The version of record is available at: http://dx.doi.org/10.1561/2400000026

Distributionally Robust Learning

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

[1]*Boston University, USA; rchen15@bu.edu*
[2]*Boston University, USA; yannisp@bu.edu*

ABSTRACT

This monograph develops a comprehensive statistical learning framework that is robust to (distributional) perturbations in the data using *Distributionally Robust Optimization (DRO)* under the Wasserstein metric. Beginning with fundamental properties of the Wasserstein metric and the DRO formulation, we explore duality to arrive at tractable formulations and develop finite-sample, as well as asymptotic, performance guarantees. We consider a series of learning problems, including (i) distributionally robust linear regression; (ii) distributionally robust regression with group structure in the predictors; (iii) distributionally robust multi-output regression and multiclass classification, (iv) optimal decision making that combines distributionally robust regression with nearest-neighbor estimation; (v) distributionally robust semi-supervised learning, and (vi) distributionally robust reinforcement learning. A tractable DRO relaxation for each problem is being derived, establishing a connection between robustness and regularization, and obtaining bounds on the prediction and estimation errors of the solution. Beyond theory, we include numerical experiments and case studies using synthetic and real data. The real data experiments are all associated with various health informatics problems, an application area which provided the initial impetus for this work.

Ruidi Chen and Ioannis Ch. Paschalidis (2020), "Distributionally Robust Learning", Foundations and Trends® in Optimization: Vol. 4, No. 1–2, pp 1–243. DOI: 10.1561/2400000026.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

1

Introduction

A central problem in *machine learning* is to learn from data ("big" or "small") how to predict outcomes of interest. Outcomes can be *binary* or *discrete*, such as an event or a category, or *continuous*, e.g., a real value. In either case, we have access to a number N of examples from which we can learn; each example is associated with a potentially large number p of *predictor* variables and the "ground truth" discrete or continuous outcome. This form of learning is called *supervised*, because it relies on the existence of known examples associating predictor variables with the outcome. In the case of a binary/discrete outcome the problem is referred to as *classification*, while for continuous outcomes we use the term *regression*.

There are many methods to solve such supervised learning problems, from ordinary (linear) least squares regression, to logistic regression, Classification And Regression Trees (CART) [1], ensembles of decision trees [2], [3], to modern deep learning models [4]. Whereas the nonlinear models (random forests, gradient boosted trees, and deep learning) perform very well in many specific applications, they have two key drawbacks: (i) they produce predictive models that lack *interpretability* and (ii) they are hard to analyze and do not give rise to rigorous

The version of record is available at: http://dx.doi.org/10.1561/2400000026

mathematical results characterizing their performance and important properties. In this monograph, we will mainly focus on the more classical linear models, allowing for some nonlinear extensions.

Clearly, there is a plethora of application areas where such models have been developed and used. A common thread throughout this monograph is formed by applications in medicine and health care, broadly characterized by the term *predictive health analytics*. While in principle these applications are not substantially different from other domains, they have important salient features that need to be considered. These include:

1. *Presence of outliers.* Medical data often contain outliers, which may be caused by medical errors, erroneous or missing data, equipment and lab configuration errors, or even different interpretation/use of a variable by different physicians who enter the data.

2. *Risk of "overfitting" from too many variables.* For any individual and any outcome we wish to predict, using all predictor variables may lead to *overfitting* and large generalization errors (out-of-sample). The common practice is to seek *sparse* models, using the fewest variables possible without significantly compromising accuracy. In some settings, especially when genetic information is included in the predictors, the number of predictors can exceed the training sample size, further stressing the need for *sparsity*. Sparse regression models originated in the seminar work on the *Least Absolute Shrinkage and Selection Operator*, better known under the acronym LASSO [5].

3. *Lack of linearity.* In some applications, the linearity of regression or logistic regression may not fully capture the relationship between predictors and outcome. While kernel methods [6] can be used to employ linear models in developing nonlinear predictors, other choices include combining linear models with nearest neighbor ideas to essentially develop *piecewise linear* models.

To formulate the learning problems of interest more concretely, let $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ denote a column vector with the predictors and

The version of record is available at: http://dx.doi.org/10.1561/2400000026

let $y \in \mathbb{R}$ be the outcome or response. In the classification problem, we have $y \in \{-1, +1\}$. We are given training data $(\mathbf{x}_i, y_i)$, $i \in [\![N]\!]$, where $[\![N]\!] \triangleq 1, \ldots, N$, from which we want to "learn" a function $f(\cdot)$ so that $f(\mathbf{x}_i) = y_i$ for most i. Further, we want $f(\cdot)$ to generalize well to new samples (i.e., to have good out-of-sample performance).

In the regression problem, we view the $\mathbf{x}_i$'s as independent variables (predictor vectors) and y_i as the real-valued dependent variable. We still want to determine a function $f(\mathbf{x})$ that predicts y. In linear regression, $f(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$, where $\boldsymbol{\beta}$ is a coefficient vector, prime denotes transpose, and we assume one of the elements of $\mathbf{x}$ is equal to one with the corresponding coefficient being the *intercept* (of the regression function at zero). Both classification and regression problems can be formulated as:

$$\min_{\boldsymbol{\beta}} \mathbb{E}^{\mathbb{P}^*}[h_{\boldsymbol{\beta}}(\mathbf{x}, y)], \tag{1.1}$$

where $\mathbb{P}^*$ is the probability distribution of $(\mathbf{x}, y)$, $\mathbb{E}^{\mathbb{P}^*}$ stands for the expectation under $\mathbb{P}^*$, and $h_{\boldsymbol{\beta}}(\mathbf{x}, y)$ is a *loss* function penalizing differences between $f(\mathbf{x})$ and y. This formulation is known as *expected risk minimization*. *Ordinary Least Squares (OLS)* uses a squared loss $h_{\boldsymbol{\beta}}(\mathbf{x}, y) = (f(\mathbf{x}) - y)^2$ while logistic regression uses the *logloss* function $h_{\boldsymbol{\beta}}(\mathbf{x}, y) = \log(1 + \exp\{-yf(\mathbf{x})\})$. Since $\mathbb{P}^*$ is typically unknown, a common practice is to approximate it using the *empirical distribution* $\hat{\mathbb{P}}_N$ which assigns equal probability to each training sample, leading to the following *empirical risk minimization* formulation:

$$\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} h_{\boldsymbol{\beta}}(\mathbf{x}_i, y_i).$$

One of the well known issues of OLS regression is that the regression function can be particularly sensitive to outliers. To illustrate this with a simple example, consider a case of regression with a single predictor; see Figure 1.1. Points in the training set are shown as blue dots. Suppose we include in the training set some outliers depicted as magenta stars. OLS regression results in the black line. Notice how much the slope of this line has shifted away from the blue dots to accommodate the outliers. This skews future predictions but also our ability to identify new outlying observations. Several approaches have been introduced to address this issue [7], [8] and we discuss them in more detail in Section 4.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

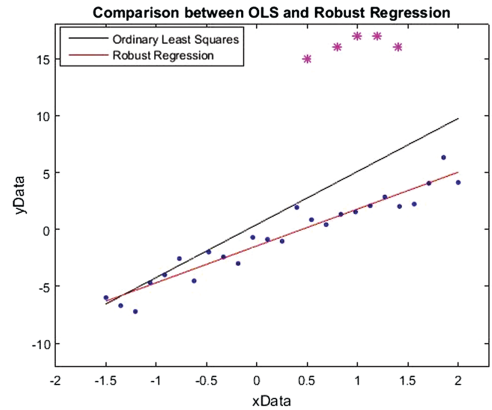


Figure 1.1: Regression example.

The main focus of this monograph is to develop *robust learning* methods for a variety of learning problems. To introduce robustness into the generic problem, we will use ideas from *robust optimization* and formulate a robust version of the expected risk minimization Problem (1.1). We will further focus on *distributional robustness*. The problems we will formulate are min-max versions of Problem (1.1) where one minimizes a worst case estimate of the loss over some appropriately defined ambiguity set. Such min-max formulations have a long history, going back to the origins of game theory [9], where one can view the problem as a game between an adversary who may affect the training set and the optimizer who responds to the worst-case selection by the adversary. They also have strong connections with $\mathcal{H}_\infty$ and robust control theory [10], [11].

To avoid being overly broad, we will restrict our attention to the intersection of statistical learning and *Distributionally Robust Optimization (DRO)* under the Wasserstein metric [12]–[14]. Even this more narrow area has generated a lot of interest and recent work. While we will cover several aspects, we will not cover a number of topics, including:

- the integration of DRO with different optimization schemes, e.g., inverse optimization [15], polynomial optimization [16], multistage optimization [17], [18], and chance-constrained optimization [19], [20];

The version of record is available at: http://dx.doi.org/10.1561/2400000026

- the application of DRO to stochastic control problems, see, e.g., [21]–[23], and statistical hypothesis testing [24];
- the combination of DRO with general estimation techniques, see, e.g., [25] for distributionally robust Minimum Mean Square Error Estimation, and [26] for distributionally robust Maximum Likelihood Estimation.

Most of the learning problems we consider, except for Section 8.2, are static *single-period* problems where the data are assumed to be independently and identically distributed. For extensions of DRO to a dynamic setting where the data come in a sequential manner, we refer to [27] for a distributionally robust Kalman filter model [23], [28], and [29] for robust dynamic programming, and [30] for a distributionally robust online adaptive algorithm.

In this monograph, we focus mainly on linear predictive models, with the exception of Section 7, where the non-linearity is captured by a non-parametric *K-Nearest Neighbors (K-NN)* model. For extensions of robust optimization to non-linear settings, we refer to [31] for robust kernel methods, [32] for distributionally robust graphical models, and [33] for distributionally robust deep neural networks.

In the remainder of this Introduction, we will present a brief outline of robust optimization in Section 1.1 and distributionally robust optimization in Section 1.2. In Section 1.3 we provide an outline of the topics covered in the rest of the monograph. Section 1.4 summarizes our notational conventions and Section 1.5 collects all abbreviations we will use.

1.1 Robust Optimization

Robust optimization [34], [35] provides a way of modeling uncertainty in the data without the use of probability distributions. It restricts data perturbations to be within a deterministic uncertainty set, and seeks a solution that is optimal for the worst-case realization of this uncertainty. Consider a general optimization problem:

$$\min_{\beta} \; h_{\beta}(\mathbf{z}), \tag{1.2}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where $\boldsymbol{\beta}$ is a vector of decision variables, $\mathbf{z}$ is a vector of given parameters, and h is a real-valued function. Assuming that the values of $\mathbf{z}$ lie within some uncertainty set $\mathcal{Z}$, a robust counterpart of Problem (1.2) can be written in the following form:

$$\min_{\boldsymbol{\beta}} \max_{\mathbf{z}\in\mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{z}). \tag{1.3}$$

Problem (1.3) is computationally tractable for many classes of uncertainty sets $\mathcal{Z}$. For a detailed overview of robust optimization we refer to [34]–[36].

There has been an increasing interest in using robust optimization to develop machine learning algorithms that are immunized against data perturbations; see, for example, [37]–[44] for classification methods. [41] considered both feature uncertainties:

$$\mathcal{Z}_{\mathbf{x}} \triangleq \{\boldsymbol{\Delta}\mathbf{X} \in \mathbb{R}^{N\times p}\colon \|\boldsymbol{\Delta}\mathbf{x}_i\|_q \le \rho, i \in [\![N]\!]\},$$

where $\boldsymbol{\Delta}\mathbf{X}$ can be viewed as a feature perturbation matrix on N samples with p features, $\|\cdot\|_q$ is the ℓ_q norm, and $\boldsymbol{\Delta}\mathbf{x}_i \in \mathbb{R}^p, i \in [\![N]\!]$, are the rows of $\boldsymbol{\Delta}\mathbf{X}$, as well as label uncertainties:

$$\mathcal{Z}_y \triangleq \left\{\boldsymbol{\Delta}\mathbf{y} \in \{0,1\}^N\colon \sum_{i=1}^{N} \Delta y_i \le \Gamma\right\},$$

where $\Delta y_i \in \{0,1\}$, with 1 indicating that the label was incorrect and has in fact been flipped, and 0 otherwise, and Γ is an integer-valued parameter controlling the number of data points that are allowed to be mislabeled. They solved various robust classification models under these uncertainty sets. As an example, the robust Support Vector Machine (SVM) [45] problem was formulated as:

$$\min_{\mathbf{w},b} \max_{\boldsymbol{\Delta}\mathbf{y}\in\mathcal{Z}_y} \max_{\boldsymbol{\Delta}\mathbf{X}\in\mathcal{Z}_{\mathbf{x}}} \sum_{i=1}^{N} \max\{1 - y_i(1-2\Delta y_i)(\mathbf{w}'(\mathbf{x}_i + \boldsymbol{\Delta}\mathbf{x}_i) - b), 0\}.$$

[39] studied a robust linear regression problem with feature-wise disturbance:

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta}\mathbf{X}\in\mathcal{Z}_{\mathbf{x}}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta}\mathbf{X})\boldsymbol{\beta}\|_2,$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients, and the uncertainty set

$$\mathcal{Z}_{\mathbf{x}} \triangleq \{\boldsymbol{\Delta}\mathbf{X} \in \mathbb{R}^{N\times p}\colon \|\boldsymbol{\Delta}\tilde{\mathbf{x}}_i\|_2 \le c_i,\ i \in [\![p]\!]\},$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where $\mathbf{\Delta}\tilde{\mathbf{x}}_i \in \mathbb{R}^N, i \in [\![p]\!]$, are the columns of $\mathbf{\Delta X}$. They showed that such a robust regression problem is equivalent to the following ℓ_1-norm regularized regression problem:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \sum_{i=1}^{p} c_i|\beta_i|.$$

1.2 Distributionally Robust Optimization

Different from robust optimization, *Distributionally Robust Optimization (DRO)* treats the data uncertainty in a probabilistic way. It minimizes a worst-case expected loss function over a probabilistic ambiguity set that is constructed from the observed samples and characterized by certain known properties of the true data-generating distribution. DRO has been an active area of research in recent years, due to its probabilistic interpretation of the uncertain data, tractability when assembled with certain metrics, and extraordinary performance observed on numerical examples, see, for example, [12]–[14], [46], [47]. DRO can be interpreted in two related ways: it refers to (i) a robust optimization problem where a worst-case loss function is being hedged against; or, alternatively, (ii) a stochastic optimization problem where the expectation of the loss function with respect to the probabilistic uncertainty of the data is being minimized. Figure 1.2 provides a schematic comparison of various optimization frameworks.

To formulate a DRO version of the expected risk minimization problem (1.1), consider the stochastic optimization problem:

$$\inf_{\boldsymbol{\beta}} \mathbb{E}^{\mathbb{P}^*}[h_{\boldsymbol{\beta}}(\mathbf{z})], \tag{1.4}$$

where we set $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{Z} \subseteq \mathbb{R}^d$ in (1.1), $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of coefficients to be learned, $h_{\boldsymbol{\beta}}(\mathbf{z})$: $\mathcal{Z} \times \mathbb{R}^p \to \mathbb{R}$ is the loss function of applying $\boldsymbol{\beta}$ on a sample $\mathbf{z} \in \mathcal{Z}$, and $\mathbb{P}^*$ is the underlying true probability distribution of $\mathbf{z}$. The DRO formulation for (1.4) minimizes the worst-case expected loss over a probabilistic ambiguity set Ω:

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h_{\boldsymbol{\beta}}(\mathbf{z})]. \tag{1.5}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

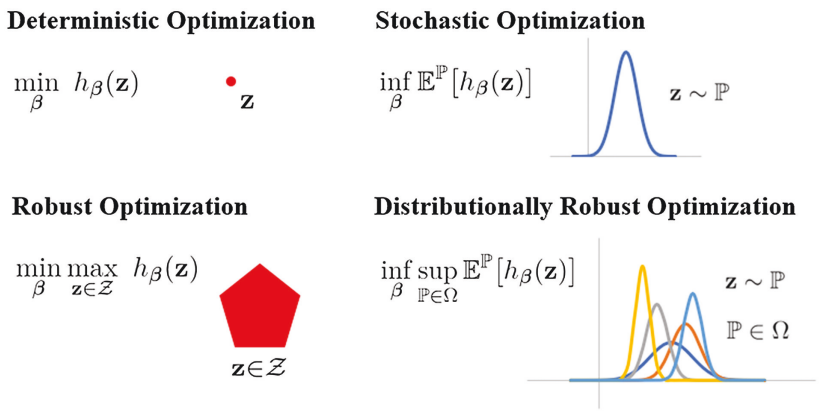


Figure 1.2: Comparison of robust optimization with distributionally robust optimization.

The existing literature on DRO can be split into two main branches, depending on the way in which Ω is defined. One is through a moment ambiguity set, which contains all distributions that satisfy certain moment constraints [48]–[53]. In many cases it leads to a tractable DRO problem but has been criticized for yielding overly conservative solutions [54]. The other is to define Ω as a ball of distributions:

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon D(\mathbb{Q},\ \mathbb{P}_0) \leq \epsilon\},$$

where $\mathcal{Z}$ is the set of possible values for $\mathbf{z}$; $\mathcal{P}(\mathcal{Z})$ is the space of all probability distributions supported on $\mathcal{Z}$; ϵ is a pre-specified radius of the set Ω; and $D(\mathbb{Q}, \mathbb{P}_0)$ is a probabilistic distance function that measures the distance between $\mathbb{Q}$ and a nominal distribution $\mathbb{P}_0$.

The nominal distribution $\mathbb{P}_0$ is typically chosen as the empirical distribution on the observed samples $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$:

$$\mathbb{P}_0 = \hat{\mathbb{P}}_N \triangleq \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{z}_i}(\mathbf{z}),$$

where $\delta_{\mathbf{z}_i}(\cdot)$ is the Dirac density assigning probability mass equal to 1 at $\mathbf{z}_i$; see [12], [13], and [55]. There are also works employing a nonparametric kernel density estimation method to obtain a continuous density function for the nominal distribution, when the underlying true

The version of record is available at: http://dx.doi.org/10.1561/2400000026

distribution is continuous, see [56], [57]. The kernel density estimator is defined as:

$$f_0(\mathbf{z}) = \frac{1}{N|\mathbf{H}|^{1/2}} \sum_{i=1}^{N} K(\mathbf{H}^{-1/2}(\mathbf{z} - \mathbf{z}_i)),$$

where f_0 represents the density function of the nominal distribution $\mathbb{P}_0$, i.e., $f_0 = \mathrm{d}\mathbb{P}_0/\mathrm{d}\mathbf{z}$, $\mathbf{H} \in \mathbb{R}^{d\times d}$ represents a symmetric and positive definite bandwidth matrix, and $K(\cdot)$: $\mathbb{R}^d \to \mathbb{R}^+$ is a symmetric kernel function satisfying $K(\cdot) \geq 0$, $\int_{\mathbb{R}^d} K(\mathbf{z})\mathrm{d}\mathbf{z} = 1$, and $\int_{\mathbb{R}^d} K(\mathbf{z})\mathbf{z}\mathrm{d}\mathbf{z} = \mathbf{0}$.

An example of the probabilistic distance function $D(\cdot, \cdot)$ is the ϕ-divergence [58]:

$$D(\mathbb{Q}, \mathbb{P}_0) = \mathbb{E}^{\mathbb{P}_0}\left[\phi\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}_0}\right)\right],$$

where $\phi(\cdot)$ is a convex function satisfying $\phi(1) = 0$. For example, if $\phi(t) = t\log t$, we obtain the Kullback–Leibler (KL) divergence [59], [60]. The definition of the ϕ-divergence requires that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}_0$. If we take the empirical measure to be the nominal distribution $\mathbb{P}_0$, this implies that the support of $\mathbb{Q}$ must be a subset of the empirical examples. This constraint could potentially hurt the generalization capability of DRO.

Other choices for $D(\cdot, \cdot)$ include the Prokhorov metric [61], and the Wasserstein distance [13], [14], [18], [62], [63]. DRO with the Wasserstein metric has been extensively studied in the machine learning community; see, for example, [12] and [64] for robustified regression models, [33] for adversarial training in neural networks, and [55] for distributionally robust logistic regression. [46] and [47] provided a comprehensive analysis of the Wasserstein-based distributionally robust statistical learning problems with a scalar (as opposed to a vector) response. In recent work, [65] proposed a DRO formulation for convex regression under an absolute error loss.

In this monograph we adopt the Wasserstein metric to define a data-driven DRO problem. Specifically, the ambiguity set Ω is defined as:

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{s,t}(\mathbb{Q}, \hat{\mathbb{P}}_N) \leq \epsilon\}, \tag{1.6}$$

where $\hat{\mathbb{P}}_N$ is the uniform empirical distribution over N training samples $\mathbf{z}_i$, $i \in [\![N]\!]$, and $W_{s,t}(\mathbb{Q},\ \hat{\mathbb{P}}_N)$ is the order-t Wasserstein distance ($t \geq 1$)

The version of record is available at: http://dx.doi.org/10.1561/2400000026

between $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$ defined as:

$$W_{s,t}(\mathbb{Q}, \hat{\mathbb{P}}_N) \triangleq \left(\min_{\pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})} \int_{\mathcal{Z} \times \mathcal{Z}} (s(\mathbf{z}_1, \mathbf{z}_2))^t \mathrm{d}\pi(\mathbf{z}_1, \mathbf{z}_2) \right)^{1/t}, \qquad (1.7)$$

where s is a metric on the data space $\mathcal{Z}$, and π is the joint distribution of $\mathbf{z}_1$ and $\mathbf{z}_2$ with marginals $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$, respectively. The Wasserstein distance between two distributions represents the cost of an optimal mass transportation plan, where the cost is measured through the metric s.

We choose the Wasserstein metric for two main reasons. 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 [13], [14]. 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 [66]. Moreover, the Wasserstein metric takes into account the closeness between support points while other metrics such as the ϕ-divergence only consider the probabilities on these points. An image retrieval example in [14] 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 the 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.

1.3 Outline

The goal of this monograph is to develop a comprehensive robust statistical learning framework using a Wasserstein-based DRO as the modeling tool. Specifically,

- we provide background knowledge on the basics of DRO and the Wasserstein metric, and show its robustness inducing property

The version of record is available at: http://dx.doi.org/10.1561/2400000026

through discussions on the Wasserstein ambiguity set and the property of the DRO solution;

- we cover a variety of predictive and prescriptive models that can be posed and solved using the Wasserstein DRO approach, and show novel problem-tailored theoretical results and real world applications, strengthening the notion of robustness through these discussions;

- we consider a variety of synthetic and real world case studies of the respective models, which validate the theory and the proposed DRO approach and highlight its advantages compared to several alternatives. This could potentially (i) ease the understanding of the model and approach; and (ii) attract practitioners from various fields to put these models into use.

Robust models can be useful when (i) the training data is contaminated with noise, and we want to learn a model that is immunized against the noise; or (ii) the training data is pure, but the test set is contaminated with outliers. In both scenarios we require the model to be insensitive to the data uncertainty/unreliability, which is characterized through a probability distribution that resides in a set consisting of all distributions that are within a pre-specified distance from a nominal distribution. The learning problems that are studied in this monograph include:

- *Distributionally Robust Linear Regression (DRLR)*, which estimates a robustified linear regression plane by minimizing the worst-case expected absolute loss over a probabilistic ambiguity set characterized by the Wasserstein metric.

- *Groupwise Wasserstein Grouped LASSO (GWGL)*, which aims at inducing sparsity at a group level when there exists a predefined grouping structure for the predictors, through defining a specially structured Wasserstein metric for DRO.

- *Distributionally Robust Multi-Output Learning*, which solves a DRO problem with a multi-dimensional response/label vector, generalizing the single-output model addressed in DRLR.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

- Optimal decision making using *DRLR informed K-Nearest Neighbors (K-NN) estimation*, which selects among a set of actions the optimal one through predicting the outcome under each action using K-NN with a distance metric weighted by the DRLR solution.

- *Distributionally Robust Semi-Supervised Learning*, which estimates a robust classifier with partially labeled data, through (i) either restricting the marginal distribution to be consistent with the unlabeled data, (ii) or modifying the structure of DRO by allowing the center of the ambiguity set to vary, reflecting the uncertainty in the labels of the unsupervised data.

- *Distributionally Robust Reinforcement Learning*, which considers *Markov Decision Processes (MDPs)* and seeks to inject robustness into the probabilistic transition model, deriving a lower bound for the *distributionally robust* value function in a regularized form.

The remainder of this monograph is organized as follows. Section 2 presents basics and key properties for the Wasserstein metric. Section 3 discusses how to solve a general Wasserstein DRO problem, the structure of the worst-case distribution, and the performance guarantees of the DRO estimator. The rest of the sections are dedicated to specific learning problems that can be posed as a DRO problem.

In Section 4, we develop the Wasserstein DRO formulation for linear regression under an absolute error loss. Section 5 discusses distributionally robust grouped variable selection, and develops the *Groupwise Wasserstein Grouped LASSO (GWGL)* formulation under the absolute error loss and log-loss. In Section 6, we generalize the single-output model and develop distributionally robust multi-output learning models under Lipschitz continuous loss functions and the multiclass log-loss. Section 7 presents an optimal decision making framework which selects among a set of actions the best one, using predictions from *K-Nearest Neighbors (K-NN)* with a metric weighted by the Wasserstein DRO solution. Section 8 covers a number of active research topics in the domain of DRO under the Wasserstein metric, including (i) DRO in *Semi-Supervised Learning (SSL)* with partially labeled datasets; (ii) DRO in

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Reinforcement Learning (RL) with temporal correlated data. We close the monograph by discussing further potential research directions in Section 9.

1.4 Notational Conventions

Vectors

- Boldfaced lowercase letters denote vectors, ordinary lowercase letters denote scalars, boldfaced uppercase letters denote matrices, and calligraphic capital letters denote sets.
- $\mathbf{e}_i$ denotes the i-th unit vector, $\mathbf{e}$ or $\mathbf{1}$ the vector of ones, and $\mathbf{0}$ a vector of zeros.
- 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 $\mathbf{x}$.

Sets and functions

- We use $\mathbb{R}$ to denote the set of real numbers, and $\mathbb{R}^+$ the set of non-negative real numbers.
- For a set $\mathcal{X}$, we use $|\mathcal{X}|$ to denote its cardinality.
- We write $\text{cone}\{\mathbf{v} \in \mathcal{V}\}$ for a cone that is generated from the set of vectors $\mathbf{v} \in \mathcal{V}$.
- $\mathbf{1}_{\mathcal{A}}(\mathbf{x})$ denotes the indicator function, i.e., $\mathbf{1}_{\mathcal{A}}(\mathbf{x}) = 1$ if $\mathbf{x} \in \mathcal{A}$, and 0 otherwise.
- For $\mathbf{z} \triangleq (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ and a function h, the notations $h(\mathbf{z})$ and $h(\mathbf{x}, y)$ are used interchangeably, and $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$.
- $\mathcal{B}(\mathcal{Z})$ denotes the set of Borel measures supported on $\mathcal{Z}$, and $\mathcal{P}(\mathcal{Z})$ denotes the set of Borel probability measures supported on $\mathcal{Z}$.
- For any integer n we write $[\![n]\!]$ for the set $\{1, \ldots, n\}$. Hence, $\mathcal{P}([\![n]\!])$ denotes the n-th dimensional probability simplex.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Matrices

- $\mathbf{I}$ denotes the identity matrix.
- Prime denotes transpose. Specifically, $\mathbf{A}'$ denotes the transpose of a matrix $\mathbf{A}$.
- For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we will denote by $\mathbf{A} = (a_{ij})_{i\in[\![m]\!]}^{j\in[\![n]\!]}$ the elements of $\mathbf{A}$, by $\mathbf{a}_1, \ldots, \mathbf{a}_m$ the rows of $\mathbf{A}$, and, with some abuse of our notation which denotes vectors by lowercase letters, we will denote by $\mathbf{A}_1, \ldots, \mathbf{A}_n$ the columns of $\mathbf{A}$.
- For a symmetric matrix $\mathbf{A}$, we write $\mathbf{A} \succ 0$ to denote a positive definite matrix, and $\mathbf{A} \succcurlyeq 0$ a positive semi-definite matrix.
- $\mathrm{diag}(\mathbf{x})$ denotes a diagonal matrix whose main diagonal consists of the elements of $\mathbf{x}$ and all off-diagonal elements are zero.
- $\mathrm{tr}(\mathbf{A})$ denotes the trace (i.e., sum of the diagonal elements) of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.
- $|\mathbf{A}|$ denotes the determinant of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.

Norms

- $\|\mathbf{x}\|_p \triangleq (\sum_i |x_i|^p)^{1/p}$ denotes the ℓ_p norm with $p \geq 1$, and $\|\cdot\|$ the general vector norm that satisfies the following properties:
  1. $\|\mathbf{x}\| = 0$ implies $\mathbf{x} = \mathbf{0}$;
  2. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, for any scalar a;
  3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$;
  4. $\|\mathbf{x}\| = \||\mathbf{x}|\|$, where $|\mathbf{x}| = (|x_1|, \ldots, |x_{\dim(\mathbf{x})}|)$;
  5. $\|(\mathbf{x}, \mathbf{0})\| = \|\mathbf{x}\|$, for an arbitrarily long vector $\mathbf{0}$.
- Any $\mathbf{W}$-weighted ℓ_p norm defined as

$$\|\mathbf{x}\|_p^{\mathbf{W}} \triangleq ((|\mathbf{x}|^{p/2})'\mathbf{W}|\mathbf{x}|^{p/2})^{1/p}$$

  with a positive definite matrix $\mathbf{W}$ satisfies the above conditions, where $|\mathbf{x}|^{p/2} = (|x_1|^{p/2}, \ldots, |x_{\dim(\mathbf{x})}|^{p/2})$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

- For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we use $\|\mathbf{A}\|_p$ to denote its induced ℓ_p norm that is defined as $\|\mathbf{A}\|_p \triangleq \sup_{\mathbf{x}\neq\mathbf{0}} \|\mathbf{A}\mathbf{x}\|_p/\|\mathbf{x}\|_p$.

Random variables

- For two random variables w_1 and w_2, we say that w_1 is stochastically dominated by w_2, denoted by $w_1 \overset{\text{D}}{\leq} w_2$, if $\mathbb{P}(w_1 \geq x) \leq \mathbb{P}(w_2 \geq x)$ for all $x \in \mathbb{R}$.
- For a dataset $\mathcal{D} \triangleq \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, we use $\hat{\mathbb{P}}_N$ to denote the empirical measure supported on $\mathcal{D}$, i.e., $\hat{\mathbb{P}}_N \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{z}_i}(\mathbf{z})$, where $\delta_{\mathbf{z}_i}(\mathbf{z})$ denotes the Dirac delta function at point $\mathbf{z}_i \in \mathcal{Z}$.
- The N-fold product of a distribution $\mathbb{P}$ on $\mathcal{Z}$ is denoted by $\mathbb{P}^N$, which represents a distribution on the Cartesian product space $\mathcal{Z}^N$. We write $\mathbb{P}^\infty$ to denote the limit of $\mathbb{P}^N$ as $N \to \infty$.
- $\mathbb{E}^{\mathbb{P}}$ denotes the expectation under a probability distribution $\mathbb{P}$.
- For a random vector $\mathbf{x}$, $\text{cov}(\mathbf{x})$ will denote its covariance.
- $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ denotes the p-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$.
- For a distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, $\mathbb{P}_{\mathcal{X}}(\cdot) \triangleq \sum_{y\in\mathcal{Y}} \mathbb{P}(\cdot, y)$ denotes the marginal distribution over $\mathcal{X}$, and $\mathbb{P}_{|\mathbf{x}} \in \mathcal{P}^{\mathcal{X}}(\mathcal{Y})$ is the conditional distribution over $\mathcal{Y}$ given $\mathbf{x} \in \mathcal{X}$, where $\mathcal{P}^{\mathcal{X}}(\mathcal{Y})$ denotes the set of all conditional distributions supported on $\mathcal{Y}$, given features in $\mathcal{X}$.
- $W_{s,t}(\mathbb{P}, \mathbb{Q})$ denotes the order-t Wasserstein distance between measures $\mathbb{P}, \mathbb{Q}$ under a cost metric s. For ease of notation and when the cost metric is clear from the context we will be writing $W_t(\mathbb{P}, \mathbb{Q})$.
- $\Omega_\epsilon^{s,t}(\mathbb{P})$ denotes the set of probability distributions whose order-t Wasserstein distance under a cost metric s from the distribution $\mathbb{P}$ is less than or equal to ϵ, i.e.,

$$\Omega_\epsilon^{s,t}(\mathbb{P}) \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{s,t}(\mathbb{Q},\ \mathbb{P}) \leq \epsilon\}.$$

  For ease of notation, when the cost metric is clear from the context and $t = 1$, we will be writing $\Omega_\epsilon(\mathbb{P})$, or simply Ω when the center distribution $\mathbb{P}$ is clear from the context.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

1.5 Abbreviations

ACE	.....	Angiotensin-Converting Enzyme
ACS	.....	American College of Surgeons
AD	.....	Absolute Deviation
ARB	.....	Angiotensin Receptor Blockers
a.s.	.....	almost surely
AUC	.....	Area Under the ROC Curve
BMI	.....	Body Mass Index
CART	.....	Classification And Regression Trees
CCA	.....	Canonical Correlation Analysis
CCR	.....	Correct Classification Rate
CI	.....	Confidence Interval
CT	.....	Computed Tomography
CTDI	.....	CT Dose Index
CVaR	.....	Conditional Value at Risk
C&W	.....	The Curds and Whey procedure
DRLR	.....	Distributionally Robust Linear Regression
DRO	.....	Distributionally Robust Optimization
EHRs	.....	Electronic Health Records
EN	.....	Elastic Net
FA	.....	False Association
FD	.....	False Disassociation
FES	.....	Factor Estimation and Selection
GLASSO	.....	Grouped LASSO
GSRL	.....	Grouped Square Root LASSO
GWGL	.....	Groupwise Wasserstein Grouped LASSO
HbA_{1c}	.....	hemoglobin A1c
HIPAA	.....	Health Insurance Portability and Accountability Act
ICD-9	.....	International Classification of Diseases, Ninth Revision
i.i.d.	.....	independently and identically distributed

The version of record is available at: http://dx.doi.org/10.1561/2400000026

IRB	.....	Institutional Review Board
IRLS	.....	Iteratively Reweighted Least Squares
KL	.....	Kullback–Leibler
K-NN	.....	K-Nearest Neighbors
LAD	.....	Least Absolute Deviation
LASSO	.....	Least Absolute Shrinkage and Selection Operator
LG	.....	Logistic Regression
LHS	.....	Left Hand Side
LMS	.....	Least Median of Squares
LOESS	.....	LOcally Estimated Scatterplot Smoothing
LTS	.....	Least Trimmed Squares
MAD	.....	Median Absolute Deviation
MCC	.....	MultiClass Classification
MDP	.....	Markov Decision Process
MeanAE	.....	Mean Absolute Error
min-max	.....	minimization-maximization
MLE	.....	Maximum Likelihood Estimator
MLG	.....	Multiclass Logistic Regression
MLR	.....	Multi-output Linear Regression
MPD	.....	Minimal Perturbation Distance
MPI	.....	Maximum Percentage Improvement
MPMs	.....	Minimax Probability Machines
MSE	.....	Mean Squared Error
NPV	.....	Negative Predictive Value
NSQIP	.....	National Surgical Quality Improvement Program
OLS	.....	Ordinary Least Squares
PCR	.....	Principal Components Regression
PPV	.....	Positive Predictive Value
PVE	.....	Proportion of Variance Explained
RBA	.....	Robust Bias-Aware
RHS	.....	Right Hand Side
RL	.....	Reinforcement Learning

The version of record is available at: http://dx.doi.org/10.1561/2400000026

ROC	.....	Receiver Operating Characteristic
RR	.....	Relative Risk
RRR	.....	Reduced Rank Regression
RTE	.....	Relative Test Error
SNR	.....	Signal to Noise Ratio
SR	.....	Squared Residuals
SSL	.....	Semi-Supervised Learning
std	.....	standard deviation
SVM	.....	Support Vector Machine
TA	.....	True Association
TD	.....	True Disassociation
TAR	.....	True Association Rate
TDR	.....	True Disassociation Rate
WGD	.....	Within Group Difference
w.h.p.	.....	with high probability
WMSE	.....	Weighed Mean Squared Error
w.p.1	.....	with probability 1
w.r.t.	.....	with respect to

The version of record is available at: http://dx.doi.org/10.1561/2400000026

2

The Wasserstein Metric

In this section, we outline basic properties of the Wasserstein distance. A definition in the case of discrete measures is provided in Section 2.1. Section 2.2 establishes that it is a proper distance metric. A dual formulation and a generalization to arbitrary measures are presented in Section 2.3. Special cases are described in Section 2.4. A discussion on how to set the Wasserstein underlying transport cost function in the context of robust learning is in Section 2.5. A related robustness-inducing property of the Wasserstein metric is shown in Section 2.6 and a discussion on how to set the radius of the Wasserstein ambiguity set is included in Section 2.7.

2.1 Basics

We start by reviewing basic properties of the Wasserstein metric defined in Section 1 (cf. Equation (1.7)). We will define the metric and establish key results, first using discrete probability distributions, and then state how the definitions and results generalize to arbitrary probability measures.

Consider two discrete probability distributions $\mathbb{P} = \{p_1, \ldots, p_m\}$ and $\mathbb{Q} = \{q_1, \ldots, q_n\}$, where $p_i, q_j \geq 0$, for all i, j, and $\sum_{i=1}^{m} p_i = \sum_{j=1}^{n} q_j = 1$. For convenience, let us write $\mathbf{p} = (p_1, \ldots, p_m)$ and $\mathbf{q} = (q_1, \ldots, q_n)$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

for the corresponding column vectors. Define a metric (cost) between points in the support of $\mathbb{P}$ and $\mathbb{Q}$ by $s(i,j)$, $i \in [\![m]\!]$, $j \in [\![n]\!]$, and collect all these quantities in an $m \times n$ matrix $\mathbf{S} = (s(i,j))$ whose (i,j) element is $s(i,j)$. Consider the *Linear Programming (LP)* problem:

$$\begin{aligned} W_{\mathbf{S},1}(\mathbb{P},\mathbb{Q}) = \min_{\boldsymbol{\pi}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} \pi(i,j)s(i,j) \\ \text{s.t.} \quad & \sum_{i=1}^{m} \pi(i,j) = q_j, \quad j \in [\![n]\!], \\ & \sum_{j=1}^{n} \pi(i,j) = p_i, \quad i \in [\![m]\!], \\ & \pi(i,j) \geq 0, \quad \forall i,j, \end{aligned} \tag{2.1}$$

where $\boldsymbol{\pi} = (\pi(i,j);\, \forall i,j)$ is the decision vector. Notice that according to the definition in Equation (1.7), the objective value is the order-1 Wasserstein distance between distributions $\mathbb{P}$ and $\mathbb{Q}$. In $W_{\mathbf{S},1}(\mathbb{P},\mathbb{Q})$ we have inserted the subscript $\mathbf{S}$ to explicitly denote the dependence on the cost matrix. Similarly, by defining a cost matrix $\mathbf{S}^t = ((s(i,j))^t)$, the order-$t$ Wasserstein distance, denoted by $W_{\mathbf{S},t}(\cdot,\cdot)$, can be obtained as the t-th root of the optimal value of the same LP with cost matrix $\mathbf{S}^t$; namely,

$$W_{\mathbf{S},t}(\mathbb{P},\mathbb{Q}) = (W_{\mathbf{S}^t,1}(\mathbb{P},\mathbb{Q}))^{1/t}. \tag{2.2}$$

The LP formulation in (2.1) is equivalent to the well-known *transportation problem* [67] and can be interpreted as the cost of transporting probability mass from the support points of $\mathbb{P}$ to those of $\mathbb{Q}$. Specifically, the problem corresponds to the bipartite graph in Figure 2.1 with nodes $\{u_1,\ldots,u_m\}$ representing the support of $\mathbb{P}$, nodes $\{v_1,\ldots,v_n\}$ representing the support of $\mathbb{Q}$, p_i being the supply at node u_i, q_j the demand at node v_j, and $\pi(i,j)$ the flow of material (probability mass) from node u_i to node v_j incurring a transportation cost of $s(i,j)$ per unit of material.

The formulation in (2.1) has a long history, starting with Monge [68] who formulated a problem of optimally transferring material extracted from a mining site to various construction sites; hence, the terms *optimal mass transport* and *earth mover's distance*. In Monge's formulation, all

The version of record is available at: http://dx.doi.org/10.1561/2400000026

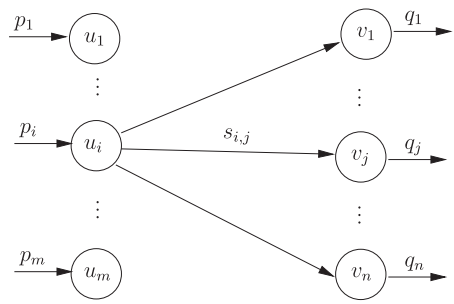


Figure 2.1: The transportation problem for computing the Wasserstein distance $W_{\mathbf{S},1}(\mathbb{P}, \mathbb{Q})$.

material from a source node u_i gets "assigned" to a destination node v_j. Kantorovich [69], [70] relaxed the problem by allowing sources to split their material to several destination nodes. For Kantorovich, this was an application of an LP he had earlier defined for production planning problems ([71], later translated in English in [72]) and a method (and a duality theorem) he had developed for these problems [73]. Definitive references on optimal mass transport are [74], and, focusing more on computational aspects, [75]. In presenting some of the key properties and duality we will follow the approach of [75] which presents the theory for discrete probability distributions.

2.2 A Distance Metric

In this section we establish that the Wasserstein distance $W_{\mathbf{S},t}(\mathbb{P}, \mathbb{Q})$ is a distance metric, assuming that the underlying cost $s(i, j)$ is a proper distance metric.

Assumption A. Let $n = m$ and assume

1. $s(i, j) \geq 0$, with $s(i, j) = 0$ if and only if $i = j$.
2. $s(i, j) = s(j, i)$ for $i \neq j$.
3. For any triplet $i, j, k \in [\![n]\!]$, $s(i, k) \leq s(i, j) + s(j, k)$.

Theorem 2.2.1. Under Assumption A, the order-t Wasserstein distance ($t \geq 1$) is a metric, i.e.,

The version of record is available at: http://dx.doi.org/10.1561/2400000026

1. $W_{\mathbf{S},t}(\mathbb{P},\mathbb{Q}) \geq 0$ for any $\mathbb{P},\mathbb{Q} \in \mathcal{P}([\![n]\!])$, with $W_{\mathbf{S},t}(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.

2. $W_{\mathbf{S},t}(\mathbb{P},\mathbb{Q}) = W_{\mathbf{S},t}(\mathbb{Q},\mathbb{P})$ for any $\mathbb{P},\mathbb{Q} \in \mathcal{P}([\![n]\!])$.

3. For any triplet $\mathbb{P},\mathbb{Q},\mathbb{V} \in \mathcal{P}([\![n]\!])$, $W_{\mathbf{S},t}(\mathbb{P},\mathbb{V}) \leq W_{\mathbf{S},t}(\mathbb{P},\mathbb{Q}) + W_{\mathbf{S},t}(\mathbb{Q},\mathbb{V})$.

Proof. Recall Equation (2.2) that relates $W_{\mathbf{S},t}(\mathbb{P},\mathbb{Q})$ to $W_{\mathbf{S}^t,1}(\mathbb{P},\mathbb{Q})$. The latter quantity can be obtained as the optimal value of the LP in (2.1) using the cost metric $\mathbf{S}^t$.

1. The non-negativity follows directly from the formulation in (2.1) since $s(i,j) \geq 0$ (by Assumption A), hence $(s(i,j))^t \geq 0$, and any feasible solution satisfies $\pi(i,j) \geq 0$. In addition, $W_{\mathbf{S},t}(\mathbb{P},\mathbb{P}) = 0$, because, in this case, the optimal solution in formulation (2.1) satisfies $\pi(i,j) = 0$, if $i \neq j$, and $\pi(i,i) = p_i$, for all i. Since $(s(i,i))^t = 0$ (due to Assumption A), the optimal value of the LP in (2.1) is zero. Further, if $\mathbb{P} \neq \mathbb{Q}$, there should be flow $\pi(i,j) > 0$ for some $i \neq j$, and since $s(i,j) > 0$ for those i,j (due to Assumption A), the optimal value of the LP is positive.

2. To establish symmetry, consider $W_{\mathbf{S},t}(\mathbb{P},\mathbb{Q})$ and compare it with $W_{\mathbf{S},t}(\mathbb{Q},\mathbb{P})$. It suffices to compare $W_{\mathbf{S}^t,1}(\mathbb{P},\mathbb{Q})$ with $W_{\mathbf{S}^t,1}(\mathbb{Q},\mathbb{P})$. To that end, notice that given an optimal solution $\pi_f(i,j)$, for all i,j, for $W_{\mathbf{S}^t,1}(\mathbb{P},\mathbb{Q})$ computed from the LP in (2.1), we can obtain an optimal solution $\pi_b(i,j)$ for $W_{\mathbf{S}^t,1}(\mathbb{Q},\mathbb{P})$ simply by reversing the flows, i.e., $\pi_b(j,i) = \pi_f(i,j)$, for all i,j. Given the symmetry of the cost $s(i,j)$ due to Assumption A, the result follows.

3. To establish the triangle inequality, fix $\mathbb{P},\mathbb{Q},\mathbb{V} \in \mathcal{P}([\![n]\!])$ and consider $W_{\mathbf{S},t}(\mathbb{P},\mathbb{Q})$ and $W_{\mathbf{S},t}(\mathbb{Q},\mathbb{V})$. Let $\mathbf{\Pi}_1 = (\pi_1(i,j))_{i,j\in[\![n]\!]}$ and $\mathbf{\Pi}_2 = (\pi_2(i,j))_{i,j\in[\![n]\!]}$ be the optimal solutions of the LPs corresponding to $W_{\mathbf{S}^t,1}(\mathbb{P},\mathbb{Q})$ and $W_{\mathbf{S}^t,1}(\mathbb{Q},\mathbb{V})$, respectively. Define a $\tilde{\mathbb{Q}}$ such that $\tilde{q}_i = q_i$, if $q_i > 0$, and $\tilde{q}_i = 1$, otherwise. Let $\tilde{\mathbf{q}}$ be the corresponding column vector. Define $\mathbf{D} = \mathrm{diag}(1/\tilde{q}_1,\ldots,1/\tilde{q}_n)$.

   Consider next $W_{\mathbf{S},t}(\mathbb{P},\mathbb{V})$ and the LP corresponding to $W_{\mathbf{S}^t,1}(\mathbb{P},\mathbb{V})$. We will first argue that $\mathbf{\Pi}_{1,2} \stackrel{\triangle}{=} \mathbf{\Pi}_1\mathbf{D}\mathbf{\Pi}_2$ forms a feasible solution

The version of record is available at: http://dx.doi.org/10.1561/2400000026

to that LP. Specifically, recalling that $\mathbf{e}$ is the vector of all ones,

$$\mathbf{\Pi}_{1,2}\mathbf{e} = \mathbf{\Pi}_1\mathbf{D}\mathbf{\Pi}_2\mathbf{e} = \mathbf{\Pi}_1\mathbf{D}\mathbf{q} = \mathbf{\Pi}_1\mathbf{e}_{\mathbb{Q}} = \mathbf{p},$$

where we used the feasibility of $\mathbf{\Pi}_1, \mathbf{\Pi}_2$, and $\mathbf{e}_{\mathbb{Q}}$ is a vector whose ith element is set to 1 if $q_i > 0$, and to zero, otherwise. Similarly, we can also show $\mathbf{e}'\mathbf{\Pi}_{1,2} = \mathbf{v}'$, where $\mathbf{v}$ is the column vector corresponding to $\mathbb{V}$.

Letting $\mathbf{\Pi}_{1,2} = (\pi_{1,2}(i,j))_{i,j\in[\![n]\!]}$, we have

$$\begin{aligned}
W_{\mathbf{S},t}(\mathbb{P},\mathbb{V}) &= (W_{\mathbf{S}^t,1}(\mathbb{P},\mathbb{V}))^{1/t} \\
&\leq \Big(\sum_{i,j}(s(i,j))^t\pi_{1,2}(i,j)\Big)^{1/t} && (2.3)\\
&= \Big(\sum_{i,j}(s(i,j))^t\sum_k\frac{\pi_1(i,k)\pi_2(k,j)}{\tilde{q}_k}\Big)^{1/t} \\
&\leq \Big(\sum_{i,j,k}(s(i,k)+s(k,j))^t\frac{\pi_1(i,k)\pi_2(k,j)}{\tilde{q}_k}\Big)^{1/t} && (2.4)\\
&= \Big(\sum_{i,j,k}\Big[s(i,k)\Big(\frac{\pi_1(i,k)\pi_2(k,j)}{\tilde{q}_k}\Big)^{1/t} \\
&\qquad + s(k,j)\Big(\frac{\pi_1(i,k)\pi_2(k,j)}{\tilde{q}_k}\Big)^{1/t}\Big]^t\Big)^{1/t} \\
&\leq \Big(\sum_{i,j,k}(s(i,k))^t\frac{\pi_1(i,k)\pi_2(k,j)}{\tilde{q}_k}\Big)^{1/t} \\
&\qquad + \Big(\sum_{i,j,k}(s(k,j))^t\frac{\pi_1(i,k)\pi_2(k,j)}{\tilde{q}_k}\Big)^{1/t} && (2.5)\\
&= \Big(\sum_{i,k}(s(i,k))^t\pi_1(i,k)\sum_j\frac{\pi_2(k,j)}{\tilde{q}_k}\Big)^{1/t} \\
&\qquad + \Big(\sum_{j,k}(s(k,j))^t\pi_2(k,j)\sum_i\frac{\pi_1(i,k)}{\tilde{q}_k}\Big)^{1/t}
\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

$$= \Big(\sum_{i,k} (s(i,k))^t \pi_1(i,k) \Big)^{1/t} + \Big(\sum_{j,k} (s(k,j))^t \pi_2(k,j) \Big)^{1/t} \tag{2.6}$$
$$= W_{\mathbf{S},t}(\mathbb{P}, \mathbb{Q}) + W_{\mathbf{S},t}(\mathbb{Q}, \mathbb{V}),$$

where (2.3) follows from the feasibility (and potential suboptimality) of $\pi_{1,2}(i,j)$, (2.4) follows from the triangle inequality for $s(i,j)$, (2.5) is due to the Minkowski inequality, and (2.6) uses the feasibility of $\mathbf{\Pi}_1, \mathbf{\Pi}_2$. □

As a final comment in this section, we note that the order-1 Wasserstein distance $W_{\mathbf{S},1}(\mathbb{P}, \mathbb{Q})$, viewed as a function of the vectors $\mathbf{p}$ and $\mathbf{q}$ corresponding to $\mathbb{P}$ and $\mathbb{Q}$, is a convex function. This follows from the LP formulation (2.1), where the optimal value is a convex function of the RHS of the constraints [67, Section 5.2].

2.3 The Dual Problem

In this section, we derive the dual of the mass transportation problem in (2.1). Let f_j be the dual variable corresponding to the flow conservation constraint for q_j and g_i the dual variable corresponding to the flow conservation constraint for p_i. We write $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^m$ for the corresponding dual vectors. Using LP duality, the dual of (2.1) takes the form:

$$\begin{aligned} W_{\mathbf{S},1}(\mathbb{P}, \mathbb{Q}) = \max_{\mathbf{f},\mathbf{g}} \quad & \sum_{i=1}^{m} g_i p_i + \sum_{j=1}^{n} f_j q_j \\ \text{s.t.} \quad & f_j + g_i \le s(i,j), \quad i \in [\![m]\!],\ j \in [\![n]\!]. \end{aligned} \tag{2.7}$$

The optimal value is equal to the primal optimal value due to the LP strong duality. The complementary slackness conditions suggest that

$$\text{if } \pi(i,j) > 0 \quad \text{ then } f_j + g_i = s(i,j). \tag{2.8}$$

Necessary and sufficient conditions for a primal solution $\mathbf{\Pi}$ to be primal optimal and dual solutions $\mathbf{f}$ and $\mathbf{g}$ to be dual optimal are: (i) primal feasibility, (ii) dual feasibility, and (iii) the complementary slackness condition in (2.8).

The version of record is available at: http://dx.doi.org/10.1561/2400000026

The primal and dual problems can be interpreted as follows. The primal problem is the problem of minimizing transportation cost for a transporter of mass across the bipartite graph in Figure 2.1. The transporter faces a cost of $s(i,j)$ per unit of mass transported on link (i,j). Suppose now that the transporter, instead of carrying out the transportation plan, hires another shipping company (e.g., a company like UPS, DHL, or Fedex). This shipping company charges a price of g_i for picking one unit of mass from node u_i and a price of f_j for delivering one unit of mass to node v_j. The dual problem is then the problem solved by the shipping company to maximize its revenue by carrying out the transportation of mass. Strong duality simply states that there should not be an "arbitrage" opportunity and the transportation cost must be the same irrespective of whether the transporter of mass hires a shipping company or not. In other words, if the price offered by the shipping company was strictly less than the transportation cost, then the mass transporter would be able to make money just by outsourcing shipping. Furthermore, the market conditions would be ripe for another middleperson to come into the market, offer the shipping company higher prices, while still making it profitable for the transporter to use the middleperson's services. More specifically, the complementary slackness conditions (2.8) suggest that if there is mass transported along link (i,j), the cost of transporting the mass through the shipping company must equal the transportation cost faced by the transporter across that link.

A different interpretation of the primal and the dual can be obtained through an analogy with electrical circuits. Let us treat p_i as current flowing *into* node u_i. Similarly, q_j is current flowing *out of* node v_j, or, equivalently, the inflow into v_j is equal to $\hat{q}_j = -q_j$. Rewriting the dual problem (2.7) using the $\hat{q}_i$'s and changing variables from f_j to $\hat{f}_j = -f_j$ yields:

$$\begin{aligned} W_{\mathbf{S},1}(\mathbb{P},\mathbb{Q}) = \max_{\hat{\mathbf{f}},\mathbf{g}} \quad & \sum_{i=1}^{m} g_i p_i + \sum_{j=1}^{n} \hat{f}_j \hat{q}_j \\ \text{s.t.} \quad & g_i - \hat{f}_j \leq s(i,j), \quad i \in [\![m]\!],\ j \in [\![n]\!]. \end{aligned} \tag{2.9}$$

In this context, the constraints of the primal can be viewed as Kirchoff's current law and the dual variables (g_i at nodes u_i and $\hat{f}_j$ at nodes v_j)

The version of record is available at: http://dx.doi.org/10.1561/2400000026

can be interpreted as electric potentials (voltages with respect to the ground) at the nodes. The complementary slackness conditions state that if there is current flowing from node u_i to v_j, the voltage, or potential difference among these nodes, must equal $s(i,j)$. More simply put, for one unit of flow (current), the voltage must be equal to the "resistor" $s(i,j)$, which corresponds to Ohm's law. These node potentials are known as Kantorovich potentials [75].

2.3.1 Arbitrary Measures and Kantorovich Duality

The primal problem we defined in (2.1) can be generalized to arbitrary measures as defined in Equation (1.7). Consider two Polish (i.e., complete, separable, metric) probability spaces $(\mathcal{Z}_1, \mathbb{P})$ and $(\mathcal{Z}_2, \mathbb{Q})$ and a lower semicontinuous cost function s: $\mathcal{Z}_1 \times \mathcal{Z}_2 \to \mathbb{R} \cup \{+\infty\}$. Then, the order-1 Wasserstein distance can be defined as the optimal value of the primal problem:

$$W_{s,1}(\mathbb{P}, \mathbb{Q}) = \min_{\pi} \int_{\mathcal{Z}_1 \times \mathcal{Z}_2} s(\mathbf{z}_1, \mathbf{z}_2) \mathrm{d}\pi(\mathbf{z}_1, \mathbf{z}_2), \tag{2.10}$$

where $\pi \in \mathcal{P}(\mathcal{Z}_1 \times \mathcal{Z}_2)$ is a joint probability distribution of $\mathbf{z}_1, \mathbf{z}_2$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. The order-t Wasserstein distance can be obtained as:

$$W_{s,t}(\mathbb{P}, \mathbb{Q}) = (W_{s^t,1}(\mathbb{P}, \mathbb{Q}))^{1/t}, \tag{2.11}$$

where $s^t(\mathbf{z}_1, \mathbf{z}_2) = (s(\mathbf{z}_1, \mathbf{z}_2))^t$.

The dual problem, known as the Kantorovich dual [74, Theorem 5.10], analogously to Problem (2.7) can be written as:

$$\begin{aligned} W_{s,1}(\mathbb{P}, \mathbb{Q}) = \sup_{f,g} \quad & \int_{\mathcal{Z}_1} g(\mathbf{z}_1) \mathrm{d}\mathbb{P}(\mathbf{z}_1) + \int_{\mathcal{Z}_2} f(\mathbf{z}_2) \mathrm{d}\mathbb{Q}(\mathbf{z}_2) \\ \text{s.t.} \quad & f(\mathbf{z}_2) + g(\mathbf{z}_1) \leq s(\mathbf{z}_1, \mathbf{z}_2), \quad \mathbf{z}_1 \in \mathcal{Z}_1, \ \mathbf{z}_2 \in \mathcal{Z}_2, \end{aligned} \tag{2.12}$$

where f and g are absolutely integrable under $\mathbb{Q}$ and $\mathbb{P}$, respectively. By the Kantorovich-Rubinstein Theorem [74], when $s(\mathbf{z}_1, \mathbf{z}_2)$ is a distance metric on a Polish space $\mathcal{Z}_1$, (2.12) can be simplified to

$$\begin{aligned} W_{s,1}(\mathbb{P}, \mathbb{Q}) = \sup_{g} \quad & \int_{\mathcal{Z}_1} g(\mathbf{z}_1) \mathrm{d}\mathbb{P}(\mathbf{z}_1) - \int_{\mathcal{Z}_2} g(\mathbf{z}_2) \mathrm{d}\mathbb{Q}(\mathbf{z}_2) \\ \text{s.t.} \quad & |g(\mathbf{z}_1) - g(\mathbf{z}_2)| \leq s(\mathbf{z}_1, \mathbf{z}_2), \quad \mathbf{z}_1 \in \mathcal{Z}_1, \ \mathbf{z}_2 \in \mathcal{Z}_2. \end{aligned} \tag{2.13}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

2.4 Some Special Cases

2.4.1 One-Dimensional Cases

Suppose $\mathbb{P}$ and $\mathbb{Q}$ are discrete distributions on $\mathbb{R}$. Let $\mathbb{P}$ have mass of $1/n$ at each of the points $x_i \in \mathbb{R},\ i \in [\![n]\!]$, where $x_1 \leq x_2 \leq \cdots \leq x_n$. Similarly, $\mathbb{Q}$ assigns mass of $1/n$ at each of the points $y_i \in \mathbb{R},\ i \in [\![n]\!]$, where $y_1 \leq y_2 \leq \cdots \leq y_n$. Then, with $s(x, y) = |x - y|$, the order-t Wasserstein distance can be obtained as:

$$W_{s,t}(\mathbb{P}, \mathbb{Q}) = \left(\frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|^t \right)^{1/t}. \tag{2.14}$$

This can be easily obtained by solving the corresponding formulation in (2.1).

For continuous one-dimensional distributions on $\mathbb{R}$, let $F_{\mathbb{P}}$ denote the *Cumulative Distribution Function (CDF)* of $\mathbb{P}$, namely,

$$F_{\mathbb{P}}(x) = \int_{-\infty}^{x} \mathrm{d}\mathbb{P}, \quad x \in \mathbb{R}.$$

Define the inverse CDF or quantile function $F_{\mathbb{P}}^{-1}(p)$ as

$$F_{\mathbb{P}}^{-1}(p) = \min\{x \in \mathbb{R} \cup \{-\infty\} \colon F_{\mathbb{P}}(x) \geq p\}, \quad p \in [0, 1].$$

Let $F_{\mathbb{Q}}$ and $F_{\mathbb{Q}}^{-1}$ be the corresponding quantities for $\mathbb{Q}$. Then, using again the metric $s(x, y) = |x - y|$, for $x, y \in \mathbb{R}$, the order-t Wasserstein distance can be computed as [75]:

$$W_{s,t}(\mathbb{P}, \mathbb{Q}) = \left(\int_0^1 \left| F_{\mathbb{P}}^{-1}(p) - F_{\mathbb{Q}}^{-1}(p) \right|^t \mathrm{d}p \right)^{1/t}. \tag{2.15}$$

2.4.2 Sliced Wasserstein Distance

The fact that Wasserstein distances can be easily computed for one-dimensional distributions on $\mathbb{R}$ has led to the following approximation of the Wasserstein distance between distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}^d$. Specifically, for any direction $\boldsymbol{\theta}$ on the ball $\mathcal{S}^d = \{\boldsymbol{\theta} \in \mathbb{R}^d \colon \|\boldsymbol{\theta}\|_2 = 1\}$, let $T_{\boldsymbol{\theta}} \colon \mathbf{x} \in \mathbb{R}^d \to \mathbb{R}$ be the projection from $\mathbb{R}^d$ to $\mathbb{R}$. Let $T_{\boldsymbol{\theta},\#\mathbb{P}}$ be the

The version of record is available at: http://dx.doi.org/10.1561/2400000026

so-called push-forward measure satisfying

$$T_{\boldsymbol{\theta},\#\mathbb{P}}(\mathcal{A}) = \mathbb{P}(\{\mathbf{x} \in \mathbb{R}^d \colon T_{\boldsymbol{\theta}}(\mathbf{x}) \in \mathcal{A}\}), \quad \mathcal{A} \subseteq \mathbb{R}.$$

Define $T_{\boldsymbol{\theta},\#\mathbb{Q}}$ similarly. Then, the so-called sliced Wasserstein distance [76], [77] can be defined as:

$$SW_{s,2} = \int_{\mathcal{S}^d} W_{s,2}(T_{\boldsymbol{\theta},\#\mathbb{P}}, T_{\boldsymbol{\theta},\#\mathbb{Q}}) \mathrm{d}\boldsymbol{\theta}, \tag{2.16}$$

where $s(x, y) = |x - y|$, for $x, y \in \mathbb{R}$. Such an integral can be approximated using Monte-Carlo integration, giving rise to a computational method for computing Wasserstein distances between distributions in $\mathbb{R}^d$.

2.4.3 Gaussian Distributions

We next consider the case of two Gaussian distributions. Let $\mathbb{P} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ be a d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}_1$ and covariance $\boldsymbol{\Sigma}_1$. Similarly, let $\mathbb{Q} \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Define the metric $s(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_2$. Then, the order-2 Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$ is given in closed-form in [78] and [79]:

$$W_{s,2}(\mathbb{P}, \mathbb{Q}) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{1/2})^{1/2}).$$

2.5 The Transport Cost Function

In this monograph, we are focusing on the use of the Wasserstein metric in the context of robust learning, specifically the DRO problem we defined in Equation (1.5). As a result, the cost function s used in defining the Wasserstein metric should reflect any implicit knowledge we have on the nature of the data $\mathbf{z} = (\mathbf{x}, y)$. Without loss of generality, suppose that the data have already been *standardized*, specifically, for all data points $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i \in [\![N]\!]$, in the training set, we have normalized every variable (coordinate) in $\mathbf{x}_i$ by subtracting the empirical mean and dividing by the sample standard deviation. Then, an element of $\mathbf{x}_i$ will have a large absolute value if the corresponding variable deviates substantially from the empirical mean. Below, we discuss a number of different scenarios on what may be known regarding the data and the implied appropriate corresponding cost function.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

1. Suppose we know that the model we are seeking is *sparse*, i.e., there are few variables, and in the extreme case one, that determine the output y. In this case, an appropriate cost function is an ℓ_∞ norm in the $\mathbf{z} = (\mathbf{x}, y)$ space. In particular, given two data points $\mathbf{z}_1 = (\mathbf{x}_1, y_1)$ and $\mathbf{z}_2 = (\mathbf{x}_2, y_2)$, if $y_1 \neq y_2$ and $\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty < |y_1 - y_2|$, the distance between $\mathbf{z}_1$ and $\mathbf{z}_2$ is equal to $|y_1 - y_2|$. If, however, $y_1 \approx y_2$, then the distance between $\mathbf{z}_1$ and $\mathbf{z}_2$ is determined by the absolute difference in the most deviating variable, that is, $\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty$.

2. Suppose now that we believe the model to be *dense*, implying that almost all variables are relevant and predictive of the output y. Then, an appropriate distance metric between two points $\mathbf{z}_1 = (\mathbf{x}_1, y_1)$ and $\mathbf{z}_2 = (\mathbf{x}_2, y_2)$ is the ℓ_2 norm $\|\mathbf{z}_1 - \mathbf{z}_2\|_2$, where all $\mathbf{x}$ coordinates and y are weighted equally. More generally, one can introduce weights and use a $\mathbf{W}$-weighted ℓ_p norm defined as $\|\mathbf{z}\|_p^{\mathbf{W}} = ((|\mathbf{z}|^{p/2})'\mathbf{W}|\mathbf{z}|^{p/2})^{1/p}$ with a positive definite weight matrix $\mathbf{W}$.

3. As one more example, suppose that the data $\mathbf{z}$ are organized into a set of (overlapping or non-overlapping) groups according to $\mathbf{z} = (\mathbf{z}^1, \ldots, \mathbf{z}^L)$. To reflect this group structure, we can define a (q, t)-norm, with $q, t \geq 1$, as:

$$\|\mathbf{z}\|_{q,t} = \left(\sum_{l=1}^{L} (\|\mathbf{z}^l\|_q)^t \right)^{1/t}.$$

Notice that the (q, t)-norm of $\mathbf{z}$ is actually the ℓ_t-norm of the vector $(\|\mathbf{z}^1\|_q, \ldots, \|\mathbf{z}^L\|_q)$, which represents each group vector $\mathbf{z}^l$ in a concise way via the ℓ_q-norm. A special case is the $(2, \infty)$-norm on the weighted predictor-response vector

$$\mathbf{z}_\mathbf{w} \triangleq \left(\frac{1}{\sqrt{p_1}} \mathbf{x}^1, \ldots, \frac{1}{\sqrt{p_L}} \mathbf{x}^L, My \right),$$

where the weight vector is

$$\mathbf{w} = \left(\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}}, M \right),$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

and M is a positive weight assigned to the response. Specifically,

$$\|\mathbf{z}_{\mathbf{w}}\|_{2,\infty} = \max\left\{\frac{1}{\sqrt{p_1}}\|\mathbf{x}^1\|_2, \ldots, \frac{1}{\sqrt{p_L}}\|\mathbf{x}^L\|_2, M|y|\right\},$$

where different groups are scaled by the number of variables they contain. The ℓ_2 norm at the individual group level reflects the intuition that all variables in a group are relevant, whereas the ℓ_∞ norm among groups reflects the intuition that there is a dominant group predictive of the response, just like the situation we outlined in Item 1 above. As we will see later, such a norm imposes a *group sparsity* structure.

2.5.1 Transport Cost Function via Metric Learning

We now discuss a metric learning approach for determining the weighted transport cost function we outlined in Item 2 above, following the line of work in [80]. The intuition is to calibrate a cost function $s(\cdot)$ which assigns a high transportation cost to a pair of data points $(\mathbf{z}_1, \mathbf{z}_2)$ if transporting mass between these locations significantly impacts the performance.

