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

Tobias Wegel
ETH Zurich

Filip Kovačević
IST Austria

Alexandru Țifrea
ETH Zurich

Fanny Yang
ETH Zurich

Abstract

Simultaneously addressing multiple objectives is becoming increasingly important in modern machine learning. At the same time, data is often high-dimensional and costly to label. For a single objective such as prediction risk, conventional regularization techniques are known to improve generalization when the data exhibits low-dimensional structure like sparsity. However, it is largely unexplored how to leverage this structure in the context of *multi-objective learning (MOL)* with multiple competing objectives. In this work, we discuss how the application of vanilla regularization approaches can fail, and propose a two-stage MOL framework that can successfully leverage low-dimensional structure. We demonstrate its effectiveness experimentally for multi-distribution learning and fairness-risk trade-offs.

1 INTRODUCTION

As machine learning systems are employed more and more broadly, they are expected to excel in different aspects: The models should not only be accurate, but also robust against perturbations (Szegedy et al., 2014) or distribution shifts (Rojas-Carulla et al., 2018), fairness-aware (Hardt et al., 2016), private (Dwork, 2006), interpretable (Belle and Papantonis, 2021) and, more recently, aligned with diverse human preferences (Ji et al., 2023), to name just a few. Similarly, in data-driven computational design such as drug discovery or materials science, the discovered compound or material should usually satisfy multiple desired properties (Ashby, 2000; Luukkonen et al., 2023). However, it is well understood that in many settings, doing well on all objectives simultaneously can be inherently impossible and that a trade-off between them is unavoidable (see, e.g., Menon and Williamson (2018); Zhang et al. (2019); Cummings et al. (2019); Sanyal et al. (2022); Guo et al. (2024)).

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

In the presence of such inherent trade-offs, we are usually interested in learning models that lie on the *Pareto front*—a concept studied in *multi-objective optimization (MOO)*—where improving in one objective must come at the expense of another. Under some conditions (Ehrgott, 2005), a point on the Pareto front of K objectives $\mathcal{L}_1, \dots, \mathcal{L}_K$ can be recovered by minimizing a scalarized objective, such as

$$\sum_{k=1}^{K} \lambda_k \mathcal{L}_k \quad \text{or} \quad \max_{k \in [K]} \lambda_k \mathcal{L}_k \tag{1}$$

using the appropriate weight vector $\boldsymbol{\lambda}$ from the simplex.

In the context of machine learning, however, the Pareto front cannot be computed since the objectives are population-level and hence unknown. Instead, the Pareto front has to be learned from data (Jin and Sendhoff, 2008)—a problem that falls under the general *multi-objective learning (MOL)* paradigm. A standard approach (Lin et al., 2019; Hu et al., 2024) to learning Pareto optimal points is to use an MOO method on the empirical versions of the losses $\widehat{\mathcal{L}}_1, \dots, \widehat{\mathcal{L}}_K$, for example by minimizing the empirical scalarized objective $\sum_{k=1}^{K} \lambda_k \widehat{\mathcal{L}}_k$ or $\max_{k \in [K]} \lambda_k \widehat{\mathcal{L}}_k$. The focus in such works is usually the optimization algorithm, rather than evaluating the generalization of the estimated Pareto frontier to the "true" Pareto frontier on test data. To date, there is only little work that formally characterizes how well such methods estimate the population-level Pareto set. Further, to the best of our knowledge, the proposed methods in existing theoretical works (Súkeník and Lampert, 2024; Cortes et al., 2020) only recover the true Pareto front if there is sufficiently much labeled data for training the models.

However, in modern overparameterized regimes with relatively little labeled data compared to the model complexity, it is crucial to leverage low-dimensional structure. In single-objective learning, efficient estimators have been proposed that successfully leverage the structural simplicity of the estimand to enjoy sample-efficiency, e.g., for recovering sparse ground truths (Bühlmann and van de Geer, 2011; Wainwright, 2019). For multi-objective learning, however, no such results exist to date. Hence, in this paper, we take a step towards addressing the following question:

How can we leverage a low-dimensional structure like sparsity in the presence of multiple competing objectives?

Our main contributions are outlined below.

- We introduce a new two-stage MOL framework that can successfully take advantage of the low-dimensional structure of some distributional parameters to estimate the entire Pareto set (Section 3).
- We prove upper bounds that illustrate how the estimation error of such distributional parameters propagates to the estimation error of all points in the Pareto set, and show minimax optimality of the procedure under mild conditions using lower bounds (Section 4). In doing so, we discover that unlabeled data plays a crucial role in MOL.
- We demonstrate the effectiveness of our two-stage estimation procedure in several applications (Section 4.3), validated by experiments (Section 5).

2 SETTING AND NOTATION

Throughout, we denote K-tuples and sets that contain K-tuples in bold, where K is the number of objectives. Let $\mathcal{F}$ be a hypothesis space of functions $f_\vartheta : \mathcal{X} \to \mathcal{Y}$ parameterized by $\vartheta \in \mathbb{R}^m$. Let $\boldsymbol{\mathcal{L}} : \mathbb{R}^m \times \boldsymbol{\mathcal{P}} \to \mathbb{R}^K$ be a non-negative vector-valued function that consists of K objectives

$$\boldsymbol{\mathcal{L}}(\vartheta, \mathbb{P}) = \left(\mathcal{L}_1(\vartheta, \mathbb{P}^1), \ldots, \mathcal{L}_K(\vartheta, \mathbb{P}^K)\right),$$

where the k-th objective depends on a joint distribution $\mathbb{P}^k$ defined on $\mathcal{X} \times \mathcal{Y}$. We denote by $\mathbb{P}^k_X$ the marginal of X, and $\boldsymbol{\mathcal{P}}$ denotes the set of all possible K-tuples $\mathbb{P} = (\mathbb{P}^1, \ldots, \mathbb{P}^K)$ of joint distributions. We sometimes use the short-hand $\boldsymbol{\mathcal{L}}(\vartheta) = \boldsymbol{\mathcal{L}}(\vartheta, \mathbb{P})$. For any k, we define the single-objective population minimizers as

$$\vartheta_k \in \operatorname*{arg\,min}_{\vartheta \in \mathbb{R}^m} \mathcal{L}_k(\vartheta, \mathbb{P}^k) \subset \mathbb{R}^m. \tag{2}$$

Our setting includes many well-known problems involving several objectives. For example, choosing objectives

$$\mathcal{L}_k(\vartheta, \mathbb{P}^k) = \mathop{\mathbb{E}}_{(X,Y)\sim\mathbb{P}^k}\left[\ell_k(f_\vartheta(X), Y)\right], \tag{3}$$

for some point-wise losses $\ell_k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, gives rise to the multi-distribution learning setting (Haghtalab et al., 2022; Zhang et al., 2024; Larsen et al., 2024). Concretely, in this paper, we often consider the example of multi-distribution learning for sparse linear regression, described below.

Example 1 (Sparse linear regression)**.** *Consider $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ and $\boldsymbol{\mathcal{P}}$ that consists of distributions $\mathbb{P}^k$ induced by*

$$Y = \langle X, \beta_k \rangle + \xi \quad \textit{with} \quad \xi \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

where $\beta_k \in \mathbb{R}^d$ are s-sparse ground truths $\|\beta_k\|_0 \leqslant s$ with bounded norm $\|\beta_k\|_2 \leqslant 1$ for all k, and X are B^2-sub-Gaussian covariate vectors with covariance matrices $\Sigma_k = \mathbb{E}_{X\sim\mathbb{P}^k_X}\left[XX^\top\right] \succeq b^2\mathbf{I}_d$. As objectives, consider the prediction risk measured by the expected squared-loss on each distribution that is defined as

$$\begin{aligned}\mathcal{L}_k(\vartheta, \mathbb{P}^k) &= \mathbb{E}_{(X,Y)\sim\mathbb{P}^k}\left[(\langle X, \vartheta\rangle - Y)^2\right] \\ &= \left\|\Sigma_k^{1/2}(\vartheta - \beta_k)\right\|_2^2 + \sigma^2.\end{aligned}$$

More generally, our formulation also captures objectives that cannot be written in the specific form (3).

Example 2 (Fairness and risk)**.** *Consider a family of joint distributions $\mathbb{P}$ of the random variables Y, X, A, where $A \in \{\pm 1\}$ is a Rademacher variable and $Y = \langle X, \beta\rangle + \xi$ with an s-sparse ground-truth $\|\beta\|_2 \leqslant 1$, noise $\xi \sim \mathcal{N}(0, \sigma^2)$ and $X|A \sim \mathcal{N}(A\mu, \mathbf{I}_d)$. The variable A may represent an observed protected group attribute. In this setting we consider two objectives: the expected squared loss*

$$\mathcal{L}_{\text{risk}}(\vartheta, \mathbb{P}) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[(\langle X, \vartheta\rangle - Y)^2\right]$$

and demographic parity via the 2-Wasserstein distance between the group-wise distributions of $(\langle X, \vartheta\rangle | A = a)$ and their barycenter (Gouic et al., 2020; Chzhen and Schreuder, 2022; Fukuchi and Sakuma, 2024), defined as

$$\begin{aligned}\mathcal{L}_{\text{fair}}(\vartheta, \mathbb{P}) = \min_{\nu\in\mathcal{P}_2(\mathbb{R})} \Big\{ &\frac{1}{2} W_2^2\left(\text{law}(\langle X, \vartheta\rangle \mid A = 1), \nu\right) \\ &+ \frac{1}{2} W_2^2\left(\text{law}(\langle X, \vartheta\rangle \mid A = -1), \nu\right) \Big\}.\end{aligned}$$

More details on this setting can be found in Appendix B, where we also demonstrate that under our assumptions, $\mathcal{L}_{\text{fair}}(\vartheta, \mathbb{P}) = \langle \mu, \vartheta\rangle^2$. Hence, unless $\langle \mu, \beta\rangle = 0$, there is a trade-off between fairness and risk.

2.1 Multi-objective optimization

In view of multiple objectives, one goal could be to find a parameter that simultaneously minimizes all objectives at once. However, often this is not possible because the sets of minimizers of the objectives do not intersect. In that case, *multi-objective optimization* (MOO) (Ehrgott, 2005) aims to find *Pareto-optimal solutions* as defined below.

Definition 1 (Pareto-optimality)**.** *A parameter $\vartheta \in \mathbb{R}^m$ is* Pareto-optimal*, if for all $\vartheta' \in \mathbb{R}^m$ and $k \in [K]$*

$$\mathcal{L}_k(\vartheta', \mathbb{P}^k) < \mathcal{L}_k(\vartheta, \mathbb{P}^k) \Rightarrow \exists j : \mathcal{L}_j(\vartheta', \mathbb{P}^j) > \mathcal{L}_j(\vartheta, \mathbb{P}^j).$$

The set $\mathfrak{F} = \{\boldsymbol{\mathcal{L}}(\vartheta, \mathbb{P}) \mid \vartheta \text{ is Pareto-optimal}\}$ is called the Pareto front *and the set $\{\vartheta \in \mathbb{R}^m \mid \vartheta \text{ is Pareto-optimal}\}$ is called* Pareto set *of $\boldsymbol{\mathcal{L}}$ and $\mathbb{P}$.*

For a toy two-objective problem, Figure 1 depicts the Pareto set in the parameter space $\mathbb{R}^m$ with "end-points" ϑ_k as the solid gray line on the left, and the Pareto front in the space of objective values on the right. The gray-shaded region on the right corresponds to the set of all achievable value pairs of the two objective functions. The Pareto set can often be recovered using what is known as *scalarization*.

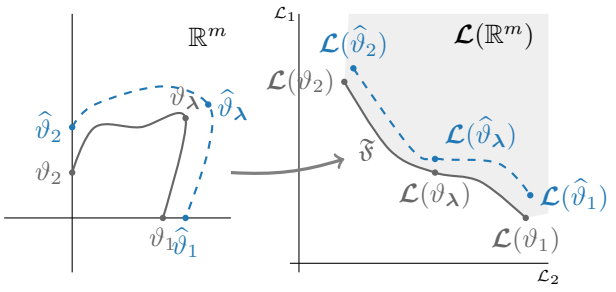


Figure 1: The parameter space $\mathbb{R}^m$ (left) parameterizes the hypothesis set $\mathcal{F}$ and contains the population Pareto set $\{\vartheta_{\boldsymbol{\lambda}} | \boldsymbol{\lambda} \in \Delta^K\}$ (gray line), and the set of the empirical estimators $\{\widehat{\vartheta}_{\boldsymbol{\lambda}} | \boldsymbol{\lambda} \in \Delta^K\}$ (dashed blue line). The right figure depicts the region of all values that can be obtained by $\boldsymbol{\mathcal{L}}(\vartheta)$ for some ϑ (gray shaded area), the population Pareto front $\mathfrak{F}$ (gray line) and estimated Pareto front (dashed blue line).

Definition 2. *A* scalarization *of $\boldsymbol{\mathcal{L}}$ is the composition $(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})$ with a function $s_{\boldsymbol{\lambda}} : \mathbb{R}^K \to \mathbb{R}$, parameterized by $\boldsymbol{\lambda}$ in the simplex $\Delta^K = \{\boldsymbol{\lambda} \in \mathbb{R}^K : \lambda_k \geqslant 0$ and $\sum_{k=1}^K \lambda_k = 1\}$.*

Equation (1) already introduced two important scalarizations, known as linear and Chebyshev scalarization. We denote by $\vartheta_{\boldsymbol{\lambda}}$ the minimizer of a scalarization, i.e.,

$$\vartheta_{\boldsymbol{\lambda}} \in \arg\min_{\vartheta \in \mathbb{R}^m} (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \mathbb{P}). \tag{5}$$

For rest of the paper, when we use $\vartheta_{\boldsymbol{\lambda}}$ to denote a minimizer of an unspecified scalarization, we implicitly assume that the scalarization parameterizes the Pareto set, that is, $\boldsymbol{\lambda} \mapsto \vartheta_{\boldsymbol{\lambda}}$ is a surjection from the simplex to the Pareto set—see Hillermeier (2001); Roy et al. (2023) for details on when this is true, and the manifold structure this map can induce.

2.2 Multi-objective learning

In practice, we may only observe finite samples from the distributions $\mathbb{P}^k$. We consider a semi-supervised setting, where we observe a set $\mathcal{D}$ consisting of n_k i.i.d. labeled samples from $\mathbb{P}^k$, denoted $\{(X_i^k, Y_i^k)\}_{i=1}^{n_k}$, as well as $N_k \in \mathbb{N}$ unlabeled i.i.d. samples $\{X_i^k\}_{i=n_k+1}^{n_k+N_k}$ from each marginal distribution $\mathbb{P}_X^k$. Then, the empirical measure $\widehat{\mathbb{P}}^k = n_k^{-1} \sum_{i=1}^{n_k} 1_{(X_i^k, Y_i^k)}$ is an estimate of the distribution $\mathbb{P}^k$ and we denote $\widehat{\mathbb{P}} = (\widehat{\mathbb{P}}^1, \ldots, \widehat{\mathbb{P}}^K)$. The aim of *multi-objective learning* (MOL) is to use the data $\mathcal{D}$ to recover the Pareto set $\{\vartheta_{\boldsymbol{\lambda}} | \boldsymbol{\lambda} \in \Delta^K\} \subset \mathbb{R}^m$, that is, to find a set of estimators $\{\widehat{\vartheta}_{\boldsymbol{\lambda}} | \boldsymbol{\lambda} \in \Delta^K\}$ with small estimation errors $\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\|_2$. In our paper, we treat the estimation of each point $\vartheta_{\boldsymbol{\lambda}}$ as a separate problem, although our statements hold for all $\boldsymbol{\lambda} \in \Delta^K$. For each $\boldsymbol{\lambda}$, we then compare our estimation procedures with the information-theoretically optimal estimation error (i.e., minimax error)

$$\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) = \inf_{\widehat{\vartheta}_{\boldsymbol{\lambda}}} \sup_{\mathbb{P} \in \boldsymbol{\mathcal{P}}} \mathbb{E}_{\mathcal{D}} \left[\left\| \widehat{\vartheta}_{\boldsymbol{\lambda}}(\mathcal{D}) - \vartheta_{\boldsymbol{\lambda}} \right\|_2 \right], \tag{6}$$

where the infimum is taken over all estimators $\widehat{\vartheta}_{\boldsymbol{\lambda}}$ that map $\mathcal{D}$ to $\mathbb{R}^m$, and the expectation is taken over draws of $\mathcal{D}$. Note that later we adopt the common abuse of notation and drop the dependence on $\mathcal{D}$, that is, we write $\widehat{\vartheta}_{\boldsymbol{\lambda}} = \widehat{\vartheta}_{\boldsymbol{\lambda}}(\mathcal{D})$.

This paper focuses on bounding the estimation error, since—under mild regularity assumptions—it bounds other metrics of interest, such as the excess scalarized objective

$$\mathcal{E}_{\boldsymbol{\lambda}}(\widehat{\vartheta}_{\boldsymbol{\lambda}}) := (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\widehat{\vartheta}_{\boldsymbol{\lambda}}, \mathbb{P}) - \min_{\vartheta \in \mathbb{R}^m} (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \mathbb{P}), \tag{7}$$

and the *hypervolume* (Zitzler and Thiele, 1999) of the estimated Pareto front, defined for $\mathcal{S} \subset [0, r]^K$, $r \geqslant 0$, as[1]

$$\mathrm{HV}_r(\mathcal{S}) := \mathrm{vol}(\{x \in [0, r]^K \mid \exists s \in \mathcal{S} : s \preceq x\}). \tag{8}$$

Here $\mathrm{vol}(\cdot)$ denotes the Lebesgue measure on $\mathbb{R}^K$. We now formalize this statement using the function

$$\varepsilon(G, \nu, \boldsymbol{\lambda}) := G \left\| \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2 + \frac{\nu}{2} \left\| \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2^2.$$

Proposition 1. *Let $G_k := \sup_{\boldsymbol{\lambda} \in \Delta^K} \|\nabla_\vartheta \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}})\|_2$, assume $\vartheta \mapsto \mathcal{L}_k(\vartheta)$ is ν_k-smooth, and define $\varepsilon_{\max} := \max_{k \in [K], \boldsymbol{\lambda} \in \Delta^K} \varepsilon(G_k, \nu_k, \boldsymbol{\lambda})$. It then holds that*

1. *for linear scalarization, $\mathcal{E}_{\boldsymbol{\lambda}}(\widehat{\vartheta}_{\boldsymbol{\lambda}}) \leqslant \varepsilon(0, s_{\boldsymbol{\lambda}}(\boldsymbol{\nu}), \boldsymbol{\lambda})$,*
2. *$\mathcal{L}_k(\widehat{\vartheta}_{\boldsymbol{\lambda}}) - \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}) \leqslant \varepsilon(G_k, \nu_k, \boldsymbol{\lambda})$,*
3. *$\mathrm{HV}_r(\widehat{\mathfrak{F}}) \geqslant (1 - 2\varepsilon_{\max}/r)^K \mathrm{HV}_r(\mathfrak{F})$,*

for $\widehat{\mathfrak{F}} = \{\boldsymbol{\mathcal{L}}(\widehat{\vartheta}_{\boldsymbol{\lambda}}) | \boldsymbol{\lambda} \in \Delta^K\}$, $\mathfrak{F} = \left\{\boldsymbol{\mathcal{L}}(\vartheta_{\boldsymbol{\lambda}}) \mid \boldsymbol{\lambda} \in \Delta^K\right\}$ and any constant $r \geqslant 2 \sup_{\boldsymbol{\lambda} \in \Delta^K} \|\boldsymbol{\mathcal{L}}(\vartheta_{\boldsymbol{\lambda}})\|_\infty$.

The proof of Proposition 1 can be found in Appendix D.1.

3 TWO ESTIMATORS

A commonly used approach to recover $\vartheta_{\boldsymbol{\lambda}}$ is to consider a *plug-in* estimator that is the minimizer of the objective (5) where $\mathbb{P}$ is replaced by $\widehat{\mathbb{P}}$ (see, e.g., (Jin, 2007)). Albeit simple, this approach suffers from the curse of dimensionality when m is large relative to the sample size (i.e., in the high-dimensional regime). We now discuss how low-dimensional structure can be leveraged for MOL.

3.1 Naive approach: direct regularization

For a single objective, a common practice in high-dimensional statistics is to use a regularizer that reflects the structural simplicity (such as sparsity) of the single objective minimizer. Since the scalarized objective in (5) can be viewed as a generic scalar loss, one may be tempted to analogously add a penalty term $\rho : \mathbb{R}^m \to \mathbb{R}$ to the empirical objective and find a minimizer of a directly regularized

[1] We write $s \preceq x$ if $s_i \leqslant x_i$ for all i.

(dr) scalarization, that it, to use[2]

$$\widehat{\vartheta}^{\mathrm{dr}}_{\boldsymbol{\lambda}} \in \operatorname*{arg\,min}_{\vartheta \in \mathbb{R}^m} (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \widehat{\mathbb{P}}) + \rho(\vartheta). \tag{9}$$

This approach can be effective if the inductive bias on the multi-objective solution (5) is the same across the Pareto set (Jin and Sendhoff, 2008; Cortes et al., 2020; Mierswa, 2007; Bieker et al., 2022; Hotegni et al., 2024). However, assuming that all points $\vartheta_{\boldsymbol{\lambda}}$ in the Pareto set have the same simple structure is less justified for a generic problem. In particular, we typically assume structural simplicity of distributional parameters, such as sparsity of β_k in Example 1. Direct regularization penalties such as (9) then help in single-objective learning because the *minimizers ϑ_k happen to coincide with distributional parameters*. But even if some distributional parameters are sparse, large parts of the Pareto set will *not* coincide with them, and hence can still be non-sparse (cf. Figure 1). Therefore, applying a regularization penalty that works for the individual objectives (such as an ℓ_1-norm penalty) does not generally improve the estimate from (9) for all $\boldsymbol{\lambda}$, except when $\lambda_k = 1$ and we are estimating ϑ_k.

