2018 IEEE Conference on Decision and Control (CDC)
Miami Beach, FL, USA, Dec. 17-19, 2018

A Distributionally Robust Optimization Approach for Outlier Detection *

Ruidi Chen[1] and Ioannis Ch. Paschalidis[2]

***Abstract*— We consider the outlier detection problem in a linear regression setting. Outlying observations can be detected by large residuals but this approach is not robust to large outliers which tend to shift the residual function. Instead, we propose a new *Distributionally Robust Optimization (DRO)* method addressing this issue. The robust optimization problem reduces to solving a second-order cone programming problem. We prove several generalization guarantees for our solution under mild conditions. Extensive numerical experiments demonstrate that our approach outperforms Huber's robust regression approach.**

I. Introduction

Outlier detection has attracted a lot of attention in recent years due to its extensive use in a wide variety of applications. Recent examples include CT radiation overdose detection, abnormal traffic jam detection and computer intrusion detection. Many techniques have been developed for detecting outliers, including both direct and indirect procedures. The authors in [1] have performed an extensive simulation study to evaluate and compare numerous outlier detection methods that use linear regression. Indirect procedures rely on the residuals or weights from robust regression estimates. Different from ordinary least squares regression, robust regression is less influenced by outliers through downweighting or ignoring aberrant data points. Some commonly used robust estimators are presented in [2].

In addition to the traditional robust estimators, researchers have also been trying to build a connection between robust regression and robust optimization. [3] formulates robust linear regression with feature-wise disturbance as a minimization of the worst-case error norm and shows that this recovers the *LASSO* (*Least Absolute Shrinkage and Selection Operator*) as a special case. [4] considers noise in both features and responses and solves the corresponding robust optimization problem in polynomial-time. [5] studies a distributionally robust least squares problem which hedges the worst-case square norm of the error over a probabilistic ambiguity set of the disturbances.

* 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 MURI grant N00014-16-1-2832, by the NIH under grant 1UL1TR001430 to the Clinical & Translational Science Institute at Boston University, and by the Boston University Digital Health Initiative.

[1]Ruidi Chen is with Division of Systems Engineering, Boston University, Boston, MA 02446, USA. rchen15@bu.edu.

[2]Ioannis Ch. Paschalidis is with Dept. of Electrical and Computer Engineering, Division of Systems Engineering, and Dept. of Biomedical Engineering, Boston University, 8 St. Mary's St., Boston, MA 02215, USA. yannisp@bu.edu, http://sites.bu.edu/paschalidis/.

Our approach also adopts a *Distributionally Robust Optimization (DRO)* formulation in a linear regression setting but is different from existing methods in the literature in three important aspects. First, instead of minimizing the worst case ℓ_2 norm of the error, we use an ℓ_1 loss function which is less sensitive to outliers. Second, a Wasserstein probabilistic ambiguity set centered at the empirical distribution of the data is used, which is rich enough to contain the true data-generating distribution with high confidence. Moreover, we are able to control the conservativeness of our formulation through adjusting the radius of the Wasserstein set. This ambiguity set is easy to construct from data and yields a second-order cone programming problem which can be solved very efficiently. Third, our formulation has the flexibility of incorporating sparsity constraints, which makes it possible to combine robust regression with the LASSO and is particularly useful in practical applications.

The rest of the paper is organized as follows. In Sec. II we derive the Wasserstein DRO formulation in a linear regression framework. Sec. III establishes out-of-sample performance guarantees. The numerical experimental results are presented in Sec. IV. We conclude the paper in Sec. V.

Notational conventions: We use boldfaced lowercase letters to denote vectors, ordinary lowercase letters to denote scalars, 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 transpose, $\|\cdot\|$ for the ℓ_2 norm, and $\|\cdot\|_1$ for the ℓ_1 norm.