Consider a classification problem where we observe N (predictor, label) pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, and $y_i \in \{-1, +1\}$. Suppose we use a weighted ℓ_2 norm as the distance metric on the space of predictors:

$$s_{\mathbf{W}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)'\mathbf{W}(\mathbf{x}_1 - \mathbf{x}_2)},$$

where the weight matrix $\mathbf{W}$ is symmetric and positive semi-definite. The goal is to inform the selection of $\mathbf{W}$ through recognizing the pairs of samples that are similar/dissimilar to each other. In a classification setting, the labels form a natural separation plane for the observed samples. We define two sets:

$$\mathcal{M} \triangleq \{(i,j)\text{: } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are close to each other and } y_i = y_j\},$$
$$\mathcal{N} \triangleq \{(i,j)\text{: } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are far away from each other}\},$$

where the closeness between $\mathbf{x}$ can be evaluated using an appropriate norm, e.g., the ℓ_2 norm. $\mathbf{x}_i$ and $\mathbf{x}_j$ are considered to be close if one is among the k nearest neighbors of the other, in the sense of the ℓ_2

The version of record is available at: http://dx.doi.org/10.1561/2400000026

norm, with k being pre-specified. We aim to automatically determine the weight $\mathbf{W}$ in a data-driven fashion through minimizing the distances on the set $\mathcal{M}$ and maximizing the distances on $\mathcal{N}$, which yields the following *Absolute Metric Learning* formulation:

$$\begin{aligned} \min_{\mathbf{W}\succcurlyeq 0} \quad & \sum_{(i,j)\in\mathcal{M}} s^2_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j) \\ \text{s.t.} \quad & \sum_{(i,j)\in\mathcal{N}} s^2_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j) \geq 1. \end{aligned} \tag{2.17}$$

A slightly different formulation considers the relative distance between predictors. Define a set

$$\mathcal{T} \triangleq \{(i,j,k)\text{: } s_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j) \text{ should be smaller than } s_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_k)\},$$

where $s_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)$ is considered to be smaller than $s_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_k)$ if any of the following holds:

1. $y_i = y_j$ and $y_i \neq y_k$;
2. $y_i = y_j = y_k$ and $\|\mathbf{x}_i - \mathbf{x}_j\|_2 < \|\mathbf{x}_i - \mathbf{x}_k\|_2$;
3. $y_i \neq y_j$ and $y_i \neq y_k$ and $\|\mathbf{x}_i - \mathbf{x}_j\|_2 < \|\mathbf{x}_i - \mathbf{x}_k\|_2$.

The *Relative Metric Learning* formulation minimizes the difference of distances on these triplets:

$$\min_{\mathbf{W}\succcurlyeq 0} \quad \sum_{(i,j,k)\in\mathcal{T}} \max(s^2_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j) - s^2_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_k) + 1, 0). \tag{2.18}$$

To hedge against potential noise in the predictors, [80] proposed to robustify (2.17) and (2.18) using robust optimization, and learn a robust data-driven transport cost function. Specifically, for the absolute metric learning formulation, suppose the sets $\mathcal{M}$ and $\mathcal{N}$ are noisy or inaccurate at level α, i.e., $\alpha \cdot 100\%$ of their elements are incorrectly assigned. We construct robust uncertainty sets $\mathcal{W}(\alpha)$ and $\mathcal{V}(\alpha)$ as follows:

$$\begin{aligned} \mathcal{W}(\alpha) = \Big\{ & \boldsymbol{\eta} = (\eta_{i,j};\ (i,j)\in\mathcal{M})\text{: } 0 \leq \eta_{i,j} \leq 1, \\ & \sum_{(i,j)\in\mathcal{M}} \eta_{i,j} \leq (1-\alpha)|\mathcal{M}| \Big\}, \\ \mathcal{V}(\alpha) = \Big\{ & \boldsymbol{\xi} = (\xi_{i,j};\ (i,j)\in\mathcal{N})\text{: } 0 \leq \xi_{i,j} \leq 1, \sum_{(i,j)\in\mathcal{N}} \xi_{i,j} \geq (1-\alpha)|\mathcal{N}| \Big\}. \end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

We then formulate the robust counterpart of the Absolute Metric Learning formulation (2.17) as:

$$\min_{\mathbf{W}\succcurlyeq 0}\max_{\lambda\geq 0}\max_{\substack{\boldsymbol{\eta}\in\mathcal{W}(\alpha)\\ \boldsymbol{\xi}\in\mathcal{V}(\alpha)}} \Bigg[\sum_{(i,j)\in\mathcal{M}}\eta_{i,j}s^2_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j) + \lambda\Bigg(1-\sum_{(i,j)\in\mathcal{N}}\xi_{i,j}s^2_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)\Bigg)\Bigg], \tag{2.19}$$

where we robustify the Lagrangian dual problem of (2.17), which is formed by bringing the constraint into the objective function via a dual variable λ, using uncertain parameters $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$. Similarly, for the relative metric learning formulation, suppose the set $\mathcal{T}$ is inaccurate at level α, the robust counterpart of the Relative Metric Learning formulation (2.18) can be formulated as:

$$\min_{\mathbf{W}\succcurlyeq 0}\max_{\mathbf{q}\in\mathcal{Q}(\alpha)}\sum_{(i,j,k)\in\mathcal{T}}q_{i,j,k}\max(s^2_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)-s^2_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_k)+1,0), \tag{2.20}$$

where the uncertainty set $\mathcal{Q}(\alpha)$ is defined as:

$$\mathcal{Q}(\alpha)=\Bigg\{\mathbf{q}=(q_{i,j,k};\ (i,j,k)\in\mathcal{T})\colon 0\leq q_{i,j,k}\leq 1,\ \sum_{(i,j,k)\in\mathcal{T}}q_{i,j,k}\leq(1-\alpha)|\mathcal{T}|\Bigg\}.$$

For solving the robust optimization problems (2.19) and (2.20), we refer the reader to [80] for a sequential iterative algorithm that alternates between optimizing over the weight matrix $\mathbf{W}$ and the uncertain parameters $\boldsymbol{\eta}$, $\boldsymbol{\xi}$ (or $\mathbf{q}$).

2.6 Robustness of the Wasserstein Ambiguity Set

The ultimate goal of using DRO is to eliminate the effect of perturbed samples and produce an estimator that is consistent with the underlying true (clean) distribution. When the data $\mathbf{z}=(\mathbf{x},y)$ are corrupted by outliers, the observed samples are not representative enough to encode the true underlying uncertainty of the data. Instead of equally

The version of record is available at: http://dx.doi.org/10.1561/2400000026

weighting all the samples as in the empirical distribution, we may wish to include more informative distributions that "drive out" the corrupted samples. DRO realizes this through hedging the expected loss against a family of distributions that include the true data-generating mechanism with a high confidence. In this section, we will provide evidence on the robustness of DRO under the Wasserstein metric, by showing that the ambiguity set defined via the Wasserstein metric is able to retain the good (clean) distribution while excluding the bad (outlying) one; thus, producing an estimator that is robust to outliers.

We make the assumption that the training data $(\mathbf{x}, y)$ are drawn from a mixture of two distributions, with probability q from the outlying distribution $\mathbb{P}_{\text{out}}$ and with probability $1-q$ from the true (clean) distribution $\mathbb{P}$. All the N training samples $(\mathbf{x}_i, y_i)$, $i \in [\![N]\!]$, are independent and identical realizations of $(\mathbf{x}, y)$. Recall that $\hat{\mathbb{P}}_N$ is the discrete uniform distribution over the N samples. 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 ϵ-Wasserstein ball Ω (cf. (1.6))

$$\Omega = \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{s,1}(\mathbb{Q}, \hat{\mathbb{P}}_N) \leq \epsilon\},$$

while the outlying distribution $\mathbb{P}_{\text{out}}$ will be excluded. Theorem 2.6.1 proves this claim.

Theorem 2.6.1. Suppose we are given two probability distributions $\mathbb{P}$ and $\mathbb{P}_{\text{out}}$, and the mixture distribution $\mathbb{P}_{\text{mix}}$ is a convex combination of the two: $\mathbb{P}_{\text{mix}} = q\mathbb{P}_{\text{out}} + (1-q)\mathbb{P}$. Then, for any cost function s,

$$\frac{W_{s,1}(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}})}{W_{s,1}(\mathbb{P}, \mathbb{P}_{\text{mix}})} = \frac{1-q}{q}.$$

Proof. As we indicated in Section 1, and for ease of notation, we will suppress the dependence of $W_{s,1}$ on the cost metric s. In addition, without loss of generality, we will assume that the probability distributions $\mathbb{P}$, $\mathbb{P}_{\text{out}}$, $\mathbb{P}_{\text{mix}}$, and any joint distributions have densities. From the definition of the Wasserstein distance, $W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}})$ is the optimal

The version of record is available at: http://dx.doi.org/10.1561/2400000026

value of the following optimization problem:

$$\begin{aligned}\min_{\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})}\ & \int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1,\mathbf{z}_2)\ \mathrm{d}\pi(\mathbf{z}_1,\mathbf{z}_2)\\ \text{s.t.}\ & \int_{\mathcal{Z}} \pi(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1), \quad \forall \mathbf{z}_1\in\mathcal{Z},\\ & \int_{\mathcal{Z}} \pi(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_1 = q\mathbb{P}_{\text{out}}(\mathbf{z}_2)+(1-q)\mathbb{P}(\mathbf{z}_2), \quad \forall \mathbf{z}_2\in\mathcal{Z}.\end{aligned} \tag{2.21}$$

Similarly, $W_1(\mathbb{P},\mathbb{P}_{\text{mix}})$ is the optimal value of the following optimization problem:

$$\begin{aligned}\min_{\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})}\ & \int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1,\mathbf{z}_2)\ \mathrm{d}\pi(\mathbf{z}_1,\mathbf{z}_2)\\ \text{s.t.}\ & \int_{\mathcal{Z}} \pi(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_2 = \mathbb{P}(\mathbf{z}_1), \quad \forall \mathbf{z}_1\in\mathcal{Z},\\ & \int_{\mathcal{Z}} \pi(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_1 = q\mathbb{P}_{\text{out}}(\mathbf{z}_2)+(1-q)\mathbb{P}(\mathbf{z}_2), \quad \forall \mathbf{z}_2\in\mathcal{Z}.\end{aligned} \tag{2.22}$$

We propose a decomposition strategy. For Problem (2.21), decompose the joint distribution π as $\pi = (1-q)\pi_1 + q\pi_2$, where π_1 and π_2 are two joint distributions of $\mathbf{z}_1$ and $\mathbf{z}_2$. The first set of constraints in Problem (2.21) can be equivalently expressed as:

$$\begin{aligned}&(1-q)\int_{\mathcal{Z}} \pi_1(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_2 + q\int_{\mathcal{Z}} \pi_2(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_2\\ &\quad = (1-q)\mathbb{P}_{\text{out}}(\mathbf{z}_1) + q\mathbb{P}_{\text{out}}(\mathbf{z}_1), \quad \forall \mathbf{z}_1\in\mathcal{Z},\end{aligned}$$

which is satisfied if

$$\int_{\mathcal{Z}} \pi_1(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1), \quad \int_{\mathcal{Z}} \pi_2(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1), \quad \forall \mathbf{z}_1\in\mathcal{Z}.$$

The second set of constraints can be expressed as:

$$\begin{aligned}&(1-q)\int_{\mathcal{Z}} \pi_1(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_1 + q\int_{\mathcal{Z}} \pi_2(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_1\\ &\quad = q\mathbb{P}_{\text{out}}(\mathbf{z}_2) + (1-q)\mathbb{P}(\mathbf{z}_2), \quad \forall \mathbf{z}_2\in\mathcal{Z},\end{aligned}$$

which is satisfied if

$$\int_{\mathcal{Z}} \pi_1(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_1 = \mathbb{P}(\mathbf{z}_2), \quad \int_{\mathcal{Z}} \pi_2(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_1 = \mathbb{P}_{\text{out}}(\mathbf{z}_2), \quad \forall \mathbf{z}_2\in\mathcal{Z}.$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

The objective function can be decomposed as:

$$\int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1,\mathbf{z}_2)\ \mathrm{d}\pi(\mathbf{z}_1,\mathbf{z}_2) = (1-q)\int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1,\mathbf{z}_2)\ \mathrm{d}\pi_1(\mathbf{z}_1,\mathbf{z}_2) + q\int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\pi_2(\mathbf{z}_1,\mathbf{z}_2).$$

Therefore, Problem (2.21) can be decomposed into the following two subproblems.

$$\text{Subproblem 1:}\quad \min_{\pi_1\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1,\mathbf{z}_2)\ \mathrm{d}\pi_1(\mathbf{z}_1,\mathbf{z}_2)$$
$$\text{s.t.} \int_{\mathcal{Z}} \pi_1(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1),\quad \forall \mathbf{z}_1\in\mathcal{Z},$$
$$\int_{\mathcal{Z}} \pi_1(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_1 = \mathbb{P}(\mathbf{z}_2),\quad \forall \mathbf{z}_2\in\mathcal{Z}.$$

$$\text{Subproblem 2:}\quad \min_{\pi_2\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1,\mathbf{z}_2)\ \mathrm{d}\pi_2(\mathbf{z}_1,\mathbf{z}_2)$$
$$\text{s.t.} \int_{\mathcal{Z}} \pi_2(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_2 = \mathbb{P}_{\text{out}}(\mathbf{z}_1),\quad \forall \mathbf{z}_1\in\mathcal{Z},$$
$$\int_{\mathcal{Z}} \pi_2(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\mathbf{z}_1 = \mathbb{P}_{\text{out}}(\mathbf{z}_2),\quad \forall \mathbf{z}_2\in\mathcal{Z}.$$

Assume that the optimal solutions to the two subproblems are π_1^* and π_2^*, respectively. We know $\pi_0 = (1-q)\pi_1^* + q\pi_2^*$ is a feasible solution to Problem (2.21). Therefore,

$$\begin{aligned} W_1(\mathbb{P}_{\text{out}},\mathbb{P}_{\text{mix}}) &\le \int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}_1,\mathbf{z}_2)\ \mathrm{d}\pi_0(\mathbf{z}_1,\mathbf{z}_2) \\ &= (1-q)W_1(\mathbb{P}_{\text{out}},\mathbb{P}) + qW_1(\mathbb{P}_{\text{out}},\mathbb{P}_{\text{out}}) \\ &= (1-q)W_1(\mathbb{P}_{\text{out}},\mathbb{P}). \end{aligned} \tag{2.23}$$

Similarly,

$$W_1(\mathbb{P},\mathbb{P}_{\text{mix}}) \le qW_1(\mathbb{P}_{\text{out}},\mathbb{P}). \tag{2.24}$$

(2.23) and (2.24) imply that

$$W_1(\mathbb{P}_{\text{out}},\mathbb{P}_{\text{mix}}) + W_1(\mathbb{P},\mathbb{P}_{\text{mix}}) \le W_1(\mathbb{P}_{\text{out}},\mathbb{P}).$$

On the other hand, using the triangle inequality for the Wasserstein metric, we have,

$$W_1(\mathbb{P}_{\text{out}},\mathbb{P}_{\text{mix}}) + W_1(\mathbb{P},\mathbb{P}_{\text{mix}}) \ge W_1(\mathbb{P}_{\text{out}},\mathbb{P}).$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

We thus conclude that

$$W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}}) + W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) = W_1(\mathbb{P}_{\text{out}}, \mathbb{P}). \tag{2.25}$$

To achieve the equality in (2.25), (2.23) and (2.24) must be equalities, i.e.,

$$W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}}) = (1-q)W_1(\mathbb{P}_{\text{out}}, \mathbb{P}),$$

and,

$$W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) = qW_1(\mathbb{P}_{\text{out}}, \mathbb{P}). \tag{2.26}$$

Thus,

$$\frac{W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}})}{W_1(\mathbb{P}, \mathbb{P}_{\text{mix}})} = \frac{(1-q)W_1(\mathbb{P}_{\text{out}}, \mathbb{P})}{qW_1(\mathbb{P}_{\text{out}}, \mathbb{P})} = \frac{1-q}{q}.$$

□

2.7 Setting the Radius of the Wasserstein Ball

Theorem 2.6.1 provides some guidance on setting the radius ϵ of the Wasserstein ball Ω. Figure 2.2 (Left) provides a graphical interpretation. As seen in the figure, the ball Ω is centered at $\mathbb{P}_{\text{mix}}$ because we assume that the training set is drawn from this distribution. According to Theorem 2.6.1, when $q < 0.5$ we have $W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) \leq \epsilon < W_1(\mathbb{P}_{\text{out}}, \mathbb{P}_{\text{mix}})$. Thus, for a large enough sample size (so that $\hat{\mathbb{P}}_N$ is a good approximation of $\mathbb{P}_{\text{mix}}$), the set Ω will include the true distribution and exclude the outlying one, which provides protection against these outliers.

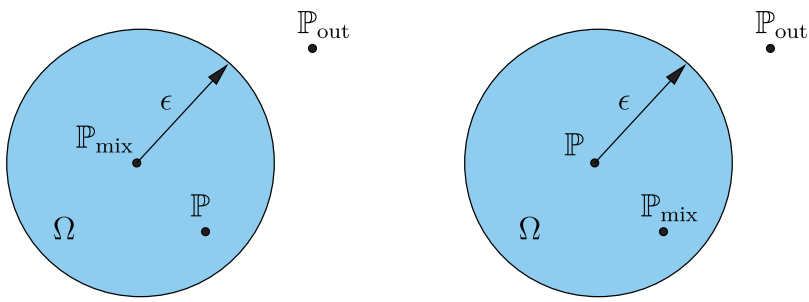


Figure 2.2: Left: Training with a contaminated training set drawn from $\mathbb{P}_{\text{mix}}$. Right: Training with a pure training set drawn from $\mathbb{P}$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

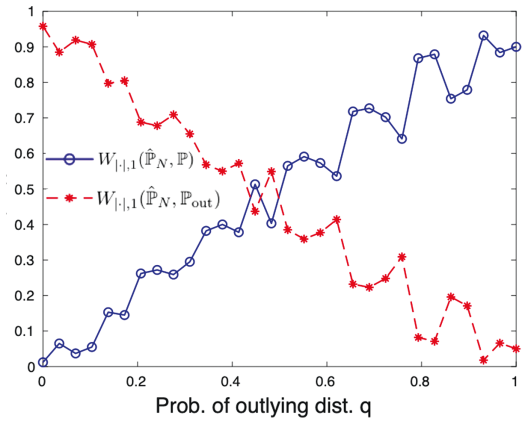


Figure 2.3: The order-1 Wasserstein distances from the empirical distribution.

To provide numerical evidence, 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 $\mathbb{P}_{\text{out}}$ assigns probability 0.5 to two data points 1 and 2. We generate 100 samples and plot the order-1 Wasserstein distances from $\hat{\mathbb{P}}_N$ for both $\mathbb{P}$ and $\mathbb{P}_{\text{out}}$, under the distance metric $s(z_1, z_2) = |z_1 - z_2|$. From Figure 2.3 we observe that for q below 0.5, the true distribution $\mathbb{P}$ is closer to $\hat{\mathbb{P}}_N$ whereas the outlying distribution $\mathbb{P}_{\text{out}}$ 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 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, if $q > 0.5$ then $\mathbb{P}_{\text{out}}$ becomes the distribution of the majority and data generated from $\mathbb{P}$ can be treated as outliers. Thus, without loss of generality, we can safely treat $\mathbb{P}_{\text{out}}$ as the distribution of the minority and assume q is always below 0.5.

An alternative use of the DRO learning approach can be seen in Figure 2.2 (Right). Here, we assume that the training set is pure, thus,

The version of record is available at: http://dx.doi.org/10.1561/2400000026

given enough samples, the empirical distribution on which the ball Ω is centered is close to $\mathbb{P}$. Consider applying the model to a test set which is contaminated with outliers. Notice from the proof of Theorem 2.6.1 that $W_1(\mathbb{P}, \mathbb{P}_{\text{mix}}) = qW_1(\mathbb{P}_{\text{out}}, \mathbb{P})$ (cf. Equation (2.26)). This implies that the smaller q is, and for a properly selected ϵ, the distribution from which the test set is drawn ($\mathbb{P}_{\text{mix}}$) is within the ball Ω and the model has the potential to generalize well in the test set, tolerating some outliers. In contrast, the outlying distribution $\mathbb{P}_{\text{out}}$ lies outside the set Ω, which suggests that the model does not "adjust" to samples generated from $\mathbb{P}_{\text{out}}$. According to this reasoning, and based on Equation (2.26), ϵ should be set so that $qW_1(\mathbb{P}_{\text{out}}, \mathbb{P}) < \epsilon < W_1(\mathbb{P}_{\text{out}}, \mathbb{P})$.

The above discussions provide some insights on the optimal selection of the radius, but could be hard to implement due to the unknown $\mathbb{P}$ and $\mathbb{P}_{\text{out}}$. In practice cross-validation is usually adopted, but could be computationally expensive. In the next two subsections we discuss two practical radius selection approaches that produce the smallest Wasserstein ball which contains the true distribution with high confidence.

2.7.1 Measure Concentration

In this subsection we study an optimal radius selection method that originates from the measure concentration theory. As will be seen in Section 3.4, it leads to an asymptotic consistent DRO estimator that generalizes well out-of-sample.

Suppose $\mathbf{z}_i,\ i \in [\![N]\!]$, are N realizations of $\mathbf{z}$ which follows an unknown distribution $\mathbb{P}^*$. One of the prerequisites for ensuring a good generalization performance of Wasserstein DRO requires that the ambiguity set $\Omega_\epsilon(\hat{\mathbb{P}}_N)$ includes the true data distribution $\mathbb{P}^*$. This implies that the radius ϵ should be chosen so that

$$W_{s,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \leq \epsilon. \tag{2.27}$$

A measure concentration result developed in [66], which characterizes the rate at which the empirical distribution $\hat{\mathbb{P}}_N$ converges to the true distribution $\mathbb{P}^*$ in the sense of the Wasserstein metric, can be used as a guidance on the optimal selection of the radius for the Wasserstein ambiguity set. In the following discussion we assume s is a norm, and

The version of record is available at: http://dx.doi.org/10.1561/2400000026

the true data distribution $\mathbb{P}^*$ satisfies the light tail condition stated in Assumption B.

Assumption B (Light-Tailed Distribution)**.** There exists an exponent $a > 1$ such that

$$A \triangleq \mathbb{E}^{\mathbb{P}^*}[\exp(\|\mathbf{z}\|^a)] = \int_{\mathcal{Z}} \exp(\|\mathbf{z}\|^a) \mathrm{d}\mathbb{P}^*(\mathbf{z}) < \infty. \tag{2.28}$$

Theorem 2.7.1 (Measure Concentration; [66], Theorem 2)**.** Suppose the Wasserstein metric is induced by some norm $\|\cdot\|$, i.e., $s(\mathbf{z}_1, \mathbf{z}_2) = \|\mathbf{z}_1 - \mathbf{z}_2\|$. Under Assumption B, we have

$$\mathbb{P}^N(W_{s,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \geq \epsilon) \leq \begin{cases} c_1 \exp(-c_2 N \epsilon^{\max(d,2)}), & \text{if } \epsilon \leq 1, \\ c_1 \exp(-c_2 N \epsilon^a), & \text{if } \epsilon > 1, \end{cases} \tag{2.29}$$

for all $N \geq 1, d \neq 2$, and $\epsilon > 0$, where N is the size of the observed training set, d is the dimension of $\mathbf{z}$, a is defined in (2.28), and c_1, c_2 are positive constants that only depend on a, A, and d.

From Theorem 2.7.1 we can derive the smallest possible ϵ so that the true distribution is contained in the Wasserstein ambiguity set with high confidence. Given some prescribed $\alpha \in (0, 1)$, it is desired that

$$\mathbb{P}^N(W_{s,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \leq \epsilon) \geq 1 - \alpha.$$

Equating the RHS of (2.29) to α and solving for ϵ yields

$$\epsilon_N(\alpha) = \begin{cases} \left(\dfrac{\log(c_1 \alpha^{-1})}{c_2 N}\right)^{1/\max(d,2)}, & \text{if } N \geq \dfrac{\log(c_1 \alpha^{-1})}{c_2}, \\ \left(\dfrac{\log(c_1 \alpha^{-1})}{c_2 N}\right)^{1/a}, & \text{if } N < \dfrac{\log(c_1 \alpha^{-1})}{c_2}. \end{cases} \tag{2.30}$$

Notice that Equation (2.30) depends on the unknown constants c_1 and c_2, and does not make use of the available training data, which could potentially result in a conservative estimation of the radius and is not of practical use [13]. By recognizing these issues, some researchers have proposed to choose the radius without relying on exogenous constants, see [57] and [81].

By using an extension of Sanov's theorem which identifies the rate function, in the form of the KL divergence, for large deviations of the

The version of record is available at: http://dx.doi.org/10.1561/2400000026

empirical measure from the true measure, [82], [81] derived a closed-form expression for computing the size of the Wasserstein ambiguity set, when the support of $\mathbf{z}$ is finite and bounded, and the true distribution is discrete. The reason for restricting to a discrete true distribution lies in that the convergence rate of the empirical measure (in the sense of the Wasserstein distance) is characterized by the KL divergence [83], which diverges when the true distribution $\mathbb{P}^*$ is continuous, and the empirical distribution $\hat{\mathbb{P}}_N$ is discrete.

Theorem 2.7.2 ([81], Theorem 2)**.** Suppose the random vector $\mathbf{z}$ is supported on a finite Polish space $(\mathcal{Z}, s)$, and is distributed according to a discrete true distribution $\mathbb{P}^*$. Assume there exists some $\mathbf{z}_0 \in \mathcal{Z}$ such that the following condition holds:

$$\log \int_{\mathcal{Z}} \exp\left(as(\mathbf{z}, \mathbf{z}_0)\right) \mathrm{d}\mathbb{P}^*(\mathbf{z}) < \infty, \quad \forall a > 0. \tag{2.31}$$

Define B as the diameter of the d-dimensional compact set $\mathcal{Z}$:

$$B \triangleq \sup\{s(\mathbf{z}_1, \mathbf{z}_2) \colon \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}\}. \tag{2.32}$$

Construct an empirical distribution $\hat{\mathbb{P}}_N$ based on N i.i.d. samples of $\mathbf{z}$. A lower bound on the probability that the Wasserstein distance between the empirical distribution $\hat{\mathbb{P}}_N$ and the true distribution $\mathbb{P}^*$ does not exceed ϵ is given by:

$$\begin{aligned}&\mathbb{P}^N(W_{s,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \leq \epsilon) \\ &\quad \geq 1 - \exp\left(-N\left(\frac{\sqrt{4\epsilon(4B+3) + (4B+3)^2}}{4B+3} - 1\right)^2\right).\end{aligned}$$

Furthermore, if

$$\epsilon \geq \left(B + \frac{3}{4}\right)\left(-\frac{1}{N}\log(\alpha) + 2\sqrt{-\frac{1}{N}\log(\alpha)}\right),$$

then

$$\mathbb{P}^N(W_{s,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \leq \epsilon) \geq 1 - \alpha.$$

[57] derived a more general formula for computing the Wasserstein set radius, without imposing the exponential integrability condition (2.31), resulting in a slower convergence rate for the radius, $\epsilon = O(\sqrt{1/N})$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Theorem 2.7.3 ([57], Proposition 3)**.** Assume the support $\mathcal{Z}$ is bounded and finite, and the true distribution $\mathbb{P}^*$ is discrete. We have,

$$\mathbb{P}^N(W_{s,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \leq \epsilon) \geq 1 - \exp\left(-\frac{N\epsilon^2}{2B^2}\right),$$

where B is as in (2.32). Moreover, if we set

$$\epsilon \geq B\sqrt{\frac{2\log(1/\alpha)}{N}},$$

then

$$\mathbb{P}^N(W_{s,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \leq \epsilon) \geq 1 - \alpha.$$

2.7.2 Robust Wasserstein Profile Inference

In this subsection we introduce a different approach proposed by [64] for optimally selecting the size of the Wasserstein ambiguity set. This method combines the information of the structure of the ambiguity set and the loss function that is being minimized. Unlike Section 2.7.1 where large deviation theory is adopted to describe the closeness between the empirical measure and the true measure, here the true measure is characterized indirectly via the first-order optimality condition of the loss function.

Recall the Wasserstein DRO formulation:

$$\inf_{\beta} \sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h_\beta(\mathbf{z})], \tag{2.33}$$

where the ambiguity set is defined as:

$$\Omega = \Omega_\epsilon^{s,t}(\hat{\mathbb{P}}_N) \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{s,t}(\mathbb{Q},\ \hat{\mathbb{P}}_N) \leq \epsilon\}.$$

We will suppress the dependence of Ω on $s, t, \epsilon, \hat{\mathbb{P}}_N$ for ease of notation. For every $\mathbb{Q} \in \Omega$, there is an optimal choice $\boldsymbol{\beta} = \boldsymbol{\beta}(\mathbb{Q})$ which minimizes the risk $\mathbb{E}^{\mathbb{Q}}[h_\beta(\mathbf{z})]$, i.e.,

$$\boldsymbol{\beta}(\mathbb{Q}) = \arg\min_{\beta} \mathbb{E}^{\mathbb{Q}}[h_\beta(\mathbf{z})].$$

We define $\mathcal{S}_\beta(\Omega) \triangleq \{\boldsymbol{\beta}(\mathbb{Q})\colon \mathbb{Q} \in \Omega\}$ to be the set of plausible selections of the parameter $\boldsymbol{\beta}$. If the true measure $\mathbb{P}^* \in \Omega$, then $\boldsymbol{\beta}^* = \boldsymbol{\beta}(\mathbb{P}^*) \in \mathcal{S}_\beta(\Omega)$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

We say that $\boldsymbol{\beta}^*$ is *plausible* with $(1-\alpha)$ confidence if $\boldsymbol{\beta}^* \in \mathcal{S}_{\boldsymbol{\beta}}(\Omega)$ with probability at least $1-\alpha$. We want to choose $\epsilon > 0$ as small as possible so that the underlying true parameter $\boldsymbol{\beta}^*$ is plausible with $(1-\alpha)$ confidence.

For any given $\mathbb{Q}$, the optimal solution $\boldsymbol{\beta}(\mathbb{Q})$ is characterized by the following first-order condition:

$$\mathbb{E}^{\mathbb{Q}}[\nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}(\mathbb{Q})}(\mathbf{z})] = \mathbf{0}, \tag{2.34}$$

where $\nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}(\mathbb{Q})}(\mathbf{z})$ is the partial derivative of $h_{\boldsymbol{\beta}}(\mathbf{z})$ w.r.t. $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}(\mathbb{Q})$. Define the *Robust Wasserstein Profile (RWP)* function associated with the estimation Equation (2.34) as:

$$R(\boldsymbol{\beta}) = \inf_{\mathbb{Q}}\{(W_{s,t}(\mathbb{Q}, \hat{\mathbb{P}}_N))^t \colon \mathbb{E}^{\mathbb{Q}}[\nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}}(\mathbf{z})] = \mathbf{0}\}.$$

$R(\boldsymbol{\beta})$ evaluates the minimal distance to the empirical distribution, for all distributions such that $\boldsymbol{\beta}$ is the minimizer of the expected loss. Note that $R(\boldsymbol{\beta})$ is a random quantity due to the randomness in the observed samples, which is reflected in $\hat{\mathbb{P}}_N$. For $\boldsymbol{\beta}^* \in \mathcal{S}_{\boldsymbol{\beta}}(\Omega)$ to hold, it is required that there exists at least one

$$\mathbb{Q} \in \{\mathbb{Q} \colon \mathbb{E}^{\mathbb{Q}}[\nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}^*}(\mathbf{z})] = \mathbf{0}\},$$

such that $\mathbb{Q} \in \Omega$. This equivalently translates into the condition that

$$R(\boldsymbol{\beta}^*) \leq \epsilon^t.$$

Therefore, $\boldsymbol{\beta}^*$ is plausible with $(1-\alpha)$ confidence if and only if

$$\mathbb{P}(R(\boldsymbol{\beta}^*) \leq \epsilon^t) \geq 1-\alpha.$$

The optimal choice of ϵ is thus $\chi_{1-\alpha}^{1/t}$, where $\chi_{1-\alpha}$ is the $1-\alpha$ quantile of $R(\boldsymbol{\beta}^*)$. Moreover,

$$\mathbb{P}(\boldsymbol{\beta}^* \in \mathcal{S}_{\boldsymbol{\beta}}(\chi_{1-\alpha})) = \mathbb{P}(R(\boldsymbol{\beta}^*) \leq \chi_{1-\alpha}) = 1-\alpha,$$

where $\mathcal{S}_{\boldsymbol{\beta}}(\chi_{1-\alpha}) \triangleq \{\boldsymbol{\beta} \colon R(\boldsymbol{\beta}) \leq \chi_{1-\alpha}\}$. Therefore, $\mathcal{S}_{\boldsymbol{\beta}}(\chi_{1-\alpha})$ is a $(1-\alpha)$ confidence region for $\boldsymbol{\beta}^*$.

The problem of optimal radius selection now reduces to finding the quantile of $R(\boldsymbol{\beta}^*)$. Since $\boldsymbol{\beta}^*$ is unknown, we need to come up with a way of estimating the distribution of the RWP function $R(\boldsymbol{\beta}^*)$. [64] developed

The version of record is available at: http://dx.doi.org/10.1561/2400000026

an asymptotic analysis of the RWP function, and established that as $N \to \infty$,

$$N^{t/2} R(\boldsymbol{\beta}^*) \xrightarrow{\mathrm{d}} \bar{R}(t),$$

for a suitably defined random variable $\bar{R}(t)$, where $\xrightarrow{\mathrm{d}}$ means convergence in distribution. We first state a number of assumptions that are needed to establish this convergence in distribution.

Assumption C. The cost function is the ℓ_r norm: $s(\mathbf{z}_1, \mathbf{z}_2) = \|\mathbf{z}_1 - \mathbf{z}_2\|_r$, where $r \geq 1$. Let s be such that $1/r + 1/s = 1$.

Assumption D. The true parameter $\boldsymbol{\beta}^*$ satisfies $\mathbb{E}^{\mathbb{P}^*}[\nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}^*}(\mathbf{z})] = \mathbf{0}$, and $\mathbb{E}^{\mathbb{P}^*} \|\nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}^*}(\mathbf{z})\|_2^2 < \infty$, where $\mathbb{P}^*$ is the underlying true distribution of $\mathbf{z}$.

Assumption E. The function $\nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}^*}(\mathbf{z})$ is continuously differentiable w.r.t. $\mathbf{z}$ with gradient $\nabla_{\boldsymbol{\beta},\mathbf{z}} h_{\boldsymbol{\beta}^*}(\mathbf{z})$.

Assumption F.

$$\mathbb{E}^{\mathbb{P}^*}[\nabla_{\boldsymbol{\beta},\mathbf{z}} h_{\boldsymbol{\beta}^*}(\mathbf{z}) \nabla_{\boldsymbol{\beta},\mathbf{z}} h_{\boldsymbol{\beta}^*}(\mathbf{z})'] \succ 0.$$

Assumption G. There exists $\kappa > 0$ such that for $\|\mathbf{z}\|_r \geq 1$,

$$\|\nabla_{\boldsymbol{\beta},\mathbf{z}} h_{\boldsymbol{\beta}^*}(\mathbf{z})\|_s \leq \kappa \|\mathbf{z}\|_r^{t-1},$$

where the LHS denotes the induced ℓ_s norm of the matrix $\nabla_{\boldsymbol{\beta},\mathbf{z}} h_{\boldsymbol{\beta}^*}(\mathbf{z})$.

Assumption H. There exists a function c: $\mathbb{R}^d \to [0, \infty)$ such that,

$$\|\nabla_{\boldsymbol{\beta},\mathbf{z}} h_{\boldsymbol{\beta}^*}(\mathbf{z} + \boldsymbol{\delta}) - \nabla_{\boldsymbol{\beta},\mathbf{z}} h_{\boldsymbol{\beta}^*}(\mathbf{z})\|_s \leq c(\mathbf{z}) \|\boldsymbol{\delta}\|_r,$$

for $\|\boldsymbol{\delta}\|_r \leq 1$, $\mathbb{E}^{\mathbb{P}^*}[c(\mathbf{z})^a] < \infty$, and $a \leq \max(2, t/(t-1))$.

Theorem 2.7.4 ([64], Theorem 3)**.** When $t > 1$, under Assumptions C–H, as $N \to \infty$,

$$N^{t/2} R(\boldsymbol{\beta}^*) \xrightarrow{\mathrm{d}} \bar{R}(t),$$

where

$$\bar{R}(t) = \max_{\boldsymbol{\zeta}} \{ t \boldsymbol{\zeta}' \mathbf{r} - (t-1) \mathbb{E}^{\mathbb{P}^*} \|\boldsymbol{\zeta}' \nabla_{\boldsymbol{\beta},\mathbf{z}} h_{\boldsymbol{\beta}^*}(\mathbf{z})\|_s^{t/(t-1)} \}.$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

When $t = 1$, suppose that $\mathbf{z}$ has a positive density almost everywhere w.r.t. the Lebesgue measure. Then, under Assumptions C–F,

$$N^{1/2} R(\boldsymbol{\beta}^*) \xrightarrow{\text{d}} \bar{R}(1),$$

where

$$\begin{aligned} \bar{R}(1) = \max_{\boldsymbol{\zeta}} \ & \boldsymbol{\zeta}'\mathbf{r} \\ \text{s.t.} \ & \mathbb{P}^*(\|\boldsymbol{\zeta}' \nabla_{\beta,\mathbf{z}} h_{\beta^*}(\mathbf{z})\|_s > 1) = 0, \end{aligned}$$

with $\mathbf{r} \sim \mathcal{N}(\mathbf{0}, \mathbb{E}[\nabla_\beta h_{\beta^*}(\mathbf{z}) \nabla_\beta h_{\beta^*}(\mathbf{z})'])$.

The proof of Theorem 2.7.4 uses the dual representation of the RWP function, and proceeds by showing that $\bar{R}(t)$ is both an asymptotic stochastic upper bound and a lower bound of $N^{t/2} R(\boldsymbol{\beta}^*)$ (refer to [64] for details). Notice that the limiting random variable $\bar{R}(t)$ still depends on the unknown parameter $\boldsymbol{\beta}^*$ and the unobservable true distribution $\mathbb{P}^*$. When using Theorem 2.7.4 in practice, some further relaxations for $\bar{R}(t)$ are needed to get rid of the unknown parameters. We next illustrate this idea using an example of distributionally robust logistic regression.

Example: Optimal Radius Selection for Wasserstein Distributionally Robust Logistic Regression Using RWP Inference

In this example we show how to use the RWP function and its limiting variable $\bar{R}(t)$ to select the optimal radius for the Wasserstein ambiguity set in a distributionally robust logistic regression problem.

Let $\mathbf{x} \in \mathbb{R}^d$ denote the predictor and $y \in \{-1, +1\}$ the associated binary label to be predicted. In logistic regression, the conditional distribution of y given $\mathbf{x}$ is modeled as

$$\mathbb{P}(y|\mathbf{x}) = (1 + \exp(-y\boldsymbol{\beta}'\mathbf{x}))^{-1},$$

where $\boldsymbol{\beta}$ is the unknown coefficient vector (classifier) to be estimated. The *Maximum Likelihood Estimator (MLE)* of $\boldsymbol{\beta}$ is found by minimizing the *negative log-likelihood (logloss)*

$$h_\beta(\mathbf{x}, y) = \log(1 + \exp(-y\boldsymbol{\beta}'\mathbf{x})).$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

We define the distance metric on the predictor-response space as follows.

$$s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)) \triangleq \begin{cases} \|\mathbf{x}_1 - \mathbf{x}_2\|_r, & \text{if } y_1 = y_2, \\ \infty, & \text{otherwise.} \end{cases} \tag{2.35}$$

The distributionally robust logistic regression problem is formulated as:

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}}[\log(1 + \exp(-y\boldsymbol{\beta}'\mathbf{x}))], \tag{2.36}$$

where the order-1 Wasserstein metric is used to define the set Ω:

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{s,1}(\mathbb{Q}, \hat{\mathbb{P}}_N) \leq \epsilon\}.$$

We apply Theorem 2.7.4 with $t = 1$ to derive the optimal radius ϵ. Note that

$$\nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}}(\mathbf{x}, y) = \frac{-y\mathbf{x}}{1 + \exp(y\boldsymbol{\beta}'\mathbf{x})}.$$

Then, for $t = 1$, as $N \to \infty$,

$$\sqrt{N} R(\boldsymbol{\beta}^*) \xrightarrow{\text{d}} \bar{R}(1),$$

where

$$\bar{R}(1) = \sup_{\boldsymbol{\zeta} \in \mathcal{A}} \boldsymbol{\zeta}'\mathbf{r},$$

and

$$\mathbf{r} \sim \mathcal{N}\left(\mathbf{0}, \mathbb{E}^{\mathbb{P}^*}\left[\frac{\mathbf{x}\mathbf{x}'}{(1 + \exp(y\mathbf{x}'\boldsymbol{\beta}^*))^2}\right]\right),$$

$$\mathcal{A} \triangleq \left\{\boldsymbol{\zeta} \in \mathbb{R}^d\colon \sup_{(\mathbf{x}, y)} \|\boldsymbol{\zeta}'\nabla_{\boldsymbol{\beta}, \mathbf{x}} h_{\boldsymbol{\beta}^*}(\mathbf{x}, y)\|_s \leq 1\right\},$$

where s satisfies $1/r + 1/s = 1$.

Note that $\bar{R}(1)$ still depends on $\boldsymbol{\beta}^*$ and $\mathbb{P}^*$ which are both unknown. We need to find a stochastic upper bound of $\bar{R}(1)$ (for a conservative selection of the radius) that is independent of the unknown quantities. By noting that $\mathcal{A}$ is a subset of

$$\{\boldsymbol{\zeta} \in \mathbb{R}^d\colon \|\boldsymbol{\zeta}\|_s \leq 1\},$$

and that

$$\mathbb{E}^{\mathbb{P}^*_{\mathcal{X}}}[\mathbf{x}\mathbf{x}'] - \mathbb{E}^{\mathbb{P}^*}\left[\frac{\mathbf{x}\mathbf{x}'}{(1 + \exp(y\mathbf{x}'\boldsymbol{\beta}^*))^2}\right]$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

is positive definite, where $\mathbb{P}^*_{\mathcal{X}}$ denotes the marginal distribution of $\mathbf{x}$ under $\mathbb{P}^*$, we have:

$$\bar{R}(1) \overset{\text{D}}{\leq} \|\tilde{\mathbf{r}}\|_r,$$

where $\tilde{\mathbf{r}} \sim \mathcal{N}(\mathbf{0}, \mathbb{E}^{\mathbb{P}^*_{\mathcal{X}}}[\mathbf{x}\mathbf{x}'])$, and $\overset{\text{D}}{\leq}$ denotes stochastic dominance.

The size of the Wasserstein ambiguity set for distributionally robust logistic regression can thus be chosen by the following procedure.

1. Estimate the $(1-\alpha)$ quantile of $\|\tilde{\mathbf{r}}\|_r$, where $\tilde{\mathbf{r}} \sim \mathcal{N}(\mathbf{0}, \mathbb{E}^{\mathbb{P}^*_{\mathcal{X}}}[\mathbf{x}\mathbf{x}'])$. Denote the estimated quantile by $\hat{\chi}_{1-\alpha}$.

2. Choose the radius ϵ to be $\epsilon = \hat{\chi}_{1-\alpha}/\sqrt{N}$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

3

Solving the Wasserstein DRO Problem

In this section we discuss how to solve the Wasserstein DRO problem, as well as the performance of the DRO estimator. A Lagrangian dual method is presented in Section 3.1, for a DRO model with an ambiguity set centered at a general nominal distribution. Section 3.2 discusses the existence and the structure of the extreme distribution that achieves the optimal value of the inner maximization problem of DRO. In Section 3.3, we apply the dual method to DRO models with an ambiguity set centered at the discrete empirical distribution. Sections 3.4 and 3.5 study the finite sample and asymptotic performance of the DRO estimator, respectively.

3.1 Dual Method

The main obstacle to solving the DRO problem (1.5) lies in the inner infinite dimensional maximization problem

$$\sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h(\mathbf{z})], \tag{3.1}$$

where we suppress the dependence of h on $\boldsymbol{\beta}$ for ease of notation, and the ambiguity set is defined as:

$$\Omega = \Omega_{\epsilon}^{s,t}(\mathbb{P}_0) \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{s,t}(\mathbb{Q},\ \mathbb{P}_0) \leq \epsilon\}.$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

We will suppress the dependence of Ω on $s, t, \epsilon, \mathbb{P}_0$ for notational convenience. To transform Problem (3.1) into a finite dimensional problem, researchers have resorted to Lagrangian duality, see [13], [14]. Write Problem (3.1) in the following form:

$$\text{Primal:} \qquad v_P = \sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} h(\mathbf{z}) \mathrm{d}\mathbb{Q}(\mathbf{z}) \colon W_{s,t}(\mathbb{Q}, \mathbb{P}_0) \leq \epsilon \right\}. \tag{3.2}$$

[14] derived the Lagrangian dual of (3.2) as follows:

$$\text{Dual:} \qquad v_D = \inf_{\lambda\geq 0} \left\{ \lambda\epsilon^t - \int_{\mathcal{Z}} \inf_{\mathbf{z}\in\mathcal{Z}} [\lambda s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z})] \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \right\}, \tag{3.3}$$

when the *growth rate* of the loss function $h(\mathbf{z})$, which, given an unbounded set $\mathcal{Z}$ and a fixed $\mathbf{z}_0 \in \mathcal{Z}$, is defined as:

$$\mathrm{GR}_h \triangleq \limsup_{s(\mathbf{z},\mathbf{z}_0)\to\infty} \frac{h(\mathbf{z}) - h(\mathbf{z}_0)}{s^t(\mathbf{z}, \mathbf{z}_0)}, \tag{3.4}$$

is finite. Note that if $\mathcal{Z}$ is bounded, by convention we set $\mathrm{GR}_h = 0$. The value of GR_h does not depend on the choice of $\mathbf{z}_0$ [14].

Remark: Define a function

$$\phi(\lambda, \mathbf{z}_0) \triangleq \inf_{\mathbf{z}\in\mathcal{Z}} [\lambda s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z})].$$

The dual objective function

$$v_D(\lambda) \triangleq \lambda\epsilon^t - \int_{\mathcal{Z}} \phi(\lambda, \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0), \quad \lambda \geq 0,$$

is the sum of a linear function and an extended real-valued convex function $-\int_{\mathcal{Z}} \phi(\lambda, \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0)$. The convexity comes from the concavity of $\phi(\lambda, \mathbf{z}_0)$ w.r.t. λ. To see this, for $q \in [0, 1]$, and a fixed $\mathbf{z}_0 \in \mathcal{Z}$,

$$\begin{aligned}
&\phi(q\lambda_1 + (1-q)\lambda_2, \mathbf{z}_0) \\
&\quad = \inf_{\mathbf{z}\in\mathcal{Z}} [(q\lambda_1 + (1-q)\lambda_2) s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z})] \\
&\quad = (q\lambda_1 + (1-q)\lambda_2) s^t(\mathbf{z}^*, \mathbf{z}_0) - h(\mathbf{z}^*) \\
&\quad = q[\lambda_1 s^t(\mathbf{z}^*, \mathbf{z}_0) - h(\mathbf{z}^*)] + (1-q)[\lambda_2 s^t(\mathbf{z}^*, \mathbf{z}_0) - h(\mathbf{z}^*)] \\
&\quad \geq q\phi(\lambda_1, \mathbf{z}_0) + (1-q)\phi(\lambda_2, \mathbf{z}_0),
\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where the first step uses the definition of ϕ, $\mathbf{z}^* = \arg\min_{\mathbf{z}\in\mathcal{Z}}[(q\lambda_1 + (1-q)\lambda_2)s^t(\mathbf{z},\mathbf{z}_0) - h(\mathbf{z})]$, and the last step is due to the fact that $\mathbf{z}^* \in \mathcal{Z}$.

Thus, $v_D(\lambda)$ is a convex function on $[0,\infty)$. Moreover, as $\lambda \to \infty$, $v_D(\lambda) \to \infty$, since $v_D(\lambda) \geq \lambda\epsilon^t + \int_{\mathbf{z}_0\in\mathcal{Z}} h(\mathbf{z}_0)\mathrm{d}\mathbb{P}_0(\mathbf{z}_0)$, where the RHS is obtained through taking $\mathbf{z} = \mathbf{z}_0$ in the definition of ϕ.

To see the necessity of having a finite growth rate, note that to ensure Problem (3.1) has a finite optimal value, it is required that

$$\mathbb{E}^{\mathbb{Q}}[h(\mathbf{z})] < \infty, \quad \forall \mathbb{Q} \in \Omega.$$

This can be equivalently expressed as

$$|\mathbb{E}^{\mathbb{Q}}[h(\mathbf{z})] - \mathbb{E}^{\mathbb{P}_0}[h(\mathbf{z})]| < \infty, \quad \forall \mathbb{Q} \in \Omega. \tag{3.5}$$

The following Theorem 3.1.1 implies that, if the growth rate of h is infinite, (3.5) will be violated. Moreover, as we will see later, when the growth rate of the loss function is infinite, strong duality for Problem (3.2) fails to hold, in which case the DRO problem becomes intractable. In the sequel, we assume h is upper semi-continuous and $\mathrm{GR}_h < \infty$.

Theorem 3.1.1. Suppose a function $h\colon \mathcal{Z} \to \mathbb{R}$ defined on two metric spaces $(\mathcal{Z}, s)$ and $(\mathbb{R}, |\cdot|)$, has a finite growth rate:

$$\frac{|h(\mathbf{z}_1) - h(\mathbf{z}_2)|}{s^t(\mathbf{z}_1,\mathbf{z}_2)} \leq L, \qquad \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}.$$

Then, for any two distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ supported on $\mathcal{Z}$,

$$|\mathbb{E}^{\mathbb{Q}_1}[h(\mathbf{z})] - \mathbb{E}^{\mathbb{Q}_2}[h(\mathbf{z})]| \leq LW^t_{s,t}(\mathbb{Q}_1,\mathbb{Q}_2).$$

Proof.

$$\begin{aligned}
&|\mathbb{E}^{\mathbb{Q}_1}[h(\mathbf{z})] - \mathbb{E}^{\mathbb{Q}_2}[h(\mathbf{z})]| \\
&\quad = \left| \int_{\mathcal{Z}} h(\mathbf{z}_1)\mathrm{d}\mathbb{Q}_1(\mathbf{z}_1) - \int_{\mathcal{Z}} h(\mathbf{z}_2)\mathrm{d}\mathbb{Q}_2(\mathbf{z}_2) \right| \\
&\quad = \left| \int_{\mathcal{Z}} h(\mathbf{z}_1) \int_{\mathbf{z}_2\in\mathcal{Z}} \mathrm{d}\pi_0(\mathbf{z}_1,\mathbf{z}_2) - \int_{\mathcal{Z}} h(\mathbf{z}_2) \int_{\mathbf{z}_1\in\mathcal{Z}} \mathrm{d}\pi_0(\mathbf{z}_1,\mathbf{z}_2) \right| \\
&\quad \leq \int_{\mathcal{Z}\times\mathcal{Z}} |h(\mathbf{z}_1) - h(\mathbf{z}_2)| \mathrm{d}\pi_0(\mathbf{z}_1,\mathbf{z}_2)
\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

$$
\begin{aligned}
&= \int_{\mathcal{Z}\times\mathcal{Z}} \frac{|h(\mathbf{z}_1) - h(\mathbf{z}_2)|}{s^t(\mathbf{z}_1, \mathbf{z}_2)} s^t(\mathbf{z}_1, \mathbf{z}_2) \mathrm{d}\pi_0(\mathbf{z}_1, \mathbf{z}_2) \\
&\leq \int_{\mathcal{Z}\times\mathcal{Z}} L s^t(\mathbf{z}_1, \mathbf{z}_2) \mathrm{d}\pi_0(\mathbf{z}_1, \mathbf{z}_2) \\
&= L W_{s,t}^t(\mathbb{Q}_1, \mathbb{Q}_2),
\end{aligned}
$$

where π_0 is the joint distribution of $\mathbf{z}_1$ and $\mathbf{z}_2$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$ that achieves the optimal value of (1.7). □

3.1.1 Weak Duality

The following Theorem 3.1.2 establishes weak duality for Problem (3.2). Later we will show that strong duality also holds, i.e., $v_P = v_D$.

Theorem 3.1.2 ([14], Proposition 1)**.** Suppose the loss function h has a finite growth rate: $\mathrm{GR}_h < \infty$. Then $v_P \leq v_D$, where v_P and v_D are defined in (3.2) and (3.3), respectively.

Proof. By weak duality, we have that:

$$
v_P \leq \inf_{\lambda \geq 0} \left\{ \lambda \epsilon^t + \sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} h(\mathbf{z}) \mathrm{d}\mathbb{Q}(\mathbf{z}) - \lambda W_{s,t}^t(\mathbb{Q}, \mathbb{P}_0) \right\} \right\}, \tag{3.6}
$$

where the RHS is the Lagrangian dual of (3.2). Using Kantorovich duality (2.12), we obtain

$$
\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} h(\mathbf{z}) \mathrm{d}\mathbb{Q}(\mathbf{z}) - \lambda W_{s,t}^t(\mathbb{Q}, \mathbb{P}_0) \right\} \\
&\quad = \sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} h(\mathbf{z}) \mathrm{d}\mathbb{Q}(\mathbf{z}) \right. \\
&\qquad - \lambda \sup_{f,g} \left\{ \int_{\mathcal{Z}} f(\mathbf{z}) \mathrm{d}\mathbb{Q}(\mathbf{z}) + \int_{\mathcal{Z}} g(\mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) : \right. \\
&\qquad\qquad \left. \left. g(\mathbf{z}_0) \leq \inf_{\mathbf{z}\in\mathcal{Z}} [s^t(\mathbf{z}, \mathbf{z}_0) - f(\mathbf{z})],\ \forall \mathbf{z}_0 \in \mathcal{Z} \right\} \right\} \\
&\quad \leq - \int_{\mathcal{Z}} \inf_{\mathbf{z}\in\mathcal{Z}} [\lambda s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z})] \mathrm{d}\mathbb{P}_0(\mathbf{z}_0),
\end{aligned}
$$

where the second inequality is obtained through setting $f(\mathbf{z}) = h(\mathbf{z})/\lambda$, for $\lambda > 0$, which is absolutely integrable due to $\mathrm{GR}_h < \infty$, and is thus a feasible solution to the inner supremum of the second line. For $\lambda = 0$,

The version of record is available at: http://dx.doi.org/10.1561/2400000026

the inequality also holds since,

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})}\left\{\int_{\mathcal{Z}} h(\mathbf{z})\mathrm{d}\mathbb{Q}(\mathbf{z})\right\} \leq \sup_{\mathbf{z}\in\mathcal{Z}} h(\mathbf{z}) = -\int_{\mathcal{Z}} \inf_{\mathbf{z}\in\mathcal{Z}}[-h(\mathbf{z})]\mathrm{d}\mathbb{P}_0(\mathbf{z}_0).$$

Combining with (3.6) we arrive at the conclusion that $v_P \leq v_D$. □

3.1.2 Strong Duality

We next show that $v_P = v_D$, through constructing a feasible solution to the primal problem (3.2) whose objective function value coincides with the dual objective. We first define the push-forward measure that will be used to construct a primal feasible distribution $\mathbb{Q}^*$.

Definition 1 (Push-Forward Measure)**.** Given measurable spaces $\mathcal{Z}$ and $\mathcal{Z}'$, a measurable function $T\colon \mathcal{Z} \to \mathcal{Z}'$, and a measure $\mathbb{P} \in \mathcal{B}(\mathcal{Z})$, define the push-forward measure of $\mathbb{P}$ through T, denoted by $T_{\#\mathbb{P}} \in \mathcal{B}(\mathcal{Z}')$, as

$$T_{\#\mathbb{P}}(\mathcal{A}) \triangleq \mathbb{P}(T^{-1}(\mathcal{A})) = \mathbb{P}\{\mathbf{z}\in\mathcal{Z}\colon T(\mathbf{z})\in\mathcal{A}\}, \qquad \mathcal{A}\subseteq\mathcal{Z}'.$$

Construct a distribution $\mathbb{Q}^*$ as a convex combination of two distributions, each of which is a perturbation of the nominal distribution $\mathbb{P}_0$:

$$\mathbb{Q}^* = q\underline{T}_{\#\mathbb{P}_0} + (1-q)\overline{T}_{\#\mathbb{P}_0}, \tag{3.7}$$

where the functions $\underline{T}, \overline{T}\colon \mathcal{Z}\to\mathcal{Z}$ produce the minimizer to $\phi(\lambda^*, \mathbf{z}_0)$, where λ^* is the optimal solution to the dual problem (3.3), i.e.,

$$\underline{T}(\mathbf{z}_0), \overline{T}(\mathbf{z}_0) \in \{\mathbf{z}\in\mathcal{Z}\colon \lambda^* s^t(\mathbf{z},\mathbf{z}_0) - h(\mathbf{z}) = \phi(\lambda^*,\mathbf{z}_0)\}, \tag{3.8}$$

and $q\in[0,1]$ is chosen such that

$$q\int_{\mathcal{Z}} s^t(\underline{T}(\mathbf{z}_0),\mathbf{z}_0)\mathrm{d}\mathbb{P}_0(\mathbf{z}_0) + (1-q)\int_{\mathcal{Z}} s^t(\overline{T}(\mathbf{z}_0),\mathbf{z}_0)\mathrm{d}\mathbb{P}_0(\mathbf{z}_0) = \epsilon^t. \tag{3.9}$$

We choose $\underline{T}, \overline{T}$ to satisfy the following conditions

$$\int_{\mathcal{Z}} s^t(\underline{T}(\mathbf{z}_0),\mathbf{z}_0)\mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \leq \epsilon^t,$$

$$\int_{\mathcal{Z}} s^t(\overline{T}(\mathbf{z}_0),\mathbf{z}_0)\mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \geq \epsilon^t,$$

in order to ensure the existence of such a q.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

We first show that $\mathbb{Q}^*$ is primal feasible. Notice that

$$
\begin{aligned}
&W_{s,t}^t(\mathbb{Q}^*, \mathbb{P}_0) \\
&= \sup_{f,g} \Big\{ \int_{\mathcal{Z}} f(\mathbf{z}) \mathrm{d}\mathbb{Q}^*(\mathbf{z}) + \int_{\mathcal{Z}} g(\mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0): \\
&\qquad\qquad f(\mathbf{z}) \le \inf_{\mathbf{z}_0 \in \mathcal{Z}} [s^t(\mathbf{z}, \mathbf{z}_0) - g(\mathbf{z}_0)],\ \forall \mathbf{z} \in \mathcal{Z} \Big\} \\
&= \sup_{f,g} \Big\{ q \int_{\mathcal{Z}} f(\mathbf{z}) \mathrm{d}\mathbb{P}_0(\underline{T}^{-1}(\mathbf{z})) + (1-q) \int_{\mathcal{Z}} f(\mathbf{z}) \mathrm{d}\mathbb{P}_0(\overline{T}^{-1}(\mathbf{z})) \\
&\qquad + \int_{\mathcal{Z}} g(\mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0):\ f(\mathbf{z}) \le \inf_{\mathbf{z}_0 \in \mathcal{Z}} [s^t(\mathbf{z}, \mathbf{z}_0) - g(\mathbf{z}_0)],\ \forall \mathbf{z} \in \mathcal{Z} \Big\} \\
&\le \sup_{g} \Big\{ q \int_{\mathcal{Z}} (s^t(\underline{T}(\mathbf{z}_0), \mathbf{z}_0) - g(\mathbf{z}_0)) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \\
&\qquad + (1-q) \int_{\mathcal{Z}} (s^t(\overline{T}(\mathbf{z}_0), \mathbf{z}_0) - g(\mathbf{z}_0)) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) + \int_{\mathcal{Z}} g(\mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \Big\} \\
&= q \int_{\mathcal{Z}} s^t(\underline{T}(\mathbf{z}_0), \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) + (1-q) \int_{\mathcal{Z}} s^t(\overline{T}(\mathbf{z}_0), \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \\
&= \epsilon^t,
\end{aligned}
$$

where the first step uses the Kantorovich duality (2.12), the second step uses the structure of $\mathbb{Q}^*$ in (3.7), the third step replaces $f(\mathbf{z})$ by its upper bound $s^t(\mathbf{z}, \mathbf{z}_0) - g(\mathbf{z}_0)$, and the last step uses the definition of q in (3.9).

Now that the feasibility of $\mathbb{Q}^*$ has been established, we next prove that $\mathbb{Q}^*$ is the primal optimal solution by showing that its objective function value matches the optimal dual value.

$$
\begin{aligned}
\int_{\mathcal{Z}} h(\mathbf{z}) \mathrm{d}\mathbb{Q}^*(\mathbf{z}) &= q \int_{\mathcal{Z}} h(\mathbf{z}) \mathrm{d}\mathbb{P}_0(\underline{T}^{-1}(\mathbf{z})) + (1-q) \int_{\mathcal{Z}} h(\mathbf{z}) \mathrm{d}\mathbb{P}_0(\overline{T}^{-1}(\mathbf{z})) \\
&= q \int_{\mathcal{Z}} (\lambda^* s^t(\underline{T}(\mathbf{z}_0), \mathbf{z}_0) - \phi(\lambda^*, \mathbf{z}_0)) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \\
&\quad + (1-q) \int_{\mathcal{Z}} (\lambda^* s^t(\overline{T}(\mathbf{z}_0), \mathbf{z}_0) - \phi(\lambda^*, \mathbf{z}_0)) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \\
&= q\lambda^* \int_{\mathcal{Z}} s^t(\underline{T}(\mathbf{z}_0), \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) - \int_{\mathcal{Z}} \phi(\lambda^*, \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \\
&\quad + (1-q)\lambda^* \int_{\mathcal{Z}} s^t(\overline{T}(\mathbf{z}_0), \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0)
\end{aligned}
$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

$$= \lambda^* \epsilon^t - \int_{\mathcal{Z}} \phi(\lambda^*, \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0)$$
$$= v_D,$$

where the first step uses the structure of $\mathbb{Q}^*$ in (3.7), the second step uses the definition of $\underline{T}, \overline{T}$ in (3.8), the fourth step uses the definition of q in (3.9), and the last step results from the optimality of λ^*. We are now ready to state the strong duality result.

Theorem 3.1.3 ([14], Theorem 1)**.** Suppose that $\mathrm{GR}_h < \infty$. The dual problem (3.3) always admits a minimizer λ^*, and strong duality holds: $v_P = v_D < \infty$.

Remark: The dual problem (3.3) admits a minimizer

$$\lambda^* \in [\max(0, \mathrm{GR}_h), \infty).$$

To see this, notice that for all $\lambda < \mathrm{GR}_h$, $\phi(\lambda, \mathbf{z}_0) = -\infty$, since

$$\begin{aligned} &\limsup_{s(\mathbf{z},\mathbf{z}_0)\to\infty} \left[\lambda s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z})\right] \\ &= \limsup_{s(\mathbf{z},\mathbf{z}_0)\to\infty} \left(\lambda - \frac{h(\mathbf{z}) - h(\mathbf{z}_0)}{s^t(\mathbf{z}, \mathbf{z}_0)}\right) s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z}_0) \\ &= -\infty, \end{aligned}$$

in which case $v_D(\lambda) = \infty$. We conclude that $\lambda^* \geq \mathrm{GR}_h$.

By using duality, [13], [14], and [18] proposed tractable convex reformulations for the DRO problem (1.5). For Lipschitz continuous loss functions, the duality result leads to an equivalent formulation for the Wasserstein DRO as a regularized empirical loss minimization problem, where the regularizer is related to the Lipschitz constant of the loss, see [46], [47]. This connection between robustness and regularization has also been established in [12] and [55]. We will discuss it in further details in Section 4.

3.2 The Extreme Distribution

Section 3.1 reveals the structure of the primal optimal solution $\mathbb{Q}^*$ (the extreme distribution) in (3.7). We summarize the discussions on the

The version of record is available at: http://dx.doi.org/10.1561/2400000026

existence and the form of the extreme distribution in the following theorem.

Theorem 3.2.1 ([14], Corollary 1)**.** Suppose $\mathcal{Z} = \mathbb{R}^d$. The worst-case distribution exists if there exists a dual minimizer λ^*, and the set $\{\mathbf{z} \in \mathcal{Z}\colon \lambda^* s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z}) = \phi(\lambda^*, \mathbf{z}_0)\}$ is non-empty $\mathbb{P}_0$-almost everywhere, and

$$\int_{\mathcal{Z}} \underline{s}^t(\lambda^*, \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \leq \epsilon^t,$$
$$\int_{\mathcal{Z}} \overline{s}^t(\lambda^*, \mathbf{z}_0) \mathrm{d}\mathbb{P}_0(\mathbf{z}_0) \geq \epsilon^t,$$

where

$$\underline{s}(\lambda, \mathbf{z}_0) \triangleq \min_{\mathbf{z} \in \mathcal{Z}} \{s(\mathbf{z}, \mathbf{z}_0)\colon \lambda s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z}) = \phi(\lambda, \mathbf{z}_0)\},$$

and,

$$\overline{s}(\lambda, \mathbf{z}_0) \triangleq \max_{\mathbf{z} \in \mathcal{Z}} \{s(\mathbf{z}, \mathbf{z}_0)\colon \lambda s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z}) = \phi(\lambda, \mathbf{z}_0)\}.$$

Whenever the worst-case distribution exists, there exists one which can be represented as a convex combination of two distributions, each of which is a perturbation of the nominal distribution:

$$\mathbb{Q} = q\underline{T}_{\#\mathbb{P}_0} + (1-q)\overline{T}_{\#\mathbb{P}_0},$$

where $q \in [0, 1]$, and $\underline{T}, \overline{T}\colon \mathcal{Z} \to \mathcal{Z}$ satisfy

$$\underline{T}(\mathbf{z}_0), \overline{T}(\mathbf{z}_0) \in \{\mathbf{z} \in \mathcal{Z}\colon \lambda^* s^t(\mathbf{z}, \mathbf{z}_0) - h(\mathbf{z}) = \phi(\lambda^*, \mathbf{z}_0)\}.$$

3.3 A Discrete Empirical Nominal Distribution

In this section we apply the strong duality result developed in previous sections to the scenario where the discrete empirical distribution $\hat{\mathbb{P}}_N$ is used as the center of the ambiguity set.

Corollary 3.3.1 ([14], Corollary 2)**.** Suppose we use the empirical distribution

$$\hat{\mathbb{P}}_N \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{z}_i}(\mathbf{z})$$

as the center of the ambiguity set, i.e., $\mathbb{P}_0 = \hat{\mathbb{P}}_N$, where $\mathbf{z}_i, i \in [\![N]\!]$, are the observed realizations of $\mathbf{z}$. Assume $\mathrm{GR}_h < \infty$. Then,

The version of record is available at: http://dx.doi.org/10.1561/2400000026

(i) The primal problem (3.2) has a strong dual problem

$$v_P = v_D = \min_{\lambda \geq 0} \left\{ \lambda \epsilon^t + \frac{1}{N} \sum_{i=1}^{N} \sup_{\mathbf{z} \in \mathcal{Z}} [h(\mathbf{z}) - \lambda s^t(\mathbf{z}, \mathbf{z}_i)] \right\}. \tag{3.10}$$

Moreover, v_P, v_D are also equal to

$$\begin{aligned} \sup_{\underline{\mathbf{z}}_i, \overline{\mathbf{z}}_i, q_1, q_2} \quad & \left\{ \frac{1}{N} \sum_{i=1}^{N} [q_1 h(\underline{\mathbf{z}}_i) + q_2 h(\overline{\mathbf{z}}_i)] \right\} \\ \text{s.t.} \quad & \frac{1}{N} \sum_{i=1}^{N} [q_1 s^t(\underline{\mathbf{z}}_i, \mathbf{z}_i) + q_2 s^t(\overline{\mathbf{z}}_i, \mathbf{z}_i)] \leq \epsilon^t, \\ & q_1 + q_2 \leq 1, \\ & q_1, q_2 \geq 0. \end{aligned} \tag{3.11}$$

(ii) When $\mathcal{Z}$ is convex and h is concave, (3.10) could be reduced to

$$\begin{aligned} \sup_{\tilde{\mathbf{z}}_i \in \mathcal{Z}} \quad & \frac{1}{N} \sum_{i=1}^{N} h(\tilde{\mathbf{z}}_i) \\ \text{s.t.} \quad & \frac{1}{N} \sum_{i=1}^{N} s^t(\mathbf{z}_i, \tilde{\mathbf{z}}_i) \leq \epsilon^t. \end{aligned} \tag{3.12}$$

(iii) Whenever the worst-case distribution exists, there exists one which is supported on at most $N+1$ points and has the form

$$\mathbb{Q}^* = \frac{1}{N} \sum_{i \neq i_0} \delta_{\mathbf{z}_i^*}(\mathbf{z}) + \frac{q}{N} \delta_{\underline{\mathbf{z}}_{i_0}^*}(\mathbf{z}) + \frac{1-q}{N} \delta_{\overline{\mathbf{z}}_{i_0}^*}(\mathbf{z}),$$

where $1 \leq i_0 \leq N$, $q \in [0,1]$, $\underline{\mathbf{z}}_{i_0}^*, \overline{\mathbf{z}}_{i_0}^* \in \arg\min_{\mathbf{z} \in \mathcal{Z}} \{\lambda^* s^t(\mathbf{z}, \mathbf{z}_{i_0}) - h(\mathbf{z})\}$, and $\mathbf{z}_i^* \in \arg\min_{\mathbf{z} \in \mathcal{Z}} \{\lambda^* s^t(\mathbf{z}, \mathbf{z}_i) - h(\mathbf{z})\}$ for all $i \neq i_0$.