Example 3 (Sparse fixed-design linear regression)**.** *Let $n \geqslant d$. For each $k \in \{1, 2\}$, we observe a fixed design matrix $\mathbf{X}_k \in \mathfrak{X}(\gamma)$, where $\mathfrak{X}(\gamma)$ is defined as*

$$\mathfrak{X}(\gamma) = \left\{ \mathbf{X} \in \mathbb{R}^{n \times d} \;\middle|\; \gamma^{-1} \mathbf{I}_d \preceq n^{-1} \mathbf{X}^\top \mathbf{X} \preceq \gamma \mathbf{I}_d \right\}$$

for some $\gamma > 1$. Further, we observe noisy responses

$$y^k = \mathbf{X}_k \beta_k + \xi^k \quad \text{with} \quad \xi^k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{I}_n),$$

where β_k is from the set $\Gamma \subset \mathbb{R}^d$ of 1-sparse vectors. With slight abuse of notation, we define the population and empirical objectives for $k \in \{1, 2\}$ as

$$\mathcal{L}_k(\vartheta, \mathbb{P}^k) = n^{-1} \left\| \mathbf{X}_k (\vartheta - \beta_k) \right\|_2^2,$$
$$\mathcal{L}_k(\vartheta, \widehat{\mathbb{P}}^k) = n^{-1} \left\| \mathbf{X}_k \vartheta - y^k \right\|_2^2.$$

Note that here $\vartheta_k = \beta_k$ are sparse and direct regularization with ℓ_1-norm (i.e., the LASSO (Tibshirani, 1996)) would mitigate the curse of dimensionality for $\boldsymbol{\lambda}$ with $\lambda_k = 1$, leading to an estimation error of order $\sigma^2 \log(d)/n$ (Bickel et al., 2009). The following proposition shows, however, that for general $\boldsymbol{\lambda}$, any estimator with direct regularization (9) *cannot* leverage the sparsity of β_k and incurs a sample complexity that is linear in the dimension d.

Proposition 2 (Insufficiency of direct regularization)**.** *Consider Example 3 and linear scalarization. For any $\boldsymbol{\lambda}$ with $\lambda_1, \lambda_2 > 0$, $\sigma^2 \leqslant 2n\gamma^2/(d+1)$ and any regularizer ρ, an estimator $\widehat{\vartheta}^{\mathrm{dr}}_{\boldsymbol{\lambda}}$ from (9) satisfies*

$$\sup_{\substack{\beta_1, \beta_2 \in \Gamma \\ \mathbf{X}_1, \mathbf{X}_2 \in \mathfrak{X}(\gamma)}} \mathbb{E}\left[\left\| \widehat{\vartheta}^{\mathrm{dr}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2^2 \right] \gtrsim \frac{\sigma^2 d}{n\gamma}.$$

[2]Alternatively, one may also add a penalty in each objective separately or view the penalty as an additional objective. For linear scalarization, any of these alternative strategies would result in a final estimator that is equivalent to (9).

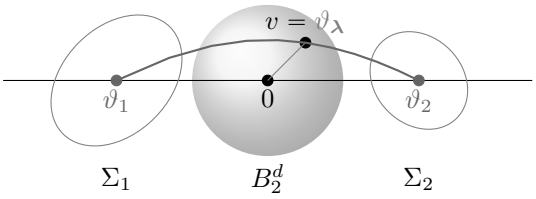


Figure 2: Illustration of the intuition for Propositions 2 and 5 in linear regression with squared loss and linear scalarization: For any $v \in B_2^d$, we can find covariance matrices Σ_1, Σ_2 with constrained condition number, and 1-sparse β_1, β_2, so that the minimizer $\vartheta_{\boldsymbol{\lambda}}$ of (5) satisfies $v = \vartheta_{\boldsymbol{\lambda}}$. This makes learning with direct regularization and without enough unlabeled data infeasible.

See Appendix D.2 for the proof. As we can see, even though the individual minimizers $\vartheta_k = \beta_k$ are 1-sparse, in a worst-case sense, direct regularization does not mitigate the curse of dimensionality for all other points in the Pareto set.

The proof of Proposition 2 uses the fact that we can choose the covariance matrices $\frac{1}{n}\mathbf{X}_k^\top \mathbf{X}_k$ adversarially within the eigenvalue constraints, so that the Pareto-optimal $\vartheta_{\boldsymbol{\lambda}}$ lies anywhere in an ℓ_2-ball of fixed radius, see Figure 2. If we did not allow for this (e.g., if the covariance matrices are scaled identities), $\vartheta_{\boldsymbol{\lambda}}$ would also be sparse, and the directly regularized estimator could achieve a fast rate. We will revisit this point later in Section 4.4 and Proposition 5.

3.2 A new two-stage estimator

The previous example suggests that in contrast to the single-objective case, learning points in the Pareto set sample-efficiently requires explicitly estimating sparse distributional parameters separately. Building on this intuition, we indeed propose such a two-stage estimator in this section. First, to formalize this, we assume that each objective depends on the distributions $\mathbb{P}^k$ via some parameter $\theta_k \in \mathbb{R}^p$.

Assumption 1. *For each $k \in [K]$, the objective $\mathcal{L}_k(\cdot, \mathbb{P}^k)$ depends on $\mathbb{P}^k$ only through $\theta_k \equiv \theta_k(\mathbb{P}^k) \in \mathbb{R}^p$, so that we can abuse notation and write $\mathcal{L}_k(\vartheta, \theta_k) = \mathcal{L}_k(\vartheta, \mathbb{P}^k)$.*

Denoting $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, we can write

$$\boldsymbol{\mathcal{L}}(\vartheta, \boldsymbol{\theta}) := (\mathcal{L}_1(\vartheta, \theta_1), \ldots, \mathcal{L}_K(\vartheta, \theta_K)),$$

and define $\boldsymbol{\Theta} \subset \mathbb{R}^{K \cdot p}$ to be the set of all possible $\boldsymbol{\theta}$ for K-tuples of distributions in $\boldsymbol{\mathcal{P}}$. Throughout, we do not distinguish between matrices and their vectorizations, unless necessary. We argue that the re-parameterization from Assumption 1 can be found in many cases. For instance, in Example 1, the parameter θ_k corresponds to the tuple (β_k, Σ_k). In Example 2, we have $\theta_{\mathrm{fair}} = \mu$ and $\theta_{\mathrm{risk}} = (\beta, \mu)$. And finally, in Example 3, we have $\theta_k = \beta_k$. Other objectives that fit this framework include the robust risk (Yin et al., 2019) and a variety of fairness losses (Berk et al., 2017). Note that an important case of Assumption 1 is when the individual minimizers ϑ_k from (2) and part of the parameters

θ_k coincide. For instance, in Example 1, ϑ_k is a component of $\theta_k = (\beta_k, \Sigma_k)$, and in Example 3, $\vartheta_k = \beta_k = \theta_k$.

Under Assumption 1, we can now introduce the two-stage estimation framework for learning Pareto-optimal solutions for any $\boldsymbol{\lambda} \in \Delta^K$ in high dimensions.

Definition 3. *We define $\widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}}$ as the final solution of the following two-stage optimization procedure.*

Stage 1: Estimation. *Use the data $\mathcal{D}$ to estimate the parameters $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \dots, \widehat{\theta}_K)^\top$ in any way.*[3]

Stage 2: Optimization. *Minimize the scalarized objective*

$$\widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}} \in \arg\min_{\vartheta \in \mathbb{R}^m} (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \widehat{\boldsymbol{\theta}}). \tag{10}$$

In its general form, this estimator first learns a *probabilistic model* of the distributions (Ng and Jordan, 2001) which it then plugs into the scalarized objective to estimate the Pareto set. Further, if the parameters θ_k coincide with the individual minimizers ϑ_k, the estimator resembles a form of *Mixture of Experts* (Dimitriadis et al., 2023; Chen and Kwok, 2024; Tang et al., 2024) with expert models $\widehat{\vartheta}_k = \widehat{\theta}_k$.

Naturally, for $\widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}}$ to be a sample-efficient estimator, the estimators $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ themselves need to be efficient. For instance, if θ_k is sparse, one could choose an ℓ_1-norm penalty (Tibshirani, 1996) with appropriate regularization strength (Bickel et al., 2009; Chatterjee and Lahiri, 2011). We now show how for Example 3, the two-stage estimator with the ℓ_1-norm penalty performs much better than any directly regularized estimator (Proposition 2).

Proposition 3. *In the setting of Example 3 and Proposition 2, consider $\widehat{\boldsymbol{\theta}}$ with $\widehat{\theta}_k = \widehat{\beta}_k$, and the corresponding two-stage estimator $\widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}}$ for linear scalarization, with*

$$\widehat{\beta}_k \in \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \left\| \mathbf{X}_k \beta - y^k \right\|_2^2 + 6\gamma\sigma \sqrt{\frac{2 \log d}{n}} \left\| \beta \right\|_1,$$

$$\widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}} \in \arg\min_{\vartheta \in \mathbb{R}^d} \frac{\lambda_1}{n} \left\| \mathbf{X}_1 (\vartheta - \widehat{\beta}_1) \right\|_2^2 + \frac{\lambda_2}{n} \left\| \mathbf{X}_2 (\vartheta - \widehat{\beta}_2) \right\|_2^2.$$

This two-stage estimator achieves estimation error

$$\sup_{\substack{\beta_1, \beta_2 \in \Gamma \\ \mathbf{X}_1, \mathbf{X}_2 \in \mathfrak{X}(\gamma)}} \left\| \widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2^2 \lesssim \frac{\gamma^7 \sigma^2 \log d}{n}$$

with probability at least $1 - 4d^{-4}$.

See Appendix D.3 for the proof. We can see that our estimator recovers the well-known rates of the LASSO (Tibshirani, 1996; Bickel et al., 2009) along the entire Pareto set.

4 THEORETICAL GUARANTEES

We now prove upper and lower bounds for more general problems. First, Theorem 1 and Proposition 4 show how the error in estimating the parameters $\boldsymbol{\theta}$ propagates to the estimation error of the two-stage estimator. We then establish a minimax lower bound in Theorem 2 that is tight in many cases. We instantiate these bounds in concrete examples to obtain explicit statistical errors for learning Pareto sets in high dimensions. Finally, we discuss the important role of unlabeled data for multi-objective learning.

4.1 Main results

Relating the error of estimating $\boldsymbol{\theta}$ with the error of estimating $\vartheta_{\boldsymbol{\lambda}}$ relies on some assumptions about the objectives. To formulate them, we introduce a new set $\widetilde{\boldsymbol{\Theta}} \supset \boldsymbol{\Theta}$ on which the objectives should satisfy these assumptions. The set $\widetilde{\boldsymbol{\Theta}}$ should be large enough so that the estimators satisfy $\widehat{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$ with high probability. We rely on two different sets of regularity assumptions: strongly convex objectives and objectives with Lipschitz parameterization.

Strongly convex objectives. We start by first stating a bound under the following two assumptions.

Assumption 2 (Strongly convex objectives)**.** *For all $\boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}}$ and $k \in [K]$, the map $\vartheta \mapsto \mathcal{L}_k(\vartheta, \theta_k)$ is differentiable and μ_k-strongly convex with $\mu_k \geqslant 0$ and $\mu_j > 0$ for at least one objective $j \in [K]$. We denote $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$.*

Note that only one of the objectives is assumed to be *strongly* convex, and $\mu_k = 0$ corresponds to regular convexity, see Appendix A for a reminder of definitions.

Assumption 3 (Locally Lipschitz-continuous gradients)**.** *For all $k \in [K]$, $\vartheta \in \mathbb{R}^m$ and all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \widetilde{\boldsymbol{\Theta}}$ it holds that*

$$\left\| \nabla_\vartheta \mathcal{L}_k(\vartheta, \theta_k) - \nabla_\vartheta \mathcal{L}_k(\vartheta, \theta'_k) \right\|_2 \leqslant \zeta_k(\vartheta) \left\| \theta_k - \theta'_k \right\|$$

with $\zeta_k(\vartheta) \geqslant 0$ depending on $\widetilde{\boldsymbol{\Theta}}$ and $\vartheta \in \mathbb{R}^m$, and where $\|\cdot\|$ is some norm. We denote $\zeta(\vartheta) = \max_{k \in [K]} \zeta_k(\vartheta)$.

Note that both Assumptions 2 and 3 are common in the (multi-objective) optimization literature (Hillermeier, 2001; Roy et al., 2023; Ehrgott, 2005; Bubeck, 2015; Boyd and Vandenberghe, 2004), and both hold for several standard settings in statistics and machine learning—including Examples 1 to 3. The next theorem provides upper bounds on the estimation error $\|\widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\|_2$ in terms of the estimation error of the parameters $\boldsymbol{\theta}$.

Theorem 1. *Let Assumptions 2 and 3 hold with $\boldsymbol{\mu}$ and ζ, respectively. Let j be the index of the strongly convex objective ($\mu_j > 0$), and $\widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}}$ be the minimizer of (10) with linear scalarization, i.e., $(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}}) = \sum_{k=1}^K \lambda_k \mathcal{L}_k$. Then, for all $\boldsymbol{\lambda} \in \Delta^K$ with $\lambda_j > 0$ and $\widehat{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$ it holds that*

$$\left\| \widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2 \leqslant \frac{\zeta(\vartheta_{\boldsymbol{\lambda}})}{s_{\boldsymbol{\lambda}}(\boldsymbol{\mu})} \sum_{k=1}^K \lambda_k \|\widehat{\theta}_k - \theta_k\|.$$

We prove Theorem 1 in Appendix D.4. Notice that if $\boldsymbol{\mathcal{P}}$ and $\zeta(\vartheta_{\boldsymbol{\lambda}})$ are such that $\zeta := \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \Delta^K} \zeta(\vartheta_{\boldsymbol{\lambda}}) < \infty$ is

[3] Note that the estimator $\widehat{\theta}_k$ of θ_k does *not* need to be the plug-in estimator of the empirical distribution $\widehat{\mathbb{P}}^k$.

a constant independent of all parameters, we obtain a bound uniformly over $\boldsymbol{\mathcal{P}}$—assumed to be true in the subsequent discussion. Note that under Assumption 2, linear scalarization is sufficient to reach the entire Pareto front (Ehrgott, 2005, Theorem 4.1), but the proof of Theorem 1 may also be extended to other scalarizations, such as (smoothed) Chebyshev scalarization (Lin et al., 2024) under other assumptions.

The upper bound in Theorem 1 has a very straight-forward interpretation: For fixed choices of $\boldsymbol{\lambda}$, the statistical rate of $\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\|_2$ inherits the rates of estimating $\boldsymbol{\theta}$. Further, using Proposition 1, we note that Theorem 1 also implies bounds on the excess scalarized objective and the hypervolume.

The proof of Theorem 1 relies on studying how the minimizer of an optimization problem changes with respect to the parameters of that problem—a problem extensively discussed in the optimization stability literature (Ito and Kunisch, 1992; Gfrerer and Klatte, 2016; Dontchev, 1995; Bonnans and Shapiro, 2000; Shvartsman, 2012). Results in that literature often rely on properties akin to Assumptions 2 and 3 for the Implicit Function Theorem to apply (see Bonnans and Shapiro (2000, §1) and Miettinen (1999, §I.3.4)). Our proof of Theorem 1 similarly relies on Assumptions 2 and 3 to show that the implicitly defined function

$$\boldsymbol{\theta} \mapsto \vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \operatorname*{arg\,min}_{\vartheta \in \mathbb{R}^m} (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \boldsymbol{\theta})$$

is Lipschitz continuous.

Lipschitz parameterization. Assumption 2 excludes examples where the objectives are (globally) Lipschitz continuous, such as the Huber loss (Huber, 1964), only convex, such as logistic loss on separable data (Ji and Telgarsky, 2019), or even non-convex. Fortunately, these cases can be addresses if the objectives are Lipschitz in their parameters using standard arguments on the excess scalarized objective. This offers an alternative to Theorem 1.

Proposition 4 (Lipschitz parameterization)**.** *Assume that the parameterization $\theta_k \mapsto \mathcal{L}_k(\cdot, \theta_k)$ is 1-Lipschitz continuous with respect to the function $\Phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, in the sense that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \widetilde{\boldsymbol{\Theta}}$ it holds*

$$\sup_{\vartheta \in \mathbb{R}^m} \left| \mathcal{L}_k(\vartheta, \theta_k) - \mathcal{L}_k(\vartheta, \theta'_k) \right| \leqslant \Phi(\theta_k, \theta'_k).$$

Then, for any scalarization of the form $s_{\boldsymbol{\lambda}}(x) = \|\boldsymbol{\lambda} \odot x\|$ with some norm $\|\cdot\|$, the excess scalarized loss of $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$—as defined in Equation (7)—is bounded by

$$\mathcal{E}_{\boldsymbol{\lambda}}(\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}) \leqslant 2 s_{\boldsymbol{\lambda}} \left((\Phi(\widehat{\theta}_k, \theta_k))_{k=1}^K \right).$$

Notably, Proposition 4 can apply to non-convex objectives, and allows the use of Chebyshev scalarization (Equation (1)). The proof, found in Appendix D.5, follows the standard uniform learning decomposition, similar to the bounds found in Súkeník and Lampert (2024). The difference here is that we may still observe the benefits from regularization and unlabeled data for the two-stage estimator.

4.2 Tightness and a minimax lower bound

We now provide a lower bound on the minimax multi-objective estimation errors from Equation (6) in terms of the minimax parameter estimation error

$$\delta_k := \inf_{\widehat{\theta}_k} \sup_{\mathbb{P} \in \boldsymbol{\mathcal{P}}} \mathbb{E}_{\mathcal{D}} \left[\left\| \widehat{\theta}_k(\mathcal{D}) - \theta_k \right\| \right],$$

where the infimum is taken over all estimators that have access to the unlabeled and labeled datasets. Our arguments rely on the following *identifiability* assumption.

Assumption 4 (Lipschitz identifiability)**.** *For all $\vartheta \in \mathbb{R}^m$, the mapping $g_k(\cdot; \vartheta) : \theta_k \mapsto \nabla_\vartheta \mathcal{L}_k(\vartheta, \theta_k)$ is injective on $\widetilde{\boldsymbol{\Theta}}$, and for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \widetilde{\boldsymbol{\Theta}}$, $\vartheta, \vartheta' \in \mathbb{R}^m$ and $u \in \operatorname{im}(g_k(\cdot; \vartheta))$, $u' \in \operatorname{im}(g_k(\cdot; \vartheta'))$, we have that*

$$\left\| g_k(\theta_k; \vartheta) - g_k(\theta'_k; \vartheta') \right\|_2 \leqslant \eta_k \left(\left\| \theta_k - \theta'_k \right\| + \left\| \vartheta - \vartheta' \right\|_2 \right),$$
$$\left\| g_k^{-1}(u; \vartheta) - g_k^{-1}(u'; \vartheta') \right\|_2 \leqslant \eta'_k \left(\left\| u - u' \right\|_2 + \left\| \vartheta - \vartheta' \right\|_2 \right).$$

Assumptions of this type are common in the inverse optimization literature, which studies the identification of optimization parameters from a minimizer; see Aswani et al. (2018); Gebken and Peitz (2021) and references therein. In particular, Assumption 4 is, e.g., satisfied by Example 3, and Examples 1 and 2 under some conditions (see Section 4.3). The intuition behind Assumption 4 is that it ensures the identifiability of optimization parameters—in our case θ_k—from the minimizer of an optimization problem—in our case $\vartheta_{\boldsymbol{\lambda}}$—in some Lipschitz manner. This allows us to lower bound the minimax estimation error.

Theorem 2. *If Assumption 4 holds and we use linear scalarization, the minimax rate is lower bounded as*

$$\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \geqslant \max_{k \in [K]} (1 + s_{\boldsymbol{\lambda}}(\boldsymbol{\eta}))^{-1} \left(\frac{\lambda_k}{\eta'_k} \delta_k - \sum_{i \neq k} \eta_i \lambda_i \delta_i \right)_+$$

where $(\cdot)_+ := \max\{0, \cdot\}$.

Theorem 2 is proved in Appendix D.6.

Minimax optimality of the two-stage estimator. We can now use the lower bound from Theorem 2 to discuss the minimax optimality of our two-stage procedure. In particular, we first note that the lower bound is tight for all $\boldsymbol{\lambda} \in \Delta^K$ that satisfy for some large enough constant $C = C(\boldsymbol{\mu}, \zeta, \boldsymbol{\eta}, \boldsymbol{\eta}') > 0$ and some $k \in [K]$ that

$$\lambda_k \delta_k \geqslant C \sum_{i \neq k} \lambda_i \delta_i. \tag{11}$$

To see this, first observe that when inequality (11) holds, the lower bound (neglecting dependence on $\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\eta}'$) reduces to $\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \gtrsim \max_{k \in [K]} \lambda_k \delta_k$. Supposing $\widehat{\theta}_k$ estimates θ_k in a minimax optimal manner so that $\sup_{\mathbb{P} \in \boldsymbol{\mathcal{P}}} \mathbb{E} \|\widehat{\theta}_k - \theta_k\| \asymp \delta_k$, under (11), the upper bound in Theorem 1 also reduces

to $\sup_{\mathbb{P}\in\boldsymbol{\mathcal{P}}} \mathbb{E}\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\|_2 \lesssim \max_{k\in[K]} \lambda_k\delta_k$. Hence, up to constants depending on $\boldsymbol{\mu}, \zeta, \boldsymbol{\eta}, \boldsymbol{\eta}'$, we obtain that

$$\sup_{\mathbb{P}\in\boldsymbol{\mathcal{P}}} \mathbb{E}\left[\left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2\right] \asymp \mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \asymp \max_{k\in[K]} \lambda_k\delta_k,$$

and the two-stage estimator is minimax optimal.

We now demonstrate that (11) holds for essentially all $\boldsymbol{\lambda} \in \Delta^K$, that is, across the majority of the Pareto front, for large enough sample size and fixed K. To that end, consider the simplified problem for $K = 2$ objectives with minimax rates δ_1, δ_2, so that (11) reduces to $\lambda_1\delta_1 \geqslant C\lambda_2\delta_2$ (or vice versa). We can make the following case distinction.

- Case 1: $\delta_2 = o(\delta_1)$ (or vice versa) as $n, d \to \infty$, that is, estimating the two parameters is unequally hard. In that case, for every fixed $\boldsymbol{\lambda} \in \Delta^2$, (11) holds when n, d are large enough. For example, in linear regression with the parameters θ_k being one sparse and one dense ground truth, the minimax rates are $\delta_1 \asymp \sqrt{d/n}$ and $\delta_2 \asymp \sqrt{\log(d)/n}$. Then (11) holds if $\lambda_1/\lambda_2 \geqslant C\sqrt{\log(d)/d}$, which constitutes a large subset of Δ^2 that expands to the entire simplex as $d \to \infty$. Therefore, for most $\boldsymbol{\lambda} \in \Delta^2$, the estimation error eventually scales with the "slower" minimax rate, and the upper bound is tight.
- Case 2: $\delta_1 \approx \delta_2$, that is, estimating the distributional parameters in both problems is approximately equally hard. In that case, (11) holds for all $\boldsymbol{\lambda} \in \Delta^2$ where either $\lambda_1 \geqslant C\lambda_2$ or $\lambda_2 \geqslant C\lambda_1$, which constitutes a large part of the simplex, unless C is very large. For example, in linear regression when both ground-truths are equally dense so that $\delta_1 = \delta_2 \asymp \sqrt{d/n}$, then for almost all $\boldsymbol{\lambda}$, we have that $\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \asymp \max\{\lambda_1, \lambda_2\}\sqrt{d/n}$.

