Journal of Machine Learning Research 19 (2018) 1-48 Submitted 6/17; Revised 4/18; Published 8/18

A Robust Learning Approach for Regression Models Based on Distributionally Robust Optimization

Ruidi Chen RCHEN15@BU.EDU
Division of Systems Engineering,
Boston University,
Boston, MA 02215, USA

Ioannis Ch. Paschalidis YANNISP@BU.EDU
Department of Electrical and Computer Engineering,
Division of Systems Engineering,
and Department of Biomedical Engineering,
Boston University,
Boston, MA 02215, USA
sites. bu. edu/ paschalidis

Editor: Edo Airoldi

Abstract

We present a *Distributionally Robust Optimization (DRO)* approach to estimate a robustified regression plane in a linear regression setting, when the observed samples are potentially contaminated with adversarially corrupted outliers. Our approach mitigates the impact of outliers by hedging against a family of probability distributions on the observed data, some of which assign very low probabilities to the outliers. The set of distributions under consideration are close to the empirical distribution in the sense of the Wasserstein metric. We show that this DRO formulation can be relaxed to a convex optimization problem which encompasses a class of models. By selecting proper norm spaces for the Wasserstein metric, we are able to recover several commonly used regularized regression models. We provide new insights into the regularization term and give guidance on the selection of the regularization coefficient from the standpoint of a confidence region. We establish two types of performance guarantees for the solution to our formulation under mild conditions. One is related to its out-of-sample behavior (prediction bias), and the other concerns the discrepancy between the estimated and true regression planes (estimation bias). Extensive numerical results demonstrate the superiority of our approach to a host of regression models, in terms of the prediction and estimation accuracies. We also consider the application of our robust learning procedure to outlier detection, and show that our approach achieves a much higher AUC (Area Under the ROC Curve) than M-estimation (Huber, 1964, 1973).
Keywords: Robust Learning, Distributionally Robust Optimization, Wasserstein Metric, Regularized Regression, Generalization Guarantees.

1. Introduction

Consider a linear regression model with response $y \in \mathbb{R}$, predictor vector $\mathbf{x} \in \mathbb{R}^{m-1}$, regression coefficient $\boldsymbol{\beta}^* \in \mathbb{R}^{m-1}$ and error $\epsilon \in \mathbb{R}$:

$$y = \mathbf{x}'\boldsymbol{\beta}^* + \epsilon.$$

©2018 Ruidi Chen and Ioannis Ch. Paschalidis.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v19/17-295.html.

Given samples $(\mathbf{x}_i, y_i), i = 1, \ldots, N$, we are interested in estimating $\boldsymbol{\beta}^*$. The *Ordinary Least Squares (OLS)* minimizes the sum of squared residuals $\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$, and works well if all the N samples are generated from the underlying true model. However, when faced with adversarial perturbations in the training data, the OLS estimator will deviate from the true regression plane to reduce large residuals. Alternatively, one can choose to minimize the sum of absolute residuals $\sum_{i=1}^{N}|y_i - \mathbf{x}_i'\boldsymbol{\beta}|$, as done in *Least Absolute Deviation (LAD)*, to mitigate the influence of large residuals. Another commonly used approach for hedging against outliers is M-estimation (Huber, 1964, 1973), which minimizes a symmetric loss function $\rho(\cdot)$ of the residuals in the form $\sum_{i=1}^{N}\rho(y_i - \mathbf{x}_i'\boldsymbol{\beta})$, which downweights the influence of samples with large absolute residuals. Several choices for $\rho(\cdot)$ include the Huber function (Huber, 1964, 1973), the Tukey's Biweight function (Rousseeuw and Leroy, 2005), the logistic function (Coleman et al., 1980), the Talwar function (Hinich and Talwar, 1975), and the Fair function (Fair, 1974).

Both LAD and M-estimation are not resistant to large deviations in the predictors. For contamination present in the predictor space, high breakdown value methods are required. Examples include the *Least Median of Squares (LMS)* (Rousseeuw, 1984), which minimizes the median of the absolute residuals, the *Least Trimmed Squares (LTS)* (Rousseeuw, 1985), which minimizes the sum of the q smallest squared residuals, and S-estimation (Rousseeuw and Yohai, 1984), which has a higher statistical efficiency than LTS with the same breakdown value. A combination of the high breakdown value method and M-estimation is the MM-estimation (Yohai, 1987). It has a higher statistical efficiency than S-estimation. We refer the reader to the book of Rousseeuw and Leroy (2005) for a detailed description of these robust regression methods.

The aforementioned robust estimation procedures focus on modifying the objective function in a heuristic way with the intent of minimizing the effect of outliers. A more rigorous line of research explores the underlying stochastic program that leads to the sample-based estimation procedures. For example, the OLS objective can be viewed as minimizing the expected squared residual under the uniform empirical distribution over the samples. It has been well recognized that optimizing under the empirical distribution yields estimators that are sensitive to perturbations in the data and suffer from overfitting. The reason is that when the data $(\mathbf{x}, y)$ are adversarially corrupted by outliers, the observed samples do not represent well the true underlying distribution of the data. Yet, the samples are typically the only information available. Instead of equally weighting all the samples as in the empirical distribution, we may wish to include more informative distributions that "drive out" the corrupted samples. One way to realize this is to hedge the expected loss against a family of distributions that include the true data-generating mechanism with a high confidence; an approach called *Distributionally Robust Optimization (DRO)*. DRO minimizes the worst-case expected loss over a probabilistic ambiguity set $\mathcal{P}$ that is constructed from the observed samples and characterized by certain known properties of the true data-generating distribution. For example, Mehrotra and Zhang (2014) study the distributionally robust least squares problem with $\mathcal{P}$ defined through either moment constraints, norm bounds with moment constraints, or a confidence region over a reference probability measure. Compared to the single distribution-based stochastic optimization, DRO often results in better out-of-sample performance due to its distributional robustness.

The existing literature on DRO can be split into two main branches according to the way in which $\mathcal{P}$ is defined. One is through a moment ambiguity set, which contains all distributions that satisfy certain moment constraints (see Popescu, 2007; Delage and Ye, 2010; Goh and Sim, 2010; Zymler et al., 2013; Wiesemann et al., 2014). In many cases, it leads to a tractable DRO problem but has been criticized for yielding overly conservative solutions (Wang et al., 2016). The other is to define $\mathcal{P}$ as a ball of distributions using some probabilistic distance functions such as the ϕ-divergences (Bayraksan and Love, 2015), which include the Kullback-Leibler (KL) divergence (Hu and Hong, 2013; Jiang and Guan, 2015) as a special case, the Prokhorov metric (Erdoğan and Iyengar, 2006), and the Wasserstein distance (Esfahani and Kuhn, 2015; Gao and Kleywegt, 2016; Zhao and Guan, 2015; Luo and Mehrotra, 2017; Blanchet and Murthy, 2016). Deviating from the stochastic setting, there are also some works focusing on deterministic robustness. El Ghaoui and Lebret (1997) consider the least squares problem with unknown but bounded, non-random disturbance and solve it in polynomial time. Xu et al. (2010) study the robust linear regression problem with norm-bounded feature perturbation and show that it is equivalent to the ℓ_1-regularized regression. See Yang and Xu (2013); Bertsimas and Copenhaver (2017) which also use a deterministic robustness approach.

In this paper we consider a DRO problem with $\mathcal{P}$ containing distributions that are close to the discrete empirical distribution in the sense of Wasserstein distance. The reason for choosing the Wasserstein metric is two-fold. On one hand, the Wasserstein ambiguity set is rich enough to contain both continuous and discrete relevant distributions, while other metrics such as the KL divergence, exclude all continuous distributions if the nominal distribution is discrete (Esfahani and Kuhn, 2015; Gao and Kleywegt, 2016). Furthermore, considering distributions within a KL distance from the empirical, does not allow for probability mass outside the support of the empirical distribution. On the other hand, measure concentration results guarantee that the Wasserstein set contains the true data-generating distribution with high confidence for a sufficiently large sample size (Fournier and Guillin, 2015). Moreover, the Wasserstein metric takes into account the closeness between support points while other metrics such as the ϕ-divergence only consider the probabilities of these points. The image retrieval example in Gao and Kleywegt (2016) suggests that the probabilistic ambiguity set constructed based on the KL divergence prefers the pathological distribution to the true distribution, whereas the Wasserstein distance does not exhibit such a problem. The reason lies in that ϕ-divergence does not incorporate a notion of closeness between two points, which in the context of image retrieval represents the perceptual similarity in color.

Our DRO problem minimizes the worst-case absolute residual over a Wasserstein ball of distributions, and could be relaxed to the following form:

$$\inf_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon\|(-\boldsymbol{\beta}, 1)\|_*, \tag{1}$$

where ϵ is the radius of the Wasserstein ball, and $\|\cdot\|_*$ is the dual norm of the norm space where the Wasserstein metric is defined on. Formulation (1) incorporates a wide class of models whose specific form depends on the notion of transportation cost embedded in the Wasserstein metric (see Section 2). Although the Wasserstein DRO formulation

simply reduces to regularized regression models, we want to emphasize a few new insights brought by this methodology. First, the regularization term controls the conservativeness of the Wasserstein set, or the amount of ambiguity in the data, which differentiates itself from the heuristically added regularizers in traditional regression models that serve the purpose of preventing overfitting, error/variance reduction, or sparsity recovery. Second, the regularization term is determined by the dual norm of the regression coefficient, which controls the *growth rate* of the ℓ_1-loss function, and the radius of the Wasserstein set. This connection provides guidance on the selection of the regularization coefficient and may lead to significant computational savings compared to cross-validation. DRO essentially enables new and more accurate interpretations of the regularizer, and establishes its dependence on the *growth rate* of the loss, the underlying metric space and the reliability of the observed samples.

The connection between robustness and regularization has been established in several works. The earliest one may be credited to El Ghaoui and Lebret (1997), who show that minimizing the worst-case squared residual within a Frobenius norm-based perturbation set is equivalent to Tikhonov regularization. In more recent works, using properly selected uncertainty sets, Xu et al. (2010) have shown the equivalence between robust linear regression with feature perturbations and the *Least Absolute Shrinkage and Selection Operator (LASSO)*. Yang and Xu (2013) extend this to more general LASSO-like procedures, including versions of the grouped LASSO. Bertsimas and Copenhaver (2017) give a comprehensive characterization of the conditions under which robustification and regularization are equivalent for regression models with deterministic norm-bounded perturbations on the features. For classification problems, Xu et al. (2009) show the equivalence between the regularized Support Vector Machines (SVMs) and a robust optimization formulation, by allowing potentially correlated disturbances in the covariates. Shafieezadeh-Abadeh et al. (2015) consider a robust version of logistic regression under the assumption that the probability distributions under consideration lie in a Wasserstein ball, and they show that the regularized logistic regression is a special case of this robust formulation. Recently, Shafieezadeh-Abadeh et al. (2017); Gao et al. (2017) have provided a unified framework for connecting the Wasserstein DRO with regularized learning procedures, for various regression and classification models.

Our work is motivated by the problem of identifying patients who receive an abnormally high radiation exposure in CT exams, given the patient characteristics and exam-related variables (Chen et al., 2018). This could be casted as an outlier detection problem; specifically, estimating a robustified regression plane that is immunized against outliers and learns the underlying true relationship between radiation dose and the relevant predictors. We focus on robust learning of the parameter in regression models under distributional perturbations residing within a Wasserstein ball. While the applicability of the Wasserstein DRO methodology is not restricted to regression analysis (Sinha et al., 2017; Gao et al., 2017; Shafieezadeh-Abadeh et al., 2017), or a particular form of the loss function (as long as it satisfies certain smoothness conditions (Gao et al., 2017)), we focus on the absolute residual loss in linear regression in light of our motivating application and for the purpose of enhancing robustness. Our contributions can be summarized as follows:

1. We develop a DRO approach to robustify linear regression using an ℓ_1 loss function and an ambiguity set around the empirical distribution of the training samples defined based on the Wasserstein metric. The formulation is general enough to in-

clude any norm-induced Wasserstein metric and incorporate additional regularization constraints on the regression coefficients (e.g., ℓ_1-norm constraints). It provides an intuitive connection between the amount of ambiguity allowed and a regularization penalty term in the robust formulation, which provides a natural way to adjust the latter.

2. We establish novel performance guarantees on both the out-of-sample loss (prediction bias) and the discrepancy between the estimated and the true regression coefficients (estimation bias). Our guarantees elucidate the role of the regularizer, which is related to the dual norm of the regression coefficients, in bounding the biases and are in concert with the theoretical foundation that leads to the regularized problem. The generalization error bound, in particular, builds a connection between the loss function and the form of the regularizer via Rademacher complexity, providing a rigorous explanation for the commonly observed good out-of-sample performance of regularized regression. On the other hand, the estimation error bound corroborates the validity of the ℓ_1-loss function, which tends to incur a lower estimation bias than other candidates such as the ℓ_2 and ℓ_∞ losses. Our results are novel in the robust regression setting and different from earlier work in the DRO literature, enabling new perspectives and interpretations of the norm-based regularization, and providing justifications for the ℓ_1-loss-based learning algorithms.

3. We empirically explore three important aspects of the Wasserstein DRO formulation, including the advantages of the ℓ_1-loss function, the selection of a proper norm for the Wasserstein metric, and the implication of penalizing the *extended regression coefficient* $(-\boldsymbol{\beta}, 1)$, by comparing with a series of regression models on a number of synthetic datasets. We show the superiority of the Wasserstein DRO approach, presenting a thorough analysis under four different experimental setups. We also consider the application of our methodology to outlier detection and compare with M-estimation in terms of the ability of identifying outliers (*ROC (Receiver Operating Characteristic)* curves). The Wasserstein DRO formulation achieves significantly higher *AUC (Area Under Curve)* values.

The rest of the paper is organized as follows. In Section 2, we introduce the Wasserstein metric and derive the general Wasserstein DRO formulation in a linear regression framework. Section 3 establishes performance guarantees for both the general formulation and the special case where the Wasserstein metric is defined on the ℓ_1-norm space. Numerical experimental results are presented in Section 4. We conclude the paper in Section 5.

Notational conventions: We use boldfaced lowercase letters to denote vectors, ordinary lowercase letters to denote scalars, boldfaced uppercase letters to denote matrices, and calligraphic capital letters to denote sets. $\mathbb{E}$ denotes expectation and $\mathbb{P}$ probability of an event. All vectors are column vectors. For space saving reasons, we write $\mathbf{x} = (x_1, \ldots, x_{\dim(\mathbf{x})})$ to denote the column vector $\mathbf{x}$, where $\dim(\mathbf{x})$ is the dimension of $\mathbf{x}$. We use prime to denote the transpose of a vector, $\|\cdot\|$ for the general norm operator, $\|\cdot\|_2$ for the ℓ_2 norm, $\|\cdot\|_1$ for the ℓ_1 norm, and $\|\cdot\|_\infty$ for the infinity norm. $\mathcal{P}(\mathcal{Z})$ denotes the set of probability measures supported on $\mathcal{Z}$. $\mathbf{e}_i$ denotes the i-th unit vector, $\mathbf{e}$ the vector of ones, $\mathbf{0}$ a vector of zeros, and $\mathbf{I}$ the identity matrix. Given a norm $\|\cdot\|$ on $\mathbb{R}^m$, the dual

norm $\|\cdot\|_*$ is defined as: $\|\boldsymbol{\theta}\|_* \triangleq \sup_{\|\mathbf{z}\|\leq 1} \boldsymbol{\theta}'\mathbf{z}$. For a function $h(\mathbf{z})$, its convex conjugate $h^*(\cdot)$ is defined as: $h^*(\boldsymbol{\theta}) \triangleq \sup_{\mathbf{z}\in \text{dom } h} \{\boldsymbol{\theta}'\mathbf{z} - h(\mathbf{z})\}$, where dom h denotes the domain of the function h.

2. Problem Statement and Justification of Our Formulation

Consider a linear regression problem where we are given a predictor/feature vector $\mathbf{x} \in \mathbb{R}^{m-1}$, and a response variable $y \in \mathbb{R}$. Our goal is to obtain an accurate estimate of the regression plane that is robust with respect to the adversarial perturbations in the data. We consider an ℓ_1-loss function $h_{\boldsymbol{\beta}}(\mathbf{x}, y) \triangleq |y - \mathbf{x}'\boldsymbol{\beta}|$, motivated by the observation that the absolute loss function is more robust to large residuals than the squared loss (see Fig. 1). Moreover, the estimation error analysis presented in Section 3.2 suggests that the ℓ_1-loss function leads to a smaller estimation bias than others. Our Wasserstein DRO problem using the ℓ_1-loss function is formulated as:

$$\inf_{\boldsymbol{\beta}\in\mathcal{B}} \sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}\big[|y - \mathbf{x}'\boldsymbol{\beta}|\big], \tag{2}$$

where $\boldsymbol{\beta}$ is the regression coefficient vector that belongs to some set $\mathcal{B}$. $\mathcal{B}$ could be $\mathbb{R}^{m-1}$, or $\mathcal{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \leq l\}$ if we wish to induce sparsity, with l being some pre-specified number. $\mathbb{Q}$ is the probability distribution of $(\mathbf{x}, y)$, belonging to some set Ω which is defined as:

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : W_p(\mathbb{Q},\ \hat{\mathbb{P}}_N) \leq \epsilon\},$$

where $\mathcal{Z}$ is the set of possible values for $(\mathbf{x}, y)$; $\mathcal{P}(\mathcal{Z})$ is the space of all probability distributions supported on $\mathcal{Z}$; ϵ is a pre-specified radius of the Wasserstein ball; and $W_p(\mathbb{Q},\ \hat{\mathbb{P}}_N)$ is the order-p Wasserstein distance between $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$ (see definition in (3)), with $\hat{\mathbb{P}}_N$ the uniform empirical distribution over samples. The formulation in (2) is robust since it minimizes over the regression coefficients the worst case expected loss, that is, the expected loss maximized over all probability distributions in the ambiguity set Ω.

Before deriving a tractable reformulation for (2), let us first define the Wasserstein metric. Let $(\mathcal{Z}, s)$ be a metric space where $\mathcal{Z}$ is a set and s is a metric on $\mathcal{Z}$. The Wasserstein metric of order $p \geq 1$ defines the distance between two probability distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ in the following way:

$$W_p(\mathbb{Q}_1, \mathbb{Q}_2) \triangleq \left(\min_{\Pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \left\{ \int_{\mathcal{Z}\times\mathcal{Z}} \big(s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2))\big)^p \ \Pi\big(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2)\big) \right\} \right)^{1/p}, \tag{3}$$

where Π is the joint distribution of $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively. The Wasserstein distance between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ represents the cost of an optimal mass transportation plan, where the cost is measured through the metric s. The order p should be selected in such a way as to ensure that the worst-case expected loss is meaningfully defined, i.e.,

$$\mathbb{E}^{\mathbb{Q}}\big[h_{\boldsymbol{\beta}}(\mathbf{x}, y)\big] < \infty,\ \forall \mathbb{Q} \in \Omega. \tag{4}$$

Notice that the ambiguity set Ω is centered at the empirical distribution $\hat{\mathbb{P}}_N$ and has radius ϵ. It may be desirable to translate (4) into:

$$\left|\mathbb{E}^{\mathbb{Q}}\big[h_{\boldsymbol{\beta}}(\mathbf{x}, y)\big] - \mathbb{E}^{\hat{\mathbb{P}}_N}\big[h_{\boldsymbol{\beta}}(\mathbf{x}, y)\big]\right| < \infty,\ \forall \mathbb{Q} \in \Omega. \tag{5}$$

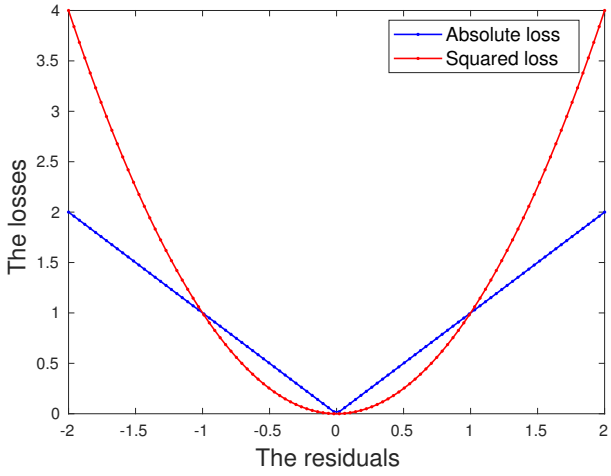


Figure 1: Comparison between ℓ_1 and ℓ_2 loss functions.