Proof. (3.10) comes directly from (3.3). For (3.11), recall that the worst-case distribution can be expressed as a convex combination of two perturbed versions of the empirical distribution, see (3.7) and (3.8). Thus, $\mathbb{Q}^*$ is supported on $2N$ points $\underline{\mathbf{z}}_i, \overline{\mathbf{z}}_i,\ i \in [\![N]\!]$, with probabilities q/N and $(1-q)/N$, respectively. Problem (3.11) finds the worst-case expected loss by imposing such a structure on the distribution $\mathbb{Q}$.

Part (ii) can be proved by noticing that Problem (3.10) is the Lagrangian dual of (3.12), which is a convex problem due to the concavity of h.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

To prove (iii), consider problem (3.11) by replacing q_1 with q_i and q_2 with $1 - q_i$, i.e., we allow q to vary over samples. (3.11) is a linear program in q_i and has an optimal solution which has at most one fractional point (i.e., $\exists\, i_0,$ s.t. $q_{i_0} > 0;\ q_i = 0,\ \forall i \neq i_0$). Therefore, there exists a worst-case distribution which is supported on at most $N + 1$ points. □

3.3.1 A Special Case

We study a special case where the loss function $h(\mathbf{z})$ is convex in $\mathbf{z}$. We will show that Problem (3.1) can be relaxed to the summation of the empirical loss and a regularizer, where the regularization strength is equal to the size of the ambiguity set, and the regularizer is defined by the dual norm.

Before we present this result, we start with two definitions and a well-known property.

Definition 2 (Dual Norm)**.** Given a norm $\|\cdot\|$ on $\mathbb{R}^d$, the dual norm $\|\cdot\|_*$ is defined as:

$$\|\boldsymbol{\theta}\|_* \triangleq \sup_{\|\mathbf{z}\| \leq 1} \boldsymbol{\theta}'\mathbf{z}. \tag{3.13}$$

It can be shown from (3.13) that for any vectors $\boldsymbol{\theta}, \mathbf{z}$, the following Hölder's inequality holds.

Theorem 3.3.2 (Hölder's Inequality)**.** Suppose we have two scalars $r, s > 1$ and $1/r + 1/s = 1$. For any two vectors $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\mathbf{z} = (z_1, \ldots, z_n)$, the following holds:

$$\sum_{i=1}^{n} |\theta_i z_i| \leq \|\boldsymbol{\theta}\|_r \|\mathbf{z}\|_s.$$

Definition 3 (Conjugate Function)**.** For a function $h(\mathbf{z})$, its convex conjugate $h^*(\cdot)$ is defined as:

$$h^*(\boldsymbol{\theta}) \triangleq \sup_{\mathbf{z} \in \text{dom } h} \{\boldsymbol{\theta}'\mathbf{z} - h(\mathbf{z})\}, \tag{3.14}$$

where dom h denotes the domain of the function h.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

If h is convex, then the convex conjugate of h^* is h, and h and h^* are called convex duals [84]. In particular,

$$h(\mathbf{z}) = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} [\boldsymbol{\theta}'\mathbf{z} - h^*(\boldsymbol{\theta})], \tag{3.15}$$

where $\boldsymbol{\Theta} \triangleq \{\boldsymbol{\theta}\colon h^*(\boldsymbol{\theta}) < \infty\}$ denotes the effective domain of the conjugate function h^*.

Theorem 3.3.3 ([13], Theorem 6.3)**.** Suppose the loss function $h(\mathbf{z})$ is convex in $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^d$, and the set $\mathcal{Z}$ is closed and convex. Define an ambiguity set around the empirical distribution which is supported on N samples $\mathbf{z}_i, i \in [\![N]\!]$, i.e.,

$$\Omega = \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{\|\cdot\|,1}(\mathbb{Q},\ \hat{\mathbb{P}}_N) \le \epsilon\},$$

where the order-1 Wasserstein metric (1.7) is induced by some norm $\|\cdot\|$. Problem (3.1) can be relaxed to:

$$\sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h(\mathbf{z})] \le \kappa\epsilon + \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{z}_i), \tag{3.16}$$

where

$$\kappa = \sup\{\|\boldsymbol{\theta}\|_*\colon h^*(\boldsymbol{\theta}) < \infty\},$$

where $\|\cdot\|_*$ stands for the dual norm as defined in (3.13), and $h^*(\cdot)$ is the convex conjugate function of $h(\mathbf{z})$ as defined in (3.14). Furthermore, (3.16) becomes an equality when $\mathcal{Z} = \mathbb{R}^d$.

Proof. Corollary 3.3.1 suggests that

$$\sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h(\mathbf{z})] = \min_{\lambda\ge 0}\left\{\lambda\epsilon + \frac{1}{N}\sum_{i=1}^{N}\sup_{\mathbf{z}\in\mathcal{Z}}[h(\mathbf{z}) - \lambda\|\mathbf{z}-\mathbf{z}_i\|]\right\}. \tag{3.17}$$

Using (3.15), we may write the inner maximization in (3.17) as:

$$\begin{aligned}
\sup_{\mathbf{z}\in\mathcal{Z}}[h(\mathbf{z}) - \lambda\|\mathbf{z}-\mathbf{z}_i\|] &= \sup_{\mathbf{z}\in\mathcal{Z}}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}[\boldsymbol{\theta}'\mathbf{z} - h^*(\boldsymbol{\theta}) - \lambda\|\mathbf{z}-\mathbf{z}_i\|] \\
&= \sup_{\mathbf{z}\in\mathcal{Z}}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\inf_{\|\mathbf{r}\|_*\le\lambda}[\boldsymbol{\theta}'\mathbf{z} - h^*(\boldsymbol{\theta}) + \mathbf{r}'(\mathbf{z}-\mathbf{z}_i)] \\
&= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\inf_{\|\mathbf{r}\|_*\le\lambda}\sup_{\mathbf{z}\in\mathcal{Z}}[(\boldsymbol{\theta}+\mathbf{r})'\mathbf{z} - h^*(\boldsymbol{\theta}) - \mathbf{r}'\mathbf{z}_i] \\
&\le \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\inf_{\|\mathbf{r}\|_*\le\lambda}\sup_{\mathbf{z}\in\mathbb{R}^d}[(\boldsymbol{\theta}+\mathbf{r})'\mathbf{z} - h^*(\boldsymbol{\theta}) - \mathbf{r}'\mathbf{z}_i],
\end{aligned} \tag{3.18}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where the second equality follows from the definition of the dual norm and the third equality uses duality. The inner maximization over $\mathbf{z} \in \mathbb{R}^d$ achieves ∞ unless $\mathbf{r} = -\boldsymbol{\theta}$.

Note that if $\sup\{\|\boldsymbol{\theta}\|_*\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\} > \lambda$, then one can pick some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\|\boldsymbol{\theta}\|_* > \lambda$, in which case the inner maximization over $\mathbf{z} \in \mathbb{R}^d$ in (3.18) achieves ∞ since $\mathbf{r} \neq -\boldsymbol{\theta}$.

When $\sup\{\|\boldsymbol{\theta}\|_*\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \leq \lambda$, by taking $\mathbf{r} = -\boldsymbol{\theta}$, we have:

$$\begin{aligned} \sup_{\mathbf{z}\in\mathcal{Z}}[h(\mathbf{z}) - \lambda\|\mathbf{z} - \mathbf{z}_i\|] &\leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}[-h^*(\boldsymbol{\theta}) + \boldsymbol{\theta}'\mathbf{z}_i] \\ &= h(\mathbf{z}_i). \end{aligned} \tag{3.19}$$

Plugging (3.19) into (3.17), we obtain

$$\sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h(\mathbf{z})] \leq \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{z}_i) + \kappa\epsilon,$$

where $\kappa = \sup\{\|\boldsymbol{\theta}\|_*\colon h^*(\boldsymbol{\theta}) < \infty\}$. □

3.4 Finite Sample Performance

In this section we discuss the finite sample out-of-sample performance of the DRO estimator. Recall the stochastic optimization problem defined in (1.4):

$$J^* \triangleq \inf_{\boldsymbol{\beta}} \mathbb{E}^{\mathbb{P}^*}[h_{\boldsymbol{\beta}}(\mathbf{z})] = \inf_{\boldsymbol{\beta}} \int_{\mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{z}) \mathrm{d}\mathbb{P}^*(\mathbf{z}). \tag{3.20}$$

Since the true measure $\mathbb{P}^*$ is unknown, Problem (3.20) is not directly solvable. We solve its DRO counterpart (1.5) using the available training data $\mathbf{z}_i, i \in [\![N]\!]$, with an effort to implicitly optimize over the true measure that is included in the ambiguity set with high confidence. Suppose $\hat{J}_N$ and $\hat{\boldsymbol{\beta}}_N$ are respectively the optimal value and optimal solution to the DRO problem (1.5), i.e.,

$$\hat{J}_N \triangleq \inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h_{\boldsymbol{\beta}}(\mathbf{z})] = \sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h_{\hat{\boldsymbol{\beta}}_N}(\mathbf{z})], \tag{3.21}$$

where the ambiguity set is defined as

$$\Omega = \Omega_\epsilon(\hat{\mathbb{P}}_N) \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{\|\cdot\|,1}(\mathbb{Q},\ \hat{\mathbb{P}}_N) \leq \epsilon\}. \tag{3.22}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

To evaluate the quality of the DRO estimator $\hat{\boldsymbol{\beta}}_N$, we study its *out-of-sample* performance on a new sample $\mathbf{z}$ drawn from $\mathbb{P}^*$,

$$\mathbb{E}^{\mathbb{P}^*}[h_{\hat{\boldsymbol{\beta}}_N}(\mathbf{z})]. \tag{3.23}$$

We want to investigate whether the *out-of-sample* loss (3.23) can be meaningfully bounded from above by some *certificate*. Specifically, if we can show that with a high probability, the *out-of-sample* loss (3.23) does not exceed the training loss $\hat{J}_N$,

$$\mathbb{P}^N\{\mathbb{E}^{\mathbb{P}^*}[h_{\hat{\boldsymbol{\beta}}_N}(\mathbf{z})] \leq \hat{J}_N\} \geq 1-\alpha,$$

where $\alpha \in (0,1)$ is a significance parameter w.r.t. the distribution $\mathbb{P}^N$, which governs both $\hat{\boldsymbol{\beta}}_N$ and $\hat{J}_N$, then we can claim that $\hat{\boldsymbol{\beta}}_N$ generalizes well out-of-sample. The following theorem, which follows directly from the measure concentration Theorem 2.7.1, establishes the result.

Theorem 3.4.1 ([13], Theorem 3.5)**.** Suppose Assumption B holds, and $\hat{J}_N$ and $\hat{\boldsymbol{\beta}}_N$ are respectively the optimal value and optimal solution to the DRO problem (1.5) with an ambiguity set specified in (3.22). Set the radius $\epsilon = \epsilon_N(\alpha)$ as defined in (2.30), where $\alpha \in (0,1)$. Then we have

$$\mathbb{P}^N\{\mathbb{E}^{\mathbb{P}^*}[h_{\hat{\boldsymbol{\beta}}_N}(\mathbf{z})] \leq \hat{J}_N\} \geq 1-\alpha.$$

Proof. The claim follows immediately from the measure concentration result presented in Theorem 2.7.1, which establishes that

$$\mathbb{P}^N(W_{\|\cdot\|,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \geq \epsilon_N(\alpha)) \leq \alpha,$$

and therefore,

$$\mathbb{P}^N\{\mathbb{E}^{\mathbb{P}^*}[h_{\hat{\boldsymbol{\beta}}_N}(\mathbf{z})] \leq \hat{J}_N\} \geq \mathbb{P}^N\{\mathbb{P}^* \in \Omega_{\epsilon_N(\alpha)}(\hat{\mathbb{P}}_N)\} \geq 1-\alpha.$$

□

Note that Theorem 3.4.1 establishes the out-of-sample performance of the DRO estimator for an order-1 Wasserstein ambiguity set. For a general Wasserstein metric with order $t > 1$, please refer to [66] for a general measure concentration result.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

3.5 Asymptotic Consistency

In addition to the finite sample result established in Section 3.4, we are also interested in the asymptotic behavior of $\hat{J}_N$ and $\hat{\boldsymbol{\beta}}_N$, as the sample size N goes to infinity. We want to establish that, if the significance level $\alpha = \alpha_N$ converges to zero at a carefully chosen rate, then the optimal value and solution of the DRO problem (1.5) with an ambiguity set of size $\epsilon = \epsilon_N(\alpha_N)$, converge to the optimal value and solution of the original stochastic optimization problem (3.20), respectively. The following Theorem 3.5.1 formalizes this statement.

Theorem 3.5.1 ([13], Theorem 3.6)**.** Suppose Assumption B holds and the significance parameter $\alpha_N \in (0, 1)$ satisfies

- $\sum_{N=1}^{\infty} \alpha_N < \infty$;
- $\lim_{N\to\infty} \epsilon_N(\alpha_N) = 0$.

Assume the loss function $h_{\boldsymbol{\beta}}(\mathbf{z})$ is Lipschitz continuous in $\mathbf{z}$ with a Lipschitz constant $L_{\boldsymbol{\beta}}$. Denote by $\hat{J}_N$ and $\hat{\boldsymbol{\beta}}_N$ the optimal value and optimal solution to the DRO problem (1.5), respectively, with an ambiguity set specified in (3.22) with $\epsilon = \epsilon_N(\alpha_N)$, where $\epsilon_N(\alpha_N)$ is defined in (2.30), and J^* is the optimal value of the original stochastic optimization problem (3.20). Then,

(i) $\hat{J}_N$ converges to J^* a.s.,

$$\mathbb{P}^{\infty}\left\{ \limsup_{N\to\infty} \hat{J}_N = J^* \right\} = 1.$$

(ii) If $h_{\boldsymbol{\beta}}(\mathbf{z})$ is lower semicontinuous in $\boldsymbol{\beta}$ for every $\mathbf{z} \in \mathcal{Z}$, and $\lim_{N\to\infty} \hat{\boldsymbol{\beta}}_N = \boldsymbol{\beta}_0$, then $\boldsymbol{\beta}_0$ is $\mathbb{P}^{\infty}$-almost surely an optimal solution to (3.20).

Proof. (i) Theorem 3.4.1 implies that

$$\mathbb{P}^N\{J^* \le \mathbb{E}^{\mathbb{P}^*}[h_{\hat{\boldsymbol{\beta}}_N}(\mathbf{z})] \le \hat{J}_N\} \ge \mathbb{P}^N\{\mathbb{P}^* \in \Omega_{\epsilon_N(\alpha_N)}(\hat{\mathbb{P}}_N)\} \ge 1 - \alpha_N. \tag{3.24}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

As $\sum_{N=1}^{\infty} \alpha_N < \infty$, the Borel–Cantelli Lemma [85], [86] implies that,

$$\mathbb{P}^{\infty}\left\{ \limsup_{N\to\infty} \hat{J}_N \geq J^* \right\} = 1.$$

It remains to show that

$$\mathbb{P}^{\infty}\left\{ \limsup_{N\to\infty} \hat{J}_N \leq J^* \right\} = 1. \tag{3.25}$$

Let $\hat{\mathbb{Q}}_N \in \Omega_{\epsilon_N(\alpha_N)}(\hat{\mathbb{P}}_N)$ be the optimal solution to the inner supremum (3.1) corresponding to $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ is the optimal solution to (3.20). Then,

$$\mathbb{E}^{\hat{\mathbb{Q}}_N}[h_{\boldsymbol{\beta}^*}(\mathbf{z})] = \sup_{\mathbb{Q}\in\Omega_{\epsilon_N(\alpha_N)}(\hat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[h_{\boldsymbol{\beta}^*}(\mathbf{z})].$$

According to Theorem 3.1.1, and due to the Lipschitz continuity of $h_{\boldsymbol{\beta}}(\mathbf{z})$, we know that

$$|\mathbb{E}^{\mathbb{Q}_1}[h_{\boldsymbol{\beta}}(\mathbf{z})] - \mathbb{E}^{\mathbb{Q}_2}[h_{\boldsymbol{\beta}}(\mathbf{z})]| \leq L_{\boldsymbol{\beta}} W_{\|\cdot\|,1}(\mathbb{Q}_1, \mathbb{Q}_2).$$

Then,

$$\begin{aligned}
\hat{J}_N &\leq \sup_{\mathbb{Q}\in\Omega_{\epsilon_N(\alpha_N)}(\hat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[h_{\boldsymbol{\beta}^*}(\mathbf{z})] \\
&= \mathbb{E}^{\hat{\mathbb{Q}}_N}[h_{\boldsymbol{\beta}^*}(\mathbf{z})] \\
&\leq \mathbb{E}^{\mathbb{P}^*}[h_{\boldsymbol{\beta}^*}(\mathbf{z})] + L_{\boldsymbol{\beta}^*} W_{\|\cdot\|,1}(\mathbb{P}^*, \hat{\mathbb{Q}}_N) \\
&= J^* + L_{\boldsymbol{\beta}^*} W_{\|\cdot\|,1}(\mathbb{P}^*, \hat{\mathbb{Q}}_N),
\end{aligned}$$

where the first step is due to the feasibility of $\boldsymbol{\beta}^*$ to (3.21), and the third step is due to Theorem 3.1.1. In order to prove (3.25), we only need to show that

$$\mathbb{P}^{\infty}\left\{ \limsup_{N\to\infty} W_{\|\cdot\|,1}(\mathbb{P}^*, \hat{\mathbb{Q}}_N) = 0 \right\} = 1. \tag{3.26}$$

The triangle inequality of the Wasserstein metric (cf. Theorem 2.2.1) ensures that

$$\begin{aligned}
W_{\|\cdot\|,1}(\mathbb{P}^*, \hat{\mathbb{Q}}_N) &\leq W_{\|\cdot\|,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) + W_{\|\cdot\|,1}(\hat{\mathbb{P}}_N, \hat{\mathbb{Q}}_N) \\
&\leq W_{\|\cdot\|,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) + \epsilon_N(\alpha_N).
\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

From Theorem 2.7.1 we know that

$$\mathbb{P}^N(W_{\|\cdot\|,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \leq \epsilon_N(\alpha_N)) \geq 1 - \alpha_N.$$

Therefore, by the Borel–Cantelli Lemma [86],

$$\mathbb{P}^\infty\left(\limsup_{N\to\infty}\{W_{\|\cdot\|,1}(\mathbb{P}^*, \hat{\mathbb{P}}_N) \leq \epsilon_N(\alpha_N)\}\right) = 1.$$

Since $\lim_{N\to\infty} \epsilon_N(\alpha_N) = 0$, (3.26) follows.

(ii) We need to show that $\boldsymbol{\beta}_0$ achieves the optimal value of (3.20), i.e.,

$$\mathbb{E}^{\mathbb{P}^*}[h_{\boldsymbol{\beta}_0}(\mathbf{z})] = J^*.$$

Note that,

$$\begin{aligned} J^* &\leq \mathbb{E}^{\mathbb{P}^*}[h_{\boldsymbol{\beta}_0}(\mathbf{z})] \\ &\leq \mathbb{E}^{\mathbb{P}^*}[\liminf_{N\to\infty} h_{\hat{\boldsymbol{\beta}}_N}(\mathbf{z})] \\ &\leq \liminf_{N\to\infty} \mathbb{E}^{\mathbb{P}^*}[h_{\hat{\boldsymbol{\beta}}_N}(\mathbf{z})] \\ &\leq \limsup_{N\to\infty} \mathbb{E}^{\mathbb{P}^*}[h_{\hat{\boldsymbol{\beta}}_N}(\mathbf{z})] \\ &\leq \limsup_{N\to\infty} \hat{J}_N \\ &= J^*, \end{aligned}$$

where the first inequality is due to the feasibility of $\boldsymbol{\beta}_0$ to (3.20), the second inequality follows from the lower semicontinuity of h in $\boldsymbol{\beta}$, the third inequality is due to Fatou's Lemma, and the fifth inequality holds $\mathbb{P}^\infty$-almost surely due to (3.24). We thus conclude that $\hat{\boldsymbol{\beta}}_N$ converges to the optimal solution of (3.20) a.s. □

The version of record is available at: http://dx.doi.org/10.1561/2400000026

4

Distributionally Robust Linear Regression

In this section, we introduce the Wasserstein DRO formulation for linear regression. The focus is to estimate a robustified linear regression plane that is immunized against potential outliers in the data. Classical approaches, such as robust regression [7], [8], remedy this problem by fitting a weighted least squares that downweights the contribution of atypical data points. By contrast, the DRO approach mitigates the impact of outliers through hedging against a family of distributions on the observed data, some of which assign very low probabilities to the outliers.

4.1 The Problem and Related Work

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

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

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Given potentially corrupted samples $(\mathbf{x}_i, y_i), i \in [\![N]\!]$, we are interested in obtaining an estimator of $\boldsymbol{\beta}^*$ that is robust with respect to the perturbations in the data. Popular robust estimators include:

- *Least Absolute Deviation (LAD)*, which minimizes the sum of absolute residuals $\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|$, and
- M-estimation [7], [8], which minimizes a symmetric loss function $\rho(\cdot)$ of the residuals in the form $\sum_{i=1}^{N} \rho(y_i - \mathbf{x}_i'\boldsymbol{\beta})$, downweighting the influence of samples with large absolute residuals.

Several choices for $\rho(\cdot)$ include the Huber function [7], [8], the Tukey's Biweight function [87], the logistic function [88], the Talwar function [89], and the Fair function [90].

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. The breakdown value is the smallest proportion of observations in the dataset that need to be replaced to carry the estimate arbitrarily far away. Examples of high breakdown value methods include the *Least Median of Squares (LMS)* [91], which minimizes the median of the absolute residuals, the *Least Trimmed Squares (LTS)* [92], which minimizes the sum of the q smallest squared residuals, and S-estimation [93], 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 [94]. It has a higher statistical efficiency than S-estimation. We refer the reader to the book [87] for an elaborate 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 optimization problem 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. Instead

The version of record is available at: http://dx.doi.org/10.1561/2400000026

of equally weighting all the samples as in the empirical distribution, one may wish to include more informative distributions that "drive out" the corrupted samples. DRO realizes this through hedging the expected loss against a family of distributions that includes the true data-generating mechanism with high confidence (cf. Theorem 2.7.1). Compared to the single distribution-based stochastic optimization, DRO often results in better out-of-sample performance due to its distributional robustness.

We consider a DRO problem with an ambiguity set containing distributions that are close to the discrete empirical distribution in the sense of Wasserstein distance. We adopt the absolute residual loss $|y-\mathbf{x}'\boldsymbol{\beta}|$ for the purpose of enhancing robustness. By exploiting duality, we relax the Wasserstein DRO formulation to a convex optimization problem which encompasses a class of regularized regression models, providing new insights into the regularizer, and establishing the connection between the amount of ambiguity allowed and a regularization penalty term. We provide justifications for the ℓ_1-loss based DRO learning by establishing 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). Extensive numerical results demonstrate the superiority of the DRO model to a host of regression models, in terms of the prediction and estimation accuracies. We also consider the application of the DRO model to outlier detection, and show that it achieves a much higher AUC (Area Under the ROC Curve) than M-estimation [7], [8].

The rest of this section is organized as follows. In Section 4.2, we introduce the Wasserstein DRO formulation in a linear regression setting. Section 4.3 establishes performance guarantees for the solution to DRO relaxation. The numerical results on the performance of DRO regression are presented in Section 4.4. An application of DRO regression to outlier detection is discussed in Section 4.5. We conclude in Section 4.6.

4.2 The Wasserstein DRO Formulation for Linear Regression

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 forgiving (hence, robust) to large residuals than the squared loss (see Figure 4.1). The Wasserstein

The version of record is available at: http://dx.doi.org/10.1561/2400000026

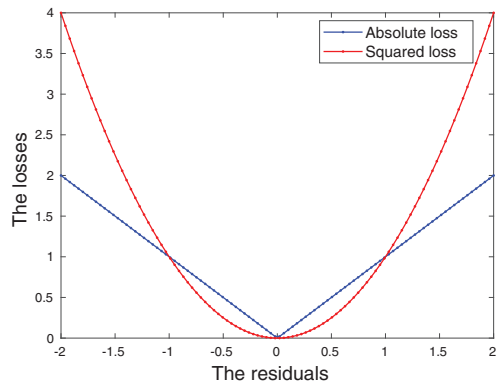


Figure 4.1: The comparison between ℓ_1 and ℓ_2 loss functions.

DRO problem using the ℓ_1-loss function is formulated as:

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[|y - \mathbf{x}'\boldsymbol{\beta}|], \tag{4.1}$$

where Ω is defined as:

$$\Omega = \Omega_{\epsilon}^{s,t}(\hat{\mathbb{P}}_N) \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{s,t}(\mathbb{Q},\ \hat{\mathbb{P}}_N) \leq \epsilon\},$$

and $W_{s,t}(\mathbb{Q},\ \hat{\mathbb{P}}_N)$ is the order-t Wasserstein distance between $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$ under a distance metric s (see definition in (1.7)), with $\hat{\mathbb{P}}_N$ the uniform empirical distribution over N samples. The formulation in (4.1) 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.

We first decide an appropriate order t for the Wasserstein metric. Based on the discussion in Section 3.1, it is required that the loss function h has a finite growth rate. Assuming that the metric s is induced by some norm $\|\cdot\|$, the bounded growth rate requirement is expressed as follows:

$$\begin{aligned}&\limsup_{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|\to\infty} \frac{|h_{\boldsymbol{\beta}}(\mathbf{x}_1,y_1) - h_{\boldsymbol{\beta}}(\mathbf{x}_2,y_2)|}{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|^t}\\ &\leq \limsup_{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|\to\infty} \frac{|y_1 - \mathbf{x}_1'\boldsymbol{\beta} - (y_2 - \mathbf{x}_2'\boldsymbol{\beta})|}{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|^t}\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

$$\leq \limsup_{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|\to\infty} \frac{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|\|(-\boldsymbol{\beta},1)\|_*}{\|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|^t} \tag{4.2}$$
$$< \infty,$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, and the second inequality is due to Hölder's inequality (cf. Theorem 3.3.2). Notice that by taking $t = 1$, (4.2) is equivalently translated into the condition that $\|(-\boldsymbol{\beta},1)\|_* < \infty$, which, as we will see in Section 4.3, 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 $t > 1$ leads to zero growth rate in the limit of (4.2), which is not desirable since it removes the Wasserstein ball structure from the formulation and renders it an optimization problem over a singleton distribution. 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 (4.1) into a tractable formulation. Suppose we have N independently and identically distributed realizations of $(\mathbf{x}, y)$, denoted by $(\mathbf{x}_i, y_i), i \in [\![N]\!]$. Since the loss function is convex in $(\mathbf{x}, y)$, using the result in Section 3.3.1, the inner supremum of (4.1) can be relaxed to the right hand side of (3.16). In Theorem 4.2.1, we compute the value of κ in (3.16) for the specific ℓ_1 loss function we use.

Theorem 4.2.1. Define $\kappa(\boldsymbol{\beta}) = \sup\{\|\boldsymbol{\theta}\|_*\colon h^*_{\boldsymbol{\beta}}(\boldsymbol{\theta}) < \infty\}$, 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 $\kappa(\boldsymbol{\beta}) = \|(-\boldsymbol{\beta},1)\|_*$.

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:

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

Consider now the two linear optimization problems A and B:

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

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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

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:

$$\exists\ r_A \leq 0, \quad \text{s.t.} \quad \tilde{\boldsymbol{\beta}} r_A = \boldsymbol{\theta} - \tilde{\boldsymbol{\beta}}, \tag{4.3}$$

$$\exists\ r_B \geq 0, \quad \text{s.t.} \quad \tilde{\boldsymbol{\beta}} r_B = \boldsymbol{\theta} + \tilde{\boldsymbol{\beta}}. \tag{4.4}$$

For all i with $\tilde{\beta}_i \leq 0$, (4.3) implies $\theta_i - \tilde{\beta}_i \geq 0$ and (4.4) implies $\theta_i \leq -\tilde{\beta}_i$. On the other hand, for all j with $\tilde{\beta}_j \geq 0$, (4.3) and (4.4) 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,

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

□

Due to Theorem 4.2.1 and (3.16), (4.1) could be formulated as the following optimization problem:

$$\inf_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon\|(-\boldsymbol{\beta}, 1)\|_*. \tag{4.5}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Notice that (4.5) coincides with the regularized LAD models [95], [96], except that it regularizes a variant of the regression coefficient. The regularization term of (4.5) is the product of the *growth rate* of the loss and the Wasserstein ball radius. 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 was discussed in Section 2.7.

The connection between robustness and regularization has been established in several works. The earliest one may be credited to [38], which shows 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, [39] has shown the equivalence between robust linear regression and the *Least Absolute Shrinkage and Selection Operator (LASSO)*. [40] extends this to more general LASSO-like procedures, including versions of the grouped LASSO. [37] gives a comprehensive characterization of the conditions under which robustification and regularization are equivalent for regression models. For classification problems, [97] shows the equivalence between the regularized support vector machines (SVMs) and a robust optimization formulation, by allowing potentially correlated disturbances in the covariates. [55] considers a robust version of logistic regression under the assumption that the probability distributions under consideration lie in a Wasserstein ball. Recently, [46], [47] have provided a unified framework for connecting the Wasserstein DRO with regularized learning procedures, for various regression and classification models.

Formulation (4.5) 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 $\|\cdot\|_2$, (4.5) is a convex quadratic problem which can be solved to optimality very efficiently. Specifically, it could be

The version of record is available at: http://dx.doi.org/10.1561/2400000026

converted to:

$$\begin{aligned} \min_{a,b_1,\ldots,b_N,\boldsymbol{\beta}} \quad & a\epsilon + \frac{1}{N}\sum_{i=1}^{N} b_i \\ \text{s.t.} \quad & \|\boldsymbol{\beta}\|_2^2 + 1 \le a^2, \\ & y_i - \mathbf{x}_i'\boldsymbol{\beta} \le b_i, \quad i \in [\![N]\!], \\ & -(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \le b_i, \quad i \in [\![N]\!], \\ & a,\ b_i \ge 0, \quad i \in [\![N]\!]. \end{aligned} \tag{4.6}$$

When the Wasserstein metric is defined using $\|\cdot\|_1$, (4.5) is a linear programming problem:

$$\begin{aligned} \min_{a,b_1,\ldots,b_N,\ \boldsymbol{\beta}} \quad & a\epsilon + \frac{1}{N}\sum_{i=1}^{N} b_i \\ \text{s.t.} \quad & a \ge \boldsymbol{\beta}'\mathbf{e}_i, \quad i \in [\![p]\!], \\ & a \ge -\boldsymbol{\beta}'\mathbf{e}_i, \quad i \in [\![p]\!], \\ & y_i - \mathbf{x}_i'\boldsymbol{\beta} \le b_i, \quad i \in [\![N]\!], \\ & -(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \le b_i, \quad i \in [\![N]\!], \\ & a \ge 1, \\ & b_i \ge 0, \quad i \in [\![N]\!]. \end{aligned} \tag{4.7}$$

More generally, when the coordinates of $(\mathbf{x}, y)$ differ from each other substantially, a properly chosen, positive definite weight matrix $\mathbf{M} \in \mathbb{R}^{(p+1)\times(p+1)}$ 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 (4.5) in this case becomes:

$$\inf_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon\sqrt{(-\boldsymbol{\beta}, 1)'\mathbf{M}^{-1}(-\boldsymbol{\beta}, 1)}. \tag{4.8}$$

We would like to highlight several novel viewpoints that are brought by the Wasserstein DRO framework and justify the value and novelty of (4.5). First, (4.5) is obtained as an outcome of a fundamental DRO formulation, which enables new interpretations of the regularizer from

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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 $\boldsymbol{\beta}$, (4.5) regularizes the vector $(-\boldsymbol{\beta}, 1)$, where the 1 takes into account the transportation cost along the y direction. Penalizing only $\boldsymbol{\beta}$ corresponds to an infinite transportation cost along y. (4.5) is more general in this sense, and establishes the connection between the metric space on the data and the form of the regularizer.

4.3 Performance Guarantees for the DRO Estimator

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 (4.5). 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 4.3.1). The other one bounds the discrepancy between the estimated and true regression planes (estimation bias), and is given in Section 4.3.2.

4.3.1 Out-of-Sample Performance

In this subsection, we investigate generalization characteristics of the solution to (4.5), which involves measuring the error generated by the DRO 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 [98]), 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

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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 the regularized formulation. Moreover, the generalization result also builds a connection between the loss function and the form of the regularizer via the 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.

Suppose the data $(\mathbf{x}, y)$ is drawn from the probability distribution $\mathbb{P}^*$. We first make several mild assumptions that are needed for the generalization result.

Assumption I. $\|(\mathbf{x}, y)\| \leq R$, a.s. under $\mathbb{P}^*$.

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

Under these two assumptions, the absolute loss could be bounded via Hölder's inequality.

Lemma 4.3.1. For every feasible $\boldsymbol{\beta}$, it follows that,

$$|y - \mathbf{x}'\boldsymbol{\beta}| \leq \bar{B}R, \quad \text{a.s. under } \mathbb{P}^*.$$

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

$$\mathcal{H} = \{(\mathbf{x}, y) \to h_{\boldsymbol{\beta}}(\mathbf{x}, y) \colon h_{\boldsymbol{\beta}}(\mathbf{x}, y) = |y - \mathbf{x}'\boldsymbol{\beta}|\}.$$

We need to show that the empirical Rademacher complexity of $\mathcal{H}$, denoted by $\mathcal{R}_N(\mathcal{H})$ and defined as:

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

is upper bounded, where $\sigma_1, \ldots, \sigma_N$ are i.i.d. uniform random variables on $\{1, -1\}$, and $(\mathbf{x}_i, y_i), i \in [\![N]\!]$, are N observed realizations of $(\mathbf{x}, y)$. The following result, similar to Lemma 3 in [99], provides a bound that is inversely proportional to the square root of the sample size.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Lemma 4.3.2.

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

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 4.3.1,

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

□

Let $\hat{\boldsymbol{\beta}}$ be an optimal solution to (4.5), obtained using the samples $(\mathbf{x}_i, y_i)$, $i \in [\![N]\!]$. Suppose we draw a new i.i.d. sample $(\mathbf{x}, y)$. In Theorem 4.3.3 we establish bounds on the error $|y - \mathbf{x}'\hat{\boldsymbol{\beta}}|$.

Theorem 4.3.3. Under Assumptions I and J, for any $0 < \delta < 1$, with probability at least $1 - \delta$ with respect to the sampling,

$$\mathbb{E}^{\mathbb{P}^*}[|y - \mathbf{x}'\hat{\boldsymbol{\beta}}|] \leq \frac{1}{N}\sum_{i=1}^N |y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}| + \frac{2\bar{B}R}{\sqrt{N}} + \bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}, \quad (4.9)$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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

$$\mathbb{P}\left(|y - \mathbf{x}'\hat{\boldsymbol{\beta}}| \geq \frac{1}{N}\sum_{i=1}^{N}|y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}| + \zeta\right) \leq \frac{\frac{1}{N}\sum_{i=1}^{N}|y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}| + \frac{2\bar{B}R}{\sqrt{N}} + \bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N}|y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}| + \zeta}. \quad (4.10)$$

Proof. We use Theorem 8 in [98], which we state for convenience as follows.

Theorem 4.3.4 (Theorem 8 in [98])**.** Consider a loss function $L\colon \mathcal{Y}\times\mathcal{A} \to [0,1]$ and a dominating cost function $\phi\colon \mathcal{Y}\times\mathcal{A} \to [0,1]$. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathcal{A}$ and let $(\mathbf{x}_i, y_i)_{i=1}^N$ be independently selected according to the probability measure $\mathbb{P}^*$. Then, for any integer N and any $0 < \delta < 1$, with probability at least $1-\delta$ over samples of length N, every f in $\mathcal{F}$ satisfies

$$\mathbb{E}^{\mathbb{P}^*}[L(y, f(\mathbf{x}))] \leq \frac{1}{N}\sum_{i=1}^{N}\phi(y_i, f(\mathbf{x}_i)) + R_N(\tilde{\phi}\circ\mathcal{F}) + \sqrt{\frac{8\log(2/\delta)}{N}},$$

where $\tilde{\phi}\circ\mathcal{F} = \{(\mathbf{x}, y) \to \phi(y, f(\mathbf{x})) - \phi(y, 0) : f \in \mathcal{F}\}$.

We set the following correspondences with the notation used in Theorem 4.3.4: $f(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$, and $L(y, f(\mathbf{x})) = \phi(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$. This yields the bound (4.9) on the expected loss. For Equation (4.10), we apply Markov's inequality to obtain:

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

□

There are two probability measures in the statement of Theorem 4.3.3. One is related to the new data $(\mathbf{x}, y)$, while the other is related to the samples $(\mathbf{x}_i, y_i), i \in [\![N]\!]$. The expectation in (4.9) (and

The version of record is available at: http://dx.doi.org/10.1561/2400000026

the probability in (4.10)) is taken w.r.t. the new data $(\mathbf{x}, y)$. For a given set of samples, (4.9) (and (4.10)) holds with probability at least $1 - \delta$ w.r.t. the measure of samples. Theorem 4.3.3 essentially says that given typical samples, the expected loss on new data using the Wasserstein DRO estimator could be bounded above by the average sample loss plus extra terms that depend on the supremum of $\|(-\boldsymbol{\beta}, 1)\|_*$ (the 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 [100] for the ℓ_2-norm-induced Wasserstein metric.

We also note that the upper bounds in (4.9) and (4.10) do not depend on the dimension of $(\mathbf{x}, y)$. This dimensionality-free characteristic implies direct applicability of the Wasserstein approach to high-dimensional settings and is particularly useful in many real applications where, potentially, hundreds of features may be present. Theorem 4.3.3 also provides guidance on the number of samples that are needed to achieve satisfactory out-of-sample performance.

Corollary 4.3.5. Suppose $\hat{\boldsymbol{\beta}}$ is the optimal solution to (4.5). For a fixed confidence level δ and some threshold parameter $\tau \geq 0$, if the sample size N satisfies

$$N \geq \left[\frac{2(1 + \sqrt{2\log(2/\delta)}\ )}{\tau} \right]^2, \tag{4.11}$$

then the percentage difference between the expected absolute loss on new data and the sample average loss is less than τ, that is,

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

Proof. The percentage difference requirement can be translated into:

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

from which (4.11) can be easily derived. □

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Corollary 4.3.6. Suppose $\hat{\boldsymbol{\beta}}$ is the optimal solution to (4.5). For a fixed confidence level δ, some $\tau \in (0,1)$ and $\gamma \geq 0$ such that $\tau\gamma + \tau - 1 > 0$, if the sample size N satisfies

$$N \geq \left[\frac{2(1+\sqrt{2\log(2/\delta)}\ )}{\tau\gamma+\tau-1}\right]^2, \tag{4.12}$$

then,

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

Proof. Based on Theorem 4.3.3, 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:

$$\frac{\gamma\bar{B}R - \frac{2\bar{B}R}{\sqrt{N}} - \bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^N |y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}| + \gamma\bar{B}R} \geq 1-\tau. \tag{4.13}$$

We cannot obtain a lower bound for N by directly solving (4.13) since N appears in a summation operator. A proper relaxation to (4.13) is:

$$\frac{\gamma - \frac{2}{\sqrt{N}} - \sqrt{\frac{8\log(2/\delta)}{N}}}{1+\gamma} \geq 1-\tau, \tag{4.14}$$

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 (4.14), we obtain (4.12). □

In Corollaries 4.3.5 and 4.3.6, 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.

4.3.2 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 subsection, we seek to bound

The version of record is available at: http://dx.doi.org/10.1561/2400000026

the difference between the estimated and true regression coefficients, under a certain distributional assumption on $(\mathbf{x}, y)$. Throughout this subsection we will use $\hat{\boldsymbol{\beta}}$ to denote the estimated regression coefficients, obtained as an optimal solution to (4.15), and $\boldsymbol{\beta}^*$ for the true (unknown) regression coefficients. The bound we will derive turns out to be related to the uncertainty in the data $(\mathbf{x}, y)$, and the geometric structure of the true regression coefficients.

To facilitate the analysis, we will use the following equivalent form of Problem (4.5):

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|(-\boldsymbol{\beta}, 1)\|_* \\ \text{s.t.} \quad & \|(-\boldsymbol{\beta}, 1)'\mathbf{Z}\|_1 \leq \gamma_N, \end{aligned} \tag{4.15}$$

where $\mathbf{Z} = [(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)]$ is the matrix with columns $(\mathbf{x}_i, y_i)$, $i \in [\![N]\!]$, and γ_N is some exogenous parameter related to ϵ. One can show that for properly chosen γ_N, (4.15) produces the same solution with (4.5) [101]. (4.15) is similar to (11) in [102], 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 [102]. 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 relevant definitions and assumptions that are needed to bound the estimation error.

Definition 4 (Sub-Gaussian Random Variable)**.** A random variable z is sub-Gaussian if the ψ_2-norm defined below is finite, i.e.,

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

We do not require sub-Gaussian variables to have zero mean values. It is though worth noting that the ψ_2-norm $|||z|||_{\psi_2}$ depends on the mean $\mathbb{E}(z)$. An equivalent property for sub-Gaussian random variables is that their tail distribution decays as fast as a Gaussian, namely,

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

for some constant C.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

A random vector $\mathbf{z} \in \mathbb{R}^{p+1}$ is sub-Gaussian if $\mathbf{z}'\mathbf{u}$ is sub-Gaussian for any $\mathbf{u} \in \mathbb{R}^{p+1}$. The ψ_2-norm of a vector $\mathbf{z}$ is defined as:

$$\|\mathbf{z}\|_{\psi_2} \triangleq \sup_{\mathbf{u}\in\mathcal{S}^{p+1}} \|\mathbf{z}'\mathbf{u}\|_{\psi_2},$$

where $\mathcal{S}^{p+1}$ denotes the unit sphere in the $(p+1)$-dimensional Euclidean space. For the properties of sub-Gaussian random variables/vectors, please refer to [103].

Definition 5 (Gaussian Width)**.** For any set $\mathcal{A} \subseteq \mathbb{R}^{p+1}$, its Gaussian width is defined as:

$$w(\mathcal{A}) \triangleq \mathbb{E}\Big[\sup_{\mathbf{u}\in\mathcal{A}} \mathbf{u}'\mathbf{g}\Big], \tag{4.16}$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a $(p+1)$-dimensional standard Gaussian random vector.

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

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

Assumption L. The true coefficient $\boldsymbol{\beta}^*$ is a feasible solution to (4.15), i.e.,

$$\|\mathbf{Z}'(-\boldsymbol{\beta}^*, 1)\|_1 \le \gamma_N.$$

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

$$\|(\mathbf{x}, y)\|_{\psi_2} = \sup_{\mathbf{u}\in\mathcal{S}^{p+1}} \|(\mathbf{x}, y)'\mathbf{u}\|_{\psi_2} \le \mu.$$

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

$$0 < \lambda_{\min} \triangleq \lambda_{\min}(\boldsymbol{\Gamma}) \le \lambda_{\max}(\boldsymbol{\Gamma}) \triangleq \lambda_{\max} < \infty.$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Notice that both $\underline{\alpha}$ in Assumption K and γ_N in Assumption L 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 [102], 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. The bound derived in Theorem 4.3.7 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 4.3.7. Suppose the true regression coefficient vector is $\boldsymbol{\beta}^*$ and the solution to (4.15) is $\hat{\boldsymbol{\beta}}$. For the set $\mathcal{A}(\boldsymbol{\beta}^*) = \text{cone}\{\mathbf{v}\colon \|(-\boldsymbol{\beta}^*, 1) + \mathbf{v}\|_* \leq \|(-\boldsymbol{\beta}^*, 1)\|_*\} \cap \mathcal{S}^{p+1}$, under Assumptions I, K, and L, we have:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{2R\gamma_N}{\underline{\alpha}}\Psi(\boldsymbol{\beta}^*), \tag{4.17}$$

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

Proof. For ease of exposition, we will adopt the notation $\mathbf{z} \triangleq (\mathbf{x}, y)$, $\mathbf{z}_i \triangleq (\mathbf{x}_i, y_i)$, $\tilde{\boldsymbol{\beta}} \triangleq (-\boldsymbol{\beta}, 1)$, $\tilde{\boldsymbol{\beta}}_{\text{est}} \triangleq (-\hat{\boldsymbol{\beta}}, 1)$, $\tilde{\boldsymbol{\beta}}_{\text{true}} \triangleq (-\boldsymbol{\beta}^*, 1)$.

Since both $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ are feasible (the latter due to Assumption L), 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 (4.15) 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 K, this yields

$$(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})'\mathbf{Z}\mathbf{Z}'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}) \geq \underline{\alpha}\|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_2^2. \tag{4.18}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

On the other hand, from Hölder's inequality:

$$\begin{aligned}
&(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})'\mathbf{Z}\mathbf{Z}'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}) \\
&\quad \le \|\mathbf{Z}'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})\|_1 \|\mathbf{Z}'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})\|_\infty \\
&\quad \le 2\gamma_N \max_i |\mathbf{z}_i'(\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}})| \\
&\quad \le 2\gamma_N \max_i \|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_* \|\mathbf{z}_i\| \\
&\quad \le 2R\gamma_N \|\tilde{\boldsymbol{\beta}}_{\text{est}} - \tilde{\boldsymbol{\beta}}_{\text{true}}\|_*.
\end{aligned} \tag{4.19}$$

Combining (4.18) and (4.19), we have:

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

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

As mentioned earlier, (4.17) 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 quantity $\underline{\alpha}$ acts as the minimum eigenvalue of the matrix $\mathbf{Z}\mathbf{Z}'$ restricted to a subspace of $\mathbb{R}^{p+1}$, and thus a proper substitute should be related to the minimum eigenvalue of the covariance matrix of $(\mathbf{x}, y)$, i.e., the $\boldsymbol{\Gamma}$ matrix (cf. Assumption N), given that $(\mathbf{x}, y)$ is zero mean. See Lemmata 4.3.8, 4.3.9 and 4.3.10 for the derivation.

Lemma 4.3.8. Consider $\mathcal{A}_{\boldsymbol{\Gamma}} = \{\mathbf{w} \in \mathcal{S}^{p+1}\colon \boldsymbol{\Gamma}^{-1/2}\mathbf{w} \in \text{cone}(\mathcal{A}(\boldsymbol{\beta}^*))\}$, where $\mathcal{A}(\boldsymbol{\beta}^*)$ is defined as in Theorem 4.3.7, and $\boldsymbol{\Gamma} = \mathbb{E}[(\mathbf{x}, y)(\mathbf{x}, y)']$. Under Assumptions M and N, when the sample size $N \ge C_1\bar{\mu}^4(w(\mathcal{A}_{\boldsymbol{\Gamma}}))^2$, where $\bar{\mu} = \mu\sqrt{\frac{1}{\lambda_{\min}}}$, and $w(\mathcal{A}_{\boldsymbol{\Gamma}})$ is the Gaussian width of $\mathcal{A}_{\boldsymbol{\Gamma}}$, with probability at least $1 - \exp(-C_2 N/\bar{\mu}^4)$, we have

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

where C_1 and C_2 are positive constants.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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

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

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

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

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

where the first inequality used Assumption M and the second inequality used Assumption N.

Applying Theorem D from [104], for any $\theta > 0$ and when

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

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

$$\begin{aligned}\sup_{f_{\mathbf{w}}\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{N} f_{\mathbf{w}}^2(\mathbf{z}_i) - \mathbb{E}[f_{\mathbf{w}}^2]\right| &= \sup_{f_{\mathbf{w}}\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{N} \mathbf{w}'\mathbf{\Gamma}^{-1/2}\mathbf{z}_i\mathbf{z}_i'\mathbf{\Gamma}^{-1/2}\mathbf{w} - 1\right| \\ &= \sup_{\mathbf{w}\in\mathcal{A}_{\mathbf{\Gamma}}} |\mathbf{w}'\mathbf{\Gamma}^{-1/2}\hat{\mathbf{\Gamma}}\mathbf{\Gamma}^{-1/2}\mathbf{w} - 1| \\ &\le \theta, \qquad (4.20)\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where $\tilde{C}_1$ is some positive constant and $\gamma_2(\mathcal{F}, \|\|\cdot\|\|_{\psi_2})$ is defined in [104] 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 [102], 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 Equation (44) in [102]. 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 $1 - \exp(-C_2 N/\bar{\mu}^4)$, for some positive constant C_2, (4.20) holds with $\theta = 1/2$. This implies that for all $\mathbf{w} \in \mathcal{A}_{\mathbf{\Gamma}}$,

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

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. □

Note that the sample size requirement stated in Lemma 4.3.8 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 4.3.9 (Lemma 4 in [102]). Let μ_0 be the ψ_2-norm of a standard Gaussian random vector $\mathbf{g} \in \mathbb{R}^{p+1}$, and $\mathcal{A}_{\mathbf{\Gamma}}$, $\mathcal{A}(\boldsymbol{\beta}^*)$ be defined as in Lemma 4.3.8. Then, under Assumption N,

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

for some positive constant C_3.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Proof. We follow the proof of Lemma 4 in [102], adapted to our setting. We include all key steps for completeness.

Recall the definition of the Gaussian width $w(\mathcal{A}_{\mathbf{\Gamma}})$ (cf. (4.16)):

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

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

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

where $\mathcal{B}^{p+1}$ is the unit ball in the $(p+1)$-dimensional Euclidean space and the inequality used Assumption N and the fact that

$$\mathbf{w}'\mathbf{\Gamma}^{-1/2}/\|\mathbf{\Gamma}^{-1/2}\mathbf{w}\|_2 \in \mathcal{B}^{p+1}, \quad \mathbf{w} \in \mathcal{A}_{\mathbf{\Gamma}}.$$

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

$$\begin{aligned}
\|\!|S_{\mathbf{v}_1} - S_{\mathbf{v}_2}|\!\|_{\psi_2} &= \Big|\!\Big|\!\Big|(\mathbf{v}_1 - \mathbf{v}_2)'\mathbf{\Gamma}^{1/2}\mathbf{g}\Big|\!\Big|\!\Big|_{\psi_2} \\
&= \|\mathbf{\Gamma}^{1/2}(\mathbf{v}_1 - \mathbf{v}_2)\|_2 \Big|\!\Big|\!\Big| \frac{(\mathbf{v}_1 - \mathbf{v}_2)'\mathbf{\Gamma}^{1/2}\mathbf{g}}{\|\mathbf{\Gamma}^{1/2}(\mathbf{v}_1 - \mathbf{v}_2)\|_2} \Big|\!\Big|\!\Big|_{\psi_2} \\
&\le \|\mathbf{\Gamma}^{1/2}(\mathbf{v}_1 - \mathbf{v}_2)\|_2 \sup_{\mathbf{u}\in\mathcal{S}^{p+1}} \|\!|\mathbf{u}'\mathbf{g}|\!\|_{\psi_2} \\
&= \mu_0 \|\mathbf{\Gamma}^{1/2}(\mathbf{v}_1 - \mathbf{v}_2)\|_2 \\
&\le \mu_0 \sqrt{\lambda_{\max}} \|\mathbf{v}_1 - \mathbf{v}_2\|_2,
\end{aligned}$$

where the last step used Assumption N.

Then, by the tail behavior of sub-Gaussian random variables (see Hoeffding bound, Theorem 2.6.2 in [103]), we have:

$$\mathbb{P}(|S_{\mathbf{v}_1} - S_{\mathbf{v}_2}| \ge \delta) \le 2\exp\Big(-\frac{C_{01}\delta^2}{\mu_0^2 \lambda_{\max} \|\mathbf{v}_1 - \mathbf{v}_2\|_2^2}\Big),$$

for some positive constant C_{01}.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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 [102],

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

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 4.3.8. Since $\mathcal{T} = \text{cone}(\mathcal{A}(\boldsymbol{\beta}^*)) \cap \mathcal{B}^{p+1} \subseteq \text{conv}(\mathcal{A}(\boldsymbol{\beta}^*) \cup \{\mathbf{0}\})$, by Lemma 2 in [105],

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

Thus,

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

□

Combining Lemmata 4.3.8 and 4.3.9, and expressing the covariance matrix $\boldsymbol{\Gamma}$ using its eigenvalues, we arrive at the following result.

Corollary 4.3.10. Under Assumptions M and N, and the conditions in Lemmata 4.3.8 and 4.3.9, when

$$N \geq \bar{C}_1\bar{\mu}^4\mu_0^2 \cdot \frac{\lambda_{\max}}{\lambda_{\min}}(w(\mathcal{A}(\boldsymbol{\beta}^*)) + 3)^2,$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

with probability at least $1 - \exp(-C_2 N/\bar{\mu}^4)$,

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

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

Proof. Combining Lemmata 4.3.8 and 4.3.9, and using the fact that for any $\mathbf{v} \in \mathcal{A}(\boldsymbol{\beta}^*)$,

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

we can derive the desired result. □

Next we derive the smallest possible value of γ_N such that $\boldsymbol{\beta}^*$ is feasible.

Lemma 4.3.11. Under Assumptions I and J, for any feasible $\boldsymbol{\beta}$,

$$\|(-\boldsymbol{\beta}, 1)'\mathbf{Z}\|_1 \leq N\bar{B}R, \quad \text{a.s. under } \mathbb{P}^*.$$

Combining Theorem 4.3.7, Corollary 4.3.10 and Lemma 4.3.11, we have the following main performance guarantee result that bounds the estimation bias of the solution to (4.15).

Theorem 4.3.12. Under Assumptions I–N, and the conditions of Theorem 4.3.7, Corollary 4.3.10 and Lemma 4.3.11, when

$$N \geq \bar{C}_1 \bar{\mu}^4 \mu_0^2 \cdot \frac{\lambda_{\max}}{\lambda_{\min}} (w(\mathcal{A}(\boldsymbol{\beta}^*)) + 3)^2,$$

with probability at least $1 - \exp(-C_2 N/\bar{\mu}^4)$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{4R^2\bar{B}}{\lambda_{\min}} \Psi(\boldsymbol{\beta}^*). \tag{4.21}$$

The estimation error bound in (4.21) depends on the variance of $(\mathbf{x}, y)$, and the geometrical structure of the true regression coefficient. It does not decay to zero as N goes to infinity. The reason is that the absolute residual $|y - \boldsymbol{\beta}'\mathbf{x}|$ has a nonzero mean, which will be propagated into the estimation bias.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

4.4 Experiments on the Performance of Wasserstein DRO

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 the Wasserstein formulation is in the following form:

$$\inf_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathbf{x}_i' \boldsymbol{\beta}| + \epsilon \|(-\boldsymbol{\beta}, 1)\|_*. \tag{4.22}$$

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 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 (4.22), and the other one induced by the $\|\cdot\|_\infty$ (Wasserstein ℓ_∞) and resulting in an ℓ_1-regularizer in (4.22).

The problem of feature selection can be formulated as an ℓ_0-norm regularized regression problem, which is NP-hard and is usually relaxed to an ℓ_1-norm regularized formulation, known as the *Least Absolute Shrinkage and Selection Operator (LASSO)*. LASSO enjoys several attractive statistical properties under various conditions on the model matrix [6], [106]. Here, in our context, we try to offer an explanation of the sparsity-inducing property of LASSO from the perspective of the Wasserstein DRO formulation, through projecting the sparsity of $\boldsymbol{\beta}^*$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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:

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

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 $\mathbf{M} \in \mathbb{R}^{(p+1)\times(p+1)}$ a diagonal matrix whose diagonal elements are $(1, \ldots, 1, c^2)$, and then applying (4.8). Alternatively, if we let $s_c(\mathbf{x}, y) = \|(\mathbf{x}, cy)\|_\infty$, Corollary 4.4.1 shows that as $c \to \infty$, the corresponding Wasserstein formulation becomes the ℓ_1-regularized LAD.

Corollary 4.4.1. If the Wasserstein metric is induced by the following metric s:

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

with c some positive constant. Then as $c \to \infty$, the Wasserstein DRO formulation (4.22) reduces to:

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

which is the ℓ_1-regularized LAD.

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

$$\|(\mathbf{x}, y)\|_{\mathbf{w},r} \triangleq \|(x_1 w_1, \ldots, x_p w_p, y w_{p+1})\|_r,$$

for some $(p+1)$-dimensional weighting vector $\mathbf{w} = (w_1, \ldots, w_{p+1})$, and $r \geq 1$. 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

The version of record is available at: http://dx.doi.org/10.1561/2400000026

of $\|\cdot\|_{\mathbf{w},\infty}$. Hölder's inequality [107] will be used for the derivation. 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 $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$:

$$\begin{aligned} \max_{\mathbf{z}} \quad & \mathbf{z}'\tilde{\boldsymbol{\beta}} \\ \text{s.t.} \quad & \|\mathbf{z}\|_{\mathbf{w},\infty} \leq 1. \end{aligned} \tag{4.23}$$

The optimal value of Problem (4.23), which is a function of $\tilde{\boldsymbol{\beta}}$, gives the dual norm evaluated at $\tilde{\boldsymbol{\beta}}$. Using Hölder's inequality, we can write

$$\begin{aligned} \mathbf{z}'\tilde{\boldsymbol{\beta}} &= \sum_{i=1}^{p+1} (w_i z_i)\left(\frac{1}{w_i}\tilde{\beta}_i\right) \\ &\leq \|\mathbf{z}\|_{\mathbf{w},\infty} \|\tilde{\boldsymbol{\beta}}\|_{\mathbf{w}^{-1},1} \\ &\leq \|\tilde{\boldsymbol{\beta}}\|_{\mathbf{w}^{-1},1}, \end{aligned}$$

where $\mathbf{w}^{-1} \triangleq (\frac{1}{w_1}, \ldots, \frac{1}{w_{p+1}})$. 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 $c \to \infty$:

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

□

It follows that regularizing over $\boldsymbol{\beta}$ implies an infinite transportation cost along y. By contrast, the 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 with several commonly used regression models that employ the SR loss function, e.g., ridge regression [100], LASSO [5], and *Elastic Net (EN)* [108]. We will also compare against M-estimation [7], [8], which uses a variant of the SR loss and is equivalent to solving a weighted least squares problem. These models will be compared under two different experimental setups, one involving

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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.

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 $\mathbf{x} \in \mathbb{R}^p$ from the normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{x}$, which is just the top left block of the matrix $\boldsymbol{\Gamma}$ in Assumption N. Specifically, $\boldsymbol{\Gamma} = \mathbb{E}[(\mathbf{x}, y)(\mathbf{x}, y)']$ is equal to
  $$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{\beta}^* \\ (\boldsymbol{\beta}^*)'\boldsymbol{\Sigma} & (\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^* + \sigma^2 \end{bmatrix},$$
  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 $\mathcal{N}(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2)$.
- Otherwise, depending on the experimental setup, generate an outlier that is either:
  - Abnormal in both $\mathbf{x}$ and y, with outlying distribution:
    1. $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) + \mathcal{N}(5\mathbf{e}, \mathbf{I})$, or $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) + \mathcal{N}(\mathbf{0}, 0.25\mathbf{I})$;
    2. $y \sim \mathcal{N}(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2) + 5\sigma$.
  - Abnormal only in $\mathbf{x}$:
    1. $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) + \mathcal{N}(5\mathbf{e}, \mathbf{I})$;
    2. $y \sim \mathcal{N}(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2)$.
- Repeat the above procedure for N times, where N is the size of the training set.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

To test the generalization ability of various formulations, we generate a test dataset containing M samples from the clean distribution. 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 $(\mathbf{x}_i, y_i),\ i \in [\![M]\!]$, 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.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

- 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.

All the regularization parameters are tuned on a separate validation dataset using the *Median Absolute Deviation (MAD)* as a selection criterion, 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 [109], where the LASSO was tuned over 50 values ranging from $\lambda_{\max} = \|\mathbf{X}'\mathbf{y}\|_\infty$ to a small fraction of $\lambda_{\max}$ on a log scale, with $\mathbf{X} \in \mathbb{R}^{N\times p}$ 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:

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

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.

4.4.1 Dense β^*, Outliers in Both $\mathbf{x}$ and y

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

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

We generate 10 datasets consisting of $N = 100, M = 60$ observations. The probability of a training sample being drawn from the outlying

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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 Figures 4.2 and 4.3. 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.

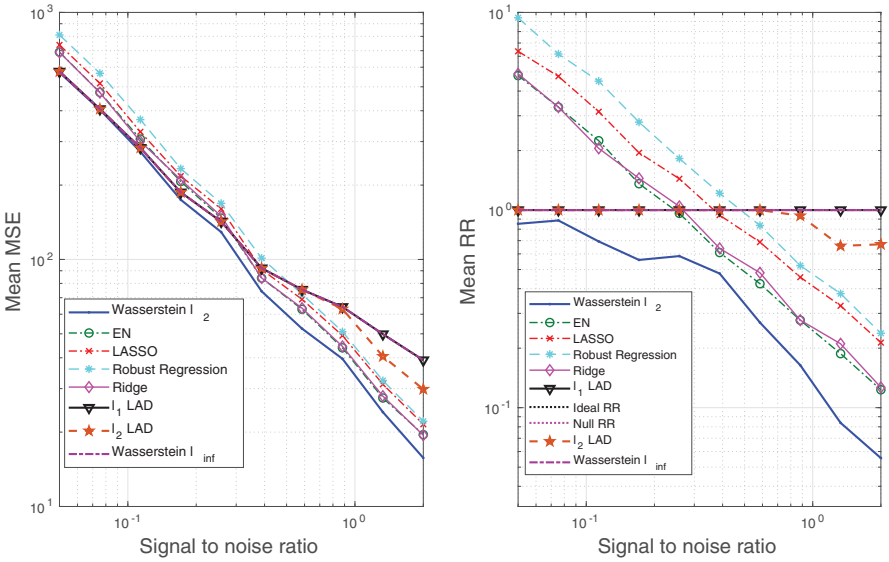
It can be seen that as the SNR decreases or the correlation between the predictors increases, the estimation problem becomes harder, and the performance of all approaches gets worse. In general the Wasserstein formulation with an ℓ_2-norm transportation cost achieves the best performance in terms of 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 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. 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.

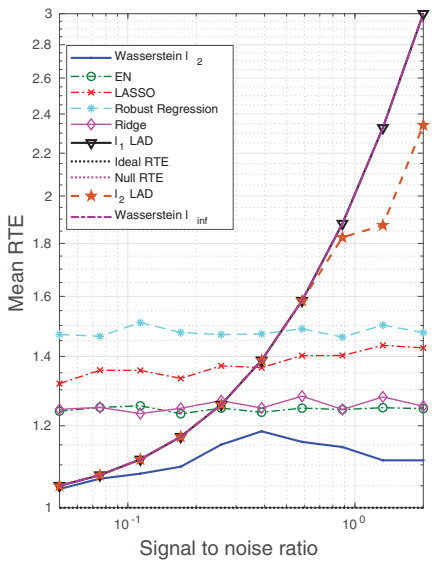
Overall the ℓ_2-regularizer outperforms the ℓ_1-regularizer, since the true regression coefficient is dense, which implies that a proper distance metric on the $(\mathbf{x}, y)$ space should take into account all the coordinates. From the perspective of the Wasserstein DRO framework, the ℓ_1-regularizer 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 4.4.3 and 4.4.4.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

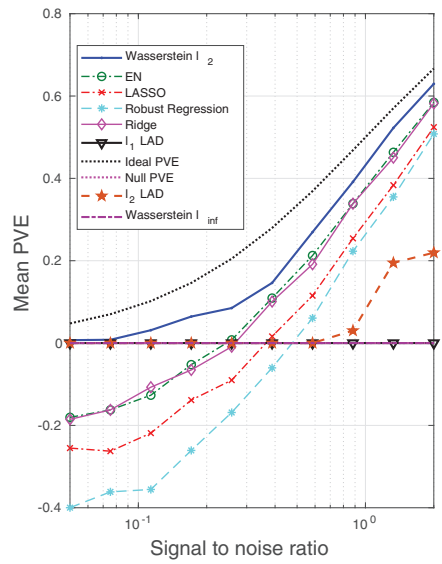


(a) Mean Squared Error.

(b) Relative risk.



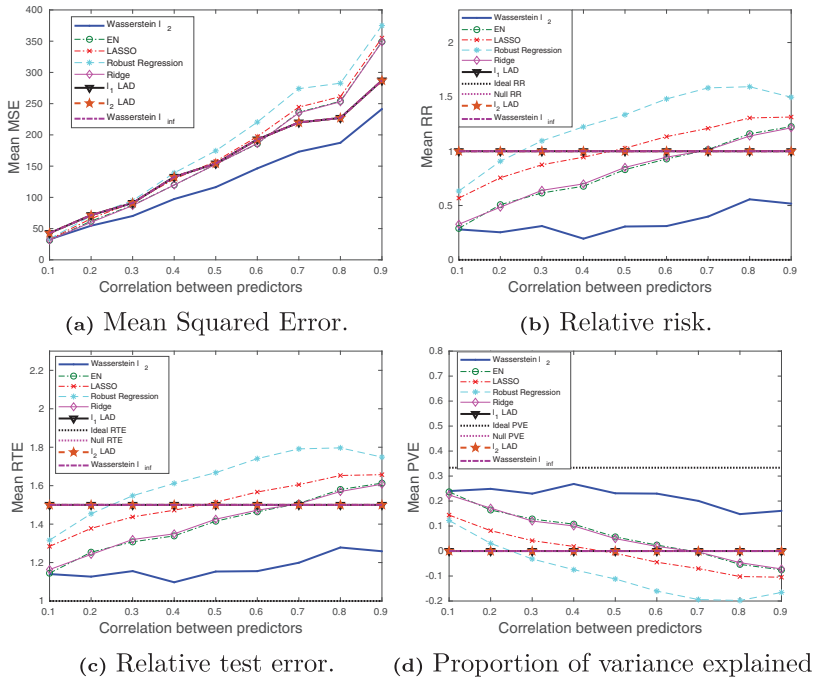
(c) Relative test error.



(d) Proportion of variance explained.

Figure 4.2: The impact of SNR on the performance metrics: Dense $\boldsymbol{\beta}^*$, outliers in both $\mathbf{x}$ and y.

The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) Mean Squared Error. (b) Relative risk.

(c) Relative test error. (d) Proportion of variance explained.

