Submitted to the Annals of Statistics

THE DISTRIBUTIONALLY ROBUST PREDICTION ERROR OF THE $\sqrt{\text{LASSO}}$ AND RELATED ESTIMATORS

BY JOSÉ LUIS MONTIEL OLEA[1,a], CYNTHIA RUSH[2,b], AMILCAR VELEZ[3,c], AND JOHANNES WIESEL[4,d]

[1]*Department of Economics, Cornell University,* [a]*montiel.olea@gmail.com*

[2]*Department of Statistics, Columbia University,* [b]*cynthia.rush@columbia.edu*

[3]*Department of Economics, Cornell University,* [c]*amilcare@cornell.edu*

[4]*Department of Mathematics, Uninversity of Copenhagen ,* [d]*wiesel@math.ku.dk*

We study the classical problem of predicting an outcome variable, Y, using a linear combination of a d-dimensional covariate vector, $\mathbf{X}$. We are interested in linear predictors whose coefficients solve:

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^d}\left(\mathbb{E}_{\mathbb{P}_n}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]\right)^{1/r}+\delta\,\rho\left(\boldsymbol{\beta}\right),$$

where $\delta>0$ is a regularization parameter, $\rho:\mathbb{R}^d\to\mathbb{R}_+$ is a convex penalty function, $\mathbb{P}_n$ is the empirical distribution of the data, and $r\geq 1$. Our main contribution is a new bound on the out-of-distribution prediction error of such estimators.

The new bound is obtained by combining three new sets of results. First, we provide conditions under which linear predictors based on these estimators solve a *distributionally robust optimization* problem: they minimize the worst-case prediction error over distributions that are close to each other in a type of *max-sliced Wasserstein metric*. Second, we provide a detailed finite-sample and asymptotic analysis of the statistical properties of the balls of distributions over which the worst-case prediction error is analyzed. Third, we present an oracle recommendation for the choice of regularization parameter, δ, that guarantees good out-of-distribution prediction error.

1. Introduction. The extent to which prediction algorithms can perform well not just on *training* data, but also on new, unseen, *testing* inputs is a central concern in machine learning. In fact, reducing a predictor's testing error—or equivalently, improving its "out-of-distribution" performance or "generalization error"—possibly at the expense of increased training error, is a typical informal motivation for introducing regularization strategies in statistical estimation; see, for example, [35, Chapter 7] and [40, Chapter 7]. More generally, the study of issues related to problems in which training and testing environments differ from one another is the subject of several recent, rapidly growing areas of research at the intersection of machine learning and statistics: transfer learning [44], distributional shifts [1, 31, 71], domain adaptation [7, 52], adversarial attacks [41, 46], learning under biased sampling [65] and cross-domain transfer performance [3] are some relevant examples.

In this paper, we study the classical problem of predicting an outcome variable, Y, using a linear combination of a d-dimensional covariate vector, $\mathbf{X}$. We focus on linear predictors whose coefficients, $\widehat{\boldsymbol{\beta}}$, solve the problem:

$$\arg\inf_{\boldsymbol{\beta}\in\mathbb{R}^d}\left(\mathbb{E}_{\mathbb{P}_n}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]\right)^{1/r}+\delta\,\rho(\boldsymbol{\beta}), \tag{1}$$

MSC2020 subject classifications: Primary 00X00, 00X00; secondary 00X00.
Keywords and phrases: First keyword, second keyword.

where $\delta > 0$ is a regularization parameter, $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is a convex penalty function, $\mathbb{P}_n$ is the empirical distribution of the data, and $r \geq 1$. We assume that both ρ and r have been determined by the statistician, and make no attempt to provide normative statements regarding their selection. The square-root LASSO (henceforth, $\sqrt{\text{LASSO}}$) [5], the square-root group LASSO [18], the square-root sorted ℓ_1 penalized estimator (SLOPE) [70], and the ℓ_1-penalized least absolute deviation estimator [76] provide examples of estimators obtained by solving (1) with different choices for r and ρ.

We are interested in studying the out-of-distribution prediction error associated with such estimators; namely

$$\mathbb{E}_{\mathbb{Q}}\left[\left|Y - \mathbf{X}^\top \widehat{\boldsymbol{\beta}}\right|^r\right], \tag{2}$$

where the expectation above is computed by fixing the estimated $\widehat{\boldsymbol{\beta}}$, and then drawing new covariates and outcomes according to some joint distribution $\mathbb{Q}$. The distribution $\mathbb{Q}$ is similar, but not necessarily equal to, the true data generating process, $\mathbb{P}$, or the empirical distribution of the data, $\mathbb{P}_n$.

Our main result is the following upper bound on the out-of-distribution prediction error (see Theorem 5.1 for a formal statement).

THEOREM (Informal). *If δ is chosen appropriately, then for all $\mathbb{Q}$ and at any $\boldsymbol{\beta}$, with high probability:*

$$\mathbb{E}_{\mathbb{Q}}\left[\left|Y - \mathbf{X}^\top \boldsymbol{\beta}\right|^r\right]^{1/r} \leq \mathbb{E}_{\mathbb{P}_n}\left[\left|Y - \mathbf{X}^\top \boldsymbol{\beta}\right|^r\right]^{1/r} + (\delta + \widehat{\mathcal{W}}_r(\mathbb{P}, \mathbb{Q}))(1 + \rho(\boldsymbol{\beta})), \tag{3}$$

where $\widehat{\mathcal{W}}_r$ denotes a type of max-sliced Wasserstein metric.

Consequently, for an appropriately chosen δ, linear predictors solving (1) have good out-of-sample performance at the true, unknown distribution of the data $\mathbb{P}$, and, also, at *testing* distributions $\mathbb{Q}$ that are close to $\mathbb{P}$ in terms of $\widehat{\mathcal{W}}_r$. In fact, we show that the objective function in (1) serves as a lower and upper bound for the out-of-distribution prediction error, up to some adjustment terms (see Corollary C.1 in Appendix C.5).

We present a formal definition of the metric $\widehat{\mathcal{W}}_r$ in (4) below, and explain how distributions that are close in this metric are required to have similar prediction errors (in a sense we make precise). The proof of the theorem above is based on three intermediate results, which bring together ideas related to *distributionally robust optimization* (DRO), finite sample analysis of the max-sliced Wasserstein metric, and empirical process theory. We believe that the three steps used to prove (3) provide results that are interesting in their own right, and in what follows, we discuss each of these steps in more detail.

First, we show that estimators constructed using (1) are equivalent to those that solve a DRO problem based on a $\widehat{\mathcal{W}}_r$-ball around $\mathbb{P}_n$ (Theorem 2.1, Section 2). The DRO representation naturally yields finite-sample bounds for (2) in terms of (1), provided that distributions $\mathbb{Q}$ are close to $\mathbb{P}_n$ in terms of our suggested metric (Section 2.3 provides examples of distributions contained in the $\widehat{\mathcal{W}}_r$-balls). Thus, our first result provides theoretical support for the claim that predictors based on estimators obtained via (1) (such as the $\sqrt{\text{LASSO}}$ and related estimators) have good out-of-distribution performance.

Second, we provide a detailed statistical analysis of the balls of distributions based on our suggested metric. More precisely, we determine the required size of a ball centered at $\mathbb{P}_n$ to guarantee that it contains $\mathbb{P}$ with high probability. We present both finite-sample results (Theorem 3.1 and Theorem 3.2 in Section 3) and large-sample approximations (Theorem 4.1 and 4.2 in Section 4). The proofs of these results are based on a novel connection between an upper bound for the Wasserstein distance in $d = 1$, and classical bounds from empirical

process theory for self-normalized processes. Its relative simplicity enables us to find the explicit constants above. Our analysis suggests that our balls are *statistically larger* than those based on the standard Wasserstein metric (Remark 4). Because the balls we consider are statistically larger, their radii can shrink to zero faster than order $n^{-1/d}$ (the usual rates for Wasserstein balls), and still contain $\mathbb{P}$ (see Figure 1).

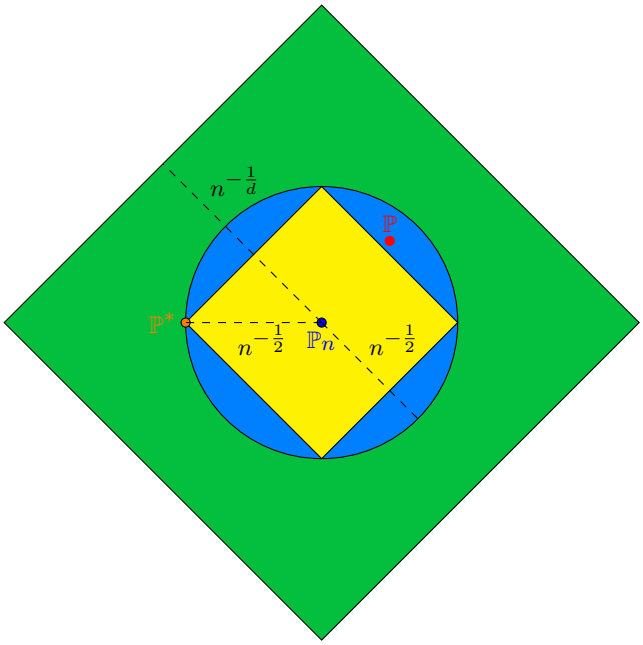


Fig 1: ρ-max-sliced Wasserstein ball of radius $n^{-1/2}$ (blue) vs. d-dimensional Wasserstein ball of radius $n^{-1/2}$ (yellow) and $n^{-1/d}$ (green). The measure $\mathbb{P}^*$ (orange) is a worst-case distribution that solves the DRO problem and is included in both Wasserstein balls of radius $n^{-1/2}$; see Remark 3 for details.

Third, we use the DRO representation of (1) and the statistical analysis of our max-sliced Wasserstein balls to i) derive oracle recommendations for the penalization parameter δ (Section 5) that guarantee good out-of-distribution prediction error (Theorem 5.1 in Section 5.1); and ii) present a test statistic to rank the out-of-distribution performance of two different linear estimators (Section 5.4). In Section 6 we present a small-scale simulation to illustrate the performance of predictions based on the $\sqrt{\text{LASSO}}$ but using our recommended parameter δ.

None of our results rely on sparsity assumptions about the true data generating process or on the relation between the sample size and the number of covariates; thus, our results broaden the scope of use of the $\sqrt{\text{LASSO}}$ and related estimators in prediction problems.

1.1. *Related Literature.* Distributionally robust optimization problems have been shown to be equivalent to various forms of penalized regression, variance-penalized estimation, and dropout training [12, 13, 30, 37, 47, 53, 55], depending on the choice of uncertainty set. It is typical to define uncertainty sets using metrics or divergences: e.g., total variation, Hellinger, Gelbrich distance [55] or Kullback-Leibler divergence [24, 63]. To the best of our knowledge, the use of the max-sliced Wasserstein metric to define uncertainty sets in DRO problems is novel.

The DRO representation of (1) has been established in [10, Theorem 1] using a ball in a modified Wasserstein metric, which is a common choice for the uncertainty set in the distributionally robust optimization literature [11, 36, 37, 45, 48, 53, 66, 69]. Relative to previous

results, beyond just changing the metric used in constructing the ball, we focus on convex penalty functions (instead of only norms) and we explicitly identify a worst-case measure in the DRO representation of (1), instead of relying on duality arguments. In this sense, our proof can be seen as a natural extension of [8, Theorem 1].

DRO representations are known to be useful in many situations, for example, those where the trained procedure will be evaluated on test data from a distribution $\widetilde{\mathbb{P}}$ that is close to that of the training data, $\mathbb{P}$, but may be different [7], when there are covariate shifts [2, 20, 60, 62, 67, 71, 72, 78], or when one experiences adversarial attacks [41, 46]. As discussed in the seminal work of [8], DRO representations *"offer a different perspective on regularization methods by identifying which adversarial perturbations the model is protected against"*. This *a fortiori* means that in any helpful DRO representation, the set of adversarial distributions that a regularization method protects against must depend on the regularizer itself. Thus, it should not be surprising that the max-sliced Wasserstein metric introduced in this paper depends on ρ. And in fact, previous uses of the Wasserstein metric for DRO representations of the $\sqrt{\text{LASSO}}$ and related estimators also depend on ρ; c.f. Proposition 2 in [10].

Starting from [27, 34], the question of establishing finite sample bounds on the Wasserstein metric and its variants has seen a spike in research activity over the last years: an incomplete list is [15, 23, 49, 57, 68, 77]; see also the references therein. When $d > 2r$, tight rates for $\mathcal{W}_r(\mathbb{P}_n, \mathbb{P})$ are of the order $n^{-1/d}$, i.e. they suffer from the curse of dimensionality. As our results show, this is not the case for the ρ-MSW distance. The faster rates of convergence for the max-sliced Wasserstein metric were first observed in [57] for subgaussian probability measures and in [51] under a projective Poincaré/Bernstein inequality. More recently, [4] have obtained sharp rates for $r = 2$ and isotropic distributions. Our rates are of the same order, up to logarithmic factors, and simultaneously hold for all $r \geq 1$ and all distributions with finite higher-order moments. Lastly, let us mention that most of the papers cited above only give explicit *rates*, while the *constants* are often non-explicit and large, cf. [33]. A notable exception is the recent work of [56] and [39]. In particular, using log-concavity, [56] derives sharp rates for the max-sliced Wasserstein metric that explicitly state the dependence on the dimension of the data. In Section 5, we further discuss how our concentration results can be used to provide a recommendation for δ based on our DRO representation.

A large part of the theoretical literature studying penalized regressions as in (1) has explicit recommendations for the choice of the penalization parameters, see [80] for a review. For example, in the case of the $\sqrt{\text{LASSO}}$, [5, 6] proposed a pivotal penalization parameter and establish asymptotic performance guarantees. For the case of the LASSO estimator, [22] present conditions under which the popular cross-validation method has nearly optimal rates of convergence in prediction norms, while [21] suggests utilizing a bootstrap approximation to estimate the penalization parameter. Our work complements these previous results by recommending a penalization parameter that explicitly controls the out-of-distribution prediction error (for a finite sample and/or asymptotically).

1.2. *Outline.* The rest of the paper is organized as follows. In Section 2, we present a detailed derivation of the DRO representation of (1). In Sections 3 and 4 we present rates for the MSW distance $\widehat{\mathcal{W}}_r$ between the true and empirical measure, both for $\mathbb{P}$ with compact support and for $\mathbb{P}$ satisfying a moment condition. Section 3 gives a finite sample analysis, while Section 4 provides asymptotics. In Section 5, we present a recommendation for the selection of regularization parameter, $\delta_{n,r}$, that guarantees good out-of-distribution prediction error. We also present a test statistic to rank the out-of-distribution performance of two different linear estimators. In Section 6, we present a small-scale simulation to illustrate the performance of predictions based on the $\sqrt{\text{LASSO}}$ but using our recommended parameter δ. All the proofs are collected in the Supplementary Material [54].

1.3. *Notation.* *Random Variables.* We use capital, bold letters—such as $\mathbf{Z}$ and $\widetilde{\mathbf{Z}}$—to denote Borel measurable random vectors in $\mathbb{R}^d$, and use Z_j to denote the j-th coordinate of $\mathbf{Z}$. We denote the set of all Borel probability measures in $\mathbb{R}^d$ by $\mathcal{P}(\mathbb{R}^d)$ and let $\mathcal{P}_r(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$ denote all Borel probability measures with finite rth moments. If the random vector $\mathbf{Z}$ has distribution or law $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$, we write $\mathbf{Z} \sim \mathbb{P}$. The expectation of $\mathbf{Z}$ is denoted as $\mathbb{E}_{\mathbb{P}}[\mathbf{Z}]$.

Covariates and outcome variables. We reserve $\mathbf{X}$ for the random column vector collecting the d covariates available for prediction, and Y for the scalar outcome variable. The realizations of covariates and outcomes are denoted as $\mathbf{x}$ and y, respectively. In a slight abuse of notation, we sometimes write $(\mathbf{X}, Y)$ to denote a random vector in $\mathbb{R}^{d+1}$ (instead of $(\mathbf{X}^\top, Y)^\top$).

Couplings. For two probability measures $\mathbb{Q}$ and $\mathbb{P}$, we define a *coupling* of $\mathbb{Q}$ and $\mathbb{P}$ as any element of $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ that preserves the marginals over $\mathbb{R}^d$. We denote the collection of all such couplings as $\Pi(\mathbb{Q}, \mathbb{P})$. By definition, if $(\widetilde{\mathbf{Z}}, \mathbf{Z})$ is an $\mathbb{R}^d \times \mathbb{R}^d$-valued random vector with distribution $\pi \in \Pi(\mathbb{Q}, \mathbb{P})$, then $\widetilde{\mathbf{Z}} \sim \mathbb{Q}$ and $\mathbf{Z} \sim \mathbb{P}$.

Penalty functions. For a function $\rho : \mathbb{R}^d \to \mathbb{R}$ we write

$$\rho^*(\boldsymbol{\beta}) := \sup_{\mathbf{x} \in \mathbb{R}^d} \left\{ \boldsymbol{\beta}^\top \mathbf{x} - \rho(\mathbf{x}) \right\}$$

for its conjugate (see [64]). If ρ is convex, a vector $\boldsymbol{\beta}^*$ is said to be a subgradient of ρ at a point $\boldsymbol{\beta}$ if

$$\rho(\mathbf{x}) \geq \rho(\boldsymbol{\beta}) + {\boldsymbol{\beta}^*}^\top (\mathbf{x} - \boldsymbol{\beta}), \qquad \forall \mathbf{x} \in \mathbb{R}^d.$$

The set of all subgradients of ρ at $\boldsymbol{\beta}$ is called the subdifferential of ρ at $\boldsymbol{\beta}$ and is denoted by $\partial\rho(\boldsymbol{\beta})$, [64, pp. 214-215].

Lastly, let us mention two important facts that will be relevant in Section 2.3. If ρ is differentiable, then its subdifferential $\partial\rho(\boldsymbol{\beta})$ is a singleton that equals the gradient of ρ at $\boldsymbol{\beta}$; see, for example, [64, Theorem 25.1]. If ρ is a norm in $\mathbb{R}^d$, then ρ^* is only equal to zero or infinity; see [17, p. 93].

2. Reformulation of equation (1) as a DRO problem.

2.1. *The ρ-max-sliced Wasserstein metric.* For any $r \in [1, \infty)$ and $\rho \colon \mathbb{R}^d \to [0, +\infty)$, we define the *$\rho$-max-sliced Wasserstein (ρ-MSW) metric*[1][2]

$$\widehat{\mathcal{W}}_r(\mathbb{P}, \widetilde{\mathbb{P}}) := \sup_{\boldsymbol{\gamma} \in \mathbb{R}^d} \Big(\inf_{\pi \in \Pi(\mathbb{P}, \widetilde{\mathbb{P}})} \frac{1}{1 + \rho(\boldsymbol{\gamma})} \Big(\mathbb{E}_\pi \big[\big| (Y - \mathbf{X}^\top \boldsymbol{\gamma}) - (\widetilde{Y} - \widetilde{\mathbf{X}}^\top \boldsymbol{\gamma}) \big|^r \big] \Big)^{1/r} \Big). \tag{4}$$

Here, for arbitrary distributions $\mathbb{P}$ and $\widetilde{\mathbb{P}}$, the set $\Pi(\mathbb{P}, \widetilde{\mathbb{P}})$ denotes the collection of probability distributions over random vectors $((\mathbf{X}, Y), (\widetilde{\mathbf{X}}, \widetilde{Y}))$, with marginal distributions $(\mathbb{P}, \widetilde{\mathbb{P}})$ (that is, the set $\Pi(\mathbb{P}, \widetilde{\mathbb{P}})$ is the collection of *couplings* of $\mathbb{P}$ and $\widetilde{\mathbb{P}}$). We refer to r as the Wasserstein exponent.

[1] Sliced Wasserstein distances [16, 61]—i.e., distances between probability distributions that consider the average or maximum of standard Wasserstein distances between one-dimensional projections—have been the subject of recent research in statistics and machine learning; see, for example, [43], [56] and the references therein. As we already mentioned, the max-sliced Wasserstein distance has been studied recently in [4, 51, 57]. Its use in the analysis of out-of-distribution prediction error of the $\sqrt{\text{LASSO}}$ and related estimators yields a hitherto unexplored connection to the field of statistical optimal transport, which we hope to be attractive from a methodological point of view.

[2] [54, Lemma B.2] shows that the ρ-MSW metric is indeed a metric.

We remark that the infimum in the definition of $\widehat{\mathcal{W}}_r(\mathbb{Q},\mathbb{P})$ given in (4) is attained for fixed $\boldsymbol{\gamma}$.[3] Furthermore, for any norm ρ, the supremum over $\boldsymbol{\gamma}$ in equation (4) is also attained.

Intuitively, $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ are close in ρ-MSW metric with Wasserstein exponent r if, for any $\boldsymbol{\gamma}$, there exists a coupling of $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ that makes the r-th norm of the *difference of their prediction errors* small, relative to $\rho(\boldsymbol{\gamma})$.

It is useful to compare the ρ-MSW metric to the d-dimensional Wasserstein metric with cost $\|\cdot\|$, defined by

$$\mathcal{W}_r(\mathbb{P},\widetilde{\mathbb{P}}) := \inf_{\pi\in\Pi(\mathbb{P},\widetilde{\mathbb{P}})} \Big(\mathbb{E}_\pi\big[\|(\widetilde{\mathbf{X}},\widetilde{Y})-(\mathbf{X},Y)\|^r\big]\Big)^{1/r}, \tag{5}$$

where $\|\cdot\|$ is an arbitrary metric on $\mathbb{R}^{d+1}$. Remark 4 in the Supplementary Material [54] shows that for a large class of penalty functions ρ, the balls constructed with (4) will typically be larger than those based on (5) when the radius is the same.

It is also useful to note that our ρ-MSW metric is a slight generalization of the *max-sliced Wasserstein metric* (MSW), first considered in [28, 43, 57, 58]. The MSW distance over probability distributions $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ on $\mathbb{R}^{d+1}$ is defined as

$$\overline{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) := \sup_{\widetilde{\boldsymbol{\gamma}}\in\mathbb{R}^{d+1}:\|\widetilde{\boldsymbol{\gamma}}\|_2=1} \mathcal{W}_r(\widetilde{\boldsymbol{\gamma}}_*\mathbb{P}\,,\widetilde{\boldsymbol{\gamma}}_*\widetilde{\mathbb{P}})\,, \tag{6}$$

where $\mathcal{W}_r$ is the one-dimensional Wasserstein metric, and $\widetilde{\boldsymbol{\gamma}}_*\mathbb{P}$ denotes the pushforward probability of $\mathbb{P}$ with respect to the linear map $\mathbf{z}\in\mathbb{R}^{d+1}\mapsto \mathbf{z}^\top\widetilde{\boldsymbol{\gamma}}\in\mathbb{R}$. The supremum is defined over all linear, one-dimensional projections generated by the vectors in the unit sphere. We provide further details about the MSW metric in the discussion following equation (17).

To further illustrate the similarities between the MSW and our ρ-MSW metric, it is convenient to assume, for the moment, that ρ is a norm on $\mathbb{R}^d$. Define the function $\|\cdot\|_\rho$ on $\mathbb{R}^{d+1}$ via $\|\widetilde{\boldsymbol{\gamma}}\|_\rho = |y|+\rho(\boldsymbol{\gamma})$, where $\widetilde{\boldsymbol{\gamma}}=(\boldsymbol{\gamma},y)$, with $\boldsymbol{\gamma}\in\mathbb{R}^d$ and $y\in\mathbb{R}$. Note that $\|\cdot\|_\rho$ is a norm. Then Lemma B.1 in the Supplementary Material [54] shows that the ρ-MSW metric can be written as

$$\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) = \sup_{\widetilde{\boldsymbol{\gamma}}\in\mathbb{R}^{d+1}:\|\widetilde{\boldsymbol{\gamma}}\|_\rho=1} \mathcal{W}_r(\widetilde{\boldsymbol{\gamma}}_*\mathbb{P}\,,\widetilde{\boldsymbol{\gamma}}_*\widetilde{\mathbb{P}})\,.$$

Thus, when $\rho(\cdot)$ is a norm, the only difference between our ρ-MSW metric and the usual MSW metric is the set of one-dimensional projections that are used to define each metric. For example, when $\rho(\cdot)=\|\cdot\|_1$, the ρ-MSW metric considers the supremum on the unit sphere defined by the ℓ_1-norm, while the MSW metric considers the supremum on the unit sphere defined by the ℓ_2-norm. In both cases, restricting the norm of the one-dimensional linear projections is needed to guarantee that these metrics are finite. This is achieved through the normalizing factor $(1+\rho(\boldsymbol{\gamma}))$ in (4).

2.2. *The DRO problem.* We define the collection of distributions

$$\begin{aligned} B_\delta(\mathbb{P}) &:= \Big\{\mathbb{Q}\in\mathcal{P}_r(\mathbb{R}^{d+1}) : \widehat{\mathcal{W}}_r(\mathbb{Q},\mathbb{P})\le\delta\Big\} \\ &= \Big\{\mathbb{Q}\in\mathcal{P}_r(\mathbb{R}^{d+1}) :\ \forall\boldsymbol{\gamma}\in\mathbb{R}^d \quad \exists \text{ a coupling } \pi(\boldsymbol{\gamma})\in\Pi(\mathbb{P},\mathbb{Q}) \\ &\qquad \text{for which } \mathbb{E}_{\pi(\boldsymbol{\gamma})}\big[|(\widetilde{Y}-Y)+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\gamma}|^r\big]\le\delta^r(1+\rho(\boldsymbol{\gamma}))^r, \\ &\qquad \text{where } \big((\mathbf{X},Y),(\widetilde{\mathbf{X}},\widetilde{Y})\big)\sim\pi(\boldsymbol{\gamma})\Big\}, \end{aligned} \tag{7}$$

[3] Indeed, note that the function $((\mathbf{x},y),(\widetilde{\mathbf{x}},\widetilde{y}))\mapsto|(\widetilde{y}-y)+(\widetilde{\mathbf{x}}-\mathbf{x})^\top\boldsymbol{\gamma}|^r$ is continuous and non-negative. The result then follows from [74, Theorem 4.1].

where we have suppressed the dependence of the ball $B_\delta(\mathbb{P})$ on r, ρ for notational simplicity.

