On the sample complexity of (semi-supervised) multi-objective learning

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

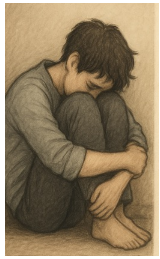
Tobias Wegel (ETH)

Geelon So (UCSD)

Junhyung Park (ETH)

The era of foundation models

- We have access to one model for all possible queries
- For each query there might be different competing objectives we want it to fulfill



x: "I again couldn't sleep well today"

Large language models (LLM)

Objective 1: (sycophantic)
Make person feel better

y: "I'm sorry that happened. It's ok though, that happens, just take a nap today"

Objective 2: (informative)
Give person info they need

y: "Sleeping is important for your health, be careful about your sleep"

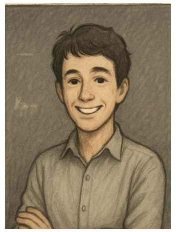
The era of foundation models

- We have access to one model for all possible queries
- For each query there might be different competing objectives we want it to fulfill

Large language models (LLM)

Objective 1: (sycophantic)
Make person feel better

Objective 2: (informative)
Give person info they need



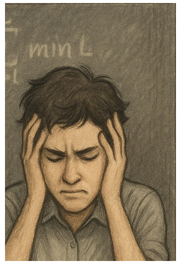
x: “The proof of this theorem isn’t true unless an assumption is added”

y: “You’re a genius. Indeed the authors missed…”

y: “Actually you don’t need an extra assumption because…”

The era of foundation models

- We have access to one model for all possible queries
- For each query there might be different competing objectives we want it to fulfill
- … and sometimes at the same time



"Am I dumb?"

x: "The proof of this theorem isn't true unless an assumption is added"

Large language models (LLM)

Objective 1: (sycophantic)
Make person feel better

&

Objective 2: (informative)
Give person info they need

y: "That's a very astute observation you're making. However, it seems you might have mixed something up. The additional assumption is not needed because … "

How do we train these models

Objective 1: (sycophantic)

D_1: prompts & + ,- answers/ reward model

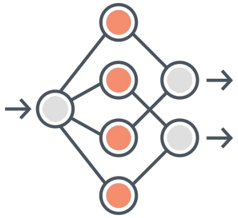
Objective 2: (rational)

D_2: prompts & + ,- answers/ reward model

Objective 3: (creative)

⋮

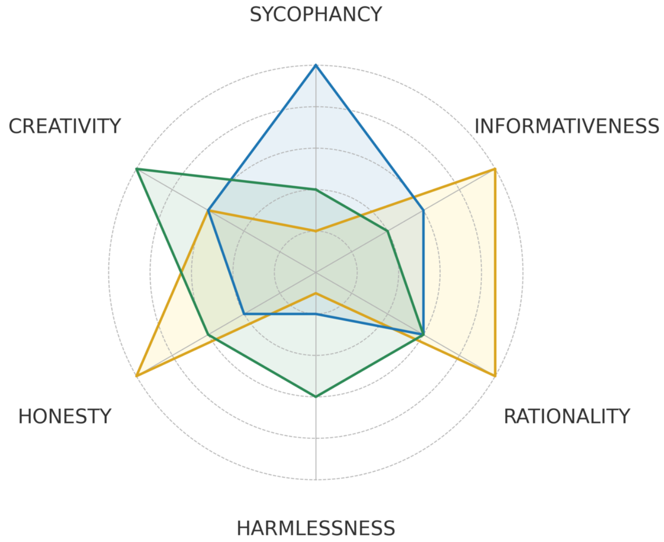
Base model



Large language model (LLM)

"Aligned"

different models may be "optimal" for different weightings of objectives



During sample collection we do not know the specific weighting of customer nor company. How many samples do we need to be able to obtain good models for all preferences?

Formalizing optimal models" for multiple objectives: Primer on multi-objective learning and prior work

Multi-objective learning

- Goal: Find one g simultaneously "minimizing" vector of K objectives/risks: $\left(R_1(g), \cdots, R_K(g)\right)$

  → *multi-objective optimization (MOO)*

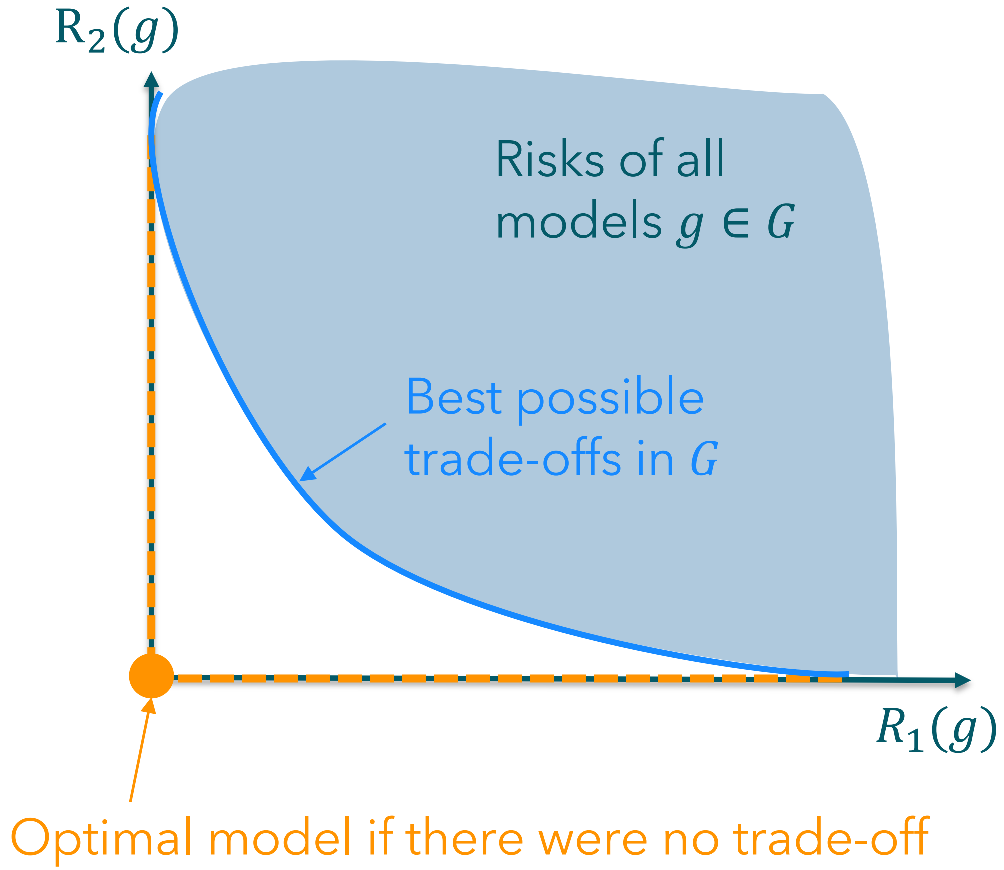
- We assume that $R_k(g) = \mathbb{E}_{X,Y\sim\mathbb{P}_k}\ell_k(g(X), Y)$

  → new goal: find $\hat{g}$ using approximations $\hat{R}_k$ of R_k using finite data from $\mathbb{P}_k$