We want to relate (5) with the Wasserstein distance $W_p(\mathbb{Q},\ \hat{\mathbb{P}}_N)$, which is no larger than ϵ for all $\mathbb{Q} \in \Omega$. The LHS of (5) could be written as:

$$\begin{aligned}
&\left|\mathbb{E}^{\mathbb{Q}}\left[h_{\boldsymbol{\beta}}(\mathbf{x}, y)\right] - \mathbb{E}^{\hat{\mathbb{P}}_N}\left[h_{\boldsymbol{\beta}}(\mathbf{x}, y)\right]\right| \\
=&\left|\int_{\mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1)\mathbb{Q}(d(\mathbf{x}_1, y_1)) - \int_{\mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2)\hat{\mathbb{P}}_N(d(\mathbf{x}_2, y_2))\right| \\
=&\left|\int_{\mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1) \int_{\mathcal{Z}} \Pi_0(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2)) - \int_{\mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2) \int_{\mathcal{Z}} \Pi_0(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2))\right| \\
\leq &\int_{\mathcal{Z}\times\mathcal{Z}} \left|h_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1) - h_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2)\right| \Pi_0(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2)),
\end{aligned} \tag{6}$$

where Π_0 is the joint distribution of $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$ with marginals $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$, respectively. Comparing (6) with (3), we see that for (5) to hold, the following quantity which characterizes the *growth rate* of the loss function needs to be bounded:

$$\mathrm{GR}_{h_{\boldsymbol{\beta}}}((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)) \triangleq \frac{\left|h_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1) - h_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2)\right|}{\left(s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2))\right)^p},\ \forall (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \in \mathcal{Z}. \tag{7}$$

A formal definition of the growth rate is due to Gao and Kleywegt (2016), which takes the limit of (7) as $s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)) \to \infty$, to eliminate its dependence on $(\mathbf{x}, y)$. One important aspect they have pointed out is that when the growth rate of the loss function is infinite, strong duality for the worst-case problem $\sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}\left[h_{\boldsymbol{\beta}}(\mathbf{x}, y)\right]$ fails to hold, in which case the DRO problem (2) becomes intractable. Assuming that the metric s is

induced by some norm $\|\cdot\|$, the bounded *growth rate* requirement is expressed as follows:

$$\limsup_{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|\to\infty} \frac{|h_{\boldsymbol{\beta}}(\mathbf{x}_1,y_1)-h_{\boldsymbol{\beta}}(\mathbf{x}_2,y_2)|}{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|^p} \leq \limsup_{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|\to\infty} \frac{|y_1-\mathbf{x}_1'\boldsymbol{\beta}-(y_2-\mathbf{x}_2'\boldsymbol{\beta})|}{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|^p}$$
$$\leq \limsup_{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|\to\infty} \frac{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|\|(-\boldsymbol{\beta},1)\|_*}{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|^p} < \infty, \tag{8}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, and the second inequality is due to the Cauchy-Schwarz inequality. Notice that by taking $p=1$, (8) is equivalently translated into the condition that $\|(-\boldsymbol{\beta},1)\|_*<\infty$, which we will see in Section 3 is an essential requirement to guarantee a good generalization performance for the Wasserstein DRO estimator. The growth rate essentially reveals the underlying metric space used by the Wasserstein distance. Taking $p>1$ leads to zero growth rate in the limit of (8), which is not desirable since it removes the Wasserstein ball structure from our formulation and renders it an optimization problem over a singleton distribution. This will be made more clear in the following analysis. We thus choose the order-1 Wasserstein metric with s being induced by some norm $\|\cdot\|$ to define our DRO problem.

Next, we will discuss how to convert (2) into a tractable formulation. Suppose we have N independently and identically distributed realizations of $(\mathbf{x},y)$, denoted by $(\mathbf{x}_i,y_i), i=1,\ldots,N$. We make the assumption that $(\mathbf{x},y)$ comes from a mixture of two distributions, with probability q from the outlying distribution $\mathbb{P}_{out}$ and with probability $1-q$ from the true distribution $\mathbb{P}$. Recall that $\hat{\mathbb{P}}_N$ is the discrete uniform distribution over the N samples. Our goal is to generate estimators that are consistent with the true distribution $\mathbb{P}$. We claim that when q is small, if the Wasserstein ball radius ϵ is chosen judiciously, the true distribution $\mathbb{P}$ will be included in the set Ω while the outlying distribution $\mathbb{P}_{out}$ will be excluded. To see this, consider a simple example where $\mathbb{P}$ is a discrete distribution that assigns equal probability to 10 data points equally spaced between 0.1 and 1, and $\mathbb{P}_{out}$ assigns probability 0.5 to two data points 1 and 2. We generate 100 samples and plot the Wasserstein distances from $\hat{\mathbb{P}}_N$ for both $\mathbb{P}$ and $\mathbb{P}_{out}$. From Fig. 2 we observe that for q below 0.5, the true distribution $\mathbb{P}$ is closer to $\hat{\mathbb{P}}_N$ whereas the outlying distribution $\mathbb{P}_{out}$ is further away. If the radius ϵ is chosen between the red $(*-)$ and blue $(\circ-)$ lines, the Wasserstein ball that we are hedging against will exclude the outlying distribution and the resulting estimator will be robust to the adversarial perturbations. Moreover, as q becomes smaller, the gap between the red and blue lines becomes larger. One implication from this observation is that as the data becomes purer, the radius of the Wasserstein ball tends to be smaller, and the confidence in the observed samples is higher. For large q values, the DRO formulation seems to fail. However, as outliers are defined to be the data points that do not conform to the majority of data, we can safely claim that $\mathbb{P}_{out}$ is the distribution of the minority and q is always below 0.5.

We now consider the inner supremum in (2). Esfahani and Kuhn (2015, Theorem 6.3) show that when the set $\mathcal{Z}$ is closed and convex, and the loss function $h_{\boldsymbol{\beta}}(\mathbf{x},y)$ is convex in $(\mathbf{x},y)$,

$$\sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h_{\boldsymbol{\beta}}(\mathbf{x},y)] \leq \kappa\epsilon + \frac{1}{N}\sum_{i=1}^{N} h_{\boldsymbol{\beta}}(\mathbf{x}_i,y_i),\ \forall \epsilon \geq 0, \tag{9}$$

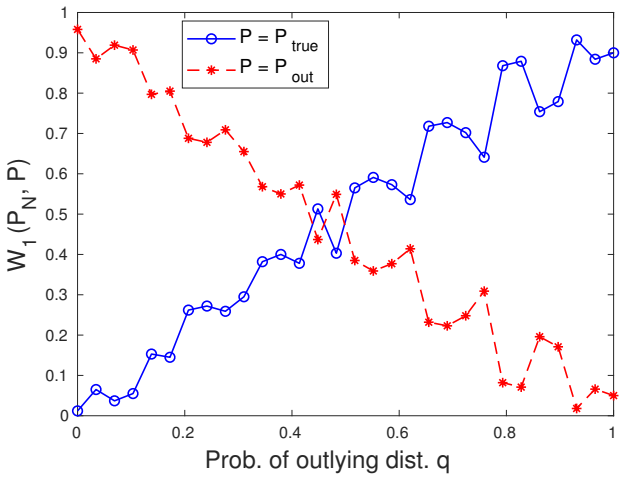


Figure 2: The order-1 Wasserstein distances from the empirical distribution.

where $\kappa(\boldsymbol{\beta}) = \sup\{\|\boldsymbol{\theta}\|_* : h^*_{\boldsymbol{\beta}}(\boldsymbol{\theta}) < \infty\}$, with $h^*_{\boldsymbol{\beta}}(\cdot)$ the convex conjugate function of $h_{\boldsymbol{\beta}}(\mathbf{x}, y)$. Through (9), we can relax problem (2) by minimizing the right hand side of (9) instead of the worst-case expected loss. Moreover, as shown in Esfahani and Kuhn (2015), (9) becomes an equality when $\mathcal{Z} = \mathbb{R}^m$. In Theorem 2.1, we compute the value of $\kappa(\boldsymbol{\beta})$ for the specific ℓ_1 loss function we use. The proof of this Theorem and all results hereafter are included in Appendix A.

Theorem 2.1 *Define $\kappa(\boldsymbol{\beta}) = \sup\{\|\boldsymbol{\theta}\|_* : h^*_{\boldsymbol{\beta}}(\boldsymbol{\theta}) < \infty\}$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, and $h^*_{\boldsymbol{\beta}}(\cdot)$ is the conjugate function of $h_{\boldsymbol{\beta}}(\cdot)$. When the loss function is $h_{\boldsymbol{\beta}}(\mathbf{x}, y) = |y - \mathbf{x}'\boldsymbol{\beta}|$, we have $\kappa(\boldsymbol{\beta}) = \|(-\boldsymbol{\beta}, 1)\|_*$.*

Due to Theorem 2.1, (2) could be formulated as the following optimization problem:

$$\inf_{\boldsymbol{\beta} \in \mathcal{B}} \epsilon \|(-\boldsymbol{\beta}, 1)\|_* + \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|. \tag{10}$$

Note that the regularization term of (10) is the product of the *growth rate* of the loss and the Wasserstein ball radius. The growth rate is closely related to the way the Wasserstein metric defines the transportation costs on the data $(\mathbf{x}, y)$. As mentioned earlier, a zero growth rate diminishes the effect of the Wasserstein distributional uncertainty set, and the resulting formulation would simply be an empirical loss minimization problem. The parameter ϵ controls the conservativeness of the formulation, whose selection depends on the sample size, the dimensionality of the data, and the confidence that the Wasserstein ball contains the true distribution (see eq. (8) in Esfahani and Kuhn, 2015). Roughly speaking, when the sample size is large enough, and for a fixed confidence level, ϵ is inversely proportional to $N^{1/m}$.

Formulation (10) incorporates a class of models whose specific form depends on the norm space we choose, which could be application-dependent and practically useful. For example, when the Wasserstein metric s is induced by $\|\cdot\|_2$ and the set $\mathcal{B}$ is the intersection of a

polyhedron with convex quadratic inequalities, (10) is a convex quadratic problem which can be solved to optimality very efficiently. Specifically, it could be converted to:

$$\begin{aligned}\min_{a,\ b_1,\ldots,b_N,\ \boldsymbol{\beta}} \quad & a\epsilon + \frac{1}{N}\sum_{i=1}^{N} b_i \\ \text{s.t.} \quad & \|\boldsymbol{\beta}\|_2^2 + 1 \leq a^2, \\ & y_i - \mathbf{x}_i'\boldsymbol{\beta} \leq b_i,\ i = 1,\ldots,N, \\ & -(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \leq b_i,\ i = 1,\ldots,N, \\ & a,\ b_i \geq 0,\ i = 1,\ldots,N, \\ & \boldsymbol{\beta} \in \mathcal{B}.\end{aligned} \tag{11}$$

When the Wasserstein metric is defined using $\|\cdot\|_1$ and the set $\mathcal{B}$ is a polyhedron, (10) is a linear programming problem:

$$\begin{aligned}\min_{a,\ b_1,\ldots,b_N,\ \boldsymbol{\beta}} \quad & a\epsilon + \frac{1}{N}\sum_{i=1}^{N} b_i \\ \text{s.t.} \quad & a \geq \boldsymbol{\beta}'\mathbf{e}_i,\ i = 1,\ldots,m-1, \\ & a \geq -\boldsymbol{\beta}'\mathbf{e}_i,\ i = 1,\ldots,m-1, \\ & y_i - \mathbf{x}_i'\boldsymbol{\beta} \leq b_i,\ i = 1,\ldots,N, \\ & -(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \leq b_i,\ i = 1,\ldots,N, \\ & a \geq 1, \\ & b_i \geq 0,\ i = 1,\ldots,N, \\ & \boldsymbol{\beta} \in \mathcal{B}.\end{aligned} \tag{12}$$

More generally, when the coordinates of $(\mathbf{x}, y)$ differ from each other substantially, a properly chosen, positive definite weight matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$ could scale correspondingly different coordinates of $(\mathbf{x}, y)$ by using the $\mathbf{M}$-weighted norm:

$$\|(\mathbf{x}, y)\|_{\mathbf{M}} = \sqrt{(\mathbf{x}, y)'\mathbf{M}(\mathbf{x}, y)}.$$

It can be shown that (10) in this case becomes:

$$\inf_{\boldsymbol{\beta}\in\mathcal{B}} \epsilon\sqrt{(-\boldsymbol{\beta}, 1)'\mathbf{M}^{-1}(-\boldsymbol{\beta}, 1)} + \frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|. \tag{13}$$

We note that this Wasserstein DRO framework could be applied to a broad class of loss functions and the tractable reformulations have been derived in Shafieezadeh-Abadeh et al. (2017); Gao et al. (2017) for regression and classification models. We adopt the absolute residual loss in this paper to enhance the robustness of the formulation, which is the focus of our work and serves the purpose of estimating robust parameters that are immunized against perturbations/outliers. Notice that (10) coincides with the regularized LAD models (Pollard, 1991; Wang et al., 2006), except that we are regularizing a variant of the regression coefficient. We would like to highlight several novel viewpoints that are

brought by the Wasserstein DRO framework and justify the value and novelty of (10). First, (10) is obtained as an outcome of a fundamental DRO formulation, which enables new interpretations of the regularizer from the standpoint of distributional robustness, and provides rigorous theoretical foundation on why the ℓ_2-regularizer prevents overfitting to the training data. The regularizer could be seen as a control over the amount of ambiguity in the data and reveals the reliability of the contaminated samples. Second, the geometry of the Wasserstein ball is embedded in the regularization term, which penalizes the regression coefficient on the dual Wasserstein space, with the magnitude of penalty being the radius of the ball. This offers an intuitive interpretation and provides guidance on how to set the regularization coefficient. Moreover, different from the traditional regularized LAD models that directly penalize the regression coefficient $\boldsymbol{\beta}$, we regularize the vector $(-\boldsymbol{\beta}, 1)$, where the 1 takes into account the transportation cost along the y direction. Penalizing only on $\boldsymbol{\beta}$ corresponds to an infinite transportation cost along y. Our model is more general in this sense, and establishes the connection between the metric space on data and the form of the regularizer.

3. Performance Guarantees

Having obtained a tractable reformulation for the Wasserstein DRO problem, we next establish guarantees on the predictive power and estimation quality for the solution to (10). Two types of results will be presented in this section, one of which bounds the prediction bias of the estimator on new, future data (given in Section 3.1). The other one that bounds the discrepancy between the estimated and true regression planes (estimation bias), is given in Section 3.2.

3.1 Out-of-Sample Performance

In this subsection we investigate generalization characteristics of the solution to (10), which involves measuring the error generated by our estimator on a new random sample $(\mathbf{x}, y)$. We would like to obtain estimates that not only explain the observed samples well, but, more importantly, possess strong generalization abilities. The derivation is mainly based on *Rademacher complexity* (see Bartlett and Mendelson, 2002), which is a measurement of the complexity of a class of functions. We would like to emphasize the applicability of such a proof technique to general loss functions, as long as their empirical Rademacher complexity could be bounded. The bound we derive for the prediction bias depends on both the sample average loss (the training error) and the dual norm of the regression coefficient (the regularizer), which corroborates the validity and necessity of our regularized formulation. Moreover, the generalization result also builds a connection between the loss function and the form of the regularizer via Rademacher complexity, which enables new insights into the regularization term and explains the commonly observed good out-of-sample performance of regularized regression in a rigorous way. We first make several mild assumptions that are needed for the generalization result.

Assumption A *The norm of the uncertainty parameter* $(\mathbf{x}, y)$ *is bounded above almost surely, i.e.,* $\|(\mathbf{x}, y)\| \leq R$.

Assumption B *The dual norm of* $(-\boldsymbol{\beta}, 1)$ *is bounded above within the feasible region, namely,*

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|(-\boldsymbol{\beta}, 1)\|_* = \bar{B}.$$

Under these two assumptions, the absolute loss could be bounded via the Cauchy-Schwarz inequality.

Lemma 3.1 *For every feasible* $\boldsymbol{\beta}$*, it follows*

$$|y - \mathbf{x}'\boldsymbol{\beta}| \leq \bar{B}R, \quad \textit{almost surely.}$$

With the above result, the idea is to bound the generalization error using the empirical *Rademacher complexity* of the following class of loss functions:

$$\mathcal{H} = \{(\mathbf{x}, y) \mapsto h_{\boldsymbol{\beta}}(\mathbf{x}, y) : h_{\boldsymbol{\beta}}(\mathbf{x}, y) = |y - \mathbf{x}'\boldsymbol{\beta}|,\ \boldsymbol{\beta} \in \mathcal{B}\}.$$

We need to show that the empirical Rademacher complexity of $\mathcal{H}$, denoted by $\mathcal{R}_N(\mathcal{H})$, is upper bounded. The following result, similar to Lemma 3 in Bertsimas et al. (2015), provides a bound that is inversely proportional to the square root of the sample size.

Lemma 3.2

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{2\bar{B}R}{\sqrt{N}}.$$

Let $\hat{\boldsymbol{\beta}}$ be an optimal solution to (10), obtained using the samples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$. Suppose we draw a new i.i.d. sample $(\mathbf{x}, y)$. In Theorem 3.3 we establish bounds on the error $|y - \mathbf{x}'\hat{\boldsymbol{\beta}}|$.

Theorem 3.3 *Under Assumptions A and B, for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$ *with respect to the sampling,*

$$\mathbb{E}[|y - \mathbf{x}'\hat{\boldsymbol{\beta}}|] \leq \frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}| + \frac{2\bar{B}R}{\sqrt{N}} + \bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}\ , \tag{14}$$

and for any $\zeta > \frac{2\bar{B}R}{\sqrt{N}} + \bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}$*,*

$$\mathbb{P}\left(|y - \mathbf{x}'\hat{\boldsymbol{\beta}}| \geq \frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}| + \zeta\right) \leq \frac{\frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}| + \frac{2\bar{B}R}{\sqrt{N}} + \bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}| + \zeta}. \tag{15}$$

There are two probability measures in the statement of Theorem 3.3. One is related to the new data $(\mathbf{x}, y)$, while the other is related to the samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$. The expectation in (14) (and the probability in (15)) is taken w.r.t. the new data $(\mathbf{x}, y)$. For a given set of samples, (14) (and (15)) holds with probability at least $1-\delta$ w.r.t. the measure of samples. Theorem 3.3 essentially says that given typical samples, the expected loss on new data using our Wasserstein DRO estimator could be bounded above by the average sample loss plus extra terms that depend on the supremum of $\|(-\boldsymbol{\beta}, 1)\|_*$ (our regularizer), and

are proportional to $1/\sqrt{N}$. This result validates the dual norm-based regularized regression from the perspective of generalization ability, and could be generalized to any bounded loss function. It also provides implications on the form of the regularizer. For example, if given an ℓ_2-loss function, the dependency on $\bar{B}$ for the generalization error bound will be of the form $\bar{B}^2$, which suggests using $\|(-\boldsymbol{\beta}, 1)\|_*^2$ as a regularizer, reducing to a variant of ridge regression (Hoerl and Kennard, 1970) for $\|\cdot\|_2$ induced Wasserstein metric.

