Distributionally Robust Optimization and its Applications in Machine Learning

Yang Kang

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Distributionally Robust Optimization and its Applications in Machine Learning

Yang Kang

The goal of Distributionally Robust Optimization (DRO) is to minimize the cost of running a stochastic system, under the assumption that an adversary can replace the underlying baseline stochastic model by another model within a family known as the distributional uncertainty region. This dissertation focuses on a class of DRO problems which are data-driven, which generally speaking means that the baseline stochastic model corresponds to the empirical distribution of a given sample.

One of the main contributions of this dissertation is to show that the class of data-driven DRO problems that we study unify many successful machine learning algorithms, including square root Lasso, support vector machines, and generalized logistic regression, among others. A key distinctive feature of the class of DRO problems that we consider here is that our distributional uncertainty region is based on optimal transport costs. In contrast, most of the DRO formulations that exist to date take advantage of a likelihood based formulation (such as Kullback-Leibler divergence, among others). Optimal transport costs include as a special case the so-called Wasserstein distance, which is popular in various statistical applications.

The use of optimal transport costs is advantageous relative to the use of divergence-based formulations because the region of distributional uncertainty contains distributions which explore samples outside of the support of the empirical measure, therefore explaining why many machine learning algorithms have the ability to improve generalization. Moreover, the DRO representations that we use to unify the previously mentioned machine learning algorithms, provide a clear interpretation of the so-called

regularization parameter, which is known to play a crucial role in controlling generalization error. As we establish, the regularization parameter corresponds exactly to the size of the distributional uncertainty region.

Another contribution of this dissertation is the development of statistical methodology to study data-driven DRO formulations based on optimal transport costs. Using this theory, for example, we provide a sharp characterization of the optimal selection of regularization parameters in machine learning settings such as square-root Lasso and regularized logistic regression.

Our statistical methodology relies on the construction of a key object which we call the robust Wasserstein profile function (RWP function). The RWP function similar in spirit to the empirical likelihood profile function in the context of empirical likelihood (EL). But the asymptotic analysis of the RWP function is different because of a certain lack of smoothness which arises in a suitable Lagrangian formulation.

Optimal transport costs have many advantages in terms of statistical modeling. For example, we show how to define a class of novel semi-supervised learning estimators which are natural companions of the standard supervised counterparts (such as square root Lasso, support vector machines, and logistic regression). We also show how to define the distributional uncertainty region in a purely data-driven way. Precisely, the optimal transport formulation allows us to inform the shape of the distributional uncertainty, not only its center (which given by the empirical distribution). This shape is informed by establishing connections to the metric learning literature. We develop a class of metric learning algorithms which are based on robust optimization. We use the robust-optimization-based metric learning algorithms to inform the distributional uncertainty region in our data-driven DRO problem. This means that we endow the adversary with additional which force him to spend effort on regions of importance to further improve generalization properties of machine learning algo-

rithms.

In summary, we explain how the use of optimal transport costs allow constructing what we call double-robust statistical procedures. We test all of the procedures proposed in this paper in various data sets, showing significant improvement in generalization ability over a wide range of state-of-the-art procedures.

Finally, we also discuss a class of stochastic optimization algorithms of independent interest which are particularly useful to solve DRO problems, especially those which arise when the distributional uncertainty region is based on optimal transport costs.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I wish to express my most sincere thanks to Dr. Jose Blanchet. He is the ideal supervisor that I can imagine before I start my Ph.D. study. Jose has unlimited energy and enthusiasm for exploring and working on research while he is extremely patient and skillful in instructing and inspiring his students. As the old saying in China, "Give a man a fish, and you feed him for a day. Teach a man to fish, and you feed him for a lifetime.", I could still remember when I start working on research, Jose gave me a baby project to teach me how to explore, come up with meaningful questions, and solve the problems in a proper scientific way. Jose knew the solution to the baby project in advance and utilized as an instructive example to work on research systematically. This is only one of the examples, among many, I benefit from working with Jose. In additional to being a great supervisor, he is more like a godfather and like a friend. His great passions in life and his deep love to his family shows me a great successful role model of finding an optimal work-life balance. I believe it is god's will that leads me to Jose. It is with his consistent and generous support that I can have a great Ph.D. life in the past four years both in academia study and everyday life in New York. I also would like to appreciate Jose's family, Lalli, Martin, and Victoria. Without their kind support and understanding, Jose could hardly spend that much time working and exploring with us.

I would like to express my greatest appreciation to my committee members. I could still remember the time when Dr. Richard Davis (former director of the graduate

study of the department of statistics at that time) sent me the email of admission to the Ph.D. program. I took the offer within an hour, and after four years I have proved that I was making the optimal decision. It is under Richard's administration that the department is becoming stronger and stronger. Dr. Paul Galsserman was also my committee of the oral exam, thanks to his thoughtful question of suggestions, we can complete some further promising research works in the last few months of my Ph.D. study. Dr. Peter Orbanz, who is my neighbor, is more than a friend rather than a supervisor. I appreciate for his time chatting with me at statistics lounge on research and everyday life. I first know Dr. Daniel Hsu from taking his class of Advanced Machine Learning, actually, that was one of the primary reasons that I would like to move my research focus to machine learning then. A lot of intuitions and motivations for my current research was learned from Daniel's class.

One of the greatest things I love Columbia is the great supportive help from the faculties members and staff of Department of Statistics and also Department of Industrial Engineering and Operations Research. First, I would like to thank Dood Kalicharan. Dood is like the mother of the department, and she helps me taking care of most of everything of my life at Columbia. She is like the angel that God sends to me. When I have the troubles of housing, application for the visa, teaching, graduating... she always says "No worries, let us fix it now!" and help me solve all the problems efficiently. I would like to show special appreciation for Dr. Mark Brown, for his supportive encouragement and instructive suggestions. I also thank Dr. Tian Zheng, for the opportunity staying and presenting in her study group during the past four years, from which I learned a lot to gain intuitions and ideas for this dissertation. In additional, I would give my gratitude to the other faculties members, Dr. Zhiliang Ying, Dr. Yang Feng, Dr. Jingcheng Liu, Dr. Arian Maleki, Dr. Garud Iyengar, Dr. Mark Brown, Dr. Victor de la Pena, Dr. Henry Lam, among others,

for their instructions and supports in both my study and life. I also appreciate the help from the other staff members at Department of Statistics and IE&OR, without their support I would have suffered much more during my Ph.D. student life. I also appreciate my supervisors, Dr. Fabio Mercurio and Dr. Gordon Ritter, during my summer internship studies. Their in-depth knowledge and a comprehensive view of the industry have a great impact on me both for my academic research and career development.

I am very grateful for being part of an active research group. I would like to thank Dr. Karthyek Murthy, Fan Zhang, Yanan Pei, Zhipeng Liu, Lin Chen, Fei He, and Chris Dolan for the motivative discussions on research and hard efforts for your joint research work. I would like to thank all my friends during my Ph.D. study, Jing Wu, Xiaopei Zhang, Ji Xu, Xiaowei Tan, Fengpei Li, Swapnil Sahai, Julia Yang, Haolei Weng, Lu Meng, Sihan Huang, Chaai Wu, Yilong Zhang, Zhangyi Hu, Fan Zhang, Karthyek Murthy, Jiqun Tu, Morgane Austern, among others, for their kind help and support. They made the years long Ph.D. study life enjoyable and colorful.

Finally, I would like to thank my parents, Hong Gao and Yong Kang, and my grandparents, Yongyu Gao and Jingsheng Fu, for their unconditional love and encouraging support. It is their more than twenty years education that makes me here, without them, I could hardly imagine how could I suffer the tough times and reach each goal. I would appreciate the support from the rest of my family, Christina, Jeanna, Mei, Yuan, and David.

To My Family.

Chapter 1

Introduction

Distributionally Robust Optimization (DRO) refers to a class of optimization problems in which the objective is to minimize the cost of running a stochastic system, under the assumption that an adversary can replace the underlying baseline stochastic model by another model within a family known as the distributional uncertainty region. More specifically, let $l(w, \beta)$ be a realized cost when a decision β is taken and some (stochastic outcome w) occurs. Consider a stochastic optimization problem of the form

$$\min_{\beta} \mathbb{E}_{P_*} [l(W, \beta)], \tag{1.1}$$

where $W \sim P_*$ (the symbol $\sim$ reads "follows the distribution P") and $\mathbb{E}_{P_*}$ is used to denote the expectation with respect to (w.r.t.) the probability measure P_*. The DRO formulation for Equation (1.1) is

$$\min_{\beta} \max_{P \in \mathcal{U}} \mathbb{E}_{P} [l(W, \beta)], \tag{1.2}$$

where we denote $\mathcal{U}$ as the distributional uncertainty set of this DRO problem (which is composed of probability models which govern the distribution of W). The intuition

is that P_* is not fully known and therefore it makes sense to choose β taking into account such ambiguity in our knowledge of P_*. DRO has been actively studied in past decades, see for example Scarf *et al.* [1958]; Ben-Tal and Nemirovski [1998]; Shapiro and Kleywegt [2002]; Iyengar [2005]; Calafiore and Ghaoui [2006]; Erdoğan and Iyengar [2006]; Delage and Ye [2010]; Goh and Sim [2010]; Bertsimas *et al.* [2010]; Ben-Tal *et al.* [2010]; Becker [2011]; Dupačová and Kopa [2012]; Ben-Tal *et al.* [2013]; Wiesemann *et al.* [2014]; Bertsimas *et al.* [2013]; Wang *et al.* [2016b]; Peyré *et al.* [2016]; Lam and Zhou [2017], and has found applications in areas such as finance and risk management (see in Calafiore [2007]; Lam and Zhou [2015]; Hall *et al.* [2015]; Glasserman and Yang [2016]), and machine learning (see for example Ruckdeschel [2010]; Zhu and Fukushima [2009]; Zymler [2010]; Shafieezadeh-Abadeh *et al.* [2015]; Blanchet *et al.* [2016b]; Blanchet and Kang [2017b,a]), among others.

The goal of this dissertation is to develop a comprehensive statistical methodology for data-driven DRO formulations such as (1.2). By data-driven DRO we understand that $\mathcal{U}$ is informed by empirical samples $\mathcal{D}_n = \{W_i\}_{i=1}^n$ of the underlying model P_* (which is unknown). A natural way to incorporate this information is to parameterize the "center" of $\mathcal{U}$ using the empirical measure $P_n = n^{-1}\sum_{i=1}^n \delta_{\{W_i\}}(dw)$. Moreover, we shall introduce a notion of discrepancy between any two probability measures P and Q and we will denote such discrepancy by $D_c(P,Q)$. Using this notation, we then let

$$\mathcal{U} = \mathcal{U}_\delta(P_n) = \{P : D_c(P, P_n) \leq \delta\}.$$

In pursuit of the stated goal, this dissertation sets as its objective to answer the following questions:

A) How to choose the discrepancy measure D_c and what are the advantages of our choice?

B) How to choose the size of the uncertainty region, δ?

C) Is there a way to inform the shape of the uncertainty region $\mathcal{U}$ in a data-driven way (not only through its center)?

D) Does the method generate new statistical insights?

E) What are the computational challenges that formulations such as (1.2) exposes, and how to address them?

F) Finally, what type of future extensions can be envisioned by this new methodology?

Throughout the rest of this Introduction, we provide a summary which explains how these questions are addressed in this dissertation and also we provide forward references to the chapters in which our discussion about these questions is elaborated.

We will introduce the optimal transport cost and briefly discuss the reason for selecting the optimal transport in Section 1.1, this addresses the point **A)** and partially point **C)**. In Section 1.2, we address **B)**, there we discuss the role of uncertainty set size δ via making connection to regularization parameters. Then we introduce an optimality criterion, rooted in statistical principles, for choosing δ. In order to optimally evaluate δ, we introduce two classes of inference procedures, which we call RWPI (Robust Wasserstein Profile Inference) and SoS (Sample-out-of-Sample) inference. In Section 1.3, we explore the flexibility of choosing optimal transport costs. We discuss by a judicious choice of such optimal transport cost, we can generate novel learning methods; for example semi-supervised learning. This discussion in Section 1.3 addresses the question **D)** and **E)**. We discuss briefly the challenges and introduce our algorithm to solve data-driven DRO problems directly in Section 1.4, which addresses **E)**. We discuss the potential future applications of our developments, for example, in multi-task learning in Section 1.5; this addresses point **F)**.

1.1 How to choose the discrepancy and why?

Most of the DRO formulations that exist to date take advantage of likelihood based constructions, such as $\phi-$divergence-based discrepancy measures, Calafiore [2007]; Ben-Tal *et al.* [2010, 2013]; Hu and Hong [2013]; Klabjan *et al.* [2013], which take the form

$$D\left(P,Q\right)=\mathbb{E}_{Q}\left[\phi\left(dP(X)/dQ(X)\right)\right],$$

for a strictly convex function satisfying $\phi\left(1\right)=0$. For example, if you take $\phi(\cdot)=-\log\left(\cdot\right)$, this is known as Kullback-Leibler divergence. For our data-driven DRO formulation, $\mathcal{U}$ is centered the empirical measure, i.e. $Q=P_n$. The definition of $\phi-$divergence discrepancy requires P to be absolute continuous w.r.t. P_n. In simple words, the support of P must be a subset of the support of P_n. This constrain on the support of the elements inside the uncertainty region $\mathcal{U}$ can potentially diminish the power of the DRO formulation, specially in statistical applications in which it is important to enhance out-of-sample performance.

In this dissertation we advocate the use of optimal transport based discrepancies. We would show via some examples that our choice of optimal transport cost as discrepancy recovers some popular algorithms in machine learning which have been studied and whose out-of-sample performance has been widely tested empirically. However, before we discuss such examples, let us introduce the concept of optimal transport cost or optimal transport discrepancy.

Introducing Optimal Transport Costs

An optimal transportation cost is also known as an earth moving distance in the image processing literature (see in Rubner *et al.* [1998, 2000]; Rubner and Tomasi [2001]; Wang *et al.* [2016a]). Intuitively speaking, as its name suggests, the optimal transport cost $D_c(P,Q)$ is measuring the cheapest way of rearranging (i.e. transport-

ing the mass of) distribution P into the distribution Q, where the cost for moving a unit from location u to w is defined as $c(u, w)$.

Normally, we assume the cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ is lower semi-continuous and we assume $c(u, w) = 0$ if and only if $u = w$. Given two probability distributions $P(\cdot)$ and $Q(\cdot)$, with supports $\mathcal{S}_P \subseteq \mathbb{R}^{d+1}$ and $\mathcal{S}_Q \subseteq \mathbb{R}^{d+1}$, respectively, one can define the optimal transport discrepancy (or optimal transport cost) between P and Q, denoted by $D_c(P, Q)$, as

$$D_c(P, Q) = \min_{\pi} \left\{ \mathbb{E}_\pi [c(U, W)] : \pi \in \mathcal{P}\left(\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}\right), \pi_U = P, \pi_W = Q \right\}. \quad (1.3)$$

We denote $\mathcal{P}\left(\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}\right)$ to be set of joint probability measures π supported on a subset of $\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}$, and π_U and π_W denote the marginals of U and W under π, respectively.

In addition to what we stated for the cost function above, if $c(\cdot)$ is symmetric, (i.e. $c(u, w) = c(w, u)$) and there exist $\varrho \geq 1$ such that the triangle inequality holds for $c^{1/\varrho}(\cdot)$, i.e.

$$c^{1/\varrho}(u, w) \leq c^{1/\varrho}(u, v) + c^{1/\varrho}(v, w),$$

for all $u, w, v \in \mathbb{R}^{d+1}$, it can be easily verified that $D_c(P, Q)^{1/\varrho}$ is a metric for probability measures supported on $\mathbb{R}^{d+1}$; this corresponds to the Wasserstein metric of order ϱ (see Villani [2003, 2008] for basic properties of optimal transport costs and other metric properties).

For example, if $c(u, w) = \|u - w\|_2^2$, where $\|\cdot\|_2$ is the Euclidean distance in $\mathbb{R}^m$, then $\rho = 2$ yields that $c(u, w)^{1/2} = \|u - w\|_2$ is symmetric, non-negative, lower semi-

continuous and it satisfies the triangle inequality. In that case,

$$\mathcal{D}_c^{1/2}(P,Q) = \inf\left\{\sqrt{\mathbb{E}_\pi \|U-W\|_2^2} : \ \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m),\ \pi_U = P,\ \pi_W = Q\right\}$$

coincides with the Wasserstein distance of order 2.

Wasserstein distances metricize weak convergence of probability measures under suitable moment assumptions, and have received immense attention in probability theory (see Rachev and Rüschendorf [1998a,b]; Villani [2008] for a collection of classical applications). More recently, optimal transport metrics and Wasserstein distances are being actively investigated for its use in various machine learning applications as well (see Seguy and Cuturi [2015]; Peyré *et al.* [2016]; Rolet *et al.* [2016]; Solomon *et al.* [2015]; Frogner *et al.* [2015]; Srivastava *et al.* [2015] and references therein for a growing list of new applications).

We can observe that optimal transport discrepancies can be obtained via solving a linear programming problem. For example, let us consider a special case, where $Q = P_n$ and we restrict the support of P, i.e. $\mathcal{S}(P)$, to be finite, then, we have that $D_c(P, P_n)$ is obtained by computing

$$\begin{aligned} &\min_{\pi} \sum_{u\in\mathcal{S}_P}\sum_{w\in\mathcal{D}_n} c(u,w)\,\pi(u,w) : \\ \text{s.t.} &\sum_{u\in\mathcal{S}_P} \pi(u,w) = \frac{1}{n}\ \forall\ w \in \mathcal{D}_n \\ &\sum_{w\in\mathcal{D}_N} \pi(u,w) = P(\{u\})\ \ \forall\ u \in \mathcal{X}_N, \\ &\pi(u,w) \geq 0\ \forall\ \ (u,w) \in \mathcal{S}_P \times \mathcal{D}_n \end{aligned} \tag{1.4}$$

For the general case (i.e. the case in which U and W are supported in arbitrary subsets of $\mathbb{R}^{d+1}$), a completely analogous linear program (LP), albeit an infinite dimensional

one, can be defined. Such an infinite dimensional LP has been extensively studied in great generality in the context of Optimal Transport under the name of Kantorovich's problem (see in Villani [2008]). Requiring $c(\cdot)$ to be lower semi-continuous guarantees the existence of an optimal solution to Kantorovich's problem. Requiring that $c(u, w) = 0$ if and only if $u = w$ implies that $D_d(P, Q) = 0$ if and only if $P = Q$.

In order to motivate the choice of optimal transport cost as a reasonable selection for data-driven DRO. We now explain discuss how, by choosing $c(\cdot)$ judiciously we can recover some well-known statistical learning methods which improving generalization (i.e. out-of-sample) performance.

Let consider a linear regression mode of the form

$$Y = \beta_*^T X + e,$$

where β_* is the true regression parameter and e is the independent mean zero random error. We assume the predictors are $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ is the response. Moreover, we have a collection of data samples $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$. A standard statistical approach is to use least squares, which consists in consider the problem

$$\min_\beta \mathbb{E}_{P_n}\left[\left(Y - \beta^T X\right)^2\right] = \min_\beta n^{-1} \sum_{i=1}^n \left(Y_i - \beta^T X_i\right)^2,$$

where

$$P_n(dx, dy) = n^{-1} \sum_{i=1}^n \delta_{\{(X_i, Y_i)\}}(dx, dy).$$

However, as it has been argued in most of the statistical learning textbooks (for example Friedman *et al.* [2001]; Bishop [2006]; James *et al.* [2013]; Goodfellow *et al.* [2016]), when the sample size is relative small relative to the dimension of the problem, direct use of least squares estimation will lead to overfitting and therefore

to poor generalization properties.

In order to enhance the generalization properties of the standard least squares estimator, let us consider a DRO formulation based on optimal transport discrepancies. We consider the cost function

$$c\big((x,y),(u,v)\big) = \begin{cases} \|x-u\|_\infty^2, & \text{if } y = v \\ \infty, & \text{otherwise.} \end{cases} . \tag{1.5}$$

This cost function $c(\cdot)$ assigns infinite cost when $y \neq v$, the minimization in Equation (1.3) is effectively over the joint distributions that do not alter the marginal distributions of Y. As a consequence, the resulting neighborhood set $\mathcal{U}_\delta(P_n) = \{P : D_c(P, P_n) \leq \delta\}$ admits distributional ambiguities only with respect to the predictors X. Intuitively, we are imposing a certain consistency property in which we predictors which are close should share the same response. Not allowing uncertainty in Y may be more sensible in cases in which Y is a categorical variable.

By taking the cost function as in Equation (1.5), we can show that the data-driven DRO formulation for linear regression is equivalent to the square-root Lasso (SR-Lasso) estimator,

$$\begin{aligned} &\min_\beta \max_{P: D_c(P,P_n)\leq\delta} \sqrt{\mathbb{E}\left[(Y - X^T\beta)^2\right]} \\ =&\min_\beta \left\{ \sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - X_i^T\beta)^2} + \sqrt{\delta}\,\|\beta\|_1 \right\}. \end{aligned}$$

SR-Lasso was introduced by Belloni *et al.* [2011] as a generalization of the Lasso method (see Tibshirani [1996]). It turns out that SR-Lasso has the benefit that the optimal choice of regularization parameter is free of the magnitude of the variance of the random error. This is particularly appearning in high dimension settings in

which the estimation of the error variance magnitude may be noisy.

A similar data-driven DRO representation could also be made for regularized logistic regression and support vector machine (SVM), among others, as we shall discuss in Chapter 2 Section 2.3. We also discuss futher generalizations, for example, we will establish explicit connections to Group Lasso and adaptive Lasso estimators. These connections will be discussed in Chapter 5 and Chapter 6.

These regularized estimators have been wildly studied and they have been shown empirically to be highly effective in improving generalization performanc. We believe that the explicit connection to a wide range of successful regularization estimators studied in this dissertation makes a strong case for the use of data-driven DRO with optimal transport costs.

1.2 How to choose the uncertainty region size δ?

Let us consider the data-driven DRO for general statistical learning model with loss function $l(\cdot)$, cost function $c(\cdot)$ and $W = (X, Y)$ for Equation (1.2), which is

$$\min_{\beta} \max_{D_c(P,P_n)\leq\delta} \mathbb{E}_P \left[l\left(X, Y; \beta\right) \right]. \tag{1.6}$$

The distributional uncertainty set, $\mathcal{U}_\delta(P_n) = \{P : \mathcal{D}_c(P, P_n) \leq \delta\}$, represents the class of models that are, in some sense, plausible variations of P_n. For every selection P in $\mathcal{U}_\delta(P_n)$, there is an optimal choice $\beta = \beta(P)$ which minimizes the risk $\mathbb{E}_P[l(X, Y; \beta)]$. We shall define $\Lambda_n(\delta) = \{\beta(P) : P \in \mathcal{U}_\delta(P_n)\}$ to be the set of plausible selections of the parameter β.

Now, for the definition of $\Lambda_n(\delta)$ to be sensible, we must have that the estimator obtained from solving (1.6) is plausible. This follows from the following result, which

is established with the aid of a min-max theorem in Chapter 2,

$$\min_{\beta\in\mathbb{R}^d} \max_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} \mathbb{E}_P\left[l(X,Y;\beta)\right] = \min_{\beta\in\Lambda_n(\delta)} \max_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} \mathbb{E}_P\left[l(X,Y;\beta)\right].$$

Then, we will say that β_* is **plausible** with $(1-\alpha)$ confidence, or simply, $(1-\alpha)$-plausible if δ is large enough so that $\beta_* \in \Lambda_n(\delta)$ with probability at least $1-\alpha$. This definition leads us to the optimality criterion that we shall consider.

Our optimal selection criterion for δ is formulated as follows: *Choose $\delta > 0$ as small as possible in order to guarantee that β_* is plausible with $(1-\alpha)$ confidence.*

As an additional desirable property, we shall verify that if β_* is $(1-\alpha)$-plausible, then $\Lambda_n(\delta)$ is a $(1-\alpha)$-confidence region for β_*.

Let us focus our discussion on linear regression model. In order to formally setup an optimization problem for the choice of $\delta > 0$, note that for any given P, by convexity, any optimal selection β is characterized by the first order optimality condition, namely,

$$\mathbb{E}_P\left[\left(Y-\beta^T X\right)X\right] = \mathbf{0}. \tag{1.7}$$

We then introduce the following object, which is the RWP (Robust Wasserstein Profile) function associated with the estimating equation (1.7),

$$R_n(\beta) = \inf\left\{D_c(P,P_n) : \mathbb{E}_P\left[\left(Y-\beta^T X\right)X\right] = \mathbf{0}\right\}.$$

Finally, we claim that the optimal choice of δ is precisely the $1-\alpha$ quantile, $\chi_{1-\alpha}$, of $R_n(\beta_*)$; that is

$$\chi_{1-\alpha} = \inf\left\{z : P\left(R_n(\beta_*) \leq z\right) \geq 1-\alpha\right\}.$$

To see this note that if $\tilde{\delta} > \chi_{1-\alpha}$ then indeed β^* is plausible with probability at least $1-\alpha$, but $\tilde{\delta}$ is not minimal. In turn, note that $R_n(\beta)$ allows to provide an explicit characterization of $\Lambda_n(\chi_{1-\alpha})$,

$$\Lambda_n(\chi_{1-\alpha}) = \{\beta : R_n(\beta) \leq \chi_{1-\alpha}\}.$$

Moreover, we clearly have

$$P(\beta_* \in \Lambda_n(\chi_{1-\alpha})) = P(R_n(\beta_*) \leq \chi_{1-\alpha}) = 1-\alpha,$$

so $\Lambda_n(\chi_{1-\alpha})$ is a $(1-\alpha)$-confidence region for β^*.

In order to further explain the role of $R_n(\beta_*)$, let us define $\mathcal{P}_{opt}(\beta_*)$ to be the set of probability measures, P, supported on a subset of $\mathbb{R}^d \times \mathbb{R}$ for which (1.7) holds with $\beta = \beta_*$. Formally,

$$\mathcal{P}_{opt}(\beta_*) := \left\{P : \mathbb{E}_P\left[\left(Y - \beta_*^T X\right)X\right] = \mathbf{0}\right\}.$$

In simple words, $\mathcal{P}_{opt}(\beta_*)$ is the set of probability measures for which β_* is an optimal risk minimization parameter. Observe that using this definition we can write

$$R_n(\beta_*) = \inf\{\mathcal{D}_c(P, P_n) : P \in \mathcal{P}_{opt}(\beta_*)\}.$$

Consequently, the set

$$\{P : \mathcal{D}_c(P, P_n) \leq R_n(\beta_*)\}$$

denotes the smallest uncertainty region around P_n (in terms of $\mathcal{D}_c$) for which one can find a distribution P satisfying the optimality condition $\mathbb{E}_P\left[(Y - \beta_*^T X)X\right] = \mathbf{0}$.

In summary, $R_n(\beta_*)$ denotes the smallest size of uncertainty that makes β_* *plausi-*

ble. If we were to choose a radius of uncertainty smaller than $R_n(\beta_*)$, then no probability measure in the neighborhood will satisfy the optimality condition $\mathbb{E}_P\left[(Y - \beta_*^T X)X\right] = \mathbf{0}$. On the other hand, if $\delta > R_n(\beta_*)$, the set

$$\left\{P : \mathbb{E}_P\big[(Y - \beta_*^T X)X\big] = \mathbf{0}, D_c\big(P, P_n\big) \leq \delta\right\}$$

is nonempty. Given the importance of $R_n(\beta_*)$ in the optimal selection of the regularization parameter λ, it is of interest to analyze its asymptotic properties as $n \to \infty$.

This discussion provides an intuitive understanding for how to pick the uncertainty size δ for $\mathcal{U}_\delta(P_n)$ optimally using the linear regression example as a motivation. A more in-depth study of the RWP function is given in Chapter 2 and further applications to machine learning settings are given in Chapter 5. Further extensions to settings in which the support of the elements in the distributional uncertainty are restricted are studied in Chapter 3 and in Chapter 4.

1.3 On shaping $\mathcal{U}$ using data and new statistical insights

One of the main advantages of considering an optimal transport discrepancy is that we have the flexibility to select a cost function which is either informed by our learning goal or which encodes additional information to improve the generalization performance.

For example, suppose that we have collection of data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ and also assume that we have unlabeled observations (i.e. observations without response Y),

which we denote as $U_{N-n} = \{X_i\}_{i=n+1}^{N}$. For simplicity, we consider binary classification problem and the response $Y \in \{-1, +1\}$. Let us further denote the set

$$\mathcal{E}_{N-n} = U_{N-n} \times \{-1, +1\} = \{X_i, 1\}_{i=n+1}^{N} \cup \{X_i, -1\}_{i=n+1}^{N},$$

in which we replicate each unlabeled data point twice, recognizing that the missing label could be any of the two available alternatives. We assume that the data must be labeled either -1 or +1. We then construct the set $\mathcal{X}_N = \mathcal{D}_n \cup \mathcal{E}_{N-n}$ which, in simple words, is obtained by just combining both the labeled data and the unlabeled data with all possible labeles which can be assigned. For a standard empirical risk minimization learning problem of the form,

$$\min_{\beta} \mathbb{E}_{P_n} \left[l \left(X, Y; \beta \right) \right],$$

we can define the semi-supervised learning DRO via

$$\min_{\beta} \max_{P \in \mathcal{P}(\mathcal{X}_N), D_c(P_n, P) \leq \delta} \mathbb{E}_{P} \left[l \left(X, Y; \beta \right) \right]. \tag{1.8}$$

We will argue that by solving the data-driven DRO problem in Equation (1.8), we may enhance the generalization error because we are using the unlabeled data to restrict the support of the members of the distributional uncertainty. The intuition is that if the predictors lie in a lower dimensional subspace of $\mathbb{R}^d$, then it suffices to enhance the out-of-sample performance of the estimator only on such lower dimensional space, which in turn might be well described by the unlabeled data set if N is sufficiently large.

The semi-supervised learning approach that we advocate in Equation (1.8) is not a robustification method that provide data-driven DRO formulation to any existing

semi-supervised learning algorithm. We provide a different and novel semi-supervised learning approach. Our semi-supervised DRO formulation utilizes the flexibility of the optimal transport discrepancy to encode the unlabeled information into the risk minimization. Further details will be discussed in Chapter 4.

In addition to restricting the support of the elements in the distributional uncertainty set, we are able to choose cost function which adapts to our learning goal. We will show that, by defining a groupwise cost function, we are able to inform the distributional uncertainty region $\mathcal{U}_\delta(P_n)$ with the side information for predictors and build up DRO representation for some popular groupwise shrinkage estimators, for example, square-root Group Lasso for linear regression and group-Lasso for logistic regression. The details of the data-driven DRO groupwise regularization estimator will be discussed in Chapter 5.

The groupwise regularization connection is based on having prior assumptions (or side-information) on the predictors. If there is no prior information available, we would like to design the cost function in a fully data-driven approach. We propose a methodology which learns such a distributional uncertainty neighborhood in a natural data-driven way. For example, we consider a parametric family of cost functions of the form $c\left(u,w\right)=\left(u-w\right)^{T}\Lambda\left(u-w\right)$ for a positive definite Λ. This choice corresponds to the so-called Mahalanobis distance. We use results from the literature on metric learning procedures to calibrate Λ in a way that is consistent with the learning task at hand. This discussion is given in Chapter 6.

Moreover, we also contribute to the metric learning literature by providing a data-driven robust optimization methodology to calibrate Λ. This additional layer of robustification, which then is used when solving our data-driven DRO formulation, justifies the name doubly robust data-driven DRO (DD-R-DRO). The DD-R-DRO methodology is also discussed in Chapter 6.

1.4 How to solve data-driven DRO problem?

For some of the data-driven DRO formulations, the dual formulation is not as easily accessible as in the case of regularized estimators as square-root Lasso, regularized logistic regression, and SVM. As we shall discussed in Chapter 4 and Chapter 6, the data-driven DRO with loss function $l(X, Y, \beta)$ and cost function $c(\cdot)$, is equivalent to solving

$$\begin{aligned}&\min_{\beta} \max_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P \left[l(X, Y, \beta) \right] \\ =&\min_{\beta} \min_{\lambda \geq 0} \frac{1}{n} \sum_{i}^{n} \max_{u} \left\{ l(u, v, \beta) - \lambda c\left((X_i, Y_i), (u, v) \right) + \lambda \delta \right\},\end{aligned}$$

where the inner-most optimization (involving $\max_u$) is taken for each sample point X_i, Y_i.

We provide a smoothing approximation technique to remove the inner maximization over u and propose an unbiased gradient estimation for the stochastic gradient algorithms to the data-driven DRO problem directly. The details of the algorithms and the smoothing approximation bound are discussed in Chapter 4 and Chapter 6. The proposed computational algorithm makes the data-driven DRO formulation applicable rather generally (beyond the setting of standard regularized estimators for which we obtain the representations discussed earlier). The optimization algorithm that we shall discuss is based on stochastic gradient descent, which is scalable to massive data sets.

1.5 Further Discussion

For the data-driven DRO formulation introduced in Section 1.1 and Section 1.3, we note that our data-driven DRO formulations can be applied to more general machine learning algorithms. Once the loss function and its gradient are accessible, we are able to apply our stochastic gradient based algorithm discussed in Chapter 6, to solve the data-driven DRO problem directly.

This is to say, even for a complex model, once the cost function is chosen properly, we can apply data-driven DRO to address the overfitting problem and to improve generalization performance. For example, as we shall discuss in Chapter 7, Section 7.1, we use multi-task training as an example to show that data-driven DRO might help in building novel learning methods to improve the generalization performance.

In Chapter 7, Section 7.2, we include a discussion on difference and connections between robustness in classical statistics and robustness in our DRO formulation. Finally, we will close the dissertation by discussing further potential research avenues, in Chapter 7, Section 7.3,

Chapter 2

Robust Wasserstein Profile Inference (RWPI)

In this chapter, we introduce RWPI (Robust Wasserstein-distance Profile-based Inference - pronounced similar to Rupee. The acronym RWPI is chosen to sound just as "RUPI", where "u" as in put and "i" as in bit. In turn, RUPI means beautiful in Sanskrit.), a novel class of statistical tools which exploits connections between Empirical Likelihood, Distributionally Robust Optimization and the Theory of Optimal Transport (via the use of Wasserstein distances). A key element of RWPI is the so-called Robust Wasserstein Profile function, whose asymptotic properties we study in this chapter. We illustrate the use of RWPI in the context of machine learning algorithms, such as the square-root Lasso (Least Absolute Shrinkage and Selection) and regularized logistic regression, among others. For these algorithms, we show how to optimally select the regularization parameter without the use of cross validation. The use of RWPI for such optimal selection requires a suitable distributionally robust representation for these machine learning algorithms, which is also novel and of independent interest. Numerical experiments are also given to validate our theoretical

findings.

2.1 Introduction

The goal of this chapter is to introduce and investigate a novel inference methodology which we call RWPI (Robust Wasserstein-distance Profile-based Inference). RWPI combines ideas from three different areas: Empirical Likelihood (EL), Distributionally Robust Optimization, and the Theory of Optimal Transport. While RWPI can be applied to a wide range of inference problems, in this chapter we use several well known algorithms in machine learning to illustrate the use and implications of this methodology.

We will explain, by means of several examples of interest, how RWPI can be used to optimally choose the regularization parameter in machine learning applications without the need of cross validation. The examples of interest that we study in this chapter include square-root Lasso (Least Absolute Shrinkage and Selection) and regularized logistic regression, among others. In order to explain RWPI let us walk through a simple application in a familiar context, namely, that of linear regression.

2.1.1 RWPI for optimal regularization of square-root Lasso

Consider a given a set of training data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$. The input $X_i \in \mathbb{R}^d$ is a vector of d predictor variables, and $Y_i \in \mathbb{R}$ is the response variable. It is postulated that

$$Y_i = \beta_*^T X_i + e_i,$$

for some $\beta_* \in \mathbb{R}^d$ and errors $\{e_1, ..., e_n\}$. Under suitable statistical assumptions (such as independence of the samples in the training data) one may be interested in estimat-

ing β_*. Underlying there is a general loss function, $l(x, y; \beta)$, which we shall take for simplicity in this discussion to be the quadratic loss, namely, $l(x, y; \beta) = \left(y - \beta^T x\right)^2$.

Over the last two decades, various regularized estimators have been introduced and studied. Many of them have gained substantial popularity because of their good empirical performance and insightful theoretical properties, (see, for example, Tibshirani [1996] for an early reference and Friedman *et al.* [2001] for a discussion on regularized estimators). One such regularized estimator, implemented, for example in the "flare" package, see Li *et al.* [2015], is the so-called square-root Lasso estimator; which is obtained by solving the following convex optimization problem in β,

$$\begin{aligned} &\min_{\beta \in \mathbb{R}^d} \left\{ \sqrt{\mathbb{E}_{P_n}\left[l\left(X, Y; \beta\right)\right]} + \lambda \left\|\beta\right\|_1 \right\} \\ &= \min_{\beta \in \mathbb{R}^d} \left\{ \sqrt{\frac{1}{n} \sum_{i=1}^{n} l\left(X_i, Y_i; \beta\right)} + \lambda \left\|\beta\right\|_1 \right\}, \end{aligned} \tag{2.1}$$

where $\left\|\beta\right\|_p$ denotes the p-th norm in the Euclidean space. The parameter λ, commonly referred to as the regularization parameter, is crucial for the performance of the algorithm and it is often chosen using cross validation.

2.1.1.1 Distributionally robust representation of square-root Lasso

We shall illustrate how to choose λ, satisfying a natural optimality criterion, as the quantile of a certain object which we call the Robust Wasserstein Profile (RWP) function evaluated at β_*. This will motivate a systematic study of the RWP function as the sample size, n, increases. However, before we define the associated RWP function, we first introduce a class of representations which are of independent interest and which are necessary to motivate the definition of the RWP function for choosing λ.

One of our contributions in this chapter is a representation of (2.1) in terms of a Distributionally Robust Optimization formulation (see Section 2.3). In particular, we construct a discrepancy measure, $D_c(P, Q)$, based on a suitable Wasserstein-type distance, between two probability measures P and Q satisfying that

$$\begin{aligned} &\min_{\beta\in\mathbb{R}^d}\left\{\sqrt{\mathbb{E}_{P_n}\left[l\left(X,Y;\beta\right)\right]}+\lambda\left\|\beta\right\|_1\right\}^2 \\ &=\min_{\beta\in\mathbb{R}^d}\max_{P:\ D_c(P,P_n)\leq\delta}\mathbb{E}_P\left[l(X,Y;\beta)\right], \end{aligned} \tag{2.2}$$

where $\delta = \lambda^{1/2}$. Observe that the regularization parameter is fully determined by the size of the uncertainty, δ, in the distributionally robust formulation on the right hand side of (2.2).

The set $\mathcal{U}_\delta(P_n) = \{P : D_c(P, P_n) \leq \delta\}$ is called the uncertainty set in the language of distributionally robust optimization, and it represents the class of models that are, in some sense, plausible variations of P_n. The estimator obtained by solving Equation (2.2) is referred as distributionally robust regression estimator, and we remark that this notion of robustness is different from the standard statistical robustness which primarily addresses data contamination with outliers (see Huber [1964]).

For every selection P in $\mathcal{U}_\delta(P_n)$, there is an optimal choice $\beta = \beta(P)$ which minimizes the risk $\mathbb{E}_P[l(X, Y; \beta)]$. We shall define $\Lambda_n(\delta) = \{\beta(P) : P \in \mathcal{U}_\delta(P_n)\}$ to be the set of plausible selections of the parameter β.

Now, for the definition of $\Lambda_n(\delta)$ to be sensible, we must have that the estimator obtained from the left hand side of (2.2) is plausible. This follows from the following result, which is established with the aid of a min-max theorem in Section 2.4,

$$\min_{\beta\in\mathbb{R}^d}\max_{P:\ D_c(P,P_n)\leq\delta}\mathbb{E}_P\left[l(X,Y;\beta)\right]=\min_{\beta\in\Lambda_n(\delta)}\max_{P:\ D_c(P,P_n)\leq\delta}\mathbb{E}_P\left[l(X,Y;\beta)\right].$$

Then, we will say that β_* is **plausible** with $(1-\alpha)$ confidence, or simply, $(1-\alpha)$-plausible if δ is large enough so that $\beta_* \in \Lambda_n(\delta)$ with probability at least $1-\alpha$. This definition leads us to the optimality criterion that we shall consider.

Our optimal selection criterion for δ is formulated as follows: *Choose $\delta > 0$ as small as possible in order to guarantee that β_* is plausible with $(1-\alpha)$ confidence.*

As an additional desirable property, we shall verify that if β_* is $(1-\alpha)$-plausible, then $\Lambda_n(\delta)$ is a $(1-\alpha)$-confidence region for β_*. A computationally efficient procedure for evaluating $\Lambda_n(\delta)$ will be studied in future work. Our focus in this chapter is on the optimal selection of δ.

2.1.1.2 The associated Robust Wasserstein Profile Function

In order to formally setup an optimization problem for the choice of $\delta > 0$, note that for any given P, by convexity, any optimal selection β is characterized by the first order optimality condition, namely,

$$\mathbb{E}_P\left[\left(Y - \beta^T X\right) X\right] = \mathbf{0}. \tag{2.3}$$

We then introduce the following object, which is the RWP function associated with the estimating equation (2.3),

$$R_n(\beta) = \inf\left\{D_c(P, P_n) : \mathbb{E}_P\left[\left(Y - \beta^T X\right) X\right] = \mathbf{0}\right\}. \tag{2.4}$$

Finally, we claim that the optimal choice of δ is precisely the $1-\alpha$ quantile, $\chi_{1-\alpha}$,

of $R_n(\beta_*)$; that is

$$\chi_{1-\alpha} = \inf\left\{z : P\left(R_n(\beta_*) \leq z\right) \geq 1-\alpha\right\}.$$

To see this note that if $\tilde{\delta} > \chi_{1-\alpha}$ then indeed β^* is plausible with probability at least $1-\alpha$, but $\tilde{\delta}$ is not minimal. In turn, note that $R_n(\beta)$ allows to provide an explicit characterization of $\Lambda_n(\chi_{1-\alpha})$,

$$\Lambda_n(\chi_{1-\alpha}) = \{\beta : R_n(\beta) \leq \chi_{1-\alpha}\}.$$

Moreover, we clearly have

$$P\left(\beta_* \in \Lambda_n(\chi_{1-\alpha})\right) = P\left(R_n(\beta_*) \leq \chi_{1-\alpha}\right) = 1-\alpha,$$

so $\Lambda_n(\chi_{1-\alpha})$ is a $(1-\alpha)$-confidence region for β^*.

In order to further explain the role of $R_n(\beta_*)$, let us define $\mathcal{P}_{opt}(\beta_*)$ to be the set of probability measures, P, supported on a subset of $\mathbb{R}^d \times \mathbb{R}$ for which (2.3) holds with $\beta = \beta_*$. Formally,

$$\mathcal{P}_{opt}(\beta_*) := \left\{P : \mathbb{E}_P\left[\left(Y - \beta_*^T X\right)X\right] = \mathbf{0}\right\}.$$

In simple words, $\mathcal{P}_{opt}(\beta_*)$ is the set of probability measures for which β_* is an optimal risk minimization parameter. Observe that using this definition we can write

$$R_n(\beta_*) = \inf\{D_c(P, P_n) : P \in \mathcal{P}_{opt}(\beta_*)\}.$$

Consequently, the set

$$\{P : D_c(P, P_n) \leq R_n(\beta_*)\}$$

denotes the smallest uncertainty region around P_n (in terms of D_c) for which one can find a distribution P satisfying the optimality condition $\mathbb{E}_P\left[(Y - \beta_*^T X)X\right] = \mathbf{0}$, see Figure 2.1 for a pictorial representation of $\mathcal{P}_{opt}(\beta_*)$ and $R_n(\beta_*)$.

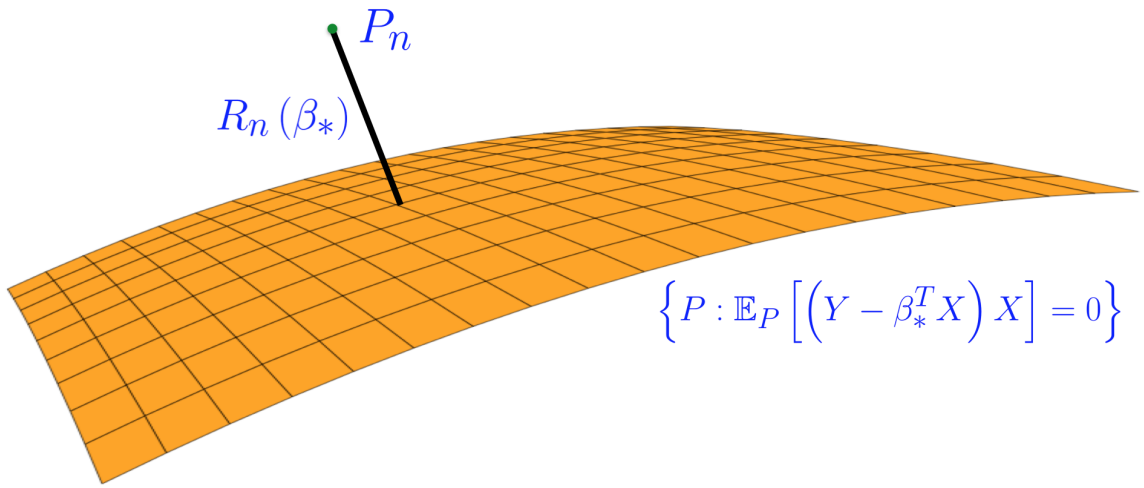

Figure 2.1: Illustration of RWP function evaluated at β_*

In summary, $R_n(\beta_*)$ denotes the smallest size of uncertainty that makes β_* *plausible*. If we were to choose a radius of uncertainty smaller than $R_n(\beta_*)$, then no probability measure in the neighborhood will satisfy the optimality condition $\mathbb{E}_P\left[(Y - \beta_*^T X)X\right] = \mathbf{0}$. On the other hand, if $\delta > R_n(\beta_*)$, the set

$$\left\{P : \mathbb{E}_P\left[(Y - \beta_*^T X)\right] = \mathbf{0}, D_c\left(P, P_n\right) \leq \delta\right\}$$

is nonempty. Given the importance of $R_n(\beta_*)$ in the optimal selection of the regularization parameter λ, it is of interest to analyze its asymptotic properties as $n \to \infty$.

It is important to note, however, that the estimating equations given in (2.3) are just one of potentially many ways in which β_* can be characterized. In the case

of Gaussian input there is an (well known) intimate connection between (2.3) and maximum likelihood estimation. In general it appears sensible, at least from the standpoint of philosophical consistency to connect the choice of estimating equation with the loss function $l(x, y; \beta)$ used in the Distributionally Robust Representation (2.2).

2.1.2 A broad perspective of the contributions of this chapter

The previous discussion in the context of linear regression highlights two key ideas: a) the RWP function as a key object of analysis, and b) the role of distributionally robust representation of regularized estimators.

The RWP function can be applied much more broadly than in the context of regularized estimators. This chapter is written with the goal of studying the RWP function for estimating equations generally and systematically. As an application, we showcase the study of the RWP function in a context of great importance, namely, the optimal selection of regularization parameters in several machine learning algorithms.

Broadly speaking, RWPI is a statistical tool which consists in building a suitable RWP function in order to estimate a parameter of interest. From a philosophical standpoint, RWPI borrows heavily from Empirical Likelihood (EL), introduced in the seminal work of Owen [1988, 1990]. There are important methodological differences, however, as we shall discuss in the sequel. In the last three decades, there have been a great deal of successful applications of Empirical Likelihood for inference [Owen, 1991; Qin and Lawless, 1994; Bravo, 2004; Hjort *et al.*, 2009; Zhou, 2015]. In principle, all of those applications can be revisited using the RWP function and its ramifications. Therefore, we spend the first part of the chapter, namely Section 2, discussing general properties of the RWP function.

The application of RWPI for the optimal selection of regularization parameters in various machine learning settings is given in Section 2.4. Once a suitable RWP function is obtained, the results in Section 2.4 are obtained directly from applications of our results in Section 2.2. In order to obtain the correct RWP function formulation for each of the machine learning settings of interest, however, we will need to derive a suitable distributionally robust representations which, analogous to those discussed in the square-root Lasso setting. These representations are given in Section 2.3 of this chapter.

We now provide a more precise description of our contributions:

A) We provide general limit theorems for the asymptotic distribution (as the sample size increases) of the RWP function defined for general estimating equations, not only those arising from linear regression problems. Hence, providing tools to apply RWPI in substantial generality (see the results in Section 2.2.4).

B) We explain how, by judiciously choosing $D_c(\cdot)$, we can define a family of regularized regression estimators (See Section 2.3). In particular, we will show how square-root Lasso (see and Theorem 2.2), and regularized logistic regression (see Theorem 2.3) arise as a particular case of a RWPI formulation.

C) The results in **B)** allow to obtain the appropriate RWP function to select an optimal regularization parameter. We then illustrate how to analyze the distribution of $R_n(\beta_*)$ using our results form **A)** (see Section 2.4).

D) We analyze our regularization selection in the high dimensional setting for square-root Lasso. Under standard regularity conditions, we show (see Theorem 2.6)

that the regularization parameter λ might be chosen so that,

$$\lambda = \frac{\pi}{\pi - 2} \frac{\Phi^{-1}\left(1 - \alpha/2d\right)}{\sqrt{n}},$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of standard normal distribution. The behavior of λ as a function of n and d is consistent with regularization selections studied in the literature motivated by different considerations.

E) We analyze the empirical performance of RWPI for the selection of the optimal regularization parameter in the context of square-root Lasso. This is done in Appendix 2.D. We apply our analysis both to simulated and real data and compare against the performance of cross validation. We conclude that our approach towards regularization parameter selection offers comparable (not worst) performance, although at a much lesser computational cost than cross validation.

We now provide a discussion on topics which are related to RWPI.

2.1.3 Connections to related inference literature

Let us first discuss the connections between RWPI and EL. In EL one builds a Profile Likelihood for an estimating equation. For instance, in the context of EL applied to estimating β satisfying (2.3), one would build a Profile Likelihood Function in which the optimization object is only defined as the likelihood (or the log-likelihood) between a given distribution P with respect to P_n. Therefore, the analogue of the uncertainty set $\{P : D_c(P, P_n) \leq \delta\}$, in the context of EL, will typically contain distributions whose support coincides with that of P_n. In contrast, the definition of the RWP function does not require the likelihood between an alternative plausible model P, and the empirical distribution, P_n, to exist. Owing to this flexibility, for

example, we are able to establish the connection between regularization estimators and a suitable profile function.

There are other potential benefits of using a profile function which does not restrict the support of alternative plausible models. For example, it has been observed in the literature that in some settings EL might exhibit low coverage Owen [2001]; Chen and Hall [1993]; Wu [2004]. It is not the goal of this chapter to examine the coverage properties of RWPI systematically, but it is conceivable that relaxing the support of alternative plausible models, as RWPI does, can translate into desirable coverage properties.

From a technical standpoint, the definition of the Profile Function in EL gives rise to a finite dimensional optimization problem. Moreover, there is a substantial amount of smoothness in the optimization problems defining the EL Profile Function. This degree of smoothness can be leveraged in order to obtain the asymptotic distribution of the Profile Function as the sample size increases. In contrast, the optimization problem underlying the definition of RWP function in RWPI is an infinite dimensional linear program. Therefore, the mathematical techniques required to analyze the associated RWP function are different (more involved) than the ones which are commonly used in the EL setting.

A significant advantage of EL, however, is that the limiting distribution of the associated Profile Function is typically chi-squared. Moreover, such distribution is self-normalized in the sense that no parameters need to be estimated from the data. Unfortunately, this is typically not the case in the case of RWPI. In many settings, however, the parameters of the limiting distribution can be easily estimated from the data itself.

Another set of tools, strongly related to RWPI, have also been studied recently by the name of SOS (Sample-Out-of-Sample) inference as we shall discuss in Chapter 3.

In this setting, also an RWP function is built, but the support of alternative plausible models is assumed to be finite (but not necessarily equal to that of P_n). Instead, the support of alternative plausible models is assumed to be generated not only by the available data, but additional samples coming from independent distributions (defined by the user). The mathematical results obtained for the RWP function in the context of SOS are different from those obtained in this chapter. For example, in the SOS setting, the rates of convergence are dimension-dependent, which is not the case in RWPI.

2.1.4 Some connections to Distributionally Robust Optimization and Optimal Transport

Connection between robust optimization and regularization procedures such as Lasso and Support Vector Machines have been studied in the literature, see Xu *et al.* [2009a,b]. The methods proposed here differ subtly: While the papers Xu *et al.* [2009a,b] add deterministic perturbations of a certain size to the predictor vectors X to quantify uncertainty, the Distributionally Robust Representations that we derive measure perturbations in terms of deviations from the empirical distribution. While this change may appear cosmetic, it brings a significant advantage: measuring deviations from empirical distribution, in turn, lets us derive suitable limit laws (or) probabilistic inequalities that can be used to choose the size of uncertainty, δ, in the uncertainty region $\mathcal{U}_\delta(P_n) = \{P : D_c(P, P_n) \leq \delta\}$.

Now, it is intuitively clear that as the number of samples n increase, the deviation of the empirical distribution from the true distribution decays to zero, as a function of n, at a specific rate of convergence. To begin with, one can simply use, as a direct approach to choosing the size of δ, a concentration inequality that measures

this rate of convergence. Such simple specification of the size of uncertainty, suitably as a function of n, does not arise naturally in the deterministic robust optimization approach. For a concentration inequality that measures such deviations in terms of the Wasserstein distance, we refer to Fournier and Guillin [2015] and references there in. For an application of these concentration inequalities to choose the size of uncertainty set in the context of distributionally robust logistic regression, refer Shafieezadeh-Abadeh *et al.* [2015]. It is important to note that, despite imposing severe tail assumptions, these concentration inequalities dictate the size of uncertainty to decay at the rate $O(n^{-1/d})$, where d is the number of covariates. Unfortunately, this prescription scales non-graciously as the dimension d increases. Since most of the modern learning problems have huge number of covariates, application of such concentration inequalities with poor rate of decay with dimensions may not be most suitable for applications.

In contrast to directly using concentration inequalities, the prescription that we advocate typically has a rate of convergence of order $O\left(n^{-1/2}\right)$ as $n \to \infty$ (for fixed d). Moreover, as we discuss in the case of Lasso, according to our results corresponding to contribution **E),** our prescription of the size of uncertainty actually can be shown (under suitable regularity conditions) to decay at rate $O(\sqrt{\log d/n})$ (uniformly over d and n), which is in agreement with the findings of compressed sensing and high-dimensional statistics literature (see Candes and Tao [2007]; Belloni *et al.* [2011]; Negahban *et al.* [2012] and references therein). Interestingly, the regularization parameter prescribed by RWPI methodology is automatically obtained without looking into the data (unlike cross-validation).

Although we have focused our discussion on the context of regularized estimators, our results are directly applicable to the area of data-driven Distributionally Robust Optimization whenever the uncertainty sets are defined in terms of a Wasserstein

distance or, more generally, an optimal transport metric. In particular, consider a given distributionally robust formulation of the form

$$\min_{\theta:G(\theta)\leq 0} \max_{P:\ D_c(P,P_n)\leq\delta} \mathbb{E}_P\left[H(W,\theta)\right],$$

for a random element W and a convex function $H(W,\cdot)$ defined over a convex region $\{\theta : G(\theta) \leq 0\}$ (assuming $G : \mathbb{R}^d \to \mathbb{R}$ convex). Here P_n is the empirical measure of the sample $\{W_1, ..., W_n\}$. One can then follow a reasoning parallel to what we advocate throughout our Lasso discussion.

Argue, by applying the corresponding KKT (Karush-Kuhn-Tucker) conditions, if possible, that an optimal solution θ_* to the problem

$$\min_{\theta:G(\theta)\leq 0} \mathbb{E}_{P_{true}}[H(W,\theta)]$$

satisfies a system of estimating equations of the form

$$\mathbb{E}_{P_{true}}[h(W,\theta_*)] = 0, \tag{2.5}$$

for a suitable $h(\cdot)$ (where P_{true} is the weak limit of the empirical measure P_n as $n \to \infty$). Then, given a confidence level $1-\alpha$, one should choose δ as the $(1-\alpha)$ quantile of the RWP function function

$$R_n(\theta_*) = \inf\{D_c(P,P_n) : \mathbb{E}_P[h(W,\theta_*)] = 0\}.$$

The results in Section 2 can then be used directly to approximate the $(1-\alpha)$ quantile of $R_n(\theta_*)$. Just as we explain in our discussion of the square-root Lasso example, the selection of δ is the smallest possible choice for which θ_* is plausible with $(1-\alpha)$

confidence.

2.1.5 Organization of this chapter

The rest of the chapter is organized as follows. Section 2.2 deals with contribution **A)** where we first revisit Wasserstein distances, which we discussed in Chapter 1 Section 1.1, and discuss the Robust Wasserstein Profile function as an inference tool in a way which is parallel to the Profile Likelihood in EL. We derive the asymptotic distribution of the RWP function for general estimating equations. Section 2.3 corresponds to contribution **B)**, namely, distributionally robust representations of some popular machine learning algorithms. Section 2.4 discusses contribution **C)**, namely the use of results from contributions **A)** for optimal regularization parameter selection. Our high-dimensional analysis of the RWP function in the case of square-root Lasso is also given in Section 2.4. The proofs for the main results along with various technical lemmas and numerical experiments are given in the Appendix.

2.2 The Robust Wasserstein Profile Function

Given an estimating equation $\mathbb{E}_{P_n}[h(W,\theta)] = \mathbf{0}$, the objective of this section is to study the asymptotic behavior of the associated RWP function $R_n(\theta)$. To do this, we first introduce some notation to define optimal transport costs and Wasserstein distances. Following this, we provide evidence, initially with a simple example, followed by results for general estimating equations, that the profile function defined using Wasserstein distances is tractable.

2.2.1 Revisit Optimal Transport Costs and Wasserstein Distances

Let us revisit the definition and properties of optimal transport discrepancy and Wasserstein Distance in this subsection.

Let $c : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty]$ be any lower semi-continuous function such that $c(u, w) = 0$ if and only if $u = w$. Given two probability distributions $P(\cdot)$ and $Q(\cdot)$ supported on $\mathbb{R}^m$, one can define the optimal transport cost or discrepancy between P and Q, denoted by $D_c(P, Q)$, as

$$D_c(P, Q) = \inf \left\{ \mathbb{E}_\pi [c(U, W)] : \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m),\ \pi_{_U} = P,\ \pi_{_W} = Q \right\}. \tag{2.6}$$

Here, $\mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m)$ is the set of joint probability distributions π of (U, W) supported on $\mathbb{R}^m \times \mathbb{R}_m$, and $\pi_{_U}$ and $\pi_{_W}$ denote the marginals of U and W under π, respectively.

Throughout this chapter, we shall select D_c for a judiciously chosen cost function $c(\cdot)$ in formulations such as (2.2). It is useful to allow $c(\cdot)$ to be lower semi-continuous and potentially be infinite in some region to accommodate some of the applications, such as regularization in the context of logistic regression, as we shall see in Section 2.3. So, our setting requires discrepancy choices which are slightly more general than standard Wasserstein distances.

2.2.2 The RWP Function for Estimating Equations and Its Use as an Inference Tool

The Robust Wasserstein Profile function's definition is inspired by the notion of the Profile Likelihood function, introduced in the pioneering work of Art Owen in the context of EL (see Owen [2001]). We provide the definition of the RWP function for

estimating $\theta_* \in \mathbb{R}^l$, which we assume satisfies

$$\mathbb{E}_{P_{true}}\left[h\left(W, \theta_*\right)\right] = \mathbf{0}, \tag{2.7}$$

for a given random variable W taking values in $\mathbb{R}^m$ and an integrable function $h : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}^r$. The parameter θ_* will typically be unique to ensure consistency, but uniqueness is not necessary for the limit theorems that we shall state, unless we explicitly indicate so.

Given a set of samples $\{W_1, ..., W_n\}$, which are assumed to be i.i.d. copies of W, we define the Wasserstein Profile function for the estimating equation (2.7) as,

$$R_n\left(\theta\right) := \inf\left\{D_c(P, P_n) : \ \mathbb{E}_P\left[h(W, \theta)\right] = \mathbf{0}\right\}. \tag{2.8}$$

Here, recall that P_n denotes the empirical distribution associated with the training samples $\{W_1, \ldots, W_n\}$ and $c(\cdot)$ is a chosen cost function. In this section, we are primarily concerned with cost functions of the form,

$$c\left(u, w\right) = \left\|w - u\right\|_q^\rho, \tag{2.9}$$

where $\rho \geq 1$ and $q \geq 1$. We remark, however, that the methods presented here can be easily adapted to more general cost functions. For simplicity, we assume that the samples $\{W_1, \ldots, W_n\}$ are distinct.

Since, as we shall see, that the asymptotic behavior of the RWP function $R_n(\theta)$ is dependent on the exponent ρ in Equation (2.9), we shall sometimes write $R_n\left(\theta; \rho\right)$ to make this dependence explicit; but whenever the context is clear, we drop ρ to avoid notational burden. Also, observe that the profile function defined in (2.4) for the linear regression example is obtained as a particular case by selecting $W = (X, Y)$,

$\beta = \theta$ and defining $h(x, y, \theta) = (y - \theta^T x)x$.

Our goal in this section is to develop an asymptotic analysis of the RWP function which parallels that of the theory of EL. In particular, we shall establish,

$$n^{\rho/2} R_n(\theta_*; \rho) \Rightarrow \bar{R}(\rho). \tag{2.10}$$

for a suitably defined random variable $\bar{R}(\rho)$ (throughout the rest of the chapter, the symbol "$\Rightarrow$" denotes convergence in distribution).

As the empirical distribution weakly converges to the underlying probability distribution from which the samples are obtained from, it follows from the definition of RWP function in Equation (2.10) that $R_n(\theta; \rho) \to 0$, as $n \to \infty$, if and only if θ satisfies $E[h(W, \theta)] = \mathbf{0}$; for every other θ, we have that $n^{\rho/2} R_n(\theta; \rho) \to \infty$. Therefore, the result in (2.10) can be used to provide confidence regions (at least conceptually) around θ_*. In particular, given a confidence level $1 - \alpha$ in (0,1), if we denote η_α as the $(1 - \alpha)$ quantile of $\bar{R}(\rho)$, that is, $P\left(\bar{R}(\rho) \leq \eta_\alpha\right) = (1 - \alpha)$, then

$$\bar{\Lambda}_n\left(\frac{\eta_\alpha}{n}\right) = \left\{\theta : R_n(\theta; \rho) \leq \frac{\eta_\alpha}{n}\right\}$$

yields an approximate $(1 - \alpha)$ confidence region for θ_*. This is because, by definition of $\bar{\Lambda}_n(\eta_\alpha/n)$, we have

$$P\left(\theta_* \in \bar{\Lambda}_n(\eta_\alpha/n)\right) = P\left(n^{\rho/2} R_n(\theta_*; \rho) \leq \eta_\alpha\right) \approx P\left(\bar{R}(\rho) \leq \eta_\alpha\right) = 1 - \alpha.$$

Throughout the development in this section, the dimension m of the underlying random vector W is kept fixed and the sample size n is sent to infinity; the function $h(\cdot)$ can be quite general. In Section 2.4.3, we extend the analysis of RWP function to the case where the ambient dimension could scale with the number of training

samples n, in the specific context of square-root Lasso for linear regression.

2.2.3 The dual formulation of RWP function

The first step in the analysis of the RWP function $R_n(\theta)$ is to use the definition of the discrepancy measure D_c to rewrite $R_n(\theta)$ as,

$$R_n(\theta) = \inf \left\{ \mathbb{E}_\pi \left[c(U,W) \right] : \pi \in \mathcal{P}\left(\mathbb{R}^m \times \mathbb{R}^m \right),\ \mathbb{E}_\pi \left[h\left(U, \theta \right) \right] = \mathbf{0},\ \pi_{_W} = P_n \right\},$$

which is a *problem of moments* of the form,

$$\begin{aligned} R_n(\theta) = \inf_{\pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m)} \Big\{ & \mathbb{E}_\pi \left[c(U,W) \right] :\ \mathbb{E}_\pi \left[h\left(U, \theta \right) \right] = \mathbf{0}, \\ & \mathbb{E}_\pi \left[\mathbb{I}(W = W_i) \right] = \frac{1}{n}, i = 1, \ldots, n \Big\}. \end{aligned} \tag{2.11}$$

The problem of moments is a classical linear programming problem for which the respective dual formulation and strong duality have been well-studied (see, for example, Isii [1962]; Smith [1995]). The linear program problem over the variable π in Equation (2.11) admits a simple dual semi-infinite linear program of form,

$$\begin{aligned} & \sup_{a_i \in \mathbb{R}, \lambda \in \mathbb{R}^r} \Bigg\{ a_0 + \frac{1}{n} \sum_{i=1}^n a_i : \\ & \qquad a_0 + \sum_{i=1}^n a_i \mathbf{1}_{\{w = W_i\}}(u,w) + \lambda^T h(u,\theta) \leq c(u,w), \forall u, w \in \mathbb{R}^m \Bigg\} \\ & = \sup_{\lambda \in \mathbb{R}^r} \left\{ \frac{1}{n} \sum_{i=1}^n \inf_{u \in \mathbb{R}^m} \left\{ c(u, W_i) - \lambda^T h(u,\theta) \right\} \right\} \\ & = \sup_{\lambda \in \mathbb{R}^r} \left\{ -\frac{1}{n} \sum_{i=1}^n \sup_{u \in \mathbb{R}^m} \left\{ \lambda^T h(u,\theta) - c(u, W_i) \right\} \right\}. \end{aligned}$$

Proposition 2.1 below states that strong duality holds under mild assumptions, and the dual formulation above indeed equals $R_n(\theta)$.