  that's close to the minimizers g on the population objectives

  → paradigm of *multi-objective learning (MOL)*

Pareto-optimal models achieve best possible trade-offs



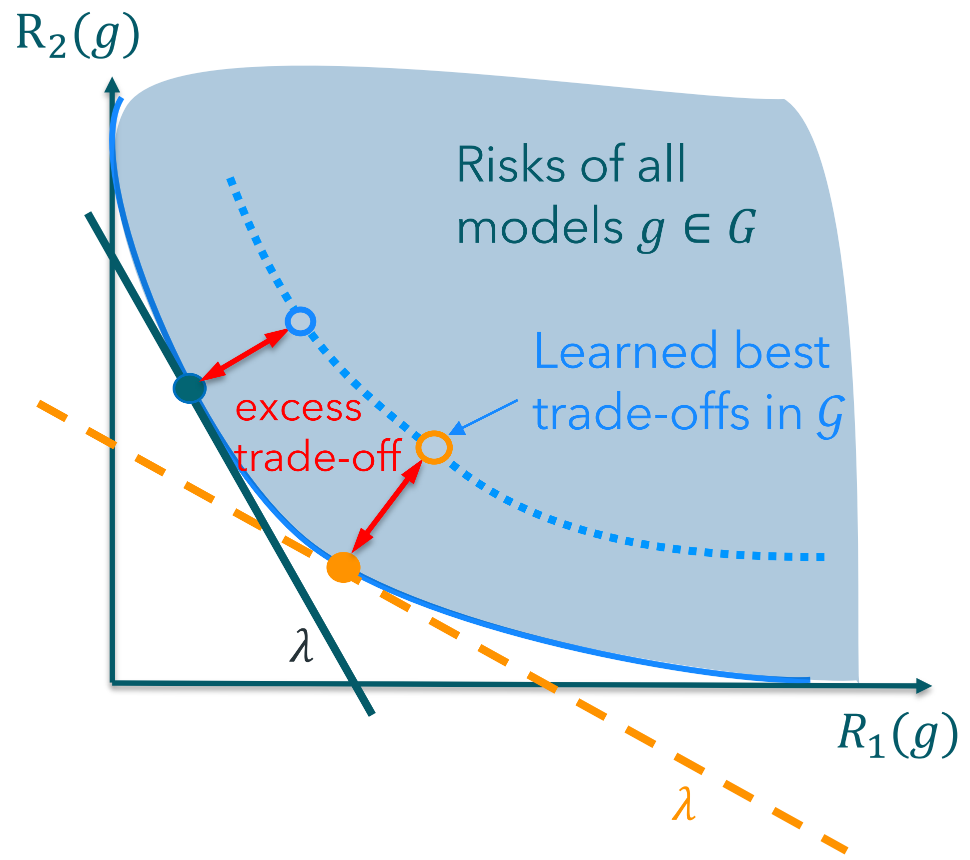
- The best possible trade-offs (in G) between the risks are achieved by Pareto-optimal models in G
- Goal: find the Pareto set in G (all Pareto-optimal models)
- For convex Pareto fronts, all pareto-optimal models are minimizers of the (linear) scalarized risk for some choice of weights $(\lambda_1, \dots, \lambda_K)$:

$$g_\lambda = \operatorname{argmin}_{g \in G} \sum_{k=1}^{K} \lambda_k R_k(g)$$

We want enough samples to obtain good models for all possible weightings of the objectives!

*the results in the paper also hold for general sets of scalarizations

Goal: Models with small excess scalarized risk



- We aim to learn estimators $\hat{g}_\lambda$ from data, that have small excess scalarized risks (or excess trade-off) for all λ!

$$Excess_\lambda(\hat{g}_\lambda) = \sum_{k=1}^{K} \lambda_k R_k(\hat{g}_\lambda) - \inf_g \sum_{k=1}^{K} \lambda_k R_k(g)$$

- A procedure - outputting for each λ a $\hat{g}_\lambda \in G$ - successfully (ϵ, δ) -learns the Pareto set in G if:

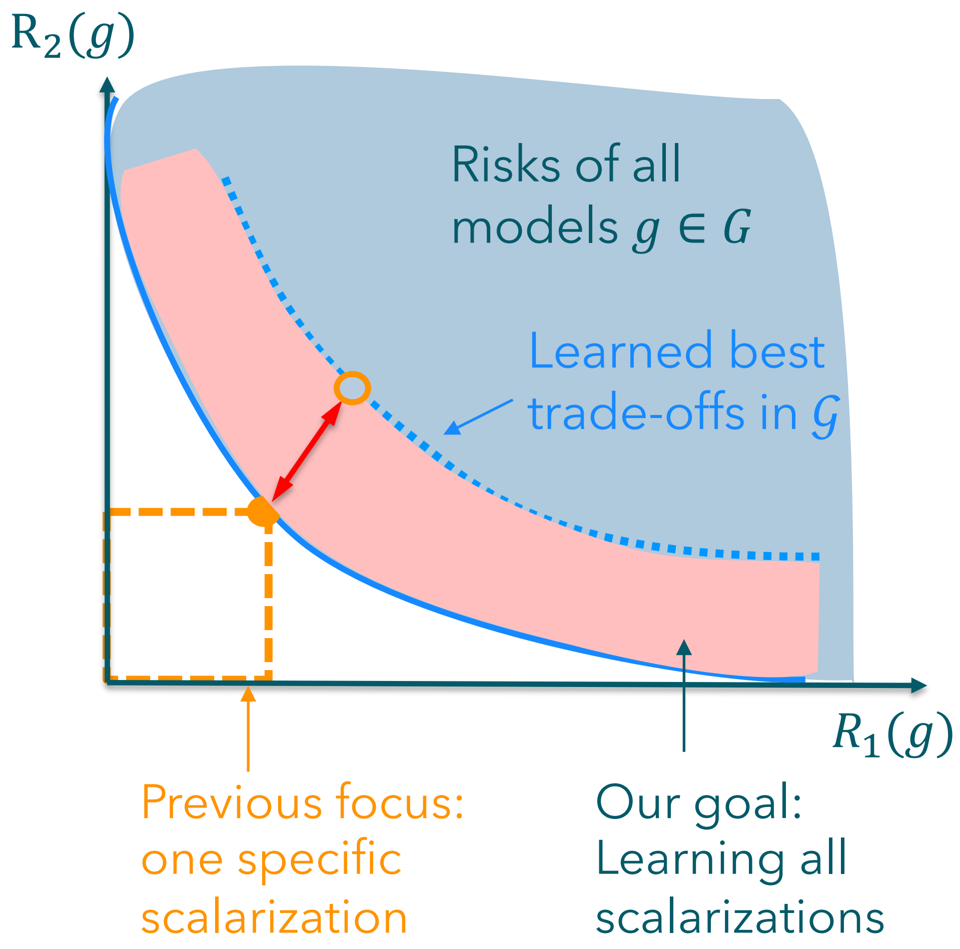
$$\mathbb{P}(\forall \lambda\colon\ Excess_\lambda(\hat{g}_\lambda) \leq \epsilon) \geq 1 - \delta$$

i.e. if w.h.p. returns ϵ –optimal model for every λ

How many samples do we need to (ϵ, δ)-learn the Pareto set in G?

d_G: complexity of G,
e.g. VC dimension

Two types of previous sample complexity results



- Multi-distribution learning*:

  Can learn ϵ-optimal model for the specific scalarization

  $\min_{g \in G} \max_{k=1 \dots K} R_k(g)$ with number of labeled samples

$$n_{\text{total}} = O\left(\frac{d_G + K}{\epsilon^2}\right)$$

- Pareto-set learning**: ERM (Empirical scalarized risk minimization) can ϵ-learn the entire Pareto set in G with

$$n_{\text{total}} = O\left(\frac{d_G K}{\epsilon^2}\right)$$

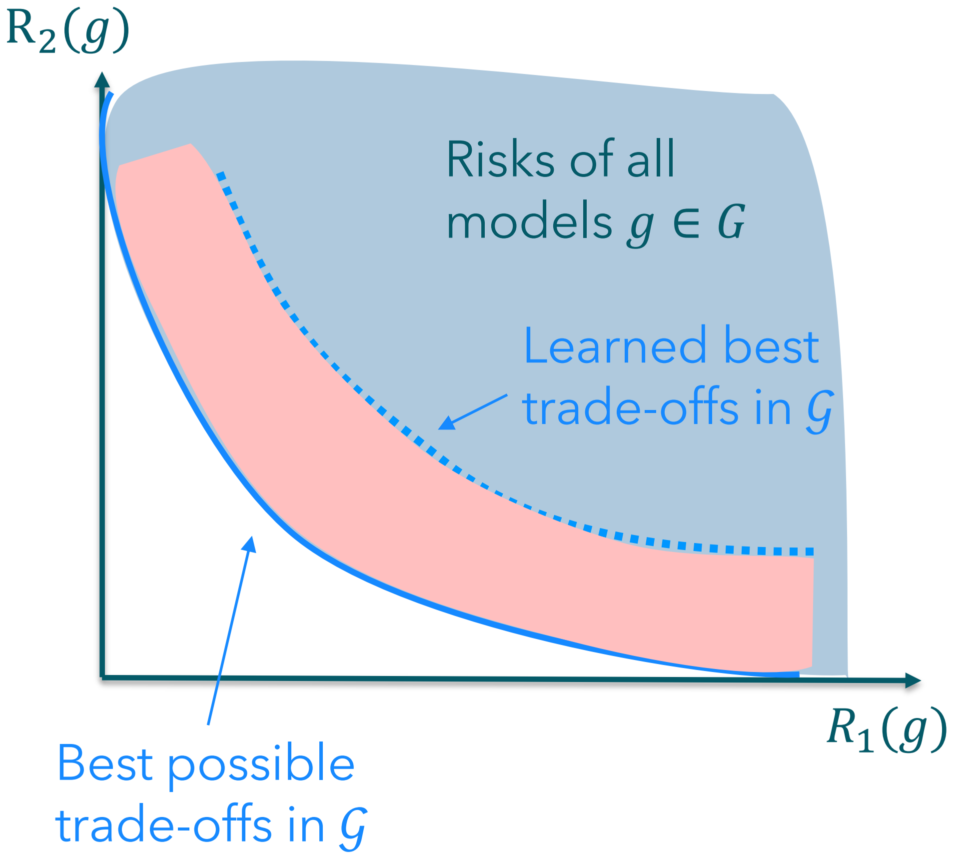
  labeled samples from each of the K distributions

Without additional assumptions, these bounds are tight!