The main result of this section establishes a formal connection between the solutions to the problems in (1) and a DRO problem.

THEOREM 2.1. *Fix $1 \le r < \infty$. Let $\rho\colon \mathbb{R}^d \to [0, +\infty)$ be a convex penalty function. Suppose that, for any $\boldsymbol{\beta} \in \mathbb{R}^d$, there exists a subgradient $\boldsymbol{\beta}^* \in \partial\rho(\boldsymbol{\beta})$ such that*

$$\left| \boldsymbol{\gamma}^\top \left(\boldsymbol{\beta}^* - \frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top \boldsymbol{\beta}} \rho^*(\boldsymbol{\beta}^*) \right) \right| \le \rho(\boldsymbol{\gamma}), \qquad \forall\, \boldsymbol{\gamma} \in \mathbb{R}^d. \tag{8}$$

Then, for any $\delta \ge 0$ and any $\boldsymbol{\beta} \in \mathbb{R}^d$ we have

$$\sup_{\widetilde{\mathbb{P}} \in B_\delta(\mathbb{P})} \mathbb{E}_{\widetilde{\mathbb{P}}} \left[\left| Y - \mathbf{X}^\top \boldsymbol{\beta} \right|^r \right] = \left(\sqrt[r]{\mathbb{E}_{\mathbb{P}} \left[|Y - \mathbf{X}^\top \boldsymbol{\beta}|^r \right]} + \delta \left(1 + \rho(\boldsymbol{\beta}) \right) \right)^r. \tag{9}$$

Theorem 2.1 shows that the worst-case, out-of-distribution performance of any linear predictor over the collection of distributions $B_\delta(\mathbb{P})$ equals the r-th power of the objective function in (1). The result in (9) thus implies

$$\arg\inf_{\boldsymbol{\beta} \in \mathbb{R}^d} \left[\sup_{\widetilde{\mathbb{P}} \in B_\delta(\mathbb{P})} \mathbb{E}_{\widetilde{\mathbb{P}}} \left[\left| Y - \mathbf{X}^\top \boldsymbol{\beta} \right|^r \right] \right] = \arg\inf_{\boldsymbol{\beta} \in \mathbb{R}^d} \sqrt[r]{\mathbb{E}_{\mathbb{P}} \left[|Y - \mathbf{X}^\top \boldsymbol{\beta}|^r \right]} + \delta \left(1 + \rho(\boldsymbol{\beta}) \right). \tag{10}$$

Our interpretation of equation (10) is that the $\sqrt{\text{LASSO}}$ and related estimators in (1) have good out-of-distribution performance for any *testing* distribution $\widetilde{\mathbb{P}}$, that is not far (in terms of the ρ-MSW metric) from the baseline *training* distribution $\mathbb{P}$. This result is independent of how the regularization parameter δ is selected and generalizes the connection between regularization and generalization performance first established in [8].

By construction, the minimax problem in (10) provides robustness in situations where i) the trained procedure will be evaluated on test data from a distribution $\widetilde{\mathbb{P}}$ that is close to that of the training data, $\mathbb{P}$, but may be different [7]; ii) where there are covariate shifts [2, 20, 60, 62, 67, 71, 72, 78]; or iii) when there is an adversarial attack [41, 46].

To the best of our knowledge, Theorem 2.1 is new. As mentioned above, a result in the spirit of (10) for the case $r = 2$, penalty $\rho(\cdot) = \|\cdot\|_p$ with $p \ge 1$, and $\widehat{\mathcal{W}}_r$ replaced by a modified Wasserstein metric (see [54, Sections C.4 and C.5] in the supplement for an extensive discussion), was first established in [10, Theorem 1], using optimal transport duality [13, 37]. This has recently been extended to more general penalty functions [25, 79]. Our balls are different to the ones considered in these papers as we focus on convex penalty functions. Moreover, our proofs do not rely on duality arguments and explicitly identify a worst-case measure $\mathbb{P}^* \in B_\delta(\mathbb{P})$ for (9), which is given by an additive perturbation of $\mathbb{P}$ (see Corollary 2.1 below). In this sense, our proof can be seen as a natural extension of the seminal results in [8, Theorem 1].

We believe that $\widehat{\mathcal{W}}_r$ is the natural metric to assess out-of-distribution performance, as it allows for the construction of neighborhoods containing a general class of *testing* distributions that are only required to generate similar prediction errors as the *training* distribution $\mathbb{P}_n$. To see this, note that the out-of-distribution prediction error, $\mathbb{Q} \mapsto \mathbb{E}_{\mathbb{Q}}[|\widetilde{Y} - \widetilde{\mathbf{X}}^\top \boldsymbol{\beta}|^r]^{1/r}$, is a Lipschitz continuous function under the ρ-MSW metric for any given $\boldsymbol{\beta}$ and penalty function ρ. Specifically, for any two distributions $\mathbb{Q}$ and $\mathbb{P}$, the definition of ρ-MSW implies

$$\left| \, \mathbb{E}_{\mathbb{Q}}[|\, \widetilde{Y} - \widetilde{\mathbf{X}}^\top \boldsymbol{\beta} \,|^r]^{1/r} - \mathbb{E}_{\mathbb{P}}[|\, Y - \mathbf{X}^\top \boldsymbol{\beta} \,|^r]^{1/r} \, \right| \le \left(1 + \rho(\boldsymbol{\beta}) \right) \widehat{\mathcal{W}}_{r,\rho}(\mathbb{P}, \mathbb{Q}). \tag{11}$$

Consequently, $1+\rho(\boldsymbol{\beta})$ can be interpreted as a Lipschitz constant that varies with $\boldsymbol{\beta}$. Thus, for fixed $\boldsymbol{\beta}$, any two distributions that are close under the ρ-MSW metric have similar prediction errors. The standard d-dimensional Wasserstein metric puts additional restrictions on the testing distributions considered. This means that two distributions can have similar prediction errors, but their Wasserstein distance could be very large, especially in high dimensions (making the associated bounds not very useful in practice). In Section C.3 of the Supplementary Material [54], we further provide an example of two Gaussian distributions for which the difference in prediction errors is small, but the standard Wasserstein metric is large.

We briefly sketch the proof of Theorem 2.1 here and refer to Section A.2 in the Supplementary Material [54] for details: first note that (11) implies

$$\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right] \leq \left(\sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]}+\delta(1+\rho(\boldsymbol{\beta}))\right)^r, \tag{12}$$

holds for any $\boldsymbol{\beta}\in\mathbb{R}^d$ and any $\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})$. We then show that for any $\boldsymbol{\beta}\in\operatorname{dom}(\rho)$, the upper bound given in (12) is tight. That is, for each $\boldsymbol{\beta}\in\operatorname{dom}(\rho)$, we explicitly construct a distribution $\mathbb{P}^*_{\boldsymbol{\beta}}\in B_\delta(\mathbb{P})$ for which the bound holds exactly. The worst-case distribution is presented in Corollary 2.1 below.

COROLLARY 2.1. *For each $\boldsymbol{\beta}\in\mathbb{R}^d$ the supremum in (9) is attained for the distribution $\mathbb{P}^*_{\boldsymbol{\beta}}$ corresponding to the random vector $(\widetilde{\mathbf{X}},\widetilde{Y})$ defined as*

$$\widetilde{\mathbf{X}}=\mathbf{X}-e\left(\boldsymbol{\beta}^*-\frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*\left(\boldsymbol{\beta}^*\right)\right), \qquad \widetilde{Y}=Y+e,$$

where

$$e:=\frac{\delta\left(Y-\mathbf{X}^\top\boldsymbol{\beta}\right)}{\sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]}}, \qquad (\mathbf{X},Y)\sim\mathbb{P}.$$

In essence, the testing distribution that attains the worst out-of-distribution performance is an additive perturbation of the baseline training distribution. The perturbation has a low-dimensional structure where a one-dimensional error, e, which is proportional to prediction error, $Y-\mathbf{X}^\top\boldsymbol{\beta}$, is added to $\mathbf{X}$ using loadings that depend on the subgradient of ρ at $\boldsymbol{\beta}$ and also on the conjugate of ρ.

REMARK 1. *It is easy to see that the minimizer in (10) is attained. Denote it by $\boldsymbol{\beta}(\mathbb{P})$. Then $\boldsymbol{\beta}(\mathbb{P})$ is also a minimizer of $\mathbb{E}_{\mathbb{P}^*_{\boldsymbol{\beta}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]$. Indeed, for any $\boldsymbol{\beta}\in\mathbb{R}^d$ we have*

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^*_{\boldsymbol{\beta}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right] &= \sup_{\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})}\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right] \\
&\geq \inf_{\boldsymbol{\beta}\in\mathbb{R}^d}\sup_{\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})}\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right] \\
&= \sup_{\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})}\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}(\mathbb{P})\right|^r\right] \\
&\geq \mathbb{E}_{\mathbb{P}^*_{\boldsymbol{\beta}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}(\mathbb{P})\right|^r\right].
\end{aligned}$$

In particular, for any linear predictor with slope $\boldsymbol{\beta}$, it is always possible to find a perturbation of $\mathbb{P}$ for which a predictor based on (10) performs better.

REMARK 2 (On condition (8)). *If ρ is a norm, then the condition in (8) is automatically satisfied; i.e., there exists a $\boldsymbol{\beta}^*\in\partial\rho(\boldsymbol{\beta})$ such that (8) is true. Thus, the conclusion of Theorem 2.1 holds for all $\rho(\cdot)=\|\cdot\|$ that are norms.*

On the other hand, Example A.1 in the Supplementary Material [54] shows that condition (8) can be satisfied by ρ that are not norms.

2.3. *Examples of distributions in the ρ-MSW ball.* In this subsection we analyze the types of testing distributions that are contained in the ball defined in (7). We do this by considering different estimators that take the form (1).

2.3.1. *$\sqrt{LASSO}$.* Let us take $r = 2$ and

$$\rho(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{d} |\beta_j|.$$

Under this choice of penalty function, the regression problem (10) is the objective function of the $\sqrt{\text{LASSO}}$ of [5], also studied in [6]. These papers have shown that the $\sqrt{\text{LASSO}}$ estimator achieves the near-oracle rates of convergence in sparse, high-dimensional regression models over data distributions that extend significantly beyond normality.

Clearly ρ is a norm; in particular, it is nonnegative and convex. Thus, Condition (8) of Theorem 2.1 is satisfied, cf. [54, Remark A.1].

One set of distributions that belongs to a neighborhood of size δ based on the ρ-MSW metric is:

$$\begin{aligned} B_\delta^{\sqrt{\text{LASSO}}}(\mathbb{P}) := \Big\{ & \mathbb{Q} \in \mathcal{P}_2(\mathbb{R}^{d+1}) \mid \exists \text{ a coupling } \pi \in \Pi(\mathbb{Q}, \mathbb{P}) \text{ for which:} \\ & \mathbb{E}_\pi \left[\left| \widetilde{X}_j - X_j \right|^2 \right] \le \delta^2, \ \forall\, j = 1, \dots, d, \text{ and } \mathbb{E}_\pi \left[\left| \widetilde{Y} - Y \right|^2 \right] \le \delta^2, \\ & \text{where } ((\mathbf{X}, Y), (\widetilde{\mathbf{X}}, \widetilde{Y})) \sim \pi \Big\}. \end{aligned} \tag{13}$$

This set of distributions contains perturbations of covariates and outcomes that are small in 2-norm. We verify that $B_\delta^{\sqrt{\text{LASSO}}}(\mathbb{P}) \subseteq B_\delta(\mathbb{P})$, where $B_\delta(\mathbb{P})$ is the set of balls used in Theorem 2.1 and defined in (7).

To see this, notice $\mathbb{E}_\pi[|\widetilde{X}_j - X_j|^2] \le \delta^2$ for all $j = 1, \dots, d$ implies condition (7), i.e.

$$\mathbb{E}_\pi \left[\left| (\widetilde{Y} - Y) + (\mathbf{X} - \widetilde{\mathbf{X}})^\top \boldsymbol{\gamma} \right|^2 \right] \le \delta^2 \left(1 + \rho(\boldsymbol{\gamma})\right)^2.$$

Indeed, the triangle inequality implies that for any $\boldsymbol{\gamma} \in \mathbb{R}^d$ and any coupling $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$ consistent with (13), we have

$$\begin{aligned} \sqrt{\mathbb{E}_\pi \left[\left| (\widetilde{Y} - Y) + (\mathbf{X} - \widetilde{\mathbf{X}})^\top \boldsymbol{\gamma} \right|^2 \right]} &\le \sqrt{\mathbb{E}_\pi \left[\left| \widetilde{Y} - Y \right|^2 \right]} + \sum_{j=1}^{d} |\gamma_j| \sqrt{\mathbb{E}_\pi \left[\left(\widetilde{X}_j - X_j \right)^2 \right]} \\ &\le \delta + \delta \sum_{j=1}^{d} |\gamma_j| = \delta(1 + \rho(\boldsymbol{\gamma})). \end{aligned}$$

Consequently, $B_\delta^{\sqrt{\text{LASSO}}}(\mathbb{P}) \subseteq B_\delta(\mathbb{P})$. We note that the other direction, namely, $B_\delta(\mathbb{P}) \subseteq B_\delta^{\sqrt{\text{LASSO}}}(\mathbb{P})$ does not hold in general.

It is worth mentioning that the set $B_\delta^{\sqrt{\text{LASSO}}}(\mathbb{P})$ contains different versions of $(\mathbf{X}, Y)$ measured with error. For example, any additive measurement error model of the form $\widetilde{X}_j = X_j + u_j$ and $\widetilde{Y} = Y + v$, where $\mathbb{E}[u_j^2] \le \delta^2$ and $\mathbb{E}[v^2] \le \delta^2$. Also, $B_\delta^{\sqrt{\text{LASSO}}}(\mathbb{P})$ contains multiplicative errors-in-variables models where $\widetilde{X}_j = X_j u_j$, and $\widetilde{Y} = Y v$, with u's independent of $(\mathbf{X}, Y)$, having mean equal to one, $\mathbb{E}_\mathbb{P}[X_j^2]\,\mathbb{E}[(u_j - 1)^2] \le \delta^2$, and independent of v having mean equal to one and $\mathbb{E}_\mathbb{P}[Y^2]\,\mathbb{E}[(v-1)^2] \le \delta^2$.

It is well known that the conjugate of ρ is

$$\rho^*(\boldsymbol{\beta}) = \begin{cases} 0 & \max\{|\beta_1|, \dots, |\beta_d|\} \le 1, \\ \infty & \text{otherwise.} \end{cases}$$

The argument is analogous to [54, Remark A.1]. Moreover, some algebra shows that $\boldsymbol{\beta}^* = (\text{sign}(\beta_1), \dots, \text{sign}(\beta_d))^\top$, is a subgradient of ρ at $\boldsymbol{\beta}$. Using these facts, we can determine the worst-case distribution for each particular $\boldsymbol{\beta}$. Indeed, Corollary 2.1 states that: $\widetilde{\mathbf{X}} = \mathbf{X} - e\,(\text{sign}(\beta_1), \dots, \text{sign}(\beta_d))^\top$, and $\widetilde{Y} = Y + e$, where

$$e := \frac{\delta\left(Y - \mathbf{X}^\top \boldsymbol{\beta}\right)}{\sqrt{\mathbb{E}_{\mathbb{P}}\left[\left(Y - \mathbf{X}^\top \boldsymbol{\beta}\right)^2\right]}}, \qquad (\mathbf{X}, Y) \sim \mathbb{P}.$$

The worst-case mean-squared error of $\sqrt{\text{LASSO}}$ is attained at distributions where there is a (possibly correlated) measurement error that has a factor structure. Note that the worst-case distribution is an element of (13).

2.3.2. *Square-root SLOPE.* Now suppose again that $r = 2$, but let

$$\rho(\boldsymbol{\beta}) = \sum_{j=1}^{d} \lambda_j |\boldsymbol{\beta}|_{(j)},$$

where $\lambda_1 \ge \cdots \ge \lambda_d \ge 0$ and $|\boldsymbol{\beta}|_{(j)}$ are the decreasing order statistics of the absolute values of the coordinates of $\boldsymbol{\beta}$. Under this penalty function—which is nonnegative—the penalized regression problem in (10) is the objective function of the square-root SLOPE of [70].

An equivalent definition for this penalty function is

$$\rho(\boldsymbol{\beta}) = \max_{pm} \sum_{j=1}^{d} \lambda_{pm(j)} |\beta_j|, \tag{14}$$

where we maximize over all permutations, pm, of the coordinates $\{1, \dots, d\}$. It follows that ρ is a norm, so Condition (8) of Theorem 2.1 is satisfied (see [54, Remark A.1]).

For a given $\boldsymbol{\beta} \in \mathbb{R}^d$, let pm^* be a permutation that solves (14). Define $\boldsymbol{\beta}^*$ by $\beta_j^* = \lambda_{pm^*(j)}\,\text{sign}(\beta_j)$. Algebra shows that $\rho(\boldsymbol{\beta}) = {\boldsymbol{\beta}^*}^\top \boldsymbol{\beta}$ and ${\boldsymbol{\beta}^*}^\top \boldsymbol{\gamma} \le \rho(\boldsymbol{\gamma})$, for any $\boldsymbol{\gamma} \in \mathbb{R}^d$. It follows that $\rho(\boldsymbol{\gamma}) \ge \rho(\boldsymbol{\beta}) + {\boldsymbol{\beta}^*}^\top \boldsymbol{\gamma} - {\boldsymbol{\beta}^*}^\top \boldsymbol{\beta}$, which implies that $\boldsymbol{\beta}^*$ is a subgradient of ρ at $\boldsymbol{\beta}$. Recall that $\rho^*(\boldsymbol{\beta}^*) = 0$; thus, (8) holds.

In this case, distributions belonging to balls of size δ based on the ρ-MSW metric are

$$\begin{aligned} B_\delta^{\text{SLOPE}}(\mathbb{P}) := \{ &\mathbb{Q} \in \mathcal{P}_2(\mathbb{R}^d) \,:\, \exists \text{ a coupling } \pi \in \Pi(\mathbb{Q}, \mathbb{P}) \text{ for which:} \\ &\mathbb{E}_\pi\left[\left|\widetilde{X}_{(j)} - X_{(j)}\right|^2\right] \le (\delta\lambda_j)^2,\ \forall\, j = 1, \dots, d, \\ &\text{and } \mathbb{E}_\pi\left[\left|\widetilde{Y} - Y\right|^2\right] \le \delta^2, \text{ where } ((\mathbf{X}, Y), (\widetilde{\mathbf{X}}, \widetilde{Y})) \sim \pi\}, \end{aligned}$$

where the decreasing order statistic is induced by the vector $\left(\mathbb{E}_\pi\left[\left|\widetilde{X}_j - X_j\right|^2\right]\right)_{j=1,\dots,d}$. As for the $\sqrt{\text{LASSO}}$, we check that $B_\delta^{\text{SLOPE}}(\mathbb{P}) \subseteq B_\delta(\mathbb{P})$. The triangle inequality implies that

for any coupling $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$:

$$\begin{aligned}\sqrt{\mathbb{E}_\pi\left[\left|(\widetilde{Y}-Y)+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\gamma}\right|^2\right]} &\le \sqrt{\mathbb{E}_\pi\left[\left|\widetilde{Y}-Y\right|^2\right]}+\sum_{j=1}^d|\gamma_j|\sqrt{\mathbb{E}_\pi\left[\left|\widetilde{X}_j-X_j\right|^2\right]}\\ &= \sqrt{\mathbb{E}_\pi\left[\left|\widetilde{Y}-Y\right|^2\right]}+\sum_{j=1}^d|\gamma_{(j)}|\sqrt{\mathbb{E}_\pi\left[\left|\widetilde{X}_{(j)}-X_{(j)}\right|^2\right]}\\ &\le \delta\left(1+\rho(\boldsymbol{\gamma})\right),\end{aligned}$$

where the last equality follows by the definition of $B_\delta^{\text{SLOPE}}(\mathbb{P})$ and (14).

Finally, we report the worst-case distribution for each particular $\boldsymbol{\beta}$. Corollary 2.1 shows that $\widetilde{\mathbf{X}} = \mathbf{X} - e\boldsymbol{\beta}^*$ and $\widetilde{Y} = Y + e$, where the j-coordinate of $\boldsymbol{\beta}^*$ is $\lambda_{pm^*(j)}\,\text{sign}(\beta_j)$ and

$$e := \frac{\delta\left(Y-\mathbf{X}^\top\boldsymbol{\beta}\right)}{\sqrt{\mathbb{E}_{\mathbb{P}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^2\right]}}, \quad (\mathbf{X},Y)\sim\mathbb{P}.$$

Note that the worst-case distribution is an element of $B_\delta^{\text{SLOPE}}(\mathbb{P})$.

3. Finite sample guarantees for the ρ-MSW-distance. Throughout this section we assume that the data $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ consists of i.i.d. draws from a true distribution $\mathbb{P}$. We denote the empirical distribution based on the available data by $\mathbb{P}_n$. Furthermore we assume that ρ satisfies

$$c_d\|(\boldsymbol{\gamma},-1)\| \le \rho(\boldsymbol{\gamma})+1, \qquad \forall\boldsymbol{\gamma}\in\mathbb{R}^d \tag{15}$$

for a constant $c_d > 0$ and a norm $\|\cdot\|$ on $\mathbb{R}^{d+1}$. E.g. for $\rho(\cdot) = \|\cdot\|_1$, (15) is satisfied with $c_d = 1$.

This section provides explicit upper bounds on the radius δ of the ball $B_\delta(\mathbb{P}_n)$ defined in (7), to guarantee that the true (and unknown) distribution, $\mathbb{P}$, belongs to the ball $B_\delta(\mathbb{P}_n)$ with a pre-specified probability. Our derivations are valid for any finite sample, which means that they hold regardless of the dimension of the covariates d, the sample size n, and the true distribution $\mathbb{P}$.

Recall from (4) that

$$\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) = \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d}\inf_{\pi\in\Pi(\mathbb{P},\widetilde{\mathbb{P}})}\frac{1}{1+\rho(\boldsymbol{\gamma})}\left(\mathbb{E}_\pi\left[\left|(\widetilde{Y}-Y)+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\gamma}\right|^r\right]\right)^{1/r}.$$

Note that we can rewrite the equation above in terms of the one-dimensional Wasserstein metric:

$$\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) = \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d}\frac{1}{1+\rho(\boldsymbol{\gamma})}\mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top\bar{\boldsymbol{\gamma}}\right]_*\mathbb{P},\left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top\bar{\boldsymbol{\gamma}}\right]_*\widetilde{\mathbb{P}}\right), \tag{16}$$

where $\bar{\boldsymbol{\gamma}}^\top = (\boldsymbol{\gamma}^\top, -1)$ and $f_*\mathbb{P}$ denotes the pushforward measure of $\mathbb{P}$ with respect to a map $f:\mathbb{R}^{d+1}\to\mathbb{R}$ and $\mathcal{W}_r$.

Using (15) we derive the following upper bound for $\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}})$:

$$\begin{aligned}\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) &= \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d}\frac{\|\bar{\boldsymbol{\gamma}}\|}{1+\rho(\boldsymbol{\gamma})}\frac{1}{\|\bar{\boldsymbol{\gamma}}\|}\mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top\bar{\boldsymbol{\gamma}}\right]_*\mathbb{P},\left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top\bar{\boldsymbol{\gamma}}\right]_*\widetilde{\mathbb{P}}\right)\\ &\le c_{\rho,d}\left(\sup_{\widetilde{\boldsymbol{\gamma}}:\|\widetilde{\boldsymbol{\gamma}}\|=1}\mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top\widetilde{\boldsymbol{\gamma}}\right]_*\mathbb{P},\left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top\widetilde{\boldsymbol{\gamma}}\right]_*\widetilde{\mathbb{P}}\right)\right) =: c_{\rho,d}\,\overline{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}),\end{aligned} \tag{17}$$

where $c_{\rho,d} := \max\{1/c_d, 1\}$.

The quantity $\overline{\mathcal{W}}_r$ defined in (17) is the max-sliced Wasserstein (MSW) distance on $(\mathbb{R}^{d+1}, \|\cdot\|)$. It is a special case of the Projection Robust Wasserstein (PRW) distance, also called the Wasserstein Projection Pursuit (WPP), see [59, Definition 1]. The work in [59, Proposition 1] shows that $\overline{\mathcal{W}}_r(\mathbb{P}, \widetilde{\mathbb{P}})$ is a metric (the proof is stated for the case $r = 2$, but carries over line by line to arbitrary $r \geq 1$).

As stated in the Introduction, it is well known that, in the worst case, $\mathcal{W}_r(\mathbb{P}_n, \mathbb{P}) \approx n^{-1/(d+1)}$. In what follows, we show that the MSW distance $\overline{\mathcal{W}}_r$ does not have this limitation. To show this, we first make a few notational simplifications. We write $\mathbb{P}_{\boldsymbol{\gamma}}$ and $F_{\boldsymbol{\gamma}}$, respectively, for the distribution and cdf of the scalar $(\mathbf{X}, Y)^\top \boldsymbol{\gamma}$ under $\mathbb{P}$. Similarly, we write $\mathbb{P}_{\boldsymbol{\gamma},n}$ and $F_{\boldsymbol{\gamma},n}$, respectively, for the probability measure and cdf of $(\mathbf{X}, Y)^\top \boldsymbol{\gamma}$ under $\mathbb{P}_n$. Note that, by (17) we have $\overline{\mathcal{W}}_r(\mathbb{P}, \widetilde{\mathbb{P}}) = \sup_{\|\boldsymbol{\gamma}\|=1} \mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma}}, \widetilde{\mathbb{P}}_{\boldsymbol{\gamma}})$.

We now provide explicit upper bounds for $\overline{\mathcal{W}}_r(\mathbb{P}, \mathbb{P}_n)$. By equation (17), for any δ we have

$$\mathbb{P}\left(\widehat{\mathcal{W}}_r(\mathbb{P}, \mathbb{P}_n) \leq c_{\rho,d} \cdot \delta\right) \geq \mathbb{P}\left(\overline{\mathcal{W}}_r(\mathbb{P}, \mathbb{P}_n) \leq \delta\right). \tag{18}$$

This means that probabilistic statements about $\overline{\mathcal{W}}_r(\mathbb{P}, \mathbb{P}_n)$ translate immediately to the ρ-MSW metric. For simplicity in the exposition, we first cover compactly supported measures $\mathbb{P}$ in Section 3.1 and then the general case in Section 3.2.