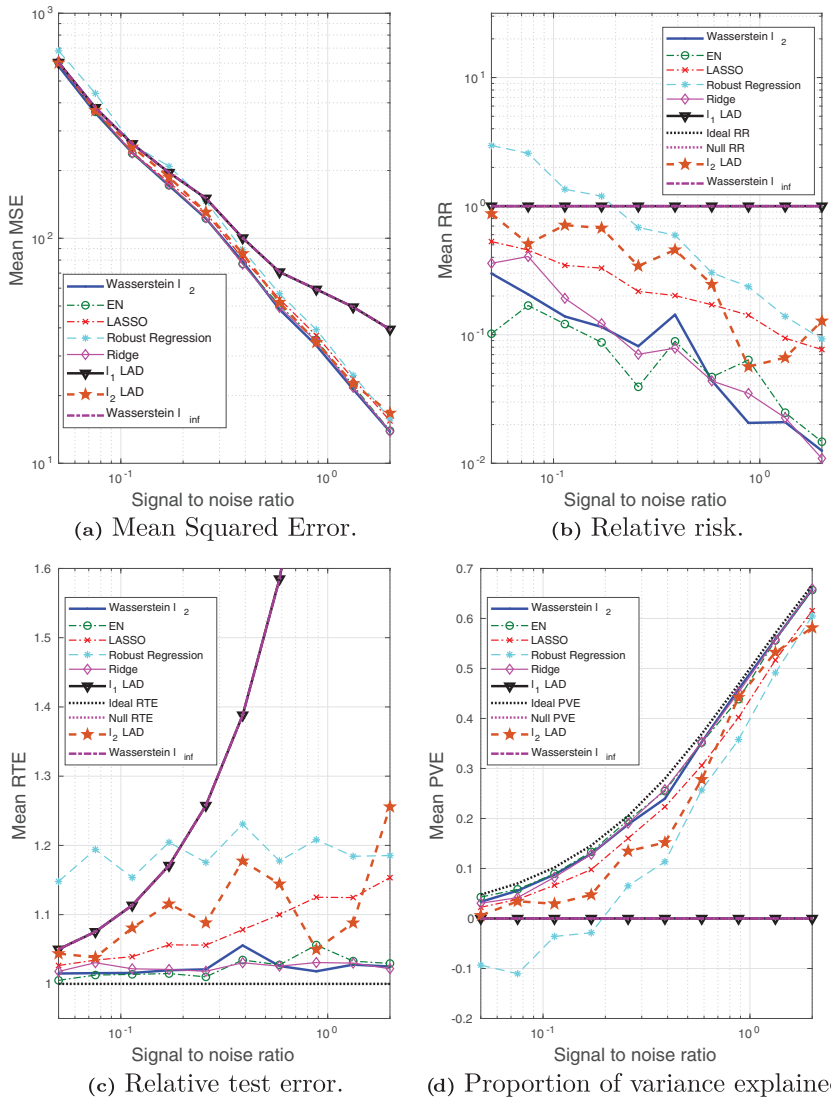
Figure 4.3: The impact of predictor correlation on the performance metrics: Dense β^*, outliers in both $\mathbf{x}$ and y.

4.4.2 Dense β^*, Outliers Only in $\mathbf{x}$

In this subsection, we will experiment with the same β^* as in Section 4.4.1, but with perturbations only in $\mathbf{x}$. Our goal is to investigate the performance of the Wasserstein formulation when the response y is not subjected to large perturbations.

Interestingly, we observe 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 Figures 4.4 and 4.5). Notably, the ℓ_2-regularized LAD, which is just slightly different from the Wasserstein ℓ_2 formulation, shows a much worse performance. This is because the ℓ_2-regularized LAD implicitly assumes an infinite transportation

The version of record is available at: http://dx.doi.org/10.1561/2400000026



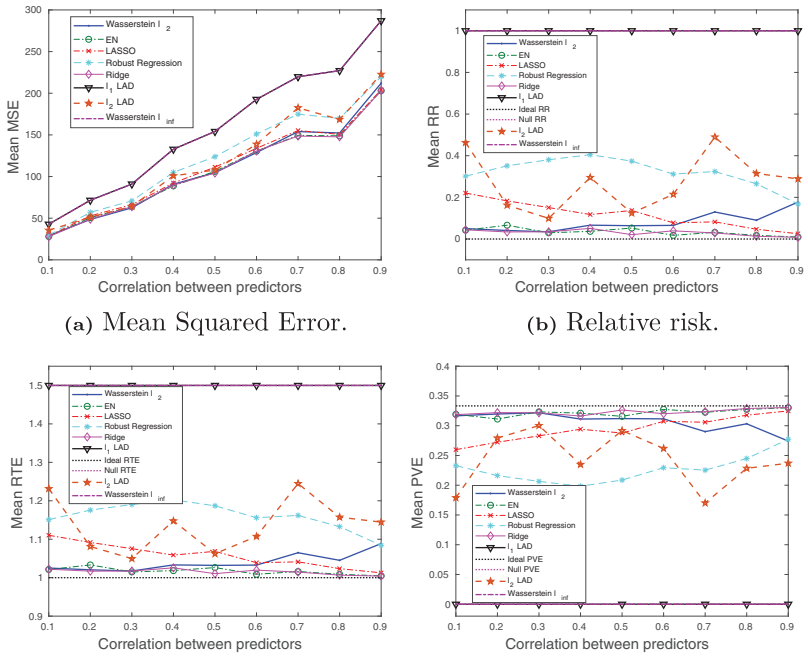
(a) Mean Squared Error. (b) Relative risk.

(c) Relative test error. (d) Proportion of variance explained.

Figure 4.4: The impact of SNR on the performance metrics: Dense $\boldsymbol{\beta}^*$, outliers only in $\mathbf{x}$.

cost along y, which gives zero tolerance to the variation in the response. Therefore, a reasonable amount of fluctuation, caused by the intrinsic randomness of y, would be overly exaggerated by the underlying metric

The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) Mean Squared Error. (b) Relative risk.

(c) Relative test error. (d) Proportion of variance explained.

Figure 4.5: The impact of predictor correlation on the performance metrics: Dense $\boldsymbol{\beta}^*$, outliers only in $\mathbf{x}$.

used by the ℓ_2-regularized LAD. In contrast, the 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 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 $\mathbf{x}$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

4.4.3 Sparse β^*, Outliers in both $\mathbf{x}$ 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 perturbations are present in both $\mathbf{x}$ and y. Specifically, the distribution of outliers is characterized by:

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

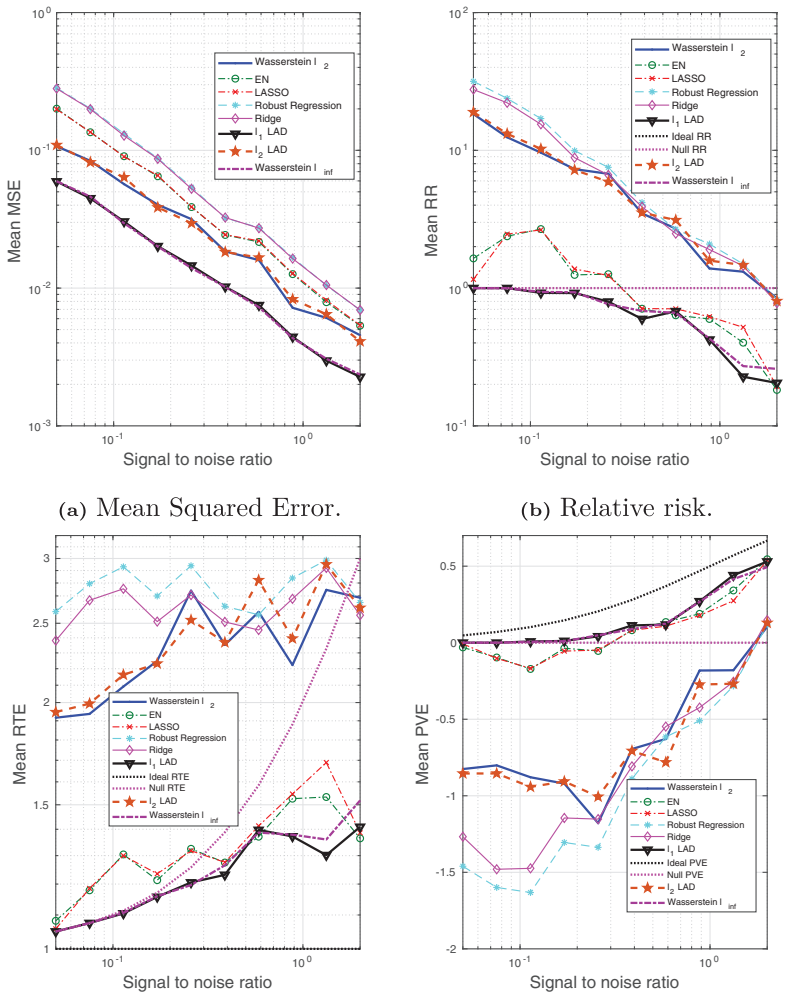
Our goal is to study the impact of the sparsity of $\boldsymbol{\beta}^*$ on the choice of the norm space for the Wasserstein metric. An intuitively appealing interpretation for the sparsity inducing property of the ℓ_1-regularizer is made available by the Wasserstein DRO framework, which we explain 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 induces the ℓ_1-regularizer is expected to outperform others.

The results are summarized in Figures 4.6 and 4.7. We note that the ℓ_1-regularized LAD achieves a similar performance, 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, the Wasserstein formulation is guaranteed to outperform all others.

4.4.4 Sparse β^*, Outliers Only in $\mathbf{x}$

In this subsection, we will use the same sparse coefficient as in Section 4.4.3, but the perturbations are present only in $\mathbf{x}$. Specifically, for

The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) Mean Squared Error. (b) Relative risk.

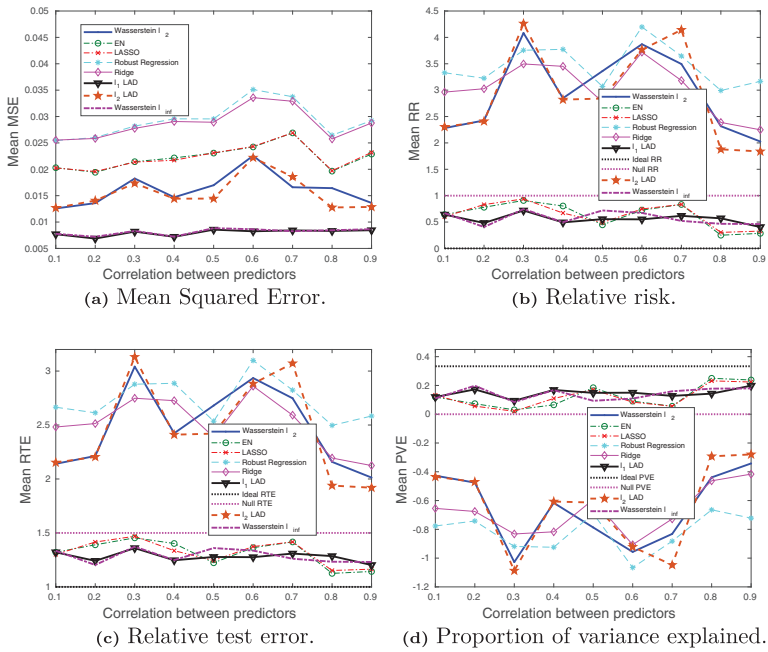
(c) Relative test error. (d) Proportion of variance explained.

Figure 4.6: The impact of SNR on the performance metrics: Sparse $\boldsymbol{\beta}^*$, outliers in both $\mathbf{x}$ and y.

outliers, their predictors and responses are drawn from the following distributions:

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

The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) Mean Squared Error.

(b) Relative risk.

(c) Relative test error.

(d) Proportion of variance explained.

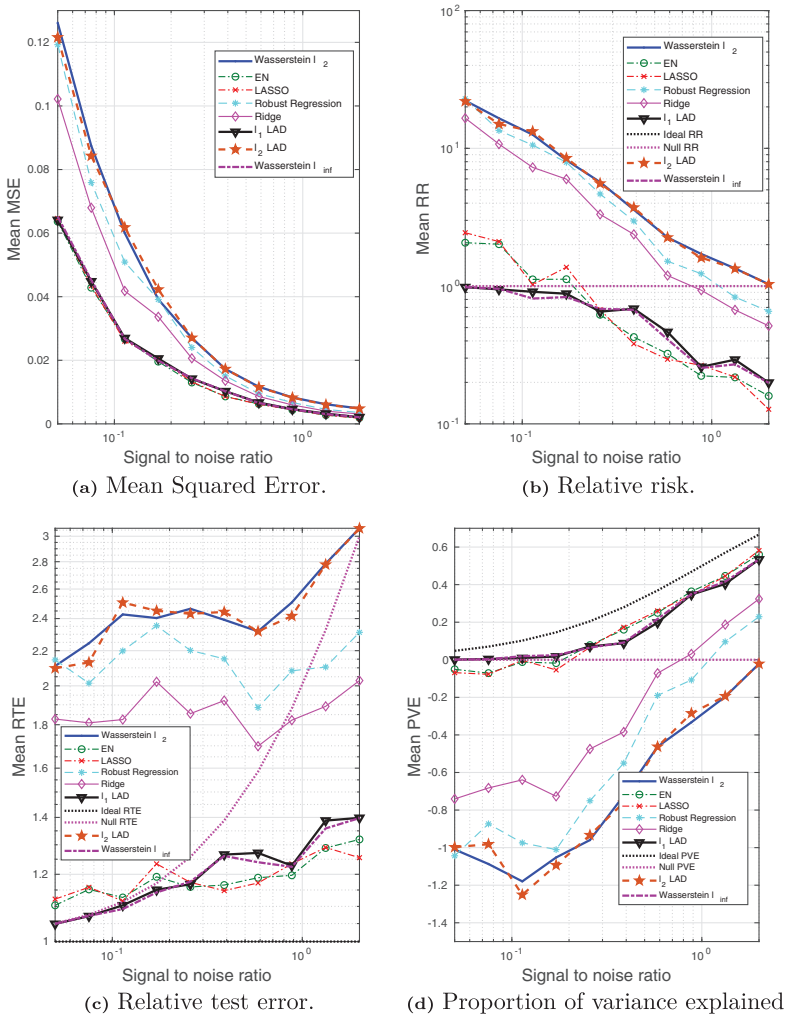
Figure 4.7: The impact of predictor correlation on the performance metrics: Sparse $\boldsymbol{\beta}^*$, outliers in both $\mathbf{x}$ and y.

The results are summarized in Figures 4.8 and 4.9. Not surprisingly, the Wasserstein ℓ_∞ and the ℓ_1-regularized LAD achieve the best performance. Notice that in Section 4.4.3, 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 Figure 4.8). For the ℓ_1-regularized formulations, the Wasserstein ℓ_∞ formulation, as well as the ℓ_1-regularized LAD, is comparable with the EN and LASSO.

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.

The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) Mean Squared Error.

(b) Relative risk.

(c) Relative test error.

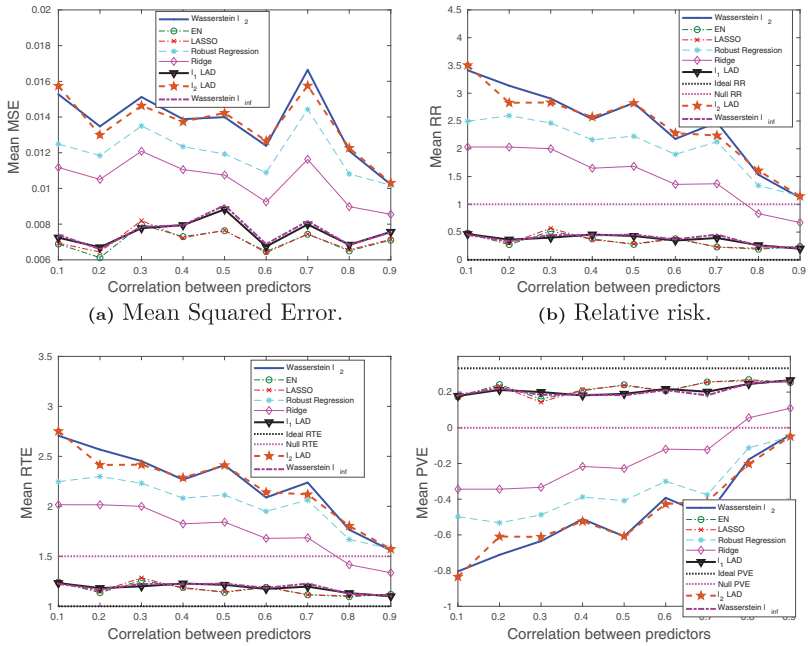
(d) Proportion of variance explained.

Figure 4.8: The impact of SNR on the performance metrics: Sparse $\boldsymbol{\beta}^*$, outliers only in $\mathbf{x}$.

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.

The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) Mean Squared Error.

(b) Relative risk.

(c) Relative test error.

(d) Proportion of variance explained.

Figure 4.9: The impact of predictor correlation on the performance metrics: Sparse β^*, outliers only in $\mathbf{x}$.

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

4.5 An Application of Wasserstein DRO to Outlier Detection

As an application, we consider an unlabeled two-class classification problem, where the 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 [7], [8], against which we will compare in terms of the outlier detection capability.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

4.5.1 Experiments on Synthetic Data

We first report results on synthetic datasets that consist of a mixture of clean and outlying examples. For clean samples, all predictors $x_i, i \in [\![30]\!]$, 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 (4.5) with the ℓ_1-regularized LAD and M-estimation with three cost functions – Huber [7] and [8], Talwar [89], and Fair [90]. The performance metrics include 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 the Wasserstein DRO model reflected in its robustness w.r.t. the choice of ϵ. In Figure 4.10 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 version of record is available at: http://dx.doi.org/10.1561/2400000026

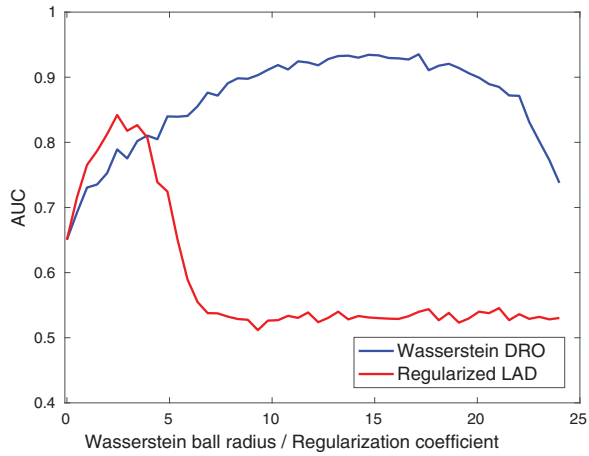
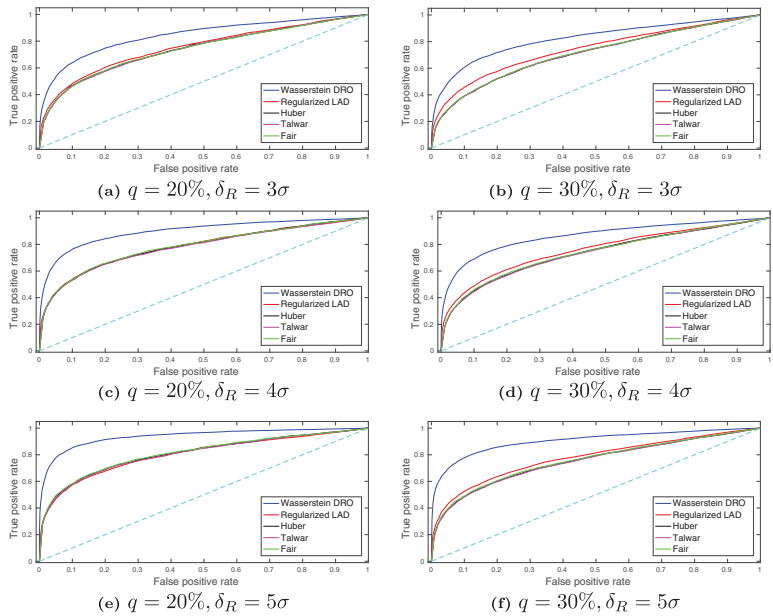


Figure 4.10: Out-of-sample AUC vs. Wasserstein ball radius (regularization coefficient).

the Wasserstein ball contains the true distribution with low confidence and thus AUC is low. On the other hand, too large ϵ makes the solution overly conservative. Note that the robustness of the Wasserstein DRO, indicated by the flatness of the 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.

In Figure 4.11 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. 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). M-estimation adopts an *Iteratively Reweighted Least Squares (IRLS)* procedure which assigns weights to data points based on the residuals from previous iterations. 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 OLS.

The version of record is available at: http://dx.doi.org/10.1561/2400000026



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

4.5.2 CT Radiation Overdose Detection

In this section we consider an application of Wasserstein DRO regression to CT radiation overdose detection [110]. The goal is to identify all CT scans with an unanticipated high radiation exposure, given the characteristics of the patient and the type of the exam. This could be cast 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. Given such a regression plane, abnormal CT scans can be identified by the residuals of the regression.

The data was obtained from a HIPAA-compliant, Institutional Review Board (IRB)-approved retrospective cohort study that was conducted at an academic medical system including a 793-bed quaternary care hospital, and two outpatient imaging facilities. The original de-identified dataset contained 28 fields for 189,959 CT exams, and the per acquisition CT Dose Index (CTDI), which measures the amount

The version of record is available at: http://dx.doi.org/10.1561/2400000026

of exposure to CT radiation. Mean patient age was 60.6 ± 17.1 years; 54.7% were females.

The data was pre-processed as follows: (i) patient visits with more than half of the corresponding variables missing, or a missing value for CTDI, were discarded; (ii) categorical variables were encoded using indicator variables, and categories present only in a small number of exams were deleted; (iii) variables that have low correlation with CTDI were removed from further consideration; (iv) missing values were imputed by the mean (for numerical predictors) or mode (for categorical predictors); (v) all predictors were standardized by subtracting the mean and dividing by the standard deviation.

After pre-processing, we were left with 606 numerically encoded predictors for 88,566 CT exams. We first applied the variable selection method LASSO to select important variables for predicting CTDI, and then employed the Wasserstein DRO regression approach (induced by the ℓ_2 norm) to learn a predictive model of CT radiation doses given important variables identified by LASSO. Patient visits whose predicted radiation dose was statistically different from the radiation dose actually received were identified as outliers.

To assess the accuracy of the outlier cohort discovery process, we conducted a manual validation in which the results of a human-expert classification were compared to those extracted by the algorithm. A validation sample size of 200 cases were reviewed, yielding specificity of 0.85 [95% CI 0.78–0.92] and sensitivity of 0.91 [95% CI 0.85–0.97] (Positive Predictive Value PPV = 0.84, Negative Predictive Value NPV = 0.92).

We compared against two alternatives on the same validation set of 200 samples that were reviewed by the human expert. The first alternative method is what we call a "cutoff" method. We computed the average and standard deviation of CTDI over a training set and identified as outlying exams where the CTDI was larger than the average plus 3 times the standard deviation. The second alternative method used OLS in lieu of the Wasserstein DRO regression, and the regression residuals (this time from OLS) were used to detect outliers. The results are reported in Table 4.1, showing an improvement of 72.5% brought by the Wasserstein DRO method in terms of the F_1 score, which is defined as the harmonic mean of sensitivity and PPV. For an additional

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Table 4.1: Comparison of Wasserstein DRO regression against OLS and the cutoff method on CT radiation data

	Sensitivity	**Specificity**	**PPV**	**NPV**	**F_1 Score**
Wasserstein ℓ_2	0.91	0.85	0.84	0.92	0.88
OLS	0.36	0.95	0.87	0.64	0.51
Cutoff	0.37	0.94	0.83	0.64	0.51

point of comparison, we considered the top-40 outliers identified by each method. Among these outliers, 7 of the top-40 OLS outliers (17.5%) were considered to be "false positives"; while all the top-40 outliers detected by Wasserstein DRO were real outliers.

4.6 Summary

In this section, we presented a novel ℓ_1-loss based robust learning procedure using *Distributionally Robust Optimization (DRO)* under the Wasserstein metric in a linear regression setting, through which a delicate connection between the metric space on data and the regularization term has been established. The 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 this approach rests in its flexibility to adjust the form of the regularizer based on the characteristics of the data.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

5

Distributionally Robust Grouped Variable Selection

In this section, we will discuss a special case of the general formulation (4.5) tailored for selecting grouped variables that are relevant to the response when there exists a predefined grouping structure for the predictors. An example of this is the encoding of a categorical predictor using a group of indicator variables. Jointly selecting/dropping all variables in a group gives rise to more interpretable models. To perform variable selection at a group level, the *Grouped LASSO (GLASSO)* proposed by [111] and [112], imposes a block-wise ℓ_2-normed penalty for the grouped coefficient vectors. We will show that by using a special norm ($\|\cdot\|_{2,\infty}$) on the data space, the Wasserstein DRO formulation recovers the GLASSO penalty under the absolute residual loss (regression) and the log-loss (classification). The resulting model offers robustness explanations for GLASSO algorithms and highlights the connection between robustification and regularization.

5.1 The Problem and Related Work

The *Grouped LASSO (GLASSO)* was first proposed by [111], [112] to induce sparsity at a group level, when there exists a predefined grouping structure for the predictors. Suppose the predictor $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L)$, and

The version of record is available at: http://dx.doi.org/10.1561/2400000026

the regression coefficient $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^L)$, where $\mathbf{x}^l, \boldsymbol{\beta}^l \in \mathbb{R}^{p_l}, l \in [\![L]\!]$, respectively represent the predictor and coefficient for group l which contains p_l predictors. GLASSO minimizes:

$$\inf_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 + \epsilon \sum_{l=1}^{L} \sqrt{p_l} \|\boldsymbol{\beta}^l\|_2,$$

where $(\mathbf{x}_i, y_i), i \in [\![N]\!]$, are N observed samples of $(\mathbf{x}, y)$. Several extensions have been explored. In particular, [113] and [114] considered grouped variable selection in nonparametric models. [115] and [116] explored GLASSO for overlapping groups. The group sparsity in general regression/classification models has also been investigated in several works, see, for example, [117]–[119] for GLASSO in logistic regression, and [120] for GLASSO in generalized linear models.

Most of the existing works endeavor to modify the GLASSO formulation heuristically to achieve various goals. As an example, [121] considers a convex combination of the GLASSO and LASSO penalties, called Sparse Grouped LASSO, to induce both group-wise and within group sparsity. [122] modified the residual sum of squares to its square root and proposed the *Grouped Square Root LASSO (GSRL)*. However, few of those works were able to provide a rigorous explanation or theoretical justification for the form of the penalty term.

In this section, we attempt to fill this gap by casting the problem of grouped variable selection into the Wasserstein DRO framework. We show that in *Least Absolute Deviation (LAD)* and *Logistic Regression (LG)*, for a specific norm-induced Wasserstein metric, the DRO model can be reformulated as a regularized empirical loss minimization problem, where the regularizer coincides with the GLASSO penalty, and its magnitude is equal to the radius of the distributional ambiguity set. Through such a reformulation we establish a connection between regularization and robustness and offer new insights into the GLASSO penalty term.

We note that such a connection between robustification and regularization has been explored in several works (see Section 4.2), but none of them considered grouped variable selection. This section sheds new light on the significance of exploring the group-wise DRO problem. It is worth noting that [123] has studied the group-wise regularization

The version of record is available at: http://dx.doi.org/10.1561/2400000026

estimator with the square root of the expected loss under the Wasserstein DRO framework and recovered the GSRL. Here, we present a more general framework that includes both the LAD and the negative log-likelihood loss functions, and recover the GLASSO penalty in both cases. Moreover, we point out the potential of generalizing such results to a class of loss functions with a finite growth rate.

The remainder of this section is organized as follows. Section 5.2 introduces the Wasserstein GLASSO formulations for LAD and LG. Section 5.3 establishes a desirable grouping effect, showing that the difference between coefficients within the same group converges to zero as $O(\sqrt{1-\rho})$, where ρ is their sample correlation. In light of this result, we use the spectral clustering algorithm to divide the predictors into a pre-specified number of groups. This renders the GLASSO algorithm completely *data-driven*, in the sense that no more information other than the data itself is needed. Section 5.4 presents numerical results on both synthetic data and a real very large dataset with surgery-related medical records. Conclusions are in Section 5.5.

5.2 The Groupwise Wasserstein Grouped LASSO

In this section we describe the model setup and derive what we call the *Groupwise Wasserstein Grouped LASSO (GWGL)* formulation. We will consider a LAD regression model for continuous responses and an LG model for binary categorical responses. In Section 5.2.3, we present a GWGL formulation for overlapping groups.

5.2.1 GWGL for Continuous Response Variables

We assume that the predictors belong to L prescribed groups with group size p_l, $l \in [\![L]\!]$, i.e., $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L)$, where $\mathbf{x}^l \in \mathbb{R}^{p_l}$ and $\sum_{l=1}^{L} p_l = p$ (no overlap among groups). The regression coefficient is $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^L)$, where $\boldsymbol{\beta}^l \in \mathbb{R}^{p_l}$ denotes the regression coefficient for group l. Similar to Section 4, we assume

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

The main assumption we make regarding $\boldsymbol{\beta}^*$ is that it is *group sparse*, i.e., $\boldsymbol{\beta}^l = \mathbf{0}$ for l in some subset of $[\![L]\!]$. Our goal is to obtain an accurate

The version of record is available at: http://dx.doi.org/10.1561/2400000026

estimate of $\boldsymbol{\beta}^*$ under perturbations on the data, when the predictors have a predefined grouping structure. We model stochastic disturbances on the data via distributional uncertainty, and apply a Wasserstein DRO framework to inject robustness into the solution. The learning problem is formulated as:

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

where $\mathbb{Q}$ is the probability distribution of $\mathbf{z} = (\mathbf{x}, y)$, belonging to some set Ω defined as:

$$\Omega = \Omega_{\epsilon}^{s,1}(\hat{\mathbb{P}}_N) \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{s,1}(\mathbb{Q},\ \hat{\mathbb{P}}_N) \leq \epsilon\}, \tag{5.1}$$

and the order-one Wasserstein distance $W_{s,1}(\mathbb{Q},\ \hat{\mathbb{P}}_N)$ is defined on the metric space $(\mathcal{Z}, s)$ associated with the data points $\mathbf{z}$. To reflect the group structure of the predictors and to take into account the group sparsity requirement, we adopt a specific notion of norm to define the metric s. Specifically, for a vector $\mathbf{z}$ with a group structure $\mathbf{z} = (\mathbf{z}^1, \ldots, \mathbf{z}^L)$, define its (q,t)-norm, with $q, t \geq 1$, as:

$$\|\mathbf{z}\|_{q,t} = \left(\sum_{l=1}^{L} (\|\mathbf{z}^l\|_q)^t\right)^{1/t}.$$

Notice that the (q,t)-norm of $\mathbf{z}$ is actually the ℓ_t-norm of the vector $(\|\mathbf{z}^1\|_q, \ldots, \|\mathbf{z}^L\|_q)$, which represents each group vector $\mathbf{z}^l$ in a concise way via the ℓ_q-norm.

Inspired by the LASSO where the ℓ_1-regularizer is used to induce sparsity on the individual level, we wish to deduce an ℓ_1-norm penalty on the group level from (4.5) to induce group sparsity on $\boldsymbol{\beta}^*$. This motivates the use of the $(2,\infty)$-norm on the weighted predictor-response vector

$$\mathbf{z}_{\mathbf{w}} \triangleq \left(\frac{1}{\sqrt{p_1}}\mathbf{x}^1, \ldots, \frac{1}{\sqrt{p_L}}\mathbf{x}^L, My\right),$$

where the weight vector is

$$\mathbf{w} = \left(\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}}, M\right),$$

and M is a positive weight assigned to the response. Specifically,

$$\|\mathbf{z}_{\mathbf{w}}\|_{2,\infty} = \max\left\{\frac{1}{\sqrt{p_1}}\|\mathbf{x}^1\|_2, \ldots, \frac{1}{\sqrt{p_L}}\|\mathbf{x}^L\|_2, M|y|\right\}. \tag{5.2}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

In (5.2) we normalize each group by the number of predictors, to prevent large groups from having a large impact on the distance metric. The $\|\cdot\|_{2,\infty}$ operator computes the maximum of the ℓ_2 norms of the (weighted) grouped predictors and the response. It essentially selects the most influential group when determining the closeness between two points in the predictor-response space, which is consistent with our group sparsity assumption in that not all groups of predictors contribute to the determination of y, and thus a metric that ignores the unimportant groups (e.g., $\|\cdot\|_{2,\infty}$) is desired.

Based on (4.5), in order to obtain the GWGL formulation, we need to derive the dual norm of $\|\cdot\|_{2,\infty}$. A general result that applies to any (q,t)-norm is presented in the following theorem. The dual norm of the $(2,\infty)$-norm is a direct application of Theorem 5.2.1.

Theorem 5.2.1. Consider a vector $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L)$, where each $\mathbf{x}^l \in \mathbb{R}^{p_l}$, and $\sum_l p_l = p$. Define the weighted (r,s)-norm of $\mathbf{x}$ with the weight vector $\mathbf{w} = (w_1, \ldots, w_L)$ to be:

$$\|\mathbf{x}_\mathbf{w}\|_{r,s} = \left(\sum_{l=1}^{L} (\|w_l \mathbf{x}^l\|_r)^s\right)^{1/s},$$

where $\mathbf{x}_\mathbf{w} = (w_1\mathbf{x}^1, \ldots, w_L\mathbf{x}^L)$, $w_l > 0, \forall l$, and $r, s \geq 1$. Then, the dual norm of the weighted (r,s)-norm with weight $\mathbf{w}$ is the (q,t)-norm with weight $\mathbf{w}^{-1}$, where $1/r + 1/q = 1$, $1/s + 1/t = 1$, and $\mathbf{w}^{-1} = (1/w_1, \ldots, 1/w_L)$.

Proof. The dual norm of $\|\cdot\|_{r,s}$ evaluated at some vector $\boldsymbol{\beta}$ is the optimal value of Problem (5.3):

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}'\boldsymbol{\beta} \\ \text{s.t.} \quad & \|\mathbf{x}_\mathbf{w}\|_{r,s} \leq 1. \end{aligned} \tag{5.3}$$

We assume that $\boldsymbol{\beta}$ has the same group structure with $\mathbf{x}$, i.e., $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^L)$. Using Hölder's inequality, we can write

$$\begin{aligned} \mathbf{x}'\boldsymbol{\beta} &= \sum_{l=1}^{L} (w_l \mathbf{x}^l)' \left(\frac{1}{w_l}\boldsymbol{\beta}^l\right) \\ &\leq \sum_{l=1}^{L} \|w_l \mathbf{x}^l\|_r \left\|\frac{1}{w_l}\boldsymbol{\beta}^l\right\|_q. \end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Define two new vectors in $\mathbb{R}^L$

$$\mathbf{x}_{\text{new}} = (\|w_1\mathbf{x}^1\|_r, \ldots, \|w_L\mathbf{x}^L\|_r),$$
$$\boldsymbol{\beta}_{\text{new}} = \left(\left\|\frac{1}{w_1}\boldsymbol{\beta}^1\right\|_q, \ldots, \left\|\frac{1}{w_L}\boldsymbol{\beta}^L\right\|_q\right).$$

Applying Hölder's inequality again to $\mathbf{x}_{\text{new}}$ and $\boldsymbol{\beta}_{\text{new}}$, we obtain:

$$\begin{aligned}\mathbf{x}'\boldsymbol{\beta} &\leq \mathbf{x}'_{\text{new}}\boldsymbol{\beta}_{\text{new}} \\ &\leq \|\mathbf{x}_{\text{new}}\|_s\|\boldsymbol{\beta}_{\text{new}}\|_t \\ &= \left(\sum_{l=1}^{L}(\|w_l\mathbf{x}^l\|_r)^s\right)^{1/s}\left(\sum_{l=1}^{L}\left(\left\|\frac{1}{w_l}\boldsymbol{\beta}^l\right\|_q\right)^t\right)^{1/t}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{x}'\boldsymbol{\beta} &\leq \|\mathbf{x}_{\mathbf{w}}\|_{r,s}\|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{q,t} \\ &\leq \|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{q,t},\end{aligned}$$

due to the constraint $\|\mathbf{x}_{\mathbf{w}}\|_{r,s} \leq 1$. The result then follows. □

Now, let us go back to (5.2), which is the weighted $(2,\infty)$-norm of $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^L, y)$ with the weight $\mathbf{w} = (\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}}, M)$. According to Theorem 5.2.1, the dual norm of the weighted $(2,\infty)$-norm with weight $\mathbf{w}$ evaluated at some $\tilde{\boldsymbol{\beta}} = (-\boldsymbol{\beta}^1, \ldots, -\boldsymbol{\beta}^L, 1)$ is:

$$\|\tilde{\boldsymbol{\beta}}_{\mathbf{w}^{-1}}\|_{2,1} = \sum_{l=1}^{L}\sqrt{p_l}\|\boldsymbol{\beta}^l\|_2 + \frac{1}{M},$$

where $\mathbf{w}^{-1} = (\sqrt{p_1}, \ldots, \sqrt{p_L}, 1/M)$. Therefore, with N i.i.d. samples $(\mathbf{x}_i, y_i)$, $i \in [\![N]\!]$, the GWGL formulation for Linear Regression (GWGL-LR) takes the following form:

$$\inf_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N}|y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \epsilon\sum_{l=1}^{L}\sqrt{p_l}\|\boldsymbol{\beta}^l\|_2, \tag{5.4}$$

where the constant term $1/M$ has been removed. We see that by using the weighted $(2,\infty)$-norm in the predictor-response space, we are able to recover the commonly used penalty term for GLASSO [111], [112]. The Wasserstein DRO framework offers new interpretations for the GLASSO penalty from the standpoint of the distance metric on the predictor-response space and establishes the connection between group sparsity and distributional robustness.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

5.2.2 GWGL for Binary Response Variables

In this subsection we will explore the GWGL formulation for binary classification problems. Let $\mathbf{x} \in \mathbb{R}^p$ denote the predictor and $y \in \{-1, +1\}$ the associated binary response/label to be predicted. In LG, the conditional distribution of y given $\mathbf{x}$ is modeled as

$$\mathbb{P}(y|\mathbf{x}) = (1 + \exp(-y\boldsymbol{\beta}'\mathbf{x}))^{-1},$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown coefficient vector (classifier) to be estimated. The *Maximum Likelihood Estimator (MLE)* of $\boldsymbol{\beta}$ is found by minimizing the *negative log-likelihood (logloss)*:

$$h_\beta(\mathbf{x}, y) = \log(1 + \exp(-y\boldsymbol{\beta}'\mathbf{x})).$$

To apply the Wasserstein DRO framework, we define the distance metric on the predictor-response space as follows.

$$s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)) \triangleq \|\mathbf{x}_1 - \mathbf{x}_2\| + M|y_1 - y_2|, \quad \forall (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \in \mathcal{Z}, \tag{5.5}$$

where M is an infinitely large positive number (different from Section 5.2.1 where M could be any positive number), and $\mathcal{Z} = \mathbb{R}^p \times \{-1, +1\}$. We use a very large weight on y to emphasize its role in determining the distance between data points, i.e., for a pair $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_j, y_j)$, if $y_i \neq y_j$, they are considered to be infinitely far away from each other; otherwise their distance is determined solely by the predictors. The robust LG problem is modeled as:

$$\inf_{\beta} \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}}[\log(1 + \exp(-y\boldsymbol{\beta}'\mathbf{x}))], \tag{5.6}$$

where Ω is defined in (5.1) with s specified in (5.5). Based on the discussion in Section 3.1, in order to derive a tractable reformulation for (5.6), we need to bound the growth rate of $h_\beta(\mathbf{x}, y)$:

$$\frac{|h_\beta(\mathbf{x}_1, y_1) - h_\beta(\mathbf{x}_2, y_2)|}{s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2))}, \quad \forall (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2).$$

To this end, we define a continuous and differentiable univariate function $l(a) \triangleq \log(1 + \exp(-a))$, and apply the mean value theorem to it, which

The version of record is available at: http://dx.doi.org/10.1561/2400000026

yields that for any $a, b \in \mathbb{R}$, $\exists c \in (a, b)$ such that:

$$\left|\frac{l(b) - l(a)}{b - a}\right| = |\nabla l(c)| = \frac{e^{-c}}{1 + e^{-c}} \leq 1.$$

By noting that $h_{\boldsymbol{\beta}}(\mathbf{x}, y) = l(y\boldsymbol{\beta}'\mathbf{x})$, we immediately have:

$$\begin{aligned} |h_{\boldsymbol{\beta}}(\mathbf{x}_1, y_1) - h_{\boldsymbol{\beta}}(\mathbf{x}_2, y_2)| &\leq |y_1\boldsymbol{\beta}'\mathbf{x}_1 - y_2\boldsymbol{\beta}'\mathbf{x}_2| \\ &\leq \|y_1\mathbf{x}_1 - y_2\mathbf{x}_2\|\|\boldsymbol{\beta}\|_* \\ &\leq s((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2))\|\boldsymbol{\beta}\|_*, \end{aligned} \tag{5.7}$$

where the second step uses Hölder's inequality, and the last step is due to the definition of the metric s and the fact that M is infinitely large. Equation (5.7) shows that the loss function $h_{\boldsymbol{\beta}}(\mathbf{x}, y)$ is Lipschitz continuous in $(\mathbf{x}, y)$ with a Lipschitz constant $\|\boldsymbol{\beta}\|_*$. Using Theorem 3.1.1 with $t = 1$, we obtain that for any $\mathbb{Q} \in \Omega$,

$$|\mathbb{E}^{\mathbb{Q}}[h_{\boldsymbol{\beta}}(\mathbf{x}, y)] - \mathbb{E}^{\hat{\mathbb{P}}_N}[h_{\boldsymbol{\beta}}(\mathbf{x}, y)]| \leq \|\boldsymbol{\beta}\|_* W_{s,1}(\mathbb{Q},\ \hat{\mathbb{P}}_N) \leq \epsilon\|\boldsymbol{\beta}\|_*.$$

Therefore, Problem (5.6) can be reformulated as:

$$\inf_{\boldsymbol{\beta}} \mathbb{E}^{\hat{\mathbb{P}}_N}[h_{\boldsymbol{\beta}}(\mathbf{x}, y)] + \epsilon\|\boldsymbol{\beta}\|_* = \inf_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} \log(1 + \exp(-y_i\boldsymbol{\beta}'\mathbf{x}_i)) + \epsilon\|\boldsymbol{\beta}\|_*. \tag{5.8}$$

We note that [46], [47], [55] arrive at a similar formulation to (5.8) by other means of derivation. Different from these existing works, we will consider specifically the application of (5.8) to grouped predictors where the goal is to induce group level sparsity on the coefficients/classifier. As in Section 5.2.1, we assume that the predictor vector $\mathbf{x}$ can be decomposed into L groups, i.e., $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L)$, each $\mathbf{x}^l$ containing p_l predictors of group l, and $\sum_{l=1}^{L} p_l = p$. To reflect the group sparse structure, we adopt the $(2, \infty)$-norm of the weighted predictor vector

$$\mathbf{x}_{\mathbf{w}} \triangleq \left(\frac{1}{\sqrt{p_1}}\mathbf{x}^1, \ldots, \frac{1}{\sqrt{p_L}}\mathbf{x}^L\right),$$

to define the metric s in (5.5), where the weight vector is:

$$\mathbf{w} = \left(\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_L}}\right).$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

According to Theorem 5.2.1, the dual norm of the weighted $(2,\infty)$-norm with weight $\mathbf{w} = (1/\sqrt{p_1}, \ldots, 1/\sqrt{p_L})$ evaluated at $\boldsymbol{\beta}$ is:

$$\|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{2,1} = \sum_{l=1}^{L} \sqrt{p_l} \|\boldsymbol{\beta}^l\|_2,$$

where $\mathbf{w}^{-1} = (\sqrt{p_1}, \ldots, \sqrt{p_L})$, and $\boldsymbol{\beta}^l$ denotes the vector of coefficients corresponding to group l. Therefore, the GWGL formulation for LG (GWGL-LG) takes the form:

$$\inf_{\beta} \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-y_i \boldsymbol{\beta}' \mathbf{x}_i)) + \epsilon \sum_{l=1}^{L} \sqrt{p_l} \|\boldsymbol{\beta}^l\|_2. \tag{5.9}$$

The above derivation techniques also apply to other loss functions whose growth rate is finite, e.g., the hinge loss used by SVM, and therefore, the GWGL SVM model can be developed in a similar fashion.

5.2.3 GLASSO with Overlapping Groups

In this subsection we will explore the GLASSO formulation with overlapping groups, and show that the Wasserstein DRO framework recovers a latent GLASSO approach that is proposed by [124] to induce a solution with support being the union of predefined overlapping groups of variables.

When the groups overlap with each other, the penalty term used by (5.4) and (5.9) leads to a solution whose support is almost surely the complement of a union of groups, see [125]. That is to say, setting one group to zero shrinks its covariates to zero even if they belong to other groups, in which case these other groups will not be entirely selected. [124] proposed a latent GLASSO approach where they introduce a set of latent variables that induce a solution vector whose support is a union of groups, so that the estimator would select entire groups of covariates. Specifically, define the latent variables $\mathbf{v}^l \in \mathbb{R}^p$ such that $\text{supp}(\mathbf{v}^l) \subset g^l, l \in [\![L]\!]$, where $\text{supp}(\mathbf{v}^l) \subset [\![p]\!]$ denotes the support of $\mathbf{v}^l$, i.e., the set of predictors $i \in [\![p]\!]$ such that $v_i^l \neq 0$, and g^l denotes the set of predictors that are in group l. Our assumption is that $\exists\ l_1, l_2$ such that $g^{l_1} \cap g^{l_2} \neq \emptyset$. The latent GLASSO formulation is in the following

The version of record is available at: http://dx.doi.org/10.1561/2400000026

form:

$$\begin{aligned} &\inf_{\boldsymbol{\beta},\mathbf{v}^1,\ldots,\mathbf{v}^L} \frac{1}{N}\sum_{i=1}^{N} h_{\boldsymbol{\beta}}(\mathbf{x}_i, y_i) + \epsilon \sum_{l=1}^{L} d_l \|\mathbf{v}^l\|_2, \\ &\text{s.t. } \boldsymbol{\beta} = \sum_{l=1}^{L} \mathbf{v}^l, \end{aligned} \tag{5.10}$$

where d_l is a user-specified penalty strength of group l. Notice that (5.4) and (5.9) are special cases of (5.10) where they require the latent vectors to have the same value at the intersecting covariates. By using the latent vectors $\mathbf{v}^l$, Formulation (5.10) has the flexibility of implicitly adjusting the support of the latent vectors such that for any $i \in \text{supp}(\hat{\mathbf{v}}^l)$ where $\hat{\mathbf{v}}^l = \mathbf{0}$, it does not belong to the support of any non-shrunk latent vectors, i.e., $i \notin \text{supp}(\hat{\mathbf{v}}^k)$ where $\hat{\mathbf{v}}^k \neq \mathbf{0}$. As a result, the covariates that belong to both shrunk and non-shrunk groups would not be mistakenly driven to zero. Formulation (5.10) favors solutions which shrink some $\mathbf{v}^l$ to zero, while the non-shrunk components satisfy $\text{supp}(\mathbf{v}^l) = g^l$, therefore leading to estimators whose support is the union of a set of groups.

To show that (5.10) can be obtained from the Wasserstein DRO framework, we consider the following weighted $(2, \infty)$-norm on the predictor space:

$$s(\mathbf{x}) = \max_l d_l^{-1} \|\mathbf{x}^l\|_2. \tag{5.11}$$

For simplicity we treat the response y as a deterministic quantity so that the Wasserstein metric is defined only on the predictor space. The scenario with stochastic responses can be treated in a similar fashion as in Sections 5.2.1 and 5.2.2 by introducing some constant. [124] showed that the dual norm of (5.11) is:

$$\Omega(\boldsymbol{\beta}) \triangleq \sum_{l=1}^{L} d_l \|\mathbf{v}^l\|_2,$$

with $\boldsymbol{\beta} = \sum_{l=1}^{L} \mathbf{v}^l$, and $\boldsymbol{\beta} \to \Omega(\boldsymbol{\beta})$ is a valid norm. By noting that (5.10) can be reformulated as:

$$\inf_{\boldsymbol{\beta}} \quad \frac{1}{N}\sum_{i=1}^{N} l_{\boldsymbol{\beta}}(\mathbf{x}_i, y_i) + \epsilon\Omega(\boldsymbol{\beta}), \tag{5.12}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

with

$$\Omega(\boldsymbol{\beta}) = \min_{\substack{\mathbf{v}^1,\ldots,\mathbf{v}^L, \\ \sum_{l=1}^L \mathbf{v}^l = \boldsymbol{\beta}}} \sum_{l=1}^{L} d_l \|\mathbf{v}^l\|_2,$$

we have shown that (5.10) can be derived as a consequence of the Wasserstein DRO formulation with the Wasserstein metric induced by (5.11). In fact, [124] pointed out that (5.12) is equivalent to a regular GLASSO in a covariate space of higher dimension obtained by duplication of the covariates belonging to several groups. For simplicity our subsequent analysis assumes non-overlapping groups.

5.3 Performance Guarantees to the DRO Groupwise Estimator

In this section we establish several performance guarantees for the solutions to GWGL-LR and GWGL-LG. We are interested in two types of performance metrics.

(1) *Prediction quality*, which measures the predictive power of the GWGL solutions on new, unseen samples.

(2) *Grouping effect*, which measures the similarity of the estimated coefficients in the same group as a function of the sample correlation between their corresponding predictors. Ideally, for highly correlated predictors in the same group, it is desired that their coefficients are close so that they can be jointly selected/dropped (group sparsity).

We note that GWGL-LR is a special case of the general Wasserstein DRO formulation (4.5), and thus the two types of performance guarantees derived in Section 4.3, one for generalization ability (Theorem 4.3.3), and the other for the estimation accuracy (Theorem 4.3.12), still apply to the GWGL-LR formulation. For GWGL-LG, we will derive its prediction performance result using similar techniques.

5.3.1 Performance Guarantees for GWGL-LR

The prediction and estimation performance of the GWGL-LR model can be described by Theorems 4.3.3 and 4.3.12, where the Wasserstein

The version of record is available at: http://dx.doi.org/10.1561/2400000026

metric is defined using the weighted $(2,\infty)$-norm with weight $\mathbf{w} = (1/\sqrt{p_1},\ldots,1/\sqrt{p_L},M)$. We thus omit the statement of these two results. With Theorem 4.3.12, we are able to provide bounds for the *Relative Risk (RR)*, *Relative Test Error (RTE)*, and *Proportion of Variance Explained (PVE)* that are introduced in Section 4.4. All these metrics evaluate the accuracy of the regression coefficient estimates on a new test sample drawn from the same probability distribution as the training samples.

Using Theorem 4.3.12, we can bound the term $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$ as follows:

$$\begin{aligned}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) &\leq \lambda_{\max}(\boldsymbol{\Sigma})\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2^2 \\ &\leq \lambda_{\max}(\boldsymbol{\Sigma})\left(\frac{4R^2\bar{B}}{\lambda_{\min}}\Psi(\boldsymbol{\beta}^*)\right)^2, \qquad (5.13)\end{aligned}$$

where $\lambda_{\max}(\boldsymbol{\Sigma})$ is the maximum eigenvalue of $\boldsymbol{\Sigma}$. Using (5.13), bounds for RR, RTE, and PVE can be readily obtained and are summarized in the following corollary.

Corollary 5.3.1. Under the specifications in Theorem 4.3.12, when the sample size

$$N \geq \bar{C}_1\bar{\mu}^4\mu_0^2\frac{\lambda_{\max}}{\lambda_{\min}}(w(\mathcal{A}(\boldsymbol{\beta}^*))+3)^2,$$

with probability at least $1-\exp(-C_2N/\bar{\mu}^4)$,

$$\mathrm{RR}(\hat{\boldsymbol{\beta}}) \leq \frac{\lambda_{\max}(\boldsymbol{\Sigma})\left(\frac{4R^2\bar{B}}{\lambda_{\min}}\Psi(\boldsymbol{\beta}^*)\right)^2}{(\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^*},$$

$$\mathrm{RTE}(\hat{\boldsymbol{\beta}}) \leq \frac{\lambda_{\max}(\boldsymbol{\Sigma})\left(\frac{4R^2\bar{B}}{\lambda_{\min}}\Psi(\boldsymbol{\beta}^*)\right)^2+\sigma^2}{\sigma^2},$$

and,

$$\mathrm{PVE}(\hat{\boldsymbol{\beta}}) \geq 1-\frac{\lambda_{\max}(\boldsymbol{\Sigma})\left(\frac{4R^2\bar{B}}{\lambda_{\min}}\Psi(\boldsymbol{\beta}^*)\right)^2+\sigma^2}{(\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^*+\sigma^2},$$

where all parameters are defined in the same way as in Theorem 4.3.12.

We next proceed to investigate the grouping effect of the GWGL-LR estimator. The next theorem provides a bound on the absolute (weighted) difference between coefficient estimates as a function of the sample correlation between their corresponding predictors.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Theorem 5.3.2. Suppose the predictors are standardized (columns of $\mathbf{X}$ have zero mean and unit variance). Let $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ be the optimal solution to (5.4). If $\mathbf{x}_{,i}$ is in group l_1 and $\mathbf{x}_{,j}$ is in group l_2, and $\|\hat{\boldsymbol{\beta}}^{l_1}\|_2 \neq 0$, $\|\hat{\boldsymbol{\beta}}^{l_2}\|_2 \neq 0$, define:

$$D(i,j) = \left| \frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2} \right|.$$

Then,

$$D(i,j) \leq \frac{\sqrt{2(1-\rho)}}{\sqrt{N}\epsilon},$$

where $\rho = \mathbf{x}'_{,i}\mathbf{x}_{,j}$ is the sample correlation, and p_{l_1}, p_{l_2} are the number of predictors in groups l_1 and l_2, respectively.

Proof. By the optimality condition associated with formulation (5.4), $\hat{\boldsymbol{\beta}}$ satisfies:

$$\mathbf{x}'_{,i}\mathrm{sgn}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = N\epsilon\sqrt{p_{l_1}}\frac{\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2}, \tag{5.14}$$

$$\mathbf{x}'_{,j}\mathrm{sgn}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = N\epsilon\sqrt{p_{l_2}}\frac{\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2}, \tag{5.15}$$

where the $\mathrm{sgn}(\cdot)$ function is applied to a vector elementwise. Subtracting (5.15) from (5.14), we obtain:

$$(\mathbf{x}_{,i} - \mathbf{x}_{,j})'\mathrm{sgn}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = N\epsilon\left(\frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2}\right).$$

Using the Cauchy–Schwarz inequality and $\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2^2 = 2(1-\rho)$, we obtain

$$\begin{aligned} D(i,j) &= \left| \frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2} \right| \\ &\leq \frac{1}{N\epsilon}\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2\|\mathrm{sgn}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\|_2 \\ &\leq \frac{\sqrt{2(1-\rho)}}{\sqrt{N}\epsilon}. \end{aligned}$$

□

The version of record is available at: http://dx.doi.org/10.1561/2400000026

When $\mathbf{x}_{,i}$ and $\mathbf{x}_{,j}$ are in the same group l and $\|\hat{\boldsymbol{\beta}}^l\|_2 \neq 0$, Theorem 5.3.2 yields

$$|\hat{\beta}_i - \hat{\beta}_j| \leq \frac{\sqrt{2(1-\rho)}\|\hat{\boldsymbol{\beta}}^l\|_2}{\epsilon\sqrt{Np_l}}. \tag{5.16}$$

From (5.16) we see that as the within group correlation ρ increases, the difference between $\hat{\beta}_i$ and $\hat{\beta}_j$ becomes smaller. In the extreme case where $\mathbf{x}_{,i}$ and $\mathbf{x}_{,j}$ are perfectly correlated, i.e., $\rho = 1$, $\hat{\beta}_i = \hat{\beta}_j$. This grouping effect enables recovery of sparsity on a group level when the correlation between predictors in the same group is high, and implies the use of predictors' correlation as a grouping criterion. One of the popular clustering algorithms, called *spectral clustering* [126]–[129], performs grouping based on the eigenvalues/eigenvectors of the Laplacian matrix of the similarity graph that is constructed using the *similarity matrix* of data (predictors). The similarity matrix measures the pairwise similarities between data points, which in our case could be the pairwise correlations between predictors.

5.3.2 Performance Guarantees for GWGL-LG

In this subsection we establish bounds on the prediction error of the GWGL-LG solution, and explore its grouping effect. We will use the *Rademacher complexity* of the class of logloss (negative log-likelihood) functions to bound the generalization error. Suppose $(\mathbf{x}, y)$ is drawn from the probability measure $\mathbb{P}^*$. Two assumptions that impose conditions on the magnitude of the regularizer and the uncertainty level of the predictor are needed.

Assumption O. The weighted $(2, \infty)$-norm of $\mathbf{x}$ is bounded above, i.e., $\|\mathbf{x}_{\mathbf{w}}\|_{2,\infty} \leq R_{\mathbf{x}}$ a.s. under $\mathbb{P}^*_{\mathcal{X}}$,, where $\mathbf{w} = (1/\sqrt{p_1}, \ldots, 1/\sqrt{p_L})$.

Assumption P. The weighted $(2,1)$-norm of $\boldsymbol{\beta}$ with weight $\mathbf{w}^{-1} = (\sqrt{p_1}, \ldots, \sqrt{p_L})$ is bounded above, namely, $\sup_{\boldsymbol{\beta}} \|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{2,1} = \bar{B}_1$.

Under these two assumptions, the logloss could be bounded via the definition of dual norm.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Lemma 5.3.3. Under Assumptions O and P, it follows that under the probability measure $\mathbb{P}^*$,

$$\log(1+\exp(-y\boldsymbol{\beta}'\mathbf{x})) \leq \log(1+\exp(R_{\mathbf{x}}\bar{B}_1)), \quad \text{a.s.}$$

Now consider the following class of loss functions:

$$\mathcal{H} = \{(\mathbf{x},y) \to h_{\boldsymbol{\beta}}(\mathbf{x},y)\colon h_{\boldsymbol{\beta}}(\mathbf{x},y) = \log(1+\exp(-y\boldsymbol{\beta}'\mathbf{x})), \\ \forall \boldsymbol{\beta} \text{ s.t. } \|\boldsymbol{\beta}_{\mathbf{w}^{-1}}\|_{2,1} \leq \bar{B}_1\}.$$

It follows from Lemma 4.3.2 that the empirical *Rademacher complexity* of $\mathcal{H}$, denoted by $\mathcal{R}_N(\mathcal{H})$, can be upper bounded by:

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{2\log(1+\exp(R_{\mathbf{x}}\bar{B}_1))}{\sqrt{N}}.$$

Then, applying Theorem 4.3.4 (Theorem 8 in [98]), we have the following result on the prediction error of the GWGL-LG estimator.

Theorem 5.3.4. Let $\hat{\boldsymbol{\beta}}$ be an optimal solution to (5.9), obtained using N training samples $(\mathbf{x}_i, y_i)$, $i \in [\![N]\!]$. Suppose we draw a new i.i.d. sample $(\mathbf{x}, y)$. Under Assumptions O and P, for any $0 < \delta < 1$, with probability at least $1-\delta$ with respect to the sampling,

$$\mathbb{E}^{\mathbb{P}^*}[\log(1+\exp(-y\mathbf{x}'\hat{\boldsymbol{\beta}}))] \leq \frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(-y_i\mathbf{x}_i'\hat{\boldsymbol{\beta}})) \\ + \frac{2\log(1+\exp(R_{\mathbf{x}}\bar{B}_1))}{\sqrt{N}} + \log(1+\exp(R_{\mathbf{x}}\bar{B}_1))\sqrt{\frac{8\log(2/\delta)}{N}},$$

and for any $\zeta > \frac{2\log(1+\exp(R_{\mathbf{x}}\bar{B}_1))}{\sqrt{N}} + \log(1+\exp(R_{\mathbf{x}}\bar{B}_1))\sqrt{\frac{8\log(2/\delta)}{N}}$,

$$\mathbb{P}\left(\log(1+\exp(-y\mathbf{x}'\hat{\boldsymbol{\beta}})) \geq \frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(-y_i\mathbf{x}_i'\hat{\boldsymbol{\beta}})) + \zeta\right) \\ \leq \frac{\frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(-y_i\mathbf{x}_i'\hat{\boldsymbol{\beta}})) + \frac{2\log(1+\exp(R_{\mathbf{x}}\bar{B}_1))}{\sqrt{N}}}{\frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(-y_i\mathbf{x}_i'\hat{\boldsymbol{\beta}})) + \zeta} \\ + \frac{\log(1+\exp(R_{\mathbf{x}}\bar{B}_1))\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(-y_i\mathbf{x}_i'\hat{\boldsymbol{\beta}})) + \zeta}.$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

The next result, similar to Theorem 5.3.2, establishes the grouping effect of the GWGL-LG estimator.

Theorem 5.3.5. Suppose the predictors are standardized (columns of $\mathbf{X}$ have zero mean and unit variance). Let $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ be the optimal solution to (5.9). If $\mathbf{x}_{,i}$ is in group l_1 and $\mathbf{x}_{,j}$ is in group l_2, and $\|\hat{\boldsymbol{\beta}}^{l_1}\|_2 \neq 0$, $\|\hat{\boldsymbol{\beta}}^{l_2}\|_2 \neq 0$, define:

$$D(i,j) = \left| \frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2} \right|.$$

Then,

$$D(i,j) \leq \frac{\sqrt{2(1-\rho)}}{\sqrt{N}\epsilon},$$

where $\rho = \mathbf{x}'_{,i}\mathbf{x}_{,j}$ is the sample correlation between predictors i and j, and p_{l_1}, p_{l_2} are the number of predictors in groups l_1 and l_2, respectively.

Proof. By the optimality condition associated with formulation (5.9), $\hat{\boldsymbol{\beta}}$ satisfies:

$$\sum_{k=1}^{N} \frac{\exp(-y_k \mathbf{x}'_k \hat{\boldsymbol{\beta}})}{1+\exp(-y_k \mathbf{x}'_k \hat{\boldsymbol{\beta}})} y_k x_{k,i} = N\epsilon\sqrt{p_{l_1}} \frac{\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2}, \tag{5.17}$$

$$\sum_{k=1}^{N} \frac{\exp(-y_k \mathbf{x}'_k \hat{\boldsymbol{\beta}})}{1+\exp(-y_k \mathbf{x}'_k \hat{\boldsymbol{\beta}})} y_k x_{k,j} = N\epsilon\sqrt{p_{l_2}} \frac{\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2}, \tag{5.18}$$

where $x_{k,i}$ and $x_{k,j}$ denote the i-th and j-th elements of $\mathbf{x}_k$, respectively. Subtracting (5.18) from (5.17), we obtain:

$$\sum_{k=1}^{N} \frac{\exp(-y_k \mathbf{x}'_k \hat{\boldsymbol{\beta}})}{1+\exp(-y_k \mathbf{x}'_k \hat{\boldsymbol{\beta}})} (y_k x_{k,i} - y_k x_{k,j}) = N\epsilon \left(\frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2} \right). \tag{5.19}$$

Note that the LHS of (5.19) can be written as $\mathbf{v}'_1\mathbf{v}_2$, where

$$\mathbf{v}_1 = \left(\frac{\exp(-y_1 \mathbf{x}'_1 \hat{\boldsymbol{\beta}})}{1+\exp(-y_1 \mathbf{x}'_1 \hat{\boldsymbol{\beta}})}, \ldots, \frac{\exp(-y_N \mathbf{x}'_N \hat{\boldsymbol{\beta}})}{1+\exp(-y_N \mathbf{x}'_N \hat{\boldsymbol{\beta}})} \right),$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

and,

$$\mathbf{v}_2 = (y_1(x_{1,i} - x_{1,j}), \ldots, y_N(x_{N,i} - x_{N,j})).$$

Using the Cauchy–Schwarz inequality and $\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2^2 = 2(1-\rho)$, we obtain

$$\begin{aligned} D(i,j) &= \left| \frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\boldsymbol{\beta}}^{l_2}\|_2} \right| \\ &\leq \frac{1}{N\epsilon}\|\mathbf{v}_1\|_2\|\mathbf{v}_2\|_2 \\ &\leq \frac{1}{N\epsilon}\sqrt{N}\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2 \\ &= \frac{\sqrt{2(1-\rho)}}{\sqrt{N}\epsilon}. \end{aligned}$$

□

We see that Theorem 5.3.5 yields the same bound with Theorem 5.3.2, and for predictors in the same group, their coefficients converge to the same value as $O(\sqrt{1-\rho})$. This encourages group level sparsity if predictor correlation is used as a grouping criterion.

5.4 Numerical Experiments

In this section we compare the GWGL formulations with other commonly used predictive models. In the linear regression setting, we compare GWGL-LR with models that either (i) use a different loss function, e.g., the traditional GLASSO with an ℓ_2-loss [112], and the Group Square-Root LASSO (GSRL) [122] that minimizes the square root of the ℓ_2-loss; or (ii) do not make use of the grouping structure of the predictors, e.g., the Elastic Net (EN) [108], and the LASSO [5]. For classification problems, we consider alternatives that minimize the empirical logloss plus penalty terms that do not utilize the grouping structure of the predictors, e.g., the ℓ_1-regularizer (LG-LASSO), ℓ_2-regularizer (LG-Ridge), and their combination (LG-EN). The results on several synthetic datasets and a real large dataset of surgery-related medical records are shown in the subsequent sections.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

5.4.1 GWGL-LR on Synthetic Datasets

In this subsection, we will compare GWGL-LR with the aforementioned models on several synthetic datasets. The data generation process is described as follows:

1. Generate $\boldsymbol{\beta}^*$ based on the following rule:

$$(\boldsymbol{\beta}^*)^l = \begin{cases} 0.5 \cdot \mathbf{e}_{p_l}, & \text{if } l \text{ is even;} \\ \mathbf{0}, & \text{otherwise,} \end{cases}$$

   where $\mathbf{e}_{p_l}$ is the p_l-dimensional vector with all ones.

2. Generate the predictor $\mathbf{x} \in \mathbb{R}^p$ from the Gaussian distribution $\mathcal{N}_p(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\sigma_{i,j})_{i,j=1}^p$ has diagonal elements equal to 1, and off-diagonal elements specified as:

$$\sigma_{i,j} = \begin{cases} \rho_w, & \text{if predictors } i \text{ and } j \text{ are in the same group;} \\ 0, & \text{otherwise.} \end{cases}$$

   Here ρ_w is the correlation between predictors in the same group, which we call *within group correlation*. The correlation between different groups is set to zero.

3. Generate the response y as follows:

$$y \sim \begin{cases} \mathcal{N}(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2), & \text{if } r \leq 1 - q; \\ \mathcal{N}(\mathbf{x}'\boldsymbol{\beta}^*, \sigma^2) + 5\sigma, & \text{otherwise,} \end{cases}$$

   where σ^2 is the intrinsic variance of y, r is a uniform random variable on $[0, 1]$, and q is the probability (proportion) of abnormal samples (outliers).

We generate 10 datasets consisting of $N = 100, M_t = 60$ observations and 4 groups of predictors, where N is the size of the training set and M_t is the size of the test set. The number of predictors in each group is: $p_1 = 1, p_2 = 3, p_3 = 5, p_4 = 7$, and $p = \sum_{i=1}^4 p_i = 16$. We are interested in studying the impact of (i) *Signal to Noise Ratio (SNR)*, and (ii) the correlation among predictors in the same group (*within group correlation*): ρ_w. The performance metrics we use are:

The version of record is available at: http://dx.doi.org/10.1561/2400000026

- *Median Absolute Deviation (MAD)* on the test dataset, which is defined to be the median value of $|y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}|,\ i \in [\![M]\!]$, with $\hat{\boldsymbol{\beta}}$ being the estimate of $\boldsymbol{\beta}^*$ obtained from the training set, and $(\mathbf{x}_i, y_i),\ i \in [\![M]\!]$, being the observations from the test dataset;
- *Relative Risk (RR)* of $\hat{\boldsymbol{\beta}}$;
- *Relative Test Error (RTE)* of $\hat{\boldsymbol{\beta}}$;
- *Proportion of Variance Explained (PVE)* of $\hat{\boldsymbol{\beta}}$.