A similar argument works for any fixed K—but (11) has to be evaluated more carefully if K grows with n, d. We conclude that, under Assumptions 1 to 4, the difficulty of MOL is dominated by the hardest individual learning problem, if *all other* individual learning tasks are easier in the sense of (11). Then the two-stage estimator is optimal.

4.3 Application to Examples 1 and 2

We now apply Theorems 1 and 2 to derive explicit bounds for the two problems from Examples 1 and 2.

Multiple sparse linear regression. Recall Example 1 and consider our two-stage estimator $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$, where in stage 1 we estimate $\theta_k = (\beta_k, \Sigma_k)$ with the sample covariance $\widehat{\Sigma}_k$ (using both the labeled and unlabeled data) and

$$\widehat{\beta}_k \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d} \frac{1}{n}\left\|\mathbf{X}_k\beta - y^k\right\|_2^2 + 136B\sigma\sqrt{\frac{\log d}{n}}\left\|\beta\right\|_1,$$

$$\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} \in \operatorname*{arg\,min}_{\vartheta\in\mathbb{R}^d} \sum_{k=1}^K \lambda_k \left\|\widehat{\Sigma}_k^{1/2}(\vartheta - \widehat{\beta}_k)\right\|_2^2. \tag{12}$$

Corollary 1. *Let Assumption D.1 (in Appendix D.7) hold. Then $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$ from (12) achieves for all $\boldsymbol{\lambda} \in \Delta^K$ and $\mathbb{P} \in \boldsymbol{\mathcal{P}}$*

$$\left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 \lesssim \frac{B^4}{b^4}\sum_{k=1}^K \lambda_k\left(\frac{\sigma}{b^2}\sqrt{\frac{s\log d}{n_k}} + \sqrt{\frac{d}{n_k+N_k}}\right)$$

with probability at least $1 - c_1K(d^{-3} + \exp(-c_2B^4d))$, where $c_1, c_2 > 0$ are some universal constants. If further $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$ is fixed and known, then

$$\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \gtrsim \max_{k\in[K]} \lambda_k \frac{b^2\sigma}{B^3}\sqrt{\frac{s\log d}{n_k}}.$$

The proof can be found in Appendix D.7. We can see that when there are enough unlabeled samples, $N_k \gg dn_k/\log d$, both bounds match up to constants.

Fairness-risk tradeoff in linear regression. Recall Example 2. As the fairness objective $\mathcal{L}_{\mathrm{fair}}$ violates strong convexity, we have to restrict ourselves to the case that $\lambda_{\mathrm{risk}} > 0$. We apply our two-stage estimator with the following natural estimators for $\theta = (\beta, \mu)$ in the first stage

$$\widehat{\beta} \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d} \frac{1}{n}\left\|\mathbf{X}\beta - y\right\|_2^2 + 136\sigma\sqrt{\frac{\log d}{n}}\left\|\beta\right\|_1,$$

$$\widehat{\mu} = \frac{1}{n+N}\sum_{i=1}^{n+N} A_iX_i,$$

and using linear scalarization, so that $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$ solves

$$\min_{\vartheta\in\mathbb{R}^d} \lambda_{\mathrm{risk}}\frac{1}{n}\left\|(\mathbf{I}_d + \widehat{\mu}\widehat{\mu}^\top)^{\frac{1}{2}}(\vartheta - \widehat{\beta})\right\|_2^2 + \lambda_{\mathrm{fair}}\langle\vartheta, \widehat{\mu}\rangle^2. \tag{13}$$

The following corollary applies Theorem 1 to this setting, and its proof can be found in Appendix D.8.

Corollary 2. *In the setting of Example 2, assume that $n \gtrsim \sigma^2 s\log d$. Then the two-stage estmator $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$ from (13) achieves for all $\boldsymbol{\lambda} \in \Delta^2$ with $\lambda_{\mathrm{risk}} > 0$*

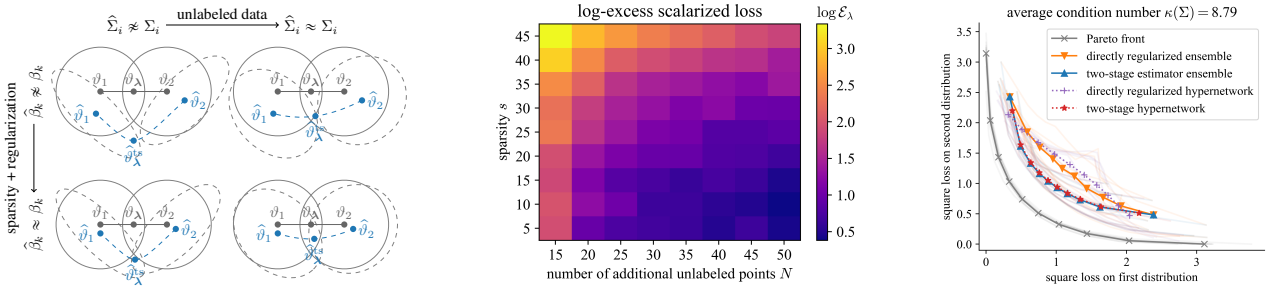
$$\left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 \lesssim \sigma\sqrt{\frac{s\log d}{n}} + \frac{1}{\lambda_{\mathrm{risk}}}\sqrt{\frac{d}{n+N}}$$

with probability at least $1 - cd^{-3}$, where $c > 0$ is some universal constant.

If μ is fixed and known, the matching minimax lower bound on the first term follows from the same arguments as in Corollary 1 without any further assumptions.

4.4 Necessity of unlabeled data

Corollaries 1 and 2 showed that provided with sufficiently much unlabeled data—so that we can estimate the covariance matrices well—and if the signal is sparse, we can learn the Pareto set efficiently. The corresponding lower bounds



(a) (b) (c)

Figure 3: The important roles of both regularization and additional unlabeled data for Example 1 illustrated on an intuitive level (a), and by evaluating the excess scalarized loss in simulations (b): Increasing sparsity together with appropriate regularization improves the estimate of the parameters β_k, while an increasing number of unlabeled datapoints N_k improves the estimate of the covariance matrices Σ_k, both improving the estimation of the Pareto front. (c): Pareto fronts for two sparse linear regression problems, using direct regularization and the two-stage approach (Section 5.1). We also plot the hypernetwork implementation.

show that this rate is tight when the covariance matrices are known. But is the unlabeled data really necessary?

We now show that this phenomenon is not an artifact of our bounds using an instance of Example 1. In particular, we prove that even when the ground truths β_k are precisely known, having enough unlabeled data to accurately estimate the covariance matrices Σ_k is crucial to efficiently estimate the Pareto set in high dimensions. To gain some intuition for this, we illustrate the effects of both regularization and the estimation of the covariance matrix in Figure 3(a). Notice how the estimators $\widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}}$ are only close to the Pareto set if both β_k and Σ_k are estimated well.

Proposition 5 (Necessity of unlabeled data)**.** *Consider the special case of the statistical model $\boldsymbol{\mathcal{P}}$ in Example 1, by restricting $\beta_1 = -\beta_2 = \beta$ for some β with $\|\beta\|_2 = 1$ that is fixed and known. Further, for $k \in \{1, 2\}$, let $\mathbb{P}^k$ be such that $\mathbb{P}^k_X = \mathcal{N}(0, \Sigma_k)$ with symmetric, unknown covariance matrices Σ_k satisfying $\frac{1}{2}\mathbf{I}_d \preceq \Sigma_k \preceq \frac{3}{2}\mathbf{I}_d$. Consider $\boldsymbol{\lambda} = (1/2, 1/2)$ and we observe $n_k = n$ labeled and $N_k = N$ unlabeled datapoints each. Then, if $d \geqslant 3$ and $\sqrt{d/(512(n+N))} \leqslant 1/(4e)$, it holds that*

$$\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \gtrsim \sqrt{\frac{d}{n+N}}.$$

Note that, of course, if the covariance matrices have sparse structure, then the amount of labeled data necessary also reduces (Wainwright, 2019, Section 6.5). The proof of Proposition 5 is provided in Appendix D.9 and follows a similar idea to Proposition 2, visualized in Figure 2.

5 EXPERIMENTS

In this section, we present some experiments on synthetic data for Example 1 and real data for Example 2 to illustrate the results from Section 4.3.

Ensembles and hypernetworks. So far, we have considered families of estimators $\{\widehat{\vartheta}_{\boldsymbol{\lambda}} | \boldsymbol{\lambda} \in \Delta^K\}$. In practice, computing such families is not possible, as it would require storing an infinite amount of estimators. Instead, in our experiments, we use two methods for approximating these sets. For one, we use what we call an *ensemble*, which is a finite set of estimators $\{\widehat{\vartheta}_{\boldsymbol{\lambda}} | \boldsymbol{\lambda} \in \Lambda\}$, where Λ is some discretization of the simplex; we use $|\Lambda| = 10$. Secondly, we also use so-called *hypernetworks* (Navon et al., 2021), which are neural networks that learn the Pareto set as a function $\widehat{h} : \Delta^K \to \mathbb{R}^d$, $\boldsymbol{\lambda} \mapsto \widehat{\vartheta}_{\boldsymbol{\lambda}}$. For example, for linear models, a prediction on x is then given by $\langle \widehat{h}(\boldsymbol{\lambda}), x \rangle$. We give implementation details of hypernetworks for the directly-regularized and two-stage estimators in Appendix C.

5.1 Multiple sparse linear regression

The first simulation is on synthetic data in the setting of Example 1 for two objectives. We present two experiments. In the first experiment, we fix $\lambda_k = 1/2$, $d = 50$, $n_k = 15$ and two arbitrarily chosen covariance matrices Σ_1, Σ_2. The covariates are then sampled from Gaussians. We vary the sparsity s of two random ground-truths (normalized in ℓ_2-norm) between 5 and 45, and the number of additional unlabeled datapoints N_i between 15 and 50. For each configuration, we repeat the experiment 10 times and show the resulting average log-excess scalarized loss $\log \mathcal{E}_{\boldsymbol{\lambda}}$ from (7) of the two-stage estimator with appropriately chosen ℓ_1-penalty in Figure 3(b). The smaller the number of non-zero entries s and the more unlabeled data are available, the better the estimator performs, as predicted by Corollary 1. In the second experiment, we compare our estimator with the directly regularized plug-in estimator for fixed dimension $d = 80$ and fixed sample sizes $n_i = 25$, $N_i = 60$, with two randomly chosen 1-sparse ground truths and covariance matrices. Figure 3(c) shows the Pareto fronts of 50 different runs and their point-wise average. As expected, the biggest benefit of our method lies in the cases where $\lambda_1 \approx \lambda_2$.

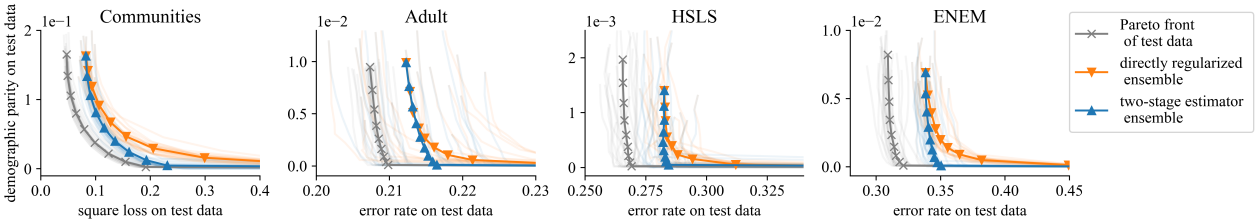


Figure 4: Pareto fronts on test data and their estimates using direct regularization (orange) and our method (blue) for the fairness experiments described in Section 5.2 and Appendix B.1, using data from Redmond (2002); Becker and Kohavi (1996); Jeong et al. (2022); Alghamdi et al. (2022). Each experiment is repeated 20 times and we plot the results (transparent), as well as their average (thick lines).

5.2 Fairness-risk trade-off in linear regression

We also apply our estimator to four fairness datasets described in Appendix B.1: The Communities and Crime dataset (Redmond, 2002), The Adult dataset (Becker and Kohavi, 1996), The HSLS dataset (Ingels et al., 2011; Jeong et al., 2022), and the ENEM dataset (Alghamdi et al., 2022). To simulate the (moderately) high-dimensional regime, we both subsample and add noisy features, as described in Appendix B.1. Our two-stage estimator and the directly regularized estimator (9) are then applied using an ℓ_1-penalty. We repeat all experiments 20 times and show the resulting estimated Pareto fronts, as well as their point-wise average in Figure 4. On all datasets the two-stage estimator outperforms the directly regularized estimator in parts of the Pareto front where the weights are bounded away from 1.

6 RELATED WORK

Multi-objective learning is a rich field of research (Jin, 2007; Jin and Sendhoff, 2008), rooted in multi-objective optimization (Deist et al., 2023; Shah and Ghahramani, 2016; Duh et al., 2012; Hu et al., 2024) and with connections to many adjacent fields of machine learning such as fairness (Yaghini et al., 2023; Ţifrea et al., 2024; Martinez et al., 2020), federated or distributed learning (Kairouz et al., 2021; Li et al., 2021b; Wang et al., 2017, 2020; Lee et al., 2017), continual learning (Parisi et al., 2019; Wang et al., 2022), reinforcement learning (Hayes et al., 2022; Van Moffaert et al., 2014), transfer learning (Li et al., 2021a) and OOD generalization (Chen et al., 2023). MOL is closely related to, but distinct from, multi-*task* learning. The latter, for example, considers settings where sharing parameters across multiple learning tasks is *helpful* for training task-specific models (Caruana, 1997; Baxter, 2000; Sener and Koltun, 2018; Li and Bilen, 2020). The use of sparsity for multi-task learning has been studied, e.g., in Lounici et al. (2009); Guo et al. (2011). While several works consider special cases of MOL, such as multi-distribution learning (Haghtalab et al., 2023; Zhang et al., 2024) or fairness (Xian et al., 2023), generalization of MOL in its general form is still rather poorly studied, with few exceptions like Súkeník and Lampert (2024); Cortes et al. (2020); Chamon and Ribeiro (2020). These works however do not yield meaningful bounds in the high-dimensional regime that we study in this paper.

Finally, we discuss the body of work on *Pareto set learning* (PSL) (Navon et al., 2021; Lin et al., 2019, 2022; Dimitriadis et al., 2023; Chen and Kwok, 2024; Tang et al., 2024) where the word "learning" does not refer to generalization as in statistical learning: Albeit often being applied in machine learning settings, the performance of PSL is usually not measured by comparing the "learned" estimators with the "true" Pareto-optimal solutions that optimize trade-offs on the test data. Instead, it aims to reduce the memory or runtime of MOO methods by harnessing, e.g., neural networks to approximate some *deterministic* Pareto set. That said, most PSL methods can be used in our two-stage framework in stage 2 (cf. Section 5.1 and Appendix C).

7 CONCLUSIONS & FUTURE WORK

In this work, we propose an estimator for the Pareto front of MOL problems that performs well in the high-dimensional regime by leveraging the sparsity of distributional parameters and unlabeled data. We investigate the estimator theoretically, with a key strength of the analysis being that it is agnostic to the estimators of the distributional parameters. Further, we prove the optimality of the proposed estimator under certain conditions. While the focus of this work is primarily on sparsity, the estimator can also, in principle, exploit other forms of low-dimensional structure. Through synthetic and real experiments, we demonstrate the good performance of our estimator in applications.

Our work opens up many exciting future research directions. For one, we leave it as future work to derive a more general framework for obtaining lower bounds in the MOL setting beyond identifiability. We also note that other results from the stability literature could relax the convexity in Assumption 2. Moreover, there are other choices of objectives where our theory can be applied, e.g., in the robustness-accuracy trade-off (Raghunathan et al., 2020). Finally, the seemingly important role of unlabeled data in multi-objective learning deserves to be investigated further.

Acknowledgments

We thank Junhyung Park for valuable feedback on the manuscript. AT was supported by a PhD fellowship from the Swiss Data Science Center. TW was supported by the SNF Grant 204439. This work was done in part while TW and FY were visiting the Simons Institute for the Theory of Computing.

References

W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. W. Michalak, S. Asoodeh, and F. P. Calmon. Beyond adult and compas: Fairness in multi-class prediction. *arXiv preprint arXiv:2206.07801*, 2022.

M. Ashby. Multi-objective optimization in material design and selection. *Acta materialia*, 48(1):359–369, 2000.

A. Aswani, Z.-J. M. Shen, and A. Siddiq. Inverse Optimization with Noisy Data. *Operations Research*, 66(3):870–892, 2018.

J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(1):149–198, 2000.

B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996.

V. Belle and I. Papantonis. Principles and practice of explainable machine learning. *Frontiers in big Data*, 4:688969, 2021.

R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4): 1705 – 1732, 2009.

K. Bieker, B. Gebken, and S. Peitz. On the Treatment of Optimization Problems With L1 Penalty Terms via Multiobjective Continuation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7797–7808, 2022.

J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer New York, 2000.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997.

L. Chamon and A. Ribeiro. Probably approximately correct constrained learning. *Advances in Neural Information Processing Systems*, 33:16722–16735, 2020.

A. Chatterjee and S. N. Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

W. Chen and J. T. Kwok. Efficient Pareto manifold learning with low-rank structure. *arXiv preprint arXiv:2407.20734*, 2024.

Y. Chen, K. Zhou, Y. Bian, B. Xie, B. Wu, Y. Zhang, M. KAILI, H. Yang, P. Zhao, B. Han, and J. Cheng. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. In *The Eleventh International Conference on Learning Representations*, 2023.

E. Chzhen and N. Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, 2022.

C. Cortes, M. Mohri, J. Gonzalvo, and D. Storcheus. Agnostic Learning with Multiple Objectives. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

A. Ţifrea, P. Lahoti, B. Packer, Y. Halpern, A. Beirami, and F. Prost. FRAPP'E: A Post-Processing Framework for Post-Processing Everything. In *International Conference on Machine Learning*, 2024.

R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern. On the Compatibility of Privacy and Fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 2019.

T. M. Deist, M. Grewal, F. J. W. M. Dankers, T. Alderliesten, and P. A. N. Bosman. Multi-objective Learning Using HV Maximization. In *Evolutionary Multi-Criterion Optimization*, 2023.

N. Dimitriadis, P. Frossard, and F. Fleuret. Pareto manifold learning: Tackling multiple tasks via ensembles of single-task models. In *International Conference on Machine Learning*, pages 8015–8052. PMLR, 2023.

A. L. Dontchev. Characterizations of Lipschitz Stability in Optimization. In R. Lucchetti and J. Revalski, editors, *Recent Developments in Well-Posed Variational Problems*, pages 95–115, Dordrecht, 1995. Springer Netherlands.

K. Duh, K. Sudoh, X. Wu, H. Tsukada, and M. Nagata. Learning to Translate with Multiple Objectives. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2012.

C. Dwork. Differential Privacy. In *Automata, Languages and Programming*, 2006.

M. Ehrgott. *Multicriteria optimization*, volume 491. Springer Science & Business Media, 2005.

R. M. Fano and D. Hawkins. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.

G. B. Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

K. Fukuchi and J. Sakuma. Demographic parity constrained minimax optimal regression under linear model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2024.

B. Gebken and S. Peitz. Inverse multiobjective optimization: Inferring decision criteria from data. *Journal of Global Optimization*, 80(1):3–29, 2021.

H. Gfrerer and D. Klatte. Lipschitz and Hölder stability of optimization problems and generalized equations. *Mathematical Programming*, 158(1):35–75, 2016.

T. L. Gouic, J.-M. Loubes, and P. Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

S. Guo, O. Zoeter, and C. Archambeau. Sparse Bayesian Multi-Task Learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Y. Guo, G. Cui, L. Yuan, N. Ding, J. Wang, H. Chen, B. Sun, R. Xie, J. Zhou, Y. Lin, et al. Controllable preference optimization: Toward controllable multi-objective alignment. *arXiv preprint arXiv:2402.19085*, 2024.

N. Haghtalab, M. Jordan, and E. Zhao. On-Demand Sampling: Learning Optimally from Multiple Distributions. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

N. Haghtalab, M. Jordan, and E. Zhao. A Unifying Perspective on Multi-Calibration: Game Dynamics for Multi-Objective Learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

M. Hardt, E. Price, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Nowé, G. Ramos, M. Restelli, P. Vamplew, and D. M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1): 26, 2022.

C. Hillermeier. Generalized homotopy approach to multiobjective optimization. *Journal of Optimization Theory and Applications*, 110(3):557–583, 2001.

S. S. Hotegni, M. Berkemeier, and S. Peitz. Multi-objective optimization for sparse deep multi-task learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024.

Y. Hu, R. Xian, Q. Wu, Q. Fan, L. Yin, and H. Zhao. Revisiting scalarization in multi-task learning: A theoretical perspective. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2024.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

S. J. Ingels, D. J. Pratt, D. R. Herget, L. J. Burns, J. A. Dever, R. Ottem, J. E. Rogers, Y. Jin, and S. Leinwand. High School Longitudinal Study of 2009 (HSLS: 09): Base-Year Data File Documentation. NCES 2011-328. *National Center for Education Statistics*, 2011.

K. Ito and K. Kunisch. Sensitivity analysis of solutions to optimization problems in Hilbert spaces with applications to optimal control and estimation. *Journal of Differential Equations*, 99 (1):1–40, 1992.

H. Jeong, H. Wang, and F. P. Calmon. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9558–9566, 2022.

J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Z. Ji and M. Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

Y. Jin. *Multi-objective Machine Learning*, volume 16. Springer Science & Business Media, 2007.

Y. Jin and B. Sendhoff. Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 38:397–415, 2008.

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

K. G. Larsen, O. Montasser, and N. Zhivotovskiy. Derandomizing multi-distribution learning. *Advances in Neural Information Processing Systems*, 37:94246–94264, 2024.

