THE OUT-OF-SAMPLE 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, Northwestern University,* [c]*amilcare@u.northwestern.edu*

[4]*Department of Statistics, Columbia University,* [d]*johannes.wiesel@columbia.edu*

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:

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

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-sample 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-sample 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-sample" 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, [31, Chapter 7] and [37, 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 [39], distributional shifts [28, 65, 1], domain adaptation [47, 7], adversarial attacks [41, 38], learning under biased sampling [59] 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:

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

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 $r \geq 1$. The square-root LASSO (henceforth, $\sqrt{\text{LASSO}}$) [5], the square-root group LASSO [17], the square-root sorted ℓ_1 penalized estimator (SLOPE) [64], and the ℓ_1-penalized least absolute deviation estimator [69] provide examples of estimators obtained by solving (1).

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

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

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-sample prediction error (see (9) and Theorem 6 for a more formal statement): If δ is chosen appropriately, then, with high probability, for any $\boldsymbol{\beta}$, we have

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

where $\widehat{\mathcal{W}}_r$ denotes a type of *max-sliced Wasserstein metric*. We present a formal definition of this metric 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 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 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 2.1 provides examples of distributions contained in our 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-sample 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 on $\mathbb{P}_n$ to guarantee that it contains $\mathbb{P}$ with high probability. We present both finite-sample results (Theorem 2 and Theorem 3 in Section 3) and large-sample approximations (Theorem 4 and 5 in Section 4). Our analysis suggests that our balls are *statistically larger* than those based on the standard Wasserstein metric (Remark 2). 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).

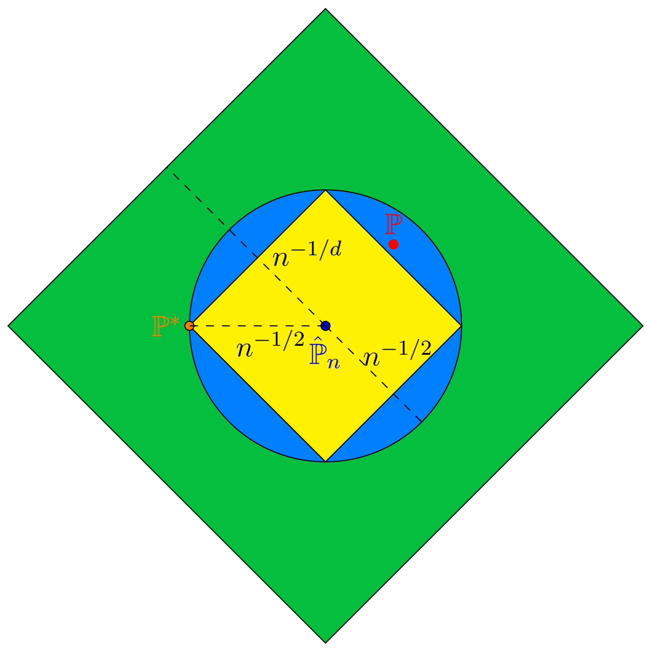


Fig 1: ρ-max-sliced Wasserstein ball of radius $n^{-1/2}$ (blue) vs. d-dimensional Wasserstein ball of radius $n^{-1/2}$ (yellow) and $n^{-1/d}$ (green). The measure $\mathbb{P}^*$ (orange) is the optimal perturbation in the DRO formulation.

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-sample prediction error (Theorem 6 in Section 5.1); and ii) present a test statistic to rank the out-of-sample 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; thus, they broaden the scope of use of the square-root lasso and related estimators in prediction problems.

We now provide an overview of the technical details of our main results.

1.1. *Main Contributions.*

1.1.1. *DRO formulation.* Our first result shows that linear predictors whose coefficients solve (1) minimize the worst-case, out-of-sample prediction error attained over a ball of distributions centered around $\mathbb{P}_n$. This ball is defined by what we call the *ρ-max-sliced Wasserstein (ρ-MSW) metric* [1]:

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

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 $((\mathbf{X}^\top,Y),(\widetilde{\mathbf{X}}^\top,\widetilde{Y}))$, 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 assume $r,\sigma\geq 1$, refer to r as the Wasserstein exponent, and take σ to be an auxiliary hyperparameter. Intuitively, $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ are close in the 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 $\rho(\boldsymbol{\gamma})$. In this sense $\widehat{\mathcal{W}}_{r,\rho,\sigma}$ imposes a metric structure on the space of probability measures, which is derived from the loss function $|\cdot - \langle\beta,\cdot\rangle|^r$ in the linear regression problem; arguably it is thus a natural metric to consider for this problem.

Formally, Theorem 1 in Section 2 shows that $\widehat{\boldsymbol{\beta}}$ solves (1) if and only if it solves the distributionally robust optimization problem

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^d}\left(\sup_{\widetilde{\mathbb{P}}\in B_\delta^{r,\rho,\sigma}(\mathbb{P}_n)}\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]\right), \tag{5}$$

where $B_\delta^{r,\rho,\sigma}(\mathbb{P}_n)$ is defined as the ball centered around the empirical distribution of the data, $\mathbb{P}_n$, collecting all the distributions $\widetilde{\mathbb{P}}$ for which $\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P}_n,\widetilde{\mathbb{P}})$ is smaller than δ. By construction, the minimax problem in (5) provides robustness in situations 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], where there are covariate shifts [61, 54, 65, 66, 2, 72, 56, 18], or when there is an adversarial attack [41, 38].

[1] Lemma 1 shows that the ρ-MSW metric is indeed a metric.

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

$$\mathcal{W}_r(\mathbb{P}, \widetilde{\mathbb{P}}) := \inf_{\pi \in \Pi(\mathbb{P}, \widetilde{\mathbb{P}})} \left(\mathbb{E}_\pi \left[\| (\widetilde{\mathbf{X}}^\top, \widetilde{Y}) - \left(\mathbf{X}^\top, Y\right) \|^r \right] \right)^{1/r}, \tag{6}$$

where $\|\cdot\|$ is an arbitrary metric in $\mathbb{R}^{d+1}$. Remark 2 in Section 2 shows that for a large class of penalty functions $\rho(\cdot)$, the balls based on (4) will typically be larger than those based on (6). We note that our ρ-MSW is a slight modification of the *max-sliced Wasserstein metric* (MSW), see equation (24). The MSW was first considered in [52, 51, 25] and it was originally derived from the sliced Wasserstein metric in [55, 15].

To the best of our knowledge, the result showing that linear predictors whose coefficients solve (1) equivalently minimize the worst-case, out-of-sample prediction error attained over a neighborhood $B_\delta^{r,\rho,\sigma}(\mathbb{P}_n)$ based on the ρ-MSW metric is new. The connection between (1) and (5) for the case $r = 2$, penalty $\rho(\cdot) = \|\cdot\|_p$, $p \geq 1$, and $\widehat{\mathcal{W}}_r$ replaced by $\mathcal{W}_r$, the Wasserstein metric, was first established in [9, Theorem 1], using optimal transport (OT) duality [33, 12]. These results have recently been extended to more general penalty functions [23, 73]. Our balls are different to the ones considered in these papers, we focus on convex penalty functions, and also our proofs do not rely on duality arguments. Instead, we explicitly identify a worst-case measure $\mathbb{P}^* \in B_\delta^{r,\rho,\sigma}(\mathbb{P}_n)$ for (5). Our results show that the measure $\mathbb{P}^*$ is given by an additive perturbation of $\mathbb{P}_n$ (see Corollary 1). In this sense, our proof can be seen as a natural extension of [8, Theorem 1]. Moreover, we believe that $\widehat{\mathcal{W}}_{r,\rho,\sigma}$ is the natural metric to consider to assess out-of-sample performance (5), as it allows for a general class of *testing* distributions that are only required to generate similar prediction errors as the *training* distribution $\mathbb{P}_n$. In contrast, the Wasserstein metric puts additional restrictions on the testing distributions considered. This in turn makes the radius of the balls very large in high dimensions (making the associated bounds not very useful in practice).

1.1.2. *Statistical Analysis of* $\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P}_n, \mathbb{P})$. Our second set of results provide a detailed analysis of the statistical properties of (4). For simplicity in the exposition, we focus on the case when ρ is a norm $\|\cdot\|$ on $\mathbb{R}^d$. Since all norms in $\mathbb{R}^d$ are equivalent to the Euclidean metric ($\|\cdot\|_2$), there exists a constant c_d such that $c_d\|\boldsymbol{\gamma}\|_2 \leq \rho(\boldsymbol{\gamma})$ for any $\boldsymbol{\gamma}$. Theorem 3 in Section 3 shows that if

$$\Gamma := \mathbb{E}_\mathbb{P}\left[\left\| \left(\mathbf{X}^\top, Y\right) \right\|_2^s \right] < \infty, \qquad \text{for some } s > 2r,$$

then with a probability greater than $1 - \alpha$ we have that

$$\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P}_n, \mathbb{P})^r \leq \left(\max\left\{ \frac{1}{c_d}, 1 \right\} \right) \frac{C \log(2n+1)^{r/s}}{\sqrt{n}}, \tag{7}$$

where C is the constant in (32) and is a function of the parameters $(r, d, \alpha, \Gamma, s)$. Furthermore, Theorem 5 in Section 4 shows there exists a constant $C := C(r, s, d)$, such that for all $x \geq 0$ we have

$$\limsup_{n\to\infty} \mathbb{P}\left(\sqrt{n}\,\widehat{\mathcal{W}}_{r,\rho,\sigma}\left(\mathbb{P}_n,\mathbb{P}\right)^r \geq x\right) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} |G_f| \geq \frac{x}{C\sqrt{\Gamma}}\right), \tag{8}$$

where $(G_f)_{f\in\mathcal{F}}$ is a zero-mean Gaussian process specified in Theorem 5.

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. We also argue that the implied rates for δ are optimal, up to logarithmic factors (cf. e.g. [30, Remark (a), p.2]). Lastly, it is worth noting that the bounds (7) and (8) also hold for the ordinary max-sliced Wasserstein distance (which can be obtained by formally setting $Y=\widetilde{Y}=\sigma=0$ in (4)).