II. Wasserstein DRO Formulation

Consider a generic DRO problem expressed as follows:

$$\hat{J}_{DRO} := \inf_{\boldsymbol{\alpha}\in\mathcal{A}} \sup_{\mathbb{Q}\in\mathcal{B}} \mathbb{E}^{\mathbb{Q}}[h(\boldsymbol{\alpha}, \mathbf{z})],$$

where $\boldsymbol{\alpha}$ is the decision variable taking values in the set $\mathcal{A}$; $\mathbf{z}$ is a vector of uncertain parameters whose distribution $\mathbb{Q}$ is unknown and can only be observed through a finite set of samples; $\mathcal{B}$ is the probabilistic ambiguity set for $\mathbf{z}$; and $h(\cdot,\cdot)$ is the loss function we seek to minimize.

The most fundamental problem in DRO is how to choose a proper ambiguity set $\mathcal{B}$. There have been some works focusing on moment ambiguity sets, which contain all distributions that satisfy certain moment constraints, see [6, 5]. Another option is to define $\mathcal{B}$ as a ball of distributions using some probability distance functions. [7] adopts this approach using the Kantorovich distance while [8] considers the Kullback-Leibler divergence. Apparently, the distance function is a

key element in this kind of DRO model. In this paper, we use the Wasserstein metric as in [9] to construct such a ball ambiguity set. Specifically, $\mathcal{B}$ is defined as:

$$\mathcal{B} := \{\mathbb{Q} \in \mathcal{M}(\mathcal{Z}) : d_W(\mathbb{Q},\ \hat{\mathbb{P}}_N) \leq \epsilon\}, \tag{1}$$

where $\mathcal{Z}$ is the set of possible values for $\mathbf{z}$; $\mathcal{M}(\mathcal{Z})$ is the space of all probability distributions supported on $\mathcal{Z}$; $\hat{\mathbb{P}}_N$ is the discrete empirical distribution constructed based on N independently and identically distributed (i.i.d.) samples of $\mathbf{z}$; ϵ is a pre-specified radius of the Wasserstein ball; and $d_W(\mathbb{Q},\ \hat{\mathbb{P}}_N)$ is the Wasserstein distance between $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$ defined as:

$$d_W(\mathbb{Q},\ \hat{\mathbb{P}}_N) \triangleq \sup_{f\in\mathcal{L}}\Big(\int_{\mathcal{Z}} f(\mathbf{z})\ \mathbb{Q}(d\mathbf{z}) - \int_{\mathcal{Z}} f(\mathbf{z})\ \hat{\mathbb{P}}_N(d\mathbf{z})\Big), \tag{2}$$

with $\mathcal{L}$ being the space of all Lipschitz continuous functions satisfying $|f(\mathbf{z}_1) - f(\mathbf{z}_2)| \leq \|\mathbf{z}_1, \mathbf{z}_2\|,\ \forall\ \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$.

Compared to other distance functions, the Wasserstein metric is advantageous in two important aspects. First, the Wasserstein ambiguity set is rich enough to contain both discrete and continuous distributions [9], while the Kullback-Leibler ambiguity set centered at $\hat{\mathbb{P}}_N$ fails to include the continuous ones. Second, the measure concentration results in [10] ensure that the Wasserstein ambiguity set contains the true data-generating distribution with a high probability, under a light tail assumption. Moreover, as proven in [9] under the same assumption, any sequence of distributions in the Wasserstein set converges weakly to the true distribution as the sample size grows to infinity.

Now let us return to the outlier detection problem. Suppose we have N i.i.d. samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$, where y_i is the i-th response variable and $\mathbf{x}_i$ is a d-dimensional vector of features. Each sample is drawn with probability p from the outlying distribution $\mathbb{P}_{out}$ and with probability $1-p$ from the true distribution $\mathbb{P}$ (clean data). $\hat{\mathbb{P}}_N$ is the discrete uniform distribution over these N samples. The ambiguity set $\mathcal{B}$ is constructed as in (1). Our goal is to first obtain an accurate estimate of the regression coefficients determined by the clean data and then detect outliers based on this estimation. Consider an ℓ_1 loss function in the linear regression setting. Using $(\mathbf{x}, y)$ to denote the feature and response variables, our Wasserstein DRO problem is formulated as:

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{\mathbb{Q}\in\mathcal{B}} \mathbb{E}^{\mathbb{Q}}\big[|y - \mathbf{x}'\boldsymbol{\beta}|\big] = \inf_{\tilde{\boldsymbol{\beta}}\in\tilde{\mathcal{D}}} \sup_{\mathbb{Q}\in\mathcal{B}} \mathbb{E}^{\mathbb{Q}}\big[|\mathbf{z}'\tilde{\boldsymbol{\beta}}|\big], \tag{3}$$

where $\boldsymbol{\beta}$ is the regression coefficient vector that belongs to some set $\mathcal{D}$; $\tilde{\boldsymbol{\beta}} \triangleq (-\boldsymbol{\beta}, 1)$; $\mathbf{z} \triangleq (\mathbf{x}, y)$; $\mathbb{Q}$ is the probability distribution of $\mathbf{z}$ belonging to some set $\mathcal{B}$ as defined in (1); and $\tilde{\mathcal{D}} = \{\tilde{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathcal{D}\}$. $\mathcal{D}$ could be $\mathbb{R}^d$, or $\mathcal{D} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \leq l\}$ for some l, if we wish to induce sparsity.

To convert (3) into a tractable optimization problem, we apply a key result in [9] [Theorem 6.3] which states that when the set $\mathcal{Z}$ is closed and convex, for any $\epsilon \geq 0$,

$$\sup_{\mathbb{Q}\in\mathcal{B}} \mathbb{E}^{\mathbb{Q}}[|\mathbf{z}'\tilde{\boldsymbol{\beta}}|] \leq \kappa\epsilon + \frac{1}{N}\sum_{i=1}^{N} |\mathbf{z}_i'\tilde{\boldsymbol{\beta}}|, \tag{4}$$

where $\mathbf{z}_i$ is the ith sample of $\mathbf{z}$; ϵ is defined as in (1); $\kappa = \sup\{\|\boldsymbol{\theta}\|_* : h^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}) < \infty\}$; $\|\cdot\|_*$ is the dual norm of the norm used in characterizing the Lipschitz continuity of functions used in (2), which is defined as $\|\boldsymbol{\theta}\|_* \triangleq \sup_{\|\mathbf{z}\|\leq 1} \boldsymbol{\theta}'\mathbf{z}$; and $h^*(\cdot, \cdot)$ is the conjugate function of the loss $h(\tilde{\boldsymbol{\beta}}, \mathbf{z}) \triangleq |\mathbf{z}'\tilde{\boldsymbol{\beta}}|$ defined as $h^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}) \triangleq \sup_{\mathbf{z}}\{\boldsymbol{\theta}'\mathbf{z} - h(\tilde{\boldsymbol{\beta}}, \mathbf{z})\}$. Note that κ is a function of $\tilde{\boldsymbol{\beta}}$. Through (4), we can relax problem (3) by minimizing the right hand side of (4) instead of the worst-case expected loss. Moreover, as shown in [9], (4) becomes an equality when $\mathcal{Z} = \mathbb{R}^{d+1}$.

What remains is then how to compute κ. In the following theorem we show that κ is equal to the ℓ_2 norm of $\tilde{\boldsymbol{\beta}}$.

Theorem II.1. *Define* $\kappa = \sup\{\|\boldsymbol{\theta}\|_* : h^*(\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}) < \infty\}$. *When the loss function* $h(\tilde{\boldsymbol{\beta}}, \mathbf{z}) = |\mathbf{z}'\tilde{\boldsymbol{\beta}}|$, $\kappa = \|\tilde{\boldsymbol{\beta}}\|$.

Proof. First rewrite κ as:

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

Consider now the two optimization problems A and B.

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

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

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

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

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

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:

$$\exists\ r_A \leq 0, \quad \text{s.t.} \quad \tilde{\boldsymbol{\beta}} r_A = \boldsymbol{\theta} - \tilde{\boldsymbol{\beta}}, \tag{5}$$

$$\exists\ r_B \geq 0, \quad \text{s.t.} \quad \tilde{\boldsymbol{\beta}} r_B = \boldsymbol{\theta} + \tilde{\boldsymbol{\beta}}. \tag{6}$$

For all i with $\tilde{\beta}_i \leq 0$, (5) implies $\theta_i - \tilde{\beta}_i \geq 0$ and (6) implies $\theta_i \leq -\tilde{\beta}_i$. On the other hand, for all j with $\tilde{\beta}_j \geq 0$, (5) and (6) imply $-\tilde{\beta}_j \leq \theta_j \leq \tilde{\beta}_j$. It is not hard to conclude that:

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

In a finite dimensional space, the dual norm of the ℓ_2 norm is the ℓ_2 norm. It follows,

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

□

Now we are ready to reformulate the Wasserstein DRO problem (3). When $\mathcal{Z} = \mathbb{R}^{d+1}$, (3) is equivalent to the following optimization problem:

$$\inf_{\tilde{\boldsymbol{\beta}} \in \tilde{\mathcal{D}}} \|\tilde{\boldsymbol{\beta}}\|\epsilon + \frac{1}{N}\sum_{i=1}^{N} |\mathbf{z}_i'\tilde{\boldsymbol{\beta}}|. \tag{7}$$

Remark 2.1 The ℓ_2-norm regularizer in (7) is related to the *growth rate* of the ℓ_1 loss function [11]. The reduction to (7) can be attributed to the structure of the Wasserstein metric. It is possible to generalize such a relaxation to other loss functions with a bounded and fixed growth rate.

Remark 2.2 The parameter ϵ controls the conservativeness of our formulation, whose selection depends on both the sample size and the confidence that the Wasserstein ball contains the true distribution (see Eq. (8) in [9]). Roughly speaking, when the sample size is large enough, for a fixed confidence level, ϵ is inversely proportional to $N^{1/(d+1)}$.

Remark 2.3 (7) is a second-order cone programming problem which can be solved to optimality very efficiently. Although it is the same with the ℓ_2 regularized *Least Absolute Deviation (LAD)* [12, 13], there exist two essential differences that manifest the value and novelty of (7). First, in the LAD literature, the regularization term (usually an ℓ_1-regularizer) is introduced to resolve the issue of ill-conditioned design matrix and to recover a sparse coefficient vector. By contrast, the ℓ_2-regularizer in (7) is a control over the amount of ambiguity in the data, whose existence is not decided by the sparsity of the coefficient or the correlation among predictors. More importantly, (7) is theoretically rooted and derived from DRO, of which the ℓ_2-regularizer is an indispensable ingredient that reveals the reliability of the contaminated samples. Second, the regularization coefficient in (7) is the radius of the Wasserstein ball, which offers an intuitive interpretation and provides guidance on how to set it. This connection is not present in the regularized LAD literature, which starts from the regularized problem rather than deriving it from a more fundamental DRO formulation.

III. Out-of-Sample Performance Guarantees

In this section we will prove that the performance of the solution to (7) on new, future data is similar to its performance on the observed samples. The proof is mainly based on a concept called Rademacher complexity [14], which is a measurement of the complexity of a class of functions. [14][Theorem 8] bounds the expected loss in terms of the Rademacher complexity. We apply this result to our specific loss function and derive the performance guarantees. Several mild assumptions are needed in this section.

Assumption A. *The uncertainty parameter* $\mathbf{z}$ *falls within a ball of radius* R *almost surely.*

Assumption B. *The* ℓ_2 *norm of* $\tilde{\boldsymbol{\beta}}$ *is bounded above within the feasible region by* $\bar{B}$.

We would like to bound the expected loss under these two assumptions. Remember that [14] provides such a bound in the form of Rademacher complexity. A natural idea is then to first bound the Rademacher complexity. The following two lemmata achieve this goal and are helpful in establishing the performance guarantees.

Lemma III.1. *For every feasible* $\tilde{\boldsymbol{\beta}}$*, it follows*

$$|\mathbf{z}'\tilde{\boldsymbol{\beta}}| \leq \bar{B}R, \quad \textit{almost surely}.$$

The proof of Lemma III.1 uses the Cauchy-Schwarz inequality. Now consider the class of functions

$$\mathcal{F} = \{\mathbf{z} \mapsto h(\tilde{\boldsymbol{\beta}}, \mathbf{z}) : h(\tilde{\boldsymbol{\beta}}, \mathbf{z}) = |\mathbf{z}'\tilde{\boldsymbol{\beta}}|, \quad \tilde{\boldsymbol{\beta}} \in \tilde{\mathcal{D}}\}.$$

