Learning Pareto manifolds in high dimensions: How can regularization help?

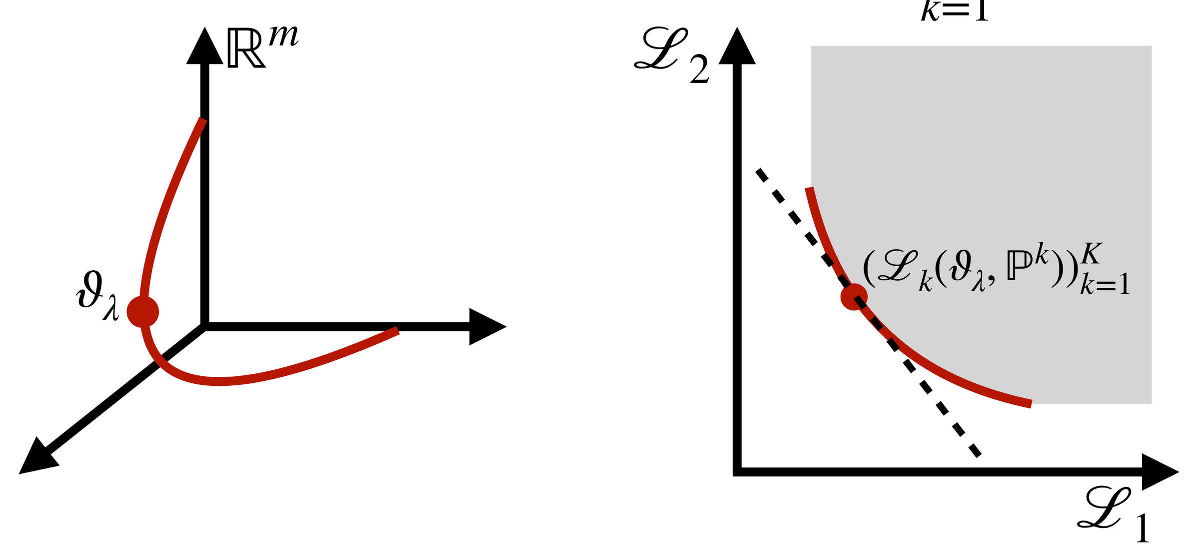
Tobias Wegel[1], Filip Kovačeić[2], Alexandru Tifrea[1], Fanny Yang[1]

[1]Department of Computer Science, ETH Zurich [2]Institute of Science and Technology Austria

Multi-objective learning

Pareto manifold of K convex objectives $\mathcal{L}_k(\cdot, \mathbb{P}^k)$:

$$(\lambda, \vartheta_\lambda) \in \Delta^{K-1} \times \mathbb{R}^m : \quad \vartheta_\lambda = \arg\min_{\vartheta \in \mathbb{R}^m} \sum_{k=1}^{K} \lambda_k \mathcal{L}_k(\vartheta, \mathbb{P}^k).$$



Goal: Estimate $\{\vartheta_\lambda : \lambda \in \Delta^{K-1}\}$ from i.i.d. data $(X_i^k, Y_i^k) \sim \mathbb{P}^k$

High dimensions: Sample sizes $= n_k \lesssim m =$ parameter dimension

$\Longrightarrow$ ***need regularization (e.g., ℓ_1-penalty)! But how?***

Failure of direct regularization

Many existing methods (e.g., [1,2]) regularize directly

$$\hat{\vartheta}_\lambda^{\text{di}} = \arg\min_{\vartheta \in \mathbb{R}^m} \sum_{k=1}^{K} \lambda_k \mathcal{L}_k(\vartheta, \hat{\mathbb{P}}^k) + \rho_\lambda(\vartheta).$$

Example: Let $\mathbf{X}_k \in \mathbb{R}^{n \times d}$, $y_k = \mathbf{X}_k \beta_k + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, $K = 2$,

$$\mathcal{L}(\vartheta, \mathbb{P}^k) = \|\mathbf{X}_k(\vartheta - \beta_k)\|_2^2 \quad \text{and} \quad \mathcal{L}(\vartheta, \hat{\mathbb{P}}^k) = \|\mathbf{X}_k \vartheta - y_k\|_2^2.$$

Then direct regularization with any penalty is lower bounded as

$$\forall \lambda_1, \lambda_2 > 0, \gamma > 1, \rho_\lambda : \quad \sup_{\substack{\gamma^{-1}\mathbf{I} \preceq \mathbf{X}_k^\top \mathbf{X}_k \preceq \gamma \mathbf{I} \\ \|\beta_k\|_0 \leq 1}} \mathbb{E}\|\hat{\vartheta}_\lambda^{\text{di}} - \vartheta_\lambda\|_2^2 \gtrsim \frac{\sigma^2 d}{n}$$

1 *Insight 1:*
Treating multi-objective learning as a single learning problem fails in high dimensions!