1.1.3. *Applications.* *Choosing δ:* Our statistical analysis provides a concrete *oracle* recommendation to select the regularization parameter $\delta_{n,r}$ in (1) to be equal to the $(1/r)$-th power of the right-hand side of (7); see Section 5.1. Of course, the oracle recommendation for $\delta_{n,r}$ is typically not feasible as it depends on the unknown parameter Γ. In Section 5.2, we present a simple strategy to normalize the sample covariates that guarantees that both (7) and (9) hold with $\Gamma=2^s$ and $\sigma=\max\{\mathbb{E}_{\mathbb{P}}[|Y|^s]^{1/s},1\}$. 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 s-th 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 (9) below. This *pivotality* was the original motivation for the use of the $\sqrt{\text{LASSO}}$ and related estimators.[2]

We also note 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 [35, 45] has argued that sparsity might not always be a compelling starting point in applications.

Bounds on out-of-sample performance: When the available sample consists of independent and identical (i.i.d.) draws from a distribution $\mathbb{P}$ for which $\Gamma<\infty$ we can show that the objective function in (1) provides explicit bounds on the out-of-sample performance of any linear estimator. In particular, Theorem 6 in Section 5.1 shows that for any *testing* distribution $\mathbb{Q}$ for which $\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P},\mathbb{Q})\leq\epsilon$, we have

$$\mathbb{E}_{\mathbb{Q}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]^{1/r} \leq \mathbb{E}_{\mathbb{P}_n}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]^{1/r} + \left(\delta_{n,r}+\epsilon\right)\left(\sigma+\rho(\boldsymbol{\beta})\right), \quad \forall\boldsymbol{\beta} \tag{9}$$

with probability at least $1-\alpha$.

[2] In Section 3.1, we consider the case where the support of $\mathbb{P}$ is compact and our recommended $\delta_{n,r}$ only depends on the diameter of the support of $\mathbb{P}$.

Ranking the out-of-sample performance of competing estimators. Finally, we present a test statistic to rank the out-of-sample performance of two different linear estimators (Section 5.4).

1.2. *Related Literature.* The distributionally robust optimization problem in (5) has been shown to be equivalent to various forms of penalized regression, variance-penalized estimation, and dropout training [48, 34, 42, 11, 12, 27, 49], 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 [49] or Kullback-Leibler divergence [57, 22]. To the best of our knowledge, the use of the max-sliced Wasserstein metric to define uncertainty sets in DRO problems is novel.

As we have discussed above, the equivalence between (1) and (5) has been established in [9, Theorem 1] using a ball in the Wasserstein metric, which is a common choice for the uncertainty set in the distributionally robust optimization literature [40, 10, 63, 43, 34, 48, 32, 60]. Relative to previous results, we focus on convex penalty functions (and not only norms), and also we explicitly identify a worst-case measure $\mathbb{P}^* \in B_\delta^{r,\rho,\sigma}(\mathbb{P}_n)$ for (5), instead of relying on duality arguments. In this sense, our proof can be seen as a natural extension of [8, Theorem 1].

DRO representations similar to (5) 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 [61, 54, 65, 66, 2, 72, 56, 18], or when one experiences adversarial attacks [41, 38].

Starting from [24, 30], 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 [14, 62, 70, 71, 44, 21]; see also the references therein. When $d > 2r$, tight rates for $\mathcal{W}_r(\mathbb{P}_n, \mathbb{P})$ are of the order $n^{-1/(rd)}$, 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 [51] for subgaussian probability measures and in [46] 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. [29]. A notable exception is the recent work of [50] and [36]. In particular, using log-concavity, [50] 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 these 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) recommends the choice of the penalization parameters to guarantee nearly oracle rates for the estimation error or prediction norm. For the case of the LASSO estimator, [20] present conditions such that the widely used cross-validation method is valid, and [19] suggest utilizing a bootstrap approximation to estimate the penalization parameter. For the case of the $\sqrt{\text{LASSO}}$, [5, 6] proposed a pivotal penalization parameter with asymptotic guarantees. In all these cases, their recommendation is valid whenever the sparsity assumption of the linear regression holds. Our work complements these previous results since we propose a penalization parameter without using any sparsity assumption. As we mentioned before, our recommendation for the penalization parameter controls the out-of-sample prediction error (for a finite sample and/or asymptotically). Therefore, we depart from the usual motivation to choose the penalization parameter based on the prediction norm since we do not assume a sparse linear regression model.

1.3. *Outline.* The rest of the paper is organized as follows. In Section 2, we present a detailed discussion of the equivalence of (1) and (5). In Sections 3 and 4 we present rates for the MSW distance $\widehat{\mathcal{W}}_{r,\rho,\sigma}$ between the true and empirical measure, both for $\mathbb{P}$ with compact support and for $\mathbb{P}$ satisfying $\Gamma < \infty$. Section 3 gives a finite sample analysis, while 4 provides asymptotics. In Section 5, we present a recommendation for the selection of regularization parameter, $\delta_{n,r}$, that guarantees good out-of-sample prediction error. We also present a test statistic to rank the out-of-sample 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. Section 7 collects the remaining proof details.

1.4. *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 [58]). 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 $\partial\rho(\boldsymbol{\beta})$, ([58]; pp. 214-215).

Lastly, let us mention two important facts that will be relevant in Section 2.1. If ρ is differentiable, then its subdifferential $\partial\rho(\boldsymbol{\beta})$ is a singleton that contains the gradient of ρ at $\boldsymbol{\beta}$; see, for example, ([58]; Theorem 25.1). If ρ is a norm in $\mathbb{R}^d$, then ρ^* is only equal to zero or infinity; see ([16]; p. 93).

2. Reformulation as a DRO problem.

For any $r, \sigma \in [1, \infty)$, and $\rho\colon \mathbb{R}^d \to [0, +\infty)$ define the collection of distributions

$$
\begin{aligned}
B_\delta^{r,\rho,\sigma}(\mathbb{P}) &:= \left\{\mathbb{Q} \in \mathcal{P}_r(\mathbb{R}^{d+1}) : \widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{Q}, \mathbb{P}) \leq \delta\right\} \\
&= \Big\{\mathbb{Q} \in \mathcal{P}_r(\mathbb{R}^{d+1}) :\ \forall\, \boldsymbol{\gamma} \in \mathbb{R}^d, \quad \exists \text{ a coupling } \pi(\boldsymbol{\gamma}) \in \Pi(\mathbb{P}, \mathbb{Q}) \\
&\qquad \text{for which } \mathbb{E}_{\pi(\boldsymbol{\gamma})}\left[\left|(\widetilde{Y} - Y) + (\mathbf{X} - \widetilde{\mathbf{X}})^\top \boldsymbol{\gamma})\right|^r\right] \leq \delta^r \left(\sigma + \rho(\boldsymbol{\gamma})\right)^r, \\
&\qquad \text{where } \big((\mathbf{X}, Y),\, (\widetilde{\mathbf{X}}, \widetilde{Y})\big) \sim \pi(\boldsymbol{\gamma})\Big\}.
\end{aligned} \tag{10}
$$

As explained in the introduction, a distribution $\mathbb{Q}$ belongs to the ball in (10) if and only if for any $\boldsymbol{\gamma}$ there exists a coupling of $\mathbb{P}$ and $\mathbb{Q}$ that makes the r-th norm of their prediction errors small, relative to $\rho(\boldsymbol{\gamma})$. We remark that the infimum in the definition of $\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{Q}, \mathbb{P})$ 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. For notational simplicity we will suppress the dependence of the ball $B_\delta^{r,\rho,\sigma}(\mathbb{P})$ on r, ρ, σ and write $B_\delta(\mathbb{P})$ instead.

The main result of this section establishes a formal connection between the solutions to the problems in (1) and (5).

THEOREM 1. *Fix $1 \leq r < \infty$ and $1 \leq \sigma < \infty$. 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*

$$
\left|\boldsymbol{\gamma}^\top \left(\boldsymbol{\beta}^* - \frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top \boldsymbol{\beta}}\, \rho^*(\boldsymbol{\beta}^*)\right)\right| \leq \rho(\boldsymbol{\gamma}), \qquad \forall\, \boldsymbol{\gamma} \in \mathbb{R}^d. \tag{11}
$$

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

$$
\sup_{\widetilde{\mathbb{P}} \in B_\delta(\mathbb{P})} \mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y - \mathbf{X}^\top \boldsymbol{\beta}\right|^r\right] = \left(\sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[|Y - \mathbf{X}^\top \boldsymbol{\beta}|^r\right]} + \delta\left(\sigma + \rho(\boldsymbol{\beta})\right)\right)^r. \tag{12}
$$

Theorem 1 shows that the worst-case, out-of-sample 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

[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 [67, Theorem 4.1].

result in (12) thus implies

$$\arg\inf_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\sup_{\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})}\mathbb{E}_{\widetilde{\mathbb{P}}}\Big[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\Big]\right]=\arg\inf_{\boldsymbol{\beta}\in\mathbb{R}^d}\sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[|Y-\mathbf{X}^\top\boldsymbol{\beta}|^r\right]}+\delta\rho(\boldsymbol{\beta}). \tag{13}$$

Our interpretation of Equation (13) is that the $\sqrt{\text{LASSO}}$ and related estimators in (1) have good out-of-sample 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].

We briefly sketch the proof of Theorem 1 here and refer to Section 7.1 for details. It proceeds in two steps:

Step 1. We use the triangle inequality and the definition of the ρ-MSW metric to show that

$$\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]\le\left(\sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[|Y-\mathbf{X}^\top\boldsymbol{\beta}|^r\right]}+\delta\left(\sigma+\rho(\boldsymbol{\beta})\right)\right)^r, \tag{14}$$

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

Step 2. We show that for any $\boldsymbol{\beta}\in\text{dom}(\rho)$, the upper bound given in **Step 1** is tight. That is, we explicitly construct, for each $\boldsymbol{\beta}\in\text{dom}(\rho)$ a distribution $\mathbb{P}^*\in B_\delta(\mathbb{P})$, for which the bound holds exactly. The worst-case distribution is presented in Corollary 1 below.