We also note that the upper bounds in (14) and (15) do not depend on the dimension of $(\mathbf{x}, y)$. This dimensionality-free characteristic implies direct applicability of our Wasserstein approach to high-dimensional settings and is particularly useful in many real applications where, potentially, hundreds of features may be present. Theorem 3.3 also provides guidance on the number of samples that are needed to achieve satisfactory out-of-sample performance.

Corollary 3.4 *Suppose $\hat{\boldsymbol{\beta}}$ is the optimal solution to (10). For a fixed confidence level δ and some threshold parameter $\tau \geq 0$, to guarantee that the percentage difference between the expected absolute loss on new data and the sample average loss is less than τ, that is,*

$$\frac{\mathbb{E}[|y - \mathbf{x}'\hat{\boldsymbol{\beta}}|] - \frac{1}{N}\sum_{i=1}^{N}|y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}|}{\bar{B}R} \leq \tau,$$

the sample size N must satisfy

$$N \geq \left\lceil \frac{2(1 + \sqrt{2\log(2/\delta)}\ )}{\tau} \right\rceil^2. \tag{16}$$

Corollary 3.5 *Suppose $\hat{\boldsymbol{\beta}}$ is the optimal solution to (10). For a fixed confidence level δ, some $\tau \in (0, 1)$ and $\gamma \geq 0$, to guarantee that*

$$\mathbb{P}\left(\frac{|y - \mathbf{x}'\hat{\boldsymbol{\beta}}| - \frac{1}{N}\sum_{i=1}^{N}|y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}|}{\bar{B}R} \geq \gamma \right) \leq \tau,$$

the sample size N must satisfy

$$N \geq \left\lceil \frac{2(1 + \sqrt{2\log(2/\delta)}\ )}{\tau \cdot \gamma + \tau - 1} \right\rceil^2, \tag{17}$$

provided that $\tau \cdot \gamma + \tau - 1 > 0$.

In Corollaries 3.4 and 3.5, the sample size is inversely proportional to both δ and τ, which is reasonable since the more confident we want to be, the more samples we need. Moreover, the smaller τ is, the stricter a requirement we impose on the performance, and thus more samples are needed.

3.2 Discrepancy between Estimated and True Regression Planes

In addition to the generalization performance, we are also interested in the accuracy of the estimator. In this section we seek to bound the difference between the estimated and true regression coefficients, under a certain distributional assumption on $(\mathbf{x}, y)$. Throughout the

section we will use $\hat{\boldsymbol{\beta}}$ to denote the estimated regression coefficients, obtained as an optimal solution to (18), and $\boldsymbol{\beta}^*$ for the true (unknown) regression coefficients. The bound we will derive turns out to be related to the Gaussian width (see definition in the Appendix) of the unit ball in $\|\cdot\|_\infty$, the sub-Gaussian norm of the uncertainty parameter $(\mathbf{x}, y)$, as well as the geometric structure of the true regression coefficients. We note that this proof technique may be applied to several other loss functions, e.g., ℓ_2 and ℓ_∞ losses, with slight modifications. However, we will see that the ℓ_1-loss function incurs a relatively low estimation bias compared to others, further demonstrating the superiority of our absolute error minimization formulation.

To facilitate the analysis, we will use the following equivalent form of problem (10):

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|(-\boldsymbol{\beta}, 1)\|_* \\ \text{s.t.} \quad & \|(-\boldsymbol{\beta}, 1)'\mathbf{Z}\|_1 \leq \gamma_N, \\ & \boldsymbol{\beta} \in \mathcal{B}, \end{aligned} \tag{18}$$

where $\mathbf{Z} = [(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)]$ is the matrix with columns $(\mathbf{x}_i, y_i), i = 1, \ldots, N$, and γ_N is some exogenous parameter related to ϵ. One can show that for properly chosen γ_N, (18) produces the same solution with (10) (Bertsekas, 1999). (18) is similar to (11) in Chen and Banerjee (2016), with the difference lying in that we impose a constraint on the error instead of the gradient, and we consider a more general notion of norm on the coefficient. On the other hand, due to their similarity, we will follow the line of development in Chen and Banerjee (2016). Still, our analysis is self-contained and the bound we obtain is in a different form, which provides meaningful insights into our specific problem. We list below the assumptions that are needed to bound the estimation error.

Assumption C *The ℓ_2 norm of $(-\boldsymbol{\beta}, 1)$ is bounded above within the feasible region, namely,*

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} \|(-\boldsymbol{\beta}, 1)\|_2 = \bar{B}_2.$$

Assumption D (Restricted Eigenvalue Condition) *For some set $\mathcal{A}(\boldsymbol{\beta}^*) = cone\{\mathbf{v} | \|(-\boldsymbol{\beta}^*, 1) + \mathbf{v}\|_* \leq \|(-\boldsymbol{\beta}^*, 1)\|_*\} \cap \mathbb{S}^m$ and some positive scalar $\underline{\alpha}$, where $\mathbb{S}^m$ is the unit sphere in the m-dimensional Euclidean space,*

$$\inf_{\mathbf{v} \in \mathcal{A}(\boldsymbol{\beta}^*)} \mathbf{v}'\mathbf{Z}\mathbf{Z}'\mathbf{v} \geq \underline{\alpha},$$

where $\mathbb{S}^m$ denotes the unit sphere in the m-dimensional Euclidean space.

Assumption E *The true coefficient $\boldsymbol{\beta}^*$ is a feasible solution to (18), i.e.,*

$$\|\mathbf{Z}'(-\boldsymbol{\beta}^*, 1)\|_1 \leq \gamma_N, \quad \boldsymbol{\beta}^* \in \mathcal{B}.$$

Assumption F *$(\mathbf{x}, y)$ is a centered sub-Gaussian random vector (see definition in the Appendix), i.e., it has zero mean and satisfies the following condition:*

$$\|\|(\mathbf{x}, y)\|\|_{\psi_2} = \sup_{\mathbf{u} \in \mathbb{S}^m} \left\|\left|(\mathbf{x}, y)'\mathbf{u}\right|\right\|_{\psi_2} \leq \mu.$$

Assumption G *The covariance matrix of* $(\mathbf{x}, y)$ *has bounded positive eigenvalues. Set* $\mathbf{\Gamma} = \mathbb{E}[(\mathbf{x}, y)(\mathbf{x}, y)']$*; then,*

$$0 < \lambda_{min} \triangleq \lambda_{min}(\mathbf{\Gamma}) \leq \lambda_{max}(\mathbf{\Gamma}) \triangleq \lambda_{max} < \infty.$$

Notice that both $\underline{\alpha}$ in Assumption D and γ_N in Assumption E are related to the random observation matrix $\mathbf{Z}$. A probabilistic description for these two quantities will be provided later. We next present a preliminary result, similar to Lemma 2 in Chen and Banerjee (2016), that bounds the ℓ_2-norm of the estimation bias in terms of a quantity that is related to the geometric structure of the true coefficients. This result gives a rough idea on the factors that affect the estimation error, and shows the advantages of using the ℓ_1-loss from the perspective of its dual norm. The bound derived in Theorem 3.6 is crude in the sense that it is a function of several random parameters that are related to the random observation matrix $\mathbf{Z}$. This randomness will be described in a probabilistic way in the subsequent analysis.

Theorem 3.6 *Suppose the true regression coefficient vector is* $\boldsymbol{\beta}^*$ *and the solution to (18) is* $\hat{\boldsymbol{\beta}}$*. For the set* $\mathcal{A}(\boldsymbol{\beta}^*) = cone\{\mathbf{v}|\ \|(-\boldsymbol{\beta}^*, 1)+\mathbf{v}\|_* \leq \|(-\boldsymbol{\beta}^*, 1)\|_*\} \cap \mathbb{S}^m$*, under Assumptions A, D, and E, we have:*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{2R\gamma_N}{\underline{\alpha}} \Psi(\boldsymbol{\beta}^*), \tag{19}$$

where $\Psi(\boldsymbol{\beta}^*) = \sup_{\mathbf{v} \in \mathcal{A}(\boldsymbol{\beta}^*)} \|\mathbf{v}\|_*$.

Notice that the bound in (19) does not explicitly depend on the sample size N. If we change to the ℓ_2-loss function, problem (18) will become:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|(-\boldsymbol{\beta}, 1)\|_* \\ \text{s.t.} \quad & \|(-\boldsymbol{\beta}, 1)'\mathbf{Z}\|_2 \leq \gamma_N, \\ & \boldsymbol{\beta} \in \mathcal{B}. \end{aligned}$$

The proof of Theorem 3.6 still applies with slight modification. We will find out that in the case of ℓ_2-loss, the estimation error bound takes the following form:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{2R\sqrt{N}\gamma_N}{\underline{\alpha}} \Psi(\boldsymbol{\beta}^*).$$

Similarly, the ℓ_∞-loss, which considers only the maximum absolute loss among the samples, turns (18) into:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|(-\boldsymbol{\beta}, 1)\|_* \\ \text{s.t.} \quad & \|(-\boldsymbol{\beta}, 1)'\mathbf{Z}\|_\infty \leq \gamma_N, \\ & \boldsymbol{\beta} \in \mathcal{B}. \end{aligned}$$

The corresponding bound becomes:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{2RN\gamma_N}{\underline{\alpha}} \Psi(\boldsymbol{\beta}^*).$$

We see that by using either ℓ_2 or ℓ_∞-loss, an explicit dependency on N is introduced. As a result, the estimation error bounds become worse. The reason is that for the ℓ_1-loss function, its dual norm operator only picks out the maximum absolute coordinate and thus avoids the dependence on the dimension, which in our case is the sample size (see Eq.(28)), whereas other norms, e.g., ℓ_2-norm, sum over all the coordinates and thus introduce a dependence on N.

As mentioned earlier, (19) provides a random upper bound, revealed in $\underline{\alpha}$ and γ_N, that depends on the randomness in $\mathbf{Z}$. We therefore would like to replace these two parameters by non-random quantities. The $\underline{\alpha}$ acts as the minimum eigenvalue of the matrix $\mathbf{ZZ}'$ restricted to a subspace of $\mathbb{R}^m$, and thus a proper substitute should be related to the minimum eigenvalue of the covariance matrix of $(\mathbf{x}, y)$, i.e., the $\mathbf{\Gamma}$ matrix (cf. Assumption G), given that $(\mathbf{x}, y)$ is zero mean. See Lemmas 3.7, 3.8 and 3.9 for the derivation.

Lemma 3.7 *Consider the set* $\mathcal{A}_{\mathbf{\Gamma}} = \{\mathbf{w} \in \mathbb{S}^m | \mathbf{\Gamma}^{-1/2}\mathbf{w} \in cone(\mathcal{A}(\boldsymbol{\beta}^*))\}$, *where* $\mathcal{A}(\boldsymbol{\beta}^*)$ *is defined as in Theorem 3.6, and* $\mathbf{\Gamma} = \mathbb{E}[(\mathbf{x}, y)(\mathbf{x}, y)']$. *Under Assumptions F and G, when the sample size* $N \geq C_1\bar{\mu}^4(w(\mathcal{A}_{\mathbf{\Gamma}}))^2$, *where* $\bar{\mu} = \mu\sqrt{\frac{1}{\lambda_{min}}}$, *and* $w(\mathcal{A}_{\mathbf{\Gamma}})$ *is the Gaussian width of* $\mathcal{A}_{\mathbf{\Gamma}}$, *with probability at least* $1 - \exp(-C_2 N/\bar{\mu}^4)$, *we have*

$$\mathbf{v}'\mathbf{ZZ}'\mathbf{v} \geq \frac{N}{2}\mathbf{v}'\mathbf{\Gamma}\mathbf{v}, \quad \forall\ \mathbf{v} \in \mathcal{A}(\boldsymbol{\beta}^*),$$

where C_1 *and* C_2 *are positive constants.*

Note that the sample size requirement stated in Lemma 3.7 depends on the Gaussian width of $\mathcal{A}_{\mathbf{\Gamma}}$, where $\mathcal{A}_{\mathbf{\Gamma}}$ relates to $\mathcal{A}(\boldsymbol{\beta}^*)$. The following lemma shows that their Gaussian widths are also related. This relation is built upon the square root of the eigenvalues of $\mathbf{\Gamma}$, which measures the extent to which $\mathcal{A}_{\mathbf{\Gamma}}$ expands $\mathcal{A}(\boldsymbol{\beta}^*)$.

Lemma 3.8 (Lemma 4 in Chen and Banerjee (2016)) *Let* μ_0 *be the* ψ_2*-norm of a standard Gaussian random vector* $\mathbf{g} \in \mathbb{R}^m$, *and* $\mathcal{A}_{\mathbf{\Gamma}}$, $\mathcal{A}(\boldsymbol{\beta}^*)$ *be defined as in Lemma 3.7. Then, under Assumption G,*

$$w(\mathcal{A}_{\mathbf{\Gamma}}) \leq C_3\mu_0\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}\Big(w(\mathcal{A}(\boldsymbol{\beta}^*)) + 3\Big),$$

for some positive constant C_3.

Combining Lemmas 3.7 and 3.8, and expressing the covariance matrix $\mathbf{\Gamma}$ using its eigenvalues, we arrive at the following result.

Corollary 3.9 *Under Assumptions F and G, and the conditions in Lemmas 3.7 and 3.8, when* $N \geq \bar{C}_1\bar{\mu}^4\mu_0^2 \cdot \frac{\lambda_{max}}{\lambda_{min}}\Big(w(\mathcal{A}(\boldsymbol{\beta}^*)) + 3\Big)^2$, *with probability at least* $1 - \exp(-C_2 N/\bar{\mu}^4)$,

$$\mathbf{v}'\mathbf{ZZ}'\mathbf{v} \geq \frac{N\lambda_{min}}{2}, \quad \forall\ \mathbf{v} \in \mathcal{A}(\boldsymbol{\beta}^*),$$

where $\bar{C}_1$ *and* C_2 *are positive constants.*

Next, we derive the smallest possible value of γ_N such that $\boldsymbol{\beta}^*$ is feasible. The derivation uses the dual norm operator of the ℓ_1-loss, resulting in a bound that depends on the Gaussian width of the unit ball in the dual norm space ($\|\cdot\|_\infty$). See Lemma 3.10 for details.

Lemma 3.10 *Under Assumptions C and F, for any feasible $\boldsymbol{\beta}$, with probability at least $1 - C_4 \exp(-\frac{C_5^2 (w(\mathcal{B}_u))^2}{4\rho^2})$,*

$$\|(-\boldsymbol{\beta}, 1)'\mathbf{Z}\|_1 \leq C\mu\bar{B}_2 w(\mathcal{B}_u),$$

where $\mathcal{B}_u$ is the unit ball of norm $\|\cdot\|_\infty$, $\rho = \sup_{\mathbf{v}\in\mathcal{B}_u} \|\mathbf{v}\|_2$, and C_4, C_5, C positive constants.

We note that for other loss functions, e.g., the ℓ_2 and ℓ_∞ losses, similar results can be obtained, where $\mathcal{B}_u$ is defined to be the unit $\|\cdot\|_*^{\text{loss}}$-ball in $\mathbb{R}^m$, with $\|\cdot\|_*^{\text{loss}}$ being the dual norm of the loss. Combining Theorem 3.6, Corollary 3.9 and Lemma 3.10, we have the following main performance guarantee result that bounds the estimation bias of the solution to (18).

Theorem 3.11 *Under Assumptions A, C, D, E, F, G, and the conditions of Theorem 3.6, Corollary 3.9 and Lemma 3.10, when $N \geq \bar{C}_1 \bar{\mu}^4 \mu_0^2 \cdot \frac{\lambda_{max}}{\lambda_{min}} \Big(w(\mathcal{A}(\boldsymbol{\beta}^*)) + 3 \Big)^2$, with probability at least $1 - \exp(-C_2 N/\bar{\mu}^4) - C_4 \exp(-C_5^2 (w(\mathcal{B}_u))^2/(4\rho^2))$,*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{\bar{C} R \bar{B}_2 \mu}{N \lambda_{min}} w(\mathcal{B}_u) \Psi(\boldsymbol{\beta}^*). \tag{20}$$

From (20) we see that the bias is decreased as the sample size increases and the uncertainty embedded in $(\mathbf{x}, y)$ (revealed in R and μ) is reduced. The estimation error bound depends on the geometric structure of the true coefficients, defined using the dual norm space of the Wasserstein metric, the Gaussian width of the unit $\|\cdot\|_*^{\text{loss}}$-ball in $\mathbb{R}^m$, and the minimum eigenvalue of the covariance matrix of $(\mathbf{x}, y)$, with a convergence rate $1/N$ for the ℓ_1-loss we applied. As mentioned earlier, other loss functions may incur a dependence on N in the numerator of the bound, thus resulting in a slower convergence rate, which substantiates the benefit of using an ℓ_1-loss function.

4. Simulation Experiments on Synthetic Datasets

In this section we will explore the robustness of the Wasserstein formulation in terms of its *Absolute Deviation (AD)* loss function and the dual norm regularizer on the *extended regression coefficient* $(-\boldsymbol{\beta}, 1)$. Recall that our Wasserstein formulation is in the following form:

$$\inf_{\boldsymbol{\beta}\in\mathcal{B}} \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon \|(-\boldsymbol{\beta}, 1)\|_*. \tag{21}$$

We will focus on the following three aspects of this formulation:

1. How to choose a proper norm $\|\cdot\|$ for the Wasserstein metric?
2. Why do we penalize the extended regression coefficient $(-\boldsymbol{\beta}, 1)$ rather than $\boldsymbol{\beta}$?
3. What is the advantage of the AD loss compared to the *Squared Residuals (SR)* loss?

To answer Question 1, we will connect the choice of $\|\cdot\|$ for the Wasserstein metric with the characteristics/structures of the data $(\mathbf{x}, y)$. Specifically, we will design two sets of experiments, one with a dense regression coefficient $\boldsymbol{\beta}^*$, where all coordinates of $\mathbf{x}$ play a role in determining the value of the response y, and another with a sparse $\boldsymbol{\beta}^*$ implying that only a few predictors are relevant/important in predicting y. Two Wasserstein formulations will be tested and compared, one induced by the $\|\cdot\|_2$ (Wasserstein ℓ_2), which leads to an ℓ_2-regularizer in (21), and the other one induced by the $\|\cdot\|_\infty$ (Wasserstein ℓ_∞) and resulting in an ℓ_1-regularizer in (21). Intuitively, and based on the past experience in implementing the regularization techniques, the Wasserstein ℓ_2 should outperform the Wasserstein ℓ_∞ in the dense setting, while in the sparse setting, the reverse is true. Researchers have well identified the sparsity inducing property of the ℓ_1-regularizer and provided a nice geometrical interpretation for it (Friedman et al., 2001). Here, we try to offer a different explanation from the perspective of the Wasserstein DRO formulation, through projecting the sparsity of $\boldsymbol{\beta}^*$ onto the $(\mathbf{x}, y)$ space and establishing a *sparse* distance metric that only extracts a subset of coordinates from $(\mathbf{x}, y)$ to measure the closeness between samples.