3.1. *The compactly supported case.*

THEOREM 3.1. *Let $\mathbb{P}$ have compact support. With probability at least $1 - \alpha$,*

$$\overline{\mathcal{W}}_r(\mathbb{P}_n, \mathbb{P})^r \leq \frac{C}{\sqrt{n}},$$

where

$$C := \left(180\sqrt{d+2} + \sqrt{2\log\left(\frac{1}{\alpha}\right)}\right)\operatorname{diam}\left(\operatorname{supp}(\mathbb{P})\right)^r, \tag{19}$$

and $\operatorname{diam}\left(\operatorname{supp}(\mathbb{P})\right) = \sup\{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_* : \mathbf{x}, \widetilde{\mathbf{x}} \in \operatorname{supp}(\mathbb{P})\}$ *is the diameter of the support of* $\mathbb{P}$ *measured with respect to the dual norm* $\|\mathbf{x}\|_* := \sup_{\mathbf{y}:\|\mathbf{y}\|=1} \mathbf{x}^\top \mathbf{y}$.

3.2. *The general case.*

We now consider a more general set-up where $\mathbb{P}$ is an arbitrary random variable that satisfies a mild moment condition; namely,

$$\Gamma := \mathbb{E}_{\mathbb{P}}\left[\|(\mathbf{X}, Y)\|_*^s\right] < \infty, \qquad \text{for some } s > 2r. \tag{20}$$

Our result generalizes the work of [57] and [51], who provide rates for $\overline{\mathcal{W}}_r(\mathbb{P}_n, \mathbb{P})$ assuming certain transport or Poincaré inequalities: we give similar rate statements with fully explicit constants under assumption (20), that is easy to verify in practice.

Our main result in this section is the following:

THEOREM 3.2. *Assume $s > 2r$ and $\Gamma < \infty$. Then, with probability greater than $1 - 3\alpha$,*

$$\overline{\mathcal{W}}_r(\mathbb{P}_n, \mathbb{P})^r \leq \frac{C\log(2n+1)^{r/s}}{\sqrt{n}}, \tag{21}$$

where

$$C := 2^r r\left(180\sqrt{d+2} + \sqrt{2\log\left(\frac{1}{\alpha}\right)} + \sqrt{\frac{\Gamma}{\alpha}}\left(\frac{8}{s/2 - r}\right)\sqrt{\log\left(\frac{8}{\alpha}\right) + (d+2)}\right). \tag{22}$$

4. Asymptotics for ρ-MSW-distance. We now provide asymptotic upper bounds for the ρ-MSW distance between the true and empirical measure. For this, it is sufficient to prove the corresponding bounds for $\overline{\mathcal{W}}_r(\mathbb{P},\mathbb{P}_n)$, as explained in (17) and (18). The following theorem provides a Donsker type result, i.e. asymptotic $(1/\sqrt{n})$-rates without logarithmic factors, as well as an inequality for the expectation without an explicit constant. One can then obtain concentration results similarly to [51, Theorem 3.7, 3.8] if a Bernstein tail condition or Poincare inequality is satisfied. As before, we relegate the proofs of these results to the Supplementary Material [54].

We first consider probability measures $\mathbb{P}$ with compact support.

THEOREM 4.1. *If $\mathbb{P}$ is compactly supported, then*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sqrt{n}\,\overline{\mathcal{W}}_r\left(\mathbb{P}_n,\mathbb{P}\right)^r \geq x\right) \leq \mathbb{P}\Big(\sup_{t\in[0,1]} |B(t)| \geq \frac{x}{c}\Big),$$

where $c = \operatorname{diam}(\operatorname{supp}(\mathbb{P}))^r$ and $(B(t))_{t\in[0,1]}$ is a standard Brownian bridge.

We now state the general result:

THEOREM 4.2. *Assume $\Gamma = \mathbb{E}_{\mathbb{P}}\left[\|(\mathbf{X},Y)\|_*^s\right] < \infty$, for some $s > 2r$, and define*

$$\mathcal{H}^+ := \left\{ |t|^s \mathbb{1}_{\{t\leq \boldsymbol{x}^\top\boldsymbol{\gamma}\}} :\ (\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times[0,\infty),\ \|\boldsymbol{\gamma}\|=1 \right\},$$

$$\mathcal{H}^0 := \left\{ \mathbb{1}_{\{\boldsymbol{x}^\top\boldsymbol{\gamma}\leq t\}} :\ (\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times\mathbb{R},\ \|\boldsymbol{\gamma}\|=1 \right\},$$

$$\mathcal{H}^- := \left\{ |t|^s \mathbb{1}_{\{t> \boldsymbol{x}^\top\boldsymbol{\gamma}\}} :\ (\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times(-\infty,0),\ \|\boldsymbol{\gamma}\|=1 \right\}.$$

Then there exists a constant $C := C(r,s,d)$, such that for all $t \geq 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left(\sqrt{n}\,\overline{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{P})^r \geq t\right) \leq \mathbb{P}\Big(\sup_{f\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-} |G_f| \geq \frac{t}{C\sqrt{\Gamma}}\Big),$$

where $(G_f)_{f\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-}$ is a zero-mean Gaussian process with covariance

$$\mathbb{E}\left[G_{f_1}G_{f_2}\right] = \mathbb{E}_{\mathbb{P}}\left[f_1f_2\right] - \mathbb{E}_{\mathbb{P}}\left[f_1\right]\mathbb{E}_{\mathbb{P}}\left[f_2\right] \qquad \forall f_1,f_2\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-. \tag{23}$$

Furthermore, for all $n\in\mathbb{N}$, we have $\mathbb{E}_{\mathbb{P}}\left[\sqrt{n}\,\overline{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{P})^r\right] \leq C\sqrt{\Gamma}$.

5. Recommendation to select the regularization parameter $\delta_{n,r}$.

5.1. *Recommendation based on finite sample bounds.* Our statistical analysis in Section 3 provides a concrete *oracle* recommendation to select the regularization parameter $\delta_{n,r}$ in (1). Our choice is based on Theorem 3.2 and guarantees that the true data generating process is contained in the ball $B_{\delta_{n,r}}(\mathbb{P}_n)$ with high probability:

$$\delta_{n,r} = \max\left\{\frac{1}{c_d},1\right\}\left[\frac{C\log(2n+1)^{r/s}}{\sqrt{n}}\right]^{1/r}, \tag{24}$$

where c_d is the constant such that $c_d\|(\boldsymbol{\gamma},-1)\| \leq \rho(\boldsymbol{\gamma})+1$ for all $\boldsymbol{\gamma}$ and C is the constant defined in (22).

If the support of $\mathbb{P}$ is compact, we can specialize our recommendation to select the regularization parameter $\delta_{n,r}$ with guidance from Theorem 3.1. This recommendation is

$$\delta_{n,r} = \max\left\{\frac{1}{c_d},1\right\}\left[\frac{C}{\sqrt{n}}\right]^{1/r}, \tag{25}$$

where C is now the constant defined in (19). In the case of compact support, our recommended regularization parameter only depends on $\mathbb{P}$ through the diameter of its support.[4]

Theorem 5.1 below shows that the objective function of the penalized regression in (1) constitutes—up to some adjustment terms—an upper bound for the expected prediction error at $\mathbb{Q}$ (provided it is close to $\mathbb{P}$).

THEOREM 5.1. *Suppose the conditions of Theorem 3.2 (or Theorem 3.1) hold. Consider $\delta_{n,r}$ defined in (24) (or (25)). Then, for any $\epsilon \geq 0$ and $\mathbb{Q}$ such that $\widehat{\mathcal{W}}_r(\mathbb{P}, \mathbb{Q}) \leq \epsilon$, with probability greater than $1 - 3\alpha$, we have for all $\boldsymbol{\beta}$ that*

$$\mathbb{E}_{\mathbb{Q}}\left[\left|Y - \mathbf{X}^\top \boldsymbol{\beta}\right|^r\right]^{1/r} \leq \mathbb{E}_{\mathbb{P}_n}\left[\left|Y - \mathbf{X}^\top \boldsymbol{\beta}\right|^r\right]^{1/r} + (\delta_{n,r} + \epsilon)\,(1 + \rho(\boldsymbol{\beta}))\,.$$

The result implies that linear predictors that solve (1) and use our recommended parameters $\delta_{n,r}$ have good out-of-sample performance at the true, unknown distribution of the data $\mathbb{P}$, and also at *testing* distributions $\mathbb{Q}$ that are close to $\mathbb{P}$ in the ρ-MSW metric.

Importantly, Theorem 5.1 (and Corollary C.1 in [54, Section C.5]) differs from other existing generalization bounds (such as Proposition 6 and Theorem 4 in [10]) in that i) our results are explicit about the possibility of a difference between the testing distribution ($\mathbb{Q}$) and the distribution that generated the training data ($\mathbb{P}$); ii) our results are derived without making reference to an underlying linear regression model for the training data; iii) the generalization error is evaluated at different values of $\boldsymbol{\beta}$—and in particular, it can be evaluated at the solution $\widehat{\boldsymbol{\beta}}$ of (1)—as opposed to the true parameter of a linear regression model; finally, iv) our results allow for a very general class of convex penalties and not only norms (although we focus on a special type of loss functions). We refer to [54, Sections C.4 and C.5] for a more extensive comparison to [10].

The oracle recommendation for the regularization parameter $\delta_{n,r}$ is typically not feasible as it depends on the unknown parameter Γ. The next section presents a normalization strategy on the covariates such that Theorem 5.1 holds with $\Gamma = 2^s$.

An interesting avenue for future work is to use the results in [56] (which assume log-concavity of the joint distribution of covariates and outcomes) to recommend a regularization parameter roughly of order: $||\Sigma||_{op}^{1/2}\sqrt{d\log(n)}/n^{1/r}$, where $r \geq 2$ and $||\cdot||_{op}$ is the operator norm of the covariance matrix of (X, Y). To do this, it would be necessary to recover the implicit constants that appear in [56, Theorem 1] (which only depend on r), and additionally provide some results for the consistent estimation of the operator norm of Σ. Because the rates in [56] are faster than ours (when d is fixed, their rates are of order $n^{-1/r}$ whereas ours are of order $n^{-1/(2r)}$), the regularization parameters based on the results of [56] will typically be smaller (making it less likely that the robust predictors ignore the available covariates). However, we remark that even for the Wasserstein metric on the real line, the rates of order $n^{-1/(2r)}$ cannot be improved upon, unless one imposes further structure on the true data generating process; see Theorem 7.11 and Corollary 7.12 in [14], and the discussion therein.

We note that our approach for choosing the regularization parameter, δ, is explicitly designed to guarantee the bound on out-of-distribution prediction error presented in Theorem 5.1. As we have explained before, a sufficient condition to obtain such a bound is to ensure that the true distribution, $\mathbb{P}$, belongs to the ball $B_{\delta_{n,r}}(\mathbb{P}_n)$ with probability at least $1 - \alpha$. Thus, Theorem 5.1 is possible thanks to the statistical analysis of the ρ-MSW metric. [10]

[4]Let us remark that estimating the support of a distribution is an intricate statistical question, going back at least to [32]. We refer to [9, 26, 81] for some recent results in support estimation. We also remark that in some applications (e.g. for discrete distributions arising in surveys) it is plausible that supp($\mathbb{P}$) is known and thus it need not be estimated.

acknowledges that a similar strategy for selecting δ using concentration inequalities for the standard Wasserstein metric would yield a recommendation of order $O(n^{-1/d})$; see their discussion after Theorem 4, p. 848. However, it is important to mention that there are other possibilities for choosing δ that do not necessarily target generalization error. For instance, if we followed the objective described in Section 1.1.2 of [10] (which the authors describe as covering the true parameter of a linear regression model with probability at least $1-\alpha$), it would be possible to recommend values for the regularization parameter of order $O(n^{-1/2})$. In particular, for the $\sqrt{\text{LASSO}}$ the authors recommend a tuning parameter equal to

$$\lambda = \frac{\pi}{\pi - 2} \frac{\Phi^{-1}(1-\alpha/2d)}{\sqrt{n}},$$

which, up to a constant, coincides with the recommendation in [5]. In Section C.5 of the Supplementary Material [54], we show that if we adopt the objective of [10] (and their assumptions), but use our DRO representation based on the ρ-MSW metric, we could recommend the same or even a smaller regularization parameter.

5.2. *Covariate Normalization.* Our statistical analysis in Section 5.1 provides a concrete *oracle* recommendation to select the regularization parameter $\delta_{n,r}$ in (1). The oracle recommendation for $\delta_{n,r}$ is typically not feasible as it depends on the unknown parameter Γ. In this section, we present a simple strategy to normalize the sample covariates that guarantees a modified version of Theorem 5.1. This means that we can turn our oracle recommendation into a simple formula that only depends on the true and unknown distribution of the data through the sth moment of the outcome (which is typically easy to estimate), while also guaranteeing robustness to perturbations of the test dataset distribution from that of the training data in the form of (3) below. This *pivotality* was the original motivation for the use of the $\sqrt{\text{LASSO}}$ and related estimators.

For this, we assume that the covariates in the data have been *normalized* to satisfy $\mathbb{E}_{\mathbb{P}_n}\left[\|(\mathbf{X},0)\|_*^s\right] = 1$. It is common practice to impose some covariate normalization to estimate the parameters of the best linear predictor using the $\sqrt{\text{LASSO}}$ and related estimators; see [5, Equation 4 p.2] for an example of a coordinate-wise, unit variance normalization.

The next theorem proposes a simple formula to select the regularization parameter $\delta_{n,r}$, that does not depend on Γ, under our suggested normalization.

THEOREM 5.2. *Suppose* $\mathbb{E}_{\mathbb{P}}\left[\|(\mathbf{X},0)\|_*^s\right] = 1$ *and* $\mathbb{E}_{\mathbb{P}}[\|(0,\ldots,0,Y)\|_*^s]^{1/s} < \infty$ *for some* $s > 2r$. *In addition, set* $\sigma = \max\{\mathbb{E}_{\mathbb{P}}[\|(0,\ldots,0,Y)\|_*^s]^{1/s},\ 1\}$ *and assume that* (15) *holds with* $c_d > 0$. *Define the* (ρ,σ)-MSW

$$\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P},\widetilde{\mathbb{P}}) := \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d} \left(\inf_{\pi\in\Pi(\mathbb{P},\widetilde{\mathbb{P}})} \frac{1}{\sigma + \rho(\boldsymbol{\gamma})} \left(\mathbb{E}_\pi\left[\left| (Y - \mathbf{X}^\top\boldsymbol{\gamma}) - (\widetilde{Y} - \widetilde{\mathbf{X}}^\top\boldsymbol{\gamma}) \right|^r \right] \right)^{1/r} \right). \tag{26}$$

and consider

$$\delta_{n,r} := \max\left\{ \frac{1}{c_d}, 1 \right\} \left[\frac{C\,\log(2n+1)^{r/s})}{\sqrt{n}} \right]^{1/r}, \tag{27}$$

where

$$C := 2^r r \left(180\sqrt{d+2} + \sqrt{2\log\left(\frac{1}{\alpha}\right)} + \sqrt{\frac{2^s}{\alpha}} \left(\frac{8}{s/2 - r} \right) \sqrt{\log\left(\frac{8}{\alpha}\right) + (d+2)} \right).$$

Then, for any $\epsilon \geq 0$ *and* $\mathbb{Q}$ *such that* $\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P},\mathbb{Q}) \leq \epsilon$, *with probability greater than* $1 - 3\alpha$,

$$\mathbb{E}_{\mathbb{Q}}\left[\left| Y - \mathbf{X}^\top\boldsymbol{\beta} \right|^r \right]^{1/r} \leq \mathbb{E}_{\mathbb{P}_n}\left[\left| Y - \mathbf{X}^\top\boldsymbol{\beta} \right|^r \right]^{1/r} + (\delta_{n,r} + \epsilon)\left(\sigma + \rho(\boldsymbol{\beta})\right), \quad \forall \boldsymbol{\beta}\,.$$

We also reiterate that our recommendation for the selection of regularization parameter does not rely on any sparsity assumption. We think this is an important point, as recent work [38, 50] has argued that sparsity might not always be a compelling starting point in applications.

5.3. *Asymptotic recommendation.* For compactly supported measures, Theorem 4.1 yields the asymptotic *oracle* recommendation

$$\delta_{n,r} = c_{\rho,d} \left[n^{-1/2} \cdot q_{1-\alpha} \right]^{1/r} \cdot \operatorname{diam}\left(\operatorname{supp}(\mathbb{P})\right), \tag{28}$$

where $c_{\rho,d}$ is as in (17), $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the Kolmogorov distribution. In the general case, Theorem 4.2 yields $\delta_{n,r} = [n^{-1/2} \cdot \Gamma^{1/2} \cdot C]^{1/r}$, for some constant $C = C(r, s, d, \alpha)$. However, the constant C is hard to determine explicitly, since it depends on α through the quantile of the zero-mean Gaussian process $(G_f)_{f \in \mathcal{H}^+ \cup \mathcal{H}^0 \cup \mathcal{H}^-}$, whose covariance structure depends on $\mathbb{P}$ and is given in (23) and is hard to bound explicitly. To the best of our knowledge, a characterization of the exact asymptotic distribution of $\widehat{\mathcal{W}}_r(\mathbb{P}, \mathbb{P}_n)$—or even of its upper bound $\overline{\mathcal{W}}_r(\mathbb{P}, \mathbb{P}_n)$—is still an open research problem. While we conjecture that the upper bounds we provide (both asymptotically and in finite samples) can be tightened, we remark that even for the Wasserstein metric on the real line, the rates of order $n^{-1/(2r)}$ that we obtain cannot be improved upon, unless one imposes further structure on the true data generating process. See Theorem 7.11 and Corollary 7.12 in [14], and the discussion therein.

5.4. *Application: ranking of estimators.* Consider two estimators $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1(\mathbb{P}_n)$ and $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2(\mathbb{P}_n)$, where $\mathbb{P}_n$ denotes the empirical distribution of i.i.d. draws from a true distribution $\mathbb{P}$. In this section, we investigate whether $\boldsymbol{\beta}_1$ has a better out-of-distribution performance than $\boldsymbol{\beta}_2$ over an uncertainty set B. That is,

$$\sup_{\mathbb{Q} \in B} \mathbb{E}_{\mathbb{Q}} \left[\left| Y - \mathbf{X}^\top \boldsymbol{\beta}_1 \right|^r \right]^{1/r} \leq \sup_{\mathbb{Q} \in B} \mathbb{E}_{\mathbb{Q}} \left[\left| Y - \mathbf{X}^\top \boldsymbol{\beta}_2 \right|^r \right]^{1/r}. \tag{29}$$

We restrict our attention to uncertainty sets B that verify two conditions:

(i) $B \subseteq B_\delta(\mathbb{P})$ for some δ, and ρ.
(ii) The supremum on the left side of (29) is achieved for $\mathbb{P}^*_{\boldsymbol{\beta}_1}$, and the supremum on the right side of (29) is achieved for $\mathbb{P}^*_{\boldsymbol{\beta}_2}$, where $\mathbb{P}^*_{\boldsymbol{\beta}_j}$ are defined according to Corollary 2.1 for $j = 1, 2$.

Examples of such sets B are given in Section 2.3. Note that we cannot evaluate (29) directly, as $\mathbb{P}$ is not observed. Instead, we propose the test statistic

$$T_n = n^{1/(2r)} \left(\frac{\mathbb{E}_{\mathbb{P}_n} \left[\left| Y - \mathbf{X}^\top \boldsymbol{\beta}_1 \right|^r \right]^{1/r} - \mathbb{E}_{\mathbb{P}_n} \left[\left| Y - \mathbf{X}^\top \boldsymbol{\beta}_2 \right|^r \right]^{1/r} + \delta\rho(\boldsymbol{\beta}_1) - \delta\rho(\boldsymbol{\beta}_2)}{2 + \rho(\boldsymbol{\beta}_1) + \rho(\boldsymbol{\beta}_2)} \right).$$

Corollary 5.1 states that T_n gives rise to a size–α test. For notational simplicity, we focus on compactly supported probability measures $\mathbb{P}$ and remark that the same reasoning can be used to derived tests for general $\mathbb{P}$ satisfying the assumptions of Theorems 3.2 and 4.2.

COROLLARY 5.1. *In the setting of Theorems 2.1 and 3.1, consider C and $c_{\rho,d}$ defined in (19) and (17). Then, for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ satisfying (29), we have $P(T_n > c_{\rho,d} C^{1/r}) \leq \alpha$.*

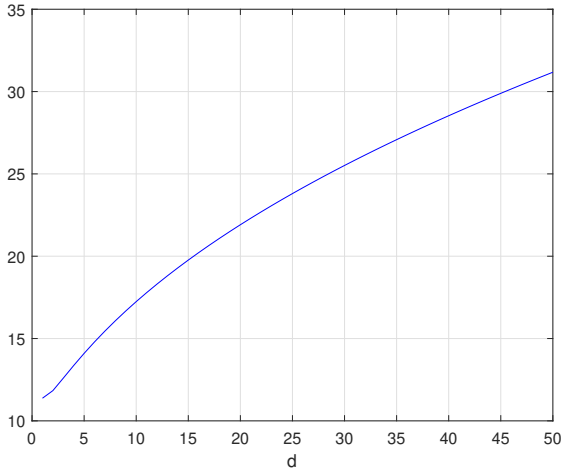


Fig 2: Ratio of the regularization parameter in (28) to that one in (30) for $\lambda = 10, \alpha = 0.05, n = 2,500$, $\boldsymbol{\beta} = [1, 0, \dots 0]^\top$, and $\sigma_\varepsilon = 1$.

6. Simulations. Suppose that the training data consists of n i.i.d. draws from a linear regression model, meaning $Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \sigma_\varepsilon \varepsilon_i$. We take ε_i to be uniformly distributed over the interval $[-1, 1]$. The vector of covariates, $\mathbf{X}_i \in \mathbb{R}^d$, is generated as $\mathbf{X}_i = \sigma_\varepsilon \lambda \widetilde{\mathbf{X}}_i$, where $\widetilde{\mathbf{X}}_i$ is a d-dimensional vector of independent uniform random variables over the $[0, 1]$ interval, independently of ε_i. The parameters controlling the simulation design are $(\boldsymbol{\beta}, \sigma_\varepsilon, \lambda, d)$.

We first focus on linear prediction using coefficients estimated via the $\sqrt{\text{LASSO}}$ ($r = 2$). Recall from (28) that our oracle recommendation for the tuning parameter δ_n is $n^{-1/4} \cdot (q_{1-\alpha})^{1/2} \cdot \text{diam}\left(\text{supp}(\mathbb{P})\right)$, where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the Kolmogorov distribution. Algebra shows (see Section C.1 of the Supplementary Material [54]) that

$$\text{diam}\left(\text{supp}(\mathbb{P})\right) = \sigma_\varepsilon \lambda \left(d + \left(\|\boldsymbol{\beta}\|_1 + (2/\lambda) \right)^2 \right)^{1/2}.$$

For comparison, the typical oracle recommendation δ_n^* for the $\sqrt{\text{LASSO}}$ based on [5], can be shown to equal

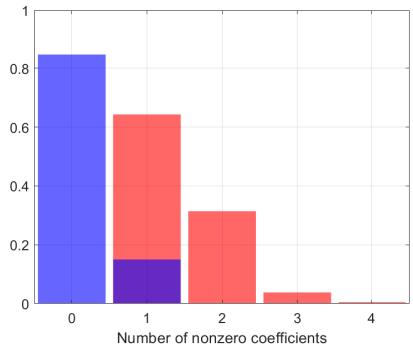
$$n^{-1/2} \cdot 3^{-1/2} \sigma_\varepsilon \lambda \cdot \Phi^{-1} \left(\frac{1}{2} + \frac{(1-\alpha)^{(1/d)}}{2} \right). \tag{30}$$

Figure 2 compares the ratio of (28) relative to (30). The figure shows that our recommendation can be more than ten times larger than the typical recommendations in the literature. Thus, one first concern is that the distributional robustness guaranteed by our choice of δ, as defined in (28), could be achieved by setting all the coefficients to zero (an adversarial nature cannot increase much the generalization error of such a predictor, as it does not rely at all on covariates). We also note that the recent work of [19] has shown that larger tuning parameters could lead to *incentive compatibility* in certain human-machine interactive environments.