Proposition 2.1. Let $h(\cdot, \theta)$ be Borel measurable, and $\Omega = \{(u, w) \in \mathbb{R}^m \times \mathbb{R}^m : c(u, w) < \infty\}$ be Borel measurable and non-empty. Further, suppose that $\mathbf{0}$ lies in the interior of the convex hull of $\{h(u, \theta) : u \in \mathbb{R}^m\}$. Then,

$$R_n(\theta) = \sup_{\lambda \in \mathbb{R}^r} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sup_{u \in \mathbb{R}^m} \left\{ \lambda^T h(u, \theta) - c(u, W_i) \right\} \right\}.$$

A proof of Proposition 2.1, along with an introduction to the problem of moments, is provided in Appendix 2.B of this Chapter.

2.2.4 Asymptotic Distribution of the RWP Function

In order to gain intuition behind (2.10), let us first consider the simple example of estimating the expectation $\theta_* = \mathbb{E}[W]$ of a real-valued random variable W, using $h(w, \theta) = w - \theta$.

Example 2.1. (RWPI for mean estimation.) Let $h(w, \theta) = w - \theta$ with $m = 1 = l = r$. First, suppose that the choice of cost function is $c(u, w) = |u - w|^\rho$ for some $\rho > 1$. As long as θ lies in the interior of convex hull of support of W, Proposition Equation (2.1) implies,

$$\begin{aligned} R_n(\theta; \rho) &= \sup_{\lambda \in \mathbb{R}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sup_{u \in \mathbb{R}} \left\{ \lambda(u - \theta) - |W_i - u|^\rho \right\} \right\} \\ &= \sup_{\lambda \in \mathbb{R}} \left\{ -\frac{\lambda}{n} \sum_{i=1}^{n} (W_i - \theta) - \frac{1}{n} \sum_{i=1}^{n} \sup_{u \in \mathbb{R}} \left\{ \lambda (u - W_i) - |W_i - u|^\rho \right\} \right\}. \end{aligned}$$

As

$$\max_{\Delta} \{\lambda \Delta - |\Delta|^\rho\} = (\rho - 1)|\lambda/\rho|^{\rho/(\rho-1)},$$

we obtain

$$\begin{aligned} R_n(\theta;\rho) &= \sup_{\lambda}\left\{-\frac{\lambda}{n}\sum_{i=1}^{n}(W_i-\theta)-(\rho-1)\left|\frac{\lambda}{\rho}\right|^{\frac{\rho}{\rho-1}}\right\} \\ &= \left|\frac{1}{n}\sum_{i=1}^{n}(W_i-\theta)\right|^{\rho}. \end{aligned}$$

Then, under the hypothesis that $\mathbb{E}[W]=\theta_*$, and assuming $\mathrm{Var}[W]=\sigma_W^2<\infty$, we obtain,

$$n^{\rho/2}R_n(\theta_*;\rho)\Rightarrow \bar{R}(\rho)\sim \sigma_W^{\rho}\left|\mathcal{N}(0,1)\right|^{\rho},$$

where $\mathcal{N}(0,1)$ denotes a standard Gaussian random variable. The limiting distribution for the case $\rho=1$ can be formally obtained by setting $\rho=1$ in the above expression for $\bar{R}(\rho)$, but the analysis is slightly different. When $\rho=1$,

$$\begin{aligned} R_n(\theta) &= \sup_{\lambda\in\mathbb{R}}\left\{-\frac{\lambda}{n}\sum_{i=1}^{n}(W_i-\theta)-\frac{1}{n}\sum_{i=1}^{n}\sup_{u\in\mathbb{R}}\left\{\lambda(u-W_i)-|u-W_i|\right\}\right\} \\ &= \sup_{\lambda}\left\{-\frac{\lambda}{n}\sum_{i=1}^{n}(W_i-\theta)-\sup_{\Delta\in\mathbb{R}}\left\{\lambda\Delta-|\Delta|\right\}\right\}. \end{aligned}$$

Following the notion that $\infty\times 0=0$,

$$\begin{aligned} R_n(\theta) &= \sup_{\lambda}\left\{\frac{\lambda}{n}\sum_{i=1}^{n}(W_i-\theta)-\infty I(|\lambda|>1)\right\} \\ &= \max_{|\lambda|\le 1}\frac{\lambda}{n}\sum_{i=1}^{n}(W_i-\theta)=\left|\frac{1}{n}\sum_{i=1}^{n}(W_i-\theta)\right|. \end{aligned}$$

So, indeed if $E[W]=\theta_*$ and $Var[W]=\sigma_W^2<\infty$, we obtain

$$n^{1/2}R_n(\theta_*)\Rightarrow \sigma_W\left|\mathcal{N}(0,1)\right|.$$

We now discuss far reaching extensions to the developments in Example 2.1 by considering estimating equations that are more general. First, we state a general asymptotic stochastic upper bound, which we believe is the most important result from an applied standpoint as it captures the speed of convergence of $R_n(\theta_*)$ to zero. Following this, we obtain an asymptotic stochastic lower bound that matches with the upper bound (and therefore the weak limit) under mild, additional regularity conditions. We discuss the nature of these additional regularity conditions, and also why the lower bound in the case $\rho = 1$ can be obtained basically without additional regularity.

For the asymptotic upper bound we shall impose the following assumptions.

Assumptions:

A1) Assume that $c(u, w) = \|u - w\|_q^\rho$ for some $q \in (1, \infty]$ and $\rho \geq 1$. For a chosen $q \in (1, \infty]$, let $p \in [1, \infty)$ be such that $1/p + 1/q = 1$.

A2) Suppose that $\theta_* \in \mathbb{R}^l$ satisfies $\mathbb{E}[h(W, \theta_*)] = \mathbf{0}$ and $\mathbb{E}\|h(W, \theta_*)\|_2^2 < \infty$. (While we do not assume that θ_* is unique, the results are stated for a fixed θ_* satisfying $E[h(W, \theta_*)] = \mathbf{0}$.)

A3) Suppose that the function $h(\cdot, \theta_*)$ is continuously differentiable with derivative $D_w h(\cdot, \theta_*)$.

A4) Suppose that for each $\zeta \neq 0$,

$$P\left(\left\|\zeta^T D_w h(W, \theta_*)\right\|_p > 0\right) > 0. \tag{2.12}$$

In order to state the theorem, let us introduce the notation for asymptotic stochastic

upper bound,

$$n^{\rho/2} R_n(\theta_*; \rho) \lesssim_D \bar{R}(\rho),$$

which expresses that for every continuous and bounded non-decreasing function $f(\cdot)$ we have that

$$\overline{\lim}_{n\to\infty} E\left[f\left(n^{\rho/2} R_n(\theta_*; \rho)\right)\right] \leq E\left[f\left(\bar{R}(\rho)\right)\right].$$

Similarly, we write $\gtrsim_D$ for an asymptotic stochastic lower bound, namely

$$\underline{\lim}_{n\to\infty} E\left[f\left(n^{\rho/2} R_n(\theta_*; \rho)\right)\right] \geq E\left[f\left(\bar{R}(\rho)\right)\right].$$

Therefore, if both stochastic upper and lower bounds hold, then $n^{\rho/2} R_n(\theta_*; \rho) \Rightarrow \bar{R}(\rho)$ as $n \to \infty$. (see, for example, Billingsley [2013]). Now we are ready to state our asymptotic upper bound.

Theorem 2.1. Under Assumptions A1) to A4) we have, as $n \to \infty$,

$$n^{\rho/2} R_n(\theta_*; \rho) \lesssim_D \bar{R}(\rho),$$

where, for $\rho > 1$,

$$\bar{R}(\rho) := \max_{\zeta \in \mathbb{R}^r} \left\{ \rho \zeta^T H - (\rho - 1) E \left\| \zeta^T D_w h(W, \theta_*) \right\|_p^{\rho/(\rho-1)} \right\},$$

and if $\rho = 1$,

$$\bar{R}(1) := \max_{\zeta : P\left(\|\zeta^T D_w h(W,\theta_*)\|_p > 1\right) = 0} \{\zeta^T H\}.$$

In both cases $H \sim \mathcal{N}(\mathbf{0}, \mathrm{Cov}[h(W, \theta_*)])$, and $\mathrm{Cov}[h(W, \theta_*)] = E\left[h(W, \theta_*) h(W, \theta_*)^T\right]$.

We remark that as $\rho \to 1$, one can verify that $\bar{R}(\rho) \Rightarrow \bar{R}(1)$, so formally one can simply keep in mind the expression $\bar{R}(\rho)$ with $\rho > 1$. In turn, it is intersting

to note that $\bar{R}(\rho)$ is Fenchel transform as a function of H_n. We now study some sufficient conditions which guarantee that $\bar{R}\,(\rho)$ is also an asymptotic lower bound for $n^{\rho/2}R_n(\theta_*;\rho)$. We consider the case $\rho = 1$ first, which will be used in applications to logistic regression discussed later in the chapter.

Proposition 2.2. In addition to assuming A1) to A4), suppose that W has a positive density (almost everywhere) with respect to the Lebesgue measure. Then,

$$n^{1/2}R_n(\theta_*;1) \Rightarrow \bar{R}\,(1)\,.$$

The following set of assumptions can be used to obtain tight asymptotic stochastic lower bounds when $\rho > 1$; the corresponding result will be applied to the context of square-root Lasso.

A5) (*Growth condition*) Assume that there exists $\kappa \in (0,\infty)$ such that for $\|w\|_q \geq 1$,

$$\|D_w h(w,\theta_*)\|_p \leq \kappa \,\|w\|_q^{\rho-1}\,, \tag{2.13}$$

and that $E\,\|W\|^\rho < \infty$.

A6) (*Locally Lipschitz continuity*) Assume that there exists $\bar{\kappa} : \mathbb{R}^m \to [0,\infty)$ such that,

$$\|D_w h(w+\Delta,\theta_*) - D_w h(w,\theta_*)\|_p \leq \bar{\kappa}\,(w)\,\|\Delta\|_q\,,$$

for $\|\Delta\|_q \leq 1$, and $E\,[\bar{\kappa}(W)^2] < \infty$.

We now summarize our last weak convergence result of this section.

Proposition 2.3. If Assumptions A1) to A6) are in force and $\rho > 1$, then

$$n^{\rho/2}R_n(\theta_*;\rho) \Rightarrow \bar{R}\,(\rho)\,.$$

Before we move on with the applications of the previous results, it is worth discussing the nature of the additional assumptions introduced to ensure that an asymptotic lower bound can be obtained which matches the upper bound in Theorem 2.1.

As we shall see in the technical development in Section 2.A.1. of the Appendix 2.A where the proofs of the above results are furnished, the dual formulation of RWP function in Proposition 2.1 can be re-expressed, assuming only A1) to A4), as,

$$n^{\rho/2} R_n(\theta_*;\rho) = \sup_{\zeta}\left\{\zeta^T H_n - \frac{1}{n}\sum_{k=1}^{n}\sup_{\Delta}\left\{\int_0^1 \zeta^T Dh\left(W_i + \Delta u/n^{1/2}, \theta_*\right)\Delta du - \|\Delta\|_q^{\rho}\right\}\right\}. \tag{2.14}$$

In order to make sure that the lower bound asymptotically matches the upper bound obtained in Theorem 2.1 we need to make sure that we rule out cases in which the inner supremum is infinite in (2.14) with positive probability in the prelimit.

In Proposition 2.2 we assume that W has a positive density with respect to the Lebesgue measure because in that case the condition

$$P\left(\left\|\zeta^T Dh\left(W,\theta_*\right)\right\|_p \leq 1\right) = 1,$$

(which appears in the upper bound obtained in Theorem 2.1) implies that $\left\|\zeta^T Dh\left(w,\theta_*\right)\right\|_p \leq 1$ almost everywhere with respect to the Lebesgue measure. Due to the appearance of the integral in the inner supremum in (2.14), an upper bound can be obtained for the inner supremum, which translates into a tight lower bound for $n^{\rho/2} R_n(\theta_*)$.

Moving to the case $\rho > 1$ studied in Proposition 2.3, condition (2.13) in A5) guarantees that (for fixed W_i and n)

$$\left\|Dh\left(W_i + \Delta u/n^{1/2}, \theta_*\right)\Delta\right\| = O\left(\|\Delta\|_q^{\rho}/n^{(\rho-1)/2}\right),$$

as $\|\Delta\|_q \to \infty$. Therefore, the cost term $\left(-\|\Delta\|_q^\rho\right)$ in (2.14) will ensure a finite optimum in the prelimit for large n. The condition that $E\|W\|_q^\rho < \infty$ is natural because we are using a optimal transport cost $c(u,w) = \|u-w\|_q^\rho$. If this condition is not satisfied, then the underlying nominal distribution is at infinite transport distance from the empirical distribution.

The local Lipschitz assumption A6) is just imposed to simplify the analysis and can be relaxed; we have opted to keep A6) because we consider it mild in view of the applications that we will study in the sequel.

2.3 Distributionally Robust Estimators for Machine Learning Algorithms

A common theme in machine learning problems is to find the best fitting parameter in a family of parameterized models that relate a vector of predictor variables $X \in \mathbb{R}^d$ to a response $Y \in \mathbb{R}$. In this section, we shall focus on a useful class of such models, namely, linear and logistic regression models. Associated with these models, we have a loss function $l(X_i, Y_i; \beta)$ which evaluates the fit of regression coefficient β for the given data points $\{(X_i, Y_i) : i = 1, \ldots, n.\}$ Then, just as we explained in the case of square-root Lasso in the Introduction, our first step will be to show that regularized linear and logistic regression estimators admit a Distributionally Robust Optimization (DRO) formulation of the form,

$$\min_{\beta \in \mathbb{R}^d} \sup_{P: D_c\left(P, P_n\right) \leq \delta} \mathbb{E}_P \left[l\big(X, Y; \beta\big) \right]. \tag{2.15}$$

Once we derive a representation such as (2.15) then we will proceed, in the next

section to find the optimal choice of δ, which, as explained in the Introduction, will immediately characterize the optimal regularization parameter.

In contrast to the empirical risk minimization that performs well only on the training data, the DRO problem (2.15) finds an optimizer β that performs uniformly well over all probability measures in the neighborhood that can be perceived as perturbations to the empirical training data distribution. Hence the solution to (2.15) is said to be "distributionally robust", and can be expected to generalize better. See Xu *et al.* [2009a,b]; Shafieezadeh-Abadeh *et al.* [2015] for works that relate robustness and generalization.

Recasting regularized regression as a DRO problem of form Equation (2.15) lets us view these regularized estimators under the lens of distributional robustness. The regularized estimators that we consider in this section, in particular, include the following.

Example 2.2. (Square-Root-Lasso) We have already started discussing this example in the Section 2.1, namely given a set of training data $\{(X_i, Y_i) : i = 1, \ldots, n\}$, with predictor $X_i \in \mathbb{R}^d$ and response $Y_i \in \mathbb{R}$, the postulated model is $Y_i = \beta_*^T X_i + e_i$ for some $\beta_* \in \mathbb{R}^d$ and errors $\{e_1, ..., e_n\}$. The underlying loss function is $l(x, y; \beta) = \left(y - \beta^T x\right)^2$ and the square-root Lasso estimator, is obtained by solving the problem,

$$\min_{\beta \in \mathbb{R}^d} \left\{ \sqrt{\mathbb{E}_{P_n} \left[l\left(X, Y; \beta\right) \right]} + \lambda \left\| \beta \right\|_1 \right\},$$

see Belloni *et al.* [2011]; Alquier [2008]; Oymak *et al.* [2013] for more on square-root Lasso. As P_n denotes the empirical distribution corresponding to training samples, $\mathbb{E}_{P_n} \left[l\left(X, Y; \beta\right) \right]$ is just the mean square training loss. In addition to the Square-Root Lasso estimator above with ℓ_1 penalty, we derive a DRO represen-

tation of the form (2.15) for ℓ_p-penalized estimators obtained by solving,

$$\min_{\beta\in\mathbb{R}^d}\left\{\sqrt{\mathbb{E}_{P_n}\left[l\left(X,Y;\beta\right)\right]}+\lambda\left\|\beta\right\|_p\right\}, \tag{2.16}$$

for any $p\in[1,\infty)$.

Example 2.3. (Regularized Logistic Regression) We next consider the context of binary classification, in which case the data is of the form $\{(X_i,Y_i):i=1,\ldots,n\}$, with $X_i\in\mathbb{R}^d$, response $Y_i\in\{-1,1\}$ and the model postulates that

$$\log\left(\frac{P\left(Y_i=1|X_i=x\right)}{1-P\left(Y_i=1|X_i=x\right)}\right)=\beta_*^Tx$$

for some $\beta_*\in\mathbb{R}^d$. In this case, the log-exponential loss function (or negative log-likelihood for binomial distribution) is

$$l\left(x,y;\beta\right)=\log\left(1+\exp(-y\cdot\beta^Tx)\right),$$

and one is interested in estimating β_* by solving

$$\min_{\beta\in\mathbb{R}^d}\left\{\mathbb{E}_{P_n}\left[l\left(X,Y;\beta\right)\right]+\lambda\left\|\beta\right\|_p\right\}, \tag{2.17}$$

for $p\in[1,\infty)$ (see Friedman *et al.* [2001] for a discussion on regularized logistic regressions).

The rest of this section is to show that square-root Lasso and Regularized Logistic Regression estimators are distributionally robust (in the sense, they admit a representation of the form (2.15)).

While these particular examples may be certainly interesting, we emphasize that the DRO formulation (2.15) should be viewed, in its entirety, as a framework for

generating distributionally robust inference procedures for different models and loss functions, without having to prove equivalences with an existing or popular algorithm.

2.3.1 Dual form of the DRO formulation (2.15)

Though the DRO formulation (2.15) involves optimizing over uncountably many probability measures, the following result ensures that the inner supremum in (2.15) over the neighborhood $\{P : D_c(P, P_n) \leq \delta\}$ admits a reformulation which is a simple, univariate optimization problem. Before stating the result, we recall that the definition of discrepancy measure $D_c(\cdot)$ (defined in (2.6)) requires the specification of cost function $c((x, y), (x', y'))$ between any two predictor-response pairs $(x, y), (x', y') \in \mathbb{R}^{d+1}$.

Proposition 2.4. Let $c(\cdot)$ be a nonnegative, lower semi-continuous cost function such that the set $\{\big((x, y), (x', y')\big) : c\big((x, y), (x', y')\big) < \infty\}$ is Borel measurable and nonempty. For $\gamma \geq 0$ and loss functions $l(x, y; \beta)$ that are upper semi-continuous in (x, y) for each β, let

$$\phi_\gamma(X_i, Y_i; \beta) = \sup_{u \in \mathbb{R}^d,\ v \in \mathbb{R}} \left\{ l(u, v; \beta) - \gamma c\big((u, v), (X_i, Y_i)\big) \right\}. \tag{2.18}$$

Then

$$\sup_{P:\ D_c(P,P_n) \leq \delta} \mathbb{E}_P\big[l(X, Y; \beta)\big] = \min_{\gamma \geq 0} \left\{ \gamma\delta + \frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(X_i, Y_i; \beta) \right\}.$$

Consequently, the DR regression problem (2.15) reduces to

$$\inf_{\beta \in \mathbb{R}^d} \sup_{P:\ D_c(P,P_n) \leq \delta} \mathbb{E}_P\big[l(X, Y; \beta)\big] = \inf_{\beta \in \mathbb{R}^d} \min_{\gamma \geq 0} \left\{ \gamma\delta + \frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(X_i, Y_i; \beta) \right\}. \tag{2.19}$$

Such reformulations have recently gained much attention in the literature of distributionally robust optimization (see Shafieezadeh-Abadeh *et al.* [2015]; Blanchet and

Murthy [2016]). For a proof of Proposition 2.4, see Appendix 2.B of this chapter.

2.3.2 Distributionally Robust Representations

2.3.2.1 Example 2.2 (continued): Recovering regularized estimators for linear regression

We examine the right-hand side of (2.19) for the square loss function for the linear regression model $Y = \beta^T X + e$, and obtain the following result without any further distributional assumptions on X, Y and the error e. For brevity, let $\bar{\beta} = (-\beta, 1)$, and recall the definition of the discrepancy measure D_c in (2.6).

Proposition 2.5 (DR linear regression with square loss). Fix $q \in (1, \infty]$. Consider the square loss function and second order discrepancy measure D_c defined using ℓ_q-norm. In other words, take $l(x, y; \beta) = (y - \beta^T x)^2$ and $c\big((x, y), (u, v)\big) = \|(x, y) - (u, v)\|_q^2$. Then,

$$\min_{\beta \in \mathbb{R}^d} \sup_{P:\ D_c(P, P_n) \le \delta} \mathbb{E}_P\big[l(X, Y; \beta)\big] = \min_{\beta \in \mathbb{R}^d} \left(\sqrt{MS\mathbb{E}_n(\beta)} + \sqrt{\delta}\ \|\bar{\beta}\|_p \right)^2, \tag{2.20}$$

where $MS\mathbb{E}_n(\beta) = \mathbb{E}_{P_n}[(Y - \beta^T X)^2] = \frac{1}{n} \sum_{i=1}^n (Y_i - \beta^T X_i)^2$ is the mean square error for the coefficient choice β, and p is such that $1/p + 1/q = 1$.

As an important special case, we consider $q = \infty$ and identify the following equivalence for DR regression applying discrepancy measure based on neighborhoods defined using ℓ_∞ norm:

$$\arg\min_{\beta \in \mathbb{R}^d} \sup_{P:\ \mathcal{D}_c(P, P_n) \le \delta} \mathbb{E}_P\big[l(X, Y; \beta)\big] \quad = \arg\min_{\beta \in \mathbb{R}^d} \left\{ \sqrt{MS\mathbb{E}_n(\beta)} + \sqrt{\delta}\ \|\beta\|_1 \right\}.$$

Here the right hand side is same as the square-root Lasso estimator with $\lambda = \sqrt{\delta}$ in Example 2.2.

The right hand side of (2.20) resembles ℓ_p-norm regularized regression (except for the fact that we have $\|\bar{\beta}\|_p$ instead of $\|\beta\|_p$). In order to obtain a closer equivalence we must introduce a slight modification to the norm $\|\cdot\|_q$ to be used as the cost function, $c(\cdot)$, in defining D_c. We define

$$N_q\big((x,y),(u,v)\big) = \begin{cases} \|x-u\|_q, & \text{if } y = v \\ \infty, & \text{otherwise.} \end{cases}, \tag{2.21}$$

to use $c(\cdot) = N_q(\cdot)$ as the cost instead of the standard ℓ_q norm $\|(x,y)-(u,v)\|_q$. Subsequently, one can consider modified cost functions of form $c((x,y),(u,v)) = (N_q((x,y),(u,v)))^a$. As this modified cost function assigns infinite cost when $y \neq v$, the infimum in (2.4) is effectively over joint distributions that do not alter the marginal distribution of Y. As a consequence, the resulting neighborhood set $\{P : D_c(P,P_n) \leq \delta\}$ admits distributional ambiguities only with respect to the predictor variables X.

The following result is essentially the same as Proposition 2.5 except for the use of the modified cost N_q and the resulting norm regularization of form $\|\beta\|_p$ (instead of $\|\bar{\beta}\|_p$ as in Proposition 2.5), thus exactly recovering the regularized regression estimators in Example 2.2.

Theorem 2.2. Consider the square loss and discrepancy measure $D_c(P,P_n)$ defined as in (2.6) using the cost function $c((x,y),(u,v)) = (N_q((x,y),(u,v)))^2$ (the function N_q is defined in (2.21)). Then,

$$\min_{\beta\in\mathbb{R}^d} \sup_{P:\ D_c(P,P_n)\leq\delta} \mathbb{E}_P\big[l(X,Y;\beta)\big] = \min_{\beta\in\mathbb{R}^d} \Big(\sqrt{MS\mathbb{E}_n(\beta)} + \sqrt{\delta}\ \|\beta\|_p\Big)^2,$$

where $MSE_n(\beta) = \mathbb{E}_{P_n}[(Y-\beta^T X)^2] = n^{-1}\sum_{i=1}^n (Y_i - \beta^T X_i)^2$ is the mean square error for the coefficient choice β, and p is such that $1/p + 1/q = 1$.

2.3.2.2 Example 2.3 (continued): Recovering regularized estimators for classification

Apart from exactly recovering well-known norm regularized estimators for linear regression, the discrepancy measure D_c based on the modified norm N_q in (2.21) is natural when our interest is in learning problems where the responses Y_i take values in a finite set – as in the binary classification problem where the response variable Y takes values in $\{-1, +1\}$.

The following result allows us to recover the DRO formulation behind the regularized logistic regression estimators discussed in Example 2.3 and support vector machine with Hinge loss function, i.e. $l(x, y, \beta) = \left(1 - y\beta^T x\right)^+$.

Theorem 2.3 (Regularized regression for Classification). Consider the discrepancy measure $D_c(\cdot)$ defined using the cost function $c((x, y), (u, v)) = N_q((x, y), (u, v))$ in (2.21). Then, for logistic regression with log-exponential loss function and Support Vector Machine (SVM) with Hinge loss,

$$\min_{\beta \in \mathbb{R}^d} \sup_{P:\ D_c(P, P_n) \leq \delta} \mathbb{E}_P\big[\log(1 + e^{-Y\beta^T X})\big] = \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-Y_i \beta^T X_i}\right) + \delta \left\|\beta\right\|_p,$$

and

$$\min_{\beta \in \mathbb{R}^d} \sup_{P:\ D_c(P, P_n) \leq \delta} \mathbb{E}_P\big[(1 - Y\beta^T X)^+\big] = \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \beta^T X_i)^+ + \ \delta \left\|\beta\right\|_p,$$

where p is such that $1/p + 1/q = 1$.

The proof of all of the results in this subsection are provided in Appendix 2.A Section 2.A.2. of this chapter.

2.4 Using RWPI for optimal regularization

Our goal in this section is to use RWP function for optimal regularization in Examples 2.2 and 2.3. As explained in the Introduction, the key step is to propose a reasonable optimality criterion for the selection of δ in the DRO formulation Equation (2.15). Then, owing to the DRO representations derived in Section 2.3.2, this would imply an automatic choice of regularization parameter $\lambda = \sqrt{\delta}$ in square-root Lasso example (following Theorem 2.2), or $\lambda = \delta$ in regularized logistic regression (following Theorem 2.3). In the development below, we follow the logic described in the Introduction for the square-root Lasso setting.

We write $\mathcal{U}_\delta(P_n)$ to denote the uncertainty set, namely $\mathcal{U}_\delta(P_n) = \{P : D_c\big(P, P_n\big) \leq \delta\}$, and β_* to denote the underlying linear or logistic regression model parameter from which the training samples $\{(X_i, Y_i) : i = 1, \ldots, n\}$ are obtained. Now, for each P, convexity considerations involving the loss functions $l\,(x, y; \beta)$, as a function of β, will allow us to conclude that the set

$$\mathcal{P}_{opt}(\beta) := \big\{P \in \mathcal{P}\big(\mathbb{R}^d \times\ \mathbb{R}\big) : \mathbb{E}_P\big[D_\beta l(X, Y; \beta_*)\big] = \mathbf{0}\big\}$$

is the set of probability measures for which β is an optimal risk minimization parameter.

As indicated in the Introduction, we shall say that β_* is plausible for a given choice of δ if,

$$\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing.$$

If this intersection is empty, we say that β_* is implausible. Moreover, we remark that

β_* is plausible with confidence at least $1-\alpha$ if,

$$P\left(\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing\right) \geq 1-\alpha.$$

We shall argue in Appendix 2.C of this chapter that the inf sup in the corresponding DRO formulation (2.15) of each of the machine learning algorithms that we consider can be exchanged as below:

Lemma 2.1. In the settings of Theorems 2.2 and 2.3, if $\mathbb{E}\|X\|_2^2 < \infty$, we have that

$$\inf_{\beta\in\mathbb{R}^d} \sup_{P\in\ \mathcal{U}_\delta(P_n)} \mathbb{E}_P\left[l\big(X,Y;\beta\big)\right] = \sup_{P\in\ \mathcal{U}_\delta(P_n)} \inf_{\beta\in\mathbb{R}^d} \mathbb{E}_P\left[l\big(X,Y;\beta\big)\right]. \tag{2.22}$$

The representation in the right hand side of (2.22) implies that

$$\begin{aligned}
&\sup_{P\in\ \mathcal{U}_\delta(P_n)} \inf_{\beta\in\mathbb{R}^d} \mathbb{E}_P\left[l\big(X,Y;\beta\big)\right] \\
&= \sup_{P\in\ \mathcal{U}_\delta(P_n)} \left\{\mathbb{E}_P\left[l\big(X,Y;\beta\big)\right] : \beta\in\mathbb{R}^d \text{ such that } \mathbb{E}_P\big[D_\beta l(X,Y;\beta)\big] = \mathbf{0}\right\} \\
&= \sup\left\{\mathbb{E}_P\left[l\big(X,Y;\beta\big)\right] : P\in\mathcal{U}_\delta(P_n), \beta\in\mathbb{R}^d \text{ such that } \mathcal{P}_{opt}(\beta)\ \cap\ \mathcal{U}_\delta(P_n) = \varnothing\right\},
\end{aligned}$$

and this motivates our interest in finding the smallest $\delta > 0$ such that

$$P\big(\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing\big) \geq 1-\alpha \tag{2.23}$$

asymptotically, as $n\to\infty$. In simple words, we wish to find the smallest value of δ for which β_* is plausible with at least $1-\alpha$ confidence (see Figure 2.1).

Observe that as

$$R_n(\beta_*) = \inf\big\{D_c(P, P_n) : P \in \mathcal{P}_{opt}(\beta_*)\big\},$$

we have,

$$P\big(\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing\big) = P\big(R_n(\beta_*) \leq \delta\big)$$

and therefore (2.23) is equivalent to

$$\inf\big\{\delta : P\left(R_n(\beta_*) \leq \delta\right) \geq 1 - \alpha\big\}, \tag{2.24}$$

thus obtaining the optimal selection of δ as the $1-\alpha$ quantile of $R_n(\beta_*)$.

Now, without knowing β_*, it is, of course, difficult to compute $R_n(\beta_*)$. However, assuming i.i.d. training data, we can obtain a limiting distribution for the quantity $nR_n(\beta_*)$ or $\sqrt{n}R_n(\beta_*)$, by applying results from Section 2.2.4.

Another consequence of Lemma 2.1 is that the set $\Lambda_n(\delta)$ of plausible values of β (i.e. β for which there exists $P \in \mathcal{U}_\delta(P_n)$ such that $\mathbb{E}_P\big[D_\beta l(X, Y; \beta)\big] = \mathbf{0}$), contains the optimal solution obtained by solving the problem in the left hand side of (2.22). (If this was not the case, the left hand side in (2.22) would be strictly smaller than the right hand side of (2.22).) The fact that the estimator for β_* obtained by solving the left hand side in (2.22) is plausible, we believe, is a property which makes our selection of δ logically consistent with the ultimate goal of the overall estimation procedure, namely, choosing β_*.

2.4.1 Linear regression models with squared loss function

In this section, we derive the asymptotic limiting distribution of suitably scaled profile function corresponding to the estimating equation

$$E[(Y - \beta^T X)X] = \mathbf{0}.$$

The chosen estimating equation describes the optimality condition for square loss function $l(x, y; \beta) = (y - \beta^T x)^2$, and therefore, the corresponding $R_n(\beta_*)$ is a suitable for choosing δ as in Equation (2.24), and the regularization parameter $\lambda = \sqrt{\delta}$ in Example 2.2.

Let H_0 denote the null hypothesis that the training samples $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ are obtained independently from the linear model $Y = \beta_*^T X + e$, where the error term e has zero mean, variance σ^2, and is independent of X. Let $\Sigma = \mathrm{Cov}[X]$.

Theorem 2.4. Consider the discrepancy measure $D_c(\cdot)$ defined as in (2.6) using the cost function $c((x, y), (u, v)) = (N_q((x, y), (u, v)))^2$ (the function N_q is defined in (2.21)). For $\beta \in \mathbb{R}^d$, let

$$R_n(\beta) = \inf \big\{ D_c(P, P_n) : \ \mathbb{E}_P\big[(Y - \beta^T X)X\big] = \mathbf{0} \big\}.$$

Then, under the null hypothesis H_0,

$$nR_n(\beta_*) \Rightarrow L_1 := \max_{\xi \in \mathbb{R}^d} \Big\{ 2\sigma \xi^T Z - \mathbb{E} \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 \Big\},$$

as $n \to \infty$. In the above limiting relationship, $Z \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Further,

$$L_1 \overset{D}{\leq} \ L_2 := \frac{\mathbb{E}[e^2]}{\mathbb{E}[e^2] - (\mathbb{E}|e|)^2} \|Z\|_q^2.$$

Specifically, if the additive error term e follows a normal distribution with zero mean, then

$$L_1 \overset{D}{\leq} L_2 := \frac{\pi}{\pi - 2}\|Z\|_q^2.$$

In the above theorem, the relationship $L_1 \overset{D}{\leq} L_2$ denotes that the limit law L_1 is stochastically dominated by L_2. We remark this notation $\overset{D}{\leq}$ for stochastic upper bound here is different from the notation $\lesssim_D$ introduced in Section 2.2.4 to denote asymptotic stochastic upper bound. A proof of Theorem 2.4 as an application of Theorem 2.1 and Proposition 2.3 is presented in Section 2.A.3. of Appendix 2.A in this chapter.

Using Theorem 2.4 to obtain regularization parameter for (2.16). Let $\eta_{1-\alpha}$ denote the $(1-\alpha)$ quantile of the limiting random variable L_1 in Theorem 2.4, or its stochastic upper bound L_2. If we choose $\delta = \eta_{1-\alpha}/n$, it follows from Theorem 2.4 that

$$P(R_n(\beta_*) \leq \delta) \geq 1 - \alpha,$$

asymptotically as $n \to \infty$, and consequently,

$$P\big(\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing\big) \geq 1 - \alpha.$$

In other words, the optimal regression coefficient β_* remains plausible (for the DRO formulation Equation (2.15)) with probability exceeding $1 - \alpha$ with this choice of δ. Due to the distributionally robust representation derived in Theorem 2.2, a prescription for the uncertainty set size δ naturally provides the prescription, $\lambda = \sqrt{\delta}$, for the regularization parameter as well. The following steps summarize the guidelines for choosing the regularization parameter in ℓ_p-penalized linear regression (2.16):

1) Draw samples Z from $\mathcal{N}(\mathbf{0}, \Sigma)$ to estimate the $1-\alpha$ quantile of one of the random variables L_1 or L_2 in Theorem 2.4. Let us use $\hat{\eta}_{1-\alpha}$ to denote the estimated quantile. While L_2 is simply the norm of Z, obtaining realizations of limit law L_1 involves solving an optimization problem for each realization of Z. If $\Sigma = \mathrm{Cov}[X]$ is not known, one can use a simple plug-in estimator for $\mathrm{Cov}[X]$ in place of Σ.

2) Choose the regularization parameter λ to be

$$\lambda = \sqrt{\delta} = \sqrt{\hat{\eta}_{1-\alpha}/n}.$$

It is interesting to note that unlike the traditional Lasso algorithm, the prescription of regularization parameter in the above procedure is self-normalizing, in the sense that it does not depend on $\mathrm{Var}[e]$.

2.4.2 Logistic Regression with log-exponential loss function

In this section, we apply results in Section 2.2.4 to prescribe regularization parameter for ℓ_p-penalized logistic regression in Example 2.3.

Let H_0 denote the null hypothesis that the training samples $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ are obtained independently from a logistic regression model satisfying

$$\log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = \beta_*^T x,$$

for predictors $X \in \mathbb{R}^d$ and corresponding responses $Y \in \{-1, 1\}$; further, under null hypothesis H_0, the predictor X has positive density almost everywhere with respect to the Lebesgue measure on $\mathbb{R}^d$. The log-exponential loss (or negative log-likelihood)

that evaluates the fit of a logistic regression model with coefficient β is given by

$$l(x, y; \beta) = -\log p(y|x; \beta) = \log\big(1 + \exp(-y\beta^T x)\big).$$

If we let

$$h(x, y; \beta) = D_\beta l(x, y; \beta) = \frac{-yx}{1 + \exp(y\beta^T x)}, \tag{2.25}$$

then the optimality condition that the coefficient β^* satisfies is $E\left[h(x, y; \beta_*)\right] = \mathbf{0}$.

Theorem 2.5. Consider the discrepancy measure $D_c(\cdot)$ defined as in (2.6) using the cost function $c((x, y), (u, v)) = N_q((x, y), (u, v))$ (the function N_q is defined in (2.21)). For $\beta \in \mathbb{R}^d$, let

$$R_n(\beta) = \inf\left\{D_c(P, P_n) : \mathbb{E}_P\big[h(x, y; \beta)\big] = \mathbf{0}\right\},$$

where $h(\cdot)$ is defined in (2.25). Then, under the null hypothesis H_0,

$$\sqrt{n}R_n(\beta_*) \Rightarrow \; L_3 := \sup_{\xi \in A} \xi^T Z$$

as $n \to \infty$. In the above limiting relationship,

$$Z \sim \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\frac{XX^T}{(1 + \exp(Y\beta_*^T X))^2}\right]\right) \text{ and}$$
$$A = \left\{\xi \in \mathbb{R}^d : \; ess\ sup_{x,y}\left\|\xi^T D_x h(x, y; \beta)\right\|_p \; \leq 1\right\}.$$

Moreover, the limit law L_3 admits the following simpler stochastic bound:

$$L_3 \overset{D}{\leq} \; L_4 := \|\tilde{Z}\|_q,$$

where $\tilde{Z} \sim \mathcal{N}(\mathbf{0}, \mathbb{E}[XX^T])$.

A proof of Theorem 2.4 as an application of Theorem 2.1 and Proposition 2.2 is presented in Section 2.A.3. of Appendix 2.A in this chapter.

Using Theorem 2.5 to obtain regularization parameter for (2.17). Similar to linear regression, the regularization parameter for Regularized Logistic Regression discussed in Example 2.3 can be chosen by the following procedure:

1) Estimate the $(1-\alpha)$ quantile of $L_4 := \|\tilde{Z}\|_q$, where $\tilde{Z} \sim \mathcal{N}(\mathbf{0}, \mathbb{E}[XX^T])$. Let us use $\hat{\eta}_{1-\alpha}$ to denote the estimate of the quantile.

2) Choose the regularization parameter λ in the norm regularized logistic regression estimator (2.17) in Example 2.3 to be

$$\lambda = \delta = \hat{\eta}_{1-\alpha}/\sqrt{n}.$$

2.4.3 Optimal regularization in high-dimensional square-root Lasso

In this section, let us restrict our attention to the square-loss function $l(x, y; \beta) = (y - \beta^T x)^2$ for the linear regression model and the discrepancy measure D_c defined using the cost function $c = N_q$ with $q = \infty$ in (2.21). Then, due to Theorem 2.2, this corresponds to the interesting case of square-root Lasso or ℓ_2-Lasso that was rather a particular example in the class of ℓ_p-penalized linear regression estimators considered in Section 2.4.1.

As an interesting byproduct of the RWP function analysis, the following theorem presents a prescription for regularization parameter even in high dimensional settings

where the ambient dimension d is larger than the number of samples n. We introduce the growth parameter,

$$C(n,d) := \frac{\mathbb{E}\,\|X\|_\infty}{\sqrt{n}} = \frac{\mathbb{E}\left[\max_{i=1,\ldots,d}|X_i|\right]}{\sqrt{n}},$$

as a function of n and d, that will be useful in stating our results. In addition, we say that the predictors X have *sub-gaussian tails* if there exists a constant $a > 0$,

$$E\left[\exp(t^T X)\right] \leq \exp(a^2\|t\|_2^2/2)$$

for every $t \in \mathbb{R}^d$.

Theorem 2.6. Let $E[X_i] = 0$ and $E[X_i^2] = 1$ for all $i = 1, \ldots, d$. Suppose the assumptions of Theorem 2.4 hold and assume the largest eigenvalue of $\Sigma = \mathrm{Cov}[X]$ be $o(nC(n,d)^2)$. In addition, suppose that β_* satisfies a weak sparsity condition that $\|\beta_*\|_1 = o(1/C(n,d))$. Then

$$nR_n(\beta_*) \lesssim_D \frac{\|Z_n\|_\infty^2}{\mathrm{Var}|e|},$$

as $n, d \to \infty$. Here, $Z_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i X_i$. In particular, if the predictors X have subgaussian tails, then we have

$$nR_n(\beta_*) \lesssim_D \frac{\mathbb{E}e^2}{\mathbb{E}e^2 - \left(\mathbb{E}|e|\right)^2}\|\tilde{Z}\|_\infty^2$$

where, $\tilde{Z}$ follows the distribution $\mathcal{N}(0, \Sigma)$. Moreover, if the additive error e is normally distributed and Σ is the identity matrix, then the above stochastic bounds simplify to

$$\sqrt{R_n(\beta_*)} \lesssim_D \frac{\pi}{\pi - 2}\frac{\Phi^{-1}(1-\alpha/2d)}{\sqrt{n}},$$

with probability asymptotically larger than $1-\alpha$. Here, $\Phi^{-1}(1-\alpha)$ denotes the quantile x of the standard normal distribution $\Phi(x) = 1-\alpha$.

The prescription of regularization parameter as

$$\lambda = \sqrt{\delta} = \frac{\pi}{\pi - 2} \frac{\Phi^{-1}(1-\alpha/2d)}{\sqrt{n}} = O\left(\sqrt{\frac{\log d}{n}}\right), \tag{2.26}$$

as in Theorem 2.6, is consistent with the findings in the literature of high-dimensional linear regression (see, for example, Belloni *et al.* [2011]; Nguyen [2013]; Zhou [2015]; Banerjee *et al.* [2014]). This agreement strengthens the interpretation of regularization parameter in regularized regression as $\sqrt{R_n(\beta_*)}$, which, in turn, corresponds to the distance of the empirical distribution P_n from the set $\{P : \mathbb{E}_P[(Y-\beta^T X)X] = \mathbf{0}\}$.

It is also interesting to note that unlike traditional Lasso algorithm, the prescription of regularization parameter as in Equation (2.26) is self-normalizing, in the sense that it does not depend on the variance of e, even if the number of predictors d is larger than n.

2.5 Conclusion

This chapter has introduced the basic principles behind the application of RWPI, we believe that the systematic use of distributionally robust optimization based on optimal transport considerations has the potential to be utilized in a wide range of settings. In addition to new applications of RWPI there are key statistical properties which remain to be studied. The well-developed literature on EL may serve as a template, not only for the development of future applications of RWPI, but also for further investigation of the RWP function, which is key in the use of this methodology. These additional developments and investigations will be reported in future research.

Additional Material TO CHAPTER 2

This additional material for the RWPI chapter is organized as follows: Proofs of all the main results in the chapter are furnished in APPENDIX 2.A. As some of the main results in the chapter utilize strong duality for problems of moments, a quick introduction to problem of moments along with a well-known strong duality result that is useful in our context is provided in APPENDIX 2.B. A technical result on exchange of sup and inf in the DRO formulation Equation (2.15) is presented in APPENDIX 2.C. Numerical experiments that compare RWPI based regularization parameter selection with cross-validation based approach are presented in APPENDIX 2.D.

APPENDIX 2.A: Proofs of main results

This section, comprising the proofs of the main results, is organized as follows. Subsection 2.A.1 contains the proofs of stochastic upper and lower bounds (and hence weak limits) presented in Section 2.2.4. While Subsection 2.A.2 is devoted to derive the results on distributionally robust representations presented in Section 2.3.2, Subsection 2.A.3 contains the proofs of Theorems 2.4 and 2.5 as applications of the stochastic upper and lower bounds presented in Section 2.2.4. Some of the useful technical results that are not central to the argument are presented in appendices 2.B and 2.C.

2.A.1. Proofs of asymptotic stochastic upper and lower bounds of RWP function in Section 2.2.4

We first use Proposition 2.1 to derive a dual formulation for $n^{\rho/2}R_n(\theta_*)$ which will be the starting point of our analysis. Due to Assumption A2) $E[h(W,\theta_*)] = \mathbf{0}$, and

therefore, $\mathbf{0}$ lies in the interior of convex hull of $\{h(u,\theta_*) : u \in \mathbb{R}^m\}$. Therefore, due to Proposition 2.1,

$$R_n(\theta_*) = \sup_{\lambda\in\mathbb{R}^r}\left\{-\frac{1}{n}\sum_{i=1}^{n}\sup_{u\in\mathbb{R}^m}\left\{\lambda^T h(u,\theta_*) - \|u-W_i\|_q^\rho\right\}\right\}.$$

In order to simplify the notation, throughout the rest of the proof we will write $h(W_i)$ instead of $h(W_i,\theta_*)$ and $Dh(W_i)$ for $D_w h(W_i,\theta_*)$.

Letting $H_n = n^{-1/2}\sum_{i=1}^n h(W_i)$ and changing variables to $\Delta = u - W_i$, we obtain

$$R_n(\theta_*) = \sup_{\lambda}\left\{-\lambda^T\frac{H_n}{n^{1/2}} - \frac{1}{n}\sum_{i=1}^{n}\sup_{\Delta}\left\{\lambda^T\big(h(W_i+\Delta)-h(W_i)\big) - \|\Delta\|_q^\rho\right\}\right\}.$$

Due to the fundamental theorem of calculus (using Assumption A3)), we have that

$$h(W_i+\Delta) - h(W_i) = \int_0^1 Dh(W_i+u\Delta)\,\Delta du.$$

Now, redefining $\zeta = \lambda n^{(\rho-1)/2}$ and $\Delta = \Delta/n^{1/2}$ we arrive at following representation

$$n^{\rho/2}R_n(\theta_*) = \sup_{\zeta}\left\{-\zeta^T H_n - M_n(\zeta)\right\}, \tag{2.27}$$

where

$$M_n(\zeta) = \frac{1}{n}\sum_{i=1}^{n}\sup_{\Delta}\left\{\zeta^T\int_0^1 Dh\left(W_i+n^{-1/2}\Delta u\right)\Delta du - \|\Delta\|_q^\rho\right\}. \tag{2.28}$$

The reformulation in Equation (2.27) is our starting point of the analysis.

To proceed further, we first state a result which will allow us to apply a localization argument in the representation of $n^{\rho/2}R_n(\theta_*)$ in Equation (2.27). Recall the definition of M_n above in Equation (2.28) and that $H_n = n^{-1/2}\sum_{i=1}^n h(W_i)$.

Lemma 2.2. Suppose that the Assumptions A2) to A4) are in force. Then, for every $\varepsilon > 0$, there exists $n_0 > 0$ and $b \in (0, \infty)$ such that

$$P\left(\sup_{\|\zeta\|_p \geq b} \left\{-\zeta^T H_n - M_n(\zeta)\right\} > 0\right) \leq \varepsilon,$$

for all $n \geq n_0$.

Proof of Lemma 2.2. For $\zeta \neq 0$, we write $\bar{\zeta} = \zeta / \|\zeta\|_p$. Let us define the vector $V_i\left(\bar{\zeta}\right) = Dh\left(W_i\right)^T \bar{\zeta}$, and put

$$\Delta_i' = \Delta_i'\left(\bar{\zeta}\right) = \left|V_i\left(\bar{\zeta}\right)\right|^{p/q} sgn\left(V_i\left(\bar{\zeta}\right)\right). \tag{2.29}$$

Define the set $C_0 = \{\|W_i\|_p \leq c_0\}$, where c_0 will be chosen large enough momentarily. Then, for any $c > 0$, plugging in $\Delta = c\Delta_i'$, we have $\zeta^T Dh(W_i)\Delta = c\|\zeta^T Dh(W_i)\|_p \|\Delta'\|_q$, and therefore,

$$\begin{aligned}
&\sup_{\Delta}\left\{\zeta^T \int_0^1 Dh(W_i + n^{-1/2}\Delta u)\Delta du - \|\Delta\|_q^\rho\right\} \\
&\quad = \sup_{\Delta}\left\{\zeta^T Dh(W_i)\Delta - \|\Delta\|_q^\rho + \zeta^T \int_0^1 \left[Dh(W_i + n^{-1/2}\Delta u) - Dh(W_i)\right]\Delta du\right\} \\
&\quad \geq \max\Bigg\{c\left\|\zeta^T Dh(W_i)\right\|_p \left\|\Delta_i'\right\|_q - c^\rho \left\|\Delta_i'\right\|_q^\rho \\
&\qquad\qquad + c\zeta^T \int_0^1 \left[Dh(W_i + cn^{-1/2}\Delta_i' u) - Dh\left(W_i\right)\right]\Delta_i' du, 0\Bigg\} I\left(W_i \in C_0\right).
\end{aligned} \tag{2.30}$$

Due to Hölder's inequality,

$$\begin{aligned}
&I\left(W_i \in C_0\right)\left|\zeta^T \int_0^1 \left[Dh(W_i + cn^{-1/2}\Delta_i' u) - Dh(W_i)\right]\Delta_i' du\right| \\
&\qquad \leq I\left(W_i \in C_0\right)\|\zeta\|_p \int_0^1 \left\|\left[Dh(W_i + cn^{-1/2}\Delta_i' u) - Dh(W_i)\right]\Delta_i'\right\|_q du.
\end{aligned}$$

Because of continuity $Dh(\cdot)$ and the fact that $W_i \in C_0$ (so the integrand is bounded), we have that the previous expression converges to zero as $n \to \infty$. Therefore, for given positive constants ε', c (note than convergence is uniform on $W_i \in C_0$), there exists n_0 such that for all $n \geq n_0$

$$cI\left(W_i \in C_0\right)\left|\zeta^T \int_0^1 \left[Dh(W_i + cn^{-1/2}\Delta_i' u) - Dh(W_i)\right]\Delta_i' du\right| \leq c\varepsilon' \left\|\zeta\right\|_p. \tag{2.31}$$

Next, as $\|\bar{\zeta}^T Dh(W_i)\|_p^{p/q} = \|\Delta_i'\|_q$ and $1 + p/q = p$,

$$c\left\|\zeta^T Dh\left(W_i\right)\right\|_p \left\|\Delta_i'\right\|_q - c^\rho \left\|\Delta_i'\right\|_q^\rho = c\left\|\zeta\right\|_p \|\bar{\zeta}^T Dh(W_i)\|_p^p - c^\rho \|\bar{\zeta}^T Dh(W_i)\|_p^{\rho\frac{p}{q}}.$$

Consequently, it follows from Equation (2.30) and Equation (2.31) that

$$M_n(\zeta) \geq \frac{1}{n}\sum_{i=1}^n \left\{c\left\|\zeta\right\|_p \|\bar{\zeta}^T Dh(W_i)\|_p^p - c^\rho \|\bar{\zeta}^T Dh(W_i)\|_p^{\rho\frac{p}{q}} - c\varepsilon'\|\zeta\|_p\right\} I\left(W_i \in C_0\right). \tag{2.32}$$

Now, since the map $\bar{\zeta} \hookrightarrow \left\|\bar{\zeta}^T Dh(W_i)\right\|_p^p$ is Lipschitz continuous on $\left\|\bar{\zeta}\right\|_p = 1$, we conclude that,

$$\frac{1}{n}\sum_{i=1}^n \left\|\bar{\zeta}^T Dh(W_i)\right\|_p^p I\left(W_i \in C_0\right) \to \mathbb{E}\left[\left\|\bar{\zeta}^T Dh\left(W\right)\right\|_p^p I\left(W \in C_0\right)\right], \tag{2.33}$$

with probability one as $n \to \infty$. Moreover, due to Fatou's lemma we have that the map $\bar{\zeta} \hookrightarrow P\left(\left\|\bar{\zeta}^T Dh\left(W\right)\right\|_p > 0\right)$ is lower semi-continuous. Therefore, by A4), we have that there exists $\delta > 0$ such that

$$\inf_{\bar{\zeta}} \mathbb{E}\left\|\bar{\zeta}^T Dh\left(W\right)\right\|_p^p > \delta. \tag{2.34}$$

Consecutively, by selecting $c_0 > 0$ large enough, we conclude from Equation (2.33) that for $n \geq N'(\delta)$,

$$\frac{1}{n}\sum_{i=1}^{n} \left\|\bar{\zeta}^T Dh(W_i)\right\|_p^p I\left(W_i \in C_0\right) > \frac{\delta}{2}. \tag{2.35}$$

Further, if we let $c_1 := \sup_{w\in C_0} \|\bar{\zeta}^T Dh(w)\|_p^{p/q} < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} \left\|\bar{\zeta}^T Dh(w)\right\|_p^{\rho\frac{p}{q}} I\left(W_i \in C_0\right) < c_1^{\rho},$$

for all $n > N'(\delta)$. As a consequence, if $n \geq N'(\delta)$, it follows from Equation (2.32) and Equation (2.35) that

$$\begin{aligned}\sup_{\|\zeta\|_p>b} \left\{-\zeta^T H_n - M_n\left(\zeta\right)\right\} &\leq \sup_{\|\zeta\|_p>b} \left\{-\zeta^T H_n - \left(\frac{c\delta\|\zeta\|_p}{2} - (cc_1)^{\rho} - c\varepsilon'\|\zeta\|_p\right)\right\} \\ &\leq \sup_{\|\zeta\|_p>b} \left\{-\zeta^T H_n - \|\zeta\|_p \left\{c\left(\frac{\delta}{2} - \varepsilon'\right) - \frac{(cc_1)^{\rho}}{b}\right\}\right\}.\end{aligned}$$

Consequently, on the set $\|H_n\|_q \leq b'$, we obtain

$$\sup_{\|\zeta\|_p>b} \left\{-\zeta^T H_n - M_n\left(\zeta\right)\right\} \leq \sup_{\|\zeta\|_p>b} \|\zeta\|_p \left[b' - \left\{c\left(\frac{\delta}{2} - \varepsilon'\right) - \frac{(cc_1)^{\rho}}{b}\right\}\right].$$

Now, if we take $c > 4(b'+1)/\delta$, $\varepsilon' = \delta/4$ and b to be large enough such that $b > (cc_1)^{\rho}$ then

$$b' - \left\{c\left(\frac{\delta}{2} - \varepsilon'\right) - \frac{(cc_1)^{\rho}}{b}\right\} < 0.$$

Therefore, if $n \geq n_0$ (see Equation (2.31)), then

$$P\left(\max_{\|\zeta\|_p > b} \left\{-2\zeta^T H_n - M_n(\zeta)\right\} > 0\right) \leq P\left(\|H_n\|_q > b'\right) + P\left(N'(\delta) > n\right).$$

The result now follows immediately from the previous inequality by choosing b' large enough so that $P(\|H_n\|_q > b') \leq \varepsilon/2$ and later n_0 so that $P(N'(\delta) > n_0) \leq \varepsilon/2$. The selection of b' is feasible due to A2). This proves the statement of Lemma 2.2. □

Lemma 2.3. For any $b > 0$ and $c_0 \in (0, \infty)$,

$$\frac{1}{n}\sum_{i=1}^{n} \left\|\zeta^T Dh(W_i)\right\|_p^{\rho/(\rho-1)} I\left(\|W_i\|_p \leq c_0\right) \to \mathbb{E}\left[\left\|\zeta^T Dh(W)\right\|_p^{\rho/(\rho-1)} I(\|W\|_p \leq c_0)\right],$$

uniformly over $\|\zeta\|_p \leq b$ in probability as $n \to \infty$.

Proof of Lemma 2.3. We first argue a suitable Lipschitz property for the map $\zeta \hookrightarrow \left\|\zeta^T Dh(W_i)\right\|_p^{\rho/(\rho-1)}$. It is elementary that for any $0 \leq a_0 < a_1$ and $\gamma > 1$

$$a_1^\gamma - a_0^\gamma = \gamma \int_{a_0}^{a_1} t^{\gamma-1} dt \leq \gamma a_1^{\gamma-1}(a_1 - a_0).$$

Applying this observation with

$$\begin{aligned} a_1 &= \max\left(\left\|\zeta_1^T Dh(W_i)\right\|_p, \left\|\zeta_0^T Dh(W_i)\right\|_p\right), \\ a_0 &= \min\left(\left\|\zeta_1^T Dh(W_i)\right\|_p, \left\|\zeta_0^T Dh(W_i)\right\|_p\right), \\ \gamma &= \rho/(\rho-1), \end{aligned}$$

and using that $\left\|\zeta^T Dh(W_i)\right\|_p \leq b\|Dh(W_i)\|_p$ for $\|\zeta\|_p \leq b$, we obtain

$$\left|\left\|\zeta_0^T Dh(W_i)\right\|_p^{\rho/(\rho-1)} - \left\|\zeta_1^T Dh(W_i)\right\|_p^{\rho/(\rho-1)}\right| \leq \frac{\rho}{\rho-1} b^{1/(\rho-1)} \|Dh(W_i)\|_p^{\rho/(\rho-1)} \|\zeta_0 - \zeta_1\|_p.$$

Consequently, we have that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left\|\zeta_0^T Dh\left(W_i\right)\right\|_p^{\frac{\rho}{\rho-1}}-\frac{1}{n}\sum_{i=1}^{n}\left\|\zeta_1^T Dh\left(W_i\right)\right\|_p^{\frac{\rho}{\rho-1}}\right|\leq\frac{\rho}{\rho-1}\left\|\zeta_0-\zeta_1\right\|_p\frac{b^{\frac{1}{\rho-1}}}{n}\sum_{i=1}^{n}\left\|Dh\left(W_i\right)\right\|_p^{\frac{\rho}{\rho-1}}.$$

Since $Dh(\cdot)$ is continuous, $\mathbb{E}\left[\left\|Dh\left(W\right)\right\|_p^{\rho/(\rho-1)}I(\left\|W\right\|_p\leq c_0)\right]<\infty$, thus yielding the tightness of

$$\frac{1}{n}\sum_{i=1}^{n}\|\zeta^T Dh(W_i)\|_p^{\rho/(\rho-1)}I\left(\left\|W_i\right\|_p\leq c_0\right),$$

under the uniform topology on compact sets. The Strong Law of Large Numbers guarantees that finite dimensional distributions converge (for any choice of $\zeta_1,\ldots,\zeta_k, k\geq 1$), and, since the limit is deterministic, we obtain the desired convergence in probability. □

Proof of Theorem 2.1. Let us first observe that $R_n(\theta_*)\geq 0$ (choosing $\zeta=0$). Then, as a consequence of Lemma 2.2, there exists $b>0$ such that the event

$$\mathcal{A}_n=\left\{n^{\rho/2}R_n(\theta_*)=\max_{\|\zeta\|_p\leq b}\left\{-2\zeta^T H_n-M_n\left(\zeta\right)\right\}\right\},\tag{2.36}$$

where the outer supremum is attained at some $\|\zeta_*\|_p\leq b$, occurs with probability at least $1-\varepsilon$, as long as $n\geq n_0$. In other words, $P(\mathcal{A}_n)\geq 1-\varepsilon$ when $n\geq n_0$.

We first consider the case $\rho>1$. For $\zeta\neq 0$, write $\bar{\zeta}=\zeta/\left\|\zeta\right\|_p$. Next, define the vector $V_i(\bar{\zeta})$ via $V_i\left(\bar{\zeta}\right)=Dh\left(W_i\right)^T\bar{\zeta}$ (that is, the j-th entry of $V_i\left(\bar{\zeta}\right)$ is the j-th entry of the vector $Dh\left(W_i\right)^T\bar{\zeta}$), and put

$$\Delta_i'=\Delta_i'\left(\bar{\zeta}\right)=\left|V_i\left(\bar{\zeta}\right)\right|^{p/q}sgn\left(V_i\left(\bar{\zeta}\right)\right).\tag{2.37}$$

Next, let $\bar{\Delta}_i = c_i \Delta_i'$ with c_i chosen so that

$$\left\|\bar{\Delta}_i\right\|_q = \left(\frac{1}{\rho}\left\|\zeta^T Dh\left(W_i\right)\right\|_p\right)^{1/(\rho-1)}.$$

In such case we have that

$$\begin{aligned}\max_{\Delta}\left\{\zeta^T Dh\left(W_i\right)\Delta - \|\Delta\|_p^\rho\right\} &= \max_{\|\Delta\|_q \geq 0}\left\{\left\|\zeta^T Dh\left(W_i\right)\right\|_p \|\Delta\|_q - \|\Delta\|_q^\rho\right\} \\ &= \zeta^T Dh\left(W_i\right)\bar{\Delta}_i - \left\|\bar{\Delta}_i\right\|_q^\rho \\ &= \left\|\zeta^T Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)}\left(\frac{1}{\rho}\right)^{1/(\rho-1)}\left(1-\frac{1}{\rho}\right). \quad (2.38)\end{aligned}$$

Pick $c_0 \in (0,\infty)$ and define $C_0 = \{\|W_i\|_p \leq c_0\}$. Note that

$$M_n(\zeta) \geq M_n'(\zeta, c_0),$$

where

$$M_n'(\zeta, c_0) = \frac{1}{n}\sum_{i=1}^n I\left(W_i \in C_0\right)\left\{\zeta^T \int_0^1 Dh\left(W_i + n_i^{-1/2}\bar{\Delta}_i u\right)\bar{\Delta}_i du - \left\|\bar{\Delta}_i\right\|_q^\rho\right\}^+.$$

Therefore

$$\max_{\|\zeta\|_p \leq b}\left\{-\zeta^T H_n - M_n(\zeta)\right\} \leq \max_{\|\zeta\|_p \leq b}\left\{-\zeta^T H_n - M_n'(\zeta, c_0)\right\}. \quad (2.39)$$

Define

$$\begin{aligned}\widehat{M}_n(\zeta, c_0) &= \frac{1}{n}\sum_{i=1}^n I\left(W_i \in C_0\right)\left\{\zeta^T Dh\left(W_i\right)\bar{\Delta}_i du - \left\|\bar{\Delta}_i\right\|_q^\rho\right\}^+ \\ &= \frac{1}{n}\sum_{i=1}^n I\left(W_i \in C_0\right)\left\|\zeta^T Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)}\left(\frac{1}{\rho}\right)^{1/(\rho-1)}\left(1-\frac{1}{\rho}\right),\end{aligned}$$

where the equality follows from (2.38). We then claim that

$$\sup_{\|\zeta\|_q \leq b} \left| \widehat{M}_n(\zeta, c_0) - M'_n(\zeta, c_0) \right| \to 0. \tag{2.40}$$

In order to verify (2.40), note, using the continuity of $Dh(\cdot)$, that for any $\varepsilon' > 0$ there exists n_0 such that if $n \geq n_0$ then (uniformly over $\|\zeta\|_p \leq b$),

$$\left| \int_0^1 I(W_i \in C_0) \left\| \zeta^T \left[Dh(W_i + n^{-1/2}\bar{\Delta}_i u) - Dh(W_i) \right] \right\|_p \left\| \bar{\Delta}_i \right\|_q du \right| \leq \varepsilon'.$$

Therefore, if $n \geq n_0$,

$$\frac{1}{n} \sum_{i=1}^n I(W_i \in C_0) \left| \zeta^T \int_0^1 \left[Dh(W_i + n^{-1/2}\bar{\Delta}_i u) - Dh(W_i) \right] \bar{\Delta}_i du \right| \leq \varepsilon'.$$

Since $\varepsilon' > 0$ is arbitrary, we conclude (2.40). Then, applying Lemma 2.3 we obtain

$$\widehat{M}_n(\zeta, c_0) \to \mathbb{E}\left(\zeta^T Dh(W_i) \bar{\Delta}_i du - \left\| \bar{\Delta}_i \right\|_q^\rho \right)^+$$

uniformly over $\|\zeta\|_p \leq b$ as $n \to \infty$, in probability. Therefore, applying the continuous mapping principle, we have that

$$\begin{aligned} &\max_{\|\zeta\|_p \leq b} \left\{ -\zeta^T H_n - M'_n(\zeta, c_0) \right\} \\ &\quad \Rightarrow \max_{\|\zeta\|_p \leq b} \left\{ -\zeta^T H - \kappa(\rho) \mathbb{E}\left[\left\| \zeta^T Dh(W) \right\|_p^{\rho/(\rho-1)} I\left(\|W\|_p \leq c_0 \right) \right] \right\}, \end{aligned} \tag{2.41}$$

as $n \to \infty$, where

$$\kappa(\rho) = \left(\frac{1}{\rho} \right)^{1/(\rho-1)} \left(1 - \frac{1}{\rho} \right),$$

and $H \sim \mathcal{N}(0, Cov[h(W, \theta_*)])$. From (2.39) and the construction of (2.36), we can

easily obtain that $n^{\rho/2}R_n(\theta_*)$ is stochastically bounded (asymptotically) by

$$\max_{\zeta}\left\{-\zeta^T H-\kappa(\rho)\,\mathbb{E}\left[\left\|\zeta^T Dh(W)\right\|_p^{\rho/(\rho-1)}\right]\right\},$$

which verifies the first part of the theorem when $\rho>1$.

