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

Anonymous Author
Anonymous Institution

Abstract

Modern machine learning methods often have to rely on high-dimensional data that is expensive to label, while unlabeled data is abundant. When the data exhibits low-dimensional structure such as sparsity, conventional regularization techniques are known to improve generalization for a single objective (e.g., prediction risk). 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 the first MOL estimator that provably yields improved performance in the presence of sparsity and unlabeled data. We demonstrate its effectiveness experimentally for multi-distribution learning and fairness-risk trade-offs.

1 INTRODUCTION

As machine learning models are employed more and more broadly, they are expected to be *trustworthy* in numerous ways: Besides being accurate, they should also be robust against (adversarial) distribution shifts (Szegedy et al., 2014; Yin et al., 2019), fairness-aware (Pedreshi et al., 2008; Hardt et al., 2016), private (Dwork, 2006; Vaidya, 2009), interpretable (Marcinkevičs and Vogt, 2023) and, more recently, aligned with human values (Ji et al., 2023), to name just a few. It is well understood that in many settings, achieving all of these objectives simultaneously can be inherently impossible and that a trade-off between them is unavoidable (Menon and Williamson, 2018; Wang et al., 2024; Zhang et al., 2019; Raghunathan et al., 2020; Cummings et al., 2019; Sanyal et al., 2022; Guo et al., 2024).

In the presence of such inherent trade-offs, we are usually interested in learning models that lie on the *Pareto front*, where improving in one objective must come at the expense of another (e.g., Ehrgott (2005)). In the optimization literature, it is well-known (Ehrgott, 2005; López et al., 2011) that under some regularity conditions, 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 λ from the simplex.

In the context of machine learning, however, the individual objectives of interest 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. To estimate the population-level Pareto front, a common approach is to compute the Pareto front of empirical plug-in estimates $\widehat{\mathcal{L}}_1, \dots, \widehat{\mathcal{L}}_K$, for example, by minimizing the empirical scalarized objective $\sum_{k=1}^{K} \lambda_k \widehat{\mathcal{L}}_k$ or $\max_{k \in [K]} \lambda_k \widehat{\mathcal{L}}_k$, thereby reducing the problem to a single-objective problem (Jin and Sendhoff, 2008; Lin et al., 2019; Hu et al., 2024). This approach can indeed perform well in regimes with sufficiently many labeled training samples, where $\widehat{\mathcal{L}} \approx \mathcal{L}$.

However, in modern overparameterized regimes with relatively little labeled data but large amounts of unlabeled data available, it is crucial to leverage low-dimensional structure. In single-objective learning, this has been established in a long line of work on estimators with the right inductive bias, e.g., for recovering sparse ground truths (Bühlmann and van de Geer, 2011; Wainwright, 2019). On the other hand, in the context of multi-objective learning, there are so far no theoretical guarantees on how regularization can counteract the curse of dimensionality in the presence of sparsity or other low-dimensional structure.

Hence, in this paper, we take a step towards addressing the following question:

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

Preliminary work. Under review by AISTATS 2025. Do not distribute.

Our main contributions are outlined below.

- We introduce a new MOL estimator that can successfully take advantage of the low-dimensional structure of the objective-specific minimizers to estimate the entire Pareto front, while leveraging unlabeled data (Section 3).
- We prove upper and lower bounds for the estimation error of the Pareto front which reveal interesting geometric peculiarities inherent to MOL problems (Section 4).
- We demonstrate the effectiveness of our method in several applications, validated by experiments on synthetic and semi-synthetic data (Section 5).

2 SETTING AND NOTATION

We begin by introducing the multi-objective learning (MOL) problem using multi-objective optimization.

2.1 Multi-objective optimization

In the context of multi-objective learning, we are ultimately interested in solving a *multi-objective optimization* (MOO) problem (Ehrgott, 2005; López et al., 2011), where the objectives are defined via distributions. Let $\mathcal{F}$ be our hypothesis space of functions $f_\vartheta : \mathcal{X} \to \mathcal{Y}$ parameterized by $\vartheta \in \mathbb{R}^m$. Further, we write $\mathcal{L} : \mathbb{R}^m \times \mathcal{P} \to \mathbb{R}^K$ for a non-negative vector-valued function that consists of K objectives

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

where the k-th objective depends on a distribution $\mathbb{P}^k$ that is defined on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{P}$ denotes all K-tuples $\mathbb{P} = (\mathbb{P}^1, \ldots, \mathbb{P}^K)$. For each joint distribution $\mathbb{P}^k$ of the random vector (X, Y), we denote the marginal distribution of X as $\mathbb{P}^k_X$. Finally, we denote by θ_k the minimizers of the individual objectives, that is,

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

For the sake of a simpler exposition, we assume uniqueness of the minimizers in this section.

Our formulation includes the case where we have different losses on the same distribution (e.g., Duh et al. (2012)), or the same loss on different distributions (i.e., multi-distribution learning (Haghtalab et al., 2022)). For concreteness, we now present a simple example for the latter that reappears in later sections of the paper, among others.