COROLLARY 1. *For each $\boldsymbol{\beta}\in\mathbb{R}^d$ the supremum in* (12) *is attained for the distribution $\mathbb{P}^*_{\boldsymbol{\beta}}$ corresponding to the random vector $(\widetilde{\mathbf{X}},\widetilde{Y})$ defined as*

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

where

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

One aspect of Corollary 1 that is worth emphasizing is that the testing distribution that attains the worst out-of-sample 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.

It is easy to see that the minimizer in (13) is attained. Denote this minimizer 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 the following:

$$\mathbb{E}_{\mathbb{P}^*_{\boldsymbol{\beta}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]=\sup_{\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})}\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]$$

$$\geq \inf_{\boldsymbol{\beta}\in\mathbb{R}^d} \sup_{\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})} \mathbb{E}_{\widetilde{\mathbb{P}}}\Big[\Big|Y - \mathbf{X}^\top\boldsymbol{\beta}\Big|^r\Big]$$

$$= \sup_{\widetilde{\mathbb{P}}\in B_\delta(\mathbb{P})} \mathbb{E}_{\widetilde{\mathbb{P}}}\Big[\Big|Y - \mathbf{X}^\top\boldsymbol{\beta}(\mathbb{P})\Big|^r\Big]$$

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

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 (13) performs better.

REMARK 1 (On condition (11)). *If ρ is a norm, then the condition in* (11) *is automatically satisfied; i.e., there exists a $\boldsymbol{\beta}^* \in \partial\rho(\boldsymbol{\beta})$ such that* (11) *is true. Thus, the conclusion of Theorem 1 holds for all $\rho(\cdot) = \|\cdot\|$ that are norms. Indeed, recalling that the dual norm of $\|\cdot\|$ is given by*

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

in that case [16, 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

$$(\boldsymbol{\beta}^*)^\top\boldsymbol{\beta} - \rho^*(\boldsymbol{\beta}^*) = \rho(\boldsymbol{\beta}). \tag{15}$$

Both facts together imply that $\rho^*(\boldsymbol{\beta}^*) = 0$; thus, $\|\boldsymbol{\beta}^*\|_* \leq 1$ for all $\boldsymbol{\beta}^* \in \partial\rho(\boldsymbol{\beta})$. Hence in (11)*, as claimed, we have*

$$\left|\boldsymbol{\gamma}^\top\left(\boldsymbol{\beta}^* - \frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\right)\right| = \left|\boldsymbol{\gamma}^\top\boldsymbol{\beta}^*\right| \leq \|\boldsymbol{\gamma}\|\,\|\boldsymbol{\beta}^*\|_* \leq \|\boldsymbol{\gamma}\|, \quad \forall\,\boldsymbol{\gamma}\in\mathbb{R}^d. \tag{16}$$

On the other hand, the following example shows that condition (11) is not only satisfied by norms:

EXAMPLE 1 (Condition (11) 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,

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

By (15) *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 2. *Take any norm* $\|\cdot\|$ *on* $\mathbb{R}^{d+1}$ *satisfying* $\|(0,\dots,0,1)\| = 1$ *and recall that its dual norm is given by*

$$\|\mathbf{x}\|_* := \sup_{\mathbf{y}:\|\mathbf{y}\|=1} \mathbf{x}^\top \mathbf{y}. \tag{17}$$

Assume that $\mathbb{E}_{\mathbb{P}}\left[\|(\mathbf{X},Y)\|_*^r\right] < \infty$ *and consider a Wasserstein ball* $\mathcal{B}_\delta(\mathbb{P})$ *with cost* $\|\cdot\|_*$*, defined as*

$$\mathcal{B}_\delta(\mathbb{P}) = \left\{ \widetilde{\mathbb{P}} \in \mathcal{P}_r(\mathbb{R}^{d+1}) :\ \mathcal{W}_r(\mathbb{P}, \widetilde{\mathbb{P}}) \le \delta \right\}, \tag{18}$$

where

$$\mathcal{W}_r(\mathbb{P}, \widetilde{\mathbb{P}}) = \inf_{\substack{\pi \in \Pi(\mathbb{P},\widetilde{\mathbb{P}}): \\ ((\mathbf{X},Y),(\widetilde{\mathbf{X}},\widetilde{Y}))\sim\pi}} \sqrt[r]{\mathbb{E}_\pi\left[\|(\mathbf{X},Y) - (\widetilde{\mathbf{X}},\widetilde{Y})\|_*^r\right]}.$$

We show that for $\rho(\boldsymbol{\gamma}) = \|\boldsymbol{\gamma}\|$*, the ball defined in* (10) *contains the ball in* (18)*, i.e.* $\mathcal{B}_\delta(\mathbb{P}) \subseteq B_\delta(\mathbb{P})$*. For this, we note that by* (17) *we have*

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

As $\sigma \ge 1$*, we conclude*

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

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

2.1. *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 (10). We do this by considering different estimators that take the form (1).

2.1.1. $\sqrt{\textit{LASSO}}$. 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 (13) 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 (11) of Theorem 1 is satisfied, cf. Remark 1.

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

$$(19) \quad \begin{aligned} B_\delta^{\sqrt{\text{LASSO}}}(\mathbb{P}) := \Big\{ & \mathbb{Q} \in \mathcal{P}_2(\mathbb{R}^{d+1}) \mid \exists \text{ a coupling } \pi \in \Pi(\mathbb{Q}, \mathbb{P}) \text{ for which:} \\ & \mathbb{E}_\pi \left[\left| \widetilde{X}_j - X_j \right|^2 \right] \le \delta^2, \ \forall\, j = 1, \ldots, d, \text{ and } \mathbb{E}_\pi \left[\left| \widetilde{Y} - Y \right|^2 \right] \le (\delta\sigma)^2, \\ & \text{where } ((\mathbf{X}, Y), (\widetilde{\mathbf{X}}, \widetilde{Y})) \sim \pi \Big\}. \end{aligned}$$

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 1 and defined in (10).

To see this, notice that $\mathbb{E}_\pi \left[\left| \widetilde{X}_j - X_j \right|^2 \right] \le \delta^2$ for all $j = 1, \ldots, d$ implies condition (10), i.e. that

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

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

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

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

$$\widetilde{X}_j = X_j + u_j, \quad \widetilde{Y} = Y + v,$$

where $\mathbb{E}[u_j^2] \leq \delta^2$ and $\mathbb{E}[v^2] \leq (\delta\sigma)^2$. Also, $B_\delta^{\sqrt{\text{LASSO}}}(\mathbb{P})$ contains multiplicative errors-in-variables models where

$$\widetilde{X}_j = X_j u_j, \quad \widetilde{Y} = Yv,$$

with u's independent of $(\mathbf{X}, Y)$, having mean equal to one, $\mathbb{E}_{\mathbb{P}}\left[X_j^2\right] \mathbb{E}\left[(u_j-1)^2\right] \leq \delta^2$, and independent of v having mean equal to one and $\mathbb{E}_{\mathbb{P}}\left[Y^2\right] \mathbb{E}\left[(v-1)^2\right] \leq (\delta\sigma)^2$.

It is well known that the conjugate of ρ is

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

The argument is analogous to Remark 1. Moreover, algebra shows that

$$\boldsymbol{\beta}^* = (\text{sign}(\beta_1), \ldots, \text{sign}(\beta_d))^\top,$$

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 1 states that:

$$\widetilde{\mathbf{X}} = \mathbf{X} - e\,(\text{sign}(\beta_1), \ldots, \text{sign}(\beta_d))^\top, \qquad \widetilde{Y} = Y + \sigma e,$$

where

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

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 (19).

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

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

where $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ 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 (13) is the objective function of the square-root SLOPE of [64].

An equivalent definition for this penalty function is

$$\rho(\boldsymbol{\beta}) = \max_{pm} \sum_{j=1}^d \lambda_{pm(j)} |\beta_j|, \tag{20}$$

where we maximize over all permutations, pm, of the coordinates $\{1,\ldots,d\}$. It follows that ρ is a norm, so Condition (11) of Theorem 1 is satisfied (see Remark 1).

For a given $\boldsymbol{\beta}\in\mathbb{R}^d$, let pm^* be a permutation that solves (20). 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, (11) holds.

In this case, a set of distributions that belongs to balls of size δ based on the ρ-MSW metric is

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

where the decreasing order statistic is induced by the vector $\left(\mathbb{E}_\pi\left[\left|\widetilde{X}_j-X_j\right|^2\right]\right)_{j=1,\ldots,d}$. 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 $\pi\in\Pi(\mathbb{P},\mathbb{Q})$:

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

where the last equality follows by the definition of $B_\delta^{\text{SLOPE}}(\mathbb{P})$ and (20).

Finally, we report the worst-case distribution for each particular $\boldsymbol{\beta}$. Corollary 1 shows that

$$
\widetilde{\mathbf{X}}=\mathbf{X}-e\boldsymbol{\beta}^*,\quad \widetilde{Y}=Y+\sigma e,
$$

where the j-coordinate of $\boldsymbol{\beta}^*$ is $\lambda_{pm^*(j)}\,\text{sign}(\beta_j)$ and

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

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 that we denote by $\mathbb{P}$. We denote the empirical distribution based on the available data by $\mathbb{P}_n$.