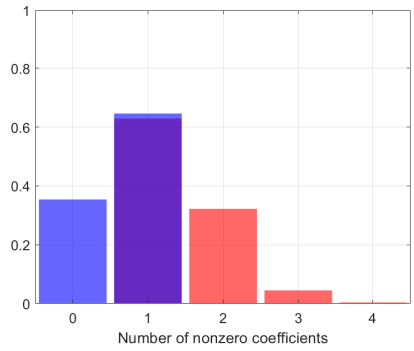
We now argue that in our simulation design it is possible to figure out the smallest sample size that would be required to avoid a "trivial" prediction. It is known, see [70], that $\boldsymbol{\beta} = 0_{d \times 1}$ is a solution to the $\sqrt{\text{LASSO}}$ problem if and only if

$$\frac{\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i y_i \|_\infty}{\sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2}} \leq \delta_{n,2}. \tag{31}$$

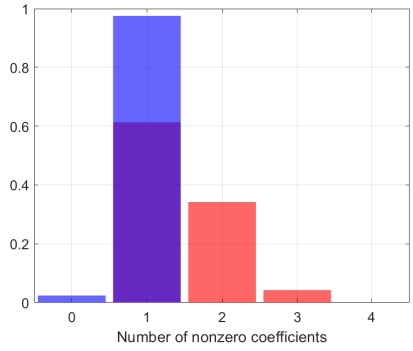
Using a Central Limit Theorem and a Law of Large of numbers, algebra shows (see Section C.2 in the Supplementary Material [54]) that (31) holds with high probability whenever



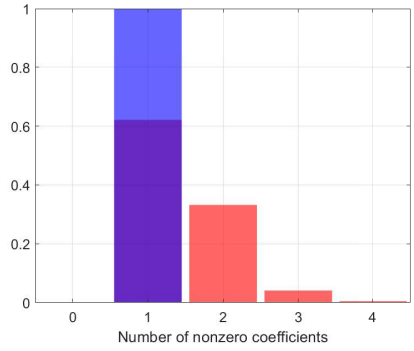
(a) $n = 2125$



(b) $n = 2250$



(c) $n = 2375$



(d) $n = 2500$

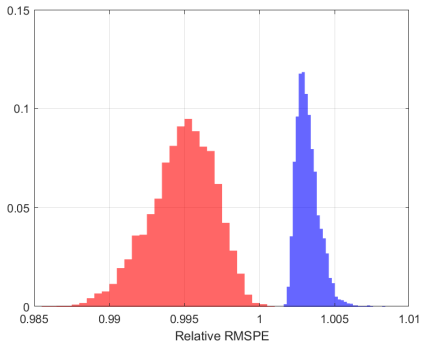
Fig 3: Fraction of simulation draws in which the $\sqrt{\text{LASSO}}$ selects 0,1,…, 4 nonzero coefficients using the regularization parameters defined in (28) (blue) and (30) (red), where $d = 10$ and $\boldsymbol{\beta} = [1, 0, \dots 0]^\top$.

$$n \leq 9 \cdot \left\| \frac{\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}} \right\|_\infty^{-4} \cdot (q_{1-\alpha})^2 \cdot \left(d + (\|\boldsymbol{\beta}\|_1 + (2/\lambda))^2 \right)^2 . \tag{32}$$

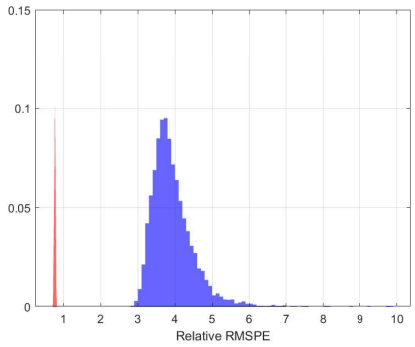
For $d = 10$, $\boldsymbol{\beta} = (1, 0, 0 \dots 0)^\top$, $\alpha = .05$ (or equivalently $q_{1-\alpha} = 1.358$) the corresponding conservative bound is of about 2,200. This means that it will take a relatively large sample size in order for our regularization parameter to select at least some covariates for prediction.

We verify this conjecture numerically. We simulate data using the $\sigma_\varepsilon = 1, \lambda = 10, d = 10, \boldsymbol{\beta} = [1, 0, \dots, 0]^\top$ and consider sample sizes $n \in \{2125, 2250, 2375, 2500\}$. Our design corresponds to a low-dimensional problem (10 covariates and at least 2,000 observations). Figure 3 reports the histogram associated to the number of nonzero coefficients selected by the $\sqrt{\text{LASSO}}$ using the regularization parameters in (28) and (30). The numerical results reported are in line with the bound derived in (32).

Training/Testing error. Figure 4 reports the training/testing root-mean squared prediction error (RMSPE) associated to the three estimators considered in our simulations: the OLS estimator, the $\sqrt{\text{LASSO}}$ with the δ_n^* as in (30), and the $\sqrt{\text{LASSO}}$ with the δ_n as in (28). The training data is generated according to the design described above for a sample size of $n = 2500$. For testing, we perturb the true data generating process according to the worst-case distribution derived in Corollary 2.1 with δ_n in (28) replacing δ. The plots report the histogram—across simulations—of the relative RMSPE in the training (or testing) data. For



(a) RMSPE ratio of the $\sqrt{\text{LASSO}}$ using (30) to the OLS estimator



(b) RMSPE ratio of the $\sqrt{\text{LASSO}}$ using (28) to the version using (30)

Fig 4: Histogram of Root Mean-Squared Prediction Error (RMSPE) ratios for two estimators using training data (blue) and adversarial testing data (red).

example, Panel a) of Figure 4 reports the RMSPE of the $\sqrt{\text{LASSO}}$, divided by the root MSPE of OLS, in both the training and testing data.

The simulation results are in line with the theoretical predictions. First, since we are considering a simulation design where n is large relative to d, the oracle δ_n^* in (30) is close to zero. This means that the predictions associated to the $\sqrt{\text{LASSO}}$ in the training sample are not very different to those obtained via OLS. Panel a) of Figure 4 indeed shows that the relative training error between the $\sqrt{\text{LASSO}}$ (with the typical δ_n^*) and OLS remains very close to one across simulations. Panel b) shows that that the difference between the regularization parameters in (30) and (28) generates a sizeable difference in training error. However, the larger value of the tuning parameter does translate to better out-of-distribution performance.

Finally, we verify the bound in Theorem 5.1. The corollary implies that with probability at least 95% the RMSPE of the $\sqrt{\text{LASSO}}$ in the *testing* set (for any distribution in the ball that is ϵ away from the true data generating process) must be bounded by the sum of i) the RMSPE of the $\sqrt{\text{LASSO}}$ in the *training* set and ii) $(\delta_n + \epsilon)(1 + \rho(\boldsymbol{\beta}))$. Figure 5 shows that the bound holds for the recommended δ_n, but not for the usual one. Additional simulations are reported in Section D of the Supplementary Material [54].

SUPPLEMENTARY MATERIAL

Proofs of theorems, additional derivations, and numerical simulations are provided in the Supplementary Material [54].

REFERENCES

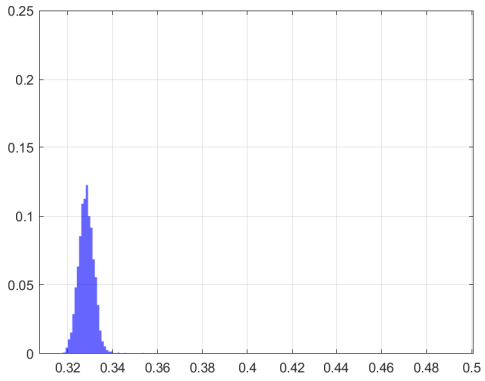
[1] ADJAHO, C. AND T. CHRISTENSEN (2022): "Externally Valid Treatment Choice," *arXiv preprint arXiv:2205.05561*.

[2] AGARWAL, D., L. LI, AND A. SMOLA (2011): "Linear-time estimators for propensity scores," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, 93–100.

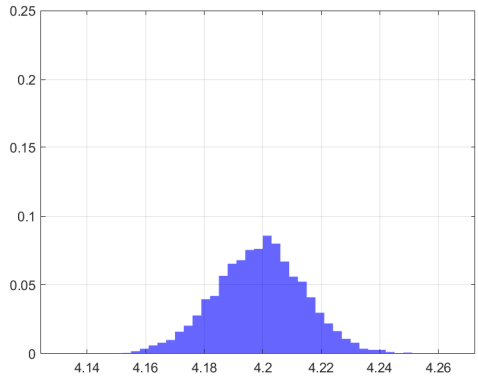
[3] ANDREWS, I., D. FUDENBERG, A. LIANG, AND C. WU (2022): "The Transfer Performance of Economic Models," *arXiv preprint arXiv:2202.04796*.

[4] BARTL, D. AND S. MENDELSON (2022): "Structure preservation via the Wasserstein distance," *arXiv preprint arXiv:2209.07058*.

[5] BELLONI, A., V. CHERNOZHUKOV, AND L. WANG (2011): "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, 98, 791–806.



(a) The $\sqrt{\text{LASSO}}$ uses (28)



(b) The $\sqrt{\text{LASSO}}$ uses (30)

Fig 5: Histogram of the ratio of the testing error of the $\sqrt{\text{LASSO}}$ to the upper bound on the out-of-distribution prediction error defined in (3).

[6] ——— (2014): "Pivotal estimation via square-root lasso in nonparametric regression," *The Annals of Statistics*, 42, 757–788.

[7] BEN-DAVID, S., J. BLITZER, K. CRAMMER, A. KULESZA, F. PEREIRA, AND J. W. VAUGHAN (2010): "A theory of learning from different domains," *Machine learning*, 79, 151–175.

[8] BERTSIMAS, D. AND M. S. COPENHAVER (2018): "Characterization of the equivalence of robustification and regularization in linear and matrix regression," *European Journal of Operational Research*, 270, 931–942.

[9] BIAU, G., B. CADRE, AND B. PELLETIER (2008): "Exact rates in density support estimation," *Journal of Multivariate Analysis*, 99, 2185–2207.

[10] BLANCHET, J., Y. KANG, AND K. MURTHY (2019): "Robust Wasserstein profile inference and applications to machine learning," *Journal of Applied Probability*, 56, 830–857.

[11] BLANCHET, J., Y. KANG, K. MURTHY, AND F. ZHANG (2019): "Data-driven optimal transport cost selection for distributionally robust optimization," in *2019 Winter simulation conference (WSC)*, IEEE, 3740–3751.

[12] BLANCHET, J., Y. KANG, J. L. M. OLEA, V. A. NGUYEN, AND X. ZHANG (2020): "Machine Learning's Dropout Training is Distributionally Robust Optimal," *arXiv preprint arXiv:2009.06111*.

[13] BLANCHET, J. AND K. MURTHY (2019): "Quantifying distributional model risk via optimal transport," *Mathematics of Operations Research*, 44, 565–600.

[14] BOBKOV, S. AND M. LEDOUX (2019): *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, vol. 261, American Mathematical Society.

[15] BOISSARD, E. AND T. LE GOUIC (2014): "On the mean speed of convergence of empirical and occupation measures in Wasserstein distance," in *Annales de l'IHP Probabilités et statistiques*, vol. 50, 539–563.

[16] BONNEEL, N., J. RABIN, G. PEYRÉ, AND H. PFISTER (2015): "Sliced and Radon Wasserstein barycenters of measures," *Journal of Mathematical Imaging and Vision*, 51, 22–45.

[17] BOYD, S. AND L. VANDENBERGHE (2004): *Convex optimization*, Cambridge university press.

[18] BUNEA, F., J. LEDERER, AND Y. SHE (2013): "The group square-root lasso: Theoretical properties and fast algorithms," *IEEE Transactions on Information Theory*, 60, 1313–1325.

[19] CANER, M. AND K. ELIAZ (2024): "Should Humans Lie to Machines? The Incentive Compatibility of Lasso and GLM Structured Sparsity Estimators," *Journal of Business & Economic Statistics*, 1–19.

[20] CHEN, X., M. MONFORT, A. LIU, AND B. D. ZIEBART (2016): "Robust covariate shift regression," in *Artificial Intelligence and Statistics*, PMLR, 1270–1279.

[21] CHERNOZHUKOV, V., D. CHETVERIKOV, K. KATO, AND Y. KOIKE (2023): "High-dimensional data bootstrap," *Annual Review of Statistics and Its Application*, 10, 427–449.

[22] CHETVERIKOV, D., Z. LIAO, AND V. CHERNOZHUKOV (2021): "On cross-validated lasso in high dimensions," *The Annals of Statistics*, 49, 1300–1317.

[23] CHIZAT, L., P. ROUSSILLON, F. LÉGER, F.-X. VIALARD, AND G. PEYRÉ (2020): "Faster Wasserstein distance estimation with the Sinkhorn divergence," *Advances in Neural Information Processing Systems*, 33, 2257–2269.

[24] CHRISTENSEN, T. AND B. CONNAULT (2023): "Counterfactual sensitivity and robustness," *Econometrica*, 91, 263–298.

[25] CHU, H. T., K.-C. TOH, AND Y. ZHANG (2022): "On regularized square-root regression problems: distributionally robust interpretation and fast computations," *Journal of Machine Learning Research*, 23, 1–39.

[26] CUEVAS, A. AND R. FRAIMAN (1997): "A plug-in approach to support estimation," *The Annals of Statistics*, 25, 2300 – 2312.

[27] DEREICH, S., M. SCHEUTZOW, AND R. SCHOTTSTEDT (2013): "Constructive quantization: Approximation by empirical measures," in *Annales de l'IHP Probabilités et statistiques*, vol. 49, 1183–1203.

[28] DESHPANDE, I., Y.-T. HU, R. SUN, A. PYRROS, N. SIDDIQUI, S. KOYEJO, Z. ZHAO, D. FORSYTH, AND A. G. SCHWING (2019): "Max-sliced Wasserstein distance and its use for GANs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10648–10656.

[29] DEVROYE, L. AND G. LUGOSI (2001): *Combinatorial methods in density estimation*, Springer Science & Business Media.

[30] DUCHI, J. C., P. W. GLYNN, AND H. NAMKOONG (2021): "Statistics of robust optimization: A generalized empirical likelihood approach," *Mathematics of Operations Research*.

[31] DUCHI, J. C. AND H. NAMKOONG (2021): "Learning models with uniform performance via distributionally robust optimization," *The Annals of Statistics*, 49, 1378–1406.

[32] FISHER, R. A., A. S. CORBET, AND C. B. WILLIAMS (1943): "The relation between the number of species and the number of individuals in a random sample of an animal population," *The Journal of Animal Ecology*, 42–58.

[33] FOURNIER, N. (2023): "Convergence of the empirical measure in expected Wasserstein distance: non-asymptotic explicit bounds in $\mathbb{R}^d$," *ESAIM: Probability and Statistics*, 27, 749–775.

[34] FOURNIER, N. AND A. GUILLIN (2015): "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, 162, 707–738.

[35] FRIEDMAN, J., T. HASTIE, AND R. TIBSHIRANI (2017): *The elements of statistical learning: data mining, inference and prediction*, vol. 1 of *Series in Statistics*, New York: Springer, second edition ed.

[36] GAO, R., X. CHEN, AND A. J. KLEYWEGT (2022): "Wasserstein distributionally robust optimization and variation regularization," *Operations Research*.

[37] GAO, R. AND A. KLEYWEGT (2023): "Distributionally robust stochastic optimization with Wasserstein distance," *Mathematics of Operations Research*, 48, 603–655.

[38] GIANNONE, D., M. LENZA, AND G. E. PRIMICERI (2021): "Economic Predictions With Big Data: The Illusion of Sparsity," *Econometrica*, 89, 2409–2437.

[39] GOLDFELD, Z., K. KATO, G. RIOUX, AND R. SADHU (2024): "Statistical inference with regularized optimal transport," *Information and Inference: A Journal of the IMA*, 13.

[40] GOODFELLOW, I., Y. BENGIO, AND A. COURVILLE (2016): *Deep Learning*, MIT Press.

[41] GOODFELLOW, I. J., J. SHLENS, AND C. SZEGEDY (2014): "Explaining and Harnessing Adversarial Examples," *CoRR*, abs/1412.6572.

[42] HASTIE, T., A. MONTANARI, S. ROSSET, AND R. J. TIBSHIRANI (2022): "Surprises in high-dimensional ridgeless least squares interpolation," *The Annals of Statistics*, 50, 949–986.

[43] KOLOURI, S., K. NADJAHI, U. SIMSEKLI, R. BADEAU, AND G. ROHDE (2019): "Generalized sliced Wasserstein distances," *Advances in neural information processing systems*, 32.

[44] KPOTUFE, S. AND G. MARTINET (2021): "Marginal singularity and the benefits of labels in covariate-shift," *The Annals of Statistics*, 49, 3299–3323.

[45] KUHN, D., P. M. ESFAHANI, V. A. NGUYEN, AND S. SHAFIEEZADEH-ABADEH (2019): "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations research & management science in the age of analytics*, Informs, 130–166.

[46] KURAKIN, A., I. GOODFELLOW, AND S. BENGIO (2016): "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*.

[47] LAM, H. (2016): "Robust sensitivity analysis for stochastic systems," *Mathematics of Operations Research*, 41, 1248–1275.

[48] LEE, J. AND M. RAGINSKY (2018): "Minimax statistical learning with Wasserstein distances," *Advances in Neural Information Processing Systems*, 31.

[49] LEI, J. (2020): "Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces," *Bernoulli*, 26, 767–798.

[50] LI, C. M. AND U. K. MÜLLER (2021): "Linear regression with many controls of limited explanatory power," *Quantitative Economics*, 12, 405–442.

[51] LIN, T., Z. ZHENG, E. CHEN, M. CUTURI, AND M. I. JORDAN (2021): "On projection robust optimal transport: Sample complexity and model misspecification," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 262–270.

[52] MANSOUR, Y., M. MOHRI, AND A. ROSTAMIZADEH (2009): "Domain Adaptation: Learning Bounds and Algorithms," in *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*, Montréal, Canada.

[53] MOHAJERIN ESFAHANI, P. AND D. KUHN (2018): "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, 171, 115–166.

[54] MONTIEL OLEA, J. L., C. RUSH, A. VELEZ, AND J. WIESEL (2024): "Supplement to "The out-of sample prediction error of the $\sqrt{\text{LASSO}}$ and related estimators"," .

[55] NGUYEN, V. A., S. S. ABADEH, D. FILIPOVIĆ, AND D. KUHN (2021): "Mean-covariance robust risk measurement," *Swiss Finance Institute Research Paper*, 21-93.

[56] NIETERT, S., Z. GOLDFELD, R. SADHU, AND K. KATO (2022): "Statistical, robustness, and computational guarantees for sliced Wasserstein distances," *Advances in Neural Information Processing Systems*, 35, 28179–28193.

[57] NILES-WEED, J. AND P. RIGOLLET (2022): "Estimation of Wasserstein distances in the spiked transport model," *Bernoulli*, 28, 2663–2688.

[58] PATY, F.-P. AND M. CUTURI (2019): "Subspace robust Wasserstein distances," in *International conference on machine learning*, PMLR, 5072–5081.

[59] ——— (2019): "Subspace robust Wasserstein distances," in *International conference on machine learning*, PMLR, 5072–5081.

[60] QUINONERO-CANDELA, J., M. SUGIYAMA, A. SCHWAIGHOFER, AND N. D. LAWRENCE (2008): *Dataset shift in machine learning*, Mit Press.

[61] RABIN, J., G. PEYRÉ, J. DELON, AND M. BERNOT (2011): "Wasserstein barycenter and its application to texture mixing," in *International Conference on Scale Space and Variational Methods in Computer Vision*, Springer, 435–446.

[62] REDDI, S., B. POCZOS, AND A. SMOLA (2015): "Doubly robust covariate shift correction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29.

[63] RÉNYI, A. (1961): "On measures of entropy and information," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Berkeley, California, USA, vol. 1.

[64] ROCKAFELLAR, R. T. (1997): *Convex analysis*, Princeton university press.

[65] SAHOO, R., L. LEI, AND S. WAGER (2022): "Learning from a biased sample," *arXiv preprint arXiv:2209.01754*.

[66] SHAFIEEZADEH ABADEH, S., P. M. MOHAJERIN ESFAHANI, AND D. KUHN (2015): "Distributionally robust logistic regression," *Advances in Neural Information Processing Systems*, 28.

[67] SHIMODAIRA, H. (2000): "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of statistical planning and inference*, 90, 227–244.

[68] SINGH, S. AND B. PÓCZOS (2018): "Minimax distribution estimation in Wasserstein distance," *arXiv preprint arXiv:1802.08855*.

[69] SINHA, A., H. NAMKOONG, AND J. C. DUCHI (2018): "Certifying Some Distributional Robustness with Principled Adversarial Training," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net.

[70] STUCKY, B. AND S. VAN DE GEER (2017): "Sharp oracle inequalities for square root regularization," *Journal of Machine Learning Research*, 18, 1–29.

[71] SUGIYAMA, M., M. KRAULEDAT, AND K.-R. MÜLLER (2007): "Covariate shift adaptation by importance weighted cross validation." *Journal of Machine Learning Research*, 8.

[72] SUGIYAMA, M. AND K.-R. MÜLLER (2005): "Input-dependent estimation of generalization error under covariate shift," *Statistics & Risk Modeling*, 23, 249–279.

[73] TIAN, X., J. R. LOFTUS, AND J. E. TAYLOR (2018): "Selective inference with unknown variance via the square-root lasso," *Biometrika*, 105, 755–768.

[74] VILLANI, C. (2008): *Optimal Transport: Old and New*, vol. 338, Springer.

[75] WAINWRIGHT, M. J. (2019): *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge University Press.

[76] WANG, H., G. LI, AND G. JIANG (2007): "Robust regression shrinkage and consistent variable selection through the LAD-Lasso," *Journal of Business & Economic Statistics*, 25, 347–355.

[77] WEED, J. AND F. BACH (2019): "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance," *Bernoulli*, 25, 2620–2648.

[78] WEN, J., C.-N. YU, AND R. GREINER (2014): "Robust learning under uncertain test distributions: Relating covariate shift to model misspecification," in *International Conference on Machine Learning*, PMLR, 631–639.

[79] WU, Q., J. LI, AND T. MAO (2022): "On Generalization and Regularization via Wasserstein Distributionally Robust Optimization," *SSRN Electronic Journal*.

[80] WU, Y. AND L. WANG (2020): "A survey of tuning parameter selection for high-dimensional regression," *Annual review of statistics and its application*, 7, 209–226.

[81] WU, Y. AND P. YANG (2019): "Chebyshev polynomials, moment matching, and optimal estimation of the unseen," *The Annals of Statistics*, 47, 857–883.

SUPPLEMENTARY MATERIAL TO "The distributionally robust prediction error of the $\sqrt{\text{LASSO}}$ and related estimators"

APPENDIX A: PROOFS OF RESULTS IN THE MAIN TEXT

A.1. Proofs of Remark 2, Example A.1 and Remark 4. We begin with a remark about the worst-case distribution.

REMARK 3. *The worst-case distribution $\mathbb{P}^*_{\boldsymbol{\beta}}$ is also included in the* Wasserstein *ball of radius $n^{-1/2}$ for any norm ρ and $\|(x,y)\| := (\rho(x)+|y|)_*$; we have included this in Figure 1. Indeed, from Corollary 2.1 and Remark 2, we have that*

$$\mathcal{W}_r(\mathbb{P},\mathbb{P}^*_{\boldsymbol{\beta}}) = \inf_{\pi\in\Pi(\mathbb{P},\mathbb{P}^*_{\boldsymbol{\beta}})} \left(\mathbb{E}_\pi\left[\left\|\left(-e\big(\boldsymbol{\beta}^* - \frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\big), e\right)\right\|^r\right]\right)^{1/r} \le \mathbb{E}[\|e(\boldsymbol{\beta}^*,1)\|^r]^{1/r}$$
$$\le \mathbb{E}[|e|^r]^{1/r} = n^{-1/2},$$

where we have used $\rho^*(\boldsymbol{\beta}^*)=0$ and $\|(\boldsymbol{\beta}^*,1)\|\le 1$, as well as $\mathbb{E}[|e|^r]^{1/r}=n^{-1/2}$.

PROOF OF REMARK 2. Recalling that the dual norm of $\|\cdot\|$ is given by

$$\|\mathbf{x}\|_* := \sup_{\mathbf{y}:\|\mathbf{y}\|=1} \mathbf{x}^\top\mathbf{y}, \tag{A.1}$$

[17, Example 3.26] states that $\rho^*(\mathbf{x}) = \infty\mathbb{1}_{\{\|\mathbf{x}\|_*>1\}}$. Recall further that $\boldsymbol{\beta}^*\in\partial\rho(\boldsymbol{\beta})$ if and only if

$$(\boldsymbol{\beta}^*)^\top\boldsymbol{\beta} - \rho^*(\boldsymbol{\beta}^*) = \rho(\boldsymbol{\beta}). \tag{A.2}$$

Both facts together imply that $\rho^*(\boldsymbol{\beta}^*)=0$; thus, $\|\boldsymbol{\beta}^*\|_*\le 1$ for all $\boldsymbol{\beta}^*\in\partial\rho(\boldsymbol{\beta})$. Hence in (8), as claimed, we have

$$\left|\boldsymbol{\gamma}^\top\left(\boldsymbol{\beta}^* - \frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\right)\right| = \left|\boldsymbol{\gamma}^\top\boldsymbol{\beta}^*\right| \le \|\boldsymbol{\gamma}\|\,\|\boldsymbol{\beta}^*\|_* \le \|\boldsymbol{\gamma}\|, \quad \forall\,\boldsymbol{\gamma}\in\mathbb{R}^d. \tag{A.3}$$

□

EXAMPLE (Condition (8) for a function ρ that is not a norm). *Fix any compact set $K\subseteq\mathbb{R}^d$ such that $-K=K$ and consider*

$$\rho(\boldsymbol{\beta}) = \sup_{\mathbf{y}\in K}\boldsymbol{\beta}^\top\mathbf{y}.$$

Then ρ is convex (as a supremum of linear functions), finite (as K is compact), non-negative (as $K=-K$), symmetric $\rho(\boldsymbol{\beta})=\rho(-\boldsymbol{\beta})$ (as $K=-K$), and homogeneous $\rho(\lambda\boldsymbol{\beta})=\lambda\rho(\boldsymbol{\beta})$. Thus,

$$\rho^*(\boldsymbol{\beta}^*) = \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d}\left(\boldsymbol{\beta}^{*\top}\boldsymbol{\gamma}-\rho(\boldsymbol{\gamma})\right) = \begin{cases}\infty & \text{if } \exists\boldsymbol{\gamma}\in\mathbb{R}^d \text{ s.t. } \boldsymbol{\beta}^{*\top}\boldsymbol{\gamma}-\rho(\boldsymbol{\gamma})>0,\\ 0 & \text{if } \boldsymbol{\beta}^{*\top}\boldsymbol{\gamma}-\rho(\boldsymbol{\gamma})\le 0 \text{ for all } \boldsymbol{\gamma}\in\mathbb{R}^d.\end{cases}$$

By (A.2) we conclude that $\rho^*(\boldsymbol{\beta}^*)=0$ for all $\boldsymbol{\beta}\in\mathbb{R}^d$; therefore, $\boldsymbol{\beta}^{*\top}\boldsymbol{\gamma}\le\rho(\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma}\in\mathbb{R}^d$. By symmetry of ρ, we also have that $|\boldsymbol{\beta}^{*\top}\boldsymbol{\gamma}|\le\rho(\boldsymbol{\gamma})$. It follows that

$$\left|\boldsymbol{\gamma}^\top\left(\boldsymbol{\beta}^* - \frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\right)\right| \le \rho(\boldsymbol{\gamma}), \quad \forall\,\boldsymbol{\gamma}\in\mathbb{R}^d.$$

REMARK 4. *Take any norm $\|\cdot\|$ on $\mathbb{R}^{d+1}$ satisfying $\|(0,\dots,0,1)\|=1$ and recall that its dual norm is given by (A.1). Assume that $\mathbb{E}_{\mathbb{P}}[\|(\mathbf{X},Y)\|_*^r]<\infty$ and consider a Wasserstein ball $\mathcal{B}^{\mathcal{W}}_\delta(\mathbb{P})$ with cost $\|\cdot\|_*$, defined as*

$$\mathcal{B}^{\mathcal{W}}_\delta(\mathbb{P}) = \left\{\widetilde{\mathbb{P}}\in\mathcal{P}_r(\mathbb{R}^{d+1}):\ \mathcal{W}_r(\mathbb{P},\widetilde{\mathbb{P}})\le\delta\right\}. \tag{A.4}$$

We show that for $\rho(\cdot) = \|\cdot\|$, the ball defined in (7) contains the ball in (A.4), i.e. $\mathcal{B}_\delta^{\mathcal{W}}(\mathbb{P}) \subseteq B_\delta(\mathbb{P})$. For this, we note that by (A.1) we have

$$\mathbb{E}_\pi\left[\left|\left(\widetilde{Y}-Y\right)+\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^\top\boldsymbol{\gamma}\right|^r\right] \leq \|(\boldsymbol{\gamma},-1)\|^r\ \mathbb{E}_\pi\left[\|(\mathbf{X},Y)-(\widetilde{\mathbf{X}},\widetilde{Y})\|_*^r\right]$$
$$\leq (1+\|\boldsymbol{\gamma}\|)^r\ \mathbb{E}_\pi\left[\|(\mathbf{X},Y)-(\widetilde{\mathbf{X}},\widetilde{Y})\|_*^r\right].$$

We conclude

$$\sup_{\boldsymbol{\gamma}\in\mathbb{R}^d}\inf_{\pi\in\Pi(\mathbb{P},\widetilde{\mathbb{P}})}\frac{1}{1+\|\boldsymbol{\gamma}\|}\sqrt[r]{\mathbb{E}_\pi\left[\left|(\widetilde{Y}-Y)+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\gamma}\right|^r\right]} \leq \inf_{\pi\in\Pi(\mathbb{P},\widetilde{\mathbb{P}})}\sqrt[r]{\mathbb{E}_\pi\left[\|(\mathbf{X},Y)-(\widetilde{\mathbf{X}},\widetilde{Y})\|_*^r\right]}.$$

The above can be applied in particular to $\|\cdot\| = \|\cdot\|_p$ and $\|\cdot\|_* = \|\cdot\|_q$, where $1/p+1/q=1$.