Example 1 (Multiple linear regression)**.** *Consider $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ and the distributions $\mathbb{P}^k$ induced by the model*

$$Y = \langle X, \theta_k \rangle + \xi \quad \textit{with} \quad \xi \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

where $\theta_k \in \mathbb{R}^m$ are (differing) s-sparse ground truths $\|\theta_k\|_0 \leqslant s$ and X are B^2-sub-Gaussian covariate vectors with non-degenerate covariance matrices $\Sigma_k = \mathbb{E}_{X \sim \mathbb{P}^k_X}\left[XX^\top\right]$. We are interested in the prediction risks using squared-loss for each k, defined as

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

In view of multiple objectives, one goal could be to find the minimizer of $\mathcal{L}$ with respect to $\vartheta \in \mathbb{R}^m$ in each coordinate simultaneously. However, in many cases, this is not possible because the sets of minimizers of the objectives do not intersect (e.g., Martinez et al. (2020); Yaghini et al. (2023)). In that case, MOO aims to either find the set of all *Pareto-optimal* solutions, or a set of points that have a small *scalarized* loss, both 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 $k \in [K]$*

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

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

In Figure 1, we illustrate the set of Pareto-optimal points and the Pareto front for a toy two-objective problem. On the left, the solid red line depicts the set of Pareto-optimal points in the parameter space $\mathbb{R}^m$ with "end-points" θ_k, while on the right, it traces the Pareto front in the two-dimensional space of objective function values. The red shaded region on the right corresponds to the set of all achievable value pairs of the two objective functions.

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

Equation (1) already introduced two important scalarizations, known as linear and Chebyshev scalarization, respectively. We denote by ϑ_λ the minimizers of a scalarization, i.e.,

$$\vartheta_\lambda \in \operatorname*{arg\,min}_{\vartheta \in \mathbb{R}^m} (s_\lambda \circ \mathcal{L})(\vartheta, \mathbb{P}). \tag{4}$$

It is well-known (López et al., 2011) that a solution to (4) is Pareto-optimal for linear or Chebyshev scalarization, if the solution is unique or $\lambda_k > 0$ for all $k \in [K]$. Hence, each minimizer ϑ_λ lies in the set of Pareto-optimal points, and $\mathcal{L}(\vartheta_\lambda, \mathbb{P})$ lies on the Pareto front, as depicted in Figure 1. Further, when (4) has a unique solution for all λ with $\lambda_k = 0$ for some k, Chebyshev scalarization recovers all Pareto-optimal points. When additionally strict convexity holds for each objective, linear scalarization also parameterizes the entire set of Pareto-optimal points, that is each λ corresponds to one Pareto-optimal point and vice versa (Miettinen, 1998; Hillermeier, 2001; Roy et al., 2023). Note that then, in the special case $\lambda_k = 1$, we obtain $\vartheta_\lambda = \theta_k$. In the rest of the paper, when we use ϑ_λ to denote a minimizer of an unspecified scalarization, we implicitly assume that the scalarization parameterizes the Pareto-optimal points.

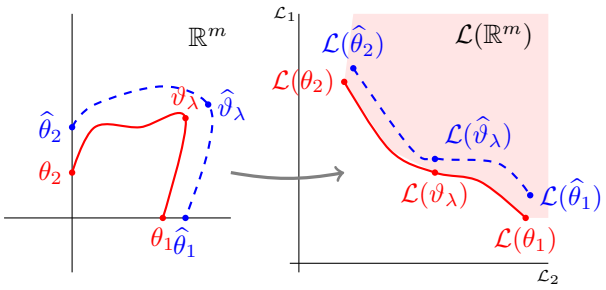


Figure 1: We use the shorthand $\mathcal{L}(\vartheta) = \mathcal{L}(\vartheta, \mathbb{P})$. The parameter space $\mathbb{R}^m$ (left) parameterizes the hypothesis set $\mathcal{F}$ and contains the set of the population Pareto-optimal points $\vartheta_\lambda, \lambda \in \Delta^K$ (red line), and the set of the empirical estimators (dashed blue line). ϑ_λ can be non-sparse even when θ_1, θ_2 are sparse. In the right figure we depict the region of all values that can be obtained by $\mathcal{L}(\vartheta, \mathbb{P})$ for some ϑ (red shaded area), the population Pareto front (red line) and estimated Pareto front (dashed blue line).

2.2 Multi-objective learning

In practice, we do not observe the distributions $\mathbb{P}^k$ but finite samples from it. In particular, we assume a semi-supervised setting, where we observe a set 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 contained in the dataset $\{X_i^k\}_{i=n_k+1}^{N_k}$ from each marginal distribution $\mathbb{P}_X^k$. We let $\mathcal{D}$ denote the entire dataset of labeled and unlabeled datapoints. Based on the labeled set of samples, we can estimate $\mathbb{P}^k$ using the empirical measure $\widehat{\mathbb{P}}^k = n_k^{-1} \sum_{i=1}^{n_k} \mathbb{1}_{(X_i^k, Y_i^k)}$ and define $\widehat{\mathbb{P}} = (\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K)$ accordingly. Moreover, we can estimate the marginal $\mathbb{P}_X^k$ via the empirical marginal using the entire dataset $\widehat{\mathbb{P}}_X^k = (n_k + N_k)^{-1} \sum_{i=1}^{n_k+N_k} \mathbb{1}_{X_i^k}$.

The aim of *multi-objective learning* (MOL) is to use the data $\mathcal{D}$ to estimate the set of Pareto optimal points $\{\vartheta_\lambda | \lambda \in \Delta^K\} \subset \mathbb{R}^m$ with a set of estimators $\{\widehat{\vartheta}_\lambda | \lambda \in \Delta^K\}$ with small squared estimation errors $\|\widehat{\vartheta}_\lambda - \vartheta_\lambda\|_2^2$. Note that here, we denote by $\widehat{\vartheta}_\lambda$ the estimator for the Pareto-optimal point ϑ_λ parameterized by λ. We compare our estimation procedures with the information-theoretically optimal estimation error (i.e., minimax error), defined as

$$\mathcal{M}_\lambda(\mathcal{P}) = \inf_{\widehat{\vartheta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\left\|\widehat{\vartheta}(\mathcal{D}) - \vartheta_\lambda\right\|_2^2\right], \tag{5}$$

where the infimum is taken over all estimators $\widehat{\vartheta}(\mathcal{D}) \in \mathbb{R}^m$, and the expectation is taken over draws of the dataset.