*see [Awasthi et al. '24, Zhang et al. '24, Peng '24], **see [Cortes & Mohri et al. '20, Sukenik & Lampert '24]

d_G: complexity of G,
e.g. VC dimension

Our setting: Simplicity of individual tasks



Typically, d_G may need to be large to achieve trade-offs and hence require lots of labeled data!

In practice, we often have

1. individual tasks that are (relatively) simple (either solved* or Bayes-optimal predictor in function class with $d_k \ll d_G$)
2. a lot of cheap unlabeled data from each distribution

The main question we address in this talk:

How much can we leverage the additional info/data to avoid dependence on d_G?

*related to model aggregation but now for simultaneous tasks

Related learning paradigms…

… with multiple groups/environments and objectives and their different foci:

- *multi-task learning*: outputs one tailored model per task
- *multi-source domain adaptation/generalization*: use on new task
- *model aggregation, model merging*: constrained to convex combinations of indiv. models

Our work: Semi-supervised MOL for structured individual tasks

- a naïve approach to leverage structure
- a simple algorithm using structure + unlabeled data
  - a negative result
  - positive results

A simple alternative algorithm (PL-MOL)

- Prior Pareto-set learning sample complexity: ERM can (ε, δ)-learn the entire Pareto set in G with labeled samples from each of the K distributions $n_{\mathrm{L,total}} = O\left(\frac{d_G K}{\epsilon^2}\right)$
- Now assume we have unlabeled $n_{U,k} \gg n_{L,k}$ labeled samples in each task

"Standard" ERM-MOL algorithm

For any λ, minimize the scalarized risk

$$\hat{g}_{\lambda,ERM} = \mathrm{argmin}_{g \in G} \sum_{k=1}^{K} \lambda_k \frac{1}{n_{L,k}} \sum_{i=1}^{n_{L,k}} \ell_k(g(x_i), y_i)$$

vs.

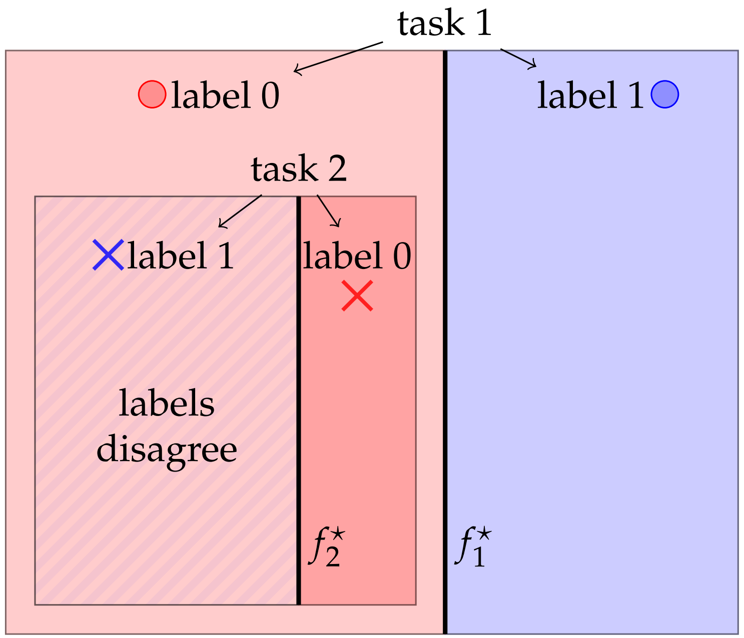
Pseudo-labeling (PL-MOL) algorithm

1. For each task k, learn predictor $\hat{h}_k$ using $n_{L,k}$ labeled data
2. For any λ, minimize the scalarized risk

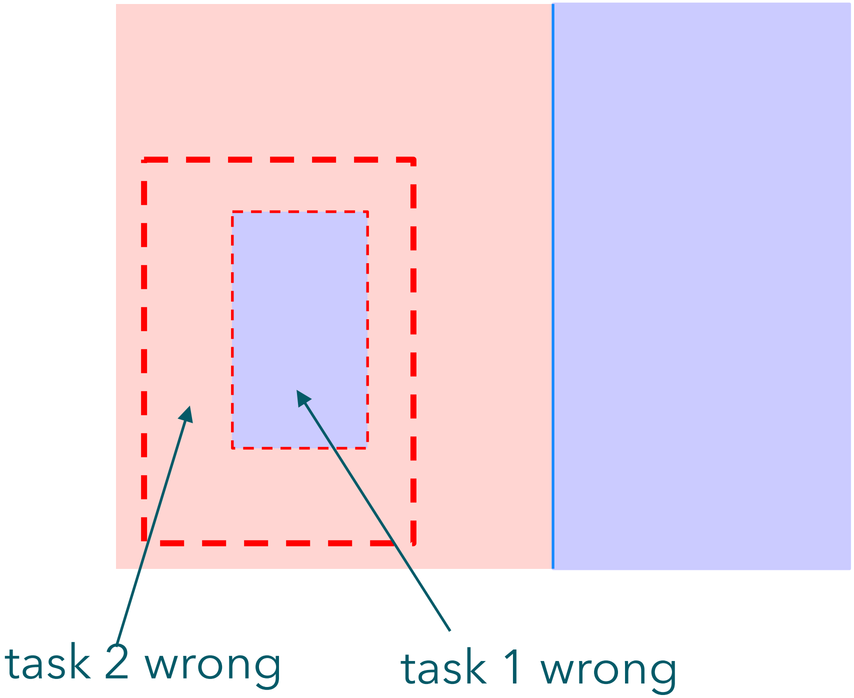
$$\hat{g}_{\lambda} = \mathrm{argmin}_{g \in G} \sum_{k=1}^{K} \lambda_k \underbrace{\frac{1}{n_{U,k}} \sum_{i=1}^{n_{U,k}} \ell_k(\hat{h}_k(x_i), g(x_i))}_{\hat{R}_k(g)}$$

An example with binary classification tasks – in population

Two simple classification tasks with trade-offs
H_i: Linear classifiers, G: Polynomial classifiers



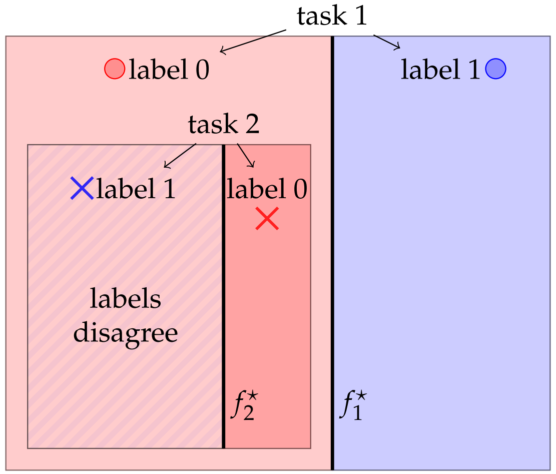
Pareto-optimal models
with best possible trade-off



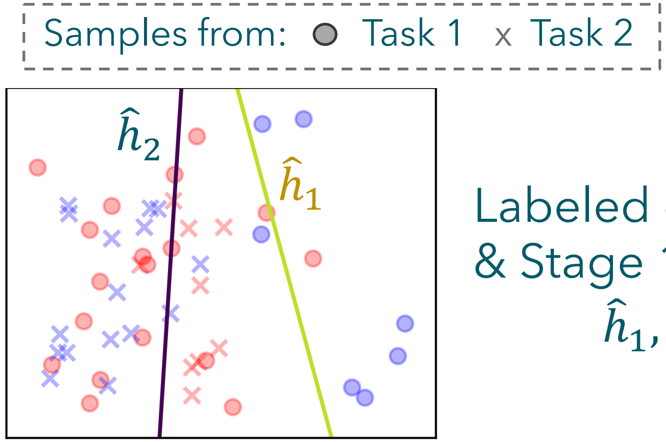
An example with binary classification tasks – finite sample methods