We note that the conditions used for the derivations in Remark 2 are sufficient, but not necessary. To make this point, consider the case in which $r=2$ and $\rho(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$. The Wasserstein distance, $\mathcal{W}_2$, between $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ is defined by

$$\mathcal{W}_2(\mathbb{P},\widetilde{\mathbb{P}}) = \inf_{\pi\in\Pi(\mathbb{P},\widetilde{\mathbb{P}})}\mathbb{E}_\pi[\|(\mathbf{X},Y)-(\widetilde{\mathbf{X}},\widetilde{Y})\|_2^2]^{1/2},$$

where $\|\cdot\|_2$ is the Euclidean distance. For any coupling π, the Cauchy-Schwarz inequality implies

$$\mathbb{E}_\pi\left[\left|(\widetilde{Y}-Y)+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\beta}\right|^2\right]^{1/2} \leq \|(\boldsymbol{\beta},-1)\|_2\ \mathbb{E}_\pi\left[\|(\mathbf{X},Y)-(\widetilde{\mathbf{X}},\widetilde{Y})\|_2^2\right]^{1/2},$$

where the right-hand side of the previous inequality is less than

$$(1+\|\boldsymbol{\beta}\|_2)\mathbb{E}_\pi\left[\|(\mathbf{X},Y)-(\widetilde{\mathbf{X}},\widetilde{Y})\|_2^2\right]^{1/2}.$$

Note that $1+\|\boldsymbol{\beta}\|_2 \leq 1+\|\boldsymbol{\beta}\|_1$, and thus

$$\widehat{\mathcal{W}}_2(\mathbb{P},\widetilde{\mathbb{P}}) \leq \mathcal{W}_2(\mathbb{P},\widetilde{\mathbb{P}}).$$

Thus, balls based on the ρ-MSW metric can be larger than balls based on the d-dimensional Wasserstein metric, even when the latter does not use a cost function based on the dual norm of ρ. Our previous derivations also hold, *mutatis mutandi*, for penalty functions $\rho(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_p$ whenever $p\in[1,2]$. Remark 2 focuses on the case in which i) ρ is a norm, and ii) the cost function used in the d-dimensional Wasserstein metric is associated to the dual norm of ρ to make our results directly comparable to those in Proposition 2 in [10].

A.2. Proof of Theorem 2.1. In Section 2 we provided a proof sketch after the statement of Theorem 2.1. Here we elaborate on the details of the proof. The statements of two steps mentioned in Section 2 are repeated below for the reader's convenience.

Step 1. We show that

$$\left(\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]\right)^{1/r} \leq \sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]}+\delta\left(1+\rho(\boldsymbol{\beta})\right), \tag{A.5}$$

holds for any $\boldsymbol{\beta}\in\mathbb{R}^d$ and any $\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})$.

Proving Step 1. Take an arbitrary $\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})$ and let $\pi(\boldsymbol{\beta})$ be an optimal coupling for $\widehat{\mathcal{W}}_{r,\rho}$. By writing $\pi(\boldsymbol{\beta})$ we emphasize that the coupling will depend on $\boldsymbol{\beta}$; though, this matters little for the proof. Namely, $((\mathbf{X},Y),(\widetilde{\mathbf{X}},\widetilde{Y}))\sim\pi(\boldsymbol{\beta})$ with $(\mathbf{X},Y)\sim\mathbb{P}$ and $(\widetilde{\mathbf{X}},\widetilde{Y})\sim\widetilde{\mathbb{P}}$. Consequently we conclude that

$$\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right] = \mathbb{E}_{\pi(\boldsymbol{\beta})}\left[\left|\widetilde{Y}-\widetilde{\mathbf{X}}^\top\boldsymbol{\beta}\right|^r\right].$$

By the triangle inequality we obtain

$$\sqrt[r]{\mathbb{E}_{\pi(\boldsymbol{\beta})}\left[\left|\widetilde{Y}-\widetilde{\mathbf{X}}^\top\boldsymbol{\beta}\right|^r\right]} = \sqrt[r]{\mathbb{E}_{\pi(\boldsymbol{\beta})}\left[\left|(\widetilde{Y}-Y)+(Y-\mathbf{X}^\top\boldsymbol{\beta})+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\beta})\right|^r\right]}$$
$$\leq \sqrt[r]{\mathbb{E}_{\pi(\boldsymbol{\beta})}\left[|Y-\mathbf{X}^\top\boldsymbol{\beta}|^r\right]} + \sqrt[r]{\mathbb{E}_{\pi(\boldsymbol{\beta})}\left[\left|(\widetilde{Y}-Y)+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\beta})\right|^r\right]}.$$

Recalling the choice of $\pi(\beta)$ we conclude that

$$\sqrt[r]{\mathbb{E}_{\pi(\boldsymbol{\beta})}\left[\left|\widetilde{Y}-\widetilde{\mathbf{X}}^\top\boldsymbol{\beta}\right|^r\right]} \leq \sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]} + \delta\left(1+\rho(\boldsymbol{\beta})\right). \tag{A.6}$$

Step 2. We show that for any $\boldsymbol{\beta}\in\mathrm{dom}(\rho)$, the upper bound given in **Step 1** is tight; i.e. we construct $\mathbb{P}^*\in B_\delta(\mathbb{P})$, for which the bound holds exactly.

Proof Step 2. Let $\boldsymbol{\beta}^*$ be an element of $\partial\rho(\boldsymbol{\beta})$ satisfying Equation (8).

Consider the distribution $\mathbb{P}^*$ corresponding to the random vector $(\widetilde{\mathbf{X}},\widetilde{Y})$ defined by

$$\widetilde{\mathbf{X}}=\mathbf{X}-e\left(\boldsymbol{\beta}^*-\frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\right), \qquad \widetilde{Y}=Y+e, \tag{A.7}$$

where

$$e:=\frac{\delta(Y-\mathbf{X}^\top\boldsymbol{\beta})}{\sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]}}, \qquad (Y,\mathbf{X})\sim\mathbb{P}.$$

The distributions $\mathbb{P}^*$ and $\mathbb{P}$ are already coupled, since $(\widetilde{\mathbf{X}},\widetilde{Y})$ are measurable functions of $(\mathbf{X},Y)\sim\mathbb{P}$. Let $\pi^*(\boldsymbol{\beta})$ denote the coupling of $(\mathbb{P}^*,\mathbb{P})$.

Next we show that the distribution $\mathbb{P}^*$ of $(\widetilde{\mathbf{X}},\widetilde{Y})$ is an element of $B_\delta(\mathbb{P})$: by construction we have

$$\mathbb{E}_{\pi^*(\boldsymbol{\beta})}\left[\left|(\widetilde{Y}-Y)+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\gamma}\right|^r\right] = \mathbb{E}_{\pi^*(\boldsymbol{\beta})}\left[|e|^r\left|1+\left(\boldsymbol{\beta}^*-\frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\right)^\top\boldsymbol{\gamma}\right|^r\right]$$
$$= \left|1+(\boldsymbol{\beta}^*)^\top\boldsymbol{\gamma}-\frac{\boldsymbol{\beta}^\top\boldsymbol{\gamma}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\right|^r \mathbb{E}_{\pi^*(\boldsymbol{\beta})}\left[|e|^r\right]$$
$$\leq \left[\delta\left(1+\rho(\boldsymbol{\gamma})\right)\right]^r,$$

where the last inequality follows because

$$\left|\left(\boldsymbol{\beta}^*-\frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\right)^\top\boldsymbol{\gamma}\right|\leq\rho(\boldsymbol{\gamma}),$$

for any $\boldsymbol{\gamma}\in\mathbb{R}^d$ by the assumption in (8) and since $\mathbb{E}_{\mathbb{P}}[|e|^r]=\delta^r$.

Thus, we only need to compute $\mathbb{E}_{\mathbb{P}^*}[|Y-\mathbf{X}^\top\boldsymbol{\beta}|^r]=\mathbb{E}_{\pi^*(\boldsymbol{\beta})}[|\widetilde{Y}-\widetilde{\mathbf{X}}^\top\boldsymbol{\beta}|^r]$. Adding and subtracting $\mathbf{X}^\top\beta$ and Y to $\widetilde{Y}-\widetilde{\mathbf{X}}^\top\boldsymbol{\beta}$ we have from (A.7)

$$\widetilde{Y}-\widetilde{\mathbf{X}}^\top\boldsymbol{\beta}=\widetilde{Y}-Y+Y-\mathbf{X}^\top\boldsymbol{\beta}+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\beta}=\left(Y-\mathbf{X}^\top\boldsymbol{\beta}\right)+e\left(1+\rho(\boldsymbol{\beta})\right), \tag{A.8}$$

where the last term applies [64, Theorem 23.5, p. 218], which shows that for any proper, convex function $\boldsymbol{\beta}^*\in\partial\rho(\boldsymbol{\beta})$ if and only if

$$(\boldsymbol{\beta}^*)^\top\boldsymbol{\beta}-\rho^*(\boldsymbol{\beta}^*)=\rho(\boldsymbol{\beta});$$

hence,

$$\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^\top\boldsymbol{\beta}=e\left(\boldsymbol{\beta}^*-\frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\right)^\top\boldsymbol{\beta}=e\rho\left(\boldsymbol{\beta}\right).$$

Therefore, using (A.8) and writing $(Y - \mathbf{X}^\top \boldsymbol{\beta})$ as $e\,\sqrt[r]{\mathbb{E}_{\mathbb{P}}[|Y - \mathbf{X}^\top \boldsymbol{\beta}|^r]}/\delta$, we have that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^*}\left[\left|Y - \mathbf{X}^\top \boldsymbol{\beta}\right|^r\right] &= \mathbb{E}_{\pi^*(\boldsymbol{\beta})}\left[\left|\left(Y - \mathbf{X}^\top \boldsymbol{\beta}\right) + e(1+\rho(\boldsymbol{\beta}))\right|^r\right] \\ &= \left|\frac{1}{\delta}\sqrt[r]{\mathbb{E}_{\mathbb{P}}[|Y - \mathbf{X}^\top \boldsymbol{\beta}|^r]} + (1+\rho(\boldsymbol{\beta}))\right|^r \mathbb{E}_{\mathbb{P}}\left[|e|^r\right] \\ &= \left|\sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[|Y - \mathbf{X}^\top \boldsymbol{\beta}|^r\right]} + \delta(1+\rho(\boldsymbol{\beta}))\right|^r.\end{aligned}$$

In the final step above, we again used that $\mathbb{E}_{\mathbb{P}}[|e|^r] = \delta^r$.

A.3. Proof of Theorem 3.1. We first recall the representations for the one-dimensional Wasserstein distance for $r \geq 1$

$$\mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}})^r = \int_0^1 \left|F^{-1}_{\boldsymbol{\gamma},n}(p) - F^{-1}_{\boldsymbol{\gamma}}(p)\right|^r dp, \tag{A.9}$$

and for $r = 1$

$$\mathcal{W}_1(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}}) = \int_{\mathbb{R}} \left|F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)\right| dt, \tag{A.10}$$

see e.g. [14, Theorem 2.9, Theorem 2.10]. We also note that $\mathcal{W}_r$ is translation invariant, which implies

$$\mathcal{W}_r\left(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}}\right) = \mathcal{W}_r\left(\left[\left((\mathbf{X},Y) - (\mathbf{x_0}, y_0)\right)^\top \boldsymbol{\gamma}\right]_* \mathbb{P}_{\boldsymbol{\gamma},n},\, \left[\left((\mathbf{X},Y) - (\mathbf{x_0}, y_0)\right)^\top \boldsymbol{\gamma}\right]_* \mathbb{P}_{\boldsymbol{\gamma}}\right),$$

for any $\mathbf{x_0} \in \mathbb{R}^d$ and $y_0 \in \mathbb{R}$. Defining $c := \operatorname{diam}(\operatorname{supp}(\mathbb{P}))$, there is no loss of generality if we assume

$$\operatorname{supp}(\mathbb{P}_{\boldsymbol{\gamma}}) \subseteq [0,\, c]\,. \tag{A.11}$$

Noting that $|F^{-1}_{\boldsymbol{\gamma},n}(p) - F^{-1}_{\boldsymbol{\gamma}}(p)| \leq c$ for all $p \in (0,1)$, we estimate

$$\begin{aligned}\overline{\mathcal{W}}_r(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}})^r 7 &= \sup_{\|\boldsymbol{\gamma}\|=1} \int_0^1 \left|F^{-1}_{\boldsymbol{\gamma},n}(p) - F^{-1}_{\boldsymbol{\gamma}}(p)\right|^r dp \\ &\leq c^{r-1} \sup_{\|\boldsymbol{\gamma}\|=1} \int_0^1 \left|F^{-1}_{\boldsymbol{\gamma},n}(p) - F^{-1}_{\boldsymbol{\gamma}}(p)\right| dp = c^{r-1} \sup_{\|\boldsymbol{\gamma}\|=1} \int_{\mathbb{R}} \left|F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)\right| dt,\end{aligned}$$

where the final inequality follows from (A.9) and (A.10). Next, recalling (A.11),

$$\sup_{\|\boldsymbol{\gamma}\|=1} \int_{\mathbb{R}} \left|F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)\right| dt \leq \sup_{\|\boldsymbol{\gamma}\|=1} \int_0^c \sup_t \left|F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)\right| ds \leq c \sup_{f \in \mathcal{H}} \left|\mathbb{E}_{\mathbb{P}_n}[f] - \mathbb{E}_{\mathbb{P}}[f]\right|,$$

where

$$\mathcal{H} := \left\{\mathbb{1}_{\{\mathbf{x}^\top \boldsymbol{\gamma} \leq t\}} : \boldsymbol{\gamma} \in \mathbb{R}^{d+1},\, t \in \mathbb{R}\right\}.$$

The claim now follows from Lemma B.3 in Section B.

A.4. Proof of Theorem 3.2. By Lemma B.4 in Section B with $k = \log(2n+1)^{1/s}$ we have

$$\overline{\mathcal{W}}_r(\mathbb{P}_n, \mathbb{P})^r \leq 2^r r \log(2n+1)^{r/s} \left(I_1 + \frac{\sqrt{\Gamma \vee \Gamma_n}}{s/2 - r} \log(2n+1)^{-1/2} I_2\right), \tag{A.12}$$

where

$$I_1 = \sup_{(\boldsymbol{\gamma},t) \in \mathbb{R}^{d+1} \times \mathbb{R}} \left|F_{\boldsymbol{\gamma}}(t) - F_{\boldsymbol{\gamma},n}(t))\right|,$$

$$I_2 = \sup_{(\boldsymbol{\gamma},t) \in \mathbb{R}^{d+1} \times \mathbb{R}} \frac{(F_{\boldsymbol{\gamma}}(t) - F_{\boldsymbol{\gamma},n}(t))^+}{\sqrt{F_{\boldsymbol{\gamma}}(t)(1 - F_{\boldsymbol{\gamma},n}(t))}} + \sup_{(\boldsymbol{\gamma},t) \in \mathbb{R}^{d+1} \times \mathbb{R}} \frac{(F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))^+}{\sqrt{F_{\boldsymbol{\gamma},n}(t)(1 - F_{\boldsymbol{\gamma}}(t))}},$$

$$\Gamma_n = \sup_{\|\boldsymbol{\gamma}\|=1} \mathbb{E}_{\mathbb{P}_n}\left[|(\mathbf{X},Y)^\top\boldsymbol{\gamma}|^s\right] = \sup_{\|\boldsymbol{\gamma}\|=1} \frac{1}{n}\sum_{i=1}^n \left|(\mathbf{X}_i,Y_i)^\top\boldsymbol{\gamma}\right|^s,$$

$$\Gamma = \mathbb{E}_{\mathbb{P}}\left[\|(\mathbf{X},Y)\|_2^s\right] = \mathbb{E}_{\mathbb{P}}\Big[\sup_{\|\boldsymbol{\gamma}\|=1} |(\mathbf{X},Y)^\top\boldsymbol{\gamma}|^s\Big].$$

Next, by Markov's inequality and the triangle inequality

$$\mathbb{P}\left(\Gamma_n \geq C\right) \leq \frac{\mathbb{E}_{\mathbb{P}}[\Gamma_n]}{C} = \frac{1}{C}\mathbb{E}_{\mathbb{P}}\left[\sup_{\|\boldsymbol{\gamma}\|=1}\frac{1}{n}\sum_{i=1}^n\left|(\mathbf{X}_i,Y_i)^\top\boldsymbol{\gamma}\right|^s\right] \leq \frac{\Gamma}{C}.$$

Setting the last expression equal to α yields $\Gamma_n \leq \Gamma/\alpha$ on a set of probability at least $1-\alpha$. Combining this with Lemma B.3 (to control I_1) and Lemma B.5 (to control I_2) yields that $\overline{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{P})^r$ is less than or equal to the following with probability greater than $1-3\alpha$:

$$2^r r\log(2n+1)^{\frac{r}{s}}\left[\frac{1}{\sqrt{n}}\left(180\sqrt{d+2}+\sqrt{2\log\left(\frac{1}{\alpha}\right)}\right)+\sqrt{\frac{\Gamma}{\alpha}}\frac{1}{s/2-r}\frac{1}{\sqrt{\log(2n+1)}}I_2\right]$$

$$\leq \frac{2^r r\log(2n+1)^{\frac{r}{s}}}{\sqrt{n}}\left[180\sqrt{d+2}+\sqrt{2\log\left(\frac{1}{\alpha}\right)}+\sqrt{\frac{\Gamma}{\alpha}}\frac{8}{s/2-r}\sqrt{\log\left(\frac{8}{\alpha}\right)+(d+2)}\right],$$

which is the claim.

A.5. Proof of Theorem 4.1. This claim follows from the estimate

$$\mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma},n},\mathbb{P}_{\boldsymbol{\gamma}})^r \leq \operatorname{diam}(\operatorname{supp}(\mathbb{P}))^r \sup_{f\in\mathcal{H}^0}|\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f]|,$$

stated in the proof of Theorem 3.1 together with

$$\sqrt{n}\sup_{f\in\mathcal{H}^0}|\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f]| \Rightarrow \sup_{f\in\mathcal{H}^0}|G_f|,$$

as in the proof of Theorem 4.2. As $\sup_{f\in\mathcal{H}^0}|G_f|$ dominates $\sup_{t\in[0,1]}|B(t)|$ in stochastic order, this concludes the proof.

A.6. Proof of Theorem 4.2. Note that again by [14, Proposition 7.14] we have

$$\begin{aligned}
\mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma},n},\mathbb{P}_{\boldsymbol{\gamma}})^r &\leq r2^{r-1}\int|t|^{r-1}|F_{\boldsymbol{\gamma},n}(t)-F_{\boldsymbol{\gamma}}(t)|\,dt\\
&= r2^{r-1}\Big(\int_0^\infty|t|^{r-1}|(1-F_{\boldsymbol{\gamma},n}(t))-(1-F_{\boldsymbol{\gamma}}(t))|\,dt+\int_{-\infty}^0|t|^{r-1}|F_{\boldsymbol{\gamma},n}(t)-F_{\boldsymbol{\gamma}}(t)|\,dt\Big)\\
&\leq r2^{r-1}\sup_{f\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-}|\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f]|\int(1\wedge|t|^{r-s-1})\,dt\\
&\leq c\sup_{f\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-}|\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f]|,
\end{aligned}$$

where $c := r2^{r-1}\int(1\wedge|t|^{r-s-1})\,dt$. We next find an envelope F for $\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-$: it is easy to see that

$$\sup_{f\in\mathcal{H}^+}|f(\boldsymbol{x})| \leq \sup_{\|\boldsymbol{\gamma}\|=1}|\boldsymbol{x}^\top\boldsymbol{\gamma}|^s \leq \|\boldsymbol{x}\|_*^s.$$

A similar argument for $\mathcal{H}^-$ yields

$$F(\boldsymbol{x}) := \sup_{f\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-}|f(\boldsymbol{x})| \leq \|\boldsymbol{x}\|_*^s\vee 1.$$

As $\mathcal{H}^0$ is VC-subgraph by Lemma B.3, [Van der Vaart, Wellner, Lemma 2.6.22] implies that $\mathcal{H}^+$ and $\mathcal{H}^-$ are also VC-subgraph: indeed note that

$$\begin{aligned}\{(\boldsymbol{x},u):\ u\le |t|^s \mathbb{1}_{\{t\le \boldsymbol{x}^\top\boldsymbol{\gamma}\}}\} &= \{(\boldsymbol{x},u):\ t\le \boldsymbol{x}^\top\boldsymbol{\gamma}, u\le |t|^s\}\cup\{(\boldsymbol{x},u):\ t>\boldsymbol{x}^\top\boldsymbol{\gamma}\}\\ &= \{(\boldsymbol{x},u):\ t\le \boldsymbol{x}^\top\boldsymbol{\gamma}\}\cap\{(\boldsymbol{x},u):\ u\le |t|^s\}\cup\{(\boldsymbol{x},u):\ t>\boldsymbol{x}^\top\boldsymbol{\gamma}\},\end{aligned}$$

so the claim follows from the fact that $\mathcal{H}$ is VC, finite dimensional vector spaces of functions are VC subgraph [Van der Vaart, Wellner, Lemma 2.6.15], and [Van der Vaart, Wellner, Lemma 2.6.17 (ii), (iii)]. Then, [Van der Vaart, Wellner, Theorem 2.6.7] states that for all $\epsilon\in(0,1)$,

$$N(\epsilon\|F\|_{\mathbb{Q},2},\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-,L_2(\mathbb{Q}))\le C_1\Big(\frac{1}{\epsilon}\Big)^{2C_2-1},$$

for universal constants $C_1,C_2>1$ and any probability measure $\mathbb{Q}$, for which $\|F\|_{\mathbb{Q},2}>0$. Thus,

$$\int_0^\infty \sup_{\mathbb{Q}}\sqrt{\log N(\epsilon\|F\|_{\mathbb{Q},2},\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-,L_2(\mathbb{Q}))}\,d\epsilon<\infty,$$

and together with $\Gamma<\infty$, [Van der Vaart, Wellner, Theorem 2.5.2] implies that $\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-$ is Donsker. Thus, the convergence in distribution

$$\sqrt{n}\sup_{f\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-}|\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f]|\Rightarrow\sup_{f\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-}|G_f|,$$

holds, where (G_f) is a zero-mean Gaussian process satisfying

$$\mathbb{E}[G_{f_1}G_{f_2}]=\mathbb{E}_{\mathbb{P}}[f_1f_2]-\mathbb{E}_{\mathbb{P}}[f_1]\mathbb{E}_{\mathbb{P}}[f_2],$$

for any $f_1,f_2\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-$. Next, from the proof of [Van der Vaart, Wellner, Theorem 2.5.2] we obtain the inequality

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}}\left[\sqrt{n}\sup_{f\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-}|\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f]|\right]\\ &\qquad\le C\sqrt{\Gamma}\int_0^\infty\sup_{\mathbb{Q}}\sqrt{\log N\left(\epsilon\|F\|_{\mathbb{Q},2},\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-,L_2(\mathbb{Q})\right)}\,d\epsilon.\end{aligned}$$

This shows the second claim.