Now, for $\rho=1$, we will follow very similar steps. Again, due to Lemma 2.2 we concentrate on the region $\|\zeta\|_p\le b$ for some $b>0$. For the upper bound, define Δ_i' as in (2.37). Using a localization technique similar to that described in the proof of Lemma 2.2 in which the set C_0 as introduced we might assume that $\|W_i\|_p\le c_0$ for some $c_0>0$. Then, for a given constant $c>0$, setting $\Delta_i=c\Delta_i'$, we obtain that

$$\begin{aligned}&\max_{\|\zeta\|_p\le b}\left\{-\zeta^T H_n-\frac{1}{n}\sum_{i=1}^n\sup_{\Delta_i}\left\{\zeta^T\int_0^1 Dh(W_i+\Delta_i u/n^{1/2})\Delta_i du-\|\Delta_i\|_q\right\}\right\}\\&\quad\le\max_{\|\zeta\|_p\le b}\left\{-\zeta^T H_n-\frac{1}{n}\sum_{i=1}^n\left(c\zeta^T\int_0^1 Dh(W_i+c\Delta_i' u/n^{1/2})\Delta_i' du-c\|\Delta_i'\|_q\right)I(W_i\in C_0)\right\}.\end{aligned}$$

As in the case $\rho>1$ we have that

$$\frac{1}{n}\sum_{i=1}^n I(W_i\in C_0)\int_0^1\zeta^T\left[Dh(W_i+c\Delta_i' u/n^{1/2})-Dh(W_i)\right]\Delta_i' du\to 0$$

in probability uniformly on ζ-compact sets. Similarly, in addition, for any $c>0$ and any $b>0$

$$\begin{aligned}&\max_{\|\zeta\|_p\le b}\left\{-\zeta^T H_n-\frac{1}{n}\sum_{i=1}^n\left(c\zeta^T Dh(W_i)\Delta_i' du-c\|\Delta_i'\|_q\right)I(W_i\in C_0)\right\}\\&\quad=\max_{\|\zeta\|\le b}\left\{-\zeta^T H_n-\frac{1}{n}\sum_{i=1}^n c\left(\left\|\zeta^T Dh(W)\right\|_p-1\right)^+\|\Delta_i'\|_q I(\|W_i\|_p\le c_0)\right\}\\&\quad\Rightarrow\max_{\|\zeta\|\le b}\left\{-\zeta^T H-c\mathbb{E}\left[\left(\left\|\zeta^T Dh(W)\right\|_p-1\right)^+\|\bar{\zeta}^T Dh(W)\|_p^{p/q}I(\|W\|_p\le c_0)\right]\right\},\end{aligned}$$

because $\|\Delta'\|_q^q = \|\bar{\zeta}^T Dh(W_i)\|_p^p$. Next, as the constant c can be arbitrarily large, we obtain a stochastic upper bound of the form

$$\max_{\|\zeta\|\leq b: P\left(\|\zeta^T Dh(W)\|_p \leq 1\right)=1} \left\{-\zeta^T H\right\} \leq \max_{\zeta: P\left(\|\zeta^T Dh(W)\|_p \leq 1\right)=1} \left\{-\zeta^T H\right\}.$$

This completes the proof of Theorem 2.1. □

Proof of Proposition 2.2. We follow the notation introduced in the proof of Theorem 2.1. Recall from Equation (2.27) and Equation (2.28) that

$$n^{1/2} R_n(\theta_*) = \sup_{\zeta} \left\{ \zeta^T H_n - \frac{1}{n} \sum_{k=1}^{n} \sup_{\Delta} \left\{ \int_0^1 \zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right) \Delta du - \|\Delta\|_q \right\} \right\}.$$

Let $A := \{\zeta : \operatorname{esssup} \left\|\zeta^T Dh(w)\right\|_p \leq 1\}$, where the essential supremum is taken with respect to the Lebesgue measure. Then, due to Hölder's inequality, if $\zeta \in A$,

$$\begin{aligned}
&\sup_{\Delta} \left\{ \int_0^1 \zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right) \Delta du - \|\Delta\|_q \right\} \\
&\quad \leq \sup_{\Delta} \left\{ \int_0^1 \left\|\zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right)\right\|_p \|\Delta\|_q du - \|\Delta\|_q \right\} \\
&\quad \leq \sup_{\Delta} \|\Delta\|_q \left\{ \int_0^1 \left(\left\|\zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right)\right\|_p - 1 \right) du \right\} \leq 0.
\end{aligned}$$

Consequently,

$$n^{1/2} R_n(\theta_*) \geq \sup_{\zeta \in A} \zeta^T H_n.$$

Letting $n \to \infty$ we conclude that

$$\sup_{\zeta \in A} \zeta^T H_n \Rightarrow \sup_{\zeta \in A} \zeta^T H.$$

Because W_i is assumed to have a density with respect to the Lebesgue measure it

follows that $P\left(\left\|\zeta^T Dh\left(W_i\right)\right\|_p \leq 1\right) = 1$ if and only if $\zeta \in A$ and the result follows. □

Finally, we provide the proof of Proposition 2.3.

Proof of Proposition 2.3. Recall from Equation (2.27) and Equation (2.28) that

$$n^{1/2} R_n\left(\theta_*\right) = \sup_{\zeta}\left\{\zeta^T H_n - \frac{1}{n}\sum_{k=1}^{n}\sup_{\Delta}\left\{\int_0^1 \zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right)\Delta du - \|\Delta\|_q^{\rho}\right\}\right\}. \tag{2.42}$$

As in the proof of Theorem 2.1, due to Lemma 2.2, we might assume that $\|\zeta\|_p \leq b$ for some $b > 0$.

The strategy will be to split the inner supremum in values of $\|\Delta\|_q \leq \delta n^{1/2}$ and values $\|\Delta\|_q > \delta n^{1/2}$ for a suitably small positive constant δ. In Step 1, we shall show that the supremum is achieved with high probability in the former region. Then, in Step 2, we analyze the region in which $\|\Delta\|_q \leq \delta n^{1/2}$ and argue that the integrals inside the summation in Equation (2.42) can be replaced by $\zeta^T Dh\left(W_i\right)\Delta$. Once this substitution is performed we can solve the inner maximization problem explicitly in Step 3 and, finally, we will apply a weak convergence result on ζ-compact sets to conclude the result. We now proceed to execute this strategy.

Execution of Step 1: Pick $\delta > 0$ small, to be chosen in the sequel, then note that A5) implies (by redefining κ if needed, due to the continuity of $Dh\left(\cdot\right)$) that

$$\left\|Dh\left(w\right)\right\|_p \leq \kappa\left(1 + \|w\|_q^{\rho-1}\right).$$

Therefore, for ζ such that $\|\zeta\|_p \leq b$,

$$\sup_{\|\Delta\|_q \geq \delta n^{1/2}} \left\{ \int_0^1 \left| \zeta^T Dh \left(W_i + \Delta u / n^{1/2} \right) \Delta \right| du - \|\Delta\|_q^\rho \right\}$$
$$\leq \sup_{\|\Delta\|_q \geq \delta n^{1/2}} \left\{ b\kappa \left(1 + \int_0^1 \left\| W_i + \Delta u / n^{1/2} \right\|_q^{\rho-1} du \right) \|\Delta\|_q - \|\Delta\|_q^\rho \right\}.$$

Note that if $\rho \in (1,2)$, then $0 < \rho - 1 < 1$, and therefore by the triangle inequality and concavity

$$\left\| W_i + \Delta u / n^{1/2} \right\|_q^{\rho-1} \leq \left(\|W_i\|_q + \left\| \Delta / n^{1/2} \right\|_q \right)^{\rho-1} \leq \|W_i\|_q^{\rho-1} + \left\| \Delta / n^{1/2} \right\|_q^{\rho-1}.$$

On the other hand, if $\rho \geq 2$, then $\rho - 1 \geq 1$ and the triangle inequality combined with Jensen's inequality applied as follows:

$$\|a + c\|^{\rho-1} \leq 2^{\rho-1} \left(\frac{1}{2} \|a\|^{\rho-1} + \frac{1}{2} \|c\|^{\rho-1} \right) = 2^{\rho-2} \left(\|a\|^{\rho-1} + \|c\|^{\rho-1} \right),$$

yields

$$\left\| W_i + \Delta u / n^{1/2} \right\|_q^{\rho-1} \leq 2^{\rho-2} \left(\|W_i\|_q^{\rho-1} + \left\| \Delta / n^{1/2} \right\|_q^{\rho-1} \right).$$

So, in both cases we can write

$$\sup_{\|\Delta\|_q \geq \delta n^{1/2}} \left\{ \int_0^1 \left| \zeta^T Dh(W_i + \Delta u / n^{1/2}) \Delta \right| du - \|\Delta\|_q^\rho \right\}$$
$$\leq \sup_{\|\Delta\|_q \geq \delta n^{1/2}} \left\{ b\kappa \left(1 + 2^{\rho-1} \left(\|W_i\|_q^{\rho-1} + \left\| \Delta / n^{1/2} \right\|_q^{\rho-1} \right) \right) \|\Delta\|_q - \|\Delta\|_q^\rho \right\}$$
$$\leq \sup_{\|\Delta\|_q \geq \delta n^{1/2}} \left\{ b\kappa \left(\|\Delta\|_q + 2^{\rho-1} \|W_i\|_q^{\rho-1} \|\Delta\|_q + 2^{\rho-1} \|\Delta\|_q^\rho / n^{(\rho-1)/2} \right) - \|\Delta\|_q^\rho \right\}.$$

Next, as $E\|W_n\|^\rho < \infty$, we have that for any $\varepsilon' > 0$,

$$P\left(\|W_n\|_q^\rho \geq \varepsilon' n \text{ i.o.}\right) = 0,$$

therefore we might assume that there exists n_0 such that for all $i \leq n$ and $n \geq n_0$, $\|W_i\|_q^{\rho-1} \leq (\varepsilon' n)^{(\rho-1)/\rho}$. Therefore, if $(\varepsilon')^{(\rho-1)/\rho} \leq \delta^{\rho-1}/\left(b\kappa 2^\rho\right)$, we conclude that if $\|\Delta\|_q \geq \delta n^{1/2}$ and $n > n_0$,

$$\begin{aligned} b\kappa 2^{\rho-1}\|W_i\|_q^{\rho-1}\|\Delta\|_q &\leq b\kappa 2^{\rho-1}(\varepsilon' n)^{(\rho-1)/\rho}\|\Delta\|_q \\ &\leq \frac{1}{2}\delta^{\rho-1}n^{(\rho-1)/2}\|\Delta\|_q \leq \frac{1}{2}\|\Delta\|_q^\rho. \end{aligned}$$

Similarly, choosing n sufficiently large we can guarantee that

$$b\kappa\left(\|\Delta\|_q + 2^{\rho-1}\|\Delta\|_q^\rho / n^{(\rho-1)/2}\right) \leq \frac{1}{2}\|\Delta\|_q^\rho.$$

Therefore, we conclude that for any fixed $\delta > 0$,

$$\sup_{\|\Delta\|_q \geq \delta\sqrt{n}} \left\{ \int_0^1 \left|\zeta^T Dh(W_i + \Delta u/n^{1/2})\Delta\right| du - \|\Delta\|_q^\rho \right\} \leq 0 \tag{2.43}$$

provided n is large enough, thus achieving the desired result over the region $\|\Delta\|_q \geq \delta\sqrt{n}$.

Execution of Step 2: Next, we let $\varepsilon'' > 0$, and note that

$$\begin{aligned}
&\sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T Dh(W_i + \Delta u/n^{1/2}) \Delta du - \|\Delta\|_q^\rho \right\} \qquad (2.44) \\
&\leq \sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T \left[Dh(W_i + \Delta u/n^{1/2}) - Dh(W_i) \right] \Delta du - \varepsilon'' \|\Delta\|_q^\rho \right\} \\
&\qquad + \sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \left\{ \zeta^T Dh\left(W_i\right) \Delta - (1 - \varepsilon'') \|\Delta\|_q^\rho \right\}.
\end{aligned}$$

We now argue locally, using A6), a bound for the first term in the right hand side of Equation (2.44):

$$\begin{aligned}
&\sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T \left[Dh(W_i + \Delta u/n^{1/2}) - Dh(W_i) \right] \Delta du - \varepsilon'' \|\Delta\|_q^\rho \right\} \qquad (2.45) \\
&\qquad \leq \sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \left\{ \|\zeta\|_p \bar{\kappa}\left(W_i\right) \|\Delta\|_q^2 / n^{1/2} - \varepsilon'' \|\Delta\|_q^\rho \right\} \\
&\qquad \leq \sup_{\|\bar{\Delta}\|_q \leq 1} \left\{ b\bar{\kappa}\left(W_i\right) \left\|\bar{\Delta}\right\|_q^2 \delta^2 n^{1/2} - \varepsilon'' \left\|\bar{\Delta}\right\|_q^\rho \left(\delta n^{1/2}\right)^\rho \right\}.
\end{aligned}$$

As $\sup_{x \in [0,1]} \{a_n x^2 - b_n x^\rho\} \leq (\rho - 2)^+ (a_n^\rho / b_n^2)^{1/(\rho-2)} / \rho$ when $b_n > a_n$, we have, for all n sufficiently large, that

$$\begin{aligned}
&\sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T \left[Dh(W_i + \Delta u/n^{1/2}) - Dh(W_i) \right] \Delta du - \varepsilon'' \|\Delta\|_q^\rho \right\} \\
&\leq \frac{(\rho - 2)^+}{\rho} \left(\frac{b\bar{\kappa}(W_i)}{\varepsilon'' \sqrt{n}} \right)^{\rho/(\rho-2)}.
\end{aligned}$$

Since $E[\bar{\kappa}(W)^2] < \infty$ (from Assumption A6)), we have that $P(\bar{\kappa}(W_i) > \varepsilon''' \sqrt{i} \text{ i.o.}) = 0$

for any $\varepsilon''' > 0$. Consecutively, $\bar{\kappa}(W_i) < \varepsilon'''\sqrt{i}$ for all i large enough, and therefore,

$$\begin{aligned}
\overline{\lim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \sup_{\|\Delta\|_q \le \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T \left[Dh\left(W_i + \Delta u/n^{1/2}\right) - Dh\left(W_i\right)\right] \Delta du - \varepsilon'' \|\Delta\|_q^\rho \right\} \\
\le \frac{(\rho-2)^+}{\rho} \overline{\lim}_{n\to\infty} \left(\frac{b}{\varepsilon''}\right)^{\rho/(\rho-2)} \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\bar{\kappa}(W_i)}{\sqrt{n}}\right)^{\rho/(\rho-2)} \\
\le \frac{(\rho-2)^+}{\rho} \left(b\frac{\varepsilon'''}{\varepsilon''}\right)^{\rho/(\rho-2)},
\end{aligned}$$

which can be made arbitrarily small by choosing ε''' arbitrarily small. Therefore, for any fixed $\varepsilon'', \delta > 0$,

$$\overline{\lim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \sup_{\|\Delta\|_q \le \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T \left[Dh\left(W_i + \Delta u/n^{1/2}\right) - Dh\left(W_i\right)\right] \Delta du - \varepsilon'' \|\Delta\|_q^\rho \right\} = 0. \tag{2.46}$$

Execution of Step 3: Next, it follows from Equation (2.43), Equation (2.44) and Equation (2.46) that for any fixed $\varepsilon'', \delta > 0$, there exists N_0 such that if $n \ge N_0$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \sup_{\Delta} \left\{ \int_0^1 \zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right) \Delta du - \|\Delta\|_q^\rho \right\} \\
\le \frac{1}{n}\sum_{i=1}^{n} \sup_{\Delta \le \delta\sqrt{n}} \left\{ \zeta^T Dh\left(W_i\right) \Delta du - (1-\varepsilon'') \|\Delta\|_q^\rho \right\} + \delta \\
\le \frac{1}{n}\sum_{i=1}^{n} \min\left\{ \kappa\left(\rho, \varepsilon''\right) \left\| \zeta^T Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)},\ c_n \right\} + \delta,
\end{aligned}$$

where

$$\kappa\left(\rho, \varepsilon''\right) = \left(\frac{1}{\rho(1-\varepsilon'')}\right)^{1/(\rho-1)} \left(1 - \frac{1}{\rho}\right),$$

and $c_n \to \infty$ as $n \to \infty$ (the exact value of c_n is not important).

Next, note that A5) implies that

$$\|Dh(W_i)\|_p^{\rho/(\rho-1)} I(\|W_i\| \geq 1) \leq \kappa I(\|W_i\| \geq 1) \|W_i\|_q^\rho \leq \kappa \|W_i\|_q^\rho$$

and, therefore, since $Dh(\cdot)$ is continuous (therefore locally bounded) and $E\|W_i\|_q^\rho < \infty$ also by A5), we have that

$$E\|Dh(W)\|_p^{\rho/(\rho-1)} < \infty.$$

Then, an argument similar to Lemma 2.3 shows that

$$\sup_{\|\zeta\|_p \leq b} \left\{ \zeta^T H_n - \frac{1}{n} \sum_{i=1}^n \left\{ \kappa(\rho, \varepsilon'') \left\|\zeta^T Dh(W_i)\right\|_q^{\rho/(\rho-1)}, c_n \right\} \right\}$$
$$\Rightarrow \sup_{\|\zeta\|_p \leq b} \left\{ \zeta^T H - \kappa(\rho, \varepsilon'') E\left\|\zeta^T Dh(W_i)\right\|_q^{\rho/(\rho-1)} \right\},$$

as $n \to \infty$ (where $\Rightarrow$ denotes weak convergence). Finally, we can send $\varepsilon'', \delta \to 0$ and $b \to \infty$ to obtain the desired asymptotic stochastic lower bound. □

2.A.2. Proofs of the distributionally robust representations in Section 2.3.2

Here, we provide proofs for results in Section 2.3.2 that recover various norm regularized regressions as a special cases of distributionally robust regression (Proposition 2.5, Theorems 2.2 and 2.3).

Proof of Proposition 2.5. We utilize the duality result in Proposition 2.4 to prove Proposition 2.5. For brevity, let $\bar{X}_i = (X_i, Y_i)$ and $\bar{\beta} = (-\beta, 1)$. Then the loss function becomes $l(X_i, Y_i; \beta) = (\bar{\beta}^T \bar{X}_i)^2$. We first decipher the function $\phi_\gamma(X_i, Y_i; \beta)$ defined

in Proposition 2.4:

$$\phi_\gamma(X_i, Y_i; \beta) = \sup_{\bar{u} \in \mathbb{R}^{d+1}} \left\{ (\bar{\beta}^T \bar{u})^2 - \gamma \|\bar{X}_i - \bar{u}\|_q^2 . \right\}$$

To proceed further, we change the variable to $\Delta = \bar{u} - \bar{X}_i$, and apply Hölder's inequality to see that $|\bar{\beta}^T \Delta| \leq \|\bar{\beta}\|_p \|\Delta\|_q$, where the equality holds for some $\Delta \in \mathbb{R}^{d+1}$. Therefore,

$$\begin{aligned}
\phi_\gamma(\bar{X}_i; \beta) &= \sup_{\Delta \in \mathbb{R}^{d+1}} \left\{ \left(\bar{\beta}^T \bar{X}_i + \bar{\beta}^T \Delta \right)^2 - \gamma \left\| \Delta \right\|_q^2 \right\} \\
&= \sup_{\Delta \in \mathbb{R}^{d+1}} \left\{ \left(\bar{\beta}^T \bar{X}_i + \operatorname{sign}\left(\bar{\beta}^T \bar{X}_i \right) \left| \bar{\beta}^T \Delta \right| \right)^2 - \gamma \left\| \Delta \right\|_q^2 \right\} \\
&= \sup_{\Delta \in \mathbb{R}^{d+1}} \left\{ \left(\bar{\beta}^T \bar{X}_i + \operatorname{sign}\left(\bar{\beta}^T \bar{X}_i \right) \left\| \Delta \right\|_q \left\| \bar{\beta} \right\|_p \right)^2 - \gamma \left\| \Delta \right\|_q^2 \right\}.
\end{aligned}$$

On expanding the squares, the above expression simplifies as below:

$$\begin{aligned}
\phi_\gamma(\bar{X}_i; \beta) &= \left(\bar{\beta}^T \bar{X}_i \right)^2 + \sup_{\Delta \in \mathbb{R}^{d+1}} \left\{ - \left(\gamma - \left\| \bar{\beta} \right\|_p^2 \right) \left\| \Delta \right\|_q^2 + 2 \left| \bar{\beta}^T \bar{X}_i \right| \left\| \bar{\beta} \right\|_p \left\| \Delta \right\|_q \right\} \\
&= \begin{cases} \left(\bar{\beta}^T \bar{X}_i \right)^2 \gamma / (\gamma - \left\| \bar{\beta} \right\|_p^2) & \text{if } \gamma > \left\| \bar{\beta} \right\|_p^2, \\ +\infty & \text{if } \gamma \leq \left\| \bar{\beta} \right\|_p^2. \end{cases}
\end{aligned} \tag{2.47}$$

With this expression for $\phi_\gamma(X_i, Y_i; \beta)$, we next investigate the right hand side of the duality relation in Proposition 2.4. As $\phi_\gamma(x, y; \beta) = \infty$ when $\gamma \leq \|\beta\|_p^2$, we obtain from the dual formulation in Proposition 2.4 that

$$\begin{aligned}
\sup_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P \left[l(X, Y; \beta) \right] &= \inf_{\gamma \geq 0} \left\{ \gamma \delta + \frac{1}{n} \sum_{i=1}^n \phi_\gamma(X_i, Y_i; \beta) \right\} \\
&= \inf_{\gamma > \|\beta\|_p^2} \left\{ \gamma \delta + \frac{\gamma}{\gamma - \|\bar{\beta}\|_p^2} \frac{1}{n} \sum_{i=1}^n (\bar{\beta}^T \bar{X}_i)^2 \right\}.
\end{aligned} \tag{2.48}$$

Now, see that $\sum_{i=1}^{n}(\bar{\beta}^T\bar{X}_i)^2/n$ is nothing but the mean square error $MS\mathbb{E}_n(\beta)$. Next, as the right hand side of (2.48) is a convex function growing to ∞ (when $\gamma \to \infty$ or $\gamma \to \|\bar{\beta}\|_p^2$), its global minimizer can be characterized uniquely via first order optimality condition. This, in turn, renders the right hand side of (2.48) as

$$\sup_{P:D_c(P,P_n)\leq\delta} \mathbb{E}_P\left[l(X,Y;\beta)\right] = \left(\sqrt{MS\mathbb{E}_n(\beta)} + \sqrt{\delta}\|\bar{\beta}\|_p\right)^2.$$

This completes the proof of Proposition 2.5. □

Outline of a proof of Theorem 2.2. The proof of Theorem 2.2 is essentially the same as the proof of Proposition 2.5, except for adjusting for ∞ in the definition of cost function $N_q((x,y),(u,v))$ when $y \neq v$ (as in the derivation leading to $\phi_\gamma(X_i,Y_i;\beta)$ defined in (2.18)). First, see that

$$\phi_\gamma(X_i,Y_i;\beta) = \sup_{x'\in\mathbb{R}^d, y'\in\mathbb{R}} \big\{(y'^T x'^2 - \gamma N_q\big((x',y'),(X_i,Y_i)\big)\big\}.$$

As $N_q((x',y'),(X_i,Y_i)) = \infty$ when $y' \neq Y_i$, the supremum in the above expression is effectively over only (x',y') such that $y' = Y_i$. As a result, we obtain,

$$\begin{aligned}\phi_\gamma(X_i,Y_i;\beta) &= \sup_{x'\in\mathbb{R}^d} \big\{(Y_i - \beta^T x'^2 - \gamma N_q\big((x',Y_i),(X_i,Y_i)\big)\big\}. \\ &= \sup_{x'\in\mathbb{R}^d} \big\{(Y_i - \beta^T x'^2 - \gamma\|x'-X_i\|_q^2\big)\big\}.\end{aligned}$$

Now, following same lines of reasoning as in the proof of Theorem 2.5 and the deriva-

tion leading to (2.47), we obtain

$$\phi_\gamma(x,y;\beta) = \begin{cases} \frac{\gamma}{\gamma - \|\beta\|_p^2} \quad (Y_i - \beta^T X_i)^2 & \text{when } \lambda > \|\beta\|_p^2, \\ +\infty & \text{otherwise.} \end{cases}$$

The rest of the proof is same as in the proof of Proposition 2.5.

Proof of Theorem 2.3. As in the proof of Proposition 2.5, we apply the duality formulation in Proposition 2.4 to write the worst case expected log-exponential loss function as:

$$\begin{aligned} &\sup_{P:\ D_c(P,P_n)\le\delta} \mathbb{E}_P\big[l(X,Y;\beta)\big] \\ &= \inf_{\lambda\ge 0}\left\{\delta\lambda + \frac{1}{n}\sum_{i=1}^n \sup_x \left\{\log\left(1+\exp(-Y_i\beta^T x)\right) - \lambda\left\|x - X_i\right\|_p\right\}\right\}. \end{aligned}$$

For each (X_i, Y_i), following Lemma 1 in Shafieezadeh-Abadeh *et al.* [2015], we obtain

$$\begin{aligned} &\sup_x \left\{\log\left(1+\exp(-Y_i\beta^T x)\right) - \lambda\left\|x - X_i\right\|_p\right\} \\ &= \begin{cases} \log\left(1+\exp(-Y_i\beta^T X_i)\right) & \text{if } \|\beta\|_q \le \lambda, \\ +\infty & \text{if } \|\beta\|_q > \lambda. \end{cases} \end{aligned}$$

Then we can write the worst case expected loss function as,

$$\begin{aligned}
&\inf_{\lambda\geq 0}\left\{\delta\lambda+\frac{1}{n}\sum_{i=1}^{n}\sup_{x}\left\{\log\left(1+\exp(-Y_i\beta^T x)\right)-\lambda\left\|x-X_i\right\|_p\right\}\right\}\\
&\quad=\inf_{\lambda\geq 0}\left\{\delta\lambda+\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(1+\exp(-Y_i\beta^T X_i)\right)1_{\{\lambda>\|\beta\|_q\}}+\infty 1_{\{\lambda\leq\|\beta\|_q\}}\right)\right\}\\
&\quad=\inf_{\lambda>\|\beta\|_q}\left\{\delta\lambda+\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp(-Y_i\beta^T X_i)\right)\right\}\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp(-Y_i\beta^T X_i)\right)+\delta\left\|\beta\right\|_q,
\end{aligned}$$

which is equivalent to regularized logistic regression in the theorem statement.

Next we move to SVM with Hinge loss function, let us apply the duality formulation in Proposition 2.4 to write the worst case expected Hinge loss function as:

$$\sup_{P:\ D_c(P,P_n)\leq\delta}\mathbb{E}_P\left[\left(1+Y\beta^T X\right)^+\right]=\inf_{\lambda\geq 0}\left\{\delta\lambda+\frac{1}{n}\sum_{i=1}^{n}\sup_{x}\left\{\left(1-Y_i\beta^T x\right)^+-\lambda\left\|x-X_i\right\|_p\right\}\right\}.$$

For each i, let us consider the the maximization problem and for simplicity we denote $\Delta u_i = x - X_i$

$$\begin{aligned}
&\sup_{\Delta u_i}\left\{\left(1-Y_i\beta^T\left(X_i+\Delta u_i\right)\right)^+-\lambda\left\|\Delta u_i\right\|_p\right\}\\
&=\sup_{\Delta u_i}\sup_{0\leq\alpha_i\leq 1}\left\{\alpha_i\left(1-Y_i\beta^T\left(X_i+\Delta u_i\right)\right)-\lambda\left\|\Delta u_i\right\|_p\right\}\\
&=\sup_{0\leq\alpha_i\leq 1}\sup_{\Delta u_i}\left\{\alpha_i Y_i\beta^T\Delta u_i-\lambda\left\|\Delta u_i\right\|_p+\alpha_i\left(1-Y_i\beta^T X_i\right)\right\}\\
&=\sup_{0\leq\alpha_i\leq 1}\sup_{\Delta u_i}\left\{\alpha_i\left\|\beta\right\|_q\left\|\Delta u_i\right\|_p-\lambda\left\|\Delta u_i\right\|_p+\alpha_i\left(1-Y_i\beta^T X_i\right)\right\}\\
&=\begin{cases}\left(1-Y_i\beta^T X_i\right)^+ & \text{if } \left\|\beta\right\|_q\leq\lambda\\ +\infty & \text{if } \left\|\beta\right\|_q>\lambda\end{cases}
\end{aligned}$$

The first equality is due to $x^+ = \sup_{0\le\alpha\le 1} x$; second equality is because the function is concave in Δu_i linear in α and α is in a compact set, we can apply minimax theorem to switch the order of maximals; third equality is due to applying Holder inequality to the first term and since the second term only depends on the norm of Δu_i we can argue the equality also holds for this maximization problem. We notice the objective function is a minimization problem thus we will require $\lambda \ge \|\beta\|_q$. Then we have

$$\inf_{\lambda\ge\|\beta\|_q}\left\{\delta\lambda + \frac{1}{n}\sum_{i=1}^{n}\left(1 - Y_i\beta^T X_i\right)^+\right\} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - Y_i\beta^T X_i\right)^+ + \delta\|\beta\|_q .$$

□

2.A.3. Proofs of RWP function limit theorems for linear and logistic regression examples

We first obtain the dual formulation of the respective RWP functions for linear and logistic regressions using Proposition 2.1. Let $E[h(x,y;\beta)] = \mathbf{0}$ be the estimating equation under consideration ($h(x,y;\beta) = (y-\beta^T x)x$ for linear regression and $h(x,y;\beta)$ as in Equation (2.25) for logistic regression). Recall that the cost function is $c(\cdot) = N_q(\cdot)$. Due to the duality result in Proposition 2.1, we obtain

$$\begin{aligned} R_n(\beta_*) &= \inf\left\{D_c(P,P_n):\ \mathbb{E}_P[h(X,Y;\beta_*)] = \mathbf{0}\right\} \\ &= \sup_{\lambda}\left\{-\frac{1}{n}\sum_{i=1}^{n}\sup_{(x',y')}\left\{\lambda^T h(x',y';\beta_*) - N_q\big((x',y'),(X_i,Y_i)\big)\right\}\right\}. \end{aligned}$$

As $N_q((x',y'),(X_i,Y_i)) = \infty$ when $y' \ne Y_i$, the above expression simplifies to,

$$R_n(\beta_*) = \sup_{\lambda}\left\{-\frac{1}{n}\sum_{i=1}^{n}\sup_{x'}\left\{\lambda^T h(x',Y_i;\beta_*) - \|x'-X_i\|_q^\rho\right\}\right\}, \tag{2.49}$$

where $\rho = 2$ for the case of linear regression (Theorem 2.4) and $\rho = 1$ for the case of logistic regression (Theorem 2.5). As RWP function here is similar to the RWP function for general estimating equation in Section 2.2.4, a similar limit theorem holds. We state here the assumptions for proving RWP limit theorems for the dual formulation in Equation (2.49).

Assumptions:

A2') Suppose that $\beta_* \in \mathbb{R}^d$ satisfies $E[h(X, Y; \beta_*)] = \mathbf{0}$ and $E\|h(X, Y; \beta_*)\|_2^2 < \infty$ (While we do not assume that β_* is unique, the results are stated for a fixed β_* satisfying $E[h(X, Y; \beta_*)] = \mathbf{0}$.)

A4') Suppose that for each $\xi \neq \mathbf{0}$, the partial derivative $D_x h(x, y; \beta_*)$ satisfies,

$$P\left(\left\|\xi^T D_x h(X, Y; \beta_*)\right\|_p > 0\right) > 0.$$

A6') Assume that there exists $\bar{\kappa} : \mathbb{R}^m \to \infty$ such that

$$\|D_x h(x + \Delta, y; \beta_*) - D_x h(x, y; \beta_*)\|_p \leq \bar{\kappa}(x, y)\|\Delta\|_q,$$

for all $\Delta \in \mathbb{R}^d$, and $\mathbb{E}[\bar{\kappa}(X, Y)^2] < \infty$.

Lemma 2.4. If $\rho \geq 2$, under Assumptions A2'), A4') and A6'), we have,

$$nR_n(\beta_*; \rho) \Rightarrow \bar{R}(\rho),$$

where

$$\bar{R}(\rho) = \sup_{\xi \in \mathbb{R}^d} \left\{ \rho \xi^T H - (\rho - 1)\mathbb{E}\left\|\xi^T D_x h(X, Y; \beta_*)\right\|_p^{\rho/(\rho-1)} \right\},$$

with $H \sim \mathcal{N}(\mathbf{0}, \mathrm{Cov}[h(X, Y; \beta_*)]$ and $1/p + 1/q = 1$.

Lemma 2.5. If $\rho = 1$, in addition to assuming A2'), A4'), suppose that $D_x h(\cdot, y; \beta_*)$ is continuous for every y in the support of probability distribution of Y. Also suppose that X has a positive probability density (almost everywhere) with respect to the Lebesgue measure. Then,

$$nR_n(\beta_*; 1) \Rightarrow \bar{R}(1),$$

where

$$\bar{R}(1) = \sup_{\xi: P(\|\xi^T D_x h(X,Y;\beta_*)\|_p > 1) = 0} \{\xi^T H\},$$

with $H \sim \mathcal{N}(\mathbf{0}, \mathrm{Cov}[h(X, Y; \beta_*])$.

The proof of Lemma 2.4 and 2.5 follows closely the proof of our results in Section 2.2 and therefore it is omitted. We prove Theorem 2.4 and 2.5 as a quick application of these lemmas.

Proof of Theorem 2.4. To show that the RWP function dual formulation in Equation (2.49) converges in distribution, we verify the assumptions of Lemma 2.4 with $h(x, y; \beta) = (y - \beta^T x)x$. Under the null hypothesis H_0, $Y - \beta_*^T X = e$ is independent of X, has zero mean and finite variance σ^2. Therefore,

$$\mathbb{E}\left[h(X, Y; \beta)\right] = \mathbb{E}\left[eX\right] = 0, \text{ and}$$
$$\mathbb{E}\|h(X, Y; \beta)\|_2^2 = \mathbb{E}\left[e^2 X^T X\right] = \sigma^2 \mathbb{E}\|X\|_2^2,$$

which is finite, because trace of the covariance matrix Σ is finite. This verifies As-

sumption A2'). Further,

$$D_x h(X, Y; \beta_*) = \big(y - \beta_*^T X\big) I_d - X\beta_*^T = eI_d - X\beta_*^T,$$

where I_d is the $d \times d$ identity matrix. For any $\xi \neq \mathbf{0}$,

$$P\left(\|\xi^T D_x h(X, Y; \beta_*)\|_p = 0\right) = P\left(e\xi = (\xi^T X)\beta\right) = 0,$$

thus satisfying Assumption A4') trivially. In addition,

$$\|D_x h(x + \Delta, y; \beta_*) - D_x h(x, y; \beta_*)\|_p = \left\|\beta_*^T \Delta I_d - \Delta \beta_*^T\right\|_p \leq c\|\Delta\|_q,$$

for some positive constant c. This verifies Assumption A6'). As all the assumptions imposed in Lemma 2.4 are easily satisfied, using $\rho = 2$, we obtain the following convergence in distribution as a consequence of Lemma 2.4.

$$R_n(\beta_*) \Rightarrow \sup_{\xi \in \mathbb{R}^d} \left\{ 2\xi^T H - \mathbb{E}\left\|e\xi - (\xi^T X)\beta_*\right\|_p^2 \right\},$$

as $n \to \infty$. Here, $H \sim \mathcal{N}(\mathbf{0}, \mathrm{Cov}[h(X, Y; \beta_*)]$. As $\mathrm{Cov}[h(X, Y; \beta_*)] = E\left[e^2 X X^T\right] = \sigma^2\Sigma$, if we let $Z = H/\sigma$, we obtain the limit law,

$$L_1 = \sup_{\xi \in \mathbb{R}^d} \left\{ 2\sigma\xi^T Z - \mathbb{E}\left\|e\xi - (\xi^T X)\beta_*\right\|_p^2 \right\},$$

where $Z = \mathcal{N}(\mathbf{0}, \Sigma)$, as in the statement of the theorem.

Proof of the stochastic upper bound in Theorem 2.4: For the stochastic upper bound, let us consider the asymptotic distribution L_1 and rewrite the maximization problem

as,

$$\begin{aligned} L_1 &= \sup_{\|\xi\|_p=1} \sup_{\alpha\geq 0} \left\{ 2\sigma\alpha\xi^T Z - \alpha^2 \mathbb{E} \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 \right\} \\ &\leq \sup_{\|\xi\|_p=1} \sup_{\alpha\geq 0} \left\{ 2\sigma\alpha \left\| Z \right\|_q - \alpha^2 \mathbb{E} \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 \right\}, \end{aligned}$$

because of Hölder's inequality. By solving the inner optimization problem in α, we obtain

$$L_1 \leq \sup_{\|\xi\|_p=1} \frac{\sigma^2 \left\| Z \right\|_q^2}{\mathbb{E} \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2} = \frac{\sigma^2 \left\| Z \right\|_q^2}{\inf_{\|\xi\|_p=1} \mathbb{E} \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2}. \tag{2.50}$$

Next, consider the minimization problem in the denominator: Due to triangle inequality,

$$\begin{aligned} &\inf_{\|\xi\|_p=1} \mathbb{E} \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 \\ &\geq \inf_{\|\xi\|_p=1} \mathbb{E} \left(|e| \left\| \xi \right\|_p - \left| \xi^T X \right| \left\| \beta_* \right\|_p \right)^2 \\ &= \mathbb{E} \left| e \right|^2 + \inf_{\|\xi\|_p=1} \left\{ \left\| \beta_* \right\|_p^2 \mathbb{E} \left| \xi^T X \right|^2 - 2 \left\| \beta_* \right\|_p \mathbb{E} \left| e \right| \mathbb{E} \left| \xi^T X \right| \right\} \\ &\geq \mathbb{E} \left| e \right|^2 + \inf_{\|\xi\|_p=1} \left\{ \left\| \beta_* \right\|_p^2 \left(\mathbb{E} \left| \xi^T X \right| \right)^2 - 2 \left\| \beta_* \right\|_p \mathbb{E} \left| e \right| \mathbb{E} \left| \xi^T X \right| \right\} \\ &= \mathbb{E} \left| e \right|^2 - \left(\mathbb{E} \left| e \right| \right)^2 + \inf_{\|\xi\|_p=1} \left(\left\| \beta_* \right\|_p \mathbb{E} \left| \xi^T X \right| - \mathbb{E} \left| e \right| \right)^2 \\ &\geq \mathbb{E} \left| e \right|^2 - \left(\mathbb{E} \left| e \right| \right)^2 = \mathrm{Var} \left[|e| \right]. \end{aligned}$$

Combining the above inequality with (2.50), we obtain,

$$\sup_{\xi\in\mathbb{R}^d} \left\{ \sigma^2 \xi^T Z - \mathbb{E} \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 \right\} \leq \frac{\sigma^2 \left\| Z \right\|_q^2}{\mathrm{Var} \left| e \right|}.$$

Consequently,

$$nR_n(\beta_*) \xrightarrow{D} L_1 := \max_{\xi\in\mathbb{R}^d}\left\{\sigma\xi^T Z - \mathbb{E}\left\|e\xi - (\xi^T X)\beta_*\right\|_p^2\right\} \overset{D}{\leq} \frac{\mathbb{E}[e^2]}{\mathbb{E}[e^2] - (\mathbb{E}\,|e|)^2}\|Z\|_q^2.$$

If random error e is normally distributed, then

$$nR_n(\beta_*) \lesssim_D \frac{\pi}{\pi - 2}\|Z\|_q^2,$$

thus establishing the desired upper bound. □

Proof of Theorem 2.5. Under null hypothesis H_0, the training samples $\mathcal{D}_n = \{X_i, Y_i\}_{i=1}^n$ are produced from the logistic regression model with parameter β_*. As β_* minimizes the expected log-exponential loss $l(x, y; \beta)$, the corresponding optimality condition is $E[h(X, Y; \beta_*)] = \mathbf{0}$, where

$$h(x, y; \beta_*) = \frac{-yx}{1 + \exp(y\beta_* x)}.$$

As $\mathbb{E}\|h(X, Y; \beta_*)\|_2^2 \leq \mathbb{E}\|X\|_2^2$ is finite, Assumption A2') is satisfied. Let I_d denote $d \times d$ identity matrix. While

$$D_x h(x, y; \beta_*) = \frac{-yI_d}{1 + \exp(y\beta_*^T x)} + \frac{x\beta_*^T}{(1 + \exp(y\beta_*^T x))(1 + \exp(-y\beta_*^T x))}$$

is continuous (as a function of x) for every y, it is also true that

$$P\left(\left\|\xi^T D_x h(X, Y; \beta_*)\right\|_p = 0\right) = P\left(Y\big(1 + \exp(-Y\beta_*^T X)\big)\xi = (\xi^T X)\beta\right) = 0,$$

for any $\xi \neq \mathbf{0}$, thus satisfying Assumption A4'). As all the conditions required for

the convergence in distribution in Lemma 2.5 are satisfied, we obtain,

$$\sqrt{n}R_n(\beta_*) \Rightarrow \sup_{\xi \in A} \xi^T Z,$$

where $Z \sim \mathcal{N}(\mathbf{0}, \mathbb{E}[XX^T/(1+\exp(Y\beta_*^T X))^2])$ as a consequence of Lemma 2.5. Here, the set $A = \{\xi \in \mathbb{R}^d : \text{ess sup}\|\xi^T D_x h(X, Y; \beta_*)\| \leq 1\}$.

Proof of the stochastic upper bound in Theorem 2.5: First, we claim that A is a subset of the norm ball $\{\xi \in \mathbb{R}^d : \|\xi\|_p \leq 1\}$. To establish this, we observe that,

$$\begin{aligned}
&\left\|\xi^T D_x h(X, Y; \beta_*)\right\|_p \qquad (2.51)\\
&\geq \left\|\frac{-Y\xi}{1+\exp(Y\beta_*^T X)}\right\|_p - \left\|\frac{(\xi^T X)\beta_*}{\big(1+\exp(Y\beta_*^T X)\big)\big(1+\exp(Y\beta_*^T X)\big)}\right\|_p \\
&\geq \left(\frac{1}{1+\exp(Y\beta_*^T X)} - \frac{\|X\|_q \|\beta_*\|_p}{\big(1+\exp(Y\beta_*^T X)\big)\big(1+\exp(-Y\beta_*^T X)\big)}\right)\|\xi\|_p, \qquad (2.52)
\end{aligned}$$

because $Y \in \{+1, -1\}$, and due to Hölder's inequality $|\xi^T X| \leq \|\xi\|_p \|X\|_q$. If $\xi \in \mathbb{R}^d$ is such that $\|\xi\|_p = (1-\epsilon)^{-2} > 1$ for a given $\epsilon > 0$, then following (2.52), $\|\xi^T D_x h(X, Y)\|_p > 1$, whenever

$$(X, Y) \in \Omega_\epsilon := \left\{(x, y) : \frac{\|x\|_q \|\beta_*\|_p}{1+\exp(-y\beta_*^T x)} \leq \frac{\epsilon}{2},\ \frac{1}{1+\exp(y\beta_*^T x)} \geq 1 - \frac{\epsilon}{2}\right\}.$$

Since X has positive density almost everywhere, the set Ω_ϵ has positive probability for every $\epsilon > 0$. Thus, if $\|\xi\|_p > 1$, $\|\xi^T D_x h(X, Y; \beta_*)\|_p > 1$ with positive probability. Therefore, A is a subset of $\{\xi : \|\xi\|_p \leq 1\}$. Consequently,

$$L_3 := \sup_{\xi \in A} \xi^T Z \overset{D}{\leq} \sup_{\xi : \|\xi\|_p \leq 1} \xi^T Z = \|Z\|_q.$$

If we let $\tilde{Z} \sim \mathcal{N}(\mathbf{0}, \mathbb{E}[XX^T])$, then $\text{Cov}[\tilde{Z}] - \text{Cov}[Z]$ is positive definite. As a result,

L_3 is stochastically dominated by $L_4 := \|\tilde{Z}\|_q$, thus verifying the desired stochastic upper bound in the statement of Theorem 2.5. □

Proof of Theorem 2.6. Instead of characterizing the exact weak limit, we will find a stochastic upper bound for $R_n(\beta_*)$. The RWP function, as in the proof of Theorem 2.4, admits the following dual representation (see Equation (2.49)):

$$
\begin{aligned}
&R_n(\beta_*) \\
&= \sup_{\lambda} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sup_{x'} \left\{ \lambda^T (Y_i - \beta_*^T x') x' - \|x' - X_i\|_\infty^2 \right\} \right\} \\
&= \sup_{\lambda} \left\{ -\lambda^T \frac{Z_n}{\sqrt{n}} - \frac{1}{n} \sum_{i=1}^{n} \sup_{\Delta} \left\{ e_i \lambda^T \Delta - (\beta_*^T \Delta)(\lambda^T X_i) - \left(\|\Delta\|_\infty^2 + (\beta_*^T \Delta)(\lambda^T \Delta) \right) \right\} \right\},
\end{aligned}
$$

where $Z_n = n^{-1/2} \sum_{i=1}^{n} e_i X_i$, $e_i = Y_i - \beta_*^T X_i$. In addition, we have changed the variable from $x' - X_i = \Delta$. If we let $\zeta = \sqrt{n}\lambda$, then

$$
\begin{aligned}
&nR_n(\beta_*) \\
&= \sup_{\zeta} \left\{ -\zeta^T Z_n - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sup_{\Delta} \left\{ e_i \zeta^T \Delta - (\beta_*^T \Delta)(\zeta^T X_i) - \left(\sqrt{n} \|\Delta\|_\infty^2 + (\beta_*^T \Delta)(\zeta^T \Delta) \right) \right\} \right\} \\
&\leq \sup_{\zeta} \left\{ -\zeta^T Z_n - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sup_{\|\Delta\|_\infty} \left\{ \left\| e_i \zeta^T - (\zeta^T X_i) \beta_*^T \right\|_1 \|\Delta\|_\infty - \sqrt{n} \left(1 - \frac{\|\beta_*\|_1 \|\zeta\|_1}{\sqrt{n}} \right) \|\Delta\|_\infty^2 \right\} \right\},
\end{aligned}
$$

where we have used Hölder's inequality thrice to obtain the upper bound. If we solve the inner supremum over the variable $\|\Delta\|$, we obtain,

$$
\begin{aligned}
nR_n(\beta_*) &\leq \sup_{\zeta} \left\{ -\zeta^T Z_n - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\left\| e_i \zeta - (\zeta^T X_i) \beta_* \right\|_1^2}{4\sqrt{n} \left(1 - \|\beta_*\|_1 \|\zeta\|_1 n^{-1/2} \right)} \right\} \\
&\leq \sup_{a \geq 0} \sup_{\zeta : \|\zeta\|_1 = 1} \left\{ -a \zeta^T Z_n - \frac{a^2}{4 \left(1 - a \|\beta_*\|_1 n^{-1/2} \right)} \frac{1}{n} \sum_{i=1}^{n} \left\| e_i \zeta - (\zeta^T X_i) \beta_* \right\|_1^2 \right\},
\end{aligned}
$$

where we have split the optimization into two parts: one over the magnitude (denoted by a), and another over all unit vectors ζ. Further, due to Hölder's inequality, we have $|\zeta^T Z_n| \leq \|Z_n\|_\infty$ as $\|\zeta\|_1 = 1$. Therefore,

$$nR_n(\beta_*) \leq \sup_{a\geq 0}\left\{c_1(n)a - \frac{a^2}{\big(1-c_2(n)a\big)}c_3(n)\right\},$$

where we have let

$$c_1(n) = \|Z_n\|_\infty, \quad c_2(n) = \|\beta_*\|_1 n^{-1/2}, \text{ and } c_3(n) = \inf_{\zeta:\|\zeta\|_1=1}\frac{1}{4n}\sum_{i=1}^n \left\|e_i\zeta - (\zeta^T X_i)\beta_*\right\|_1^2.$$

As $\|\beta_*\|_1 n^{-1/2} \to 0$ when $n \to \infty$, we have $c_2(n) \to 0$. Therefore, the above supremum over a is attained at $a = c_1(n)/2c_3(n) + o(1)$ when $n \to \infty$. Consequently,

$$nR_n(\beta_*) \leq \frac{\|Z_n\|_\infty^2}{\inf_{\{\zeta:\|\zeta\|_1=1\}}\frac{1}{n}\sum_{i=1}^n \left\|e_i\zeta - (\zeta^T X_i)\beta_*\right\|_1^2} + o(1). \tag{2.53}$$

The infimum in the denominator can be lower bounded as in the proof of Theorem 2.4. In particular, due to triangle inequality,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \left\|e_i\zeta - (\zeta^T X_i)\beta_*\right\|_1^2 &\geq \frac{1}{n}\sum_{i=1}^n \left(|e_i|\,\|\zeta\|_1 - \left|\zeta^T X_i\right|\,\|\beta_*\|_1\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n |e_i|^2 + \|\beta_*\|_1^2\,\frac{1}{n}\sum_{i=1}^n \left|\zeta^T X_i\right|^2 - 2\,\|\beta_*\|_1\,\frac{1}{n}\sum_{i=1}^n |e_i|\left|\zeta^T X_i\right| \\
&= \frac{1}{n}\sum_{i=1}^n |e_i|^2 + \|\beta_*\|_1^2\frac{1}{n}\sum_{i=1}^n \left|\zeta^T X_i\right|^2 - 2\,\|\beta_*\|_1\,\mathbb{E}\,|e_i|\,\frac{1}{n}\sum_{i=1}^n \left|\zeta^T X_i\right| - \epsilon_n(\zeta),
\end{aligned}$$

where $\epsilon_n(\zeta) = 2\,\|\beta_*\|_1\,\frac{1}{n}\sum_{j=1}^n(|e_i| - E|e_i|)|\zeta^T X_i|$. If we let $\tilde{e}_i = |e_i| - E\,|e_i|$, then $\mathbb{E}[\tilde{e}_i] = 0$ and $\mathrm{Var}[\tilde{e}_i] \leq \mathrm{Var}[e_i]$. As $\tilde{e}_i$ is independent of X_i, $\mathbb{E}[\tilde{e}_i|\zeta^T X_i|] = 0$. In addition,

$$\mathrm{Var}\left[\tilde{e}_i|\zeta^T X_i|\right] = \mathrm{Var}[\tilde{e}_i]\zeta^T\Sigma\zeta \leq \mathrm{Var}[e_i]\zeta^T\Sigma\zeta,$$

where we recall that $\Sigma = \mathrm{Cov}[X]$. With the assumption that on the largest eigen value of Σ, denoted by $\lambda_{\max}(\Sigma)$, is $o(nC(n,d)^2)$, we have

$$\sup_{\|\zeta\|_1=1} \zeta^T \Sigma \zeta \leq \sup_{|\zeta|_1=1} \lambda_{\max}(\Sigma) \|\zeta\|_2^2 \leq \lambda_{\max}(\Sigma) = o(nC(n,d)^2).$$

Consequently, the variance of $\frac{1}{n}\sum_{j=1}^n \tilde{e}_i \left|\zeta^T X_i\right|$ is of order $o\left(C(n,d)^2\right)$ uniformly in ζ for $\|\zeta\|_1 = 1$. Combining this with the assumption that $\|\beta_*\|_1 = o(1/C(n,d))$ we have

$$\epsilon_n(\zeta) = 2\|\beta_*\|_1 \frac{1}{n}\sum_{j=1}^n (|e_i| - \mathbb{E}|e_i|)\left|\zeta^T X_i\right| = o_p(1).$$

Since the bound is uniformly in ζ such that $\|\zeta\|_1 = 1$, we have for sufficiently large n,

$$\begin{aligned}
&\inf_{\zeta:\|\zeta\|_1=1} \frac{1}{n}\sum_{i=1}^n \left\|e_i\zeta - \zeta^T X_i \beta_*\right\|_1^2 \\
&\quad \geq \frac{1}{n}\sum_{i=1}^n |e_i|^2 + \inf_{\zeta:\|\zeta\|_1=1} \left\{ \|\beta_*\|_1^2 \frac{1}{n}\sum_{i=1}^n \left|\zeta^T X_i\right|^2 - 2\|\beta_*\|_1 \mathbb{E}|e_i| \frac{1}{n}\sum_{i=1}^n \left|\zeta^T X_i\right| \right\} + o_p(1) \\
&\quad \geq \frac{1}{n}\sum_{i=1}^n |e_i|^2 - \left(\mathbb{E}|e_i|\right)^2 + \inf_{\zeta:\|\zeta\|_1=1} \left(\|\beta_*\|_1 \frac{1}{n}\sum_{i=1}^n \left|\zeta^T X_i\right| - \mathbb{E}|e_i| \right)^2 + o_p(1) \\
&\quad \geq \mathrm{Var}\,|e_i| + o_p(1).
\end{aligned}$$

Then, as $n \to \infty$, it follows from Equation (2.53),

$$nR_n(\beta_*) \leq \frac{\|Z_n\|_\infty^2}{\mathrm{Var}\,|e|} + o_p(1).$$

The second claim is a direct consequence of Corollary 2.1 in Chernozhukov *et al.* [2013] when X has sub-Gaussian tails. Finally, the last claim is the special example of computing the $(1-\alpha)$ quantile of $\|Z\|_\infty$ for $Z \sim \mathcal{N}(0, I_d)$. Here, the distribution of maximum of d i.i.d. standard normal random variables have $\Phi^{-1}(1-\alpha/2d)$ as its

$(1-\alpha)$ quantile, and

$$\frac{\mathbb{E}[e^2]}{\mathbb{E}[e^2]-(\mathbb{E}|e|)^2}=\pi/(\pi-2),$$

when the additive error e is normally distributed. □

APPENDIX 2.B: Strong duality of the linear semi-infinite programs in the chapter

In this chapter, we have utilized strong duality of linear semi-infinite programs in two contexts: 1) to derive a dual representation of the RWP function in order to perform asymptotic analysis (see Proposition 2.1), and 2) to derive distributional robust representations (see Proposition 2.4). Establishing these strong dualities rely on the following well-known result on problem of moments (Isii [1962]; Newey and Smith [2004]).

The problem of moments. Let Ω be a nonempty Borel measurable subset of $\mathbb{R}^m$, which, in turn, is endowed with the Borel sigma algebra $\mathcal{B}_\Omega$. Let X be a random vector taking values in the set Ω, and $f=(f_1,\ldots,f_k):\Omega\to\mathbb{R}^k$ be a vector of moment functionals. Let $\mathcal{P}_\Omega$ and $\mathcal{M}_\Omega^+$ denote, respectively, the set of probability and non-negative measures, respectively on $(\Omega,\mathcal{B}_\Omega)$ such that the Borel measurable functionals $\phi, f_1, f_2, \ldots, f_k$, defined on Ω, are all integrable. Given a real vector $q=(q_1,\ldots,q_k)$, the objective of the problem of moments is to find the worst-case bound,

$$v(q):=\sup\left\{\mathbb{E}_\mu[\phi(X)]:\mathbb{E}_\mu[f(X)]=q,\ \mu\in\mathcal{P}_\Omega\right\}. \tag{2.54}$$

If we let $f_0=\mathbf{1}_\Omega$, it is convenient to add the constraint, $\mathbb{E}_\mu[f_0(X)]=1$, by appending $\tilde{f}=(f_0,f_1,\ldots,f_k)$, $\tilde{q}=(1,q_1,\ldots,q_k)$, and consider the following reformulation of

the above problem:

$$v(q) := \sup \left\{ \int \phi(x) d\mu(x) : \int \tilde{f}(x) d\mu(x) = \tilde{q},\ \mu \in \mathcal{M}_{\Omega}^{+} \right\}. \tag{2.55}$$

Then, under the assumption that a certain Slater's type of condition is satisfied, one has the following equivalent dual representation for the moment problem Equation (2.55). See Theorem 1 (and the discussion of Case [I] following Theorem 1) in Isii [1962] for a proof of the following result:

Proposition 2.6. Let $\mathcal{Q}_{\tilde{f}} = \left\{ \int \tilde{f}(x) d\mu(x) : \mu \in \mathcal{M}_{\Omega}^{+} \right\}$. If $\tilde{q} = (1, q_1, \ldots, q_k)$ is an interior point of $\mathcal{Q}_{\tilde{f}}$, then

$$v(q) = \inf \left\{ \sum_{i=0}^{k} a_i q_i :\ a_i \in \mathbb{R},\ \sum_{i=0}^{k} a_i \tilde{f}_i(x) \geq \phi(x) \text{ for all } x \in \Omega \right\}.$$

In the rest of this section, we recast the dual reformulation of RWP function (in Equation (2.1)) and the dual reformulation of the distributional representation in Proposition 2.4 as particular cases of the dual representation of the problem of moments in Proposition 2.6.

Dual representation of RWP function Recall from Section 2.2.3 that W is a random vector taking values in $\mathbb{R}^m$ and $h(\cdot, \theta)$ is Borel measurable.

Proof of Proposition 2.1. For simplicity, we do not write the dependence on parameter θ in $h(u, \theta)$ and $R_n(\theta)$ in this proof; nevertheless, we should keep in mind that the RWP function is a function of parameter θ. Given estimating equation $E[h(W)] = \mathbf{0}$. Recall the definition of the corresponding RWP function,

$$\begin{aligned} R_n &:= \inf \left\{ D_c(P, P_n) :\ \mathbb{E}_P\big[h(W)\big] = \mathbf{0} \right\} \\ &= \inf \left\{ \mathbb{E}_\pi\big[c(U, W)\big] :\ \mathbb{E}_\pi\big[h(U)\big] = \mathbf{0},\ \pi_{_W} = P_n,\ \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m) \right\}, \end{aligned}$$

where $\pi_{_W}$ denotes the marginal distribution of W. To recast this as a problem of moments as in Equation (2.54), let $\Omega = \{(u,w) \in \mathbb{R}^m \times \mathbb{R}^m : c(u,w) < \infty\}$,

$$f(u,w) = \begin{bmatrix} \mathbf{1}_{\{w=W_1\}}(u,w) \\ \mathbf{1}_{\{w=W_2\}}(u,w) \\ \vdots \\ \mathbf{1}_{\{w=W_n\}}(u,w) \\ h(u) \end{bmatrix} \qquad \text{and} \qquad q = \begin{bmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \\ \mathbf{0} \end{bmatrix}.$$

Further, let $\phi(u,w) = -c(u,w)$, for all $(u,w) \in \Omega$. Then,

$$R_n = -\sup\left\{\mathbb{E}_\pi\big[\phi(U,W)\big] : \ \mathbb{E}_\pi\big[f(U,W)\big] = q, \ \pi \in \mathcal{P}_\Omega\right\},$$

is of the same form as Equation (2.54). Let H denote the convex hull of the range $\{h(u) : (u,w) \in \Omega\}$. Then, following the definition of $Q_{\tilde{f}}$ in the abstract formulation in Proposition 2.6, we obtain $Q_{\tilde{f}} = \mathbb{R}^{n+1}_+ \times H$. As $\{\mathbf{0}\}$ lies in the interior of convex hull H, it is immediate that the Slater's condition, $\tilde{q} = (1,q)$ lying in the interior of the $Q_{\tilde{f}}$, is satisfied. Consequently, we obtain the following dual representation of R_n due to Proposition 2.6:

$$\begin{aligned} R_n &= -\inf_{a_i \in \mathbb{R}} \left\{ a_0 + \frac{1}{n}\sum_{i=1}^n a_i : \ a_0 + \sum_{i=1}^n a_i \mathbf{1}_{\{w=W_i\}}(u,w) \right. \\ &\qquad \left. + \sum_{i=n+1}^k a_i h_i(u) \geq -c(u,w), \text{ for all } (u,w) \in \Omega \right\} \\ &= -\inf_{a_i \in \mathbb{R}} \left\{ a_0 + \frac{1}{n}\sum_{i=1}^n a_i : \ a_0 + a_i \geq \sup_{u : c(u,W_i) < \infty} \left\{ -c(u,W_i) - \sum_{i=n+1}^k a_i h_i(u) \right\} \right\}. \end{aligned}$$

As the inner supremum is not affected even if we take supremum over $\{u : c(u,W_i) =$

$\infty\}$, after letting $\lambda = (a_{n+1}, \ldots, a_k)$ for notational convenience, we obtain

$$R_n = \sup_{\lambda} \left\{ \frac{1}{n} \sum_{i=1}^{n} \inf_{u \in \mathbb{R}^m} \left\{ c(u, W_i) + \lambda^T h(u) \right\} \right\}. \tag{2.56}$$

As λ is a free variable, we flip the sign of λ to arrive at the statement of Proposition 2.1. This completes the proof. □

Dual representation of the DRO formulation in Equation (2.15) Here, we provide a proof for the dual representation in Proposition 2.4 that has been instrumental in establishing the distributional robust representations of Lasso and regularized logistic regression.

Proof of Proposition 2.4. Given a Borel measurable g, our first objective is to prove that the worst-case loss $\sup\{\mathbb{E}_P[g(W)] : D_c(P, P_n) \leq \delta\}$ admits the dual representation,

$$\sup\{\mathbb{E}_P[g(W)] : D_c(P, P_n) \leq \delta\} = \inf_{\lambda \geq 0} \left\{ \lambda\delta + \frac{1}{n} \sum_{i=1}^{n} \phi_\lambda(W_i) \right\}, \tag{2.57}$$

with $\phi_\lambda(W_i) = \sup_u\{g(u) - \lambda c(u, w)\}$. This would essentially prove Proposition 2.4 if we let $W = (X, Y)$, $g(W) = l(X, Y; \beta)$ and $\phi_\lambda(X, Y; \beta) = \phi_\lambda(W)$.