For the second question, we first note that if the Wasserstein metric is induced by the following metric s_c:

$$s_c(\mathbf{x}, y) = \|(\mathbf{x}, cy)\|_2,$$

for a positive constant c, then as $c \to \infty$, the resulting Wasserstein DRO formulation becomes:

$$\inf_{\boldsymbol{\beta}\in\mathcal{B}} \frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon\|\boldsymbol{\beta}\|_2,$$

which is the ℓ_2-regularized LAD. This can be proved by recognizing that $s_c(\mathbf{x}, y) = \|(\mathbf{x}, y)\|_{\mathbf{M}}$, with $\mathbf{M} \in \mathbb{R}^{m\times m}$ a diagonal matrix whose diagonal elements are $(1, \ldots, 1, c^2)$, and then applying (13). Alternatively, if we let

$$s_c(\mathbf{x}, y) = \|(\mathbf{x}, cy)\|_\infty,$$

it can be shown that as $c \to \infty$, the corresponding Wasserstein formulation becomes:

$$\inf_{\boldsymbol{\beta}\in\mathcal{B}} \frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon\|\boldsymbol{\beta}\|_1,$$

which is the ℓ_1-regularized LAD (see proof in the Appendix). It follows that regularizing over $\boldsymbol{\beta}$ implies an infinite transportation cost along y. In other words, for two data points $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$, if $y_1 \neq y_2$, then they are considered to be infinitely far away. By contrast, our Wasserstein formulation, which regularizes over the extended regression coefficient $(-\boldsymbol{\beta}, 1)$, stems from a finite cost along y that is equally weighted with $\mathbf{x}$. We will see the disadvantages of penalizing only $\boldsymbol{\beta}$ in the analysis of the experimental results.

To answer Question 3, we will compare against several commonly used regression models that employ the SR loss function, e.g., ridge regression (Hoerl and Kennard, 1970), LASSO (Tibshirani, 1996), and *Elastic Net (EN)* (Zou and Hastie, 2005). We will also compare against M-estimation (Huber, 1964, 1973), which uses a variant of the SR loss and is equivalent to solving a weighted least squares problem, where the weights are determined

by the residuals. These models will be compared under two different experimental setups, one involving adversarial perturbations in both $\mathbf{x}$ and y, and the other with perturbations only in $\mathbf{x}$. The purpose is to investigate the behavior of these approaches when the noise in y is substantially reduced. As shown by Fig. 1, compared to the SR loss, the AD loss is less vulnerable to large residuals, and hence, it is advantageous in the scenarios where large perturbations appear in y. We are interested in studying whether its performance is consistently good when the corruptions appear mainly in $\mathbf{x}$.

We next describe the data generation process. Each training sample has a probability q of being drawn from the outlying distribution, and a probability $1-q$ of being drawn from the true (clean) distribution. Given the true regression coefficient $\boldsymbol{\beta}^*$, we generate the training data as follows:

- Generate a uniform random variable on $[0,1]$. If it is no larger than $1-q$, generate a clean sample as follows:

  1. Draw the predictor $\mathbf{x} \in \mathbb{R}^{m-1}$ from the normal distribution $N_{m-1}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{x}$, which is just the top left block of the matrix $\boldsymbol{\Gamma}$ in Assumption G. Specifically, $\boldsymbol{\Gamma} = \mathbb{E}[(\mathbf{x}, y)(\mathbf{x}, y)']$ is equal to

  $$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{\beta}^* \\ (\boldsymbol{\beta}^*)'\boldsymbol{\Sigma} & (\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^* + \sigma^2 \end{pmatrix},$$

  with σ^2 being the variance of the noise term. In our implementation, $\boldsymbol{\Sigma}$ has diagonal elements equal to 1 (unit variance) and off-diagonal elements equal to ρ, with ρ the correlation between predictors.

  2. Draw the response variable y from $N(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2)$.

- Otherwise, depending on the experimental setup, generate an outlier that is either:

  – Abnormal in both $\mathbf{x}$ and y, with outlying distribution:

    1. $\mathbf{x} \sim N_{m-1}(\mathbf{0}, \boldsymbol{\Sigma}) + N_{m-1}(5\mathbf{e}, \mathbf{I})$, or $\mathbf{x} \sim N_{m-1}(\mathbf{0}, \boldsymbol{\Sigma}) + N_{m-1}(\mathbf{0}, 0.25\mathbf{I})$;
    2. $y \sim N(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2) + 5\sigma$.

  – Abnormal only in $\mathbf{x}$:

    1. $\mathbf{x} \sim N_{m-1}(\mathbf{0}, \boldsymbol{\Sigma}) + N_{m-1}(5\mathbf{e}, \mathbf{I})$;
    2. $y \sim N(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2)$.

- Repeat the above procedure for N times, where N is the size of the training set.

To test the generalization ability of various formulations, we generate a test dataset containing M samples from the clean distribution. It is worth noting that only clean samples are included in the test set, since we only care about the prediction accuracy on clean data points, and our estimator is supposed to be consistent with the clean distribution and stay away from the outlying one. We are interested in studying the performance of various methods as the following factors are varied:

- *Signal to Noise Ratio (SNR)*, defined as:

$$\text{SNR} = \frac{(\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^*}{\sigma^2},$$

  which is equally spaced between 0.05 and 2 on a log scale.

- The correlation between predictors: ρ, which takes values in $(0.1, 0.2, \ldots, 0.9)$.

The performance metrics we use include:

- *Mean Squared Error (MSE)* on the test dataset, which is defined to be $\sum_{i=1}^{M}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2/M$, with $\hat{\boldsymbol{\beta}}$ being the estimate of $\boldsymbol{\beta}^*$ obtained from the training set, and $(\mathbf{x}_i, y_i),\ i = 1, \ldots, M$, being the observations from the test dataset;

- *Relative Risk (RR)* of $\hat{\boldsymbol{\beta}}$ defined as:

$$\text{RR}(\hat{\boldsymbol{\beta}}) \triangleq \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{(\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^*}.$$

- *Relative Test Error (RTE)* of $\hat{\boldsymbol{\beta}}$ defined as:

$$\text{RTE}(\hat{\boldsymbol{\beta}}) \triangleq \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sigma^2}{\sigma^2}.$$

- *Proportion of Variance Explained (PVE)* of $\hat{\boldsymbol{\beta}}$ defined as:

$$\text{PVE}(\hat{\boldsymbol{\beta}}) \triangleq 1 - \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sigma^2}{(\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^* + \sigma^2}.$$

For the metrics that evaluate the accuracy of the estimator, i.e., the RR, RTE and PVE, we list below two types of scores, one achieved by the best possible estimator $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, called the perfect score, and the other one achieved by the null estimator $\hat{\boldsymbol{\beta}} = 0$, called the null score.

- RR: a perfect score is 0 and the null score is 1.

- RTE: a perfect score is 1 and the null score is SNR+1.

- PVE: a perfect score is $\frac{\text{SNR}}{\text{SNR}+1}$, and the null score is 0.

During the training process, all the regularization parameters are tuned on a separate validation dataset. Specifically, we divide all the N training samples into two sets, dataset 1 and dataset 2 (validation set). For a pre-specified range of values for the penalty parameters, dataset 1 is used to train the models and derive $\hat{\boldsymbol{\beta}}$, and the performance of $\hat{\boldsymbol{\beta}}$ is evaluated on dataset 2. We choose the regularization parameter that yields the minimum *Median Absolute Deviation (MAD)* on the validation set. Using MAD as a selection criterion serves to hedge against the potentially large noise in the validation samples. As to the range of values for the tuned parameters, we borrow ideas from Hastie et al. (2017), where the

LASSO was tuned over 50 values ranging from $\lambda_m = \|\mathbf{X}'\mathbf{y}\|_\infty$ to a small fraction of λ_m on a log scale, with $\mathbf{X} \in \mathbb{R}^{N\times(m-1)}$ the design matrix whose i-th row is $\mathbf{x}_i'$, and $\mathbf{y} = (y_1, \ldots, y_N)$ the response vector. In our experiments, this range is properly adjusted for procedures that use the AD loss. Specifically, for Wasserstein ℓ_2 and ℓ_∞, ℓ_1- and ℓ_2-regularized LAD, the range of values for the regularization parameter is:

$$\sqrt{\exp\Big(\text{lin}\Big(\log(0.005 * \|\mathbf{X}'\mathbf{y}\|_\infty), \log(\|\mathbf{X}'\mathbf{y}\|_\infty), 50\Big)\Big)},$$

where $\text{lin}(a, b, n)$ is a function that takes in scalars a, b and n (integer) and outputs a set of n values equally spaced between a and b; the exp function is applied elementwise to a vector. The square root operator is in consideration of the AD loss that is the square root of the SR loss if evaluated on a single sample.

The regularization coefficient ϵ in formulation (10), which is the radius of the Wasserstein ball, allows for a more efficient tuning procedure. It has been noted in Esfahani and Kuhn (2015) that for a large enough sample size, ϵ is inversely proportional to $N^{1/m}$. This proportionality could be used as a guidance on setting ϵ, where only the proportional factor needs to be tuned (using cross-validation or a separate validation dataset as described earlier). In our implementation, given the small size of the simulated datasets, we will still adopt the validation dataset approach to tune the regularization parameter.

4.1 Dense $\boldsymbol{\beta}^*$, outliers in both $\mathbf{x}$ and y

In this subsection, we choose a dense regression coefficient $\boldsymbol{\beta}^*$, set the intercept $\beta_0^* = 0.3$, and the coefficient for each predictor x_i to be $\beta_i^* = 0.5, i = 1, \ldots, 20$. The adversarial perturbations are present in both $\mathbf{x}$ and y. Specifically, the outlying distribution is described by:

1. $\mathbf{x} \sim N_{m-1}(\mathbf{0}, \boldsymbol{\Sigma}) + N_{m-1}(5\mathbf{e}, \mathbf{I})$;
2. $y \sim N(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2) + 5\sigma$.

We generate 10 datasets consisting of $N = 100, M = 60$ observations. The probability of a training sample being drawn from the outlying distribution is $q = 30\%$. The mean values of the performance metrics (averaged over the 10 datasets), as we vary the SNR and the correlation between predictors, are shown in Figs. 3 and 4. Note that when SNR is varied, the correlation between predictors is set to 0.8 times a random noise uniformly distributed on the interval $[0.2, 0.4]$. When the correlation ρ is varied, the SNR is fixed to 0.5.

It can be seen that as the SNR decreases or the correlation between the predictors increases, the estimation problem becomes harder, and the performance of all approaches gets worse. In general the Wasserstein ℓ_2 achieves the best performance in terms of all four metrics. Specifically,

- It is better than the ℓ_2-regularized LAD, which assumes an infinite transportation cost along y.
- It is better than the Wasserstein ℓ_∞ and ℓ_1-regularized LAD which use the ℓ_1-regularizer.

- It is better than the approaches that use the SR loss function.

Empirically we have found out that in most cases, the approaches that use the AD loss, including the ℓ_1- and ℓ_2-regularized LAD, and the Wasserstein ℓ_∞ formulation, drive all the coordinates of $\boldsymbol{\beta}$ to zero, due to the relatively small magnitude of the AD loss compared to the norm of the coefficient, so that the regularizer dominates the solution. The approaches that use the SR loss, e.g., ridge regression and EN, do not exhibit such a problem, since the squared residuals weaken the dominance of the regularization term.

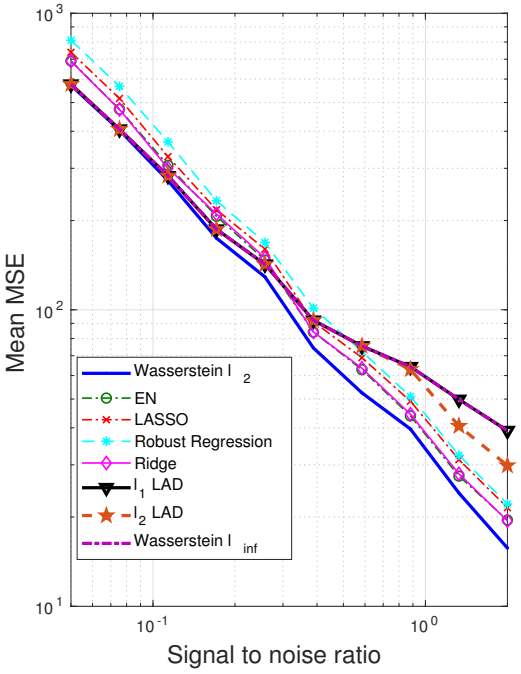
Overall the ℓ_2-regularizer outperforms the ℓ_1-regularizer, since the true regression coefficient is dense, which implies that a proper distance metric on the $(\mathbf{x}, y)$ space should take into account all the coordinates. From the perspective of the Wasserstein DRO framework, the ℓ_1-regularizer corresponds to an $\|\cdot\|_\infty$-based distance metric on the $(\mathbf{x}, y)$ space that only picks out the most influential coordinate to determine the closeness between data points, which in our case is not reasonable since every coordinate plays a role (reflected in the dense $\boldsymbol{\beta}^*$). In contrast, if $\boldsymbol{\beta}^*$ is sparse, using the $\|\cdot\|_\infty$ as a distance metric on $(\mathbf{x}, y)$ is more appropriate. A more detailed discussion of this will be presented in Sections 4.3 and 4.4.

4.2 Dense $\boldsymbol{\beta}^*$, outliers only in x

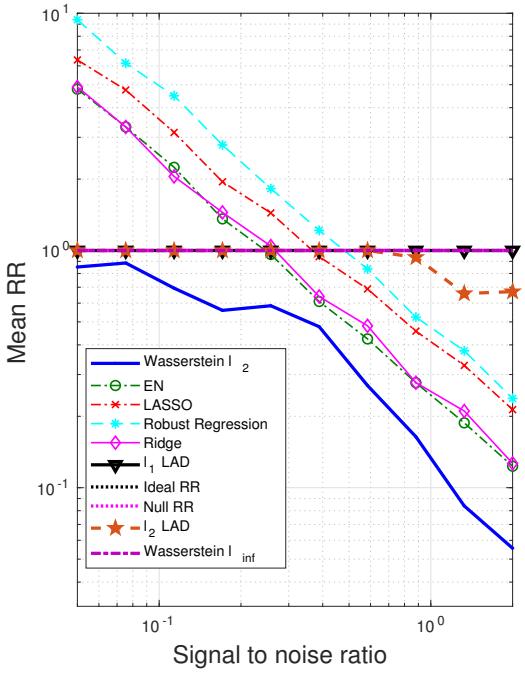
In this subsection we will experiment with the same $\boldsymbol{\beta}^*$ as in Section 4.1, but with perturbations only in $\mathbf{x}$, i.e., for a given $\mathbf{x}$ of the outlier, the corresponding y value is drawn in the same way as the clean samples. Our goal is to investigate the performance of the Wasserstein formulation when the response y is not subjected to large perturbations. The motivation for introducing the AD loss in the Wasserstein formulation is to hedge against large residuals, as illustrated in Fig. 1. We are interested in comparing the AD and SR loss functions when the residuals have moderate magnitudes.

Interestingly, we have observed that although the ℓ_1- and ℓ_2-regularized LAD, as well as the Wasserstein ℓ_∞ formulation, exhibit unsatisfactory performance, the Wasserstein ℓ_2, which shares the same loss function with them, is able to achieve a comparable performance with the best among all – EN and ridge regression (see Figs. 5 and 6). Notably, the ℓ_2-regularized LAD, which is just slightly different from our Wasserstein ℓ_2 formulation, shows a much worse performance. This is because the ℓ_2-regularized LAD implicitly assumes an infinite transportation cost along y, which gives zero tolerance to the variation in the response. For example, given two data points $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$, as long as $y_1 \neq y_2$, the distance between them is infinity. Therefore, a reasonable amount of fluctuation, caused by the intrinsic randomness of y, would be overly exaggerated by the underlying metric used by the ℓ_2-regularized LAD. In contrast, our Wasserstein approach uses a proper notion of norm to evaluate the distance in the $(\mathbf{x}, y)$ space and is able to effectively distinguish abnormally high variations from moderate, acceptable noise.

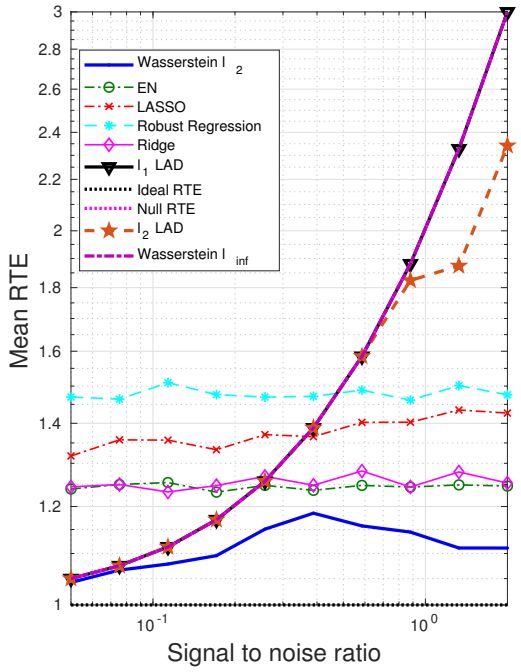
It is also worth noting that the formulations with the AD loss, e.g., ℓ_2- and ℓ_1-regularized LAD, and the Wasserstein ℓ_∞, perform worse than the approaches with the SR loss. One reasonable explanation is that the AD loss, introduced primarily for hedging against large perturbations in y, is less useful when the noise in y is moderate, in which case the sensitivity to response noise is needed. Although the AD loss is not a wise choice, penalizing



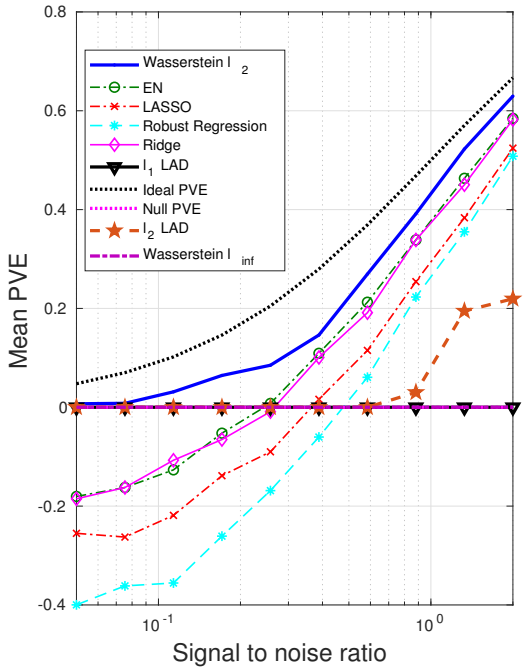
(a) Mean Squared Error.



(b) Relative risk.

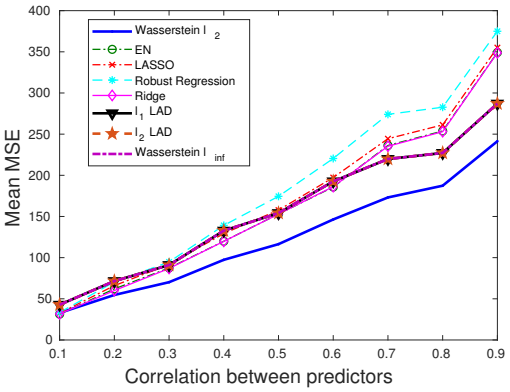


(c) Relative test error.

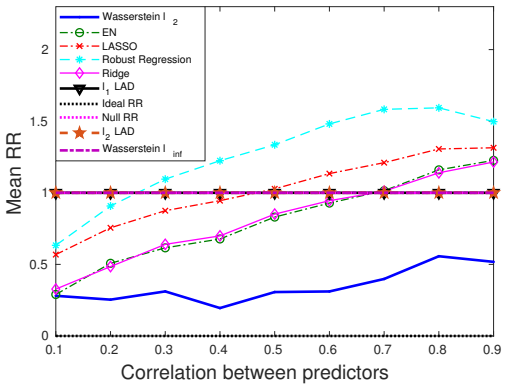


(d) Proportion of variance explained.