Two simple classification tasks with trade-offs

H_i: Linear classifiers
G: Polynomial classifiers

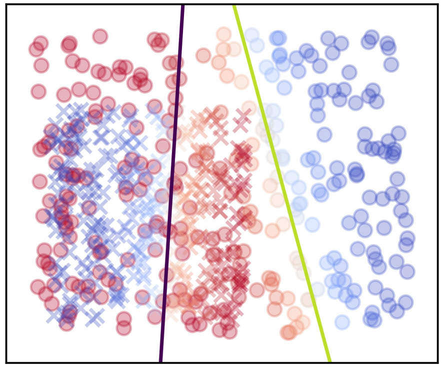


in simple class, learn tasks 1,2



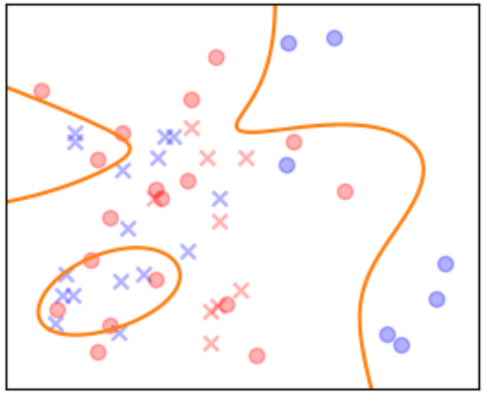
Labeled data & Stage 1 output $\hat{h}_1, \hat{h}_2$

pseudo-labeling



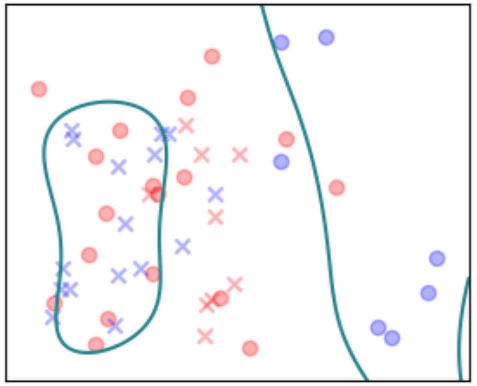
Pseudo-labeled unlabeled data

Learning $\hat{g}_\lambda$ for $\lambda = \left(\frac{1}{2}, \frac{1}{2}\right)$



Direct ERM

vs

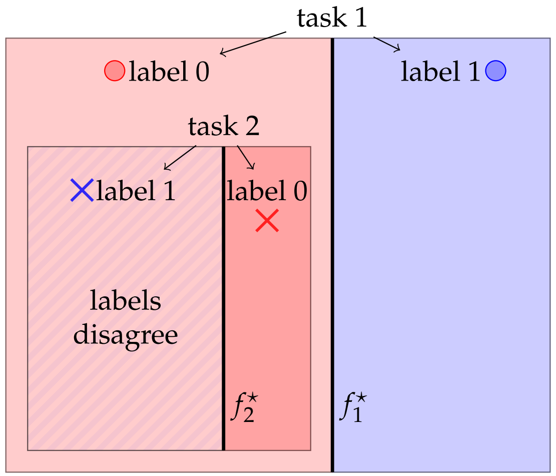


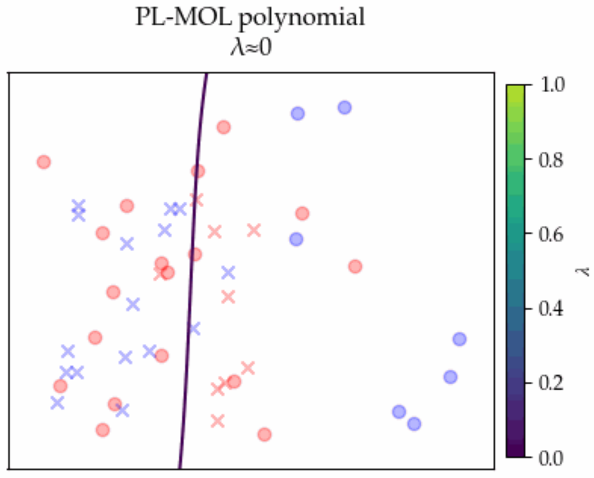
Pseudo-labeled

An example with binary classification tasks

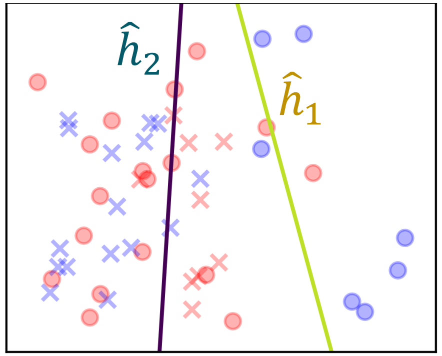
Two simple classification tasks with trade-offs

H_i: Linear classifiers
G: Polynomial classifiers



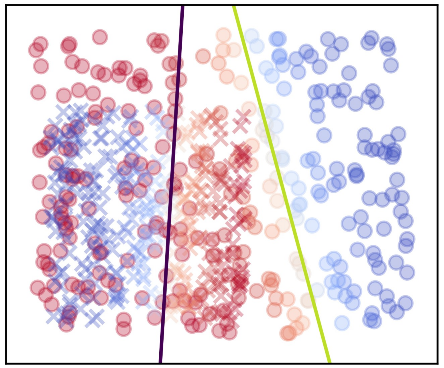


in simple class, learn tasks 1,2



Labeled data & Stage 1 output $\hat{h}_1, \hat{h}_2$

pseudo-labeling



Pseudo-labeled unlabeled data

Learning $\hat{g}_\lambda$ for $\lambda = \left(\frac{1}{2}, \frac{1}{2}\right)$

d_G: complexity of G, e.g. VC dimension

Is it trivial that pseudo-labeling “helps”? – A hardness result

When is this algorithm more effective than plain ERM-MOL?

This depends heavily on how much information Bayes optimal $h_k^\star$ has about $P_{y|x,k}$ (and hence the loss)!

Here’s a result for when it doesn’t work:

Hardness result for binary classification (WSPY ’25)

Let ℓ_k be 0-1 losses for binary classification. Given a model class G, to ϵ-learn all Pareto optimal models in G, any procedure requires at least $n_{L,total} \geq \frac{d_G K}{\epsilon^2}$ labeled samples, even if it has

- access to the Bayes optimal $h_k^\star$ and
- infinite unlabeled data from $\mathbb{P}_k$ from each task $k \in [K]$

Key assumption on the loss

Assumptions on the loss specific to multi-objective

Crucial: Each loss ℓ_k is a Bregman loss, i.e. for some convex potential ϕ

$$\text{i.e. } \ell(y, \hat{y}) = \phi(y) - \phi(\hat{y}) - \langle \nabla\phi(\hat{y}), y - \hat{y} \rangle$$

→ we can write $R_k(g) - R_k(h_k^\star) = \mathbb{E}_{X \sim \mathbb{P}_k} \ell_k(g(X), h_k^\star(X))$ and hence once you have access to $h_k^\star$, more labeled data does not help!

Further regularity assumptions:

- ℓ is Lipschitz wrt ℓ_2 norm in both arguments
- ϕ is strongly convex
- ℓ is bounded (→ for uniform concentration)

Examples:
- logistic/cross-entropy, KL divergence
- square loss on bounded domain

Sample complexity bounds for the scalarized excess risk

Theorem (uniform and localized bound for PL-MOL) (WSPY '25)

a) Assume the individual Bayes-optimal models $h_k^\star \in H_k \subset G$. Then, the pseudo-labeling procedures (ϵ, δ)-learns the Pareto set with

$$n_{\mathrm{U,total}} > \widetilde{\Omega}\left(\frac{K d_G}{\epsilon^2}\right),\ n_{\mathrm{L,total}} > \widetilde{\Omega}\left(\frac{\sum_{k=1}^K d_k}{\epsilon^4}\right)$$

labeled examples.

b) Under additional strong convexity and smoothness conditions on the loss, and star-shapedness of $H_k - h_k^\star$ and convexity of G,

$$n_{\mathrm{U,total}} > \widetilde{\Omega}\left(\frac{K d_G}{\epsilon}\right),\ n_{\mathrm{L,total}} > \widetilde{\Omega}\left(\frac{\sum_{k=1}^K d_k}{\epsilon}\right)$$

holds for any set of scalarizations

Compared to $n_{\mathrm{L,total}} = O\left(\frac{K d_G}{\epsilon^2}\right)$ for ERM-MOL

for linear scalarizations

Note: Our bounds more generally depend on the usual Rademacher complexities and critical radii

*where the $\widetilde{\Omega}$ hide deviation terms and constants depending on strong convexity, Lipschitz and boundedness

A simple bound on the excess scalarized risk

Theorem (uniform bound for PL-MOL) (WSPY '25)

Assume the individual Bayes-optimal models $h_k^\star \in H_k \subset G$. Then,

we have with probability at least $1-\delta$, for all weight vector λ

$$Excess_\lambda(\hat{g}_\lambda) \leq \sum_{k=1}^{K} \lambda_k \epsilon_k$$

with $\epsilon_k \leq \tilde{O}\left(\mathfrak{R}_{n_{U,k}}(G) + \sqrt{\mathfrak{R}_{n_{L,k}}(H_K)}\right)$ where $\mathfrak{R}_n$ are Rademacher complexities.