Since the problem $\sup\{\mathbb{E}_P[g(W)] : D_c(P, P_n) \leq \delta\}$ has inequality constraints, one way is to proceed exactly as in RWP dual formulation above except for restricting the Lagrange multiplier corresponding to the equality constraint to be non-negative. Alternatively, one can recast the problem as in Equation (2.54) with the introduction

of a slack variable S as below:

$$\begin{aligned} & \sup\big\{\mathbb{E}_P[g(W)] : D_c(P, P_n) \le \delta\big\} \\ = \; & \sup\big\{\mathbb{E}_\pi\big[g(U)\big] : \; \mathbb{E}_\pi\big[c(U,W) + S\big] = \delta, \; \pi_{_W} = P_n, \\ & \qquad \pi(S = v) = 1, \; \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}_+)\big\}. \end{aligned}$$

In the context of notation introduced for the problem of moments described at the beginning of this appendix, let $\Omega = \{(u, w, s) : c(u, w) < \infty, s \ge 0\}$,

$$f(u,w,s) = \begin{bmatrix} \mathbf{1}_{\{w=W_1\}}(u,w,s) \\ \mathbf{1}_{\{w=W_2\}}(u,w,s) \\ \vdots \\ \mathbf{1}_{\{w=W_n\}}(u,w,s) \\ \mathbf{1}_{\{s=v\}} \\ c(u,w) + s \end{bmatrix} \qquad \text{and} \qquad q = \begin{bmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \\ 1 \\ \delta \end{bmatrix}.$$

In addition, if we let $\phi(u, w, s) = g(u)$, then

$$\sup\big\{\mathbb{E}_P[g(W)] : \; D_c(P, P_n) \le \delta\big\} = \sup\big\{\mathbb{E}_\pi[\phi(U,W,S)] : \; \mathbb{E}_\pi[f(U,W,S)] = q, \; \pi \in \mathcal{P}_\Omega\big\},$$

is a problem of moments of the form Equation (2.54). Similar to the RWP dual formulation discussed earlier in the section, $\tilde{q} = (1, q)$ lies in the interior of $Q_{\tilde{f}} = \mathbb{R}_+^{n+3}$, thus satisfying Slater's condition for all $\delta > 0$. Then, due to Proposition 2.6, we obtain

$$\sup\big\{\mathbb{E}_P[g(W)] : \; D_c(P, P_n) \le \delta\big\} = \inf_{a \in \mathbb{A}} \left\{ a_0 + \frac{1}{n}\sum_{i=1}^{n} a_i + a_{n+1} + a_{n+2}\delta \right\},$$

where the set $\mathbb{A}$ is the collection of $a = (a_0, a_1, \ldots, a_{n+1}) \in \mathbb{R}^{n+3}$ such that

$$a_0 + \sum_{i=1}^{n} a_i \mathbf{1}_{\{w=W_i\}}(u, w, s) + a_{n+1}\mathbf{1}_{\{s=v\}}(u, w, s) + a_{n+2}\big(c(u, w) + s\big) \geq g(u),$$

for all $(u, w, s) \in \Omega$. Further, observe that the value of the optimization problem above does not change, even if we consider only the following constraints:

$$\begin{aligned} a_0 + a_i + a_{n+1} &\geq \sup\Big\{g(u) - a_{n+2}\big(c(u, W_i) + s\big) : u \in \mathbb{R}^m,\ s \geq 0\Big\} \\ &= \begin{cases} \sup_{u\in\mathbb{R}^m}\Big\{g(u) - a_{n+2}c(u, W_i)\Big\} & \text{if } a_{n+2} \geq 0, \\ \infty & \text{if } a_{n+2} < 0. \end{cases} \end{aligned}$$

If we recall the notation that $\phi_\lambda(W_i) = \sup_{u\in\mathbb{R}^m}\{g(u) - \lambda c(u, W_i)\}$, then

$$\begin{aligned} & \sup\big\{\mathbb{E}_P[g(W)] :\ D_c(P, P_n) \leq \delta\big\} \\ = & \inf_{\substack{a_{n+2}\geq 0 \\ a_i\in\mathbb{R},}} \left\{a_0 + \frac{1}{n}\sum_{i=1}^{n} a_i + a_{n+1} + a_{n+2}\delta : a_0 + a_i + a_{n+1} \geq \phi_{a_{n+2}}(W_i)\right\} \\ = & \inf_{a_{n+2}\geq 0}\big\{\phi_{a_{n+2}}(W_i) + a_{n+2}\delta\big\}, \end{aligned}$$

thus proving Equation (2.57). As explained earlier, letting $W = (X, Y)$, $g(W) = l(X, Y; \beta)$ and $\phi_\lambda(X, Y; \beta)$ in Equation (2.57) verifies the proof of Proposition 2.4. $\square$

APPENDIX 2.C: Exchange of sup and inf in the DRO formulation Equation (2.15)

Proposition 2.7. Let us write

$$\mathcal{U}_\delta = \left\{P : D_c\big(P, P_n\big) \leq \delta\right\},$$

and define

$$g\left(\beta\right) = \sup_{P \in \mathcal{U}_\delta} \mathbb{E}_P\left[l\big(X, Y; \beta\big)\right].$$

Suppose that $g\left(\cdot\right)$ is convex and assume that there exists $b \in \mathbb{R}$ such that $\kappa_b = \{\beta : g\left(\beta\right) \leq b\}$ is compact and non-empty. In addition, suppose that $\mathbb{E}_P\left[l\big(X, Y; \beta\big)\right]$ is lower semi-continuous and convex as a function of β throughout κ_b. Then,

$$\min_{\beta \in \mathbb{R}^d} \sup_{P : D_c\left(P, P_n\right) \leq \delta} \mathbb{E}_P\left[l\big(X, Y; \beta\big)\right] = \sup_{P : D_c\left(P, P_n\right) \leq \delta} \min_{\beta \in \mathbb{R}^d} \mathbb{E}_P\left[l\big(X, Y; \beta\big)\right].$$

Proof. By definition of κ_b we have that

$$\min_{\beta \in \mathbb{R}^d} \sup_{P : D_c\left(P, P_n\right) \leq \delta} \mathbb{E}_P\left[l\big(X, Y; \beta\big)\right] = \min_{\beta \in \kappa_b} \sup_{P : D_c\left(P, P_n\right) \leq \delta} \mathbb{E}_P\left[l\big(X, Y; \beta\big)\right].$$

By a min-max result of Terkelsen (see Corollary 2 in Terkelsen [1973]), since both $\mathcal{U}_\delta(P_n)$ and κ_b are convex, κ_b is compact, $\mathbb{E}_P\left[l\big(X, Y; \beta\big)\right]$ is lower semi-continuous and convex throughout κ_b as a function of β, and $\mathbb{E}_P\left[l\big(X, Y; \beta\big)\right]$ is concave as a function of P, then

$$\min_{\beta \in \kappa_b} \sup_{P : D_c\left(P, P_n\right) \leq \delta} \mathbb{E}_P\left[l\big(X, Y; \beta\big)\right] = \sup_{P : D_c\left(P, P_n\right) \leq \delta} \min_{\beta \in \kappa_b} \mathbb{E}_P\left[l\big(X, Y; \beta\big)\right].$$

The proof is complete if are able to argue the identity

$$\sup_{P:D_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\kappa_b}\mathbb{E}_P\left[l\big(X,Y;\beta\big)\right]=\sup_{P:D_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\mathbb{R}^d}\mathbb{E}_P\left[l\big(X,Y;\beta\big)\right].$$

To see this, note that we always have

$$\sup_{P:D_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\kappa_b}\mathbb{E}_P\left[l\big(X,Y;\beta\big)\right]\geq\sup_{P:D_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\mathbb{R}^d}\mathbb{E}_P\left[l\big(X,Y;\beta\big)\right]. \tag{2.58}$$

Let us assume that the strict inequality holds. If this is the case then we must have that there exists $\beta'\notin\kappa_b$ such that

$$\begin{aligned} b &< g\left(\beta'\right)=\sup_{P:D_c\left(P,P_n\right)\leq\delta}\mathbb{E}_P\left[l\big(X,Y;\beta'\big)\right] \\ &< \sup_{P:D_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\kappa_b}\mathbb{E}_P\left[l\big(X,Y;\beta\big)\right] \\ &\leq b, \end{aligned}$$

where the second inequality follows because we are assuming that (2.58) holds with strict inequality.We therefore contradict the assumption that the strict inequality in (2.58) holds. Hence, the proof is complete. □

Proof of Lemma 2.1. Let us consider linear regression loss function first. Under null hypothesis, $\mathbb{E}\|X\|_2^2<\infty$ and $\mathbb{E}[e^2]<\infty$. Therefore, for any $\beta\in\mathbb{R}^d$, $\mathbb{E}[l(X,Y;\beta)]=\mathbb{E}[(Y-\beta^TX)^2]<\infty$. Further, as the loss function $l(x,y;\beta)$ is a convex function of β,

$$g(\beta)=\sup_{P\in\ \mathcal{U}_\delta(P_n)}\mathbb{E}_P[l(X,Y;\beta)]=\left(\sqrt{\mathbb{E}_{P_n}[(Y-\beta^TX)^2]}+\sqrt{\delta}\|\beta\|_p\right)^2$$

is convex as well and finite for all β in $\mathbb{R}^d$ (the second equality follows from the

distributional robust representation in Theorem 2.2). Further, as $g(\beta) \to \infty$ when $\|\beta\|_p \to \infty$ and $g(\beta)$ is convex and continuous in $\mathbb{R}^d$, the level sets $\kappa_b = \{\beta : g(\beta) \leq b\}$ are compact and nonempty as long as $b > (\sqrt{\mathbb{E}_{P_n}[(Y - \beta_*^T X)^2]} + \sqrt{\delta}\|\beta_*\|)^2$. Finally, due to the convexity and finiteness of $E[l(X, Y; \beta)]$ lower semi-continuity of $\mathbb{E}[l(X, Y; \beta)]$ is immediate as well. As all the conditions in Proposition 2.7 are satisfied, the sup and inf in the DRO formulation (2.15) can be exchanged in the linear regression example as a consequence of Proposition 2.7.

It is straightforward to check that exactly same argument applies for logistic regression loss function as well when $\mathbb{E}\|X\|_2^2$ is finite. □

APPENDIX 2.D: Numerical Examples

In this section, we consider two examples that demonstrate the numerical performance of the square-root Lasso (SR-Lasso) algorithm (see Example 2.2) when the regularization parameter λ is selected as described in Section 2.4.1 using a suitable quantile of the RWPI limiting distribution.

Example 2.4. (RWPI on Sparse-regression) Consider the linear model $Y = 3X_1 + 2X_2 + 1.5X_4 + e$ where the vector of predictor variables $X = (X_1, \ldots, X_d)$ is distributed according to the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma_{k,j} = 0.5^{|k-j|}$ and additive error e is normally distributed with mean 0 and standard deviation $\sigma = 10$. Letting n denote the number of training samples, we illustrate the effectiveness of the RWPI based SR-Lasso procedure for various values of d and n by computing the mean square loss / error (MSE) over a simulated test data set of size $N = 10000$. Specifically, we take the number of predictors to be $d = 300$ and 600, the number of standardized i.i.d. training samples to range from $n = 350, 700, 3500, 10000$, and the desired confidence level

to be 95%, that is, $1-\alpha = 0.95$. In each instance, we run the SR-Lasso algorithm using the 'flare' package proposed in Li *et al.* [2015] (available as a library in R) with regularization parameter λ chosen as prescribed in Section 2.4.1.

Repeating each experiment 100 times, we report the average training and test MSE in Tables 2.1 and 2.2, along with the respective results for ordinary least squares regression (OLS) and SR-Lasso algorithm with regularization parameter chosen as prescribed by cross-validation (denoted as SR-Lasso CV in the tables.) We also report the average ℓ_1 and ℓ_2 error of the regression coefficients in Tables 2.1 and 2.2. In addition, we report the empirical coverage probability that the optimal error $\mathbb{E}[(Y - \beta_*^T X)^2] = \sigma^2 = 100$ is smaller than the worst case expected loss computed by the DRO formulation Equation (2.15). As this empirical coverage probability reported in Table 2.3 is closer to the desired confidence $1 - \alpha = 0.95$, the worst case expected loss computed by Equation (2.15) can be seen as a tight upper bound of the optimal loss $\mathbb{E}[l(X, Y; \beta_*)]$ (thus controlling generalization) with probability at least $1 - \alpha = 0.95$.

Example 2.5. (RWPI on Diabetes data) Consider the diabetes data set from the 'lars' package in R (see Efron *et al.* [2004]), where there are 64 predictors (including 10 baseline variables and other 54 possible interactions) and 1 response. After standardizing the variables, we split the entire data set of 442 observations into $n = 142$ training samples (chosen uniformly at random) and the remaining $N = 300$ samples as test data for each experiment, in order to compute training and test mean square errors using the generalized Lasso algorithm with regularization parameter picked as in Section 2.4.1. After repeating the experiment 100 times, we report the average training and test errors in Table 2.4, and compare the performance of RWPI based regularization parameter selection with other

standard procedures such as OLS and SR- Lasso algorithm with regularization parameter chosen according to cross-validation.

Training data size, n	Method	Training Error	Test Error	ℓ_1 loss $\|\beta - \beta_*\|_1$	ℓ_2 loss $\|\beta - \beta_*\|_2$
350	RWPI	101.16(±8.11)	122.59(±6.64)	4.08(±0.69)	5.23(±0.76)
	SR-Lasso CV	92.23(±7.91)	117.25(±6.07)	3.91(±0.42)	5.02(±1.28)
	OLS	13.95(±2.63)	702.73(±188.05)	31.59(±3.64)	436.19(±50.55)
700	RWPI	101.81(±3.01)	117.96(±4.80)	3.31(±0.40)	4.38(±0.48)
	SR-Lasso CV	99.66(±4.64)	115.46(±4.36)	2.96(±0.37)	3.98(±0.66)
	OLS	56.82(±3.94)	178.44(±21.74)	10.99(±0.57)	152.04(±8.25)
3500	RWPI	102.55(±2.39)	108.44(±2.54)	2.18(±0.16)	3.28(±1.66)
	SR-Lasso CV	100.74(±2.35)	113.83(±2.33)	2.66(±0.14)	3.91(±2.18)
	OLS	90.37(±2.17)	114.78(±5.50)	3.96(±0.20)	54.67(±3.09)
10000	RWPI	102.12(±8.11)	105.97(±0.88)	1.13(±0.08)	1.63(±0.11)
	SR-Lasso CV	100.69(±7.91)	112.82(±0.71)	1.15(±0.07)	1.94(±0.12)
	OLS	95.91(±1.11)	107.74(±2.96)	2.23(±0.10)	30.91(±1.43)

Table 2.1: Sparse linear regression for $d = 300$ predictor variables in Example 2.4. The training and test mean square errors of RWPI based SR- Lasso regularization parameter selection is compared with ordinary least squares estimator (written as OLS) and cross-validation based SR- Lasso estimator (written as SR-Lasso CV)

Training data size, n	Method	Training Error	Test Error	ℓ_1 loss $\|\beta - \beta_*\|_1$	ℓ_2 loss $\|\beta - \beta_*\|_2$
350	RWPI	108.05(±8.38)	109.46(±4.68)	4.02(±0.71)	4.08(±0.70)
	SR-Lasso CV	93.17(±10.83)	104.51(±4.76)	2.23(±0.38)	6.89(±2.35)
	OLS	–	–	–	–
700	RWPI	104.33(±5.03)	103.18(±2.14)	2.91(±0.42)	2.99(±0.43)
	SR-Lasso CV	100.50(±4.70)	99.92(±2.18)	1.45(±0.28)	2.82(±0.64)
	OLS	14.27(±2.02)	699.06(±137.45)	31.66(±2.21)	518.02(±44.87)
3500	RWPI	101.52(±2.52)	96.38(±0.80)	1.23(±0.24)	1.32(±0.24)
	SR-Lasso CV	102.58(±2.49)	98.55(±0.94)	1.18(±0.15)	1.94(±0.24)
	OLS	82.22(±2.31)	102.01(±6.14)	6.76(±0.23)	114.05(±5.73)
10000	RWPI	101.36(±1.11)	94.86(±0.36)	0.75(±0.13)	0.81(±0.14)
	SR-Lasso CV	103.00(±1.11)	98.55(±0.49)	1.16(±0.08)	1.94(±0.13)
	OLS	95.11(±1.10)	99.53(±4.83)	3.26(±0.11)	63.67(±2.16)

Table 2.2: Sparse linear regression for $d = 600$ predictor variables in Example 2.4. The training and test mean square errors of RWPI based SR- Lasso regularization parameter selection is compared with ordinary least squares estimator (written as OLS) and cross-validation based SR-Lasso estimator (written as SR-Lasso CV). As $n < d$ when $n = 350$, OLS estimation is not applicable in that case (denoted by a blank)

No. of predictors	Training sample size			
d	350	700	3500	10000
300	0.974	0.977	0.975	0.969
600	0.963	0.966	0.970	0.968

Table 2.3: Coverage Probability of empirical worst case expected loss in Example 2.4

	Training Error	Testing Error
RWPI	0.58(±0.05)	0.60(±0.04)
SR- Lasso CV	0.44(±0.06)	0.57(±0.03)
OLS	0.26(±0.05)	1.38(±0.68)

Table 2.4: Linear Regression for Diabetes data in Example 2.5 with 142 training samples and 300 test samples. The training and test mean square errors of RWPI based SR- Lasso regularization parameter selection is compared with ordinary least squares estimator (written as OLS) and cross-validation based SR-Lasso estimator (written as SR-Lasso CV).

Chapter 3

Sample-out-of-Sample (SoS) Inference

In this chapter, we present another novel inference approach which we call Sample Out-of-Sample (or SoS) inference. SoS method is the analogue of RWPI method as we introduced in Chapter 2, while we restrict the support of distributional uncertainty set $\mathcal{U}_\delta(P_n)$ to be finite (but not restricted on observed samples). Our motivation is to propose a method which is well suited for data-driven stress testing, in which emphasis is placed on measuring the impact of (plausible) out-of-sample scenarios on a given performance measure of interest (such as a financial loss). The methodology is inspired by Empirical Likelihood (EL), but we optimize the empirical Wasserstein distance (instead of the empirical likelihood) induced by observations. From a methodological standpoint, our analysis of the asymptotic behavior of the induced Wasserstein-distance profile function shows dramatic qualitative differences relative to EL. For instance, in contrast to EL, which typically yields chi-squared weak convergence limits, our asymptotic distributions are often not chi-squared. Also, the rates of convergence that we obtain have some dependence on the dimension in a non-trivial way but which remains controlled as the dimension increases.

3.1 Introduction

The goal of this chapter is to introduce a novel methodology for non-parametric inference which allows to measure the adverse impact of out-of-sample scenarios. We call the procedure Sample Out-of-Sample inference or SoS inference.

In order to motivate our goal and the mathematical development that follows, consider the following stress-testing exercise. An insurance company wishes to estimate a certain expectation of interest, $\mathbb{E}(L(X))$, where X might represent a risk factor and $L(X)$ the corresponding financial loss. The insurance company may estimate $\mathbb{E}(L(X))$ based on n i.i.d. (independent and identically distributed) empirical samples $X_1, ..., X_n \in \mathbb{R}^l$. However, the regulator (or auditor) is also interested in quantifying the potential financial loss based on stress scenarios, say an i.i.d. sample $\tilde{X}_1, ..., \tilde{X}_m \in \mathbb{R}^l$ (for simplicity we let $m = n$). The scenarios provided by the regulator may or may not come from the same distribution as the X_i's.

The methodology developed in this chapter allows to incorporate both the empirical sample and the stress scenarios provided by the regulator in a meaningful way using what we call "the SoS profile function" (or SoS function) which we describe next in the stress-testing setting.

Define $Z_k = X_k$ and $Z_{n+k} = \tilde{X}_k$ for $k = 1, ..., n$ (i.e. merge both the empirical samples and the stress scenarios into a set $\{Z_1, ..., Z_{2n}\}$). The corresponding SoS function in the current context, $R_n(\cdot)$, is defined as

$$R_n(\theta) = \min\{\sum_{i,k} \|X_i - Z_k\|_2^2 \pi(i,k) : \tag{3.1}$$
$$\text{s.t. } \sum_k \pi(i,k) = 1/n \ \forall i,\ \pi(i,k) \geq 0 \ \forall i,k,\ \sum_{i,k} L(Z_k)\pi(i,k) = \theta\}\ .$$

We can easy observe that, SoS function is an analogue of RWP function as we defined

in Equation (2.8), where RWP function is solving a semi-infinite linear programming problem while $R_n(\theta)$ is obtained by solving a regular linear programming problem. There is a strong connection between the SOS function and the Wasserstein's distance of order two. This is discussed in the next section.

The results of this chapter characterize, in particular, the asymptotic distribution of $R_n(\mathbb{E}(L(X)))$ (i.e. assuming $\theta = \mathbb{E}(L(X))$) under reasonable assumptions (e.g. the existence of a density with respect the Lebesgue measure and finite variances for both the $L(X_i)$'s and $L(Y_k)$'s). For example, in the one dimensional case (i.e. $\theta \in \mathbb{R}$ and $l = 1$), we will show that

$$nR_n(\mathbb{E}(L(X))) \Rightarrow \upsilon R, \tag{3.2}$$

where $\upsilon > 0$ is explicitly characterized, and $R \sim \chi^2$ (i.e. chi-squared with one degree of freedom). (Here and thorough the chapter we use $\Rightarrow$ to denote weak convergence.) Therefore, if $\delta_n = \delta/n$ is chosen so that $\mathbb{P}(\chi^2 \leq \delta/\upsilon) \approx .95$ then the set

$$\{\theta : R_n(\theta) \leq \delta_n\} \tag{3.3}$$

(which is easily seen to be an interval) is an approximate 95% confidence interval which uses the stress scenarios in a meaningful way.

It is important to stress that the confidence interval designed via (3.2) contains estimates corresponding to all probability distributions which recognize the possibility of the stress scenarios, but which are also plausible given the available empirical evidence.

Let us provide additional motivation for the study of $R_n(\theta)$ by establishing a connection to distributional robust performance analysis of stochastic systems (see, for

example, Lam and Zhou [2015]; Ben-Tal *et al.* [2013]; Goh and Sim [2010]). To illustrate such connection we continue working with the stress-testing situation introduced earlier. A distributional robust estimate of $\mathbb{E}\left(L(X)\right)$ is obtained by evaluating

$$\mathbb{U}_n\left(\Delta\right) = \max\{\sum_{i,k} L\left(Z_k\right)\pi\left(i,k\right) : \tag{3.4}$$
$$\text{s.t. } \sum_k \pi\left(i,k\right) = 1/n \ \forall i,\ \pi\left(i,k\right) \geq 0 \ \forall i,k,\ \sum_{i,k} \|X_i - Z_k\|_2^2 \pi\left(i,k\right) \leq \Delta\}\}.$$

In simple words, $\mathbb{U}_n\left(\Delta\right)$ provides the worst estimate of the expected loss among all distributions that incorporate both the empirical data and the stress scenarios, and that are within distance Δ (in the corresponding Wasserstein metric) of the empirical distribution. By judiciously choosing Δ, we can guarantee that $\mathbb{U}_n\left(\Delta\right)$ is an upper bound for the actual expected loss, $\mathbb{E}\left(L(X)\right)$, with high probability. Naturally, in order to avoid extremely conservative estimates, it is of interest to find Δ in an optimal way. It is precisely here that the formulation of $R_n\left(\theta\right)$ is useful.

Observe that if $\delta_n = \delta/n$

$$\mathbb{U}_n\left(\delta_n\right) = \max\{\theta : R_n\left(\theta\right) \leq \delta_n\}.$$

To see this equality, let $\theta_n^+ = \max\{\theta : R_n\left(\theta\right) \leq \delta_n\}$ and let $\pi^R\left(\theta_n^+\right)$ be the optimizer of (3.1) (taking $\theta = \theta_n^+$) then, because $\pi^R\left(\theta_n^+\right)$ is feasible for (3.4), we have that $\mathcal{U}_n\left(\delta_n\right) \geq \theta_n^+$. Likewise, let $\pi^U\left(\delta_n\right)$ be the optimizer of (3.4) (taking $\Delta = \delta_n$) then, since $\pi^U\left(\delta_n\right)$ is feasible for (3.1) we obtain that $R_n\left(\mathbb{U}_n\left(\delta_n\right)\right) \leq \delta_n$ and therefore, by definition of θ_n^+ we must have $\mathbb{U}_n\left(\delta_n\right) \leq \theta_n^+$.

Therefore, our study of confidence intervals such as (3.3), and the asymptotic analysis of $R_n\left(\theta\right)$, as we indicate in (3.2) provide the means for optimally choosing δ_n in the context of distributional robust performance analysis. Similar connections

to Empirical Likelihood had been noted in the literature (see Lam and Zhou [2015, 2017]; Blanchet and Murthy [2016]). Additional connections to distributional robust optimization are discussed in Section 3.4.

The main methodological objective of this chapter is to study the asymptotic behavior of general SoS functions for estimating equations (which we define in subsequent sections in the chapter). That is, we wish to estimate θ_* such that

$$\mathbb{E}\left(h\left(\theta_*, X\right)\right) = 0, \tag{3.5}$$

where $h\left(\theta, X\right) = \left(h_1\left(\theta, X\right), ..., h_q\left(\theta, X\right)\right)^T$ (a column vector of functions) and $\theta \in \mathbb{R}^d$ (for $q \leq d$), under standard assumptions which make the inference problem of finding θ_* well posed using suitable SoS functions. Note that the particular case leading to (3.2) is obtained by letting $q = 1 = d$ and $h\left(\theta, x\right) = L(x) - \theta$.

The theory that we develop in this chapter parallels the main fundamental results obtained in the context of Empirical Likelihood (EL), introduced by Art Owen in Owen [1988, 1990, 2001]. In fact, as the reader might appreciate, we borrow a great deal of inspiration from the EL inference paradigm (and its extensions based on divergence criteria, rather than the likelihood function, Owen [2001]). There are, however, several important characteristics of our framework that, we believe, add significant value to the non-parametric inference literature.

First, from a conceptual standpoint, the EL framework restricts the support of the outcomes only to the observed empirical sample and, therefore, there is no reason to expect particularly good out of sample performance of estimates based on EL, for example, in settings similar to the stress testing exercise discussed earlier. In fact, the potentially out-of-sample problems which arise from using divergence criteria for data-driven distributional robust optimization (closely related to EL) are noted in the

stochastic optimization literature, see Esfahani and Kuhn [2015]; Wang *et al.* [2009]; Ben-Tal *et al.* [2013], for related work.

Second, from a methodological standpoint, the mathematical techniques needed to understand the asymptotic behavior of $R_n(\theta)$ are qualitatively different from those arising typically in the context of EL. We will show that if $l \geq 3$, then the following weak convergence limit holds (under suitable assumptions on $h(\cdot)$),

$$n^{1/2+3/(2l+2)} R_n(\theta_*) \Rightarrow R(\theta_*),$$

as $n \to \infty$. Note that the scaling depends on the dimension in a very particular way. In contrast, the Empirical Likelihood Profile function is always scaled linearly in n and the asymptotic limiting distribution is generally a chi-squared distribution with appropriate degrees of freedom and a constant scaling factor. In our case $R(\theta_*)$ can be explicitly characterized, depending on the dimension in a non-trivial way, but it is no longer a suitably scaled chi-squared distribution. As mentioned earlier in (3.2), when $l = 1$, we obtain a similar limiting distribution as in the EL case. The case $l = 2$, interestingly, requires a special analysis. In this case the scaling remains linear in n (as in the case $l = 1$), although the limiting distribution is not exactly chi-squared, but a suitable quadratic form of a multivariate Gaussian random vector. For the case $l \geq 3$ the limiting distribution is not a quadratic transformation of a multivariate Gaussian, but a more complex (yet still explicit) polynomial function depending on the dimension.

At a high level, these qualitative distinctions in the form of the asymptotic arise because of the linear programming formulation underlying the SOS function, which will typically lead to corner solutions (i.e. basic feasible solutions in the language of linear programming). In contrast, in the EL analysis of the profile function, the

optimal solutions are amenable to a perturbation analysis as $n \to \infty$ using a Taylor expansion of higher order terms. The lack of a continuously differentiable derivative (of the optimal solution as a function of θ) requires a different type of analysis relative to the approach (traced back to the classical Wilks theorem, Wilks [1938]) which lies at the core of EL analysis. We believe that the proof techniques that we develop here might have wider applicability.

Let us now provide a precise description of our contributions in this chapter:

a) We characterize the asymptotic distribution of $R_n(\theta_*)$ defined in (3.5) as $n \to \infty$ (see Theorem 3.1).

b) We introduce two forms of the SoS inference framework for estimating equations. We call these the implicit and the explicit SoS formulations, respectively. These formulations, as we shall discuss, are motivated by different types of applications (see Theorem 3.2 and Theorem 3.3).

c) Writing $\theta_* = (\gamma_*, v_*)$ we develop the asymptotic distribution of $R_n(\gamma_*, \bar{v}_n)$, where $\bar{v}_n$ is a suitable consistent plug-in estimator for v_* as $n \to \infty$. This extension is particularly useful to reduce the computational burden involved in solving the optimization problem underlying the use of the SoS function for inference (see Corollary 3.1 and Corollary 3.2).

d) We apply our SOS inference framework in the context of stochastic optimization and stress testing (see Section 3.4).

e) Possible extensions and applications of our framework are given in our conclusions section, namely, Section 3.5. We also discuss results in Chapter 2, which include connections to machine learning, extensions beyond the Wasserstein distance

of order two, and more general distributions for out-of-sample evaluation (beyond those supported on finitely many scenarios as discussed here).

We have discussed the qualitative features of our contributions in a) and b).

About item c), its analysis parallels, in a way, the extensions developed by Hjort *et al.* [2009] in the context of EL. The applications to stochastic optimization, in particular, highlight the need for the general from of SoS function.

Regarding item d). A recent paper of Esfahani and Kuhn [2015] proposes Wasserstein's distance in the context of distributional robust stochastic optimization. In Esfahani and Kuhn [2015], the authors take advantage of recently developed concentration inequalities for the Wasserstein distance (see Fournier and Guillin [2015]) to guarantee an asymptotically correct confidence level for the obtained stochastic programming bounds. In particular, given a certain degree of confidence (say 95%), if one wishes to estimate a plausible distributional robust feasible region within ε error, their bound implies $O\left(\varepsilon^{-l}\right)$ number of samples. In contrast, applying our results to the problems in Esfahani and Kuhn [2015] we can see that $O\left(\varepsilon^{-\min(l,2)}\right)$ samples suffice. In simple words, the bounds obtained in Esfahani and Kuhn [2015] appear to be rather pessimistic; while the bounds in Esfahani and Kuhn [2015] suggest that estimating the distributional uncertain region suffers from the curse of dimensionality, our results show that this is not the case. We believe that our results here might be helpful when estimating Wasserstein's distances in high dimensions.

The rest of the chapter is organized as follows. In Section 3.2 we present and discuss our methodological results, in particular the contributions related to items a) to c) above. In Section 3.3 we provide the proofs of our results. Section 3.4 contains applications to stochastic optimization and stress testing (corresponding to item d) above), and including an empirical example. As mentioned earlier in item e), Section

3.5 contains final considerations and further applications.

3.2 Basic Definitions and Main Results

In this section we present our results for the analysis of the SoS profile function for means first and later we move to estimating equations. As we shall observe, the SoS function is an analogue of RWP function as we defined in Equation (2.8) of Chapter 2.

3.2.1 SoS Function for Means

We state the following underlying assumption throughout this subsection.

A1): Let us write $\mathcal{X}_n = \{X_1, ..., X_n\} \subset \mathbb{R}^l$ to denote an i.i.d. sample from a continuous distribution. So, the cardinality of the set $\mathcal{X}_n$ is n.

A2): We also consider an independent i.i.d. sample $\mathcal{Y}_m = \{\tilde{X}_1, ..., \tilde{X}_m\} \subset \mathbb{R}^l$ from a continuous distribution. Throughout our discussion we shall assume that $m = [\kappa n]$ with $\kappa \in [0, \infty)$.

A3): Assume that $\mathbb{E}\left\|X_1\right\|_2^2 + \mathbb{E}\left\|\tilde{X}_1\right\|_2^2 < \infty$.

A4): If $l = 1$ we assume that X_i and Y_i have positive densities $f_X(\cdot)$ and $f_{\tilde{X}}(\cdot)$. If $l \geq 2$ we assume that X_i and Y_i have differentiable positive densities $f_X(\cdot)$ and $f_{\tilde{X}}(\cdot)$, with bounded gradients.

Define $\mathcal{Z}_{n+m} = \{Z_1, ..., Z_{n+m}\} = \mathcal{X}_n \cup \mathcal{Y}_m$, with $Z_k = X_k$ for $k = 1, ..., n$, and $Z_{n+j} = \tilde{X}_j$ for $j = 1, ..., m$. For any closed set $\mathcal{C}$ let us write $\mathcal{P}(\mathcal{C})$ to denote the set of probability measures supported on $\mathcal{C}$. So, in particular, a typical element

$\upsilon_n \in \mathcal{P}(\mathcal{Z}_{n+m})$ takes the form

$$\upsilon_n(dz) = \sum_{k=1}^{n+m} v(k)\, \delta_{Z_k}(dz),$$

where $\delta_{Z_k}(dz)$ is a Dirac measure centered at Z_k. Now, we shall use $\mu_n \in \mathcal{P}(\mathcal{X}_n)$ to denote the empirical measure associated to $\mathcal{X}_n$. Given any $\pi \in \mathcal{P}\left(\mathcal{X}_n \times \mathcal{Z}_{(n+m)}\right)$ we write $\pi_X \in \mathcal{P}(\mathcal{X}_n)$ to denote the marginal distribution with respect to the first coordinate, namely $\pi_X(dx) = \int_{z \in \mathcal{Z}_{(n+m)}} \pi(dx, dz)$ and, likewise, we define $\pi_Z \in \mathcal{P}(\mathcal{Z}_n)$ as $\pi_Z(dz) = \int_{x \in \mathcal{X}_n} \pi(dx, dz)$.

We have the following formal definition of the SoS function for estimating means.

Definition 1.

The SoS function, $R_n(\cdot)$, to estimate $\theta_* = E(X)$ is defined as

$$\begin{aligned} R_n(\theta_*) = \inf\{ & \int\int \|x - z\|_2^2 \, \pi(dx, dz) : \\ & \text{s.t.} \;\; \pi \in \mathcal{P}\left(\mathcal{X}_n \times \mathcal{Z}_{(n+m)}\right), \pi_X = \mu_n, \pi_Z = \upsilon_n, \int z \upsilon_n(dz) = \theta_* \}, \\ = \inf\{ & \int\int \|x - z\|_2^2 \, \pi(dx, dz) : \\ & \text{s.t.} \;\; \pi \in \mathcal{P}\left(\mathcal{X}_n \times \mathcal{Z}_{(n+m)}\right), \pi_X = \mu_n, \int z \pi_Z(dz) = \theta_* \}. \end{aligned} \tag{3.6}$$

Remark 3.1. The connection to the Wasserstein distance (of order 2), $d_2(\mu_n, \upsilon_n)$, can be directly appreciated by recalling that

$$d_2(\mu_n, \upsilon_n)^2 = \inf\{\int\int \|x - z\|_2^2 \, \pi(dx, dz) : \pi \in \mathcal{P}\left(\mathcal{X}_n \times \mathcal{Z}_{(n+m)}\right), \pi_X = \mu_n, \pi_Z = \upsilon_n\}.$$

In simple words, $R_n(\theta_*)$ is obtained by minimizing the (squared) Wasserstein distance to the empirical measure among all distributions υ_n supported on $\mathcal{Z}_{(n+m)}$ with

expected value equal to θ_* (i.e. $\mathbb{E}_{v_n}(Z) = \int z v_n(dz) = \theta_*$).

We now state the following asymptotic distributional result for the SoS function.

Theorem 3.1 (SoS Profile Function Analysis for Means). In addition to Assumptions A1)-A3), suppose that the covariance matrix of X, $Var(X)$. The following asymptotic result follows

- When $l = 1$,

$$nR_n(\theta_*) \Rightarrow \sigma^2 \chi_1^2$$

  where $\sigma^2 = Var(X)$.

- When $l = 2$,

$$nR_n(\theta_*) \Rightarrow \rho\left(\tilde{Z}\right)\left(2 - \tilde{\eta}\left(\tilde{Z}\right)\rho\left(\tilde{Z}\right)\right)\left\|\tilde{Z}\right\|_2^2$$

  where $\rho\left(\tilde{Z}\right)$ is the unique solution to the equation

$$\frac{1}{\rho} = \tilde{g}\left(\rho\tilde{Z}\right),$$

  and $\tilde{g} : \mathbb{R}^l \to \mathbb{R}$ is a deterministic function defined as

$$\tilde{g}(x) = \mathbb{P}\left(\|x\|_2^2 \geq \tau\right).$$

  The function $\tilde{\eta} : \mathbb{R}^l \to \mathbb{R}$ is a deterministic function given as

$$\tilde{\eta}(x) = \mathbb{E}\left[\max\left(1 - \tau/\|x\|_2^2, 0\right)\right].$$

  Also, $\tilde{Z} \sim N(0, Var(X)) \in \mathbb{R}^l$ and τ is independent of $\tilde{Z}$ satisfying

$$\mathbb{P}(\tau > t) = \mathbb{E}\left[\exp\left(-\left(f_X(X_1) + \kappa f_{\tilde{X}}(X_1)\right)\pi t\right)\right].$$

- When $l \geq 3$,

$$n^{1/2+\frac{3}{2l+2}} R_n(\theta_*) \Rightarrow \frac{2l+2}{l+2} \frac{\left\| \tilde{Z} \right\|_2^{1+\frac{1}{l+1}}}{\left(\mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)} \left(f_X(X_1) + \kappa f_{\tilde{X}}(X_1) \right) \right] \right)^{\frac{1}{l+1}}}$$

  where $\tilde{Z} \sim N(0, Var(X)) \in \mathbb{R}^l$.

3.2.2 SoS Function for Estimating Equations

Throughout this subsection we assume that **A1)** and **A2)** are in force. Let us assume that $h : R^d \times R^l \to R^q$, we assume that $q \leq d$. We impose the following assumptions.

B1) Assume $\theta_* \in R^d$ satisfies

$$\mathbb{E}(h(\theta_*, X)) = 0.$$

B2) Furthermore, suppose that

$$\mathbb{E} \|h(\theta_*, X)\|_2^2 < \infty.$$

Our goal is to estimate θ_* under two reasonable SoS function formulations, which we shall discuss. These are "implicit" or "indirect" and "explicit" or "direct" formulations, we will explain their nature next.

3.2.2.1 Implicit SoS Formulation for Estimating Equations

The first SoS function form for estimating equations is the following, we call it Implicit SoS or Indirect SoS function because the Wasserstein distance is applied to $h(\theta, X_i)$ and $h(\theta, Z_k)$ and thus it implicitly or indirectly induces a notion of proximity among the samples.

Definition 2. ***Implicit SoS Profile Function for Estimating Equations***
Let us write $\mathcal{X}_n^h(\theta_*) = \{h(\theta_*, X_i) : X_i \in \mathcal{X}_n\}$ and $\mathcal{Z}_n^h(\theta_*) = \{h(\theta_*, Z_k) : Z_k \in \mathcal{Z}_n\}$ then

$$R_n(\theta_*) = \inf\{\int\int \|h(\theta_*, x) - h(\theta_*, z)\|_2^2 \pi(dx, dz) : \quad (3.7)$$
$$\text{s.t. } \pi \in \mathcal{P}\left(\mathcal{X}_n^h(\theta_*) \times \mathcal{Z}_n^h(\theta_*)\right), \pi_X = \mu_n, \int h(\theta_*, z)\, \pi_Z(dz) = 0\}.$$

The Implicit SoS formulation might lead to dimension reductions if l (the dimension of the ambient space of X) is large. In addition, the presence of $h(\cdot)$ in the distance evaluation allows the procedure to use the available information in a more efficient way. For instance, if $h(\theta, x) = |x| - \theta$, then the sign of x is irrelevant for the estimation problem and this will have the effect of increasing the power of the Implicit SoS function relative to the explicit counterpart.

The analysis of the Implicit SoS function follows as a direct consequence of Theorem 3.1; just redefine $X_i \leftarrow h(\theta_*, X_i)$, $Z_k \leftarrow h(\theta_*, Z_k)$, and apply Theorem 3.1 directly. Thus the proof of the next result is omitted.

Theorem 3.2 (Implicit SoS Profile Function Analysis). Let us use denote $g_X(\cdot)$ is the density for $h(\theta_*, X_i) \in R^q$ and $g_Y(\cdot)$ for the density of $h(\theta_*, Y_i) \in R^q$. Then, the Wasserstein profile function defined in Equation (3.7) have following asymptotic results:

- When $q = 1$,
$$nR_n(\theta_*) \Rightarrow Var(h(\theta_*, X_1))\, \chi_1^2$$

- When $q = 2$,
$$nR_n(\theta_*) \Rightarrow \rho\left(\tilde{Z}\right)\left[2 - \eta\left(\tilde{Z}\right)\rho\left(\tilde{Z}\right)\right]\left\|\tilde{Z}\right\|_2^2,$$

where $\rho\left(\tilde{Z}\right)$ is the unique solution to the equation

$$\frac{1}{\rho} = \tilde{g}\left(\rho\tilde{Z}\right),$$

and $\tilde{g}: \mathbb{R}^q \to \mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{g}(x) = \mathbb{P}\left(\|x\|_2^2 \geq \tau\right).$$

The function $\tilde{\eta}: \mathbb{R}^q \to \mathbb{R}$ is a deterministic continuous function given as

$$\tilde{\eta}(x) = \mathbb{E}\left[\max\left(1 - \tau/\|x\|_2^2, 0\right)\right].$$

Moreover, $\tilde{Z} \sim N\left(0, Var\left(h\left(\theta_*, X\right)\right)\right) \in R^q$ and τ is independent of $\tilde{Z}$ satisfying

$$\mathbb{P}\left[\tau > t\right] = \mathbb{E}\left[\exp\left(-\left[g_X\left(h\left(\theta_*, X_1\right)\right) + \kappa g_{\tilde{X}}\left(h\left(\theta_*, X_1\right)\right)\right]\pi t\right)\right].$$

- When $q \geq 3$,

$$n^{1/2+\frac{3}{2q+2}} R_n(\theta_*) \Rightarrow \frac{2q+2}{q+2} \frac{\left\|\tilde{Z}\right\|_2^{1+\frac{1}{q+1}}}{\left(\mathbb{E}\left[\frac{\pi^{q/2}}{\Gamma(q/2+1)}\left(g_X\left(h\left(\theta_*, X_1\right)\right) + \kappa g_{\tilde{X}}\left(h\left(\theta_*, X_1\right)\right)\right)\right]\right)^{\frac{1}{q+1}}}$$

where $\tilde{Z} \sim N\left(0, Var\left(h\left(\theta_*, X\right)\right)\right) \in \mathbb{R}^q$.

3.2.2.2 Explicit SoS Formulation for Estimating Equations

The second SoS function form we call Explicit SoS function because the Wasserstein distance is explicitly or directly applied to the samples and the scenarios.

Definition 3. ***Explicit SoS Profile Function for Estimating Equations***

$$R_n(\theta_*) = \inf\{\int\int \|x-z\|_2^2 \,\pi\,(dx,dz) : \tag{3.8}$$
$$\text{s.t. } \pi \in \mathcal{P}\left(\mathcal{X}_n \times \mathcal{Z}_{(n+m)}\right), \pi_X = \mu_n, \int h\,(\theta_*, z)\,\pi_Z\,(dz) = 0\} \,.$$

Both the implicit and explicit SoS have their merits. We have discussed the merit of the implicit SoS formulation. For the Explicit SoS formulation, consider the stress testing application discussed in the Introduction. The interest of an auditor or a regulator might be on the impact of scenarios on a specific performance measure of interest. One might think that the regulator applies the same stress scenarios to different insurance companies or banks, and therefore the function $h\,(\cdot)$ is unique to each insurance company. The regulator is interested in the impact of stress testing scenarios on the structure of the bank (modeled by $h\,(\cdot)$). In this setting, the Explicit SoS formulation appears more appropriate.

While the analysis of the Explicit SoS formulation is also largely based on the techniques developed for Theorem 3.1, it does require some additional assumptions that are not immediately clear without examining the proof of Theorem 3.1. In particular, in addition to **A1)**, **A2)**, **B1)** and **B2)**, here we impose the following assumptions.

BE1) Assume that the derivative of $h\,(\theta_*, x)$ with respect to (w.r.t.) x, $D_x h\,(\theta_*, \cdot) : R^l \to R^{q\times l}$, is continuous function of x and the second derivative w.r.t. x is bounded, i.e. $\|D_x^2 h\,(\theta_*, \cdot)\| < \tilde{K}$ for all x.

BE2) Define $V_i = D_x h\,(\theta_*, X_i) \cdot D_x h\,(\theta_*, X_i)^T \in R^{q\times q}$ and assume that $\Upsilon = \mathbb{E}\,(V_i)$ is strictly positive definite.

We provide the proof of the next result in our technical Section 3.3.

Theorem 3.3 (Explicit SoS Profile Function Analysis). Under assumptions A1)-A2),

B1)-B2) and BE1)-BE2), we have that (3.8) satisfies

- When $l = 1$,

$$nR_n(\theta_*) \Rightarrow \tilde{Z}^T \Upsilon^{-1} \tilde{Z}$$

where $\tilde{Z} \sim N\left(0, Var\left(h\left(\theta_*, X\right)\right)\right) \in \mathbb{R}^q$.

- Assume that $l = 2$. It is possible to uniquely define deterministic continuous mapping, $\tilde{\zeta} : \mathbb{R}^q \to \mathbb{R}^q$, such that

$$z = -\mathbb{E}\left[V_1 I\left(\tau \leq \tilde{\zeta}^T\left(z\right) V_1 \tilde{\zeta}\left(z\right)\right)\right] \tilde{\zeta}\left(z\right),$$

where τ is independent of $\tilde{Z}$ satisfying

$$\mathbb{P}\left(\tau > t\right) = \mathbb{E}\left(\exp\left(-\left[f_X\left(X_1\right) + \kappa f_{\tilde{X}}\left(X_1\right)\right]\pi t\right)\right).$$

Moreover, we have that,

$$nR_n(\theta_*) \Rightarrow -2\tilde{Z}^T \tilde{\zeta}\left(\tilde{Z}\right) - \tilde{\zeta}^T\left(\tilde{Z}\right) \tilde{G}\left(\tilde{\zeta}\left(\tilde{Z}\right)\right) \tilde{\zeta}\left(\tilde{Z}\right),$$

where $\tilde{G} : \mathbb{R}^q \to \mathbb{R}^{q \times q}$ is a deterministic continuous mapping defined as,

$$\tilde{G}\left(\zeta\right) = \mathbb{E}\left[V_1 \max\left(1 - \tau / \left(\zeta^T V_1 \zeta\right), 0\right)\right],$$

and $\tilde{Z} \sim N\left(0, Var\left(h\left(\theta_*, X\right)\right)\right) \in \mathbb{R}^q$.

- Suppose that $l \geq 3$. It is possible to uniquely define deterministic continuous mapping $\tilde{\zeta} : \mathbb{R}^q \to \mathbb{R}^q$, such that

$$z = -\mathbb{E}\left[\frac{\pi^{l/2}\left(f_X\left(X_1\right) + \kappa f_{\tilde{X}}\left(X_1\right)\right)}{\Gamma(l/2+1)} V_1 \cdot \left(\tilde{\zeta}^T\left(z\right) V_1 \tilde{\zeta}\left(z\right)\right)^l\right] \tilde{\zeta}\left(z\right),$$

(note that $\bar{V}_1$ is a function of X_1). Moreover,

$$n^{1/2+\frac{3}{2l+2}} R_n(\theta_*) \;\Rightarrow -2\tilde{Z}^T\tilde{\zeta}\left(\tilde{Z}\right) - \frac{2}{l+2}\tilde{G}\left(\tilde{\tilde{Z}}\right),$$

where $\tilde{G}:\mathbb{R}^q \to \mathbb{R}$ is a deterministic continuous function defined as,

$$\tilde{G}(\zeta) = \mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left(f_X(X_1) + \kappa f_{\tilde{X}}(X_1)\right)\left(\zeta^T V_1 \zeta\right)^{l/2+1}\right],$$

and $\tilde{Z} \sim N\left(0, Var\left(h\left(\theta_*, X\right)\right)\right) \in \mathbb{R}^q$ independent of X_1.

We should observe that unlike implicit formulation, the rate of convergence will only depend on the dimension of data $X_i \in \mathbb{R}^l$, but the shape of asymptotic distribution is determined by the estimating functions $h\left(\theta_*, X_i\right) \in \mathbb{R}^q$.

3.2.3 Plug-in Estimators for SoS Functions

In many situations, for example in the context of stochastic optimization, we are interested in a specific parameter $\theta_* = (\gamma_*, \nu_*) \in R^{d+p}$ such that $\mathbb{E}\left[h\left(\gamma_*, \nu_*, X\right)\right] = 0$, where $\nu_* \in R^p$ is the nuisance parameter. We shall discuss a method that allows us to deal with the nuisance parameter using a plug-in estimator, while taking advantage of the SoS framework for the estimation of γ_*. After we state our assumptions we will provide the results in this section and the proofs, which follow closely those of Theorems 3.3 and 3.2 will be given in Section 3.3.

Throughout this subsection, let us suppose that $h\left(\gamma, \nu, x\right) \in \mathbb{R}^q$. In addition, we impose the following assumptions.

C1) Given γ_* there is a unique $\nu_* \in R^p$ such that

$$\mathbb{E}\left[h\left(\gamma_*, \nu, X\right)\right] = 0 \tag{3.9}$$

and, given ν_*, we also assume that γ_* satisfies

$$\mathbb{E}\left[h\left(\gamma, \nu_*, X\right)\right] = 0. \tag{3.10}$$

C2) We have access to a suitable estimator v_n such that the sequence

$$\left\{n^{1/2}\left(v_n - \nu_*\right)\right\}_{n=1}^{\infty} \text{ is tight,}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h\left(\gamma_*, v_n, X_i\right) \Rightarrow \tilde{Z}',$$

for some random variable $\tilde{Z}'$, as $n \to \infty$.

C3) Assume that $h\left(\gamma, \cdot, x\right)$ is continuously differentiable a.e. (almost everywhere with respect to the Lebesgue measure) in some neighborhood $\mathcal{V}$ around v_*.

C4) Suppose that there is a function $M\left(\cdot\right) : R^l \to (0, \infty)$ satisfying that

$$\left\|h\left(\gamma_*, \nu, x\right)\right\|_2^2 \leq M(x) \text{ for a.e. } \nu \in \mathcal{V},$$
$$\left\|D_\nu h\left(\gamma_*, \nu, x\right)\right\|_2^2 \leq M(x) \text{ for a.e. } \nu \in \mathcal{V},$$

and $E\left(M\left(X_1\right)\right) < \infty$.

3.2.3.1 Plug-in Estimators for Implicit SoS Functions

We are interested in studying the plug-in implicit SoS function (or implicit pseudo-SoS profile function) given by

$$R_n(\gamma_*) = \inf\{\int\int \|h(\gamma_*, v_n, x) - h(\gamma_*, v_n, z)\|_2^2 \, \pi(dx, dz) : \tag{3.11}$$
$$\text{s.t. } \pi \in \mathcal{P}\left(\mathcal{X}_n^h(\gamma_*, v_n) \times \mathcal{Z}_{(n+m)}^h(\gamma_*, v_n)\right), \pi_X = \mu_n, \int h(\gamma_*, v_n, z) \, \pi_Z(dz) = 0\},$$

where,

$$\mathcal{X}_n^h(\gamma_*, v_n) = \{h(\gamma_*, v_n, x) : x \in \mathcal{X}_n\}, \ \mathcal{Z}_{(n+m)}^h(\gamma_*, v_n) = \{h(\gamma_*, v_n, z) : z \in \mathcal{Z}_{(n+m)}\}.$$

We typically will use (3.9) to find a plug-in estimator v_n. Under suitable assumptions on the consistency and convergence rate of the plug-in estimator we have an asymptotic result for (3.11), as we indicate next.

Corollary 3.1 (Plug-in for Implicit SoS Formulation). Assuming **A1)-A2)**, and **C1)-C4)** hold. Moreover, suppose denote $g_X(\cdot)$ as the density for $h(\gamma_*, v_*, X_i) \in R^q$ and $g_Y(\cdot)$ for the density of $h(\gamma_*, v_*, Y_i) \in R^q$. We notice $\tilde{Z}' \in \mathbb{R}^q$ is defined in C2). We obtain that (3.11) has following asymptotic behavior

- When $q = 1$,

$$nR_n(\gamma_*) \Rightarrow \left(\tilde{Z}'\right)^2.$$

- When $q = 2$,

$$nR_n(\gamma_*) \Rightarrow \rho\left(\tilde{Z}'\right)\left[2 - \tilde{\eta}\left(\tilde{Z}'\right)\rho\left(\tilde{Z}'\right)\right]\left\|\tilde{Z}'\right\|_2^2$$

where $\rho\left(\tilde{Z}\right)$ is the unique solution to the equation

$$\frac{1}{\rho} = \tilde{g}\left(\rho\tilde{Z}\right),$$

and $\tilde{g}:\mathbb{R}^q \to \mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{g}(x) = \mathbb{P}\left(\|x\|_2^2 \geq \tau\right).$$

The function $\tilde{\eta}:\mathbb{R}^q \to \mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{\eta}(x) = \mathbb{E}\left[\max\left(1-\tau/\|x\|_2^2, 0\right)\right].$$

Moreover, $\tilde{Z}'$ is defined in assumption C2) and τ is independent of $\tilde{Z}'$ satisfying

$$\mathbb{P}[\tau > t] = \mathbb{E}\left[\exp\left(-\left[g_X\left(h\left(\gamma_*,\nu_*,X_1\right)\right) + \kappa g_{\tilde{X}}\left(h\left(\gamma_*,\nu_*,X_1\right)\right)\right]\pi t\right)\right].$$

- When $q \geq 3$,

$$n^{1/2+\frac{3}{2q+2}} R_n(\gamma_*) \Rightarrow$$
$$\frac{2q+2}{q+2} \frac{\left\|\tilde{Z}'\right\|_2^{1+\frac{1}{q+1}}}{\left(\mathbb{E}\left[\frac{\pi^{q/2}}{\Gamma(q/2+1)}\left(g_X\left(h\left(\gamma_*,\nu_*,X_1\right)\right) + \kappa g_{\tilde{X}}\left(h\left(\gamma_*,\nu_*,X_1\right)\right)\right)\right]\right)^{\frac{1}{q+1}}}.$$

3.2.3.2 Plug-in Estimators for Explicit SoS Functions

We can also analyze plug-in estimators for Explicit SoS profile functions. We now define the explicit plug-in (or pseudo) SoS function based on (3.8) as simply plugging-

in the nuisance parameter:

$$R_n(\gamma_*) = \inf \Big\{ \int\int \|x-z\|_2^2 \, \pi\,(dx, dz) : \tag{3.12}$$
$$\text{s.t.} \qquad \pi \in \mathcal{P}\left(\mathcal{X}_n^h\,(\gamma_*, v_n) \times \mathcal{Z}_{(n+m)}^h\,(\gamma_*, v_n)\right),$$
$$\pi_X = \mu_n, \int h\,(\gamma_*, v_n, z)\,\pi_Z\,(dz) = 0\Big\}.$$

In addition to **C1)** to **C4)** introduced at the beginning of this subsection, we shall impose the following additional assumptions:

C5) Define $\bar{V}_i\,(v_*) = D_x h\,(\gamma_*, \nu_*, X_i) \cdot D_x h\,(\gamma_*, \nu_*, X_i)^T$ and assume that $\bar{\Upsilon} = E\left(\bar{V}_i\right)$ is strictly positive definite.

C6) The function $M\,(\cdot)$ from condition **C4)** also satisfies

$$\|D_x h\,(\gamma_*, \nu, x)\|_2^2 \leq M(x) \text{ for a.e. } \nu \in \mathcal{V}.$$
$$\|D_\nu D_x h\,(\gamma_*, \nu, x)\|_2^2 \leq M(x) \text{ for a.e. } \nu \in \mathcal{V}.$$

C7) The second derivative w.r.t. x exist and bounded, i.e. $\|D_x^2 h\,(\gamma_*, \nu, x)\| < \tilde{K}$ for a.e. $\nu \in \mathcal{V}$ and all x.

Corollary 3.2 (Plug-in for Explicit SoS Formulation). $X_i \in \mathbb{R}^l$, $h\,(\gamma, \nu, x) \in \mathbb{R}^q$. Assume that A1)-A2) and C1)-C7) hold. We notice $\tilde{Z}'$ is defined in C2). Then, the SoS profile function defined in Equation (3.12) has the following asymptotic properties.

- When $l = 1$,
$$nR_n(\gamma_*) \Rightarrow \tilde{Z}'^T \bar{\Upsilon}^{-1} \tilde{Z}'.$$

- Suppose that $l = 2$. It is possible to uniquely define deterministic continuous

mapping $\tilde{\zeta}:\mathbb{R}^q\to\mathbb{R}^q$, such that

$$z=-\mathbb{E}\left[\bar{V}_1 I\left(\tau\leq\tilde{\zeta}^T(z)\,\bar{V}_1\tilde{\zeta}(z)\right)\right]\tilde{\zeta}(z),$$

where τ is independent of $\tilde{Z}'$ satisfying

$$\mathbb{P}(\tau>t)=\mathbb{E}\left(\exp\left(-\left[f_X(X_1)+\kappa f_{\tilde{X}}(X_1)\right]\pi t\right)\right).$$

Furthermore,

$$nR_n(\theta_*)\Rightarrow -2\tilde{\zeta}^T\left(\tilde{Z}'\right)\tilde{Z}'-\tilde{\zeta}^T\left(\tilde{Z}'\right)\tilde{G}\left(\tilde{\zeta}\left(\tilde{Z}'\right)\right)\tilde{\zeta}\left(\tilde{Z}'\right),$$

where $\tilde{G}:\mathbb{R}^q\to\mathbb{R}^{q\times q}$ is a deterministic continuous mapping defined as,

$$\tilde{G}(\zeta)=\mathbb{E}\left[\bar{V}_1\max\left(1-\tau/\left(\zeta^T\bar{V}_1\zeta\right),0\right)\right],$$

and $\tilde{Z}'$ is independent with $\bar{V}_1$ and τ.

- Assume that $l\geq 3$. A deterministic and continuous mapping $\tilde{\zeta}:\mathbb{R}^q\to\mathbb{R}^q$ can be defined uniquely so that

$$z=-\mathbb{E}\left[\frac{\pi^{l/2}\left(f_X(X_1)+\kappa f_{\tilde{X}}(X_1)\right)}{\Gamma(l/2+1)}\bar{V}_1\left(\tilde{\zeta}^T(z)\,\bar{V}_1\tilde{\zeta}(z)\right)^l\right]\tilde{\zeta}(z)$$

(note that $\bar{V}_1$ is a function of X_1). Moreover,

$$n^{1/2+\frac{3}{2l+2}}R_n(\theta_*)\Rightarrow -2\tilde{\zeta}^T\left(\tilde{Z}'\right)\tilde{Z}'-\frac{2}{l+2}\tilde{G}\left(\tilde{\zeta}\left(\tilde{Z}'\right)\right),$$

where $\tilde{G}:\mathbb{R}^q\to\mathbb{R}$ is a deterministic continuous function defined as,

$$\tilde{G}(\zeta)=\mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left(f_X(X_1)+\kappa f_{\tilde{X}}(X_1)\right)\left(\zeta^T\bar{V}_1\zeta\right)^{l/2+1}\right],$$

and $\tilde{Z}'$ and X_1 are independent.

3.3 Methodological Development

We shall analyze the limiting distribution of the SoS profile function for means first. In order to gain some intuition let us perform some basic manipulations. First, without loss of generality we assume $\theta_*=0$, otherwise, we can let $X_i^*=X_i-\theta_*$ and apply the analysis to the $X_i{}^*$'s.

3.3.1 The Dual Problem and High-Level Understanding of Results

The Dual Problem Let us revisit the definition of (3.6) and write it as a linear programming problem,

$$R_n(\theta_*)=\min_{\pi(i,j)\geq 0}\quad\sum_{i=1}^{n}\sum_{j=1}^{m+n}\pi(i,j)\left\|X_i-Z_j\right\|_2^2 \tag{3.13}$$
$$\text{s.t.}\quad\begin{cases}\sum_{j=1}^{(m+n)}\pi(i,j)=1/n,\ \text{for all } i\\ \sum_{j=1}^{(m+n)}\left(\sum_{i=1}^{n}\pi(i,j)\right)Z_j=0\end{cases}.$$

We know with probability 1 when $n\to\infty$, $\vec{0}$ is in the convex hull of Z_j, thus the original linear programming problem is feasible for all n large enough with probability one. Applying the strong duality theorem for linear programming problem, see for

example, Luenberger [1973a], we can write (3.13) in the dual formulation as

$$R_n(\theta_*) = \max_{\lambda,\tilde{\gamma}_i} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \tilde{\gamma}_i \right\}$$
$$s.t. \quad \tilde{\gamma}_i + \|X_i - Z_j\|_2^2 - \lambda^T Z_j \geq 0 \text{ for all } i, j.$$

Let us define $\gamma_i = \tilde{\gamma}_i - \lambda^T Z_i$. By the constraint in the above optimization problem, if we take $i = j$, we have $\tilde{\gamma}_i \geq \lambda^T Z_i$, which is equivalent to $\gamma_i \geq 0$. Then, we can write the optimization problem in γ_i's as

$$R_n(\theta_*) = \max_{\lambda,\gamma_i \geq 0} \left\{ -\lambda^T \bar{X}_n - \frac{1}{n}\sum_{i=1}^{n} \gamma_i \right\}$$
$$\text{s.t. } -\lambda^T X_i - \gamma_i \leq -\lambda^T Z_j + \|X_i - Z_j\|_2^2, \text{ for all } i, j.$$

We can further simplify the constraints by minimizing over j, while keeping i fixed, therefore arriving to the simplified dual formulation

$$R_n(\theta_*) = \max_{\lambda,\gamma_i \geq 0} \left\{ -\lambda^T \bar{X}_n - \frac{1}{n}\sum_{i=1}^{n} \gamma_i \right\} \tag{3.14}$$
$$\text{s.t. } -\lambda^T X_i - \gamma_i \leq \inf_j \left\{ -\lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\}, \text{ for all } i.$$

High-Level Intuitive Analysis At this point we can perform a high-level and intuitive analysis which can help us guide our intuition about our result. First, consider an approximation performed by freeing the Z_j in the constraints of (3.14), in this portion the reader can appreciate that the assumption that X_j has a density yields

$$\inf_j \left\{ \|Z_j - (X_i + \lambda/2)\|_2^2 \right\} = \epsilon_n(i), \tag{3.15}$$

where error $\epsilon_n(i)$ is small and it will be discussed momentarily. Now, observe that the optimal $a_*(i) = X_i + \lambda/2$, therefore

$$\inf_j \left\{-\lambda^T Z_j + \|X_i - Z_j\|_2^2\right\} = -\lambda^T X_i - \|\lambda\|_2^2/4 + \epsilon_n(i).$$

Hence, the i-th constraint in (3.14) takes the form

$$-\lambda^T X_i - \gamma_i \leq -\lambda^T X_i - \|\lambda\|_2^2/4 + \epsilon_n(i),$$

and thus (3.14) can ultimately be written as

$$\begin{aligned} R_n(\theta_*) = -\min_{\lambda,\gamma_i \geq 0} &\left\{\lambda^T \bar{X}_n + \frac{1}{n}\sum_{i=1}^n \gamma_i\right\} \\ \text{s.t. } \gamma_i &\geq (1 - \epsilon_n(i))\|\lambda\|_2^2/4 \text{ for all } i. \end{aligned} \tag{3.16}$$

Consider the case $l = 1$, in this case it is not difficult to convince ourselves (because of the existence of a density) that $\epsilon_n(i) = O_p(1/n)$ as $n \to \infty$ (basically with a probability which is bounded away from zero there will be a point in the sample $\{Z_1, ..., Z_{m+n}\} \backslash X_i$ which is within $O_p(1/n)$ distance of $a_*(i)$). Then it is intuitive to expect the approximation

$$R_n(\theta_*) = -\min_{\lambda} \left\{\lambda \bar{X}_n + (1 + O_p(1/n))\lambda^2/4\right\},$$

which formally yields an optimal selection

$$\lambda_* = -\frac{\bar{X}_n}{(1/2 + O_p(1/n))} = -2\bar{X}_n + O_p\left(1/n^{3/2}\right),$$

and therefore we expect, due to the Central Limit Theorem (CLT), that

$$nR_n(\theta_*) = n\bar{X}_n^2 + nO_p\left(1/n^{3/2}\right) \Rightarrow Var\left(X\right)\chi_1^2, \tag{3.17}$$

as $n \to \infty$. This analysis will be made rigorous in the next subsection.

Let us continue our discussion in order to elucidate why the rate of convergence in the asymptotic distribution of $R_n(\theta_*)$ depends on the dimension. Such dependence arises due to the presence of the error term $\epsilon_n(i)$. Note that in dimension $l = 2$, we expect $\epsilon_n(i) = O_p\left(1/n^{1/2}\right)$; this time, with positive probability (uniformly as $n \to \infty$) we must have that a point in the sample $\{Z_1, ..., Z_{m+n}\}\backslash X_i$ is within $O_p\left(1/n^{1/2}\right)$ distance of $a_*(i)$ (because the probability that X_i lies inside a ball of size $1/n^{1/2}$ around a point a is of order $O\left(1/n^{1/2}\right)$). Therefore, in the case $l = 2$ we formally have $\lambda_*(n) = -\bar{X}_n + O_p\left(n^{-1/2}\right)$, but we know from the CLT that $\bar{X}_n = O_p\left(n^{-1/2}\right)$ so this time contribution of $\epsilon_i(n)$ is non-negligible.

Similarly, when $l \geq 3$ this simple analysis allows us to conclude that the contribution of $\epsilon_i(n) = O\left(n^{-1/l}\right)$ will actually dominate the behavior of $\lambda_*(n)$ and this explains why the rate of convergence depends on the dimension of the vector X_i, namely, l. The specific rate depends on a delicate analysis of the error being $\epsilon_i(n)$ which is performed in the next section. A key technical device introduced in our proof technique is a Poisson point process which approximates the number of points in $\{Z_1, ..., Z_{m+n}\}\backslash X_i$ which are within a distance of size $O\left(n^{-1/l}\right)$ from the free optimizer $a_*(i)$ arising in (3.15).

The introduction of this point process, which in turn is required to analyze $\epsilon_i(n)$, makes the proof of our result substantially different from the standard approach used in the theory of Empirical Likelihood (see Owen [1988]; Qin and Lawless [1994], which builds on Wilks [1938]).

3.3.2 Proof of Theorem 3.1

The proof of Theorem 3.1 is divided in several steps which we will carefully record so that we can build from these steps in order to prove the remaining results in the chapter.

3.3.2.1 Step 1 (Dual Formulation and Lower Bound):

Using the same transformations introduced in (3.13) we can obtain the dual formulation of the SOS profile function (3.6), which is a natural adaptation of (3.14), namely

$$R_n(\theta_*) = \max_{\lambda,\gamma_i \geq 0} \left\{ -\lambda \bar{X}_n - \frac{1}{n}\sum_{i=1}^{n} \gamma_i \right\}$$
$$\text{s.t. } -\lambda^T X_i - \gamma_i \leq \inf_j \left\{ -\lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\}, \text{ for all } i.$$

Observe that the following lower bound applies by optimizing over $a \in \mathbb{R}^l$ instead of $a = Z_j \in \mathcal{Z}_n$, therefore obtaining the lower bound

$$\begin{aligned} \inf_j \left\{ -\lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\} &\geq \inf_a \left\{ -\lambda^T a + \|X_i - a\|_2^2 \right\} \\ &= -\lambda^T X_i - \|\lambda\|_2^2 / 4, \end{aligned}$$

with the optimizer $a_*(X_i, \lambda) = X_i + \lambda/2$.