We next establish a result for the empirical Rademacher complexity of this class of functions, denoted by $\mathcal{R}_N(\mathcal{F})$ and defined as:

$$\mathcal{R}_N(\mathcal{F}) \triangleq \mathbb{E}\left[\sup_{h\in\mathcal{F}} \frac{2}{N}\left|\sum_{i=1}^{N}\sigma_i h(\tilde{\boldsymbol{\beta}}, \mathbf{z}_i)\right| \middle| \mathbf{z}_1, \cdots, \mathbf{z}_N\right],$$

where $\sigma_1, \ldots, \sigma_N$ are i.i.d. uniform random variables on $\{1, -1\}$. The proof is similar to the proof of Lemma 3 in [15].

Lemma III.2. *Under Assumptions A and B, it holds that,*

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

Proof. By Lemma III.1, we have:

$$\begin{aligned}
\mathcal{R}_N(\mathcal{F}) &\leq \frac{2\bar{B}R}{N}\mathbb{E}\left[\left|\sum_{i=1}^{N}\sigma_i\right|\right] \\
&\leq \frac{2\bar{B}R}{N}\mathbb{E}\left[\sqrt{\sum_{i=1}^{N}\sigma_i^2}\right] \\
&= \frac{2\bar{B}R}{\sqrt{N}}.
\end{aligned}$$

□

We are now ready to state the main result guaranteeing out-of-sample performance in the following theorem. Let $\tilde{\boldsymbol{\beta}}^* = (-\boldsymbol{\beta}^*, 1)$ be an optimal solution to (7), obtained using the samples $\mathbf{z}_i$, $i = 1, \ldots, N$. Suppose we draw a new i.i.d. sample $\mathbf{z} = (\mathbf{x}, y)$. We establish bounds on the error $|y - \mathbf{x}'\boldsymbol{\beta}^*|$.

Theorem III.3. *Under Assumptions A and B, for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$ *with respect to the sampling,*

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

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

$$\begin{aligned}
&\mathbb{P}\left(|\mathbf{z}'\tilde{\boldsymbol{\beta}}^*| \geq \frac{1}{N}\sum_{i=1}^{N}|\mathbf{z}_i'\tilde{\boldsymbol{\beta}}^*| + \zeta\right) \\
&\qquad \leq \frac{\frac{1}{N}\sum_{i=1}^{N}|\mathbf{z}_i'\tilde{\boldsymbol{\beta}}^*| + \frac{2\bar{B}R}{\sqrt{N}} + \bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N}|\mathbf{z}_i'\tilde{\boldsymbol{\beta}}^*| + \zeta}.
\end{aligned} \tag{9}$$

Proof. Applying Lemma 1 from [15], which is a specialization of Theorem 8 in [14], using $h(\tilde{\boldsymbol{\beta}}, \mathbf{z}) = |\mathbf{z}'\tilde{\boldsymbol{\beta}}|$ and the bound on $\mathcal{R}_N(\mathcal{F})$ presented in Lemma III.1, we are able to prove the bound on the expected loss. Eq. (9) can be proved using Markov's inequality. □

Remark 3.1 There are two probability measures in the statement of Theorem III.3. One is related to the new data $\mathbf{z}$, while the other is related to the samples $\mathbf{z}_1, \cdots, \mathbf{z}_N$. The expectation in (8) (and the probability in (9)) is taken w.r.t the new data $\mathbf{z}$. For a fixed set of samples, (8) (and (9)) holds with probability at least $1-\delta$ w.r.t. the measure of samples. Theorem III.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 are proportional to $1/\sqrt{N}$. Since a term of (7) minimizes the sample average loss, it is ensured that our estimator achieves a good out-of-sample performance.

Next we will show two corollaries that provide a guidance on how many samples are needed to achieve satisfactory performance.

