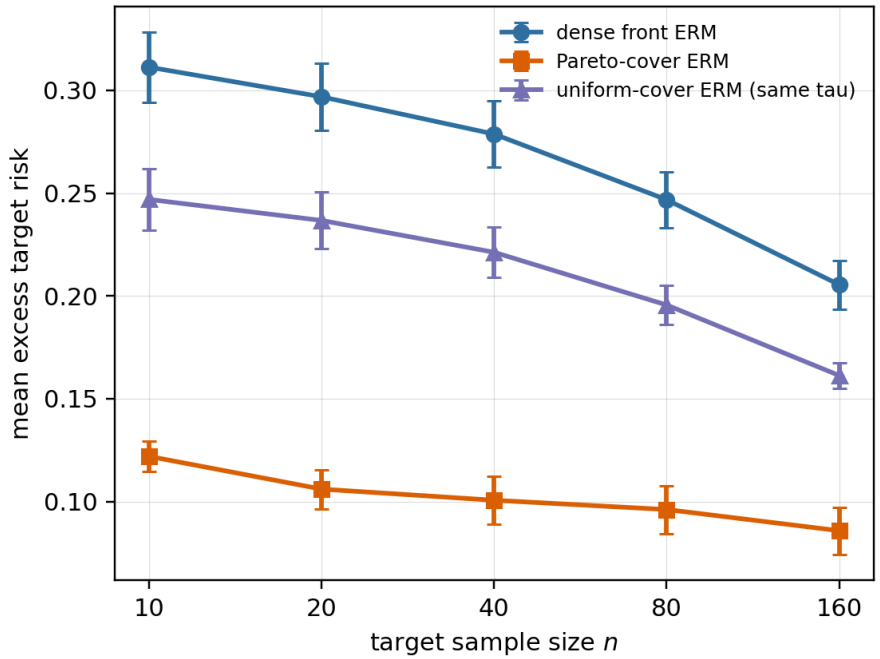
Finite H: Efficiency gain from only searching in Pareto set



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

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





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

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

Next step: Efficient computation of Pareto covers for usage on real data

Takeaway and references

- In the case of monotonicity, we can focus learning on the Pareto set
- The more "proper trade-offs" the harder learning is

- 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

Sample complexity of transfer with invariances

- 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