Under weak assumptions such as smoothness, a bound on the estimator error $\|\widehat{\vartheta}_\lambda - \vartheta_\lambda\|_2^2$ implies a bounds on the excess in each individual objective, $\mathcal{L}_k(\widehat{\vartheta}_\lambda, \mathbb{P}) - \mathcal{L}_k(\vartheta_\lambda, \mathbb{P})$, and the excess scalarized objective

$$\mathcal{E}_\lambda(\widehat{\vartheta}_\lambda) := (s_\lambda \circ \mathcal{L})(\widehat{\vartheta}_\lambda, \mathbb{P}) - \min_{\vartheta \in \mathbb{R}^m} (s_\lambda \circ \mathcal{L})(\vartheta, \mathbb{P}). \tag{6}$$

3 TWO ESTIMATORS

A commonly used approach in the multi-objective learning literature is to consider a *plug-in* estimator that is the minimizer of the objective (4) where $\mathbb{P}$ is replaced by $\widehat{\mathbb{P}}$ (Jin, 2006). However, the high-dimensional regime requires introducing regularization to this approach.

3.1 A directly regularized estimator

Since the scalarized objective can be viewed as a generic scalar loss, we may be tempted to follow common practice for high-dimensional (single-objective) learning and add a penalty term $\rho : \mathbb{R}^m \to \mathbb{R}$ to the empirical objective and solve

$$\min_{\vartheta \in \mathbb{R}^m} (s_\lambda \circ \mathcal{L})(\vartheta, \widehat{\mathbb{P}})) + \rho(\vartheta) \tag{7}$$

or constrain the parameter space to a subset of $\mathbb{R}^m$.[1] Such approaches are often used if the inductive bias on the multi-objective solution from (4) is the same across the Pareto front, or there is only one objective besides the penalty (Jin and Sendhoff, 2008; Cortes et al., 2020; Mierswa, 2007; Bieker et al., 2022; Hotegni et al., 2024).

The rationale behind regularization in high-dimensional learning is the assumption that the minimizer exhibits a certain structure that is captured by the penalty ρ (e.g., sparsity by the ℓ_1-norm). Hence, estimators of the form (7) could work well if the corresponding Pareto-optimal points adhere to this structure. However, as we aim to recover the entire Pareto-optimal set, asserting that the structure is present across the entire set of Pareto-optimal solutions might be rather unrealistic depending on the problem at hand. For example, it may be natural to assume that the minimizers θ_k of the individual objectives are sparse, but generally, large parts of the Pareto-optimal 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 (7) for all λ, except when $\lambda_k = 1$ and we are estimating θ_k.

3.2 A new two-stage estimator

We now introduce our two-stage estimator that is able to leverage the structure of individual minimizers. For this purpose, we assume that each objective depends on the distributions $\mathbb{P}^k$ via two parameter vectors.

Assumption 1. *Objective $\mathcal{L}_k(\cdot, \mathbb{P}^k)$ depends on $\mathbb{P}^k$ only through $\theta_k \in \mathbb{R}^m$ as defined in Equation (2) and a parameter $\omega_k \in \mathbb{R}^M$ that only depends on the marginal $\mathbb{P}_X^k$.*

In Example 1, the parameter ω_k corresponds to a vectorization of the covariance matrices, which only depends on

[1] 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 (7)

the marginal. This means we can estimate ω_k using unlabeled data. We argue that such a re-parameterization holds in many cases, which we also demonstrate in a number of applications in Sections 4 and 5.

Denoting $\theta = (\theta_1 \dots \theta_K) \in \mathbb{R}^{m \times K}$ and $\omega = (\omega_1 \dots \omega_K) \in \mathbb{R}^{M \times K}$, we abuse notation and write $\mathcal{L}_k(\vartheta, \theta_k, \omega_k)$ for $\mathcal{L}_k(\vartheta, \mathbb{P}^k)$ and correspondingly the vector-valued function

$$\mathcal{L}(\vartheta, \theta, \omega) := (\mathcal{L}_1(\vartheta, \theta_1, \omega_1), \dots, \mathcal{L}_K(\vartheta, \theta_K, \omega_K)).$$

This lets us define the *two-stage regularized multi-objective estimator* $\widehat{\vartheta}^{\text{ts}}_\lambda$ as the solution to a two-stage procedure.

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

Stage 1: Estimation. *Use the data $\mathcal{D}$ to estimate the parameter matrices $\widehat{\theta} = (\widehat{\theta}_1 \dots \widehat{\theta}_K), \widehat{\omega} = (\widehat{\omega}_1 \dots \widehat{\omega}_K)$, e.g., with*

$$\widehat{\theta}_k \in \arg\min_{\vartheta \in \mathbb{R}^m} \mathcal{L}_k(\vartheta, \widehat{\mathbb{P}}^k) + \rho_k(\vartheta).$$

Stage 2: Optimization. *Minimize the scalarized objective*

$$\widehat{\vartheta}^{\text{ts}}_\lambda \in \arg\min_{\vartheta \in \mathbb{R}^m} (s_\lambda \circ \mathcal{L})(\vartheta, \widehat{\theta}, \widehat{\omega}). \tag{8}$$

Note that when $\lambda_k = 1$, we recover the directly regularized estimator from (7). The results in Section 4 shed light on the relevance and unconventional benefits of using unlabeled data to estimate ω for MOL.