This section provides explicit upper bounds on the radius δ of the ball $B_\delta(\mathbb{P}_n)$ defined in (10), 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

$$\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P},\widetilde{\mathbb{P}}) = \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d} \inf_{\substack{\pi\in\Pi(\mathbb{P},\widetilde{\mathbb{P}}):\\ ((\mathbf{X},Y),(\widetilde{\mathbf{X}},\widetilde{Y}))\sim\pi}} \frac{1}{\sigma+\rho(\boldsymbol{\gamma})} \sqrt[r]{\mathbb{E}_\pi\left[\left|(\widetilde{Y}-Y)+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\gamma}\right|^r\right]}. \tag{21}$$

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

$$\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P},\widetilde{\mathbb{P}}) = \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d} \frac{1}{\sigma+\rho(\boldsymbol{\gamma})} \mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top\bar{\boldsymbol{\gamma}}\right]_*\mathbb{P},\ \left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top\bar{\boldsymbol{\gamma}}\right]_*\widetilde{\mathbb{P}}\right), \tag{22}$$

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 $\mathcal{W}_r$. The one-dimensional Wasserstein metric is simply defined as

$$\mathcal{W}_r(\mathbb{Q},\widetilde{\mathbb{Q}}) = \inf_{\substack{\pi\in\Pi(\mathbb{Q},\widetilde{\mathbb{Q}}):\\ (X,\widetilde{X})\sim\pi}} \sqrt[r]{\mathbb{E}_\pi\left[\left|X-\widetilde{X}\right|^r\right]}. \tag{23}$$

We focus on the case $\rho(\cdot)=\|\cdot\|$, where $\|\cdot\|$ is an arbitrary norm in $\mathbb{R}^d$. It is well known that since all norms in $\mathbb{R}^d$ are equivalent to the Euclidean norm $\|\cdot\|_2$, there exists a constant $c_d>0$ such that

$$c_d\|\mathbf{x}\|_2 \le \|\mathbf{x}\|, \qquad \text{for all } \mathbf{x}\in\mathbb{R}^d.$$

Using the definition in (22) we derive the following upper bound for $\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P},\widetilde{\mathbb{P}})$:

$$\begin{aligned}\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P},\widetilde{\mathbb{P}}) &= \sup_{\boldsymbol{\gamma}\in\mathbb{R}^d} \frac{\|\bar{\boldsymbol{\gamma}}\|_2}{\sigma+\|\boldsymbol{\gamma}\|}\frac{1}{\|\bar{\boldsymbol{\gamma}}\|_2}\mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top\bar{\boldsymbol{\gamma}}\right]_*\mathbb{P},\ \left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top\bar{\boldsymbol{\gamma}}\right]_*\widetilde{\mathbb{P}}\right)\\ &\le c_{\rho,d}\left(\sup_{\widetilde{\boldsymbol{\gamma}}:\|\widetilde{\boldsymbol{\gamma}}\|_2=1}\mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top\widetilde{\boldsymbol{\gamma}}\right]_*\mathbb{P},\ \left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top\widetilde{\boldsymbol{\gamma}}\right]_*\widetilde{\mathbb{P}}\right)\right)\\ &=: c_{\rho,d}\,\overline{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}),\end{aligned} \tag{24}$$

where $\bar{\boldsymbol{\gamma}}^\top=(\boldsymbol{\gamma}^\top,-1)$ and $c_{\rho,d}:=\max\{1/c_d,1\}$.

The quantity $\overline{\mathcal{W}}_r$ defined in (24) is known as the max-sliced Wasserstein (MSW) distance on $(\mathbb{R}^{d+1},\|\cdot\|_2)$. Moreover, it is a special case of the Projection Robust Wasserstein (PRW) distance, also called the Wasserstein Projection Pursuit (WPP), see [53, Definition 1]. The work in [53, 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 $r\ge 1$).

As stated in the Introduction, it is well known that, in the worst case, $\mathcal{W}_r(\mathbb{P}_n,\mathbb{P})^r \sim n^{-1/(d+1)}$. 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, in particular, by (24) we have

$$\overline{\mathcal{W}}_r(\mathbb{P}, \widetilde{\mathbb{P}}) = \sup_{\|\boldsymbol{\gamma}\|_2 = 1} \mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma}}, \widetilde{\mathbb{P}}_{\boldsymbol{\gamma}}).$$

We now provide explicit upper bounds for $\overline{\mathcal{W}}_r(\mathbb{P}, \mathbb{P}_n)$. By equation (24), for any δ we have

$$\mathbb{P}\left(\widehat{\mathcal{W}}_{r,\rho,\sigma}(\mathbb{P}, \mathbb{P}_n) \leq c_{\rho,d} \cdot \delta\right) \geq \mathbb{P}\left(\overline{\mathcal{W}}_r(\mathbb{P}, \mathbb{P}_n) \leq \delta\right). \tag{25}$$

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 2. *Let $\mathbb{P}$ be a distribution with compact support. With probability at least $1 - \alpha$, we have*

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

where

$$C := \left(180\sqrt{d+2} + \sqrt{2 \log\left(\frac{1}{\alpha}\right)}\right) \operatorname{diam}\left(\operatorname{supp}(\mathbb{P})\right)^r, \tag{26}$$

and $\operatorname{diam}(\operatorname{supp}(\mathbb{P}))$ *is the diameter of the support of* $\mathbb{P}$ *measured with respect to the Euclidean norm.*

PROOF. We first recall the representations for the one-dimensional Wasserstein distance

$$\mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}})^r = \int_0^1 \left|F_{\boldsymbol{\gamma},n}^{-1}(p) - F_{\boldsymbol{\gamma}}^{-1}(p)\right|^r \, dp, \tag{27}$$

for $r \geq 1$ and

$$\mathcal{W}_1(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}}) = \int_{\mathbb{R}} \left|F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)\right| \, dt, \tag{28}$$

see e.g. [13, Theorem 2.9, Theorem 2.10]. We also note that $\mathcal{W}_r$ is translation invariant, which implies in particular that

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

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

$$\text{supp}(\mathbb{P}_{\boldsymbol{\gamma}}) \subseteq [0,\, c]\,. \tag{29}$$

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

$$\begin{aligned}
\overline{\mathcal{W}}_r(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}})^r &= \sup_{\|\boldsymbol{\gamma}\|_2=1} \left(\int_0^1 \left| F^{-1}_{\boldsymbol{\gamma},n}(p) - F^{-1}_{\boldsymbol{\gamma}}(p) \right|^r \, dp \right) \\
&\leq c^{r-1} \sup_{\|\boldsymbol{\gamma}\|_2=1} \left(\int_0^1 \left| F^{-1}_{\boldsymbol{\gamma},n}(p) - F^{-1}_{\boldsymbol{\gamma}}(p) \right| \, dp \right) \\
&= c^{r-1} \sup_{\|\boldsymbol{\gamma}\|_2=1} \left(\int_{\mathbb{R}} |F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)| \, dt \right),
\end{aligned}$$

where the final inequality follows from (27) and (28). Next, recalling (29),

$$\begin{aligned}
\sup_{\|\boldsymbol{\gamma}\|_2=1} \int_{\mathbb{R}} |F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)| \, dt &\leq \sup_{\|\boldsymbol{\gamma}\|_2=1} \int_0^c \sup_t |F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)| \, ds \\
&\leq c \sup_{f \in \mathcal{H}} |\mathbb{E}_{\mathbb{P}_n}[f] - \mathbb{E}_{\mathbb{P}}[f]|\,,
\end{aligned}$$

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 2 in Section 7.2. □

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,

$$\Gamma := \mathbb{E}_{\mathbb{P}}\left[\|(\mathbf{X}, Y)\|_2^s\right] < \infty, \qquad \text{for some } s > 2r. \tag{30}$$

Our result generalizes the work of [51] and [46], 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 (30), that is easy to verify in practice.

Our main result in this section is the following:

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

$$\overline{\mathcal{W}}_r(\mathbb{P}_n, \mathbb{P})^r \leq \frac{C \, \log{(2n+1)}^{r/s}}{\sqrt{n}}, \tag{31}$$

where

$$C := 2^r r\left(180\sqrt{d+2} + \sqrt{2\log\left(\frac{1}{\alpha}\right)} + \sqrt{\frac{\Gamma}{\alpha}}\frac{8}{s/2-r}\sqrt{\log\left(\frac{8}{\alpha}\right)+(d+2)}\right). \tag{32}$$

PROOF. By Lemma 3 in Section 7 with $k = \log(2n+1)^{1/s}$ we have

$$\overline{\mathcal{W}}_r(\mathbb{P}_n,\mathbb{P})^r \le 2^r r \log(2n+1)^{r/s}\left(I_1 + \frac{\sqrt{\Gamma\vee\Gamma_n}}{s/2-r}\log(2n+1)^{-1/2} I_2\right), \tag{33}$$

where

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

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

$$\Gamma_n = \sup_{\|\boldsymbol{\gamma}\|_2=1} \mathbb{E}_{\mathbb{P}_n}\left[\left|(\mathbf{X},Y)^\top\boldsymbol{\gamma}\right|^s\right] = \sup_{\|\boldsymbol{\gamma}\|_2=1}\frac{1}{n}\sum_{i=1}^n \left|(\mathbf{X}_i,Y_i)^\top\boldsymbol{\gamma}\right|^s,$$

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

Next, by Markov's inequality and the triangle inequality

$$\mathbb{P}(\Gamma_n \ge C) \le \frac{\mathbb{E}_{\mathbb{P}}[\Gamma_n]}{C} = \frac{1}{C}\mathbb{E}_{\mathbb{P}}\left[\sup_{\|\boldsymbol{\gamma}\|_2=1}\frac{1}{n}\sum_{i=1}^n\left|(\mathbf{X}_i,Y_i)^\top\boldsymbol{\gamma}\right|^s\right] \le \frac{\Gamma}{C}.$$

Setting the last expression equal to α yields $\Gamma_n \le \Gamma/\alpha$ on a set of probability at least $1-\alpha$. Combining this with Lemma 2 (to control I_1) and Lemma 4 (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 $1-3\alpha$:

$$2^r r\log(2n+1)^{\frac{r}{s}}\left[\frac{1}{\sqrt{n}}\left(180\sqrt{d+2}+\sqrt{2\log\left(\frac{1}{\alpha}\right)}\right) + \sqrt{\frac{\Gamma}{\alpha}}\frac{1}{s/2-r}\frac{1}{\sqrt{\log(2n+1)}}I_2\right]$$
$$\le \frac{2^r r\log(2n+1)^{\frac{r}{s}}}{\sqrt{n}}\left[180\sqrt{d+2}+\sqrt{2\log\left(\frac{1}{\alpha}\right)}+\sqrt{\frac{\Gamma}{\alpha}}\frac{8}{s/2-r}\sqrt{\log\left(\frac{8}{\alpha}\right)+(d+2)}\right],$$

which is the claim. □

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 (24) and (25). The following theorem provides a Donsker type result, i.e. asymptotic $\sqrt{n}$-rates without logarithmic factors, as well as an inequality for the expectation without an explicit constant. One can then obtain concentration results similarly to [46, 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 appendix. We first consider probability measures $\mathbb{P}$ with compact support.