3.3.2.2 Step 2 (Auxiliary Poisson Point Processes):

Then, for each i let us define a point process,

$$N_n^{(i)}(t, \lambda) = \# \left\{ Z_j : \|Z_j - a_*(X_i, \lambda)\|_2^2 \leq t^{2/l}/n^{2/l}, Z_j \neq X_i \right\},$$

(recall that $Z_j \in \mathbb{R}^l$). Observe that, actually, we have

$$N_n^{(i)}(t,\lambda) = N_n^{(i)}(t,\lambda,1) + N_n^{(i)}(t,\lambda,2),$$

where

$$\begin{aligned} N_n^{(i)}(t,\lambda,1) &= \#\left\{X_j : \|X_j - a_*(X_i,\lambda)\|_2^2 \le t^{2/l}/n^{2/l}, X_j \ne X_i\right\}, \\ N_n^{(i)}(t,\lambda,2) &= \#\left\{Y_j : \|Y_j - a_*(X_i,\lambda)\|_2^2 \le t^{2/l}/n^{2/l}\right\}. \end{aligned}$$

For any X_j with $j \ne i$, conditional on X_i, due to the assumption of density and the formula for the volume of $l-$*dimensional* ball (Rudin [1964]), we have,

$$\begin{aligned} &\mathbb{P}\left[\|X_j - a_*(X_i,\lambda)\|_2^2 \le t^{2/l}/n^{2/l} \middle| X_i\right] \\ &= f_X(a_*(X_i,\lambda)) \frac{\pi^{l/2}}{\Gamma(l/2+1)}\frac{t}{n} + o_p(t/n) = f_X(X_i + \lambda/2)\frac{\pi^{l/2}}{\Gamma(l/2+1)}\frac{t}{n} + o_p(t/n). \end{aligned}$$

Similarly,

$$\mathbb{P}\left[\left\|\tilde{X}_j - a_*(X_i,\lambda)\right\|_2^2 \le t^{2/l}/n^{2/l} \middle| X_i\right] = f_{\tilde{X}}(X_i + \lambda/2)\frac{\pi^{l/2}}{\Gamma(l/2+1)}\frac{t}{n} + o_p(t/n).$$

Since we have i.i.d. structure for the data points, thus we know, $N_n^{(i)}(t,\lambda,1)$ and $N_n^{(i)}(t,\lambda,2)$ conditional on X_i follow binomial distributions,

$$\begin{aligned} N_n^{(i)}(t,\lambda,1)|X_i &\sim \text{Bin}\left(f_X(X_i + \lambda/2)\frac{\pi^{l/2}}{\Gamma(d/2+1)}\frac{t}{n} + o_p(t/n), n-1\right), \\ N_n^{(i)}(t,\lambda,2)|X_i &\sim \text{Bin}\left(f_{\tilde{X}}(X_i + \lambda/2)\frac{\pi^{l/2}}{\Gamma(l/2+1)}\frac{t}{n} + o_p(t/n), [\kappa n]\right), \\ N_n^{(i)}(t,\lambda) &= N_n^{(i)}(t,\lambda,1) + N_n^{(i)}(t,\lambda,2). \end{aligned}$$

Moreover, we have as $n \to \infty$,

$$f_X\left(X_i + \lambda/2\right) \frac{\pi^{l/2}}{\Gamma(l/2+1)} \frac{t}{n} \times (n-1) \to f_X\left(X_i + \lambda/2\right) \frac{\pi^{l/2}}{\Gamma(l/2+1)} t.$$

Thus, by Poisson approximation to binomial distribution, we have the weak convergence result

$$N_n^{(i)}(\cdot, \lambda, 1) | X_i \Rightarrow \text{Poi}\left(f_X\left(X_i + \lambda/2\right) \frac{\pi^{l/2}}{\Gamma(l/2+1)} \cdot \right),$$

in $D[0, \infty)$.

So we have that $N_n^{(i)}(\cdot, \lambda, 1)$, conditional on X_i, is asymptotically a time homogeneous Poisson process with rate $f_X\left(X_i + \lambda/2\right) \pi^{d/2} / \Gamma(d/2+1)$. Similar considerations apply to $N_n^{(i)}(\cdot, \lambda, 2) | X_i$ which yield that

$$N_n^{(i)}(\cdot, \lambda) | X_i \Rightarrow \text{Poi}\left(\Lambda\left(X_i, \lambda\right) \cdot\right),$$

where

$$\Lambda\left(X_i, \lambda\right) = \left[f_X\left(X_i + \lambda/2\right) + \kappa f_{\tilde{X}}\left(X_i + \lambda/2\right) \right] \frac{\pi^{l/2}}{\Gamma(l/2+1)}.$$

Let us write $T_i(n)$ to denote the first arrival time of $N_n^{(i)}(\cdot, \lambda)$, that is,

$$T_i(n) = \inf\left\{ t \geq 0 : N_n^{(i)}(t, \lambda) \geq 1 \right\}$$

Then, we can specify the survival function for $T_i(n)$ to be:

$$\mathbb{P}\left[T_i(n) > t \mid X_i\right] = \mathbb{P}\left[N_n^{(i)}(t, \lambda) = 0 \;\middle|\; X_i \right] = \exp\left(-\Lambda\left(X_i, \lambda\right) t\right) \left(1 + O\left(1/n^{1/l}\right)\right), \tag{3.18}$$

uniformly on t over compact sets. The error rate $O\left(1/n^{1/l}\right)$ is obtained by a simple Taylor expansion of the exponential function applied to the middle term in the pre-

vious string of equalities. Motivated by the form in the right hand side of (3.18) we define $\tau_i(X_i)$ to be a random variable such that

$$\mathbb{P}[\tau_i(X_i) > t | X_i] = \exp(-\Lambda(X_i, \lambda)\, t),$$

and we drop the dependence on X_i and the subindex i when we refer to the unconditional version of $\tau_i(X_i)$, namely

$$\mathbb{P}[\tau > t] = \mathbb{E}[\exp(-\Lambda(X_1, \lambda)\, t)].$$

We finish Step 2 with the statement of two technical lemmas. The first provides a rate of convergence for the Glivenko-Cantelli theorem associated to the sequence $\{T_i(n)\}_{i=1}^n$.

Lemma 3.1. For any $T \in (0, \infty)$ (deterministic) and $\alpha \in (0, 2]$, we have that

$$\overline{\lim}_{n\to\infty} \mathbb{E}\left(\sup_{t\in[0,T]} \left| \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left(I(T_i(n) \leq t) - \mathbb{P}[T_i(n) \leq t] \right) \right| \right) < \infty,$$

and

$$\overline{\lim}_{n\to\infty} \mathbb{E}\left(\sup_{t\in[0,T]} \left| \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left(\max\left(t^2 - T_i(n)^\alpha, 0\right) - \mathbb{E}\left[\max\left(t^2 - T_i(n)^\alpha, 0\right)\right] \right) \right| \right) < \infty.$$

The second technical lemma deals with local properties of the distribution of $T_i(n)$. The proofs of both of these technical results are given at the end of the proof of Theorem 3.1, in Section 3.3.2.7.

Lemma 3.2. For $X_i \in \mathbb{R}^l$ and any finite t, we have the Poisson approximation to binomial as:

$$\mathbb{P}[T_i(n) \leq t] - \mathbb{P}[\tau \leq t] = O(t^{1+1/l}/n^{1/l}),$$

and

$$\mathbb{P}\left[T_i\left(n\right) \leq t\right] - \mathbb{P}\left[\tau \leq t\right] = \mathbb{P}\left[\tau > t\right] O\left(1/n^l\right).$$

3.3.2.3 Step 3 (Closest Point and SoS Function Simplification):

Note that the i-th constraint, namely,

$$-\gamma_i \leq \lambda^T X_i + \inf_j \left\{-\lambda^T Z_j + \left\|X_i - Z_j\right\|_2^2\right\},$$

can be written as

$$\begin{aligned}
-\gamma_i &\leq \inf_j \left\{-\lambda^T (Z_j - X_i) + \left\|X_i - Z_j\right\|_2^2\right\} \\
&= -\left\|\lambda\right\|_2^2/4 + \inf_j \left\{\left\|Z_j - (\lambda/2 + X_i)\right\|_2^2\right\} \\
&= -\left\|\lambda\right\|_2^2/4 + T_i^{2/l}\left(n\right)/n^{2/l}.
\end{aligned}$$

However, since $\gamma_i \geq 0$ we must have that

$$-\gamma_i \leq -\left\|\lambda\right\|_2^2/4 + \min\left(T_i^{2/l}\left(n\right)/n^{2/l}, \left\|\lambda\right\|_2^2/4\right).$$

Therefore, the SoS profile function takes the form

$$R_n(\theta_*) = \max_{\lambda} \left\{-\lambda^T \bar{X}_n - \left\|\lambda\right\|_2^2/4 + \frac{1}{n}\sum_{i=1}^{n} \min\left(\frac{T_i^{2/l}\left(n\right)}{n^{2/l}}, \left\|\lambda\right\|_2^2/4\right)\right\}.$$

To simplify the notation, let us redefine $\lambda \longleftarrow 2\lambda$ then we have that the simplified SoS profile function becomes:

$$R_n(\theta_*) = \max_{\lambda} \left\{-2\lambda^T \bar{X}_n - \frac{1}{n}\sum_{i=1}^{n} \max\left(\left\|\lambda\right\|_2^2 - \frac{T_i^{2/l}\left(n\right)}{n^{2/l}}, 0\right)\right\}. \tag{3.19}$$

3.3.2.4 Step 4 (Case $l = 1$):

When $l = 1$, let's denote $\sqrt{n}\bar{X}_n = Z_n$ and $\sqrt{n}\lambda = \zeta$, where by CLT we can show $Z_n \Rightarrow \tilde{Z} \sim N(0, \sigma^2)$, where when $l = 1$ we have $\sigma^2 = \Sigma$. Then, as $n \to \infty$, we have:

$$\begin{aligned} nR_n(\theta_*) &= \max_{\zeta} \left\{ -2\zeta Z_n - \frac{1}{n} \sum_{i=1}^{n} \max\left(\zeta^2 - T_i^2(n)\, n^{-1}, 0\right) \right\} \\ &= \max_{\zeta} \left\{ -2\zeta Z_n - \mathbb{E}\left[\max\left(\zeta^2 - T_i^2(n)\, n^{-1}, 0\right)\right] \right\} + o_p(1) \end{aligned}$$

The second equation follows the estimate in (Lemma 3.1). We know the objective function as a function of ζ is a strictly convex function. Since as $\zeta = b\,|Z_n|$ with $b \to \pm\infty$ implies that the objective function will tend to $-\infty$, we conclude that the sequence of global optimizers is compact and each optimizer (i.e. for each n) could be characterized by the first order optimality condition almost surely. To make the analysis more clear, let us denote the expectation in the maximization problem to be $g(\zeta, n)$, as a function of ζ, i.e.

$$G(\zeta, n) = \mathbb{E}\left[\max\left(\zeta^2 - T_i^2(n)\, n^{-1}, 0\right)\right],$$

which is a deterministic function of ζ and for any n it is convex. Moreover, the derivative of $G(\zeta, n)$ is,

$$g(\zeta, n) = \nabla_\zeta G(\zeta, n) = 2\zeta \mathbb{P}\left(T_i(n) \leq n\zeta^2\right).$$

We need to notice that while taking the derivative we require exchanging the derivative and expectation, this can be done true hereby the dominating convergence the-

orem since

$$\delta^{-1}\left|\max\left(\left(\zeta+\delta\right)^{2}-T_{i}^{2}\left(n\right)n^{-1},0\right)-\max\left(\zeta^{2}-T_{i}^{2}\left(n\right)n^{-1},0\right)\right|\leq 2|\zeta|,$$

for all $\delta > 0$. We can take the derivative with respect to ζ in $-2\zeta Z_n - G\left(\zeta, n\right)$ and set it to zero, as $n \to \infty$ we obtain

$$Z_n = -\zeta P\left(T_i(n) \leq n\zeta^2\right) = -\zeta P\left(\tau \leq n\zeta^2\right) + o_p(1) = -\zeta + o_p(1).$$

This estimate follows the second result of Lemma 3.2. Therefore, the optimizer ζ_n^*, satisfies $\zeta_n^* = -Z_n + o_p(1)$, as $n \to \infty$. Then, we plug it into the objective function to obtain that the scaled SoS profile function satisfies

$$nR_n(\theta_*) = 2Z_n^2 - G\left(Z_n, n\right) + o_p\left(1\right) \text{ as } n \to \infty.$$

We should notice $G\left(Z_n, n\right)$ is a function defined via expectation and evaluated at Z_n, thus it is a random variable depends on Z_n. By definition and $E\left[|X|\right] = \int_0^\infty \mathbb{P}\left[|X| \geq t\right] dt$, we know as $n \to \infty$,

$$\begin{aligned} G\left(\zeta, n\right) &= \int_0^{\zeta^2} \mathbb{P}\left[T_i^2\left(n\right) \leq n\left(\zeta^2 - t\right)\right] dt \\ &= \int_0^{\zeta^2} \mathbb{P}\left[\tau^2\left(n\right) \leq n\left(\zeta^2 - t\right)\right] dt + o(1) \\ &= \int_0^{\zeta^2} 1dt + o(1) = \zeta^2 + o(1), \end{aligned}$$

where the second equality is derived from the second argument of Lemma 3.2. Then for the SoS profile function, it becomes,

$$nR_n(\theta_*) = 2Z_n^2 - Z_n^2 + o_p(1) = Z_n^2 + o_p(1) \text{ as } n \to \infty.$$

It is easy to see by applying continuous mapping theorem and central limitation for Z_n, we have

$$nR_n(\theta_*) \Rightarrow \sigma^2 \chi_1^2.$$

3.3.2.5 Step 5 (Case $l = 2$):

Once again we introduce the substitution $\zeta = \sqrt{n}\lambda$ and $\sqrt{n}\bar{X}_n = Z_n$ into (3.19). Then, scaling the profile function by n, as $n \to \infty$ we have

$$\begin{aligned} nR_n(\theta_*) =& \max_{\zeta} \left\{ -2\zeta^T Z_n - \frac{1}{n} \sum_{i=1}^{n} \max \left(\|\zeta\|_2^2 - T_i(n), 0 \right) \right\} \\ =& \max_{\zeta} \left\{ -2\zeta^T Z_n - \mathbb{E} \left[\max \left(\|\zeta\|_2^2 - T_i(n), 0 \right) \right] \right\} + o_p(1), \end{aligned} \tag{3.20}$$

where the previous estimate follows by applying Lemma 3.1 (the error is obtained by localizing ζ on a compact set, which is valid because the sequence of global optimizers is easily seen to be tight). The objective function is strictly convex as a function of ζ and we know when $\|\zeta\|_2 \to \infty$ the objective function tends to $-\infty$, thus each global maximizer (for each n) can be characterized by the first order optimality condition almost surely. Similar as Case $l = 1$, let us denote

$$G(\zeta, n) = \mathbb{E} \left[\max \left(\|\zeta\|_2^2 - T_i(n), 0 \right) \right].$$

It is a continuous differentiable and convex function in ζ and with derivative equals

$$g(\zeta, n) = \nabla_\zeta G(\zeta, n) = 2\zeta \mathbb{P}\left[\|\zeta\|_2^2 \geq T_i(n)\right] = 2\zeta \mathbb{P}\left[\|\zeta\|_2^2 \geq \tau\right] + o(1) \text{ as } n \to \infty,$$

where the first equality requires applying dominating convergence theorem as for $l = 1$ and second estimate follows the first argument in Lemma 3.2. Combining the above estimation, we have the first order optimality condition becomes

$$Z_n = -\zeta \mathbb{P}\left[\|\zeta\|_2^2 \geq \tau\right] + o_p(1) = -\zeta \tilde{g}(\zeta) + o_p(1) \text{ as } n \to \infty, \tag{3.21}$$

where $\tilde{g}(\zeta) = \mathbb{P}\left[\|\zeta\|_2^2 \geq \tau\right]$ is a deterministic function of ζ. Using equation (3.21), we conclude that the optimizer ζ_n^*, satisfies $\zeta_n^* = -\rho Z_n + o_p(1)$, for some ρ. In turn, plugging in this representation into equation (3.21), as $n \to \infty$ we have

$$\|\zeta_n^*\|_2 \tilde{g}(\zeta_n^*) + o_p(1) = \|Z_n\|_2 .$$

Sending $n \to \infty$, we conclude that ρ is the unique solution to

$$\frac{1}{\rho} = \tilde{g}\left(\rho \tilde{Z}\right). \tag{3.22}$$

Since the objective function is strict convex and the above equation is derived from first order optimality condition, we know the solution exists and is unique (alternatively we can use the continuity and monotonicity of left and right hand side of (3.22), to argue the existence and uniqueness). Let us plug in the optimizer back to the objective function and we can see the scaled SoS profile function becomes

$$nR_n(\theta_*) = 2\rho \left(\left\|\tilde{Z}\right\|_2^2\right) \|Z_n\|_2^2 - G(\zeta_n^*, n) + o_p(1).$$

For a positive random variable Y, we have: $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}[Y \geq t]\, dt$. Therefore, for ζ in a compact set, as $n \to \infty$ we have the following estimate

$$\begin{aligned} G(\zeta, n) &= \int_0^{\|\zeta\|_2^2} \mathbb{P}\left[\|\zeta\|_2^2 - T_i(n) \geq t\right] dt \\ &= \int_0^{\|\zeta\|_2^2} \mathbb{P}\left[\|\zeta\|_2^2 - \tau \geq t\right] dt + o(1) \\ &= \|\zeta\|_2^2 \int_0^1 \mathbb{P}\left[1 - \tau/\|\zeta\|_2^2 \geq s\right] ds + o(1) \\ &= \|\zeta\|_2^2\, \mathbb{E}\left[\max\left(1 - \tau/\|\zeta\|_2^2, 0\right)\right] + o(1) \\ &= \|\zeta\|_2^2\, \tilde{\eta}(\zeta) + o(1), \end{aligned}$$

where we define $\tilde{\eta}(\zeta) = \mathbb{E}\left[\max\left(1 - \tau/\|\zeta\|_2^2, 0\right)\right]$ is a deterministic continuous function of ζ. The second equation follows the first result of Lemma 3.2. Finally combine $G(\zeta, n)$ and the first term, using the CLT and continuous mapping theorem, where we denote $Z_n \Rightarrow \tilde{Z} \sim N(0, Var(X))$, as $n \to \infty$ we have:

$$\begin{aligned} nR_n(\theta_*) &= 2\rho\left(\tilde{Z}\right)\|Z_n\|_2^2 - \rho\left(\tilde{Z}\right)^2 \tilde{\eta}(Z_n)\|Z_n\|_2^2 + o_p(1) \\ &\Rightarrow 2\rho\left(\tilde{Z}\right)\left\|\tilde{Z}\right\|_2^2 - \rho\left(\tilde{Z}\right)^2 \tilde{\eta}\left(\tilde{Z}\right)\left\|\tilde{Z}\right\|_2^2. \end{aligned}$$

3.3.2.6 Step 6 (Case $l \geq 3$):

For simplicity, let us write $\sqrt{n}\bar{X}_n = Z_n$ and $n^{\frac{3}{2l+2}}\lambda = \zeta$, then as $n \to \infty$ we have

$$
\begin{aligned}
&n^{1/2+\frac{3}{2l+2}} R_n(\theta_*) \\
&= \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} \frac{1}{n} \sum_{i=1}^{n} \max\left(\left\| \frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}} \right\|_2^2 - T_i^{2/l}(n), 0 \right) \right\} \\
&= \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} \mathbb{E}\left[\max\left(\left\| \frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}} \right\|_2^2 - T_1^{2/l}(n), 0 \right) \right] \right\} + o_p(1).
\end{aligned}
$$

The estimate in the previous display is due to an application of Lemma 3.1. Similar as for the lower dimensional case, let us denote

$$
G(\zeta, n) = n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} \mathbb{E}\left[\max\left(\left\| \frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}} \right\|_2^2 - T_1^{2/l}(n), 0 \right) \right],
$$

being a deterministic function continuous and differentiable as a function of ζ. As we discussed for the case $l = 2$ case, the objective function is strictly convex in ζ, the global optimizers are not only tight, but each optimizer is also characterized by first order optimality conditions almost surely. We can apply the dominating convergence as we discussed for $l = 1$ and the gradient of $G(\zeta, n)$ has the following estimate as $n \to \infty$,

$$
\begin{aligned}
g(\zeta, n) = \nabla_\zeta G(\zeta, n) &= 2n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} \zeta \mathbb{P}\left[T_i(n) \leq \left\| \zeta n^{-\left(\frac{3}{2l+2}-\frac{1}{l}\right)} \right\|_2^l \right] \\
&= 2n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} \zeta \mathbb{P}\left[\tau(n) \leq \left\| \zeta n^{-\left(\frac{3}{2l+2}-\frac{1}{l}\right)} \right\|_2^l \right] + o(1).
\end{aligned}
$$

The second equality estimate is considering ζ within a compact set and the derivation follows the first argument in Lemma 3.2. Then the first order optimality condition

for the SoS profile function becomes,

$$Z_n = -n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)}\zeta\mathbb{P}\left[\tau\left(n\right) \leq \left\|\zeta n^{-\left(\frac{3}{2l+2}-\frac{1}{l}\right)}\right\|_2^l\right] + o(1) \text{ as } n \to \infty.$$

For notation simplicity, let us define

$$\kappa_n = \zeta n^{-\left(\frac{3}{2l+2}-\frac{1}{l}\right)}.$$

We can observe for ζ in a compact set, $\left\|\zeta n^{-\left(\frac{3}{2l+2}-\frac{1}{l}\right)}\right\|_2^l = \|\kappa_n\|_2^l \to 0$, as $n \to \infty$, then we can write

$$\begin{aligned}
&\mathbb{P}\left[\tau \leq \ \|\kappa_n\|_2^l\right] = 1 - \mathbb{P}\left[\tau > \|\kappa_n\|_2^l\right] = 1 - \mathbb{E}\left[\mathbb{P}\left[\tau > \|\kappa_n\|_2^l \ \middle|\ X_1\right]\right] \\
&= \mathbb{E}\left[1 - \exp\left(-\frac{\pi^{l/2}\left(f_X\left(X_1+\kappa_n\right) + f_{\tilde{X}}\left(X_1+\kappa_n\right)\right)}{\Gamma(l/2+1)}\|\kappa_n\|_2^l\right)\right] \\
&= \mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left[f_X\left(X_1\right) + f_{\tilde{X}}\left(X_1\right)\right]\|\kappa_n\|_2^l\right] + o_p\left(n^{-\left(\frac{3l}{2l+2}-1\right)}\right) \\
&= C\|\kappa_n\|_2^l + o_p\left(n^{-\left(\frac{3l}{2l+2}-1\right)}\right),
\end{aligned}$$

where we denote

$$C = \frac{\pi^{l/2}}{\Gamma(l/2+1)}\mathbb{E}\left[f_{\tilde{X}}\left(X_1\right) + f_Y\left(X_1\right)\right].$$

Plug it back into the optimizer, and as $n \to \infty$ we have:

$$Z_n = -Cn^{\left(1/2-\frac{3}{2l+2}\right)}n^{\left(-\frac{3l}{2l+2}+1\right)}\zeta\|\zeta\|_2^l + o_p(1) = -C\zeta\|\zeta\|_2^l + o_p(1).$$

We know that within the objective function, the second term is only based on the L_2 norm of ζ, thus to maximize the objective function we will asymptotically select $\zeta_n^* = -c_* Z_n\left(1 + o\left(1\right)\right)$, where $c_* > 0$ is suitably chosen, thus, we conclude that the

optimizer takes the form,

$$\zeta_n^* = -Z_n \|Z_n\|_2^{\left(\frac{1}{l+1}-1\right)} / C^{\frac{1}{l+1}} + o_p(1).$$

Plugging-in the optimizer back into the objective function, as $n \to \infty$we have:

$$n^{1/2+\frac{3}{2l+2}} R_n(\theta_*) = -2{\zeta_n^*}^T Z_n - G\left(\zeta_n^*, n\right) + o_p(1).$$

Let us focus on the analysis of $G\left(\zeta, n\right)$ in a compact set. By definition, we can notice that inside the previous expectation there is a positive random variable bounded by $\left\| \frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}} \right\|_2^2 = \|\kappa_n\|_2^2$, thus as $n \to \infty$ we have the following estimate for the expectation as.

$$\begin{aligned}
&\mathbb{E}\left[\max\left(\|\kappa_n\|_2^2 - T_1^{2/l}\left(n\right), 0\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\max\left(\|\kappa_n\|_2^2 - T_1^{2/l}\left(n\right), 0\right) \;\middle|\; X_1\right]\right] \\
&= \mathbb{E}\left[\int_0^{\kappa_n} \mathbb{P}\left[T_1\left(n\right) \leq \left(\kappa_n - t\right)^{l/2} \;\middle|\; X_1\right] dt\right] \\
&= \mathbb{E}\left[\int_0^{\|\kappa_n\|_2^2} \mathbb{P}\left[\tau \leq \left(\|\kappa_n\|_2^2 - t\right)^{l/2} \;\middle|\; X_1\right] + O\left(1/n^{-1/2+1/l}\right) dt\right] \\
&= \mathbb{E}\left[\int_0^{\|\kappa_n\|_2^2} \left(1 - e^{-\frac{\pi^{l/2}\left(f_X\left(X_1 + \frac{\zeta}{n^{\frac{3}{2l+2}}}\right) + f_{\tilde{X}}\left(X_1 + \frac{\zeta}{n^{\frac{3}{2l+2}}}\right)\right)}{\Gamma(l/2+1)}\left(\|\kappa_n\|_2^2 - t\right)^{l/2}}\right) dt\right] \\
&\qquad + O\left(1/n^{-1/2+3/l-\frac{6}{2l+2}}\right) \\
&= C\frac{2}{l+2}\left\| \frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}} \right\|^{l+2} + O\left(1/n^{-1/2+3/l-\frac{6}{2l+2}}\right)
\end{aligned}$$

The estimate in third equation follows by applying the first argument in Lemma 3.2. The final equality estimate is due to $\|\kappa_n\|_2^2 = \left\|\zeta n^{-\left(\frac{3}{2l+2}-\frac{1}{l}\right)}\right\|_2^2 \to 0$ as $n \to \infty$. Then,

owing to the previous results, as $n \to \infty$ we have estimate for $G(\zeta, n)$ as

$$\begin{aligned} G(\zeta, n) &= -\frac{2C}{l+2} n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} n^{\left(-\frac{3l+6}{2l+2}+\frac{l+2}{l}\right)} \|\zeta\|_2^{l+2} + o(1) \\ &= -\frac{2C}{l+2} \|\zeta\|_2^{l+2} + o(1). \end{aligned}$$

Finally, we can know that, as $n \to \infty$, by the CLT we have $Z_n \Rightarrow \tilde{Z}$, then using continuous mapping theorem, we have that the scaled SoS profile function has the asymptotic distribution given by

$$\begin{aligned} n^{1/2+\frac{5}{4l+2}} R_n(\theta_*) &= 2\|Z_n\|_2^2 \frac{\|Z_n\|_2^{\left(\frac{1}{l+1}-1\right)}}{C^{\frac{1}{l+1}}} - \frac{2}{l+2} \frac{\|Z_n\|_2^{1+\frac{1}{l+1}}}{C^{\frac{1}{l+1}}} + o_p(1) \\ &= \frac{2l+2}{l+2} \frac{\|Z_n\|_2^{1+\frac{1}{l+1}}}{C^{\frac{1}{l+1}}} + o_p(1) \Rightarrow \frac{2l+2}{l+2} \frac{\left\|\tilde{Z}\right\|_2^{1+\frac{1}{l+1}}}{C^{\frac{1}{l+1}}}. \end{aligned}$$

3.3.2.7 Proofs of Technical Lemmas in Step 2

Proof of Lemma 3.1. We shall introduce some notation which will be convenient throughout our development. Define for $t \geq 0$,

$$\begin{aligned} F_n(t) &= P(T_i(n) \leq t), \\ D_i(t) &= I(T_i(n) \leq t), \quad \bar{D}_i(t) = I(T_i(n) \leq t) - F_n(t), \\ \bar{F}_n(t) &= 1 + n^{-1/2} \sum_{i=1}^{n} \bar{D}_i(t). \end{aligned}$$

Therefore, we are interested in studying

$$\bar{F}_n(t) - 1 = \frac{1}{n^{1/2}} \sum_{i=1}^{n} (I(T_i(n) \leq t) - F_n(t)).$$

We will start by studying

$$\mathbb{E}[\sup\{\bar{F}_n(t) : t \in [0,T]\}].$$

First, we define

$$h_n(t) = \frac{\bar{F}_n(t_-)}{\left(\bar{F}_n^*(t_-)^2 + [\bar{F}_n](t_-)\right)^{1/2}},$$

where, for a given function $\{g(t) : t \in [0,T]\}$, we define

$$g^*(t) = \sup\{g(s) : s \in [0,t]\},$$
$$[g](t) = \int_0^t (dg(s))^2.$$

In addition, $[g](t)$ is defined as the quadratic variational process, i.e.,

$$[g](t) = \lim_{n\to\infty} \sum_{i=1}^{n} \left[g\left(\frac{i \times t}{n}\right) - g\left(\frac{(i-1) \times t}{n}\right)\right]^2.$$

In particular,

$$[\bar{F}_n](t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i(n) \leq t).$$

We observe that $\bar{F}_n^*(t) \geq 1$, therefore $h_n(t)$ is well defined; moreover, note that

$$h_n(t)^2 \leq 1.$$

We invoke Theorem 1.2 of Beiglböck and Siorpaes [2015] and conclude that

$$\sup_{0\leq t\leq T} \bar{F}_n(t) \leq 6\sqrt{[\bar{F}_n](T)} + 2\int_0^T h_n(t)\, d\bar{F}_n(t).$$

Now we analyze the integral in the right hand side of the previous display. Observe

that

$$\mathbb{E}\left(\int_0^T h_n(t)\, d\bar{F}_n(t)\right) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \mathbb{E}\left(\int_0^T h_n(t)\, d\bar{D}_i(t)\right)$$
$$= n^{1/2}\mathbb{E}\left(\int_0^T h_n(t)\, d\bar{D}_1(t)\right). \tag{3.23}$$

Let us write

$${}_1\bar{F}_n(t) = \bar{F}_n(t) - \bar{D}_1(t)/n^{1/2},$$

that is, we simply remove the last term in the sum defining $\bar{F}_n(t)$. We have that

$$h_n(t) = \frac{{}_1\bar{F}_n(t_-) + \bar{D}_1(t_-)/n^{1/2}}{\left(\bar{F}_n^*(t_-)^2 + [{}_1\bar{F}_n](t_-) + [D_1](t_-)/n\right)^{1/2}},$$

moreover,

$$\left|{}_1\bar{F}_n^*(t) - \bar{F}_n^*(t)\right| \leq 1/n^{1/2}.$$

We then can write

$$h_n(t) = \frac{{}_1\bar{F}_n(t_-) + \bar{D}_1(t_-)/n^{1/2}}{\left(\bar{F}_n^*(t_-)^2 + [{}_1\bar{F}_n](t_-) + [D_1](t_-)/n\right)^{1/2}} \tag{3.24}$$
$$= \frac{{}_1\bar{F}_n(t_-)}{\left({}_1\bar{F}_n^*(t_-)^2 + [{}_1\bar{F}_n](t_-)\right)^{1/2}}\left(1 + \frac{L_n(t_-)}{n^{1/2}}\right),$$

where we can select a deterministic constant $c \in (0,\infty)$ such that $|L_n(t)| \leq c$ for $j = 0$ and 1 assuming $n \geq 4$ (this constrain in n is imposed so that a Taylor expansion for the function $1/(1-x)$ can be developed for $x \in (0,1)$). We now insert (3.24) into (3.23) and conclude that if we define

$$\bar{h}_n(t) = \frac{{}_1\bar{F}_n(t_-)}{\left({}_1\bar{F}_n^*(t_-)^2 + [{}_1\bar{F}_n](t_-)\right)^{1/2}},$$

it suffices to verify that

$$n^{1/2}\mathbb{E}\left(\int_0^T \bar{h}_n(t)\, d\bar{D}_1(t)\right) < \infty.$$

Define $\widetilde{h}_n(t)$ to be a copy of $\bar{h}_n(t)$, independent of X_1 and $T_1(n)$. In particular, $\widetilde{h}_n(t)$ is constructed by using all of the X_j's except for X_1, which might be replaced by an independent copy, X_1', of X_1. Observe that the number of processes $\{\bar{D}_i(t) : t \leq T\}$ that depend on $T_1(n)$ and X_1 is smaller than $N_n(T, \lambda, 1)$. Therefore, similarly as we obtained from the analysis leading to the definition of $\bar{h}_n(\cdot)$, we have that a random variable $\bar{L}_{N_n(T,\lambda,1)}$ can be defined so that $\left|\bar{L}_{N_n(T,\lambda,1)}\right| \leq c(1 + N_n(T, \lambda, 1))$ for some (deterministic) $c > 0$ and $n \geq 4$ and satisfying

$$\begin{aligned}
&\mathbb{E}\left(\int_0^T \bar{h}_n(t)\, d\bar{D}_1(t)\right) \\
&= \mathbb{E}\left(\bar{h}_n(T_1(n))\, I(T_1(n) \leq T)\right) - \mathbb{E}\left(\widetilde{h}_n(T_1(n))\, I(T_1(n) \leq T)\right) \\
&= \mathbb{E}\left(\widetilde{h}_n(T_1(n))\, I(T_1(n) \leq T)\right) - \mathbb{E}\left(\widetilde{h}_n(\tau_i(X_i))\, I(\tau_i(X_i) \leq T)\right) \\
&+ \mathbb{E}\left(\bar{L}_{N_n(T,\lambda,1)}/n^{1/2}\right) \\
&= \mathbb{E}\left(\bar{L}_{N_n(T,\lambda,1)}/n^{1/2}\right).
\end{aligned}$$

We have that

$$\left|\mathbb{E}\left(\bar{L}_{N_n(T,\lambda,1)}/n^{1/2}\right)\right| \leq \left|\mathbb{E}\left(c(1 + N_n(T, \lambda, 1))\right)\right|/n^{1/2} = O\left(1/n^{1/2}\right).$$

Consequently, we conclude that

$$n^{1/2}\mathbb{E}\left(\int_0^T h_n(t)\, d\bar{D}_1(t)\right) = O(1),$$

as $n \to \infty$, as required. Thus we proved that the first part of the lemma holds. For the second part, we observe that

$$\begin{aligned}
&\overline{\lim}_{n\to\infty}\mathbb{E}\left(\sup_{t\in[0,T]}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(\max\left(t^2-T_i(n)^{\alpha},0\right)-\mathbb{E}\left[\max\left(t^2-T_i(n)^{\alpha},0\right)\right]\right)\right|\right)\\
&=\overline{\lim}_{n\to\infty}\mathbb{E}\left(\sup_{t\in[0,T]}\left|\int_0^t\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(2sI\left(T_i^{\alpha}(n)\leq s^2\right)-2s\mathbb{P}[T_i^{\alpha}(n)\leq s^2]\right)ds\right|\right)\\
&\leq\overline{\lim}_{n\to\infty}\int_0^T\mathbb{E}\left(\sup_{t\in[0,T]}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(2tI\left(T_i^{\alpha}(n)\leq t^2\right)-2t\mathbb{P}[T_i^{\alpha}(n)\leq t^2]\right)\right|\right)dt\\
&\leq 2T^2\overline{\lim}_{n\to\infty}\mathbb{E}\left(\sup_{t\in[0,T]}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(I\left(T_i(n)\leq t\right)-\mathbb{P}[T_i(n)\leq t]\right)\right|\right)<\infty.
\end{aligned}$$

Hence, applying the result for the first part of the lemma, we conclude the second part as well. □

Proof of Lemma 3.2.

$$\begin{aligned}
\mathbb{P}\left[T_i(n)\leq t\right]&=\mathbb{P}\left(Bin\left(\mathbb{P}\left(\left\|X_i-a\left(X_i,\lambda\right)\right\|_2\leq t^{1/l}/n^{1/l}\right),n-1\right)\geq 1\right)\\
&=1-\left(1-\mathbb{P}\left(\left\|X_i-a\left(X_i,\lambda\right)\right\|_2\leq t^{1/l}/n^{1/l}\right)\right)^n.
\end{aligned}$$

Then, as $n\to\infty$ and $t\to 0^+$

$$\mathbb{P}\left(\left\|X_j-a\left(X_i,\lambda\right)\right\|_2\leq t^{1/l}/n^{1/l}\right)=c_0t/n+c_1t/n\cdot t^{1/l}/n^{1/l}+o\left(t^{1+1/l}/n^{1+1/l}\right).$$

Therefore by the Poisson approximation to the Binomial distribution we know:

$$\begin{aligned}
\mathbb{P}\left[T_i(n)\leq t\right]&=1-\exp\left(-c_0t\right)+O\left(t^{1+1/l}/n^{1/l}\right),\\
\mathbb{P}\left[\tau\leq t\right]&=1-\exp\left(-c_0t\right).
\end{aligned}$$

Thus we proved the first claim:

$$\mathbb{P}\left[T_i\left(n\right) \leq t\right] - \mathbb{P}\left[\tau \leq t\right] = O\left(t^{1+1/l}/n^{1/l}\right).$$

The second claim follows the definition of τ and equation Equation (3.18), where as $n \to \infty$ we have

$$\begin{aligned} & \mathbb{P}\left[T_i\left(n\right) \leq t\right] - \mathbb{P}\left[\tau \leq t\right] = \mathbb{P}\left[T_i\left(n\right) > t\right] - \mathbb{P}\left[\tau > t\right] \\ = \; & \mathbb{E}\left[\exp\left(-\Lambda\left(\lambda, X_1\right)\right)\right]\left(1 + O\left(1/n^l\right)\right) - \mathbb{E}\left[\exp\left(-\Lambda\left(\lambda, X_1\right)\right)\right] \\ = \; & \mathbb{P}\left[\tau > t\right] O\left(1/n^l\right). \end{aligned}$$

□

3.3.3 Proofs of Additional Theorems

In this subsection, we are going to provide the proofs of the remaining theorems and corollaries (Theorem 3.2, Theorem 3.3, Corollary 3.1 and Corollary 3.2). We are going to follow closely the proof of Theorem 3.1 and discuss the differences inside each of its steps.

3.3.3.1 Proofs of SoS Theorems for General Estimation

We will first prove the corresponding theorems for general estimating equations. As we discussed before, Theorem 3.2 is the direct generalization of Theorem 3.1 and we are going to only discuss the proof of Theorem 3.3 in this part.

Proof of Theorem 3.3. Let us first denote $\bar{h}_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} h\left(\theta, X_i\right)$. The analogue

of **Step 1**, namely, the dual formulation takes the form

$$\begin{aligned} & R_n(\theta_*) \\ &= \max_{\lambda} \left\{ -\lambda^T \bar{h}_n(\theta_*) - \frac{1}{n}\sum_{i=1}^{n} \max_{j} \left\{ \lambda^T h(\theta_*, Z_j) - \lambda^T h(\theta_*, X_i) - \|X_i - Z_j\|_2^2 \right\}^+ \right\}. \end{aligned}$$

Step 2 and 3 are given as follows, for $l = 1$ and $l = 2$, let us denote $\sqrt{n}\bar{h}_n(\theta_*) = Z_n$ and $\sqrt{n}\lambda = 2\zeta$, we can scale the SOS profile function by n, arriving to

$$\begin{aligned} & nR_n(\theta_*) \\ &= \max_{\zeta} \left\{ -2\zeta^T Z_n - \frac{1}{n}\sum_{i=1}^{n} n \max_{j} \left\{ 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, Z_j) - 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, X_i) - \|X_i - Z_j\|_2^2 \right\}^+ \right\}. \end{aligned}$$

For each i, let us consider the maximization problem

$$\max_{j} \left\{ 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, Z_j) - 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, X_i) - \|X_i - Z_j\|_2^2 \right\}. \tag{3.25}$$

Similar as Step 1 of the proof for Theorem 3.1, we would like to solve the maximization problem (3.25) by first minimizing over z (as a free variable), instead of over j and then quantify the gap. Observe that the uniform bound $\|D_x^2 h(\theta_*, \cdot)\| < \tilde{K}$ stated in BE1) implies that for all n large enough (in particular, $n^{1/2} > 2\tilde{K}\|\zeta\|$) implies that

$$\max_{z} \left\{ 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, z) - 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, X_i) - \|X_i - z\|_2^2 \right\}, \tag{3.26}$$

has an optimizer in the interior. Therefore, by the differentiability assumption stated

in BE1) we know that any global minimizer, $\bar{a}_*(X_i, \zeta)$, of the problem (3.26) satisfies

$$\begin{aligned}\bar{a}_*(X_i,\zeta) &= X_i + D_x h(\theta_*, \bar{a}_*(X_i,\zeta))^T \cdot \frac{\zeta}{n^{1/2}} \\ &= X_i + D_x h(\theta_*, X_i)^T \cdot \frac{\zeta}{n^{1/2}} + O\left(\frac{\|\zeta\|_2^2}{n} \|D_x h(\theta_*, \bar{a}_*(X_i,\zeta))\|_2\right). \quad (3.27)\end{aligned}$$

Moreover, owing to BE1), we obtain that

$$\|D_x h(\theta_*, \bar{a}_*(X_i,\zeta)) - D_x h(\theta_*, X_i)\|_2 \leq \tilde{K} \frac{\|\zeta\|_2}{n^{1/2}}. \quad (3.28)$$

Consequently, if we define

$$a_*(X_i,\zeta) = X_i + D_x h(\theta_*, X_i)^T \cdot \frac{\zeta}{n^{1/2}},$$

we obtain due to (3.27) and (3.28) that

$$\|a_*(X_i,\zeta) - \bar{a}_*(X_i,\zeta)\|_2 = O\left(\frac{\|\zeta\|_2^2}{n}\left(\|D_x h(\theta_*, X_i)\|_2 + \frac{\|\zeta\|_2}{n^{1/2}}\right)\right).$$

Then, after performing a Taylor expansion and applying inequality (3.28) we obtain that

$$\begin{aligned}&2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, X_i) - 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, \bar{a}_*(X_i,\zeta)) + \|X_i - \bar{a}_*(X_i,\zeta)\|_2^2 \\ &= 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, X_i) - 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, a_*(X_i,\zeta)) + \|X_i - a_*(X_i,\zeta)\|_2^2 \\ &\quad + O\left(\frac{\|\zeta\|^3}{n^{3/2}}\right) + O\left(\frac{\|D_x h(\theta_*, X_i)\|_2^2 \|\zeta\|_2^3}{n^{3/2}}\right).\end{aligned}$$

In turn, a direct calculation gives that, as $n \to \infty$

$$-\frac{\zeta^T V_i \zeta}{n} = 2\frac{\zeta^T}{\sqrt{n}} h\left(\theta_*, X_i\right) - 2\frac{\zeta^T}{\sqrt{n}} h\left(\theta_*, a_*\left(X_i, \zeta\right)\right) + \left\|X_i - a_*\left(X_i, \zeta\right)\right\|_2^2 + O\left(\frac{\left\|D_x h\left(\theta_*, X_i\right)\right\|^2 \left\|\zeta\right\|^3}{n^{3/2}}\right).$$

Similarly as in Step 2 of the proof of Theorem 3.1 we can define the point process $N^{(i)}\left(t, \zeta\right)$ and $T_i\left(n\right)$. We know the gap between freeing the variable z and restricting the maximization over the Z_j's (i.e. the difference between (3.26) and (3.25)) is

$$\max_j \left\{\frac{1}{n}\zeta^T V_i \zeta - \left(2\frac{\zeta^T}{\sqrt{n}} h\left(\theta_*, Z_j\right) - 2\frac{\zeta^T}{\sqrt{n}} h\left(\theta_*, X_i\right) - \left\|X_i - Z_j\right\|_2^2\right)\right\} + O\left(\frac{\left\|D_x h\left(\theta_*, X_i\right)\right\|^2 \left\|\zeta\right\|^3}{n^{3/2}}\right).$$

By the definition of $T_i\left(n\right)$, we can write the profile function for $l = 1$ as

$$nR_n(\theta_*) = \max_\zeta \left\{-2\zeta^T Z_n - \frac{1}{n}\sum_{i=1}^n \max\left(\zeta^T V_i \zeta - \frac{T_i^2\left(n\right)}{n} + O\left(\frac{\left\|D_x h\left(\theta_*, X_i\right)\right\|^2 \left\|\zeta\right\|^3}{n^{1/2}}\right), 0\right)\right\}.$$

Note that the sequence of global optimizers is tight as $n \to \infty$ because $\mathbb{E}\left(V_i\right)$ is assumed to be strictly positive definite with probability one. In turn, from the previous expression we obtain, following a similar derivation as in the proof of Theorem 3.1 (invoking Lemma 3.1) and using the fact that ζ can be restricted to compact sets, that as $n \to \infty$

$$nR_n(\theta_*) = \max_\zeta \left\{-2\zeta^T Z_n - \mathbb{E}\left[\max\left(\zeta^T V_1 \zeta - \frac{T_1^2\left(n\right)}{n}\right)\right]\right\} + o_p\left(1\right).$$

Then, for $l = 2$, as $n \to \infty$ we have estimate for the profile function as

$$nR_n(\theta_*) = \max_{\zeta} \left\{ -2\zeta^T Z_n - \mathbb{E}\left[\max\left(\zeta^T V_1 \zeta - T_1^2(n) \right) \right] \right\} + o_p(1).$$

When $l \geq 3$, let us denote $\sqrt{n}\bar{h}_n(\theta_*) = Z_n$ and $n^{\frac{3}{2l+2}}\lambda = 2\zeta$, we can scale profile function by $n^{\frac{1}{2}+\frac{3}{2l+2}}$ and write it as

$$\begin{aligned}
&n^{\frac{1}{2}+\frac{3}{2l+2}} R_n(\theta_*) \\
&= \max_{\zeta} \Big\{ -2\zeta^T Z_n \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} n^{\frac{1}{2}+\frac{3}{2l+2}} \max_j \left\{ 2\frac{\zeta^T}{n^{\frac{3}{2l+2}}} h(\theta_*, Z_j) - 2\frac{\zeta^T}{n^{\frac{3}{2l+2}}} h(\theta_*, X_i) - \|X_i - Z_j\|_2^2 \right\}^+ \Big\}.
\end{aligned}$$

By applying same derivation as for $l = 1$ and 2 above, we can define a point process $N^{(i)}(t, \zeta)$ and $T_i(n)$ as in the proof of Theorem 3.1. As $n \to \infty$, we have the estimate for profile function becomes

$$\begin{aligned}
&n^{\frac{1}{2}+\frac{3}{2l+2}} R_n(\theta_*) \\
&= \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{\frac{1}{2}+\frac{3}{2l+2}-\frac{2}{l}} \frac{1}{n}\sum_{i=1}^{n} \max\left(n^{-\left(\frac{6}{2l+2}-\frac{2}{l}\right)} \zeta^T V_i \zeta - T_i^{2/l}(n), 0 \right) \right\} + o_p(1) \\
&= \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{\frac{1}{2}+\frac{3}{2l+2}-\frac{2}{l}} \mathbb{E}\left[\max\left(n^{-\left(\frac{6}{2l+2}-\frac{2}{l}\right)} \zeta^T V_1 \zeta - T_1^{2/l}(n), 0 \right) \right] \right\} + o_p(1).
\end{aligned}$$

The final estimation follows as in the proof for Theorem 3.1 (i.e. applying Lemma 3.1).

In **Step 4** for $l = 1$, as $n \to \infty$ the objective function is

$$nR_n(\theta_*) = \max_{\zeta} \left\{ -2\zeta^T Z_n(\theta_*) - \mathbb{E}\left[\max\left(\zeta^T V_1 \zeta - \frac{T_1^2(n)}{n}, 0 \right) \right] \right\} + o_p(1).$$

Let us denote $G : \mathbb{R}^l \to \mathbb{R}$ to be a deterministic continuous function, defined as

$$G(\zeta, n) = \mathbb{E}\left[\max\left(\zeta^T V_1 \zeta - \frac{T_1^2(n)}{n}, 0\right)\right].$$

We know $\Upsilon = \mathbb{E}[V_1]$ is symmetric strictly positive definite matrix, then the objective function is strictly convex and differentiable in ζ. Thus the (unique) global maximizer is characterized by the first order optimality condition almost surely. We take derivative w.r.t. ζ and set it to be 0, applying the same estimation in the original proof the first order optimality condition becomes

$$Z_n = -\Upsilon\zeta + o_p(1) \text{ as } n \to \infty. \tag{3.29}$$

Since Υ is invertible, for any n we can solve optimal $\zeta_n^* = -\Upsilon^{-1} Z_n + o_p(1)$. Plugging ζ_n^* in the objective function, as $n \to \infty$ we have

$$nR_n(\theta_*) = 2Z_n^T \Upsilon^{-1} Z_n - G\left(-\Upsilon^{-1} Z_n, n\right) + o_p(1).$$

As $n \to \infty$, we can apply the same estimation in the proof of Theorem 3.1, it becomes

$$nR_n(\theta_*) \Rightarrow \tilde{Z}^T \Upsilon^{-1} \tilde{Z}.$$

Thus we proof the claim for $l = 1$.

In **Step 5** for $l = 2$, as $n \to \infty$ the objective function has estimate

$$nR_n(\theta_*) = \max_{\zeta} \left\{-2\zeta^T Z_n(\theta_*) - \mathbb{E}\left[\max\left(\zeta^T V_1 \zeta - T_1(n), 0\right)\right]\right\} + o_p(1).$$

Still, we denote $G(\zeta, n)$ to be a deterministic function given as,

$$G(\zeta, n) = \mathbb{E}\left[\max\left(\zeta^T V_1 \zeta - T_1(n), 0\right)\right].$$

Same as discussed in for $l = 1$, the objective function is strictly convex and differentiable in ζ, thus the (unique) global maximizer could be characterized via first order optimality condition almost surely. We take derivative w.r.t. ζ and set it to be 0, applying same estimation in the proof of Theorem 3.1 the first order optimality condition becomes

$$Z_n = -\mathbb{E}\left[V_1 1_{(\tau \leq \zeta^T V_1 \zeta)}\right]\zeta + o_p(1) \text{ as } n \to \infty. \tag{3.30}$$

We know the objective function is strictly convex differentiable, then for fixed Z_n there is a unique ζ_n^* that satisfies the first order optimality condition (3.30). We plug in the optimizer and the objective function becomes

$$nR_n(\theta_*) = -2Z_n^T \zeta_n^* - G(\zeta_n^*, n) + o_p(1) \text{ as } n \to \infty.$$

As $n \to \infty$, we can apply the same estimation in the proof of Theorem 3.1, we have

$$nR_n(\theta_*) \Rightarrow -2\tilde{Z}^T\tilde{\zeta} - \tilde{\zeta}^T \tilde{G}\left(\tilde{\zeta}\right)\tilde{\zeta},$$

where $\tilde{G} : \mathbb{R}^q \to \mathbb{R}^q \times \mathbb{R}^q$ is a deterministic continuous mapping defined as,

$$\tilde{G}(\zeta) = \mathbb{E}\left[V_1 \max\left(1 - \tau/\left(\zeta^T V_1 \zeta\right), 0\right)\right],$$

and $\tilde{\zeta} := \tilde{\zeta}\left(\tilde{Z}\right)$ is the unique solution to

$$\tilde{Z} = -\zeta \mathbb{E}\left[V_1 1_{(\tau \leq \zeta^T V_1 \zeta)}\right].$$

Then we proved the claim for $l = 2$.

Finally, in **Step 6** for $l \geq 3$, as $n \to \infty$ the objective function is

$$\begin{aligned} &n^{1/2+\frac{3}{2l+2}} R_n(\theta_*) \\ &= \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} \mathbb{E}\left[\max\left(n^{-\left(\frac{6}{2l+2}-\frac{2}{l}\right)} \zeta^T V_1 \zeta - T_1^{2/l}(n), 0\right)\right] \right\} + o_p(1). \end{aligned}$$

We denote $G(\zeta, n)$ to be a deterministic function defined as,

$$G(\zeta, n) = n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} \mathbb{E}\left[\max\left(n^{-\left(\frac{6}{2l+2}-\frac{2}{l}\right)} \zeta^T V_1 \zeta - T_1^{2/l}(n), 0\right)\right].$$

Follows the same discussion above for $l = 1$ and 2, we know the objective function is strictly convex differentiable in ζ and the global maximizer is characterized by first order optimality condition almost surely. We take derivative of the objective function w.r.t. ζ and set it to be 0. We apply the same technique as in the proof of Theorem 3.1, the first order optimality condition becomes

$$Z_n = -\mathbb{E}\left[V_1 \frac{\pi^{l/2}\left(f_X(X_1) + \kappa f_{\tilde{X}}(X_1)\right)}{\Gamma(l/2+1)} V_1 \left(\zeta^T V_1 \zeta\right)^l\right] \zeta + o_p(1). \text{ as } n \to \infty \quad (3.31)$$

The objective condition is strictly convex differentiable and for fixed Z_n there is a unique ζ_n^* satisfying the first optimality condition (3.31). We plug ζ_n^* into the objective function and it becomes

$$n^{1/2+\frac{3}{2l+2}} R_n(\theta_*) = -2Z_n^T \zeta_n^* - G(\zeta_n^*, n) + o_p(1) \text{ as } n \to \infty.$$

As $n \to \infty$, we can apply same estimate in the proof of Theorem 3.1, we have

$$n^{1/2+\frac{3}{2l+2}} R_n(\theta_*) \Rightarrow -2\tilde{Z}^T\tilde{\zeta} - \frac{2}{l+2}\tilde{G}\left(\tilde{\zeta}\right),$$

where $\tilde{G} : \mathbb{R}^q \to \mathbb{R}$ is a deterministic continuous function given as,

$$\tilde{G}(\zeta) = \mathbb{E}\left[\frac{\pi^{l/2}\left(f_X(X_1) + \kappa f_{\tilde{X}}(X_1)\right)}{\Gamma(l/2+1)}\left(\zeta^T V_1 \zeta\right)^{l/2+1}\right],$$

and $\tilde{\zeta} := \tilde{\zeta}\left(\tilde{Z}\right)$ is the unique solution to

$$\tilde{Z} = -\mathbb{E}\left[V_1 \frac{\pi^{l/2}\left(f_X(X_1) + \kappa f_{\tilde{X}}(X_1)\right)}{\Gamma(l/2+1)} V_1 \left(\zeta^T V_1 \zeta\right)^l\right]\zeta.$$

We proved the claim for $l \geq 3$ and finish the proof for Theorem 3.3. □

3.3.3.2 Proofs of SoS Theorems for General Estimation with Plug-In

The proofs of the plug-in version of SoS theorems for general estimation equation also mainly follows the proof of Theorem 3.1, we are going to discuss the different steps here.

Proof of Corollary 3.1. For implicit formulation, as we discussed for Theorem 3.2, we can redefine $X_i \leftarrow h(\gamma_*, \nu_n, X_i)$, $Z_k \leftarrow h(\gamma_*, \nu_n, Z_k)$, $X_i(*) \leftarrow h(\gamma_*, \nu_*, X_i)$ and $Z_k(*) \leftarrow h(\gamma_*, \nu_*, X_i)$. Then the proof for the implicit formulation with plug-in goes as follows.

In **Step 1**, the dual formulation is similar given as

$$R_n(\gamma_*) = \max_{\lambda,\gamma_i \geq 0} \left\{ -\lambda \bar{X}_n - \frac{1}{n}\sum_{i=1}^{n} \gamma_i \right\}$$
$$\text{s.t. } -\gamma_i \leq \min_j \left\{ \lambda^T X_i - \lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\}, \text{ for all } i.$$

We can apply first order Taylor expansion to $h\left(\gamma_*, \nu_n, X_i\right)$ w.r.t. ν, then we have

$$h\left(\gamma_*, \nu_n, X_i\right) = h\left(\gamma_*, \nu_*, X_i\right) + O_p\left(\frac{\|D_\nu h\left(\gamma_*, \bar{\nu}_n, X_i\right)\|}{n^{1/2}}\right),$$

where $\bar{\nu}_n$ is a point between ν_n and ν_*. By our change of notation for X_i, $X_i(*)$, Z_k and $Z_k(*)$ and the above Taylor expansion, we can observe

$$Z_k = Z_k(*) + \epsilon_n\left(Z_k\right),$$

where $\epsilon_n\left(Z_k\right) = O_p\left(\|D_\nu h\left(\gamma_*, \bar{\nu}_n, Z_k\right)\| / n^{1/2}\right)$.

In **Step 2** we can define a point process $N_n^{(i)}\left(t, \lambda\right)$ and $T_i\left(n\right)$ as in the proof of Theorem 3.1, but the rate becomes

$$\Lambda\left(X_i, \lambda\right) = \left[f_X\left(X_i + \lambda/2 + \epsilon_n\left(X_i\right)\right) + \kappa f_{\tilde{X}}\left(X_i + \lambda/2 + \epsilon_n\left(X_i\right)\right)\right] \frac{\pi^{l/2}}{\Gamma\left(l/2+1\right)}.$$

As $n \to \infty$, same as in the proof of Theorem 3.1 and Theorem 3.3 we can argue $\lambda \to 0$. Then we can define τ same as in the proof of Theorem 3.1 and has the with same distribution

$$\mathbb{P}\left[\tau \geq t\right] = \mathbb{E}\left[\exp\left(-\left(f_X\left(X_1\right) + \kappa f_{\tilde{X}}\left(X_1\right)\right) \frac{\pi^{l/2}}{\Gamma\left(l/2+1\right)}\right)\right].$$

Then the rest of the proof in **Step 3, 4, 5 and 6** stays the same as that of Theorem

3.1, but replacing the CLT for Z_n by asymptotic distribution given in C2). □

Proof of Corollary 3.2. For explicit formulation, the proof is more close to the proof of Theorem 3.3 and we are discussing the difference as follows.

In **Step 1**, the dual formulation takes the form

$$R_n(\theta_*) = \max_{\lambda} \Big\{ - \lambda^T \bar{h}_n(\gamma_*, \nu_n) - \frac{1}{n}\sum_{i=1}^{n} \max_j \left\{ \lambda^T h(\gamma_*, \nu_n, Z_j) - \lambda^T h(\gamma_*, \nu_n, X_i) - \|X_i - Z_j\|_2^2 \right\}^+ \Big\}.$$

Step 2 and 3 Follows the same as for the proof of Theorem 3.3 however we need to notice that difference is the definition of $\bar{a}_*(X_i, \zeta)$, for $l = 1$ and 2 we have

$$\begin{aligned}
\bar{a}_*(X_i, \zeta) &= X_i + D_x h(\gamma_*, \nu_n, \bar{a}_*(X_i, \zeta)) \cdot \frac{\zeta}{n^{1/2}} \\
&= X_i + D_x h(\gamma_*, \nu_n, X_i) \cdot \frac{\zeta}{n^{1/2}} + O\left(\frac{\|\zeta\|_2^2}{n} \|D_x h(\gamma_*, \nu_n, \bar{a}_*(X_i, \zeta))\|_2 \right) \\
&= X_i + D_x h(\gamma_*, \nu_*, X_i) \cdot \frac{\zeta}{n^{1/2}} + O\left(\frac{\|\zeta\|_2^2}{n} \|D_x h(\gamma_*, \nu_n, \bar{a}_*(X_i, \zeta))\|_2 \right) \\
&\quad + O\left(\frac{\|\zeta\|_2}{n^{1/2}} \|\nu_n - \nu_*\|_2 \|D_x h(\gamma_*, \nu_n, \bar{a}_*(X_i, \zeta))\|_2 \|D_\nu D_x h(\gamma_*, \bar{\nu}_n, \bar{a}_*(X_i, \zeta))\|_2 \right),
\end{aligned}$$

where $\bar{\nu}_n$ is a point between ν_n and ν_*. By assumption C5)-C7) we can notice the rest of step 2 and 3 stay the same as in the proof of Theorem 3.3. In **Step 4, 5 and 6** we use $Z_n = \frac{1}{n^{1/2}} \sum_{i=1}^{n} h(\gamma_*, \nu_n, X_i) \Rightarrow \tilde{Z}'$ given in C2) instead of CLT. □

3.4 Application to Stochastic Optimization and Stress Testing

We are going to provide an application of the SoS inference framework to quantify model uncertainty in the context of stochastic programming. As a motivating application we consider the problem of evaluating Conditional Value at Risk (C-VaR). More examples of applying the SoS inference methods will be discussed in Chapter 4, where we consider the support is a combination of the labeled and unlabeled data to encode the unsupervised information into modeling to propose a semi-supervised algorithm.

We are interested in the value function of a stochastic programming problem formulated via

$$\begin{aligned} C_* = \min_{\theta} \quad & \mathbb{E}\left[m(\theta, X)\right] \\ & s.t.\ \mathbb{E}[\phi(\theta, X)] \leq 0. \end{aligned} \tag{3.32}$$

We assume that the objective function $\psi(\theta) = \mathbb{E}\left[m(\theta, X)\right]$ is a convex function in θ; while the constraints $\mathbb{E}[\phi(\theta, X)] \leq 0$ specify a convex region in θ, for example we can assume $\phi(\theta, X)$ is a convex function in θ for any X.

Following Lam and Zhou [2015], the goal is to estimate the optimal value function using the SOS formulation and we will apply a plug-in estimator for θ_* (which is treated as a nuisance parameter). Subsequently, when introducing the Lagrangian relaxation of (3.32) we will be able to also introduce a plug-in estimator for the associated Lagrange multiplier. Therefore, for simplicity we shall focus on the unconstrained minimization problem $C_* = \min_\theta \left\{\mathbb{E}\left[m(\theta, X)\right]\right\}$.

The authors in Lam and Zhou [2015] provide a discussion for some potential ap-

proaches to derive nonparametric confidence interval (including Empirical Likelihood, a Bayesian approach, Bootstrap and the Delta method). In Lam and Zhou [2015] it is argued that the Empirical Likelihood method tends to have best finite sample performance, and Lam and Zhou [2015] provides an optimal (in certain sense) specification for Empirical Likelihood approach. More importantly, in Lam and Zhou [2015] an approach combining empirical likelihood and a plug-in estimator for optimizer is introduced, which avoids solving a non-convex optimization problem introduced in the discussion of Lam and Zhou [2015].

Our goal in this section is to derive a plug-in estimator based on the SOS inference approach introduced in Section 3.2. The approach that we introduce next is the analog of the plug-in strategy discussed in Blanchet *et al.* [2016a] in order to find a robust confidence interval for C_*.

The following result is a direct extension of Corollary 3.1 and Corollary 3.2. This corollary plays the key role in specifying confidence interval for C_*. To ensure the corollary hold, we need some assumptions:

D1): Assume $\psi(\cdot)$ is convex differentiable in θ and there is a unique optimizer θ_*.

D2): Assume that $\psi(\cdot)$ is strongly convex at θ_*, that is, for every θ there exist $\delta > 0$, such that

$$M(\theta) \geq M(\theta_*) + \delta \|\theta - \theta_*\|_2^2 .$$

Corollary 3.3. [Plug-in for Implicit/Explicit SoS Function for Stochastic Optimization] Let us consider stochastic programming problem $C_* = \min_\theta M(\theta) = \min_\theta \mathbb{E}[m(\theta, X)]$. We assume assumption D1)-D2) hold. We consider the estimating equations to be

the derivative condition and value function condition

$$\mathbb{E}\left[m(\theta_*, X) - C_*\right] = 0, \text{ and } \mathbb{E}\left[D_\theta m\left(\theta_*, X\right)\right] = 0.$$

For simplicity, let us denote $h\left(\theta_*, C_*, x\right) = \left(m(\theta_*, x) - C_*, D_\theta m\left(\theta_*, X\right)^T\right)^T$. We are interested in C_* only and consider a sample average approximation (SAA) estimator for θ_* to be $\hat{\theta}_{SAA}$. For $h\left(\cdot, C_*, x\right)$ we assume C1)-C7) hold. Let us denote $U \sim N\left(0, \text{Var}\left(m\left(\theta_*, X\right)\right)\right) \in \mathbb{R}$ and $U(0) = \left(U, \vec{0}\right)^T \in \mathbb{R}^{d+1}$. Recall the implicit and explicit formulations for general estimating equation SoS function defined in Definition 2 and Definition 3, we have the following asymptotic results.
For the implicit SoS formulation, we have

- When $d = 1$ (estimating equation dimension is $d + 1 = 2$)

$$nR_n^W(C_*) \Rightarrow \rho\left(U\right)\left[2 - \tilde{\eta}\left(U\right)\rho\left(U\right)\right]U^{2,}$$

  where $\rho\left(U\right)$ is the unique solution to

$$\frac{1}{\rho} = \tilde{g}\left(\rho U\right),$$

  and $\tilde{g} : \mathbb{R} \to \mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{g}\left(x\right) = \mathbb{P}\left[x^2 \geq \tau\right].$$

  $\tilde{\eta}\left(x\right)$ is also a deterministic function, defined as

$$\tilde{\eta}\left(x\right) = \mathbb{E}\left[\max\left(1 - \tau/x^2, 0\right)\right],$$

and τ is independent of U satisfying

$$\mathbb{P}\left[\tau > t\right] = E\left(\exp\left(-g\left(h\left(\theta_*, C_*, X_1\right)\right)\pi t\right)\right).$$

- When $d \geq 2$,

$$n^{1/2+\frac{3}{2d+4}} R_n^W(C_*) \Rightarrow \frac{2d+4}{d+3} \frac{||U||^{1+\frac{1}{d+2}}}{\mathbb{E}\left[\frac{\pi^{(d+1)/2}}{\Gamma((d+3)/2)} g_X\left(h\left(\theta_*, C_*, X_1\right)\right)\right]^{\frac{1}{d+2}}}.$$

For the explicit formulation, we have following asymptotic results (we use $\zeta_{[1]}$ denote first element of vector ζ)

- When $l = 1$,

$$nR_n^W(C_*) \Rightarrow v_{1,1}U^2,$$

where $v_{1,1}$ is the $(1,1)$ element of matrix Υ^{-1}.

- Suppose that $l = 2$. It is possible to uniquely define deterministic continuous mapping $\tilde{\zeta} : \mathbb{R}^q \to \mathbb{R}^q$, such that

$$z = -\mathbb{E}\left[\bar{V}_1 I\left(\tau \leq \tilde{\zeta}^T(z)\,\bar{V}_1\tilde{\zeta}(z)\right)\right]\tilde{\zeta}(z),$$

where τ is independent of U satisfying

$$\mathbb{P}\left(\tau > t\right) = \mathbb{E}\left(\exp\left(-\left[f_X\left(X_1\right) + \kappa f_{\tilde{X}}\left(X_1\right)\right]\pi t\right)\right).$$

Furthermore,

$$nR_n(\theta_*) \Rightarrow -2U\tilde{\zeta}_{[1]} - \tilde{\zeta}^T\left(U(0)\right)\tilde{G}\left(\tilde{\zeta}\right)\tilde{\zeta}\left(U(0)\right),$$

where $\tilde{G}:\mathbb{R}^q \to \mathbb{R}^{q\times q}$ is a deterministic continuous mapping defined as

$$\tilde{G}\left(\zeta\right) = \mathbb{E}\left[\bar{V}_1 \max\left(1 - \frac{\tau}{\zeta^T\bar{V}_1\zeta}, 0\right)\right],$$

and U is independent with $\bar{V}_1$ and τ.