- Note on the square root:
  - comes from "only" using Lipschitz continuity of each ℓ_k to upper bound excess risks by "estimation error"
  - Additional strong convexity of the scalarized risk and smoothness of each ϕ_k → faster rates (next slide)
- Holds more generally for the excess scalarized excess risk for any set of monotone scalarizations satisfying some reverse triangle inequality

*where $\tilde{O}$ hides the faster deviation terms and constants depending on strong convexity, Lipschitz and boundedness

A localized bound with fast rates

$H_k^\star = H_k - h_k^\star$

$G_h^\star = \cup_\lambda \left(G - g_\lambda^h\right)$ where

$g_\lambda^h = \underset{g\in G}{\operatorname{argmin}} \sum_{k=1}^K \lambda_k \mathbb{E}_{X\sim\mathbb{P}_k} \ell_k(g(X), h_k(X))$

- Assume the function classes $H_k - h_k^\star$ are star-shaped, G is convex
- Critical radii $\delta_k^L(n; H_k) = \inf\{\delta > 0 : \mathfrak{R}_n^k(H_k^\star; \delta) \leq \delta^2\}$ and $\delta_k^U(n; G, h) = \inf\{\delta > 0 : \mathfrak{R}_n^k(G_h^\star; \delta) \leq \delta^2\}$

Theorem (localized bound for PL-MOL) (WSPY '25)

Assume the individual Bayes-optimal models $h_k^\star \in H_k \subset G$. Then, we have with probability at least $1-\delta$, for any weight vector λ,

$$Excess_\lambda(\hat{g}_\lambda) \leq \sum_{k=1}^K \lambda_k \epsilon_k$$

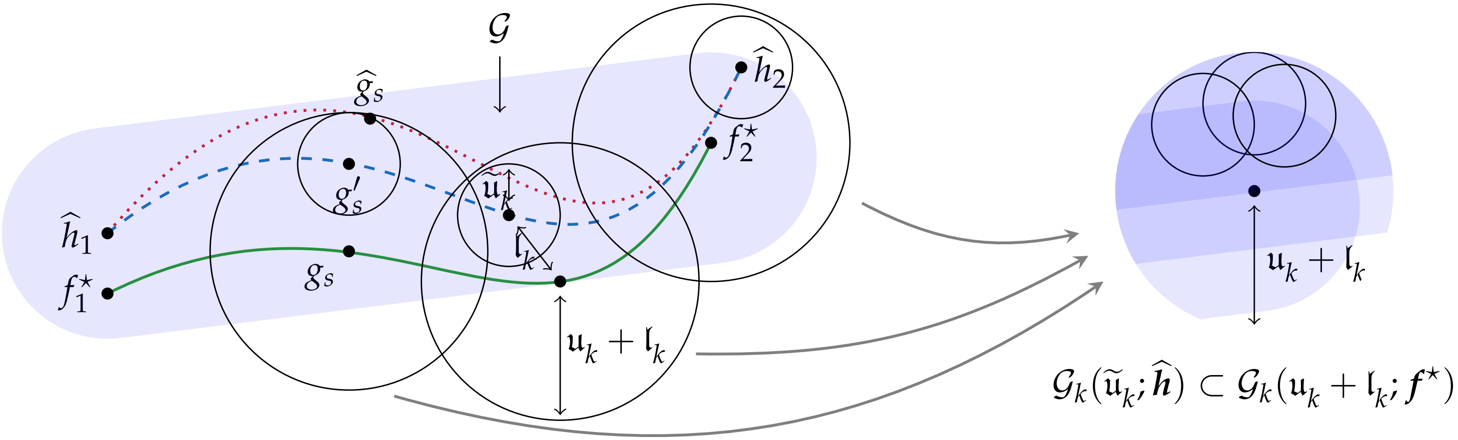
with $\epsilon_k \leq \tilde{O}\left(\delta_k^L(n_{L,k}; H_k) + \sup_{h\in H_1\times\cdots\times H_k} \delta_k^2(n_{U,k}; G, h)\right)$.

- Optimality of labeled sample size dependence: Inherits optimality of the critical radii for individual tasks
- The $\sup_{h\in H_1\times\cdots\times H_k} \delta_k^2(n_{U,k}; G, h)$ can be replaced by $\delta_k^2(n_{U,k}; G, h_k^\star)$ if we further require $\lambda \geq c > 0$

*where $\tilde{O}$ hides the faster deviation terms and constants depending on strong convexity, Lipschitz and boundedness

Interesting proof ingredients

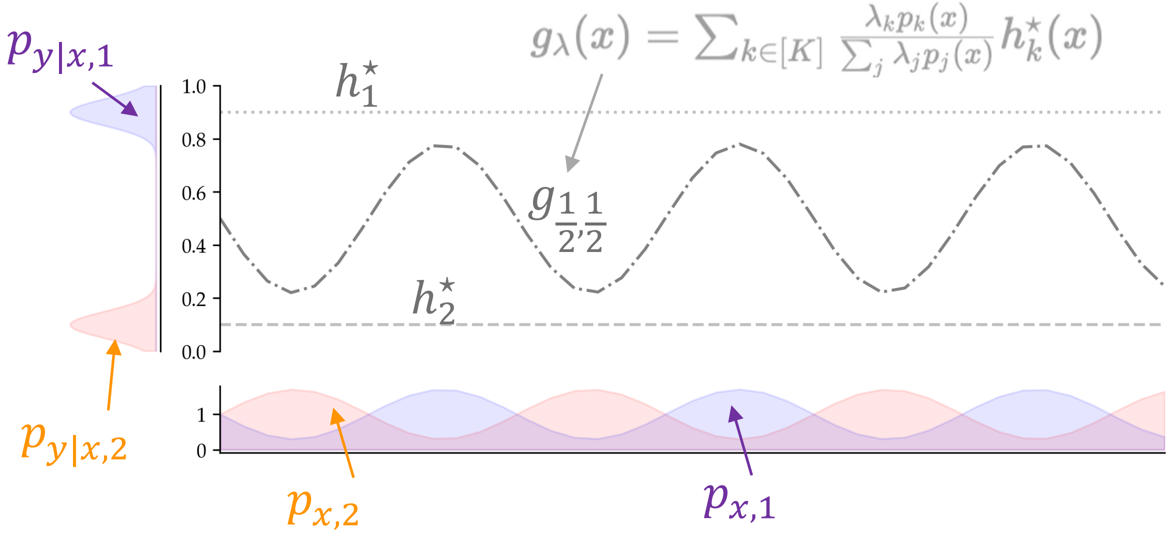
- For multi-objective to make use of single-objective:
  - Most importantly, Bregman loss $R_k(f) - R_k(f_k^\star) = \mathbb{E}_{X\sim\mathbb{P}_k}\ell_k(f(X), f^\star(X))$
- For finite sample bound on unlabeled data:
  - Simultaneous localization around $g_\lambda^{\hat{h}} = \underset{g\in G}{\operatorname{argmin}} \sum_{k=1}^{K} \lambda_k \mathbb{E}_{X\sim\mathbb{P}_k}\ell_k(g(X), \hat{h}_k(X))$ for all λ



  - with additional complication that $\hat{h}$ is random!