THEOREM 4. *If $\mathbb{P}$ is compactly supported, then*

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

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 5. *Assume*

$$\Gamma = \mathbb{E}_{\mathbb{P}}\left[\|(\mathbf{X},Y)\|_2^s\right] < \infty, \qquad \textit{for some } s > 2r,$$

and define

$$\mathcal{H}^+ := \left\{|t|^s\,\mathbb{1}_{\{t\leq \boldsymbol{x}^\top\boldsymbol{\gamma}\}} :\ (\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times[0,\infty),\ \|\boldsymbol{\gamma}\|_2 = 1\right\},$$
$$\mathcal{H}^0 := \left\{\mathbb{1}_{\{\boldsymbol{x}^\top\boldsymbol{\gamma}\leq t\}} :\ (\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times\mathbb{R},\ \|\boldsymbol{\gamma}\|_2 = 1\right\},$$
$$\mathcal{H}^- := \left\{|t|^s\,\mathbb{1}_{\{t\geq \boldsymbol{x}^\top\boldsymbol{\gamma}\}} :\ (\boldsymbol{\gamma},t)\in\mathbb{R}^{d+1}\times(-\infty,0),\ \|\boldsymbol{\gamma}\|_2 = 1\right\}.$$

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

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

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

$$\mathbb{E}\left[G_{f_1}G_{f_2}\right] = \mathbb{E}_{\mathbb{P}}\left[f_1 f_2\right] - \mathbb{E}_{\mathbb{P}}\left[f_1\right]\mathbb{E}_{\mathbb{P}}\left[f_2\right] \qquad \forall f_1,f_2\in\mathcal{H}^+\cup\mathcal{H}^0\cup\mathcal{H}^-. \tag{34}$$

Furthermore, for all $n\in\mathbb{N}$,

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

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 and guarantees that the true data generating process is contained in the ball $B_{\delta_{n,r}}(\mathbb{P}_n)$ with high probability:

$$\delta_{n,r} = \left[\left(\max\left\{\frac{1}{c_d},1\right\}\right)\frac{C\,\log(2n+1)^{r/s}}{\sqrt{n}}\right]^{1/r}, \tag{35}$$

where c_d is the constant such that $c_d\|\boldsymbol{\gamma}\|_2 \leq \rho(\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma}$ and C is the constant defined in (32).

In the case where the support of $\mathbb{P}$ is compact, we can specialize our recommendation to select the regularization parameter $\delta_{n,r}$ with guidance from Theorem 2. This recommendation is given in the following:

$$\delta_{n,r} = \left[\left(\max\left\{\frac{1}{c_d}, 1\right\}\right)\frac{C}{\sqrt{n}}\right]^{1/r}, \tag{36}$$

where C is now the constant defined in (26). Note that in the case of compact support, our recommended regularization parameter only depends on $\mathbb{P}$ through the diameter of the support of $\mathbb{P}$ measured with respect to the Euclidean norm.

The next corollary guarantees that the corresponding linear predictors that 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.

THEOREM 6. *Suppose the conditions of Theorem 3 (or 2) holds. Consider $\delta_{n,r}$ defined in* (35) *(or* (36)*). 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*

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

PROOF. Note that $\widehat{\mathcal{W}}_r(\mathbb{P}_n, \mathbb{Q}) \leq \widehat{\mathcal{W}}_r(\mathbb{P}_n, \mathbb{P}) + \widehat{\mathcal{W}}_r(\mathbb{P}, \mathbb{Q})$ by the triangle inequality because $\widehat{\mathcal{W}}_r$ is a metric for any norm ρ. 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 (or 2). In the equation above, $c_{\rho,d}$ is defined via (24). Finally, on the event E_n^c, we have

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

where the last equality follows from Theorem 1. □

Theorem 6 shows that the loss function of a penalized regression evaluated on the training sample $\mathbb{P}_n$ constitutes an upper bound for the expected loss function of a linear regression evaluated at $\mathbb{Q}$ (provided it is close to $\mathbb{P}$).

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 6 holds with $\Gamma = 2^s$ and $\sigma = \max\left\{\mathbb{E}_{\mathbb{P}}\left[|Y|^s\right]^{1/s},\ 1\right\}$.

An interesting avenue for future work is to use the results in [50] (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 Theorem 1 of [50] (which only depend on r), and additionally provide some results for the consistent estimation of the operator norm of Σ. Because the rates in [50] are faster than ours (when d is fixed, their rates are of order $n^{1/2}$ wheras ours are of order $n^{1/4}$), this means that the regularization parameters based on the results of [50] will typically be smaller (making it less likely that the robust predictors ignore the available covariates).

5.2. *Covariate Normalization.* In this section, we assume that the covariates in the data have been *normalized* to satisfy $\mathbb{E}_{\mathbb{P}_n}[\|\mathbf{X}\|_2^s] = 1$. This means that under minimal regularity conditions we can assume that the true data generating process satisfies $\mathbb{E}_{\mathbb{P}}[\|\mathbf{X}\|_2^s] = 1$. 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 $\delta_{n,r}$ under our suggested normalization.

THEOREM 7. *Suppose* $\mathbb{E}_{\mathbb{P}}[\|\mathbf{X}\|_2^s] = 1$ *and* $\mathbb{E}_{\mathbb{P}}[|Y|^s]^{1/s} < +\infty$ *for some* $s > 2r$. *In addition, assume that* $\rho(\cdot) = \|\cdot\|$ *is an arbitrary norm in* $\mathbb{R}^d$ *such that* $c_d\|\boldsymbol{\gamma}\| \leq \rho(\boldsymbol{\gamma})$, *and* $\sigma = \max\left\{\mathbb{E}_{\mathbb{P}}[|Y|^s]^{1/s},\ 1\right\}$. *Consider*

$$\delta_{n,r} := \left[\left(\max\left\{\frac{1}{c_d}, 1\right\}\right)\frac{C\,\log(2n+1)^{r/s})}{\sqrt{n}}\right]^{1/r} \tag{37}$$

where

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

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

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

PROOF. 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 1 and 3.

Step 1: Let us compare the ρ-MSW metric to the MSW metric using reasoning that is similar to our derivations in (22) – (24). Defining $\bar{\gamma}_\sigma^\top = (\gamma^\top, -\sigma)$, we obtain

$$\begin{aligned}\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}}) &= \sup_{\gamma\in\mathbb{R}^d} \frac{\|\bar{\gamma}_\sigma\|_2}{\sigma+\|\gamma\|_2}\frac{1}{\|\bar{\gamma}_\sigma\|_2}\mathcal{W}_r\left(\left[(\mathbf{X},Y/\sigma)^\top\bar{\gamma}_\sigma\right]_*\mathbb{P},\, [(\widetilde{\mathbf{X}},\widetilde{Y}/\sigma)^\top\bar{\gamma}_\sigma]_*\widetilde{\mathbb{P}}\right)\\ &\le (\max\{1/c_d,1\})\sup_{\|\gamma\|_2=1}\mathcal{W}_r\left(\left[(\mathbf{X},Y/\sigma)^\top\gamma\right]_*\mathbb{P},\,[(\widetilde{\mathbf{X}},\widetilde{Y}/\sigma)^\top\gamma]_*\widetilde{\mathbb{P}}\right)\\ &= (\max\{1/c_d,1\})\,\overline{\mathcal{W}}_r\left(\mathbb{P}^\sigma,\widetilde{\mathbb{P}}^\sigma\right),\end{aligned}$$

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 to compute the rates for $\overline{\mathcal{W}}_r(\mathbb{P}^\sigma,\widetilde{\mathbb{P}}^\sigma)$, which depend on $\mathbb{E}_{\mathbb{P}^\sigma}[\|(\mathbf{X},Y)\|_2^s]$. Consider the following derivation

$$\mathbb{E}_{\mathbb{P}^\sigma}[\|(\mathbf{X},Y)\|_2^s] \le 2^{s-1}\left(\mathbb{E}_{\mathbb{P}^\sigma}[\|\mathbf{X}\|_2^s]+\mathbb{E}_{\mathbb{P}^\sigma}[|Y|^s]\right)\,.$$

Note that $\mathbb{E}_{\mathbb{P}^\sigma}[\|\mathbf{X}\|_2^s] = \mathbb{E}_{\mathbb{P}}[\|\mathbf{X}\|_2^s] = 1$ and $\mathbb{E}_{\mathbb{P}^\sigma}[|Y|^s] = \mathbb{E}_{\mathbb{P}}[|Y/\sigma|^s] \le 1$ due to our assumptions. This implies that

$$\mathbb{E}_{\mathbb{P}^\sigma}[\|(\mathbf{X},Y)\|_2^s] \le 2^s\,.$$

Step 3: We note that Theorem 3 still holds for any Γ larger than $\mathbb{E}_{\mathbb{P}}[\|(\mathbf{X},Y)\|_2^s]$. In particular, we can consider $\Gamma = 2^s$ due to Step 2. In addition, we note that the conclusion of Theorem 1 is unaffected by the choice of $\sigma \ge 1$. These observations and the same argument presented in the proof of Theorem 6 conclude our proof. □

5.3. *Asymptotic recommendation.* For compactly supported measures, Theorem 4 yields the asymptotic *oracle* recommendation

$$\delta_{n,r} = \left[c_{\rho,d}\, n^{-1/2}\cdot q_{1-\alpha}\right]^{1/r}\cdot\operatorname{diam}(\operatorname{supp}(\mathbb{P}))\,, \tag{38}$$

where $c_{\rho,d}$ is as (24), $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the Kolmogorov distribution. In the general case, Theorem 5 yields

$$\delta_{n,r} = \left[(\Gamma n)^{-1/2}\cdot C\right]^{1/r}$$

for some constant $C = C(r,s,d,\alpha)$. However, the constant C is hard to determine explicitly. In particular 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 (34) and is hard to bound explicitly. We leave this issue for future research.