Corollary III.4. *For a fixed confidence level δ and some threshold parameter $\tau \geq 0$, to guarantee that the percentage difference between the expected absolute loss and the sample average loss is less than τ, that is,*

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

the sample size N must satisfy

$$N \geq \left[\frac{2(1+\sqrt{2\log(2/\delta)}\ )}{\tau}\right]^2. \tag{10}$$

Corollary III.5. *For a fixed confidence level δ, some $\tau \in (0,1)$ and $\gamma \geq 0$, to guarantee that*

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

the sample size N must satisfy

$$N \geq \left[\frac{2(1+\sqrt{2\log(2/\delta)}\ )}{\tau\cdot\gamma+\tau-1}\right]^2, \tag{11}$$

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

IV. Simulation Experiments

In this section we apply our Wasserstein DRO approach to a number of synthetic datasets and compare its performance with a traditional robust regression method (M-estimation) due to Huber [16, 17]. The experimental scenarios are designed according to [1]. Specifically, we consider interior $\mathbf{x}$-space outliers, which are observations that are abnormal only in the y direction, but have $\mathbf{x}$ values that are within the normal range. Our approach will be tested in three different scenarios differentiated by the location of outliers.

The datasets are constructed based on a linear regression model. Specifically, suppose there are K predictors; the response y for clean observations is obtained through: $y = \beta_0+\beta_1 x_1+\cdots+\beta_K x_K+\eta$, where η is the noise term normally distributed with mean 0 and variance σ_η^2. The response values for outlying observations are placed at a distance δ_R off the regression plane: $y = \beta_0+\beta_1 x_1+\cdots+\beta_K x_K+\delta_R$.

We set $\beta_0 = 0.3, \beta_1 = \cdots = \beta_K = 0.5$ throughout this section. For clean observations, all features $x_1, \ldots, x_K$ come from a normal distribution with mean 7.5 and standard deviation 4.0. The experiments are conducted in different factor settings, where the factors considered are: *percentage of outliers p:* 20%, 30%; *outlying distance δ_R:* $3\sigma_\eta, 4\sigma_\eta, 5\sigma_\eta$; *the number of regressors K:* 6, 30.

For $K=6$, $\sigma_\eta = 1$; for $K=30$, $\sigma_\eta = 5$. We compare our approach with the commonly used robust regression method called M-estimation with four cost functions – Tukey's Biweight [18, 19], Huber [16, 17], Talwar [20], and Fair [21]. The performance metrics we use are: (i) $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$, where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient and $\boldsymbol{\beta}$ is the true coefficient determined by the clean data; and (ii) the *Receiver Operating Characteristic (ROC)* curve which plots the true positive rate against the false positive rate.

In all experiments, the size of the training dataset is $N=60$ for $K=6$ and $N=300$ for $K=30$, including both clean and outlying observations. The size of the test dataset is $M=36$ for $K=6$ and $M=180$ for $K=30$. We run 500 replications and take the average of the performance metrics. The radius ϵ is chosen to be inversely proportional to $N^{1/(d+1)}$ by a factor decided based on the *Area Under the ROC Curve (AUC)*. Before providing the details of the experimental results, we summarize our major findings below.

1) In terms of AUC, all approaches have better performance when p is lower, δ_R is larger, and K is smaller.
2) In terms of the ROC curve, our approach performs better than M-estimation with the difference being larger when the percentage of outliers p is lower and the outlying distance δ_R is larger, especially when the number of features K is larger.
3) Increasing the size of the training dataset N could greatly improve the performance of our approach.

For all scenarios, we will show ROC curves only for the most challenging factor setting due to limited space. According to the findings stated above, the most challenging factor combination is $p=30\%, \delta_R = 3\sigma_\eta$. The scalar performance metrics for $K=30$ are properly summarized in Tables I–III.

A. Randomly Scattered Outliers

In this subsection, we consider outliers that are randomly scattered in the interior of the $\mathbf{x}$-space. The feature variables for outlying observations have the same distribution as that of the clean data, but the response values are placed at a distance δ_R off the regression plane.

The Wasserstein radius ϵ is set to be proportional to $1/N^{1/(d+1)}$, where the constant factor is chosen to maximize the out-of-sample AUC. In Fig. 1 we plot the out-of-sample AUC as the radius is changed. 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. It is clear that for $N = 60$, the optimal $\epsilon = 15.10$. For $N = 300$, we set $\epsilon = 21.08$.