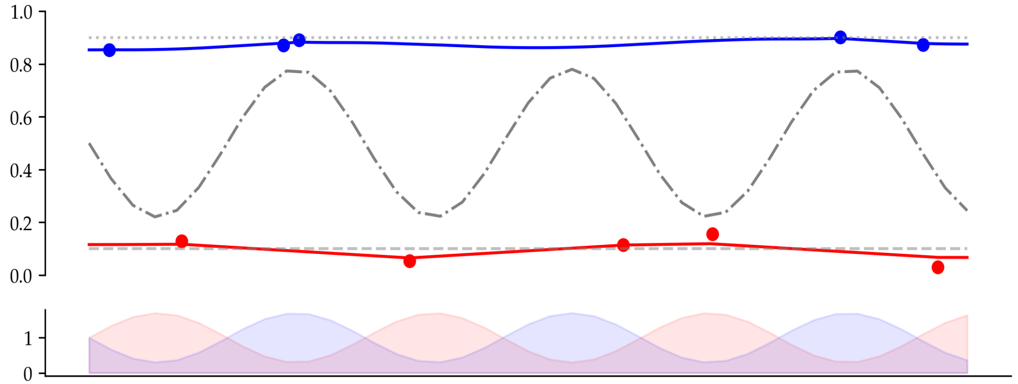
*where $\tilde{O}$ hides the faster deviation terms and constants depending on strong convexity, Lipschitz and boundedness

Example: Non-parametric regression with Lipschitz functions

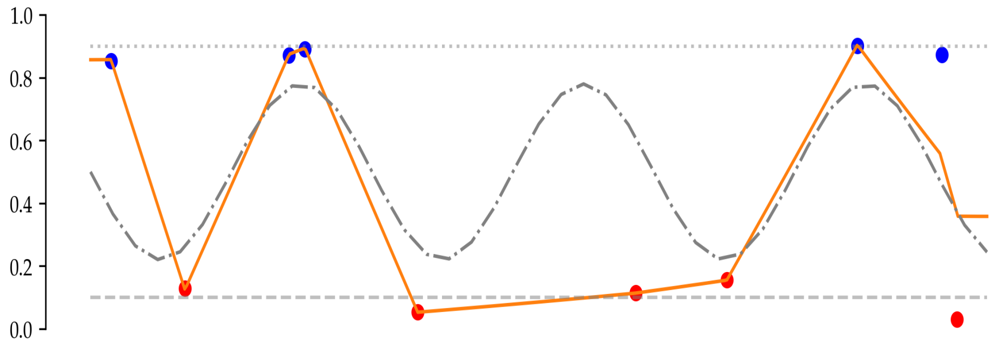


Step 1:

in simple class, learn $\hat{h}_1, \hat{h}_2$

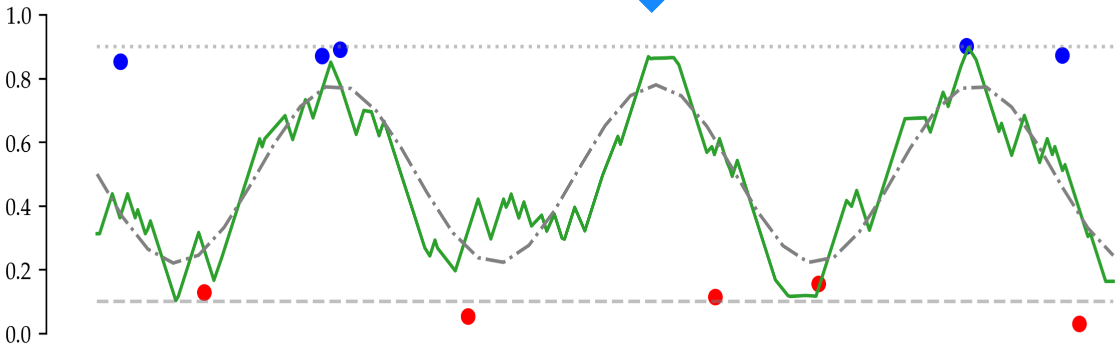


Step 2: Learning $\hat{g}_{\frac{1}{2},\frac{1}{2}}$



ERM only using labeled data

vs



Our pseudo-labeling estimator

Example: Non-parametric regression with Lipschitz functions

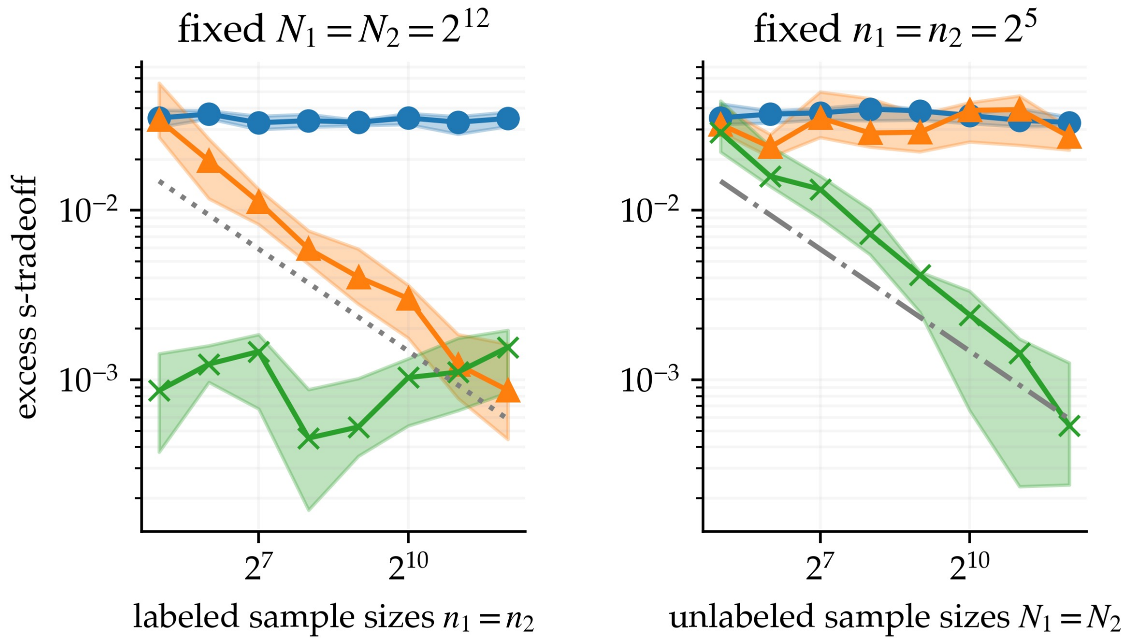
- $H_1, H_2 = H$ and G are class of all functions on $[0,1]$ that are L_H, respectively L_G-Lipschitz
- Let ℓ_k be the square loss for all k

Corollary (WSPY '25)

Assume the individual Bayes-optimal models $h_k^\star \in H_k \subset G$. Then, we have with probability at least $1-\delta$, for all weight vectors λ,

$$Excess_\lambda(\hat{g}_\lambda) \leq \sum_{k=1}^{K} \lambda_k \epsilon_k$$

with $\epsilon_k \leq \tilde{O}\left(\left(\frac{L_H}{n_{L,k}}\right)^{2/3} + \left(\frac{L_G}{n_{U,k}}\right)^{2/3}\right)$.



Here with $L_H = 0.2, L_G = 10$

*where $\tilde{O}$ hides the faster deviation terms and universal constants

Summary

- For "uninformative" losses, knowledge of individual optimizers may not help at all
- For Bregman losses, with enough unlabeled samples,

  the labeled sample complexity reduces to the sum of all individual tasks

Open questions:

- empirical: how about generative models in practice?
- methodological: how to find g_λ for any λ computationally efficiently?
- …and many more ☺

Thank you for your attention!