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-sample performance than $\boldsymbol{\beta}_2$ over an uncertainty set B. That is,

$$\sup_{\mathbb{Q}\in B} \mathbb{E}_{\mathbb{Q}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}_1\right|^r\right]^{1/r} \;\leq\; \sup_{\mathbb{Q}\in B} \mathbb{E}_{\mathbb{Q}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}_2\right|^r\right]^{1/r}. \tag{39}$$

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

(i) $B \subseteq B_\delta(\mathbb{P}) = B_\delta^{r,\rho,\sigma}(\mathbb{P})$ for some σ, δ, and $\rho(\cdot)$.
(ii) The supremum on the left side of (39) is achieved for $\mathbb{P}^*_{\boldsymbol{\beta}_1}$, and the supremum on the right side of (39) is achieved for $\mathbb{P}^*_{\boldsymbol{\beta}_2}$, where $\mathbb{P}^*_{\boldsymbol{\beta}_j}$ are defined according to Corollary 1, $j=1,2$.

Examples of such sets B are given in Section 2.1. Note that we cannot evaluate (39) directly, as $\mathbb{P}$ is not observed. Instead we propose the test statistic

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

The next corollary states that T_n gives rise to a size–α test. For notational simplicity we focus here on compactly supported probability measures $\mathbb{P}$, and simply remark that the same reasoning can be used to derived tests for general $\mathbb{P}$ satisfying the assumptions of Theorems 3 and 5.

COROLLARY 2. *In the setting of Theorems 1 and 2, consider C and $c_{\rho,d}$ defined in* (26) *and* (24). *Then we have*

$$P\left(T_n > c_{\rho,d}C^{1/r}\right) \leq \alpha$$

for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ satisfying (39).

PROOF. By Theorem 1 and conditions (i), (ii) above, (39) is equivalent to

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

The previous expression is equivalent to

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

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

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

Therefore we have

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

Using $\widehat{\mathcal{W}}_r(\mathbb{P}, \mathbb{P}_n) \le c_{\rho,d} \overline{\mathcal{W}}_r(\mathbb{P}_n, \mathbb{P})$ from (24), we derive that

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

Finally, Theorem 2 implies that the above probability is bounded by α. □

6. Simulations. Suppose that the training data consists of n i.i.d. draws from a linear regression model, in other words

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \sigma \varepsilon_i.$$

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 according to

$$\mathbf{X}_i = \sigma \lambda \widetilde{\mathbf{X}}_i,$$

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 $(\boldsymbol{\beta}, \sigma, \lambda, d)$.

We first focus on linear prediction using coefficients estimated via the $\sqrt{\text{LASSO}}$ $(r = 2)$. Recall from (38) that our oracle recommendation for the tuning parameter 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 7.4.1) that

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

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

$$n^{-1/2} \cdot 3^{-1/2} \sigma\lambda \cdot \Phi^{-1}\left(\frac{1}{2} + \frac{(1-\alpha)^{(1/d)}}{2}\right). \tag{40}$$

Figure 2 compares the ratio of (38) relative to (40). 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_n could be achieved by setting all the coefficients to be zero (an adversarial nature cannot increase much the generalization error of such a predictor, as it does not rely at all on covariates).

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 that $\boldsymbol{\beta} = 0_{d\times 1}$ is a solution to the $\sqrt{\text{LASSO}}$ problem if and only if

$$\frac{\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i y_i\|_\infty}{\sqrt{\frac{1}{n}\sum_{i=1}^n y_i^2}} \le \delta_{n,2}, \tag{41}$$

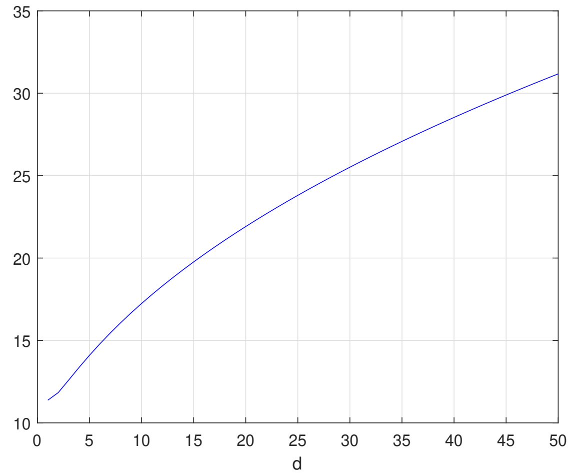


Fig 2: (Blue, Solid Line) Ratio of the oracle recommendations in (38) and (40): $\sigma = 1, \lambda = 10, \alpha = 0.05, n = 2,500$ and $\boldsymbol{\beta} = [1, 0, \ldots 0]^\top$.

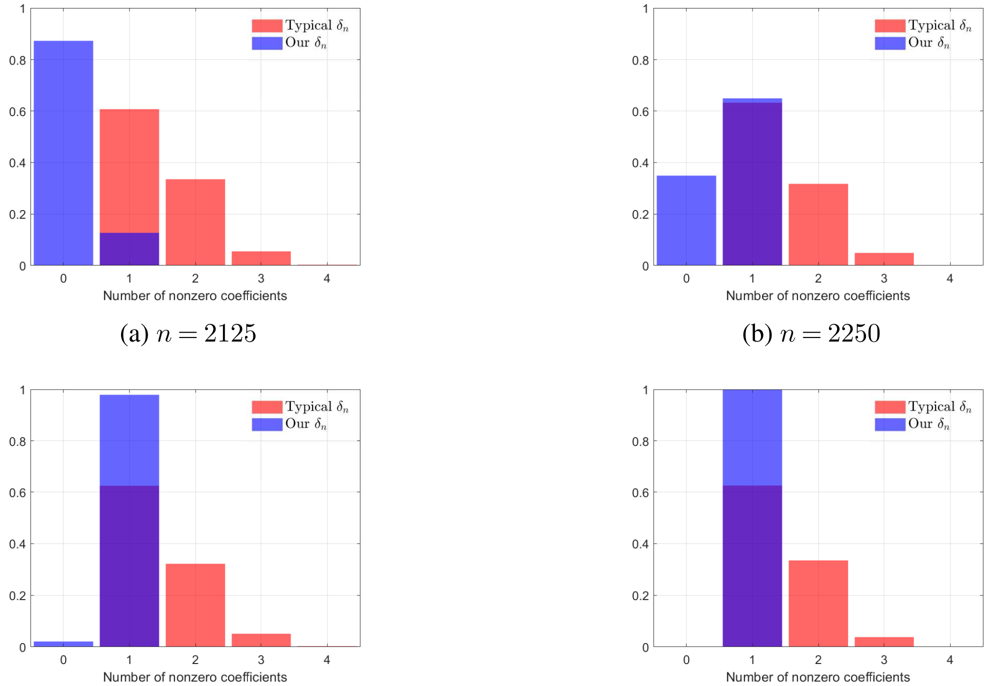
see [64]. Using a Central Limit Theorem and a Law of Large of numbers, algebra shows (see Appendix 7.4.2) that the event (41) holds with high probability whenever

$$n \leq 9 \cdot \left\| \frac{\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}} \right\|_\infty^{-4} \cdot (q_{1-\alpha})^2 \cdot \left(d + (\|\boldsymbol{\beta}\|_1 + (2/\lambda))^2 \right)^2. \tag{42}$$

For $d = 10$, $\boldsymbol{\beta} = (1, 0, 0 \ldots 0)^\top$, $\alpha = .05$ (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 $\sigma = 1, \lambda = 10, d = 10, \boldsymbol{\beta} = [1, 0, \ldots, 0]^\top$ 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 below reports the histogram associated to the number of nonzero coefficients selected by the $\sqrt{\text{LASSO}}$ using the regularization parameters in (38) and (40). The numerical results reported below are in line with the bound derived in (42).

Training/Testing error. Figure 4 below 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 in (40), and the $\sqrt{\text{LASSO}}$ with the δ_n in (38). 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 1 with δ_n in (38) replacing δ. The plots report the histogram—across simulations—of the "relative" root mean-squared prediction error in the training (or testing) data. Each figure compares the estimators indicated in the legend below them. For example, Panel a) of Figure 4



(a) $n = 2125$ (b) $n = 2250$ (c) $n = 2375$ (d) $n = 2500$

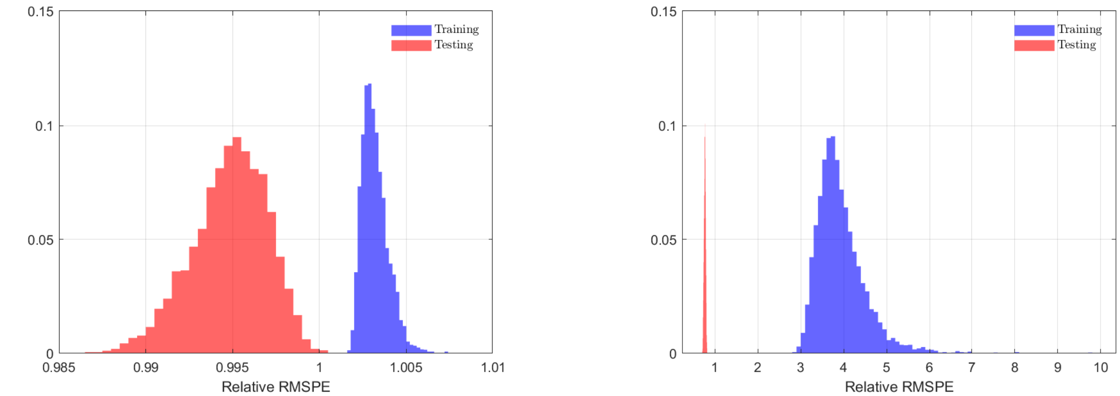
Fig 3: Fraction of simulation draws in which the $\sqrt{\text{LASSO}}$ selects 0,1,. . . , 4 nonzero coefficients: (Blue) oracle δ_n in (40); (Red) our oracle recommendation of δ_n in (38).

reports the root MSPE of the $\sqrt{\text{LASSO}}$ (SQL), 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 (40) 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 (40) and (38) generates a sizeable difference in training error. However, that the larger value of δ_n does translate to better out-of-sample performance.