- Assume that $l \geq 3$. A continuous function $\tilde{\zeta}:\mathbb{R}^q \to \mathbb{R}^q$ can be defined uniquely so that

$$z = -\mathbb{E}\left[\frac{\pi^{l/2}\left(f_X\left(X_1\right) + \kappa f_{\tilde{X}}\left(X_1\right)\right)}{\Gamma(l/2+1)}\bar{V}_1\left(\tilde{\zeta}^T\left(z\right)\bar{V}_1\tilde{\zeta}\left(z\right)\right)^l\right]\tilde{\zeta}\left(z\right)$$

(note that $\bar{V}_1$ is a function of X_1). Moreover,

$$n^{1/2+\frac{3}{2l+2}}R_n(\theta_*) \Rightarrow -2U\tilde{\zeta}_{[1]} - \frac{2}{l+2}\tilde{G}\left(\tilde{\zeta}\left(U(0)\right)\right),$$

where $\tilde{G}:\mathbb{R}^q \to \mathbb{R}$ is a deterministic function given as

$$\tilde{G}\left(\zeta\right) = \mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}f_X\left(X_1\right)\left(\zeta^T\bar{V}_1\zeta\right)^{l/2+1}\right],$$

and U and X_1 are independent.

This corollary is a special case of plug-in theorem for SoS formulation is a special case of Corollary 3.1 and Corollary 3.2. The estimating equations correspond to the first order optimality condition (i.e. the first derivative equal to zero), condition and the corresponding optimal value equation. We use sample average approximation

estimator as the underlying plug-in estimator.

We notice for sample average approximation algorithm, guaranteed by assumptions D1)-D3, it has been shown in Ruszczynski and Shapiro [2003]; Shapiro and Dentcheva [2014a] the optimizer $\hat{\theta}_{SAA}$ and optimal value function $\frac{1}{n}\sum_{i=1}^{n} m\left(\hat{\theta}_{SAA}, X_i\right)$ have

$$\begin{aligned}
&\hat{\theta}_{SAA} - \theta_* = O\left(1/n^{1/2}\right) \\
&\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta m\left(\hat{\theta}_{SAA}, X_i\right) = 0, \\
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m\left(\hat{\theta}_{SAA}, X_i\right) - C_*\right) \Rightarrow N\left(0, \mathrm{Var}\left(m\left(\theta_*, X\right)\right)\right).
\end{aligned}$$

Since Corollary 3.3 follows as a direct application of Corollary 3.2 and Corollary 3.1, its proof is omitted.

Similar as the derivation in Blanchet *et al.* [2016a] for empirical likelihood, for the plug-in estimator derived from sample average approximation, if we denote

$$n^{1/2+3/(2d+4)} R_n^{W(implicit)}(C_*) \Rightarrow R_0^{(implicit)} \text{ and } n^{1/2+3/(2l+2)} R_n^{W(explicit)}(C_*) \Rightarrow R_0^{(explicit)},$$

we can specify a robust 95% confidence interval for C_* under both explicit and implicit formulation by:

$$\mathrm{CI}^{(\cdot)}\left(C_*\right) = \left\{ C \in \mathbb{R} \middle| n^{\alpha} R_n^{W(\cdot)}\left(C\right) \leq R_0^{(\cdot)}\left(95\%\right) \right\}$$

where α depends on the formulation and dimension as in Corollary 3.3 and $R_0^{(\cdot)}\left(95\%\right)$ is the upper 95% quantile for $R_0^{(explicit)}$ (or $R_0^{(implicit)}$). The upper/lower bound of con-

fidence interval $(C_{up}^{(\cdot)}/C_{l0}^{(\cdot)})$ can be found by solving the linear programming problem

$$C_{up}^{(\cdot)}/C_{lo}^{(\cdot)} = \max_{\pi(i,j)} \ / \ \min_{\pi(i,j)} \{ \sum_{i,j=1}^{n} \pi(i,j) m(\hat{\theta}_{SAA}, X_i)$$
$$s.t. \ \pi(i,j) \geq 0 \ \sum_{j=1}^{n} \pi(i,j) = 1/n; \sum_{i,j=1}^{n} \pi(i,j) \left\| X_i - X_j \right\|_2^2 \leq R_0^{(\cdot)}(95\%) \}.$$

Next, we are going to provide a numerical example in quantifying C-VaR using the methodology we developed above.

Example 3.1. (Quantify the uncertainty of Conditional Value at Risk (C–VaR)) In this example we would like to consider find a SoS based 95% confidence interval for conditional value at risk with 90% level. The conditional value at risk with $\alpha-$level is given as solving the stochastic programming problem:

$$\text{C-VaR}(\alpha) = \inf_{\theta} \left\{ \theta + \frac{1}{1-\alpha} \mathbb{E} \left[\left(\sum_{k=1}^{l} X^{(k)} - \theta \right)^{+} \right] \right\}.$$

We shall test our method using simulated data under different distributional assumptions. We a sample i.i.d. observations $\{X_i\}_{i=1}^{n} \subset \mathbb{R}^l$. We will apply the SoS inference procedure to provide a non-parametric confidence interval for C-VaR(90%). In order to verify the coverage probability we use data simulated from normal distribution and Laplace (double exponential) distributions. We consider the case $l = 4$. For the normal distribution setting we assume $X_i \sim N(0, I_{4\times 4})$, while for Laplace distribution we consider for each $k = 1, ..., 4$, $X_i^k \sim Laplace(0, 1)$ and all of these random variables are independent. For these two cases, we can calculate the solution in closed form; for the normal setting the optimizer is $\theta^\star = 2.5632$ and optimal value function is C-VaR$(0.9) = 3.510$; for Laplace setting the optimizer is $\theta^\star = 3.497$ with optimal value function equal to

C-VaR$(0.9) = 5.066$.

As for this example, we have three approaches in which our SoS procedure can be applied: 1) implicit SoS formulation (ISOS); 2) explicit SoS formulation while assume underline data is l dimension (ESOS-O), i.e. $X_i = \left(X_i^{(1)}, \ldots, X_i^{(l)}\right)^T \in \mathbb{R}^l$; 3) explicit formulation while assume underline data is 1 dimension (ESOS-C), i.e. $X_i = X_i^{(1)} + \ldots + X_i^{(l)} \in \mathbb{R}$. We compare our methods with empirical likelihood method (EL) inBlanchet *et al.* [2016a], nonparametric bootstrap method (BT), and CLT based Delta method (CLT) discussed in Theorem 5.7 Shapiro and Dentcheva [2014a]. We consider four settings $n = 20, 50, 100$ and 500. For each setting, we repeat the experiment $N = 1000$ times, and note down the empirical coverage probability, mean of upper and lower bounds, and the mean and standard deviation of the interval width for each method. The results are summarized in Table 1 for Normal distribution and Table 2 for Laplace distribution below.

We can observe that, the three SOS-based approaches tend to have better coverage probabilities in all cases for both distributions comparing to EL, bootstrap and the Delta method. Especially for small sample situations ($n = 10, 20$) EL and all of the SOS-based approaches appear to perform better than everything else. It is discussed in Lam and Zhou [2015] that EL has better finite sample performance compared to the Delta method and bootstrap. We can also notice that all empirical SoS methods tend to have smaller variance compared to others, especially for relatively large sample sizes ($n = 100, 500$). Between the three SoS methods, we can see that explicit formulations work better comparing to implicit, which follows our discussion after Definition 3. For the two explicit-formulation methods, since we know the data affects the objective function in the form $X_i^{(1)} + \ldots + X_i^{(l)}$, we would expect better performance if we combine

the data in a single dimension. The numerical results validate our intuition.

n	Method	Coverage Probability	Mean Lower Bound	Mean Upper Bound	Mean Interval Length	S.D. of Length
20	ESOS-C	79.8%	2.59	4.68	2.09	0.79
	ESoS-O	73.4%	2.55	4.65	2.10	1.21
	ISoS	70.8%	2.34	4.87	2.53	0.82
	EL	71.7%	2.61	5.18	2.57	1.92
	BT	55.6%	1.76	3.88	2.12	1.23
	CLT	71.8%	2.01	4.52	2.51	1.87
50	ESOS-C	93.3%	2.67	4.57	1.90	0.30
	ESoS-O	91.0%	2.63	4.54	1.91	0.57
	ISoS	87.3%	2.70	4.75	2.05	0.56
	EL	89.2%	2.81	4.78	1.96	0.83
	BT	82.7%	2.30	4.25	1.95	0.77
	CLT	86.6%	2.47	4.44	1.97	0.78
100	ESOS-C	92.8%	2.84	4.20	1.36	0.08
	ESoS-O	92.4%	2.80	4.22	1.42	0.23
	ISoS	91.3%	2.89	4.32	1.53	0.25
	EL	91.4%	2.94	4.46	1.52	0.43
	BT	90.1%	2.67	4.16	1.49	0.41
	CLT	90.4%	2.75	4.17	1.42	0.39
500	ESOS-C	95.3%	3.16	3.85	0.69	0.01
	ESoS-O	94.9%	3.14	3.77	0.63	0.05
	ISoS	91.2%	3.19	3.88	0.79	0.03
	EL	93.9%	3.20	3.93	0.73	0.08
	BT	94.2%	3.16	3.84	0.68	0.07
	CLT	94.7%	3.17	3.84	0.67	0.08

Table 3.1: $\alpha = 0.9-$**Conditional Value at Risk with Gaussian Data.** The data X is simulated from 4-dim standard Gaussian distribution, while each dimension is independent. We consider sample size $n = 20, 50, 100,$ and 500. We repeat the experiments $N = 1000$ times and record the coverage probability for the confidence interval (CI), the average upper and lower bound for CI, also the average length and standard deviation for CI. ESoS-C is the explicit formulation of SoS with combined data, ESoS-O stands for explicit-SoS with original data, ISoS is the implicit SOS, EL stands for empirical likelihood, BT is short for nonparametric bootstrap, and CLT is the asymptotic CI method.

3.5 Conclusions and Discussion

This chapter introduces a methodology inspired by Empirical Likelihood, but in which the likelihood ratio function is replaced by a Wasserstein distance. The methodology that we propose is motivated by the problem of systematically finding estimators which are incorporate out-of-sample performance in their design. In turn, as a motivation for the need of finding these types of estimators we discussed applications

n	Method	Coverage Probability	Mean Lower Bound	Mean Upper Bound	Mean Interval Length	S.D. of Length
20	ESOS-C	78.2%	3.57	6.89	3.32	1.10
	ESoS-O	73.8%	3.48	7.10	3.62	1.91
	ISoS	73.1%	3.87	7.55	3.68	1.16
	EL	72.3%	3.56	8.00	4.44	3.30
	BT	58.1%	2.40	6.01	3.61	2.40
	CLT	70.5%	2.53	6.90	4.37	3.24
50	ESOS-C	89.4%	3.78	6.64	2.86	0.42
	ESoS-O	89.3%	3.69	6.78	3.09	0.89
	ISoS	80.1%	4.21	7.17	2.96	0.63
	EL	86.2%	3.89	7.43	3.53	1.66
	BT	80.5%	3.15	6.58	3.43	1.54
	CLT	83.6%	3.29	6.64	3.35	1.54
100	ESOS-C	91.9%	3.93	6.22	2.29	0.14
	ESoS-O	90.8%	3.88	6.30	2.42	0.43
	ISoS	86.6%	4.30	6.78	2.44	0.36
	EL	89.9%	4.10	6.66	2.56	0.86
	BT	86.2%	3.71	6.16	2.45	0.81
	CLT	87.6%	3.76	6.17	2.41	0.79
500	ESOS-C	94.7%	4.53	5.62	1.09	0.06
	ESoS-O	94.3%	4.46	5.59	1.13	0.08
	ISoS	92.1%	4.43	5.61	1.17	0.13
	EL	94.0	4.53	5.78	1.25	0.18
	BT	92.2%	4.46	5.58	1.12	0.16
	CLT	93.1%	4.45	5.48	1.13	0.15

Table 3.2: $\alpha = 0.9-$**Conditional Value at Risk with Laplace Data.** The data X is simulated from 4-dim standard Laplace distribution, while each dimension is independent. We consider sample size $n = 20, 50, 100,$ and 500. We repeat the experiments $N = 1000$ times and record the coverage probability for the confidence interval (CI), the average upper and lower bound for CI, also the average length and standard deviation for CI. ESoS-C is the explicit formulation of SoS with combined data, ESoS-O stands for explicit-SoS with original data, ISoS is the implicit SOS, EL stands for empirical likelihood, BT is short for nonparametric bootstrap, and CLT is the asymptotic CI method.

to stress testing. We envision this chapter as the first installment on this research area and we plan to explore more deeply applications not only in stress testing but also in machine learning. For example, in Chapter 2, we study a connection between the estimation procedure that we introduce here and statistical techniques such as LASSO and support vector machine (SVM) which are popular in machine learning.

In Chapter 2 we also explore the limiting distribution obtained for the SoS function when we compare the empirical distribution against any other distribution, as opposed to only distributions supported on a finite set of scenarios and, in this case, we show that the distribution is typically chi-squared (so this case is, in some sense, closer to the Empirical Likelihood setting).

In addition, given the parallel philosophy underpinning the method that we proposed (based on Empirical Likelihood), the results on this chapter open up a significant amount of research opportunities which are parallel to the substantial literature produced in the area of Empirical Likelihood during the last three decades. We mention, in particular, applications to regression problems (see Owen [1991]; Chen [1993, 1994]; Wang and Rao [2001]; Zhao and Wang [2008]; Chen and Keilegom [2009]), survival analysis (see Murphy [1995]; Li *et al.* [1996]; Hollander and McKeague [1997]; Li *et al.* [1997]; Einmahl and McKeague [1999]; Wang *et al.* [2009]; Zhou [2015]), econometrics (see Newey and Smith [2004]; Bravo [2004]; Kitamura [2006]; Antoine *et al.* [2007]; Guggenberger [2008]; Imbens [2012]) and additional recent work on stochastic optimization (see Lam and Zhou [2015]; Blanchet *et al.* [2016b]). The methodology we propose could be extended to the above applications by simply replacing the Empirical Likelihood function by the SoS function and by applying asymptotic theorems developed in this chapter (or natural extensions).

Chapter 4

Semi-Supervised Learning based on Distributionally Robust Optimization

Starting from this chapter and in the following two chapters, namely, Chapter 5 and Chapter 6, we are going to discuss the generalization and application of the data-driven DRO formulation and the RWPI and SoS inference methods. We also start provide algorithms to solve data-driven DRO problems directly.

In this chapter, we propose a novel method for semi-supervised learning (SSL) based on data-driven distributionally robust optimization (DRO) using optimal transport metrics. Our proposed method enhances generalization error by using the unlabeled data to restrict the support of the worst case distribution in our DRO formulation. We enable the implementation of our DRO formulation by proposing a stochastic gradient descent algorithm which allows to easily implement the training procedure. We demonstrate that our Semi-supervised DRO method is able to improve the generalization error over natural supervised procedures and state-of-the-art SSL estimators. Finally, we include a discussion on the large sample behavior of the optimal uncertainty region in the DRO formulation. Our discussion exposes important

aspects such as the role of dimension reduction in SSL.

4.1 Introduction

We propose a novel method for semi-supervised learning (SSL) based on data-driven distributionally robust optimization (DRO) using an optimal transport metric.

Our approach enhances generalization error by using the unlabeled data to restrict the support of the models which lie in the region of distributional uncertainty. The intuition is that our mechanism for fitting the underlying model is automatically tuned to generalize beyond the training set, but only over potential instances which are relevant. The expectation is that predictive variables often lie in lower dimensional manifolds embedded in the underlying ambient space; thus, the shape of this manifold is informed by the unlabeled data set (see Figure 4.1 for an illustration of this intuition).

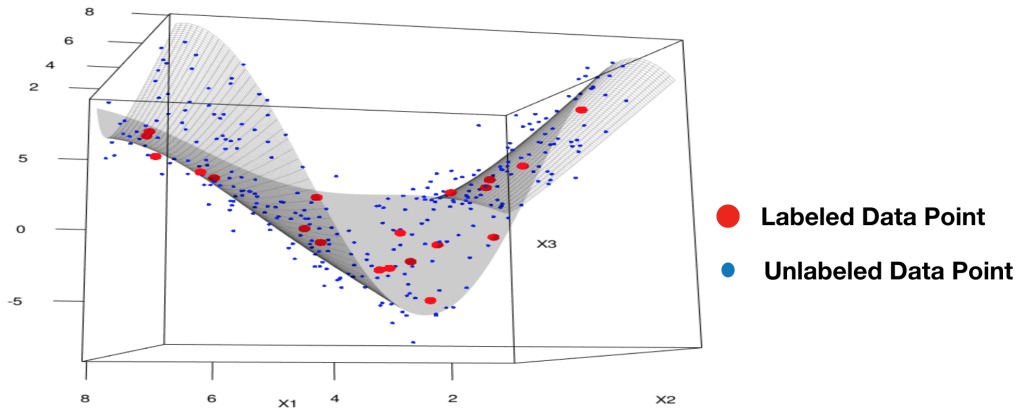

Figure 4.1: Idealization of the way in which the unlabeled predictive variables provide a proxy for an underlying lower dimensional manifold. Large red dots represent labeled instances and small blue dots represent unlabeled instances.

To enable the implementation of the DRO formulation we propose a stochastic

gradient descent (SGD) algorithm which allows to implement the training procedure at ease. Our SGD construction includes a procedure of independent interest which, we believe, can be used in more general stochastic optimization problems.

We focus our discussion on semi-supervised classification but the modeling and computational approach that we propose can be applied more broadly as we shall illustrate in Section 4.4.

We now explain briefly the formulation of our learning procedure. Suppose that the training set is given by $\mathcal{D}_n = \{(Y_i, X_i)\}_{i=1}^n$, where $Y_i \in \{-1, 1\}$ is the label of the i-th observation and we assume that the predictive variable, X_i, takes values in $\mathbb{R}^d$. We use n to denote the number of labeled data points.

In addition, we consider a set of unlabeled observations, $\{X_i\}_{i=n+1}^N$. We build the set $\mathcal{E}_{N-n} = \{(1, X_i)\}_{i=n+1}^N \cup \{(-1, X_i)\}_{i=n+1}^N$. That is, we replicate each unlabeled data point twice, recognizing that the missing label could be any of the two available alternatives. We assume that the data must be labeled either -1 or 1.

We then construct the set $\mathcal{X}_N = \mathcal{D}_n \cup \mathcal{E}_{N-n}$ which, in simple words, is obtained by just combining both the labeled data and the unlabeled data with all the possible labels that can be assigned. The cardinality of $\mathcal{X}_N$, denoted as $|\mathcal{X}_N|$, is equal to $2(N-n)+n$ (for simplicity we assume that all of the data points and the unlabeled observations are distinct).

Let us define $\mathcal{P}(\mathcal{X}_N)$ to be the space of probability measures whose support is contained in $\mathcal{X}_N$. We use P_n to denote the empirical measure supported on the set $\mathcal{D}_n$, so $P_n \in \mathcal{P}(\mathcal{X}_N)$. In addition, we write $E_P(\cdot)$ to denote the expectation associated with a given probability measure P.

Let us assume that we are interested in fitting a classification model by minimizing a given expected loss function $l(X, Y, \beta)$, where β is a parameter which uniquely characterizes the underlying model. We shall assume that $l(X, Y, \cdot)$ is a convex function

for each fixed (X, Y). The empirical risk associated to the parameter β is

$$E_{P_n}\left(l\left(X, Y, \beta\right)\right) = \frac{1}{n}\sum_{i=1}^{n} l\left(X_i, Y_i, \beta\right).$$

In this paper, we propose to estimate β by solving the DRO problem

$$\min_{\beta} \max_{P \in \mathcal{P}(\mathcal{X}_N): D_c(P, P_n) \leq \delta^*} E_P[l\left(X, Y, \beta\right)], \tag{4.1}$$

where $D_c\left(\cdot\right)$ is the optimal transport distance introduced in Chapter 1 Section 1.1.

So, intuitively, (4.1) represents the value of a game in which the outer player (we) will choose β and the adversary player (nature) will rearrange the support and the mass of P_n within a budget measured by δ^*. We then wish to minimize the expected risk regardless of the way in which the adversary might corrupt (within the prescribed budget) the existing evidence. In formulation (4.1), the adversary is crucial to ensure that we endow our mechanism for selecting β with the ability to cope with the risk impact of out-of-sample (i.e. out of the training set) scenarios. We denote the formulation in Equation (4.1) as semi-supervised distributionally robust optimization (SSL-DRO) or semi-supervised learning based on distributionally robust optimization.

The criterion that we use to define $D_c\left(\cdot\right)$ is based on the theory of optimal transport and it is closely related to the concept of Wasserstein distance, see Section 4.3. The choice of $D_c\left(\cdot\right)$ is motivated by recent results which show that popular estimators such as regularized logistic regression, Support Vector Machines (SVM) and square-root Lasso (SR-Lasso) admit a DRO representation *exactly equal to* (4.1) in which the support $\mathcal{X}_N$ is replaced by $\mathbb{R}^{d+1}$ (see Chapter 2 and Equation (4.10) in this chapter.)

In view of these representation results for supervised learning algorithms, the

inclusion of $\mathcal{X}_N$ in our DRO formulation (4.1) provides a natural SSL approach in the context of classification and regression. The goal of this chapter is to enable the use of the distributionally robust training framework (4.1) as a SSL technique. We will show that estimating β via (4.1) may result in a significant improvement in generalization relative to natural supervised learning counterparts (such as regularized logistic regression and SR-Lasso). The potential improvement is illustrated in Section 4.4. Moreover, we show via numerical experiments in Section 4.5, that our method is able to improve upon state-of-the-art SSL algorithms.

As a contribution of independent interest, we construct a stochastic gradient descent algorithm to approximate the optimal selection, β_N^*, minimizing (4.1).

An important parameter when applying (4.1) is the size of the uncertainty region, which is parameterized by δ^*. We apply cross-validation to calibrate δ^*, but we also discuss the non-parametric behavior of an optimal selection of δ^* (according to a suitably defined optimality criterion explained in Section 4.6) as $n, N \to \infty$.

In Section 4.2, we provide a broad overview of alternative procedures in the SSL literature, including recent approaches which are related to robust optimization. A key role in our formulation is played by δ^*, which can be seen as a regularization parameter. This identification is highlighted in the form of (4.1) and the DRO representation of regularized logistic regression which we recall in Equation (4.10). The optimal choice of δ^* ensures statistical consistency as $n, N \to \infty$.

We close this Introduction with a few important notes. First, our SSL-DRO is not a robustifying procedure for a given SSL algorithm. Instead, our contribution is in showing how to use unlabeled information on top of DRO to enhance traditional supervised learning methods. In addition, our SSL-DRO formulation, as stated in Equation (4.1) , is not restricted to logistic regression, instead DRO counterpart could be formulated for general supervised learning methods with various choice of

loss function.

The rest of the chapter is structured as follows. We will quickly review the alternative related state-of-the-art SSL algorithms. In Section 4.3 we discuss the elements of our DRO formulation, including the definition of optimal transport metric and the implementation of a stochastic gradient descent algorithm for the solution of (4.1). In Section 4.4 we explore the improvement in out-of-sample performance of our method relative to regularized logistic regression. In Section 4.5, we compare our procedure against alternative SSL estimators, both in the context of some binary classification real data sets. In Section 4.6, we explore the behavior of the optimal uncertainty size δ^* as the sample size increases, especially we discuss certain asymptotic results on how to pick up the distributional uncertainty size optimally with asymptotic consistency. Section 4.7 contains final considerations and further discussions. In Appendix **??**, we provide more technical details for the asymptotic results stated in Section 4.6.

4.2 Alternative Semi-supervised Learning Procedures

We shall briefly discuss alternative procedures which are known in the SSL literature, which is quite substantial. We refer the reader to the excellent survey of Zhu *et al.* [2005] for a general overview of the area. Our goal here is to expose the similarities and connections between our approach and some of the methods that have been adopted in the community.

For example, broadly speaking graph-based methods Blum and Chawla [2001]; Chapelle *et al.* [2009] attempt to construct a graph which represents a sketch of a lower dimensional manifold in which the predictive variables lie. Once the graph is constructed a regularization procedure is performed which seeks to enhance generalization error along the manifold while ensuring continuity in the prediction in terms

of an intrinsic metric. Our approach by-passes the construction of the graph, which we see as a significant advantage of our procedure. However, we believe that the construction of the graph can be used to inform the choice of cost function $c(\cdot)$ which should reflect high transportation costs for moving mass away from the manifold sketched by the graph.

Some recent SSL estimators are based on robust optimization, such as the work of Balsubramani and Freund [2015]. The difference between data-driven DRO and robust optimization is that the inner maximization in (4.1) for robust optimization is not over probability models which are variations of the empirical distribution. Instead, in robust optimization, one attempts to minimize the risk of the worst case performance of potential outcomes inside a given uncertainty set.

In Balsubramani and Freund [2015], the robust uncertainty set is defined in terms of constraints obtained from the testing set. The problem with the approach in Balsubramani and Freund [2015] is that there is no clear mechanism which informs an optimal size of the uncertainty set (which in our case is parameterized by δ^*). In fact, in the last paragraph of Section 2.3, Balsubramani and Freund [2015] point out that the size of the uncertainty could have a significant detrimental impact in practical performance.

We conclude with a short discussion on the the work of Loog [2016], which is related to our approach. In the context of linear discriminant analysis, Loog [2016] also proposes a distributionally robust optimization estimator, although completely different to the one we propose here. More importantly, we provide a way (both in theory and practice) to study the size of the distributional uncertainty (i.e. δ^*), which allows us to achieve asymptotic consistency of our estimator.

4.3 Semi-supervised Learning based on DRO

This section is divided into two parts. First, we provide the elements of our DRO formulation. Then we will explain how to solve the SSL-DRO problem, i.e. find optimal β in (4.1).

4.3.1 Revisit the optimal transport discrepancy:

Assume that the cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ is lower semi-continuous. As mentioned in the Chapter 1 Section 1.1, we also assume that $c(u, v) = 0$ if and only if $u = v$.

Now, given two distributions P and Q, with supports $\mathcal{S}_P \subseteq \mathcal{X}_N$ and $\mathcal{S}_Q \subseteq \mathcal{X}_N$, respectively, we define the optimal transport discrepancy, D_c, via

$$D_c\left(P, Q\right) = \inf\{E_\pi\left[c\left(U, V\right)\right] : \pi \in \mathcal{P}\left(\mathcal{S}_P \times \mathcal{S}_Q\right),\ \pi_U = P,\ \pi_V = Q\}, \tag{4.2}$$

where $\mathcal{P}\left(\mathcal{S}_P \times \mathcal{S}_Q\right)$ is the set of probability distributions π supported on $\mathcal{S}_P \times \mathcal{S}_Q$, and π_U and π_V denote the marginals of U and V under π, respectively.

Observe that (4.2) is obtained by solving a linear programming problem. For example, suppose that $Q = P_n$, and let $P \in \mathcal{P}\left(\mathcal{X}_N\right)$ then, using $U = (X, Y)$, we have that $D_c\left(P, P_n\right)$ is obtained by computing

$$\min_\pi \Big\{ \sum_{u \in \mathcal{X}_N} \sum_{v \in \mathcal{D}_n} c\left(u, v\right) \pi\left(u, v\right) :\ \text{s.t.} \sum_{u \in \mathcal{X}_N} \pi\left(u, v\right) = \frac{1}{n}\ \forall\ v \in \mathcal{D}_n, \tag{4.3}$$

$$\sum_{v \in \mathcal{D}_N} \pi\left(u, v\right) = P\left(\{u\}\right)\ \forall\ u \in \mathcal{X}_N, \pi\left(u, v\right) \geq 0\ \forall\ \ (u, v) \in \mathcal{X}_N \times \mathcal{D}_n \Big\}$$

4.3.2 Solving the SSL-DRO formulation:

A direct approach to solve (4.1) would involve alternating between minimization over β, which can be performed by, for example, stochastic gradient descent and maximization which is performed by solving a linear program similar to (4.3). Unfortunately, the large scale of the linear programming problem, which has $O(N)$ variables and $O(n)$ constraints, makes this direct approach rather difficult to apply in practice.

So, our goal here is to develop a direct stochastic gradient descent approach which can be used to approximate the solution to (4.1).

First, it is useful to apply linear programming duality to simplify (4.1). Note that, given β, the inner maximization in (4.1) is simply

$$\max_{\pi} \Big\{ \sum_{u\in\mathcal{X}_N} \sum_{v\in\mathcal{D}_N} l\left(u,\beta\right)\pi\left(u,v\right) \,:\text{s.t.}\quad \sum_{u\in\mathcal{X}_N} \pi\left(u,v\right) = \frac{1}{n}\ \forall\ v\in\mathcal{D}_n \qquad (4.4)$$
$$\sum_{u\in\mathcal{X}_N} \sum_{v\in\mathcal{D}_n} c\left(u,v\right)\pi\left(u,v\right) \leq \delta\ \ \pi\left(u,v\right)\geq 0\ \forall\ \ \left(u,v\right)\in\mathcal{X}_N\times\mathcal{D}_n\ \Big\}.$$

Of course, the feasible region in this linear program is always non-empty because the probability distribution $\pi\left(u,v\right) = I\left(u=v\right)I\left(v\in\mathcal{D}_n\right)/n$ is a feasible choice. Also, the feasible region is clearly compact, so the dual problem is always feasible and by strong duality its optimal value coincides with that of the primal problem, see Bertsimas *et al.* [2011, 2013] and Appendix in Chapter 2.

The dual problem associated to (4.4) is given by

$$\begin{aligned} &\min \sum_{v \in \mathcal{D}_N} \gamma(v)/n + \lambda\delta \\ &\text{s.t. } \gamma(v) \geq l(u,\beta) - \lambda c(u,v) \ \ \forall \ \ (u,v) \in \mathcal{X}_N \times \mathcal{D}_n \\ &\gamma(v) \in \mathbb{R} \ \forall \ v \in \mathcal{D}_n, \ \lambda \geq 0 \ . \end{aligned} \tag{4.5}$$

Maximizing over $u \in \mathcal{X}_N$ in the inequality constraint, for each v, and using the fact that we are minimizing the objective function, we obtain that (4.5) can be simplified to

$$\mathbb{E}_{P_n}[\max_{u \in \mathcal{X}_N} \{l(u,\beta) - \lambda c(u,(X,Y)) + \lambda\delta^*\}] \ .$$

Consequently, defining $\phi(X,Y,\beta,\lambda) = \max_{u \in \mathcal{X}_N} \{l(u,\beta) - \lambda c(u,(X,Y)) + \lambda\delta^*\}$, we have that (4.1) is equivalent to

$$\min_{\lambda \geq 0, \beta} \mathbb{E}_{P_n}[\phi(X,Y,\beta,\lambda)] \ . \tag{4.6}$$

Moreover, if we assume that $l(u,\cdot)$ is a convex function, then we have that the mapping $(\beta,\lambda) \hookrightarrow l(u,\beta) - \lambda c(u,(X,Y)) + \lambda\delta^*$is convex for each u and therefore, $(\beta,\lambda) \hookrightarrow \phi(X,Y,\beta,\lambda)$, being the maximum of convex mappings is also convex.

A natural approach consists in directly applying stochastic sub-gradient descent (see Boyd and Vandenberghe [2004]; Ram *et al.* [2010]). Unfortunately, this would involve performing the maximization over all $u \in \mathcal{X}_N$ in each iteration. This approach could be prohibitively expensive in typical machine learning applications where N is large.

So, instead, we perform a standard smoothing technique, namely, we introduce

$\epsilon > 0$ and define

$$\phi_\epsilon(X, Y, \beta, \lambda) = \lambda\delta^* + \epsilon \log\left(\sum_{u \in \mathcal{X}_N} \exp\left(\{l(u, \beta) - \lambda c(u, (X, Y))\}/\epsilon\right)\right).$$

It is easy to verify (using Hölder inequality) that $\phi_\epsilon(X, Y, \cdot)$ is convex and it also follows that

$$\phi(X, Y, \beta, \lambda) \leq \phi_\epsilon(X, Y, \beta, \lambda) \leq \phi(X, Y, \beta, \lambda) + \log(|\mathcal{X}_N|)\epsilon.$$

Hence, we can choose $\epsilon = O(1/\log N)$ in order to control the bias incurred by replacing ϕ by ϕ_ϵ. Then, defining

$$\tau_\epsilon(X, Y, \beta, \lambda, u) = \exp\left(\{l(u, \beta) - \lambda c(u, (X, Y))\}/\epsilon\right),$$

we have (assuming differentiability of $l(u, \beta)$) that

$$\begin{aligned} \nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda) = & \quad \frac{\sum_{u \in \mathcal{X}_N} \tau_\epsilon(X, Y, \beta, \lambda, u) \nabla_\beta l(u, \beta)}{\sum_{v \in \mathcal{X}_N} \tau_\epsilon(X, Y, \beta, \lambda, v)}, \\ \frac{\partial \phi_\epsilon(X, Y, \beta, \lambda)}{\partial \lambda} = & \quad \delta^* - \frac{\sum_{u \in \mathcal{X}_N} \tau_\epsilon(X, Y, \beta, \lambda, u) c(u, (X, Y))}{\sum_{v \in \mathcal{X}_N} \tau_\epsilon(X, Y, \beta, \lambda, v)}. \end{aligned} \tag{4.7}$$

In order to make use of the gradient representations (4.7) for the construction of a stochastic gradient descent algorithm, we must construct unbiased estimators for $\nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda)$ and $\partial \phi_\epsilon(X, Y, \beta, \lambda)/\partial\lambda$, given (X, Y). This can be easily done if we assume that one can simulate directly $u \in \mathcal{X}_N$ with probability proportional to $\tau(X, Y, \beta, \lambda, u)$. Because of the potential size of $\mathcal{X}_N$ and specially because such distribution depends on (X, Y) sampling with probability proportional to $\tau(X, Y, \beta, \lambda, u)$ can be very time consuming.

So, instead, we apply a strategy discussed in Blanchet and Glynn [2015] and explained in Section 2.2.1, which produces random variables $\Lambda(X, Y, \beta, \lambda)$ and $\Gamma(X, Y, \beta, \lambda)$, which can be simulated easily by drawing i.i.d. samples from the uniform distribution over $\mathcal{X}_N$, and such that

$$\mathbb{E}(\Lambda(X, Y, \beta, \lambda) | X, Y) = \partial_\lambda \phi_\epsilon(X, Y, \beta, \lambda),$$
$$\mathbb{E}(\Gamma(X, Y, \beta, \lambda) | X, Y) = \nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda).$$

Using this pair of random variables, then we apply the stochastic gradient descent recursion

$$\begin{aligned} \beta_{k+1} &= \beta_k - \alpha_{k+1}\Gamma(X_{k+1}, Y_{k+1}, \beta_k, \lambda_k), \\ \lambda_{k+1} &= (\lambda_k - \alpha_{k+1}\Lambda(X_{k+1}, Y_{k+1}, \beta_k, \lambda_k))^+, \end{aligned} \tag{4.8}$$

where learning sequence, $\alpha_k > 0$ satisfies the standard conditions, namely, $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, see Shapiro and Dentcheva [2014b].

We apply a technique from Blanchet and Glynn [2015], which originates from Multilevel Monte Carlo introduced in Giles [2008, 2015], and associated randomization methods McLeish [2011]; Rhee and Glynn [2015].

First, define $\bar{P}_N$ to be the uniform measure on $\mathcal{X}_N$ and let W be a random variable with distribution $\bar{P}_N$. Note that, given (X, Y),

$$\begin{aligned} \nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda) &= \frac{\mathbb{E}_{\bar{P}_N}(\tau_\epsilon(X, Y, \beta, \lambda, W) \nabla_\beta l(W, \beta) \mid X, Y)}{\mathbb{E}_{\bar{P}_N}(\tau_\epsilon(X, Y, \beta, \lambda, W) \mid X, Y)}, \\ \partial_\lambda \phi_\epsilon(X, Y, \beta, \lambda) &= \delta^* - \frac{\mathbb{E}_{\bar{P}_N}(\tau_\epsilon(X, Y, \beta, \lambda, W) c(W, (X, Y)) \mid X, Y)}{\mathbb{E}_{\bar{P}_N}(\tau_\epsilon(X, Y, \beta, \lambda, W) \mid X, Y)}. \end{aligned}$$

Note that both gradients can be written in terms of the ratios of two expectations. The

following results from Blanchet and Glynn [2015] can be used to construct unbiased estimators of functions of expectations. The function of interest in our case is the ratio of expectations.

Let us define:

$$h_0(W) = \tau_\epsilon(X, Y, \beta, \lambda, W),$$
$$h_1(W) = h_0(W)\, c(W, (X, Y)),$$
$$h_2(W) = h_0(W)\, \nabla_\beta l(W, \beta).$$

Then, we can write the gradient estimator as

$$\partial_\lambda \phi_\epsilon(X, Y, \beta, \lambda) = \frac{E_{\bar{P}_N}(h_1(W) \mid X, Y)}{E_{\bar{P}_N}(h_0(W) \mid X, Y)}, \text{ and } \nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda) = \frac{E_{\bar{P}_N}(h_2(W) \mid X, Y)}{E_{\bar{P}_N}(h_0(W) \mid X, Y)}.$$

The procedure developed in Blanchet and Glynn [2015] proceeds as follows. First, define for a given $h(W)$, and $n \geq 0$, the average over odd and even labels to be

$$\bar{S}^E_{2^n}(h) = \frac{1}{2^n}\sum_{i=1}^{2^n} h(W_{2i}), \quad \bar{S}^O_{2^n}(h) = \frac{1}{2^n}\sum_{i=1}^{2^n} h(W_{2i-1}),$$

and the total average to be $\bar{S}_{2^{n+1}}(h) = \frac{1}{2}\left(\bar{S}^E_{2^n}(h) + \bar{S}^O_{2^n}(h)\right)$. We then state the following algorithm for sampling unbiased estimators of $\partial_\lambda \phi_\epsilon(X, Y, \beta, \lambda)$ and $\nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda)$ in Algorithm 4.1.

Algorithm 4.1 Unbiased Gradient

1: Given (X, Y, β) the function outputs (Λ, Γ) such that $E(\Lambda) = \partial_\lambda \phi_\epsilon(X, Y, \beta, \lambda)$ and $E(\Gamma) = \nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda)$.
2: **Step1:** Sample G from geometric distribution with success parameter $p_G = 1 - 2^{-3/2}$.
3: **Step2:** Sample $W_0, W_1, ..., W_{2^{G+1}}$ i.i.d. copies of W independent of G.
4: **Step3:** Compute

$$\Delta^\lambda = \frac{\bar{S}_{2^{G+1}}(h_1)}{\bar{S}_{2^{G+1}}(h_0)} - \frac{1}{2}\left(\frac{\bar{S}^O_{2^{G+1}}(h_1)}{\bar{S}^O_{2^{G+1}}(h_0)} + \frac{\bar{S}^E_{2^G}(h_1)}{\bar{S}^E_{2^G}(h_0)}\right),$$
$$\Delta^\beta = \frac{\bar{S}_{2^{G+1}}(h_2)}{\bar{S}_{2^{G+1}}(h_0)} - \frac{1}{2}\left(\frac{\bar{S}^O_{2^{G+1}}(h_2)}{\bar{S}^O_{2^{G+1}}(h_0)} + \frac{\bar{S}^E_{2^G}(h_2)}{\bar{S}^E_{2^G}(h_0)}\right).$$

5: **Output:**

$$\Lambda = \delta^* - \frac{\Delta^\lambda}{p_G(1-p_G)^G} - \frac{h_1(W_0)}{h_0(W_0)}, \qquad \Gamma = \frac{\Delta^\beta}{p_G(1-p_G)^G} + \frac{h_2(W_0)}{h_0(W_0)}.$$

4.4 Error Improvement of Our SSL-DRO Formulation

Our goal in this section is to intuitively discuss why, owing to the inclusion of the constraint $P \in \mathcal{P}(\mathcal{X}_N)$, we expect desirable generalization properties of the SSL-DRO formulation (4.1). Moreover, our intuition suggests strongly why our SSL-DRO formulation should possess better generalization performance than natural supervised counterparts. We restrict the discussion for logistic regression due to the simple form of regularization connection we will make in Equation (4.10), however, the error improvement discussion should also apply to general supervised learning setting.

As discussed in the Introduction using the game-theoretic interpretation of (4.1), by introducing $\mathcal{P}(\mathcal{X}_N)$, the SSL-DRO formulation provides a mechanism for choosing β which focuses on potential out-of-sample scenarios which are more relevant based on available evidence

Suppose that the constraint $P \in \mathcal{P}(\mathcal{X}_N)$ was not present in the formulation. So, the inner maximization in (4.1) is performed over all probability measures $\mathcal{P}\left(\mathbb{R}^{d+1}\right)$ (supported on some subset of $\mathbb{R}^{d+1}$). As indicated earlier, we assume that $l(X, Y; \cdot)$ is strictly convex and differentiable, so the first order optimality condition

$$\mathbb{E}_P\left(\nabla_\beta l(X, Y; \beta)\right) = 0$$

characterizes the optimal choice of β assuming the validity of the probabilistic model P. It is natural to assume that there exists an actual model underlying the generation of the training data, which we denote as P_∞. Moreover, we may also assume that there exists a unique β^* such that $\mathbb{E}_{P_\infty}\left(\nabla_\beta l(X, Y; \beta^*)\right) = 0$.

The set

$$\mathcal{M}(\beta_*) = \left\{P \in \mathcal{P}\left(\mathbb{R}^{d+1}\right) : \mathbb{E}_P\left(\nabla_\beta l(X, Y; \beta^*)\right) = 0\right\}$$

corresponds to the family of all probability models which correctly estimate β^*. Clearly, $P_\infty \in \mathcal{M}(\beta_*)$, whereas, typically, $P_n \notin \mathcal{M}(\beta_*)$. Moreover, if we write $\mathcal{X}_\infty = supp(P_\infty)$ we have that

$$P_\infty \in m(N, \beta^*) := \left\{P \in \mathcal{P}(\mathcal{X}_\infty) : \quad \mathbb{E}_P\left(\nabla_\beta l(X, Y; \beta^*)\right) = 0\right\} \subset \mathcal{M}(\beta_*).$$

Since $\mathcal{X}_N$ provides a sketch of $\mathcal{X}_\infty$, then we expect to have that the extremal (i.e. worst case) measure, denoted by P_N^*, will be in some sense a better description of P_∞. Figure 4.2 provides a pictorial representation of the previous discussion. In the absence of the constraint $P \in \mathcal{P}(\mathcal{X}_N)$, the extremal measure chosen by nature can be interpreted as a projection of P_n onto $\mathcal{M}(\beta_*)$. In the presence of the constraint $P \in \mathcal{P}(\mathcal{X}_N)$, we can see that P_N^* may bring the learning procedure closer to P_∞. Of course, if N is not large enough, the schematic may not be valid because one may

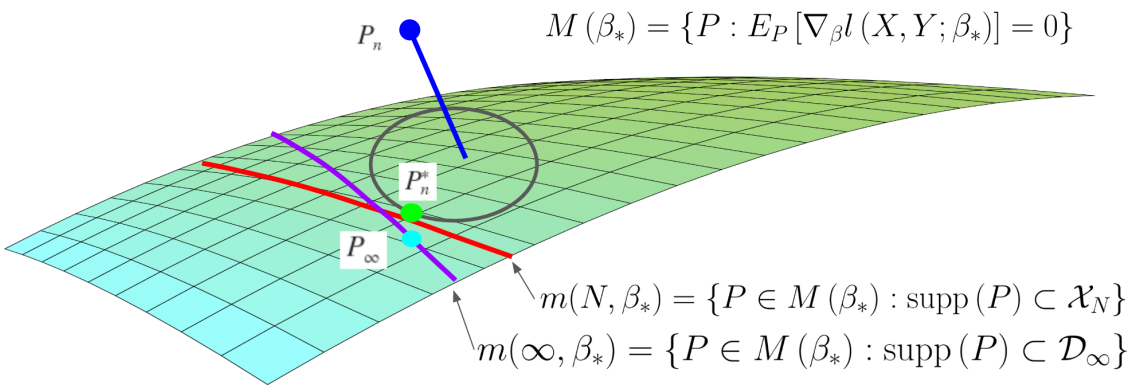

Figure 4.2: Pictorial representation of the role that the support constraint plays in the SSL-DRO approach and how its presence enhances the out-of-sample performance.

actually have $m(N, \beta^*) = \varnothing$.

The previous discussion is useful to argue that our SSL-DRO formulation should be superior to the data-driven DRO formulation which is not informed by the unlabeled data. But this comparison may not directly apply to alternative supervised procedures that are mainstream in machine learning, which should be considered as the natural benchmark to compare with. Fortunately, replacing the constraint that $P \in \mathcal{P}(\mathcal{X}_N)$ by $P \in \mathcal{P}\left(\mathbb{R}^{d+1}\right)$ in the data-driven DRO formulation recovers exactly supervised learning algorithms such as regularized logistic regression.

Recall from Chapter 2 that if $l(x, y, \beta) = \log(1 + \exp(-y \cdot \beta^T x))$ and if we define

$$c\left((x, y), (x', y')\right) = \begin{cases} \|x - x'\|_q, & \text{if } y = y' \\ \infty, & \text{otherwise.} \end{cases}, \tag{4.9}$$

for $q \geq 1$ then, according to Theorem 2.2 in Chapter 2, we have that

$$\min_{\beta} \max_{D_c(P, P_n) \leq \bar{\delta}} \mathbb{E}_P[l(X, Y, \beta)] = \min_{\beta \in \mathbb{R}^d} \left\{ \mathbb{E}_{P_n}[l(X, Y, \beta)] + \bar{\delta} \|\beta\|_p \right\}, \tag{4.10}$$

where q satisfies $1/p + 1/q = 1$. Formulation (4.1) is, therefore, the natural SSL extension of the standard regularized logistic regression estimator.

We conclude that, for logistic regression, SSL-DRO as formulated in (4.1), is a natural SSL extension of the standard regularized logistic regression estimator, which would typically induce superior generalization abilities over its supervised counterparts, and similar discussion should apply to most supervised learning methods.

4.5 Numerical Experiments

We proceed to numerical experiments to verify the performance of our SSL-DRO method empirically using six binary classification real data sets from UCI machine learning data base Lichman [2013].

We consider our SSL-DRO formulation based on logistic regression and compare with other state-of-the-art logistic regression based SSL algorithms, entropy regularized logistic regression with L_1 regulation (ERLRL1) Grandvalet and Bengio [2005] and regularized logistic regression based self-training (STLRL1) Li *et al.* [2008]. In addition, we also compare with its supervised counterpart, which is regularized logistic regression (LRL1). For each iteration of a data set, we randomly split the data into labeled training, unlabeled training and testing set, we train the models on training sets and evaluate the testing error and accuracy with testing set. We report the mean and standard deviation for training and testing error using log-exponential loss and the average testing accuracy, which are calculated via 200 independent experiments for each data set. We summarize the detailed results, the basic information of the data sets, and our data split setting in Table 4.1.

We can observe that our SSL-DRO method has the potential to improve upon these state-of-the-art SSL algorithms.

		breast cancer	banknote	qsar	magic	minibone	spambase
LRL1	Train	$.185 \pm .123$	$.080 \pm .030$	$.614 \pm .038$	$.548 \pm .087$	$.401 \pm .167$	$.470 \pm .040$
	Test	$.428 \pm .338$	$.340 \pm .228$	$.755 \pm .019$	$.610 \pm .050$	$.910 \pm .131$	$.588 \pm .141$
	Accur	$.929 \pm .023$	$.930 \pm .042$	$.646 \pm .036$	$.665 \pm .045$	$.717 \pm .041$	$.811 \pm .034$
ERLRL1	Train	$.019 \pm .010$	$.032 \pm .030$	$.249 \pm .050$	$2.37 \pm .987$	$.726 \pm .353$	$.008 \pm .028$
	Test	$.265 \pm .146$	$.793 \pm .611$	$.720 \pm .029$	4.28 ± 1.51	$1.98 \pm .678$	$.505 \pm .108$
	Accur	$.944 \pm .018$	$.920 \pm .047$	$.731 \pm .026$	$.721 \pm .056$	$.708 \pm .071$	$.883 \pm .018$
STLRL1	Train	$.089 \pm .019$	$.115 \pm .113$	$.498 \pm .120$	$3.05 \pm .987$	$1.50 \pm .706$	$.370 \pm .082$
	Test	$.672 \pm .034$	4.00 ± 2.78	$2.37 \pm .860$	8.03 ± 1.51	$4.81 \pm .732$	$1.465 \pm .316$
	Accur	$.955 \pm .023$	$.919 \pm .004$	$.694 \pm .038$	$.692 \pm .056$	$.704 \pm .033$	$.843 \pm .023$
SSL-DRO	Train	$.045 \pm .023$	$.101 \pm .035$	$.402 \pm .039$	$.420 \pm .075$	$.287 \pm .047$	$.221 \pm .028$
	Test	$.120 \pm .029$	$.194 \pm .067$	$.555 \pm .025$	$.561 \pm .039$	$.609 \pm .054$	$.333 \pm .012$
	Accur	$.956 \pm .016$	$.930 \pm .037$	$.734 \pm .025$	$.733 \pm .034$	$.710 \pm .032$	$.892 \pm .009$
Num Predictors		30	4	30	10	20	56
Labeled Size		40	20	80	30	30	150
Unlabeled Size		200	600	500	9000	5000	1500
Testing Size		329	752	475	9990	125034	2951

Table 4.1: Numerical Experiments on real data sets for SSL.

4.6 Discussion on the Size of the Uncertainty Set

One of the advantages of DRO formulations such as Equation (4.1) and Equation (4.10) is that they lead to a natural criterion for the optimal choice of the parameter δ^* or, in the case of Equation (4.10), the choice of $\bar{\delta}$ (which incidentally corresponds to the regularization parameter). The optimality criterion that we use to select the size of δ^* is motivated by Figure 4.2.

First, interpret the uncertainty set

$$\mathcal{U}_\delta\left(P_n, \mathcal{X}_N\right) = \left\{P \in \mathcal{P}\left(\mathcal{X}_N\right) : D_c\left(P, P_n\right) \leq \delta\right\}$$

as the set of plausible models which are consistent with the empirical evidence encoded in P_n and $\mathcal{X}_N$. Then, for every plausible model P, we can compute

$$\beta\left(P\right) = \arg\min \mathbb{E}_P[l\left(X, Y, \beta\right)]$$

and therefore the set

$$\Lambda_\delta\left(P_n, \mathcal{X}_N\right) = \{\beta\left(P\right) = \arg\min \mathbb{E}_P[l\left(X, Y, \beta\right)] : P \in \mathcal{U}_\delta\left(P_n, \mathcal{X}_N\right)\}$$

can be interpreted as a confidence region. It is then natural to select a confidence level $\alpha \in (0, 1)$ and compute $\delta^* := \delta^*_{N,n}$ by solving

$$\min\{\delta : P\left(\beta^* \in \Lambda_\delta\left(P_n, \mathcal{X}_N\right)\right) \geq 1 - \alpha\}. \tag{4.11}$$

Similarly, for the supervised version, we can select $\bar{\delta} = \bar{\delta}_n$ by solving the problem

$$\min\{\delta : P\left(\beta^* \in \Lambda_\delta\left(P_n, \mathbb{R}^{d+1}\right)\right) \geq 1 - \alpha\}. \tag{4.12}$$

It is easy to see that $\bar{\delta}_n \leq \delta^*_{N,n}$. Now, we let $N = \gamma n$ for some $\gamma > 0$ and consider $\delta^*_{N,n}$, $\bar{\delta}_n$ as $n \to \infty$. This analysis is relevant because we are attempting to sketch $supp\left(P_\infty\right)$ using the set $\mathcal{X}_N$, while considering large enough plausible variations to be able to cover β^* with $1 - \alpha$ confidence.

More precisely, following the discussion in Chapter 2 for the supervised case in finding $\bar{\delta}_n$ in Equation (4.11) using Robust Wasserstein Profile (RWP) function, solving Equation (4.12) for $\delta^*_{N,n}$ is equivalent to finding the $1-\alpha$ quantile of the asymptotic distribution of the RWP function, defined as

$$R_n\left(\beta\right) = \min_\pi \Big\{ \sum_{u \in \mathcal{X}_n} \sum_{v \in \mathcal{D}_n} c(u, v)\pi(u, v), \sum_{u \in \mathcal{X}_n} \pi(u, v) = \frac{1}{n}, \forall v \mathcal{D}_n, \tag{4.13}$$
$$\pi \subset \mathcal{P}\left(\mathcal{X}_n \times \mathcal{D}_n\right), \sum_{u \in \mathcal{X}_n} \sum_{v \in \mathcal{D}_n} \nabla_\beta l\left(u; \beta\right) \pi(u, v) = 0. \Big\}.$$

The RWP function is the distance, measured by the optimal transport cost func-

tion, between the empirical distribution and the manifold of probability measures for which β_* is the optimal parameter. A pictorial representation is given in Figure 4.2. Additional discussion on the RWP function and its interpretations can be found in Chapter 2 and Chapter 3.

In the setting of the DRO formulation for Equation (4.10) it is shown in Chapter 2, that $\bar{\delta}_n = O\left(n^{-1}\right)$ for Equation (4.10) as $n \to \infty$. Intuitively, we expect that if the predictive variables possess a positive density supported in a lower dimensional manifold of dimension $\bar{d} < d$, then sketching $supp\left(P_\infty\right)$ with $O\left(n\right)$ data points will leave relatively large portions of the manifold unsampled (since, on average, $O\left(n^{\bar{d}}\right)$ sampled points are needed to be within distance $O\left(1/n\right)$ of a given point in box of unit size in $\bar{d}$ dimensions). The optimality criterion will recognize this type of discrepancy between $\mathcal{X}_N$ and $supp\left(P_\infty\right)$. Therefore, we expect that $\delta^*_{\gamma n,n}$ will converge to zero at a rate which might deteriorate slightly as $\bar{d}$ increases.

This intuition is given rigorous support in Theorem 4.1 for the case of linear regression with square loss function and L_2 cost function for DRO. In turn, Theorem 4.1 follows as a corollary to the results in Chapter 3. Detailed assumptions are given in the appendix.

Theorem 4.1. Assume the linear regression model $Y = \beta^* X + e$ with square loss function, i.e. $l\left(X, X; \beta\right) = \left(Y - \beta^T X\right)^2$, and transport cost

$$c\left(\left(x, y\right), \left(x', y'\right)\right) = \left\|x - x'\right\|_2^2 I_{y=y'} + \infty I_{y \neq y'}.$$

Assume $N = \gamma n$ and under mild assumptions on (X, Y), if we denote $\tilde{Z} \sim \mathcal{N}\left(0, E[V_1]\right)$, we have:

- When $d = 1$,

$$nR_n(\beta_*) \Rightarrow \kappa_1 \chi_1^2.$$

- When $d = 2$,

$$nR_n(\beta_*) \Rightarrow F_2\left(\tilde{Z}\right),$$

where $F_2(\cdot)$ is a continuous function and $F_2(z) = O(\|z\|_2^2)$ as $\|z\|_2 \to \infty$.

- When $d \geq 3$,

$$n^{1/2+\frac{3}{2d+2}} R_n(\beta_*) \Rightarrow F_d\left(\tilde{Z}\right),$$

where $F_d(\cdot)$ is a continuous function (depending on d) and $F_d(z) = O\big(\|z\|_2^{d/2+1}\big)$.

4.7 Conclusions

We have shown that our SSL-DRO, as a semi-supervised method, is able to enhance the generalization predicting power versus its supervised counterpart. Our numerical experiments show superior performance of our SSL-DRO method when compared to state-of-the-art SSL algorithms such as ERLRL1 and STLRL1. We would like to emphasize that our SSL-DRO method is not restricted to linear and logistic regressions. As we can observe from the DRO formulation and the algorithm. If a learning algorithm has an accessible loss function and the loss gradient can be computed, we are able to formulate the SSL-DRO problem and benefit from unlabeled information. Finally, we discussed a stochastic gradient descent technique for solving DRO problems such as (4.1), which we believe can be applied to other settings in which the gradient is a non-linear function of easy-to-sample expectations.

ADDITIONAL MATERIAL TO CHAPTER 4 In this additional material for SSL-DRO chapter, we will provide technical details for Theorem 4.1. In Section APPENDIX 4.A, we first state the general assumptions to guarantee the validity of the asymptotically optimal selection for the distributional uncertainty size in

Section 4.A.1, and in Section 4.A.2 we provide a roadmap for the proof of Theorem 4.1. In Section 4.B, we revisit Theomre 4.1 and provide a more formal statement in Section 4.B.1 and a detailed proof using the techniques in Chapter 3 in Section 4.B.2.

4.A: Technical Details for Theorem 4.1

In this appendix section, we first state the general assumptions to guarantee the validity of the asymptotically optimal selection for the distributional uncertainty size in Section 4.A.1. In Section 4.A.2 we provide a roadmap for the proof of Theorem 4.1.

4.A.1: Assumptions of Theorem 4.1

For linear regression model, let us assume we have a collection of labeled data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ and a collection of unlabeled data $\{X_i\}_{i=n+1}^N$. We consider the set $\mathcal{X}_N = \{X_i\}_{i=1}^N \times \{Y_i\}_{i=1}^n$, to be the cross product of all the predictors from labeled and unlabeled data and the labeled responses. In order to have proper asymptotic results holds for the RWP function, we require some mild assumptions on the density and moments of (X, Y) and estimating equation $\nabla_\beta l\left(X, Y; \beta\right) = \left(Y - \beta_*^T\right) X$. We state them explicitly as follows:

A) We assume the predictors X_i's for the labeled and unlabeled data are i.i.d. from the same distribution with positive differentiable density $f_X(\cdot)$ with bounded bounded gradients.

B) We assume the $\beta_* \in \mathbb{R}^d$ is the true parameter and under null hypothesis of the linear regression model satisfying $Y = \beta_*^T X + e$, where e is a random error independent of X.

C) We assume $\mathbb{E}\left[X^T X\right]$ exists and is positive definite and $\mathbb{E}\left[e^2\right] < \infty$.

D) For the true model of labeled data, we have $\mathbb{E}_{P_*}\left[X\left(Y-\beta_*^T X\right)\right]=0$ (where P_* denotes the actual population distribution which is unknown).

The first two assumptions, namely Assumption A and B, are elementary assumptions for linear regression model with an additive independent random error. The requirements for the differentiable positive density for the predictor X, is because when $d \geq 3$, the density function appears in the asymptotic distribution. Assumption C is a mild requirement on the moments exist for predictors and error, and Assumption D is to guarantee true parameter β_* could be characterized via first order optimality condition, i.e. the gradient of the square loss function. Due to the simple structure of the linear model, with the above four assumptions, we can prove Theorem 4.1 and we show a sketch in the following subsection.

4.A.2: Sketch of the Proof of Theorem 4.1

Theorem 4.1 is a corollary of Theorem 3.3 in Chapter 3, although its proof requires some adaptations. The proof of Theorem 4.1 follows the 6-step procedure explained in Section 3.3 of Chapter 3. We highlight the main differences in deriving the duality of the RWP function in this section. To make the chapter more self-contained, we include more technical details borrowed from Chapter 3 in the Section 4.B.

Sketch of the Proof of Theorem 4.1. **Deriving Strong Duality From for RWP Function.** For $u \in \mathcal{D}_n$ and $v \in \mathcal{X}_N$, let us denote u_x, u_y and v_x, v_y to be its subvectors for the predictor and response. By the definition of RWP function as in Equation

(4.13), we can write it as a linear program (LP), given as

$$R_n(\beta_*) = \min_{\pi} \Big\{ \sum_{u\in\mathcal{D}_n} \sum_{v\in\mathcal{X}_N} \pi(u,v) \left(\|u_x - v_x\|_2^2 I_{v_y=u_y} + \infty I_{v_y\neq u_y} \right) \ \textbf{s.t.} \quad \pi \in \mathcal{P}(\mathcal{X}_N \times \mathcal{D}_n),$$
$$\sum_{u\in\mathcal{D}_n} \sum_{v\in\mathcal{X}_N} \pi(u,v)\, v_x \left(v_y - \beta_*^T v_x \right) = 0, \sum_{v\in\mathcal{X}_N} \pi(u,v) = 1/n, \forall u \in \mathcal{D}_n. \Big\}$$

For as n large enough the LP is finite and feasible (because P_n approaches P_*, and P_* is feasible). Thus, for n large enough we can write

$$R_n(\beta_*) = \min_{\pi} \Big\{ \sum_{u\in\mathcal{D}_n} \sum_{v_x\in\{X_i\}_{i=1}^N} \pi(u,v_x) \|u_x - v_x\|_2^2 \ \ \textbf{s.t.} \quad \pi \in \mathcal{P}(\mathcal{X}_N \times \mathcal{D}_n)$$
$$\sum_{u\in\mathcal{D}_n} \sum_{v\in\mathcal{X}_N} \pi(u,v)\, v_x \left(u_y - \beta_*^T v_x \right) = 0, \sum_{v\in\mathcal{X}_N} \pi(u,v) = 1/n, \forall u \in \mathcal{D}_n. \Big\}$$

We can apply strong duality theorem for LP, see Luenberger [1973b], and write the RWP function in dual form:

$$\begin{aligned} R_n(\beta_*) &= \max_{\lambda} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{j=\overline{1,N}} \left\{ -\lambda^T X_j \left(Y_i - \beta_*^T X_j \right) + \|X_i - X_j\|_2^2 \right\} \right\}, \\ &= \max_{\lambda} \Big\{ \frac{1}{n} \sum_{i=1}^n -\lambda^T X_i \left(Y_i - \beta_*^T X_i \right) \\ &\qquad + \min_{j=\overline{1,N}} \left\{ \lambda^T X_i \left(Y_i - \beta_*^T X_j \right) - \lambda^T X_j \left(Y_i - \beta_*^T X_j \right) + \|X_i - X_j\|_2^2 \right\} \Big\}, . \end{aligned}$$

This finishes Step 1 as in the 6-step proving technique introduced in Section 3.3 of Chapter 3.

In Step 2 and Step 3, after rescaling the RWP function by n for $d = 1$ and 2 and rescaling by $n^{\frac{1}{2}+\frac{3}{2d+2}}$ for $d \geq 3$, we can quantify the difference between the inner

minimization problem for each i,

$$\min_{j=\overline{1,N}} \left\{ \lambda^T X_i \left(Y_i - \beta_*^T X_i \right) - \lambda^T X_j \left(Y_i - \beta_*^T X_j \right) + \| X_i - X_j \|_2^2 \right\}$$

and its lower bound,

$$\min_{a} \left\{ \lambda^T X_i \left(Y_i - \beta_*^T X_i \right) - \lambda^T a \left(Y_i - \beta_*^T a \right) + \| X_i - a \|_2^2 \right\},$$

by defining a family of auxiliary, weakly dependent, Poisson point processes (indexed by i).

Applying the results in Step 3, we can prove the asymptotic distribution for $d = 1$ in Step 4, $d = 2$ in Step 5, and $d \geq 3$ in Step 6 using the Central Limit Theorem (CLT) and the Continuous Mapping Theorem. More details are shown in the Section 4.B.2.

□

4.B: Additional Techinical Details for Theorem 4.1

In this supplementary material, we will restate Theorem 4.1 more explicitly to show how the asymptotic distribution varies for different dimension d in Section 4.B.1. In Section 4.B.2, we will feed more technical details in proving Theorem 4.1.

4.B.1: Revisit Theorem 4.1

In this section, we revisit the asymptotic result for optimally choosing uncertainty size for semi-supervised learning for the linear regression model. We assume that, under the null hypothesis, $Y = \beta_*^T X + e$, where $X \in \mathbb{R}^d$ is the predictors, e is independent

of X as random error, and $\beta_* \in \mathbb{R}^d$ is the true parameter. We consider the square loss function and assume that β_* is the minimizer to the square loss function, i.e.

$$\beta_* = \arg\min_{\beta} \mathbb{E}\left[\left(Y - \beta^T X\right)^2\right].$$

If we can assume the second-moment exists for X and e, then we can switch the order of expectation and derivative w.r.t. β, then optimal β could be uniquely characterized via the first order optimality condition,

$$\mathbb{E}\left[X\left(Y - \beta_*^T X\right)\right] = 0.$$

As we discussed in Section 4.6, the optimal distributional uncertainty size $\delta_{n,N}^*$ at confidence level $1-\alpha$, is simply the $1-\alpha$ quantile of the RWP function defined in Equation (4.13). In turn, the asymptotic limit of the RWP function is characterized in Theorem 1, which we restate more explicitly here.