All the regularization parameters are tuned using a separate validation dataset. As to the range of values for the tuned parameters, we adopt the idea from Section 4.4 and adjust properly for the GLASSO estimators. Specifically,

- For GWGL and GSRL, the range of values for ϵ or λ is:

$$\sqrt{\exp(\mathrm{lin}(\log(0.005 \cdot \|\mathbf{X}'\mathbf{y}\|_\infty), \log(\|\mathbf{X}'\mathbf{y}\|_\infty), 50))} / \max_{l \in [\![L]\!]} p_l,$$

  where $\mathrm{lin}(a, b, n)$ 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. Compared to LASSO [109], the values are scaled by $\max_{l \in [\![L]\!]} p_l$, and the square root operation is due to the ℓ_1-loss function, or the square root of the ℓ_2-loss used in these formulations.

- For the GLASSO with ℓ_2-loss, the range of values for λ is:

$$\exp(\mathrm{lin}(\log(0.005 \cdot \|\mathbf{X}'\mathbf{y}\|_\infty), \log(\|\mathbf{X}'\mathbf{y}\|_\infty), 50)) / \sqrt{\max_{l \in [\![L]\!]} p_l}.$$

We note that before solving for the regression coefficients using various GLASSO formulations, the grouping of predictors needs to be determined. Unlike most of the existing works where the grouping structure is assumed to be known or can be obtained from expert knowledge [112], [122], [130], we propose to use a data-driven clustering algorithm to group the predictors based on their sample correlations, as suggested by Theorem 5.3.2. Specifically, we consider the *spectral clustering* [126]–[128] algorithm with the following Gaussian similarity

The version of record is available at: http://dx.doi.org/10.1561/2400000026

function

$$\mathrm{Gs}(\mathbf{x}_{,i}, \mathbf{x}_{,j}) \triangleq \exp(-\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2^2/(2\sigma_s^2)), \tag{5.20}$$

where σ_s is some scale parameter whose selection will be explained later. Notice that for standardized predictors, (5.20) captures the sample pairwise correlations between predictors, since $\|\mathbf{x}_{,i} - \mathbf{x}_{,j}\|_2^2 = 2(1 - \mathrm{cor}(\mathbf{x}_{,i}, \mathbf{x}_{,j}))$, where $\mathrm{cor}(\mathbf{x}_{,i}, \mathbf{x}_{,j}) \triangleq \mathbf{x}'_{,i}\mathbf{x}_{,j}$. Using (5.20), we can transform the set of predictors into a *similarity graph*, whose Laplacian matrix will be used for spectral clustering. In our implementation, the k-nearest neighbor similarity graph is constructed, where we connect $\mathbf{x}_{,i}$ and $\mathbf{x}_{,j}$ with an undirected edge if $\mathbf{x}_{,i}$ is among the k-nearest neighbors of $\mathbf{x}_{,j}$ (in the sense of Euclidean distance) *or* if $\mathbf{x}_{,j}$ is among the k-nearest neighbors of $\mathbf{x}_{,i}$. The parameter k is chosen such that the resulting graph is connected. The scale parameter σ_s in (5.20) is set to the mean distance of a point to its k-th nearest neighbor [131]. We assume that the number of clusters is known in order to perform spectral clustering, but in case it is unknown, the eigengap heuristic [131] can be used, where the goal is to choose the number of clusters c such that all eigenvalues $\lambda_1, \ldots, \lambda_c$ of the graph Laplacian are very small, but λ_{c+1} is relatively large. The implementation of spectral clustering uses the Matlab package[1] developed according to the tutorial [131].

We next present the experimental results. For a percentage of outliers $q = 20\%, 30\%$, we plot two sets of graphs.

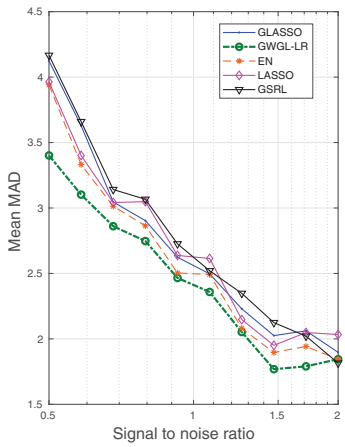
- The performance metrics, i.e., out-of-sample MAD, RR, RTE, and PVE, vs. SNR, where the SNR values are equally spaced between 0.5 and 2 on a log scale. Note that when SNR is varied, the within group correlation between predictors is set to 0.8 times a random noise uniformly distributed on the interval $[0.2, 0.4]$.

- The performance metrics vs. within group correlation ρ_w, where ρ_w takes values in $(0.1, 0.2, \ldots, 0.9)$. When ρ_w is varied, SNR is fixed to 1.

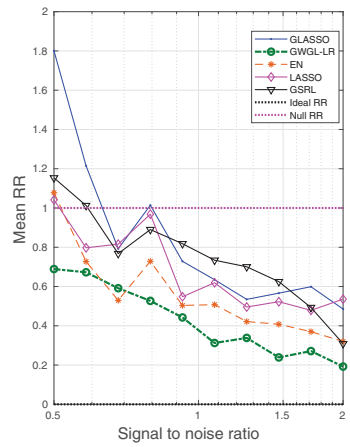
Results for varying the SNR are shown in Figures 5.1 and 5.2. Results for varying the within group correlation are shown in Figures 5.3 and 5.4.

[1] https://www.mathworks.com/matlabcentral/fileexchange/34412-fast-and-efficient-spectral-clustering.

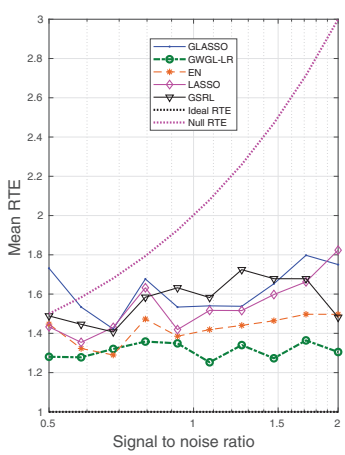
The version of record is available at: http://dx.doi.org/10.1561/2400000026



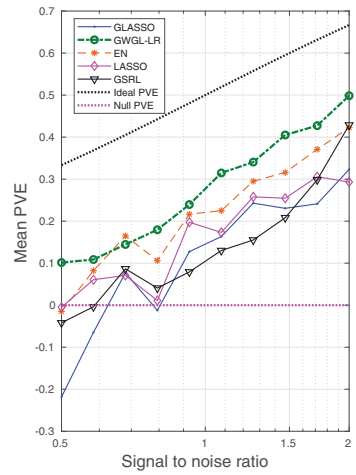
(a) Median Absolute Deviation.



(b) Relative risk.



(c) Relative test error.

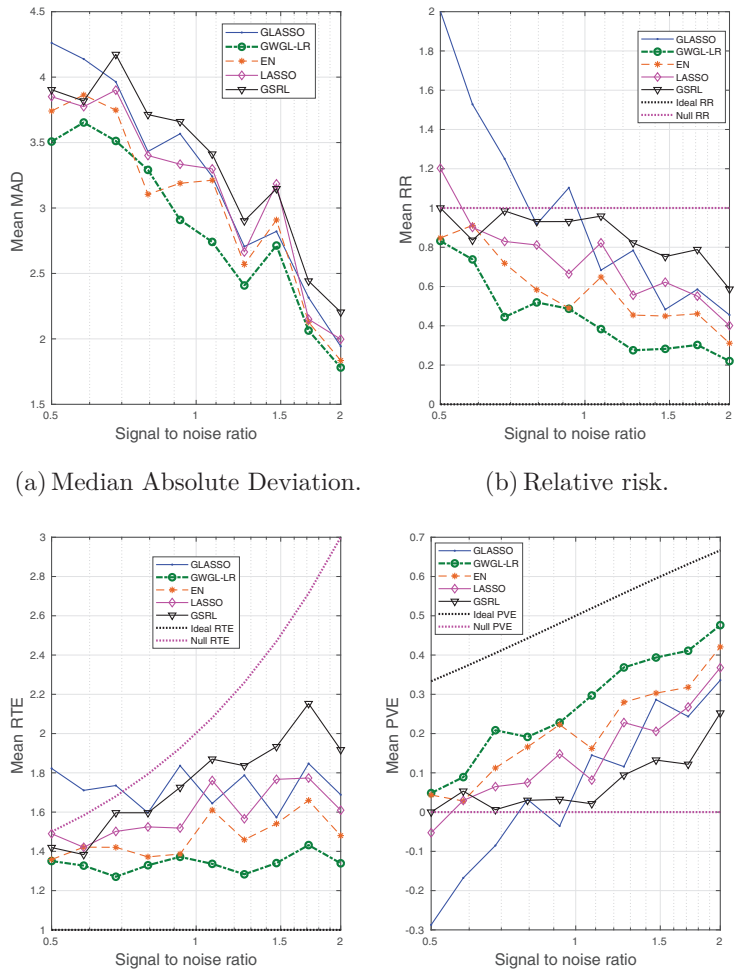


(d) Proportion of variance explained.

Figure 5.1: The impact of SNR on the performance metrics, $q = 20\%$.

To better highlight the benefits of GWGL-LR, in Tables 5.1 and 5.2 we summarize the *Maximum Percentage Improvement (MPI)* brought about by our methods compared to other procedures, when varying the SNR and ρ_w, respectively. In all tables, the number outside the parentheses is the MPI value corresponding to each metric,

The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) Median Absolute Deviation.

(b) Relative risk.

(c) Relative test error.

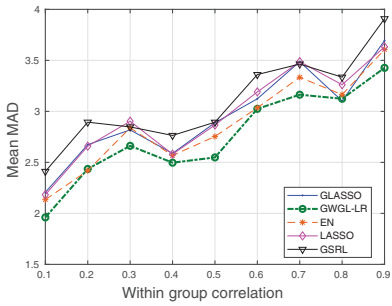
(d) Proportion of variance explained.

Figure 5.2: The impact of SNR on the performance metrics, $q = 30\%$.

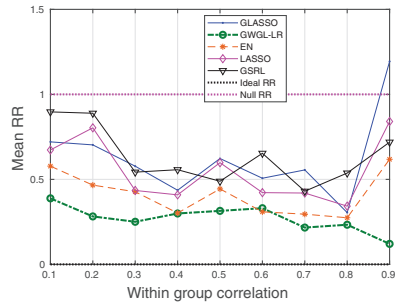
while the number in the parentheses indicates the value of SNR/ρ_w at which the MPI is attained. For each performance metric, the MPI is defined as the maximum percentage difference of the performance between GWGL-LR and the best among all others.

We summarize below our main findings from the results we have presented.

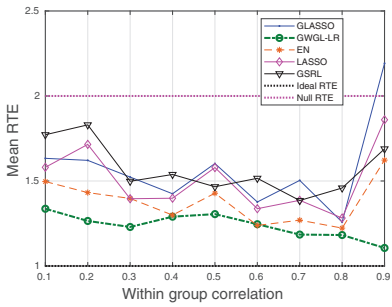
The version of record is available at: http://dx.doi.org/10.1561/2400000026



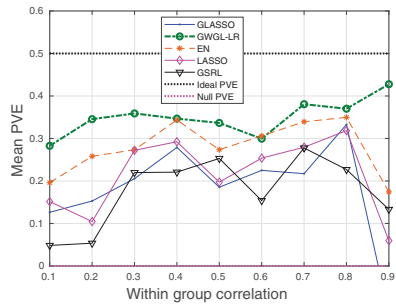
(a) Median Absolute Deviation.



(b) Relative risk.



(c) Relative test error.

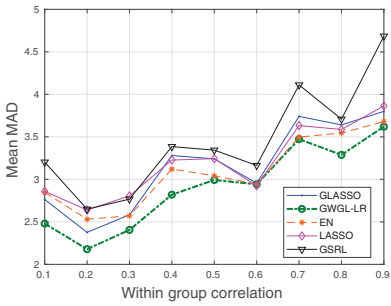


(d) Proportion of variance explained.

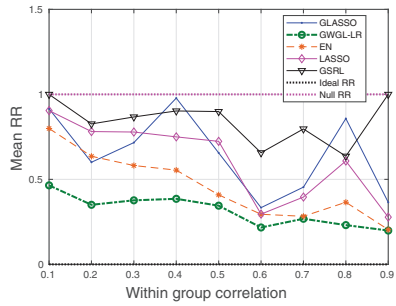
Figure 5.3: The impact of within group correlation on the performance metrics, $q = 20\%$.

- For all approaches under consideration, MAD and RR decrease as the data becomes less noisy. PVE increases when the noise is reduced.
- The GWGL-LR formulation has better prediction and estimation performances than all other approaches under consideration.
- The relative improvement of GWGL-LR over GLASSO (with an ℓ_2-loss) is more significant for highly noisy data (with low SNR values or a high percentage of outliers), which can be attributed to the ℓ_1-loss function it uses. Moreover, GWGL-LR generates more stable estimators than GLASSO.
- When the within group correlation is varied, GWGL-LR shows a more stable performance than others.

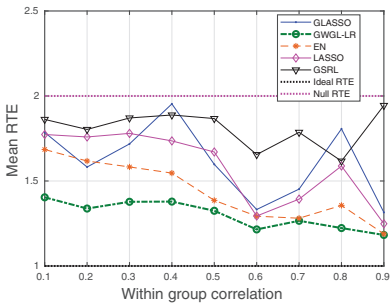
The version of record is available at: http://dx.doi.org/10.1561/2400000026



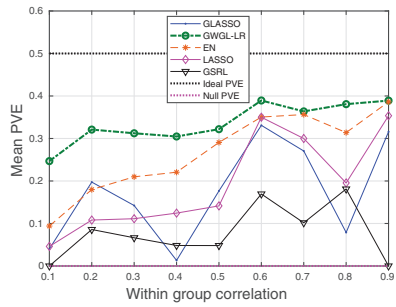
(a) Median Absolute Deviation.



(b) Relative risk.



(c) Relative test error.



(d) Proportion of variance explained.

Figure 5.4: The impact of within group correlation on the performance metrics, $q = 30\%$.

Table 5.1: Maximum percentage improvement of all metrics when varying the SNR

	MAD	**RR**	**RTE**	**PVE**
$q = 20\%$	13.7 (0.5)	41.4 (1.47)	13.1 (1.47)	68.9 (0.79)
$q = 30\%$	14.7 (1.08)	40.9 (1.08)	17 (1.08)	85.7 (0.68)

Table 5.2: Maximum percentage improvement of all metrics when varying the within group correlation

	MAD	**RR**	**RTE**	**PVE**
$q = 20\%$	8.2 (0.1)	80.5 (0.9)	31.8 (0.9)	145.4 (0.9)
$q = 30\%$	10.2 (0.1)	41.9 (0.1)	16.7 (0.1)	162.5 (0.1)

The version of record is available at: http://dx.doi.org/10.1561/2400000026

5.4.2 GWGL-LG on Synthetic Datasets

In this subsection we explore the GWGL-LG formulation on synthetic datasets. The data generation process is described as follows:

1. Generate $\boldsymbol{\beta}^*$ based on the following rule:

$$\beta_k^* = \begin{cases} \mathcal{U}[2.5, 7], & \text{if } \beta_k^* \in (\boldsymbol{\beta}^*)^l \text{ where } l \text{ is even;} \\ 0, & \text{otherwise,} \end{cases}$$

   where $\mathcal{U}[2.5, 7]$ stands for a random variable that is uniformly distributed on the interval $[2.5, 7]$.

2. Generate the predictor $\mathbf{x} \in \mathbb{R}^p$ from the Gaussian distribution $\mathcal{N}_p(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\sigma_{i,j})_{i,j=1}^p$ has diagonal elements equal to 1, and off-diagonal elements specified as:

$$\sigma_{i,j} = \begin{cases} 0.9, & \text{if predictors } i \text{ and } j \text{ are in the same group;} \\ 0, & \text{otherwise.} \end{cases}$$

3. Generate the response y as follows:

$$y \sim \begin{cases} \mathcal{B}([1 + e^{-(\mathbf{x}'\boldsymbol{\beta}^* + \mathcal{N}(0,\sigma^2))}]^{-1}), & \text{if } r \leq 1 - q; \\ \mathcal{B}(0.5), & \text{otherwise,} \end{cases}$$

   where $\mathcal{B}(p)$ stands for the Bernoulli distribution with the probability of success p; $\sigma^2 = (\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}^*$; r is a uniform random variable on $[0, 1]$; and q is the probability (proportion) of abnormal samples (outliers).

We generate 10 datasets consisting of 100 observations and 4 groups of predictors, 80% of which constitute the training dataset, and the remaining forming the test set. The number of predictors in each group is: $p_1 = 3, p_2 = 4, p_3 = 6, p_4 = 7$, and $p = \sum_{i=1}^4 p_i = 20$. The following performance metrics will be used to evaluate the prediction and estimation quality of the solutions.

- The Correct Classification Rate (CCR) on the test dataset, which is defined to be the proportion of test set samples that are correctly classified by the classifier $\hat{\boldsymbol{\beta}}$, with a threshold 0.5 on the predicted probability of success.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

- The *AUC (Area Under the ROC Curve)* on the test dataset.
- The average logloss on the test set.
- The *Within Group Difference (WGD)* of the classifier $\hat{\boldsymbol{\beta}}$, defined as:

$$\mathrm{WGD}(\hat{\boldsymbol{\beta}}) \triangleq \frac{1}{|\{l\colon p_l \geq 2\}|} \sum_{l\colon p_l \geq 2} \frac{1}{\binom{p_l}{2}} \sum_{x_i, x_j \in \mathbf{x}^l} \left| \frac{\hat{\beta}_i - \hat{\beta}_j}{\mathbf{x}'_{\cdot,i}\mathbf{x}_{\cdot,j}} \right|,$$

  where $|\{l\colon p_l \geq 2\}|$ denotes the cardinality of the set $\{l\colon p_l \geq 2\}$, and $\mathbf{x}'_{\cdot,i}\mathbf{x}_{\cdot,j}$ measures the sample correlation between predictors x_i and x_j ($\mathbf{x}_{\cdot,i}$ and $\mathbf{x}_{\cdot,j}$ are standardized). $\mathrm{WGD}(\hat{\boldsymbol{\beta}})$ essentially evaluates the ability of $\hat{\boldsymbol{\beta}}$ to induce group level sparsity. It is desired that the coefficients in the same group are close so that they can be jointly selected/dropped. Theorem 5.3.5 implies that the higher the correlation, the smaller the difference between the coefficients, and thus, a smaller WGD value would suggest a stronger ability of grouped variable selection.

Notice that the first three metrics mentioned above evaluate the *prediction* quality of the classifier, while the last one evaluates its *estimation* quality. If the true coefficient vector $\boldsymbol{\beta}^*$ is known, we will also use the following *confusion matrix* which summarizes the number of zero/nonzero elements in $\hat{\boldsymbol{\beta}}$ that are zero/nonzero in the true coefficient $\boldsymbol{\beta}^*$.

Two ratios will be computed using Table 5.3, the *True Association Rate (TAR)* defined as:

$$\mathrm{TAR} = \frac{\mathrm{TA}}{\mathrm{TA} + \mathrm{FD}},$$

Table 5.3: Confusion matrix

	$\boldsymbol{\beta}^*$	
$\hat{\boldsymbol{\beta}}$	**Nonzero**	**Zero**
Nonzero	True Association (TA)	False Association (FA)
Zero	False Disassociation (FD)	True Disassociation (TD)

The version of record is available at: http://dx.doi.org/10.1561/2400000026

which calculates the proportion of nonzero coefficients that are correctly discovered by the estimator, and the *True Disassociation Rate (TDR)* defined as:

$$\text{TDR} = \frac{\text{TD}}{\text{FA} + \text{TD}},$$

which calculates the proportion of zero coefficients that are correctly identified as zero by the estimator.

We compare GWGL-LG with four formulations: the vanilla logistic regression (LG) that minimizes the empirical logloss on the training samples (without penalty), LG-LASSO that imposes an ℓ_1-norm regularizer on $\boldsymbol{\beta}$, LG-Ridge that uses an ℓ_2-norm regularizer, and LG-EN that uses both the ℓ_1- and ℓ_2-regularizers. All the penalty (regularization) parameters are tuned in the same way as Section 5.4.1. The penalty parameter for GWGL-LG is tuned over 50 values ranging from $\lambda_m = \max_{l \in [\![L]\!]}(\|(\mathbf{X}^l)'(\mathbf{y} - \bar{y}\mathbf{1})\|_2/\sqrt{p_l})$ to a small fraction of λ_m on a log scale [118], where $\mathbf{X}^l$ consists of the columns of the design matrix $\mathbf{X}$ corresponding to group l, $\mathbf{y}$ is the vector of training set labels, and $\bar{y} = \mathbf{y}'\mathbf{1}/N$. The maximum penalty parameter λ_m for LG-LASSO is computed by recognizing it as a special case of GWGL-LG where each group contains only one predictor. For LG-Ridge, λ_m is set to be the square root of the maximum penalty parameter for LG-LASSO, due to the fact that we penalize the square of the ℓ_2-norm regularizer in LG-Ridge. The range of penalty parameters for LG-EN is set in a similar way.

Similar to Section 5.4.1, the spectral clustering algorithm with the Gaussian similarity function (5.20) is used to perform grouping on the predictors. We experiment with two scenarios: (i) $q = 20\%$, and (ii) $q = 30\%$. The results are shown in Tables 5.4 and 5.5, where the number outside the parentheses is the mean value across 10 repetitions, and the number in the parentheses is the corresponding standard deviation.

We see that in general, the penalized formulations perform significantly better than the vanilla logistic regression. LG-EN has very similar prediction performance (i.e., CCR, AUC and logloss on the test set) to GWGL-LG, better than LG-LASSO and LG-Ridge. Regarding the estimation quality, the penalized formulations achieve much lower WGD values than LG. LG-Ridge does not induce sparsity, and therefore has

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Table 5.4: The performances of different classification formulations on synthetic datasets, $q = 20\%$

	CCR	AUC	Logloss	WGD	TAR	TDR
LG	0.62 (0.14)	0.67 (0.13)	0.87 (0.24)	1.71 (0.32)	1.00 (0.00)	0.00 (0.00)
LG-LASSO	0.69 (0.14)	0.77 (0.12)	0.60 (0.13)	0.33 (0.19)	0.54 (0.19)	0.42 (0.23)
LG-Ridge	0.67 (0.12)	0.72 (0.14)	0.69 (0.19)	0.82 (0.42)	1.00 (0.00)	0.01 (0.04)
LG-EN	0.70 (0.15)	0.77 (0.13)	0.59 (0.13)	0.23 (0.08)	0.57 (0.21)	0.42 (0.24)
GWGL-LG	0.68 (0.15)	0.79 (0.12)	0.59 (0.14)	0.13 (0.07)	0.98 (0.04)	0.28 (0.34)

Table 5.5: The performances of different classification formulations on synthetic datasets, $q = 30\%$

	CCR	AUC	Logloss	WGD	TAR	TDR
LG	0.63 (0.09)	0.68 (0.08)	0.99 (0.21)	2.68 (0.51)	1.00 (0.00)	0.00 (0.00)
LG-LASSO	0.73 (0.08)	0.73 (0.08)	0.65 (0.11)	0.56 (0.37)	0.58 (0.21)	0.43 (0.23)
LG-Ridge	0.72 (0.06)	0.74 (0.06)	0.64 (0.10)	0.78 (0.58)	0.98 (0.04)	0.00 (0.00)
LG-EN	0.74 (0.08)	0.77 (0.06)	0.59 (0.08)	0.24 (0.09)	0.48 (0.14)	0.42 (0.23)
GWGL-LG	0.73 (0.08)	0.78 (0.06)	0.60 (0.09)	0.21 (0.16)	0.99 (0.03)	0.18 (0.09)

the highest WGD among the four regularized models. LG-LASSO shows a relatively small WGD, due to the sparsity inducing (at the individual level) property of the ℓ_1-regularizer. GWGL-LG achieves the smallest WGD among all (significantly lower than that of LG-EN), which provides empirical evidence on its group sparsity inducing property, and is consistent with our earlier discussion in Theorem 5.3.5 that the GLASSO penalty tends to drive the coefficients in the same group to the same value if the within group correlation is high. Moreover, GWGL-LG successfully drops out all the coefficients in the first group, while other formulations are not able to drop any of the four groups.

Regarding the TAR and TDR, we notice that GWGL-LG obtains very high TAR values, and compared to other formulations that achieve almost perfect TARs (e.g., LG-Ridge and LG), it has a significantly higher TDR. LG-LASSO and LG-EN achieve the highest TDRs, but their TARs are significantly worse. Note that a dense estimator would result in a perfect TAR but a zero TDR, as in LG and LG-Ridge. The higher the TDR, the more parsimonious the model is, but on the other hand, a higher TAR is more appreciated as we do not want to leave out any of the important (effective) predictors. A low TAR means that a

The version of record is available at: http://dx.doi.org/10.1561/2400000026

substantial proportion of the meaningful predictors are dropped, the cost of which is usually much higher than the cost of wrongly selecting the unimportant ones. Therefore, taking into account both the parsimony and effectiveness of the model, GWGL-LG outperforms all others.

We also want to highlight the robustness of GWGL-LG to misspecified groups. For example, in the scenario with $q = 30\%$ outliers, even though spectral clustering outputs a wrong grouping structure (it divides the data into four groups with group size being $p_1 = 2,\ p_2 = 3,\ p_3 = 5,\ p_4 = 10$, and the correct group size is $p_1 = 3,\ p_2 = 4,\ p_3 = 6,\ p_4 = 7$), GWGL-LG is still able to achieve a satisfactory prediction performance and an almost perfect TAR with a sparse model (nonzero TDR).

5.4.3 An Application to Hospital Readmission

In this section we test the GWGL formulations on a real dataset containing medical records of patients who underwent a general surgical procedure. In 2005, the American College of Surgeons (ACS) established the National Surgical Quality Improvement Program (NSQIP), which collects detailed demographic, laboratory, clinical, procedure and postoperative occurrence data in several surgical subspecialties. The dataset includes (i) baseline demographics; (ii) pre-existing comorbidity information; (iii) preoperative variables; (iv) index admission-related diagnosis and procedure information; (v) postoperative events and complications, and (vi) additional socioeconomic variables.

In our study, patients who underwent a general surgery procedure over 2011–2014 and were tracked by the NSQIP were identified. We will focus on two supervised learning models: (i) a linear regression model whose objective is to predict the post-operative hospital length of stay using pre- and intra-operative variables, and (ii) an LG model whose objective is to predict the re-hospitalization of patients within 30 days after discharge using the same set of explanatory variables. Both models are extremely useful as they allow hospital staff to predict post-operative bed occupancy and prevent costly 30-day readmissions.

Data were pre-processed as follows: (i) categorical variables (such as race, discharge destination, insurance type) were numerically encoded and units homogenized; (ii) missing values were replaced by the mode;

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Table 5.6: The mean and standard deviation of out-of-sample MAD on the surgery data

	Mean	Standard Deviation
GLASSO with ℓ_2-loss	0.17	0.0007
GWGL-LR	0.16	0.001
EN	0.17	0.0009
LASSO	0.17	0.0009
GSRL	0.17	0.0009

(iii) all variables were normalized by subtracting the mean and divided by the standard deviation; (iv) patients who died within 30 days of discharge or had a postoperative length of stay greater than 30 days were excluded. After pre-processing, we were left with a total of 2,275,452 records.

After encoding the categorical predictors using indicator variables, we have 131 numerical predictors for the regression model and 132 for the classification model (the post-operative hospital length of stay is used as a predictor for the 30-day re-hospitalization prediction). The spectral clustering algorithm is used to perform grouping on the predictors, with the number of groups specified as 67 based on a preliminary analysis of the data. (The eigengap heuristic [131] mentioned in Section 5.4.1 was used.)

For predicting the post-operative hospital length of stay, we report the out-of-sample MAD in Table 5.6, i.e., the median of the absolute difference between the predicted and actual length of stay on the test set. The mean and standard deviation of the MAD are computed across 5 repetitions, each with a different training set. We see that the GWGL-LR formulation achieves the lowest mean MAD with a small variation. Compared to the best among others (GLASSO with ℓ_2-loss), it improves the mean MAD by 7.30%. For longer hospital length of stay, this could imply 1 or 2 days improvement in prediction accuracy, which is both clinically and economically meaningful and significant.

For predicting the 30-day re-hospitalization of patients, we notice that the dataset is highly unbalanced, with only 6% of patients being re-hospitalized. To obtain a balanced training set, we randomly draw

The version of record is available at: http://dx.doi.org/10.1561/2400000026

20% patients from the positive class (re-hospitalized patients), and sample the same number of patients from the negative class, resulting in a training set of size 53,616. All the remaining patients go to the test dataset. It turns out that the prediction capabilities of all approaches are very similar. All formulations achieve an average *out-of-sample CCR* around 0.62, an average *out-of-sample AUC* of 0.83, and an average *logloss* on the test set ranging from 0.84 to 0.87. From Table 5.7 we see that GWGL-LG obtains a significantly smaller WGD than others, which implies that the GWGL-LG formulation encourages group level sparsity. This can also be revealed by the number of groups that are dropped by various formulations (see Table 5.8). Notice that though LG-EN and LG-LASSO obtain the most parsimonious models in terms of the number of dropped features (sparsity at an individual level), GWGL-LG has a stronger ability to induce group level sparsity.

Table 5.7: The Within Group Difference (WGD) of the estimators on the surgery data

	Mean	**Standard Deviation**
LG	23.93	1.28
LG-LASSO	16.28	0.72
LG-Ridge	23.38	1.15
LG-EN	16.26	0.74
GWGL-LG	5.04	0.45

Table 5.8: The number of groups/features dropped by various formulations on the surgery data

	Number of Dropped Groups	**Number of Dropped Features**
LG	1	2
LG-LASSO	6	24
LG-Ridge	2	2
LG-EN	10	25
GWGL-LG	16	19

The version of record is available at: http://dx.doi.org/10.1561/2400000026

5.5 Summary

In this section we presented a *Distributionally Robust Optimization (DRO)* formulation under the Wasserstein metric that recovers the GLASSO penalty for *Least Absolute Deviation (LAD)* and LG, through which we have established a connection between group-sparse regularization and robustness and offered new insights into the group sparsity penalty term. We provided insights on the grouping effect of the estimators, which suggests the use of spectral clustering with the Gaussian similarity function to perform grouping on the predictors. We established finite-sample bounds on the prediction errors, which justify the form of the regularizer and provide guidance on the number of training samples needed in order to achieve specific out-of-sample accuracy.

We reported results from several experiments, using both synthetic data and a real dataset with surgery-related medical records. It has been observed that the GWGL formulations (i) achieve more accurate and stable estimates compared to others, especially when the data are noisy, or potentially contaminated with outliers; (ii) have a stronger ability of inducing group-level sparsity, and thus producing more interpretable models, and (iii) successfully identify most of the effective predictors with a reasonably parsimonious model.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

6

Distributionally Robust Multi-Output Learning

In this section, we focus on robust multi-output learning where a multi-dimensional response/label vector is to be learned. The difference from previous sections lies in that we need to estimate a coefficient matrix, rather than a coefficient vector, to explain the dependency of each response variable on the set of predictors. We develop *Distributionally Robust Optimization (DRO)* formulations under the Wasserstein metric for *Multi-output Linear Regression (MLR)* and *Multiclass Logistic Regression (MLG)*, when both the covariates and responses/labels may be contaminated by outliers. Through defining a new notion of matrix norm, we relax the DRO formulation into a regularized learning problem whose regularizer is the norm of the coefficient matrix, establishing a connection between robustness and regularization and generalizing the single-output results presented in Section 4.

6.1 The Problem and Related Work

We consider the multi-output learning problem under the framework of *Distributionally Robust Optimization (DRO)* where the ambiguity set is defined via the Wasserstein metric [13], [14], [46], [47]. The term multi-output learning refers to scenarios where multiple correlated responses

The version of record is available at: http://dx.doi.org/10.1561/2400000026

are to be predicted – *Multi-output Linear Regression (MLR)*, or one of multiple classes is to be assigned – *MultiClass Classification (MCC)*, based on a linear combination of a set of predictors. Both involve learning a target vector $\mathbf{y}$ from a vector of covariates $\mathbf{x}$. MLR has many applications in econometrics [132], health care [133], [134], and finance [135], [136], for modeling multiple measurements of a single individual [137], or evaluating a group of interdependent variables [138]. MCC has seen wide applications in image segmentation [139], text classification [140], and bioinformatics [141].

Unlike a single-output learning problem where the response variable is scalar and a coefficient vector representing the dependency of the response on the predictors is to be learned, in the multi-output setting the decision variable is a coefficient matrix $\mathbf{B} \in \mathbb{R}^{p\times K}$ whose k-th column explains the variation in the k-th coordinate of $\mathbf{y} \in \mathbb{R}^K$ that can be attributed to the predictors $\mathbf{x} \in \mathbb{R}^p$, for $k \in [\![K]\!]$. Inspired by the DRO relaxation derived in Section 4 for the single-output case, which adds a dual norm regularizer to the empirical loss, we obtain a novel **matrix norm** regularizer for the multi-output case through reformulating the Wasserstein DRO problem. The matrix norm exploits the geometrical structure of the coefficient matrix, and provides a way of associating the coefficients for the potentially correlated responses through the dual norm of the distance metric in the data space.

As the simplest MLR model, the multi-output extension of OLS regresses each response variable against the predictors independently, which does not take into account the potential correlation between the responses, and is vulnerable to high correlations existing among the predictors. A class of methods that are used in the literature to overcome this issue is called linear factor regression, where the response $\mathbf{y}$ is regressed against a small number of linearly transformed predictors (factors). Examples include reduced rank regression [142], [143], principal components regression [144], and *Factor Estimation and Selection (FES)* [145]. Another type of methods applies multivariate shrinkage by either estimating a linear transformation of the OLS predictions [138], or solving a regularized MLR problem, e.g., ridge regression [146], [147], and FES [145], whose regularizer is the coefficient matrix's Ky Fan norm defined as the sum of its singular values.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

As for the popular MCC models, [148] provided a thorough survey on the existing MCC techniques which can be categorized into: (i) transformation to binary, e.g., one vs. rest and one vs. one; (ii) extension from binary, e.g., decision trees [1], neural networks [149], K-Nearest Neighbor [150], Naive Bayes classifiers [151], and Support Vector Machine (SVM) [45]; and (iii) hierarchical classification [152].

The research on robust classification has mainly focused on binary classifiers. For example, to robustify logistic regression, [153] proposed to optimize a robustified linear correlation between the response y and a linear function of $\mathbf{x}$; [154] introduced T-logistic regression which replaces the exponential distribution in LG by the t-exponential distribution family; [155] introduced a shift parameter for each data point to account for the label error; and [156] modeled the label error through flipping probabilities, which can be extended to multiclass LG. Another line of research uses a modified loss function that gives less influence to points far from the boundary, e.g., [157] used a tangent loss, and [158] proposed an M-estimator like loss metric which, however, is not robust to outliers with high leverage covariates.

None of the aforementioned works, however, explore distributionally robust learning problems with multiple responses, with the exception of [159], which considered distributionally robust multiclass classification models under the ϕ-divergence metric. We fill this gap by developing DRO formulations for both MLR and MCC under the Wasserstein metric. To the best of our knowledge, we are the first to study the robust multi-output learning problem from the standpoint of distributional robustness. Our approach is completely optimization-based, without the need to explicitly model the complicated relationship between different responses, leading to compact and computationally solvable models. It is interesting that a purely optimization-based method that is completely agnostic to the covariate and response correlation structure can be used as a better-performing alternative to statistical approaches that explicitly model this correlation structure.

The rest of this section is organized as follows. In Section 6.2, we develop the DRO-MLR and DRO-MLG formulations and introduce the matrix norm that is used to define the regularizer. Section 6.3 establishes the out-of-sample performance guarantees for the solutions

The version of record is available at: http://dx.doi.org/10.1561/2400000026

to DRO-MLR and DRO-MLG. The numerical experimental results are presented in Section 6.4. We conclude in Section 6.5.

6.2 Distributionally Robust Multi-Output Learning Models

In this section we introduce the Wasserstein DRO formulations for MLR and MLG, and offer a dual norm interpretation for the regularization terms.

6.2.1 Distributionally Robust Multi-Output Linear Regression

We assume the following model for the MLR problem:

$$\mathbf{y} = \mathbf{B}'\mathbf{x} + \boldsymbol{\eta},$$

where $\mathbf{y} = (y_1, \ldots, y_K)$ is the vector of K responses, potentially correlated with each other; $\mathbf{x} = (x_1, \ldots, x_p)$ is the vector of p predictors; $\mathbf{B} = (B_{ij})_{i\in[\![p]\!]}^{j\in[\![K]\!]}$ is the $p \times K$ matrix of coefficients, the j-th column of which describes the dependency of y_j on the predictors; and $\boldsymbol{\eta}$ is the random error. Suppose we observe N realizations of the data, denoted by $(\mathbf{x}_i, \mathbf{y}_i), i \in [\![N]\!]$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}), \mathbf{y}_i = (y_{i1}, \ldots, y_{iK})$. The Wasserstein DRO formulation for MLR minimizes the following worst-case expected loss:

$$\inf_{\mathbf{B}} \sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h_{\mathbf{B}}(\mathbf{x}, \mathbf{y})], \tag{6.1}$$

where $h_{\mathbf{B}}(\mathbf{x}, \mathbf{y}) \triangleq l(\mathbf{y} - \mathbf{B}'\mathbf{x})$, with l: $\mathbb{R}^K \to \mathbb{R}$ an L-Lipschitz continuous function on the metric spaces $(\mathcal{D}, \|\cdot\|_r)$ and $(\mathcal{C}, |\cdot|)$, where $\mathcal{D}, \mathcal{C}$ are the domain and codomain of $l(\cdot)$, respectively; and $\mathbb{Q}$ is the probability distribution of the data $(\mathbf{x}, \mathbf{y})$, belonging to a set Ω defined as

$$\Omega = \Omega_\epsilon^{s,1}(\hat{\mathbb{P}}_N) \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z})\colon W_{s,1}(\mathbb{Q},\ \hat{\mathbb{P}}_N) \leq \epsilon\},$$

where the order-1 Wasserstein distance $W_{s,1}(\mathbb{Q},\ \hat{\mathbb{P}}_N)$ is induced by the metric $s(\mathbf{z}_1, \mathbf{z}_2) \triangleq \|\mathbf{z}_1 - \mathbf{z}_2\|_r$. Notice that we use the same norm to define the Wasserstein metric and the metric space on the domain $\mathcal{D}$ of $l(\cdot)$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Write the loss function as $h_{\tilde{\mathbf{B}}}(\mathbf{z}) \triangleq l(\tilde{\mathbf{B}}\mathbf{z})$, where $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, and $\tilde{\mathbf{B}} = [-\mathbf{B}', \mathbf{I}_K]$. From Theorem 3.1.1, we know that to derive a tractable reformulation for (6.1), the key is to bound the following *growth rate* of the loss:

$$\mathrm{GR}(h_{\tilde{\mathbf{B}}}) \triangleq \limsup_{\|\mathbf{z}_1 - \mathbf{z}_2\|_r \to \infty} \frac{|h_{\tilde{\mathbf{B}}}(\mathbf{z}_1) - h_{\tilde{\mathbf{B}}}(\mathbf{z}_2)|}{\|\mathbf{z}_1 - \mathbf{z}_2\|_r}.$$

Let us first consider the numerator. By the Lipschitz continuity of $l(\cdot)$, we have:

$$|h_{\tilde{\mathbf{B}}}(\mathbf{z}_1) - h_{\tilde{\mathbf{B}}}(\mathbf{z}_2)| = |l(\tilde{\mathbf{B}}\mathbf{z}_1) - l(\tilde{\mathbf{B}}\mathbf{z}_2)| \leq L\|\tilde{\mathbf{B}}(\mathbf{z}_1 - \mathbf{z}_2)\|_r.$$

The key is to bound $\|\tilde{\mathbf{B}}(\mathbf{z}_1 - \mathbf{z}_2)\|_r$ in terms of $\|\mathbf{z}_1 - \mathbf{z}_2\|_r$. The following lemmata provide three types of bounds whose tightness will be analyzed in the sequel.

Lemma 6.2.1. For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and any vector $\mathbf{x} \in \mathbb{R}^n$, we have:

$$\|\mathbf{A}\mathbf{x}\|_r \leq \|\mathbf{x}\|_r \left(\sum_{i=1}^m \|\mathbf{a}_i\|_1^r \right)^{1/r},$$

for any $r \geq 1$, where $\mathbf{a}_i, i \in [\![m]\!]$, are the rows of $\mathbf{A}$.

Proof. Suppose $\mathbf{A} = (a_{ij})_{i \in [\![m]\!]}^{j \in [\![n]\!]}$. Then,

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_r^r &= \left\| \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{bmatrix} \right\|_r^r \\
&= \sum_{i=1}^m |a_{i1}x_1 + \cdots + a_{in}x_n|^r \\
&\leq \|\mathbf{x}\|_r^r \left(\sum_{i=1}^m (|a_{i1}| + \cdots + |a_{in}|)^r \right),
\end{aligned}$$

where in the second step we use the fact that $|x_i| \leq \|\mathbf{x}\|_r$. □

Lemma 6.2.2. For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and any vector $\mathbf{x} \in \mathbb{R}^n$, we have:

$$\|\mathbf{A}\mathbf{x}\|_r \leq \|\mathbf{x}\|_r \left(\sum_{i=1}^m \|\mathbf{a}_i\|_s^r \right)^{1/r},$$

for any $r \geq 1$, where $\mathbf{a}_i, i \in [\![m]\!]$, are the rows of $\mathbf{A}$, and $1/r + 1/s = 1$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Proof.

$$\begin{aligned}\|\mathbf{A}\mathbf{x}\|_r^r &= \sum_{i=1}^m |\mathbf{a}_i'\mathbf{x}|^r \\ &\leq \sum_{i=1}^m \|\mathbf{x}\|_r^r \|\mathbf{a}_i\|_s^r \\ &= \|\mathbf{x}\|_r^r \sum_{i=1}^m \|\mathbf{a}_i\|_s^r,\end{aligned}$$

where $r, s \geq 1$, $1/r + 1/s = 1$, and the second step uses Hölder's inequality. □

Lemma 6.2.3. Given an $m \times n$ matrix $\mathbf{A} = (a_{ij})_{i\in[\![m]\!]}^{j\in[\![n]\!]}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, the following holds:

$$\|\mathbf{A}\mathbf{x}\|_r \leq \|\mathbf{x}\|_r \|\mathbf{v}\|_s,$$

where $r, s \geq 1$, and $1/r + 1/s = 1$; $\mathbf{v} = (v_1, \ldots, v_n)$, with $v_j = \sum_{i=1}^m |a_{ij}|$.

Proof. Suppose $\mathbf{A} = (a_{ij})_{i\in[\![m]\!]}^{j\in[\![n]\!]}$. Then,

$$\begin{aligned}\|\mathbf{A}\mathbf{x}\|_r &= \left\| \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{bmatrix} \right\|_r \\ &= \left(\sum_{i=1}^m |a_{i1}x_1 + \cdots + a_{in}x_n|^r \right)^{1/r} \\ &\leq \left(\left(\sum_{i=1}^m |a_{i1}x_1 + \cdots + a_{in}x_n| \right)^r \right)^{1/r} \\ &= \sum_{i=1}^m |a_{i1}x_1 + \cdots + a_{in}x_n| \\ &\leq |x_1| \sum_{i=1}^m |a_{i1}| + \cdots + |x_n| \sum_{i=1}^m |a_{in}| \\ &= |\mathbf{x}|'\mathbf{v} \\ &\leq \|\mathbf{x}\|_r \|\mathbf{v}\|_s,\end{aligned}$$

where $r, s \geq 1$, $1/r + 1/s = 1$, $|\mathbf{x}| = (|x_1|, \ldots, |x_n|)$, and $\mathbf{v} = (v_1, \ldots, v_n)$, with $v_j = \sum_{i=1}^m |a_{ij}|$. The last step uses Hölder's inequality and the fact that $\|\mathbf{x}\| = \||\mathbf{x}|\|$. □

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Note that for any $s \geq 1$, we know $\|\mathbf{a}_i\|_s \leq \|\mathbf{a}_i\|_1$, implying that $\sum_{i=1}^m \|\mathbf{a}_i\|_s^r \leq \sum_{i=1}^m \|\mathbf{a}_i\|_1^r$. Therefore, Lemma 6.2.2 provides a tighter bound than Lemma 6.2.1. The vector $\mathbf{v}$ in the statement of Lemma 6.2.3 can be written as $\mathbf{v} = \sum_{i=1}^m |\mathbf{a}_i|$, where the $|\cdot|$ is applied element-wise to $\mathbf{a}_i$. We thus have,

$$\|\mathbf{v}\|_s = \left\| \sum_{i=1}^m |\mathbf{a}_i| \right\|_s \leq \sum_{i=1}^m \|\mathbf{a}_i\|_s.$$

It is clear that when $r = 1$, Lemma 6.2.3 gives a tighter bound than Lemma 6.2.2. However, when $r \geq s$, we claim that Lemma 6.2.2 yields a better bound. To see this, notice that

$$\begin{aligned}
\|\mathbf{v}\|_s^r &= \left(\sum_{j=1}^n v_j^s \right)^{r/s} \\
&= \left(\sum_{j=1}^n (|a_{1j}| + \cdots + |a_{mj}|)^s \right)^{r/s} \\
&\geq \left(\sum_{j=1}^n (|a_{1j}|^s + \cdots + |a_{mj}|^s) \right)^{r/s} \\
&= \left(\sum_{i=1}^m (|a_{i1}|^s + \cdots + |a_{in}|^s) \right)^{r/s} \\
&\geq \sum_{i=1}^m (|a_{i1}|^s + \cdots + |a_{in}|^s)^{r/s} \\
&= \sum_{i=1}^m \|\mathbf{a}_i\|_s^r,
\end{aligned}$$

where in the derivation we have used Lemma 6.2.4.

Lemma 6.2.4. For any $k \geq 1$ and $c_i \geq 0$, the following holds:

$$\left(\sum_{i=1}^m c_i \right)^k \geq \sum_{i=1}^m c_i^k.$$

Proof. If all $c_i = 0$, the result obviously holds. Without loss of generality, we assume not all c_i are equal to zero. Let $\lambda = \sum_{i=1}^m c_i$; then $\lambda > 0$. Set $b_i = c_i/\lambda$; then $b_i \in [0, 1]$, and $b_i^k \leq b_i$. Together with the fact that

The version of record is available at: http://dx.doi.org/10.1561/2400000026

$\sum_i b_i = 1$, we have:

$$\begin{aligned}\sum_{i=1}^{m} c_i^k &= \lambda^k \sum_{i=1}^{m} b_i^k \\ &\leq \lambda^k \sum_{i=1}^{m} b_i \\ &= \lambda^k \\ &= \Big(\sum_{i=1}^{m} c_i\Big)^k,\end{aligned}$$

for any $k \geq 1, c_i \geq 0$. □

We now proceed to obtain a tractable relaxation to formulation (6.1). Using Lemma 6.2.2 and Theorem 3.1.1, we have:

$$\begin{aligned}&\big|\mathbb{E}^{\mathbb{Q}}[h_{\tilde{\mathbf{B}}}(\mathbf{z})] - \mathbb{E}^{\hat{\mathbb{P}}_N}[h_{\tilde{\mathbf{B}}}(\mathbf{z})]\big| \\ &\quad\leq \int_{\mathcal{Z}\times\mathcal{Z}} \frac{|h_{\tilde{\mathbf{B}}}(\mathbf{z}_1) - h_{\tilde{\mathbf{B}}}(\mathbf{z}_2)|}{\|\mathbf{z}_1 - \mathbf{z}_2\|_r} \|\mathbf{z}_1 - \mathbf{z}_2\|_r \mathrm{d}\pi_0(\mathbf{z}_1, \mathbf{z}_2) \\ &\quad\leq \int_{\mathcal{Z}\times\mathcal{Z}} \frac{L\|\tilde{\mathbf{B}}(\mathbf{z}_1 - \mathbf{z}_2)\|_r}{\|\mathbf{z}_1 - \mathbf{z}_2\|_r} \|\mathbf{z}_1 - \mathbf{z}_2\|_r \mathrm{d}\pi_0(\mathbf{z}_1, \mathbf{z}_2) \\ &\quad\leq L\Big(\sum_{i=1}^{K} \|\mathbf{b}_i\|_s^r\Big)^{1/r} \int_{\mathcal{Z}\times\mathcal{Z}} \|\mathbf{z}_1 - \mathbf{z}_2\|_r \mathrm{d}\pi_0(\mathbf{z}_1, \mathbf{z}_2) \\ &\quad= L\Big(\sum_{i=1}^{K} \|\mathbf{b}_i\|_s^r\Big)^{1/r} W_{s,1}(\mathbb{Q},\ \hat{\mathbb{P}}_N) \\ &\quad\leq \epsilon L\Big(\sum_{i=1}^{K} \|\mathbf{b}_i\|_s^r\Big)^{1/r}, \quad \forall \mathbb{Q} \in \Omega,\end{aligned}$$

where $\mathbf{b}_i = (-B_{1i}, \ldots, -B_{pi}, \mathbf{e}_i)$ is the i-th row of $\tilde{\mathbf{B}}$, with $\mathbf{e}_i$ the i-th unit vector in $\mathbb{R}^K$. The above derivation implies that when the Wasserstein metric is induced by $\|\cdot\|_r$,

$$\sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h_{\tilde{\mathbf{B}}}(\mathbf{z})] \leq \mathbb{E}^{\hat{\mathbb{P}}_N}[h_{\tilde{\mathbf{B}}}(\mathbf{z})] + \epsilon L\Big(\sum_{i=1}^{K} \|\mathbf{b}_i\|_s^r\Big)^{1/r},$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where $1/r + 1/s = 1$. This directly yields the following relaxation to (6.1):

$$\inf_{\mathbf{B}} \frac{1}{N} \sum_{i=1}^{N} h_{\mathbf{B}}(\mathbf{x}_i, \mathbf{y}_i) + \epsilon L \Big(\sum_{i=1}^{K} \|\mathbf{b}_i\|_s^r \Big)^{1/r}, \tag{6.2}$$

which we call the MLR-SR relaxation. The regularization term in (6.2) penalizes the aggregate of the dual norm of the regression coefficients corresponding to each of the K responses. Notice that when $r \neq 1$, (6.2) cannot be decomposed into K independent terms. When $s = r = 2$, the regularizer is just the Frobenius norm of $\tilde{\mathbf{B}}$. Using a similar derivation, Lemma 6.2.3 yields the following relaxation to (6.1):

$$\inf_{\mathbf{B}} \frac{1}{N} \sum_{i=1}^{N} h_{\mathbf{B}}(\mathbf{x}_i, \mathbf{y}_i) + \epsilon L \|\mathbf{v}\|_s, \tag{6.3}$$

where $\mathbf{v} \triangleq (v_1, \ldots, v_p, 1, \ldots, 1)$, with $v_i = \sum_{j=1}^{K} |B_{ij}|$, i.e., v_i is a condensed representation of the coefficients for predictor i through summing over the K coordinates. We call (6.3) the MLR-1S relaxation (the naming convention will be more clear after introducing the $L_{r,s}$ matrix norm in Section 6.2.2). When $s \neq 1$, it cannot be decomposed into K subproblems due to the entangling of coefficients in the regularization term.

Note that when $K = 1$, with an 1-Lipschitz continuous loss function, the two regularizers in MLR-SR and MLR-1S reduce to $\epsilon\|(-\boldsymbol{\beta}, 1)\|_s$, which coincides with the Wasserstein DRO formulation derived in Section 4 with an absolute error loss. In both relaxations for MLR, the Wasserstein ball radius ϵ and the Lipschitz constant L determine the strength of the penalty term. Recall that we assume the loss function is Lipschitz continuous on the same norm space with the one used by the Wasserstein metric. This assumption can be relaxed by allowing a different norm space for the Lipschitz continuous loss function, and the derivation technique can be easily adapted to obtain relaxations to (6.1). On the other hand, however, the norm space used by the Wasserstein metric can provide implications on what loss function to choose. For example, if we restrict the class of loss functions $l(\cdot)$ to the norms, our assumption suggests that $l(\mathbf{z}) = \|\mathbf{z}\|_r$, which is a reasonable choice since it reflects the distance metric on the data space.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

6.2.2 A New Perspective on the Formulation

In this subsection we will present a matrix norm interpretation for the two relaxations (6.2) and (6.3). Different from the commonly used matrix norm definitions in the literature, e.g., the vector norm-induced matrix norm $\|\mathbf{A}\| \triangleq \max_{\|\mathbf{x}\|\leq 1} \|\mathbf{A}\mathbf{x}\|$, the entrywise norm that treats the matrix as a vector, and the Schatten–von-Neumann norm that defines the norm on the vector of singular values [160], we adopt the $L_{r,s}$ norm, which summarizes each column by its ℓ_r norm, and then computes the ℓ_s norm of the aggregate vector. The formal definition is described as follows.

Definition 6 ($L_{r,s}$ Matrix Norm)**.** For any $m \times n$ matrix $\mathbf{A} = (a_{ij})_{i\in[\![m]\!]}^{j\in[\![n]\!]}$, define its $L_{r,s}$ norm as:

$$\|\mathbf{A}\|_{r,s} \triangleq \left(\sum_{j=1}^{n}\left(\sum_{i=1}^{m}|a_{ij}|^r\right)^{s/r}\right)^{1/s},$$

where $r, s \geq 1$.

Note that $\|\mathbf{A}\|_{r,s}$ can be viewed as the ℓ_s norm of a newly defined vector $\mathbf{v} = (v_1, \ldots, v_n)$, where $v_j = \|\mathbf{A}_j\|_r$, with $\mathbf{A}_j$ the j-th column of $\mathbf{A}$. When $r = s = 2$, the $L_{r,s}$ norm is the Frobenius norm. Moreover, $\|\mathbf{A}\|_{r,s}$ is a convex function in $\mathbf{A}$, which can be shown as follows.

Proof. For two matrices $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_n]$, $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_n]$, where $\mathbf{A}_i, \mathbf{B}_i$ are the columns of $\mathbf{A}$ and $\mathbf{B}$, respectively, consider their convex combination $\lambda\mathbf{A} + (1-\lambda)\mathbf{B}$, where $\lambda \in [0, 1]$. Its $L_{r,s}$ norm can be expressed as:

$$\begin{aligned}
&\|\lambda\mathbf{A} + (1-\lambda)\mathbf{B}\|_{r,s} \\
&\quad = \|(\|\lambda\mathbf{A}_1 + (1-\lambda)\mathbf{B}_1\|_r, \ldots, \|\lambda\mathbf{A}_n + (1-\lambda)\mathbf{B}_n\|_r)\|_s \\
&\quad \leq \|(\lambda\|\mathbf{A}_1\|_r + (1-\lambda)\|\mathbf{B}_1\|_r, \ldots, \lambda\|\mathbf{A}_n\|_r + (1-\lambda)\|\mathbf{B}_n\|_r)\|_s \\
&\quad = \|\lambda(\|\mathbf{A}_1\|_r, \ldots, \|\mathbf{A}_n\|_r) + (1-\lambda)(\|\mathbf{B}_1\|_r, \ldots, \|\mathbf{B}_n\|_r)\|_s
\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

$$\leq \lambda\|(\|\mathbf{A}_1\|_r, \ldots, \|\mathbf{A}_n\|_r)\|_s + (1-\lambda)\|(\|\mathbf{B}_1\|_r, \ldots, \|\mathbf{B}_n\|_r)\|_s$$
$$= \lambda\|\mathbf{A}\|_{r,s} + (1-\lambda)\|\mathbf{B}\|_{r,s}.$$

Therefore, the $L_{r,s}$ norm is convex. □

The $L_{r,s}$ matrix norm depends on the structure of the matrix, and transposing a matrix changes its norm. For example, given $\mathbf{A} \in \mathbb{R}^{n\times 1}$, $\|\mathbf{A}\|_{r,s} = \|\mathbf{a}\|_r$, $\|\mathbf{A}'\|_{r,s} = \|\mathbf{a}\|_s$, where $\mathbf{a}$ represents the vectorization of $\mathbf{A}$. To show the validity of the $L_{r,s}$ norm, we need to verify the following properties:

1. $\|\mathbf{A}\|_{r,s} \geq 0$.
2. $\|\mathbf{A}\|_{r,s} = 0$ if and only if $\mathbf{A} = 0$.
3. $\|\alpha\mathbf{A}\|_{r,s} = |\alpha|\|\mathbf{A}\|_{r,s}$.
4. $\|\mathbf{A}+\mathbf{B}\|_{r,s} \leq \|\mathbf{A}\|_{r,s} + \|\mathbf{B}\|_{r,s}$.

The first three properties are straightforward. To show the sub-additivity property (triangle inequality), assume $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_n]$ and $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_n]$, where $\mathbf{A}_j, \mathbf{B}_j, j \in [\![n]\!]$, are the columns of $\mathbf{A}$ and $\mathbf{B}$, respectively. Define two vectors $\mathbf{v} \triangleq (\|\mathbf{A}_1\|_r, \ldots, \|\mathbf{A}_n\|_r)$, and $\mathbf{t} \triangleq (\|\mathbf{B}_1\|_r, \ldots, \|\mathbf{B}_n\|_r)$, we have:

$$\begin{aligned}
\|\mathbf{A}\|_{r,s} + \|\mathbf{B}\|_{r,s} &= \|\mathbf{v}\|_s + \|\mathbf{t}\|_s \\
&\geq \|\mathbf{v}+\mathbf{t}\|_s \\
&= \left(\sum_{i=1}^{n}(\|\mathbf{A}_i\|_r + \|\mathbf{B}_i\|_r)^s\right)^{1/s} \\
&\geq \left(\sum_{i=1}^{n}\|\mathbf{A}_i + \mathbf{B}_i\|_r^s\right)^{1/s} \\
&= \|\mathbf{A}+\mathbf{B}\|_{r,s}.
\end{aligned}$$

The $L_{r,s}$ norm also satisfies the following *sub-multiplicative* property:

$$\|\mathbf{AB}\|_{r,s} \leq \|\mathbf{A}\|_{1,u}\|\mathbf{B}\|_{t,s}, \tag{6.4}$$

for $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times K}$, and any $t, u \geq 1$ satisfying $1/t + 1/u = 1$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Proof. Assume $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times K}$, and $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_K]$, where $\mathbf{B}_j, j \in [\![K]\!]$, are the columns of $\mathbf{B}$. Then, $\mathbf{AB} = [\mathbf{AB}_1, \ldots, \mathbf{AB}_K]$, and $\|\mathbf{AB}\|_{r,s} = \|\mathbf{w}\|_s$, where $\mathbf{w} = (w_1, \ldots, w_K)$ with $w_j = \|\mathbf{AB}_j\|_r$. From the proof of Lemma 6.2.3, we immediately have:

$$w_j = \|\mathbf{AB}_j\|_r \leq \|\mathbf{B}_j\|_t \|\mathbf{A}\|_{1,u},$$

where $1/t + 1/u = 1$. We thus have,

$$\begin{aligned}
\|\mathbf{AB}\|_{r,s} &= \left(\sum_{j=1}^{K} w_j^s\right)^{1/s} \\
&\leq \left(\sum_{j=1}^{K} \|\mathbf{B}_j\|_t^s \|\mathbf{A}\|_{1,u}^s\right)^{1/s} \\
&= \|\mathbf{A}\|_{1,u} \left(\sum_{j=1}^{K} \|\mathbf{B}_j\|_t^s\right)^{1/s} \\
&= \|\mathbf{A}\|_{1,u} \|\mathbf{B}\|_{t,s},
\end{aligned}$$

for any $t, u \geq 1$ satisfying $1/t + 1/u = 1$. □

Next we will reformulate the two relaxations (6.2) and (6.3) using the $L_{r,s}$ norm. When the Wasserstein metric is defined by $\|\cdot\|_r$, the MLR-SR relaxation can be written as:

$$\inf_{\mathbf{B}} \frac{1}{N} \sum_{i=1}^{N} h_{\mathbf{B}}(\mathbf{x}_i, \mathbf{y}_i) + \epsilon L \|\tilde{\mathbf{B}}'\|_{s,r}.$$

Similarly, the MLR-1S relaxation can be written as:

$$\inf_{\mathbf{B}} \frac{1}{N} \sum_{i=1}^{N} h_{\mathbf{B}}(\mathbf{x}_i, \mathbf{y}_i) + \epsilon L \|\tilde{\mathbf{B}}\|_{1,s},$$

where $r, s \geq 1$ and $1/r + 1/s = 1$. When the loss function is convex, e.g., $h_{\mathbf{B}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{y} - \mathbf{B}'\mathbf{x}\|$, it is obvious that both MLR-SR and MLR-1S are convex optimization problems. By using the $L_{r,s}$ matrix norm, we are able to express the two relaxations in a compact way, which reflects the role of the norm space induced by the Wasserstein metric on the regularizer, and demonstrates the impact of the size of the Wasserstein ambiguity set and the Lipschitz continuity of the loss function on the regularization strength.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

6.2.3 Distributionally Robust Multiclass Logistic Regression

In this subsection we apply the Wasserstein DRO framework to the problem of *Multiclass Logistic Regression (MLG)*. Suppose there are K classes, and we are given a predictor vector $\mathbf{x} \in \mathbb{R}^p$. Our goal is to predict its class label, denoted by a K-dimensional binary label vector $\mathbf{y} \in \{0,1\}^K$, where $\sum_k y_k = 1$, and $y_k = 1$ if and only if $\mathbf{x}$ belongs to class k. The conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is modeled as

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{K} p_i^{y_i},$$

where $p_i = e^{\mathbf{w}_i'\mathbf{x}} / \sum_{k=1}^{K} e^{\mathbf{w}_k'\mathbf{x}}$, and $\mathbf{w}_i, i \in [\![K]\!]$, are the coefficient vectors to be estimated that account for the contribution of $\mathbf{x}$ in predicting the class labels. The log-likelihood can be expressed as:

$$\begin{aligned}
\log p(\mathbf{y}|\mathbf{x}) &= \sum_{i=1}^{K} y_i \log(p_i) \\
&= \sum_{i=1}^{K} y_i \log \frac{e^{\mathbf{w}_i'\mathbf{x}}}{\sum_{k=1}^{K} e^{\mathbf{w}_k'\mathbf{x}}} \\
&= \sum_{i=1}^{K} y_i \mathbf{w}_i'\mathbf{x} - \left(\log \sum_{k=1}^{K} e^{\mathbf{w}_k'\mathbf{x}} \right) \sum_{i=1}^{K} y_i \\
&= \sum_{i=1}^{K} y_i \mathbf{w}_i'\mathbf{x} - \log \sum_{k=1}^{K} e^{\mathbf{w}_k'\mathbf{x}} \\
&= \mathbf{y}'\mathbf{B}'\mathbf{x} - \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}},
\end{aligned}$$

where $\mathbf{B} \triangleq [\mathbf{w}_1, \ldots, \mathbf{w}_K]$, $\mathbf{1}$ is the vector of ones, and the exponential operator is applied element-wise to the exponent vector. The log-loss is defined to be the negative log-likelihood, i.e., $h_{\mathbf{B}}(\mathbf{x}, \mathbf{y}) \triangleq \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}} - \mathbf{y}'\mathbf{B}'\mathbf{x}$. The Wasserstein DRO formulation for MLG minimizes the following worst-case expected loss:

$$\inf_{\mathbf{B}} \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}}[\log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}} - \mathbf{y}'\mathbf{B}'\mathbf{x}], \tag{6.5}$$

where Ω is defined using the order-1 Wasserstein metric induced by:

$$s(\mathbf{z}_1, \mathbf{z}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_r + M s_{\mathbf{y}}(\mathbf{y}_1, \mathbf{y}_2),$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{y}_1)$, $\mathbf{z}_2 = (\mathbf{x}_2, \mathbf{y}_2)$, $s_{\mathbf{y}}(\cdot, \cdot)$ could be any metric, and M is a very large positive constant. To make (6.5) tractable, we need to derive an upper bound for the growth rate of the loss function, which involves bounding the following difference

$$\begin{aligned} & |h_{\mathbf{B}}(\mathbf{x}_1, \mathbf{y}_1) - h_{\mathbf{B}}(\mathbf{x}_2, \mathbf{y}_2)| \\ = \; & |\log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_1} - \mathbf{y}_1' \mathbf{B}' \mathbf{x}_1 - \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_2} + \mathbf{y}_2' \mathbf{B}' \mathbf{x}_2| \\ \leq \; & |\log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_1} - \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_2}| + |\mathbf{y}_1' \mathbf{B}' \mathbf{x}_1 - \mathbf{y}_2' \mathbf{B}' \mathbf{x}_2|, \end{aligned} \tag{6.6}$$

in terms of $s(\mathbf{z}_1, \mathbf{z}_2)$. Let us examine the two terms in (6.6) separately. For the first term, define a function $g(\mathbf{a}) = \log \mathbf{1}' e^{\mathbf{a}}$, where $\mathbf{a} \in \mathbb{R}^K$. Using the mean value theorem, we know for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, there exists some $t \in (0, 1)$ such that

$$|g(\mathbf{b}) - g(\mathbf{a})| \leq \|\nabla g((1-t)\mathbf{a} + t\mathbf{b})\|_s \|\mathbf{b} - \mathbf{a}\|_r \leq K^{1/s} \|\mathbf{b} - \mathbf{a}\|_r, \tag{6.7}$$

where $r, s \geq 1$, $1/r + 1/s = 1$, the first inequality is due to Hölder's inequality, and the second inequality is due to the fact that $\nabla g(\mathbf{a}) = e^{\mathbf{a}}/\mathbf{1}' e^{\mathbf{a}}$, which implies that each element of $\nabla g(\mathbf{a})$ is smaller than 1. Based on (6.7) we have:

$$|\log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_1} - \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_2}| \leq K^{1/s} \|\mathbf{B}'(\mathbf{x}_1 - \mathbf{x}_2)\|_r.$$

We can use Lemma 6.2.2 or 6.2.3 to bound $\|\mathbf{B}'(\mathbf{x}_1 - \mathbf{x}_2)\|_r$, which respectively leads to the following two results:

$$\begin{aligned} |\log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_1} - \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_2}| &\leq K^{1/s} \|\mathbf{x}_1 - \mathbf{x}_2\|_r \Big(\sum_{i=1}^{K} \|\mathbf{w}_i\|_s^r \Big)^{1/r} \\ &= K^{1/s} \|\mathbf{x}_1 - \mathbf{x}_2\|_r \|\mathbf{B}\|_{s,r}, \end{aligned} \tag{6.8}$$

and

$$|\log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_1} - \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}_2}| \leq K^{1/s} \|\mathbf{x}_1 - \mathbf{x}_2\|_r \|\mathbf{B}'\|_{1,s}, \tag{6.9}$$

where $r, s \geq 1$ and $1/r + 1/s = 1$. By noting that $\|\mathbf{x}_1 - \mathbf{x}_2\|_r \leq s(\mathbf{z}_1, \mathbf{z}_2)$, we obtain the upper bound for the first term in (6.6) in terms of $s(\mathbf{z}_1, \mathbf{z}_2)$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

For the second term, we have,

$$\begin{aligned}|\mathbf{y}_1'\mathbf{B}'\mathbf{x}_1 - \mathbf{y}_2'\mathbf{B}'\mathbf{x}_2| &= \Big|\sum_{i=1}^{K}\mathbf{w}_i'(y_{1i}\mathbf{x}_1 - y_{2i}\mathbf{x}_2)\Big| \\ &\le \sum_{i=1}^{K}|\mathbf{w}_i'(y_{1i}\mathbf{x}_1 - y_{2i}\mathbf{x}_2)| \\ &\le \sum_{i=1}^{K}\|\mathbf{w}_i\|_s\|y_{1i}\mathbf{x}_1 - y_{2i}\mathbf{x}_2\|_r \\ &\le s(\mathbf{z}_1,\mathbf{z}_2)\sum_{i=1}^{K}\|\mathbf{w}_i\|_s \\ &= s(\mathbf{z}_1,\mathbf{z}_2)\|\mathbf{B}\|_{s,1}, \end{aligned} \tag{6.10}$$

where $\mathbf{y}_1 = (y_{11},\ldots,y_{1K}), \mathbf{y}_2 = (y_{21},\ldots,y_{2K})$, $1/s+1/r = 1$, the second inequality uses the Hölder's inequality, and the last inequality can be proved by noting that if $y_{1i} = y_{2i}$, $\|y_{1i}\mathbf{x}_1 - y_{2i}\mathbf{x}_2\|_r \le s(\mathbf{z}_1,\mathbf{z}_2)$; otherwise $s(\mathbf{z}_1,\mathbf{z}_2)$ goes to infinity. Suppose π_0 is the optimal transportation plan that moves the probability mass from $\mathbb{Q}$ to $\hat{\mathbb{P}}_N$, combining (6.8) with (6.10), we have:

$$\begin{aligned}&|\mathbb{E}^{\mathbb{Q}}[h_{\mathbf{B}}(\mathbf{x},\mathbf{y})] - \mathbb{E}^{\hat{\mathbb{P}}_N}[h_{\mathbf{B}}(\mathbf{x},\mathbf{y})]| \\ &\le \int_{\mathcal{Z}\times\mathcal{Z}}|h_{\mathbf{B}}(\mathbf{x}_1,\mathbf{y}_1) - h_{\mathbf{B}}(\mathbf{x}_2,\mathbf{y}_2)|\mathrm{d}\pi_0(\mathbf{z}_1,\mathbf{z}_2) \\ &= \int_{\mathcal{Z}\times\mathcal{Z}}\frac{|h_{\mathbf{B}}(\mathbf{x}_1,\mathbf{y}_1) - h_{\mathbf{B}}(\mathbf{x}_2,\mathbf{y}_2)|}{s(\mathbf{z}_1,\mathbf{z}_2)}s(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\pi_0(\mathbf{z}_1,\mathbf{z}_2) \\ &\le \int_{\mathcal{Z}\times\mathcal{Z}}(K^{1/s}\|\mathbf{B}\|_{s,r} + \|\mathbf{B}\|_{s,1})s(\mathbf{z}_1,\mathbf{z}_2)\mathrm{d}\pi_0(\mathbf{z}_1,\mathbf{z}_2) \\ &\le \epsilon(K^{1/s}\|\mathbf{B}\|_{s,r} + \|\mathbf{B}\|_{s,1}),\end{aligned}$$

which yields the following MLG-SR relaxation to (6.5):

$$\inf_{\mathbf{B}}\frac{1}{N}\sum_{i=1}^{N}(\log\mathbf{1}'e^{\mathbf{B}'\mathbf{x}_i} - \mathbf{y}_i'\mathbf{B}'\mathbf{x}_i) + \epsilon(K^{1/s}\|\mathbf{B}\|_{s,r} + \|\mathbf{B}\|_{s,1}).$$

Similarly, combining (6.9) with (6.10) produces the following MLG-1S relaxation:

$$\inf_{\mathbf{B}}\frac{1}{N}\sum_{i=1}^{N}(\log\mathbf{1}'e^{\mathbf{B}'\mathbf{x}_i} - \mathbf{y}_i'\mathbf{B}'\mathbf{x}_i) + \epsilon(K^{1/s}\|\mathbf{B}'\|_{1,s} + \|\mathbf{B}\|_{s,1}).$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

We note that both MLG-SR and MLG-1S are convex optimization problems. The convexity of the regularizer has been shown in Section 6.2.2. The convexity of the log-loss is shown in the following theorem.