A.7. Proof of Theorem 5.1. Note that $\widehat{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{Q})\le\widehat{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{P})+\widehat{\mathcal{W}}_r(\mathbb{P},\mathbb{Q})$ by the triangle inequality because $\widehat{\mathcal{W}}_r$ is a metric. Then,

$$E_n:=\left\{\widehat{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{Q})>\delta_{n,r}+\epsilon\right\}\subset\left\{\widehat{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{P})>\delta_{n,r}\right\}\subset\left\{\overline{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{P})>\delta_{n,r}/c_{\rho,d}\right\},$$

which implies that the probability of E_n^c is greater than $1-3\alpha$ due to Theorem 3.2 (or 3.1). In the equation above, $c_{\rho,d}$ is defined via (17). Finally, on the event E_n^c, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]^{1/r}&\le\sup_{\widetilde{\mathbb{P}}\in B_{\delta_{n,r}+\epsilon}(\mathbb{P}_n)}\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]^{1/r}\\ &=\mathbb{E}_{\mathbb{P}_n}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]^{1/r}+\left(\delta_{n,r}+\epsilon\right)\left(1+\rho(\boldsymbol{\beta})\right),\qquad\forall\boldsymbol{\beta},\end{aligned}$$

where the last equality follows from Theorem 2.1.

A.8. Proof of Theorem 5.2. The proof has three steps. The first two steps adapt what we learn in Section 3 to our particular setup. The last step concludes based on observations about Theorems 2.1 and 3.2.

Step 1: Let us compare the (ρ,σ)-MSW metric to the MSW metric using reasoning that is similar to our derivations in (16) – (17). Defining $\bar{\boldsymbol{\gamma}}_\sigma^\top = (\boldsymbol{\gamma}^\top, -\sigma)$, we obtain

$$\begin{aligned}\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P},\widetilde{\mathbb{P}}) &= \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d} \frac{\|\bar{\boldsymbol{\gamma}}_\sigma\|}{\sigma+\rho(\boldsymbol{\gamma})}\frac{1}{\|\bar{\boldsymbol{\gamma}}_\sigma\|}\mathcal{W}_r\left(\left[(\mathbf{X},Y/\sigma)^\top\bar{\boldsymbol{\gamma}}_\sigma\right]_*\mathbb{P},\,[(\widetilde{\mathbf{X}},\widetilde{Y}/\sigma)^\top\bar{\boldsymbol{\gamma}}_\sigma]_*\widetilde{\mathbb{P}}\right)\\ &\le (\max\{1/c_d,1\})\sup_{\|\boldsymbol{\gamma}\|=1}\mathcal{W}_r\left(\left[(\mathbf{X},Y/\sigma)^\top\boldsymbol{\gamma}\right]_*\mathbb{P},\,[(\widetilde{\mathbf{X}},\widetilde{Y}/\sigma)^\top\boldsymbol{\gamma}]_*\widetilde{\mathbb{P}}\right)\\ &= (\max\{1/c_d,1\})\,\overline{\mathcal{W}}_r\left(\mathbb{P}^\sigma,\widetilde{\mathbb{P}}^\sigma\right),\end{aligned}$$

where $\mathbb{P}^\sigma := (\mathbf{X},Y/\sigma)_*\mathbb{P}$ and $\widetilde{\mathbb{P}}^\sigma := (\mathbf{X},Y/\sigma)_*\widetilde{\mathbb{P}}$.

Step 2: Let us apply Theorem 3.2 to compute the rates for $\overline{\mathcal{W}}_r(\mathbb{P}^\sigma,\widetilde{\mathbb{P}}^\sigma)$, which depend on $\mathbb{E}_{\mathbb{P}^\sigma}\left[\|(\mathbf{X},Y)\|_*^s\right]$. Consider the following derivation

$$\mathbb{E}_{\mathbb{P}^\sigma}\left[\|(\mathbf{X},Y)\|_*^s\right] \le 2^{s-1}\left(\mathbb{E}_{\mathbb{P}^\sigma}\left[\|(\mathbf{X},0)\|_*^s\right] + \mathbb{E}_{\mathbb{P}^\sigma}\left[\|(0,\ldots,0,Y)\|_*^s\right]\right)\,.$$

Note that $\mathbb{E}_{\mathbb{P}^\sigma}\left[\|(\mathbf{X},0)\|_*^s\right] = \mathbb{E}_{\mathbb{P}}\left[\|(\mathbf{X},0)\|_*^s\right] = 1$ and

$$\mathbb{E}_{\mathbb{P}^\sigma}\left[\|(0,\ldots,0,Y)\|_*^s\right] = \mathbb{E}_{\mathbb{P}}\left[\|(0,\ldots,0,Y/\sigma)\|_*^s\right] \le 1,$$

due to our assumptions. This implies that

$$\mathbb{E}_{\mathbb{P}^\sigma}\left[\|(\mathbf{X},Y)\|_*^s\right] \le 2^s\,.$$

Step 3: We note that Theorem 3.2 still holds for any Γ larger than $\mathbb{E}_{\mathbb{P}}\left[\|(\mathbf{X},Y)\|_*^s\right]$. In particular, we can consider $\Gamma = 2^s$ due to Step 2. In addition, we note that the conclusion of Theorem 2.1 is unaffected when replacing $\widehat{\mathcal{W}}_r$ by $\widehat{\mathcal{W}}_{r,\rho,\sigma}$ and the choice of $\sigma \ge 1$. These observations and the argument presented in the proof of Theorem 5.1 conclude our proof.

A.9. Additional remark on selection of ρ. While we are able to provide a recommendation for δ, our current results do not allow us to say anything concrete about the selection of penalty function, ρ. This is in part due to the fact that, in our framework, we have too much flexibility making this choice. To illustrate this point, suppose that we wanted to pick the penalty function to optimize the out-of-distribution prediction error of a linear predictor based on (1) at a known distribution $\mathbb{Q}$. If n denotes the sample size, we could always pick the convex penalty function

$$\rho^n(\boldsymbol{\beta}) := n\left(\mathrm{E}_{\mathbb{Q}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]\right)^{1/r}.$$

As n grows to infinity, the relevance of the penalty function increases, and thus the solution of (1) converges to

$$\arg\inf_{\boldsymbol{\beta}\in\mathbb{R}^d}\left(\mathbb{E}_{\mathbb{Q}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]\right)^{1/r}.$$

This just formalizes the obvious point that if we know the testing distribution $\mathbb{Q}$ at which we would like to have good performance, then it is better to use the best predictor based on such a distribution.

A.10. Proof of Corollary 5.1. By Theorem 1 and conditions (i), (ii) above, (29) is equivalent to

$$\mathbb{E}_{\mathbb{P}}\left[|Y-\mathbf{X}^\top\boldsymbol{\beta}_1|^r\right]^{1/r} + \delta\rho(\boldsymbol{\beta}_1) \le \mathbb{E}_{\mathbb{P}}\left[|Y-\mathbf{X}^\top\boldsymbol{\beta}_2|^r\right]^{1/r} + \delta\rho(\boldsymbol{\beta}_2)\,.$$

The previous expression is equivalent to

$$n^{-1/(2r)}\left(2+\rho(\boldsymbol{\beta}_1)+\rho(\boldsymbol{\beta}_2)\right)T_n \le \Delta_n(\boldsymbol{\beta}_2) - \Delta_n(\boldsymbol{\beta}_1)\,,$$

where $\Delta_n(\boldsymbol{\beta}) = \mathbb{E}_{\mathbb{P}}[|Y-\mathbf{X}^\top\boldsymbol{\beta}|^r]^{1/r} - \mathbb{E}_{\mathbb{P}_n}[|Y-\mathbf{X}^\top\boldsymbol{\beta}|^r]^{1/r}$. By definition of $\widehat{\mathcal{W}}_r$ and Theorem 2.1, it follows that

$$\frac{\Delta_n(\boldsymbol{\beta}_2)}{1+\rho(\boldsymbol{\beta}_2)} \le \widehat{\mathcal{W}}_r(\mathbb{P},\mathbb{P}_n), \quad \text{and} \quad \frac{-\Delta_n(\boldsymbol{\beta}_1)}{1+\rho(\boldsymbol{\beta}_1)} \le \widehat{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{P}).$$

Therefore, we have

$$n^{-1/(2r)}\left(2+\rho(\boldsymbol{\beta}_1)+\rho(\boldsymbol{\beta}_2)\right)T_n \leq \widehat{\mathcal{W}}_r(\mathbb{P},\mathbb{P}_n)\left(2+\rho(\boldsymbol{\beta}_1)+\rho(\boldsymbol{\beta}_2)\right) .$$

Using $\widehat{\mathcal{W}}_r(\mathbb{P},\mathbb{P}_n) \leq c_{\rho,d}\overline{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{P})$ from (17), we derive that

$$\mathbb{P}\left(T_n > c_{\rho,d}C^{1/r}\right) \leq \mathbb{P}\left(c_{\rho,d}\overline{\mathcal{W}}_r(\mathbb{P},\mathbb{P}_n) > c_{\rho,d}\,C^{1/r}n^{-1/(2r)}\right) .$$

Finally, Theorem 3.1 implies that the above probability is bounded by α.

APPENDIX B: AUXILIARY RESULTS

We start with a preliminary discussion of $\widehat{\mathcal{W}}_r$.

LEMMA B.1. *Suppose that ρ is a norm on $\mathbb{R}^d$. Define the norm $\|\cdot\|_\rho$ via*

$$\|\widetilde{\boldsymbol{\gamma}}\|_\rho := |\gamma_{d+1}| + \rho(\boldsymbol{\gamma}),$$

for $\widetilde{\boldsymbol{\gamma}} = (\boldsymbol{\gamma},\gamma_{d+1})$, where $\boldsymbol{\gamma}\in\mathbb{R}^d$ and $\gamma\in\mathbb{R}$. Then

$$\widehat{W}_{r,\rho}(\mathbb{P},\widetilde{\mathbb{P}}) = \sup_{\widetilde{\boldsymbol{\gamma}}\in\mathbb{R}^{d+1}:\|\widetilde{\boldsymbol{\gamma}}\|_\rho=1} \mathcal{W}_r(\widetilde{\boldsymbol{\gamma}}_*\mathbb{P}\,,\widetilde{\boldsymbol{\gamma}}_*\widetilde{\mathbb{P}}) ,$$

and there exist positive constants c_1 and c_2 (may depend on the dimension d) such that

$$c_1\widehat{W}_{r,\rho}(\mathbb{P},\widetilde{\mathbb{P}}) \leq \overline{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) \leq c_2\widehat{W}_{r,\rho}(\mathbb{P},\widetilde{\mathbb{P}}).$$

PROOF. Denote $\mathbf{Z} = (\mathbf{X}^\top, Y)$ and $\widetilde{\mathbf{Z}} = (\widetilde{\mathbf{X}}^\top, \widetilde{Y})$. By definition

$$\begin{aligned}
\widehat{W}_{r,\rho}(\mathbb{P},\widetilde{\mathbb{P}}) &= \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d}\frac{1}{1+\rho(\boldsymbol{\gamma})}\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\mathbb{E}_\pi\left[|\mathbf{Z}^\top(\boldsymbol{\gamma},-1)-\widetilde{\mathbf{Z}}^\top(\boldsymbol{\gamma},-1)|^r\right]^{1/r}\\
&= \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d}\frac{1}{\|(\boldsymbol{\gamma},-1)\|_\rho}\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\mathbb{E}_\pi\left[|\mathbf{Z}^\top(\boldsymbol{\gamma},-1)-\widetilde{\mathbf{Z}}^\top(\boldsymbol{\gamma},-1)|^r\right]^{1/r}\\
&= \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d}\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\mathbb{E}_\pi\left[\left|(\mathbf{Z}-\widetilde{\mathbf{Z}})^\top\left(\frac{\boldsymbol{\gamma}}{\|(\boldsymbol{\gamma},-1)\|_\rho},\frac{-1}{\|(\boldsymbol{\gamma},-1)\|_\rho}\right)\right|^r\right]^{1/r}\\
&\leq \sup_{\widetilde{\boldsymbol{\gamma}}=(\boldsymbol{\gamma},\gamma_{d+1})\in\mathbb{R}^{d+1}:\|\widetilde{\boldsymbol{\gamma}}\|_\rho=1}\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\mathbb{E}_\pi\left[|\mathbf{Z}^\top\widetilde{\boldsymbol{\gamma}}-\widetilde{\mathbf{Z}}^\top\widetilde{\boldsymbol{\gamma}}|^r\right]^{1/r}\\
&= \sup_{\widetilde{\boldsymbol{\gamma}}=(\boldsymbol{\gamma},\gamma_{d+1})\in\mathbb{R}^{d+1}:\|\widetilde{\boldsymbol{\gamma}}\|_\rho=1}\mathcal{W}_r(\widetilde{\boldsymbol{\gamma}}_*\mathbb{P}\,,\widetilde{\boldsymbol{\gamma}}_*\widetilde{\mathbb{P}})\\
&= \sup_{\widetilde{\boldsymbol{\gamma}}=(\boldsymbol{\gamma},\gamma_{d+1})\in\mathbb{R}^{d+1}}\frac{1}{|\gamma_{d+1}|(1+\rho(\boldsymbol{\gamma}/|\gamma_{d+1}|))}\inf_\pi\mathbb{E}_\pi\left[|\mathbf{Z}^\top\widetilde{\boldsymbol{\gamma}}-\widetilde{\mathbf{Z}}^\top\widetilde{\boldsymbol{\gamma}}|^r\right]^{1/r}\\
&= \sup_{(\boldsymbol{\gamma},\gamma_{d+1})\in\mathbb{R}^{d+1}}\frac{1}{(1+\rho(\boldsymbol{\gamma}/|\gamma_{d+1}|))}\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\mathbb{E}_\pi\left[\left|(\mathbf{Z}-\widetilde{\mathbf{Z}})^\top\left(\frac{\boldsymbol{\gamma}}{|\gamma_{d+1}|},\frac{\gamma_{d+1}}{|\gamma_{d+1}|}\right)\right|^r\right]^{1/r}\\
&\leq \widehat{W}_{r,\rho}(\mathbb{P},\widetilde{\mathbb{P}}) .
\end{aligned}$$

This proves our first result. Now, to prove the second result we rely on the following representations.

$$\overline{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) = \sup_{\widetilde{\boldsymbol{\gamma}}\in\mathbb{R}^{d+1}:\|\widetilde{\boldsymbol{\gamma}}\|_2=1}\mathcal{W}_r(\widetilde{\boldsymbol{\gamma}}_*\mathbb{P}\,,\widetilde{\boldsymbol{\gamma}}_*\widetilde{\mathbb{P}}) = \sup_{\widetilde{\boldsymbol{\gamma}}\in\mathbb{R}^{d+1}}\frac{1}{\|\widetilde{\boldsymbol{\gamma}}\|_2}\mathcal{W}_r(\widetilde{\boldsymbol{\gamma}}_*\mathbb{P}\,,\widetilde{\boldsymbol{\gamma}}_*\widetilde{\mathbb{P}}),$$

and

$$\widehat{W}_{r,\rho}(\mathbb{P},\widetilde{\mathbb{P}}) = \sup_{\widetilde{\boldsymbol{\gamma}}\in\mathbb{R}^{d+1}}\frac{1}{\|\widetilde{\boldsymbol{\gamma}}\|_\rho}\mathcal{W}_r(\widetilde{\boldsymbol{\gamma}}_*\mathbb{P}\,,\widetilde{\boldsymbol{\gamma}}_*\widetilde{\mathbb{P}}) .$$

Because $\|\cdot\|_2$ and $\|\cdot\|_\rho$ are norms on $\mathbb{R}^{d+1}$, it follows that there exist positive constants c_1 and c_2 such that

$$\frac{c_1}{\|\widetilde{\boldsymbol{\gamma}}\|_\rho} \leq \frac{1}{\|\widetilde{\boldsymbol{\gamma}}\|_2} \leq \frac{c_2}{\|\widetilde{\boldsymbol{\gamma}}\|_\rho}\,, \quad \forall \widetilde{\boldsymbol{\gamma}} \in \mathbb{R}^{d+1}\,.$$

We conclude the second result by using the previous inequality and the representations for $\overline{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}})$ and $\widehat{W}_{r,\rho}(\mathbb{P},\widetilde{\mathbb{P}})$ presented above. □

More generally, $\widehat{\mathcal{W}}_r$ is always a metric, as the following lemma shows:

LEMMA B.2. *The ρ-max-sliced Wasserstein $\widehat{\mathcal{W}}_{r,\rho}$ distance is a metric.*

PROOF. Recall from (16) that

$$\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) = \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d} \frac{1}{1+\rho(\boldsymbol{\gamma})}\, \mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top \bar{\boldsymbol{\gamma}}\right]_* \mathbb{P},\, \left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top \bar{\boldsymbol{\gamma}}\right]_* \widetilde{\mathbb{P}}\right),$$

where $\bar{\boldsymbol{\gamma}}^\top = (\boldsymbol{\gamma}^\top, -1)$. Because the one-dimensional Wasserstein metric, $\mathcal{W}_r$, is non-negative, symmetric and satisfies the triangle inequality, the same is true for $\widehat{\mathcal{W}}_r$. It remains to show that $\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) = 0$ implies $\mathbb{P} = \widetilde{\mathbb{P}}$. For this, we first see that because $1+\rho(\boldsymbol{\gamma}) > 0$, it follows that $\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) = 0$ implies

$$\mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top \bar{\boldsymbol{\gamma}}\right]_* \mathbb{P},\, \left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top \bar{\boldsymbol{\gamma}}\right]_* \widetilde{\mathbb{P}}\right) = 0, \qquad \forall \boldsymbol{\gamma} \in \mathbb{R}^d. \tag{A.13}$$

Now, for any $\widetilde{\boldsymbol{\gamma}} \in \mathbb{R}^{d+1}$ satisfying $\|\widetilde{\boldsymbol{\gamma}}\|_2 = 1$ and $\widetilde{\gamma}_{d+1} \leq 0$, there exists a sequence $\{\boldsymbol{\gamma}_n\}_{n\in\mathbb{N}}$ in $\mathbb{R}^d$ such that

$$\lim_{n\to\infty} \frac{\bar{\boldsymbol{\gamma}}_n}{\|\bar{\boldsymbol{\gamma}}_n\|_2} = \widetilde{\boldsymbol{\gamma}},$$

where again $\bar{\boldsymbol{\gamma}}_n^\top := (\boldsymbol{\gamma}_n^\top, -1)$. By continuity, this implies

$$\mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top \widetilde{\boldsymbol{\gamma}}\right]_* \mathbb{P},\, \left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top \widetilde{\boldsymbol{\gamma}}\right]_* \widetilde{\mathbb{P}}\right) = 0, \qquad \forall \widetilde{\boldsymbol{\gamma}} \in \mathbb{R}^{d+1} \text{ with } \widetilde{\gamma}_{d+1} \leq 0;$$

and because $\left[(\mathbf{X},Y)^\top \widetilde{\boldsymbol{\gamma}}\right]_* \mathbb{P} = \left[(\mathbf{X},Y)^\top \widetilde{\boldsymbol{\gamma}}\right]_* \widetilde{\mathbb{P}}$ implies $\left[-(\mathbf{X},Y)^\top \widetilde{\boldsymbol{\gamma}}\right]_* \mathbb{P} = \left[-(\mathbf{X},Y)^\top \widetilde{\boldsymbol{\gamma}}\right]_* \widetilde{\mathbb{P}}$, we have

$$\mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top \widetilde{\boldsymbol{\gamma}}\right]_* \mathbb{P},\, \left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top \widetilde{\boldsymbol{\gamma}}\right]_* \widetilde{\mathbb{P}}\right) = 0, \qquad \forall \widetilde{\boldsymbol{\gamma}} \in \mathbb{R}^{d+1}.$$

Positivity of $\widehat{\mathcal{W}}_r$ now follows from the fact that $\mathcal{W}_r$ is positive. This concludes the proof. □

LEMMA B.3 (cf. [75, Theorem 4.10], [29, Chapter 4]). *Let*

$$\mathcal{H} := \left\{\mathbb{1}_{\{\mathbf{x}^\top \boldsymbol{\gamma} \leq t\}} : \boldsymbol{\gamma} \in \mathbb{R}^{d+1},\, t \in \mathbb{R}\right\} \tag{A.14}$$

be the set of indicator functions of half spaces. Then, with probably at least $1-\alpha$,

$$\sup_{(\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times\mathbb{R}} \left|F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)\right| = \sup_{f\in\mathcal{H}} \left|\mathbb{E}_{\mathbb{P}_n}[f] - \mathbb{E}_{\mathbb{P}}[f]\right| \leq 180\sqrt{\frac{d+2}{n}} + \sqrt{\frac{2}{n}\log\left(\frac{1}{\alpha}\right)}.$$

PROOF. By [75, Theorem 4.10], we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}} \left|\mathbb{E}_{\mathbb{P}_n}[f] - \mathbb{E}_{\mathbb{P}}[f]\right| > 2\mathcal{R}_n(\mathcal{H}) + \epsilon\right) \leq e^{-n\epsilon^2/2},$$

where

$$\mathcal{R}_n(\mathcal{H}) := \mathbb{E}_{\mathbb{P},\varepsilon}\left[\sup_{f\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(\mathbf{X}_i)\right|\right],$$

is the Rademacher complexity of $\mathcal{H}$. Next, following [29, statement and proof of Theorem 3.2], we obtain

$$\mathcal{R}_n(\mathcal{H}) \leq \frac{12}{\sqrt{n}} \max_{\mathbf{x}_1,\dots,\mathbf{x}_n \in \mathbb{R}^{d+1}} \int_0^1 \sqrt{2 \log N\left(\mathsf{r}, \mathcal{H}(\mathbf{x}_1^n)\right)}\, d\mathsf{r}, \tag{A.15}$$

where $\mathbf{x}_1^n := \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ and

$$\mathcal{H}(\mathbf{x}_1^n) := \{(f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)) : f \in \mathcal{H}\},$$

and $N(\mathsf{r}, B)$ is defined as the cardinality of the smallest cover for any set $B \subseteq \{0,1\}^n$ of radius r with respect to the distance

$$\rho(\mathbf{b}, \mathbf{d}) := \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{b_i \neq d_i\}}},$$

where in the above, vectors $\mathbf{b}, \mathbf{d} \in B$. [29, Theorem 4.3] states that

$$N(r, \mathcal{H}(\mathbf{x}_1^n)) \leq \left(\frac{4e}{r^2}\right)^{V/(1-1/e)} = \left(\frac{4e}{r^2}\right)^{Ve/(e-1)}, \tag{A.16}$$

where V is the VC-dimension of $\mathcal{H}$. Furthermore, by [29, Corollary 4.2], the VC-dimension of $\mathcal{H}$ is bounded by $d+2$, i.e. $V \leq d+2$. In conclusion, using (A.16),

$$\log N\left(\mathsf{r}, \mathcal{H}(\mathbf{x}_1^n)\right) \leq \frac{eV}{e-1} \log\left(\frac{4e}{\mathsf{r}^2}\right) \leq \frac{e(d+2)}{e-1} \log\left(\frac{4e}{\mathsf{r}^2}\right). \tag{A.17}$$

Following [29, proof of Theorem 3.3] we estimate

$$\int_0^1 \sqrt{\log\left(\frac{4e}{\mathsf{r}^2}\right)}\, d\mathsf{r} \leq \sqrt{2\pi e}, \tag{A.18}$$

so that from (A.17) and (A.18), we have

$$\int_0^1 \sqrt{2 \log N(\mathsf{r}, \mathcal{H}(\mathbf{x}_1^n))}\, d\mathsf{r} \leq 2e \sqrt{\frac{(d+2)\pi}{e-1}} \leq 7.5\sqrt{d+2}.$$

Using (A.15), this yields

$$\mathcal{R}_n(\mathcal{H}) \leq 90 \sqrt{\frac{d+2}{n}}.$$

□

LEMMA B.4. *Define*

$$\Gamma_n := \sup_{\|\boldsymbol{\gamma}\|=1} \mathbb{E}_{\mathbb{P}_n}\left[\left|(\mathbf{X}, Y)^\top \boldsymbol{\gamma}\right|^s\right] = \sup_{\|\boldsymbol{\gamma}\|=1} \frac{1}{n} \sum_{i=1}^n \left|(\mathbf{X}, Y)_i^\top \boldsymbol{\gamma}\right|^s.$$

For any $k \in \mathbb{R}^+$, we have

$$\overline{\mathcal{W}}_r(\mathbb{P}_n, \mathbb{P})^r \leq r(2k)^r \sup_{(\boldsymbol{\gamma},t) \in \mathbb{R}^{d+1} \times \mathbb{R}} \left|F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))\right|$$

$$+ \frac{2^r r \sqrt{\Gamma \vee \Gamma_n}}{s/2 - r} k^{r - \frac{s}{2}} \left[\sup_{(\boldsymbol{\gamma},t) \in \mathbb{R}^{d+1} \times \mathbb{R}} \frac{(F_{\boldsymbol{\gamma}}(t) - F_{\boldsymbol{\gamma},n}(t))^+}{\sqrt{F_{\boldsymbol{\gamma}}(t)(1 - F_{\boldsymbol{\gamma},n}(t))}} + \sup_{(\boldsymbol{\gamma},t) \in \mathbb{R}^{d+1} \times \mathbb{R}} \frac{(F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))^+}{\sqrt{F_{\boldsymbol{\gamma},n}(t)(1 - F_{\boldsymbol{\gamma}}(t))}}\right],$$

with the convention that $0/0 = 0$ and the notation $x^+ := \max\{0, x\}$.