Restate of Theorem 4.1 in Section 4.6: For linear regression model we defined above and square loss function, if we take cost function for DRO formulation to be

$$c\left((x,y),(x',y')\right) = \|x - x'\|_2^2 I_{y=y'} + \infty I_{y\neq y'}.$$

If we assume Assumptions A,B, and D stated in Section **??** to be true and number of unlabeled data satisfing $N = \gamma n$. Furthermore, let us denote: $V_i = \left(e_i I - X_i \beta_*^T\right)\left(e_i I - \beta_* X_i^T\right)$, where $e_i = Y_i - \beta_*^T X_i$ being the residual under the null hypothesis. Then, we have:

- When $d = 1$,

$$nR_n(\beta_*) \Rightarrow \frac{\mathbb{E}\left[X_1^2 e_1^2\right]}{\mathbb{E}\left[\left(e_1 - \beta_*^T X_1\right)^2\right]}\chi_1^2.$$

- When $d = 2$,

$$nR_n(\beta_*) \Rightarrow 2\tilde{\zeta}(\tilde{Z})^T\tilde{Z} - \tilde{\zeta}\left(\tilde{Z}\right)^T \tilde{G}_2\left(\tilde{\zeta}\left(\tilde{Z}\right)\right)\tilde{\zeta}\left(\tilde{Z}\right),$$

where $\tilde{Z} \sim \mathcal{N}(0, E[V_1])$, $\tilde{G}_2 : \mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}^2$ is a continuous mapping defined as

$$\tilde{G}_2(\zeta) = \mathbb{E}\left[V_1 \max\left(1 - \tau/(\zeta^T V_1 \zeta), 0\right)\right],$$

and $\tilde{\zeta} : \mathbb{R}^2 \to \mathbb{R}^2$ is a continuous mapping, such that $\tilde{\zeta}(\tilde{Z})$ is the unique solution to

$$\tilde{Z} = -\mathbb{E}\left[V_1 I_{(\tau \le \zeta^T V_1 \zeta)}\right]\zeta.$$

- When $d \ge 3$,

$$n^{1/2+\frac{3}{2d+2}} R_n(\beta_*) \Rightarrow -2\tilde{\zeta}(\tilde{Z})^T\tilde{Z} - \frac{2}{d+2}\tilde{G}_3\left(\tilde{\zeta}(\tilde{Z})\right),$$

where $\tilde{Z} \sim \mathcal{N}(0, E[V_1])$, $\tilde{G}_2 : \mathbb{R}^d \to \mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{G}_2(\zeta) = \mathbb{E}\left[\frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma(d/2+1)}\left(\zeta^T V_1 \zeta\right)^{d/2+1}\right],$$

and $\tilde{\zeta} : \mathbb{R}^d \to \mathbb{R}^d$ js a continuous mapping, such that $\tilde{\zeta}(\tilde{Z})$ is the unique solution to

$$\tilde{Z} = -\mathbb{E}\left[V_1 \frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma(d/2+1)}\left(\zeta^T V_1 \zeta\right)^d\right]\zeta.$$

4.B.2: Proof of Theorem 4.1

In this section, we complete the proof of Theorem 4.1 in addition to the scratch in Section **??**. As we discussed before, Theorem 4.1 could be treated as a non-trivial

corollary of Theorem 3.3 in Chapter 3 and the proving techniques follow the 6-step proof for Sample-out-of-Sample (SoS) Theorem, namely Theorem 3.1 and Theorem 3.3 in Chapter 3.

Proof of Theorem 4.1. We derived the duality formulation for RWP function in Section 4.A.2 as the Step 1 of the proof.

Step 2 and Step 3, When $d = 1$ and 2, we consider scaling the RWP function by n and let define $\zeta = \sqrt{n}\lambda/2$ and denote $W_n = n^{-1/2}\sum_{i=1}^{n} X_i e_i$, we have the scaled RWP function becomes,

$$\begin{aligned} nR_n(\beta_*) = \max_{\zeta} \big\{ & -\zeta^T W_n \\ & + \sum_{i=1}^{n} \min_{j=\overline{1,N}} \{-2\frac{\zeta^T}{\sqrt{n}} X_j \left(Y_i - \beta_*^T X_j\right) + 2\frac{\zeta^T}{\sqrt{n}} X_i \left(Y_i - \beta_*^T X_i\right) + \|X_i - X_j\|_2^2\}\big\}. \end{aligned}$$

For each fixed i, let us consider the inner minimization problem,

$$\min_{j=\overline{1,N}} \{-2\frac{\zeta^T}{\sqrt{n}} X_j \left(Y_i - \beta_*^T X_j\right) + 2\frac{\zeta^T}{\sqrt{n}} X_i \left(Y_i - \beta_*^T X_i\right) + \|X_i - X_j\|_2^2\}$$

Similar to Section 3.3 in Chapter 3, we would like to solve the minimization problem by first replacing X_j by a, which is a free variable without support constraint in $\mathbb{R}^d$, then quantify the gap. We then obtain a lower bound for the optimization problem via

$$\min_{a} \{-2\frac{\zeta^T}{\sqrt{n}} a \left(Y_i - \beta_*^T a\right) + 2\frac{\zeta^T}{\sqrt{n}} X_i \left(Y_i - \beta_*^T X_i\right) + \|X_i - a\|_2^2\}. \tag{4.14}$$

As we can observe in Equation (4.14), the coefficient of second order of a is of order $O\left(1/\sqrt{n}\right)$ for any fixed ζ, and the coefficients for the last term is always 1, it is easy to observe that, as n large enough, Equation (4.14) has an optimizer in the interior.

We can solve for the optimizer $a = \bar{a}_*(X_i, Y_i, \zeta)$ of the lower bound in Equation

(4.14) satisfying the first order optimality condition as

$$\begin{aligned}\bar{a}_*\left(X_i, Y_i, \zeta\right) - X_i = &\left(e_i I - \beta_*^T X_i\right) \frac{\zeta}{\sqrt{n}} \\ &+ \left(\beta_*^T \left(\bar{a}_*\left(X_i, Y_i, \zeta\right) - X_i\right) I - \left(\bar{a}_*\left(X_i, Y_i, \zeta\right) - X_i\right) \beta_*^T\right) \frac{\zeta}{\sqrt{n}}.\end{aligned} \tag{4.15}$$

Since the optimizer $\bar{a}_*\left(X_i, Y_i, \zeta\right)$ is in the interior, it is easy to notice from Equation (4.15) that $\bar{a}_*\left(X_i, Y_i, \zeta\right) - X_i = O\left(\frac{\|\zeta\|_2}{\sqrt{n}}\right)$. Plug in the estimate back into Equation (4.15) obtain

$$\bar{a}_*\left(X_i, Y_i, \zeta\right) = X_i + \left(e_i I - \beta_*^T X_i\right) \frac{\zeta}{\sqrt{n}} + O\big(\frac{\|\zeta\|_2^2}{n}\big). \tag{4.16}$$

Let us define $a_*\left(X_i, Y_i, \zeta\right) = X_i + \left(e_i I - \beta_*^T X_i\right) \frac{\zeta}{\sqrt{n}}$. Using Equation (4.16), we have

$$\left\|a_*\left(X_i, Y_i, \zeta\right) - \bar{a}_*\left(X_i, Y_i, \zeta\right)\right\|_2 = O\big(\frac{\|\zeta\|_2^2}{n}\big). \tag{4.17}$$

Then, for the optimal value function of lower bound of the inner optimization problem, we have:

$$\begin{aligned}&- 2\frac{\zeta^T}{\sqrt{n}} \bar{a}_*\left(X_i, Y_i, \zeta\right)\left(Y_i - \beta_*^T a\right) + 2\frac{\zeta^T}{\sqrt{n}} X_i\left(Y_i - \beta_*^T X_i\right) + \left\|X_i - \bar{a}_*\left(X_i, Y_i, \zeta\right)\right\|_2^2 \\ &= -2\frac{\zeta^T}{\sqrt{n}} a_*\left(X_i, Y_i, \zeta\right)\left(Y_i - \beta_*^T a\right) + 2\frac{\zeta^T}{\sqrt{n}} X_i\left(Y_i - \beta_*^T X_i\right) + \left\|X_i - a_*\left(X_i, Y_i, \zeta\right)\right\|_2^2 + O\big(\frac{\|\zeta\|_2^3}{n^{3/2}}\big) \\ &= \frac{\zeta^T V_i \zeta}{n} + O\big(\frac{\|\zeta\|_2^3}{n^{3/2}}\big).\end{aligned} \tag{4.18}$$

For the above equation, first equality is due to Equation (4.17) and the second equality is by the estimation of $\bar{a}_*\left(X_i, Y_i, \zeta\right)$ in Equation (4.16).

Then for each fixed i, let us define a point process

$$N_n^{(i)}(t,\zeta) = \#\left\{X_j : \|X_j - a_*(X_i, Y_i, \zeta)\|_2^2 \le t^{2/d}/n^{2/d}, X_j \neq X_i\right\}.$$

We denote $T_i(n)$ to be the first jump time of $N_n^{(i)}(t,\zeta)$, i.e.

$$T_i(n) = \inf\left\{t \ge 0 : N_n^{(i)}(t,\zeta) \ge 1\right\}.$$

It is easy to observe that, as n goes to infinity, we have

$$N_n^{(i)}(t,\zeta)\,|X_i \Rightarrow Poi\left(\Lambda(X_i,\zeta), t\right),$$

where $Poi\left(\Lambda(X_i,\zeta), t\right)$ denotes a Poisson point process with rate

$$\Lambda(X_i,\zeta) = \gamma f_X\left(X_i + \frac{\zeta}{2\sqrt{\zeta}}\right)\frac{\pi^{d/2}}{\Gamma(d/2+1)}.$$

Then, the conditional survival function for $T_i(n)$, i.e. $P\left(T_i(n) \ge t|X_i\right)$ is

$$P\left(T_i(n) \ge t|X_i\right) = \exp\left(-\Lambda\left(X_i,\zeta\right)t\right)\left(1 + O\left(1/n^{1/d}\right)\right),$$

and we can define τ_i to be the random variable with survival function being

$$P\left(\tau_i(n) \ge t|X_i\right) = \exp\left(-\Lambda\left(X_i,\zeta\right)t\right).$$

We can also integrate the dependence on X_i and define τ satisfying

$$P\left(\tau \ge t\right) = \mathbb{E}\left[\exp\left(-\Lambda\left(X_1,\zeta\right)t\right)\right].$$

Therefore,for $d = 1$ by the definition of $T_i(n)$ and the estimation in Equation (4.18), we have the scaled RWP function becomes

$$nR_n(\beta_*) = \max_{\zeta} \Big\{ -2\zeta W_n - \frac{1}{n}\sum_{i=1}^{n} \max\big(\zeta^T V_i \zeta - T_i(n)^2/n + O\big(\frac{\|\zeta\|_2^3}{n^{3/2}}\big), 0\big) \Big\}$$

The sequence of global optimizers is tight as $n \to \infty$, because according to Assumption C, $E(V_i)$ is assumed to be strictly positive definite with probability one. In turn, from the previous expression we can apply Lemma 3.1 in Chapter 3 and use the fact that the variable ζ can be restricted to compact sets for all n sufficiently large. We are then able to concludee

$$nR_n(\beta_*) = \max_{\zeta} \Big\{ -2\zeta^T W_n - \mathbb{E}\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2/n, 0\right)\right] \Big\} + o_p(1). \tag{4.19}$$

When $d = 2$, a similar estimation applies as for the case $d = 1$. the scaled RWP function becomes

$$nR_n(\beta_*) = \max_{\zeta} \Big\{ -2\zeta^T W_n - \mathbb{E}\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2, 0\right)\right] \Big\} + o_p(1). \tag{4.20}$$

For the case when $d \geq 3$, let us define $\zeta = \lambda/(2n^{\frac{3}{2d+2}})$. We follow a similar estimation procedure as in the cases $d = 1, 2$. We also define identical auxiliary Poisson point process, we can write the scaled RWP function to be

$$\begin{aligned} n^{\frac{1}{2}+\frac{3}{2d+2}} R_n(\beta_*) = \max_{\zeta} \Big\{ & -2\zeta^T W_n \\ & - n^{\frac{1}{2}+\frac{3}{2+2d}-\frac{2}{d}} \mathbb{E}\left[\max\left(n^{\frac{2}{2}-\frac{6}{2d+2}} \zeta^T V_i \zeta - T_i(n)^{3/d}, 0\right)\right] \Big\} + o_p(1). \end{aligned} \tag{4.21}$$

This addresses Step 2 and 3 in the proof.

Step 4: when $d = 1$, as $n \to \infty$, we have the scaled RWP function given in Equation (4.19). Let us use $G_1 : \mathbb{R} \to \mathbb{R}$ to denote a deterministic continuous function defined as

$$G_1(\zeta, n) = \mathbb{E}\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2/n, 0\right)\right].$$

By Assumption C, we know $\mathbb{E}V_i$ is positive, thus G_1 as a function of ζ is strictly convex. Thus the optimizer for the scaled RWP function could be uniquely characterized via the first order optimality condition, which is equivalent to

$$\zeta_n^* = -\frac{W_n}{\mathbb{E}[V_i]} + o_p(1), \text{ as } n \to \infty. \tag{4.22}$$

We plug in Equation (4.22) into Equation (4.19) and let $n \to \infty$. Applying the CLT for W_n and the continuous mapping theorem, we have

$$nR_n(\beta_*) = 2W_n^2/\mathbb{E}[V_1] - G_1\left(-\frac{W_n}{\mathbb{E}[V_1]}, n\right) + o_p(1) \Rightarrow \frac{\tilde{Z}^2}{E[V_1]} = \frac{\mathbb{E}[X_1^2 e_1^2]}{\mathbb{E}\left[(e_1 - \beta * X_1)^2\right]}\chi_1^2,$$

where $W_n \Rightarrow \tilde{Z}$ and $\tilde{Z} \sim \mathcal{N}\left(0, E\left[(e_1 - \beta * X_1)^2\right]\right)$.

We conclude the stated convergence for $d = 1$.

Step 5: when $d = 2$, as $n \to \infty$, we have the scaled RWP function given in Equation (4.20). Let us use $G_2 : \mathbb{R} \times \mathbb{N} \to \mathbb{R}$ to denote a deterministic continuous function defined as

$$G_2(\zeta, n) = \mathbb{E}\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2, 0\right)\right].$$

Following the same discussion as in Step 4 for the case $d = 1$, we know that the optimizer ζ_n^* can be uniquely characterized via first order optimality condition given

as

$$W_n = -\mathbb{E}\left[V_1 I_{(\tau \leq \zeta^T V_1 \zeta)}\right] \zeta + o_p(1), \text{ as } n \to \infty.$$

Since we know that the objective function is strictly convex there exist a continuous mapping, $\tilde{\zeta} : \mathbb{R}^2 \to \mathbb{R}^2$, such that $\tilde{\zeta}(W_n)$ is the unique solution to

$$W_n = -\mathbb{E}\left[V_1 I_{(\tau \leq \zeta^T V_1 \zeta)}\right] \zeta.$$

Then, we can plug-in the first order optimality condition to the value function, and the scaled RWP function becomes,

$$n\mathbb{R}_n(\beta_*) = 2\tilde{\zeta}(W_n)^T W_n - G_2\left(\tilde{\zeta}(W_n), n\right) + o_p(1).$$

Applying Lemma 3.2 in Chapter 3 we can show that as $n \to \infty$,

$$n\mathbb{R}_n(\beta_*) \Rightarrow 2\tilde{\zeta}(\tilde{Z})^T \tilde{Z} - \tilde{\zeta}\left(\tilde{Z}\right)^T \tilde{G}_2\left(\tilde{\zeta}\left(\tilde{Z}\right)\right) \tilde{\zeta}\left(\tilde{Z}\right)$$

where $\tilde{G}_2 : \mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}^2$ is a continuous mapping defined as

$$\tilde{G}_2(\zeta) = \mathbb{E}\left[V_1 \max\left(1 - \tau/(\zeta^T V_1 \zeta), 0\right)\right].$$

This concludes the claim for $d = 2$.

Step 6: when $d = 3$, as $n \to \infty$, we have the scaled RWP function given in Equation (4.21). Let us write $G_3 : \mathbb{R} \times \mathbb{N} \to \mathbb{R}$ to denote a deterministic continuous function defined as

$$G_3(\zeta, n) = n^{\frac{1}{2} + \frac{3}{2+2d} - \frac{2}{d}} \mathbb{E}\left[\max\left(n^{\frac{2}{2} - \frac{6}{2d+2}} \zeta^T V_i \zeta - T_i(n)^{3/d}, 0\right)\right].$$

Same as discussed in Step 4 and 5, the objective function is strictly convex and the optimizer could be uniquely characterized via first order optimality condition, i.e.

$$W_n = -\mathbb{E}\left[V_1 \frac{\pi^{d/2}\gamma f_X\left(X_1\right)}{\Gamma\left(d/2+1\right)}\left(\zeta^T V_1 \zeta\right)^d\right]\zeta + o_p(1), \text{ as } n \to \infty.$$

Since we know that the objective function is strictly convex, there exist a continuous mapping, $\tilde{\zeta} : \mathbb{R}^d \to \mathbb{R}^d$, such that $\tilde{\zeta}(W_n)$ is the unique solution to

$$W_n = -\mathbb{E}\left[V_1 \frac{\pi^{d/2}\gamma f_X\left(X_1\right)}{\Gamma\left(d/2+1\right)}\left(\zeta^T V_1 \zeta\right)^d\right]\zeta.$$

Let us plug-in the optimality condition and the scaled RWP function becomes

$$n^{\frac{1}{2}+\frac{3}{2d+2}} R_n\left(\beta_*\right) = -2\tilde{\zeta}(W_n)^T W_n - G_3\left(\tilde{\zeta}(W_n, n)\right) + o_p(1).$$

As $n \to \infty$, we can apply Lemma 3.2 in Chapter 3 to derive estimation for the RWP function and it leads to

$$n^{\frac{1}{2}+\frac{3}{2d+2}} R_n\left(\beta_*\right) \Rightarrow -2\tilde{\zeta}(\tilde{Z})^T \tilde{Z} - \frac{2}{d+2}\tilde{G}_3\left(\tilde{\zeta}(\tilde{Z})\right),$$

where $\tilde{G}_2 : \mathbb{R}^d \to \mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{G}_2\left(\zeta\right) = \mathbb{E}\left[\frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma\left(d/2+1\right)}\left(\zeta^T V_1 \zeta\right)^{d/2+1}\right].$$

This concludes the case when $d \geq 3$ and for Theorem 4.1. □

Chapter 5

Distributionally Robust Groupwise Regularization Estimator

In this Chapter, we will discuss a generalization of data-driven DRO method by exploring the flexibility of the choice of cost function. In Chapter 4, we were considering the flexibility of data-driven DRO formulation in restricting the candidate probability measures in constructing the distributional uncertainty set. The optimal transport discrepancy cost function considered in the former chapters, i.e. Chapter 2, Chapter 3, and Chapter 4, is using the Euclidean norm, i.e. $\|\cdot\|_p$. In this chapter, we will propose a groupwise norm, as we shall define in Equation (5.1), as cost function, which is trying to encode the side information of the predictors into data-driven DRO modeling.

Regularized estimators in the context of group variables have been applied successfully in model and feature selection in order to preserve interpretability. We formulate a data-driven Distributionally Robust Optimization (DRO) problem which recovers popular estimators, such as Group Square Root Lasso (GSRL). Our data-driven DRO formulation allows us to interpret GSRL as a game, in which we learn a regression

parameter while an adversary chooses a perturbation of the data. We wish to pick the parameter to minimize the expected loss under any plausible model chosen by the adversary - who, on the other hand, wishes to increase the expected loss. The regularization parameter turns out to be precisely determined by the amount of perturbation on the training data allowed by the adversary. In this chapter, we introduce a data-driven (statistical) criterion for the optimal choice of regularization, which we evaluate asymptotically, in closed form, as the size of the training set increases. Our easy-to-evaluate regularization formula is compared against cross-validation, showing good (sometimes superior) performance.

5.1 Introduction

Group Lasso (GR-Lasso) estimator is a generalization of the Lasso estimator (see Tibshirani [1996]). The method focuses on variable selection in settings where some predictive variables, if selected, must be chosen as a group. For example, in the context of the use of dummy variables to encode a categorical predictor, the application of the standard Lasso procedure might result in the algorithm including only a few of the variables but not all of them, which could make the resulting model difficult to interpret. Another example, where the GR-Lasso estimator is particularly useful, arises in the context of feature selection. Once again, a particular feature might be represented by several variables, which often should be considered as a group in the variable selection process.

The GR-Lasso estimator was initially developed for the linear regression case (see Yuan and Lin [2006]), but a similar group-wise regularization was also applied to logistic regression in Meier *et al.* [2008]. A brief summary of GR-Lasso technique type of methods can be found in Friedman *et al.* [2010].

Recently, Bunea *et al.* [2014] developed a variation of the GR-Lasso estimator, called the Group-Square-Root-Lasso (GSRL) estimator, which is very similar to the GR-Lasso estimator. The GSRL is to the GR-Lasso estimator what sqrt-Lasso, introduced in Belloni *et al.* [2011], is to the standard Lasso estimator. In particular, GSRL has a superior advantage over GR-Lasso, namely, that the regularization parameter can be chosen independently from the standard deviation of the regression error in order to guarantee the statistical consistency of the regression estimator (see Belloni *et al.* [2011], and Bunea *et al.* [2014]).

Our contribution in this chapter is to provide a data-driven DRO representation for the GSRL estimator, which is rich in interpretability and which provides insights to optimally select (using a natural criterion) the regularization parameter without the need of time-consuming cross-validation. We compute the optimal regularization choice (based on a simple formula we derive in this chapter) and evaluate its performance empirically. We will show that our method for the regularization parameter is comparable, and sometimes superior, to cross-validation.

In order to describe our contributions more precisely, let us briefly describe the GSRL estimator. We choose the context of linear regression to simplify the exposition, but an entirely analogous discussion applies to the context of logistic regression. Consider a given a set of training data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. The input $X_i \in \mathbb{R}^d$ is a vector of d predicting variables, and $Y_i \in \mathbb{R}$ is the response variable. We use (X, Y) to denote a generic sample from the training data set. It is postulated that

$$Y_i = X_i^T \beta^* + e_i,$$

for some $\beta^* \in \mathbb{R}^d$ and errors $\{e_1, ..., e_n\}$. Under suitable statistical assumptions (such as independence of the samples in the training data), one may be interested in

estimating β^*.

Underlying, we consider the square loss function, i.e. $l(x, y; \beta) = \left(y - \beta^T x\right)^2$, for the purpose of this discussion but this choice, as we shall see, is not necessary.

Throughout the chapter we will assume the following group structure for the space of predictors. There are $\bar{d} \leq d$ mutually exclusive groups, which form a partition. More precisely, suppose that $G_1, \ldots, G_{\bar{d}}$ satisfies that $G_i \cap G_j = \varnothing$ for $i \neq j$, that $G_1 \cup ... \cup G_{\bar{d}} = \{1, ..., d\}$, and the G_i's are non-empty. We will use g_i to denote the cardinality of G_i and shall write G for a generic set in the partition and let g denote the cardinality of G.

We shall denote by $x(G) \in \mathbb{R}^g$ the sub-vector $x \in \mathbb{R}^d$ corresponding to G. So, if $G = \{i_1, ..., i_g\}$, then $x(G) = \left(X_{i_1}, \ldots, X_{i_g}\right)^T$.

Next, given $p, s \geq 1$, and $\alpha \in \mathbb{R}^{\bar{d}}_{++}$ (i.e. $\alpha_i > 0$ for $1 \leq i \leq \bar{d}$) we define for each $x \in \mathbb{R}^d$,

$$\|x\|_{\alpha\text{-}(p,s)} = \left(\sum_{i=1}^{\bar{d}} \alpha_i^s \|x(G_i)\|_p^s \right)^{1/s}, \tag{5.1}$$

where $\|x(G_i)\|_p$ denotes the p-norm of $x(G_i)$ in $\mathbb{R}^{g_i}$. (We will study fundamental properties of $\|x\|_{\alpha\text{-}(p,s)}$ as a norm in Proposition 5.1.)

The GSRL estimator takes the form

$$\min_{\beta} \sqrt{\frac{1}{n} \sum_{i=1}^{n} l(X_i, Y_i; \beta)} + \lambda \|\beta\|_{\tilde{g}^{-1}\text{-}(2,1)} = \min_{\beta} \left(\mathbb{E}_{P_n}^{1/2} \left[l(X, Y; \beta) \right] + \lambda \|\beta\|_{\sqrt{\tilde{g}}\text{-}(2,1)} \right),$$

where λ is the so-called regularization parameter. The previous optimization problem can be easily solved using standard convex optimization techniques as explained in Belloni *et al.* [2011] and Bunea *et al.* [2014].

Our contributions in this chapter can now be explicitly stated. We introduce a notion of discrepancy, $\mathcal{D}_c(P, P_n)$, discussed in Section 5.2, between P_n and any other

probability measure P, such that

$$\min_{\beta} \max_{P:D_c(P,P_n)\leq\delta} \mathbb{E}_P^{1/2}\left[l\left(X,Y;\beta\right)\right] = \min_{\beta}\left(\mathbb{E}_{P_n}^{1/2}\left[l\left(X,Y;\beta\right)\right] + \delta^{1/2}\left\|\beta\right\|_{\alpha-(p,s)}\right). \quad (5.2)$$

Using this representation, which we formulate, together with its logistic regression analogue, in Section 5.2.2.1 and Section 5.2.2.2, we are able to draw the following insights:

I) GSRL can be interpreted as a game in which we choose a parameter (i.e. β) and an adversary chooses a "plausible" perturbation of the data (i.e. P); the parameter δ controls the degree in which P_n is allowed to be perturbed to produce P. The value of the game is dictated by the expected loss, under E_P, of the decision variable β.

II) The set $\mathcal{U}_\delta\left(P_n\right) = \left\{P : \mathcal{D}_c\left(P,P_n\right) \leq \delta\right\}$ denotes the set of distributional uncertainty. It represents the set of plausible variations of the underlying probabilistic model which are reasonably consistent with the data.

III) The DRO representation (5.2) exposes the role of the regularization parameter. In particular, because $\lambda = \delta^{1/2}$, we conclude that λ directly controls the size of the distributionally uncertainty and should be interpreted as the parameter which dictates the degree to which perturbations or variations of the available data should be considered.

IV) As a consequence of I) to III), the DRO representation (5.2) endows the GSRL estimator with desirable generalization properties. The GSRL aims at choosing a parameter, β, which should perform well for *all* possible probabilistic descriptions which are plausible given the data.

In the rest of the chapter we answer the following questions. First, in Section 5.2 we will revisit the the definition of $D_c\left(P,Q\right)$ as the optimal transport cost.

Intuitively, $D_c\left(P,P_n\right)$ represents the minimal transportation cost for moving the

mass encoded by P_n into a sinkhole which is represented by P. The cost of moving mass from location $u = (x, y)$ to $w = (x', y')$ is encoded by a cost function $c(u, w)$ which we shall discuss and this will depend on the α-(p, s) norm that we defined in (5.1). The subindex c in $\mathcal{D}_c(P, P_n)$ represents the dependence on the chosen cost function.

The next item of interest is the choice of δ, again the discussion of items I) to III) of the DRO formulation (5.2) provides a natural way to optimally choose δ. The idea is that every model $P \in \mathcal{U}_\delta(P_n)$ should intuitively represent a plausible variation of P_n and therefore $\beta^P = \arg\min\{\mathbb{E}_P[l(X, Y; \beta)] : \beta\}$ is a plausible estimate of β^*. The set $\{\beta^P : P \in \mathcal{U}_\delta(P_n)\}$ therefore yields a confidence region for β^* which is increasing in size as δ increases. Hence, it is natural to minimize δ to guarantee a target confidence level (say 95%). In Section 5.3 we explain how this optimal choice can be asymptotically computed as $n \to \infty$.

Finally, it is of interest to investigate if the optimal choice of δ (and thus of λ) actually performs well in practice. We compare performance of our (asymptotically) optimal choice of λ against cross-validation empirically in Section 5.4. We conclude that our choice is quite comparable to cross validation.

5.2 Optimal Transport and DRO

5.2.1 Revisit the optimal transport discrepancy

Let $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ be lower semicontinuous and we assume that $c(u, w) = 0$ if and only if $u = w$. For reasons that will become apparent in the sequel, we will refer to $c(\cdot)$ as a cost function.

Given two distributions P and Q, with supports $\mathcal{S}_P \subseteq \mathbb{R}^{d+1}$ and $\mathcal{S}_Q \subseteq \mathbb{R}^{d+1}$,

respectively, we define the optimal transport discrepancy, $\mathcal{D}_c$, via

$$\mathcal{D}_c\left(P,Q\right)=\inf_{\pi}\{E_{\pi}\left[c\left(U,W\right)\right]:\pi\in\mathcal{P}\left(\mathcal{S}_P\times\mathcal{S}_Q\right),\ \pi_U=P,\ \pi_W=Q\},\tag{5.3}$$

where $\mathcal{P}\left(\mathcal{S}_P\times\mathcal{S}_Q\right)$ is the set of probability distributions π supported on $\mathcal{S}_P\times\mathcal{S}_Q$, and π_U and π_W denote the marginals of U and W under π, respectively.

We shall discuss in details in next section, how to choose $c\left(\cdot\right)$ to recover (5.2) and the corresponding logistic regression formulation of GR-Lasso.

5.2.2 DRO Representation of GSRL Estimators

In this section, we will construct a cost function $c\left(\cdot\right)$ to obtain the GSRL (or GR-Lasso) estimators. We will follow an approach introduced in Chapter 2 for the context of square-root Lasso (SR-Lasso) and regularized logistic regression estimators.

5.2.2.1 GSRL Estimators for Linear Regression

We start by assuming precisely the linear regression setup described in the Introduction and leading to (5.2). Given $\alpha=\left(\alpha_1,...,\alpha_{\bar{d}}\right)^T\in R^{\bar{d}}_{++}$ define $\alpha^{-1}=\left(\alpha_1^{-1},...,\alpha_{\bar{d}}^{-1}\right)^T$. Now, underlying there is a partition $G_1,...,G_{\bar{d}}$ of $\{1,...,d\}$ and given $q,t\in[1,\infty]$ we introduce the cost function

$$c\left(\left(x,y\right),\left(x',y'\right)\right)=\begin{cases}\left\|x-x'\right\|^{\varrho}_{\alpha^{-1}\text{-}(q,t)} & \text{if}\quad y=y'\\ \infty & \text{if}\quad y\neq y'\end{cases},\tag{5.4}$$

where, following (5.1), we have that

$$\left\|x-x'\right\|^{\varrho}_{\alpha^{-1}\text{-}(q,t)}=\left(\sum_{i=1}^{\bar{d}}\alpha_i^{-t}\left\|x\left(G_i\right)-x'\left(G_i\right)\right\|_q^t\right)^{\varrho/t}.$$

Then, we obtain the following result.

Theorem 5.1 (DRO Representation for Linear Regression GSRL). Suppose that $q, t \in [1, \infty]$ and $\alpha \in R_{++}^{\bar{d}}$ are given and $c(\cdot)$ is defined as in (5.4) for $\varrho = 2$. Then, if $l(x, y; \beta) = \left(y - x^T\beta\right)^2$ we obtain

$$\min_{\beta \in \mathbb{R}^d} \sup_{P: D_c(P, P_n) \leq \delta} \left(\mathbb{E}_P \left[l(X, Y; \beta)\right]\right)^{1/2} = \min_{\beta \in \mathbb{R}^d} \left(E_{P_n} \left[l(X, Y; \beta)\right]\right)^{1/2} + \sqrt{\delta} \left\|\beta\right\|_{\alpha\text{-}(p,s)},$$

where $1/p + 1/q = 1$, and $1/s + 1/t = 1$.

We remark that choosing $p = q = 2$, $t = \infty$, $s = 1$, and $\alpha_i = \sqrt{g_i}$ for $i \in \{1, ..., \bar{d}\}$ we end up obtaining the GSRL estimator formulated in Bunea *et al.* [2014]).

We note that the cost function $c(\cdot)$ only allows mass transportation on the predictors (i.e X), but no mass transportation is allowed on the response variable Y. This implies that the GSRL estimator implicitly assumes that distributional uncertainty is only present on prediction variables (i.e. variations on the data only occurs through the predictors).

5.2.2.2 GR-Lasso Estimators for Logistic Regression

We now discuss GR-Lasso for classification problems. We consider a training data set of the form $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Once again, the input $X_i \in \mathbb{R}^d$ is a vector of d predictor variables, but now the response variable $Y_i \in \{-1, 1\}$ is a categorical variable. In this section we shall consider as our loss function the log-exponential function, namely,

$$l(x, y; \beta) = \log\left(1 + \exp\left(-y\beta^T x\right)\right). \tag{5.5}$$

This loss function is motivated by a logistic regression model which we shall review in the sequel. But for the DRO representation formulation it is not necessary to impose

any statistical assumption. We then obtain the following theorem.

Theorem 5.2 (DRO Representation for Logistic Regression GR-Lasso). Suppose that $q, t \in [1, \infty]$ and $\alpha \in R_{++}^{\bar{d}}$ are given and $c(\cdot)$ is defined as in (5.4) for $\varrho = 1$. Then, if $l(x, y; \beta)$ is defined as in (5.5) we obtain

$$\min_{\beta \in \mathbb{R}^d} \sup_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P \left[l(X, Y; \beta) \right] = \min_{\beta \in \mathbb{R}^d} E_{P_n} \left(l(X, Y; \beta) \right) + \delta \left\| \beta \right\|_{\alpha\text{-}(p,s)},$$

where $1 \leq q, t \leq \infty$, $1/p + 1/q = 1$ and $1/s + 1/t = 1$.

We note that by taking $p = q = 2$, $t = \infty$, $s = 1$, $\alpha_i = \sqrt{g_i}$ for $i \in \{1, ..., \bar{d}\}$, and $\lambda = \delta$ we recover the GR-Lasso logistic regression estimator from Meier *et al.* [2008].

As discussed in the previous subsection, the choice of $c(\cdot)$ implies that the GR-Lasso estimator implicitly assumes that distributionally uncertainty is only present on prediction variables.

5.3 Optimal Choice of Regularization Parameter

Let us now discuss the mathematical formulation of the optimal criterion that we discussed for choosing δ (and therefore the regularization parameter λ). We define

$$\Lambda_\delta(P_n) = \{\beta^P : P \in \mathcal{U}_\delta(P_n)\},$$

as discussed in the Introduction, $\Lambda_\delta(P_n)$ is a natural confidence region for β^* because each element P in the distributional uncertainty set $\mathcal{U}_\delta(P_n)$ can be interpreted as a plausible variation of the empirical data P_n. Then, given a confidence level $1 - \chi$ (say $1 - \chi = .95$) we wish to choose

$$\delta_n^* = \inf\{\delta : P(\beta^* \in \Lambda_\delta(P_n)) > 1 - \chi\}.$$

Note that in the evaluation of $P\left(\beta^* \in \Lambda_\delta\left(P_n\right)\right)$ the random element is P_n. So, we shall impose natural probabilistic assumptions on the data generating process in order to *asymptotically evaluate* δ_n^* as $n \to \infty$.

5.3.1 Revisit The Robust Wasserstein Profile Function

In order to asymptotically evaluate δ_n^* we must recall basic properties of the so-called Robust Wassertein Profile function (RWP function) introduced in Section 2.4 of Chapter 2.

Suppose for each (x, y), the loss function $l\left(x, y; \cdot\right)$ is convex and differentiable, then under natural moment assumptions which guarantee that expectations are well defined, we have that for

$$P \in \mathcal{U}_\delta\left(P_n\right) = \{P : D_c\left(P, P_n\right) \leq \delta\},$$

the parameter β^P must satisfy

$$\mathbb{E}_P\left[\nabla_\beta l\left(X, Y; \beta^P\right)\right] = 0. \tag{5.6}$$

Now, for any given β, let us define

$$\mathcal{M}\left(\beta\right) = \left\{P : \mathbb{E}_P\left[\nabla_\beta l\left(X, Y; \beta\right)\right] = 0\right\},$$

which is the set of probability measures P, under which β is the optimal risk minimization parameter. We would like to choose δ as small as possible so that

$$\mathcal{U}_\delta\left(P_n\right) \cap \mathcal{M}\left(\beta^*\right) \neq \varnothing \tag{5.7}$$

with probability at least $1-\chi$. But note that (5.7) holds if and only if there exists P such that $D_c(P, P_n) \leq \delta$ and $\mathbb{E}_P[\nabla_\beta l(X, Y; \beta^*)] = 0$.

The RWP function is defined

$$R_n(\beta) = \min\{D_c(P, P_n) : \mathbb{E}_P[\nabla_\beta l(X, Y; \beta)] = 0\}. \tag{5.8}$$

In view of our discussion following (5.7), it is immediate that $\beta^* \in \Lambda_\delta(P_n)$ if and only if $R_n(\beta^*) \leq \delta$, which then implies that

$$\delta_n^* = \inf\{\delta : P(R_n(\beta^*) \leq \delta) > 1-\chi\}.$$

Consequently, we conclude that δ_n^* can be evaluated asymptotically in terms of the $1-\chi$ quantile of $R_n(\beta^*)$ and therefore we must identify the asymptotic distribution of $R_n(\beta^*)$ as $n \to \infty$. We illustrate intuitively the role of the RWP function and $\mathcal{M}(\beta)$ in Figure 5.1, where RWP function $R_n(\beta^*)$ could be interpreted as the discrepancy distance between empirical measure P_n and the manifold $\mathcal{M}(\beta^*)$ associated with β^*.

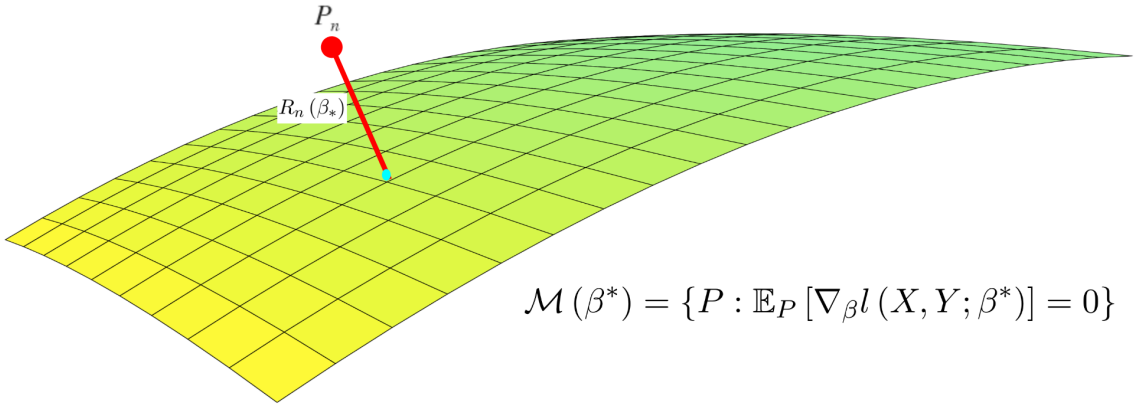

Figure 5.1: Intuitive Plot for the RWP function $R_n(\beta)$ and the set $\mathcal{M}(\beta)$.

Typically, under assumptions supporting the underlying model (as in the gener-

alized linear setting we considered), we will have that β^* is characterized by the estimating equation (5.6). Therefore, under natural statistical assumptions one should expect that $R_n(\beta^*) \to 0$ as $n \to \infty$ at a certain rate and therefore $\delta_n^* \to 0$ at a certain (optimal) rate. This then yields an optimal rate of convergence to zero for the underlying regularization parameter. The next subsections will investigate the precise rate of convergence analysis of δ_n^*.

5.3.2 Optimal Regularization for GSRL Linear Regression

We assume, for simplicity, that the training data set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is i.i.d. and that the linear relationship $Y_i = \beta^{*T} X_i + e_i$, holds with the errors $\{e_1, ..., e_n\}$ being i.i.d. and independent of $\{X_1, \ldots, X_n\}$. Moreover, we assume that both the entries of X_i and the errors have finite second moment and the errors have zero mean. Since in our current setting $l(x, y; \beta) = \left(y - x^T\beta\right)^2$, then the RWP function (5.8) for linear regression model is given as,

$$R_n(\beta) = \min_P \left\{ D_c(P, P_n) : \mathbb{E}_P \left[X \left(Y - X^T\beta \right) \right] = 0 \right\}. \tag{5.9}$$

Theorem 5.3 (RWP Function Asymptotic Results: Linear Regression). Under the assumptions imposed in this subsection and the cost function as given in Equation (5.4), with $\varrho = 2$,

$$nR_n(\beta^*) \Rightarrow L_1 := \max_{\zeta \in \mathbb{R}^d} \left\{ 2\sigma\zeta^T Z - \mathbb{E}\left[\left\| e\zeta - \left(\zeta^T X\right)\beta^* \right\|^2_{\alpha\text{-}(p,s)} \right] \right\},$$

as $n \to \infty$, where $\Rightarrow$ means convergence in distribution and $Z \sim \mathcal{N}(0, \Sigma)$ with

$\Sigma = Var(X)$. Moreover, we can observe the more tractable stochastic upper bound,

$$L_1 \overset{D}{\leq} L_2 := \frac{\mathbb{E}\left[e^2\right]}{\mathbb{E}\left[e^2\right] - \left(\mathbb{E}\left[|e|\right]\right)^2} \left\|Z\right\|^2_{\alpha^{-1}\text{-}(q,t)} .$$

We now explain how to use Theorem 5.3 to set the regularization parameter in GSRL linear regression:

1. Estimate the $1-\chi$ quantile of $\left\|Z\right\|^2_{\alpha^{-1}\text{-}(q,t)}$. We use use $\hat{\eta}_{1-\chi}$ to denote the estimator for this quantile. This step involves estimating Σ from the training data.

2. The regularization parameter λ in the GSRL linear regression takes the form

$$\lambda = \sqrt{\delta} = \sqrt{\frac{\hat{\eta}^{1/2}_{1-\chi}}{n(1 - \left(\mathbb{E}\left|e\right|\right)^2 / \mathbb{E}e^2)}}.$$

   Note that the denominator in the previous expression must be estimated from the training data.

Note that the regularization parameter for GSRL for linear regression chosen via our RWPI asymptotic result does not depends on the magnitude of error e (see also the discussion in Bunea *et al.* [2014]).

5.3.3 Optimal Regularization for GR-Lasso Logistic Regression

We assume that the training data set $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is i.i.d.. In addition, we assume that the X_i's have a finite second moment and also that they

possess a density with respect to the Lebesgue measure. Moreover, we assume a logistic regression model; namely,

$$P\left(Y_i = 1|X_i\right) = 1/\left(1 + \exp\left(-X_i^T\beta^*\right)\right), \tag{5.10}$$

and $P\left(Y_i = -1|X_i\right) = 1 - P\left(Y_i = 1|X_i\right)$.

In the logistic regression setting, we consider the log-exponential loss defined in Equation (5.5). Therefore, the RWP function, (5.8), for logistic regression is

$$R_n\left(\beta\right) = \min\left\{D_c\left(P, P_n\right) : \mathbb{E}_P\left[\frac{YX}{1 + \exp\left(YX^T\beta\right)}\right] = 0\right\}. \tag{5.11}$$

Theorem 5.4 (RWP Function Asymptotic Results: Logistic Regression). Under the assumptions imposed in this subsection and the cost function as given in Equation (5.4) with $\varrho = 1$,

$$\sqrt{n}R_n\left(\beta^*\right) \Rightarrow L_3 := \sup_{\zeta \in A} \quad \zeta^T Z,$$

as $n \to \infty$, where

$$Z \sim \mathcal{N}\left(0, \mathbb{E}\left[\frac{XX^T}{\left(1 + \exp\left(YX^T\beta^*\right)\right)^2}\right]\right)$$

and

$$A = \left\{\zeta \in \mathbb{R}^d : \operatorname*{ess\,sup}_{X,Y}\left\|\zeta^T \frac{y\left(1 + \exp\left(YX^T\beta^*\right)\right) I_{d\times d} - XX^T}{\left(1 + \exp\left(YX^T\beta^*\right)\right)^2}\right\|_{\alpha\text{-}(p,s)} \leq 1\right\}.$$

Further, the limit law L_3 follows the simpler stochastic bound,

$$L_3 \overset{D}{\leq} L_4 := \left\|\tilde{Z}\right\|_{\alpha^{-1}\text{-}(q,t)},$$

where $\tilde{Z} \sim \mathcal{N}\left(0, \Sigma\right)$.

We now explain how to use Theorem 5.4 to set the regularization parameter in GR-Lasso logistic regression.

1. Estimate the $1-\chi$ quantile of L_4. We use use $\hat{\eta}_{1-\chi}$ to denote the estimator for this quantile. This step involves estimating Σ from the training data.

2. We choose the regularization parameter λ in the GR-Lasso problem to be,

$$\lambda = \delta = \hat{\eta}_{1-\chi}/\sqrt{n}.$$

5.4 Numerical Experiments

We proceed to numerical experiments on both simulated and real data to verify the performance of our method for choosing the regularization parameter. We apply "grpreg" in R, from Breheny and Breheny [2016], to solve GR-Lasso for logistic regression. For GSRL for linear regression, we consider apply the "grpreg" solver for the GR-Lasso problem combined with the iterative procedure discussed in Section 2 of Sun and Zhang [2011] (see also Section 5 of Li *et al.* [2015] for the Lasso counterpart of such numerical procedure).

Data preparation for simulated experiments: We borrow the setting from example III in Yuan and Lin [2006], where the group structure is determined by the third order polynomial expansion. More specifically, we assume that we have 17 random variables $Z_1, \ldots, Z_{16}$ and W, they are i.i.d. and follow the normal distribution. The covariates $X_1, \ldots, X_{16}$ are given as $X_i = (Z_i + W)/\sqrt{2}$. For the predictors, we consider each covariate and its second and third order polynomial, i.e. X_i, X_i^2 and X_i^3. In total, we have 48 predictors.

For linear regression: The response Y is given by

$$Y = \beta_{3,1}X_3 + \beta_{3,2}X_3^2 + \beta_{3,3}X_3^3 + \beta_{5,1}X_5 + \beta_{5,2}X_5^2 + \beta_{5,3}X_5^3 + e,$$

where $\beta_{(\cdot,\cdot)}$ coefficients draw randomly and e represents an independent random error.

For classification: We consider Y simulated by a Bernoulli distribution, i.e.

$$Y \sim Ber\left(1/\left[1+\exp\left(-\left(\beta_{3,1}X_3 + \beta_{3,2}X_3^2 + \beta_{3,3}X_3^3 + \beta_{5,1}X_5 + \beta_{5,2}X_5^2 + \beta_{5,3}X_5^3\right)\right)\right]\right).$$

We compare the following methods for linear regression and logistic regression: 1) groupwise regularization with asymptotic results (in Theorem 5.3, 5.4) selected tuning parameter (RWPI GRSL and RWPI GR-Lasso), 2) groupwise regularization with cross-validation (CV GRSL and CV GR-Lasso), and 3) ordinary least square and logistic regression (OLS and LR).

We report the error as the square loss for linear regression and log-exponential loss for logistic regression. The training error is calculated via the training data. The size of the training data is taken to be $n = 50, 100, 500$ and 1000. The testing error is evaluated using a simulated data set of size 1000 using the same data generating process described earlier. The mean and standard deviation of the error are reported via 200 independent runs of the whole experiment, for each sample size n.

The detailed results are summarized in Table 5.1 for linear regression and Table 5.2 for logistic regression. We can see that our procedure is very comparable to cross validation, but it is significantly less time consuming and all of the data can be directly used to estimate the model parameter, by-passing significant data usage in the estimation of the regularization parameter via cross validation

We also validated our method using the Breast Cancer classification problem with

	RWPI GSRL		CV GSRL		OLS	
Sample Size	Training	Testing	Training	Testing	Training	Testing
$n = 50$	5.64 ± 1.16	9.15 ± 3.58	3.18 ± 1.07	7.66 ± 2.69	0.07 ± 0.09	80.98 ± 30.53
$n = 100$	4.67 ± 0.70	5.83 ± 1.38	3.61 ± 0.74	5.22 ± 1.05	2.09 ± 0.44	73.35 ± 16.51
$n = 500$	4.09 ± 0.29	4.16 ± 0.27	3.93 ± 0.3	4.12 ± 0.27	3.63 ± 0.27	73.08 ± 10.40
$n = 1000$	4.02 ± 0.19	4.11 ± 0.26	3.95 ± 0.19	4.11 ± 0.26	3.82 ± 0.19	72.28 ± 8.05

Table 5.1: Linear Regression Simulation Results.

	RWPI GR-Lasso		CV GR-Lasso		Logistic Regression	
Sample Size	Training	Testing	Training	Testing	Training	Testing
$n = 50$	$.683 \pm .016$	$.702 \pm .014$	$.459 \pm .118$	$.628 \pm .099$	$.002 \pm .001$	5.288 ± 1.741
$n = 100$	$.593 \pm .038$	$.618 \pm .029$	$.450 \pm .061$	$.551 \pm .037$	$.042 \pm .041$	4.571 ± 1.546
$n = 500$	$.513 \pm .021$	$.518 \pm .019$	$.461 \pm .025$	$.493 \pm .018$	$.083 \pm .057$	$1.553 \pm .355$
$n = 1000$	$.492 \pm .016$	$.488 \pm .017$	$.491 \pm .017$	$.488 \pm .019$	$.442 \pm .018$	$.510 \pm .028$

Table 5.2: Logistic Regression Simulation Results.

data from the UCI machine learning database discussed in Lichman [2013]. The data set contains 569 samples with one binary response and 30 predictors. We consider all the predictors and their first, second, and third order polynomial expansion. Thus, we end up having 90 predictors divided into 30 groups. For each iteration, we randomly split the data into a training set with 112 samples and the rest in the testing set. We repeat the experiment 500 times to observe the log-exponential loss function for the training and testing error. We compare our asymptotic results based GR-Lasso logistic regression (RWPI GR-Lasso), cross-validation based GR-Lasso logistic regression (CV GR-Lasso), vanilla logistic regression (LR), and regularized logistic regression (LRL1). We can observe, even when the sample size is small as in the example, our method still provides very comparable results (see in Table 5.3).

LR		LRL1		RWPI GR-Lasso		CV GR-Lasso	
Training	Testing	Training	Testing	Training	Testing	Training	Testing
0.0 ± 0.0	15.267 ± 5.367	$.510 \pm .215$	$.414 \pm .173$	$.186 \pm .032$	$.240 \pm .098$	$.198 \pm .041$	$.213 \pm .041$

Table 5.3: Numerical results for breast cancer data set.

5.5 Conclusion and Extensions

Our discussion of GSRL as a DRO problem has exposed rich interpretations which we have used to understand GSRL's generalization properties by means of a game theoretic formulation. Moreover, our DRO representation also elucidates the crucial role of the regularization parameter in measuring the distributional uncertainty present in the data. Finally, we obtained asymptotically valid formulas for optimal regularization parameters under a criterion which is naturally motivated, once again, thanks to our DRO formulation. Our easy-to-implement formulas are shown to perform well compared to cross validation.

We strongly believe that our discussion in this chapter can be easily extended to a wide range of machine learning estimators. We envision formulating the DRO problem considering different types of models and cost functions. We plan to investigate algorithms which solve the DRO problem directly (even if no direct regularization representation, as the one we considered here, exists). Moreover, it is natural to consider different types of cost functions which might improve upon the simple choice which, as we have shown, implies the GSRL estimator. Questions related to alternative choices of cost functions are also under current research investigations, and our progress will be reported in the next chapter.

APPENDIX 5.A: Technical Proofs

We will first derive some properties for α-(p, s) norm (in Section 5.A.1) we defined in Equation (5.1), then we move to the proof for DRO problem in Section 5.A.2 and the optimal selection of regularization parameter in Section 5.A.3. We will focus on the proof for linear regression and leave the part for logistic regression, which follows

the similar techniques, in the Additional Technical Results, namely APPENDIX B.

5.A.1: Basic Properties of the α-(p, s) Norm

The following Proposition, which describes basic properties of the α-(p, s) norm, will be very useful in our proofs.

Proposition 5.1. For $\alpha - (p, s)$ norm defined for $\mathbb{R}^d$ as in Equation (5.1) and the notations therein, we have the following properties:

I) The dual norm of $\alpha-(p, s)$ norm is α^{-1}-(q, t) norm, where $\alpha^{-1} = (1/\alpha_1, \ldots, 1/\alpha_{\bar{d}})^T$, $1/p + 1/q = 1$, and $1/s + 1/t = 1$ (i.e. p, q are conjugate and s, t are conjugate).

II) The Hölder inequality holds for the α-(p, s) norm, i.e. for $a, b \in \mathbb{R}^d$, we have,

$$a^T b \leq \|a\|_{\alpha\text{-}(p,s)} \|b\|_{\alpha^{-1}\text{-}(q,t)},$$

where the equality holds if and only if $sign(a(G_j)_i) = sign(b(G_j)_i)$ and

$$|\alpha_j a(G_j)_i| \left\| \frac{1}{\alpha_j} b(G_j) \right\|_q^{q/p-t/s} \|b\|^{t/s}_{\alpha^{-1}\text{-}(q,t)} = \left| \frac{1}{\alpha_j} b(G_j)_i \right|^{q/p}.$$

is true for all $j = 1, \ldots, \bar{d}$ and $i = 1, \ldots, g_j$.

The triangle inequality holds, i.e. for $a, b \in \mathbb{R}^d$ and $a \neq 0$, we have

$$\|a\|_{\alpha\text{-}(p,s)} + \|b\|_{\alpha\text{-}(p,s)} \geq \|a + b\|_{\alpha\text{-}(p,s)},$$

where the equality holds if and only if, there exists nonnegative τ, such that $\tau a = b$.

Proof of Proposition 5.1. We first proceed to prove II). Let us consider any $a, b \in \mathbb{R}^d$. We can assume $a, b \neq 0$, otherwise the claims are immediate. The inner product (or

dot product) of a and b an be written as:

$$a^T b = \sum_{j=1}^{\bar{d}} \left[\sum_{i=1}^{g_j} a(G_j)_i b(G_j)_i \right] \leq \sum_{j=1}^{\bar{d}} \left[\sum_{i=1}^{g_j} |a(G_j)_i| \cdot |b(G_j)_i| \right].$$

The equality holds for the above inequality if and only if $a(G_j)_i$ and $b(G_j)_i$ shares the same sign. For each fixed $j = 1, \ldots, \bar{d}$, we consider the term in the bracket,

$$\sum_{i=1}^{g_j} |a(G_j)_i| \cdot |b(G_j)_i| = \sum_{i=1}^{g_j} \alpha_j \left| a(G_j)_i \right| \cdot \left| b(G_j)_i \right| / \alpha_j \leq \left\| \alpha_j a(G_j) \right\|_p \cdot \left\| \frac{1}{\alpha_j.} b(G_j) \right\|_q .$$

The above inequality is due to Hölder's inequality for $p-$norm and the equality holds if and only if

$$\left\| \frac{1}{\alpha_j.} b(G_j) \right\|_q^q \left| \alpha_j a(G_j)_i \right|^p = \left\| \alpha_j a(G_j) \right\|_p^p \left| \frac{1}{\alpha_j} b(G_j)_i \right|^q ,$$

is true for all $i = \overline{1, g_j}$. Combining the above result for each $j = 1, \ldots, \bar{d}$, we have,

$$a^T b \leq \sum_{j=1}^{\bar{d}} \left\| \alpha_j a(G_j) \right\|_p \cdot \left\| \frac{1}{\alpha_j} b(G_j) \right\|_q \leq \left\| a \right\|_{\alpha\text{-}(p,s)} \cdot \left\| b \right\|_{\alpha^{-1}\text{-}(q,t)} ,$$

where the final inequality is due to Hölder inequality applied to the vectors

$$\tilde{a} = \left(\alpha_1 \left\| a(G_1) \right\|_p , \ldots, \alpha_{\bar{d}} \left\| a(G_{\bar{d}}) \right\|_p \right)^T , \text{ and } \tilde{b} = \left(\frac{1}{\alpha_1} \left\| b_{G_1} \right\|_q , \ldots, \frac{1}{\alpha_{\bar{d}}} \left\| b(G_{\bar{d}}) \right\|_q \right)^T . \tag{5.12}$$

This proves the Hölder type inequality stated in the theorem. We can further observe that the final inequality becomes equality if and only if

$$\left\| b \right\|_{\alpha^{-1}-(q,t)}^t \left\| \alpha_j a(G_j) \right\|_p^s = \left\| a \right\|_{\alpha-(p,s)}^s \left\| \frac{1}{\alpha_j} b(G_j) \right\|_q^t ,$$

holds for all $j = 1, \ldots, \bar{d}$. Combining the conditions for equalities hold for each inequality, we conclude condition II) in the statement of the proposition.

Now we proceed to prove I). Recall the definition of a dual norm, i.e.

$$\|b\|^*_{\alpha\text{-}(p,s)} = \sup_{a:\|a\|_{\alpha\text{-}(p,s)}=1} a^T b$$

. Now, choose $b \in \mathbb{R}^d$, $b \neq 0$, and let us take a satisfying, $\|a\|_{\alpha-(p,s)} = 1$ and

$$a(G_j)_i = \frac{sign(b(G_j)_i)}{\alpha_j} \frac{\left|\frac{1}{\alpha_j} b(G_j)_i\right|^{q/p}}{\left\|\frac{1}{\alpha_j} b(G_j)\right\|_q^{q/p-t/s} \|b\|^{t/s}_{\alpha^{-1}\text{-}(q,t)}}.$$

By part II), we have that

$$\|b\|^*_{\alpha-(p,s)} = \sup_{a:\|a\|_{\alpha-(p,s)}=1} a^T b = \|a\|_{\alpha-(p,s)} \|b\|_{\alpha^{-1}-(q,t)} = \|b\|_{\alpha^{-1}-(q,t)}.$$

Thus we proved part I). Finally, let us discuss the triangle inequality. For any $a, b \in \mathbb{R}^d$ and $a, b \neq 0$ we have

$$\begin{aligned}
&\|a\|_{\alpha\text{-}(p,s)} + \|b\|_{\alpha\text{-}(p,s)} \\
&= \left[\sum_{j=1}^{\bar{d}} \alpha_j \|a(G_j)\|_p^s\right]^{1/s} + \left[\sum_{j=1}^{\bar{d}} \alpha_j \|b(G_j)\|_p^s\right]^{1/s} \\
&\geq \left[\sum_{j=1}^{\bar{d}} \alpha_j \left(\|a(G_j)\|_p^s + \|b(G_j)\|_p^s\right)\right]^{1/s} \\
&\geq \left[\sum_{j=1}^{\bar{d}} \alpha_j \|a(G_j) + b(G_j)\|_p^s\right]^{1/s} \\
&= \|a+b\|_{\alpha\text{-}(p,s)}.
\end{aligned}$$

For the above derivation, the first equality is due to definition in Equation (5.1), Second equality is applying the triangle inequality of s-norm for $\tilde{a}$ and $\tilde{b}$ defined in Equation (5.12), where the equality holds if and only if, there exist positive number $\tilde{\tau}$, such that $\tilde{\tau}\tilde{a} = \tilde{b}$. Third inequality is due to triangle equality of p-norm to $a(G_j)$ and $b(G_j)$ for each $j = 1, \ldots, \bar{d}$, where the equality holds if and only if, there exists nonnegative numbers τ_j, such that $\tau_j a(G_j) = b(G_j)$. Combining the equality condition for second and third estimate above, we can conclude the equality condition for the triangle inequality for α-(p, s) norm is if and only if there exists a non-negative number τ, such that $\tau a = b$. □

5.A.2: Proof of DRO for Linear Regression

Proof of Theorem 5.1. Let us apply the strong duality results, as in the Appendix of Chapter 2, for worst-case expected loss function, which is a semi-infinity linear programming problem, and write the worst-case loss as,

$$\begin{aligned}&\sup_{P:D_c(P,P_n)\leq\delta} \mathbb{E}_P\left[\left(Y - X^T\beta\right)^2\right] \\ =&\min_{\gamma\geq 0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^{n}\sup_{u}\left\{\left(y_i - u^T\beta\right)^2 - \gamma\left\|x_i - u\right\|^2_{\alpha^{-1}\text{-}(q,t)}\right\}\right\}.\end{aligned}$$

For each i, let us consider the inner optimization problem over u. We can denote $\Delta = u - x_i$ and $e_i = y_i - x_i^T\beta$ for notation simplicity, then the $i-$th inner optimization

problem becomes,

$$
\begin{aligned}
& e_i^2 + \sup_{\Delta} \left\{ \left(\Delta^T \beta\right)^2 - 2 e_i \Delta^T \beta - \gamma \left\| \Delta \right\|^2_{\alpha^{-1}-(q,t)} \right\} \\
=& e_i^2 + \sup_{\Delta} \left\{ \left(\sum_j \left|\Delta_j\right| \left|\beta_j\right| \right)^2 + 2 \left|e_i\right| \sum_j \left|\Delta_j\right| \left|\beta_j\right| - \gamma \left\| \Delta \right\|^2_{\alpha^{-1}-(q,t)} \right\} \\
=& e_i^2 \sup_{\left\|\Delta\right\|_{\alpha^{-1}\text{-}(q,t)}} \left\{ \left\|\beta\right\|^2_{\alpha\text{-}(p,s)} \left\|\Delta\right\|^2_{\alpha^{-1}\text{-}(q,t)} + 2 \left\|\beta\right\|_{\alpha\text{-}(p,s)} \left|e_i\right| \left\|\Delta\right\|_{\alpha^{-1}\text{-}(q,t)} - \gamma \left\|\Delta\right\|^2_{\alpha^{-1}\text{-}(q,t)} \right\} \\
=& \begin{cases} e_i^2 \frac{\gamma}{\gamma - \left\|\beta\right\|^2_{\alpha\text{-}(p,s)}} & \text{if } \gamma > \left\|\beta\right\|^2_{\alpha\text{-}(p,s)}\,, \\ +\infty & \text{if } \gamma \leq \left\|\beta\right\|^2_{\alpha\text{-}(p,s)}\,. \end{cases}
\end{aligned}
\tag{5.13}
$$

where the development uses the duality results developed in Proposition 5.1. The last equality is optimize over Δ for two different cases of λ.

Since optimization over γ is a minimization, the outer player will always select γ that avoids an infinite value of the game. Then we can write the worst-case expected loss function as,

$$
\begin{aligned}
& \sup_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P \left[\left(Y - X^T \beta \right)^2 \right] \\
=& \min_{\gamma > \left\|\beta\right\|^2_{\alpha\text{-}(p,s)}} \left\{ \gamma\delta - \gamma \frac{E_{P_n} l\left(X, Y; \beta\right)}{\gamma - \left\|\beta\right\|^2_{\alpha\text{-}(p,s)}} \right\} \\
=& \left(\sqrt{E_{P_n} l\left(X, Y; \beta\right)} + \sqrt{\delta} \left\|\beta\right\|_{\alpha\text{-}(p,s)} \right)^2 .
\end{aligned}
\tag{5.14}
$$

The first equality in (5.14) is a plug-in from the result in (5.13). For the second equality, we can observe the target function is convex and differentiable and as $\gamma \to \infty$ and $\gamma \to \left\|\beta\right\|^2_{\alpha\text{-}(p,s)}$, the value function will be infinity. We can solve this convex optimization problem which leads to the result above. We further take square root and take minimization over β on both sides, we proved the claim of the theorem. $\square$

5.A.3: Proof for Optimal Selection of Regularization for Linear Regression

Proof for Theorem 5.3. For linear regression with the square loss function, if we apply the strong duality result for semi-infinity linear programming problem as in Appendix B of Chapter 2, we can write the scaled RWP function for linear regression as

$$nR_n\left(\beta^*\right)=\sup_{\zeta}\left\{-\zeta^T Z_n-\mathbb{E}_{P_n}\phi\left(X_i,Y_i,\beta^*,\zeta\right)\right\}, \tag{5.15}$$

where $Z_n=\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i X_i$ and

$$\begin{aligned}&\phi\left(X_i,Y_i,\beta^*,\zeta\right)\\ =&\sup_{\Delta}\left\{e_i\zeta^T\Delta-\left(\beta^{*\,T}\Delta\right)\left(\zeta^T X_i\right)-\left(\|\Delta\|^2_{\alpha^{-1}\text{-}(q,t)}+n^{-1/2}\left(\beta^{*\,T}\Delta\right)\left(\zeta^T\Delta\right)\right)\right\}.\end{aligned}$$

Follow the similar discussion in the proof of Theorem 2.1 in Chapter 2. Applying Lemma 2.2 in Chapter 2, we can argue that the optimizer ζ can be restrict on a compact set asymptotically with high probability. We can apply the uniform law of large number estimate as in Lemma 2.3 of Chapter 2 to the second term in (5.15) and we obtain

$$nR_n\left(\beta^*\right)=\sup_{\zeta}\{-\zeta^T Z_n-\mathbb{E}\phi\left(X,Y,\beta,\zeta\right)]\}+o_P(1). \tag{5.16}$$

For any fixed X,Y, as $n\to\infty$, we can simplify the contribution of $\phi\left(\cdot\right)$ inside sup in (5.16). This is done by applying the duality result (Hölder-type inequality) in Proposition 5.1 and noting that $\phi\left(\cdot\right)$ becomes quadratic in $\|\Delta\|_{\alpha^{-1}\text{-}(q,t)}$. This results

in the simplified expression

$$nR_n(\beta^*) = \sup_{\zeta} \left\{ -\zeta^T Z_n - \mathbb{E}\left[\left\| e\zeta - (\zeta^T X)\beta^* \right\|^2_{\alpha\text{-}(p,s)} \right] \right\} + o_P(1).$$

Since we can observe that, $Z_n \Rightarrow \sigma Z$, then as $n \to \infty$ we proved the first argument. For this step we need to show that the feasible region can be compactified with high probability. This compactification argument is done using a technique similar to Lemma 2.2 in Chapter 2.