Theorem 6.2.5. The log-loss $h_{\mathbf{B}}(\mathbf{x}, \mathbf{y}) \triangleq \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}} - \mathbf{y}'\mathbf{B}'\mathbf{x}$ is convex in $\mathbf{B}$.

Proof. Since the linear function is convex, we only need to show the convexity of $\log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}}$. The following result will be used.

Corollary 6.2.6. The function $f(\mathbf{x}) = \log(\sum_{i=1}^{n} e^{x_i})$ is a convex function of $\mathbf{x} \in \mathbb{R}^n$.

By Corollary 6.2.6, we have for any $\lambda \in [0, 1]$, and any two matrices $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_K]$ and $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_K]$,

$$\begin{aligned}
\log \mathbf{1}' e^{(\lambda\mathbf{B}+(1-\lambda)\mathbf{C})'\mathbf{x}} &= \log \left(\sum_{i=1}^{K} e^{\lambda\mathbf{B}_i'\mathbf{x}+(1-\lambda)\mathbf{C}_i'\mathbf{x}} \right) \\
&= f(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2) \\
&\leq \lambda f(\mathbf{v}_1) + (1-\lambda) f(\mathbf{v}_2) \\
&= \lambda \log \left(\sum_{i=1}^{K} e^{\mathbf{B}_i'\mathbf{x}} \right) + (1-\lambda) \log \left(\sum_{i=1}^{K} e^{\mathbf{C}_i'\mathbf{x}} \right) \\
&= \lambda \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}} + (1-\lambda) \log \mathbf{1}' e^{\mathbf{C}'\mathbf{x}},
\end{aligned}$$

where $\mathbf{v}_1 = (\mathbf{B}_1'\mathbf{x}, \ldots, \mathbf{B}_K'\mathbf{x})$, and $\mathbf{v}_2 = (\mathbf{C}_1'\mathbf{x}, \ldots, \mathbf{C}_K'\mathbf{x})$. Therefore the log-loss is convex. □

When $K = 2$, by taking one of the two classes as a reference, we can set one column of $\mathbf{B}$ to zero, in which case all three regularizers $\|\mathbf{B}\|_{s,r}$, $\|\mathbf{B}\|_{s,1}$ and $\|\mathbf{B}'\|_{1,s}$ reduce to $\|\boldsymbol{\beta}\|_s$, where $\mathbf{B} \triangleq [\boldsymbol{\beta}, \mathbf{0}]$, and the MLG-SR and MLG-1S relaxations coincide with the regularized logistic regression formulation derived in (5.8).

We also note that the number of classes K, along with the Wasserstein set radius ϵ, determines the regularization magnitude in the two MLG relaxations. There are two terms in the regularizer, one accounting for the predictor/feature uncertainty, and the other accounting for

The version of record is available at: http://dx.doi.org/10.1561/2400000026

the label uncertainty. In the MLG-SR regularizer, we summarize each column of $\mathbf{B}$ by its dual norm, and aggregate them by the ℓ_r and ℓ_1 norms to reflect the predictor and label uncertainties, respectively.

6.3 The Out-of-Sample Performance Guarantees

In this section we will show the out-of-sample performance guarantees for the solutions to the MLR and MLG relaxations, i.e., given a new test sample, what is the expected prediction bias/log-loss. The results are established using the Rademacher complexity [98], following the line of proof presented in Section 4.3.1. The resulting bounds shed light on the role of the regularizer in inducing a low prediction error.

6.3.1 Performance Guarantees for MLR Relaxations

In this subsection we study the out-of-sample predictive performance of the solutions to (6.2) and (6.3). Suppose the data $(\mathbf{x}, \mathbf{y})$ is drawn from the probability measure $\mathbb{P}^*$. We first make the following assumptions that are essential for deriving the bounds.

Assumption Q. The ℓ_r-norm of the data $(\mathbf{x}, \mathbf{y})$ is bounded above a.s. under the probability measure $\mathbb{P}^*$, i.e., $\|(\mathbf{x}, \mathbf{y})\|_r \leq R,$ a.s.

Assumption R. For any feasible solution to MLR-SR, it holds that $\|\tilde{\mathbf{B}}'\|_{s,r} \leq \bar{B}_{s,r}$.

Assumption S. For any feasible solution to MLR-1S, it holds that $\|\tilde{\mathbf{B}}\|_{1,s} \leq \bar{B}_{1,s}$.

Assumption T. The loss resulting from $(\mathbf{x}, \mathbf{y}) = (\mathbf{0}, \mathbf{0})$ is 0, i.e., $h_{\tilde{\mathbf{B}}}(\mathbf{0}) = 0$.

Note that Assumption Q bounds the magnitude of the data in terms of its ℓ_r-norm, and R can be assumed to be reasonably small with standardized data input. Assumptions R and S impose restrictions on the norm of the coefficient matrix, which are a result of adding appropriate regularizers into the formulation as in (6.2) and (6.3). Assumption T easily holds when the loss function is defined via some norm, i.e., $h_{\mathbf{B}}(\mathbf{x}, \mathbf{y}) \triangleq \|\mathbf{y} - \mathbf{B}'\mathbf{x}\|$. Under Assumptions Q, R and T,

The version of record is available at: http://dx.doi.org/10.1561/2400000026

using Lemma 6.2.2 and the Lipschitz continuity of the loss function, we have

$$h_{\tilde{\mathbf{B}}}(\mathbf{z}) \leq L\|\tilde{\mathbf{B}}\mathbf{z}\|_r \leq L\|\mathbf{z}\|_r\|\tilde{\mathbf{B}}'\|_{s,r} \leq LR\bar{B}_{s,r}. \tag{6.11}$$

Similarly, under Assumptions Q, S and T, Lemma 6.2.3 yields the following:

$$h_{\tilde{\mathbf{B}}}(\mathbf{z}) \leq L\|\tilde{\mathbf{B}}\mathbf{z}\|_r \leq L\|\mathbf{z}\|_r\|\tilde{\mathbf{B}}\|_{1,s} \leq LR\bar{B}_{1,s}. \tag{6.12}$$

With the above results, the idea is to bound the out-of-sample prediction error using the empirical *Rademacher complexity* $\mathcal{R}_N(\cdot)$ of the class of loss functions: $\mathcal{H} = \{\mathbf{z} \to h_{\tilde{\mathbf{B}}}(\mathbf{z})\}$, denoted by $\mathcal{R}_N(\mathcal{H})$. Using Lemma 4.3.2 and the upper bounds in (6.11) and (6.12), we arrive at the following result.

Lemma 6.3.1. Under Assumptions Q, R and T,

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{2LR\bar{B}_{s,r}}{\sqrt{N}}.$$

Under Assumptions Q, S and T,

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{2LR\bar{B}_{1,s}}{\sqrt{N}}.$$

Using the Rademacher complexity of the class of loss functions, the out-of-sample prediction bias of the solutions to (6.2) and (6.3) can be bounded by applying Theorem 8 in [98].

Theorem 6.3.2. Suppose the solution to (6.2) is $\hat{\mathbf{B}}_{s,r}$. Under Assumptions Q, R and T, for any $0 < \delta < 1$, with probability at least $1 - \delta$ with respect to the sampling,

$$\mathbb{E}[h_{\hat{\mathbf{B}}_{s,r}}(\mathbf{x}, \mathbf{y})] \leq \frac{1}{N}\sum_{i=1}^{N} h_{\hat{\mathbf{B}}_{s,r}}(\mathbf{x}_i, \mathbf{y}_i) + \frac{2LR\bar{B}_{s,r}}{\sqrt{N}} + LR\bar{B}_{s,r}\sqrt{\frac{8\log(\frac{2}{\delta})}{N}},$$

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

$$\mathbb{P}\left(h_{\hat{\mathbf{B}}_{s,r}}(\mathbf{x}, \mathbf{y}) \geq \frac{1}{N}\sum_{i=1}^{N} h_{\hat{\mathbf{B}}_{s,r}}(\mathbf{x}_i, \mathbf{y}_i) + \zeta\right) \leq \frac{\frac{1}{N}\sum_{i=1}^{N} h_{\hat{\mathbf{B}}_{s,r}}(\mathbf{x}_i, \mathbf{y}_i) + \frac{2LR\bar{B}_{s,r}}{\sqrt{N}} + LR\bar{B}_{s,r}\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N} h_{\hat{\mathbf{B}}_{s,r}}(\mathbf{x}_i, \mathbf{y}_i) + \zeta}.$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Theorem 6.3.3. Suppose the solution to (6.3) is $\hat{\mathbf{B}}_{1,s}$. Under Assumptions Q, S and T, for any $0 < \delta < 1$, with probability at least $1 - \delta$ with respect to the sampling,

$$\mathbb{E}[h_{\hat{\mathbf{B}}_{1,s}}(\mathbf{x}, \mathbf{y})] \leq \frac{1}{N} \sum_{i=1}^{N} h_{\hat{\mathbf{B}}_{1,s}}(\mathbf{x}_i, \mathbf{y}_i) + \frac{2LR\bar{B}_{1,s}}{\sqrt{N}} + LR\bar{B}_{1,s} \sqrt{\frac{8 \log(\frac{2}{\delta})}{N}},$$

and for any $\zeta > \frac{2LR\bar{B}_{1,s}}{\sqrt{N}} + LR\bar{B}_{1,s}\sqrt{\frac{8\log(2/\delta)}{N}}$,

$$\mathbb{P}\left(h_{\hat{\mathbf{B}}_{1,s}}(\mathbf{x}, \mathbf{y}) \geq \frac{1}{N} \sum_{i=1}^{N} h_{\hat{\mathbf{B}}_{1,s}}(\mathbf{x}_i, \mathbf{y}_i) + \zeta \right) \leq \frac{\frac{1}{N}\sum_{i=1}^{N} h_{\hat{\mathbf{B}}_{1,s}}(\mathbf{x}_i, \mathbf{y}_i) + \frac{2LR\bar{B}_{1,s}}{\sqrt{N}} + LR\bar{B}_{1,s}\sqrt{\frac{8\log(2/\delta)}{N}}}{\frac{1}{N}\sum_{i=1}^{N} h_{\hat{\mathbf{B}}_{1,s}}(\mathbf{x}_i, \mathbf{y}_i) + \zeta}.$$

Theorems 6.3.2 and 6.3.3 present bounds on the out-of-sample prediction errors of the solutions to (6.2) and (6.3), respectively. The expectations/probabilities are taken w.r.t. the new sample $(\mathbf{x}, \mathbf{y})$. The magnitude of the regularizer plays a role in controlling the bias, and a smaller upper bound on the matrix norm leads to a smaller prediction error, suggesting the superiority of MLR-SR for $r \geq 2$, and the superiority of MLR-1S for $r = 1$ (see Section 6.2.1). But on the other hand, the prediction error also depends on the sample average loss over the training set, for which there is no guarantee on which model wins out. In practice we suggest trying both models and selecting the one that yields a smaller error on a validation set.

6.3.2 Performance Guarantees for MLG Relaxations

In this subsection, we study the out-of-sample log-loss of the solutions to MLG-SR and MLG-1S. Suppose the data $(\mathbf{x}, \mathbf{y})$ is drawn from the probability measure $\mathbb{P}^*$. We first make several assumptions that are needed to establish the results.

Assumption U. The ℓ_r norm of the predictor $\mathbf{x}$ is bounded above a.s. under the probability measure $\mathbb{P}^*_{\mathcal{X}}$, i.e., $\|\mathbf{x}\|_r \leq R_{\mathbf{x}}$, a.s.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Assumption V. For any feasible solution to MLG-SR, the following holds:

$$K^{1/s}\|\mathbf{B}\|_{s,r} + \|\mathbf{B}\|_{s,1} \leq \bar{C}_{s,r}.$$

Assumption W. For any feasible solution to MLG-1S, the following holds:

$$K^{1/s}\|\mathbf{B}'\|_{1,s} + \|\mathbf{B}\|_{s,1} \leq \bar{C}_{1,s}.$$

With standardized predictors, $R_{\mathbf{x}}$ in Assumption U can be assumed to be small. The form of the constraints in Assumptions V and W is consistent with the form of the regularizers in MLG-SR and MLG-1S, respectively. We will see later that the bounds $\bar{C}_{s,r}$ and $\bar{C}_{1,s}$ respectively control the out-of-sample log-loss of the solutions to MLG-SR and MLG-1S, which validates the role of the regularizer in improving the out-of-sample performance. Under Assumptions U and V, using (6.6), (6.8) and (6.10), we have,

$$|h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) - h_{\mathbf{B}}(\mathbf{0},\mathbf{y})| \leq K^{1/s}\|\mathbf{x}\|_r\|\mathbf{B}\|_{s,r} + \|\mathbf{x}\|_r\|\mathbf{B}\|_{s,1} \leq R_{\mathbf{x}}\bar{C}_{s,r}.$$

By noting that $h_{\mathbf{B}}(\mathbf{0},\mathbf{y}) = \log K$, we immediately have,

$$\log K - R_{\mathbf{x}}\bar{C}_{s,r} \leq h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) \leq R_{\mathbf{x}}\bar{C}_{s,r} + \log K. \tag{6.13}$$

Similarly, under Assumptions U and W, using (6.6), (6.9) and (6.10), we have,

$$|h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) - h_{\mathbf{B}}(\mathbf{0},\mathbf{y})| \leq K^{1/s}\|\mathbf{x}\|_r\|\mathbf{B}'\|_{1,s} + \|\mathbf{x}\|_r\|\mathbf{B}\|_{s,1} \leq R_{\mathbf{x}}\bar{C}_{1,s},$$

which implies that

$$\log K - R_{\mathbf{x}}\bar{C}_{1,s} \leq h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) \leq R_{\mathbf{x}}\bar{C}_{1,s} + \log K. \tag{6.14}$$

Using (6.13) and (6.14), we can now proceed to bound the out-of-sample log-loss using the empirical *Rademacher complexity* $\mathcal{R}_N(\cdot)$ of the following class of loss functions:

$$\mathcal{H} = \{(\mathbf{x},\mathbf{y}) \to h_{\mathbf{B}}(\mathbf{x},\mathbf{y})\colon h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) = \log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}} - \mathbf{y}'\mathbf{B}'\mathbf{x}\}.$$

Lemma 6.3.4. Under Assumptions U and V,

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{2(R_{\mathbf{x}}\bar{C}_{s,r} + \log K)}{\sqrt{N}}.$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Under Assumptions U and W,

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{2(R_{\mathbf{x}}\bar{C}_{1,s} + \log K)}{\sqrt{N}}.$$

Using Lemma 6.3.4, we are able to bound the out-of-sample log-loss of the solutions to MLG-SR and MLG-1S by applying Theorem 8 in [98].

Theorem 6.3.5. Suppose the solution to MLG-SR is $\hat{\mathbf{B}}_{s,r}$. Under Assumptions U and V, for any $0 < \delta < 1$, with probability at least $1 - \delta$ with respect to the sampling,

$$\begin{aligned}\mathbb{E}[\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{s,r}\mathbf{x}} - \mathbf{y}'\hat{\mathbf{B}}'_{s,r}\mathbf{x}] \leq\ & \frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i} - \mathbf{y}_i'\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i) \\ & + \frac{2(R_{\mathbf{x}}\bar{C}_{s,r} + \log K)}{\sqrt{N}} + (R_{\mathbf{x}}\bar{C}_{s,r} + \log K)\sqrt{\frac{8\log(\frac{2}{\delta})}{N}},\end{aligned}$$

and for any $\zeta > \frac{2(R_{\mathbf{x}}\bar{C}_{s,r}+\log K)}{\sqrt{N}} + (R_{\mathbf{x}}\bar{C}_{s,r} + \log K)\sqrt{\frac{8\log(\frac{2}{\delta})}{N}}$,

$$\begin{aligned}&\mathbb{P}\left(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{s,r}\mathbf{x}} - \mathbf{y}'\hat{\mathbf{B}}'_{s,r}\mathbf{x} \geq \frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i} - \mathbf{y}_i'\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i) + \zeta\right) \\ &\quad\leq \frac{\frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i} - \mathbf{y}_i'\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i) + \frac{2(R_{\mathbf{x}}\bar{C}_{s,r}+\log K)}{\sqrt{N}}}{\frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i} - \mathbf{y}_i'\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i) + \zeta} \\ &\qquad + \frac{(R_{\mathbf{x}}\bar{C}_{s,r} + \log K)\sqrt{\frac{8\log(\frac{2}{\delta})}{N}}}{\frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i} - \mathbf{y}_i'\hat{\mathbf{B}}'_{s,r}\mathbf{x}_i) + \zeta}.\end{aligned}$$

Theorem 6.3.6. Suppose the solution to MLG-1S is $\hat{\mathbf{B}}_{1,s}$. Under Assumptions U and W, for any $0 < \delta < 1$, with probability at least $1 - \delta$ with respect to the sampling,

$$\begin{aligned}\mathbb{E}[\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{1,s}\mathbf{x}} - \mathbf{y}'\hat{\mathbf{B}}'_{1,s}\mathbf{x}] \leq\ & \frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i} - \mathbf{y}_i'\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i) \\ & + \frac{2(R_{\mathbf{x}}\bar{C}_{1,s} + \log K)}{\sqrt{N}} + (R_{\mathbf{x}}\bar{C}_{1,s} + \log K)\sqrt{\frac{8\log(\frac{2}{\delta})}{N}},\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

and for any $\zeta > \frac{2(R_{\mathbf{x}}\bar{C}_{1,s}+\log K)}{\sqrt{N}} + (R_{\mathbf{x}}\bar{C}_{1,s}+\log K)\sqrt{\frac{8\log(\frac{2}{\delta})}{N}}$,

$$\mathbb{P}\Bigg(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{1,s}\mathbf{x}} - \mathbf{y}'\hat{\mathbf{B}}'_{1,s}\mathbf{x} \geq \frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i} - \mathbf{y}'_i\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i) + \zeta\Bigg)$$
$$\leq \frac{\frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i} - \mathbf{y}'_i\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i) + \frac{2(R_{\mathbf{x}}\bar{C}_{1,s}+\log K)}{\sqrt{N}}}{\frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i} - \mathbf{y}'_i\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i) + \zeta}$$
$$+ \frac{(R_{\mathbf{x}}\bar{C}_{1,s}+\log K)\sqrt{\frac{8\log(\frac{2}{\delta})}{N}}}{\frac{1}{N}\sum_{i=1}^{N}(\log \mathbf{1}'e^{\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i} - \mathbf{y}'_i\hat{\mathbf{B}}'_{1,s}\mathbf{x}_i) + \zeta}.$$

We note that the expected log-loss on a new test sample depends both on the sample average log-loss on the training set, and the magnitude of the regularizer in the formulation. The form of the bounds in Theorems 6.3.5 and 6.3.6 demonstrates the validity of MLG-SR and MLG-1S in leading to a good out-of-sample performance. For $r \geq 2$, $\bar{C}_{s,r}$ can be considered smaller than $\bar{C}_{1,s}$, while for $r = 1$, the reverse holds. We can decide which model to use on a case-by-case basis, by computing their out-of-sample error on a validation set.

6.4 Numerical Experiments

In this section, we will test the out-of-sample performance of the MLR and MLG relaxations on a number of synthetic datasets, and compare with several commonly used multi-output regression/classification models.

6.4.1 MLR Relaxations

In this subsection we will first explore the selection of a proper norm for the regularizer based on an appropriate notion of distance in the data space. To this end, we design two different structures for the true coefficient matrix denoted by $\mathbf{B}^*$ in order to reflect different distance metrics.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

1. $\mathbf{B}^*$ is drawn from a standard multivariate normal distribution, which corresponds to an ℓ_2-norm induced Wasserstein metric ($r = 2$).

2. We first generate $\mathbf{B}^*$ from a standard multivariate normal distribution, and then normalize each row using the softmax function while keeping the sign of each element unchanged. The normalization guarantees an equal row absolute sum for $\mathbf{B}^*$. This can be thought of as standardizing the effect of each predictor, which is represented by the absolute sum over the K columns of $\mathbf{B}^*$. Such a coefficient matrix implies an ℓ_1-norm distance metric in the data space ($r = 1$). The reason is that in the dual space ($\|\cdot\|_\infty$), the vertex of the constraint set has each coordinate being the same in absolute value, and in our setting each coordinate is represented by the absolute sum over the K columns of $\mathbf{B}^*$.

The predictor $\mathbf{x}$ is generated from a multivariate normal distribution with mean zero and covariance $\mathbf{\Sigma}_\mathbf{x} = (\sigma_{ij}^\mathbf{x})_{i,j\in[\![p]\!]}$, where $\sigma_{ij}^\mathbf{x} = 0.9^{|i-j|}$. The response vector $\mathbf{y}$ is generated as

$$\mathbf{y} = (\mathbf{B}^*)'\mathbf{x} + \boldsymbol{\eta},$$

where $\boldsymbol{\eta}$ is a standard normal random vector. Throughout the experiments we set $p = 5$, and $K = 3$.

We adopt a loss function $h_\mathbf{B}(\mathbf{x}, \mathbf{y}) = \|\mathbf{y} - \mathbf{B}'\mathbf{x}\|_r$ that is 1-Lipschitz continuous on $\|\cdot\|_r$. Note that we use the same norm to define the loss function and the Wasserstein metric. We will compare the MLR-SR and MLR-1S relaxations induced by $r = 1$ and $r = 2$, respectively, in terms of their out-of-sample *Weighted Mean Squared Error (WMSE)*, defined as:

$$\text{WMSE} \triangleq \frac{1}{M}\sum_{i=1}^{M}(\mathbf{y}_i - \hat{\mathbf{y}}_i)'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i),$$

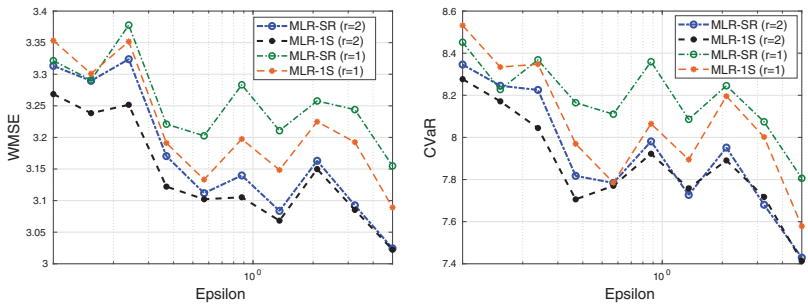
where M is the size of the test set, $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are the true and predicted response vectors for the i-th test sample, respectively, and $\hat{\mathbf{\Sigma}}$ is the covariance matrix of the prediction error on the training set,

$$\hat{\mathbf{\Sigma}} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})/(N - pK),$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

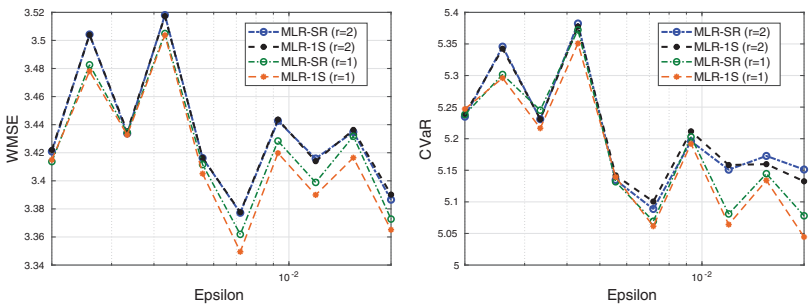
where $\mathbf{Y}, \hat{\mathbf{Y}} \in \mathbb{R}^{N \times K}$ are the true and estimated response matrices of the training set, respectively, and N is the size of the training set. We will also look at the *Conditional Value at Risk (CVaR)* of the WMSE (at the confidence level $\alpha = 0.8$) that quantifies its tail behavior.

Figures 6.1 and 6.2 show the comparison of MLR-SR and MLR-1S formulations derived from the Wasserstein metric induced by the ℓ_r norm, with $r = 1$ and $r = 2$, when the radius of the Wasserstein ball ϵ is varied. As expected, when $\mathbf{B}^*$ is a dense matrix, the ℓ_2 norm is a proper distance metric in the data space, and as a result, the two relaxations with $r = 2$ achieve a lower out-of-sample prediction bias. On



(a) WMSE. (b) CVaR of WMSE.

Figure 6.1: The out-of-sample performance of MLR-SR and MLR-1S with normally distributed $\mathbf{B}^*$.



(a) WMSE. (b) CVaR of WMSE.

Figure 6.2: The out-of-sample performance of MLR-SR and MLR-1S when $\mathbf{B}^*$ has an equal row absolute sum.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Table 6.1: The out-of-sample performance of MLR-SR and MLR-1S with cross-validated ϵ when $\mathbf{B}^*$ is normally distributed

	WMSE	**CVaR**
MLR-SR ($r = 1$)	3.26 (0.48)	4.91 (0.70)
MLR-1S ($r = 1$)	3.21 (0.40)	4.93 (0.65)
MLR-SR ($r = 2$)	3.11 (0.36)	4.74 (0.62)
MLR-1S ($r = 2$)	3.11 (0.35)	4.75 (0.64)

Table 6.2: The out-of-sample performance of MLR-SR and MLR-1S with cross-validated ϵ when $\mathbf{B}^*$ has an equal row absolute sum

	WMSE	**CVaR**
MLR-SR ($r = 1$)	3.04 (0.52)	4.56 (0.83)
MLR-1S ($r = 1$)	3.05 (0.52)	4.60 (0.89)
MLR-SR ($r = 2$)	3.05 (0.52)	4.62 (0.88)
MLR-1S ($r = 2$)	3.05 (0.52)	4.62 (0.89)

the other hand, when the structure of $\mathbf{B}^*$ implies an ℓ_1-norm distance metric on the data (Figure 6.2), the formulations with $r = 1$ have a better performance.

We also compare the four MLR relaxations with an optimal ϵ chosen by cross-validation. Tables 6.1 and 6.2 show the mean WMSE and CVaR over 100 repetitions (the numbers inside the parentheses indicate the corresponding standard deviations). Similar conclusions can be drawn from the results in the tables. With a proper choice of r, the MLR relaxations are able to achieve a lower prediction error with a smaller variance. For example, in Table 6.1, compared to MLR-SR ($r = 1$), the two relaxations with $r = 2$ improved the WMSE by 4.6%.

We next compare the MLR-SR and MLR-1S formulations with several other popular methods for MLR, including *OLS*, *Reduced Rank Regression (RRR)* [142], [143], *Principal Components Regression (PCR)* [144], *Factor Estimation and Selection (FES)* [145], the *Curds and Whey (C&W)* procedure [138], and *Ridge Regression (RR)* [146], [147]. We provide a brief outline of these methods. RRR restricts the rank of $\mathbf{B}$, and its solution is obtained by a *Canonical Correlation Analysis (CCA)* of the response and predictor matrices that finds a sequence of

The version of record is available at: http://dx.doi.org/10.1561/2400000026

uncorrelated linear combinations of the predictors and a corresponding sequence of uncorrelated linear combinations of the responses such that their correlations are successively maximized. PCR converts the predictors into a set of linearly uncorrelated variables and applies OLS on the transformed variables. Both RRR and PCR form linear combinations of predictors and responses (in the case of RRR) which hurts interpretability since it is not possible to explain an original response via the original predictors. FES penalizes the sum of the singular values of $\mathbf{B}$. The C&W procedure shrinks the canonical variates between $\mathbf{x}$ and $\mathbf{y}$. RR penalizes the sum of the squared elements in $\mathbf{B}$ (equivalent to multiple independent ridge regression of each coordinate of $\mathbf{y}$).

To test the robustness of various methods, we inject outliers to the training datasets whose distribution differs from the majority by a normally distributed random quantity. Specifically, the response of outliers is generated as

$$\mathbf{y} = (\mathbf{B}^*)'\mathbf{x} + \boldsymbol{\eta} + \mathbf{o}_{\boldsymbol{\eta}},$$

where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{o}_{\boldsymbol{\eta}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{y}})$, where $\boldsymbol{\Sigma}_{\mathbf{y}} = (\sigma_{ij}^{\mathbf{y}})_{i,j\in[\![K]\!]}$, with $\sigma_{ij}^{\mathbf{y}} = (-0.9)^{|i-j|}$. Note that the perturbation occurs only on the response variables.

We generate 20 datasets with a training size of 100 and a test size of 60, and compare the WMSE and CVaR of various models on a clean test set. All the regularization coefficients are tuned through cross-validation. Table 6.3 shows the average performance on datasets with 20% and 30% outliers, respectively, when $\mathbf{B}^*$ is generated from a standard normal distribution. We see that as the proportion of outliers increases, the WMSE and its CVaR increase, and in both scenarios, the MLR-SR and MLR-1S relaxations achieve the smallest out-of-sample prediction error with a small variance. They improve the WMSE by 1%–5% and 3%–7% when the proportion of outliers is 20% and 30%, respectively. PCR and FES achieve a slightly worse performance, but with a considerably higher variance in the scenario with 20% outlier. PCR works well with linearly correlated predictors, but could possibly fail when there exists a highly nonlinear relationship among the predictors.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Table 6.3: The out-of-sample performance of different MLR models, mean (std)

	WMSE	**CVaR**	**AUC**
Proportion of Outliers		**20%**	
MLR-SR ($r = 2$)	2.55 (0.26)	3.91 (0.51)	0.89 (0.04)
MLR-1S ($r = 2$)	2.59 (0.21)	3.93 (0.43)	0.89 (0.03)
OLS	2.66 (0.44)	4.11 (0.83)	0.85 (0.06)
RR	2.64 (0.42)	4.12 (0.82)	0.87 (0.03)
RRR	2.68 (0.36)	4.03 (0.61)	0.80 (0.05)
FES	2.58 (0.29)	4.01 (0.55)	0.89 (0.03)
C&W	2.65 (0.42)	4.10 (0.80)	0.86 (0.06)
PCR	2.61 (0.29)	4.01 (0.50)	0.86 (0.03)
Proportion of Outliers		**30%**	
MLR-SR ($r = 2$)	2.63 (0.33)	4.07 (0.66)	0.83 (0.09)
MLR-1S ($r = 2$)	2.57 (0.31)	4.02 (0.60)	0.83 (0.09)
OLS	2.75 (0.33)	4.14 (0.71)	0.73 (0.11)
RR	2.72 (0.32)	4.14 (0.60)	0.78 (0.10)
RRR	2.76 (0.44)	4.22 (0.80)	0.71 (0.12)
FES	2.66 (0.33)	4.02 (0.61)	0.83 (0.09)
C&W	2.74 (0.33)	4.11 (0.65)	0.74 (0.11)
PCR	2.68 (0.32)	4.02 (0.62)	0.73 (0.11)

To further characterize the robustness of various approaches, we compute outlier detection rates on the test set, and draw the *Receiver Operating Characteristic (ROC)* curves obtained from varying the threshold values in the outlier detection rule. Note that in this case both the training and test datasets contain outliers. The response of outliers is generated as

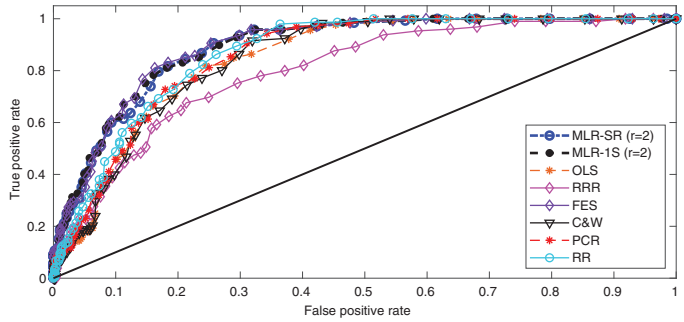
$$\mathbf{y} = (\mathbf{B}^*)'\mathbf{x} + \mathbf{o}_{\boldsymbol{\eta}},$$

where $\mathbf{o}_{\boldsymbol{\eta}} \sim \mathcal{N}(4 * \mathbf{1}_K, \boldsymbol{\Sigma}_{\mathbf{y}})$, with $\mathbf{1}_K$ the K-dimensional vector of all ones, and $\boldsymbol{\Sigma}_{\mathbf{y}} = (\sigma_{ij}^{\mathbf{y}})_{i,j\in[\![K]\!]}$, with $\sigma_{ij}^{\mathbf{y}} = (-0.9)^{|i-j|}$. The outlier detection criterion is described as follows:

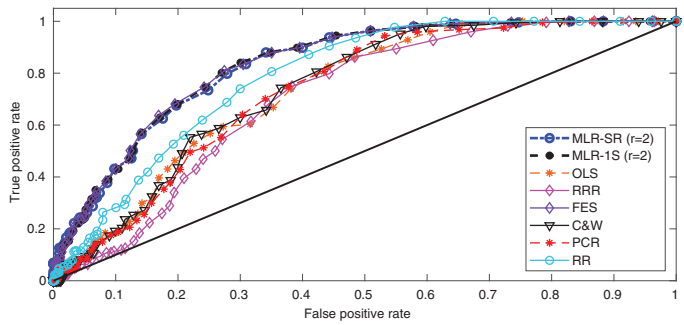
$$(\mathbf{x}_i, \mathbf{y}_i) = \begin{cases} \text{outlier}, & \text{if } \mathbf{r}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{r}_i \geq c, \\ \text{not an outlier}, & \text{otherwise}, \end{cases}$$

where $\mathbf{r}_i = \mathbf{y}_i - \hat{\mathbf{B}}'\mathbf{x}_i$ is the estimated residual, $\hat{\boldsymbol{\Sigma}}$ is the covariance matrix of the prediction error on the training set, and c is the threshold

The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) 20% outliers.



(b) 30% outliers.

Figure 6.3: The ROC curves of different MLR models.

value that is varied between 0 and $\chi^2_{0.99}(K)$ (0.99 percentile of the chi-square distribution with K degrees of freedom) to produce the ROC curves. Table 6.3 shows the average *Area Under the ROC Curve (AUC)* on the test set over 20 repetitions, and Figure 6.3 shows the ROC curves for different methods with 20% and 30% outliers, where the true positive rates and false positive rates are averaged over 20 repetitions. Compared to other methods except FES, the MLR-SR and MLR-1S models improve the AUC by 2%–11% when we have 20% outliers, and 6%–17% when we have 30% outliers, with a relatively small variability. Notice that FES also achieves a high AUC, but with a worse out-of-sample predictive performance.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

6.4.2 MLG Relaxations

In this subsection we study the performance of the two DRO-MLG relaxations, and compare them with a number of MLG variants on simulated datasets, in terms of their out-of-sample log-loss and classification accuracy.

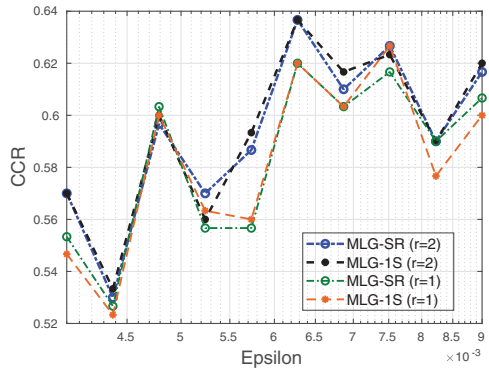
We first study the problem of selecting the right regularizer based on the distance metric in the data space. Similar to Section 6.4.1, we experiment with two types of $\mathbf{B}^*$, one coming from a multivariate normal distribution, and the other normalized to have an equal row absolute sum. They respectively correspond to an ℓ_2 and ℓ_1-norm distance metric in the data space. The predictor is drawn according to $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mathbf{x})$, where $\mathbf{\Sigma}_\mathbf{x} = (\sigma_{ij}^\mathbf{x})_{i,j\in[\![p]\!]}$, and $\sigma_{ij}^\mathbf{x} = 0.9^{|i-j|}$. The label vector $\mathbf{y} \in \{0,1\}^K$ is generated from a multinomial distribution with probabilities specified by the softmax normalization of $(\mathbf{B}^*)'\mathbf{x} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$.

We set $p = 5, K = 3$, and conduct 20 simulation runs, each with a training size of 100 and a test size of 60. The performance metrics we use include: (i) the average log-loss, (ii) the *Correct Classification Rate (CCR)*, and (iii) the *Conditional Value at Risk (CVaR)* (at the confidence level 0.8) of log-loss, which computes the expectation of extreme log-loss values. The average performance metrics on the test set over 20 replications are reported.

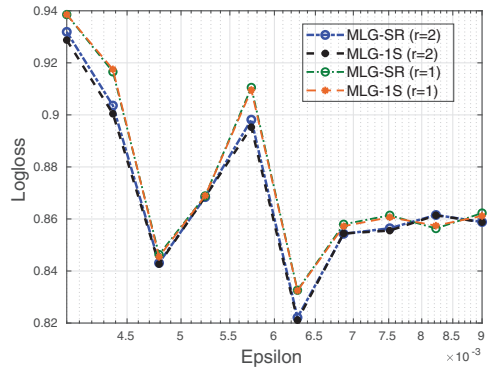
Figures 6.4 and 6.5 show the comparison of the four models as the Wasserstein radius ϵ is varied. We see that when $\mathbf{B}^*$ is a dense matrix, the MLG-SR and MLG-1S induced by the ℓ_2-norm have a higher classification accuracy and a lower log-loss. By contrast, when the structure of $\mathbf{B}^*$ implies an ℓ_1-norm distance metric in the data space, the MLG-SR and MLG-1S with $r = 1$ perform better. We also validate this conclusion in Tables 6.4 and 6.5 where the optimal Wasserstein set radius ϵ is chosen through cross-validation. With normally distributed $\mathbf{B}^*$, the formulations with $r = 2$ improve the CCR and log-loss by 5% compared to the ones induced by $r = 1$.

Next we will compare with a number of MLG models, including: (i) Vanilla MLG which minimizes the empirical log-loss with no penalty term, and (ii) Ridge MLG which penalizes the trace of $\mathbf{B}'\mathbf{B}$ (the squared Frobenius norm of $\mathbf{B}$) as in ridge regression [146], [147]. In addition

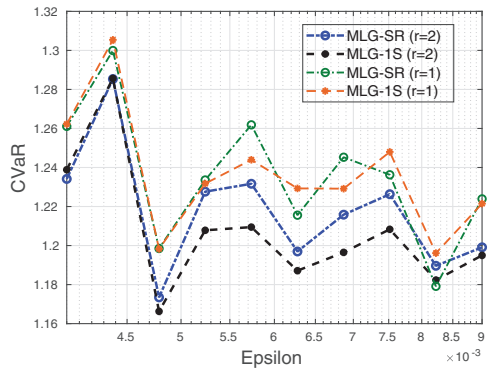
The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) CCR.



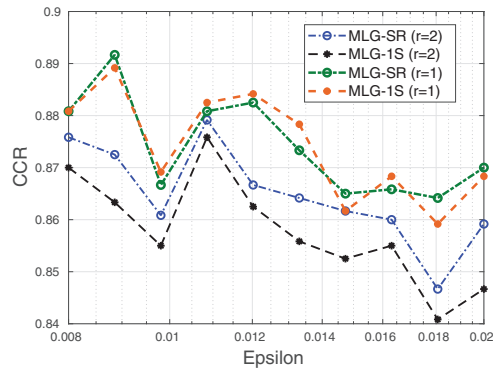
(b) Log-loss.



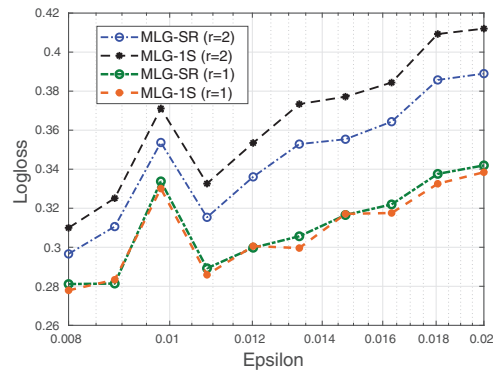
(c) CVaR of log-loss.

Figure 6.4: The out-of-sample performance of MLG-SR and MLG-1S when $\mathbf{B}^*$ is normally distributed.

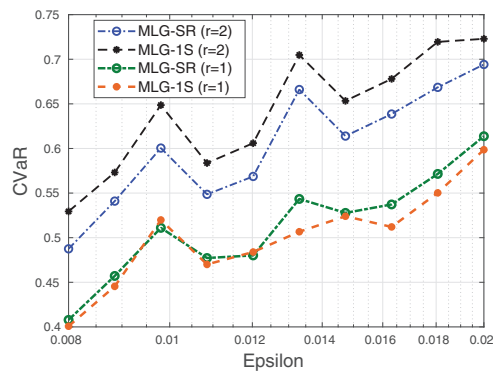
The version of record is available at: http://dx.doi.org/10.1561/2400000026



(a) CCR.



(b) Log-loss.



(c) CVaR of log-loss.

Figure 6.5: The out-of-sample performance of MLG-SR and MLG-1S when $\mathbf{B}^*$ has an equal row absolute sum.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Table 6.4: The out-of-sample performance of MLG-SR and MLG-1S with cross-validated ϵ when $\mathbf{B}^*$ is normally distributed

	CCR	**Log-Loss**	**CVaR**
MLG-SR ($r = 1$)	0.73 (0.06)	0.59 (0.09)	1.17 (0.24)
MLG-1S ($r = 1$)	0.75 (0.03)	0.60 (0.10)	1.16 (0.19)
MLG-SR ($r = 2$)	0.76 (0.05)	0.57 (0.08)	1.16 (0.21)
MLG-1S ($r = 2$)	0.76 (0.04)	0.57 (0.09)	1.11 (0.13)

Table 6.5: The out-of-sample performance of MLG-SR and MLG-1S with cross-validated ϵ when $\mathbf{B}^*$ has an equal row absolute sum

	CCR	**Log-Loss**	**CVaR**
MLG-SR ($r = 1$)	0.84 (0.04)	0.35 (0.05)	0.54 (0.18)
MLG-1S ($r = 1$)	0.84 (0.04)	0.34 (0.07)	0.55 (0.26)
MLG-SR ($r = 2$)	0.82 (0.05)	0.35 (0.12)	0.58 (0.32)
MLG-1S ($r = 2$)	0.83 (0.05)	0.35 (0.06)	0.58 (0.25)

to the three performance metrics used earlier, we introduce another robustness measure that calculates the minimal perturbation needed to fool the classifier. For a given $\mathbf{x}$ with label k, for any $j \neq k$, consider the following optimization problem:

$$\begin{aligned} \min_{\tilde{\mathbf{x}}} \quad & \|\mathbf{x} - \tilde{\mathbf{x}}\|_1 \\ \text{s.t.} \quad & P_j(\tilde{\mathbf{x}}) \geq P_k(\tilde{\mathbf{x}}), \\ & k = \arg\max_i P_i(\mathbf{x}), \end{aligned} \tag{6.15}$$

where $P_i(\mathbf{x})$ denotes the probability of assigning class label i to $\mathbf{x}$, which is a function of the trained classifier. Problem (6.15) measures the minimal perturbation distance (in terms of the ℓ_1-norm) that is needed to change the label of $\mathbf{x}$. Its optimal value evaluates the robustness of a given classifier in terms of the perturbation magnitude. The more robust the classifier, the larger the required perturbation to switch the label, and thus the larger the optimal value. We solve Problem (6.15) for every test point $\mathbf{x}$ and any $j \neq k$, and take the minimum of the optimal values to be the *Minimal Perturbation Distance (MPD)* of the classifier.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

We experiment with two types of $\mathbf{B}^*$:

1. **type-1**: each column of $\mathbf{B}^*$, with probability 0.4, is generated from a multivariate normal distribution, and with probability 0.6, it is generated as a sparse vector where only one randomly picked element is set to nonzero;

2. **type-2**: we first generate a $\mathbf{B}^*$ that is normalized to have an equal row absolute sum as before, and then set each column to a sparse vector with probability 0.6.

To test the robustness of various methods, we inject outliers to the training datasets. The predictors of the outliers have the same distribution as the clean samples, but their label vector is generated from a multinomial distribution with probabilities specified by the softmax normalization of $1/(\mathbf{x}'\mathbf{B}^* + \boldsymbol{\eta})$, where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$. The test set does not contain any outlier.

Tables 6.6 and 6.7 show the average performance of various models over 20 repetitions under different experimental settings. For type-1 $\mathbf{B}^*$, the MLG-1S ($r = 2$) achieves the highest classification accuracy and the largest MPD, while for type-2 $\mathbf{B}^*$, the MLG-1S ($r = 1$) excels. Notice that in both cases, the DRO-MLG models lose to vanilla MLG in terms of the log-loss. The reason is that vanilla MLG focuses solely on minimizing the sample average log-loss, while the MLG-SR and MLG-1S models balance between maintaining a low log-loss and achieving a high robustness level. By allowing for a slightly larger log-loss, the DRO-MLG models achieve a considerably higher MPD and a higher classification

Table 6.6: The out-of-sample performance of different MLG models trained on datasets with 30% outliers and type-1 $\mathbf{B}^*$

	CCR	**Log-Loss**	**CVaR**	**MPD**
MLG-SR ($r = 1$)	0.81 (0.05)	0.65 (0.09)	0.97 (0.10)	0.04 (0.15)
MLG-1S ($r = 1$)	0.81 (0.05)	0.65 (0.08)	0.99 (0.10)	0.04 (0.13)
MLG-SR ($r = 2$)	0.81 (0.05)	0.64 (0.04)	0.94 (0.11)	0.07 (0.02)
MLG-1S ($r = 2$)	0.82 (0.05)	0.66 (0.07)	0.95 (0.14)	0.10 (0.03)
Vanilla MLG	0.78 (0.07)	0.61 (0.08)	0.94 (0.17)	0.02 (0.01)
Ridge MLG	0.81 (0.04)	0.68 (0.08)	0.95 (0.13)	0.06 (0.05)

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Table 6.7: The out-of-sample performance of different MLG models trained on datasets with 40% outliers and type-2 $\mathbf{B}^*$

	CCR	Log-Loss	CVaR	MPD
MLG-SR ($r = 1$)	0.69 (0.14)	0.80 (0.06)	1.10 (0.17)	0.01 (0.007)
MLG-1S ($r = 1$)	0.69 (0.14)	0.82 (0.06)	1.11 (0.17)	0.03 (0.02)
MLG-SR ($r = 2$)	0.66 (0.13)	0.84 (0.05)	1.08 (0.12)	0.02 (0.01)
MLG-1S ($r = 2$)	0.68 (0.13)	0.82 (0.05)	1.10 (0.13)	0.02 (0.02)
Vanilla MLG	0.66 (0.06)	0.79 (0.06)	1.13 (0.14)	0.004 (0.002)
Ridge MLG	0.65 (0.09)	0.84 (0.04)	1.10 (0.09)	0.02 (0.01)

accuracy. They improve the CCR by 1%–5% and 5%–6%, the MPD by 67%–400% and 50%–650% in Tables 6.6 and 6.7, respectively. Compared to ridge MLG, they improve the log-loss by 6% and 5% in the two tables.

6.5 Summary

In this section, we developed a *Distributionally Robust Optimization (DRO)* based approach under the Wasserstein metric to robustify *Multi-output Linear Regression (MLR)* and *Multiclass Logistic Regression (MLG)*, leading to matrix-norm regularized formulations that establish a connection between robustness and regularization in the multi-output scenario. We established out-of-sample performance guarantees for the solutions to the DRO-MLR and DRO-MLG extracts, illustrating the role of the regularizer in controlling the out-of-sample prediction error. We provided empirical evidence showing that the DRO-MLR and DRO-MLG models achieve a comparable (slightly better) out-of-sample predictive performance to others, but a significantly higher robustness to outliers.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

7

Optimal Decision Making via Regression Informed K-NN

In this section, we will develop a prediction-based prescriptive model for optimal decision making that (i) predicts the outcome under each possible action using a robust nonlinear model, and (ii) adopts a randomized prescriptive policy determined by the predicted outcomes. The predictive model combines the Wasserstein DRO regression with the *K-Nearest Neighbors (K-NN)* regression that helps to capture the nonlinearity embedded in the data. We apply the proposed methodology in making recommendations for medical prescriptions, using a diabetes and a hypertension dataset extracted from the Electronic Health Records (EHRs) of a major safety-net hospital in New England.

7.1 The Problem and Related Work

Suppose we are given a set of M actions, and our goal is to choose $m \in [\![M]\!]$ such that the future outcome y is optimized. We are interested in finding the optimal decision with the aid of auxiliary data $\mathbf{x} \in \mathbb{R}^p$ that is concurrently observed, and correlated with the uncertain outcome y. A main challenge with learning from observational data lies in the lack of counterfactual information. One solution is to predict the effects of counterfactual policies by learning an action-dependent *predictive*

The version of record is available at: http://dx.doi.org/10.1561/2400000026

model that groups the training samples based on their actions, and fits a model in each group between the outcome y and the feature $\mathbf{x}$. The predictions from this composite model can be used to determine the optimal action to take. The performance of the prescribed decision hinges on the quality of the predictive model. We have observed that (i) there is often significant "noise" in the data caused by recording errors, missing values, and large variability across individuals, and (ii) the underlying relationship we try to learn is usually nonlinear and its parametric form is not known a priori. To deal with these issues, a nonparametric robust learning procedure is in need.

Motivated by the observation that individuals with similar features $\mathbf{x}$ would have similar outcomes y if they were to take the same action, we propose a predictive model that makes predictions based on the outcomes of similar individuals – to be called *neighbors* – in each group of the training set. It is a nonlinear and nonparametric estimator which constructs locally linear (constant) curves based on the similarity between individuals. To find reasonable neighbors, we need to accurately identify the set of features that are correlated with the outcome. We use the Wasserstein DRO regression for this task in consideration of the noise that could potentially bias the estimation. Our prescriptive methodology is established on the basis of a regression informed *K-Nearest Neighbors (K-NN)* model [161] that evaluates the importance of features through Wasserstein DRO regression, and estimates the outcome by averaging over the neighbors identified by a regression coefficients-weighted distance metric.

Our framework uses both parametric (Wasserstein DRO regression) and nonparametric (K-NN) predictive models, producing robust predictions immunized against significant noise and capturing the underlying nonlinearity by utilizing the information of neighbors. It is more information-efficient and more interpretable than the vanilla K-NN. We then develop a randomized prescriptive policy that chooses each action m with probability $e^{-\xi \hat{y}_m(\mathbf{x})} / \sum_{j=1}^{M} e^{-\xi \hat{y}_j(\mathbf{x})}$, for some pre-specified positive constant ξ, where $\hat{y}_m(\mathbf{x})$ is the predicted future outcome for $\mathbf{x}$ under action $m \in [\![M]\!]$. We show that this randomized strategy leads to a nearly optimal future outcome by an appropriate choice of ξ.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

In recent years there has been an emerging interest in combining ideas from machine learning with operations research to develop a framework that uses data to prescribe optimal decisions [162]–[164]. Current research focus has been on applying machine learning methodologies to predict the counterfactuals, based on which optimal decisions can be made. Local learning methods such as K-Nearest Neighbors [161], LOESS (LOcally Estimated Scatterplot Smoothing) [165], CART (Classification And Regression Trees) [1], and Random Forests [2], have been studied in [162], [166]–[169]. Extensions to continuous and multi-dimensional decision spaces with observational data were considered in [170]. To prevent overfitting, [171] proposed two robust prescriptive methods based on Nadaraya-Watson and nearest-neighbors learning. Deviating from such a predict-optimize paradigm, [172] presented a new bandit algorithm based on the LASSO to learn a model of decision rewards conditional on individual-specific covariates.

Our problem is closely related to contextual bandits [173]–[176], where an agent learns a sequence of decisions conditional on the contexts with the aim of maximizing its cumulative reward. It has recently found applications in learning personalized treatment of long-term diseases from mobile health data [177]–[179]. However, we learn the interaction between the context and rewards in each action group across similar individuals, not over the history of the same individual as in contextual bandits. A contextual bandits framework is most suitable for learning sequential strategies through repeated interactions with the environment, which requires a substantial amount of historical data for exploring the reward function and exploiting the promising actions. In contrast, our method does not require the availability of historical data, but instead learns the payoff function from similar individuals. This can be viewed as a different type of exploration, i.e., when little information can be acquired for the past states of an individual, investigating the behavior of similar subjects may be beneficial. This is essential for learning from the Electronic Health Records (EHRs), where rapid and continuous collection of patient data is not possible. We may observe a very short treatment history for some patients, and the lag between patient visits is usually large.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Our method is similar to K-NN regression with an OLS-weighted metric used in [166] to learn the optimal treatment for type-2 diabetic patients. The key differences lie in that: (i) we adopt a robustified regression procedure that is immunized against outliers and is thus more stable and reliable; (ii) we propose a randomized prescriptive policy that adds robustness to the methodology whereas [166] deterministically prescribed the treatment with the best predicted outcome; (iii) we establish theoretical guarantees on the quality of the predictions and the prescribed actions, and (iv) the prescriptive rule in [166] was activated when the improvement of the recommended treatment over the standard of care exceeded a certain threshold, whereas our method looks into the improvement over the previous regimen. This distinction makes our algorithm applicable in the scenario where the standard of care is unknown or ambiguous. Further, we derive a closed-form expression for the threshold level, which greatly improves the computational efficiency compared to [166], where a threshold was selected by cross-validation.

The remainder of this section is organized as follows. In Section 7.2, we introduce the robust nonlinear predictive model and present the performance guarantees on its predictive power. Section 7.3 develops the randomized prescriptive policy and proves its optimality in terms of the expected true outcome. The numerical experimental results are presented in Section 7.4. We conclude in Section 7.5.

7.2 Robust Nonlinear Predictive Model

Given a feature vector $\mathbf{x} \in \mathbb{R}^p$, and a set of M available actions, our goal is to predict the future outcome $y_m(\mathbf{x})$ under each possible action $m \in [\![M]\!]$. Assume the following relationship between the feature and outcome:

$$y_m = \mathbf{x}_m' \boldsymbol{\beta}_m^* + h_m(\mathbf{x}_m) + \epsilon_m,$$

where $(\mathbf{x}_m, y_m)$ represents the feature-outcome pair of an individual taking action m; $\boldsymbol{\beta}_m^*$ is the coefficient that captures the linear trend; $h_m(\cdot)$ is a nonlinear function (whose form is unknown) describing the nonlinear fluctuation in y_m, and ϵ_m is the noise term with zero mean and standard deviation η_m that expresses the intrinsic randomness of y_m and is assumed to be independent of $\mathbf{x}_m$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Suppose for each $m \in [\![M]\!]$, we observe N_m independently and identically distributed (i.i.d.) training samples $(\mathbf{x}_{mi}, y_{mi}), i \in [\![N_m]\!]$, that take action m. To estimate $\boldsymbol{\beta}_m^*$, we adopt the ℓ_2-norm induced Wasserstein DRO formulation. A robust model could lead to an improved out-of-sample performance, and accommodate the potential nonlinearity that is not explicitly revealed by the linear coefficient $\boldsymbol{\beta}_m^*$, thus, resulting in a more accurate assessment of the features. Solving the Wasserstein DRO regression model gives us a robust estimator of the linear regression coefficient $\boldsymbol{\beta}_m^*$, which we denote by $\hat{\boldsymbol{\beta}}_m \triangleq (\hat{\beta}_{m1}, \ldots, \hat{\beta}_{mp})$. The elements of $\hat{\boldsymbol{\beta}}_m$ measure the relative significance of the predictors in determining the outcome y_m. We feed the estimator into the nonlinear non-parametric K-NN regression model, by considering the following $\hat{\boldsymbol{\beta}}_m$-weighted metric:

$$\|\mathbf{x} - \mathbf{x}_{mi}\|_{\hat{\mathbf{W}}_m} = \sqrt{(\mathbf{x} - \mathbf{x}_{mi})' \hat{\mathbf{W}}_m (\mathbf{x} - \mathbf{x}_{mi})}, \tag{7.1}$$

where $\hat{\mathbf{W}}_m = \operatorname{diag}(\hat{\boldsymbol{\beta}}_m^2)$, and $\hat{\boldsymbol{\beta}}_m^2 = (\hat{\beta}_{m1}^2, \ldots, \hat{\beta}_{mp}^2)$. For a new test sample $\mathbf{x}$, within each action group m, we find its K_m nearest neighbors using the weighted distance function (7.1). The predicted future outcome for $\mathbf{x}$ under action m, denoted by $\hat{y}_m(\mathbf{x})$, is computed by

$$\hat{y}_m(\mathbf{x}) = \frac{1}{K_m} \sum_{i=1}^{K_m} y_{m(i)}, \tag{7.2}$$

where $y_{m(i)}$ is the outcome of the i-th closest individual to $\mathbf{x}$ in the training set who takes action m. In essence, we compute a K-NN estimate of the future outcome by using the regression weighted distance function, which can be viewed as a locally smoothed estimator in the neighborhood of $\mathbf{x}$. Notice that due to (7.1), the nearest neighbors are similar to $\mathbf{x}$ in the features that are most predictive of the outcome. Therefore, their corresponding response values should serve as a good approximation for the future outcome of $\mathbf{x}$.

We next show that (7.2) provides a good prediction in the sense of *Mean Squared Error (MSE)*. The bias-variance decomposition implies

The version of record is available at: http://dx.doi.org/10.1561/2400000026

the following:

$$
\begin{aligned}
&\mathrm{MSE}(\hat{y}_m(\mathbf{x})|\mathbf{x}, \mathbf{x}_{mi}, i \in [\![N_m]\!]) \\
&\quad \triangleq \mathbb{E}[(\hat{y}_m(\mathbf{x}) - y_m(\mathbf{x}))^2 | \mathbf{x}, \mathbf{x}_{mi}, i \in [\![N_m]\!]] \\
&\quad = \mathbb{E}\Bigg[\Bigg(\frac{1}{K_m}\sum_{j=1}^{K_m}(\mathbf{x}'_{m(j)}\boldsymbol{\beta}^*_m + h_m(\mathbf{x}_{m(j)}) + \epsilon_{m(j)}) \\
&\qquad - (\mathbf{x}'\boldsymbol{\beta}^*_m + h_m(\mathbf{x}) + \epsilon_m)\Bigg)^2 \Bigg| \mathbf{x}, \mathbf{x}_{mi}, i \in [\![N_m]\!]\Bigg] \\
&\quad = \Bigg(\mathbf{x}'\boldsymbol{\beta}^*_m + h_m(\mathbf{x}) - \frac{1}{K_m}\sum_{i=1}^{K_m}(\mathbf{x}'_{m(i)}\boldsymbol{\beta}^*_m + h_m(\mathbf{x}_{m(i)}))\Bigg)^2 + \frac{\eta_m^2}{K_m} + \eta_m^2 \\
&\quad = \Bigg(\frac{1}{K_m}\sum_{i=1}^{K_m}((\mathbf{x} - \mathbf{x}_{m(i)})'\boldsymbol{\beta}^*_m + h_m(\mathbf{x}) - h_m(\mathbf{x}_{m(i)}))\Bigg)^2 + \frac{\eta_m^2}{K_m} + \eta_m^2,
\end{aligned}
\tag{7.3}
$$

where $y_m(\mathbf{x})$ is the *true* future outcome for $\mathbf{x}$ under action m, and $\mathbf{x}_{m(i)}$, $\epsilon_{m(i)}$ are the feature vector and noise term corresponding to the i-th closest sample to $\mathbf{x}$ within group m, respectively. For each $m \in [\![M]\!]$, we aim at providing a probabilistic bound for (7.3) w.r.t. the measure of the N_m training samples. According to (7.3), for the MSE to be small the following three conditions suffice:

1. $\|\boldsymbol{\beta}^*_m - \hat{\boldsymbol{\beta}}_m\|_2$ is small;
2. $\|\mathbf{x} - \mathbf{x}_{m(i)}\|_{\hat{\mathbf{W}}_m}$ is small for $i \in [\![K_m]\!]$;
3. $h_m(\mathbf{x}) - h_m(\mathbf{x}_{m(i)})$ is small for $i \in [\![K_m]\!]$.

In other words, to ensure an accurate prediction of the outcome, we require an accurate estimate of the linear trend and a smooth nonlinear fluctuation, with the selected neighbors close enough to $\mathbf{x}$. An upper bound for the MSE follows from bounding these three quantities. We note that Theorem 4.3.12 provides an upper bound on the estimation bias $\|\boldsymbol{\beta}^*_m - \hat{\boldsymbol{\beta}}_m\|_2$ in a linear model. If we view the nonlinear term $h_m(\mathbf{x}_m)$ as part of the noise, then the bound provided by Theorem 4.3.12 applies to our case. The increased variance of the noise (due to $h_m(\mathbf{x}_m)$) is reflected in the eigenvalues of the covariance matrix, which play a

The version of record is available at: http://dx.doi.org/10.1561/2400000026

role in the estimation error bound. We provide a simplified version of Theorem 4.3.12 in the following theorem.

Theorem 7.2.1. Under Assumptions I–N, when the sample size $N_m \geq n_m$, with probability at least δ_m,

$$\|\boldsymbol{\beta}_m^* - \hat{\boldsymbol{\beta}}_m\|_2 \leq \tau_m.$$

We next show that the distance between $\mathbf{x}$ and its K_m nearest neighbors $\mathbf{x}_{m(i)}$ could be upper bounded probabilistically. All predictors are assumed to be centered, and independent from each other. In Theorem 7.2.2 we present a lower bound for $\mathbb{P}(\|\mathbf{x} - \mathbf{x}_{m(i)}\|_{\mathbf{W}} \leq \bar{w}_m, i \in [\![K_m]\!])$, for any positive definite diagonal matrix $\mathbf{W}$.

Theorem 7.2.2. Suppose we are given N_m i.i.d. samples $(\mathbf{x}_{mi}, y_{mi})$, $i \in [\![N_m]\!]$, drawn from some unknown probability distribution with finite fourth moment. Every $\mathbf{x}_{mi}$ has independent, centered coordinates:

$$\mathbb{E}(\mathbf{x}_{mi}) = \mathbf{0}, \quad \operatorname{cov}(\mathbf{x}_{mi}) = \operatorname{diag}(\sigma_{m1}^2, \ldots, \sigma_{mp}^2), \quad \forall i \in [\![N_m]\!].$$

For a fixed predictor $\mathbf{x}$, and for any given positive definite diagonal matrix $\mathbf{W} \in \mathbb{R}^{p\times p}$ with diagonal elements w_j, $j \in [\![p]\!]$, and $|w_j| \leq \bar{B}^2$, suppose:

$$|(x_{mij} - x_j)^2 - (\sigma_{mj}^2 + x_j^2)| \leq T_m, \quad \text{a.s.}, \ \forall i \in [\![N_m]\!], \ j \in [\![p]\!],$$

where x_{mij}, x_j are the j-th components of $\mathbf{x}_{mi}$ and $\mathbf{x}$, respectively. Under the condition that $\bar{w}_m^2 > \bar{B}^2 \sum_{j=1}^p (\sigma_{mj}^2 + x_j^2)$, with probability at least $1 - I_{1-p_{m0}}(N_m - K_m + 1, K_m)$,

$$\|\mathbf{x} - \mathbf{x}_{m(i)}\|_{\mathbf{W}} \leq \bar{w}_m, \quad i \in [\![K_m]\!],$$

where $g(u) = (1+u)\log(1+u) - u$,

$$I_{1-p_{m0}}(N_m - K_m + 1, K_m) \triangleq (N_m - K_m + 1)\binom{N_m}{K_m - 1}\int_0^{1-p_{m0}} t^{N_m - K_m}(1-t)^{K_m - 1}dt,$$

$$p_{m0} = 1 - \exp\left(-\frac{\sigma_m^2}{T_m^2} g\left(\frac{T_m(\bar{w}_m^2/\bar{B}^2 - \sum_j(\sigma_{mj}^2 + x_j^2))}{\sigma_m^2}\right)\right),$$

and,

$$\sigma_m^2 = \sum_{j=1}^p \operatorname{var}((x_{mij} - x_j)^2).$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Proof. To simplify the notation, we will omit the subscript m in all proofs, e.g., using $\mathbf{x}_i$ and $\mathbf{x}_{(i)}$ for $\mathbf{x}_{mi}$ and $\mathbf{x}_{m(i)}$, respectively, and N for N_m. Define the event $\mathcal{A}_i := \{\|\mathbf{x}_i - \mathbf{x}\|_{\bar{B}^2\mathbf{I}} \leq \bar{w}\}$. As long as we can calculate the probability that at least K of $\mathcal{A}_i$, $i \in [\![N]\!]$, occur, we are able to provide a lower bound on $\mathbb{P}(\|\mathbf{x} - \mathbf{x}_{(i)}\|_{\mathbf{W}} \leq \bar{w},\ i \in [\![K]\!])$. Note that given $\mathbf{x}$, $\mathcal{A}_i$, $i \in [\![N]\!]$, are independent and equiprobable, since $\mathbf{x}_i$, $i \in [\![N]\!]$, are i.i.d. Based on Bennett's inequality [103], we have:

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_i) &= \mathbb{P}(\|\mathbf{x}_i - \mathbf{x}\|^2_{\bar{B}^2\mathbf{I}} \leq \bar{w}^2) \\
&= \mathbb{P}(\bar{B}^2(x_{i1} - x_1)^2 + \cdots + \bar{B}^2(x_{ip} - x_p)^2 \leq \bar{w}^2) \\
&= \mathbb{P}(t_1 + \cdots + t_p \leq \bar{w}^2/\bar{B}^2) \\
&= \mathbb{P}\bigg(\sum_j (t_j - (\sigma_j^2 + x_j^2)) \leq \bar{w}^2/\bar{B}^2 - \sum_j (\sigma_j^2 + x_j^2)\bigg) \\
&\geq 1 - \exp\bigg(-\frac{\sigma^2}{T^2} g\bigg(\frac{T(\bar{w}^2/\bar{B}^2 - \sum_j (\sigma_j^2 + x_j^2))}{\sigma^2}\bigg)\bigg) \\
&\triangleq p_0,
\end{aligned}$$

where $t_j = (x_{ij} - x_j)^2$, $j \in [\![p]\!]$; $\sigma^2 = \sum_j \mathrm{var}(t_j)$. In the above derivation, we used the fact that t_j, $j \in [\![p]\!]$, are independent, and $|t_j - \mathbb{E}[t_j]| \leq T$, a.s., $\forall j$.