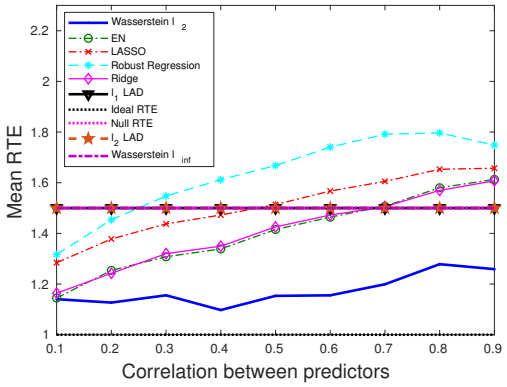
Figure 3: The impact of SNR on the performance metrics: dense $\boldsymbol{\beta}^*$, outliers in both $\mathbf{x}$ and y.



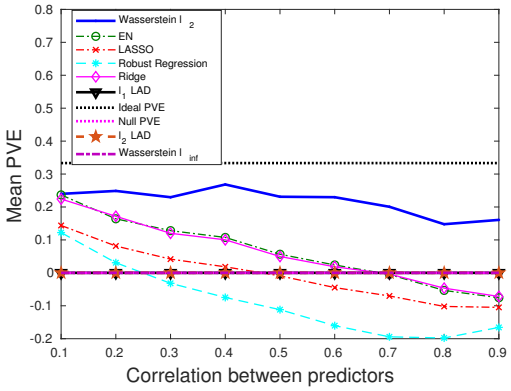
(a) Mean Squared Error.



(b) Relative risk.



(c) Relative test error.



(d) Proportion of variance explained.

Figure 4: The impact of predictor correlation on the performance metrics: dense $\boldsymbol{\beta}^*$, outliers in both $\mathbf{x}$ and y.

the extended coefficient vector $(-\boldsymbol{\beta}, 1)$ seems to make up, making the Wasserstein ℓ_2 a competitive method even when the perturbations appear only in $\mathbf{x}$.

4.3 Sparse $\boldsymbol{\beta}^*$, outliers in both x and y

In this subsection we will experiment with a sparse $\boldsymbol{\beta}^*$. The intercept is set to $\beta_0^* = 3$, and the coefficients for the 20 predictors are set to $\boldsymbol{\beta}^* = (0.05, 0, 0.006, 0, -0.007, 0, 0.008, 0, \ldots, 0)$. The adversarial perturbations are present in both $\mathbf{x}$ and y. Specifically, the distribution of outliers is characterized by:

1. $\mathbf{x} \sim N_{m-1}(\mathbf{0}, \boldsymbol{\Sigma}) + N_{m-1}(\mathbf{0}, 0.25\mathbf{I})$;
2. $y \sim N(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2) + 5\sigma$.

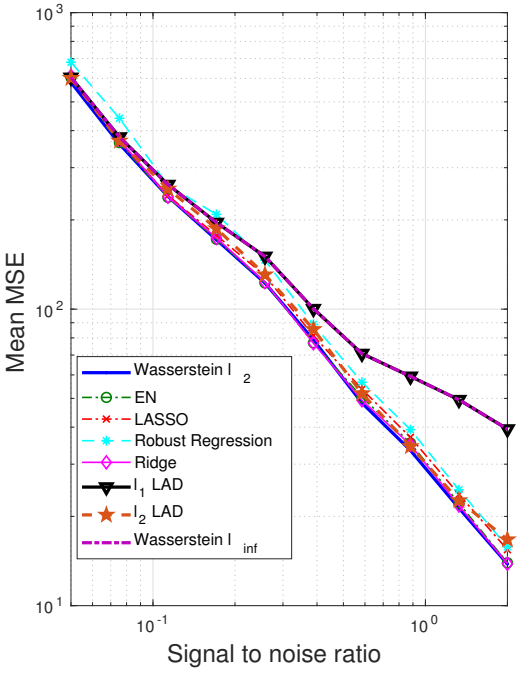
Our goal is to study the impact of the sparsity of $\boldsymbol{\beta}^*$ on the choice of the norm space for the Wasserstein metric. We know that the ℓ_1-regularizer works better than the ℓ_2-regularizer for sparse data, which has been validated by our results in Figs. 7 and 8. We will see that the Wasserstein ℓ_∞ formulation significantly outperforms the Wasserstein ℓ_2. An intuitively appealing interpretation for the sparsity inducing property of the ℓ_1-regularizer is made available by the Wasserstein DRO framework, which we explain as follows. The sparse regression coefficient $\boldsymbol{\beta}^*$ implies that only a few predictors are relevant to the regression model, and thus when measuring the distance in the $(\mathbf{x}, y)$ space, we need a metric that only extracts the subset of relevant predictors. The $\|\cdot\|_\infty$, which takes only the most influential coordinate of its argument, roughly serves this purpose. Compared to the $\|\cdot\|_2$ which takes into account all the coordinates, most of which are redundant due to the sparsity assumption, $\|\cdot\|_\infty$ results in a better performance, and hence, the Wasserstein ℓ_∞ formulation that stems from the $\|\cdot\|_\infty$ distance metric on $(\mathbf{x}, y)$ and induces the ℓ_1-regularizer is expected to outperform others.

We note that the ℓ_1-regularized LAD achieves similar performance to ours, since replacing $\|\boldsymbol{\beta}\|_1$ by $\|(-\boldsymbol{\beta}, 1)\|_1$ only adds a constant term to the objective function. The generalization performance (mean MSE) of the AD loss-based formulations is consistently better than those with the SR loss, since the AD loss is less affected by large perturbations in y. Also note that choosing a wrong norm for the Wasserstein metric, e.g., the Wasserstein ℓ_2, could lead to an enormous estimation error, whereas with a right norm space, we are guaranteed to outperform all others. Even when the SNR is very low, our performance is at least as good as the null estimator (see Fig. 7). Although EN and LASSO achieve similar performance to ours for moderate SNR values, they have a chance of performing even worse than the null estimator when there is little signal/information to learn from.

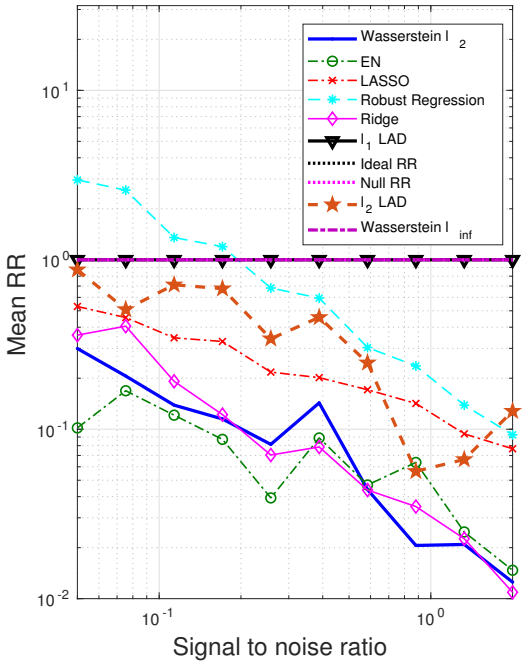
4.4 Sparse $\boldsymbol{\beta}^*$, outliers only in x

In this subsection, we will use the same sparse coefficient as in Section 4.3, but the perturbations are present only in $\mathbf{x}$. Specifically, for outliers, their predictors and responses are drawn from the following distributions:

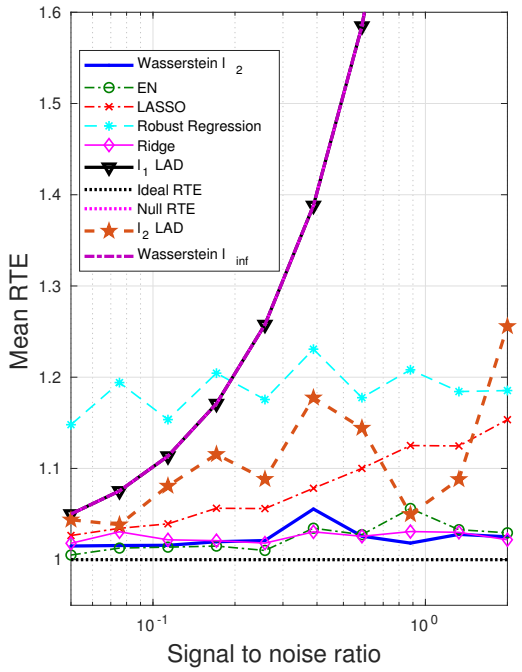
1. $\mathbf{x} \sim N_{m-1}(\mathbf{0}, \boldsymbol{\Sigma}) + N_{m-1}(5\mathbf{e}, \mathbf{I})$;
2. $y \sim N(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2)$.



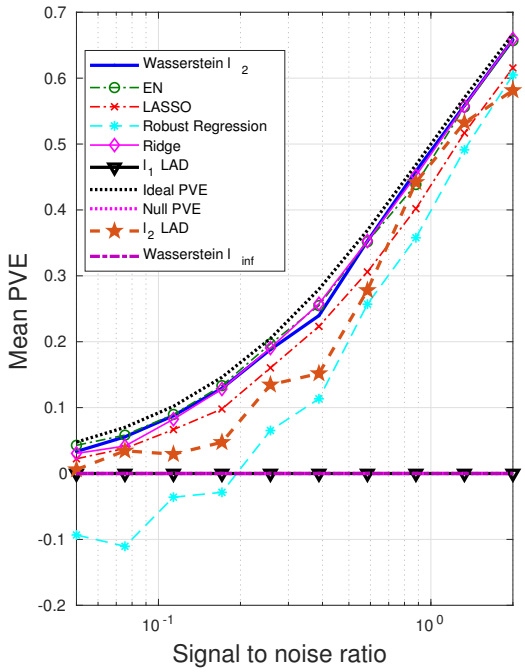
(a) Mean Squared Error.



(b) Relative risk.

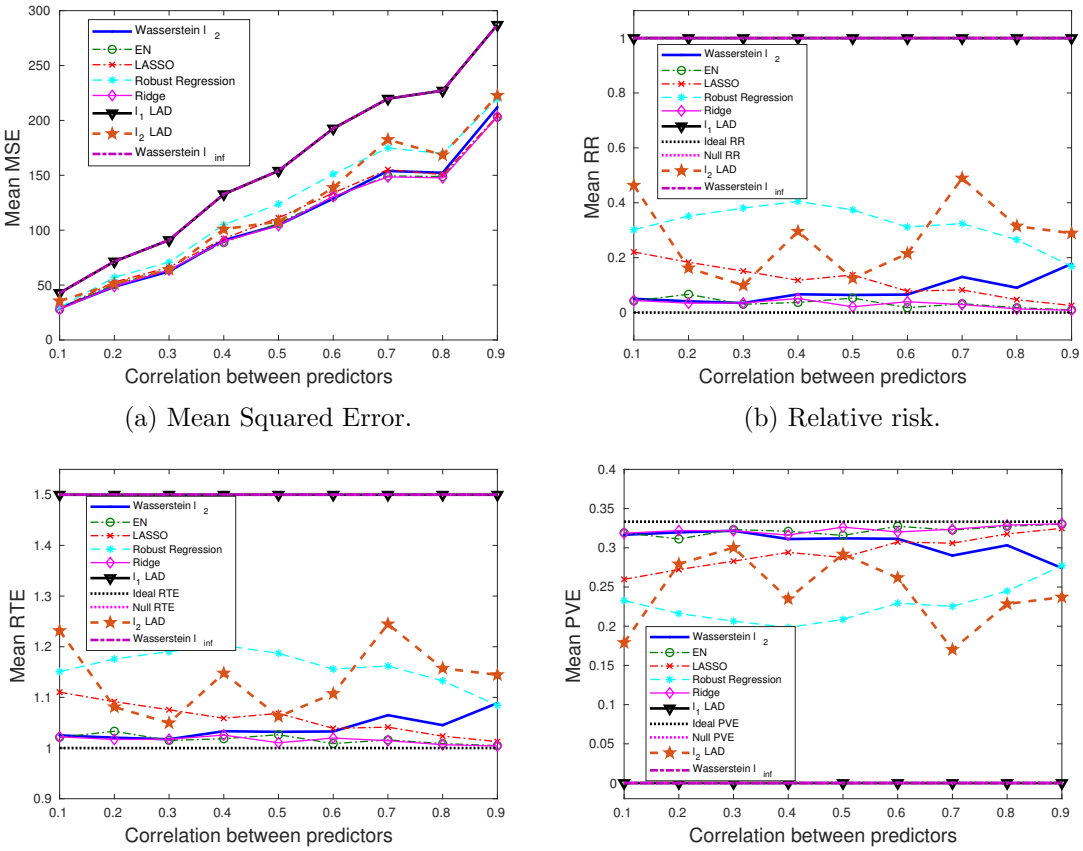


(c) Relative test error.



(d) Proportion of variance explained.

Figure 5: The impact of SNR on the performance metrics: dense $\boldsymbol{\beta}^*$, outliers only in $\mathbf{x}$.



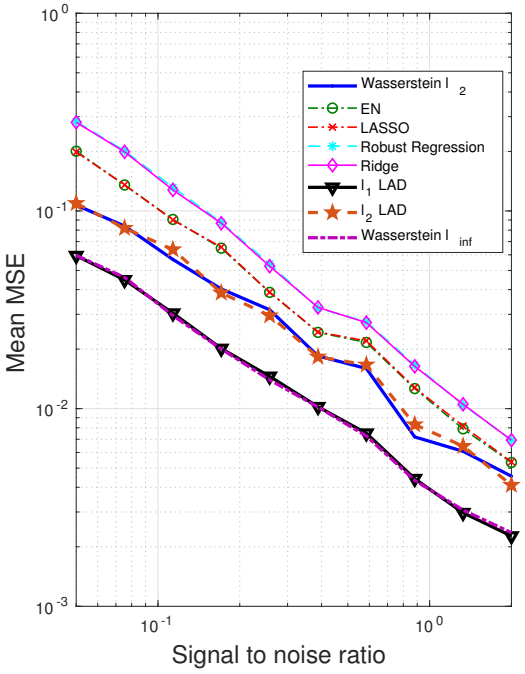
(a) Mean Squared Error.

(b) Relative risk.

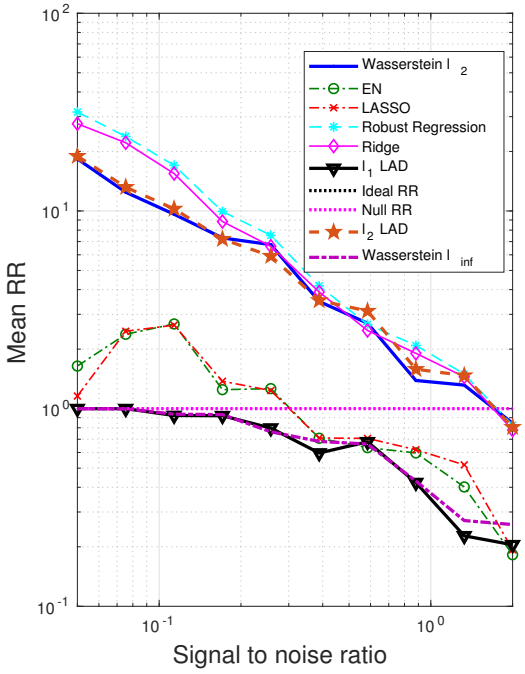
(c) Relative test error.

(d) Proportion of variance explained.

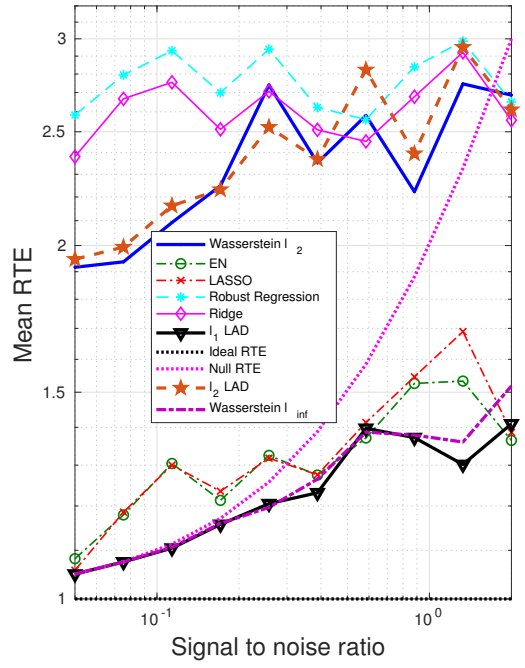
Figure 6: The impact of predictor correlation on the performance metrics: dense $\boldsymbol{\beta}^*$, outliers only in $\mathbf{x}$.



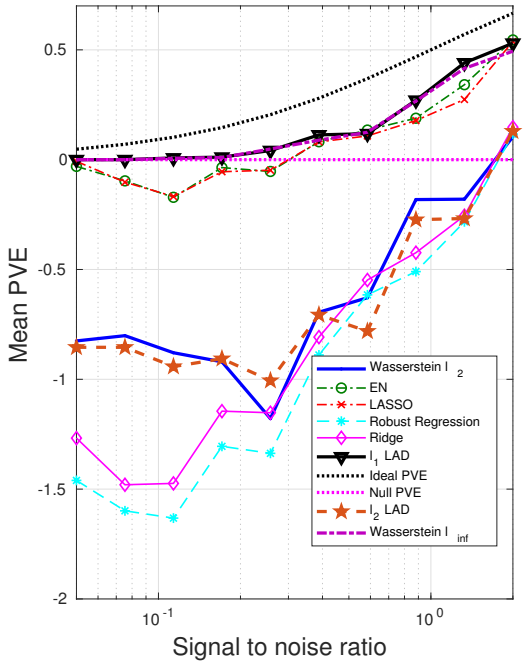
(a) Mean Squared Error.



(b) Relative risk.

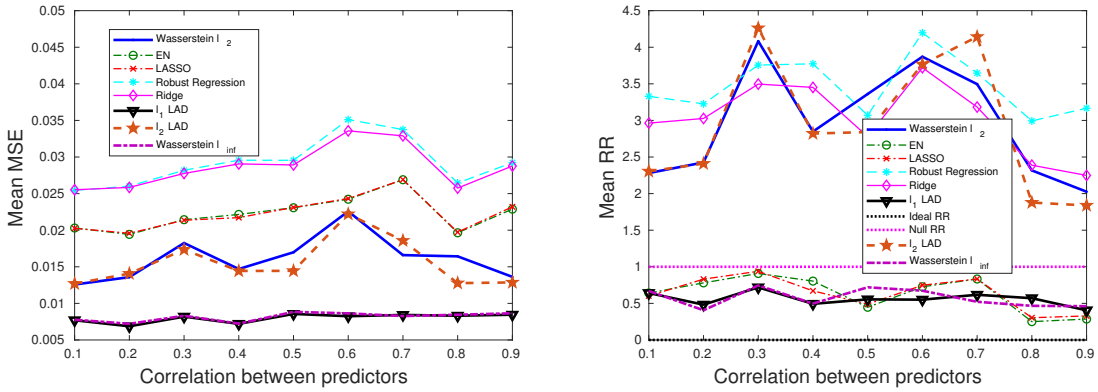


(c) Relative test error.



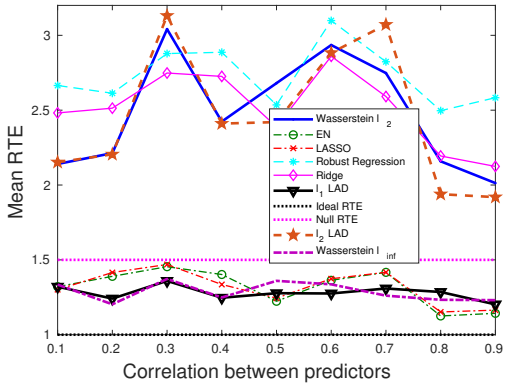
(d) Proportion of variance explained.

Figure 7: The impact of SNR on the performance metrics: sparse $\boldsymbol{\beta}^*$, outliers in both $\mathbf{x}$ and y.