The estimator is similar to probabilistic modeling pipelines (Ng and Jordan, 2001) in that it learns all (necessary) parameters first, and then the estimation of *any* Pareto-optimal point ϑ_λ can reap the benefits of the efficiency of the parameter estimation of θ, ω. But of course, for $\widehat{\vartheta}^{\text{ts}}_\lambda$ to be an efficient estimator, the estimators for θ, ω have to be chosen well. In particular, the penalty ρ_k should induce the correct inductive bias towards the population minimizers. For instance, if θ_k is sparse, one could choose ρ_k to be the ℓ_1-norm (Tibshirani, 1996).

We now show how the two-stage estimator with the ℓ_1-norm penalty can outperform all directly regularized estimators of the form in (7) in a concrete example.

3.3 A simple comparison

Consider the fixed-design linear regression setting, where we observe matrices $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{n \times d}$ and noisy responses

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

With slight abuse of notation, we define the population and empirical objectives for $k = 1, 2$ as

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

For any constant $c > 0$, let $\Gamma(c)$ denote the set of all matrices $\mathbf{X} \in \mathbb{R}^{n \times d}$ such that $\frac{1}{n}\mathbf{X}^\top \mathbf{X}$ has eigenvalues lower-bounded by $c/2$ and upper bounded by $2c$, and $\Theta \subset \mathbb{R}^m$ the set of 1-sparse vectors θ. The following proposition shows that any estimator in the form of (7) generally *cannot* use the sparsity of θ_k to mitigate the curse of dimensionality.

Proposition 1 (Insufficiency of direct regularization)**.** *Suppose that $c \geqslant 2\sigma^2$, $\lambda_1 = \lambda_2 = 1/2$, $n \geqslant d$ and we use linear scalarization. Then, for any regularizer ρ, an estimator $\widehat{\vartheta}$ in the form of (7) satisfies*

$$\sup_{\substack{\theta_1, \theta_2 \in \Theta \\ \mathbf{X}_1, \mathbf{X}_2 \in \Gamma(c)}} \mathbb{E}\left[\left\| \widehat{\vartheta} - \vartheta_\lambda \right\|_2^2 \right] \gtrsim \frac{\sigma^2 d}{n},$$

whereas the two-stage estimator $\widehat{\vartheta}^{\text{ts}}_\lambda$ with penalty $\rho_k(\vartheta) = \alpha \left\| \vartheta \right\|_1$ and optimally chosen α achieves

$$\sup_{\substack{\theta_1, \theta_2 \in \Theta \\ \mathbf{X}_1, \mathbf{X}_2 \in \Gamma(c)}} \mathbb{E}\left[\left\| \widehat{\vartheta}^{\text{ts}}_\lambda - \vartheta_\lambda \right\|_2^2 \right] \lesssim \frac{\sigma^2 \log d}{n}.$$

See Appendix C.1 for the proof. As we can see, any directly penalized estimator may not be able to overcome the curse of being in high dimensions, e.g., when $d \gtrsim n$, even though the individual minimizers are 1-sparse. Our estimator, on the other hand, recovers the well-known rates of the LASSO (Tibshirani, 1996; Wainwright, 2019).

The proof of Proposition 1 relies on being able to choose the covariance matrices adversarially (within the eigenvalue constraints), so that the Pareto-optimal ϑ_λ lies anywhere in an ℓ_2-ball of fixed radius. If we did not allow for this (i.e., the covariance matrices are scaled identities), the Pareto-optimal ϑ_λ would have to be Ks-sparse, and the directly regularized estimator would work.

4 THEORETICAL GUARANTEES

We now provide estimation error bounds for our two-stage estimation procedure for a general set of problems.

4.1 Main results

In this section we state theoretical guarantees for objectives that satisfy the following two regularity assumptions.

Assumption 2 (Strong convexity, smoothness and local Lipschitz-continuity)**.**

1. *The function $\vartheta \mapsto \mathcal{L}_k(\vartheta, \theta_k, \omega_k)$ is convex, twice continuously differentiable, ν_k-smooth for all $k \in [K]$ and $(\theta_k, \omega_k) \in \mathbb{R}^m \times \mathbb{R}^M$. Further, for at least one $j \in [K]$ it is strongly convex w.r.t. the ℓ_2-norm,*
2. *The function $(\theta_k, \omega_k) \mapsto \nabla_\vartheta \mathcal{L}_k(\vartheta, \theta_k, \omega_k)$ is locally Lipschitz-continuous for all $k \in [K]$ and $\vartheta \in \mathbb{R}^m$.*

The definitions of strong convexity, smoothness and local Lipschitz-continuity are commonly used in the convex optimization literature; we recall them in Appendix A for completeness. They are standard assumptions in large parts of (multi-objective) optimization (Hillermeier, 2001; Roy et al., 2023; Ehrgott, 2005; Rockafellar, 1970), and are naturally satisfied for many standard losses in machine learning. For example, it is easily verified that the objectives from Example 1, and other objectives we discuss later, satisfy Assumption 2. For our lower bound, we additionally rely on the following *identifiability* assumption.

Assumption 3 (Lipschitz identifiability)**.** *For all $k \in [K], \omega_k \in \mathbb{R}^M$ and $\vartheta \in \mathbb{R}^m$ the function*

$$\theta_k \mapsto \nabla_\vartheta \mathcal{L}_k(\vartheta, \theta_k, \omega_k)$$

is C_k-Lipschitz, injective, and its inverse is C_k'-Lipschitz.

