Distributionally Robust Learning

Other titles in Foundations and Trends® in Optimization

Atomic Decomposition via Polar Alignment: The Geometry of Structured Optimization
Zhenan Fan, Halyun Jeong, Yifan Sun and Michael P. Friedlander
ISBN: 978-1-68083-742-1

Optimization Methods for Financial Index Tracking: From Theory to Practice
Konstantinos Benidis, Yiyong Feng and Daniel P. Palomar
ISBN: 978-1-68083-464-2

The Many Faces of Degeneracy in Conic Optimization
Dmitriy Drusvyatskiy and Henry Wolkowicz
ISBN: 978-1-68083-390-4

Distributionally Robust Learning

Ruidi Chen
Boston University
USA
rchen15@bu.edu

Ioannis Ch. Paschalidis
Boston University
USA
yannisp@bu.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Optimization

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

R. Chen and I. Ch. Paschalidis. *Distributionally Robust Learning*. Foundations and Trends® in Optimization, vol. 4, no. 1–2, pp. 1–243, 2020.

ISBN: 978-1-68083-773-5

Foundations and Trends® in Optimization
Volume 4, Issue 1-2, 2020

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Optimization publishes survey and tutorial articles in the following topics:

- algorithm design, analysis, and implementation (especially, on modern computing platforms
- models and modeling systems, new optimization formulations for practical problems
- applications of optimization in machine learning, statistics, and data analysis, signal and image processing, computational economics and finance, engineering design, scheduling and resource allocation, and other areas

Information for Librarians

Foundations and Trends® in Optimization, 2020, Volume 4, 4 issues. ISSN paper version 2167-3888. ISSN online version 2167-3918. Also available as a combined paper and online subscription.

Contents

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.

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

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

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.

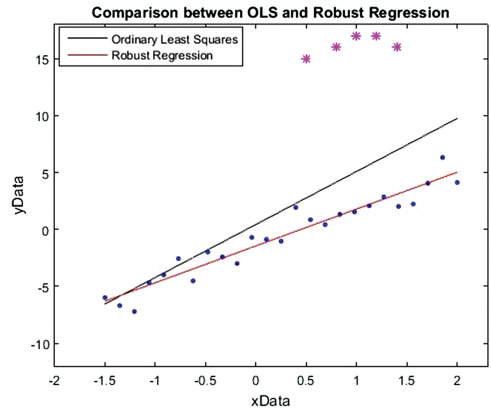

Figure 1.1: Regression example.

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 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:

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

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

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

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:

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

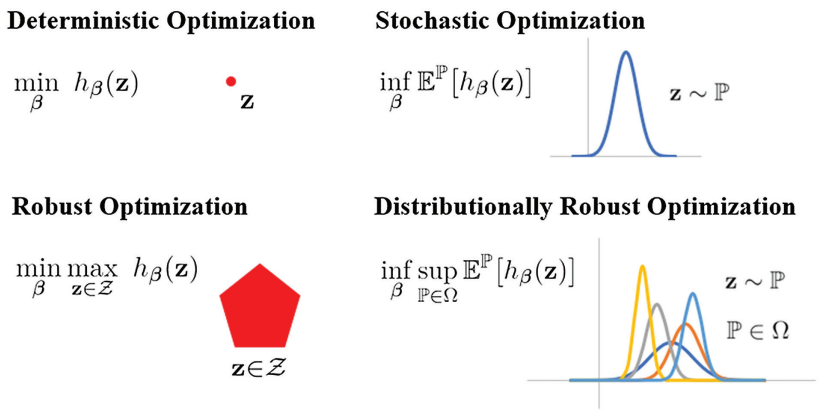

Figure 1.2: Comparison of robust optimization with distributionally robust optimization.

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

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 ($t \geq 1$)

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

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.

- 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

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.

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 $|\mathbf{x}|^{p/2} = (|x_1|^{p/2}, \ldots, |x_{\dim(\mathbf{x})}|^{p/2})$.

- 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.

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

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

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

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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.