PROOF. We first note that [14, Proposition 7.14] yields, for any $k > 0$,

(A.19)
$$\begin{aligned}\mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}})^r &\le r2^{r-1} \int |t|^{r-1} |F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)|\, dt \\ &\le r(2k)^r \sup_t |F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))| \\ &\quad + r2^{r-1} \int_{\mathbb{R}\setminus[-k,k]} |t|^{r-1} \sqrt{F_{\boldsymbol{\gamma}}(t)(1-F_{\boldsymbol{\gamma},n}(t))}\, \frac{(F_{\boldsymbol{\gamma}}(t) - F_{\boldsymbol{\gamma},n}(t))^+}{\sqrt{F_{\boldsymbol{\gamma}}(t)(1-F_{\boldsymbol{\gamma},n}(t))}}\, dt \\ &\quad + r2^{r-1} \int_{\mathbb{R}\setminus[-k,k]} |t|^{r-1} \sqrt{F_{\boldsymbol{\gamma},n}(t)(1-F_{\boldsymbol{\gamma}}(t))}\, \frac{(F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))^+}{\sqrt{F_{\boldsymbol{\gamma},n}(t)(1-F_{\boldsymbol{\gamma}}(t))}}\, dt.\end{aligned}$$

By Markov's inequality, we have for any $s \ge 1$ and any $t \in \mathbb{R} \setminus \{0\}$,

$$\sqrt{F_{\boldsymbol{\gamma}}(t)(1-F_{\boldsymbol{\gamma},n}(t))} \vee \sqrt{F_{\boldsymbol{\gamma},n}(t)(1-F_{\boldsymbol{\gamma}}(t))} \le \sqrt{\frac{\mathbb{E}_{\mathbb{P}}[|(\mathbf{X},Y)^\top \boldsymbol{\gamma}|^s] \vee \mathbb{E}_{\mathbb{P}_n}[|(\mathbf{X},Y)^\top \boldsymbol{\gamma}|^s]}{|t|^s}}.$$

Plugging these bounds into (A.19), we obtain

(A.20)
$$\begin{aligned}&\mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}})^r \\ &\le r(2k)^r \sup_t |F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))| \\ &+ r2^{r-1} \int_{\mathbb{R}\setminus[-k,k]} |t|^{r-1-s/2} \sqrt{\mathbb{E}_{\mathbb{P}}[|(\mathbf{X},Y)^\top \boldsymbol{\gamma}|^s] \vee \mathbb{E}_{\mathbb{P}_n}[|(\mathbf{X},Y)^\top \boldsymbol{\gamma}|^s]}\, \frac{(F_{\boldsymbol{\gamma}}(t) - F_{\boldsymbol{\gamma},n}(t))^+}{\sqrt{F_{\boldsymbol{\gamma}}(t)(1-F_{\boldsymbol{\gamma},n}(t))}}\, dt \\ &+ r2^{r-1} \int_{\mathbb{R}\setminus[-k,k]} r|t|^{r-1-s/2} \sqrt{\mathbb{E}_{\mathbb{P}}[|(\mathbf{X},\mathbf{Y})^\top \boldsymbol{\gamma}|^s] \vee \mathbb{E}_{\mathbb{P}_n}[|(\mathbf{X},Y)^\top \boldsymbol{\gamma}|^s]}\, \frac{(F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))^+}{\sqrt{F_{\boldsymbol{\gamma},n}(t)(1-F_{\boldsymbol{\gamma}}(t))}}\, dt.\end{aligned}$$

Recall that we have assumed $s/2 > r$ where $r \ge 1$. In particular, this means that $|t|^{r-1-s/2}$ is integrable on $\mathbb{R} \setminus [-k,k]$ and

$$r2^{r-1} \int_{\mathbb{R}\setminus[-k,k]} |t|^{r-1-s/2}\, dt = \frac{2^r r}{s/2 - r} k^{r-s/2}.$$

Taking the supremum over $\boldsymbol{\gamma}$ and t in (A.20) thus yields the claim. □

LEMMA B.5. *With probability greater than* $1-\alpha$ *we have*

$$\begin{aligned}&\sup_{(\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times\mathbb{R}} \frac{(F_{\boldsymbol{\gamma}}(t) - F_{\boldsymbol{\gamma},n}(t))^+}{\sqrt{F_{\boldsymbol{\gamma}}(t)(1-F_{\boldsymbol{\gamma},n}(t))}} \vee \sup_{(\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times\mathbb{R}} \frac{(F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))^+}{\sqrt{F_{\boldsymbol{\gamma},n}(t)(1-F_{\boldsymbol{\gamma}}(t))}} \\ &\le 4\sqrt{\frac{\log(8/\alpha) + (d+2)\log(2n+1)}{n}}.\end{aligned}$$

PROOF. We first define

$$\mathcal{J} = \left\{ \mathbb{1}_{\{\mathbf{x}^\top \boldsymbol{\gamma} \le t\}}, \mathbb{1}_{\{\mathbf{x}^\top \boldsymbol{\gamma} > t\}} : (\boldsymbol{\gamma}, t) \in \mathbb{R}^{d+1} \times \mathbb{R} \right\} \supseteq \mathcal{H},$$

where $\mathcal{H}$ was defined in Lemma B.3. Considering the cases $F_{\boldsymbol{\gamma},n}(t) < 1/2$ and $F_{\boldsymbol{\gamma},n}(t) \ge 1/2$ separately—noting that e.g. $\mathbb{E}_{\mathbb{P}_n}[\mathbb{1}_{\{\mathbf{x}^\top\boldsymbol{\gamma}>t\}}] = 1 - \mathbb{E}_{\mathbb{P}_n}[\mathbb{1}_{\{\mathbf{x}^\top\boldsymbol{\gamma}\le t\}}] = 1 - F_{\boldsymbol{\gamma},n}(t)$—one can check

(A.21)
$$\sup_{(\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times\mathbb{R}} \frac{(F_{\boldsymbol{\gamma}}(t) - F_{\boldsymbol{\gamma},n}(t))^+}{\sqrt{F_{\boldsymbol{\gamma}}(t)(1-F_{\boldsymbol{\gamma},n}(t))}} \le 2 \left(\sup_{f\in\mathcal{J}} \frac{(\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{P}_n}[f])^+}{\sqrt{\mathbb{E}_{\mathbb{P}}[f]}} \vee \sup_{f\in\mathcal{J}} \frac{(\mathbb{E}_{\mathbb{P}_n}[f] - \mathbb{E}_{\mathbb{P}}[f])^+}{\sqrt{\mathbb{E}_{\mathbb{P}_n}[f]}} \right).$$

By symmetry,

(A.22)
$$\sup_{(\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times\mathbb{R}} \frac{(F_{\boldsymbol{\gamma},n}(t)-F_{\boldsymbol{\gamma}}(t))^+}{\sqrt{F_{\boldsymbol{\gamma},n}(t)(1-F_{\boldsymbol{\gamma}}(t))}} \leq 2\left(\sup_{f\in\mathcal{J}} \frac{(\mathbb{E}_{\mathbb{P}}[f]-\mathbb{E}_{\mathbb{P}_n}[f])^+}{\sqrt{\mathbb{E}_{\mathbb{P}}[f]}} \vee \sup_{f\in\mathcal{J}} \frac{(\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f])^+}{\sqrt{\mathbb{E}_{\mathbb{P}_n}[f]}}\right).$$

Concentration for the terms on the right hand side of equations (A.21) and (A.22) is well studied: indeed, e.g. by [29, Exercises 3.3 & 3.4] we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{J}} \frac{\mathbb{E}_{\mathbb{P}}[f]-\mathbb{E}_{\mathbb{P}_n}[f]}{\sqrt{\mathbb{E}_{\mathbb{P}}[f]}} > \epsilon\right) \leq 4S_{\mathcal{J}}(2n)e^{-n\epsilon^2/4},$$
$$\mathbb{P}\left(\sup_{f\in\mathcal{J}} \frac{\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f]}{\sqrt{\mathbb{E}_{\mathbb{P}_n}[f]}} > \epsilon\right) \leq 4S_{\mathcal{J}}(2n)e^{-n\epsilon^2/4}.$$

for all $\epsilon > 0$, where $S_{\mathcal{J}}(2n)$ is the shattering coefficient of $\mathcal{J}$. Note that by [29, Theorem 4.1] we have $S_{\mathcal{J}}(2n) \leq 2S_{\mathcal{H}}(2n)$. As the VC-dimension of $\mathcal{H}$ is bounded by $d+2$, Sauer's lemma [29, Theorem Corollary 4.1] yields

$$\log(S_{\mathcal{J}}(2n)) \leq (d+2)\log(2n+1).$$

The claim follows by solving the above expression for ϵ. □

APPENDIX C: ADDITIONAL DERIVATIONS

C.1. Diameter of the support of $\mathbb{P}$ in the simulation. Notice that $\widetilde{\mathbf{X}}_i^\top \boldsymbol{\beta} \geq -\|\boldsymbol{\beta}^-\|_1$, where equality holds for $\widetilde{\mathbf{X}}_i$ that has ones in the entries corresponding to negative values of $\boldsymbol{\beta}$, and zeros otherwise. Similarly, $\widetilde{\mathbf{X}}_i^\top \boldsymbol{\beta} \leq \|\boldsymbol{\beta}^+\|_1$. Since $\mathbf{X}_i = \sigma_\varepsilon\lambda\widetilde{\mathbf{X}}_i$, it follows that $\inf_{\mathbf{X}_i} X_i^\top\boldsymbol{\beta} = -\sigma_\varepsilon\lambda\|\boldsymbol{\beta}^-\|_1$ and $\sup_{\mathbf{X}_i} \mathbf{X}_i^\top\boldsymbol{\beta} = \sigma_\varepsilon\lambda\|\boldsymbol{\beta}^+\|_1$. Therefore, $Y_i \in [-\sigma_\varepsilon(\lambda\|\boldsymbol{\beta}^-\|_1+1), \sigma_\varepsilon(\lambda\|\boldsymbol{\beta}^+\|_1+1)]$. Then, diameter of the support equals

$$\sqrt{d(\sigma_\varepsilon\lambda)^2 + \sigma_\varepsilon^2(\lambda\|\boldsymbol{\beta}^+\|_1 + \lambda\|\boldsymbol{\beta}^-\|_1+2)^2} = \sigma_\varepsilon\lambda\sqrt{d+(\|\boldsymbol{\beta}\|_1+2/\lambda)^2}.$$

C.2. Derivation of Equation (32). The tuning parameter $\delta_{n,2}$ used in the simulation is

$$\delta_{n,2} = n^{-1/4}\cdot C_{\text{sim}}, \quad \text{where } C_{\text{sim}} \equiv (q_{1-\alpha})^{1/2}\cdot\sigma_\varepsilon\lambda\left(d+(\|\boldsymbol{\beta}\|_1+(2/\lambda))^2\right)^{1/2}.$$

According to equation (31) it is known that $\boldsymbol{\beta} = 0_{d\times 1}$ is a solution to the $\sqrt{\text{LASSO}}$ problem if and only if

(A.23)
$$\frac{\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i y_i\|_\infty}{\sqrt{\frac{1}{n}\sum_{i=1}^n y_i^2}} \leq \delta_{n,2} = n^{-1/4}\cdot C_{\text{sim}}.$$

Because

$$Y_i = \mathbf{X}_i^\top\boldsymbol{\beta} + \sigma_\varepsilon\varepsilon_i,$$

equation (A.23) holds if and only if

(A.24)
$$\frac{\|\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\right)\boldsymbol{\beta} + \frac{\sigma_\varepsilon}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{X}_i\varepsilon_i\|_\infty}{\sqrt{\frac{1}{n}\sum_{i=1}^n y_i^2}} \leq n^{-1/4}\cdot C_{\text{sim}}.$$

Because $\mathbf{X}_i = \sigma_\varepsilon\lambda\widetilde{\mathbf{X}}_i$ where $\widetilde{\mathbf{X}}_i$ is a d-dimensional vector of independent uniform random variables over the $[0,1]$, then

$$\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top \xrightarrow{p} \mathbb{E}[\mathbf{X}_i\mathbf{X}_i^\top] = \frac{1}{3}\sigma_\varepsilon^2\lambda^2\mathbb{I}_d,$$

where we have used that $\mathbb{E}[\widetilde{\mathbf{X}}_{i,j}^2] = 1/3$ because $\widetilde{\mathbf{X}}_{i,j}$ is a uniform distribution on the $[0,1]$ interval. The Continuous Mapping Theorem and Central Limit Theorem then imply that

$$\left\| \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right) \boldsymbol{\beta} + \frac{\sigma_\varepsilon}{\sqrt{n}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{X}_i \varepsilon_i \right\|_\infty \xrightarrow{p} \frac{1}{3} \sigma_\varepsilon^2 \lambda^2 \|\boldsymbol{\beta}\|_\infty.$$

For the denominator,

$$\frac{1}{n} \sum_{i=1}^{n} y_i^2 \xrightarrow{p} \mathbb{E}[Y_i^2] = \boldsymbol{\beta}^\top \mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top] \boldsymbol{\beta} + \sigma_\varepsilon^2 \mathbf{V}(\varepsilon_i) = \frac{1}{3} \sigma_\varepsilon^2 \lambda^2 \boldsymbol{\beta}^\top \boldsymbol{\beta} + \sigma_\varepsilon^2 \mathbf{V}(\varepsilon_i),$$

where we have used the fact that ϵ_i is mean zero and independent of $\widetilde{\mathbf{X}}_i$. The left-hand side of (A.24) is thus bounded above with high probability by

$$\frac{(1/3)\sigma_\varepsilon^2 \lambda^2 \|\boldsymbol{\beta}\|_\infty}{\sqrt{(1/3)\sigma_\varepsilon^2 \lambda^2} \sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}} = \sqrt{1/3} \cdot \sigma_\varepsilon \cdot \lambda \cdot \left\| \frac{\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}} \right\|_\infty.$$

This means that the event in (A.23) occurs with high probability if

$$\sqrt{1/3} \cdot \sigma_\varepsilon \cdot \lambda \cdot \left\| \frac{\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}} \right\|_\infty \leq \frac{1}{n^{1/4}} C_{\text{sim}}. \tag{A.25}$$

Using the definition of C_{sim}, the event in (A.25) occurs if and only if

$$n \leq 9 \cdot \left\| \frac{\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}} \right\|_\infty^{-1/4} (q_{1-\alpha})^2 \cdot \left(d + (\|\boldsymbol{\beta}\|_1 + (2/\lambda))^2 \right)^2.$$

Thus, a sample size smaller than the right-hand side of the equation above implies that, with high probability, $\boldsymbol{\beta} = 0_{d \times 1}$ will be a solution to the $\sqrt{\text{LASSO}}$ problem.

C.3. Gaussian distributions with similar prediction errors, but large Wasserstein distance. Suppose $(\mathbf{X}, Y) \sim N_{d+1}(0, \mathbb{I}_{d+1})$. Because $(\mathbf{X}, Y)$ are, by assumption, independent and have mean zero, the prediction error of any linear predictor $\mathbf{X}^\top \boldsymbol{\gamma}$ equals the variance of Y plus the variance of $\mathbf{X}^\top \boldsymbol{\gamma}$; that is

$$\mathbb{E}[(Y - \mathbf{X}^\top \boldsymbol{\gamma})^2] = 1 + \|\boldsymbol{\gamma}\|_2^2.$$

The prediction error scales with $\|\boldsymbol{\gamma}\|_2$, so it makes sense to restrict this norm. Let us focus on predictors for which $\|\boldsymbol{\gamma}\|_2 = 1$.

Consider now a random vector $(\widetilde{\mathbf{X}}, \widetilde{Y})$. Assume $\widetilde{\mathbf{X}} = \mathbf{X} + \mathbf{V}$ where $\mathbf{V} \sim N_d(0, \sigma_v^2 \mathbb{I}_d)$ is an independent source of measurement error. Set $\widetilde{Y} = Y$. For any $\boldsymbol{\gamma}$ such that $\|\boldsymbol{\gamma}\| = 1$:

$$\mathbb{E}\left[(\widetilde{Y} - \widetilde{\mathbf{X}}^\top \boldsymbol{\gamma})^2 \right] = 2 + \sigma_v^2.$$

Thus, in this example

$$\left(\mathbb{E}\left[(Y - \mathbf{X}^\top \boldsymbol{\gamma})^2 \right] - \mathbb{E}\left[(\widetilde{Y} - \widetilde{\mathbf{X}}^\top \boldsymbol{\gamma})^2 \right] \right) = \sigma_v^2.$$

Consequently, the difference in prediction errors equals σ_v^2.

Let $\mathbb{P}$ denote the distribution of $(\mathbf{X}, Y)$ and, analogously, let $\widetilde{\mathbb{P}}$ denote the distribution of $(\widetilde{\mathbf{X}}, \widetilde{Y})$. The distance between $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ can be considerably large when measured using the standard d-dimensional Wasserstein metric. In fact, algebra shows that

$$\mathcal{W}_2(\mathbb{P}, \widetilde{\mathbb{P}}) = (\sqrt{1 + \sigma_v^2} - 1) d^{1/2},$$

Thus, when d is large, the standard d-dimensional Wasserstein distance suggests that $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ are very different from one another. This stands in contrast with the magnitude of the difference in prediction errors associated to $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ which is equal to σ_v^2.

In this example, one can further show that if we set $\rho(\cdot) = \|\cdot\|_1$, then

$$\widehat{\mathcal{W}}_2(\mathbb{P}, \widetilde{\mathbb{P}}) \leq \sigma_v \ .$$

Therefore, the example further shows that two distributions can be close in ρ-MSW metric, even when their standard d-dimensional Wasserstein distance is large.

C.4. Comparison to Proposition 6 and Theorem 4 in [10]. In this section we explain that there are a few major differences between our bounds for the out-of-distribution prediction error and those provided in [10] (in particular, Proposition 6 and Theorem 4 therein).

First, [10, Proposition 6] assumes that $(\mathbf{X}, Y) \sim \mathbb{P}$ and their result makes reference to an underlying linear regression model $Y = \mathbf{X}^T\boldsymbol{\beta}_* + \epsilon$, where $\boldsymbol{\beta}_* := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{E}_{\mathbb{P}}[|Y - \mathbf{X}^T\boldsymbol{\beta}|^2]$ can be interpreted as the population's best linear predictor. The authors then provide bounds for $\mathbb{E}_{\mathbb{P}_n}[|Y - \mathbf{X}^T\boldsymbol{\beta}_*|^2]$: the *in-sample* prediction error at $\boldsymbol{\beta}_*$. We emphasize that in our setting we never focus on generalization bounds that are *only* valid for $\boldsymbol{\beta}_*$. Moreover, since our focus is *out-of-distribution* prediction error, we are naturally interested in distributions $\mathbb{Q}$ that are close, but different, than $\mathbb{P}$ (and also different than $\mathbb{P}_n$).

Second, [10, Theorem 4] states an asymptotic confidence bound for $\mathbb{E}_{\mathbb{P}}[|Y - \mathbf{X}^T\boldsymbol{\beta}_*|^2]$ that is valid with high probability as $n \to \infty$. Our Theorem 5.1 holds for finite $n \in \mathbb{N}$, with high probability, and is uniform in $\boldsymbol{\beta}$. In particular, we give generalization bounds for $\mathbb{E}_{\mathbb{Q}}[|Y - \mathbf{X}^T\widehat{\boldsymbol{\beta}}|^2]$, where $\widehat{\boldsymbol{\beta}}$ is an estimator constructed using the training data (in our results, we focus on the estimators that solve (1)). It is not clear to us how to recover results of this type from [10, Theorem 4], even for $n \to \infty$.

It is important to mention that in [10], the regularization parameter is not selected based on the Wasserstein distance $\mathcal{W}_r(\mathbb{P}_n, \mathbb{P})$ between the empirical distribution and the true data-generating distribution. Instead, they consider the distance

$$R_n(\boldsymbol{\beta}) = \min\{\mathcal{W}^B(\mathbb{P}_n, \mathbb{Q}) : \boldsymbol{\gamma} \mapsto \mathbb{E}_{\mathbb{Q}}[|Y - \mathbf{X}^\top\boldsymbol{\gamma}|^r] \text{ has minimizer } \boldsymbol{\beta}\}, \tag{A.26}$$

where $\mathcal{W}^B$ is a type of Wasserstein-like distance (see [10, eq. (16)]; more discussion below) and then consider the regularization $\delta > R_n(\boldsymbol{\beta}_*)$, where $\boldsymbol{\beta}_*$ (as defined above) is the coefficient of an underlying linear regression model.

The work in [10] derives bounds for this distance: [10, Theorem 7 and Remark 1] states that, under Gaussian additive errors ϵ, an appropriate normalization of $\mathbf{X}$, and if $Y = \mathbf{X}^T\boldsymbol{\beta}^* + \epsilon$ under $\mathbb{P}$ (see beginning of [10, Section 4.2]), then one has

$$\sqrt{R_n(\boldsymbol{\beta}^*)} \leq \frac{\pi}{\pi - 2}\frac{\Phi^{-1}(1 - \alpha/2d)}{\sqrt{n}}, \tag{A.27}$$

which aligns with recommendations of [6].

As mentioned above, $\mathcal{W}^B$ is *not* the standard $(d+1)$-dimensional Wasserstein distance. Instead, it only allows for fluctuations of the $\mathbf{X}$-values, not the Y-values (see [10, eq. (14)]). In consequence, a ball in $\mathcal{W}^B$ around $\mathbb{P}_n$ will *never* contain the true distribution $\mathbb{P}$ if Y has a continuous distribution under $\mathbb{P}$ (as all measures in the $\mathcal{W}^B$-ball must have the same Y values). This is the reason why no generalization bounds at deviations from $\mathbb{P}$ (analogous to Theorem 5.1 in the main body of the paper) can be derived from the results of [10].

Lastly, the distribution $\mathbb{Q}$ achieving the minimum in (A.26) is usually *not* $\mathbb{P}$, but can be arbitrarily far away from $\mathbb{P}$ in the usual Wasserstein distance. See [10, Section 1.1.3] for an extended discussion of $R_n(\boldsymbol{\beta})$. In that sense, bounds for $R_n(\boldsymbol{\beta})$ are concerned with the recovery of β_* in the linear model $Y = \mathbf{X}^T\boldsymbol{\beta}_* + \epsilon$. In fact, [10] acknowledges that a similar strategy as the one we suggest for selecting δ, but using the standard Wasserstein metric, would only yield a recommendation of order $O(n^{-1/d})$; see their discussion after Theorem 4, p. 848 and [48, Theorem 4].

C.5. Alternative criterion for choosing δ and a new generalization-type bound. As mentioned in Section C.4, there are a couple of decisive differences between [10] and our results. In this section, we present a brief discussion of the differences between our recommendation for selecting δ (which specifically targets out-of-distribution performance) and the recommendation in [10]. We show that if we use the same criterion as in [10], our optimal δ would be upper bounded by their

recommendation. As we explain below, this has to do with the fact that our ρ-MSW balls are larger than those based on the standard Wasserstein metric.

Distributionally robust representation: First, it is worth mentioning that both our paper and [10] present a distributionally robust representation of the $\sqrt{\text{LASSO}}$ and related estimators. The key difference is that we define our class of testing distributions using $\widehat{\mathcal{W}}_r$, instead of the Wasserstein metric $\mathcal{W}^B$ defined in Section C.4. Recall that [10] take the same testing and training distributions of the outcome variable; see their Proposition 2, Equation (14), Theorem 1. Thus, to make our results comparable to them we set $\sigma = 0$ and use the notation $\widehat{\mathcal{W}}_{r,\rho,0}$.

Criterion: [10] recommend δ_n^* as the $1-\alpha$ quantile of the profile function $R_n(\boldsymbol{\beta})$ defined in (A.26), and (A.27) holds with probability asymptotically larger than $1-\alpha$, as $n \to \infty$, see [10, Theorem 7 and Remark 1]. This aligns with the recommendation of [5].

It is important to note that even though Theorem 5 in [10] characterizes the exact asymptotic distribution of $nR_n(\boldsymbol{\beta}_*)$, their recommended tuning parameter is based on a stochastic upper bound for this distribution (see their Remark 1). Note that using the exact asymptotic distribution presents at least three complications. First, simulating the quantile of the exact distribution involves solving repeatedly an optimization problem in $\mathbb{R}^d$ that in principle requires estimators of $\boldsymbol{\beta}_*$ and the variance of e (solving these optimization problems could be computationally demanding). Second, the quantiles of the asymptotic distribution could very well depend on the variance of ϵ, which would mean that the recommended tuning parameters based on the exact asymptotic distribution need not be pivotal (hence, deviate from the recommendation of [5]). Third, since the selection of tuning parameters is intended to "cover" the true parameter $\boldsymbol{\beta}_*$, it is difficult to obtain bounds on the out-of-distribution prediction error as the ones we present in Theorems 5.1.

We now consider a modification of the profile function $R_n(\boldsymbol{\beta})$ based on the ρ-MSW metric as follows:

$$\widetilde{R}_n(\boldsymbol{\beta}) = \min\{\widehat{\mathcal{W}}_{r,\rho,0}(\mathbb{P}_n, \mathbb{Q}) : \boldsymbol{\gamma} \mapsto \mathbb{E}_{\mathbb{Q}}[|Y - \mathbf{X}^\top \boldsymbol{\gamma}|^r] \text{ has minimizer } \boldsymbol{\beta}\} .$$

If we define by $\widetilde{\delta}_n^*$ the $1-\alpha$ quantile of the modified profile function $\widetilde{R}_n(\boldsymbol{\beta})$, the event $\{R_n(\boldsymbol{\beta}) \leq z\}$ is included in the event $\{\widetilde{R}_n(\boldsymbol{\beta}) \leq z\}$ because for any $\mathbb{Q}$ such that $\boldsymbol{\beta}$ is a minimizer of $\boldsymbol{\gamma} \mapsto \mathbb{E}_{\mathbb{Q}}[|Y - \mathbf{X}^\top \boldsymbol{\gamma}|^r]$ we have that $\widehat{\mathcal{W}}_{r,\rho,0}(\mathbb{P}_n, \mathbb{Q}) \leq \mathcal{W}^B(\mathbb{P}_n, \mathbb{Q})$ due to Remark 4. This implies that

$$\mathbb{P}(\widetilde{R}_n(\boldsymbol{\beta}) \leq z) \geq \mathbb{P}(R_n(\boldsymbol{\beta}) \leq z) ,$$

which allows us to conclude $\widetilde{\delta}_n^* \leq \delta_n^*$. That is, if we were interested in the criterion used in [10] (and not in the out-of-distribution prediction error) we could use our results to recommend a smaller regularization parameter.

Next, we complement the result in Theorem 5.1 with upper and lower bounds on the out-of-distribution prediction error of linear predictors that solve (1) using their worst-case prediction error in the training data. To this end, for any given $\mathbb{Q}$ and $\boldsymbol{\beta}$ (which could be data dependent; i.e. $\boldsymbol{\beta} \equiv \boldsymbol{\beta}(\mathbb{P}_n)$) define

$$\Delta_{n,r}(\mathbb{Q}, \boldsymbol{\beta}) := \mathbb{E}_{\mathbb{Q}}\left[\left|Y - \mathbf{X}^\top \boldsymbol{\beta}\right|^r\right]^{1/r} - \left(\sup_{\widetilde{\mathbb{P}} \in B_{\delta_{n,r}}(\mathbb{P}_n)} \mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y - \mathbf{X}^\top \boldsymbol{\beta}\right|^r\right]\right)^{1/r} ,$$

where $B_{\delta_{n,r}}(\mathbb{P}_n)$ denotes a ball (based on the ρ-MSW metric) of radius $\delta_{n,r}$ around the empirical distribution.