Given the lower bound for $\mathbb{P}(\mathcal{A}_i)$, we can derive a lower bound for the probability that exactly K of $\mathcal{A}_i$, $i \in [\![N]\!]$, occur. For a given $\mathbf{x}$, $\mathcal{A}_i$, $i \in [\![N]\!]$, are independent, and thus,

$$\begin{aligned}
&\mathbb{P}(\|\mathbf{x} - \mathbf{x}_{(i)}\|_{\mathbf{W}} \leq \bar{w}, \quad i \in [\![K]\!]) \\
&\quad \geq \mathbb{P}(\text{at least } K \text{ of } \mathcal{A}_i, i \in [\![N]\!] \text{ occur}) \\
&\quad = \sum_{k=K}^{N} \binom{N}{k} (\mathbb{P}(\mathcal{A}_i))^k (1 - \mathbb{P}(\mathcal{A}_i))^{N-k} \\
&\quad \geq \sum_{k=K}^{N} \binom{N}{k} p_0^k (1 - p_0)^{N-k} \\
&\quad = 1 - I_{1-p_0}(N - K + 1, K),
\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where $I_{1-p_0}(N-K+1,K)$ is the *regularized incomplete beta function* defined as:

$$I_{1-p_0}(N-K+1,K) \triangleq (N-K+1)\binom{N}{K-1}\int_0^{1-p_0} t^{N-K}(1-t)^{K-1}dt.$$

□

Note that p_{m0} is nonnegative, due to the assumption that $\bar{w}_m^2 > \bar{B}^2\sum_j(\sigma_{mj}^2+x_j^2)$, and the non-decreasing property of the function $g(\cdot)$ when its argument is non-negative. We also note that a lower bound on $\mathbb{P}(\|\mathbf{x}-\mathbf{x}_{m(i)}\|_2 \le \bar{w}_m,\ i \in [\![K_m]\!])$ can be obtained by setting $\bar{B}=1$.

By now we have shown results on the accuracy of $\hat{\boldsymbol{\beta}}_m$ and the similarity between $\mathbf{x}$ and its neighbors. Notice that for a Lipschitz continuous function $h_m(\cdot)$ with a Lipschitz constant L_m, the difference between $h_m(\mathbf{x})$ and $h_m(\mathbf{x}_{m(i)})$ can be bounded by $L_m\|\mathbf{x}-\mathbf{x}_{m(i)}\|_2$. With these results we are ready to bound the MSE of $\hat{y}_m(\mathbf{x})$.

Theorem 7.2.3. Suppose we are given N_m i.i.d. copies of $(\mathbf{x}_m, y_m)$, denoted by $(\mathbf{x}_{mi}, y_{mi}), i \in [\![N_m]\!]$, where $\mathbf{x}_m$ has independent, centered coordinates, and $\mathrm{cov}(\mathbf{x}_m) = \mathrm{diag}(\sigma_{m1}^2,\ldots,\sigma_{mp}^2)$. We are given a fixed predictor $\mathbf{x}=(x_1,\ldots,x_p)$, a scalar $\bar{w}_m$, and we assume:

1. $h_m(\cdot)$ is Lipschitz continuous with a Lipschitz constant L_m on the metric spaces $(\mathcal{X}_m, \|\cdot\|_2)$ and $(\mathcal{Y}_m, |\cdot|)$, where $\mathcal{X}_m, \mathcal{Y}_m$ are the domain and codomain of $h_m(\cdot)$, respectively;

2. $\bar{w}_m^2 > \bar{B}_m^2\sum_{j=1}^p(\sigma_{mj}^2+x_j^2)$, where $\bar{B}_m$ is the upper bound on $\|(-\boldsymbol{\beta}_m,1)\|_2$ for any feasible $\boldsymbol{\beta}_m$ to (4.5);

3. $|(x_{mij}-x_j)^2-(\sigma_{mj}^2+x_j^2)|$ is upper bounded a.s. under the probability measure $\mathbb{P}^*_{\mathcal{X}_m}$ for any i,j, where x_{mij} is the j-th component of $\mathbf{x}_{mi}$, and $\mathbb{P}^*_{\mathcal{X}_m}$ is the underlying true probability distribution of $\mathbf{x}_m$;

4. the coordinates of any feasible solution to (4.5) have absolute values greater than or equal to some positive number b_m (dense estimators).

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Under Assumptions I–N, when $N_m \geq n_m$, with probability at least $\delta_m - I_{1-p_{m0}}(N_m - K_m + 1, K_m)$ w.r.t. the measure of samples,

$$\begin{aligned}&\mathbb{E}[(\hat{y}_m(\mathbf{x}) - y_m(\mathbf{x}))^2 \,|\, \mathbf{x}, \mathbf{x}_{mi}, i \in [\![N_m]\!]] \\ &\quad \leq \left(\frac{\bar{w}_m \tau_m}{b_m} + \sqrt{p}\bar{w}_m + \frac{L_m \bar{w}_m}{\bar{B}_m}\right)^2 + \frac{\eta_m^2}{K_m} + \eta_m^2, \end{aligned} \tag{7.4}$$

and for any $a \geq (\bar{w}_m \tau_m / b_m + \sqrt{p}\bar{w}_m + L_m \bar{w}_m / \bar{B}_m)^2 + \eta_m^2 / K_m + \eta_m^2$,

$$\begin{aligned}&\mathbb{P}\Big((\hat{y}_m(\mathbf{x}) - y_m(\mathbf{x}))^2 \geq a \,|\, \mathbf{x}, \mathbf{x}_{mi}, i \in [\![N_m]\!]\Big) \\ &\quad \leq \frac{(\frac{\bar{w}_m \tau_m}{b_m} + \sqrt{p}\bar{w}_m + \frac{L_m \bar{w}_m}{\bar{B}_m})^2 + \frac{\eta_m^2}{K_m} + \eta_m^2}{a}, \end{aligned} \tag{7.5}$$

where all parameters are set in the same way as in Theorems 7.2.1 and 7.2.2.

Proof. We omit the subscript m for simplicity. By Theorems 7.2.1 and 7.2.2, we know that,

$$\begin{aligned}|(\mathbf{x} - \mathbf{x}_{(i)})'(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})| &= |(\mathbf{x} - \mathbf{x}_{(i)})' \hat{\mathbf{W}}^{\frac{1}{2}} \hat{\mathbf{W}}^{-\frac{1}{2}} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})| \\ &\leq \|(\mathbf{x} - \mathbf{x}_{(i)})' \hat{\mathbf{W}}^{\frac{1}{2}}\|_2 \|\hat{\mathbf{W}}^{-\frac{1}{2}} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})\|_2 \\ &\leq \frac{\bar{w}\tau}{b}, \end{aligned}$$

where the second inequality used the fact that $\|\hat{\mathbf{W}}^{-\frac{1}{2}} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})\|_2 \leq \tau / b$ if $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 \leq \tau$, which can be verified by the Courant-Fischer Theorem, and the fact that $\hat{\mathbf{W}}$ is diagonal with elements $\hat{\beta}_j^2, j \in [\![p]\!]$, and $|\hat{\beta}_j| \geq b$. Based on the inequality $(\sum_{i=1}^n a_i)^2 \leq n(\sum_{i=1}^n a_i^2)$, we know:

$$\begin{aligned}|(\mathbf{x} - \mathbf{x}_{(i)})' \hat{\boldsymbol{\beta}}| &= \left|\sum_{j=1}^p \hat{\beta}_j (\mathbf{x} - \mathbf{x}_{(i)})_j\right| \\ &\leq \sqrt{p \sum_{j=1}^p \left(\hat{\beta}_j (\mathbf{x} - \mathbf{x}_{(i)})_j\right)^2} \\ &= \sqrt{p(\mathbf{x} - \mathbf{x}_{(i)})' \hat{\mathbf{W}} (\mathbf{x} - \mathbf{x}_{(i)})} \\ &\leq \sqrt{p}\bar{w}. \end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Therefore,

$$\begin{aligned}|(\mathbf{x}-\mathbf{x}_{(i)})'\boldsymbol{\beta}^*| &= |(\mathbf{x}-\mathbf{x}_{(i)})'(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}})+(\mathbf{x}-\mathbf{x}_{(i)})'\hat{\boldsymbol{\beta}}| \\ &\le |(\mathbf{x}-\mathbf{x}_{(i)})'(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}})|+|(\mathbf{x}-\mathbf{x}_{(i)})'\hat{\boldsymbol{\beta}}| \\ &\le \frac{\bar{w}\tau}{b}+\sqrt{p}\bar{w}.\end{aligned}$$

Thus, for a given $\mathbf{x}$,

$$\begin{aligned}&\mathbb{E}[(\hat{y}(\mathbf{x})-y(\mathbf{x}))^2\,|\,\mathbf{x},\mathbf{x}_i] \\ &= \left(\frac{1}{K}\sum_{i=1}^{K}\left((\mathbf{x}-\mathbf{x}_{(i)})'\boldsymbol{\beta}^*+h(\mathbf{x})-h(\mathbf{x}_{(i)})\right)\right)^2+\frac{\eta^2}{K}+\eta^2 \\ &\le \left(\frac{1}{K}\sum_{i=1}^{K}\left(|(\mathbf{x}-\mathbf{x}_{(i)})'\boldsymbol{\beta}^*|+|h(\mathbf{x})-h(\mathbf{x}_{(i)})|\right)\right)^2+\frac{\eta^2}{K}+\eta^2 \\ &\le \left(\frac{\bar{w}\tau}{b}+\sqrt{p}\bar{w}+\frac{L\bar{w}}{\bar{B}}\right)^2+\frac{\eta^2}{K}+\eta^2.\end{aligned}$$

The probability bound can be easily derived using Markov's inequality. □

The expectation in (7.4) and the probability in (7.5) are taken w.r.t. the measure of the noise ϵ_m. Theorem 7.2.3 essentially says that for any given predictor $\mathbf{x}$, with a high probability (w.r.t. the measure of samples), the prediction from our model is close to the true future outcome. The prediction bias depends on the sample size, the variation in the predictors and response, and the smoothness of the nonlinear fluctuation.

The dependence on b_m in the upper bound provided by (7.4) is due to the fact that $\hat{\mathbf{W}}_m$ has diagonal elements $\hat{\beta}^2_{mj}, j \in [\![p]\!]$, which are assumed to be at least b_m^2. If we multiply $\hat{\mathbf{W}}_m$ by a very large number, the neighbor selection criterion is not affected, since the relative significance of the predictors stays unchanged, but the b_m appearing in (7.4) would be replaced by a very large number, diminishing the effect of the first term in the parentheses, at the price of increasing $\bar{B}_m$ and $\bar{w}_m$, which in turn have an effect on the number of neighbors needed. It might be interesting to explore this implicit trade-off and optimize $\hat{\mathbf{W}}_m$ to achieve the smallest MSE.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

7.3 Prescriptive Policy Development

We now proceed to develop the prescriptive policy with the aim of minimizing the future outcome. A natural idea is to take the action that yields the minimum predicted outcome. To allow for flexibility in exploring alternatives that have a comparable performance, and also to correct for potential prediction errors that might mislead the ranking of actions, we propose a randomized policy that prescribes each action with a probability inversely proportional to its exponentiated predicted outcome. It can be viewed as an offline Hedge algorithm [180] that increases the robustness of our method through exploration.

Specifically, given an individual with a feature vector $\mathbf{x}$, and her predicted future outcome under each action m, denoted by $\hat{y}_m(\mathbf{x})$, we consider a randomized policy that chooses action m with probability $e^{-\xi\hat{y}_m(\mathbf{x})}/\sum_{j=1}^{M} e^{-\xi\hat{y}_j(\mathbf{x})}$, with ξ some pre-specified positive constant. The randomness in making decisions might hurt the interpretability of the model. But on the other hand, it presents a range of comparable options that can be assessed subjectively by the decision maker based on her expertise. As ξ goes to infinity, the randomized policy will converge to a deterministic one which selects the action with the lowest predicted outcome. We next establish a related property of the randomized policy in terms of its expected *true* outcome.

Theorem 7.3.1. Given any fixed predictor $\mathbf{x} \in \mathbb{R}^p$, denote its predicted and true future outcome under action m by $\hat{y}_m(\mathbf{x})$ and $y_m(\mathbf{x})$, respectively. Assume that we adopt a randomized strategy that prescribes action m with probability $e^{-\xi\hat{y}_m(\mathbf{x})}/\sum_{j=1}^{M} e^{-\xi\hat{y}_j(\mathbf{x})}$, for some $\xi \geq 0$. Assume $\hat{y}_m(\mathbf{x})$ and $y_m(\mathbf{x})$ are non-negative, $\forall m \in [\![M]\!]$. The expected true outcome under this policy satisfies:

$$\sum_{m=1}^{M} \frac{e^{-\xi\hat{y}_m(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}} y_m(\mathbf{x}) \leq y_k(\mathbf{x}) + \left(\hat{y}_k(\mathbf{x}) - \frac{1}{M}\sum_{m=1}^{M} \hat{y}_m(\mathbf{x})\right) + \xi\left(\frac{1}{M}\sum_{m=1}^{M} \hat{y}_m^2(\mathbf{x}) + \sum_{m=1}^{M} \frac{e^{-\xi\hat{y}_m(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}} y_m^2(\mathbf{x})\right) + \frac{\log M}{\xi}, \quad (7.6)$$

for any $k \in [\![M]\!]$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Proof. The proof borrows ideas from Theorem 1.5 in [180]. Define $W_m \triangleq e^{-\xi \hat{y}_m(\mathbf{x})} / \sum_{j=1}^M e^{-\xi \hat{y}_j(\mathbf{x})}$, and $\phi \triangleq \sum_{m=1}^M e^{-\xi \hat{y}_m(\mathbf{x})} e^{-\xi y_m(\mathbf{x})}$. Then,

$$\begin{aligned}
\phi &= \left(\sum_{j=1}^M e^{-\xi \hat{y}_j(\mathbf{x})}\right) \sum_{m=1}^M W_m e^{-\xi y_m(\mathbf{x})} \\
&\leq \left(\sum_{j=1}^M e^{-\xi \hat{y}_j(\mathbf{x})}\right) \sum_{m=1}^M W_m (1 - \xi y_m(\mathbf{x}) + \xi^2 y_m^2(\mathbf{x})) \\
&= \left(\sum_{j=1}^M e^{-\xi \hat{y}_j(\mathbf{x})}\right) \left(1 - \xi \sum_{m=1}^M W_m y_m(\mathbf{x}) + \xi^2 \sum_{m=1}^M W_m y_m^2(\mathbf{x})\right) \\
&\leq \left(\sum_{j=1}^M e^{-\xi \hat{y}_j(\mathbf{x})}\right) e^{-\xi \sum_{m=1}^M W_m y_m(\mathbf{x}) + \xi^2 \sum_{m=1}^M W_m y_m^2(\mathbf{x})},
\end{aligned}$$

where the first inequality uses the fact that for $x \geq 0$, $e^{-x} \leq 1 - x + x^2$, and the last inequality is due to the fact that $1 + x \leq e^x$. Next let us examine the sum of exponentials:

$$\begin{aligned}
\sum_{j=1}^M e^{-\xi \hat{y}_j(\mathbf{x})} &\leq \sum_{j=1}^M (1 - \xi \hat{y}_j(\mathbf{x}) + \xi^2 \hat{y}_j^2(\mathbf{x})) \\
&= M\left(1 - \xi \frac{1}{M} \sum_{j=1}^M \hat{y}_j(\mathbf{x}) + \xi^2 \frac{1}{M} \sum_{j=1}^M \hat{y}_j^2(\mathbf{x})\right) \\
&\leq M e^{-\xi \frac{1}{M} \sum_{j=1}^M \hat{y}_j(\mathbf{x}) + \xi^2 \frac{1}{M} \sum_{j=1}^M \hat{y}_j^2(\mathbf{x})}.
\end{aligned}$$

On the other hand, for any $k \in [\![M]\!]$,

$$\begin{aligned}
e^{-\xi \hat{y}_k(\mathbf{x}) - \xi y_k(\mathbf{x})} \leq & \phi \\
\leq & M \exp\left\{ - \frac{\xi \sum_{j=1}^M \hat{y}_j(\mathbf{x})}{M} + \frac{\xi^2 \sum_{j=1}^M \hat{y}_j^2(\mathbf{x})}{M} \right. \\
& \left. - \xi \sum_{m=1}^M W_m y_m(\mathbf{x}) + \xi^2 \sum_{m=1}^M W_m y_m^2(\mathbf{x}) \right\}.
\end{aligned} \tag{7.7}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Taking the logarithm on both sides of (7.7) and dividing by ξ, we obtain

$$\frac{1}{M}\sum_{m=1}^{M}\hat{y}_m(\mathbf{x}) + \sum_{m=1}^{M}\frac{e^{-\xi\hat{y}_m(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}y_m(\mathbf{x}) \leq \hat{y}_k(\mathbf{x}) + y_k(\mathbf{x}) + \xi\left(\frac{1}{M}\sum_{m=1}^{M}\hat{y}_m^2(\mathbf{x}) + \sum_{m=1}^{M}\frac{e^{-\xi\hat{y}_m(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}y_m^2(\mathbf{x})\right) + \frac{\log M}{\xi}.$$

□

Theorem 7.3.1 says that the expected *true* outcome of the randomized policy is no worse than the true outcome of any action k plus two components, one accounting for the gap between the *predicted* outcome under k and the average predicted outcome, and the other depending on the parameter ξ. Thinking about choosing $k = \arg\min_m y_m(\mathbf{x})$, if $\hat{y}_k(\mathbf{x})$ is below the average predicted outcome (which should be true if we have an accurate prediction), it follows from (7.6) that the randomized policy leads to a nearly optimal future outcome by an appropriate choice of ξ.

In the medical applications, when determining the *future* prescription for a patient, we usually have access to some auxiliary information such as the *current* prescription that she is receiving, and her *current* lab results. In consideration of the health care costs and treatment transients, it is not desired to switch patients' treatments too frequently. We thus set a threshold level for the expected improvement in the outcome, below which the randomized strategy will be "frozen" and the current therapy will be continued. Specifically,

$$m_{\mathrm{f}}(\mathbf{x}) = \begin{cases} m, \text{ w.p. } \dfrac{e^{-\xi\hat{y}_m(\mathbf{x})}}{\sum_{j=1}^{M} e^{-\xi\hat{y}_j(\mathbf{x})}}, & \text{if } \sum_k \dfrac{e^{-\xi\hat{y}_k(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}\hat{y}_k(\mathbf{x}) \leq x_{\mathrm{co}} - T(\mathbf{x}), \\ m_{\mathrm{c}}(\mathbf{x}), & \text{otherwise,} \end{cases}$$

where $m_{\mathrm{f}}(\mathbf{x})$ and $m_{\mathrm{c}}(\mathbf{x})$ are the future and current prescriptions for patient $\mathbf{x}$, respectively; x_{co} represents the current observed outcome (e.g., current blood pressure), which is assumed to be one of the components of $\mathbf{x}$, and $T(\mathbf{x})$ is some threshold level which will be determined later. This prescriptive rule basically says that the randomized strategy will

The version of record is available at: http://dx.doi.org/10.1561/2400000026

be activated only if the expected improvement relative to the current observed outcome is significant.

Theorem 7.3.2. Assume that the distribution of the predicted outcome $\hat{y}_m(\mathbf{x})$ conditional on $\mathbf{x}$, is sub-Gaussian, and its ψ_2-norm is equal to $\sqrt{2}C_m(\mathbf{x})$, for any $m \in [\![M]\!]$ and any $\mathbf{x}$. Given a small $0 < \bar{\epsilon} < 1$, to satisfy

$$\mathbb{P}\left(\sum_k \frac{e^{-\xi\hat{y}_k(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}\hat{y}_k(\mathbf{x}) > x_{\text{co}} - T(\mathbf{x})\right) \leq \bar{\epsilon},$$

it suffices to set a threshold

$$T(\mathbf{x}) = \max\left(0,\ \min_m\left(x_{\text{co}} - \mu_{\hat{y}_m}(\mathbf{x}) - \sqrt{-2C_m^2(\mathbf{x})\log(\bar{\epsilon}/M)}\right)\right),$$

where $\mu_{\hat{y}_m}(\mathbf{x}) = \mathbb{E}[\hat{y}_m(\mathbf{x})|\mathbf{x}]$.

Proof. By the sub-Gaussian assumption we have:

$$\begin{aligned}
&\mathbb{P}\Big(\sum_k \frac{e^{-\xi\hat{y}_k(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}\hat{y}_k(\mathbf{x}) > x_{\text{co}} - T(\mathbf{x})\Big)\\
&\leq \mathbb{P}\Big(\max_k \hat{y}_k(\mathbf{x}) > x_{\text{co}} - T(\mathbf{x})\Big)\\
&= \mathbb{P}\Big(\bigcup_k \{\hat{y}_k(\mathbf{x}) > x_{\text{co}} - T(\mathbf{x})\}\Big)\\
&\leq \sum_k \mathbb{P}(\hat{y}_k(\mathbf{x}) > x_{\text{co}} - T(\mathbf{x}))\\
&\leq \sum_k \exp\left(-\frac{\left(x_{\text{co}} - T(\mathbf{x}) - \mu_{\hat{y}_k}(\mathbf{x})\right)^2}{2C_k^2(\mathbf{x})}\right). \qquad (7.8)
\end{aligned}$$

Note that the probability in (7.8) is taken with respect to the measure of the training samples. We essentially want to find the largest threshold $T(\mathbf{x})$ such that the probability of the expected improvement being less than $T(\mathbf{x})$ is small. Given a small $0 < \bar{\epsilon} < 1$ and due to (7.8), to satisfy

$$\mathbb{P}\left(\sum_k \frac{e^{-\xi\hat{y}_k(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}\hat{y}_k(\mathbf{x}) > x_{\text{co}} - T(\mathbf{x})\right) \leq \bar{\epsilon},$$

it suffices to set:

$$\sum_k \exp\left(-\frac{\left(x_{\text{co}} - T(\mathbf{x}) - \mu_{\hat{y}_k}(\mathbf{x})\right)^2}{2C_k^2(\mathbf{x})}\right) \leq \bar{\epsilon}. \qquad (7.9)$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

A sufficient condition for (7.9) is:

$$\exp\left(-\frac{\left(x_{\text{co}}-T(\mathbf{x})-\mu_{\hat{y}_m}(\mathbf{x})\right)^2}{2C_m^2(\mathbf{x})}\right)\leq\frac{\bar{\epsilon}}{M},\qquad\forall m\in[\![M]\!],$$

which yields that,

$$T(\mathbf{x})\leq x_{\text{co}}-\mu_{\hat{y}_m}(\mathbf{x})-\sqrt{-2C_m^2(\mathbf{x})\log(\bar{\epsilon}/M)},\qquad\forall m\in[\![M]\!].\quad(7.10)$$

Given that $T(\mathbf{x})$ is non-negative, we set the largest possible threshold satisfying (7.10) to:

$$T(\mathbf{x})=\max\left(0,\ \min_m(x_{\text{co}}-\mu_{\hat{y}_m}(\mathbf{x})-\sqrt{-2C_m^2(\mathbf{x})\log(\bar{\epsilon}/M)})\right).$$

When using a deterministic policy ($\xi\to\infty$), for any $m\in[\![M]\!]$, we have

$$\begin{aligned}\mathbb{P}\Big(\min_m\hat{y}_m(\mathbf{x})>x_{\text{co}}-T(\mathbf{x})\Big)&=\mathbb{P}\Big(\bigcap_m\{\hat{y}_m(\mathbf{x})>x_{\text{co}}-T(\mathbf{x})\}\Big)\\&\leq\mathbb{P}(\hat{y}_m(\mathbf{x})>x_{\text{co}}-T(\mathbf{x}))\\&\leq\exp\left(-\frac{\left(x_{\text{co}}-T(\mathbf{x})-\mu_{\hat{y}_m}(\mathbf{x})\right)^2}{2C_m^2(\mathbf{x})}\right).\end{aligned}$$

Similarly, to make

$$\mathbb{P}(\min_m\hat{y}_m(\mathbf{x})>x_{\text{co}}-T(\mathbf{x}))\leq\bar{\epsilon},$$

we set:

$$T(\mathbf{x})=\max\left(0,\ \min_m\left(x_{\text{co}}-\mu_{\hat{y}_m}(\mathbf{x})-\sqrt{-2C_m^2(\mathbf{x})\log\bar{\epsilon}}\right)\right),$$

which establishes the desired result. □

Theorem 7.3.2 finds the largest threshold $T(\mathbf{x})$ such that the probability of the expected improvement being less than $T(\mathbf{x})$ is small. The parameters $\mu_{\hat{y}_m}(\mathbf{x})$ and $C_m(\mathbf{x})$, for $m\in[\![M]\!]$, can be estimated by simulation through random sampling a subset of the training examples. Algorithm 1 provides the details.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Algorithm 1 Estimating the conditional mean and standard deviation of the predicted outcome.

Input: a feature vector $\mathbf{x}$; a_m: the number of subsamples used to compute $\hat{\boldsymbol{\beta}}_m$, $a_m < N_m$; d_m: the number of repetitions.
for $i = 1, \ldots, d_m$ **do**

Randomly pick a_m samples from group m, and use them to estimate a robust regression coefficient $\hat{\boldsymbol{\beta}}_{m_i}$ through solving (4.5).

The future outcome for $\mathbf{x}$ under action m is predicted as $\hat{y}_{m_i}(\mathbf{x}) = \mathbf{x}'\hat{\boldsymbol{\beta}}_{m_i}$.
end for
Output: Estimate the conditional mean of $\hat{y}_m(\mathbf{x})$ as:

$$\mu_{\hat{y}_m}(\mathbf{x}) = \frac{1}{d_m}\sum_{i=1}^{d_m}\hat{y}_{m_i}(\mathbf{x}),$$

and the conditional standard deviation as:

$$C_m(\mathbf{x}) = \sqrt{\frac{1}{d_m - 1}\sum_{i=1}^{d_m}(\hat{y}_{m_i}(\mathbf{x}) - \mu_{\hat{y}_m}(\mathbf{x}))^2}.$$

A Special Case As $\xi \to \infty$, the randomized policy will assign probability 1 to the action with the lowest predicted outcome, which is equivalent to the following deterministic policy:

$$m_{\mathrm{f}}(\mathbf{x}) = \begin{cases} \arg\min_m \hat{y}_m(\mathbf{x}), & \text{if } \min_m \hat{y}_m(\mathbf{x}) \le x_{\mathrm{co}} - T(\mathbf{x}), \\ m_{\mathrm{c}}(\mathbf{x}), & \text{otherwise.} \end{cases}$$

A slight modification to the threshold level $T(\mathbf{x})$ is given as follows:

$$T(\mathbf{x}) = \max\left(0,\ \min_m\left(x_{\mathrm{co}} - \mu_{\hat{y}_m}(\mathbf{x}) - \sqrt{-2C_m^2(\mathbf{x})\log\bar{\epsilon}}\right)\right).$$

7.4 Developing Optimal Prescriptions for Patients

In this section, we apply our method to develop optimal prescriptions for patients with type-2 diabetes and hypertension. The data used for the study come from the Boston Medical Center – the largest safety-net

The version of record is available at: http://dx.doi.org/10.1561/2400000026

hospital in New England – and consist of *Electronic Health Records (EHRs)* containing the patients' medical history in the period 1999–2014. The medical history of each patient includes demographics, diagnoses, prescriptions, lab tests, and past admission records. We build two datasets from the EHRs, one containing the medical records of patients with type-2 diabetes and the other for patients with hypertension. For diabetic patients, we want to determine the treatment (drug regimen) that leads to the lowest future HbA_{1c}[1] based on the medical histories, while for hypertension patients, our goal is to find the treatment that minimizes the future systolic blood pressure.[2]

7.4.1 Description of the Datasets

The patients that meet the following criteria are included in the diabetes dataset:

- patients present in the system for at least 1 year;
- received at least one blood glucose regulation agent, including injectable (e.g., insulin) and oral (e.g., metformin) drugs, etc., and had at least one medical record 100 days before this prescription;
- had at least three measurements of HbA_{1c} in the system; and,
- were not diagnosed with type-1 diabetes.

Similarly, for the hypertension dataset, the patients that meet the following criteria are included:

- patients present in the system for at least 1 year;
- received at least one type of cardiovascular medications, including ACE inhibitors, Angiotensin Receptor Blockers (ARB), calcium channel blockers, diuretics, α-blockers and β-blockers, and had at least one medical record 10 days before this prescription;

[1] HbA_{1c} measures the percentage of glycosylated hemoglobin in the total amount of hemoglobin present in the blood. It reflects average blood glucose levels over the past 6–8 weeks. The normal range is below 5.7%.

[2] Systolic blood pressure is the maximum arterial pressure during contraction of the left ventricle of the heart. It is measured in mmHg (millimeters of mercury) and the normal range is below 120.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

- had at least one recorded diagnosis of hypertension (corresponding to the ICD-9 diagnosis codes 401–405);
- had at least three measurements of the systolic blood pressure.

We have identified 11,230 patients for the diabetes dataset and 49,401 patients for the hypertension dataset. Each patient may have multiple entries in her/his medical record. We define the *line of therapy* as a time period (between 200 and 500 days) during which the combination of drugs prescribed to the patient does not change. Each line of therapy is characterized by a drug regimen which is defined as the combination of drugs prescribed to the patient within the first 200 days. The line of therapy intends to capture the period when the patient was experiencing the effect of the drug regimen.

We define *patient visits* within each line of therapy to reflect changes in the features and outcomes. For the diabetic patients, we consider four possible drug regimens (combinations of oral and injectable drugs), while for the hypertension patients, we consider the most frequent 19 of the 32 combinations of drugs and merge all others into one class.

Diabetic Patients. During each line of therapy, we assume that the patient visits every 100 days, beginning from the start of the therapy and continuing until at least 80 days prior to the end of the therapy. The measurements, lab tests are averaged over the 100 days prior to the visit. We define the *current prescription* of each visit as the combination of drugs that was given during the 100 days immediately preceding the visit, and the *standard of care* as the drug regimen that is prescribed by the doctors at the time of the visit. If no value exists over the 100 days, we use the neighboring visits to determine the measurements/lab tests (through linear interpolation) and the current prescription. The *future* outcome for each visit is computed as the average HbA_{1c} 75 to 200 days after the visit. Patient visits that contain missing values for the outcome are dropped. We end up with 12,016 valid visits, which are divided into four groups based on their standard of care.

Hypertension Patients. During each line of therapy, the patient visits are considered occurring every 70 days, beginning from the start of the

The version of record is available at: http://dx.doi.org/10.1561/2400000026

therapy and continuing until at least 180 days prior to the end of the therapy. The measurements, lab tests are averaged over the 10 days prior to the visit. We define the *current prescription* of each visit as the combination of drugs that was given during the 10 days immediately preceding the visit, and the *standard of care* as the drug regimen that is prescribed by the doctors at the time of the visit. We narrow down the time window due to the fact that the blood pressure is usually much more noisy than the HbA_{1c}, and thus the features within a smaller time window tend to be more relevant. The *future* outcome of the visit is the average systolic blood pressure 70 to 180 days after it. Linear interpolation is used to replace the missing values of the measurements and lab tests. We have obtained 26,128 valid visits, which are divided into 20 groups based on their standard of care.

Prescriptions. The prescriptions are used to group the patient visits. For the diabetic patients, we consider two types of prescriptions: one includes oral medications, e.g., metformin, pioglitazone, and sitagliptin, etc., and the other type includes injectable medications, e.g., insulin. Typically, injectable medications are prescribed for patients with more advanced disease. For the hypertension patients, six types of prescriptions are considered: ACE inhibitor, Angiotensin Receptor Blockers (ARB), calcium channel blockers, thiazide and thiazide-like diuretics, α-blockers and β-blockers.

The following sets of features are considered for building the predictive model. The number of features included in both datasets is 63. All features are standardized before fed into our algorithm.

Demographic information. Includes sex (male, female and other), age and race (10 types). We consider the three most frequent races: Caucasian, Black, and Hispanic, and group all others into one category "other".

Measurements. Systolic/diastolic blood pressure (mmHg), Body Mass Index (BMI) and pulse.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Lab tests. Two types of tests considered: blood chemistry tests such as calcium, carbon dioxide, chloride, potassium, sodium, creatinine, and urea nitrogen; and hematology tests such as blood glucose, hematocrit, hemoglobin, leukocyte count, platelet count, and mean corpuscular volume.

Diagnosis history. The ICD-9 coding system is used to record diagnoses.

7.4.2 Model Development and Results

We will compare our prescriptive algorithm with several alternatives that replace our *Distributionally Robust Linear Regression (DRLR)* informed K-NN with a different predictive model such as LASSO, CART, and OLS informed K-NN [166]. Both deterministic and randomized prescriptive policies are considered using predictions from these models. We note a very recent tree-based algorithm called Optimal Prescription Tree (OPT) developed in [167], that uses either constant or linear models in the leaves of the tree in order to predict the counterfactuals and to assign optimal treatments to new samples. We do not include it as a comparison in this work, yet, it would be interesting to do in subsequent work.

Parameter tuning. Within each prescription group, we randomly split the patient visits into three sets: a training set (80%), a validation set (10%), and a test set (10%). To reflect the dependency of the number of neighbors on the number of training samples, we perform a linear regression between these two quantities, which will be used to determine the number of neighbors needed in different settings.

To tune the exponent ξ for the randomized strategy, it is necessary to evaluate the effects of counterfactual treatments. We assess the predictive power of a series of robust predictive models in terms of the following metrics:

- R-square:

$$\mathrm{R}^2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{N_t}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_t}(y_i - \bar{y})^2},$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where $\mathbf{y} = (y_1, \ldots, y_{N_t})$ and $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_{N_t})$ are the vectors of the true (observed) and predicted outcomes, respectively, with N_t the size of the test set, and $\bar{y} = (1/N_t) \sum_{i=1}^{N_t} y_i$.

- *Mean Squared Error (MSE)*:

$$\text{MSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - \hat{y}_i)^2.$$

- *Mean Absolute Error (MeanAE)*, which is more robust to large deviations than the MSE in that the absolute value function increases more slowly than the square function over large (absolute) values of the argument.

$$\text{MeanAE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N_t} \sum_{i=1}^{N_t} |y_i - \hat{y}_i|.$$

- MAD, which can be viewed as a robust measure of the MeanAE, computing the median of the absolute deviations:

$$\text{MAD}(\mathbf{y}, \hat{\mathbf{y}}) = \text{Median}\left(|y_i - \hat{y}_i|, i \in [\![N_t]\!]\right).$$

The out-of-sample performance metrics of the various models on the two datasets are shown in Tables 7.1 and 7.2, where the numbers in the parentheses show the improvement of DRLR informed K-NN compared against other methods. Huber refers to the robust regression method proposed in [7], [8], and CART refers to the *Classification And Regression Trees*. Huber/OLS/LASSO + K-NN means fitting a K-NN regression model with a Huber/OLS/LASSO-weighted distance metric. We note that in order to produce well-defined and meaningful predictive performance metrics, the dataset used to generate Tables 7.1 and 7.2 did not group the patients by their prescriptions. A universal model was fit to all patients with prescription information being used as one of the predictors. Nevertheless, it would still be considered as a fair comparison as all models were evaluated on the same dataset. The results provide supporting evidence for the validity of our DRLR+K-NN model that outperforms all others in all metrics, and is thus used to impute the outcome for an unobservable treatment m, through averaging over the

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Table 7.1: Performance of different models for predicting future HbA_{1c} for diabetic patients

Methods	R^2	MSE	MeanAE	MAD
OLS	0.52 (2%)	1.36 (2%)	0.81 (4%)	0.55 (11%)
LASSO	0.52 (2%)	1.37 (2%)	0.80 (3%)	0.54 (9%)
Huber	0.36 (47%)	1.81 (26%)	0.96 (19%)	0.70 (30%)
RLAD	0.50 (4%)	1.40 (4%)	0.78 (1%)	0.50 (1%)
K-NN	0.25 (109%)	2.11 (37%)	1.07 (27%)	0.81 (39%)
OLS+K-NN	0.52 (0%)	1.34 (0%)	0.79 (1%)	0.51 (3%)
LASSO+K-NN	0.52 (1%)	1.36 (1%)	0.79 (1%)	0.50 (1%)
Huber+K-NN	0.51 (3%)	1.38 (3%)	0.81 (3%)	0.53 (7%)
DRLR+K-NN	0.52 (N/A)	1.34 (N/A)	0.78 (N/A)	0.49 (N/A)
CART	0.49 (7%)	1.43 (7%)	0.81 (3%)	0.50 (2%)

Table 7.2: Performance of different models for predicting future systolic blood pressure for hypertension patients

Methods	R^2	MSE	MeanAE	MAD
OLS	0.31 (14%)	170.80 (6%)	10.09 (7%)	8.15 (9%)
LASSO	0.31 (14%)	170.83 (6%)	10.08 (7%)	8.22 (10%)
Huber	0.22 (62%)	193.54 (17%)	10.70 (12%)	8.61 (14%)
RLAD	0.30 (18%)	173.32 (8%)	10.11 (7%)	8.28 (11%)
K-NN	0.33 (10%)	167.41 (5%)	9.62 (2%)	7.50 (2%)
OLS+K-NN	0.35 (1%)	160.22 (0%)	9.42 (0%)	7.49 (1%)
LASSO+K-NN	0.32 (12%)	169.50 (6%)	9.74 (3%)	7.73 (5%)
Huber+K-NN	0.32 (10%)	167.92 (5%)	9.71 (3%)	7.84 (6%)
DRLR+K-NN	0.36 (N/A)	159.74 (N/A)	9.42 (N/A)	7.38 (N/A)
CART	0.25 (43%)	186.23 (14%)	10.34 (9%)	8.22 (10%)

most similar patient visits who have received the prescription m in the *validation set*, where the number of neighbors is selected to fit the size of the validation set. Note that using DRLR+K-NN as an imputation model might cause bias in evaluating the performance of different methods, since it is in favor of the framework that uses the same model (DRLR+K-NN) to predict the future outcome. Using a weighted combination of several different predictive models may alleviate the bias. This could be done in future work.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Model training. We solve the predictive models on the whole training set with the best tuned parameters, the output of which is used to develop the optimal prescriptions for the test set patients. The parameter $\bar{\epsilon}$ in the threshold $T(\mathbf{x})$ is set to 0.1. For estimating the conditional mean and standard deviation of the predicted outcome using Algorithm 1, we set $a_m = 0.9N_m$, and $d_m = 100$. We compute the average improvement (reduction) in outcomes for patients in the test set, which is defined to be the difference between the (expected) *future* outcome under the recommended therapy and the *current* observed outcome. If the recommendation does not match the standard of care, its future outcome is estimated through the imputation model that was discussed earlier, where K_m should be selected to fit the size of the *test set*.

Results and discussions. The reductions in outcomes (future minus current) for various models are shown in Table 7.3. The columns indicate the prescriptive policies (deterministic or randomized); the rows represent the predictive models whose outcomes $\hat{y}_m(\mathbf{x})$ serve as inputs to the prescriptive algorithm. We test the performance of all algorithms over five repetitions, each with a different training set. The numbers outside the parentheses are the mean reductions in the outcome and the numbers inside the parentheses are the corresponding standard deviations. We note that HbA_{1c} is measured in percentage while systolic blood pressure in mmHg. We also list the reductions in outcomes resulted from the *standard of care*, and the *current prescription* which prescribes $m_f(\mathbf{x}) = m_c(\mathbf{x})$ with probability one, i.e., always continuing the current drug regimen.

Table 7.3: The reduction in HbA_{1c}/systolic blood pressure for various models

	Diabetes		Hypertension	
	Deterministic	**Randomized**	**Deterministic**	**Randomized**
LASSO	−0.51 (0.16)	−0.51 (0.16)	−4.71 (1.09)	−4.72 (1.10)
CART	−0.45 (0.13)	−0.42 (0.14)	−4.84 (0.62)	−4.87 (0.66)
OLS+K-NN	−0.53 (0.13)	−0.53 (0.13)	−4.33 (0.46)	−4.33 (0.47)
DRLR+K-NN	−0.56 (0.06)	−0.55 (0.08)	−6.98 (0.86)	−7.22 (0.82)
Current prescription	−0.22 (0.04)		−2.52 (0.19)	
Standard of care	−0.22 (0.03)		−2.37 (0.11)	

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Several observations are in order: (i) all models outperform the current prescription and the standard of care; (ii) the DRLR-informed K-NN model leads to the largest reduction in outcomes with a relatively stable performance; and (iii) the randomized policy achieves a similar performance (slightly better on the hypertension dataset) to the deterministic one. We expect the randomized strategy to win when the effects of several treatments do not differ much, in which case the deterministic algorithm might produce misleading results. The randomized policy could potentially improve the out-of-sample (generalization) performance, as it gives the flexibility of exploring options that are suboptimal on the training set, but might be optimal on the test set. The advantages of the DRLR+K-NN model are more prominent in the hypertension dataset, due to the fact that we considered a finer classification of the prescriptions for patients with hypertension, while for diabetic patients, we only distinguish between oral and injectable prescriptions.

7.4.3 Refinement on the DRLR+K-NN Model

Up to now, we used a patient-independent parameter K_m (the number of neighbors in group m) to predict the effects of treatments on different individuals. Such a strategy might improperly utilize less relevant information and lead to inadequate predictions. For example, denote by d_i^m the distance between the patient in question and her i-th closest neighbor in group m, and assume there exists a "big jump" at d_j^m, i.e., $d_j^m - \sum_{i=1}^{j-1} d_i^m/(j-1)$ is large. If $K_m \geq j$, we would include the j-th closest neighbor in computing the K-NN average, resulting in a biased estimate given its dissimilarity to the patient of interest.

We thus propose a patient-specific rule to determine the appropriate number of neighbors. Specifically, using the notations d_i^m defined above, we know $d_1^m \leq d_2^m \leq \cdots \leq d_{K_m}^m$. Define

$$j_m^* = \arg\max_j \left(d_j^m - \sum_{i=1}^{j-1} \frac{d_i^m}{j-1} \right).$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

The number of neighbors K'_m will be determined as follows:

$$K'_m = \begin{cases} j^*_m - 1, & \text{if } \frac{d^m_{j^*_m} - \sum_{i=1}^{j^*_m - 1} \frac{d^m_i}{j^*_m - 1}}{\sum_{i=1}^{j^*_m - 1} \frac{d^m_i}{j^*_m - 1}} > \tilde{T}, \\ K_m, & \text{otherwise,} \end{cases}$$

where $\tilde{T}$ is some threshold that can be tuned using cross-validation. This strategy discards the neighbors that are relatively far away from the patient under consideration. We test this strategy on the two datasets, using a cross-validated threshold $\tilde{T} = 2.5$ and 1 for diabetes and hypertension, respectively, and show the results in Tables 7.4 and 7.5. Notice that such a truncation strategy could affect both the training of DRLR+K-NN and the imputation model that is used to evaluate the effects of counterfactual treatments. To compare with the original strategy of using a uniform K_m for every patient, we list in the left halves of the tables the results from adopting the truncation strategy to both training and imputation, and in the right halves the results

Table 7.4: The reduction in HbA_{1c} for various models ($\tilde{T} = 2.5$)

	Training with K'_m		**Training with K_m**	
	Deterministic	**Randomized**	**Deterministic**	**Randomized**
LASSO	−0.54 (0.19)	−0.54 (0.20)	−0.50 (0.17)	−0.49 (0.17)
CART	−0.62 (0.32)	−0.57 (0.27)	−0.56 (0.19)	−0.53 (0.15)
OLS+K-NN	−0.65 (0.25)	−0.64 (0.25)	−0.61 (0.16)	−0.61 (0.17)
DRLR+K-NN	−0.68 (0.20)	−0.67 (0.23)	−0.61 (0.10)	−0.59 (0.10)
Current prescription	−0.23 (0.05)		−0.22 (0.05)	
Standard of care	−0.22 (0.03)		−0.22 (0.03)	

Table 7.5: The reduction in systolic blood pressure for various models ($\tilde{T} = 1$)

	Training with K'_m		**Training with K_m**	
	Deterministic	**Randomized**	**Deterministic**	**Randomized**
LASSO	−4.34 (0.28)	−4.33 (0.28)	−4.22 (0.20)	−4.22 (0.19)
CART	−4.46 (0.46)	−4.49 (0.50)	−4.48 (0.55)	−4.51 (0.49)
OLS+K-NN	−4.30 (0.35)	−4.30 (0.32)	−4.27 (0.32)	−4.29 (0.31)
DRLR+K-NN	−7.42 (0.46)	−7.58 (0.51)	−6.58 (0.70)	−6.78 (0.73)
Current prescription	−2.56 (0.14)		−2.50 (0.16)	
Standard of care	−2.37 (0.11)		−2.37 (0.11)	

The version of record is available at: http://dx.doi.org/10.1561/2400000026

from applying the truncation only to the imputation/evaluation model. We see that using a patient-specific K'_m in general leads to a larger reduction in outcomes.

7.5 Summary

We proposed an interpretable robust predictive method by combining ideas from distributionally robust optimization with the local learning procedure K-Nearest Neighbors, and established theoretical guarantees on its out-of-sample predictive performance. We also developed a randomized prescriptive policy based on the robust predictions, and proved its optimality in terms of the expected true outcome. In conjunction, we derived a closed-form expression for a clinically meaningful threshold that is used to activate the randomized prescriptive policy. We applied the proposed methodology to a diabetes and a hypertension dataset obtained from a major safety-net hospital, providing numerical evidence for the predicted improvement on outcomes due to our algorithm.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

8

Advanced Topics in Distributionally Robust Learning

In this section, we will cover a number of active research topics in the domain of DRO under the Wasserstein metric. Different from previous sections, where we focused on traditional supervised learning models with identically and independently distributed labeled data, here we want to explore how to adapt the DRO framework to more complex data and model regimes. Specifically, we will study:

- Distributionally Robust *Semi-Supervised Learning (SSL)*, which estimates a robust classifier with partially labeled data, through (i) either restricting the marginal distribution to be consistent with the unlabeled data, (ii) or modifying the structure of DRO by allowing the center of the ambiguity set to vary, reflecting the uncertainty in the labels of the unsupervised data.

- DRO in *Reinforcement Learning (RL)* with temporally correlated data, which considers *Markov Decision Processes (MDPs)* and seeks to inject robustness into the probabilistic transition model. We will derive a lower bound for the *distributionally robust* value function in a regularized form.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

8.1 Distributionally Robust Learning with Unlabeled Data

In this section we study a Distributionally Robust Optimization (DRO) model with the availability of unlabeled data. This problem can be approached with two types of model architectures. One assumes a setting where supervised DRO with labeled data does not ensure a good generalization performance, and explores the role of unlabeled data in enhancing the performance of conventional supervised DRO, while the other is set up in a semi-supervised setting with potential noise on both labeled and unlabeled data, and aims to robustify SSL algorithms by employing the DRO framework.

Note that the role of the unlabeled data in the two modeling schemes is different, so are the learning objectives. One seeks to utilize the additional information contained in the unlabeled data, while the other seeks immunity to perturbations on both labeled and unlabeled data. As we will see in the subsequent sections, the former objective is realized through confining the elements of the DRO formulation, i.e., the ambiguity set Ω, to digest the additional information brought by the unlabeled data. By contrast, the latter requires modification of the underlying infrastructure of DRO so that it can be adapted to existing SSL algorithms.

Examples of past works that use unlabeled data to improve adversarial robustness include [181]–[184]. For inducing robustness to SSL, [185] proposed an ensemble learning approach through label aggregation. Previous works that fall into the intersection of DRO and SSL include [186]–[188], where the first two study the role of unlabeled data in improving the generalization performance, while the third one focuses on robustifying a well-known SSL framework, called self-training, by using the DRO.

Throughout this section, we consider a K-class classification problem with a dataset $\mathcal{D}$ of size N consisting of two non-overlapping sets $\mathcal{D}_l$ (labeled) and $\mathcal{D}_{ul}$ (unlabeled), with size N_l and N_{ul}, respectively, and $N_l + N_{ul} = N$. Denote by $\mathcal{I}_l$ and $\mathcal{I}_{ul}$ the index sets corresponding to the labeled and unlabeled data points, respectively. Thus, $\mathcal{D}_l = \{\mathbf{z}_i \triangleq (\mathbf{x}_i, y_i) \colon i \in \mathcal{I}_l\}$, where $y_i \in [\![K]\!]$, and $\mathcal{D}_{ul} = \{\mathbf{x}_i \colon i \in \mathcal{I}_{ul}\}$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

8.1.1 Incorporating Unlabeled Data into Distributionally Robust Learning

One of the prerequisites for ensuring a good generalization performance of Wasserstein DRO requires that the ambiguity set includes the true data distribution. In a "medium-data" regime, where the observed data may be far from the true data distribution, the Wasserstein ball must be extremely large to contain the true data distribution (cf. Theorem 2.7.1). As a result, the learner has to be robust to an enormous variety of data distributions, preventing it from making a prediction with any confidence [187]. To address this problem, a number of works have proposed to use unlabeled data to further constrain the adversary, see [186], [187]. Recall the general Wasserstein DRO formulation for a supervised learning problem with feature vector $\mathbf{x}$ and label y:

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q}\in\Omega} \mathbb{E}^{\mathbb{Q}}[h_{\boldsymbol{\beta}}(\mathbf{x}, y)], \tag{8.1}$$

where $h_{\boldsymbol{\beta}}(\mathbf{x}, y)$ is the loss function evaluated at some hypothesis $\boldsymbol{\beta}$, and $\mathbb{Q}$ is the probability distribution of $(\mathbf{x}, y)$ belonging to some set Ω that constrains the distribution to be close to the empirical distribution of the labeled data, denoted by $\hat{\mathbb{P}}_{N_l}$, in the sense of the order-1 Wasserstein metric induced by a cost metric s:

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})\colon W_{s,1}(\mathbb{Q},\ \hat{\mathbb{P}}_{N_l}) \leq \epsilon\}.$$

To overcome the problem of overwhelmingly-large ambiguity set Ω, [187] proposed to remove from Ω the distributions that are unrealistic in the sense that their marginals in feature space do not resemble the unlabeled data. Specifically, they define the uncertainty set to be

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{U}(\mathbb{P}_{\mathcal{X}}, \underline{\mathbb{P}}_{\mathcal{Y}}, \overline{\mathbb{P}}_{\mathcal{Y}})\colon W_{s,1}(\mathbb{Q},\ \hat{\mathbb{P}}_{N_l}) \leq \epsilon\}, \tag{8.2}$$

where $\underline{\mathbb{P}}_{\mathcal{Y}}$ and $\overline{\mathbb{P}}_{\mathcal{Y}}$ are two distributions on the label y with probability vectors $\underline{\mathbf{p}} \triangleq (\underline{p}_1, \ldots, \underline{p}_K)$ and $\overline{\mathbf{p}} \triangleq (\overline{p}_1, \ldots, \overline{p}_K)$, respectively, and $\mathcal{U}(\mathbb{P}_{\mathcal{X}}, \underline{\mathbb{P}}_{\mathcal{Y}}, \overline{\mathbb{P}}_{\mathcal{Y}})$ is the set of probability measures whose $\mathbf{x}$-marginal is $\mathbb{P}_{\mathcal{X}}$ and y-marginal is constrained by $[\underline{\mathbf{p}}, \overline{\mathbf{p}}]$, i.e., the class i probability $p_i \in [\underline{p}_i, \overline{p}_i], i \in [\![K]\!]$. They choose $\mathbb{P}_{\mathcal{X}}$ to be consistent with the

The version of record is available at: http://dx.doi.org/10.1561/2400000026

unlabeled data $\mathbf{x} \in \mathcal{D}_{ul}$. The constraint on $\mathbb{P}_{\mathcal{Y}}$ could come from prior knowledge, or could be implied by the labeled training data.

[186] also constrained the uncertainty set Ω by incorporating the information of the unlabeled data. Different from (8.2) where the marginals are enforced to be consistent with the unlabeled data, they set the joint support of the feature and labels to be confined to the empirical observations. Specifically, they build a "complete" unlabeled set by assigning all possible labels to each unlabeled data point: $\mathcal{C}_{ul} \triangleq \bigcup_{y=1}^{K} \{(\mathbf{x}_i, y) \colon i \in \mathcal{I}_{ul}\}$, and then construct the full dataset $\mathcal{C} = \mathcal{D}_l \cup \mathcal{C}_{ul}$. The uncertainty set is restricted to be supported on $\mathcal{C}$, namely,

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{C}) \colon W_{s,1}(\mathbb{Q},\ \hat{\mathbb{P}}_{N_l}) \leq \epsilon\}. \tag{8.3}$$

Compared to (8.2), (8.3) is more restrictive in the sense that it imposes constraints on the joint distribution of the feature and labels, while (8.2) only restricts the marginals. Furthermore, it does not allow support points outside the empirical observations, which eliminates one of the major advantages of the Wasserstein metric. In the absence of the unlabeled data, (8.3) essentially asks the learner to be robust only to distributions with support on $\mathcal{D}_l$, which could hurt the generalization capability on unseen data. By contrast, (8.2) guarantees robustness to distributions with support on the whole data space.

Note that the DRO formulation with an uncertainty set defined through either (8.2) or (8.3) does not serve the purpose of robustifying an existing SSL model. Rather, it explores ways of improving the generalization performance of a DRO model by utilizing the unlabeled data information.

In the remainder of this section, we will discuss a *Stochastic Gradient Descent (SGD)* algorithm proposed in [187], in order to solve the Wasserstein DRO formulation assembled with the ambiguity set (8.2). The key is to transform the inner infinite-dimensional maximization problem in (8.1) into its finite-dimensional dual. Define the worst-case expected loss as

$$v_P(\boldsymbol{\beta}) \triangleq \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}}[h_{\boldsymbol{\beta}}(\mathbf{x}, y)]. \tag{8.4}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Rewrite (8.4) by casting it as an optimal transportation problem with a transport plan $\pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$:

$$\begin{aligned} v_P(\boldsymbol{\beta}) = \sup_{\pi \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \quad & \int_{(\mathcal{X}\times\mathcal{Y})\times\mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{x}, y) \mathrm{d}\pi((\mathbf{x}, y), \mathbf{z}') \\ \text{s.t.} \quad & \int_{\mathcal{Z}\times\mathcal{Z}} s(\mathbf{z}, \mathbf{z}') \mathrm{d}\pi(\mathbf{z}, \mathbf{z}') \le \epsilon, \\ & \int_{\mathcal{Z}\times\mathcal{Z}} \delta_{\mathbf{z}_i}(\mathbf{z}') \mathrm{d}\pi(\mathbf{z}, \mathbf{z}') = \frac{1}{N_l}, \quad \forall i \in \mathcal{I}_l, \\ & \int_{(\mathcal{A}\times\mathcal{Y})\times\mathcal{Z}} \mathrm{d}\pi((\mathbf{x}, y), \mathbf{z}') = \mathbb{P}_{\mathcal{X}}(\mathcal{A}), \quad \forall \mathcal{A} \subseteq \mathcal{X}, \\ & \int_{(\mathcal{X}\times\mathcal{Y})\times\mathcal{Z}} \delta_i(y) \mathrm{d}\pi((\mathbf{x}, y), \mathbf{z}') \le \overline{p}_i, \quad \forall i \in [\![K]\!], \\ & \int_{(\mathcal{X}\times\mathcal{Y})\times\mathcal{Z}} \delta_i(y) \mathrm{d}\pi((\mathbf{x}, y), \mathbf{z}') \ge \underline{p}_i, \quad \forall i \in [\![K]\!], \end{aligned} \tag{8.5}$$

where we use $\mathbf{z} \triangleq (\mathbf{x}, y)$ to index the support of the worst-case measure and $\mathbf{z}'$ to index the support of $\hat{\mathbb{P}}_{N_l}$. Notice that the constraint on the $\mathbf{x}$-marginal is infinite dimensional. Through translating (8.5) to its dual one can move the infinite dimensional constraint to an expectation under $\mathbb{P}_{\mathcal{X}}$ in the objective. The dual to (8.5) can be formulated as

$$\begin{aligned} v_D(\boldsymbol{\beta}) = \inf_{\alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}}} \quad & \alpha\epsilon + \frac{1}{N_l}\sum_{i=1}^{N_l} \gamma_i + \sum_{k=1}^{K} (\overline{\lambda}_k \overline{p}_k - \underline{\lambda}_k \underline{p}_k) \\ & + \mathbb{E}^{\mathbb{P}_{\mathcal{X}}}[\phi(\mathbf{x}; \boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}})] \\ \text{s.t.} \quad & \alpha, \underline{\lambda}_k, \overline{\lambda}_k \ge 0, \ \forall k \in [\![K]\!], \end{aligned} \tag{8.6}$$

where

$$\phi(\mathbf{x}; \boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}}) = \max_{i \in [\![N_l]\!], k \in [\![K]\!]} \phi^{i,k}(\mathbf{x}; \boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}}),$$

$$\phi^{i,k}(\mathbf{x}; \boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}}) \triangleq h_{\boldsymbol{\beta}}(\mathbf{x}, k) - (\alpha s((\mathbf{x}, k), \mathbf{z}_i) + \gamma_i) - (\overline{\lambda}_k - \underline{\lambda}_k).$$

It can be shown that strong duality holds if the primal problem (8.5) is feasible. We refer the reader to Theorem 2 of [187] for a detailed proof. The DRO problem (8.1) reduces to minimizing $v_D(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$, which can be solved via the stochastic gradient method. The main obstacle to deriving the gradient lies in the expectation in the objective of $v_D(\boldsymbol{\beta})$. By applying the Reynolds Transport Theorem [189], one can obtain

The version of record is available at: http://dx.doi.org/10.1561/2400000026

that

$$\frac{\partial}{\partial\alpha}\mathbb{E}^{\mathbb{P}_{\mathcal{X}}}[\phi(\mathbf{x};\boldsymbol{\beta},\alpha,\boldsymbol{\gamma},\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{\lambda}})] = \mathbb{E}^{\mathbb{P}_{\mathcal{X}}}\left[\frac{\partial}{\partial\alpha}\phi(\mathbf{x};\boldsymbol{\beta},\alpha,\boldsymbol{\gamma},\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{\lambda}})\right]. \tag{8.7}$$

Notice that ϕ is defined to be the maximum of a series of functions $\phi^{i,k}$. To evaluate its derivative, we need to partition the feature space $\mathcal{X}$ to recognize the set of points $\mathbf{x}$ where the maximum is achieved at each (i,k). Define

$$\mathcal{V}^{i,k} \triangleq \{\mathbf{x}\in\mathcal{X}\colon \phi^{i,k}(\mathbf{x};\boldsymbol{\beta},\alpha,\boldsymbol{\gamma},\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{\lambda}}) \geq \phi^{i',k'}(\mathbf{x};\boldsymbol{\beta},\alpha,\boldsymbol{\gamma},\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{\lambda}}), \forall i',k'\}.$$

The derivative of ϕ can be evaluated as

$$\frac{\partial}{\partial\alpha}\phi(\mathbf{x};\boldsymbol{\beta},\alpha,\boldsymbol{\gamma},\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{\lambda}}) = -\sum_{i=1}^{N_l}\sum_{k=1}^{K}\mathbf{1}_{\mathcal{V}^{i,k}}(\mathbf{x})s((\mathbf{x},k),\mathbf{z}_i),$$

where $\mathbf{1}_{\mathcal{V}^{i,k}}(\mathbf{x})$ denotes the indicator function of the event $\mathbf{x}\in\mathcal{V}^{i,k}$. Similarly, the gradients w.r.t. other parameters are computed as follows

$$\frac{\partial}{\partial\gamma_i}\phi(\mathbf{x};\boldsymbol{\beta},\alpha,\boldsymbol{\gamma},\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{\lambda}}) = -\sum_{k=1}^{K}\mathbf{1}_{\mathcal{V}^{i,k}}(\mathbf{x}),$$

$$\frac{\partial}{\partial\underline{\lambda}_k}\phi(\mathbf{x};\boldsymbol{\beta},\alpha,\boldsymbol{\gamma},\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{\lambda}}) = \sum_{i=1}^{N_l}\mathbf{1}_{\mathcal{V}^{i,k}}(\mathbf{x}),$$

$$\frac{\partial}{\partial\overline{\lambda}_k}\phi(\mathbf{x};\boldsymbol{\beta},\alpha,\boldsymbol{\gamma},\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{\lambda}}) = -\sum_{i=1}^{N_l}\mathbf{1}_{\mathcal{V}^{i,k}}(\mathbf{x}),$$

$$\frac{\partial}{\partial\beta_j}\phi(\mathbf{x};\boldsymbol{\beta},\alpha,\boldsymbol{\gamma},\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{\lambda}}) \in \sum_{i=1}^{N_l}\sum_{k=1}^{K}\mathbf{1}_{\mathcal{V}^{i,k}}(\mathbf{x})\frac{\partial}{\partial\beta_j}h_{\boldsymbol{\beta}}(\mathbf{x},k).$$

For $\mathbf{x}$ lying on the boundary between two of the sets $\mathcal{V}^{i,k}$, we can obtain a subgradient by arbitrarily selecting only one of these $\mathcal{V}^{i,k}$ to contain $\mathbf{x}$ when evaluating $\mathbf{1}_{\mathcal{V}^{i,k}}(\mathbf{x})$. To evaluate the expectation of the gradient under $\mathbb{P}_{\mathcal{X}}$ on the RHS of (8.7), one can simulate a series of $\mathbf{x}$ values, say $\mathbf{x}_1,\ldots,\mathbf{x}_{N_b}$, from $\mathbb{P}_{\mathcal{X}}$, and compute the above gradients by taking the sample average. This is summarized in Algorithm 2.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Algorithm 2 SGD for distributionally robust learning with unlabeled data under uncertainty set (8.2).

Input: $\epsilon \geq 0, \underline{p}_i, \overline{p}_i \in [0,1], i \in [\![K]\!]$, feasible solution $\boldsymbol{\beta}_0$, step size $\delta > 0$, batch size N_b.
$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}_0, \alpha \leftarrow 0, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}} \leftarrow \mathbf{0}$.
while not converged **do**
  Sample $\mathbf{x}_1, \ldots, \mathbf{x}_{N_b} \sim \mathbb{P}_{\mathcal{X}}$.
  $\boldsymbol{\beta} \leftarrow \text{Proj}_{\mathcal{B}}[\boldsymbol{\beta} - \frac{\delta}{N_b} \sum_{j=1}^{N_b} \nabla_{\boldsymbol{\beta}} \phi(\mathbf{x}_j; \boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}})]$
  $\alpha \leftarrow \max(0, \alpha - \delta[\epsilon + \frac{1}{N_b} \sum_{j=1}^{N_b} \nabla_{\alpha} \phi(\mathbf{x}_j; \boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}})])$
  $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} - \delta[\frac{1}{N_l}\mathbf{e} + \frac{1}{N_b} \sum_{j=1}^{N_b} \nabla_{\boldsymbol{\gamma}} \phi(\mathbf{x}_j; \boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}})]$
  $\underline{\boldsymbol{\lambda}} \leftarrow \max(0, \underline{\boldsymbol{\lambda}} - \delta[-\underline{\mathbf{p}} + \frac{1}{N_b} \sum_{j=1}^{N_b} \nabla_{\underline{\boldsymbol{\lambda}}} \phi(\mathbf{x}_j; \boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}})])$
  $\overline{\boldsymbol{\lambda}} \leftarrow \max(0, \overline{\boldsymbol{\lambda}} - \delta[\overline{\mathbf{p}} + \frac{1}{N_b} \sum_{j=1}^{N_b} \nabla_{\overline{\boldsymbol{\lambda}}} \phi(\mathbf{x}_j; \boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}, \underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}})])$
end while

8.1.2 Distributionally Robust Semi-Supervised Learning

In this subsection we discuss the problem of robustifying existing SSL algorithms via DRO. Different from Section 8.1.1, the goal here is to induce robustness into conventional SSL models, which requires modification of the DRO infrastructure in order to fit the characteristics of the problem at hand. Note that DRO cannot readily be applied to the partially-labeled setting, since it needs complete knowledge of all the feature-label pairs.

A well-known family of SSL models is called *self-learning*, which first trains a classifier on the labeled portion of a dataset, and then assigns pseudo-labels to the remaining unlabeled samples using the learned rules. The enlarged dataset consisting of both the supervised and artificially-labeled unsupervised samples is used in the final stage of training. To prevent overfitting, instead of assigning a deterministic hard label to the unsupervised data points, one can apply a soft labeling scheme that maintains a level of uncertainty through specifying a probability distribution of the labels.

To use DRO in a semi-supervised setting, we need to address the uncertainty embedded in the unknown labels of the unsupervised samples. This can be resolved by soft-labeling. Define the consistent set of

The version of record is available at: http://dx.doi.org/10.1561/2400000026

probability distributions $\hat{\mathcal{P}}(\mathcal{D}) \subseteq \mathcal{P}(\mathcal{Z})$ w.r.t. a partially-labeled dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_{ul}$ as

$$\hat{\mathcal{P}}(\mathcal{D}) \triangleq \left\{ \left(\frac{N_l}{N}\right)\hat{\mathbb{P}}_{N_l} + \left(\frac{N_{ul}}{N}\right)\hat{\mathbb{P}}_{N_{ul}} \cdot \mathbb{Q} \colon \mathbb{Q} \in \mathcal{P}^{\mathcal{X}}(\mathcal{Y}) \right\},$$

where $\mathbb{Q}$ encodes the uncertainty in the labels for the unsupervised dataset $\mathcal{D}_{ul}$, and $\mathcal{P}^{\mathcal{X}}(\mathcal{Y})$ denotes the set of all conditional distributions supported on $\mathcal{Y}$, given features in $\mathcal{X}$. Note that the distributions in $\hat{\mathcal{P}}(\mathcal{D})$ differ from each other only in the way they assign soft labels to the unlabeled data, and the empirical measure corresponding to the true complete dataset is a member of $\hat{\mathcal{P}}(\mathcal{D})$.