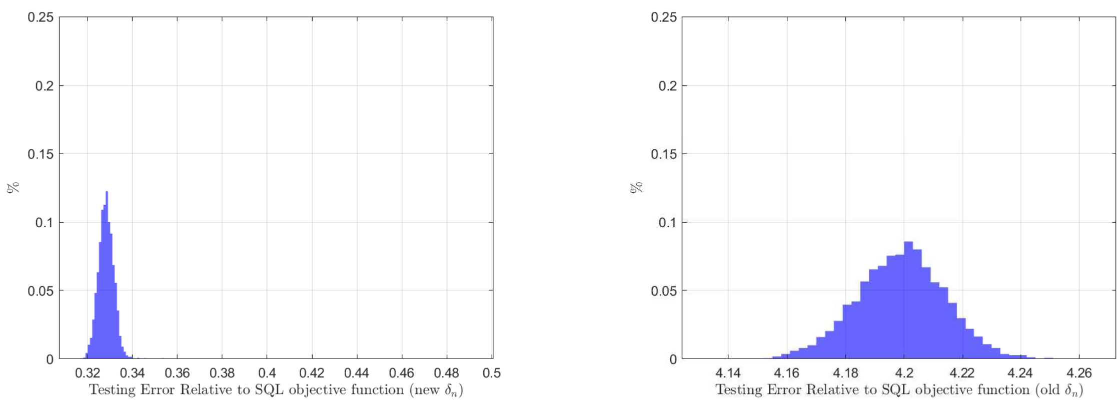
Finally, we verify the bound in Theorem 6. The corollary implies that with probability at least 95% the root-MSPE of the $\sqrt{\text{LASSO}}$ in the *testing* set (for any distribution in the ball that is ϵ away from the true DGP) must be bounded by the sum of a) the root-MSPE of the $\sqrt{\text{LASSO}}$ in the *training* set and ii) $(\delta_n + \epsilon)(\sigma + \rho(\boldsymbol{\beta}))$. The figure below shows that the bound holds for the new δ_n, but not for the old one.

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



(a) SQL/OLS (b) SQL (new δ_n) / SQL

Fig 4: Relative Training/Testing Root Mean-Squared Prediction Error.



(a) SQL (new δ_n) (b) SQL (old δ_n)

Fig 5: Testing error and Objective function

LEMMA 1. *The ρ-max-sliced Wasserstein $\widehat{\mathcal{W}}_r$ distance is a metric.*

PROOF. Recall from (22) that

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

where $\bar{\gamma}^\top = (\gamma^\top, -1)$. 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 realize that because $\sigma + \rho(\gamma) > 0$, it follows that

$\widehat{\mathcal{W}}_r(\mathbb{P},\widetilde{\mathbb{P}})=0$ implies

$$\mathcal{W}_r\left(\left[(\mathbf{X},Y)^\top\bar{\boldsymbol{\gamma}}\right]_*\mathbb{P},\ \left[(\widetilde{\mathbf{X}},\widetilde{Y})^\top\bar{\boldsymbol{\gamma}}\right]_*\widetilde{\mathbb{P}}\right)=0 \qquad \forall\boldsymbol{\gamma}\in\mathbb{R}^d. \tag{43}$$

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

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

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

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

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

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

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

7.1. *Proof of Theorem 1.*

PROOF. In Section 2 we provide a proof sketch after the statement of Theorem 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

$$\left(\mathbb{E}_{\widetilde{\mathbb{P}}}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]\right)^{1/r}\le\sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[|Y-\mathbf{X}^\top\boldsymbol{\beta}|^r\right]}+\delta\left(\sigma+\rho(\boldsymbol{\beta})\right), \tag{44}$$

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 $\widehat{\mathcal{W}}_{r,\rho,\sigma}$. 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

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

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

$$\sqrt[r]{\mathbb{E}_{\pi(\boldsymbol{\beta})}\left[\left|\widetilde{Y}-\widetilde{\mathbf{X}}^\top\boldsymbol{\beta}\right|^r\right]} \leq \sqrt[r]{\mathbb{E}_{\mathbb{P}}\left[|Y-\mathbf{X}^\top\boldsymbol{\beta}|^r\right]}+\delta\left(\sigma+\rho(\boldsymbol{\beta})\right). \tag{45}$$

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 (11).

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

$$\widetilde{\mathbf{X}}=\mathbf{X}-e\left(\boldsymbol{\beta}^*-\frac{\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}\rho^*(\boldsymbol{\beta}^*)\right), \qquad \widetilde{Y}=Y+\sigma e, \tag{46}$$

where

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

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

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

where the last inequality follows since

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

for any $\boldsymbol{\gamma}\in\mathbb{R}^d$ by the assumption in (11) 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 (46)

$$\begin{aligned}\widetilde{Y}-\widetilde{\mathbf{X}}^\top\boldsymbol{\beta}&=\widetilde{Y}-Y+Y-\mathbf{X}^\top\boldsymbol{\beta}+(\mathbf{X}-\widetilde{\mathbf{X}})^\top\boldsymbol{\beta}\\ &=\left(Y-\mathbf{X}^\top\boldsymbol{\beta}\right)+e\left(\sigma+\rho(\boldsymbol{\beta})\right),\end{aligned} \tag{47}$$

where the last term applies [58, 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 (47) 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

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

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

□

7.2. *Proofs for Lemmas used in Theorems 2 and 3.*

LEMMA 2 (cf. [68, Theorem 4.10], [26, Chapter 4]). *Let*

$$\mathcal{H} := \left\{\mathbb{1}_{\{\mathbf{x}^\top\boldsymbol{\gamma} \le t\}} : \boldsymbol{\gamma} \in \mathbb{R}^{d+1},\, t \in \mathbb{R}\right\}, \tag{48}$$

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

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

PROOF. By [68, Theorem 4.10], we have

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

where

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

is the Rademacher complexity of $\mathcal{H}$. Next, following [26, statement and proof of Theorem 3.2], we obtain

$$\mathcal{R}_n(\mathcal{H}) \leq \frac{12}{\sqrt{n}} \max_{\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^{d+1}} \int_0^1 \sqrt{2 \log N\left(\mathsf{r}, \mathcal{H}(\mathbf{x}_1^n)\right)}\, d\mathsf{r}, \tag{49}$$

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

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

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 $\mathbf{b}, \mathbf{d} \in B$. [26, Theorem 4.3] states that

$$N(r, \mathcal{H}(\mathbf{x}_1^n)) \leq \left(\frac{4e}{r^2}\right)^{V/(1-1/e)} = \left(\frac{4e}{r^2}\right)^{Ve/(e-1)}, \tag{50}$$

where V is the VC-dimension of $\mathcal{H}$. Furthermore, by [26, Corollary 4.2], the VC-dimension of $\mathcal{H}$ is bounded by $d+2$, i.e. $V \leq d+2$. In conclusion, using (50),

$$\log N\left(\mathsf{r}, \mathcal{H}(\mathbf{x}_1^n)\right) \leq \frac{eV}{e-1} \log\left(\frac{4e}{\mathsf{r}^2}\right) \leq \frac{e(d+2)}{e-1} \log\left(\frac{4e}{\mathsf{r}^2}\right). \tag{51}$$

Following [26, proof of Theorem 3.3] we estimate

$$\int_0^1 \sqrt{\log\left(\frac{4e}{\mathsf{r}^2}\right)}\, d\mathsf{r} \leq \sqrt{2\pi e}, \tag{52}$$

so that from (51) and (52), 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 (49), this yields

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

□

LEMMA 3. *Define*

$$\Gamma_n := \sup_{\|\boldsymbol{\gamma}\|_2 = 1} \mathbb{E}_{\mathbb{P}_n}\left[\left|(\mathbf{X}, Y)^\top \boldsymbol{\gamma}\right|^s\right] = \sup_{\|\boldsymbol{\gamma}\|_2 = 1} \frac{1}{n} \sum_{i=1}^n \left|(\mathbf{X}, Y)_i^\top \boldsymbol{\gamma}\right|^s.$$

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

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

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

PROOF. We first note that [13, Proposition 7.14] yields, for any $k > 0$,

(53)
$$\begin{aligned}\mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}})^r &\leq r2^{r-1}\int |t|^{r-1}|F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t)|\,dt \\ &\leq r(2k)^r \sup_t |F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))| \\ &\quad + r2^{r-1}\int_{\mathbb{R}\setminus[-k,k]} |t|^{r-1}\sqrt{F_{\boldsymbol{\gamma}}(t)(1-F_{\boldsymbol{\gamma},n}(t))}\,\frac{(F_{\boldsymbol{\gamma}}(t) - F_{\boldsymbol{\gamma},n}(t))^+}{\sqrt{F_{\boldsymbol{\gamma}}(t)(1-F_{\boldsymbol{\gamma},n}(t))}}\,dt \\ &\quad + r2^{r-1}\int_{\mathbb{R}\setminus[-k,k]} |t|^{r-1}\sqrt{F_{\boldsymbol{\gamma},n}(t)(1-F_{\boldsymbol{\gamma}}(t))}\,\frac{(F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))^+}{\sqrt{F_{\boldsymbol{\gamma},n}(t)(1-F_{\boldsymbol{\gamma}}(t))}}\,dt.\end{aligned}$$

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

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

Plugging these bounds into (53), we obtain

(54)
$$\begin{aligned}&\mathcal{W}_r(\mathbb{P}_{\boldsymbol{\gamma},n}, \mathbb{P}_{\boldsymbol{\gamma}})^r \\ &\leq r(2k)^r \sup_t |F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))| \\ &+ r2^{r-1}\int_{\mathbb{R}\setminus[-k,k]} |t|^{r-1-s/2}\sqrt{\mathbb{E}_{\mathbb{P}}[|(\mathbf{X}, Y)^\top\boldsymbol{\gamma}|^s] \vee \mathbb{E}_{\mathbb{P}_n}[|(\mathbf{X}, Y)^\top\boldsymbol{\gamma}|^s]}\,\frac{(F_{\boldsymbol{\gamma}}(t) - F_{\boldsymbol{\gamma},n}(t))^+}{\sqrt{F_{\boldsymbol{\gamma}}(t)(1-F_{\boldsymbol{\gamma},n}(t))}}\,dt \\ &+ r2^{r-1}\int_{\mathbb{R}\setminus[-k,k]} r|t|^{r-1-s/2}\sqrt{\mathbb{E}_{\mathbb{P}}[|(\mathbf{X}, \mathbf{Y})^\top\boldsymbol{\gamma}|^s] \vee \mathbb{E}_{\mathbb{P}_n}[|(\mathbf{X}, Y)^\top\boldsymbol{\gamma}|^s]}\,\frac{(F_{\boldsymbol{\gamma},n}(t) - F_{\boldsymbol{\gamma}}(t))^+}{\sqrt{F_{\boldsymbol{\gamma},n}(t)(1-F_{\boldsymbol{\gamma}}(t))}}\,dt.\end{aligned}$$