Assumptions of this type are common in the inverse optimization literature, which studies the identification of optimization parameters from a minimizer; cf. Aswani et al. (2018); Gebken and Peitz (2021) and references therein. Assumption 3 is satisfied by many common losses, such as Example 1 (under some assumptions). In Section 4.2, we discuss how identifiability relates to MOL.

The next theorem provides upper bounds on the estimation error $\|\widehat{\vartheta}_\lambda^{\text{ts}} - \vartheta_\lambda\|_2^2$ in terms of the estimation error of the parameters θ, ω. Additionally, we provide a lower bound on the minimax multi-objective estimation errors (5) in terms of the minimax single-objective ℓ_2-estimation error

$$\delta_k := \mathcal{M}_k(\mathcal{P}_k) = \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\left[\left\|\widehat{\theta}(\mathcal{D}) - \theta_k\right\|_2\right],$$

where the infimum is taken over all estimators that have access to the unlabeled and labeled datasets.

Theorem 1. *Let Assumption 2 hold and j be the index of the strongly convex function, and $\widehat{\vartheta}_\lambda^{\text{ts}}$ be the minimizer of the linear scalarization $(s_\lambda \circ \mathcal{L}) = \sum_{k=1}^K \lambda_k \mathcal{L}_k$ in Equation (8), with $\lambda_j > 0$. Then there exists a constant $C_{K,\lambda} > 0$ that only depends on K and λ such that*

$$\left\|\widehat{\vartheta}_\lambda^{\text{ts}} - \vartheta_\lambda\right\|_2^2 \leqslant C_{K,\lambda} \sum_{k=1}^K \left(\|\widehat{\theta}_k - \theta_k\|_2^2 + \|\widehat{\omega}_k - \omega_k\|_2^2\right).$$

If, additionally, Assumption 3 holds and ω is known, i.e. $\widehat{\omega} = \omega$, we have for $|\cdot|_+ = \max\{0, \cdot\}$ that

$$\mathcal{M}_\lambda(\mathcal{P}) \geqslant \max_{k \in [K]} s_\lambda^{-2}(\nu) \left| \frac{\lambda_k}{C_k'} \delta_k - \sum_{i \neq k} C_i \lambda_i \delta_i \right|_+^2.$$

We prove Theorem 1 in Appendix C.2. Notice that using $\max_{\lambda \in \Delta^K} C_{K,\lambda}$ we get a uniform bound on the estimation error. The upper bound in Theorem 1 has a very straightforward interpretation: The terms $\|\widehat{\theta}_k - \theta_k\|_2^2$ improve whenever the θ_k have a low-dimensional structure like sparsity, and the terms $\|\widehat{\omega}_k - \omega_k\|_2^2$ vanish when the number of unlabeled data points increase, recalling that estimating ω_k can be done purely with unlabeled data (Assumption 1).

On the other hand, the lower bound characterizes the "limits" of Pareto-optimal set estimation in the best case when we have a lot of unlabeled data (see also discussion in Section 4.4), in which case we can essentially think of $\widehat{\omega}_k = \omega_k$. This is for simplicity of exposition only, and could be readily adapted to capture the estimation error of $\widehat{\omega}$. We can now compare the upper and lower bound in the case when $\delta_i \ll \delta_k$ for some k and all $i \neq k$. Then, the lower bound reduces to $\mathcal{M}_\lambda(\mathcal{P}) \gtrsim \max_{k \in [K]} \delta_k^2$ if $\lambda_k > 0$. Moreover, since we consider $\widehat{\omega}_k = \omega_k$, the upper bound only consists of the terms $\|\widehat{\theta}_k - \theta_k\|_2^2$. Supposing $\widehat{\theta}_k$ estimates θ_k in a minimax optimal manner so that $\mathbb{E}\|\widehat{\theta}_k - \theta_k\|_2^2 \asymp \delta_k^2$ (e.g., via the correct choice of ρ), we obtain that

$$\max_{k \in [K]} \delta_k^2 \lesssim \mathcal{M}_\lambda(\mathcal{P}) \leqslant \mathbb{E}\left[\left\|\widehat{\vartheta}_\lambda^{\text{ts}} - \vartheta_\lambda\right\|_2^2\right] \lesssim \max_{k \in [K]} \delta_k^2,$$

and hence the two bounds are tight up to constant factors. We conclude that, under Assumptions 2 and 3, the difficulty of MOL is dominated by the hardest individual learning problem, if *all other* individual learning tasks are easier. In such a case our estimator is optimal.

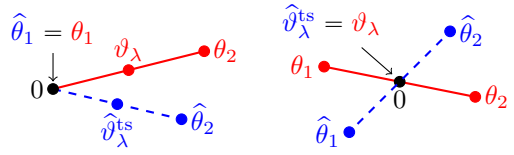
4.2 Discussion of the lower bound and Assumption 3

Intuitively, one might think that MOL should *always* be as hard as the hardest individual problem. However, this is not exactly what the lower bound predicts: If all individual minimax rates are of similar order ($\delta_k \asymp \delta_i$ for all k, j), the lower bound can vanish. Does this mean the lower bound is too loose, or can this actually happen? We now show in a simple setting how, indeed, this can happen for $\delta_k \asymp \delta_i$ for all k, j, and how the Lipschitz identifiability assumption prevents such a setting when exactly one parameter θ_k is harder to learn, i.e., $\delta_k \gg \delta_i$ for all $i \neq k$. Hence, the lower bound is not as loose as it might appear at first glance and captures some counterintuitive peculiarities of MOL. We use the following toy example, visualized in Figure 2.