COROLLARY C.1. *Suppose the conditions of Theorem 3.2 (or 3.1) hold. Consider $\delta_{n,r}$ defined in (24) (or (25)). Then, for any $\epsilon \geq 0$ and $\mathbb{Q}$ such that $\widehat{\mathcal{W}}_r(\mathbb{P}, \mathbb{Q}) \leq \epsilon$, with probability greater than $1 - 3\alpha$, we have for all $\boldsymbol{\beta}$*

$$-(2\delta_{n,r} + \epsilon)(1 + \rho(\boldsymbol{\beta}))) \leq \Delta_{n,r}(\mathbb{Q}, \boldsymbol{\beta}) \leq \epsilon(1 + \rho(\boldsymbol{\beta})).$$

In particular, the inequality above holds for $\widehat{\boldsymbol{\beta}}$ that solve (1).

PROOF. By the proof of Theorem 5.1, there is an event with probability greater than $1-3\alpha$ such that $\widehat{\mathcal{W}}_r(\mathbb{P},\mathbb{P}_n) \leq \delta_{n,r}$ holds. Conditional on this event, consider first the following derivations,

$$\begin{aligned}|\Delta_{n,r}(\mathbb{Q},\boldsymbol{\beta})| \leq & \underbrace{|\mathbb{E}_{\mathbb{Q}}[|Y-\mathbf{X}^T\boldsymbol{\beta}|^r]^{1/r} - \mathbb{E}_{\mathbb{P}}[|Y-\mathbf{X}^T\boldsymbol{\beta}|^r]^{1/r}|}_{\text{(I)}} \\ & + \underbrace{|\mathbb{E}_{\mathbb{P}}[|Y-\mathbf{X}^T\boldsymbol{\beta}|^r]^{1/r} - \mathbb{E}_{\mathbb{P}_n}[|Y-\mathbf{X}^T\boldsymbol{\beta}|^r]^{1/r}|}_{\text{(II)}} \\ & + \underbrace{|\mathbb{E}_{\mathbb{P}_n}[|Y-\mathbf{X}^T\boldsymbol{\beta}|^r]^{1/r} - \sup_{\widetilde{\mathbb{P}}\in B_{\delta_{n,r}(\mathbb{P}_n)}} \mathbb{E}_{\widetilde{\mathbb{P}}}[|Y-\mathbf{X}^T\boldsymbol{\beta}|^r]^{1/r}|}_{\text{(III)}} .\end{aligned}$$

Note (I) $\leq \widehat{\mathcal{W}}_r(\mathbb{Q},\mathbb{P})(1+\rho(\boldsymbol{\beta})) \leq \epsilon(1+\rho(\boldsymbol{\beta}))$ by (11), (II) $\leq \widehat{\mathcal{W}}_r(\mathbb{P},\mathbb{P}_n)(1+\rho(\boldsymbol{\beta})) \leq \delta_{n,r}(1+\rho(\boldsymbol{\beta}))$ by (11), and (III) $\leq \delta_{n,r}(1+\rho(\boldsymbol{\beta}))$ by Theorem 2.1.

Conditional on the same event as above, we now consider the following derivations,

$$\Delta_{n,r}(\mathbb{Q},\boldsymbol{\beta}) \overset{(1)}{\leq} \mathbb{E}_{\mathbb{Q}}[|Y-\mathbf{X}^T\boldsymbol{\beta}|^r]^{1/r} - \mathbb{E}_{\mathbb{P}}[|Y-\mathbf{X}^T\boldsymbol{\beta}|^r]^{1/r} \overset{(2)}{\leq} \widehat{\mathcal{W}}_r(\mathbb{Q},\mathbb{P})(1+\rho(\boldsymbol{\beta})) \leq \epsilon(1+\rho(\boldsymbol{\beta})),$$

where (1) holds since $\widehat{\mathcal{W}}_r(\mathbb{P},\mathbb{P}_n) \leq \delta_{n,r}$ and (2) holds by (11). The claim follows. □

APPENDIX D: ADDITIONAL SIMULATIONS

Suppose that the training data consists of n i.i.d. draws from a Gaussian, homoskedastic, linear regression model. In other words,

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \sigma_\varepsilon \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0,1)$ and $\mathbf{X}_i \sim \mathcal{N}_d(\mathbf{0},\mathbb{I}_d)$, with $\varepsilon_i \perp \mathbf{X}_i$. The parameters controlling the simulation design are $(\boldsymbol{\beta},\sigma_\epsilon,d)$.

We are interested in comparing the out-of-distribution performance of a linear predictor that uses coefficients estimated via the $\sqrt{\text{LASSO}}$ ($r=2$), with other popular regularization procedures (Ridge regression and the LASSO). We use the standard tuning parameter for the $\sqrt{\text{LASSO}}$ in [6]; namely,

$$\widetilde{\delta}_n \equiv (1.1)\cdot\Phi^{-1}\left(1-\frac{\alpha}{2d}\right)\cdot n^{-1/2}, \tag{A.28}$$

and we take $\alpha = 0.05$. For Ridge regression, we use the approximately optimal oracle recommendation in Corollary 6 of [42].[5] For the LASSO, we use the oracle recommendation of [5], which equals $\widetilde{\delta}_n$ multiplied by the unknown parameter σ_ϵ.[6]

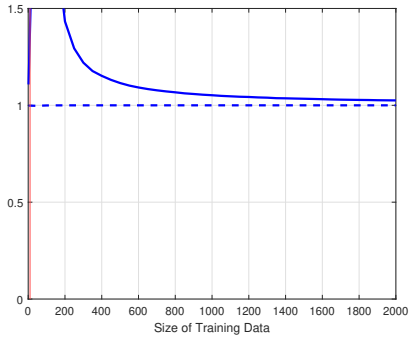
We consider different sizes of the training data $n_{\text{train}} \in \{2,50,100,150,\ldots,2000\}$. We set the size of the testing data to be $n_{\text{testing}} = 1000$, and we benchmark the performance of each of the regularized estimators relative to the root mean-squared prediction error of a predictor based on ordinary least-squares (OLS). When $d > n_{\text{train}}$ we use the "ridgeless" estimator in [42] as a benchmark.

For the testing data, we consider two different distributions. First, we consider n_{testing} new draws from the true data generating process and report results in Figure D.1. Second, we perturb the true data generating process according to the worst-case distribution derived in Corollary 2.1 with δ_n equal to the tuning parameter used by the $\sqrt{\text{LASSO}}$ and report results in Figure D.2. According to the first theorem in the paper, the $\sqrt{\text{LASSO}}$ offers robustness against perturbations of the training data.

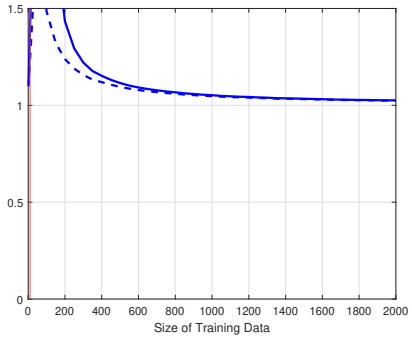
The simulation results provide two interesting findings. First, as shown in Figure D.1, when the testing distribution and the data generating process are the same, the out-of-distribution prediction error of the optimally tuned Ridge, LASSO, and $\sqrt{\text{LASSO}}$ estimators are larger than the OLS estimator. In this case, neither of the penalized estimators seem to have any attractive property in terms of out-of-distribution performance. Second, as shown in Figure D.2, the simulation results are markedly different

[5] In our set-up this equals $\sigma_\epsilon d/\|\beta\|^2$.

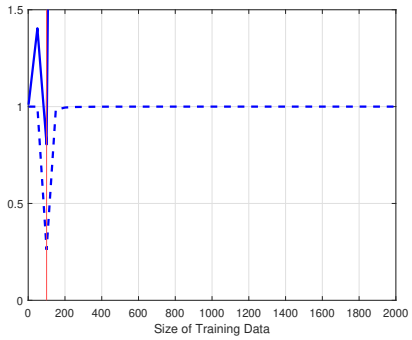
[6] We implement the LASSO using the Matlab function `lasso`.



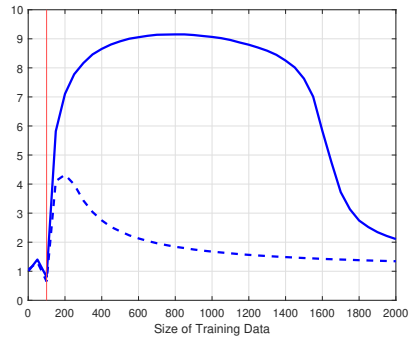
(a) Ridge: $d = 10$



(b) LASSO: $d = 10$



(c) Ridge: $d = 100$



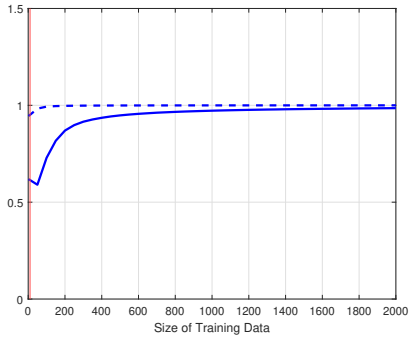
(d) LASSO: $d = 100$

Fig D.1: out-of-distribution prediction error of the Ridge, LASSO, and $\sqrt{\text{LASSO}}$ relative to that of the OLS estimator. The testing data are independently drawn from the same distribution as the training data. The solid line corresponds to the $\sqrt{\text{LASSO}}$ with the standard tuning parameter given by (A.28). The dashed line corresponds to the other estimators (Ridge/LASSO) using their oracle tuning parameters. A red vertical line indicates d. $\alpha = 0.05$, $\boldsymbol{\beta} = (1, \ldots, 1)^\top$ (vector of d ones), and $\sigma_\varepsilon = 1$.

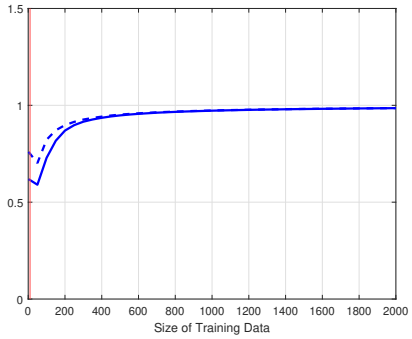
when the testing distribution and the true data generating process are allowed to differ—and, in particular, the testing distribution is adversarial—then the OLS estimator has a larger out-of-distribution prediction error in comparison to the penalized estimators. Moreover, the $\sqrt{\text{LASSO}}$ estimator—with the tuning parameter recommended by [5]—reports the lowest out-of-distribution prediction error among the estimators. Importantly, in this case, the $\sqrt{\text{LASSO}}$ seems to deliver a clearly superior performance (up to 25%) relative to optimally tuned Ridge and LASSO.

- *$\sqrt{\text{LASSO}}$ vs. LASSO*: As we discussed in Section 6, the oracle recommendation for the regularization parameter of $\sqrt{\text{LASSO}}$ (based on our analysis of the ρ-MSW metric) can be more than 10 times larger than the standard recommendation in [5]. This raises the question of whether the out-of-distribution performance of the LASSO reported in Panel d) of Figure D.2 can be improved by also using a larger regularization parameter. Lemma 2 in [73]—which shows that the $\sqrt{\text{LASSO}}$ and the LASSO share an explicit reparameterization of their solution paths conditional on the data—imply that this is indeed possible. In fact, [73, Lemma 2] shows that, for each data realization and each possible regularization parameter for the $\sqrt{\text{LASSO}}$, for example $\widetilde{\delta}_n$ as in (A.28), one can find a regularization parameter for the LASSO, denoted by $\delta_{n,\text{LASSO}}$, such that both the LASSO and the $\sqrt{\text{LASSO}}$ estimators coincide. As a consequence, using $\delta_{n,\text{LASSO}}$, guarantees that the out-of-distribution performance of the two procedures must coincide.

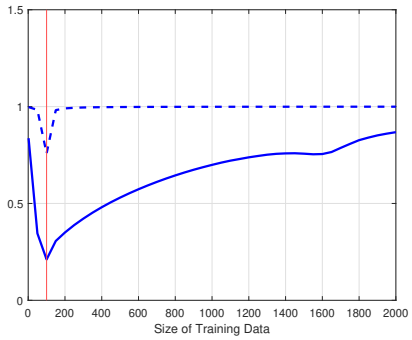
To investigate this, we consider the same Gaussian, homoskedastic, linear regression model described at the beginning of this section. We set $d = 100$ and set $\boldsymbol{\beta} = (1, \ldots, 1)^\top$ and $\sigma_\epsilon = 1$. We



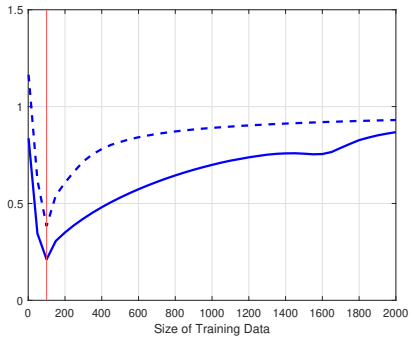
(a) Ridge: $d = 10$



(b) LASSO: $d = 10$



(c) Ridge: $d = 100$



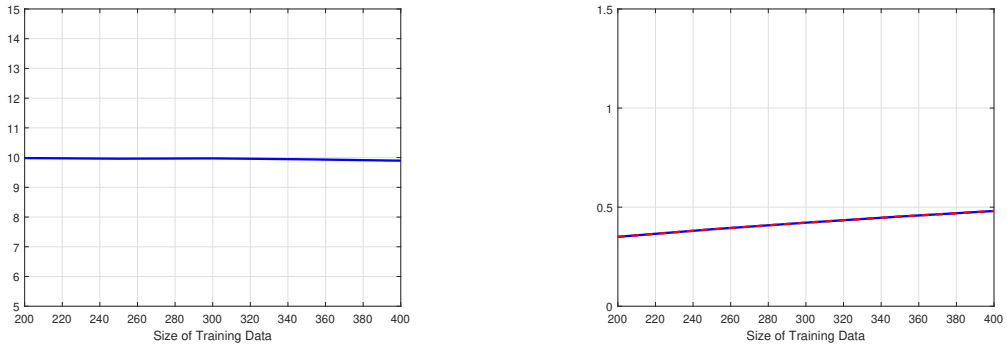
(d) LASSO: $d = 100$

Fig D.2: out-of-distribution prediction error of the Ridge, LASSO, and $\sqrt{\text{LASSO}}$ relative to that of the OLS estimator. The testing data are independently drawn from a distribution that is adversarial to the training data. The solid line corresponds to the $\sqrt{\text{LASSO}}$ with the standard tuning parameter given by (A.28). The dashed line corresponds to the other estimators (Ridge/LASSO) using their oracle tuning parameters. A red vertical line indicates d. $\alpha = 0.05$, $\boldsymbol{\beta} = (1, \ldots, 1)^\top$ (vector of d ones), and $\sigma_\varepsilon = 1$.

consider different training sizes $n_{\text{train}} \in \{200, 250, 300, \ldots, 2000\}$. For each data realization, we implement the formula in [73, Lemma 2] to obtain a new regularization parameter $\delta_{n,\text{LASSO}}$. For the $\sqrt{\text{LASSO}}$, we once again use the tuning parameter in [5]. The testing distribution is the worst-case distribution derived in Corollary 2.1.

Panel a) in Figure D.3 reports the ratio of $\delta_{n,\text{LASSO}}$ relative the oracle regularization parameter for the LASSO. For each sample size n_{train}, the figure reports the average ratio across data realizations. The figure shows that in order for the LASSO to have the same out-of-distribution performance as the $\sqrt{\text{LASSO}}$, the new regularization parameter needs to be, on average, 10 times larger than the standard tuning parameter. Panel b) in Figure D.3 confirms that the new regularization parameter indeed aligns the out-of-distribution performance of both procedures.

• *Cross-validated $\sqrt{\text{LASSO}}$*: As we have discussed, our Theorem 2.1 implies that the $\sqrt{\text{LASSO}}$ solves a distributionally robust optimization (DRO) problem: it minimizes the worst-case prediction error over data distributions that yield similar prediction errors to the training data (which we formally express as a ball around the empirical distribution using a type of max-sliced Wasserstein metric). The size of the ball in the DRO formulation is determined by the regularization parameter δ_n. This means that larger values of the regularization parameter for $\sqrt{\text{LASSO}}$ will correspond to a larger set of distributions under consideration when evaluating the worst-case prediction error.



(a) (b)

Fig D.3: (a) Ratio of the tuning parameter $\delta_{n,\text{LASSO}}$ to the oracle tuning parameter for the LASSO, where both the LASSO with $\delta_{n,\text{LASSO}}$ and $\sqrt{\text{LASSO}}$ with $\widetilde{\delta}_n$ coincide. (b) out-of-distribution prediction error of the LASSO with $\delta_{n,\text{LASSO}}$ (dashed red line) and the $\sqrt{\text{LASSO}}$ with $\widetilde{\delta}_n$ (solid blue line) relative to that of the OLS estimator. $\alpha = 0.05$, $d = 100$, $\boldsymbol{\beta} = (1, \ldots, 1)^\top$ (vector of d ones) and $\sigma_\varepsilon = 1$.

We now present numerical evidence suggesting that the K-fold cross-validated regularization parameter for the $\sqrt{\text{LASSO}}$ tends to be smaller than the standard recommendation given in [5], which we implement in (A.28).[7] Our interpretation of this result is that there will be perturbations of the true generating process ($\mathbb{P}$), for which the prediction error of the $\sqrt{\text{LASSO}}$ (tuned as in [5]) will be smaller than the prediction error of the cross-validated $\sqrt{\text{LASSO}}$.

Once again, we use the Gaussian, homoskedastic, linear regression model described at the beginning of this section to calculate the regularization (or tuning) parameter for the $\sqrt{\text{LASSO}}$ based on 5-fold cross-validation. In our exercises, we consider different values for the dimension of the vector of covariates: $d \in \{2, 5, 10, 20, 50, 100, 150, 200, 300, 400, 500\}$. Figure D.4 reports the average of 500 tuning parameters obtained via cross-validation (where the average is over data realizations, since each of the cross-validated tuning parameters depends on the data). Figure D.4 also includes the standard recommendation in [5]—and implemented as in Equation (A.28)—which in our simulations ends up being much larger than the average tuning parameter based on cross-validation. Let us recall that Section 6 already reported that our recommended tuning parameter is larger than the one based on the standard recommendation of [5].

We make two brief remarks about the simulation results in Figure D.4. First, we note that the comparison of tuning parameter for $\sqrt{\text{LASSO}}$ obtained via cross-validation and the tuning parameter recommended in [5] is consistent with other results in the literature. For instance, as is discussed in Chetverikov et al. [22, Remark 4.3], the cross-validation tuning parameter for the LASSO is often smaller than its oracle recommendation.

Second, we think that low values for the tuning parameter of the $\sqrt{\text{LASSO}}$ obtained via cross-validation reported in Figure D.4 are consistent with the idea that cross-validation aims to maximize performance at the true data generating process, $\mathbb{P}$, not some unknown, nearby distribution $\mathbb{Q}$. For example, note that in Figure D.4 the cross-validated tuning parameter is close to zero when d is lower than or similar to the training sample $n_{\text{train}} = 100$. Such a low value of the tuning parameter will

[7] The K-fold cross-validation regularization parameter is defined as the minimizer of $CV(\delta)$ in a given set Δ of possible values. We use Wu and Wang [80, Algorithm 1, page 212], but with the loss function of the $\sqrt{\text{LASSO}}$, to define the cross-validation error $CV(\delta)$ for a given tuning parameter $\delta \in \Delta$. We use $\Delta = \{C_\delta a^j : j = 0, \ldots, 24\}$ with $a = 0.5$ and $C_\delta = 8\widetilde{\delta}_n$, where $\widetilde{\delta}_n$ is as in (A.28).

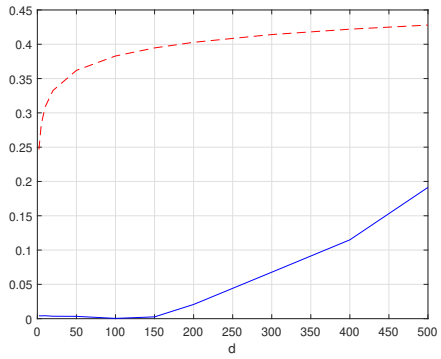


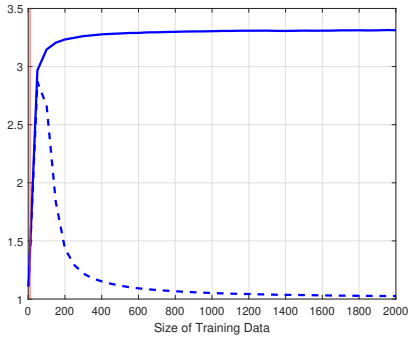
Fig D.4: Tuning parameter δ for the $\sqrt{\text{LASSO}}$. The dashed red line corresponds to the standard tuning parameter given by (A.28), while the solid blue line shows the average of 500 tuning parameters selected via 5-fold cross-validation. $\alpha = 0.05$ and $n_{\text{train}} = 100$.

make the cross-validated $\sqrt{\text{LASSO}}$ very similar to the OLS estimator. And as reported in Figure D.1, when no perturbations of the true data generating process are considered (and hence, there are no concerns about genuine out-of-distribution performance), then OLS has a better prediction error relative to $\sqrt{\text{LASSO}}$ (and other penalized estimators).

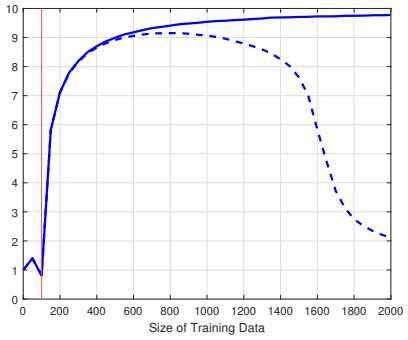
• $\sqrt{\textit{LASSO}}$ *with our recommended tuning parameter*: Finally, we present a new figure, analogous to Figure D.2, where we compare the out-of-distribution performance of two versions of the $\sqrt{\text{LASSO}}$ using (i) the tuning parameter given by (28), which is computed using the diameter of the empirical support of the data as an estimator of $\text{diam}(\text{supp}(\mathbb{P}))$, and (ii) the tuning parameter given by (A.28), which is the standard recommendation based on [5].

The results of the simulations are presented in Figure D.5 below. The figure illustrates that the relative performance of the $\sqrt{\text{LASSO}}$ estimators—with respect to the OLS estimator—depends on the testing distribution used to compute out-of-distribution prediction error. More concretely, if the testing distribution is constructed using our Corollary 2.1 (applied to the true data generating process and using the recommended tuning parameter), then the $\sqrt{\text{LASSO}}$ with our recommended tuning parameter has a better out-of-distribution performance than the $\sqrt{\text{LASSO}}$ with the usual tuning parameter.

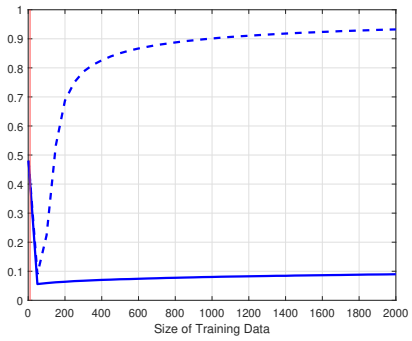
Figure D.6 further shows that a feasible version of our recommended tuning parameter, given by (28) and an estimator of the diameter of the support of the data distribution, is several times larger than the standard tuning parameter based on [6], given by (A.28).



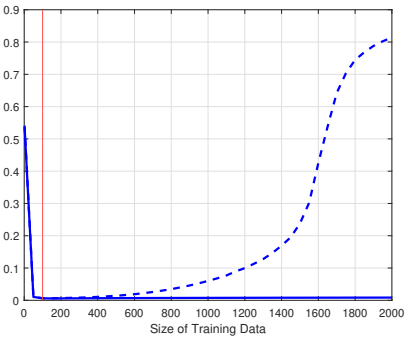
(a) True DGP: $d = 10$



(b) True DGP: $d = 100$



(c) Adversarial DGP: $d = 10$



(d) Adversarial DGP: $d = 100$

Fig D.5: out-of-distribution prediction error of the $\sqrt{\text{LASSO}}$ relative to that of the OLS estimator. The testing data in (a) and (b) are independently drawn from the same distribution as the training data, while those in (c) and (d) are independently drawn from a distribution that is adversarial to the training data. The solid line corresponds to the $\sqrt{\text{LASSO}}$ with the tuning parameter given by (28), which is computed using the diameter of the empirical support of the data as an estimator of $\text{diam}(\text{supp}(\mathbb{P}))$. The dashed line corresponds to the $\sqrt{\text{LASSO}}$ with the standard tuning parameter given by (A.28). A red vertical line indicates d. $\alpha = 0.05$, $\boldsymbol{\beta} = (1, \ldots, 1)^\top$ (vector of d ones), and $\sigma_\varepsilon = 1$.

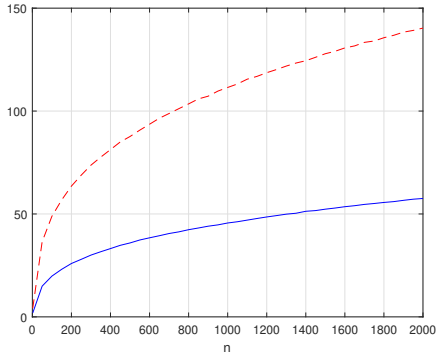


Fig D.6: Ratio of the regularization parameter in (28), which is computed using the diameter of the empirical support of the data as an estimator of $\text{diam}(\text{supp}(\mathbb{P}))$, to that one in (A.28). The solid blue line corresponds to $d = 10$, and the dashed red line corresponds to $d = 100$. $\alpha = 0.05$.