J. D. Lee, Q. Liu, Y. Sun, and J. E. Taylor. Communication-efficient Sparse Regression. *Journal of Machine Learning Research*, 18 (5):1–30, 2017.

S. Li, T. T. Cai, and H. Li. Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2021a.

T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and Robust Federated Learning Through Personalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 2021b.

W.-H. Li and H. Bilen. Knowledge Distillation for Multi-task Learning. In *Computer Vision – ECCV 2020 Workshops*, 2020.

X. Lin, H.-L. Zhen, Z. Li, Q.-F. Zhang, and S. Kwong. Pareto Multi-Task Learning. *Advances in Neural Information Processing Systems*, 32, 2019.

X. Lin, Z. Yang, X. Zhang, and Q. Zhang. Pareto set learning for expensive multi-objective optimization. *Advances in neural information processing systems*, 35:19231–19247, 2022.

X. Lin, X. Zhang, Z. Yang, F. Liu, Z. Wang, and Q. Zhang. Smooth tchebycheff scalarization for multi-objective optimization. *arXiv preprint arXiv:2402.19078*, 2024.

K. Lounici, M. Pontil, A. Tsybakov, and S. Van De Geer. Taking advantage of sparsity in multi-task learning. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

S. Luukkonen, H. W. van den Maagdenberg, M. T. Emmerich, and G. J. van Westen. Artificial intelligence in multi-objective drug design. *Current Opinion in Structural Biology*, 79:102537, 2023.

N. Martinez, M. Bertran, and G. Sapiro. Minimax Pareto Fairness: A Multi Objective Perspective. In *Proceedings of Machine Learning Research*, volume 119, 2020.

A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, 2018.

I. Mierswa. Controlling overfitting with multi-objective support vector machines. In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, 2007.

K. Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 1999.

A. Navon, A. Shamsian, G. Chechik, and E. Fetaya. Learning the Pareto Front with Hypernetworks. In *International Conference on Learning Representations*, 2021.

M. Neykov. On the Minimax Rate of the Gaussian Sequence Model Under Bounded Convex Constraints. *IEEE Transactions on Information Theory*, 69(2):1244–1260, 2023.

A. Ng and M. Jordan. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang. Understanding and Mitigating the Tradeoff between Robustness and Accuracy. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 2020.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

M. Redmond. Communities and Crime. UCI Machine Learning Repository, 2002.

I. Rivin. Another Simple Proof of a Theorem of Chandler Davis. *arXiv preprint math/0208223*, 2002.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

A. Roy, G. So, and Y.-A. Ma. Optimization on Pareto sets: On a theory of multi-objective optimization. *arXiv preprint arXiv:2308.02145*, 2023.

A. Sanyal, Y. Hu, and F. Yang. How unfair is private learning? In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180, 2022.

O. Sener and V. Koltun. Multi-task learning as multi-objective optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

A. Shah and Z. Ghahramani. Pareto frontier learning with expensive correlated objectives. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48, 2016.

I. Shvartsman. On stability of minimizers in convex programming. *Nonlinear Analysis: Theory, Methods & Applications*, 75(3): 1563–1571, 2012. Variational Analysis and Its Applications.

P. Súkeník and C. Lampert. Generalization in multi-objective machine learning. *Neural Computing and Applications*, pages 1–15, 2024.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations*, 2014.

A. Tang, L. Shen, Y. Luo, S. Liu, H. Hu, and B. Du. Towards efficient pareto set approximation via mixture of experts based model fusion. *arXiv preprint arXiv:2406.09770*, 2024.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

K. Van Moffaert, T. Brys, A. Chandra, L. Esterle, P. R. Lewis, and A. Nowé. A novel adaptive weight selection algorithm for multi-objective multi-agent reinforcement learning. In *International Joint Conference on Neural Networks*, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

J. Wang, M. Kolar, N. Srebro, and T. Zhang. Efficient Distributed Learning with Sparsity. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2017.

J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in neural information processing systems*, volume 33, 2020.

Z. Wang, Z. Zhan, Y. Gong, G. Yuan, W. Niu, T. Jian, B. Ren, S. Ioannidis, Y. Wang, and J. Dy. SparCL: Sparse Continual Learning on the Edge. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

R. Xian, L. Yin, and H. Zhao. Fair and Optimal Classification via Post-Processing. In *Proceedings of the International Conference on Machine Learning*, 2023.

M. Yaghini, P. Liu, F. Boenisch, and N. Papernot. Learning to Walk Impartially on the Pareto Frontier of Fairness, Privacy, and Utility. In *NeurIPS 2023 Workshop on Regulatable ML*, 2023.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

D. Yin, R. Kannan, and P. Bartlett. Rademacher Complexity for Adversarially Robust Generalization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 2019.

B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan. Theoretically Principled Trade-off between Robustness and Accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 2019.

R. Zhang and D. Golovin. Random hypervolume scalarizations for provable multi-objective black box optimization. In *International conference on machine learning*, pages 11096–11105. PMLR, 2020.

Z. Zhang, W. Zhan, Y. Chen, S. S. Du, and J. D. Lee. Optimal Multi-Distribution Learning. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247, 2024.

E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE transactions on Evolutionary Computation*, 3(4):257–271, 1999.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes (Section 2)
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes (Section 4)
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes (Assumptions 1,2,4)
   (b) Complete proofs of all theoretical results. Yes (Appendix D)
   (c) Clear explanations of any assumptions. Yes (Section 4)

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes (Section 5)
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes (Section 5)
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). No (The experiments are very small-scale, run on a standard MacBook Pro in under 5 minutes)

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes (Section 5)
   (b) The license information of the assets, if applicable. Not Applicable (We are not releasing new or existing assets)
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable (We are not releasing new or existing assets)
   (d) Information about consent from data providers/curators. Not Applicable (All datasets are licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.)
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

A PRELIMINARIES FROM CONVEX OPTIMIZATION

In this section, we briefly recall some basic concepts from convex optimization. A general introduction to convex optimization can be found, for example, in Boyd and Vandenberghe (2004); Bubeck (2015).

Let $f : \mathbb{R}^d \to \mathbb{R}$ denote a differentiable function with gradient $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$. We recall the definitions of convexity, strict convexity, strong convexity and smoothness: For some $\mu \geqslant 0$, the function f is called *μ-strongly convex*, if for all $x, y \in \mathbb{R}^d$ it holds that $f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geqslant \frac{\mu}{2} \|x - y\|_2^2$. If f is 0-strongly convex, it is simply convex. For some $0 \leqslant \nu < \infty$, the function f is called *ν-smooth*, if its gradient is ν-Lipschitz, that is, for all $x, y \in \mathbb{R}^d$ it holds that $\|\nabla f(x) - \nabla f(y)\|_2 \leqslant \nu \|x - y\|_2$.

We use the following well-known facts multiple times throughout the proofs of our results in Appendix D. Let $f_1, \dots, f_K : \mathbb{R}^d \to \mathbb{R}$ be differentiable, $\boldsymbol{\lambda} \in \Delta^K$ and denote $f_{\boldsymbol{\lambda}} = \sum_{k=1}^K \lambda_k f_k$.

(A.1) If f is μ-strongly convex, then $\forall x, y \in \mathbb{R}^d,\ \|\nabla f(x) - \nabla f(y)\|_2 \geqslant \mu \|x - y\|_2$.

(A.2) If f is ν-smooth, then $\forall x, y \in \mathbb{R}^d,\ |f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \leqslant \frac{\nu}{2} \|x - y\|_2^2$.

(A.3) If $f_1, \dots, f_K$ are μ_k-strongly convex functions, then the function $f_{\boldsymbol{\lambda}}$ is $\sum_{k=1}^K \lambda_k \mu_k$-strongly convex.

(A.4) If $f_1, \dots, f_K$ are ν_k-smooth functions, then the function $f_{\boldsymbol{\lambda}}$ is $\sum_{k=1}^K \lambda_k \nu_k$-smooth.

(A.5) If for $k \in [K]$ we have $f_k(x) = (x - y_k)^\top M_k (x - y_k)$ with $M_k \in \mathbb{R}^{d \times d}$ and $y_k \in \mathbb{R}^d$, then it holds that

$$\operatorname*{arg\,min}_{x \in \mathbb{R}^d} f_{\boldsymbol{\lambda}}(x) = \left\{ \left(\sum_{k=1}^K \lambda_k M_k \right)^\dagger \left(\sum_{k=1}^K \lambda_k M_k y_k \right) + z \;\middle|\; z \in \ker \left(\sum_{k=1}^K \lambda_k M_k \right) \right\},$$

where † denotes the Moore–Penrose inverse.

B DETAILS FOR THE FAIRNESS-RISK TRADEOFF IN LINEAR REGRESSION

In this section, we describe in more detail the fairness metric from Example 2 that used in Section 4.3 and Section 5, and elaborate on its interpretation.

Recall the setting described in Example 2: The random variables Y, X, A are distributed according to $Y = \langle X, \beta \rangle + \xi$ with an s-sparse ground-truth $\beta \in \mathbb{R}^d$, where $\xi \sim \mathcal{N}(0, \sigma^2)$, and for $a \in \{-1, 1\}$ we have $(X|A = a) \sim \mathcal{N}(a\mu, \mathbf{I}_d)$, where A is a Rademacher random variable that is uniformly distributed on $\{-1, 1\}$. A represents an observed binary protected group attribute (such as gender or ethnicity), of two groups that have different covariate means, which for simplicity we model as $\mathbb{E}[X|A = \pm 1] = \pm\mu$.

In this context, a commonly used fairness metric is called *demographic parity*, which asserts that predictions of the machine learning model f_ϑ are independent of the group attribute A, that is, the model f_ϑ satisfies demographic parity, if

$$(f_\vartheta(X)|A = 1) \overset{\text{law}}{=} (f_\vartheta(X)|A = -1)$$

where $\overset{\text{law}}{=}$ denotes equality in distribution. However, the assertion of demographic parity is quite strong, as it imposes an exact equality. Instead, often it is preferred to measure how much the predictor f_ϑ violates this constraint.

To measure this, we consider a notion of unfairness introduced in a recent line of work (Gouic et al., 2020; Chzhen and Schreuder, 2022; Fukuchi and Sakuma, 2024). The demographic parity score that is based on the 2-Wasserstein distance. Most of this appendix section is based on Chzhen and Schreuder (2022) and references therein. Denote $\mathcal{P}_2(\mathbb{R})$ the space of probability distributions ν on $\mathbb{R}$ with bounded second moment, that is, if $X \sim \nu$, then $\mathbb{E}\left[X^2\right] < \infty$. The 2-Wasserstein distance on $\mathcal{P}_2(\mathbb{R})$, denoted W_2^2, is defined as

$$W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^2} |x - y|^2 \, d\pi(x, y)$$

where $\Pi(\mu, \nu)$ denotes the set of measures on $\mathbb{R}^2$ with marginals μ and ν. See (Chzhen and Schreuder, 2022, Appendix A.1) for more details. The fairness notion that measures violation of demographic parity is then defined as the 2-Wasserstein

distance between the group-wise distributions of $\langle X, \vartheta \rangle | A = a$, denoted $\text{law}\left(\langle X, \vartheta \rangle | A = a\right)$ and their barycenter;

$$\mathcal{L}_{\text{fair}}(\vartheta, \mathbb{P}) = \min_{\nu \in \mathcal{P}_2(\mathbb{R})} \left\{ \frac{1}{2} W_2^2 \left(\text{law}(\langle X, \vartheta \rangle \mid A = 1), \nu\right) + \frac{1}{2} W_2^2 \left(\text{law}(\langle X, \vartheta \rangle \mid A = -1), \nu\right) \right\}. \tag{14}$$

We now show that under our assumptions, this fairness metric can be rewritten in the simple form $\mathcal{L}_{\text{fair}}(\vartheta, \mathbb{P}) = \langle \mu, \vartheta \rangle^2$.

Lemma B.1. *Under the distribution described above, we have that*

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

Proof. Recall that for $a \in \{-1, 1\}$ we have $(X|A = a) \sim \mathcal{N}(a\mu, \mathbf{I}_d)$, and hence $(\langle X, \vartheta \rangle | A = a) \sim \mathcal{N}(\langle a\mu, \vartheta \rangle, \|\vartheta\|_2^2)$. Therefore, by Chzhen and Schreuder (2022, Lemma A.2), the optimization problem in Equation (14) is solved by $\nu = \mathcal{N}(0, \|\vartheta\|_2^2)$. Further, by Chzhen and Schreuder (2022, Lemma A.1), we can plug this into $\mathcal{L}_{\text{fair}}$ and get

$$\begin{aligned}
\mathcal{L}_{\text{fair}}(\vartheta, \mathbb{P}) = &= \min_{\nu \in \mathcal{P}_2(\mathbb{R})} \left\{ \frac{1}{2} W_2^2 \left(\text{law}(\langle X, \vartheta \rangle \mid A = 1), \nu\right) + \frac{1}{2} W_2^2 \left(\text{law}(\langle X, \vartheta \rangle \mid A = -1), \nu\right) \right\} \\
&= \frac{1}{2} W_2^2 \left(\mathcal{N}(\langle \mu, \vartheta \rangle, \|\vartheta\|_2^2), \mathcal{N}(0, \|\vartheta\|_2^2) \right) + \frac{1}{2} W_2^2 \left(\mathcal{N}(\langle -\mu, \vartheta \rangle, \|\vartheta\|_2^2), \mathcal{N}(0, \|\vartheta\|_2^2) \right) \\
&= \frac{1}{2} \langle \mu, \vartheta \rangle^2 + \frac{1}{2} \langle -\mu, \vartheta \rangle^2 \\
&= \langle \mu, \vartheta \rangle^2 ,
\end{aligned}$$

which concludes the proof. □

Based on this reformulation, it is easy to see that unless $\langle \mu, \vartheta \rangle = 0$, there is a trade-off between fairness and risk, since the risk is minimized at ϑ and the demographic parity score is minimized by any vector ϑ so that $\langle \mu, \vartheta \rangle = 0$. Moreover, this reformulation significantly speeds up the computation, which is why we use it in our experiments in Section 5. Hence, we are implicitly modeling the data to come from a Gaussian model as described above.

B.1 Description of fairness datasets

We briefly describe the four datasets and their usage in our experiments in Section 5.2.

- For the Communities and Crime dataset (Redmond, 2002), the task is to predict the number of violent crimes in a community. The protected attribute is a quantization of the share of white residents of the community. To simulate the (moderately) high-dimensional regime for the Communities and Crime dataset (data dimension $d = 145$), we subsample uniformly $n = 150$ labeled and $N = 350$ unlabeled datapoints, and use the remaining samples as test samples to estimate risk and fairness scores from Section 4.3.
- For the Adult dataset (Becker and Kohavi, 1996), the task is to predict the income of individuals, and the protected attribute is their gender. Since the Adult dataset only has dimension $d = 13$, additionally to subsampling, we add 1000 noisy features (sampled from a Gaussian) to artificially increase the data dimension to $d = 1013$. We then uniformly sample $n = 1000$ labeled and $N = 2000$ unlabeled examples, with the remaining samples serving as the test set.
- For the high-school longitudinal study (HSLS) dataset (Ingels et al., 2011; Jeong et al., 2022), the target is to predict math test performance of 9th-grade high-school students, and the protected attribute is the students' ethnicity. As the data dimension is $d = 59$, we subsample $n = 1000$ labeled and $N = 4000$ unlabeled datapoints.
- In the ENEM dataset (Alghamdi et al., 2022), the aim is to predict Brazilian college entrance exam scores based on the students' demographic information and socio-economic questionnaire answers, and the protected attribute is also their ethnicity. Here, the dimension is $d = 139$, and we subsample $n = 2000$ labeled and $N = 8000$ unlabeled datapoints.

Note that the risk in Example 2 and Sections 4.3 and 5.2 is for regression. The datasets Adult, HSLS and ENEM are usually classification tasks. Hence, for comparability with other works, we report the error rate rather than the mean squared error in Figure 4, while using the squared loss risk during training—similar to previous work such as Berk et al. (2017). The benefit of the two-stage estimator empirically transfers between these two metrics.

C IMPLEMENTATION OF HYPERNETWORKS

The hypernetworks h used in Section 5 all have the following architecture. The first layer is linear, where the weight matrix has dimension $128 \times K$, the second layer is a component-wise ReLU layer, and the third layer is again linear of dimension $d \times 128$. To select the regularization strengths, we perform validation on held-out sets. To train the hypernetworks, we use an adapted version of Algorithm 1 in Navon et al. (2021). We specify the two versions below (Algorithms 1 and 2), where $\mathrm{Dir}(\cdot)$ denotes the Dirichlet distribution. We use the PyTorch implementation of Adam (Kingma, 2014) to optimize.

Algorithm 1 Training directly regularized hypernetwork

1: Input: number of iterations T, candidate regularization strength α.
2: **for** $t = 1, \ldots, T$ **do**
3: Sample $\boldsymbol{\lambda} \sim \mathrm{Dir}(\frac{1}{K}, \ldots, \frac{1}{K})$
4: Adam step on $(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(h(\boldsymbol{\lambda}), \widehat{\mathbb{P}}) + \alpha \left\| h(\boldsymbol{\lambda}) \right\|_1$ with respect to the weights of h, using learning rate 10^{-3}.
5: **end for**
6: **return** hypernetwork h

Algorithm 2 Training two-stage hypernetwork

1: Input: number of iterations T, estimates $\widehat{\boldsymbol{\theta}}$.
2: **for** $t = 1, \ldots, T$ **do**
3: Sample $\boldsymbol{\lambda} \sim \mathrm{Dir}(\frac{1}{K}, \ldots, \frac{1}{K})$
4: Adam step on $(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(h(\boldsymbol{\lambda}), \widehat{\boldsymbol{\theta}})$ with respect to the weights of h, using learning rate 10^{-3}.
5: **end for**
6: **return** hypernetwork h

D DEFERRED PROOFS

In this section we provide the deferred proofs.

D.1 Proof of Proposition 1

Proposition 1. *Let* $G_k := \sup_{\boldsymbol{\lambda} \in \Delta^K} \left\| \nabla_\vartheta \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}) \right\|_2$, *assume* $\vartheta \mapsto \mathcal{L}_k(\vartheta)$ *is* ν_k*-smooth, and define* $\varepsilon_{\max} := \max_{k \in [K], \boldsymbol{\lambda} \in \Delta^K} \varepsilon(G_k, \nu_k, \boldsymbol{\lambda})$. *It then holds that*

1. *for linear scalarization,* $\mathcal{E}_{\boldsymbol{\lambda}}(\widehat{\vartheta}_{\boldsymbol{\lambda}}) \leqslant \varepsilon(0, s_{\boldsymbol{\lambda}}(\boldsymbol{\nu}), \boldsymbol{\lambda})$,
2. $\mathcal{L}_k(\widehat{\vartheta}_{\boldsymbol{\lambda}}) - \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}) \leqslant \varepsilon(G_k, \nu_k, \boldsymbol{\lambda})$,
3. $\mathrm{HV}_r(\widehat{\mathfrak{F}}) \geqslant (1 - 2\varepsilon_{\max}/r)^K \, \mathrm{HV}_r(\mathfrak{F})$,

for $\widehat{\mathfrak{F}} = \{\boldsymbol{\mathcal{L}}(\widehat{\vartheta}_{\boldsymbol{\lambda}}) | \boldsymbol{\lambda} \in \Delta^K\}$, $\mathfrak{F} = \left\{\boldsymbol{\mathcal{L}}(\vartheta_{\boldsymbol{\lambda}}) \;\middle|\; \boldsymbol{\lambda} \in \Delta^K\right\}$ *and any constant* $r \geqslant 2 \sup_{\boldsymbol{\lambda} \in \Delta^K} \left\| \boldsymbol{\mathcal{L}}(\vartheta_{\boldsymbol{\lambda}}) \right\|_\infty$.

Proof of Proposition 1. The proof of the first two claims are straight-forward consequences of smoothness and the definition of G_k. For the last bound, we use a representation via an expected random hypervolume scalarization from Zhang and Golovin (2020).

We begin with the first bound. By (A.3), when $s_{\boldsymbol{\lambda}}$ is linear, that is, $s_{\boldsymbol{\lambda}}(x) = \sum_{i=1}^K \lambda_i x_i$, ν_k-smoothness of $\mathcal{L}_k$ implies $s_{\boldsymbol{\lambda}}(\boldsymbol{\nu})$-smoothness of $(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})$ and so by Item (A.2) it holds

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{\lambda}}(\widehat{\vartheta}_{\boldsymbol{\lambda}}) &= (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\widehat{\vartheta}_{\boldsymbol{\lambda}}) - (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta_{\boldsymbol{\lambda}}) \\
&\leqslant \left\langle \nabla_\vartheta (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta_{\boldsymbol{\lambda}}), \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\rangle + \frac{s_{\boldsymbol{\lambda}}(\boldsymbol{\nu})}{2} \left\| \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2^2 \\
&= \frac{s_{\boldsymbol{\lambda}}(\boldsymbol{\nu})}{2} \left\| \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2^2 \\
&= \varepsilon(0, s_{\boldsymbol{\lambda}}(\boldsymbol{\nu}), \boldsymbol{\lambda}).
\end{aligned}$$

where we used the stationarity condition $\nabla_\vartheta (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta_{\boldsymbol{\lambda}}) = 0$.

Smoothness, the definition of G_k, and Cauchy-Schwarz also imply that

$$\begin{aligned}\mathcal{L}_k(\widehat{\vartheta}_{\boldsymbol{\lambda}}) - \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}) &\leqslant \left\langle \nabla_\vartheta \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}), \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\rangle + \frac{\nu_k}{2} \left\| \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2^2 \\ &\leqslant G_k \left\| \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2 + \frac{\nu_k}{2} \left\| \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2^2 \\ &=: \varepsilon(G_k, \nu_k, \boldsymbol{\lambda}).\end{aligned}$$

Finally, we use a representation of hypervolume as an expected random scalarizations to prove the bound on the hypervolume. To that end, we use a version of the arguments in Zhang and Golovin (2020), which we prove after finishing the main proof.