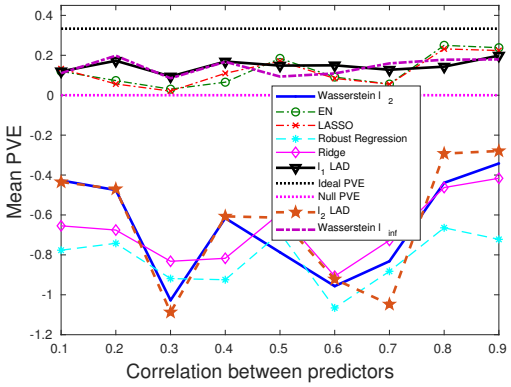


(a) Mean Squared Error.

(b) Relative risk.



(c) Relative test error.



(d) Proportion of variance explained.

Figure 8: The impact of predictor correlation on the performance metrics: sparse $\boldsymbol{\beta}^*$, outliers in both $\mathbf{x}$ and y.

Not surprisingly, the Wasserstein ℓ_∞ and the ℓ_1-regularized LAD achieve the best performance. Notice that in Section 4.3, where perturbations appear in both $\mathbf{x}$ and y, the AD loss-based formulations have smaller generalization and estimation errors than the SR loss-based formulations. When we reduce the variation in y, the SR loss seems superior to the AD loss, if we restrict attention to the improperly regularized (ℓ_2-regularizer) formulations (see Fig. 9). For the ℓ_1-regularized formulations, our Wasserstein ℓ_∞ formulation, as well as the ℓ_1-regularized LAD, is comparable with the EN and LASSO. Moreover, when there is little information to utilize (low SNR), EN and LASSO are worse than the null estimator, whereas our performance is at least as good as the null estimator.

We summarize below our main findings from all sets of experiments we have presented:

1. When a proper norm space is selected for the Wasserstein metric, the Wasserstein DRO formulation outperforms all others in terms of the generalization and estimation qualities.

2. Penalizing the extended regression coefficient $(-\boldsymbol{\beta}, 1)$ implicitly assumes a more reasonable distance metric on $(\mathbf{x}, y)$ and thus leads to a better performance.

3. The AD loss is remarkably superior to the SR loss when there is large variation in the response y.

4. The Wasserstein DRO formulation shows a more stable estimation performance than others when the correlation between predictors is varied.

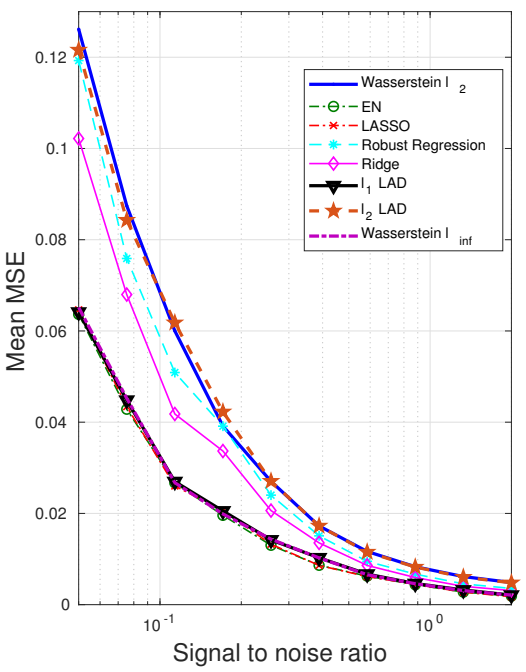
4.5 An outlier detection example

As an application, we consider an unlabeled two-class classification problem, where our goal is to identify the abnormal class of data points based on the predictor and response information using the Wasserstein formulation. We do not know a priori whether the samples are normal or abnormal, and thus classification models do not apply. The commonly used regression model for this type of problem is the M-estimation (Huber, 1964, 1973), against which we will compare in terms of the outlier detection capability.

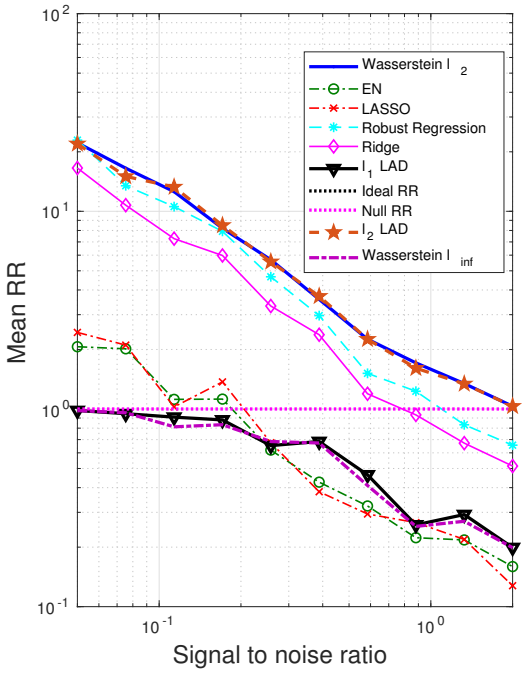
The data are generated in the same fashion as before. For clean samples, all predictors $x_1, \ldots, x_{30}$ come from a normal distribution with mean 7.5 and standard deviation 4.0. The response is a linear function of the predictors with $\beta_0^* = 0.3$, $\beta_1^* = \cdots = \beta_{30}^* = 0.5$, plus a Gaussian distributed noise term with zero mean and standard deviation σ. The outliers concentrate in a cloud that is randomly placed in the interior of the $\mathbf{x}$-space. Specifically, their predictors are uniformly distributed on $(u - 0.125, u + 0.125)$, where u is a uniform random variable on $(7.5 - 3 \times 4, 7.5 + 3 \times 4)$. The response values of the outliers are at a δ_R distance off the regression plane.

$$y = \beta_0^* + \beta_1^* x_1 + \cdots + \beta_{30}^* x_{30} + \delta_R.$$

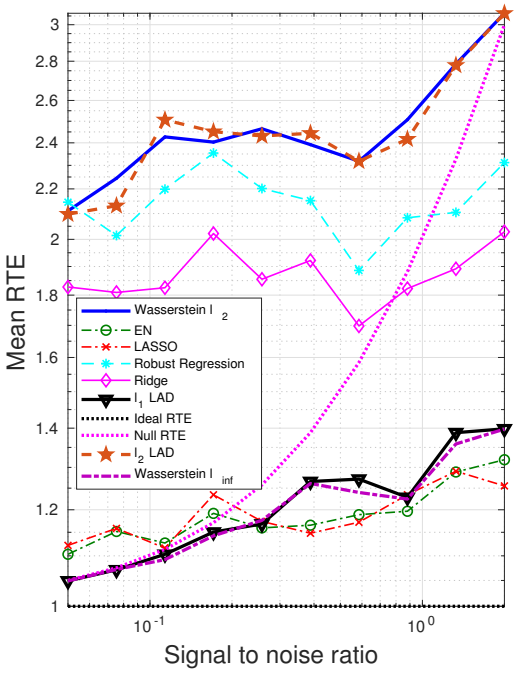
We will compare the performance of the Wasserstein ℓ_2 formulation (10) with the ℓ_1-regularized LAD and M-estimation with three cost functions – Huber (Huber, 1964, 1973), Talwar (Hinich and Talwar, 1975), and Fair (Fair, 1974). The performance metrics include



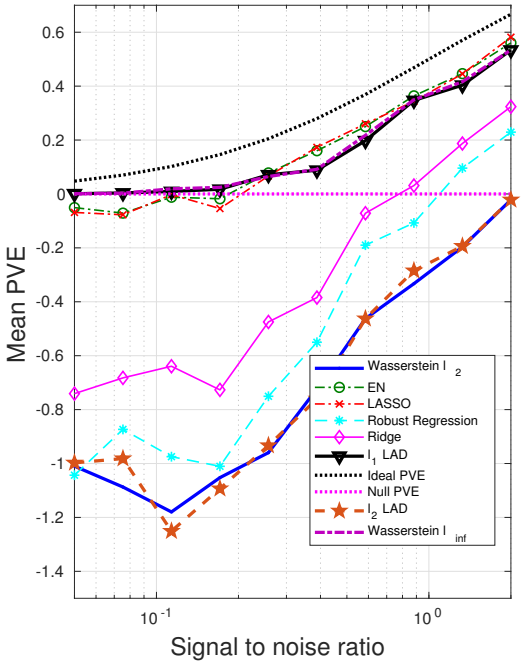
(a) Mean Squared Error.



(b) Relative risk.

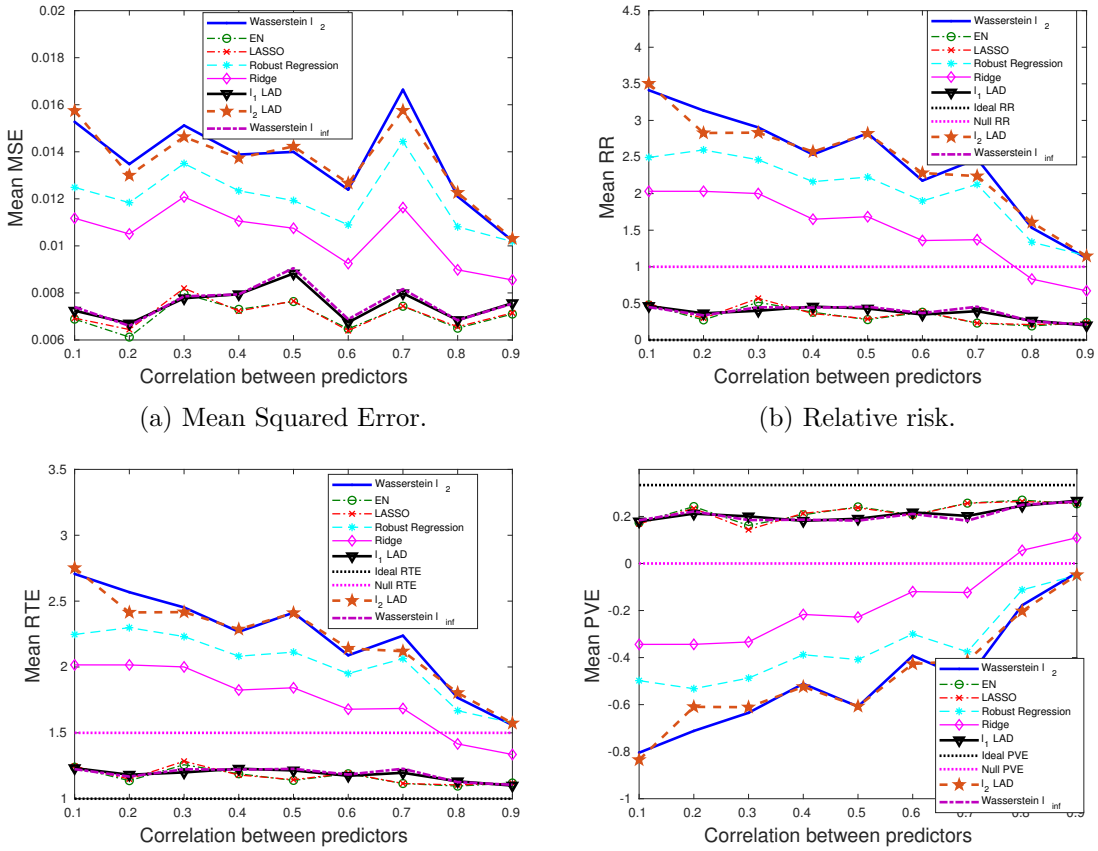


(c) Relative test error.



(d) Proportion of variance explained.

Figure 9: The impact of SNR on the performance metrics: sparse $\boldsymbol{\beta}^*$, outliers only in $\mathbf{x}$.



(a) Mean Squared Error.

(b) Relative risk.

(c) Relative test error.

(d) Proportion of variance explained.

Figure 10: The impact of predictor correlation on the performance metrics: sparse $\boldsymbol{\beta}^*$, outliers only in $\mathbf{x}$.

the *Receiver Operating Characteristic (ROC)* curve which plots the true positive rate against the false positive rate, and the related *Area Under Curve (AUC)*.

Notice that all the regression methods under consideration only generate an estimated regression coefficient. The identification of outliers is based on the residual and estimated standard deviation of the noise. Specifically,

$$\text{Outlier} = \begin{cases} \text{YES}, & \text{if } |\text{residual}| > \text{threshold} \times \hat{\sigma}, \\ \text{NO}, & \text{otherwise}, \end{cases}$$

where $\hat{\sigma}$ is the standard deviation of residuals in the entire training set. ROC curves are obtained through adjusting the threshold value.

The regularization parameters for Wasserstein DRO and regularized LAD are tuned using a separate validation set as done in previous sections. We would like to highlight a salient advantage of our approach reflected in its robustness w.r.t. the choice of ϵ. In Fig. 11 we plot the out-of-sample AUC as the radius ϵ (regularization parameter) varies, for the ℓ_2-induced Wasserstein DRO and the ℓ_1-regularized LAD. For the Wasserstein DRO curve, when ϵ is small, the Wasserstein ball contains the true distribution with low confidence and thus AUC is low. On the other hand, too large ϵ makes our solution overly conservative. Note that the robustness of our approach, indicated by the flatness of the Wasserstein DRO curve, constitutes another advantage, whereas the performance of LAD dramatically deteriorates once the regularizer deviates from the optimum. Moreover, the maximal achievable AUC for Wasserstein DRO is significantly higher than LAD.

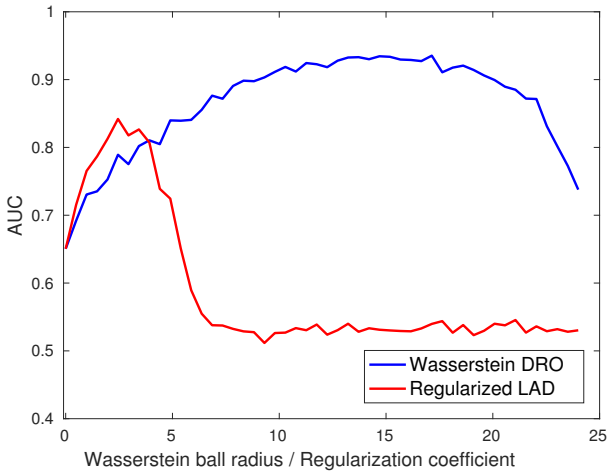
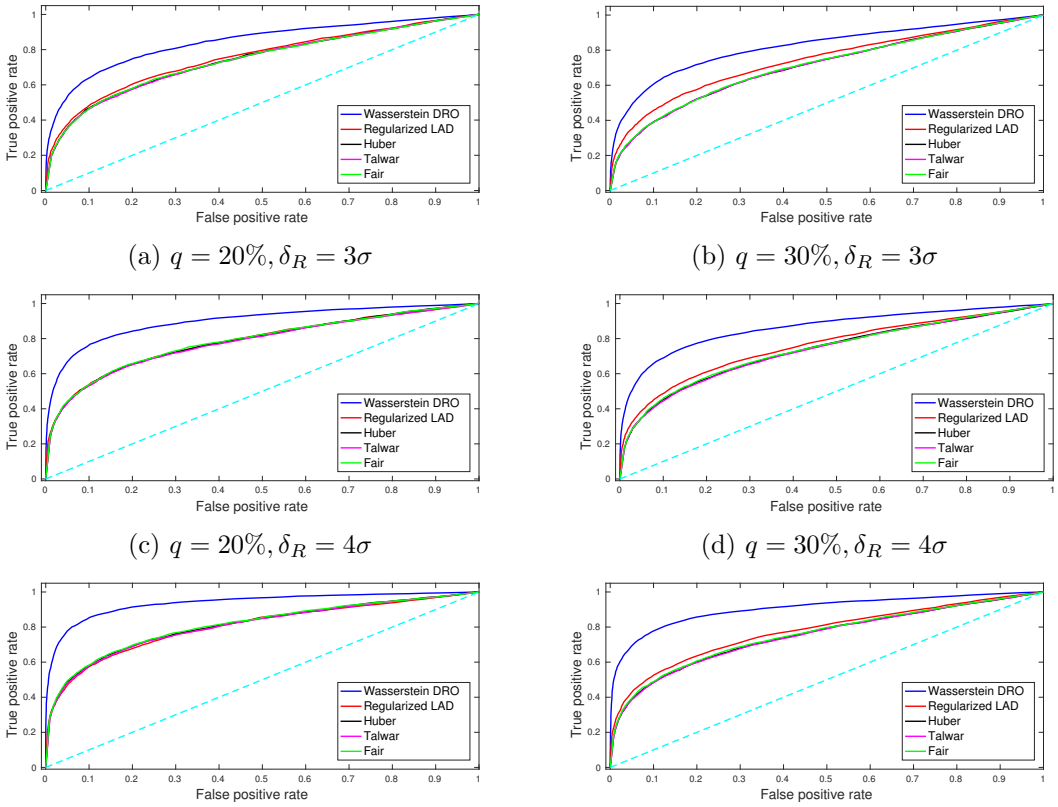


Figure 11: Out-of-sample AUC v.s. Wasserstein ball radius (regularization coefficient).

In Fig. 12 we show the ROC curves for different approaches, where q represents the percentage of outliers, and δ_R the outlying distance along y. We see that the Wasserstein DRO formulation consistently outperforms all other approaches, with its ROC curve lying well above others. In general, all approaches have better performance when the percentage of outliers is lower, and the outlying distance is larger. The approaches that use the AD loss function (e.g., Wasserstein DRO and regularized LAD) tend to outperform those that adopt

the SR loss (e.g., M-estimation which uses a variant of the SR loss). The superiority of our formulation could be attributed to the AD loss function, and the distributional robustness since we hedge against a family of plausible distributions, including the true distribution with high confidence. By contrast, M-estimation adopts an *Iteratively Reweighted Least Squares (IRLS)* procedure which assigns weights to data points based on the residuals from previous iterations, and then solves a weighted least squares estimation problem. With such an approach, there is a chance of exaggerating the influence of outliers while downplaying the importance of clean observations, especially when the initial residuals are obtained through *Ordinary Least Squares (OLS)*.



(a) $q = 20\%, \delta_R = 3\sigma$ (b) $q = 30\%, \delta_R = 3\sigma$

(c) $q = 20\%, \delta_R = 4\sigma$ (d) $q = 30\%, \delta_R = 4\sigma$

(e) $q = 20\%, \delta_R = 5\sigma$ (f) $q = 30\%, \delta_R = 5\sigma$

Figure 12: ROC curves for outliers in a randomly placed cloud, $N = 60, \sigma = 0.5$.

5. Conclusions

We presented a novel ℓ_1-loss based robust learning procedure using *Distributionally Robust Optimization (DRO)* in a linear regression framework, through which a delicate connection between the metric space on data and the regularization term has been established. The Wasserstein metric was utilized to construct the ambiguity set and a tractable reformulation was derived. It is worth noting that the linear law assumption does not necessarily limit the applicability of our model. In fact, by appropriately pre-processing the data, one can

often find a roughly linear relationship between the response and transformed explanatory variables. Our Wasserstein formulation incorporates a class of models whose specific form depends on the norm space that the Wasserstein metric is defined on. We provide out-of-sample generalization guarantees, and bound the estimation bias of the general formulation. Extensive numerical examples demonstrate the superiority of the Wasserstein formulation and shed light on the advantages of the ℓ_1-loss, the implication of the regularizer, and the selection of the norm space for the Wasserstein metric. We also presented an outlier detection example as an application of this robust learning procedure. A remarkable advantage of our approach rests in its flexibility to adjust the form of the regularizer based on the characteristics of the data.