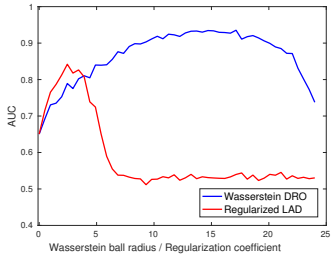


Fig. 1: Out-of-sample AUC v.s. Wasserstein ball radius.

It is worth mentioning that our formulation (7), as well as robust regression, only generates an estimated regression coefficient. The identification of outliers is based on the residual and estimated standard deviation of the noise. Specifically,

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

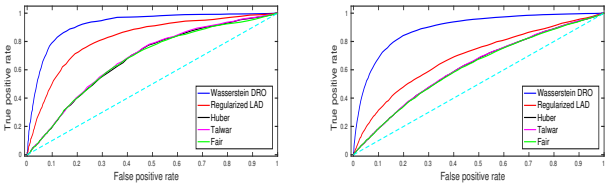
where $\hat{\sigma}$ is the standard deviation of residuals in the entire training set. ROC curves are obtained through adjusting the threshold values. We note that for all tables, the numbers in parentheses are the results from ℓ_1-regularized LAD, M-estimation with Huber, Talwar, and Fair cost functions, respectively, while the number outside parentheses is the result from our approach.

TABLE I: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ for randomly scattered outliers with 30 features.

Outlying distance	Percentage of outliers	
	20%	30%
3	0.81 (1.63, 2.06, 2.04, 2.10)	0.87 (1.81, 2.18, 2.16, 2.23)
4	0.82 (1.80, 2.45, 2.42, 2.48)	0.90 (2.14, 2.76, 2.73, 2.82)
5	0.82 (2.06, 3.06, 3.05, 3.06)	0.90 (2.46, 3.41, 3.38, 3.50)

From the performance comparison results we see that our approach consistently outperforms M-estimation with all four cost functions. The Wasserstein DRO approach achieves higher AUC and smaller $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ in all factor settings. Moreover, even in the most challenging cases, its ROC curve still lies well above the ROC curves of M-estimation, among which Fair's cost slightly outperforms the other three cost functions.

The superiority of our approach could be attributed to the distributional robustness, which means we hedge against a family of plausible distributions, including the true distribution with high confidence, when minimizing the cost. 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. 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) $K = 6$ (b) $K = 30$

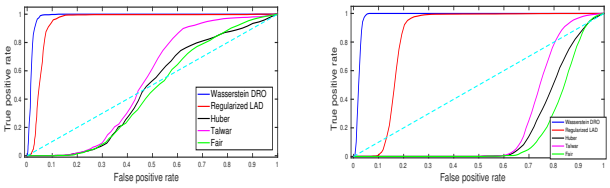
Fig. 2: ROC curves for randomly scattered outliers, where $p = 30\%, \delta_R = 3\sigma_\eta$.

B. Outliers in A Cloud at the Centroid of the x-Space

In this subsection we consider outliers that are gathered in a cloud at the centroid of the $\mathbf{x}$-space. The features for outlying observations are uniformly distributed on the interval $[7.375, 7.625]$ since clean observations have features centered around 7.5. The response values are still at a δ_R distance off the regression plane.

TABLE II: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ for outliers in a centroid cloud with 30 features.

Outlying distance	Percentage of outliers	
	20%	30%
3	0.70 (1.24, 2.96, 2.83, 3.08)	0.69 (1.45, 5.46, 4.87, 5.55)
4	0.70 (1.24, 3.71, 3.47, 3.87)	0.69 (1.50, 7.22, 6.44, 7.36)
5	0.70 (1.22, 4.31, 4.13, 4.49)	0.69 (1.50, 8.77, 7.76, 8.95)



(a) $K = 6$ (b) $K = 30$

Fig. 3: ROC curves for outliers in a centroid cloud, where $p = 30\%, \delta_R = 3\sigma_\eta$.