Lemma D.1. *Denote the positive $(K-1)$-dimensional sphere as $\mathbb{S}_+^{K-1} = \left\{v \in \mathbb{R}^K \mid \|v\|_2 = 1, \forall i \in [K]\ v_i \geqslant 0\right\}$. Moreover, define U to be the uniform probability measure on $\mathbb{S}_+^{K-1}$ with Borel σ-algebra (also known as Haar measure), that is, for all Borel sets $A \subset \mathbb{S}_+^{K-1}$ we define $U(A) = c_K^{-1} \cdot \mu_K\left(\{tv \mid t \in [0,1], v \in A\}\right)$ where μ_K is the Lebesgue measure on $\mathbb{R}^K$ and $c_K = \pi^{K/2}/(2^K \Gamma(K/2+1))$. Here Γ denotes the gamma function. Let $r > 0$ and $\mathcal{S} \subset [0,r]^K$. Then*

$$\mathrm{HV}_r(\mathcal{S}) = c_K \mathbb{E}_{u\sim U}\left[\max_{s\in\mathcal{S}} \min_{k\in[K]} \left(\frac{r - s_k}{u_k}\right)^K\right].$$

Plugging our sets $\mathfrak{F}$ and $\widehat{\mathfrak{F}}$ in as $\mathcal{S}$ from Lemma D.1, we have for the constant c_K from Lemma D.1 and $\delta = \min_{k\in[K]} \min_{\boldsymbol{\lambda}\in\Delta^K} (r - \mathcal{L}_k(\widehat{\vartheta}_{\boldsymbol{\lambda}}))/(r - \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}))$ that

$$\begin{aligned}\mathrm{HV}_r(\widehat{\mathfrak{F}}) &= c_K \mathbb{E}_{u\sim U}\left[\max_{\boldsymbol{\lambda}\in\Delta^K} \min_{k\in[K]} \left(\frac{r - \mathcal{L}_k(\widehat{\vartheta}_{\boldsymbol{\lambda}})}{u_k}\right)^K\right] \\ &\geqslant c_K \mathbb{E}_{u\sim U}\left[\max_{\boldsymbol{\lambda}\in\Delta^K} \min_{k\in[K]} \left(\frac{r - \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}})}{u_k}\right)^K\right] \delta^K \\ &= \mathrm{HV}_r(\mathfrak{F})\delta^K.\end{aligned}$$

By definition of r, which implies that $r - \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}) \geqslant r/2$, and the definition of $\varepsilon(G, \nu, \boldsymbol{\lambda})$ we have that

$$\begin{aligned}\delta &= \min_{k\in[K]} \min_{\boldsymbol{\lambda}\in\Delta^K} \frac{r - \mathcal{L}_k(\widehat{\vartheta}_{\boldsymbol{\lambda}})}{r - \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}})} \\ &\geqslant \min_{k\in[K]} \min_{\boldsymbol{\lambda}\in\Delta^K} \frac{r - \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}) - \varepsilon(G_k, \nu_k, \boldsymbol{\lambda})}{r - \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}})} \\ &\geqslant 1 - \max_{k\in[K]} \max_{\boldsymbol{\lambda}\in\Delta^K} \frac{2\varepsilon(G_k, \nu_k, \boldsymbol{\lambda})}{r} \\ &= 1 - \frac{2\varepsilon_{\max}}{r}.\end{aligned}$$

Plugging this in yields the last bound and finishes the proof. □

Proof of Lemma D.1. The proof is analogous to the proof of Zhang and Golovin (2020, Lemma 5) with minor adjustments. Considering the hyper-rectangle R_s with corners $\mathbf{r} = (r, \ldots, r)^\top \in \mathbb{R}^K$ and $s \in \mathcal{S}$, that is, $R_s = \times_{k=1}^K [s_k, r]$. Take any $v \in \mathbb{S}_+^{k-1}$ and consider the ray $\{\mathbf{r} - tv \mid t \geqslant 0\}$. Let $p \in \mathbb{R}^K$ be the point where the ray exits the rectangle. It is easy to see that $p = \mathbf{r} - \min_{k\in[K]} \left(\frac{r_k - s_k}{v_k}\right) v$ because $p = r - tv$ and t must be maximal so that $p_i = r - tv_i \geqslant s_i$. Now, if we extend this argument to the entire dominated set,

$$\mathcal{D}_{\mathcal{S}} = \left\{x \in [0,r]^K \mid \exists s \in \mathcal{S} : x \succeq s\right\} = \bigcup_{s\in\mathcal{S}} R_s,$$

which is the union over such rectangles, we see that the point where the ray exits is given by $p = \mathbf{r} - t_v v$, where

$$t_v = \max_{s\in\mathcal{S}} \min_{k\in[K]} \left(\frac{r_k - s_k}{v_k}\right) \quad \text{and hence} \quad 1_{\mathcal{D}_{\mathcal{S}}}(\mathbf{r} - tv) = \begin{cases} 1 & \text{if } t \in [0, t_v], v \in \mathbb{S}_+^{K-1}, \\ 0 & \text{else.} \end{cases}$$

Denote μ_K the Lebesgue measure on $\mathbb{R}^K$. By Folland (1999, Theorem 2.49), the Borel measure σ on $\mathbb{S}^{K-1}$, defined for any Borel measurable $A \subset \mathbb{S}^{K-1}$ as $\sigma(A) = K \cdot \mu_K(\{tv \mid t \in (0,1], v \in A\})$—an unnormalized uniform measure on $\mathbb{S}^{K-1}$—satisfies

$$\begin{aligned}
\mu_K(\mathcal{D}_\mathcal{S}) &= \int_{\mathbb{R}^K} 1_{\mathcal{D}_\mathcal{S}}(x) d\mu_K(x) \\
&= \int_{\mathbb{R}^K} 1_{\mathcal{D}_\mathcal{S}}(\mathbf{r} - x) d\mu_K(x) \\
&= \int_{(0,\infty)} \int_{\mathbb{S}^{K-1}} t^{K-1} 1_{\mathcal{D}_\mathcal{S}}(\mathbf{r} - tv) d\sigma(v) d\mu_1(t) \\
&= \int_{\mathbb{S}_+^{K-1}} \int_0^{t_v} t^{K-1} d\mu_1(t) d\sigma(v) \\
&= \frac{1}{K} \int_{\mathbb{S}_+^{K-1}} \max_{s \in \mathcal{S}} \min_{k \in [K]} \left(\frac{r_k - s_k}{v_k} \right)^K d\sigma(v).
\end{aligned}$$

Now, since $\sigma(\mathbb{S}^{K-1}) = 2\pi^{K/2}/\Gamma(K/2)$ (Folland, 1999, Proposition 2.54), we have that $\sigma(\mathbb{S}_+^{K-1}) = 2\pi^{K/2}/(\Gamma(K/2)2^K)$. Hence, $U = \sigma \cdot \Gamma(K/2)2^{K-1}/\pi^{K/2}$ is a probability measure on $\mathbb{S}_+^{K-1}$, and we can write

$$\mu_K(\mathcal{D}_\mathcal{S}) = \frac{\pi^{K/2}}{K\Gamma(K/2)2^{K-1}} \mathbb{E}_{S \sim U} \left(\max_{s \in \mathcal{S}} \min_{k \in [K]} \left(\frac{r_k - s_k}{v_k} \right)^K \right)$$

Noting that by $\Gamma(x+1) = x\Gamma(x)$ we see that $\frac{\pi^{K/2}}{K\Gamma(K/2)2^{K-1}} = \frac{\pi^{K/2}}{\Gamma(K/2+1)2^K} = c_K$, and since $\mathrm{HV}_r(\mathcal{S}) = \mu_K(\mathcal{D}_\mathcal{S})$, this finishes the proof. □

D.2 Proof of Proposition 2

Proposition 2 (Insufficiency of direct regularization)**.** *Consider Example 3 and linear scalarization. For any $\boldsymbol{\lambda}$ with $\lambda_1, \lambda_2 > 0$, $\sigma^2 \leqslant 2n\gamma^2/(d+1)$ and any regularizer ρ, an estimator $\widehat{\vartheta}^{\mathrm{dr}}_{\boldsymbol{\lambda}}$ from (9) satisfies*

$$\sup_{\substack{\beta_1, \beta_2 \in \Gamma \\ \mathbf{X}_1, \mathbf{X}_2 \in \mathfrak{X}(\gamma)}} \mathbb{E}\left[\left\| \widehat{\vartheta}^{\mathrm{dr}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2^2 \right] \gtrsim \frac{\sigma^2 d}{n\gamma}.$$

Proof. We prove this lower bound by showing that any estimator in the form of Equation (9) is equivalent to a penalized least-squares estimator in a Gaussian sequence model over the ℓ_2-ball in d dimensions. Using this reduction, we can use standard lower bounds on this sequence model, for example from Neykov (2023). Importantly, this is *not* a minimax lower bound on the original problem and only applies to the directly regularized estimator $\widehat{\vartheta}^{\mathrm{dr}}_{\boldsymbol{\lambda}}$ because we first prove the reduction.

We begin the proof by stating the following auxiliary fact that is also visualized in Figure 2, which we prove after concluding the main proof. Denote $B_2^d \subset \mathbb{R}^d$ the set of vectors $v \in \mathbb{R}^d$ with $\|v\|_2 \leqslant 1$.

Claim D.1. *For all $v \in B_2^d$, $\lambda_1, \lambda_2 > 0$ with $\lambda_1 + \lambda_2 = 1$, $\gamma > 1$ and $d, n \in \mathbb{N}$ with $n \geqslant d$, there exist matrices $\mathbf{X}_1, \mathbf{X}_2 \in \mathfrak{X}(\gamma) \subset \mathbb{R}^{n \times d}$ and 1-sparse vectors $\beta_1, \beta_2 \in \Gamma \subset \mathbb{R}^d$ so that for $\Sigma_k := \frac{1}{n}\mathbf{X}_k^\top \mathbf{X}_k$, and $\vartheta_{\boldsymbol{\lambda}} = \arg\min_\vartheta \sum_{i=k}^K \lambda_k \|\mathbf{X}_k(\vartheta - \beta_k)\|_2^2$ (from Equation (5)) it holds*

1. $\lambda_1 \Sigma_1 + \lambda_2 \Sigma_2 = \mathbf{I}_d$,
2. $\lambda_1 \Sigma_1 \beta_1 + \lambda_2 \Sigma_2 \beta_2 = v = \vartheta_{\boldsymbol{\lambda}}$.

We use this claim in the main part of the proof. Fix some $\lambda_1, \lambda_2 > 0$ with $\lambda_1 + \lambda_2 = 1$. For an arbitrary choice of $v \in B_2^d$, take the corresponding design matrices $\mathbf{X}_k \in \mathfrak{X}(\gamma) \subset \mathbb{R}^{n \times d}$, where $n \geqslant d$, and 1-sparse ground-truths $\beta_1, \beta_2 \in \Gamma$ from Claim D.1. Denote $\Sigma_k = \frac{1}{n}\mathbf{X}_k^\top \mathbf{X}_k$ and recall that the eigenvalues of Σ_k are bounded to lie in $[\gamma^{-1}, \gamma]$.

Recalling that we use linear scalarization, and using the auxiliary fact that $\lambda_1\Sigma_1 + \lambda_2\Sigma_2 = \mathbf{I}_d$ and $\lambda_1\Sigma_1\beta_1 + \lambda_2\Sigma_2\beta_2 = v = \vartheta_{\boldsymbol{\lambda}}$, the scalarized empirical objective of Equation (9) reduces to

$$\begin{aligned}(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \widehat{\mathbb{P}}) &= \lambda_1 \frac{1}{n} \left\|\mathbf{X}_1\vartheta - y^1\right\|_2^2 + \lambda_2 \frac{1}{n} \left\|\mathbf{X}_2\vartheta - y^2\right\|_2^2 \\ &= \vartheta^\top\vartheta - 2\vartheta^\top \left(\frac{\lambda_1}{n}\mathbf{X}_1^\top y^1 + \frac{\lambda_2}{n}\mathbf{X}_2^\top y^2\right) + \frac{\lambda_1}{n}\left\|y^1\right\|_2^2 + \frac{\lambda_2}{n}\left\|y^2\right\|_2^2 \\ &= \vartheta^\top\vartheta - 2\vartheta^\top \left(\vartheta_{\boldsymbol{\lambda}} + \frac{\lambda_1}{n}\mathbf{X}_1^\top \xi^1 + \frac{\lambda_2}{n}\mathbf{X}_2^\top \xi^2\right) + C(\xi^1, \xi^2)\end{aligned}$$

where $C(\xi^1, \xi^2)$ is a (random) constant independent of ϑ. Since the following random vector is Gaussian

$$\xi := \frac{\lambda_1}{n}\mathbf{X}_1^\top \xi^1 + \frac{\lambda_2}{n}\mathbf{X}_2^\top \xi^2 \sim \mathcal{N}(0, \mathbf{M}) \quad \text{with} \quad \mathbf{M} = \frac{\sigma^2}{n}\left(\lambda_1^2\Sigma_1 + \lambda_2^2\Sigma_2\right),$$

we can define $y = \vartheta_{\boldsymbol{\lambda}} + \xi$ and it follows that we can write

$$(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \widehat{\mathbb{P}}) = \vartheta^\top\vartheta - 2\vartheta^\top y + C(\xi^1, \xi^2) = \|\vartheta - y\|_2^2 + C'(\xi^1, \xi^2).$$

As the auxiliary fact shows, v could lie anywhere in B_2^d, and hence, from the point of view of the empirical scalarized objective, the problem has fully reduced to a Gaussian sequence model. In particular, the empirical objective from Equation (9) is that of a penalized least-squares estimator in a Gaussian sequence model in d dimensions with ground-truth $v = \vartheta_{\boldsymbol{\lambda}}$, using the noisy observations

$$y = v + \xi \quad \text{with} \quad v \in B_2^d, \quad \xi \sim \mathcal{N}(0, \mathbf{M}). \tag{15}$$

If $\lambda_1 = \lambda_2 = 1/2$, then ξ is isotropic (since $\mathbf{M} = \sigma^2\mathbf{I}_d/(2n)$) and we could directly apply known lower bounds for the minimax rate in this setting (Neykov, 2023), but in the general form we use the following bound on the eigenvalues of $\mathbf{M}$

$$\widetilde{\sigma}^2 := \frac{\sigma^2}{2n\gamma} \preceq \mathbf{M}$$

This can easily be seen by using $\lambda_1^2 + \lambda_2^2 \geqslant 1/2$ and hence

$$\lambda_{\min}(\mathbf{M}) = \frac{\sigma^2}{n}\lambda_{\min}\left(\lambda_1^2\Sigma_1 + \lambda_2^2\Sigma_2\right) \geqslant \frac{\sigma^2}{n}\frac{\lambda_1^2 + \lambda_2^2}{\gamma} \geqslant \frac{\sigma^2}{2n\gamma}$$

We show that one may bound the minimax rate in the model (15) from below in terms of the minimax rate of the Gaussian sequence model

$$\widetilde{y} = v + \widetilde{\xi} \quad \text{with} \quad v \in B_2^d, \quad \widetilde{\xi} \sim \mathcal{N}(0, \widetilde{\sigma}^2\mathbf{I}_d) \tag{16}$$

where $\widetilde{\sigma}^2 = \sigma^2/(2n\gamma)$.

To see this, first, note that we may write $\xi = Z + W$, where $Z \sim \mathcal{N}(0, \lambda_{\min}(\mathbf{M})\mathbf{I}_d)$ and $W \sim \mathcal{N}(0, \mathbf{M} - \lambda_{\min}(\mathbf{M})\mathbf{I}_d)$ are two independent Gaussian vectors. Then, for any estimator $\widehat{v}(v + Z + W)$ in the Gaussian sequence model (15), we can define a corresponding estimator $v' : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$v'(v + Z) = \mathbb{E}_W\left[\widehat{v}(v + Z + W)\right]$$

that is an estimator in the the Gaussian sequence model with only noise Z. By a standard bias-variance decomposition, the following lower bound holds for all $v \in B_2^d, z \in \mathbb{R}^d$:

$$\mathbb{E}\left[\left\|\widehat{v}(v + Z + W) - v\right\|_2^2 \mid Z = z\right] \geqslant \left\|v'(v + z) - v\right\|_2^2.$$

The claim then directly follows by taking expectation with respect to Z, taking the supremum over v and infimum over estimators, and using $\lambda_{\min}(\mathbf{M}) \geqslant \widetilde{\sigma}^2$. Therefore, the error of any estimator $\widehat{\vartheta}_{\boldsymbol{\lambda}}^{\mathrm{dr}}$ of the form (9) using any penalty ρ is lower bounded by the minimax rate in the Gaussian sequence model of (16).

If $\widetilde{\sigma}^2 \leqslant 1/(d+1)$, the minimax rate in this setting is lower bounded by $\widetilde{\sigma}^2 d$ (up to constant factors), see for example (Neykov, 2023, Corollary 3.3). Hence, from plugging in the definition of $\widetilde{\sigma}^2 = \sigma^2/(2n\gamma)$, we see that under our assumption that $\sigma^2 \leqslant 2n\gamma/(d+1)$, the minimax rate is lower bounded by $\sigma^2 d/(n\gamma)$ (up to constant factors). Putting everything together, we have for any estimator $\widehat{\vartheta}^{\mathrm{dr}}_{\boldsymbol{\lambda}}$ in the form of (9)

$$\begin{aligned}
\sup_{\substack{\beta_1,\beta_2\in\Gamma \\ \mathbf{X}_1,\mathbf{X}_2\in\mathfrak{X}(\gamma)}} \mathbb{E}\left[\left\|\widehat{\vartheta}^{\mathrm{dr}}_{\boldsymbol{\lambda}}(y) - \vartheta_{\boldsymbol{\lambda}}\right\|_2^2\right] &\geqslant \sup_{v\in B_2^d} \mathbb{E}\left[\left\|\widehat{\vartheta}^{\mathrm{dr}}_{\boldsymbol{\lambda}}(y) - v\right\|_2^2\right] && \text{(Claim D.1)} \\
&\geqslant \inf_{\widehat{v}} \sup_{v\in B_2^d} \mathbb{E}\left[\|\widehat{v}(\widetilde{y}) - v\|_2^2\right] && \text{(lower bound from (16))} \\
&\gtrsim \widetilde{\sigma}^2 d = \frac{\sigma^2 d}{n\gamma}. && \text{(Neykov, 2023, Corollary 3.3)}
\end{aligned}$$

This concludes the lower bound. □

Proof of Claim D.1. To see this, first note that the optimization problem is minimized by

$$\vartheta_{\boldsymbol{\lambda}} = (\lambda_1\Sigma_1 + \lambda_2\Sigma_2)^{-1}(\lambda_1\Sigma_1\beta_1 + \lambda_2\Sigma_2\beta_2) \tag{17}$$

because the problem is strongly convex (A.3) and the first-order stationarity condition

$$\nabla_\vartheta \left(\frac{\lambda_1}{n}\|\mathbf{X}_1(\vartheta-\beta_1)\|_2^2 + \frac{\lambda_2}{n}\|\mathbf{X}_2(\vartheta-\beta_2)\|_2^2\right) = 2\lambda_1\Sigma_1(\vartheta-\beta_1) + 2\lambda_2\Sigma_2(\vartheta-\beta_2) = 0$$

is satisfied by $\vartheta_{\boldsymbol{\lambda}}$, cf. (A.5). Moreover, because for every symmetric positive definite Σ with eigenvalues in $[\gamma^{-1},\gamma]$, there exists a $\mathbf{X}\in\mathfrak{X}(\gamma)$ so that $\Sigma = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, we may simply construct Σ_1,Σ_2 directly. This is because $n \geqslant d$ and we may choose

$$\mathbf{X} = \sqrt{n}\begin{pmatrix}\Sigma^{1/2} \\ \mathbf{0}_{(d-n)\times d}\end{pmatrix} \implies \frac{1}{n}\mathbf{X}^\top\mathbf{X} = \Sigma.$$

The proof of the auxiliary fact then follows by explicit construction: Let $\mathbf{e}_i$ denotes the i-th standard basis vector. Fix any $v\in\mathbb{R}^d$ with $\|v\|_2 \leqslant 1$ and, without loss of generality, assume that $v_1 \neq 0$ (where $v = (v_1,\dots,v_d)^\top$). Otherwise change index 1 with any other index (the case $v=0$ is trivially solved by $\beta_1 = \mathbf{e}_1, \beta_2 = -(\lambda_1/\lambda_2)\mathbf{e}_1$ and $\Sigma_1 = \Sigma_2 = \mathbf{I}_d$). We choose the vectors

$$\beta_1 = \frac{1}{v_1\lambda_1}\frac{\gamma}{\gamma-1}\mathbf{e}_1 \quad \text{and} \quad \beta_2 = -\frac{1}{v_1\lambda_2}\frac{\gamma}{\gamma-1}\mathbf{e}_1,$$

and we further choose the matrices

$$\Sigma_1 = \lambda_2\frac{\gamma-1}{\gamma}vv^\top + \mathbf{I}_d \quad \text{and} \quad \Sigma_2 = -\lambda_1\frac{\gamma-1}{\gamma}vv^\top + \mathbf{I}_d.$$

The covariance matrices and vectors satisfy all requirements, as β_1 and β_2 are clearly 1-sparse, and the matrices Σ_1,Σ_2 are symmetric, and their eigenvalues are

$$\begin{aligned}
&\lambda_{\min}(\Sigma_1) = 1 > \gamma^{-1}, && \lambda_{\max}(\Sigma_1) = 1 + \lambda_2\|v\|_2^2\frac{\gamma-1}{\gamma} \leqslant \frac{2\gamma-1}{\gamma} \leqslant \gamma, \\
&\lambda_{\min}(\Sigma_2) = 1 - \lambda_1\|v\|_2^2\frac{\gamma-1}{\gamma} \geqslant \frac{\gamma-\gamma+1}{\gamma} = \gamma^{-1}, && \lambda_{\max}(\Sigma_2) = 1 \leqslant \gamma,
\end{aligned}$$

also implying they are positive definite. The claims then follow directly for these choices of Σ_k,β_k. Specifically, we have by Equation (17) that

$$\vartheta_{\boldsymbol{\lambda}} = (\lambda_1\Sigma_1 + \lambda_2\Sigma_2)^{-1}(\lambda_1\Sigma_1\beta_1 + \lambda_2\Sigma_2\beta_2) = \mathbf{I}_d^{-1}v = v$$

and the proof is complete. □

D.3 Proof of Proposition 3

Proposition 3. *In the setting of Example 3 and Proposition 2, consider $\widehat{\boldsymbol{\theta}}$ with $\widehat{\theta}_k = \widehat{\beta}_k$, and the corresponding two-stage estimator $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$ for linear scalarization, with*

$$\widehat{\beta}_k \in \arg\min_{\beta\in\mathbb{R}^d} \frac{1}{n}\left\|\mathbf{X}_k\beta - y^k\right\|_2^2 + 6\gamma\sigma\sqrt{\frac{2\log d}{n}}\left\|\beta\right\|_1,$$
$$\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} \in \arg\min_{\vartheta\in\mathbb{R}^d} \frac{\lambda_1}{n}\left\|\mathbf{X}_1(\vartheta-\widehat{\beta}_1)\right\|_2^2 + \frac{\lambda_2}{n}\left\|\mathbf{X}_2(\vartheta-\widehat{\beta}_2)\right\|_2^2.$$

This two-stage estimator achieves estimation error

$$\sup_{\substack{\beta_1,\beta_2\in\Gamma\\ \mathbf{X}_1,\mathbf{X}_2\in\mathfrak{X}(\gamma)}} \left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2^2 \lesssim \frac{\gamma^7\sigma^2\log d}{n}$$

with probability at least $1-4d^{-4}$.