Recall that we have assumed $s/2 > r$ where $r \geq 1$. 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 (54) thus yields the claim. □

LEMMA 4. *With probability greater than* $1-\alpha$ *we have*

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

PROOF OF LEMMA 4. We first define

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

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

$$\begin{aligned}\sup_{(\gamma,t)\in\mathbb{R}^{d+1}\times\mathbb{R}} &\frac{(F_{\gamma}(t)-F_{\gamma,n}(t))^+}{\sqrt{F_{\gamma}(t)(1-F_{\gamma,n}(t))}} \\ &\leq 2\left(\sup_{f\in\mathcal{J}}\frac{(\mathbb{E}_{\mathbb{P}}[f]-\mathbb{E}_{\mathbb{P}_n}[f])^+}{\sqrt{\mathbb{E}_{\mathbb{P}}[f]}} \vee \sup_{f\in\mathcal{J}}\frac{(\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f])^+}{\sqrt{\mathbb{E}_{\mathbb{P}_n}[f]}}\right).\end{aligned} \tag{55}$$

By symmetry,

$$\begin{aligned}\sup_{(\gamma,t)\in\mathbb{R}^{d+1}\times\mathbb{R}} &\frac{(F_{\gamma,n}(t)-F_{\gamma}(t))^+}{\sqrt{F_{\gamma,n}(t)(1-F_{\gamma}(t))}} \\ &\leq 2\left(\sup_{f\in\mathcal{J}}\frac{(\mathbb{E}_{\mathbb{P}}[f]-\mathbb{E}_{\mathbb{P}_n}[f])^+}{\sqrt{\mathbb{E}_{\mathbb{P}}[f]}} \vee \sup_{f\in\mathcal{J}}\frac{(\mathbb{E}_{\mathbb{P}_n}[f]-\mathbb{E}_{\mathbb{P}}[f])^+}{\sqrt{\mathbb{E}_{\mathbb{P}_n}[f]}}\right).\end{aligned} \tag{56}$$

Concentration for the terms on the right hand side of equations (55) and (56) is well studied: indeed, e.g. by [26, 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 [26, 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 [26, 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. □

7.3. *Proofs of asymptotic results.*

PROOF OF THEOREM 5. Note that again by [13, Proposition 7.14] we have

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

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

$$\sup_{f \in \mathcal{H}^+} |f(\boldsymbol{x})| \le |\boldsymbol{x}^\top \boldsymbol{\gamma}|^s \le \|\boldsymbol{x}\|_2^s.$$

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

$$F(\boldsymbol{x}) := \sup_{f \in \mathcal{H}^+ \cup \mathcal{H}^0 \cup \mathcal{H}^-} |f(\boldsymbol{x})| \le \|\boldsymbol{x}\|_2^s \vee 1.$$

As $\mathcal{H}^0$ is VC-subgraph by Lemma 2, [Van der Vaart, Wellner, Lemma 2.6.22] implies that $\mathcal{H}^+$ and $\mathcal{H}^-$ are also VC-subgraph: indeed note that

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

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

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

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

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

This shows the second claim. □

PROOF OF THEOREM 4. This claim follows from the estimate

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

stated in the proof of Theorem 2 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 5. As $\sup_{f \in \mathcal{H}^0} |G_f|$ is dominated in stochastic order by $\sup_{t \in [0,1]} |B(t)|$, this concludes the proof. □

7.4. *Additional Derivations.*

7.4.1. *Diameter of the support of $\mathbb{P}$ in the simulation.* Notice that $\widetilde{\mathbf{X}}_i^\top \boldsymbol{\beta} \ge -\|(\boldsymbol{\beta})^-\|_1$ where the equality holds, making equal to ones the entries of $\widetilde{\mathbf{X}}_i$ that corresponds negative values of $\boldsymbol{\beta}$. Similarly, $\widetilde{\mathbf{X}}_i^\top \boldsymbol{\beta} \le \|(\boldsymbol{\beta})^+\|_1$. Since $\mathbf{X}_i = \sigma\lambda\widetilde{\mathbf{X}}_i$, it follows that $\inf_{\mathbf{X}_i} X_i^\top \boldsymbol{\beta} = -\sigma\lambda\|(\boldsymbol{\beta})^-\|_1$ and $\sup_{\mathbf{X}_i} \mathbf{X}_i^\top \boldsymbol{\beta} = \sigma\lambda\|(\boldsymbol{\beta})^+\|_1$. Therefore, $Y_i \in [-\sigma(\lambda\|(\boldsymbol{\beta})^-\|_1 + 1), \sigma(\lambda\|(\boldsymbol{\beta})^+\|_1 + 1)]$. Then, diameter of the support equals

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

7.4.2. *Derivation of Equation* (42). The tuning parameter $\delta_{n,2}$ used in the simulation is

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

According to equation (41) it is known that $\boldsymbol{\beta} = 0_{d\times 1}$ is a solution to the $\sqrt{\text{LASSO}}$ problem if and only if

$$\frac{\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i y_i\|_\infty}{\sqrt{\frac{1}{n}\sum_{i=1}^n y_i^2}} \le \delta_{n,2} = n^{-1/4} \cdot C_{\text{sim}}. \tag{57}$$

Since

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

Equation (57) holds if and only if

$$\frac{\| \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\right)\boldsymbol{\beta} + \frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{X}_i\varepsilon_i\|_\infty}{\sqrt{\frac{1}{n}\sum_{i=1}^n y_i^2}} \leq n^{-1/4}\cdot C_{\text{sim}}. \tag{58}$$

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

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

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

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

For the denominator,

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

where we have used the fact that ϵ_i is mean zero and independent of $\widetilde{\mathbf{X}}_i$. The left-hand side of (58) is thus bounded above with high probability by

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

This means that the event in (57) occurs with high probability if

$$\sqrt{1/3}\cdot\sigma\cdot\lambda\cdot\left\|\frac{\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top\boldsymbol{\beta}}}\right\|_\infty \leq \frac{1}{n^{1/4}}C_{\text{sim}}. \tag{59}$$

Using the definition of C_{sim}, the event in (59) 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.

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.

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

[10] 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.

[11] 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*.

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

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

[14] 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.

[15] 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.

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

[17] 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.

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

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

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

[21] 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.

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

[23] CHU, H. T. M., K.-C. TOH, AND Y. ZHANG (2021): “On Regularized Square-root Regression Problems: Distributionally Robust Interpretation and Fast Computations,” .

[24] 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.

[25] 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.

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

[27] DUCHI, J., P. GLYNN, AND H. NAMKOONG (2016): “Statistics of robust optimization: A generalized empirical likelihood approach,” *arXiv preprint arXiv:1610.03425*.

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

[29] FOURNIER, N. (2022): “Convergence of the empirical measure in expected Wasserstein distance: non asymptotic explicit bounds in $\mathbb{R}^d$,” *arXiv preprint arXiv:2209.00923*.

[30] 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.

[31] 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.

[32] GAO, R., X. CHEN, AND A. J. KLEYWEGT (2017): “Distributional robustness and regularization in statistical learning,” *arXiv preprint arXiv:1712.06050*.

[33] GAO, R. AND A. KLEYWEGT (2022): “Distributionally robust stochastic optimization with Wasserstein distance,” *Mathematics of Operations Research*.

[34] GAO, R. AND A. J. KLEYWEGT (2016): “Distributionally robust stochastic optimization with Wasserstein distance,” *arXiv preprint arXiv:1604.02199*.

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

[36] GOLDFELD, Z., K. KATO, G. RIOUX, AND R. SADHU (2022): “Statistical inference with regularized optimal transport,” *arXiv preprint arXiv:2205.04283*.

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

[38] GOODFELLOW, I. J., J. SHLENS, AND C. SZEGEDY (2014): “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*.

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

[40] 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.

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

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

[43] LEE, J. AND M. RAGINSKY (2018): “Minimax statistical learning with wasserstein distances,” *Advances in Neural Information Processing Systems*, 31.

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

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

[46] 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.

[47] MANSOUR, Y., M. MOHRI, AND A. ROSTAMIZADEH (2009): "Domain adaptation: Learning bounds and algorithms," *arXiv preprint arXiv:0902.3430*.

[48] 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.

[49] NGUYEN, V. A., S. S. ABADEH, D. FILIPOVIĆ, AND D. KUHN (2021): "Mean-covariance robust risk measurement," *arXiv preprint arXiv:2112.09959*.

[50] 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.

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

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

[53] ——— (2019): "Subspace Robust Wasserstein distances," *ArXiv*, abs/1901.08949.

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

[55] 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.

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

[57] 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.

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

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

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

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

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

[63] SINHA, A., H. NAMKOONG, R. VOLPI, AND J. DUCHI (2017): "Certifying some distributional robustness with principled adversarial training," *arXiv preprint arXiv:1710.10571*.

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

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

[66] SUGIYAMA, M. AND K.-R. MÜLLER (2005): "Input-dependent estimation of generalization error under covariate shift," .

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

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

[69] 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.

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

[71] WEED, J. AND Q. BERTHET (2019): "Estimation of smooth densities in Wasserstein distance," in *Conference on Learning Theory*, PMLR, 3118–3119.

[72] 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.

[73] WU, Q., J. Y.-M. LI, AND T. MAO (2022): "On Generalization and Regularization via Wasserstein Distributionally Robust Optimization," *arXiv preprint arXiv:2212.05716*.