We will illustrate the idea proposed in [188] for introducing DRO to SSL, where they select a suitable measure from $\hat{\mathcal{P}}(\mathcal{D})$, and use it as a proxy of the true empirical probability measure that serves as the center of the Wasserstein ball. The learner essentially aims to hedge against a set of distributions centered at some $\mathbb{S} \in \hat{\mathcal{P}}(\mathcal{D})$ that is induced by a soft-label distribution $\mathbb{Q}$, so that the resulting classification rule would show low sensitivity to adversarial perturbations around the soft-label distribution. The criterion for choosing $\mathbb{S}$ is to make the worst-case expected loss as small as possible. Specifically, the *Semi-Supervised Distributionally Robust Learning (SSDRO)* model proposed by [188] can be formulated as

$$\inf_{\boldsymbol{\beta}} \inf_{\mathbb{S} \in \hat{\mathcal{P}}(\mathcal{D})} \left\{ \sup_{\mathbb{P} \in \Omega_\epsilon(\mathbb{S})} \mathbb{E}^{\mathbb{P}}[h_{\boldsymbol{\beta}}(\mathbf{x}, y)] + \left(\frac{1 - N_l/N}{\lambda}\right) \mathbb{E}^{\hat{\mathbb{P}}_{N_{ul}}}[H(\mathbb{S}_{|\mathbf{x}})] \right\}, \quad (8.8)$$

where $\Omega_\epsilon(\mathbb{S})$ denotes the set of probability distributions that are close to $\mathbb{S}$ by a distance at most ϵ, i.e.,

$$\Omega_\epsilon(\mathbb{S}) \triangleq \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon W_{s,1}(\mathbb{P},\ \mathbb{S}) \leq \epsilon\},$$

In (8.8), $\mathbb{S}_{|\mathbf{x}}$ is the conditional distribution over $\mathcal{Y}$ given $\mathbf{x} \in \mathcal{X}$, $\lambda < 0$ is a user-defined parameter, and $H(\cdot)$ denotes the Shannon entropy.

Notice that for a fixed $\boldsymbol{\beta}$, the inner infimum of (8.8) guides the learner to pick a soft label distribution that tends to reduce the loss function, which [188] refers to as an *optimistic* learner. Alternatively, one can choose to be *pessimistic*, i.e., choosing a $\boldsymbol{\beta}$ that hedges against the maximum loss over all possible choices of $\mathbb{S}$. To prevent hard labeling

The version of record is available at: http://dx.doi.org/10.1561/2400000026

of the unsupervised data, λ is set to be negative for *optimistic* learning, and positive for *pessimistic* learning.

Note also that compared to conventional DRO models, in (8.8) we have an additional regularization term that penalizes the Shannon entropy of the conditional label distribution of the unlabeled data. When $\lambda < 0$, the regularization term $(\frac{1-N_l/N}{\lambda})\mathbb{E}^{\hat{\mathbb{P}}_{Nul}}[H(\mathbb{S}_{|\mathbf{x}})]$ becomes negative. The formulation (8.8) essentially promotes softer labels for the unlabeled data by encouraging a larger entropy, implying a higher level of uncertainty in the labels.

We next discuss how to solve Problem (8.8). Using duality, [188] was able to transform the inner min-max formulation to an analytic form whose gradient can be efficiently computed. A Lagrangian relaxation to (8.8) is given in the following theorem.

Theorem 8.1.1 ([188], Theorem 1)**.** Consider a continuous loss function h, and a continuous transportation cost s. For a partially-labeled dataset $\mathcal{D}$ with size N, define the empirical Semi-Supervised Adversarial Risk (SSAR), denoted by $\hat{R}_{\text{SSAR}}(\boldsymbol{\beta}; \mathcal{D})$, as

$$\hat{R}_{\text{SSAR}}(\boldsymbol{\beta}; \mathcal{D}) \triangleq \frac{1}{N}\sum_{i\in\mathcal{I}_l} \phi_\gamma(\mathbf{x}_i, y_i; \boldsymbol{\beta}) + \frac{1}{N}\sum_{i\in\mathcal{I}_{ul}} \underset{y\in\mathcal{Y}}{\overset{(\lambda)}{\text{softmin}}}\{\phi_\gamma(\mathbf{x}_i, y; \boldsymbol{\beta})\} + \gamma\epsilon, \tag{8.9}$$

where $\gamma \geq 0$, and the adversarial loss $\phi_\gamma(\mathbf{x}, y; \boldsymbol{\beta})$ and the soft-minimum operator are defined as:

$$\phi_\gamma(\mathbf{x}, y; \boldsymbol{\beta}) \triangleq \sup_{\mathbf{z}'\in\mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{z}') - \gamma s(\mathbf{z}', (\mathbf{x}, y)), \tag{8.10}$$

and

$$\underset{y\in\mathcal{Y}}{\overset{(\lambda)}{\text{softmin}}}\{q(y)\} \triangleq \frac{1}{\lambda}\log\left(\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} e^{\lambda q(y)}\right),$$

respectively. Let $\boldsymbol{\beta}^*$ be a minimizer of (8.8) for some given $\epsilon \geq 0$ and $\lambda < 0$. Then, there exists $\gamma \geq 0$ such that $\boldsymbol{\beta}^*$ is also a minimizer of (8.9) with the same parameters ϵ and λ.

According to Theorem 8.1.1 our problem is now translated to solving for a $\boldsymbol{\beta}$ that minimizes $\hat{R}_{\text{SSAR}}(\boldsymbol{\beta}; \mathcal{D})$. To apply SGD, the key is to derive the gradient of the adversarial loss function $\phi_\gamma(\mathbf{x}, y; \boldsymbol{\beta})$, which itself is the

The version of record is available at: http://dx.doi.org/10.1561/2400000026

output of an optimization problem. The gradient of ϕ w.r.t. $\boldsymbol{\beta}$ relies on the optimal solution of Problem (8.10), i.e., $\nabla_{\boldsymbol{\beta}}\phi_\gamma(\mathbf{x}, y; \boldsymbol{\beta}) = \mathbf{g}_{\boldsymbol{\beta}}(\mathbf{z}^*(\boldsymbol{\beta}))$, where $\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{z}) \triangleq \nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}}(\mathbf{z})$ and $\mathbf{z}^*(\boldsymbol{\beta})$ is the optimal solution to (8.10). The following lemma specifies a set of sufficient conditions to ensure the uniqueness of the solution.

Lemma 8.1.2 ([188], Lemma 1)**.** Assume the loss function h to be differentiable w.r.t. $\mathbf{z}$, and $\nabla_{\mathbf{z}} h_{\boldsymbol{\beta}}(\mathbf{z})$ is L_z-Lipschitz w.r.t. $\boldsymbol{\beta}$. Also, the cost metric s is 1-strongly convex in its first argument. If $\gamma > L_z$, then Problem (8.10) is $(\gamma - L_z)$-strongly concave for all $(\mathbf{x}, y) \in \mathcal{Z}$.

Lemma 8.1.2 guarantees the existence and uniqueness of the solution to (8.10). We can thus express the gradients of ϕ and $\hat{R}_{\text{SSAR}}$ explicitly as a function of the solution. An efficient computation of the gradient of $\hat{R}_{\text{SSAR}}(\boldsymbol{\beta}; \mathcal{D})$ w.r.t. $\boldsymbol{\beta}$ is given in the following theorem.

Theorem 8.1.3 ([188], Lemma 2)**.** Under conditions of Lemma 8.1.2, assume the loss function h to be differentiable w.r.t. $\boldsymbol{\beta}$, and let $\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{z}) \triangleq \nabla_{\boldsymbol{\beta}} h_{\boldsymbol{\beta}}(\mathbf{z})$. For a fixed $\boldsymbol{\beta}$, define

$$\mathbf{z}_i^*(\boldsymbol{\beta}) = \arg\max_{\mathbf{z}' \in \mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{z}') - \gamma s(\mathbf{z}', (\mathbf{x}_i, y_i)), \quad i \in \mathcal{I}_l, \tag{8.11}$$

and,

$$\mathbf{z}_i^*(y; \boldsymbol{\beta}) = \arg\max_{\mathbf{z}' \in \mathcal{Z}} h_{\boldsymbol{\beta}}(\mathbf{z}') - \gamma s(\mathbf{z}', (\mathbf{x}_i, y)), \quad y \in \mathcal{Y},\ i \in \mathcal{I}_{ul}. \tag{8.12}$$

Then, the gradient of (8.9) w.r.t. $\boldsymbol{\beta}$ can be obtained as

$$\nabla_{\boldsymbol{\beta}} \hat{R}_{\text{SSAR}}(\boldsymbol{\beta}; \mathcal{D}) = \frac{1}{N} \sum_{i \in \mathcal{I}_l} \mathbf{g}_{\boldsymbol{\beta}}(\mathbf{z}_i^*(\boldsymbol{\beta})) + \frac{1}{N} \sum_{i \in \mathcal{I}_{ul}} \sum_{y \in \mathcal{Y}} q(y; \boldsymbol{\beta}) \mathbf{g}_{\boldsymbol{\beta}}(\mathbf{z}_i^*(y; \boldsymbol{\beta})), \tag{8.13}$$

where $q(y; \boldsymbol{\beta}) \triangleq e^{\lambda \phi_\gamma(\mathbf{x}_i, y; \boldsymbol{\beta})} / (\sum_{y' \in \mathcal{Y}} e^{\lambda \phi_\gamma(\mathbf{x}_i, y'; \boldsymbol{\beta})})$.

Using Theorem 8.1.3, we can apply SGD to solve (8.9), or equivalently, the SSDRO model (8.8). This is summarized in Algorithm 3. [188] proved a convergence rate of $O(T^{-1/2})$ for Algorithm 3, if we assume $\mathbf{z}_i^*(\boldsymbol{\beta})$ and $\mathbf{z}_i^*(y; \boldsymbol{\beta})$ can be computed exactly. Nonetheless, the optimality gap δ can be set infinitesimally small due to the strong

The version of record is available at: http://dx.doi.org/10.1561/2400000026

concavity of (8.11) and (8.12) that is shown in Lemma 8.1.2. The parameters γ and λ can be tuned via cross-validation.

Algorithm 3 Stochastic Gradient Descent for SSDRO.

Inputs: $\mathcal{D}, \gamma, \lambda, k \leq N, \delta, \alpha, T$.
Initialize: $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}_0$, $t \leftarrow 0$.
for $t = 0 \rightarrow T - 1$ **do**
  Randomly select index set $\mathcal{I} \subseteq [\![N]\!]$ with size k.
  for $i \in \mathcal{I}_l \cap \mathcal{I}$ **do**
    Compute a δ-approx of $\mathbf{z}_i^*(\boldsymbol{\beta}_t)$ from (8.11).
  end for
  for $(i, y) \in (\mathcal{I}_{ul} \cap \mathcal{I}) \times \mathcal{Y}$ **do**
    Compute a δ-approx of $\mathbf{z}_i^*(y; \boldsymbol{\beta}_t)$ from (8.12).
  end for
  Compute the sub-gradient of $\hat{R}_{\text{SSAR}}(\boldsymbol{\beta}_t; \mathcal{D})$ from (8.13).
  Update: $\boldsymbol{\beta}_{t+1} \leftarrow \boldsymbol{\beta}_t - \alpha \nabla_{\boldsymbol{\beta}} \hat{R}_{\text{SSAR}}(\boldsymbol{\beta}_t; \mathcal{D})$.
end for
Output: $\boldsymbol{\beta}^* \leftarrow \boldsymbol{\beta}_T$.

8.2 Distributionally Robust Reinforcement Learning

So far in this monograph, we considered learning problems where the objective is to predict an output variable (or vector in the setting of Section 6). These learning problems were cast as distributionally robust *single-period* optimization problems. Even in the applications of Section 7 involving medical prescriptions, where we considered information from multiple past time periods to learn actions that optimize an outcome in the next time period, the resulting optimization problem was single-period. In this section, we will discuss *multi-period* optimization motivated by learning a *policy* for a *Markov Decision Process (MDP)*. We will restrict ourselves to *model-based* settings, where there is an explicit model of how the MDP transitions from state to state under some policy, and seek to inject robustness into this transition model. The development follows the work in [190].

The version of record is available at: http://dx.doi.org/10.1561/2400000026

We start by defining a discrete-time MDP. Consider an MDP with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a deterministic reward function $r\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, and a transition probability model p that, given a state s_1 and an action a, determines the probability $p(s_2|s_1,a)$ of landing to the next state s_2. A *policy* π maps states to actions; specifically, $\pi(a|s)$ denotes the probability of selecting action a in state s. The state of the MDP evolves dynamically as follows. Suppose that at time t the MDP is in state s_t. According to the policy π, it selects some action a_t, receives a reward $r(s_t,a_t)$, and transitions to the next state s_{t+1} with probability $p(s_{t+1}|s_t,a_t)$. In an infinite-horizon discounted reward setting, the objective is to select a policy π that maximizes the *expected total discounted reward*

$$\mathbb{E}_p^{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right],$$

where $\gamma\in[0,1)$ is the discount factor and $\tau\sim\pi$ represents a random trajectory $\tau=(s_0,a_0,s_1,a_1,\ldots)$ sampled by selecting the initial state s_0 according to some probability distribution $\rho_0(\cdot)\in\mathcal{P}(\mathcal{S})$, sampling actions according to $a_t\sim\pi(\cdot|s_t)$, and states according to $s_{t+1}\sim p(\cdot|s_t,a_t)$ (hence, the subscript p in the expectation to denote dependence on the transition model p).

We can now define the state *value*, or reward-to-go function, which equals the future total discounted reward when starting from state s, namely,

$$v_p^\pi(s)=\mathbb{E}_p^{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\middle|\, s_0=s\right].$$

The value function can be obtained as a solution to the following *Bellman equation*:

$$v(s)=T_p^\pi v(s)\stackrel{\triangle}{=}\sum_{a\in\mathcal{A}}\pi(a|s)\Big(r(s,a)+\gamma\sum_{q\in\mathcal{S}}p(q|s,a)v(q)\Big).$$

The operator T_p^π satisfies a contraction property with respect to the sup-norm, implying that the Bellman equation has a unique fixed point denoted by $v_p^\pi(s)$. This can for instance be obtained by successive application of T_p^π to some arbitrary initial solution – a method known as *value iteration*.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

8.2.1 Deterministically Robust Policies

A number of results in the literature examined how to introduce robustness with respect to uncertainty on the transition probability model, starting with [191]–[193]. A more complete theory of *robust dynamic programming* has been developed in [194] and [195]. In this work, the transition probability vector $\mathbf{p}_{s,a} = (p(q|s,a);\ q \in \mathcal{S})$ at any state-action pair (s,a) belongs to some ambiguity or uncertainty set $\mathcal{U}_{s,a} \subseteq \mathcal{P}(\mathcal{S})$. It is assumed that every time a state-action pair (s,a) is encountered, a potentially different measure $\mathbf{p}_{s,a} \in \mathcal{U}_{s,a}$ could be applied; this has been termed the *rectangularity assumption* in [194].

In this robust setting, one can define a *robust value function* as the worst-case value function over the uncertainty set, that is,

$$v_{\mathcal{U}}^{\pi}(s) = \inf_{p \in \mathcal{U}} \mathbb{E}_{p}^{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right], \tag{8.14}$$

where the uncertainty set $\mathcal{U}$ is the cartesian product of the transition probability uncertainty sets encountered throughout the trajectory, i.e., $\mathcal{U} = \prod_{t=0}^{\infty} \mathcal{U}_{s_t,a_t}$.

[194] and [195] show that a robust Bellman equation can be written as:

$$\begin{aligned} v(s) =& T_{\mathcal{U}}^{\pi} v(s) \\ \triangleq& \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \inf_{\mathbf{p}_{s,a} \in \mathcal{U}_{s,a}} \sum_{q \in \mathcal{S}} p(q|s,a) v(q) \right). \end{aligned} \tag{8.15}$$

As with the non-robust case, the operator $T_{\mathcal{U}}^{\pi}$ satisfies a contraction property, implying that the robust Bellman equation has a unique fixed point which can be computed by successive application of $T_{\mathcal{U}}^{\pi}$.

8.2.2 Distributionally Robust Policies

Distributionally robust MDPs can be thought of as a generalization of deterministically robust MDPs. Instead of selecting transition probabilities out of the ambiguity set $\mathcal{U}$ defined earlier, we can view the transition probability model as being sampled according to some distribution $\mu \in \mathcal{M} \subseteq \mathcal{P}(\mathcal{U})$, i.e., μ is the probability distribution of the

The version of record is available at: http://dx.doi.org/10.1561/2400000026

transition probability model p. Making the same rectangularity assumption as before, that is, requiring that μ is a product of independent distributions over $\mathcal{U}_{s,a}$, we can define a *distributionally robust value function* similarly to (8.14) as:

$$v^{\pi}_{\mathcal{M}}(s) = \inf_{\mu\in\mathcal{M}} \mathbb{E}^{\tau\sim\pi}_{p\sim\mu}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\middle|\, s_0 = s\right]. \tag{8.16}$$

[190] introduces Wasserstein distributionally robust MDPs by defining the set of distributions $\mathcal{M}$ as a Wasserstein ball around some nominal distribution. More specifically, for any state-action pair (s,a), let $\hat{\mu}_{s,a} \in \mathcal{P}(\mathcal{U}_{s,a})$ be some *nominal* distribution over $\mathcal{U}_{s,a}$. For any distribution $\mu_{s,a} \in \mathcal{P}(\mathcal{U}_{s,a})$, define the order-1 Wasserstein distance induced by some norm $\|\cdot\|$, and denote it by $W_1(\hat{\mu}_{s,a}, \mu_{s,a})$. A Wasserstein ball around the nominal distribution can be defined as:

$$\Omega_{\epsilon_{s,a}}(\hat{\mu}_{s,a}) = \{\mu_{s,a} \in \mathcal{P}(\mathcal{U}_{s,a})\colon W_1(\hat{\mu}_{s,a}, \mu_{s,a}) \le \epsilon_{s,a}\}\,. \tag{8.17}$$

Under a rectangularity assumption as in Section 8.2.1, we define the cartesian product of the sets $\Omega_{\epsilon_{s,a}}(\hat{\mu}_{s,a})$ over all state-action pairs and denote it by $\Omega_{\boldsymbol{\epsilon}}(\hat{\mu}) = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Omega_{\epsilon_{s,a}}(\hat{\mu}_{s,a})$, where $\boldsymbol{\epsilon}$ is a vector defined as $\boldsymbol{\epsilon} = (\epsilon_{s,a};\ (s,a) \in \mathcal{S}\times\mathcal{A})$, and $\hat{\mu} = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \hat{\mu}_{s,a}$.

Analogously to (8.15), the *distributionally robust Bellman equation* can be written as:

$$\begin{aligned} v(s) =& T^{\pi}_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu})} v(s) \\ \triangleq& \sum_{a\in\mathcal{A}} \pi(a|s)\Bigg(r(s,a) \\ &+ \gamma \inf_{\mu_{s,a}\in\Omega_{\epsilon_{s,a}}(\hat{\mu}_{s,a})} \int_{\mathbf{p}_{s,a}\in\mathcal{U}_{s,a}} \sum_{q\in\mathcal{S}} p(q|s,a) v(q) d\mu_{s,a}(\mathbf{p}_{s,a})\Bigg). \end{aligned} \tag{8.18}$$

The operator $T^{\pi}_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu})}$ satisfies a contraction property with respect to the sup-norm, implying that the Bellman equation has a unique fixed point denoted by $v^{\pi}_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu})}(s)$. To find an optimal policy, consider the operator

$$T_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu})} = \sup_{\pi(\cdot|s)\in\mathcal{P}(\mathcal{A})} T^{\pi}_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu})}. \tag{8.19}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

As shown in [190], [196], there exists a *distributionally robust optimal policy* π^* and a unique value function $v^*_{\Omega_\epsilon(\hat{\mu})}(s)$ which is a fixed point of the operator defined by (8.19). In particular, for every $s \in \mathcal{S}$,

$$v^*_{\Omega_\epsilon(\hat{\mu})}(s) = \sup_{\pi} \inf_{\mu \in \Omega_\epsilon(\hat{\mu})} \mathbb{E}^{\tau \sim \pi}_{p \sim \mu} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right] = v^{\pi^*}_{\Omega_\epsilon(\hat{\mu})}(s).$$

The optimal value function can be obtained by value iteration, i.e., successive application of $T_{\Omega_\epsilon(\hat{\mu})}$ to some arbitrary initial value function.

Selecting a Nominal Distribution

The nominal distribution $\hat{\mu}$ that serves as the center of the Wasserstein balls in (8.17) can be determined as the empirical distribution computed from a set of different independent episodes of the MDP. Suppose we have in our disposal n such episodes. Then, for each episode $i \in [\![n]\!]$, and using the observed sequence of states and actions during the episode, we can compute the empirical transition probability $\hat{p}^{(i)}(q|s,a)$ of transitioning into state q when applying action a in state s. The resulting empirical distribution $\hat{\mu}^n_{s,a}$ assigns mass $1/n$ to each $\hat{p}^{(i)}(\cdot|s,a)$, namely,

$$\hat{\mu}^n_{s,a} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\hat{p}^{(i)}(\cdot|s,a)},$$

where $\delta_{\hat{p}^{(i)}(\cdot|s,a)}$ is a Dirac function assigning mass 1 to the model $\hat{p}^{(i)}(\cdot|s,a)$. Defining a product distribution for each episode i by $\delta_i = \prod_{(s,a) \in \mathcal{S} \times \mathcal{A}} \delta_{\hat{p}^{(i)}(\cdot|s,a)}$, we can define the empirical distribution

$$\hat{\mu}^n = \frac{1}{n} \sum_{i=1}^{n} \delta_i.$$

The model above requires computing an empirical transition probability for each state-action pair. When the state-action space is very large, this is not practical. Instead, one can employ some approximation architecture. One possibility is to use an architecture of the following type

$$\hat{p}^{(i)}(q|s,a) = \frac{\exp\{\boldsymbol{\xi}_i' \boldsymbol{\psi}(s,a,q)\}}{\sum_{y \in \mathcal{S}} \exp\{\boldsymbol{\xi}_i' \boldsymbol{\psi}(s,a,y)\}},$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

for some vector of feature functions $\boldsymbol{\psi}(s,a,q)$ and a parameter vector $\boldsymbol{\xi}_i$; the latter can be learned from the sequence of state-actions corresponding to episode i by solving a logistic regression problem.

A Regularization Result for the Distributionally Robust MDP

[190] obtains a regularization result for the Wasserstein distributionally robust MDP that is analogous to the dual-norm regularization we obtained in Section 4. We will outline some of the key steps, referring the reader to [190] for the full details. The result obtains a lower bound on the value function $v^{\pi}_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu}^n)}(s)$.

To that end, define first the *conjugate robust value function* at state s and under policy π. Specifically, let $\mathbf{p} = (p(q|s,a); \forall q, s \in \mathcal{S}, a \in \mathcal{A})$ denote a vectorized form of the transition probability model. For any $\mathbf{z} = (z(q|s,a); \forall q, s \in \mathcal{S}, a \in \mathcal{A})$, we define the conjugate robust value function as

$$v^{*,\pi}_s(\mathbf{z}) \stackrel{\triangle}{=} \inf_{\mathbf{p}} (v^{\pi}_{\mathbf{p}}(s) - \mathbf{z}'\mathbf{p}), \tag{8.20}$$

and let $\mathcal{D}_s = \{\mathbf{z}\colon v^{*,\pi}_s(\mathbf{z}) > -\infty\}$ be its effective domain. Note that as defined, $v^{*,\pi}_s(\mathbf{z})$ is the negative of the convex conjugate of the value function as a function of $\mathbf{p}$ [84].

A key result from [190] is in the following theorem. As discussed earlier, suppose we have data from n episodes from the MDP and we have constructed the empirical transition probabilities for each episode. Let $\hat{\mathbf{p}}^{(i)} = (\hat{p}^{(i)}(q|s,a); \forall q, s \in \mathcal{S}, a \in \mathcal{A})$ be the corresponding vector.

Theorem 8.2.1 [190]. For any policy π, it holds that

$$v^{\pi}_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu}^n)}(s) \geq \frac{1}{n}\sum_{i=1}^{n} v^{\pi}_{\hat{\mathbf{p}}^{(i)}}(s) - \kappa\alpha_s,$$

where $\alpha_s = \sum_{a\in\mathcal{A}} \pi(a|s)\epsilon_{s,a}$, $\kappa = \sup_{\mathbf{z}\in\mathcal{D}_s} \|\mathbf{z}\|_*$, and $\|\cdot\|_*$ is the dual norm to the norm used in defining the Wasserstein uncertainty set (cf. (8.17)).

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Proof. We will provide an outline of the key steps. We start by expressing $v^{\pi}_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu}^n)}(s)$ using the Bellman equation (8.18). We have

$$
\begin{aligned}
&v^{\pi}_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu}^n)}(s) \\
&\quad= \sum_{a\in\mathcal{A}} \pi(a|s)\Bigg(r(s,a) \\
&\qquad + \gamma \inf_{\mu_{s,a}\in\Omega_{\epsilon_{s,a}}(\hat{\mu}^n_{s,a})} \int_{\mathbf{p}_{s,a}} \sum_{q\in\mathcal{S}} p(q|s,a)v(q)d\mu_{s,a}(\mathbf{p}_{s,a})\Bigg) \\
&\quad= \sum_{a\in\mathcal{A}} \pi(a|s)\Bigg(r(s,a) \\
&\qquad + \gamma \inf_{\{\mu_{s,a}:\ W_1(\mu_{s,a},\hat{\mu}^n_{s,a})\le\epsilon_{s,a}\}} \int_{\mathbf{p}_{s,a}} \sum_{q\in\mathcal{S}} p(q|s,a)v(q)d\mu_{s,a}(\mathbf{p}_{s,a})\Bigg) \\
&\quad= \sum_{a\in\mathcal{A}} \pi(a|s)\Bigg(r(s,a) \\
&\qquad + \gamma \inf_{\mu_{s,a}} \sup_{\lambda\ge 0} \Bigg[\int_{\mathbf{p}_{s,a}} \sum_{q\in\mathcal{S}} p(q|s,a)v(q)d\mu_{s,a}(\mathbf{p}_{s,a}) \\
&\qquad + \lambda W_1(\mu_{s,a},\hat{\mu}^n_{s,a}) - \lambda\epsilon_{s,a}\Bigg]\Bigg) \\
&\quad\ge \sum_{a\in\mathcal{A}} \pi(a|s)\Bigg(r(s,a) \\
&\qquad + \gamma \sup_{\lambda\ge 0} \inf_{\mu_{s,a}} \Bigg[\int_{\mathbf{p}_{s,a}} \sum_{q\in\mathcal{S}} p(q|s,a)v(q)d\mu_{s,a}(\mathbf{p}_{s,a}) \\
&\qquad + \lambda W_1(\mu_{s,a},\hat{\mu}^n_{s,a}) - \lambda\epsilon_{s,a}\Bigg]\Bigg), \qquad (8.21)
\end{aligned}
$$

where the last inequality used weak duality.

Next, using the structure of $\hat{\mu}^n$ as an average over n episodes and the fact (due to the rectangularity assumption) that the empirical distribution is a product distribution over state-action pairs, we can deduce from (8.21) that

$$
v^{\pi}_{\Omega_{\boldsymbol{\epsilon}}(\hat{\mu}^n)}(s) \ge \sup_{\lambda\ge 0}\Bigg[\frac{1}{n}\sum_{i=1}^{n} \inf_{\mathbf{p}}(v^{\pi}_{\mathbf{p}}(s) + \lambda\|\mathbf{p}-\hat{\mathbf{p}}^{(i)}\|) - \lambda\alpha_s\Bigg], \qquad (8.22)
$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where $\alpha_s = \sum_{a\in\mathcal{A}} \pi(a|s)\epsilon_{s,a}$. This derivation used similar techniques as in Theorem 3.1.2.

Using the definition of the dual norm and for any $\lambda \geq 0$ we have

$$\begin{aligned}\lambda\|\mathbf{p} - \hat{\mathbf{p}}^{(i)}\| &= \max_{\|\mathbf{z}_i\|_* \leq 1} \lambda \mathbf{z}_i'(\mathbf{p} - \hat{\mathbf{p}}^{(i)}) \\ &= \max_{\|\lambda\mathbf{z}_i\|_* \leq \lambda} \lambda \mathbf{z}_i'(\mathbf{p} - \hat{\mathbf{p}}^{(i)}) \\ &= \max_{\|\mathbf{u}_i\|_* \leq \lambda} \mathbf{u}_i'(\mathbf{p} - \hat{\mathbf{p}}^{(i)}). \end{aligned} \tag{8.23}$$

Denote by $\nu_s^\pi\colon \mathbf{p} \to v_{\mathbf{p}}^\pi(s)$ the function that maps the transition probability vector $\mathbf{p}$ to the value function $v_{\mathbf{p}}^\pi(s)$. Let $\breve{\text{cl}}(\nu_s^\pi)$ be its convex closure, i.e., the greatest closed and convex function upper bounded by ν_s^π at any $\mathbf{p}$. Since $\breve{\text{cl}}(\nu_s^\pi)$ is a lower bound on $v_{\mathbf{p}}^\pi(s)$ and using (8.23) and (8.22) we obtain:

$$v_{\Omega_\epsilon(\hat{\mu}^n)}^\pi(s) \geq \sup_{\lambda\geq 0}\left[\frac{1}{n}\sum_{i=1}^n \inf_{\mathbf{p}} \max_{\|\mathbf{u}_i\|_*\leq\lambda} (\breve{\text{cl}}(\nu_s^\pi)(\mathbf{p}) + \mathbf{u}_i'(\mathbf{p} - \hat{\mathbf{p}}^{(i)})) - \lambda\alpha_s\right]. \tag{8.24}$$

Using the fact that the convex closure of a function has the same convex dual as the function itself, it follows that

$$\begin{aligned}\breve{\text{cl}}(\nu_s^\pi)(\mathbf{p}) =& \breve{\text{cl}}(\nu_s^\pi)^{**}(\mathbf{p}) \\ =& \max_{\mathbf{z}\in\mathcal{D}_s}[\mathbf{z}'\mathbf{p} - \breve{\text{cl}}(\nu_s^\pi)^*(\mathbf{z})] \\ =& \max_{\mathbf{z}\in\mathcal{D}_s}[\mathbf{z}'\mathbf{p} - (\nu_s^\pi)^*(\mathbf{z})] \\ =& \max_{\mathbf{z}\in\mathcal{D}_s}[\mathbf{z}'\mathbf{p} + v_s^{*,\pi}(\mathbf{z})], \end{aligned} \tag{8.25}$$

where the last equation used the definition of the conjugate robust value function (8.20).

Then, using (8.25), the term inside the summation in the RHS of (8.24) can be written as:

$$\begin{aligned}&\inf_{\mathbf{p}} \max_{\|\mathbf{u}_i\|_*\leq\lambda} (\breve{\text{cl}}(\nu_s^\pi)(\mathbf{p}) + \mathbf{u}_i'(\mathbf{p} - \hat{\mathbf{p}}^{(i)})) \\ &= \inf_{\mathbf{p}} \max_{\mathbf{z}_i\in\mathcal{D}_s} \max_{\|\mathbf{u}_i\|_*\leq\lambda} (v_s^{*,\pi}(\mathbf{z}_i) + \mathbf{z}_i'\mathbf{p} + \mathbf{u}_i'(\mathbf{p} - \hat{\mathbf{p}}^{(i)})) \\ &= \max_{\mathbf{z}_i\in\mathcal{D}_s} \max_{\|\mathbf{u}_i\|_*\leq\lambda} [v_s^{*,\pi}(\mathbf{z}_i) - \mathbf{u}_i'\hat{\mathbf{p}}^{(i)} + \inf_{\mathbf{p}} \mathbf{p}'(\mathbf{z}_i + \mathbf{u}_i)], \end{aligned} \tag{8.26}$$

The version of record is available at: http://dx.doi.org/10.1561/2400000026

where the last equality used duality. Note that the minimization in the RHS of the above is over transition probability vectors. We can relax this minimization over all real vectors, which would render a lower bound and result in the infimum being $-\infty$ unless $\mathbf{u}_i = -\mathbf{z}_i$. Note that if $\sup\{\|\mathbf{z}_i\|_*\colon \mathbf{z}_i \in \mathcal{D}_s\} > \lambda$, then one can pick some $\mathbf{z}_i \in \mathcal{D}_s$ such that $\|\mathbf{z}_i\|_* > \lambda$, in which case the inner minimization in (8.26) achieves $-\infty$ since $\mathbf{u}_i \neq -\mathbf{z}_i$. When $\sup\{\|\mathbf{z}_i\|_*\colon \mathbf{z}_i \in \mathcal{D}_s\} \leq \lambda$, we have

$$\begin{aligned}
&\inf_{\mathbf{p}} \max_{\|\mathbf{u}_i\|_* \leq \lambda} (\breve{\mathrm{cl}}(\nu_s^\pi)(\mathbf{p}) + \mathbf{u}_i'(\mathbf{p} - \hat{\mathbf{p}}^{(i)})) \\
&\geq \max_{\mathbf{z}_i \in \mathcal{D}_s} \max_{\|\mathbf{z}_i\|_* \leq \lambda} [v_s^{*,\pi}(\mathbf{z}_i) + \mathbf{z}_i' \hat{\mathbf{p}}^{(i)}] \\
&= v^\pi_{\hat{\mathbf{p}}^{(i)}}(s),
\end{aligned}$$

where the second step follows from the fact that $v_s^{*,\pi}(\mathbf{z}_i)$ is the negative of the convex dual of the value function. It follows that

$$\begin{aligned}
&\inf_{\mathbf{p}} \max_{\|\mathbf{u}_i\|_* \leq \lambda} (\breve{\mathrm{cl}}(\nu_s^\pi)(\mathbf{p}) + \mathbf{u}_i'(\mathbf{p} - \hat{\mathbf{p}}^{(i)})) \\
&\geq \begin{cases} v^\pi_{\hat{\mathbf{p}}^{(i)}}(s), & \text{if } \sup\{\|\mathbf{z}_i\|_*\colon \mathbf{z}_i \in \mathcal{D}_s\} \leq \lambda, \\ -\infty, & \text{otherwise.} \end{cases}
\end{aligned} \tag{8.27}$$

Plugging (8.27) in (8.24) it follows that

$$v^\pi_{\Omega_\epsilon(\hat{\mu}^n)}(s) \geq \frac{1}{n} \sum_{i=1}^n v^\pi_{\hat{\mathbf{p}}^{(i)}}(s) - \kappa \alpha_s,$$

where $\kappa = \sup_{\mathbf{z} \in \mathcal{D}_s} \|\mathbf{z}\|_*$. □

The result of Theorem 8.2.1 provides a lower bound on the distributionally robust value function, which can be used in the RHS of the Bellman equation and in a value iteration scheme. It can also be used in the same manner in obtaining a distributionally robust optimal policy. However, this strategy is applicable in settings where the state-action space is relatively small. For large state-action spaces, one typically approximates either the value function or the policy. To that end, the regularization result Theorem 8.2.1 can be extended to cases where the value function is approximated by a linear function.

In particular, suppose we approximate the value function by $v^\pi_{\mathbf{p}}(s) \approx \boldsymbol{\phi}(s)' \mathbf{w}_{\mathbf{p}}$, where $\boldsymbol{\phi}(s)$ is some feature vector and $\mathbf{w}_{\mathbf{p}}$ a parameter vector.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Similar to (8.20) we can define an *approximate conjugate robust value function* at state s and under policy π as:

$$w_s^{*,\pi}(\mathbf{z}) \stackrel{\triangle}{=} \inf_{\mathbf{p}} (\boldsymbol{\phi}(s)'\mathbf{w}_{\mathbf{p}} - \mathbf{z}'\mathbf{p}), \tag{8.28}$$

and let $\mathcal{W}_s = \{\mathbf{z}\colon w_s^{*,\pi}(\mathbf{z}) > -\infty\}$ be its effective domain.

[190] provides a result analogous to Theorem 8.2.1.

Theorem 8.2.2 [190]. For any policy π, it holds that

$$\inf_{\mu \in \Omega_{\boldsymbol{\epsilon}}(\hat{\mu}^n)} \mathbb{E}_{p\sim\mu}^{\tau\sim\pi}[\boldsymbol{\phi}(s)'\mathbf{w}_{\mathbf{p}}] \geq \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(s)'\mathbf{w}_{\hat{\mathbf{p}}^{(i)}} - \eta\alpha_s, \tag{8.29}$$

where $\alpha_s = \sum_{a\in\mathcal{A}} \pi(a|s)\epsilon_{s,a}$ and $\eta = \sup_{\mathbf{z}\in\mathcal{W}_s} \|\mathbf{z}\|_*$.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

9

Discussion and Conclusions

In this monograph, we developed a Wasserstein-based distributionally robust learning framework for a comprehensive list of predictive and prescriptive problems, including (i) Distributionally Robust Linear Regression (DRLR), (ii) Groupwise Wasserstein Grouped LASSO (GWGL), (iii) Distributionally Robust Multi-Output Learning, (iv) Optimal decision making via DRLR informed K-Nearest Neighbors (K-NN), (v) Distributionally Robust Semi-Supervised Learning, and (vi) Distributionally Robust Reinforcement Learning.

Starting with the basics of the Wasserstein metric and the DRO formulation, we explored its robustness inducing properties, discussed approaches for solving the DRO formulation, and investigated the properties of the DRO solution. Then, we turned our attention into specific learning problems that can be posed and solved using the Wasserstein DRO approach. In each case, we derived equivalent regularized empirical loss minimization formulations and established the robustness of the solutions both theoretically and empirically. We showed novel theoretical results tailored to each setting and validated the methods using real world medical applications, strengthening the notion of robustness through these discussions.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

The robustness of the Wasserstein DRO approach hinges on the fact that a family of distributions that are different from, but close to the empirical measure, are being hedged against. This data-driven formulation not only utilizes the information contained in the observed samples, but also generalizes beyond that by allowing distributions with out-of-sample support. This is a distinguishing feature from DRO approaches based on alternative distance functions, such as ϕ-divergences, which only consider distributions whose support is a subset of the observed samples. Such a limitation could potentially hurt the generalization power of the model. Another salient advantage of the Wasserstein metric lies in its structure, in particular, encoding a distance metric in the data space, which makes it possible to link the form of the regularizer with the growth rate of the loss function and establish a connection between *robustness* and *regularization*.

Our results on Wasserstein DRO and its connection to regularization are not restricted to linear and logistic regression. From the analysis presented in Section 3, we see that as long as the growth rate of the loss function is bounded, the corresponding Wasserstein DRO problem can be made tractable. We consider both static settings, where all the samples are readily accessible when solving for the model (Sections 4–6), and a dynamic setting where the samples come in a sequential manner (Section 8.2). Another example of a dynamic DRO problem is [27], which proposed a distributionally robust Kalman filter that hedges against model risk; in that setting, the Wasserstein ambiguity set contains only normal distributions.

More broadly, researchers have proposed distributionally robust versions for general estimation problems, see, for example, [25] for distributionally robust Minimum Mean Square Error Estimation, [26] for distributionally robust Maximum Likelihood Estimation, which was adopted to estimate the inverse covariance matrix of a Gaussian random vector. We refer the reader to [75] for computational aspects related to Wasserstein distances and optimal transport. [197] and [198] also provided nice overviews of DRO, the former focusing specifically on the Wasserstein DRO, covering in detail the theoretical aspects of the general formulation with a brief discussion on some machine learning

The version of record is available at: http://dx.doi.org/10.1561/2400000026

applications, while the latter covered DRO models with all kinds of ambiguity sets. We summarize our key novel contributions as follows.

- We considered a comprehensive list of machine learning problems, not only predictive models, but also prescriptive models, that can be posed and solved using the Wasserstein DRO framework.
- We presented novel performance guarantees tailored to each problem, reflecting the particularity of the specific problem and providing justifications for using a Wasserstein DRO approach. This is very different from [197], where a universal performance guarantee result was derived. Their result is in general applicable to every single DRO problem, but may miss the individual characteristics of the problem at hand.
- The Wasserstein prescriptive model we presented in Section 7 is novel. We showed the power of Wasserstein DRO through the K-NN insertion in a decision making problem, and demonstrated the benefit of robustness through a novel out-of-sample MSE result.
- The non-trivial extension to multi-output DRO has implications on training robust neural networks, e.g., the robustness of the multiclass logistic regression classifier to optimized perturbations that are designed to fool the classifier, see Section 6.2.3.
- Finally, we considered a variety of synthetic and real world case studies of the respective models, demonstrating the applications of the DRO framework and its superior performance compared to other alternatives, which adds to the accessibility and appeal of this work to an application-oriented reader.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

Acknowledgments

The authors are grateful to Dimitris Bertsimas, Theodora Brisimi, Christos Cassandras, David Castañón, Alex Olshevsky, Venkatesh Saligrama, and Wei Shi, for their insightful comments and constructive suggestions.

We are thankful to the Network Optimization and Control Lab at Boston University for providing computational resources and expertise for some of the case studies. Collaborations on a number of application fronts have involved Michael Caramanis and Pirooz Vakili.

We are grateful to many clinicians and researchers in Boston area hospitals who provided access to data and collaborated in parts of the work, including: Hiroto Hatabu, George Kasotakis, Fania Mela, Rebecca Mishuris, Jenifer Siegelman, Vladimir Valtchinov, and George Velmahos. Particular mention is due to Bill Adams, at Boston Medical Center, whose efforts to make data available for research have been nothing short of extraordinary and who was instrumental in engaging the authors in health analytics research.

We are thankful to the series editors Garud Iyengar, Stephen Boyd, and to the anonymous reviewers for valuable feedback.

RC is grateful to ICP and David Castañón who have provided constant support and encouragement for her, and have been inspirational role models as excellent researchers and teachers with endless positivity and passion. She is also grateful to her parents, Xudong and Shouzhen, and her cousins Yingying, Qianqian, and Chunlei, for their unconditional

The version of record is available at: http://dx.doi.org/10.1561/2400000026

love, support and company, which have given her the strength and determination to overcome difficulties and complete this work.

ICP is grateful to Dimitris Bertsimas and John Tsitsiklis for all they have taught him and for being such inspirational role models for research and the good exposition of research ideas. He is also grateful to his family (Gina, Aris, Phevos, and Alexandros) for their love, support, and giving him the time to work on this project.

The authors are also grateful to Ulrike Fischer, who designed the style files, and Neal Parikh, who laid the groundwork for these style files.

Part of the research included in this monograph has been supported by the NSF under grants IIS-1914792, DMS-1664644, CNS-1645681, CCF-1527292, and IIS-1237022, by the ARO under grant W911NF-12-1-0390, by the ONR under grant N00014-19-1-2571, by the NIH under grants R01 GM135930 and UL54 TR004130, by the DOE under grant DE-AR-0001282, by 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.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

References

[1] L. Breiman, *Classification and Regression Trees*. Routledge, 2017.

[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SigKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[4] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge: MIT Press, 2016.

[5] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[6] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. New York: Springer Series in Statistics, 2001.

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

[8] P. J. Huber, "Robust regression: Asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, vol. 1, no. 5, pp. 799–821, 1973.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[9] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[10] G. Zames, "Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses," *IEEE Transactions on Automatic Control*, vol. 26, no. 2, pp. 301–320, 1981.

[11] K. Zhou and J. C. Doyle, *Essentials of Robust Control*, vol. 104. Upper Saddle River, NJ: Prentice Hall, 1998.

[12] R. Chen and I. C. Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 517–564, 2018.

[13] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1–2, pp. 115–166, 2018.

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

[15] P. M. Esfahani, S. Shafieezadeh-Abadeh, G. A. Hanasusanto, and D. Kuhn, "Data-driven inverse optimization with imperfect information," *Mathematical Programming*, vol. 167, no. 1, pp. 191–234, 2018.

[16] M. Mevissen, E. Ragnoli, and J. Y. Yu, "Data-driven distributionally robust polynomial optimization," in *Advances in Neural Information Processing Systems*, pp. 37–45, 2013.

[17] G. A. Hanasusanto and D. Kuhn, "Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls," *Operations Research*, vol. 66, no. 3, pp. 849–869, 2018.

[18] C. Zhao and Y. Guan, "Data-driven risk-averse stochastic optimization with Wasserstein metric," *Operations Research Letters*, vol. 46, no. 2, pp. 262–267, 2018.

[19] R. Ji and M. Lejeune, "Data-driven distributionally robust chance-constrained optimization with Wasserstein metric," Available at *SSRN 3201356*, 2020.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[20] W. Xie, "On distributionally robust chance constrained programs with Wasserstein distance," *Mathematical Programming*, pp. 1–41, 2019.

[21] B. P. Van Parys, D. Kuhn, P. J. Goulart, and M. Morari, "Distributionally robust control of constrained stochastic systems," *IEEE Transactions on Automatic Control*, vol. 61, no. 2, pp. 430–442, 2015.

[22] I. Yang, "Wasserstein distributionally robust stochastic control: A data-driven approach," *arXiv preprint arXiv:1812.09808*, 2018.

[23] I. Yang, "A dynamic game approach to distributionally robust safety specifications for stochastic systems," *Automatica*, vol. 94, pp. 94–101, 2018.

[24] R. Gao, L. Xie, Y. Xie, and H. Xu, "Robust hypothesis testing using Wasserstein uncertainty sets," in *Advances in Neural Information Processing Systems*, pp. 7902–7912, 2018.

[25] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization," *arXiv preprint arXiv:1911.03539*, 2019.

[26] V. A. Nguyen, D. Kuhn, and P. M. Esfahani, "Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator," *arXiv preprint arXiv:1805.07194*, 2018.

[27] S. S. Abadeh, V. A. Nguyen, D. Kuhn, and P. M. M. Esfahani, "Wasserstein distributionally robust Kalman filtering," in *Advances in Neural Information Processing Systems*, pp. 8474–8483, 2018.

[28] D. Duque and D. P. Morton, "Distributionally robust stochastic dual dynamic programming," *SIAM Journal on Optimization*, vol. 30, no. 4, pp. 2841–2865, 2020.

[29] G. A. Hanasusanto and D. Kuhn, "Robust data-driven dynamic programming," in *Advances in Neural Information Processing Systems*, pp. 827–835, 2013.

[30] A. Sinha, M. O'Kelly, H. Zheng, R. Mangharam, J. Duchi, and R. Tedrake, "FormulaZero: Distributionally robust online adaptation via offline population synthesis," in *International Conference on Machine Learning*, 2020.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[31] C. Shang, X. Huang, and F. You, "Data-driven robust optimization based on kernel learning," *Computers & Chemical Engineering*, vol. 106, pp. 464–479, 2017.

[32] R. Fathony, A. Rezaei, M. A. Bashiri, X. Zhang, and B. Ziebart, "Distributionally robust graphical models," in *Advances in Neural Information Processing Systems*, pp. 8344–8355, 2018.

[33] A. Sinha, H. Namkoong, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," in *International Conference on Learning Representations*, 2018.

[34] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*, vol. 28. Princeton University Press, 2009.

[35] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and applications of robust optimization," *SIAM Review*, vol. 53, no. 3, pp. 464–501, 2011.

[36] A. Ben-Tal and A. Nemirovski, "Selected topics in robust convex optimization," *Mathematical Programming*, vol. 112, no. 1, pp. 125–158, 2008.

[37] D. Bertsimas and M. S. Copenhaver, "Characterization of the equivalence of robustification and regularization in linear and matrix regression," *European Journal of Operational Research*, vol. 270, no. 3, pp. 931–942, 2018.

[38] 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.

[39] H. Xu, C. Caramanis, and S. Mannor, "Robust regression and LASSO," in *Advances in Neural Information Processing Systems*, pp. 1801–1808, 2009.

[40] W. Yang and H. Xu, "A unified robust regression model for LASSO-like algorithms," in *International Conference on Machine Learning*, pp. 585–593, 2013.

[41] D. Bertsimas, J. Dunn, C. Pawlowski, and Y. D. Zhuo, "Robust classification," *INFORMS Journal on Optimization*, vol. 1, no. 1, pp. 2–34, 2018.

[42] A. Liu and B. Ziebart, "Robust classification under sample selection bias," in *Advances in Neural Information Processing Systems*, pp. 37–45, 2014.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[43] L. El Ghaoui, G. R. G. Lanckriet, and G. Natsoulis, "Robust classification with interval data," *Tech. Rep. UCB/CSD-03-1279,* EECS Department, University of California, Berkeley, 2003. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2003/5772.html.

[44] T. B. Trafalis and R. C. Gilbert, "Robust classification and regression using support vector machines," *European Journal of Operational Research*, vol. 173, no. 3, pp. 893–909, 2006.

[45] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[46] R. Gao, X. Chen, and A. J. Kleywegt, "Wasserstein distributional robustness and regularization in statistical learning," *arXiv preprint arXiv:1712.06050*, 2017.

[47] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Regularization via mass transportation," *arXiv preprint arXiv: 1710.10016*, 2017.

[48] J. Goh and M. Sim, "Distributionally robust optimization and its tractable approximations," *Operations Research*, vol. 58, no. 4-part-1, pp. 902–917, 2010.

[49] I. Popescu, "Robust mean-covariance solutions for stochastic optimization," *Operations Research*, vol. 55, no. 1, pp. 98–112, 2007.

[50] 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.

[51] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Research*, vol. 58, no. 3, pp. 595–612, 2010.

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

[53] S. Zymler, D. Kuhn, and B. Rustem, "Distributionally robust joint chance constraints with second-order moment information," *Mathematical Programming*, vol. 137, no. 1–2, pp. 167–198, 2013.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[54] Z. Wang, P. W. Glynn, and Y. Ye, "Likelihood robust optimization for data-driven problems," *Computational Management Science*, vol. 13, no. 2, pp. 241–261, 2016.

[55] S. S. Abadeh, P. M. M. Esfahani, and D. Kuhn, "Distributionally robust logistic regression," in *Advances in Neural Information Processing Systems*, pp. 1576–1584, 2015.

[56] R. Jiang and Y. Guan, "Risk-averse two-stage stochastic program with distributional ambiguity," *Operations Research*, vol. 66, no. 5, pp. 1390–1405, 2018.

[57] C. Zhao and Y. Guan, "Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics," Available on *Optimization Online,* http://www.optimization-online.org/DB_FILE/2015/07/5014.pdf, 2015.

[58] G. Bayraksan and D. K. Love, "Data-driven stochastic programming using phi-divergences," *Tutorials in Operations Research*, pp. 1–19, 2015.

[59] Z. Hu and L. J. Hong, "Kullback–Leibler divergence constrained distributionally robust optimization," Available at *Optimization Online,* http://www.optimization-online.org/DB_FILE/2012/11/3677.pdf, 2013.

[60] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Mathematical Programming*, pp. 1–37, 2015.

[61] E. Erdoğan and G. Iyengar, "Ambiguous chance constrained problems and robust optimization," *Mathematical Programming*, vol. 107, no. 1–2, pp. 37–61, 2006.

[62] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *Mathematics of Operations Research*, vol. 44, no. 2, pp. 565–600, 2019.

[63] F. Luo and S. Mehrotra, "Decomposition algorithm for distributionally robust optimization using Wasserstein metric," *arXiv preprint arXiv:1704.03920*, 2017.

[64] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," *Journal of Applied Probability*, vol. 56, no. 3, pp. 830–857, 2019.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[65] J. Blanchet, P. W. Glynn, J. Yan, and Z. Zhou, "Multivariate distributionally robust convex regression under absolute error loss," *arXiv preprint arXiv:1905.12231*, 2019.

[66] 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.

[67] D. Bertsimas and J. Tsitsiklis, *Introduction to Linear Optimization*. Belmont, MA: Athena Scientific, 1997.

[68] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[69] L. Kantorovich, "On the transfer of masses (in Russian)," in *Doklady Akademii Nauk*, vol. 37, pp. 227–229, 1942.

[70] L. Kantorovich, "On the Monge problem," *Uspekhi Mat. Nauk*, vol. 3, no. 2, pp. 225–226, 1948.

[71] L. Kantorovich, "Mathematical methods of organizing production planning," *Leningrad: Leningrad State University*, 1939.

[72] L. V. Kantorovich, "Mathematical methods of organizing and planning production," *Management Science*, vol. 6, no. 4, pp. 366–422, 1960.

[73] L. V. Kantorovich, "On one effective method of solving certain classes of extremal problems," in *Dokl. Akad. Nauk. USSR*, vol. 28, pp. 212–215, 1940.

[74] C. Villani, *Optimal Transport: Old and New*, vol. 338. Springer Science & Business Media, 2008.

[75] G. Peyré and M. Cuturi, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5–6, pp. 355–607, 2019.

[76] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, "Sliced and Radon Wasserstein barycenters of measures," *Journal of Mathematical Imaging and Vision*, vol. 51, no. 1, pp. 22–45, 2015.

[77] A. Liutkus, U. Simsekli, S. Majewski, A. Durmus, and F.-R. Stöter, "Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions," in *International Conference on Machine Learning*, pp. 4104–4113, 2019.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[78] J. Delon and A. Desolneux, "A Wasserstein-type distance in the space of Gaussian mixture models," *SIAM Journal on Imaging Sciences*, vol. 13, no. 2, pp. 936–970, 2020.

[79] D. Dowson and B. Landau, "The Fréchet distance between multivariate normal distributions," *Journal of Multivariate Analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[80] J. Blanchet, Y. Kang, K. Murthy, and F. Zhang, "Data-driven optimal transport cost selection for distributionally robust optimization," in *2019 Winter Simulation Conference (WSC)*, IEEE, pp. 3740–3751, 2019.

[81] R. Ji and M. Lejeune, "Data-driven optimization of reward-risk ratio measures," Available at *SSRN 2707122*, 2018.

[82] I. N. Sanov, *On the Probability of Large Deviations of Random Variables*. United States Air Force, Office of Scientific Research, 1958.

[83] R. Wang, X. Wang, and L. Wu, "Sanov's theorem in the Wasserstein distance: A necessary and sufficient condition," *Statistics & Probability Letters*, vol. 80, no. 5–6, pp. 505–512, 2010.

[84] R. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[85] M. É. Borel, "Les probabilités dénombrables et leurs applications arithmétiques," *Rendiconti del Circolo Matematico di Palermo (1884–1940)*, vol. 27, no. 1, pp. 247–271, 1909.

[86] F. P. Cantelli, "Sulla probabilità come limite della frequenza," *Atti Accad. Naz. Lincei*, vol. 26, no. 1, pp. 39–45, 1917.

[87] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. John Wiley & Sons, 2005.

[88] D. Coleman, P. Holland, N. Kaden, V. Klema, and S. C. Peters, "A system of subroutines for iteratively reweighted least squares computations," *ACM Transactions on Mathematical Software (TOMS)*, vol. 6, no. 3, pp. 327–336, 1980.

[89] 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.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

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

[91] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.

[92] P. J. Rousseeuw, "Multivariate estimation with high breakdown point," *Mathematical Statistics and Applications*, vol. 8, pp. 283–297, 1985.

[93] P. Rousseeuw and V. Yohai, "Robust regression by means of S-estimators," in *Robust and Nonlinear Time Series Analysis*, pp. 256–272, 1984.

[94] V. J. Yohai, "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, pp. 642–656, 1987.

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

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

[97] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1485–1510, 2009.

[98] 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.

[99] 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.

[100] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[101] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific Belmont, 1999.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[102] S. Chen and A. Banerjee, "Alternating estimation for structured high-dimensional multi-response models," *arXiv preprint arXiv:1606.08957*, 2016.

[103] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2017.

[104] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Reconstruction and sub-Gaussian operators in asymptotic geometric analysis," *Geometric and Functional Analysis*, vol. 17, no. 4, pp. 1248–1282, 2007.

[105] A. Maurer, M. Pontil, and B. Romera-Paredes, "An inequality with applications to structured sparsity and multitask dictionary learning," in *Conference on Computational Learning Theory*, pp. 440–460, 2014.

[106] R. Tibshirani, "Regression shrinkage and selection via the LASSO: A retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[107] L. J. Rogers, "An extension of a certain theorem in inequalities," *Messenger of Math*, vol. 17, no. 2, pp. 145–150, 1888.

[108] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[109] T. Hastie, R. Tibshirani, and R. J. Tibshirani, "Extended comparisons of best subset selection, forward stepwise selection, and the LASSO," *arXiv preprint arXiv:1707.08692*, 2017.

[110] R. Chen, I. C. Paschalidis, H. Hatabu, V. I. Valtchinov, and J. Siegelman, "Detection of unwarranted CT radiation exposure from patient and imaging protocol meta-data using regularized regression," *European Journal of Radiology Open*, vol. 6, pp. 206–211, 2019.

[111] S. Bakin, "Adaptive regression and model selection in data mining problems," Ph.D. thesis. The Australian National University, 1999.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[112] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[113] Y. Lin and H. H. Zhang, "Component selection and smoothing in smoothing spline analysis of variance models," *Annals of Statistics*, vol. 34, no. 5, pp. 2272–2297, 2006.

[114] J. Yin, X. Chen, and E. P. Xing, "Group sparse additive models," in *International Conference on Machine Learning*, NIH Public Access, vol. 2012, p. 871, 2012.

[115] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, pp. 3468–3497, 2009.

[116] L. Jacob, G. Obozinski, and J.-P. Vert, "Group LASSO with overlap and graph LASSO," in *International Conference on Machine Learning*, pp. 433–440, 2009.

[117] Y. Kim, J. Kim, and Y. Kim, "Blockwise sparse regression," *Statistica Sinica*, pp. 375–390, 2006.

[118] L. Meier, S. Van De Geer, and P. Bühlmann, "The group LASSO for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[119] D. Bertsimas and A. King, "Logistic regression: From art to science," *Statistical Science*, vol. 32, no. 3, pp. 367–384, 2017.

[120] V. Roth and B. Fischer, "The group-LASSO for generalized linear models: Uniqueness of solutions and efficient algorithms," in *International Conference on Machine Learning*, pp. 848–855, 2008.

[121] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group LASSO," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[122] F. Bunea, J. Lederer, and Y. She, "The group square-root LASSO: Theoretical properties and fast algorithms," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1313–1325, 2014.

[123] J. Blanchet and Y. Kang, "Distributionally robust groupwise regularization estimator," *arXiv preprint arXiv:1705.04241*, 2017.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[124] G. Obozinski, L. Jacob, and J.-P. Vert, "Group LASSO with overlaps: The latent group LASSO approach," *arXiv:1110.0413*, 2011.

[125] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2777–2824, 2011.

[126] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[127] M. Meila and J. Shi, "Learning segmentation by random walks," in *Advances in Neural Information Processing Systems*, pp. 873–879, 2001.

[128] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, pp. 849–856, 2002.

[129] C. Ding, "A tutorial on spectral clustering," in *Talk Presented at International Conference on Machine Learning*, 2004.

[130] S. Ma, X. Song, and J. Huang, "Supervised group LASSO with applications to microarray data analysis," *BMC Bioinformatics*, vol. 8, no. 1, p. 60, 2007.

[131] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[132] H. Zhang, H. Zhao, J. Sun, D. Wang, and K. Kim, "Regression analysis of multivariate panel count data with an informative observation process," *Journal of Multivariate Analysis*, vol. 119, pp. 71–80, 2013.

[133] B. Hidalgo and M. Goodman, "Multivariate or multivariable regression?" *American Journal of Public Health*, vol. 103, no. 1, pp. 39–40, 2013.

[134] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang, "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," *The Annals of Applied Statistics*, vol. 4, no. 1, p. 53, 2010.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[135] F. Islam, M. Shahbaz, A. U. Ahmed, and M. M. Alam, "Financial development and energy consumption nexus in Malaysia: A multivariate time series analysis," *Economic Modelling*, vol. 30, pp. 435–441, 2013.

[136] R. S. Tsay, *Multivariate Time Series Analysis: With R and Financial Applications*. John Wiley & Sons, 2013.

[137] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[138] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multiple linear regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 1, pp. 3–54, 1997.

[139] N. Plath, M. Toussaint, and S. Nakajima, "Multi-class image segmentation using conditional random fields and global classification," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 817–824, 2009.

[140] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1289–1305, 2003.

[141] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.

[142] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," *Journal of Multivariate Analysis*, vol. 5, no. 2, pp. 248–264, 1975.

[143] R. Velu and G. C. Reinsel, *Multivariate Reduced-Rank Regression: Theory and Applications*, vol. 136. Springer Science & Business Media, 2013.

[144] W. F. Massy, "Principal components regression in exploratory statistical research," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 234–256, 1965.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[145] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, "Dimension reduction and coefficient estimation in multivariate linear regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 3, pp. 329–346, 2007.

[146] P. J. Brown and J. V. Zidek, "Adaptive multivariate ridge regression," *The Annals of Statistics*, vol. 8, no. 1, pp. 64–74, 1980.

[147] Y. Haitovsky, "On multivariate ridge regression," *Biometrika*, vol. 74, no. 3, pp. 563–570, 1987.

[148] M. Aly, "Survey on multiclass classification methods," *Neural Networks*, vol. 19, pp. 1–9, 2005.

[149] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[150] S. D. Bay, "Combining nearest neighbor classifiers through multiple feature subsets," in *International Conference on Machine Learning*, vol. 98, pp. 37–45, 1998.

[151] I. Rish, "An empirical study of the naive Bayes classifier," in *International Joint Conferences on Artificial Intelligence (IJCAI) Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, pp. 41–46, 2001.

[152] S. Kumar, J. Ghosh, and M. M. Crawford, "Hierarchical fusion of multiple classifiers for hyperspectral data analysis," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 210–220, 2002.

[153] J. Feng, H. Xu, S. Mannor, and S. Yan, "Robust logistic regression and classification," in *Advances in Neural Information Processing Systems*, pp. 253–261, 2014.

[154] N. Ding, S. Vishwanathan, M. Warmuth, and V. S. Denchev, "T-logistic regression for binary and multiclass classification," *The Journal of Machine Learning Research*, vol. 5, pp. 1–55, 2013.

[155] J. Tibshirani and C. D. Manning, "Robust logistic regression using shift parameters," *arXiv preprint arXiv:1305.4987*, 2013.

[156] J. Bootkrajang and A. Kabán, "Label-noise robust logistic regression and its applications," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 143–158, 2012.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[157] H. Masnadi-Shirazi, V. Mahadevan, and N. Vasconcelos, "On the design of robust classifiers for computer vision," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 779–786, 2010.

[158] D. Pregibon, "Resistant fits for some commonly used logistic models with medical application," *Biometrics*, vol. 38, no. 2, pp. 485–498, 1982.

[159] W. Hu, G. Niu, I. Sato, and M. Sugiyama, "Does distributionally robust supervised learning give robust classifiers?" *arXiv preprint arXiv:1611.02041*, 2016.

[160] R. Tomioka and T. Suzuki, "Convex tensor decomposition via structured Schatten norm regularization," in *Advances in Neural Information Processing Systems*, pp. 1331–1339, 2013.

[161] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[162] D. Bertsimas and N. Kallus, "From predictive to prescriptive analytics," *Management Science*, 2019. DOI: 10.1287/mnsc.2018.3253.

[163] D. Den Hertog and K. Postek, "Bridging the gap between predictive and prescriptive analytics-new optimization methodology needed," Tech. Rep., Technical report, Tilburg University, Netherlands, 2016. Available at: http://www.optimization-online.org/DB_HTML/2016/12/5779.html, 2016.

[164] F. Bravo and Y. Shaposhnik, "Mining optimal policies: A pattern recognition approach to model analysis," Available at *SSRN 3069690*, 2018.

[165] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.

[166] D. Bertsimas, N. Kallus, A. M. Weinstein, and Y. D. Zhuo, "Personalized diabetes management using electronic medical records," *Diabetes Care*, vol. 40, no. 2, pp. 210–217, 2017.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[167] D. Bertsimas, J. Dunn, and N. Mundru, "Optimal prescriptive trees," *INFORMS Journal on Optimization*, vol. 1, no. 2, pp. 91–183, 2019.

[168] J. Dunn, "Optimal trees for prediction and prescription," Ph.D. thesis. Massachusetts Institute of Technology, 2018.

[169] M. Biggs and R. Hariss, "Optimizing objective functions determined from random forests," Available at *SSRN 2986630*, 2018.

[170] D. Bertsimas and C. McCord, "Optimization over continuous and multi-dimensional decisions with observational data," *arXiv preprint arXiv:1807.04183*, 2018.

[171] D. Bertsimas and B. Van Parys, "Bootstrap robust prescriptive analytics," *arXiv preprint arXiv:1711.09974*, 2017.

[172] H. Bastani and M. Bayati, "Online decision making with high-dimensional covariates," *Operations Research*, vol. 68, no. 1, pp. 276–294, 2020.

[173] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire, "Taming the monster: A fast and simple algorithm for contextual bandits," in *International Conference on Machine Learning*, pp. 1638–1646, 2014.

[174] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

[175] A. Slivkins, "Contextual bandits with similarity information," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2533–2568, 2014.

[176] H. Wu, R. Srikant, X. Liu, and C. Jiang, "Algorithms with logarithmic or sublinear regret for constrained contextual bandits," in *Advances in Neural Information Processing Systems*, pp. 433–441, 2015.

[177] A. Tewari and S. A. Murphy, "From ads to interventions: Contextual bandits in mobile health," in *Mobile Health*, pp. 495–517, 2017.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[178] I. Xia, "The price of personalization: An application of contextual bandits to mobile health," Ph.D. thesis. Harvard University, 2018.

[179] F. Zhu, J. Guo, R. Li, and J. Huang, "Robust actor-critic contextual bandit for mobile health (mhealth) interventions," in *ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 492–501, 2018.

[180] E. Hazan, "Introduction to online convex optimization," *Foundations and Trends® in Optimization*, vol. 2, no. 3–4, pp. 157–325, 2016.

[181] J.-B. Alayrac, J. Uesato, P.-S. Huang, A. Fawzi, R. Stanforth, and P. Kohli, "Are labels required for improving adversarial robustness?" In *Advances in Neural Information Processing Systems*, pp. 12 214–12 223, 2019.

[182] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," in *Advances in Neural Information Processing Systems*, pp. 11 192–11 203, 2019.

[183] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang, "Adversarial training can hurt generalization," *arXiv preprint arXiv:1906.06032*, 2019.

[184] R. Zhai, T. Cai, D. He, C. Dan, K. He, J. Hopcroft, and L. Wang, "Adversarially robust generalization just requires more unlabeled data," *arXiv preprint arXiv:1906.00555*, 2019.

[185] Y. Yan, Z. Xu, I. W. Tsang, G. Long, and Y. Yang, "Robust semi-supervised learning through label aggregation," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[186] J. Blanchet and Y. Kang, "Distributionally robust semi-supervised learning," *arXiv preprint arXiv:1702.08848*, 2017.

[187] C. Frogner, S. Claici, E. Chien, and J. Solomon, "Incorporating unlabeled data into distributionally robust learning," *arXiv preprint arXiv:1912.07729*, 2019.

[188] A. Najafi, S.-I. Maeda, M. Koyama, and T. Miyato, "Robustness to adversarial perturbations in learning from incomplete data," in *Advances in Neural Information Processing Systems*, pp. 5541–5551, 2019.

The version of record is available at: http://dx.doi.org/10.1561/2400000026

[189] O. Reynolds, A. W. Brightmore, and W. H. Moorby, *Papers on Mechanical and Physical Subjects: The Sub-Mechanics of the Universe*, vol. 3. The University Press, 1903.

[190] E. Derman and S. Mannor, "Distributional robustness and regularization in reinforcement learning," *arXiv preprint arXiv:2003.02894*, 2020.

[191] J. K. Satia and R. E. Lave Jr, "Markovian decision processes with uncertain transition probabilities," *Operations Research*, vol. 21, no. 3, pp. 728–740, 1973.

[192] C. C. White III and H. K. Eldeib, "Markov decision processes with imprecise transition probabilities," *Operations Research*, vol. 42, no. 4, pp. 739–749, 1994.

[193] J. Bagnell, A. Y. Ng, and J. Schneider, "Solving uncertain Markov decision problems," *Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-01-25*, 2001.

[194] G. Iyengar, "Robust dynamic programming," *Math. Operations Research*, vol. 30, no. 2, pp. 1–21, 2005.

[195] A. Nilim and L. E. Ghaoui, "Robust solutions to Markov decision problems with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[196] Z. Chen, P. Yu, and W. B. Haskell, "Distributionally robust optimization for sequential decision-making," *Optimization*, vol. 68, no. 12, pp. 2397–2426, 2019.

[197] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations Research & Management Science in the Age of Analytics*, pp. 130–166, INFORMS, 2019.

[198] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019.