Two-stage estimator

Separate learning and optimization using re-parametrization:

Assume $\exists \theta_k \equiv \theta_k(\mathbb{P}^k)$: $\quad \mathcal{L}_k(\vartheta, \mathbb{P}^k) = \mathcal{L}_k(\vartheta, \theta_k)$

Stage 1: estimate $\hat{\theta}_1, \ldots, \hat{\theta}_K$

Stage 2: optimize $\hat{\vartheta}_\lambda^{\text{ts}} = \arg\min_{\vartheta \in \mathbb{R}^p} \sum_{k=1}^{K} \lambda_k \mathcal{L}_k(\vartheta, \hat{\theta}_k)$

Theoretical guarantees

Theorem: Under (strong) convexity in $\vartheta \mapsto \mathcal{L}_k(\vartheta, \theta_k)$ and locally Lipschitz parameterization $\theta_k \mapsto \nabla_\vartheta \mathcal{L}_k(\vartheta, \theta_k)$,

$$\forall \lambda \in \Delta^{K-1} : \quad \|\hat{\vartheta}_\lambda^{\text{ts}} - \vartheta_\lambda\|_2 \lesssim \sum_{k=1}^{K} \lambda_k \|\hat{\theta}_k - \theta_k\|.$$

Theorem: Denote $\delta_k = \inf_{\hat{\theta}} \sup_{\mathbb{P}} \mathbb{E}\|\hat{\theta} - \theta_k\|$. Under convexity and „Lipschitz identifiability", the minimax estimation error is at least

$$\inf_{\hat{\vartheta}_\lambda} \sup_{\mathbb{P}} \mathbb{E}\|\hat{\vartheta}_\lambda - \vartheta_\lambda\|_2 \gtrsim \max_{k \in [K]} \left(\lambda_k \delta_k - \sum_{i \neq k} \lambda_i \delta_i\right)_+.$$

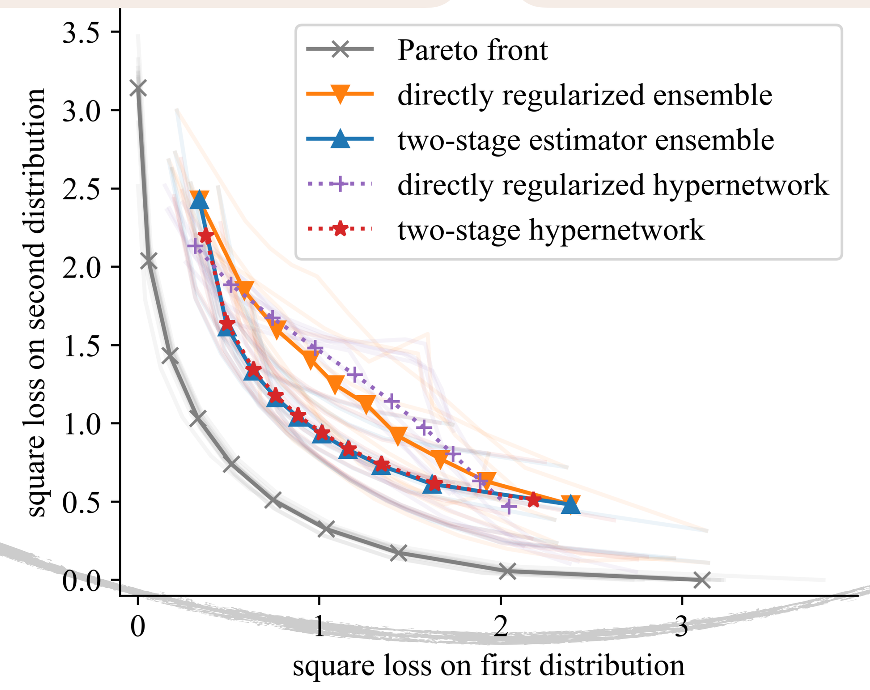
$\Longrightarrow$In many cases our procedure achieves minimax rate $\max_{k \in [K]} \lambda_k \delta_k$!