Example 2. *Consider Example 1 with two objectives $\mathcal{L}_k(\vartheta, \theta_k) = \|\vartheta - \theta_k\|_2^2$, $k = 1, 2$, where the covariance matrices are the identity. Assume that the set of minimizers is known to satisfy one of the two:*

1. *$\theta_1 = -\theta_2$ (Figure 2(b)),*
2. *$\theta_1 = 0$ (Figure 2(a)).*

In the first case of Example 2, estimating θ_1 and θ_2 is *equally* hard (i.e., $\delta_1 = \delta_2$), but for the choice $\lambda = (1/2, 1/2)$, it holds that $\vartheta_\lambda = 0$ irrespective of θ_1 and θ_2. Hence, one can always estimate $\widehat{\vartheta}_\lambda^{\text{ts}} = 0$, and achieve estimation error $\|\widehat{\vartheta} - \vartheta_\lambda\|_2^2 = 0$. Thus, the lower bound *must* be 0, even though both individual learning tasks may be hard.



(a) $\delta_1 \ll \delta_2$ (b) $\delta_1 = \delta_2$

Figure 2: Two MOL problems to exemplify the counterintuitive lower bound from Theorem 1. (a): A problem where estimating one individual minimizer $\theta_1 = 0$ is easy, but estimating the other θ_2 and all other ϑ_λ is hard. (b): A problem where estimating both individual minimizers $\theta_1 = -\theta_2$ is hard, but estimating a parameter $\vartheta_\lambda = 0$ is easy.

It is now natural to wonder why the lower bound does not vanish for the case $\delta_1 \ll \delta_2$, such as in the second case of Example 2. Specifically, as discussed in Section 4.1, in such a case our lower bound is tight and predicts that learning ϑ_λ is as hard as learning θ_2. To understand this, interpreting the identifiability assumption Assumption 3 is crucial. On a high level, it implies that knowing ϑ_λ and θ_1, we can fully identify θ_2. Specifically, in the second case of Example 2, given the minimizer ϑ_λ for some λ, we can compute $\theta_2 = \vartheta_\lambda / \lambda_2$. Hence, anticipating a contradition, if we assume there existed a λ for which it were easier to learn ϑ_λ than θ_2, i.e., $\sup_{\theta_2} \mathbb{E}\|\widehat{\vartheta}_\lambda - \vartheta_\lambda\|_2 < \lambda_2 \delta_2$, then, by identifiability, we could also estimate θ_2 using $\widehat{\theta}_2 = \widehat{\vartheta}_\lambda / \lambda_2$, yielding $\sup_{\theta_2} \mathbb{E}\|\widehat{\theta}_2 - \theta_2\|_2 < \delta_2$ – a contradiction of the definition of the minimax rate.

This leads us to the following counterintuitive conclusion that is captured in the lower bound of Theorem 1: *Under Lipschitz indentifiability, a MOL problem may be easy if individual learning tasks are hard, but a MOL cannot be easy if all but one individual learning tasks are easy*.

4.3 Discussion of the upper bound and Assumption 2

We now turn to Assumption 2. The key ingredient for deriving guarantees for the two-stage estimator is to understand how improvements in estimating the parameters θ, ω translate into improvements to estimating the Pareto-optimal ϑ_λ. Assumption 2 is used in Theorem 1 to apply tools from optimization stability (Ito and Kunisch, 1992; Gfrerer and Klatte, 2016; Dontchev, 1995; Bonnans and Shapiro, 2000; Shvartsman, 2012), which is the study of how the minimizer of an optimization problem changes with respect to the parameters of that problem. In our setting, Assumption 2 ensures that the implicitly defined function

$$(\theta, \omega) \mapsto \vartheta_\lambda(\theta, \omega) = \arg\min_{\vartheta \in \mathbb{R}^m} (s_\lambda \circ \mathcal{L})(\vartheta, \theta, \omega)$$

is Lipschitz continuous (Shvartsman, 2012). Results in this literature often rely on assumptions similar to Assumption 2 for the Implicit Function Theorem to apply, see (Bonnans and Shapiro, 2000, §1). A literature review on sensitivity analysis for MOO is outlined in (Miettinen, 1998, §I.3.4) .

Assumption 2 excludes examples where all objectives are Lipschitz continuous, such as the Huber loss (Huber, 1964), and hence there is not one (globally) strongly convex function. Fortunately, the case when the loss function is Lipschitz in its parameters can easily be addressed using standard arguments on the excess scalarized objective, and the following proposition offers an alternative to Theorem 1.

Proposition 2 (Lipschitz parameterization)**.** *Assume that the parameterization* $(\theta_k, \omega_k) \mapsto \mathcal{L}_k(\cdot, \theta_k, \omega_k)$ *is* 1*-Lipschitz-continuous w.r.t.* $\Phi : (\mathbb{R}^m \times \mathbb{R}^M)^2 \to \mathbb{R}$*, i.e.*

$$\sup_{\vartheta \in \mathbb{R}^m} \left| \mathcal{L}_k(\vartheta, \theta_k, \omega_k) - \mathcal{L}_k(\vartheta, \theta'_k, \omega'_k) \right| \leqslant \Phi(\theta_k, \omega, \theta'_k, \omega'_k). \tag{9}$$

Then, for any *scalarization of the form* $s_\lambda(x) = \|\lambda \odot x\|$ *with some norm* $\|\cdot\|$*, the excess scalarized loss of* $\widehat{\vartheta}^{\text{ts}}_\lambda$*—as defined in Equation* (6)*—is bounded by*

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

Notably, Proposition 2 can apply to non-convex objectives, and allows the use of Chebyshev scalarization (Equation (1)). The proof, found in Appendix C.3, follows excess-risk style arguments similar to those found in Súkeník and Lampert (2022). The difference here is that we may still observe the effect of regularization and unlabeled data.