C. Outliers in A Cloud Randomly Placed in the Interior x-Space

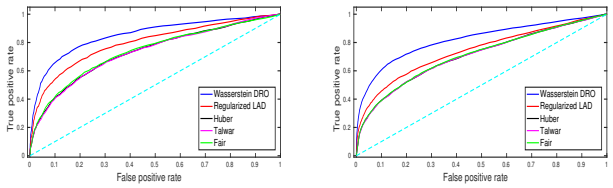
We now consider outliers concentrated in a cloud that is randomly placed in the interior of the $\mathbf{x}$-space. The features for outlying observations are uniformly distributed on $(u - 0.125, u + 0.125)$, where u is a uniform random variable on $(7.5 - 3 \times 4, 7.5 + 3 \times 4)$. The response values are at a δ_R distance off the regression plane.

V. Conclusions

We considered the problem of estimating the true regression plane when the dataset is possibly contaminated with outliers and provided a novel formulation using a distributionally robust optimization approach with a Wasserstein ambiguity set. We established rigorous guarantees on the ability of our solution to perform well out-of-sample. A host

TABLE III: $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$ for outliers in a randomly placed cloud with 30 features.

Outlying distance	Percentage of outliers	
	20%	30%
3	0.88 (1.36, 1.77, 1.77, 1.77)	0.92 (1.42, 1.80, 1.78, 1.80)
4	0.94 (1.54, 2.09, 2.09, 2.07)	1.00 (1.62, 2.23, 2.22, 2.21)
5	0.97 (1.72, 2.56, 2.59, 2.52)	1.08 (1.87, 2.73, 2.73, 2.72)



(a) $K = 6$ (b) $K = 30$

Fig. 4: ROC curves for outliers in a randomly placed cloud, where $p = 30\%, \delta_R = 3\sigma_\eta$.

of simulation experimental results suggest that our approach outperforms the commonly used robust regression method called M-estimation in terms of both outlier detection probabilities and the accuracy of the estimated coefficients.

REFERENCES

[1] J. W. Wisnowski, D. C. Montgomery, and J. R. Simpson, "A comparative analysis of multiple outlier detection procedures in the linear regression model," *Computational Statistics & Data Analysis*, vol. 36, no. 3, pp. 351–382, 2001.

[2] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. John Wiley & Sons, 2005.

[3] H. Xu, C. Caramanis, and S. Mannor, "Robust regression and lasso," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3561–3574, 2010.

[4] L. El Ghaoui and H. Lebret, "Robust solutions to least-squares problems with uncertain data," *SIAM Journal on Matrix Analysis and Applications*, vol. 18, no. 4, pp. 1035–1064, 1997.

[5] S. Mehrotra and H. Zhang, "Models and algorithms for distributionally robust least squares problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 123–141, 2014.

[6] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Operations Research*, vol. 62, no. 6, pp. 1358–1376, 2014.

[7] G. Pflug and D. Wozabal, "Ambiguity in portfolio selection," *Quantitative Finance*, vol. 7, no. 4, pp. 435–442, 2007.

[8] Z. Hu and L. J. Hong, "Kullback-Leibler divergence constrained distributionally robust optimization," *Available at Optimization Online*, 2013.

[9] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations," *Available at Optimization Online*, 2015.

[10] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3-4, pp. 707–738, 2015.

[11] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[12] D. Pollard, "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, vol. 7, no. 02, pp. 186–199, 1991.

[13] L. Wang, M. D. Gordon, and J. Zhu, "Regularized least absolute deviations regression and an efficient algorithm for parameter tuning," in *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, 2006, pp. 690–700.

[14] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[15] D. Bertsimas, V. Gupta, and I. C. Paschalidis, "Data-driven estimation in equilibrium using inverse optimization," *Mathematical Programming*, vol. 153, no. 2, pp. 595–633, 2015.

[16] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[17] ——, "Robust regression: asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, vol. 1, no. 5, pp. 799–821, 1973.

[18] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics*, vol. 16, no. 2, pp. 147–185, 1974.

[19] S. Morgenthaler, "Fitting redescending m-estimators in regression," *Robust Regression: Analysis and Applications*, vol. 108, p. 105, 1989.

[20] M. J. Hinich and P. P. Talwar, "A simple method for robust regression," *Journal of the American Statistical Association*, vol. 70, no. 349, pp. 113–119, 1975.

[21] R. C. Fair, "On the robust estimation of econometric models," in *Annals of Economic and Social Measurement, Volume 3, number 4*. NBER, 1974, pp. 667–677.