Acknowledgments

Research partially supported by the NSF under grants DMS-1664644, CNS-1645681, CCF-1527292, and IIS-1237022, by the ARO under grant W911NF-12-1-0390, by the ONR under grant MURI N00014-16-1-2832, by the NIH under grant 1UL1TR001430 to the Clinical & Translational Science Institute at Boston University, by the Boston University Digital Health Initiative and the Center for Information and Systems Engineering, and by the joint Boston University and Brigham & Women's Hospital program in Engineering and Radiology. We thank Jenifer Siegelman and Vladimir Valtchinov for useful motivating discussions. We also thank the Editor and an anonymous reviewer whose comments helped us improve and better position this work.

Appendix A. Omitted Definitions and Proofs

This section includes proofs for the theorems and lemmas, in the order they appear in the paper.

A.1 Proof of Theorem 2.1

Proof We will adopt the notation $\mathbf{z} \triangleq (\mathbf{x}, y), \tilde{\boldsymbol{\beta}} \triangleq (-\boldsymbol{\beta}, 1)$ for ease of analysis. First rewrite $\kappa(\boldsymbol{\beta})$ as:

$$\kappa(\boldsymbol{\beta}) = \sup\Big\{\|\boldsymbol{\theta}\|_* : \sup_{\mathbf{z}|\mathbf{z}'\tilde{\boldsymbol{\beta}}\geq 0} \{(\boldsymbol{\theta}-\tilde{\boldsymbol{\beta}})'\mathbf{z}\} < \infty, \quad \sup_{\mathbf{z}|\mathbf{z}'\tilde{\boldsymbol{\beta}}\leq 0} \{(\boldsymbol{\theta}+\tilde{\boldsymbol{\beta}})'\mathbf{z}\} < \infty\Big\}.$$

Consider now the two linear optimization problems A and B:

$$\text{Problem A:} \quad \begin{array}{ll} \max & (\boldsymbol{\theta}-\tilde{\boldsymbol{\beta}})'\mathbf{z} \\ \text{s.t.} & \mathbf{z}'\tilde{\boldsymbol{\beta}} \geq 0. \end{array}$$

$$\text{Problem B:} \quad \begin{array}{ll} \max & (\boldsymbol{\theta}+\tilde{\boldsymbol{\beta}})'\mathbf{z} \\ \text{s.t.} & \mathbf{z}'\tilde{\boldsymbol{\beta}} \leq 0. \end{array}$$

Form the dual problems using dual variables r_A and r_B, respectively:

$$\text{Dual-A:} \quad \begin{array}{ll} \min & 0 \cdot r_A \\ \text{s.t.} & \tilde{\boldsymbol{\beta}} r_A = \boldsymbol{\theta} - \tilde{\boldsymbol{\beta}}, \\ & r_A \leq 0, \end{array}$$

$$\text{Dual-B:} \quad \begin{array}{ll} \min & 0 \cdot r_B \\ \text{s.t.} & \tilde{\boldsymbol{\beta}} r_B = \boldsymbol{\theta} + \tilde{\boldsymbol{\beta}}, \\ & r_B \geq 0. \end{array}$$

We want to find the set of $\boldsymbol{\theta}$ such that the optimal values of problems A and B are finite. Then, Dual-A and Dual-B need to have non-empty feasible sets, which implies the following two conditions:

$$\exists\ r_A \leq 0, \quad \text{s.t.} \quad \tilde{\boldsymbol{\beta}} r_A = \boldsymbol{\theta} - \tilde{\boldsymbol{\beta}}, \tag{22}$$

$$\exists\ r_B \geq 0, \quad \text{s.t.} \quad \tilde{\boldsymbol{\beta}} r_B = \boldsymbol{\theta} + \tilde{\boldsymbol{\beta}}. \tag{23}$$

For all i with $\tilde{\beta}_i \leq 0$, (22) implies $\theta_i - \tilde{\beta}_i \geq 0$ and (23) implies $\theta_i \leq -\tilde{\beta}_i$. On the other hand, for all j with $\tilde{\beta}_j \geq 0$, (22) and (23) imply $-\tilde{\beta}_j \leq \theta_j \leq \tilde{\beta}_j$. It is not hard to conclude that:

$$|\theta_i| \leq |\tilde{\beta}_i|, \quad \forall\ i.$$

It follows,

$$\kappa(\boldsymbol{\beta}) = \sup\{\|\boldsymbol{\theta}\|_* : |\theta_i| \leq |\tilde{\beta}_i|,\ \forall i\} = \|\tilde{\boldsymbol{\beta}}\|_*.$$

■

A.2 Proof of Lemma 3.2

Proof Suppose that $\sigma_1, \ldots, \sigma_N$ are i.i.d. uniform random variables on $\{1, -1\}$. Then, by the definition of the Rademacher complexity and Lemma 3.1,

$$\begin{aligned} \mathcal{R}_N(\mathcal{H}) &= \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{2}{N} \left| \sum_{i=1}^N \sigma_i h_{\boldsymbol{\beta}}(\mathbf{x}_i, y_i) \right| \middle| (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N) \right] \\ &\leq \frac{2\bar{B}R}{N} \mathbb{E}\left[\left| \sum_{i=1}^N \sigma_i \right| \right] \\ &\leq \frac{2\bar{B}R}{N} \mathbb{E}\left[\sqrt{\sum_{i=1}^N \sigma_i^2} \right] \\ &= \frac{2\bar{B}R}{\sqrt{N}}. \end{aligned}$$

■

A.3 Proof of Theorem 3.3

Proof We use Theorem 8 in Bartlett and Mendelson (2002), setting the following correspondences with the notation used there: $\mathcal{L}(\mathbf{x}, y) = \phi(\mathbf{x}, y) = |y - \mathbf{x}'\boldsymbol{\beta}|$. This yields the

bound (14) on the expected loss. For Eq. (15), we apply Markov's inequality to obtain:

$$\mathbb{P}\left(|y-\mathbf{x}'\hat{\boldsymbol{\beta}}| \geq \frac{1}{N}\sum_{i=1}^{N}|y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}|+\zeta\right) \leq \frac{\mathbb{E}[|y-\mathbf{x}'\hat{\boldsymbol{\beta}}|]}{\frac{1}{N}\sum_{i=1}^{N}|y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}|+\zeta} \leq \frac{\frac{1}{N}\sum_{i=1}^{N}|y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}|+\frac{2\bar{B}R}{\sqrt{N}}+\bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N}|y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}|+\zeta}.$$

■

A.4 Proof of Corollary 3.4

Proof The percentage difference requirement can be translated into:

$$\frac{2}{\sqrt{N}}+\sqrt{\frac{8\log(2/\delta)}{N}} \leq \tau,$$

from which (16) can be easily derived. ■

A.5 Proof of Corollary 3.5

Proof Based on Theorem 3.3, we just need the following inequality to hold:

$$\frac{\frac{1}{N}\sum_{i=1}^{N}|y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}|+\frac{2\bar{B}R}{\sqrt{N}}+\bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N}|y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}|+\gamma\bar{B}R} \leq \tau,$$

which is equivalent to:

$$\frac{\gamma\bar{B}R-\frac{2\bar{B}R}{\sqrt{N}}-\bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N}|y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}|+\gamma\bar{B}R} \geq 1-\tau. \tag{24}$$

We cannot obtain a lower bound for N by directly solving (24) since N appears in a summation operator. A proper relaxation to (24) is:

$$\frac{\gamma-\frac{2}{\sqrt{N}}-\sqrt{\frac{8\log(2/\delta)}{N}}}{1+\gamma} \geq 1-\tau, \tag{25}$$

due to the fact that $\frac{1}{N}\sum_{i=1}^{N}|y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}| \leq \bar{B}R$. By solving (25), we obtain (17). ■

A.6 Sub-Gaussian Random Variables and Gaussian Width

Definition 1 (Sub-Gaussian random variable) *A random variable z is sub-Gaussian if the ψ_2-norm defined below is finite, i.e.,*

$$\|\!|z|\!\|_{\psi_2} \triangleq \sup_{q\geq 1} \frac{\mathbb{E}|z|^q}{\sqrt{q}} < +\infty.$$

An equivalent property for sub-Gaussian random variables is that their tail distribution decays as fast as a Gaussian, namely,

$$\mathbb{P}(|z| \geq t) \leq 2\exp\{-t^2/C^2\}, \quad \forall t \geq 0,$$

for some constant C.

A random vector $\mathbf{z} \in \mathbb{R}^m$ is sub-Gaussian if $\mathbf{z}'\mathbf{u}$ is sub-Gaussian for any $\mathbf{u} \in \mathbb{R}^m$. The ψ_2-norm of a vector $\mathbf{z}$ is defined as:

$$\|\!|\mathbf{z}|\!\|_{\psi_2} \triangleq \sup_{\mathbf{u}\in\mathbb{S}^m} \big\|\!\big|\mathbf{z}'\mathbf{u}\big|\!\big\|_{\psi_2},$$

where $\mathbb{S}^m$ denotes the unit sphere in the m-dimensional Euclidean space. For the properties of sub-Gaussian random variables/vectors, please refer to the book by Vershynin (2017).

Definition 2 (Gaussian width) *For any set $\mathcal{A} \subseteq \mathbb{R}^m$, its Gaussian width is defined as:*

$$w(\mathcal{A}) \triangleq \mathbb{E}\Big[\sup_{\mathbf{u}\in\mathcal{A}} \mathbf{u}'\mathbf{g}\Big], \tag{26}$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an m-dimensional standard Gaussian random vector.

A.7 Proof of Theorem 3.6

In all the following proofs related to Section 3.2, we will adopt the notation $\mathbf{z} \triangleq (\mathbf{x}, y)$, $\mathbf{z}_i \triangleq (\mathbf{x}_i, y_i)$, $\tilde{\boldsymbol{\beta}} \triangleq (-\boldsymbol{\beta}, 1)$, $\tilde{\boldsymbol{\beta}}_{\text{est}} \triangleq (-\hat{\boldsymbol{\beta}}, 1)$, $\tilde{\boldsymbol{\beta}}_{\text{true}} \triangleq (-\boldsymbol{\beta}^*, 1)$ for ease of exposition.

Proof Since both $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ are feasible (the latter due to Assumption E), we have:

$$\|\mathbf{Z}'\tilde{\boldsymbol{\beta}}_{\text{est}}\|_1 \leq \gamma_N,$$
$$\|\mathbf{Z}'\tilde{\boldsymbol{\beta}}_{\text{true}}\|_1 \leq \gamma_N,$$

from which we derive that $\|\mathbf{Z}'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})\|_1 \leq 2\gamma_N$. Since $\hat{\boldsymbol{\beta}}$ is an optimal solution to (18) and $\boldsymbol{\beta}^*$ a feasible solution, it follows that $\|\tilde{\boldsymbol{\beta}}_{\text{est}}\|_* \leq \|\tilde{\boldsymbol{\beta}}_{\text{true}}\|_*$. This implies that $\boldsymbol{\nu} = \tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}$ satisfies the condition $\|\tilde{\boldsymbol{\beta}}_{\text{true}} + \mathbf{v}\|_* \leq \|\tilde{\boldsymbol{\beta}}_{\text{true}}\|_*$ included in the definition of $\mathcal{A}(\boldsymbol{\beta}^*)$ and, furthermore, $(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})/\|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_2 \in \mathcal{A}(\boldsymbol{\beta}^*)$. Together with Assumption D, this yields

$$(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})'\mathbf{Z}\mathbf{Z}'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}) \geq \underline{\alpha}\|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_2^2. \tag{27}$$

On the other hand, from the Cauchy-Schwarz inequality:

$$\begin{aligned}
(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})'\mathbf{Z}\mathbf{Z}'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}) &\leq \|\mathbf{Z}'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})\|_1 \|\mathbf{Z}'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})\|_\infty \\
&\leq 2\gamma_N \max_i |\mathbf{z}_i'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})| \\
&\leq 2\gamma_N \max_i \|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_* \|\mathbf{z}_i\| \\
&\leq 2R\gamma_N \|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_*.
\end{aligned} \tag{28}$$

Combining (27) and (28), we have:

$$\begin{aligned}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &= \|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_2 \\ &\leq \frac{2R\gamma_N}{\underline{\alpha}} \frac{\|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_*}{\|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_2} \\ &\leq \frac{2R\gamma_N}{\underline{\alpha}} \Psi(\boldsymbol{\beta}^*),\end{aligned}$$

where the last step follows from the fact that $(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})/\|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_2 \in \mathcal{A}(\boldsymbol{\beta}^*)$. ■

A.8 Proof of Lemma 3.7

Proof Define $\hat{\boldsymbol{\Gamma}} = \frac{1}{N}\sum_{i=1}^N \mathbf{z}_i\mathbf{z}_i'$. Consider the set of functions $\mathcal{F} = \{f_{\mathbf{w}}(\mathbf{z}) = \mathbf{z}'\boldsymbol{\Gamma}^{-1/2}\mathbf{w} | \mathbf{w} \in \mathcal{A}_{\boldsymbol{\Gamma}}\}$. Then, for any $f_{\mathbf{w}} \in \mathcal{F}$,

$$\begin{aligned}\mathbb{E}[f_{\mathbf{w}}^2] &= \mathbb{E}[\mathbf{w}'\boldsymbol{\Gamma}^{-1/2}\mathbf{z}\mathbf{z}'\boldsymbol{\Gamma}^{-1/2}\mathbf{w}] \\ &= \mathbf{w}'\boldsymbol{\Gamma}^{-1/2}\mathbb{E}[\mathbf{z}\mathbf{z}']\boldsymbol{\Gamma}^{-1/2}\mathbf{w} \\ &= \mathbf{w}'\mathbf{w} \\ &= 1,\end{aligned}$$

where we used $\boldsymbol{\Gamma} = \mathbb{E}[\mathbf{z}\mathbf{z}']$ and the fact that $\mathbf{w} \in \mathcal{A}_{\boldsymbol{\Gamma}}$.

For any $f_{\mathbf{w}} \in \mathcal{F}$ we have

$$\begin{aligned}|||f_{\mathbf{w}}|||_{\psi_2} &= \left|\left|\left|\mathbf{z}'\boldsymbol{\Gamma}^{-1/2}\mathbf{w}\right|\right|\right|_{\psi_2} \\ &= \left|\left|\left|\mathbf{z}'\boldsymbol{\Gamma}^{-1/2}\mathbf{w}\right|\right|\right|_{\psi_2} \frac{\|\boldsymbol{\Gamma}^{-1/2}\mathbf{w}\|_2}{\|\boldsymbol{\Gamma}^{-1/2}\mathbf{w}\|_2} \\ &= \left|\left|\left|\mathbf{z}'\frac{\boldsymbol{\Gamma}^{-1/2}\mathbf{w}}{\|\boldsymbol{\Gamma}^{-1/2}\mathbf{w}\|_2}\right|\right|\right|_{\psi_2} \|\boldsymbol{\Gamma}^{-1/2}\mathbf{w}\|_2 \\ &\leq \mu\sqrt{\mathbf{w}'\boldsymbol{\Gamma}^{-1}\mathbf{w}} \\ &\leq \mu\sqrt{\frac{1}{\lambda_{\min}}\|\mathbf{w}\|_2^2} \\ &= \mu\sqrt{\frac{1}{\lambda_{\min}}} = \bar{\mu},\end{aligned}$$

where the first inequality used Assumption F and the second inequality used Assumption G.

Applying Theorem D from Mendelson et al. (2007), for any $\theta > 0$ and when

$$\tilde{C}_1\bar{\mu}\gamma_2(\mathcal{F}, |||\cdot|||_{\psi_2}) \leq \theta\sqrt{N},$$

with probability at least $1-\exp(-\tilde{C}_2\theta^2 N/\bar{\mu}^4)$ we have

$$\begin{aligned}\sup_{f_{\mathbf{w}}\in\mathcal{F}}\Big|\frac{1}{N}\sum_{i=1}^{N} f_{\mathbf{w}}^2(\mathbf{z}_i)-\mathbb{E}[f_{\mathbf{w}}^2]\Big| &= \sup_{f_{\mathbf{w}}\in\mathcal{F}}\Big|\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}'\mathbf{\Gamma}^{-1/2}\mathbf{z}_i\mathbf{z}_i'\mathbf{\Gamma}^{-1/2}\mathbf{w}-1\Big| \\ &= \sup_{\mathbf{w}\in\mathcal{A}_{\mathbf{\Gamma}}}\Big|\mathbf{w}'\mathbf{\Gamma}^{-1/2}\hat{\mathbf{\Gamma}}\mathbf{\Gamma}^{-1/2}\mathbf{w}-1\Big| \\ &\leq \theta, \end{aligned} \tag{29}$$

where $\tilde{C}_1$ is some positive constant and $\gamma_2(\mathcal{F},\|\cdot\|_{\psi_2})$ is defined in Mendelson et al. (2007) as a measure of the size of the set $\mathcal{F}$ with respect to the metric $\|\cdot\|_{\psi_2}$. Using $\theta=1/2$, and properties of $\gamma_2(\mathcal{F},\|\cdot\|_{\psi_2})$ outlined in Chen and Banerjee (2016), we can set N to satisfy

$$\begin{aligned}\tilde{C}_1\bar{\mu}\gamma_2(\mathcal{F},\|\cdot\|_{\psi_2}) &\leq \tilde{C}_1\bar{\mu}^2\gamma_2(\mathcal{A}_{\mathbf{\Gamma}},\|\cdot\|_2) \\ &\leq \tilde{C}_1\bar{\mu}^2 C_0 w(\mathcal{A}_{\mathbf{\Gamma}}) \\ &\leq \frac{1}{2}\sqrt{N},\end{aligned}$$

for some positive constant C_0, where we used Eq. (44) in Chen and Banerjee (2016). This implies

$$N \geq C_1\bar{\mu}^4(w(\mathcal{A}_{\mathbf{\Gamma}}))^2$$

for some positive constant C_1. Thus, for such N and with probability at least $1-\exp(-C_2N/\bar{\mu}^4)$, for some positive constant C_2, (29) holds with $\theta=1/2$. This implies that for all $\mathbf{w}\in\mathcal{A}_{\mathbf{\Gamma}}$,

$$\Big|\mathbf{w}'\mathbf{\Gamma}^{-1/2}\hat{\mathbf{\Gamma}}\mathbf{\Gamma}^{-1/2}\mathbf{w}-1\Big|\leq\frac{1}{2}$$

or

$$\mathbf{w}'\mathbf{\Gamma}^{-1/2}\hat{\mathbf{\Gamma}}\mathbf{\Gamma}^{-1/2}\mathbf{w}\geq\frac{1}{2}=\frac{1}{2}\mathbf{w}'\mathbf{\Gamma}^{-1/2}\mathbf{\Gamma}\mathbf{\Gamma}^{-1/2}\mathbf{w}.$$

By the definition of $\mathcal{A}_{\mathbf{\Gamma}}$, for any $\mathbf{v}\in\mathcal{A}(\boldsymbol{\beta}^*)$,

$$\mathbf{v}'\hat{\mathbf{\Gamma}}\mathbf{v}\geq\frac{1}{2}\mathbf{v}'\mathbf{\Gamma}\mathbf{v}.$$

Noting that $\hat{\mathbf{\Gamma}}=(1/N)\mathbf{Z}\mathbf{Z}'$ yields the desired result. ■

A.9 Proof of Lemma 3.8

We follow the proof of Lemma 4 in Chen and Banerjee (2016), adapted to our setting. We include all key steps for completeness.