Proof of Proposition 3. The bound follows by combining known bounds on the estimation error of the LASSO, e.g., from Bickel et al. (2009); Wainwright (2019), with a closed-form solution of the multi-objective optimization step.

Specifically, note that similar to our previous derivations (A.5), by first-order optimality, our estimator $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$ takes the form

$$\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} = (\lambda_1\Sigma_1+\lambda_2\Sigma_2)^{-1}\left(\lambda_1\Sigma_1\widehat{\beta}_1 + \lambda_2\Sigma_2\widehat{\beta}_2\right)$$

where $\widehat{\beta}_1, \widehat{\beta}_2$ are the LASSO estimates with ℓ_1-penalty. Since the eigenvalues of Σ_k are larger than γ^{-1}, the eigenvalues of $(\lambda_1\Sigma_1+\lambda_2\Sigma_2)^{-1}$ are smaller than γ, and so

$$\begin{aligned}
\left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2^2 &= \left\|(\lambda_1\Sigma_1+\lambda_2\Sigma_2)^{-1}\left(\lambda_1\Sigma_1\widehat{\beta}_1+\lambda_2\Sigma_2\widehat{\beta}_2\right) - (\lambda_1\Sigma_1+\lambda_2\Sigma_2)^{-1}(\lambda_1\Sigma_1\beta_1+\lambda_2\Sigma_2\beta_2)\right\|_2^2\\
&\leqslant \left\|(\lambda_1\Sigma_1+\lambda_2\Sigma_2)^{-1}\right\|_2^2\left\|\lambda_1\Sigma_1(\beta_1-\widehat{\beta}_1)+\lambda_2\Sigma_2(\beta_2-\widehat{\beta}_2)\right\|_2^2\\
&\leqslant \gamma\left(\lambda_1\left\|\Sigma_1\right\|_2\left\|\beta_1-\widehat{\beta}_1\right\|_2 + \lambda_2\left\|\Sigma_2\right\|_2\left\|\beta_2-\widehat{\beta}_2\right\|_2\right)^2\\
&\leqslant \gamma^3\left(\left\|\beta_1-\widehat{\beta}_1\right\|_2 + \left\|\beta_1-\widehat{\beta}_1\right\|_2\right)^2\\
&\leqslant 2\gamma^3\left(\left\|\beta_1-\widehat{\beta}_1\right\|_2^2 + \left\|\beta_2-\widehat{\beta}_2\right\|_2^2\right).
\end{aligned} \tag{18}$$

To bound the error on the LASSO estimates, we can invoke standard results, such as Wainwright (2019, Theorem 7.13(a) or Theorem 7.19). To do so, we verify that the matrix $\mathbf{X}_k$ satisfies the restricted eigenvalue condition with constant γ^{-1}, and is column-normalized with constant γ: For any (1-sparse) vector $\beta_k \in \mathbb{R}^d$, we have

$$\frac{1}{n}\left\|\mathbf{X}_k\beta_k\right\|_2^2 = \beta_k^\top\Sigma_k\beta_k \geqslant \gamma^{-1}\left\|\beta_k\right\|_2^2$$

because we have assumed that $\mathbf{X}_k \in \mathfrak{X}(\gamma)$, and for the standard basis vectors $\mathbf{e}_i$ we have

$$\frac{1}{n}\left\|\mathbf{X}_k\mathbf{e}_i\right\|_2^2 = \mathbf{e}_i^\top\Sigma_k\mathbf{e}_i \leqslant \gamma\left\|\mathbf{e}_i\right\|_2^2 = \gamma.$$

Furthermore, choosing the penalty strength

$$\alpha = 6\gamma\sigma\sqrt{\frac{2\log d}{n}}$$

so that the event $\left\{\frac{1}{n}\left\|\mathbf{X}_k\xi^k\right\|_\infty \leqslant \alpha/2\right\}$ has probability at least at least $1-2d^{-4}$ (Wainwright, 2019, Example 7.14 with $t = 2\sqrt{2\log(d)/n}$). Hence, by Wainwright (2019, Theorem 7.13(a)), we get that

$$\left\|\beta_k-\widehat{\beta}_k\right\|_2 \leqslant 3\gamma\alpha = 18\gamma^2\sigma\sqrt{\frac{2\log d}{n}}$$

with probability at least $1 - 2d^{-4}$. Putting this together with (18), we get our upper bound

$$\left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2^2 \leqslant 2\gamma^3 \left(36\gamma^2\sigma\sqrt{\frac{2\log d}{n}}\right)^2 = 2592\frac{\gamma^7\sigma^2\log d}{n}$$

with probability at least $1 - 4d^{-4}$ where we take the union bound for both estimators $k \in \{1, 2\}$. Note that the upper bound can also be shown indirectly using Theorem 1 with known $\boldsymbol{\omega}$. This concludes the proof of Proposition 2. □

D.4 Proof of Theorem 1

Theorem 1. *Let Assumptions 2 and 3 hold with $\boldsymbol{\mu}$ and ζ, respectively. Let j be the index of the strongly convex objective ($\mu_j > 0$), and $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$ be the minimizer of (10) with linear scalarization, i.e., $(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}}) = \sum_{k=1}^K \lambda_k \mathcal{L}_k$. Then, for all $\boldsymbol{\lambda} \in \Delta^K$ with $\lambda_j > 0$ and $\widehat{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$ it holds that*

$$\left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 \leqslant \frac{\zeta(\vartheta_{\boldsymbol{\lambda}})}{s_{\boldsymbol{\lambda}}(\boldsymbol{\mu})} \sum_{k=1}^K \lambda_k \|\widehat{\theta}_k - \theta_k\|.$$

Proof. The proof of the theorem follows from the regularity assumptions and Pareto stationarity. It is based on arguments akin to using the Implicit Function Theorem for the stability of minimizers, similar to (Shvartsman, 2012, Theorem 3.1) or (Bonnans and Shapiro, 2000, Proposition 4.32).

By Assumption 2 and (A.3), we know that $\vartheta \mapsto (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \boldsymbol{\theta})$ is $s_{\boldsymbol{\lambda}}(\boldsymbol{\mu})$-strongly convex for all $\boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}}$, and because there is one j such that $\lambda_j\mu_j > 0$, we have $s_{\boldsymbol{\lambda}}(\boldsymbol{\mu}) > 0$. Hence, the map $\boldsymbol{\theta} \mapsto \vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \arg\min_{\vartheta}(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \boldsymbol{\theta})$ is well-defined on $\widetilde{\boldsymbol{\Theta}}$. We now show that this function is locally Lipschitz.

Define the function $F_{\boldsymbol{\lambda}} : \mathbb{R}^m \times \mathbb{R}^{Kp} \to \mathbb{R}^m$

$$F_{\boldsymbol{\lambda}}(\vartheta, \boldsymbol{\theta}) = \nabla_{\vartheta}(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \boldsymbol{\theta}) = \sum_{k=1}^K \lambda_k \nabla_{\vartheta}\mathcal{L}_k(\vartheta, \theta_k).$$

By first-order Pareto stationarity (Roy et al., 2023), we know that $F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}), \boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}}$. Let $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \widetilde{\boldsymbol{\Theta}} \subset \mathbb{R}^{Kp}$, so that $F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}), \boldsymbol{\theta}) = 0$ and $F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'), \boldsymbol{\theta}') = 0$ and hence

$$\left\|F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}), \boldsymbol{\theta}) - F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'), \boldsymbol{\theta})\right\|_2 = \left\|F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'), \boldsymbol{\theta}') - F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'), \boldsymbol{\theta})\right\|_2.$$

Using $s_{\boldsymbol{\lambda}}(\boldsymbol{\mu})$-strong convexity of $\vartheta \mapsto (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \boldsymbol{\theta})$ and (A.1), we can now lower-bound the left-hand side as

$$\left\|F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}), \boldsymbol{\theta}) - F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'), \boldsymbol{\theta})\right\|_2 \geqslant s_{\boldsymbol{\lambda}}(\boldsymbol{\mu}) \left\|\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) - \vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}')\right\|_2$$

and using local Lipschitz-continuity of $\theta_k \mapsto \nabla_{\vartheta}\mathcal{L}_k(\vartheta, \theta_k)$ (Assumption 3), we can upper bound the right-hand side as

$$\begin{aligned}\left\|F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'), \boldsymbol{\theta}') - F_{\boldsymbol{\lambda}}(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'), \boldsymbol{\theta})\right\|_2 &= \left\|\sum_{k=1}^K \lambda_k \left(\nabla_{\vartheta}\mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'), \theta_k') - \nabla_{\vartheta}\mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'), \theta_k)\right)\right\|_2 \\ &\leqslant \sum_{k=1}^K \lambda_k \zeta_k(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}')) \left\|\theta_k' - \theta_k\right\|_2\end{aligned}$$

where $\zeta_k(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'))$ denotes the local Lipschitz constant. Combining the two and letting $\zeta(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}')) = \max_k \zeta_k(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'))$, we get that

$$\left\|\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) - \vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}')\right\|_2 \leqslant \frac{\zeta(\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta}'))}{s_{\boldsymbol{\lambda}}(\boldsymbol{\mu})} \sum_{k=1}^K \lambda_k \left\|\theta_k' - \theta_k\right\|_2.$$

Since we assumed that the true parameters satisfy $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \widetilde{\boldsymbol{\Theta}}$ and the estimators satisfy $\widehat{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$ we can plug them in, that is, $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} = \vartheta_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})$ and $\vartheta_{\boldsymbol{\lambda}} = \vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$, we get

$$\left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 \leqslant \frac{\zeta(\vartheta_{\boldsymbol{\lambda}})}{s_{\boldsymbol{\lambda}}(\boldsymbol{\mu})} \sum_{k=1}^K \lambda_k \left\|\widehat{\theta}_k - \theta_k\right\|_2$$

which is the desired upper bound. □

D.5 Proof of Proposition 4

Proposition 4 (Lipschitz parameterization)**.** *Assume that the parameterization $\theta_k \mapsto \mathcal{L}_k(\cdot, \theta_k)$ is 1-Lipschitz continuous with respect to the function $\Phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, in the sense that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \widetilde{\boldsymbol{\Theta}}$ it holds*

$$\sup_{\vartheta \in \mathbb{R}^m} \left| \mathcal{L}_k(\vartheta, \theta_k) - \mathcal{L}_k(\vartheta, \theta_k') \right| \leqslant \Phi(\theta_k, \theta_k').$$

Then, for any scalarization of the form $s_{\boldsymbol{\lambda}}(x) = \|\boldsymbol{\lambda} \odot x\|$ with some norm $\|\cdot\|$, the excess scalarized loss of $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$—as defined in Equation (7)—is bounded by

$$\mathcal{E}_{\boldsymbol{\lambda}}(\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}) \leqslant 2 s_{\boldsymbol{\lambda}} \left((\Phi(\widehat{\theta}_k, \theta_k))_{k=1}^K \right).$$

Proof. We use the standard uniform learning bound

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{\lambda}}(\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}) &= (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}, \boldsymbol{\theta}) - (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta_{\boldsymbol{\lambda}}, \boldsymbol{\theta}) \\
&= (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}, \boldsymbol{\theta}) - (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}}) + \underbrace{(s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}}) - (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta_{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})}_{\leqslant 0} + (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta_{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}}) - (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta_{\boldsymbol{\lambda}}, \boldsymbol{\theta}) \\
&\leqslant 2 \sup_{\vartheta \in \mathbb{R}^d} \left| (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \boldsymbol{\theta}) - (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta, \widehat{\boldsymbol{\theta}}) \right| \\
&= 2 \sup_{\vartheta \in \mathbb{R}^d} \left| \|\boldsymbol{\lambda} \odot \boldsymbol{\mathcal{L}}(\vartheta, \boldsymbol{\theta})\| - \left\|\boldsymbol{\lambda} \odot \boldsymbol{\mathcal{L}}(\vartheta, \widehat{\boldsymbol{\theta}})\right\| \right|.
\end{aligned}$$

Using the reverse triangle inequality, we can further bound this as

$$\mathcal{E}_{\boldsymbol{\lambda}}(\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}) \leqslant 2 \sup_{\vartheta \in \mathbb{R}^d} \left\| \boldsymbol{\lambda} \odot \boldsymbol{\mathcal{L}}(\vartheta, \boldsymbol{\theta}) - \boldsymbol{\lambda} \odot \boldsymbol{\mathcal{L}}(\vartheta, \widehat{\boldsymbol{\theta}}) \right\| = 2 \sup_{\vartheta \in \mathbb{R}^d} \left\| \boldsymbol{\lambda} \odot \left(\boldsymbol{\mathcal{L}}(\vartheta, \boldsymbol{\theta}) - \boldsymbol{\mathcal{L}}(\vartheta, \widehat{\boldsymbol{\theta}}) \right) \right\| \leqslant 2 \left\| \boldsymbol{\lambda} \odot \left(\Phi(\theta_k, \widehat{\theta}_k) \right)_{k=1}^K \right\|$$

where the last inequality holds by assumption, and hence, we have the result. □

D.6 Proof of Theorem 2

Theorem 2. *If Assumption 4 holds and we use linear scalarization, the minimax rate is lower bounded as*

$$\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \geqslant \max_{k \in [K]} (1 + s_{\boldsymbol{\lambda}}(\boldsymbol{\eta}))^{-1} \left(\frac{\lambda_k}{\eta_k'} \delta_k - \sum_{i \neq k} \eta_i \lambda_i \delta_i \right)_+$$

where $(\cdot)_+ := \max\{0, \cdot\}$.

Proof of Theorem 2. The proof of Theorem 2 makes use of the identifiability assumption from Assumption 4. We prove the lower bound by contradiction: Assuming that we could estimate $\vartheta_{\boldsymbol{\lambda}}$ better than the stated bound (in a minimax sense), we show that the identifiability assumption implies that we could also estimate one of the parameters θ_k faster than the minimax rate δ_k. As this is impossible by definition, the lower bound follows.

Note that by first-order optimality, we have that

$$\nabla_\vartheta (s_{\boldsymbol{\lambda}} \circ \boldsymbol{\mathcal{L}})(\vartheta_{\boldsymbol{\lambda}}, \boldsymbol{\theta}) = 0,$$

which is known as Pareto stationarity (Roy et al., 2023). Expanding and rearranging yields for every k

$$\nabla_\vartheta \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}, \theta_k) = -\frac{1}{\lambda_k} \sum_{i \neq k} \lambda_i \nabla_\vartheta \mathcal{L}_i(\vartheta_{\boldsymbol{\lambda}}, \theta_i).$$

Since by Assumption 4, the map $g_k(\cdot; \vartheta_{\boldsymbol{\lambda}}) : \theta_k \mapsto \nabla_\vartheta \mathcal{L}_k(\vartheta_{\boldsymbol{\lambda}}, \theta_k)$ is injective for every $\vartheta_{\boldsymbol{\lambda}}$, we can write

$$\theta_k = g_k^{-1} \left(-\frac{1}{\lambda_k} \sum_{i \neq k} \lambda_i \nabla_\vartheta \mathcal{L}_i(\vartheta_{\boldsymbol{\lambda}}, \theta_i); \vartheta_{\boldsymbol{\lambda}} \right).$$

Now, given any estimators $\widehat{\vartheta}_{\boldsymbol{\lambda}}$ of $\vartheta_{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ we define the new estimator

$$\widehat{\theta}_k^{\text{new}} = g_k^{-1}\left(-\frac{1}{\lambda_k}\sum_{i\neq k}\lambda_i \nabla_\vartheta \mathcal{L}_i(\widehat{\vartheta}_{\boldsymbol{\lambda}}, \widehat{\theta}_i); \widehat{\vartheta}_{\boldsymbol{\lambda}}\right).$$

We can bound the error of this new estimator using the Lipschitz-continuity of the left-inverse (Assumption 4) and get

$$\begin{aligned}\left\|\widehat{\theta}_k^{\text{new}} - \theta_k\right\| &\leqslant \eta_k'\left(\left\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 + \left\|\frac{1}{\lambda_k}\sum_{i\neq k}\lambda_i\left(\nabla_\vartheta \mathcal{L}_i(\widehat{\vartheta}_{\boldsymbol{\lambda}}, \widehat{\theta}_i) - \nabla_\vartheta \mathcal{L}_i(\vartheta_{\boldsymbol{\lambda}}, \theta_i)\right)\right\|_2\right)\\ &\leqslant \eta_k'\left(\left\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 + \frac{1}{\lambda_k}\sum_{i\neq k}\lambda_i\left\|\nabla_\vartheta \mathcal{L}_i(\widehat{\vartheta}_{\boldsymbol{\lambda}}, \widehat{\theta}_i) - \nabla_\vartheta \mathcal{L}_i(\vartheta_{\boldsymbol{\lambda}}, \theta_i)\right\|_2\right).\end{aligned}$$

Further, by Assumption 4 we have for all estimators $\widehat{\theta}_i$

$$\left\|\nabla_\vartheta \mathcal{L}_i(\widehat{\vartheta}_{\boldsymbol{\lambda}}, \widehat{\theta}_i) - \nabla_\vartheta \mathcal{L}_i(\vartheta_{\boldsymbol{\lambda}}, \theta_i)\right\|_2 \leqslant \eta_k\left(\left\|\widehat{\theta}_i - \theta_i\right\| + \left\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2\right).$$

Consequently, we get

$$\begin{aligned}\left\|\widehat{\theta}_k^{\text{new}} - \theta_k\right\| &\leqslant \eta_k'\left(\left\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 + \frac{1}{\lambda_k}\sum_{i\neq k}\lambda_i\left\|\nabla_\vartheta \mathcal{L}_i(\widehat{\vartheta}_{\boldsymbol{\lambda}}, \widehat{\theta}_i) - \nabla_\vartheta \mathcal{L}_i(\vartheta_{\boldsymbol{\lambda}}, \theta_i)\right\|_2\right)\\ &\leqslant \eta_k'\left(\left\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 + \frac{1}{\lambda_k}\sum_{i\neq k}\lambda_i\eta_i\left(\left\|\widehat{\theta}_i - \theta_i\right\| + \left\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2\right)\right)\\ &\leqslant (1 + s_{\boldsymbol{\lambda}}(\boldsymbol{\eta}))\frac{\eta_k'}{\lambda_k}\left\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 + \frac{\eta_k'}{\lambda_k}\sum_{i\neq k}\lambda_i\eta_i\left\|\widehat{\theta}_i - \theta_i\right\|\end{aligned}$$

which we can rearrange to get

$$\left\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 \geqslant (1 + s_{\boldsymbol{\lambda}}(\boldsymbol{\eta}))^{-1}\left(\frac{\lambda_k}{\eta_k'}\left\|\widehat{\theta}_k^{\text{new}} - \theta_k\right\| - \sum_{i\neq k}\eta_i\lambda_i\left\|\widehat{\theta}_i - \theta_i\right\|\right).$$

Notice that this lower bound holds for *all* choices of estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\vartheta}_{\boldsymbol{\lambda}}$. Choosing $\widehat{\theta}_k$ to be minimax optimal for all $k \in [K]$, we then obtain for all $k \in [K]$ that

$$\sup_{\mathbb{P}\in\boldsymbol{\mathcal{P}}}\mathbb{E}\left\|\widehat{\theta}_k^{\text{new}} - \theta_k\right\| \geqslant \sup_{\mathbb{P}\in\boldsymbol{\mathcal{P}}}\mathbb{E}\left[\left\|\widehat{\theta}_k - \theta_k\right\|\right] = \delta_k.$$

Hence, using the notation $\widehat{\theta}_k^{\text{new}} = \widehat{\theta}_k^{\text{new}}(\widehat{\vartheta}_{\boldsymbol{\lambda}})$ to highlight the dependence, this yields the lower bound

$$\begin{aligned}\inf_{\widehat{\vartheta}_{\boldsymbol{\lambda}}}\sup_{\mathbb{P}\in\boldsymbol{\mathcal{P}}}\mathbb{E}\left[\left\|\widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2\right] &\geqslant \inf_{\widehat{\vartheta}_{\boldsymbol{\lambda}}}\sup_{\mathbb{P}\in\boldsymbol{\mathcal{P}}}\mathbb{E}\left[(1 + s_{\boldsymbol{\lambda}}(\boldsymbol{\eta}))^{-1}\left(\frac{\lambda_k}{\eta_k'}\left\|\widehat{\theta}_k^{\text{new}}(\widehat{\vartheta}_{\boldsymbol{\lambda}}) - \theta_k\right\| - \sum_{i\neq k}\eta_i\lambda_i\left\|\widehat{\theta}_i - \theta_i\right\|\right)\right]\\ &\geqslant (1 + s_{\boldsymbol{\lambda}}(\boldsymbol{\eta}))^{-1}\left(\frac{\lambda_k}{\eta_k'}\inf_{\widehat{\vartheta}_{\boldsymbol{\lambda}}}\sup_{\mathbb{P}\in\boldsymbol{\mathcal{P}}}\mathbb{E}\left[\left\|\widehat{\theta}_k^{\text{new}}(\widehat{\vartheta}_{\boldsymbol{\lambda}}) - \theta_k\right\|\right] - \sum_{i\neq k}\eta_i\lambda_i\sup_{\mathbb{P}\in\boldsymbol{\mathcal{P}}}\mathbb{E}\left[\left\|\widehat{\theta}_i - \theta_i\right\|\right]\right)\\ &\geqslant (1 + s_{\boldsymbol{\lambda}}(\boldsymbol{\eta}))^{-1}\left(\frac{\lambda_k}{\eta_k'}\delta_k - \sum_{i\neq k}\eta_i\lambda_i\delta_i\right).\end{aligned}$$

And of course, 0 is a trivial lower bound. Since the argument was valid for any k, we can take the maximum over k. This concludes the proof of Theorem 1. □

D.7 Proof of Corollary 1

Assumption D.1 (Scalings). *1. We let* $b \lesssim 1 \lesssim \sigma$.