4.4 More examples

We now apply Theorem 1 and derive upper bounds for two concrete multi-objective learning problems: The problem described in Example 1, and a setting where we have a fairness-notion as the second objective.

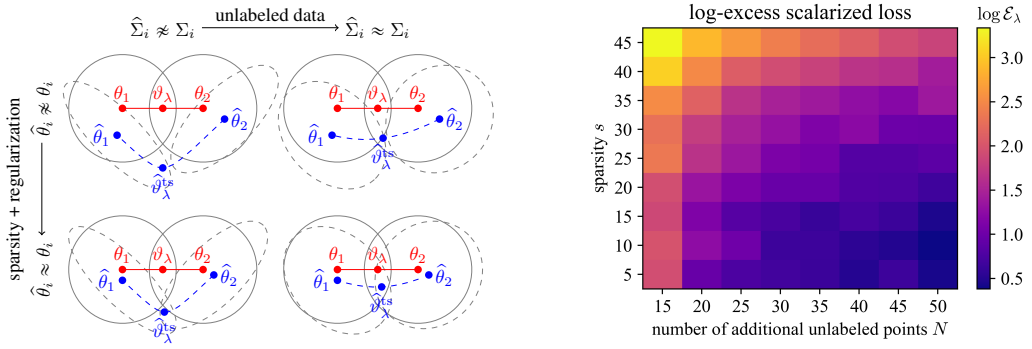
Multiple linear regression continued. We now apply Theorem 1 to the setting described in Example 1.

Corollary 1. *In the setting of Example 1, assume* $n_k = n, N_k = N$ *for all* i *and* $n + N \geqslant d$*. Then our estimator* $\widehat{\vartheta}^{\text{ts}}_\lambda$*, with* $\rho_k(\vartheta) = \alpha_k \|\vartheta\|_1$ *and optimally chosen* α_k*, achieves for all* $\lambda \in \Delta^K$

$$\mathbb{E}\left[\left\| \widehat{\vartheta}^{\text{ts}}_\lambda - \vartheta_\lambda \right\|_2^2 \right] \lesssim \frac{\sigma^2 s \log d}{n} + \frac{B^2 d^2}{n + N}.$$

The proof can be found in Appendix C.4.

Example 1 and Corollary 1 offer good intuition why it is sensible that a small upper bound requires both individual minimizers θ_k and the marginal quantities ω_k to be estimated well. We illustrate the effects of both regularization and the estimation of the marginal distribution (in terms of the covariance) in Figure 3 for the case that the true covariances are the identity. Notice how the estimators $\widehat{\vartheta}^{\text{ts}}_\lambda$ of the Pareto-optimal points depend both on the individual minimizers $\widehat{\theta}_k$ and the estimated covariance matrices. It is hence important for finding the entire Pareto set to estimate both θ and the covariance matrices well.



(a) (b)

Figure 3: The important roles of both regularization and additional unlabeled data for Example 1 illustrated on a geometric and 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 θ_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.

Fairness-risk tradeoff in linear regression. We now consider a setting where one of the losses is a group-wise fairness loss. Specifically, consider the random variables Y, X, A distributed according to $Y = \langle X, \theta \rangle + \xi$ with a s-sparse ground-truth θ, where $\xi \sim \mathcal{N}(0, \sigma^2)$, and $X|A \sim \mathcal{N}(\mu_A, \mathbf{I}_d)$, and $A \sim \text{Bernoulli}(1/2)$. A represents an observed protected group-attribute (such as gender or ethnicity), and covariates means of the groups differ $\mu_1 \neq \mu_2$.

As the first objective, we consider the usual prediction risk with square loss, as defined in Example 1, and as the second objective, we choose a recently introduced notion of unfairness (Gouic et al., 2020; Chzhen and Schreuder, 2022; Fukuchi and Sakuma, 2024) that measures *demographic parity* via the 2-Wasserstein distance between the group-wise distributions of $\langle X, \vartheta \rangle | A = a$ and their barycenter;

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

Details on this, including definitions, can be found in Appendix B, where we also demonstrate that under our assumptions, $\mathcal{L}_{\text{fair}}(\vartheta, \mathbb{P}) = \mathcal{L}_{\text{fair}}(\vartheta, \mu_1, \mu_2) = \frac{1}{4} \langle \mu_1 - \mu_2, \vartheta \rangle^2$. This violates the assumption of having a unique minimizer, which is why we have to restrict ourselves to the case that $\lambda_{\text{fair}} > 0$. Unless $\langle \mu_1 - \mu_2, \theta \rangle = 0$, there is a trade-off between fairness and risk. The following corollary applies Theorem 1 to this setting with the proof in Appendix C.5.

Corollary 2. *In the setting described above, assume that $\mu_1 = -\mu_2$, $n_k = n$, $N_k = N$. Then our estimator with the appropriate ℓ_1-penalty for estimating θ achieves for all $\lambda \in \Delta^2$ with $\lambda_{fair} > 0$*

$$\mathbb{E}\left[\left\| \widehat{\vartheta}_\lambda^{\text{ts}} - \vartheta_\lambda \right\|_2^2 \right] \lesssim \frac{\sigma^2 s \log d}{n} + \frac{d}{n+N}.$$

5 EXPERIMENTS

In this section we present some experiments on synthetic and real data to confirm the theoretical results from Section 4.4.