Proof Recall the definition of the Gaussian width $w(\mathcal{A}_{\mathbf{\Gamma}})$ (cf. (26)):

$$w(\mathcal{A}_{\mathbf{\Gamma}})=\mathbb{E}\Big[\sup_{\mathbf{u}\in\mathcal{A}_{\mathbf{\Gamma}}}\mathbf{u}'\mathbf{g}\Big],$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We have:

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{A}_{\mathbf{\Gamma}}} \mathbf{w}'\mathbf{g} &= \sup_{\mathbf{w}\in\mathcal{A}_{\mathbf{\Gamma}}} \mathbf{w}'\mathbf{\Gamma}^{-1/2}\mathbf{\Gamma}^{1/2}\mathbf{g} \\ &= \sup_{\mathbf{w}\in\mathcal{A}_{\mathbf{\Gamma}}} \|\mathbf{\Gamma}^{-1/2}\mathbf{w}\|_2 \frac{\mathbf{w}'\mathbf{\Gamma}^{-1/2}}{\|\mathbf{\Gamma}^{-1/2}\mathbf{w}\|_2}\mathbf{\Gamma}^{1/2}\mathbf{g} \\ &\leq \sqrt{\frac{1}{\lambda_{\min}}} \sup_{\mathbf{v}\in\mathrm{cone}(\mathcal{A}(\boldsymbol{\beta}^*))\cap\mathbb{B}^m} \mathbf{v}'\mathbf{\Gamma}^{1/2}\mathbf{g},\end{aligned}$$

where $\mathbb{B}^m$ is the unit ball in the m-dimensional Euclidean space and the inequality used Assumption G and the fact that $\mathbf{w}'\mathbf{\Gamma}^{-1/2}/\|\mathbf{\Gamma}^{-1/2}\mathbf{w}\|_2 \in \mathbb{B}^m$ and $\mathbf{w} \in \mathcal{A}_{\mathbf{\Gamma}}$.

Define $\mathcal{T} = \mathrm{cone}(\mathcal{A}(\boldsymbol{\beta}^*)) \cap \mathbb{B}^m$, and consider the stochastic process $\{S_{\mathbf{v}} = \mathbf{v}'\mathbf{\Gamma}^{1/2}\mathbf{g}\}_{\mathbf{v}\in\mathcal{T}}$. For any $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{T}$,

$$\begin{aligned}\|\!|S_{\mathbf{v}_1} - S_{\mathbf{v}_2}|\!\|_{\psi_2} &= \left|\!\left|\!\left|(\mathbf{v}_1 - \mathbf{v}_2)'\mathbf{\Gamma}^{1/2}\mathbf{g}\right|\!\right|\!\right|_{\psi_2} \\ &= \|\mathbf{\Gamma}^{1/2}(\mathbf{v}_1 - \mathbf{v}_2)\|_2 \left|\!\left|\!\left| \frac{(\mathbf{v}_1 - \mathbf{v}_2)'\mathbf{\Gamma}^{1/2}\mathbf{g}}{\|\mathbf{\Gamma}^{1/2}(\mathbf{v}_1 - \mathbf{v}_2)\|_2} \right|\!\right|\!\right|_{\psi_2} \\ &\leq \|\mathbf{\Gamma}^{1/2}(\mathbf{v}_1 - \mathbf{v}_2)\|_2 \sup_{\mathbf{u}\in\mathbb{S}^m} \left|\!\left|\!\left|\mathbf{u}'\mathbf{g}\right|\!\right|\!\right|_{\psi_2} \\ &= \mu_0\|\mathbf{\Gamma}^{1/2}(\mathbf{v}_1 - \mathbf{v}_2)\|_2 \\ &\leq \mu_0\sqrt{\lambda_{\max}}\|\mathbf{v}_1 - \mathbf{v}_2\|_2,\end{aligned}$$

where the last step used Assumption G.

Then, by the tail behavior of sub-Gaussian random variables (see Hoeffding bound, Thm. 2.6.2 in (Vershynin, 2017)), we have:

$$\mathbb{P}(|S_{\mathbf{v}_1} - S_{\mathbf{v}_2}| \geq \delta) \leq 2\exp\left(-\frac{C_{01}\delta^2}{\mu_0^2\lambda_{\max}\|\mathbf{v}_1 - \mathbf{v}_2\|_2^2}\right),$$

for some positive constant C_{01}.

To bound the supremum of $S_{\mathbf{v}}$, we define the metric $s(\mathbf{v}_1, \mathbf{v}_2) = \mu_0\sqrt{\lambda_{\max}}\|\mathbf{v}_1 - \mathbf{v}_2\|_2$. Then, by Lemma B in Chen and Banerjee (2016),

$$\begin{aligned}\mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{T}} \mathbf{v}'\mathbf{\Gamma}^{1/2}\mathbf{g}\right] &\leq C_{02}\gamma_2(\mathcal{T}, s) \\ &= C_{02}\mu_0\sqrt{\lambda_{\max}}\gamma_2(\mathcal{T}, \|\cdot\|_2) \\ &\leq C_3\mu_0\sqrt{\lambda_{\max}}w(\mathcal{T}),\end{aligned}$$

for positive constants C_{02}, C_3, where $\gamma_2(\mathcal{T}, s)$ is the γ_2-functional we referred to in the proof of Lemma 3.7. Since $\mathcal{T} = \mathrm{cone}(\mathcal{A}(\boldsymbol{\beta}^*)) \cap \mathbb{B}^m \subseteq \mathrm{conv}(\mathcal{A}(\boldsymbol{\beta}^*) \cup \{\mathbf{0}\})$, by Lemma 2 in Maurer et al. (2014),

$$\begin{aligned}w(\mathcal{T}) &\leq w(\mathrm{conv}(\mathcal{A}(\boldsymbol{\beta}^*) \cup \{\mathbf{0}\})) \\ &= w(\mathcal{A}(\boldsymbol{\beta}^*) \cup \{\mathbf{0}\}) \\ &\leq \max\{w(\mathcal{A}(\boldsymbol{\beta}^*)), w(\{\mathbf{0}\})\} + 2\sqrt{\ln 4} \\ &\leq w(\mathcal{A}(\boldsymbol{\beta}^*)) + 3.\end{aligned}$$

Thus,

$$\begin{aligned} w(\mathcal{A}_{\mathbf{\Gamma}}) &= \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{A}_{\mathbf{\Gamma}}} \mathbf{w}'\mathbf{g}\Big] \\ &\leq \sqrt{\frac{1}{\lambda_{\min}}}\mathbb{E}\Big[\sup_{\mathbf{v}\in\mathcal{T}} \mathbf{v}'\mathbf{\Gamma}^{1/2}\mathbf{g}\Big] \\ &\leq C_3\sqrt{\frac{1}{\lambda_{\min}}}\mu_0\sqrt{\lambda_{\max}}w(\mathcal{T}) \\ &\leq C_3\mu_0\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\Big(w(\mathcal{A}(\boldsymbol{\beta}^*))+3\Big). \end{aligned}$$

■

A.10 Proof of Corollary 3.9

Proof Combining Lemmas 3.7 and 3.8, and using the fact that for any $\mathbf{v}\in\mathcal{A}(\boldsymbol{\beta}^*)$,

$$\frac{N}{2}\mathbf{v}'\mathbf{\Gamma}\mathbf{v} \geq \frac{N\lambda_{\min}}{2},$$

we can derive the desired result. ■

A.11 Proof of Lemma 3.10

Proof By the definition of dual norm, we know that:

$$\|\tilde{\boldsymbol{\beta}}'\mathbf{Z}\|_1 = \sup_{\mathbf{v}\in\mathcal{B}_u} \tilde{\boldsymbol{\beta}}'\mathbf{Z}\mathbf{v} = \sup_{\mathbf{v}\in\mathcal{B}_u}\sum_{i=1}^N v_i\tilde{\boldsymbol{\beta}}'\mathbf{z}_i.$$

Since $v_i\tilde{\boldsymbol{\beta}}'\mathbf{z}_i,\ i=1,\ldots,N$ are independent centered sub-Gaussian random variables, and

$$\Big|\Big|\Big|v_i\tilde{\boldsymbol{\beta}}'\mathbf{z}_i\Big|\Big|\Big|_{\psi_2} \leq \mu\|v_i\tilde{\boldsymbol{\beta}}\|_2,$$

we have that $\sum_{i=1}^N v_i\tilde{\boldsymbol{\beta}}'\mathbf{z}_i$ is also a centered sub-Gaussian random variable with

$$\begin{aligned} \Big|\Big|\Big|\sum_{i=1}^N v_i\tilde{\boldsymbol{\beta}}'\mathbf{z}_i\Big|\Big|\Big|_{\psi_2}^2 &\leq C_{03}^2\sum_{i=1}^N \mu^2\|v_i\tilde{\boldsymbol{\beta}}\|_2^2 \\ &= C_{03}^2\mu^2\|\tilde{\boldsymbol{\beta}}\|_2^2\|\mathbf{v}\|_2^2, \end{aligned}$$

for a positive constant C_{03}.

Consider the stochastic process $\{S_{\mathbf{v}} = \tilde{\boldsymbol{\beta}}'\mathbf{Z}\mathbf{v}\}_{\mathbf{v}\in\mathcal{B}_u}$. As in the proof of Lemma 3.8,

$$|||S_{\mathbf{v}_1} - S_{\mathbf{v}_2}|||_{\psi_2} \leq C_{03}\mu\|\tilde{\boldsymbol{\beta}}\|_2\|\mathbf{v}_1-\mathbf{v}_2\|_2.$$

By the tail behavior of sub-Gaussian random variables (Vershynin, 2017), we know:

$$\mathbb{P}(|S_{\mathbf{v}_1} - S_{\mathbf{v}_2}| \geq \delta) \leq 2\exp\bigg(-\frac{C_{04}\delta^2}{\mu^2\|\tilde{\boldsymbol{\beta}}\|_2^2\|\mathbf{v}_1 - \mathbf{v}_2\|_2^2}\bigg),$$

for a positive constant C_{04}.

Define the metric $s(\mathbf{v}_1, \mathbf{v}_2) = \mu\|\tilde{\boldsymbol{\beta}}\|_2\|\mathbf{v}_1 - \mathbf{v}_2\|_2$. Then, by Lemma B in Chen and Banerjee (2016),

$$\mathbb{P}\bigg(\sup_{\mathbf{v}_1,\mathbf{v}_2\in\mathcal{B}_u} |S_{\mathbf{v}_1} - S_{\mathbf{v}_2}| \geq C_{05}\big(\gamma_2(\mathcal{B}_u, s) + \delta\cdot\mathrm{diam}(\mathcal{B}_u, s)\big)\bigg) \leq C_4\exp(-\delta^2),$$

for positive constants C_{05}, C_4. Also,

$$\begin{aligned}
\gamma_2(\mathcal{B}_u, s) &= \mu\|\tilde{\boldsymbol{\beta}}\|_2\gamma_2(\mathcal{B}_u, \|\cdot\|_2) \leq C_5\mu\|\tilde{\boldsymbol{\beta}}\|_2 w(\mathcal{B}_u),\\
\mathrm{diam}(\mathcal{B}_u, s) &= \sup_{\mathbf{v}_1,\mathbf{v}_2\in\mathcal{B}_u} s(\mathbf{v}_1, \mathbf{v}_2)\\
&= \mu\|\tilde{\boldsymbol{\beta}}\|_2 \sup_{\mathbf{v}_1,\mathbf{v}_2\in\mathcal{B}_u} \|\mathbf{v}_1 - \mathbf{v}_2\|_2\\
&\leq 2\mu\|\tilde{\boldsymbol{\beta}}\|_2 \sup_{\mathbf{v}\in\mathcal{B}_u} \|\mathbf{v}\|_2\\
&= 2\mu\|\tilde{\boldsymbol{\beta}}\|_2\rho,
\end{aligned}$$

for positive constants C_5. Therefore, noting that $\sup_{\mathbf{v}_1,\mathbf{v}_2\in\mathcal{B}_u} |S_{\mathbf{v}_1} - S_{\mathbf{v}_2}| \geq 2\sup_{\mathbf{v}\in\mathcal{B}_u} S_{\mathbf{v}}$, we obtain

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{\mathbf{v}\in\mathcal{B}_u} S_{\mathbf{v}} \geq C_{05}\Big(\frac{C_5}{2}\mu\|\tilde{\boldsymbol{\beta}}\|_2 w(\mathcal{B}_u) + \delta\mu\|\tilde{\boldsymbol{\beta}}\|_2\rho\Big)\bigg)\\
\leq\ &\mathbb{P}\bigg(\sup_{\mathbf{v}_1,\mathbf{v}_2\in\mathcal{B}_u} |S_{\mathbf{v}_1} - S_{\mathbf{v}_2}| \geq C_{05}\big(\gamma_2(\mathcal{B}_u, s) + \delta\mathrm{diam}(\mathcal{B}_u, s)\big)\bigg)\\
\leq\ &C_4\exp(-\delta^2).
\end{aligned}$$

Set $\delta = \frac{C_5 w(\mathcal{B}_u)}{2\rho}$; then with probability at least $1 - C_4\exp(-\frac{C_5^2(w(\mathcal{B}_u))^2}{4\rho^2})$,

$$\sup_{\mathbf{v}\in\mathcal{B}_u} S_{\mathbf{v}} \leq C\mu\bar{B}_2 w(\mathcal{B}_u).$$

The result follows. ■

A.12 Proof of the Result in Section 4

We will show that if the Wasserstein metric is defined by the following metric s_c:

$$s_c(\mathbf{x}, y) = \|(\mathbf{x}, cy)\|_\infty,$$

then as $c \to \infty$, the corresponding Wasserstein DRO formulation becomes:

$$\inf_{\boldsymbol{\beta}\in\mathcal{B}} \frac{1}{N}\sum_{i=1}^N |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon\|\boldsymbol{\beta}\|_1,$$

which is the ℓ_1-regularized LAD.

Proof We first define a new notion of norm on $(\mathbf{x}, y)$ where $\mathbf{x} = (x_1, \ldots, x_{m-1})$:

$$\|(\mathbf{x}, y)\|_{\mathbf{w},p} \triangleq \|(x_1 w_1, \ldots, x_{m-1} w_{m-1}, y w_m)\|_p,$$

for some m-dimensional weighting vector $\mathbf{w} = (w_1, \ldots, w_m)$, and $p \geq 1$. Then, $s_c(\mathbf{x}, y) = \|(\mathbf{x}, y)\|_{\mathbf{w},\infty}$ with $\mathbf{w} = (1, \ldots, 1, c)$. To obtain the Wasserstein DRO formulation, the key is to derive the dual norm of $\|\cdot\|_{\mathbf{w},\infty}$. Hölder's inequality (Rogers, 1888) will be used for the derivation. We state it below for convenience.

Theorem 1 (Hölder's inequality) *Suppose we have two scalars $p, q > 1$ and $1/p + 1/q = 1$. For any two vectors $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$, the following holds.*

$$\sum_{i=1}^{n} |a_i b_i| \leq \|\mathbf{a}\|_p \|\mathbf{b}\|_q.$$

We will use the notation $\mathbf{z} \triangleq (\mathbf{x}, y)$. Based on the definition of dual norm, we are interested in solving the following optimization problem for $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^m$:

$$\begin{aligned} \max_{\mathbf{z}} \quad & \mathbf{z}' \tilde{\boldsymbol{\beta}} \\ \text{s.t.} \quad & \|\mathbf{z}\|_{\mathbf{w},\infty} \leq 1. \end{aligned} \tag{30}$$

The optimal value of problem (30), which is a function of $\tilde{\boldsymbol{\beta}}$, gives the dual norm evaluated at $\tilde{\boldsymbol{\beta}}$. Using Hölder's inequality, we can write

$$\mathbf{z}' \tilde{\boldsymbol{\beta}} = \sum_{i=1}^{m} (w_i z_i) \Big(\frac{1}{w_i} \tilde{\beta}_i \Big) \leq \|\mathbf{z}\|_{\mathbf{w},\infty} \|\tilde{\boldsymbol{\beta}}\|_{\mathbf{w}^{-1},1} \leq \|\tilde{\boldsymbol{\beta}}\|_{\mathbf{w}^{-1},1},$$

where $\mathbf{w}^{-1} \triangleq (\frac{1}{w_1}, \ldots, \frac{1}{w_m})$. The last inequality is due to the constraint $\|\mathbf{z}\|_{\mathbf{w},\infty} \leq 1$. It follows that the dual norm of $\|\cdot\|_{\mathbf{w},\infty}$ is just $\|\cdot\|_{\mathbf{w}^{-1},1}$. Back to our problem setting, using $\mathbf{w} = (1, \ldots, 1, c)$, and evaluating the dual norm at $(-\boldsymbol{\beta}, 1)$, we have the following Wasserstein DRO formulation as $c \to \infty$:

$$\lim_{c \to \infty} \inf_{\boldsymbol{\beta} \in \mathcal{B}} \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathbf{x}_i' \boldsymbol{\beta}| + \epsilon \|(-\boldsymbol{\beta}, 1)\|_{\mathbf{w}^{-1},1} = \inf_{\boldsymbol{\beta} \in \mathcal{B}} \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathbf{x}_i' \boldsymbol{\beta}| + \epsilon \|\boldsymbol{\beta}\|_1.$$

■

References

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Güzin Bayraksan and David K Love. Data-driven stochastic programming using ϕ-divergences. *Tutorials in Operations Research*, pages 1–19, 2015.

Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

Dimitris Bertsimas and Martin S Copenhaver. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 2017.

Dimitris Bertsimas, Vishal Gupta, and Ioannis Ch Paschalidis. Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming*, 153(2):595–633, 2015.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. Technical Report arXiv:1604.01446, 2016.

Ruidi Chen, Ioannis Ch Paschalidis, Hiroto Hatabu, Vladimir Valtchinov, and Jenifer Siegelman. Detection of unwarranted CT radiation exposure from patient and imaging protocol meta-data using regularized regression. *Working paper*, 2018.

Sheng Chen and Arindam Banerjee. Alternating estimation for structured high-dimensional multi-response models. *arXiv preprint arXiv:1606.08957*, 2016.

David Coleman, Paul Holland, Neil Kaden, Virginia Klema, and Stephen C Peters. A system of subroutines for iteratively reweighted least squares computations. *ACM Transactions on Mathematical Software (TOMS)*, 6(3):327–336, 1980.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

E Erdoğan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Available at Optimization Online*, 2015.

Ray C Fair. On the robust estimation of econometric models. In *Annals of Economic and Social Measurement, Volume 3, number 4*, pages 667–677. NBER, 1974.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917, 2010.

Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the LASSO. *arXiv preprint arXiv:1707.08692*, 2017.

Melvin J Hinich and Prem P Talwar. A simple method for robust regression. *Journal of the American Statistical Association*, 70(349):113–119, 1975.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Zhaolin Hu and L Jeff Hong. Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Peter J Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, pages 1–37, 2015.

Fengqiao Luo and Sanjay Mehrotra. Decomposition algorithm for distributionally robust optimization using Wasserstein metric. *arXiv preprint arXiv:1704.03920*, 2017.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. An inequality with applications to structured sparsity and multitask dictionary learning. In *COLT*, pages 440–460, 2014.

Sanjay Mehrotra and He Zhang. Models and algorithms for distributionally robust least squares problems. *Mathematical Programming*, 146(1-2):123–141, 2014.

Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subGaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(02):186–199, 1991.

Ioana Popescu. Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1):98–112, 2007.

Leonhard James Rogers. An extension of a certain theorem in inequalities. *Messenger of Math*, 17(2):145–150, 1888.

Peter Rousseeuw and Victor Yohai. Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer, 1984.

Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

Peter J Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297, 1985.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John Wiley & Sons, 2005.

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *arXiv preprint arXiv:1710.10016*, 2017.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press (to appear), 2017.

Li Wang, Michael D Gordon, and Ji Zhu. Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 690–700. IEEE, 2006.

Zizhuo Wang, Peter W Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, 2016.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(Jul):1485–1510, 2009.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and LASSO. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.

Wenzhuo Yang and Huan Xu. A unified robust regression model for LASSO-like algorithms. In *International Conference on Machine Learning*, pages 585–593, 2013.

Victor J Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, pages 642–656, 1987.

C Zhao and Y Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Available on optimization online*, 2015.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013.