2. *For all* $k \in [K]$, $n_k + N_k \gtrsim d(B^4/b^4)$.
3. *For all* $k \in [K]$, $n_k \gtrsim (B^2/b^2)\sigma^2 s \log d$.
4. *For the lower bound we assume that for some large enough universal constant* $C > 0$ *and some* $k \in [K]$, *it holds that* $\lambda_k \frac{b^3}{B^2} n_k^{-1/2} \geqslant C \sum_{i \neq k} \lambda_i n_i^{-1/2}$. *This assumption corresponds to* (11).

Corollary 1. *Let Assumption D.1 (in Appendix D.7) hold. Then* $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$ *from* (12) *achieves for all* $\boldsymbol{\lambda} \in \Delta^K$ *and* $\mathbb{P} \in \boldsymbol{\mathcal{P}}$

$$\left\| \widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2 \lesssim \frac{B^4}{b^4} \sum_{k=1}^{K} \lambda_k \left(\frac{\sigma}{b^2} \sqrt{\frac{s \log d}{n_k}} + \sqrt{\frac{d}{n_k + N_k}} \right)$$

with probability at least $1 - c_1 K(d^{-3} + \exp(-c_2 B^4 d))$, *where* $c_1, c_2 > 0$ *are some universal constants. If further* $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$ *is fixed and known, then*

$$\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \gtrsim \max_{k \in [K]} \lambda_k \frac{b^2 \sigma}{B^3} \sqrt{\frac{s \log d}{n_k}}.$$

Proof. To prove the upper bound of Corollary 1, we apply Theorem 1, to prove the lower bound, we apply Theorem 2.

By definition of $\boldsymbol{\mathcal{P}}$ in Example 1, we have (neglecting vectorization) that $\boldsymbol{\Theta} = \Theta^K$, where

$$\Theta = \left\{ (\beta, \Sigma) \in \mathbb{R}^d \times \mathbb{R}^{d \times d} \;\middle|\; \|\beta\|_0 \leqslant s, \|\beta\|_2 \leqslant 1, \Sigma \text{ is symmetric and } b^2 \mathbf{I}_d \preceq \Sigma \preceq B^2 \mathbf{I}_d \right\}.$$

The lower bound on Σ was assumed, and the upper bound holds because X was assumed to be B^2-sub-Gaussian, implying that the largest eigenvalue is bounded as

$$\lambda_{\max}(\mathbb{E}\left[XX^\top\right]) = \sup_{\|v\|_2 = 1} v^\top \mathbb{E}\left[XX^\top\right] v = \sup_{\|v\|_2 = 1} \mathbb{E}\left[\langle X, v \rangle^2\right] \leqslant B^2.$$

To apply Theorems 1 and 2, we show that Assumption 2, 3 and 4 hold with the sets $\widetilde{\boldsymbol{\Theta}} = \widetilde{\Theta}^K$, where

$$\widetilde{\Theta} = \left\{ (\beta, \Sigma) \in \mathbb{R}^d \times \mathbb{R}^{d \times d} \;\middle|\; \|\beta\|_2 \leqslant 2, \Sigma \text{ is symmetric and } (1/2) b^2 \mathbf{I}_d \preceq \Sigma \preceq 2B^2 \mathbf{I}_d \right\},$$

and then use that our estimates lie in these sets with high probability.

Upper bound. It is easy to show that Assumption 2 holds. Recall that the objectives take the form

$$\mathcal{L}_k(\vartheta, \beta_k, \Sigma_k) = \left\| \Sigma_k^{1/2}(\vartheta - \beta_k) \right\|_2^2 + \sigma^2 \quad \text{where} \quad (1/2) b^2 \mathbf{I}_d \preceq \Sigma_k \preceq 2B^2 \mathbf{I}_d.$$

Hence $\vartheta \mapsto \mathcal{L}_k(\vartheta, \vartheta_k, \Sigma_k)$ is strongly convex with parameter $\mu_k = (1/2)b^2$ (Bubeck, 2015, §3.4) (and smooth with parameter $4B^2$). Recall Assumption 3. Now, the map $(\beta_k, \Sigma_k) \mapsto \nabla_\vartheta \mathcal{L}_k(\vartheta, \beta_k, \Sigma_k)$ is locally Lipschitz on $\widetilde{\Theta}$, since

$$\begin{aligned}
\left\| \nabla_\vartheta \mathcal{L}_k(\vartheta, \beta_k, \Sigma_k) - \nabla_\vartheta \mathcal{L}_k(\vartheta, \beta_k', \Sigma_k') \right\|_2 &= 2 \left\| \Sigma_k(\vartheta - \beta_k) - \Sigma_k'(\vartheta - \beta_k') \right\|_2 \\
&= 2 \left\| \left(\Sigma_k - \Sigma_k'\right)(\vartheta - \beta_k) + \Sigma_k'(\beta_k - \beta_k') \right\|_2 \\
&\leqslant 2 \left\| \vartheta - \beta_k \right\|_2 \left\| \Sigma_k - \Sigma_k' \right\|_2 + 2 \left\| \Sigma_k' \right\|_2 \left\| \beta_k - \beta_k' \right\|_2 \\
&\leqslant 2 \max\left\{ \left\| \vartheta - \beta_k \right\|_2, \left\| \Sigma_k' \right\|_2 \right\} \left(\left\| \Sigma_k - \Sigma_k' \right\|_2 + \left\| \beta_k - \beta_k' \right\|_2 \right) \\
&\leqslant \underbrace{2 \max\left\{ \left\| \vartheta \right\|_2 + 2, 2B^2 \right\}}_{\zeta_k(\vartheta)} \left(\left\| \Sigma_k - \Sigma_k' \right\|_2 + \left\| \beta_k - \beta_k' \right\|_2 \right)
\end{aligned}$$

where the local Lipschitz constant $\zeta_k(\vartheta)$ depends on ϑ and B. Hence, Assumption 3 is satisfied with the norm $\|\theta_k\| = \|(\beta_k, \Sigma_k)\| = \|\beta_k\|_2 + \|\Sigma_k\|_2$.

We now prove that the estimated parameters $\widehat{\theta}_k = (\widehat{\beta}_k, \widehat{\Sigma}_k)$, where $\widehat{\Sigma}_k$ is the sample covariance matrix and $\widehat{\beta}_k$ the LASSO estimate, are contained in $\widetilde{\Theta}$ with high probability for all k. Recall the definition of $\widehat{\Sigma}_k$: Denote $\mathbf{X}_k$ the covariate sample matrix that has the labeled samples X_i^k, $i = 1, \ldots, n_k$ as its rows, $\widetilde{\mathbf{X}}_k$ the sample matrix that has the unlabeled samples X_i^k, $i = n_k + 1, \ldots, N_k$ as its rows, and $y^k = (Y_1^k, \ldots, Y_n^k)^\top$ according to (4). We define the sample covariance matrix as

$$\widehat{\Sigma}_k = \frac{1}{n_k + N_k} \begin{pmatrix} \mathbf{X}_k^\top & \widetilde{\mathbf{X}}_k^\top \end{pmatrix} \begin{pmatrix} \mathbf{X}_k \\ \widetilde{\mathbf{X}}_k \end{pmatrix}.$$

We let $\widehat{\beta}_k$ be the LASSO estimates with ℓ_1-penalty $\alpha \|\cdot\|_1$, where $\alpha_k = 136B\sigma\sqrt{\log(d)/n_k}$, that is,

$$\widehat{\beta}_k = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d} \frac{1}{n_k} \left\|\mathbf{X}_k\beta - y^k\right\|_2^2 + \alpha_k \|\beta\|_1 .$$

For some universal constants $c_1 > 0$, we define the event

$$\mathcal{E}_1 = \left\{\forall k \in [K] : \left\|\widehat{\Sigma}_k - \Sigma_k\right\|_2 \leqslant c_1 B^2 \sqrt{\frac{d}{n_k + N_k}}\right\}.$$

By Wainwright (2019, Theorem 6.5), or Vershynin (2010, Corollary 5.50), and a union bound, $\mathcal{E}_1$ has probability at least $1 - c_2 K \cdot \exp\left(-c_3 B^4 d\right)$, where $c_2, c_3 > 0$ are two more universal constants. On $\mathcal{E}_1$, it holds that

$$\lambda_{\min}(\widehat{\Sigma}_k) \geqslant \lambda_{\min}(\Sigma_k) - \left\|\widehat{\Sigma}_k - \Sigma_k\right\|_2 \geqslant b^2 - c_1 B^2 \sqrt{\frac{d}{n_k + N_k}} \geqslant \frac{b^2}{2},$$
$$\lambda_{\max}(\widehat{\Sigma}_k) \leqslant \lambda_{\max}(\Sigma_k) + \left\|\widehat{\Sigma}_k - \Sigma_k\right\|_2 \leqslant B^2 + c_1 B^2 \sqrt{\frac{d}{n_k + N_k}} \leqslant 2B^2,$$

where we used the assumption that $c_1 B^2\sqrt{d/(n_k + N_k)} \leqslant b^2/2 \wedge B^2$ (the second follows from $n_k + N_k \geqslant d/c_1^2$). We also define for some universal constant $c_4 > 0$ the event

$$\mathcal{E}_2 = \left\{\forall k \in [K] : \left\|\widehat{\beta}_k - \beta_k\right\|_2 \leqslant c_4 \frac{B\sigma}{b}\sqrt{\frac{s\log d}{n_k}}\right\}$$

By Wainwright (2019, Theorem 7.19), this event holds whenever for all $k \in [K]$ we have

$$\frac{2}{n}\left\|\mathbf{X}_k^\top \xi^k\right\|_\infty \leqslant \alpha_k = 136B\sigma\sqrt{\frac{\log d}{n_k}}. \tag{19}$$

which, we now show holds with probability at least $1 - K/(4d^3)$. Since $\mathbf{X}_k$ and ξ^k are independent, we have that $(\mathbf{X}_k^\top \xi^k | \mathbf{X}_k) \sim \mathcal{N}(0, \sigma^2 \left\|X_{:,j}^k\right\|_2^2)$ (where $X_{:,j}^k$ is the j-th column of $\mathbf{X}_k$), and denoting $\kappa = \max_{j\in[d]} \left\|X_{:,j}^k\right\|_2/\sqrt{n_k}$, we get

$$\mathbb{P}\left(\frac{2}{n}\left\|\mathbf{X}_k^\top \xi^k\right\|_\infty \leqslant 8\sigma\left(\frac{\kappa}{\sqrt{n_k}}\right)\sqrt{\frac{\log d}{n_k}} \,\middle|\, \mathbf{X}_k\right) \geqslant 1 - \frac{1}{8d^3}.$$

It remains to bound $\kappa/\sqrt{n_k}$, which we can do using $\rho(\Sigma_k) = \max_{j\in[d]}\sqrt{(\Sigma_k)_{jj}} \leqslant \sqrt{\lambda_{\max}(\Sigma_k)} \leqslant B$ as

$$\mathbb{P}\left(\frac{\kappa}{\sqrt{n_k}} \leqslant B\left(1 + 16\sqrt{\frac{\log d}{n_k}}\right)\right) \geqslant 1 - d\exp\left(-4\sqrt{n\log d}\right),$$

cf. Raskutti et al. (2011, Equation (23) and Appendix I). Therefore, if $n_k \geqslant \log d$, we have with probability at least $1 - 1/(4d^3)$ that Equation (19) holds for every $k \in [K]$ separately, and taking union bound yields $\mathbb{P}(\mathcal{E}_2) \geqslant 1 - K/(4d^3)$.

Thus, on $\mathcal{E}_2$, we also have that for all $k \in [K]$, $\|\widehat{\beta}_k\|_2 \leqslant 2$, because we assumed that $c_4\frac{B\sigma}{b}\sqrt{\frac{s\log d}{n_k}} \leqslant 1$.

Consequently, on $\mathcal{E}_1 \cap \mathcal{E}_2$ we have that for all $k \in [K]$, $\widehat{\theta}_k = (\widehat{\beta}_k, \widehat{\Sigma}_k) \in \widetilde{\Theta}$. Hence, we may apply the upper bound from Theorem 1 with $s_{\boldsymbol{\lambda}}(\boldsymbol{\mu}) = (1/2)b^2$, that is,

$$\begin{aligned}\left\|\widehat{\vartheta}_{\boldsymbol{\lambda}}^{\mathrm{ts}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 &\leqslant \frac{2\zeta(\vartheta_{\boldsymbol{\lambda}})}{b^2}\sum_{k=1}^K \lambda_k\left(\|\widehat{\beta}_k - \beta_k\|_2 + \left\|\widehat{\Sigma}_k - \Sigma_k\right\|_2\right)\\ &\lesssim \frac{\zeta(\vartheta_{\boldsymbol{\lambda}})}{b^2}\sum_{k=1}^K \lambda_k\left(\frac{B\sigma}{b}\sqrt{\frac{s\log d}{n_k}} + B^2\sqrt{\frac{d}{n_k + N_k}}\right)\end{aligned}$$

Notably, we can turn it into an explicit uniform upper bound by bounding $\vartheta_{\boldsymbol{\lambda}} = \vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ as

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \zeta(\vartheta_{\boldsymbol{\lambda}}) \leqslant \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} 2\max\left\{\|\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\|_2 + 2, 2B^2\right\} = 2\max\left\{\frac{B^2}{b^2} + 2, 2B^2\right\}$$

where we used the simple bound from (A.5)

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\vartheta_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\|_2 = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\| \left(\sum_{k=1}^K \lambda_k \Sigma_k\right)^{-1} \left(\sum_{k=1}^K \lambda_k \Sigma_k \beta_k\right) \right\|_2 \leqslant \frac{B^2}{b^2}.$$

Hence, using $2\max\left\{\frac{B^2}{b^2}+2, 2B^2\right\}/b^2 \leqslant 6\frac{B^2}{b^2}\max\left\{\frac{1}{b^2},1\right\} = 6\frac{B^2}{b^4}$ since $b \leqslant 1$, the bound becomes

$$\begin{aligned}\left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 &\lesssim \frac{B^2}{b^4}\sum_{k=1}^K \lambda_k \left(\frac{B\sigma}{b}\sqrt{\frac{s\log d}{n_k}} + B^2\sqrt{\frac{d}{n_k+N_k}}\right) \\ &\leqslant \frac{B^4}{b^4}\sum_{k=1}^K \lambda_k \left(\frac{\sigma}{b^2}\sqrt{\frac{s\log d}{n_k}} + \sqrt{\frac{d}{n_k+N_k}}\right)\end{aligned}$$

which concludes the first part of the proof.

Lower bound. Throughout, denote $\boldsymbol{\Sigma}$ as the tuple of fixed covariance matrices that are known to the algorithm, each satisfying $b^2\mathbf{I}_d \preceq \Sigma_k \preceq B^2\mathbf{I}_d$. We first have to check that Assumption 4 holds. Note that $\nabla_\vartheta \mathcal{L}_k(\vartheta, \beta_k, \Sigma_k) = 2\Sigma_k(\vartheta - \beta_k)$, so that $g_k^{-1}(y,\vartheta) = \vartheta - \frac{1}{2}\Sigma_k^{-1}y$ and hence

$$\begin{aligned}\left\|g_k(\beta_k,\vartheta) - g_k(\beta_k',\vartheta')\right\|_2 &= \left\|2\Sigma_k(\vartheta-\beta_k) - 2\Sigma_k(\vartheta'-\beta_k')\right\|_2 \leqslant 2B^2\left(\left\|\vartheta-\vartheta'\right\|_2 + \left\|\beta_k-\beta_k'\right\|_2\right), \\ \left\|g_k^{-1}(y,\vartheta) - g_k^{-1}(y',\vartheta')\right\|_2 &= \left\|\vartheta - \frac{1}{2}\Sigma_k^{-1}y - \vartheta' + \frac{1}{2}\Sigma_k^{-1}y'\right\|_2 \leqslant \left\|\vartheta-\vartheta'\right\|_2 + \frac{1}{2b^2}\left\|y-y'\right\|_2,\end{aligned}$$

and therefore Assumption 4 holds with $\eta_k = 2B^2$ and $\eta_k' = \max\left\{1, 1/(2b^2)\right\} = 1/(2b^2)$ as we assume $b^2 \leqslant 1/2$. We can hence apply the lower bound from Theorem 1. For that, we use Wainwright (2019, Example 15.16), which is an application of Fano's method (Yu, 1997) with local packings. Conditioned on $\mathbf{X}_k$, it yields the minimax lower bound

$$\inf_{\widehat{\beta}_k}\sup_{\substack{\|\beta_k\|_0\leqslant s,\\ \|\beta_k\|_2\leqslant 1}} \mathbb{E}\left[\left\|\widehat{\beta}_k - \beta_k\right\|_2 \,\middle|\, \mathbf{X}_k\right] \gtrsim \frac{\sigma}{\gamma(\mathbf{X}_k)}\sqrt{\frac{s\log d}{n_k}}$$

where $\gamma(\mathbf{X}_k) = \max_{|S|=2s}\sigma_{\max}((\mathbf{X}_k)_S/\sqrt{n})$ and $\sigma_{\max}$ denotes the largest singular value. For any $|S| = 2s$, we can bound this using (Wainwright, 2019, Equation (6.16))

$$\mathbb{E}\left[\sigma_{\max}((\mathbf{X}_k)_S/\sqrt{n})\right] \leqslant \sqrt{\lambda_{\max}((\Sigma_k)_S)} + \sqrt{\frac{\mathrm{tr}\left((\Sigma_k)_S\right)}{n_k}}$$

As $\Sigma_k \preceq B^2\mathbf{I}_d$, this is bounded by $B + B\sqrt{s/n_k} = B(1+\sqrt{s/n_k})$. Hence, if $n_k \geqslant s$, we have by Jensen's inequality

$$\inf_{\widehat{\beta}_k}\sup_{\substack{\|\beta_k\|_0\leqslant s,\\ \|\beta_k\|_2\leqslant 1}} \mathbb{E}\left[\mathbb{E}\left[\left\|\widehat{\beta}_k - \beta_k\right\|_2 \,\middle|\, \mathbf{X}_k\right]\right] \gtrsim \frac{\sigma}{\mathbb{E}\left[\gamma(\mathbf{X}_k)\right]}\sqrt{\frac{s\log d}{n_k}} \gtrsim \frac{\sigma}{B}\sqrt{\frac{s\log d}{n_k}}.$$

Therefore, combining this lower bound with the upper bound on the minimax rate from the previous section of the proof, Theorem 1 yields that if for some k, with large enough constant $C > 0$ it holds

$$\lambda_k b^2 \frac{\sigma}{B}\sqrt{\frac{s\log d}{n_k}} \geqslant C\sum_{i\neq k}\lambda_i \frac{\sigma B}{b}\sqrt{\frac{s\log d}{n_i}},$$

we get using $(1+s_{\boldsymbol{\lambda}}(\boldsymbol{\eta}))^{-1} \geqslant 1/(1+2B^2) \gtrsim 1/B^2$ that

$$\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \gtrsim \max_{k\in[K]} \lambda_k \frac{b^2\sigma}{B^3}\sqrt{\frac{s\log d}{n_k}}.$$

As we assumed the above condition, up to canceling and rearranging terms, this finished the proof. □

D.8 Proof of Corollary 2

Corollary 2. *In the setting of Example 2, assume that $n \gtrsim \sigma^2 s \log d$. Then the two-stage estmator $\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}}$ from (13) achieves for all $\boldsymbol{\lambda} \in \Delta^2$ with $\lambda_{\mathrm{risk}} > 0$*

$$\left\|\widehat{\vartheta}^{\mathrm{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 \lesssim \sigma\sqrt{\frac{s\log d}{n}} + \frac{1}{\lambda_{\mathrm{risk}}}\sqrt{\frac{d}{n+N}}$$

with probability at least $1 - cd^{-3}$, where $c > 0$ is some universal constant.

Proof. As before, to be able to apply the upper bound from Theorem 1, we need to check that Assumption 2 and 3 hold.