5.1 Multiple linear regression

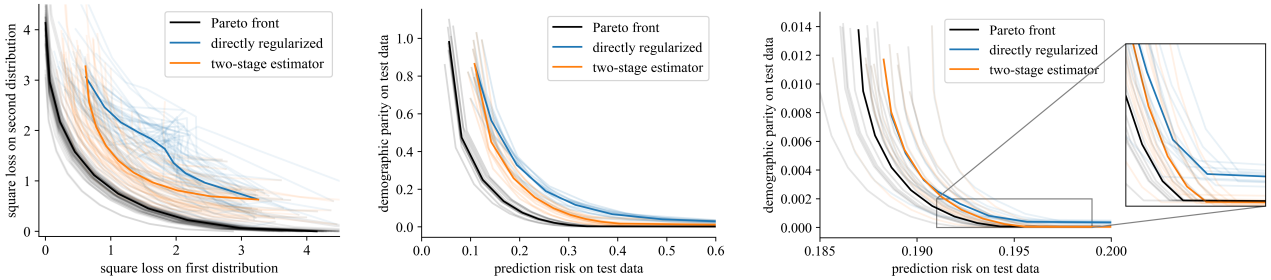
The first simulation is on synthetic data in the setting of Example 1 for two objectives. We present two experiments.

In the first experiment, we fix $\lambda_i = 1/2$, $d = 50$, $n_i = 15$ and two arbitrarily chosen covariance matrices Σ_1, Σ_2. The covariates are 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 (i.e., $\log \mathcal{E}_\lambda$) of our estimator with appropriately chosen ℓ_1-penalty in Figure 3. The smaller the 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 4(a) shows the Pareto fronts of 50 different runs and their point-wise average. As expected, the benefit of our method lies in the cases where $\lambda_1 \approx \lambda_2$.

5.2 Fairness-risk trade-off in linear regression

We also apply our estimator to two real fairness datasets: The Communities and Crime dataset (Redmond, 2002) where the task is to predict the number of violent crimes in a community and ethnicity is a protected attribute, and the Adult dataset (Becker and Kohavi, 1996) where the task is



(a) Multiple linear regression. (b) Communities and Crime dataset. (c) Adult dataset.

Figure 4: True Pareto fronts and their estimates using directly regularized estimators (blue) and our method (orange) for the experiments described in Section 5. On synthetic data (a) from sparse multiple linear regression and on the fairness datasets (Redmond, 2002; Becker and Kohavi, 1996), our estimator outperforms direct regularization methods.

to predict income and the protected attribute is gender.

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 4.4. Since the Adult dataset only has dimension $d = 13$, additional to subsampling, we add 1000 noisy features (sampled from a Gaussian) to artificially increase the data dimensionality to $d = 1013$. We then uniformly sample $n = 1000$ labeled and $N = 2000$ unlabeled examples, with the remaining samples serving as the test set. Our two-stage estimator and the directly regularized estimator (7) are then applied using an ℓ_1-penalty.

We repeat all experiments 10 times and show the resulting estimated Pareto fronts, as well as their point-wise average in Figures 4(b) and 4(c). On the Communities and Crime dataset, our estimator outperforms the directly regularized estimator across the entire Pareto front. On the Adult dataset, the difference is smaller, with the biggest difference being when the fairness is weighed more than the risk (zoomed-in box). Overall, these experiments confirm our approach.

6 RELATED WORK

Multi-objective learning is a rich field of research (Jin, 2006; 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; Ţifrea et al., 2024) and with connections to many adjacent fields machine learning such as 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), and transfer learning (Li et al., 2021a).

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 general MOL is still rather poorly studied, with the exceptions of Súkeník and Lampert (2022); Cortes et al. (2020). These works however do not yield meaningful bounds in the high-dimensional regime that we study in this paper.

7 CONCLUSIONS & FUTURE WORK

In this work, we propose an estimator for the Pareto front of a MOL problem that performs well in the high-dimensional regime by leveraging both the sparsity of the task-specific minimizers, as well as readily available unlabeled data. We investigate the estimator theoretically, and prove the optimality of the proposed estimator under certain conditions. While the focus of this work is primarily on sparsity, the proposed 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.

We leave it as future work to use arguments such as Fano's method to obtain different lower bounds that apply beyond the identifiability argument considered in Section 4, aiming to include settings similar to the weighted median (Fletcher et al., 2008). 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 (Yin et al., 2019). It remains an exciting direction for future research to demonstrate that the proposed estimator performs well in other multi-objective problems.

References

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.

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.

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

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

E. Chzhen and N. Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *Annals of Statistics*, 50: 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.

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*. 2005.

P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. Robust statistics on riemannian manifolds via the geometric median. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

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 multitask 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.

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.

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.

Y. Jin, editor. *Multi-Objective Machine Learning*, volume 16. 2006.

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.

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.

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.

J. A. López, S. Zapotecas-Martínez, and C. Coello. *An Introduction to Multiobjective Optimization Techniques*. 2011.

R. Marcinkevičs and J. E. Vogt. Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *WIREs Data Mining and Knowledge Discovery*, 13 (3):e1493, 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*. 1998.

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.

D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

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.

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

R. T. Rockafellar. *Convex Analysis*. 1970.

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. H. Lampert. Generalization in multi-objective machine learning. *arXiv preprint arXiv:2208.13499*, 2022.

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.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

J. Vaidya. *Privacy Preserving Data Mining*. 2009.

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.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. 2019.

H. Wang, L. L. He, R. Gao, and F. P. Calmon. Aleatoric and epistemic discrimination: fundamental limits of fairness interventions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2024.

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.

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.

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.

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.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

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 1,2,3)
   (b) Complete proofs of all theoretical results. Yes (Appendix C)
   (c) Clear explanations of any assumptions. Yes (Section 4.3)

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