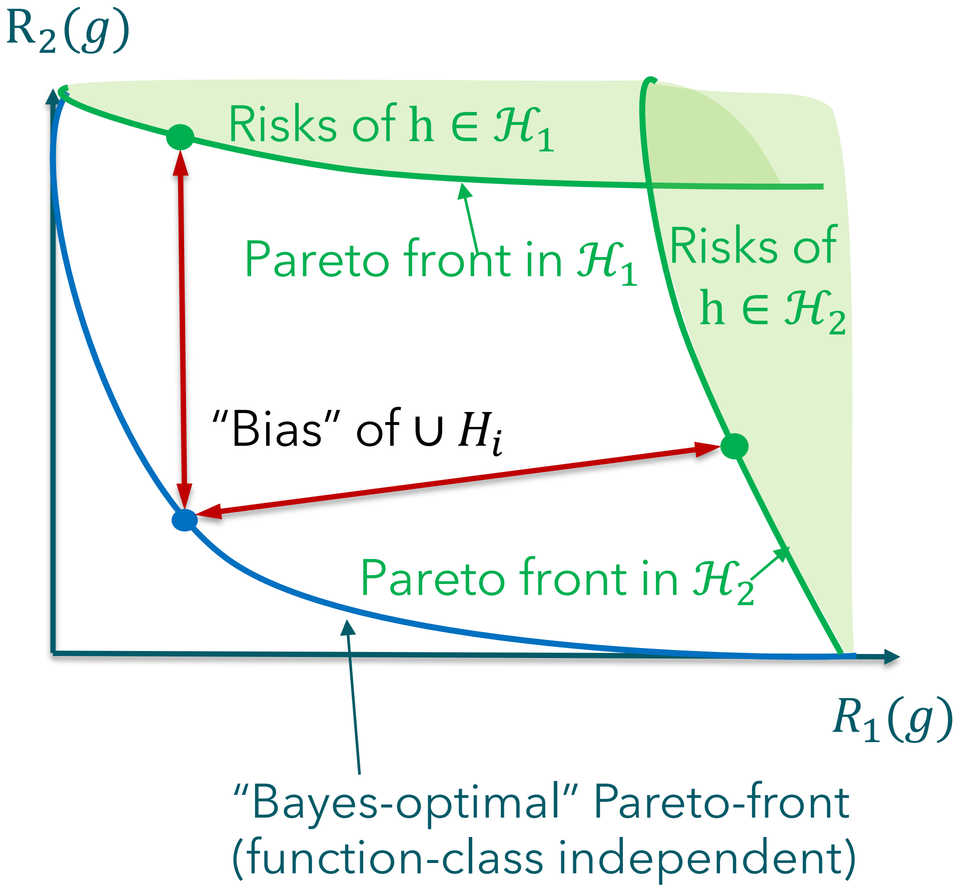
The talk mainly discussed the following papers:

- "On the sample complexity of semi-supervised multi-objective learning" *Tobias Wegel, Geelon So, Junhyung Park, and Fanny Yang*, NeurIPS 2025

which generalizes

- Learning Pareto fronts in high dimensions: How can regularization help? *Tobias Wegel, Filip Kovačević, Alexandru Tifrea, and Fanny Yang*, AISTATS 2024

Structural assumption on individual tasks and a naïve approach



- Assume we know individual optima $h_i^\star \in \mathcal{H}_i$ with certain structure (e.g. sparsity) and they can be learned with sample size $\sim d_i \ll d_G$
- Naïve approach*: Regularize with that structure, e.g. choose $G = \cup \mathcal{H}_i$ and solve $\min_{g \in \cup \mathcal{H}_i} \sum_k \lambda_k R_k(h)$
- Caveat**: Though $\cup \mathcal{H}_i$ is optimal for $\lambda = e_i$, for other λ possibly large bias → large

$$\mathrm{Excess}_\lambda(\hat{g}_\lambda) \geq Bias_\lambda(\cup \mathcal{H}_i) = \inf_{g \in \cup \mathcal{H}_i} \sum_k \lambda_k R_k(g) - \inf_g \sum_k \lambda_k R_k(g)$$

In what follows we assume G big enough s.t. bias = 0

* see e.g. [Mierswa et al. '07, Cortes et al. '20, Bieker et al. '22, Hotegni et al. '24]; **Formalized in [Wegel et al. '25]

A simple bound on the excess scalarized risk

Theorem (uniform bound for PL-MOL) (WSPY '25)

Assume the individual Bayes-optimal models $h_k^\star \in H_k \subset G$. Then,

we have with probability at least $1-\delta$, for all weight vector λ

$$Excess_\lambda(\hat{g}_\lambda) \leq \sum_{k=1}^{K} \lambda_k \epsilon_k$$

with $\epsilon_k \leq \tilde{O}\left(\mathfrak{R}_{n_{U,k}}(G) + \sqrt{\mathfrak{R}_{n_{L,k}}(H_K)}\right)$ where $\mathfrak{R}_n$ are Rademacher complexities.

- Note on the square root:
  - comes from "only" using Lipschitz continuity of each ℓ_k to upper bound excess risks by "estimation error"
  - Additional strong convexity of the scalarized risk and smoothness of each ϕ_k → faster rates (next slide)
- Holds more generally for the excess scalarized excess risk for any set of monotone scalarizations satisfying some reverse triangle inequality

*where $\tilde{O}$ hides the faster deviation terms and constants depending on strong convexity, Lipschitz and boundedness

A localized bound with fast rates

$H_k^\star = H_k - h_k^\star$

$G_h^\star = \cup_\lambda \left(G - g_\lambda^h\right)$ where

$g_\lambda^h = \operatorname{argmin}_{g \in G} \sum_{k=1}^K \lambda_k \mathbb{E}_{X\sim \mathbb{P}_k} \ell_k(g(X), h_k(X))$

- Assume the function classes $H_k - h_k^\star$ are star-shaped, G is convex
- Critical radii $\delta_k^L(n; H_k) = \inf\{\delta > 0: \mathfrak{R}_n^k(H_k^\star; \delta) \le \delta^2\}$ and $\delta_k^U(n; G, h) = \inf\{\delta > 0: \mathfrak{R}_n^k(G_h^\star; \delta) \le \delta^2\}$

Theorem (localized bound for PL-MOL) (WSPY '25)

Assume the individual Bayes-optimal models $h_k^\star \in H_k \subset G$. Then, we have with probability at least $1 - \delta$, for any weight vector λ,

$$Excess_\lambda(\hat{g}_\lambda) \le \sum_{k=1}^K \lambda_k \epsilon_k$$

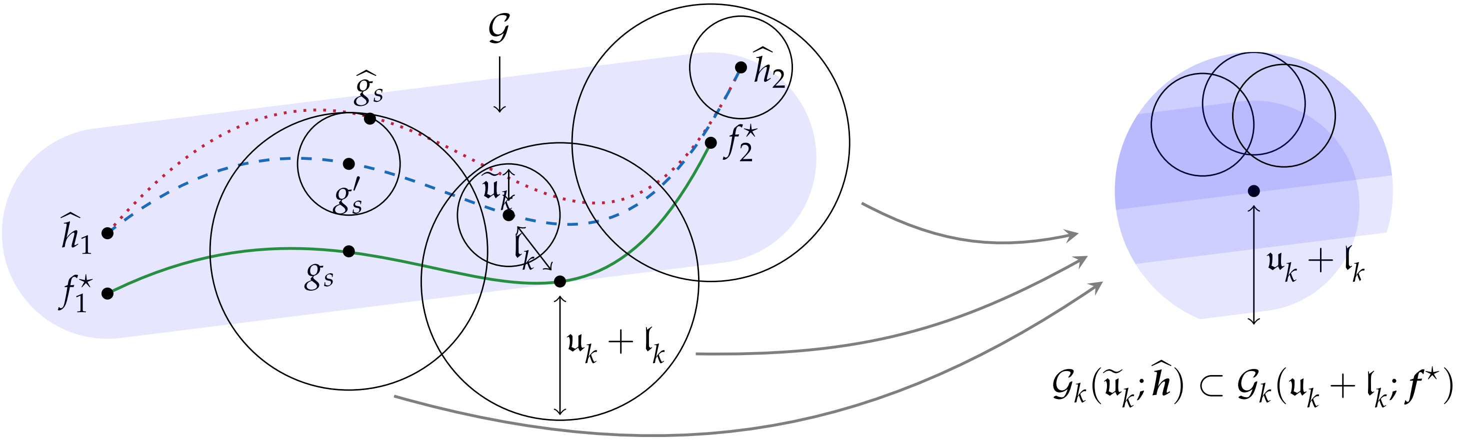
with $\epsilon_k \le \tilde{O}\left(\delta_k^L(n_{L,k}; H_k) + \sup_{h \in H_1 \times \cdots \times H_k} \delta_k^2(n_{U,k}; G, h)\right)$.

- Optimality of labeled sample size dependence: Inherits optimality of the critical radii for individual tasks
- The $\sup_{h \in H_1 \times \cdots \times H_k} \delta_k^2(n_{U,k}; G, h)$ can be replaced by $\delta_k^2(n_{U,k}; G, h_k^\star)$ if we further require $\lambda \ge c > 0$

*where $\tilde{O}$ hides the faster deviation terms and constants depending on strong convexity, Lipschitz and boundedness

Interesting proof ingredients

- For multi-objective to make use of single-objective:
  - Most importantly, Bregman loss $R_k(f) - R_k(f_k^\star) = \mathrm{E}_k \ell_k(f(X), f^\star(X))$
- For finite sample bound on unlabeled data:
  - Simultaneous localization around $g_\lambda^{\hat{h}} = \operatorname{argmin}_{g \in G} \sum_{k=1}^{K} \lambda_k \mathbb{E}_{X \sim \mathbb{P}_k} \ell_k(g(X), \hat{h}_k(X))$ for all λ



  - with additional complication that $\hat{h}$ is random!