Example continued:

Stage 1: estimate $\hat{\beta}_k = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{n}\|\mathbf{X}_k \beta - y_k\|_2^2 + \alpha_k \|\beta\|_1$

Stage 2: optimize $\hat{\vartheta}_\lambda^{\text{ts}} = \arg\min_{\vartheta \in \mathbb{R}^d} \sum_{k=1}^{K} \lambda_k \|\mathbf{X}_k(\vartheta - \hat{\beta}_k)\|_2^2$

$$\forall \lambda_1, \lambda_2 > 0, \gamma > 1, \rho_\lambda : \quad \sup_{\substack{\gamma^{-1}\mathbf{I} \preceq \mathbf{X}_k^\top \mathbf{X}_k \preceq \gamma \mathbf{I} \\ \|\beta_k\|_0 \leq 1}} \mathbb{E}\|\hat{\vartheta}_\lambda^{\text{ts}} - \vartheta_\lambda\|_2^2 \lesssim \gamma^7 \frac{\sigma^2 \log d}{n}$$



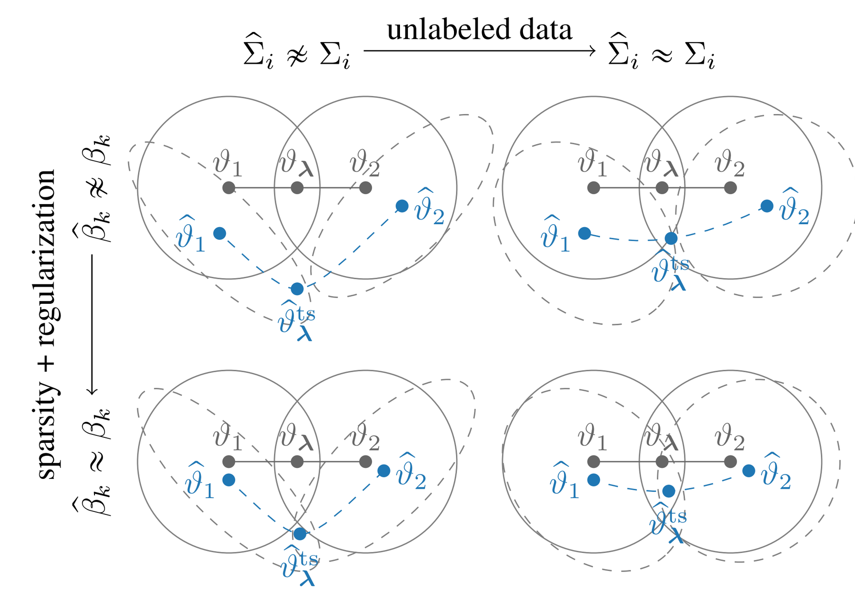
2 *Insight 2:*
By separating optimization and learning we can mitigate the curse of dimensionality!

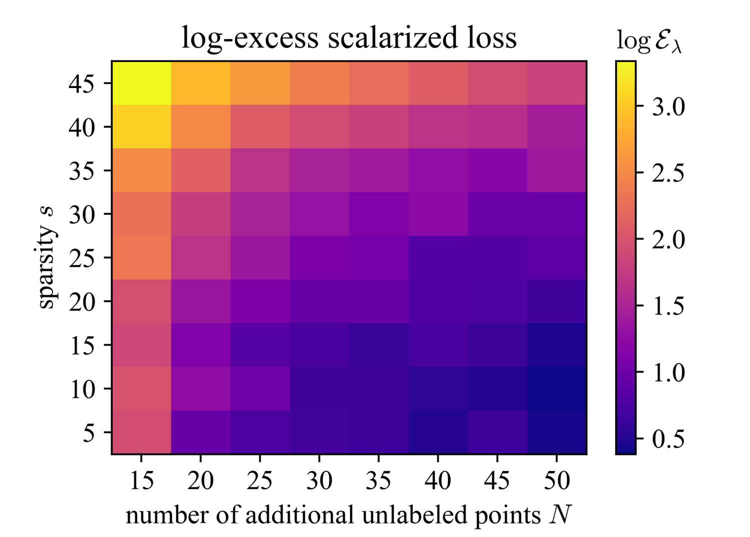
Necessity of unlabeled data

Random design? Use N unlabeled data to estimate covariance!