First, notice that in the specific distribution that we are considering we have

$$\begin{aligned}
\mathbb{E}\left[XX^\top\right] &= \frac{1}{2}\mathbb{E}\left[XX^\top | A = 1\right] + \frac{1}{2}\mathbb{E}\left[XX^\top | A = -1\right] \\
&= \frac{1}{2}\left(\mathbb{E}\left[(X-\mu)(X-\mu)^\top | A = 1\right] + \mu_1\mu_1^\top\right) + \frac{1}{2}\left(\mathbb{E}\left[(X+\mu)(X+\mu)^\top | A = -1\right] + \mu\mu^\top\right) \\
&= \mathbf{I}_d + \mu\mu^\top.
\end{aligned}$$

and hence we can write

$$\mathcal{L}_{\mathrm{risk}}(\vartheta, \beta, \mu) = \left\|(\mathbf{I}_d + \mu\mu^\top)^{1/2}(\vartheta - \beta)\right\|_2^2 + \sigma^2 \implies \nabla_\vartheta \mathcal{L}_{\mathrm{risk}}(\vartheta, \beta, \mu) = 2(\mathbf{I}_d + \mu\mu^\top)(\vartheta - \beta).$$

Moreover, by Lemma B.1, we can write $\mathcal{L}_{\mathrm{fair}}(\vartheta, \mathbb{P})$ as

$$\mathcal{L}_{\mathrm{fair}}(\vartheta, \mu) = \langle \mu, \vartheta\rangle^2 \implies \nabla\mathcal{L}_{\mathrm{fair}}(\vartheta, \mu) = 2\langle \mu, \vartheta\rangle \mu.$$

We may define $\boldsymbol{\Theta} = \Theta_{\mathrm{risk}} \times \Theta_{\mathrm{fair}}$ and $\widetilde{\boldsymbol{\Theta}} = \widetilde{\Theta}_{\mathrm{risk}} \times \widetilde{\Theta}_{\mathrm{fair}}$, where

$$\begin{aligned}
\boldsymbol{\Theta} &= \left\{(\theta_{\mathrm{risk}}, \theta_{\mathrm{fair}}) = ((\beta, \mu), \mu) \in \mathbb{R}^{3d} \;\middle|\; \|\beta\|_0 \leqslant s, \|\beta\|_2 \leqslant 1, \|\mu\|_2 \leqslant 1\right\}, \\
\widetilde{\boldsymbol{\Theta}} &= \left\{(\theta_{\mathrm{risk}}, \theta_{\mathrm{fair}}) = ((\beta, \mu), \mu) \in \mathbb{R}^{3d} \;\middle|\; \|\beta\|_2 \leqslant 2, \|\mu\|_2 \leqslant 2\right\}.
\end{aligned}$$

Clearly, $\vartheta \mapsto \mathcal{L}_{\mathrm{risk}}(\vartheta, \beta, \mu)$ is 1-strongly convex (Bubeck, 2015, §3.4) and continuously differentiable for all $((\beta, \mu), \mu) \in \widetilde{\boldsymbol{\Theta}}$, cf. proof of Corollary 1. Moreover, the map $(\beta, \mu) \mapsto \nabla_\vartheta \mathcal{L}_{\mathrm{risk}}(\vartheta, \beta, \mu)$ is locally Lipschitz continuous on $\widetilde{\boldsymbol{\Theta}}$, as

$$\begin{aligned}
\left\|\nabla_\vartheta \mathcal{L}_{\mathrm{risk}}(\vartheta, \beta, \mu) - \nabla_\vartheta \mathcal{L}_{\mathrm{risk}}(\vartheta, \beta', \mu')\right\|_2 &= \left\|2(\mathbf{I}_d + \mu\mu^\top)(\vartheta - \beta) - 2(\mathbf{I}_d + \mu'\mu'^\top)(\vartheta - \beta')\right\|_2 \\
&= 2\left\|(\mu\mu^\top - \mu'\mu'^\top)(\vartheta - \beta) + (\mathbf{I}_d + \mu'\mu'^\top)(\beta' - \beta)\right\|_2 \\
&\leqslant 2(\left\|\mu\mu^\top - \mu'\mu'^\top\right\|_2 \|\vartheta - \beta\|_2 + \left\|\mathbf{I}_d + \mu'\mu'^\top\right\|_2 \|\beta' - \beta\|_2) \\
&\leqslant \underbrace{(16 + 8\|\vartheta\|_2)}_{=:\zeta_{\mathrm{risk}}(\vartheta)} \left(\|\mu - \mu'\|_2 + \|\beta - \beta'\|_2\right)
\end{aligned}$$

where we used $\|\vartheta - \beta\|_2 \leqslant 2 + \|\vartheta\|_2$ and $\left\|\mathbf{I}_d + \mu'\mu'^\top\right\|_2 \leqslant 3$, as well as $\left\|\mu\mu^\top - \mu'\mu'^\top\right\|_2 \leqslant \|\mu - \mu'\|_2 (\|\mu\|_2 + \|\mu'\|_2) \leqslant 4\|\mu - \mu'\|_2$. Therefore, the risk objective satisfies the conditions from Assumption 2 and, in particular, is strongly convex with $\lambda_{\mathrm{risk}} > 0$ (by assumption), so that the fairness objective does not need to be strongly convex.

Now, notice that $\vartheta \mapsto \mathcal{L}_{\mathrm{fair}}(\vartheta, \mu)$ is clearly convex and twice continuously differentiable, however, it is not strongly convex. Moreover, the map $\mu \mapsto \nabla_\vartheta \mathcal{L}_{\mathrm{fair}}(\vartheta, \mu) = 2\langle \mu, \vartheta\rangle \mu$ is locally Lipschitz, since

$$\left\|\nabla_\vartheta \mathcal{L}_{\mathrm{fair}}(\vartheta, \mu) - \nabla_\vartheta \mathcal{L}_{\mathrm{fair}}(\vartheta, \mu')\right\|_2 = 2\left\|\langle \mu, \vartheta\rangle \mu - \langle \mu', \vartheta\rangle \mu'\right\|_2 \leqslant \underbrace{8\|\vartheta\|_2}_{\zeta_{\mathrm{fair}}(\vartheta)} \|\mu - \mu'\|_2.$$

Therefore, Assumption 2 and 3 are satisfied.

It remains to show that $\widehat{\boldsymbol{\theta}} = ((\widehat{\beta}, \widehat{\mu}), \widehat{\mu}) \in \widetilde{\boldsymbol{\Theta}}$ with high probability. Recall that the two-stage estimator uses the estimator $\widehat{\boldsymbol{\theta}}$ chosen like this: Denoting $\mathbf{X}$ as the design matrix and y as the vector of noisy responses, $\widehat{\beta}$ is defined as the LASSO

$$\widehat{\vartheta} \in \operatorname*{arg\,min}_{\vartheta \in \mathbb{R}^d} \|\mathbf{X}\vartheta - y\|_2^2 + 136\sigma\sqrt{\frac{\log d}{n}} \|\vartheta\|_1.$$

To estimate μ the two-stage estimator uses the standard mean estimation

$$\widehat{\mu} := \frac{1}{n+N} \sum_{i=1}^{n+N} A_i X_i.$$

Now, define the event

$$\mathcal{E}_1 = \left\{ \|\widehat{\mu} - \mu\|_2 \leqslant c_1 \sqrt{\frac{d}{n+N}} \right\}$$

which holds with probability at least $1 - c_2 \exp(-c_3 d)$ by concentration of a χ^2_d-distribution with d degrees of freedom. Furthermore, define the event

$$\mathcal{E}_2 = \left\{ \left\|\widehat{\beta} - \beta\right\|_2 \leqslant c_4 \sigma \sqrt{\frac{s \log d}{n}} \right\}$$

which holds with probability at least $1 - c_5 d^{-3}$ by the derivations in the proof of Corollary 1. Hence, if $n + N \gtrsim d$, on $\mathcal{E}_1 \cap \mathcal{E}_2$ we have that $\widehat{\boldsymbol{\theta}} = ((\widehat{\beta}, \widehat{\mu}), \widehat{\mu}) \in \widetilde{\boldsymbol{\Theta}}$.

Consequently, on $\mathcal{E}_1 \cap \mathcal{E}_2$, noting that μ is shared across objectives (i.e., we can combine the two error terms), the upper bound in Theorem 1 applies and since $\lambda_{\text{risk}} + \lambda_{\text{fair}} = 1$, we obtain

$$\begin{aligned}
\left\|\widehat{\vartheta}^{\text{ts}}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2 &\leqslant \frac{16 + 8\|\vartheta_{\boldsymbol{\lambda}}\|_2}{\lambda_{\text{risk}}} \left(\lambda_{\text{risk}} \left\|\widehat{\beta} - \beta\right\|_2 + \|\widehat{\mu} - \mu\|_2 \right) \\
&\lesssim (1 + \|\vartheta_{\boldsymbol{\lambda}}\|_2) \left(\sigma \sqrt{\frac{s \log d}{n}} + \frac{1}{\lambda_{\text{risk}}} \sqrt{\frac{d}{n+N}} \right),
\end{aligned}$$

which holds with probability at least $1 - c_2 \exp(-c_3 d) - c_5 d^{-3} \geqslant 1 - c d^{-3}$ by the union bound. Finally, we conclude the proof since by (A.5) we have

$$\|\vartheta_{\boldsymbol{\lambda}}\|_2 = \left\| \left(\lambda_{\text{fair}} \mu\mu^\top + \lambda_{\text{risk}}(\mathbf{I}_d + \mu\mu^\top)\right)^{-1} \lambda_{\text{risk}}(\mathbf{I}_d + \mu\mu^\top)\beta \right\|_2 \leqslant 1.$$

□

D.9 Proof of Proposition 5

Proposition 5 (Necessity of unlabeled data)**.** *Consider the special case of the statistical model $\boldsymbol{\mathcal{P}}$ in Example 1, by restricting $\beta_1 = -\beta_2 = \beta$ for some β with $\|\beta\|_2 = 1$ that is fixed and known. Further, for $k \in \{1, 2\}$, let $\mathbb{P}^k$ be such that $\mathbb{P}^k_X = \mathcal{N}(0, \Sigma_k)$ with symmetric, unknown covariance matrices Σ_k satisfying $\frac{1}{2}\mathbf{I}_d \preceq \Sigma_k \preceq \frac{3}{2}\mathbf{I}_d$. Consider $\boldsymbol{\lambda} = (1/2, 1/2)$ and we observe $n_k = n$ labeled and $N_k = N$ unlabeled datapoints each. Then, if $d \geqslant 3$ and $\sqrt{d/(512(n+N))} \leqslant 1/(4e)$, it holds that*

$$\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) \gtrsim \sqrt{\frac{d}{n+N}}.$$

Proof of Proposition 5. The proof follows from Fano's method (Fano and Hawkins, 1961; Yang and Barron, 1999; Yu, 1997) using local packings, applied to our setting. Specifically, similar to the proof of Proposition 2, we will use that for any vector v in $C\delta B_2^d$ (for some constant $C > 1$), we can find covariance matrices Σ_k so that the corresponding solution $\vartheta_{\boldsymbol{\lambda}}$ equals v. We then consider a δ-packing of $C\delta B_2^d$ and apply Fano's method: To that end, we bound the mutual information between Gaussian distributions with the covariance matrices that correspond to the packing. Some calculations yield the lower bound.

We write N instead of $n + N$ for brevity throughout the proof. For $t \in (0, 1/4]$ recall the definition of the scaled ℓ_2-ball $tB_2^d = \left\{ v \in \mathbb{R}^d \,\middle|\, \|v\|_2 \leqslant t \right\}$. Then for all $v, v' \in tB_2^d$ it holds $-t \leqslant \langle v, \beta \rangle \leqslant t$ and $\|v - v'\|_2 \leqslant 2t$. Let $v_1 \ldots, v_M$ be a δ-packing of tB_2^d in ℓ_2-norm. By standard arguments (Wainwright, 2019, Example 5.8), there exists such a packing with $\log M \geqslant d \log(t/\delta)$.

For any vector $v \in tB_2^d$, we now define the matrix

$$A(v) = v\beta^\top + \beta v^\top - \langle v, \beta \rangle \beta\beta^\top$$

and notice that $A(v)\beta = v + \langle v, \beta\rangle \beta - \langle v, \beta\rangle \beta = v$. Based on $A(v)$ we define the pairs of covariance matrices $\mathbf{\Sigma}^j = (\Sigma_1^j, \Sigma_2^j)$ with

$$\Sigma_1^j = \mathbf{I}_d + A(v_j) \qquad \text{and} \qquad \Sigma_2^j = \mathbf{I}_d - A(v_j).$$

For $\mathbf{\Sigma} = (\Sigma_1, \Sigma_2)$, recall from (A.5) that the multi-objective solution is given by

$$\vartheta_{\boldsymbol{\lambda}}(\mathbf{\Sigma}) = (\Sigma_1 + \Sigma_2)^{-1} (\Sigma_1 \beta - \Sigma_2 \beta).$$

and hence we immediately see that

$$\begin{aligned}
\vartheta_{\boldsymbol{\lambda}}(\mathbf{\Sigma}^j) &= (\Sigma_1^j + \Sigma_2^j)^{-1}(\Sigma_1^j \beta - \Sigma_2^j \beta) \\
&= (\mathbf{I}_d + A(v_j) + \mathbf{I}_d - A(v_j))^{-1}((\mathbf{I}_d + A(v_j))\beta - (\mathbf{I}_d - A(v_j))\beta) \\
&= (2\mathbf{I}_d)^{-1}(2v_j) \\
&= v_j
\end{aligned}$$

as well as the eigenvalues of Σ_1^j, Σ_2^j lying within $1 + 2\langle \beta, v_j\rangle \leqslant 1 + 2t \leqslant 3/2$ and $1 - \langle v_j, \beta\rangle \geqslant 1 - 2t \geqslant 1/2$. Hence, for every $i \in [M]$, there exists a pair of distributions in $\boldsymbol{\mathcal{P}}$ so that $\vartheta_{\boldsymbol{\lambda}} = v_i$.

We now apply a version of Fano's method (Wainwright, 2019, Section 15.3) that is based on local packing. Define $J \sim U([M])$ and $(Z|J = j) \sim \mathcal{N}(0, \Sigma_1^j) \times \mathcal{N}(0, \Sigma_2^j)$. By Wainwright (2019, Proposition 15.12), we have for any $\delta > 0$ that

$$\mathfrak{M}_{\boldsymbol{\lambda}}(\boldsymbol{\mathcal{P}}) = \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \boldsymbol{\mathcal{P}}} \mathbb{E}_{\mathbb{P}} \left[\left\| \widehat{\theta} - \vartheta_{\boldsymbol{\lambda}} \right\|_2 \right] \geqslant \delta \left(1 - \frac{I(Z; J) + \log 2}{\log M} \right)$$

and $I(Z; J)$ is the mutual information between Z and J. We bound $I(Z; J)$ using Wainwright (2019, Lemma 15.17 and Equation (15.45)):

$$I(Z; J) \leqslant \frac{N}{2} \left(\log\det \operatorname{cov}(Z) - \frac{1}{M} \sum_{i=1}^{M} \log\det(\Sigma_1^i \otimes \Sigma_2^i) \right).$$

Note that unconditionally, $Z \sim \mathcal{N}(0, \frac{1}{M} \sum_{i=1}^{M} \Sigma_1^i \otimes \Sigma_2^i)$, where $A \otimes B$ denotes the matrix with the blocks A and B on its diagonal. Hence,

$$\begin{aligned}
&\log\det \operatorname{cov}(Z) - \frac{1}{M} \sum_{i=1}^{M} \log\det(\Sigma_1^i \otimes \Sigma_2^i) \\
&= \log\det \left(\frac{1}{M} \sum_{i=1}^{M} \Sigma_1^i \otimes \Sigma_2^i \right) - \frac{1}{M} \sum_{i=1}^{M} \log\det(\Sigma_1^i \otimes \Sigma_2^i) \\
&= \log\det \left(\frac{1}{M} \sum_{i=1}^{M} \Sigma_1^i \right) + \log\det \left(\frac{1}{M} \sum_{i=1}^{M} \Sigma_2^i \right) - \frac{1}{M} \sum_{i=1}^{M} \log\det(\Sigma_1^i) - \frac{1}{M} \sum_{i=1}^{M} \log\det(\Sigma_2^i)
\end{aligned}$$

We now apply the following lemma, which is proved after concluding the main proof.

Lemma D.2. *For $\|\beta\|_2 = 1$, $v \in tB_2^d$ with $t \in (0, 1/4]$, define the matrix*

$$A(v) = v\beta^\top + \beta v^\top - \langle v, \beta\rangle \beta\beta^\top$$

and define for $v_i \in tB_2^d$, $i \in \{1, \dots, M\}$ the matrices $\Sigma_1^i = \mathbf{I}_d + A(v_i)$ and $\Sigma_2^i = \mathbf{I}_d - A(v_i)$. Then, for $j \in \{1, 2\}$, it holds that

$$0 \leqslant \log\det \left(\frac{1}{M} \sum_{i=1}^{M} \Sigma_j^i \right) - \frac{1}{M} \sum_{i=1}^{M} \log\det \left(\Sigma_j^i \right) \leqslant 32t^2.$$

Therefore, we know that $I(Z, J) \leqslant 32Nt^2$. Combining this with $\log M \geqslant d \log(t/\delta)$ we have that

$$\begin{aligned}
\inf_{\widehat{\vartheta}_{\boldsymbol{\lambda}}} \sup_{\mathbb{P} \in \boldsymbol{\mathcal{P}}} \mathbb{E}_{\mathbb{P}} \left[\left\| \widehat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}} \right\|_2 \right] &\geqslant \delta \left(1 - \frac{I(Z; J) + \log 2}{\log M} \right) \\
&\geqslant \delta \left(1 - \frac{32Nt^2 + \log 2}{d \log(t/\delta)} \right)
\end{aligned}$$

Choosing $\delta = \sqrt{\frac{d}{512N}}$ and $t = e\delta$ (which, by assumption, is smaller than $1/4$), since $d \geqslant 3$ we have

$$\begin{aligned}\inf_{\hat{\vartheta}_{\boldsymbol{\lambda}}} \sup_{\mathbb{P}\in\boldsymbol{\mathcal{P}}} \mathbb{E}_{\mathbb{P}}\left[\left\|\hat{\vartheta}_{\boldsymbol{\lambda}} - \vartheta_{\boldsymbol{\lambda}}\right\|_2\right] &\geqslant \delta\left(1 - \frac{32Ne^2\frac{d}{512N}}{d} - \frac{\log 2}{d}\right)\\ &\geqslant \delta\left(1 - \frac{e^2}{16} - \frac{1}{4}\right)\\ &\geqslant \frac{1}{4}\delta \asymp \sqrt{\frac{d}{N}}\end{aligned}$$

which concludes the proof by recalling that we replaced $N + n$ by just N. □

Proof of Lemma D.2. Denote the function $f(A) = \log\det A$.

We begin by noting that f is concave on the positive semidefinite cone (Rivin, 2002), directly implying the lower bound, as it is merely Jensen's inequality.

We now show the upper bound. Since f is twice differentiable on the set of positive definite matrices, the second-order Taylor expansion in integral form yields that for any two such matrices A and B,

$$f(A) = f(B) + \langle \nabla f(B), A - B\rangle + \int_0^1 (1-s)\nabla^2 f(B + s(A-B))[A-B, A-B]ds.$$

A computation shows that the gradient and Hessian of f are given by

$$\nabla f(A) = A^{-1} \qquad \text{and} \qquad \nabla^2 f(A)[H,H] = -\operatorname{tr}\left(A^{-1}HA^{-1}H\right).$$

If we take $A = \Sigma_1^i$, $B = \bar{\Sigma}_1$ and $H = \Sigma_1^i - \bar{\Sigma}_1$, we get that

$$\begin{aligned}f(\Sigma_1^i) = f(\bar{\Sigma}_1) &+ \left\langle \nabla f(\bar{\Sigma}_1), \Sigma_1^i - \bar{\Sigma}_1\right\rangle\\ &+ \int_0^1 (1-s)\operatorname{tr}\left((\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}(\Sigma_1^i - \bar{\Sigma}_1)(\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}(\Sigma_1^i - \bar{\Sigma}_1)\right) ds\end{aligned} \tag{20}$$

Notice how $\frac{1}{M}\sum_{i=1}^M \left\langle \nabla f(\bar{\Sigma}_1), \Sigma_1^i - \bar{\Sigma}_1\right\rangle = \left\langle \nabla f(\bar{\Sigma}_1), \bar{\Sigma}_1 - \bar{\Sigma}_1\right\rangle = 0$. Hence, averaging over i on both sides of Equation (20), we get that

$$\frac{1}{M}\sum_{i=1}^M f(\Sigma_1^i) = f(\bar{\Sigma}_1) - \frac{1}{M}\sum_{i=1}^M \int_0^1 (1-s)\operatorname{tr}\left((\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}(\Sigma_1^i - \bar{\Sigma}_1)(\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}(\Sigma_1^i - \bar{\Sigma}_1)\right) ds \tag{21}$$

Therefore, to prove our bound, we need to bound the integral. To that end, we bound the trace as

$$\operatorname{tr}\left((\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}(\Sigma_1^i - \bar{\Sigma}_1)(\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}(\Sigma_1^i - \bar{\Sigma}_1)\right) \leqslant \left\|(\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}\right\|_F^2 \left\|\Sigma_1^i - \bar{\Sigma}_1\right\|_F^2.$$

We bound this Frobenius norm for our specific choice of covariance matrices. Writing $v = \langle v, \beta\rangle\beta + v_\perp$ with $\langle v_\perp, \beta\rangle = 0$, one easily verifies that

$$A(v) = v\beta^\top + \beta v^\top - \langle v,\beta\rangle\beta\beta^\top = (\langle v,\beta\rangle\beta + v_\perp)\beta^\top + \beta v^\top - \langle v,\beta\rangle\beta\beta^\top = v_\perp\beta^\top + \beta v^\top$$

Thus, we can bound for $v \in B_2^d$

$$\|A(v)\|_F \leqslant \left\|v_\perp\beta^\top\right\|_F + \left\|\beta v^\top\right\|_F = \|v_\perp\|_2\|\beta\|_2 + \|\beta\|_2\|v\|_2 \leqslant 2\|v\|_2 \leqslant 2t,$$

and hence also $1 - 2t \leqslant \left\|\Sigma_1^i\right\|_F \leqslant 1 + 2t$ and $1 - 2t \leqslant \left\|\Sigma_2^i\right\|_F \leqslant 1 + 2t$. Now, for any two vectors $v, w \in tB_2^d$, by linearity

$$\|A(v) - A(w)\|_F = \|A(v-w)\|_F \leqslant 2\|v - w\| \leqslant 4t.$$

If we define $\bar{v} = \frac{1}{M}\sum_{i=1}^M v_i \in tB_2^d$ and set $\bar{\Sigma}_1 = \mathbf{I}_d + A(\bar{v})$, then for each i,

$$\left\|\Sigma_1^i - \bar{\Sigma}_1\right\|_F = \|A(v_i) - A(\bar{v})\|_F \leqslant 2\|v_i - \bar{v}\| \leqslant 4t,$$

and analogously for $\bar{\Sigma}_2 = \mathbf{I}_d - A(\bar{v})$. Thus, we can bound the Hessian term, using $t \leqslant 1/4$, as

$$\left\|(\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}\right\|^2 \left\|\Sigma_1^i - \bar{\Sigma}_1\right\|_F^2 \leqslant \frac{1}{(1-2t)^2} \left\|\Sigma_1^i - \bar{\Sigma}_1\right\|_F^2 \leqslant 4 \left\|\Sigma_1^i - \bar{\Sigma}_1\right\|_F^2 \leqslant 64t^2.$$

Hence, finally, we get from Equation (21) that

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^M f(\Sigma_1^i) &= f(\bar{\Sigma}_1) - \frac{1}{M}\sum_{i=1}^M \int_0^1 (1-s)\,\mathrm{tr}\left((\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}(\Sigma_1^i - \bar{\Sigma}_1)(\bar{\Sigma}_1 + s(\Sigma_1^i - \bar{\Sigma}_1))^{-1}(\Sigma_1^i - \bar{\Sigma}_1)\right) ds \\
&\geqslant f(\bar{\Sigma}_1) - \frac{64t^2}{M}\sum_{i=1}^M \int_0^1 (1-s) ds \\
&= f(\bar{\Sigma}_1) - 32t^2
\end{aligned}$$

or, in other words,

$$f\left(\frac{1}{M}\sum_{i=1}^M \Sigma_1^i\right) - \frac{1}{M}\sum_{i=1}^M f(\Sigma_1^i) \leqslant 32t^2$$

which concludes the proof. □