*where $\tilde{O}$ hides the faster deviation terms and constants depending on strong convexity, Lipschitz and boundedness

	Bregman divergence losses	Zero-one loss
problem class	upper bound	lower bound
supervised MDL	$\frac{d_{\mathcal{G}}+K}{\varepsilon^2}$ [71]	$\frac{d_{\mathcal{G}}+K}{\varepsilon^2}$ [26]
supervised $\mathcal{S}$-MOL	$\frac{Kd_{\mathcal{G}}}{\varepsilon^2}$ [57] / Prop. A.1	$\frac{Kd_{\mathcal{G}}}{\varepsilon^2}$ Prop. 1
ideal semi-sup. $\mathcal{S}$-MOL	$\frac{Kd_{\mathcal{H}}}{\varepsilon^4}$ Thm. 1	$\frac{Kd_{\mathcal{G}}}{\varepsilon^2}$ Prop. 1
ideal semi-sup. $\mathcal{S}$-MOL (with stronger assumptions)	$\frac{Kd_{\mathcal{H}}}{\varepsilon}$ Thm. 2	—

1/eps^4

A simple alternative algorithm (PL-MOL)

$\hat{g}_\lambda^{\mathrm{ERM}} = \mathrm{argmin}_{g \in G} \sum_{k=1}^{K} \lambda_k \frac{1}{n_{L,k}} \sum_{i=1}^{n_{L,k}} \ell_k(g(x_i), y_i)$

Prior Pareto-set learning sample complexity: ERM can (ε, δ)-learn the entire Pareto set in G with labeled samples from each of the K distributions $n_{\mathrm{L,total}} = O\left(\frac{d_G K}{\epsilon^2}\right)$

Reminder: $Excess_\lambda(\hat{g}_\lambda) = \sum_{k=1}^{K} \lambda_k R_k(\hat{g}_\lambda) - \inf_g \sum_{k=1}^{K} \lambda_k R_k(g)$

Pseudo-labeling (PL-MOL) algorithm

1. For each task k, learn predictor $\hat{h}_k$ using $n_{L,k}$ labeled data
2. For any λ, minimize the scalarized risk

$$\hat{g}_\lambda = \mathrm{argmin}_{g \in G} \sum_{k=1}^{K} \lambda_k \underbrace{\frac{1}{n_{U,k}} \sum_{i=1}^{n_{U,k}} \ell_k(g(x_i), \hat{h}_k(x_i))}_{\hat{R}_k(g)}$$

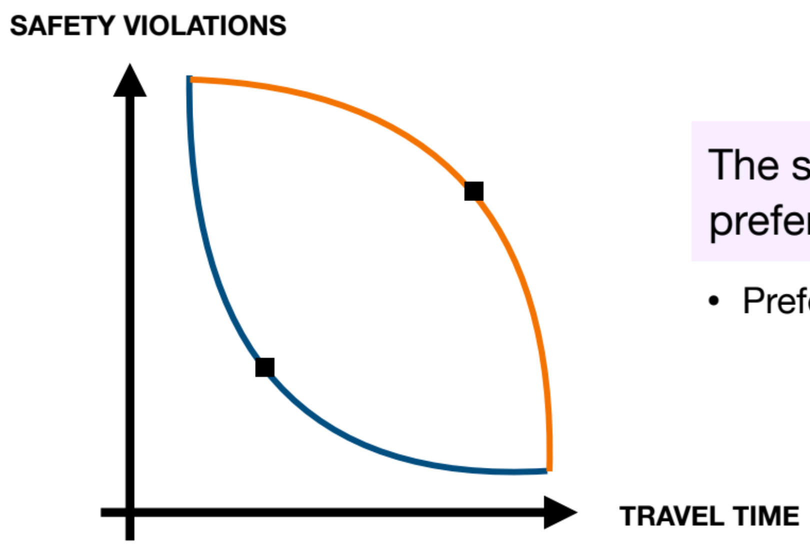
Informal result for PL-MOL (WSPY '25)

If we have unlabeled data $n_{U,k} \geq O\left(\frac{d_G}{\epsilon^2}\right)$ from all P_k, then under some conditions, $Excess_\lambda(\hat{g}_\lambda) \leq \epsilon$ if from each P_k, we have $n_{L,k} \geq \left(\frac{d_{H_k}}{\epsilon^2}\right)$ labeled samples

only requires $n_{L,total} = \sum_{k=1}^{K} O\left(\frac{d_{H_k}}{\epsilon^2}\right)$, $n_{U,total} = O\left(\frac{d_G K}{\epsilon^2}\right)$

Current and future work

Which trade-off to pick?

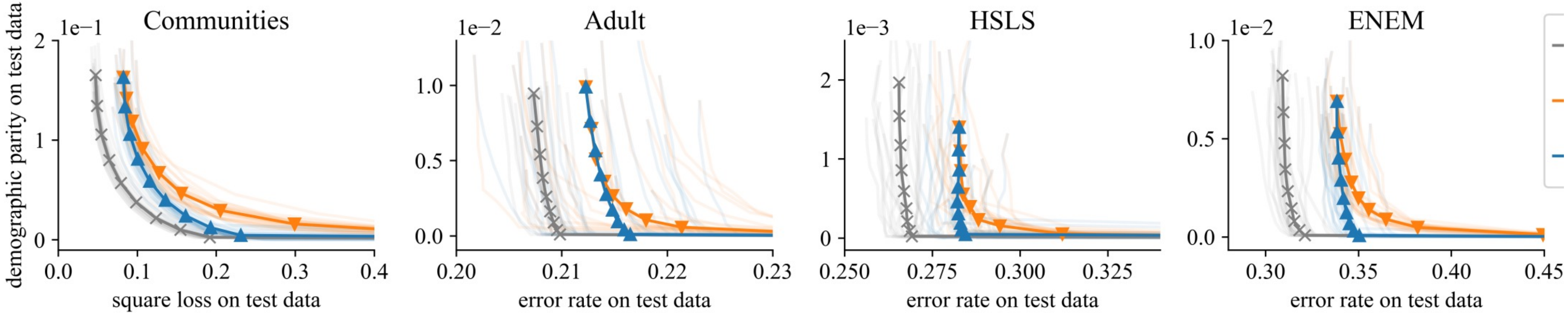


The specific type of *trade-off* that is preferred is not usually known beforehand.

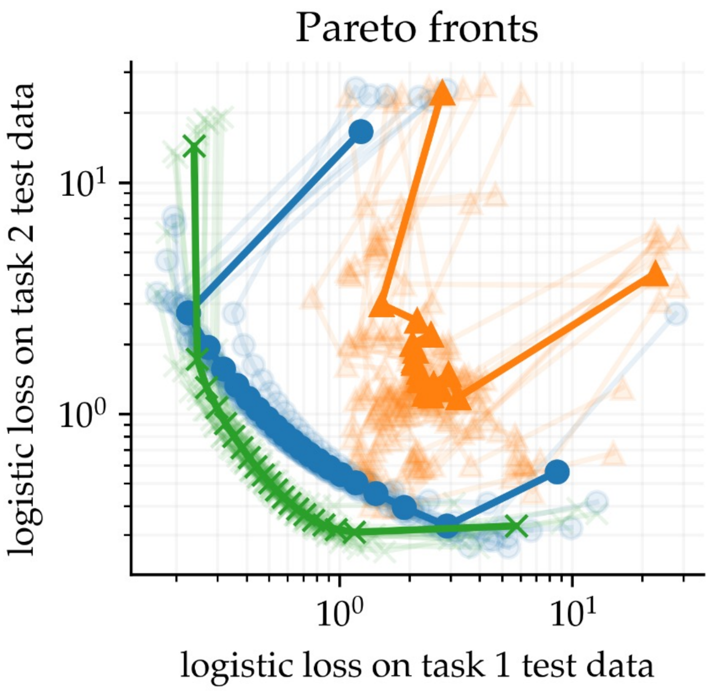
- Preference may depend on the Pareto set itself.

Empirical comparisons

- Blue: Learning on larger function class G (using pseudo-labeling)
- Orange: Learning directly on small function class H



Unlabeled data helps



Orange: Using only labeled data, learning in G

Blue: Using only labeled data, learning in H

Green: Using unlabeled data, learning in G