Example continued: If β_k are known, but covariances Σ_k unknown:

$$\sqrt{\frac{d}{n+N}} \lesssim \inf_{\hat{\vartheta}_\lambda} \sup_{1/2 \preceq \Sigma_k \preceq 3/2} \mathbb{E}\|\hat{\vartheta}_\lambda - \vartheta_\lambda\|_2 \leq \sup_{1/2 \preceq \Sigma_k \preceq 3/2} \mathbb{E}\|\hat{\vartheta}_\lambda^{\text{ts}} - \vartheta_\lambda\|_2 \lesssim \sqrt{\frac{d}{n+N}}$$



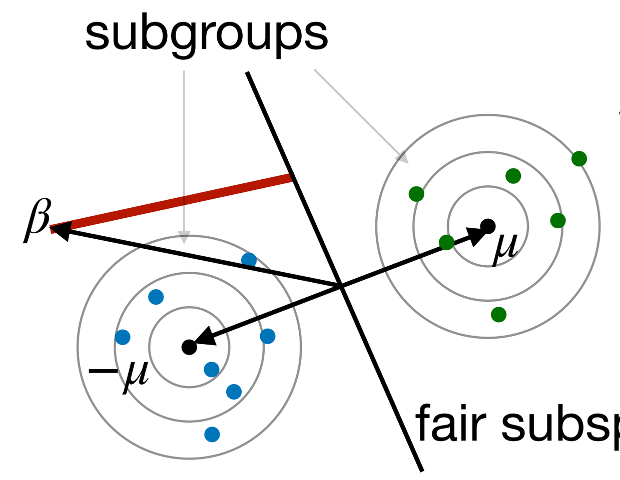


3 *Insight 3:*
Separating optimization and learning requires enough unlabeled data!

Application: fairness-risk trade-off

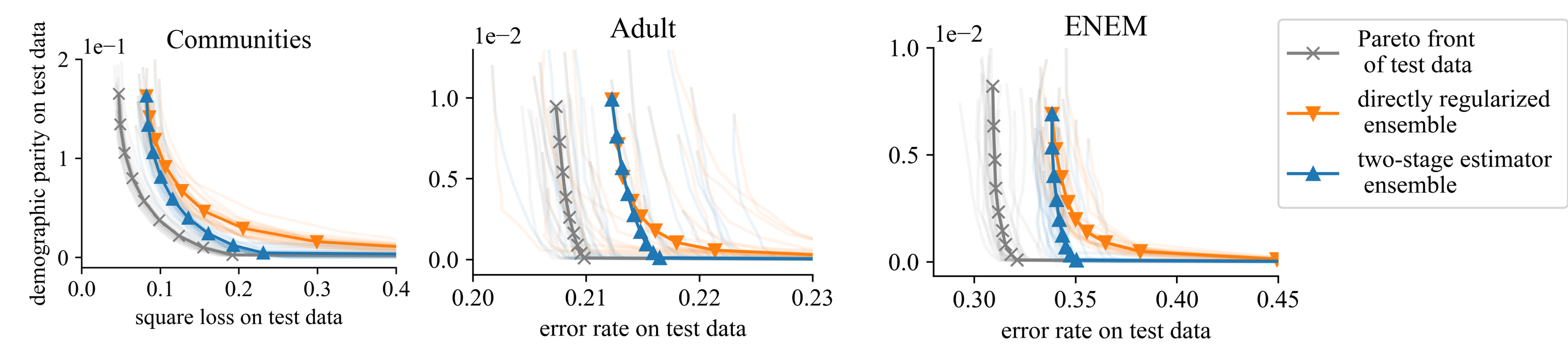
$Y = \langle X, \beta \rangle + \xi$

$[X | A] \sim \mathcal{N}(A\mu, \mathbf{I})$



$\mathcal{L}_{\text{risk}}(\vartheta, \beta, \mu) = \mathbb{E}(\langle X, \vartheta \rangle - Y)^2$

$\mathcal{L}_{\text{fair}}(\vartheta, \mu) = \langle \mu, \vartheta \rangle^2$



Related work

1. Súkeník, P., & Lampert, C. (2024). Generalization in multi-objective machine learning. *Neural Computing and Applications*, 1-15.
2. C. Cortes, M. Mohri, J. Gonzalvo, and D. Storcheus. Agnostic learning with multiple objectives. In Advances in Neural Information Processing Systems, volume 33, 2020.