By the definition of L_1, we can apply Hölder inequality to the first term, and split the optimization into optimizing over direction $\|\zeta'\|_{\alpha\text{-}(p,s)} = 1$ and magnitude $a \geq 0$. Thus, we have

$$L_1 \leq \max_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1} \max_{a \geq 0} \left\{ 2a\sigma \|Z\|_{\alpha^{-1}\text{-}(q,t)} - a^2 \mathbb{E}\left[\left\| e\zeta' - (\zeta'^T X)\beta^* \right\|^2_{\alpha\text{-}(p,s)} \right] \right\}.$$

It is easy to solve the quadratic programming problem in a and we conclude that

$$L_1 \leq \frac{\sigma^2 \|Z\|^2_{\alpha^{-1}\text{-}(q,t)}}{\min_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1} \mathbb{E}\left[\left\| e\zeta' - (\zeta'^T X)\beta^* \right\|^2_{\alpha\text{-}(p,s)} \right]}.$$

For the denominator, we have estimates as follows:

$$\begin{aligned}
&\min_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1} \mathbb{E}\left[\left\| e\zeta' - (\zeta'^T X)\beta^* \right\|^2_{\alpha\text{-}(p,s)} \right] \\
&\geq \min_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1} \mathbb{E}\left[|e| - \left| \zeta^T X \right| \|\beta^*\|_{\alpha\text{-}(p,s)} \right]^2 \\
&\geq Var(|e|) + \min_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1} \left(\|\beta^*\|_{\alpha\text{-}(p,s)} \mathbb{E}\left| \zeta'^T X \right| - \mathbb{E}|e| \right)^2 \geq Var(|e|).
\end{aligned}$$

The first estimate is due to the triangle inequality in Proposition 5.1, the second estimate follows using Jensen's inequality, the last inequality is immediate. Combining

these inequalities we conclude

$$L_1 \leq \sigma^2 \left\|Z\right\|^2_{\alpha^{-1}\text{-}(q,t)} / Var(|e|).$$

□

APPENDIX 5.B: Additional Materials

In this Section, we will provide the proofs for DRO representation and asymptotic result for logistic regression, which were discussed in Theorem 5.2 and Theorem 5.4, in Section 5.B.1 and Section 5.B.2.

5.B.1: Proof of DRO for Logistic Regression

Proof for Theorem 5.2. By applying strong duality results for semi-infinity linear programming problem in Chapter 2, we can write the worst case expected loss function as,

$$\begin{aligned}
&\sup_{P:D_c(P,P_n)\leq\delta} \mathbb{E}_P\left[\log\left(1+\exp\left(-Y\beta^T X\right)\right)\right] \\
&= \min_{\gamma\geq 0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^{n}\sup_{u}\left\{\log\left(1+\exp\left(-Y_i\beta^T u\right)\right) - \gamma\left\|X_i-u\right\|_{\alpha^{-1}\text{-}(q,t)}\right\}\right\}.
\end{aligned}$$

For each i, we can apply Lemma 1 in Shafieezadeh-Abadeh *et al.* [2015] and the dual norm result in Proposition 5.1 to deal with the inner optimization problem. It gives

us,

$$\begin{aligned}
&\sup_u \left\{ \log\left(1+\exp\left(-Y_i\beta^T u\right)\right) - \gamma \left\| X_i - u \right\|_{\alpha^{-1}\text{-}(q,t)} \right\} \\
&= \begin{cases} \log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right) & \text{if} \quad \|\beta\|_{\alpha\text{-}(p,s)} \leq \gamma, \\ \infty & \text{if} \quad \|\beta\|_{\alpha\text{-}(p,s)} > \gamma. \end{cases}
\end{aligned}$$

Moreover, since the outer player wishes to minimize, γ will be chosen to satisfy $\gamma \geq \|\beta\|_{\alpha\text{-}(p,s)}$. We then conclude

$$\begin{aligned}
&\min_{\gamma\geq 0} \left\{ \gamma\delta - \frac{1}{n}\sum_{i=1}^n \sup_u \left\{ \log\left(1+\exp\left(-Y_i\beta^T u\right)\right) - \gamma \left\| X_i - u \right\|_{\alpha^{-1}\text{-}(q,t)} \right\} \right\} \\
&= \min_{\gamma\geq \|\beta\|_{\alpha\text{-}(p,s)}} \left\{ \delta\gamma + \frac{1}{n}\sum_{i=1}^n \log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right) \right\} \\
&= \frac{1}{n}\sum_{i=1}^n \log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right) + \delta \|\beta\|_{\alpha\text{-}(p,s)},
\end{aligned}$$

where the last equality is obtained by noting that the objective function is continuous and monotone increasing in γ, thus $\gamma = \|\beta\|_{\alpha\text{-}(p,s)}$ is optimal. Hence, we conclude the DRO formulation for GR-Lasso logistic regression. □

5.B.2: Proof of Optimal Selection of Regularization for Logistic Regression

Proof of Theorem 5.4. We can apply strong duality result for semi-infinite linear programming problem in Appendix B of Chapter 2, and write the scaled RWP function evaluated at β^* in the dual form as,

$$\sqrt{n} R_n(\beta^*) = \max_{\zeta} \left\{ \zeta^T Z_n - \mathbb{E}_{P_n} \phi(X, Y, \beta^*, \zeta) \right\},$$

where $Z_n = \frac{1}{n}\sum_i^n \frac{Y_i X_i}{1+\exp\left(Y_i X_i^T \beta^*\right)}$ and

$$\phi\left(X, Y, \beta^*, \zeta\right) = \max_u \left\{ Y\zeta^T \left(\frac{X}{1+\exp\left(YX^T\beta^*\right)} - \frac{u}{1+\exp\left(Yu^T\beta^*\right)} \right) - \|X-u\|_{\alpha^{-1}\text{-}(q,t)} \right\}.$$

We proceed as in our proof of Theorem 5.3 in this chapter and also adapting the case $\rho = 1$ for Theorem 2.1 in Chapter 2. We can apply Lemma 2.2 in Chapter 2 and conclude that the optimizer ζ can be taken to lie within a compact set with high probability as $n \to \infty$. We can combine the uniform law of large number estimate as in Lemma 2.3 of Chapter 2 and obtain

$$\sqrt{n} R_n\left(\beta\right) = \max_{\zeta} \left\{ \zeta^T Z_n - \mathbb{E}_P \phi\left(X, Y, \beta^*, \zeta\right) \right\} + o_P(1).$$

For the optimization problem defining $\phi\left(\cdot\right)$, we can apply results in Lemma 2.5 in Section A.3 of Chapter 2, we know, for any choice of $\tilde{\zeta}$, if,

$$\operatorname*{ess\,sup}_{X,Y} \left\| \tilde{\zeta}^T \frac{y\left(1+\exp\left(YX^T\beta^*\right)\right) I_{d\times d} - XX^T}{\left(1+\exp\left(YX^T\beta^*\right)\right)^2} \right\|_{\alpha\text{-}(p,s)} > 1,$$

we have $\mathbb{E}\left[\phi\left(X, Y, \beta^*, \tilde{\zeta}\right)\right] = \infty$. Since the outer optimization problem is maximization over ζ, the player will restrict ζ within the set A, where

$$A = \left\{ \zeta \in \mathbb{R}^d : \operatorname*{ess\,sup}_{X,Y} \left\| \zeta^T \frac{y\left(1+\exp\left(YX^T\beta^*\right)\right) I_{d\times d} - XX^T}{\left(1+\exp\left(YX^T\beta^*\right)\right)^2} \right\|_{\alpha\text{-}(p,s)} \leq 1 \right\}.$$

Moreover, it is easy to calculate, if $\zeta \in A$, we have $\mathbb{E}[\phi\left(X, Y, \beta^*, \zeta\right)] = 0$, thus we have the scaled RWP function has the following estimate, as $n \to \infty$

$$\sqrt{n} R_n\left(\beta\right) = \max_{\zeta \in A} \zeta^T Z_n + o_P(1).$$

Letting $n \to \infty$, we obtain the exact asymptotic result.

For the stochastic upper bound, let us recall for the definition of the set A and consider the following estimate

$$\begin{aligned}
&\left\| \zeta^T \frac{y\left(1+\exp\left(YX^T\beta^*\right)\right) I_{d\times d} - XX^T}{\left(1+\exp\left(YX^T\beta^*\right)\right)^2} \right\|_{\alpha\text{-}(p,s)} \\
&\geq \left\| \frac{Y\zeta}{1+\exp\left(Y\beta^{*\,T}X\right)} \right\|_{\alpha\text{-}(p,s)} - \left\| \frac{\zeta^T X\beta^*}{\left(1+\exp\left(Y\beta^{*\,T}X\right)\right)^2} \right\|_{\alpha\text{-}(p,s)} \\
&\geq \left(\frac{1}{1+\exp\left(Y\beta^{*\,T}X\right)} - \frac{\|X\|_{\alpha^{-1}\text{-}(q,t)} \left\|\beta^*\right\|_{\alpha\text{-}(p,s)}}{\left(1+\exp\left(Y\beta^{*\,T}X\right)\right)\left(1+\exp\left(-Y\beta^{*\,T}X\right)\right)} \right) \|\zeta\|_{\alpha\text{-}(p,s)} .
\end{aligned}$$

The first inequality is due to application of triangle inequality in Proposition 5.1, while the second estimate follows from Hölder's inequality and $Y \in \{-1, +1\}$. Since we assume positive probability density for the predictor X, we can argue that, if $\|\zeta\|_{\alpha\text{-}(p,s)} = (1-\epsilon)^{-2} > 1$ and $\epsilon > 0$ is chosen arbitrarily small, we can conclude from the above estimate that, we have

$$\left\| \zeta^T \frac{y\left(1+\exp\left(YX^T\beta^*\right)\right) I_{d\times d} - XX^T}{\left(1+\exp\left(YX^T\beta^*\right)\right)^2} \right\|_{\alpha\text{-}(p,s)} > 1.$$

Thus, we proved the claim that $A \subset \left\{\zeta, \|\zeta\|_{\alpha\text{-}(p,s)} \leq 1\right\}$. The stochastic upper bound is derived by replacing A by $\left\{\zeta, \|\zeta\|_{\alpha\text{-}(p,s)} \leq 1\right\}$, i.e.

$$L_3 = \sup_{\zeta \in A} \zeta^T Z \leq \sup_{\|\zeta\|_{\alpha\text{-}(p,s)} \leq 1} \zeta^T Z = \|Z\|_{\alpha^{-1}\text{-}(q,t)} ,$$

where the final estimation is due to dual norm structure in Proposition 5.1. Since we know, $\frac{1}{1+\exp(YX^T\beta)} \leq 1$, it is easy to argue, $Var(\tilde{Z}) - Var(Z)$ is positive semidefinite, thus, we know $\|Z\|_{\alpha^{-1}\text{-}(q,t)}$ is stochastic dominated by $L_4 := \left\|\tilde{Z}\right\|_{\alpha^{-1}\text{-}(q,t)}$. Hence, we obtain $L_3 \leq L_4$. □

Chapter 6

Data-Driven Optimal Transport Cost Selection for Distributionally Robust Optimization

In the former chapter, namely Chapter 5, we consider the generalization of cost function from regular Euclidean norm to groupwise norm to encode the information on natural structure of the predictors. In this chapter, we will further explore flexibility of optimal transport cost, more specifically, we are going to apply metric learning techniques to show how to pick the cost function in a fully data-driven way.

In Chapter 2 and Chapter 5, we showed that several machine learning algorithms, such as square-root Lasso, Group Lasso, and regularized logistic regression, among many others, can be represented exactly as data-driven distributionally robust optimization (DRO) problems. The distributional uncertainty is defined as a neighborhood centered at the empirical distribution. In this chapter, we propose a methodology which learns such neighborhood in a natural data-driven way. We show rigorously that our framework encompasses adaptive regularization as a particular case. In ad-

dition, we also propose a data-driven robust optimization methodology to inform the transportation cost underlying the definition of the distributional uncertainty. Moreover, we demonstrate empirically that our proposed methodology is able to improve upon a wide range of popular machine learning estimators.

6.1 Introduction

A Distributionally Robust Optimization (DRO) problem takes the general form

$$\min_{\beta} \max_{P \in \mathcal{U}_\delta} \mathbb{E}_P \left[l\left(X, Y, \beta\right) \right], \tag{6.1}$$

where β is a decision variable, (X, Y) is a random element, and $l(x, y, \beta)$ measures a suitable loss or cost incurred when $(X, Y) = (x, y)$ and the decision β is taken. The set $\mathcal{U}_\delta$ is called the distributional uncertainty set and it is indexed by the parameter $\delta > 0$, which measures the size of the distributional uncertainty.

The DRO problem is said to be *data-driven* if the uncertainty set $\mathcal{U}_\delta$ is informed by empirical observations. One natural way to supply this information is by letting the "center" of the uncertainty region be placed at the empirical measure, P_n, induced by a data set $\mathcal{D}_n = \{X_i, Y_i\}_{i=1}^n$, which represents an empirical sample of realizations of (X, Y). In order to emphasize the data-driven nature of a DRO formulation such as (6.1), when the uncertainty region is informed by an empirical sample, we write $\mathcal{U}_\delta = \mathcal{U}_\delta(P_n)$. To the best of our knowledge, the available data is utilized in the DRO literature only by defining the center of the uncertainty region $\mathcal{U}_\delta(P_n)$ as the empirical measure P_n.

Our goal in this chapter is to discuss a data-driven framework to inform the *shape* of $\mathcal{U}_\delta(P_n)$. Throughout this paper, we assume that the class of functions to fit,

indexed by β, is given and that a sensible loss function $l(x, y, \beta)$ has been selected for the problem at hand. Our contribution concerns the construction of the uncertainty region in a fully data-driven way and the implications of this design in machine learning applications. Before providing our construction, let us revisit a example of logistic regression to show the significance of data-driven DRO in the context of machine learning.

In the context of generalized logistic regression, i.e. linear model with log exponential loss,

$$l(x, y, \beta) = \log\left(1 + \exp\left(-y\beta^T x\right)\right),$$

and given empirical samples $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ with $Y_i \in \{-1, 1\}$ and a judicious choice of the distributional uncertainty $\mathcal{U}_\delta(P_n)$ via optimal transport cost, Theorem 2.3 in Chapter 2 shows that

$$\min_\beta \max_{P \in \mathcal{U}_\delta(P_n)} \mathbb{E}_P[l(X, Y, \beta)] = \min_\beta \left(\mathbb{E}_{P_n}[l(X, Y, \beta)] + \delta \|\beta\|_p\right), \tag{6.2}$$

where $\|\cdot\|_p$ is the l_p norm in $\mathbb{R}^d$ for $p \in [1, \infty)$.

The definition of $\mathcal{U}_\delta(P_n)$ turns out to be informed by the dual norm $\|\cdot\|_q$ with $1/p + 1/q = 1$. In simple words, the *shape* of the distributional uncertainty $\mathcal{U}_\delta(P_n)$ directly implies the *type of regularization*; and the *size* of the distributional uncertainty, δ, dictates the regularization parameter.

Similar connections are made for square-root Lasso and SVMs in Chapter 2. In summary, data-driven DRO via optimal transport has been shown to encompass a wide range of prevailing machine learning estimators. However, so far the cost function $c(\cdot)$ has been taken as a given, and not chosen in a data-driven way.

Our main contribution in this chapter is to propose a comprehensive approach

for designing the uncertainty region $\mathcal{U}_\delta(P_n)$ in a fully data-driven way, using the convenient role of $c(\cdot)$ in the definition of the optimal transport discrepancy $D_c(P, P_n)$. Our modeling approach further underscores, beyond the existence of representations such as (6.2), the convenience of working with an optimal transport discrepancy for the design of data-driven DRO machine learning estimators. In other words, because one can select $c(\cdot)$ in a data driven way, it is sensible to use our data-driven DRO formulation even if one is not able to simplify the inner optimization in order to achieve a representation such as (6.2).

Our idea is to apply metric-learning procedures to estimate $c(\cdot)$ from the training data. Then, use such data-driven $c(\cdot)$ in the definition of $D_c(P, P_n)$ and the construction $\mathcal{U}_\delta(P_n)$. Finally, solve the DRO problem (6.1), using cross-validation to choose δ.

The intuition behind our proposal is the following. By using a metric learning procedure we are able to calibrate a cost function $c(\cdot)$ which attaches relatively high transportation costs to (u, w) if transporting mass between these locations substantially impacts performance (e.g. in the response variable, so increasing the expected risk). In turn, the adversary, with a given budget δ, will carefully choose the data which is to be transported. The mechanism will then induce enhanced out-of-sample performance focusing precisely on regions of relevance, while improving generalization error.

One of the challenges for the implementation of our idea is to efficiently solve (6.1). We address this challenge by proposing a stochastic gradient descent algorithm which takes advantage of a duality representation and fully exploits the nature of the LP structure embedded in the definition of $D_c(P, P_n)$, together with a smoothing technique.

Another challenge consists in selecting the type of cost $c(\cdot)$ to be used in practice,

and the methodology to fit such cost. To cope with this challenge, we rely on metric-learning procedures. We do not contribute any novel metric learning methodology; rather, we discuss various parametric cost functions and methods developed in the metric-learning literature. And we discuss their use in the context of fully data-drive DRO formulations for machine learning problems – which is what we propose in this paper. The choice of $c(\cdot)$ ultimately will be influenced by the nature of the data and the application at hand. For example, in the setting of image recognition, it might be natural to use a cost function related to similarity notions.

In addition to discussing intuitively the benefits of our approach in Section 6.2, we also show that our methodology provides a way to naturally estimate various parameters in the setting of adaptive regularization. For example, Theorem 6.1 below, shows that choosing $c(\cdot)$ using a suitable weighted norm, allows us to recover an adaptive regularized ridge regression estimator Ishwaran and Rao [2014]. In turn, using standard techniques from metric learning we can estimate $c(\cdot)$. Hence, our technique connects metric learning tools to estimate the parameters of adaptive regularized estimators.

However, the metric learning based method to choose cost function is not robustified. In addition to applying DRO to improve generalization error, we consider applying robust optimization method in training a data-driven cost function as an additional layer of robustification. One of the driving points of using robust optimization techniques in machine learning is that the introduction of an adversary can be seen as a tool to control the testing error. While the data-driven procedure discussed above is rich in the use of information, and hence it is able to improve the generalization performance, the lack of robustification exposes the testing error to potentially high variability. So, another contribution in this chapter is to design an robust optimization procedure for choosing the shape of $\mathcal{U}_\delta(P_n)$ using a suitable

parametric family. In the context of logistic regression, for example, the parametric family that we consider includes the type of choice leading to (6.2) as a particular case. In turn, the choice of $\mathcal{U}_\delta(P_n)$ is applied to formulation (6.1) in order to obtain a *doubly-robustified* estimator. We call this method to be doubly robust data-driven distributionally robust optimization (DD-R-DRO).

Figure 6.1 shows the various combinations of information and robustness which have been studied in the literature so far. The figure shows four diagrams. Diagram (A) represents standard empirical risk minimization (ERM); which fully uses the information but often leads to high variability in testing error and, therefore, poor out-of-sample performance. Diagram (B) represents DRO where only the center, P_n, and the size of the uncertainty, δ, are data driven; this choice controls out-of-sample performance but does not use data to shape the type of perturbation, thus potentially resulting in testing error bounds which might be pessimistic. Diagram (C) represents DRO with data-driven shape information for perturbation type using metric learning techniques; this construction can reduce the testing error bounds at the expense of increase in the variability of the testing error estimates. Diagram (D) represents DD-R-DRO, the shape of the perturbation allowed for the adversary player is estimated using a robust optimization procedure; this double robustification, as we shall show in the numerical experiments is able to control the variability present in the third diagram.

In the diagrams, the straight arrows represent the use of a robustification procedure. A wide arrow represents the use of high degree of information. A wiggly arrow indicates potentially noisy testing error estimates.

The contributions of this chapter can be stated, in order of importance, as follows:

1) The third diagram, illustrates the first main contribution of this paper, namely, a data-driven approach using metric learning techniques to inform the cost function,

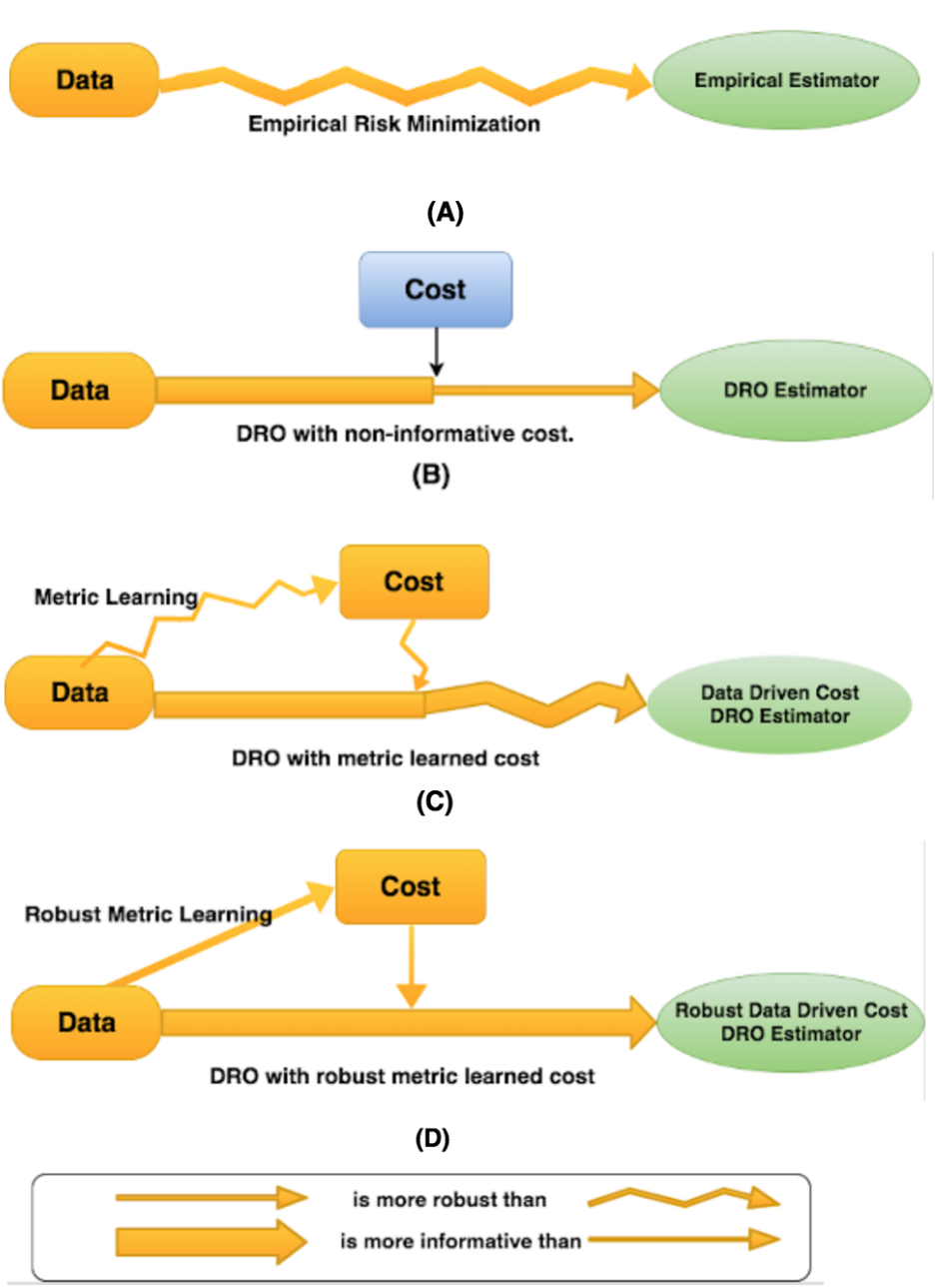

Figure 6.1: Four diagrams illustrating information on robustness.

which could reduces the generalization error.

2) We propose a stochastic gradient based algorithm to solve the DRO problem directly, as we shall observe, the algorithm place very limited assumptions on the loss function, which could be applied for more general machine learning algorithms with DRO formulation.

3) Another main contribution, namely the double robustification approach, as illustrated in the forth diagram, which reduces the generalization error, utilizes information efficiently and controls variability.

4) We also provide an explicit robust optimization formulation for metric learning tasks.

5) In addition, we show an iterative procedures for the solution of these robust optimization problems.

6.2 Data-Driven DRO: Intuition and Interpretations

One of the main benefits of DRO formulations such as (6.1) and (6.2) is their interpretability. For example, we can readily see from the left hand side of (6.2) that the regularization parameter corresponds precisely to the size of the *data-driven* distributional uncertainty.

The data-driven aspect is important because we can employ statistical thinking to optimally characterize the size of the uncertainty, δ. This readily implies an optimal choice of the regularization parameter, as explained in Chapter 2, in settings such as (6.2). Elaborating, we can interpret $\mathcal{U}_\delta(P_n)$ as the set of plausible variations of the empirical data, P_n. Consequently, for instance, in the linear regression setting leading to (6.2), the estimate $\beta_P = \arg\min_\beta \mathbb{E}_P(l(X, Y, \beta))$ is a plausible estimate of

the regression parameter β_* as long as $P \in \mathcal{U}_\delta(P_n)$. Hence, the set

$$\Lambda_\delta(P_n) = \{\beta_P : P \in \mathcal{U}_\delta(P_n)\}$$

is a natural confidence region for β_* which is non-decreasing in δ. Thus, a statistically minded approach for choosing δ is to fix a confidence level, say $(1-\alpha)$, and choose an optimal δ $(\delta_*(n))$ via

$$\delta_*(n) := \inf\{\delta : P(\beta_* \in \Lambda_\delta(P_n)) \geq 1-\alpha\}. \tag{6.3}$$

Note that the random element in $P(\beta_* \in \Lambda_\delta(P_n))$ is given by P_n. In Chapter 2 this optimization problem is solved asymptotically as $n \to \infty$ under standard assumptions on the data generating process. If the underlying model is correct, one would typically obtain, as in Theorem 2.1 of Chapter 2, that $\delta_*(n) \to 0$ at a suitable rate. For instance, in the linear regression setting corresponding to (6.2), if the data is i.i.d. with finite variance and the linear regression model holds then $\delta_*(n) = \chi_{1-\alpha}(1+o(1))/n$ as $n \to \infty$ (where χ_α is the α quantile of an explicitly characterized distribution).

In practice, one can also choose δ by cross-validation. The works in Chapter 2 and Chapter 5 compare the asymptotically optimal choice $\delta_*(n)$ against cross-validation, concluding that the performance is comparable in the experiments performed. In this paper, we use cross validation to choose δ, but the insights behind the limiting behavior of (6.3) are useful, as we shall see, to inform the design of our algorithms.

More generally, the DRO formulation (6.1) is appealing because the distributional uncertainty endows the estimation of β directly with a mechanism to enhance generalization properties. To wit, we can interpret (6.1) as a game in which we (the outer player) choose a decision β, while the adversary (the inner player) selects a model

which is a perturbation, P, of the data (encoded by P_n). The amount of the perturbation is dictated by the size of δ which, as discussed earlier, is data driven. But the type of perturbation and how the perturbation is measured is dictated by $D_c(P, P_n)$. It makes sense to inform the design of $D_c(\cdot)$ using a data-driven mechanism, which is our goal in this paper. The intuition is to allow the types of perturbations which focus the effort and budget of the adversary mostly on out-of-sample exploration over regions of relevance.

The type of benefit that is obtained by informing $D_c\left(P, P_n\right)$ with data is illustrated in Figure 6.2 below. Figure 6.2 illustrates a classification task. The data roughly lies

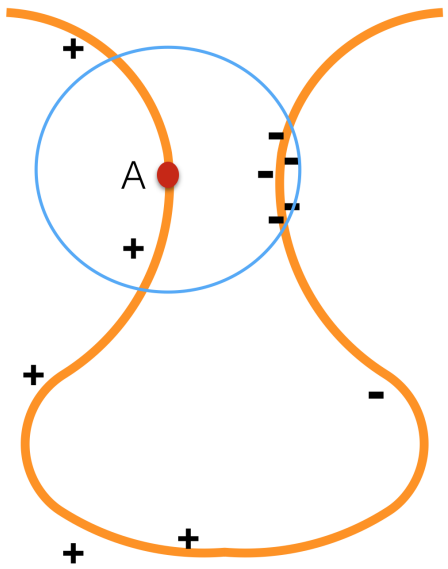

Figure 6.2: Stylized example illustrating the need for data-driven cost function. The data is observed in $\mathbb{R}^2$ but lie in a one dimensional bottle-shaped manifold as marked in orange and $+$ and $-$ are the response labels. formulation and the optimal transport distance, for point A marked in red, if regularized norm in $\mathbb{R}^2$ is applied, we will more likely to assign $-$ pseudo label. However, if we are able to learn the metric alone the manifold, we will more like to transport mass to the points close to A alone the manifold and would be expected to increas learning power.

on a lower dimensional non-linear manifold. Some data which is classified with a negative label is seen to be "close" to data which is classified with a positive label when seeing the whole space (i.e. $\mathbb{R}^2$) as the natural ambient domain of the data. However, if we use a distance similar to the geodesic distance intrinsic to the manifold we would see that the negative instances are actually far from the positive instances. So, the generalization properties of the algorithm would be enhanced relative to using a standard metric in the ambient space, because with a given budget δ the adversarial player would be allowed perturbations mostly along the intrinsic manifold where the data lies and instances which are surrounded (in the intrinsic metric) by similarly classified examples will naturally carry significant impact in testing performance. A quantitative example to illustrate this point will be discussed in the sequel.

6.3 Data-Driven Selection of Optimal Transport Cost Function

In this section we quickly review basic notions on optimal transport and metric learning methods so that we can define $D_c(P, P_n)$ and explain how to calibrate the function $c(\cdot)$.

6.3.1 Revisiting Optimal Transport Distances and Discrepancies

Assume that the cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ is lower semicontinuous. We also assume that $c(u, v) = 0$ if and only if $u = v$. Given two distributions P and Q, with supports $\mathcal{S}_P$ and $\mathcal{S}_Q$, respectively, we define the optimal transport discrepancy,

D_c, via

$$D_c\left(P,Q\right)=\inf_{\pi}\left\{\mathbb{E}_{\pi}\left[c(U,V)\right]:\pi\in\mathcal{P}\left(\mathcal{S}_{_P}\times\mathcal{S}_{_Q}\right),\ \pi_{_U}=P,\ \pi_{_V}=Q\right\},\tag{6.4}$$

where $\mathcal{P}(\mathcal{S}_{_P}\times\mathcal{S}_{_Q})$ is the set of probability distributions π supported on $\mathcal{S}_P\times\mathcal{S}_{_Q}$, and $\pi_{_U}$ and $\pi_{_V}$ denote the marginals of U and V under π, respectively. We can observe that (6.4) is a linear program in the variable π.

6.3.2 On Metric Learning Procedures

In order to keep the discussion focused, we use a few metric learning procedures, but we emphasize that our approach can be used in combination with virtually any method in the metric learning literature, see the survey paper Bellet *et al.* [2013] that contains additional discussion on metric learning procedures. The procedures that we consider, as we shall see, can be seen to already improve significantly upon natural benchmarks. Moreover, as we shall see, these metric families can be related to adaptive regularization. This connection will be useful to further enhance the intuition of our procedure.

6.3.2.1 The Mahalanobis Distance

The Mahalanobis metric is defined as

$$d_{\Lambda}\left(x,x'\right)=\left(\left(x-x'\right)^{T}\Lambda\left(x-x'\right)\right)^{1/2},$$

where Λ is symmetric and positive semi-definite and we write $\Lambda\succeq 0$. Note that $d_\Lambda(x,x')$ is the metric induced by the norm $\|x\|_\Lambda=\sqrt{x^T\Lambda x}$.

For a discussion, assume that our data is of the form $\mathcal{D}_n=\{(X_i,Y_i)\}_{i=1}^n$ and

$Y_i \in \{-1, +1\}$. The prediction variables are assumed to be standardized. Motivated by applications such as social networks, in which there is a natural graph which can be used to connect instances in the data, we assume that one is given sets $\mathcal{M}$ and $\mathcal{N}$, where $\mathcal{M}$ is the set of the pairs that should be close (so that we can connect them) to each other, and $\mathcal{N}$, on contrary, is characterizing the relations that the pairs should be far away (not connected), we define them as

$$\mathcal{M} := \{(i, j) \mid X_i \text{ and } X_j \text{ must connect}\},$$
$$\mathcal{N} := \{(i, j) \mid X_i \text{ and } X_j \text{ should not connect}\}.$$

While it is typically assumed that $\mathcal{M}$ and $\mathcal{N}$ are given, one may always resort to k-Nearest-Neighbor (k-NN) method for the generation of these sets. This is the approach that we follow in our numerical experiments. But we emphasize that choosing any criterion for the definition of $\mathcal{M}$ and $\mathcal{N}$ should be influenced by the learning task in order to retain both interpretability and performance. In our experiments we let (X_i, X_j) belong to $\mathcal{M}$ if, in addition to being sufficiently close (i.e. in the k-NN criterion), $Y_i = Y_j$. If $Y_i \neq Y_j$, then we have that $(X_i, X_j) \in \mathcal{N}$.

In addition, we consider the relative constraint set $\mathcal{R}$ containing data triplets with relative relation defined as

$$\mathcal{R} = \{(i, j, k) \,|\, d_\Lambda(X_i, X_j) \text{ should be smaller than } d_\Lambda(X_i, X_k)\}.$$

Let us consider the following two formulations of metric learning, the so-called Absolute Metric Learning formulation

$$\min_{\Lambda \succeq 0} \sum_{(i,j)\in\mathcal{M}} d^2_\Lambda(X_i, X_j) \quad \text{s.t.} \quad \sum_{(i,j)\in\mathcal{N}} d^2_A(X_i, X_j) \geq 1, \tag{6.5}$$

and the Relative Metric Learning formulation,

$$\min_{\Lambda \succeq 0} \sum_{(i,j,k)\in\mathcal{R}} \left(d_\Lambda^2\left(X_i, X_j\right) - d_\Lambda^2\left(X_i, X_k\right) + 1\right)_+ . \tag{6.6}$$

Both formulations have their merits, Equation (6.5) exploits both the constraint sets $\mathcal{M}$ and $\mathcal{N}$, while Equation (6.6) is only based on information in $\mathcal{R}$. Further intuition or motivation of those two formulations can be found in Xing *et al.* [2002] and Weinberger and Saul [2009], respectively. We will show how to formulate and solve the robust counterpart of those two representative examples by robustifying a single constraint set or two sets simultaneously For simplicity we only discuss these two formulations, but many metric learning algorithms are based on natural generalizations of those two forms, as mentioned in the survey Bellet *et al.* [2013]. In the next two subsections, we will focus on illustrating how to train a data-driven cost function considering the absolute constraints. But it would be easy to notice, same techniques should also apply relative constraint analogues.

6.3.2.2 Using Mahalanobis Distance in Data-Driven DRO

Let us focus on the absolute constraint set case for simplicity. We consider the optimization problem in Equation (6.5), it is a optimization minimizes the total distance between pairs that should be connect, while keeping the pairs that should not connect well separated.

The optimization problem (6.5) is an LP problem on the convex cone PSD (positive semidefinite) and it has been widely studied. Since $\Lambda \succeq 0$, we can always write $\Lambda = LL^T$, and therefore $d_\Lambda(X_i, X_j) = \|X_i - X_j\|_\Lambda := \|LX_i - LX_j\|_2$. There are various techniques which can be used to exploit the PSD structure to efficiently solve (6.5); see, for example, Xing *et al.* [2002] for a projection-based algorithm; or Schultz

and Joachims [2004], which uses a factorization-based procedure; or the survey paper Bellet *et al.* [2013] for the discussion of a wide range of techniques.

We have chosen formulation (6.5) to estimate Λ because it is intuitive and easy to state, but the topic of learning Mahalanobis distances is an active area of research and there are different algorithms which can be implemented (see Li *et al.* [2016]).

Let us assume that the underlying data takes the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in R^d$ and $Y_i \in R$ and the loss function, depending on a decision variable $\beta \in R^m$, is given by $l(x, y, \beta)$. Note that we are not imposing any linear structure on the underlying model or in the loss function. Then, motivated by the cost function $N_q(\cdot)$ introduced in Equation (2.21) of Chapter 2, we may consider

$$c_\Lambda\big((x,y),(x',y')\big) = d_\Lambda^2(x,x')\, I(y=y') + \infty I(y \neq y'), \tag{6.7}$$

for $\Lambda \succeq 0$. The infinite contribution in the definition of c_Λ (i.e. $\infty \cdot I(y \neq y')$) indicates that the adversarial player in the DRO formulation is not allowed to perturb the response variable.

Even in this case, since the definitions of $\mathcal{M}$ and $\mathcal{N}$ depend on (X_i, Y_i) (in particular, on the response variable), cost function $c_\Lambda(\cdot)$ (once Λ is calibrated using, for example, the method discussed in the previous subsection), will be informed by the Y_is. Then, we estimate β via

$$\min_\beta \sup_{P: D_{c_\Lambda}(P, P_n) \leq \delta} \mathbb{E}[l(X, Y, \beta)]. \tag{6.8}$$

It is important to note that Λ has been applied only to the definition of the cost function.

The intuition behind the formulation can be gained in the context of a logistic

regression setting, see the example in Figure 6.3: Suppose that $d = 2$, and that Y depends only on x_1 (i.e. the first coordinate of x). Then, the metric learning procedure in (6.5) will induce a relatively low transportation cost across the x_2 direction and a relatively high transportation cost in the x_1 direction, whose contribution, being highly informative, is reasonably captured by the metric learning mechanism. Since the x_1 direction is most impactful, from the standpoint of expected loss estimation, the adversarial player will reach a compromise, between transporting (which is relatively expensive) and increasing the expected loss (which is the adversary's objective). Out of this compromise the DRO procedure localizes the out-of-sample enhancement, and yet improves generalization.

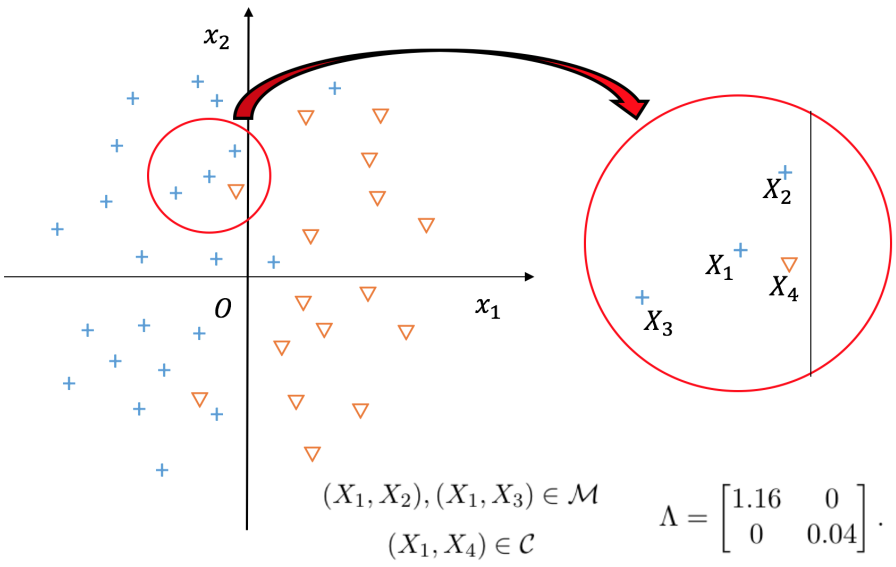

Figure 6.3: Further intuition for data-driven cost based DRO. The figure illustrates an example where the pairs in sets $\mathcal{M}$ and $\mathcal{N}$ get determined, typically, based on the first coordinate of $x = (x_1, x_2)$, and the learned metric $c(x, x') = (x^T \Lambda x')^{1/2} = (1.16(x_1 - x_1')^2 + 0.04(x_2 - x_2')^2)^{1/2}$., where Λ is the learned diagonal matrix.

6.3.2.3 Mahalanobis Metrics on a Non-Linear Feature Space

In this section, we consider the case in which the cost function is defined after applying a non-linear transformation, $\Phi : R^d \to R^l$, to the data. Assume that the data takes the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ and the loss function, depending on decision variable $\beta \in R^m$, is given by $l(x, y, \beta)$. Once again, motivated by the cost function $N_q(\cdot)$ we considered in Equation (2.21) of Chapter 2, we may define

$$c_\Lambda^\Phi\big((x,y),(x',y')\big) = d_\Lambda^2\left(\Phi(x), \Phi(x')\right) I(y = y') + \infty I(y \neq y'), \tag{6.9}$$

for $\Lambda \succeq 0$. To preserve the properties of a cost function (i.e. non-negativity, lower semi-continuity and $c_\Lambda^\Phi(u, w) = 0$ implies $u = w$), we assume that $\Phi(\cdot)$ is continuous and that $\Phi(w) = \Phi(u)$ implies that $w = u$. Then we can apply a metric learning procedure, such as the one described in (6.5), to calibrate Λ. The same observation given in (6.7), regarding the dependence of Λ on the response variables, is applicable here as well (via the definition of $\mathcal{M}$ and $\mathcal{N}$). Once Λ is calibrated our DRO problem becomes

$$\min_\beta \sup_{P: D_{c_\Lambda^\Phi}(P, P_n) \leq \delta} E\left(l(X, Y, \beta)\right).$$

It is important to note that $\Phi(\cdot)$ has been applied only to the definition of the cost function. The intuition is the same as the one provided in the linear case in Section 6.3.2.2.

6.4 Data Driven Cost Selection and Adaptive Regularization

Before we moving forward to consider additional layer of robustification in learning the data-driven cost function, let us we establish a direct connection between our fully data-driven DRO procedure and adaptive regularization. Moreover, our main result here, together with our discussion from the previous section, provides a direct connection between the metric learning literature and adaptive regularized estimators. As a consequence, the methods from the metric learning literature can be used to estimate the parameter of adaptively regularized estimators.

Throughout this section we consider again a data set of the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. Let us define the cost function $c_{_\Lambda}(\cdot)$ as in (6.7).

Using (6.7) we obtain the following result, which is proved in the Appendix 6.A.

Theorem 6.1 (DRO Representation for Generalized Adaptive Regularization). Assume that $\Lambda \in R^{d\times d}$ in (6.7) is positive definite. Given the data set $\mathcal{D}_n$, we obtain the following representation

$$\begin{aligned}&\min_{\beta} \max_{P: D_{c_{_\Lambda}}(P, P_n)\leq \delta} \mathbb{E}_P^{1/2}\left[\left(Y - X^T\beta\right)^2\right] \\ =&\min_{\beta} \sqrt{\frac{1}{n}\sum_{i=1}^n \left(Y_i - X_i^T\beta\right)^{1/2}} + \sqrt{\delta}\left\|\beta\right\|_{\Lambda^{-1}}. \end{aligned} \tag{6.10}$$

Moreover, if $Y \in \{-1, +1\}$ in the context of adaptive regularized logistic regression,

we obtain the following representation

$$\begin{aligned}&\min_{\beta}\max_{P:D_{c_\Lambda}(P,P_n)\leq\delta}\mathbb{E}\left[\log\left(1+e^{-Y(X^T\beta)}\right)\right]\\&=\min_{\beta}\frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{-Y_i(X_i^T\beta)}\right)+\delta\left\|\beta\right\|_{\Lambda^{-1}}.\end{aligned}\tag{6.11}$$

In order to recover a more familiar setting in adaptive regularization, assume that Λ is a diagonal positive definite matrix. In which case we obtain, in the setting of (6.1),

$$\begin{aligned}&\min_{\beta}\max_{P:D_{c_\Lambda}(P,P_n)\leq\delta}\mathbb{E}_P^{1/2}\left[\left(Y-X^T\beta\right)^2\right]\\&=\min_{\beta}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-X_i^T\beta\right)^2}+\sqrt{\delta}\sqrt{\sum_{i=1}^{d}\beta_i^2/\Lambda_{ii}}.\end{aligned}\tag{6.12}$$

The adaptive regularization method was first derived as a generalization for ridge regression in Hoerl and Kennard [1970b,a]. Recent work shows that adaptive regularization can improve the predictive power of its non-adaptive counterpart, specially in high-dimensional settings (see in Zou [2006]; Ishwaran and Rao [2014]).

In view of (6.12), our discussion in Section 6.3.2.1 uncovers tools which can be used to estimate the coefficients $\{1/\Lambda_{ii} : 1 \leq i \leq d\}$ using the connection to metric learning procedures. To complement the intuition given in Figure 1(b), note that in the adaptive regularization literature one often choose $\Lambda_{ii} \approx 0$ to induce $\beta_i \approx 0$ (i.e., there is a high penalty to variables with low explanatory power). This, in our setting, would correspond to transport costs which are low in such low explanatory directions.

6.5 Robust Optimization for Metric Learning

In this section, we review a robust optimization method to metric learning optimization problem to learn a robust data-driven cost function. Robust optimization is a family of optimization techniques that deals with uncertainty or misspecification in the objective function and constraints. Robust Optimization was first proposed in Ben-Tal and Nemirovski [1998, 2002] and has attracted increasing attentions in the recent decades El Ghaoui and Lebret [1997]; Bertsimas *et al.* [2011]. Robust optimization has been applied in machine learning to regularize statistical learning procedures, for example, in Xu *et al.* [2009a,b] robust optimization was employed for SR-Lasso and support vector machines. We apply robust optimization, as we shall demonstrate, to reduce the variability in testing error when implementing DRO.

6.5.1 Robust Optimization for Relative Metric Learning

The robust optimization formulation that we shall use for Equation (6.6) is based on the work of Huang *et al.* [2012]. In order to motivate this formulation, suppose that we know that only α level, e.g. $\alpha = 90\%$, of the constraints are satisfied, but we do not have information on exactly which of them are ultimately satisfied. The value of α may be inferred using cross validation.

Instead of optimizing over all subsets of constraints, we try to minimize the worst case loss function over all possible $\alpha |\mathcal{R}|$ constraints (where $|\cdot|$ is cardinality of a set) and obtain the following min-max formulation

$$\min_{\Lambda \succeq 0} \max_{\tilde{q} \in \mathcal{T}(\alpha)} \sum_{(i,j,k) \in \mathcal{R}} q_{i,j,k} \left(d_\Lambda^2 (X_i, X_j) - d_\Lambda^2 (X_i, X_k) + 1 \right)_+ , \tag{6.13}$$

where $\mathcal{T}(\alpha)$ is a robust uncertainty set of the form

$$\mathcal{T}(\alpha) = \left\{ \tilde{q} = \{q_{i,j,k} | (i,j,k) \in \mathcal{R}\} \, | 0 \leq q_{i,j,k} \leq 1, \sum_{(i,j,k)\in\mathcal{R}} q_{i,j,k} \leq \alpha \times |\mathcal{R}| \right\},$$

which is a convex and compact set.

In addition, the objective function in Equation (6.6) is convex in Λ and concave (linear) in $\tilde{q}$, so we can switch the order of min-max by resorting to Sion's min-max theorem (Terkelsen [1973]). This important observation suggests an iterative algorithm. For a fixed $\Lambda \succeq 0$, the inner maximization is linear in $\tilde{q}$, and the optimal $\tilde{q}$ satisfy $\tilde{q}_{i,j,k} = 1$ whenever $\left(d_\Lambda(X_i, X_j) - d_\Lambda(X_i, X_k) + 1\right)_+$ ranks in the top $\alpha|\mathcal{R}|$ largest values and equals $\tilde{q}_{i,j,k}$ otherwise, i.e.

$$\tilde{q}_{i,j,k} = \begin{cases} 1 & \text{if } \left(d_\Lambda(X_i, X_j) - d_\Lambda(X_i, X_k) + 1\right)_+ \text{ ranks tops} \alpha \times |\mathcal{R}| \text{ within } \mathcal{R} \\ 0, & \text{otherwise.} \end{cases}.$$

Let us use $\mathcal{R}_\alpha(\Lambda)$ to denote the subset of constraints satisfying that the corresponding loss function $\left(d_\Lambda(X_i, X_j) - d_\Lambda(X_i, X_k) + 1\right)_+$ ranks at the top $\alpha|\mathcal{R}|$ largest values among the corresponding loss function values of the triplets in $\mathcal{R}$.

For fixed $\tilde{q}$, the optimization problem is convex in Λ, we can solve this problem using sub-gradient or smoothing approximation algorithms (Nesterov [2005]). Particularly, as we discussed above, if $\tilde{q}$ is the solution for fixed Λ, we know, solving Λ is equivalent to solving its non-robust counterpart Equation (6.6), replacing $\mathcal{R}$ by $\mathcal{R}_\alpha(\Lambda)$, where $\mathcal{R}_\alpha(\Lambda)$ is a subset of $\mathcal{R}$ that contains the constraints have top $\alpha|\mathcal{R}|$ violation, i.e.

$$\mathcal{R}_\alpha(\Lambda) = \left\{ (i,j,k) \in \mathcal{R} | \left(d^2_\Lambda(X_i, X_j) - d^2_\Lambda(X_i, X_k) + 1\right)_+ \text{ ranks top } \alpha \text{ within } \mathcal{R} \right\}.$$

We summarize the sub-gradient based sequentially update algorithm as in Algorithm 6.1.

Algorithm 6.1 Sequential Coordinate-wise Metric Learning Using Relative Relations

1: **Initialize** $\Lambda = I_d$, learning rate $\alpha = 0.01$ tracking error $Error = 1000$ as a large number. Then randomly sample α proportion of elements from $\mathcal{R}$ to construct $\mathcal{R}_\alpha(\Lambda)$.
2: **while** $Error > 10^{-3}$ **do**
3: Update Λ using projected (projected to positive semidefinite matrix cone) subgradient descent technique.

$$\Lambda = \pi_{\mathbb{S}_+}\left(\Lambda - \alpha \sum_{(i,j,k)\in\mathcal{R}_\alpha(\Lambda)} \nabla_\Lambda \left(d^2_\Lambda\left(X_i, X_j\right) - d^2_\Lambda\left(X_i, X_k\right) + 1\right)_+\right)$$

4: Update tracking error $Error$ as the norm of difference between latest matrix Λ and average of last 50 iterations.
5: Every few steps (5 or 10 iterations), compute $\left(d^2_\Lambda\left(X_i, X_j\right) - d^2_\Lambda\left(X_i, X_k\right) + 1\right)_+$ for all $(i, j, k) \in \mathcal{R}$, then update $\mathcal{R}_\alpha(\Lambda)$.
6: **end while**
7: **Output** Λ.

As a remark, we would like to highlight the following observations. While we focus on metric learning simply as a loss minimization procedure as in Equation (6.6) and Equation (6.13) for simplicity, in practice people usually add a regularization term (such as $\|\Lambda\|_F$) to the loss minimization, as is common in metric learning literature (see Bellet *et al.* [2013]). It is easy to observe our discussion above regarding the min-max exchange uses Sion's min-max theorem and everything else remains largely intact if we consider regularization. Likewise, one can use a more general loss functions than the hinge loss used in Equation (6.6) and Equation (6.13).

6.5.2 Robust Optimization for Absolute Metric Learning

The robust optimization formulation that we present here for Equation (6.5) appears to be novel in the literature. Note that Equation (6.5) can be written into the La-

grangian form,

$$\min_{\Lambda\succeq 0} \quad \max_{\lambda\geq 0} \quad \sum_{(i,j)\in\mathcal{M}} d_{\Lambda}^{2}\left(X_i, X_j\right) + \lambda\big(1 - \sum_{(i,j)\in\mathcal{N}} d_{\Lambda}^{2}\left(X_i, X_j\right)\big).$$

Following similar discussion for $\mathcal{R}$, let us assume that 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 can construct robust uncertainty sets $\mathcal{W}(\alpha)$ and $\mathcal{V}(\alpha)$ from the constraints in $\mathcal{M}$ and $\mathcal{N}$ as follows,

$$\mathcal{W}(\alpha) = \big\{\tilde{\eta} = \{\eta_{ij} : (i,j) \in \mathcal{M}\}\,|0 \leq \eta_{ij} \leq 1, \sum_{(i,j)\in\mathcal{M}} \eta_{ij} \leq \alpha \times |\mathcal{M}|\big\},$$

$$\mathcal{V}(\alpha) = \big\{\tilde{\xi} = \{\xi_{ij} : (i,j) \in \mathcal{N}\}\,|0 \leq \xi_{ij} \leq 1, \sum_{(i,j)\in\mathcal{N}} \xi_{ij} \geq \alpha \times |\mathcal{N}|\big\}.$$

Then we can write the robust optimization counterpart for the loss minimization problem of metric learning as

$$\min_{\Lambda\succeq 0} \quad \max_{\lambda\geq 0} \quad \max_{\tilde{\eta}\in\mathcal{W}(\alpha),\tilde{\xi}\in\mathcal{V}(\alpha)} \sum_{(i,j)\in\mathcal{M}} \eta_{i,j} d_{\Lambda}^{2}\left(X_i, X_j\right) + \lambda\big(1 - \sum_{(i,j)\in\mathcal{N}} \xi_{i,j} d_{\Lambda}^{2}\left(X_i, X_j\right)\big) \quad (6.14)$$

Note that the Cartesian product $\mathcal{W}\left(\alpha\right) \times \mathcal{V}\left(\alpha\right)$ is a compact set, and the objective function is convex in Λ and concave (linear) in pair $(\tilde{\eta}, \tilde{\xi})$, so we can apply Sion's min-max Theorem again (see in Terkelsen [1973]) to switch the order of $\min_{\Lambda}$-$\max_{(\tilde{\eta},\tilde{\xi})}$ (after switching $\max_{\lambda}$ and $\max_{(\tilde{\eta},\tilde{\xi})}$, which can be done in general). Then we can develop a sequential iterative algorithm to solve this problem as we describe next.

At the k-th step, given fixed $\Lambda_{k-1} \succeq 0$ and $\lambda_{k-1} > 0$ (it is easy to observe that optimal solution λ is positive, i.e. the constraint is active so we may safely assume

$\lambda_{k-1} > 0$), the inner maximization problem, becomes,

$$\max_{\tilde{\eta}\in\mathcal{W}(\alpha)} \sum_{(i,j)\in\mathcal{M}} \eta_{i,j} d^2_{\Lambda_{k-1}}(X_i, X_j) + \lambda\big(1 - \min_{\tilde{\xi}\in\mathcal{V}(\alpha)} \sum_{(i,j)\in\mathcal{N}} \xi_{i,j} d^2_{\Lambda_{k-1}}(X_i, X_j)\big).$$

As we discussed for relative constraints case, the optimal solution for $\tilde{\eta}^{(k)}$ and $\tilde{\xi}^{(k)}$ is, $\tilde{\eta}^{(k)}_{i,j}$ is 1, if $d^2_{\Lambda_{k-1}}(X_i, X_j)$ ranks top α within $\mathcal{M}$ and equals 0 otherwise; while, on the contrary, $\tilde{\xi}^{(k)}_{i,j} = 1$ if $d^2_{\Lambda_{k-1}}(X_i, X_j)$ ranks bottom α within $\mathcal{N}$ and equals 0 otherwise.

Similar as $\mathcal{R}_\alpha(\Lambda)$, we can define $\mathcal{M}_\alpha(\Lambda_{k-1})$ as subset of $\mathcal{M}$, which contains the constraints with largest α percent of $d_{\Lambda_{k-1}}(\cdot)$ within in $\mathcal{M}$; and $\mathcal{N}_\alpha(\Lambda_{k-1})$ as subset of $\mathcal{N}$, which contains the constraints with smallest α percent of $d_{\Lambda_{k-1}}(\cdot)$ within in $\mathcal{N}$. As we observe that the optimal $\tilde{\eta}_{i,j} = 1$ if $(i,j) \in \mathcal{M}_\alpha(\Lambda_{k-1})$ and $\tilde{\xi}_{i,j} = 1$ if $(i,j) \in \mathcal{N}_\alpha(\Lambda_{k-1})$, thus for fixed $\tilde{\eta}$ and $\tilde{\xi}$, we can write the optimization problem over Λ in the constrained case as

$$\min_{\Lambda\succeq 0} \quad \sum_{(i,j)\in\mathcal{M}_\alpha(\Lambda_{k-1})} d^2_\Lambda(X_i, X_j) \quad \text{s.t.} \quad \sum_{(i,j)\in\mathcal{N}_\alpha(\Lambda_{k-1})} d^2_\Lambda(X_i, X_j) \geq 1.$$

This formulation falls within the setting of the problem stated in Equation (6.5) and thus it can be solved by using techniques discussed in Xing *et al.* [2002]. We summarize the details in Algorithm 6.2.

Other robust methods have also been considered in the metric learning literature, see Zha *et al.* [2009]; Lim *et al.* [2013] although the connections to robust optimization are not fully exposed.

Algorithm 6.2 Sequential Coordinate-wise Metric Learning Using Absolute Constraints

1: **Initialize** $A = I_d$, tracking error Error = 1000 as a large number. Then randomly sample α proportion of elements from $\mathcal{M}$ (resp. $\mathcal{N}$) to construct $\mathcal{M}_\alpha(A)$ (resp. $\mathcal{N}_\alpha(A)$).
2: **while** $Error > 10^{-3}$ **do**
3: Update A using procedure provided in Xing *et al.* [2002].
4: Update tracking error $Error$ as the norm of difference between latest matrix A and average of last 50 iterations.
5: Every few steps (5 or 10 iterations), compute $d_A(W_i, W_j)$ for all $(i,j) \in \mathcal{M} \cup \mathcal{N}$, then update $\mathcal{M}_\alpha(A)$ and $\mathcal{N}_\alpha(A)$.
6: **end while**
7: **Output** A.

6.6 Solving Data Driven DRO Based on Optimal Transport Discrepancies

In order to fully take advantage of the combination synergies between metric learning methodology and our DRO formulation, it is crucial to have a methodology which allows us to efficiently estimate β in DRO problems such as (6.1). In the presence of a simplified representation such as (6.2) or (6.12), we can apply standard stochast-LRic optimization results (see Lei and Jordan [2016]).

Our objective in this section is to study algorithms which can be applied for more general loss and cost functions, for which a simplified representation might not be accessible.

Throughout this section, once again we assume that the data is given in the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^{d+1}$. The loss function is written as $\{l(x, y, \beta) : (x, y) \in \mathbb{R}^{d+1}, \beta \in \mathbb{R}^m\}$. We assume that for each (x, y), the function $l(x, y, \cdot)$ is convex and continuously differentiable. Further, we shall consider cost functions of the form

$$\bar{c}((x, y), (x', y')) = c(x, x') I(y = y') + \infty I(y \neq y'),$$

as this will simplify the form of the dual representation in the inner optimization of our DRO formulation. To ensure boundedness of our DRO formulation, we impose the following assumption.

Assumption A1. There exists $\Gamma(\beta, y) \in (0, \infty)$ such that $l(u, y, \beta) \leq \Gamma(\beta, y) \cdot (1 + c(u, x))$, for all $(x, y) \in \mathcal{D}_n$, Under Assumption A1, we can guarantee that

$$\max_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P \left[l \left(X, Y, \beta \right) \right] \leq (1 + \delta) \max_{i=1,\ldots,n} \Gamma \left(\beta, Y_i \right) < \infty.$$

Using the strong duality theorem for semi-infinity linear programming problem in Appendix 2.B of Chapter 2,

$$\max_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P \left[l \left(X, Y, \beta \right) \right] = \min_{\lambda \geq 0} \frac{1}{n} \sum_{i=1}^{n} \phi \left(X_i, Y_i, \beta, \lambda \right), \tag{6.15}$$

where

$$\psi(u, X, Y, \beta, \lambda) := l(u, Y, \beta) - \lambda(c(u, X) - \delta),$$

and

$$\phi \left(X, Y, \beta, \lambda \right) := \max_{u \in \mathbb{R}^d} \psi(u, X, Y, \beta, \lambda).$$

Therefore,

$$\min_{\beta} \max_{P: D_{c_\Lambda}(P, P_n) \leq \delta} \mathbb{E}_P \left[l \left(X, Y, \beta \right) \right] = \min_{\lambda \geq 0, \beta} \left\{ \mathbb{E}_{P_n} \left[\phi \left(X, Y, \beta, \lambda \right) \right] \right\}. \tag{6.16}$$

The optimization in Equation (6.16) is minimize over β and λ, which we can consider stochastic approximation algorithm if the gradient of $\phi(\cdot)$ with respect to β and λ exist. However, $\phi(\cdot)$ is given in the form of the value function of a maximization problem, of which the gradient is not easy accessible. We will discuss the detailed algorithm and the validity of the smoothing approximation below.

We consider a smoothing approximation technique to remove the maximization problem $\phi(\cdot)$ using soft-max counterpart, $\phi_{\epsilon,f}(\cdot)$. The smoothing soft-max approximation has been explored and applied to approximately solve the DRO problem for the discrete case, where we restrict the distributionally uncertainty set only contains probability measures support on finite set (i.e., labeled training data and unlabeled training data with pseudo labels), we refer Chapter 4 for further details.

However, due to the continuous-infinite support constraint, the soft-max approximation is a non-trivial generalization of the finite-discrete analogue. The smoothing approximation for $\phi(\cdot)$ is defined as,

$$\phi_{\epsilon,f}(X,Y,\beta,\lambda) = \epsilon \log\left(\int_{\mathbb{R}^d} \exp\left(\left[\psi(u,X,Y,\beta,\lambda)\right]/\epsilon\right) f(u)\, du\right),$$

where $f(\cdot)$ is a probability density in $\mathbb{R}^d$; for example, we can consider a multivariate normal distribution and ϵ is a small positive number regarded as smoothing parameter.

Let us quantify the error induced by replacing $\phi(\cdot)$ with $\phi_{\epsilon,f}(\cdot)$. To this end, we introduce some notations and assumptions. For any set S, the r-neighborhood of S is defined as the set of all points in $\mathbb{R}^d$ that are at distance less than r from S, i.e. $S_r = \cup_{u\in S}\{\bar{u} : \|\bar{u}-u\|_2 \le r\}$. In addition, we write $f(\cdot)$ as the density of an absolutely continuous probability measure $f(\cdot)$.

Assumption A2. $\psi(\cdot, X, Y, \beta, \lambda)$ is twice continuously differentiable and the Hessian of $\psi(\cdot, X, Y, \beta, \lambda)$ evaluated at u^*, $D_u^2\psi(u^*, X, Y, \beta, \lambda)$, is positive definite. In particular, we can find $\theta > 0$ and $\eta > 0$, such that

$$\psi(u,X,Y,\beta,\lambda) \ge \psi(u^*,X,Y,\beta,\lambda) - \frac{\theta}{2}\|u-u^*\|_2^2, \quad \forall u \text{ with } \|u-u^*\|_\infty \le \eta.$$

Assumption A3. For a constant $\lambda_0 > 0$ such that $\phi(X,Y,\beta,\lambda_0) < \infty$, let

$K = K(X, Y, \beta, \lambda_0)$ be any upper bound for $\phi(X, Y, \beta, \lambda_0)$.

Assumption A4. In addition to the lower semi-continuity of $c(\cdot) \geq 0$, we assume that $c(\cdot, X)$ is coercive in the sense that $c(u, X) \to \infty$ whenever $\|u\|_2 \to \infty$.

Then, under Assumptions 3 and 4, we can define the compact set

$$\mathcal{C} = \mathcal{C}(X, Y, \beta, \lambda) = \{u : c(u, X) \leq l(X, Y, \beta) - K + \lambda_0/(\lambda - \lambda_0)\}.$$

It is easy to check that $\arg\max\{\psi(u, X, Y, \lambda)\} \subset \mathcal{C}$. Theorem 6.2 below allows to quantify the error due to smoothing approximation.

Theorem 6.2. Under Assumption A1-A4, there exists $\epsilon_0 > 0$ such that for every $\epsilon < \epsilon_0$, we have

$$\phi(X, Y, \beta, \lambda) \geq \phi_{\epsilon,f}(X, Y, \beta, \lambda) \geq \phi(X, Y, \beta, \lambda) - d\epsilon \log(1/\epsilon)$$

Proof of Theorem 6.2 is included in the Appendix 6.B. Assumptions A2 and A4 are easily verified if once chooses $c_A(\cdot)$ in terms of the Mahalanobis distance. The bound $K(X, Y, \beta, \lambda_0)$ introduced in Assumption 3 can be easily obtained in order to construct $\mathcal{C}(X, Y)$ containing $\arg\max\{\psi(u, X, Y, \lambda)\}$. For instance, consider the setting for adaptive regularized logistic regression as in Theorem 6.1, we can verify that $\lambda_0 = \|\beta\|_{\Lambda^{-1}}$ and $K(X, Y, \beta, \lambda_0) = \log\left(1 + \exp\left(-Y\beta^T X\right)\right)$ are valid choices which satisfy Assumption A3.

After applying smooth approximation, the optimization problem turns into a standard stochastic optimization problem and we can apply mini-batch based stochastic approximation (SA) algorithm to solve it. As we can notice, as a function and β and

λ, the gradient of $\phi_{\epsilon,f}(\cdot)$ satisfies

$$\nabla_\beta \phi_{\epsilon,f}(X,Y,\beta,\lambda) = \frac{\mathbb{E}_{U\sim f}\left[\exp\left(\psi(U,X,Y,\beta,\lambda)/\epsilon\right)\nabla_\beta l\left(f_\beta(U),Y\right)\right]}{\mathbb{E}_{U\sim f}\left[\exp\left(\psi(U,X,Y,\beta,\lambda)/\epsilon\right)\right]},$$
$$\nabla_\lambda \phi_{\epsilon,f}(X,Y,\beta,\lambda) = \frac{\mathbb{E}_{U\sim f}\left[\exp\left(\psi(u,X,Y,\beta,\lambda)/\epsilon\right)\left(\delta - c_{\mathcal{D}_n}(u,X)\right)\right]}{\mathbb{E}_{U\sim f}\left[\exp\left(\psi(U,X,Y,\beta,\lambda)/\epsilon\right)\right]}.$$

However, since the gradients are still given in the form of expectation, we can apply a simple Monte Carlo sampling algorithm to approximate the gradient, i.e., we sample U_i's from $f(\cdot)$ and evaluate the numerators and denominators of the gradient using Monte Carlo separately. For more details of the SA algorithm, please see in Algorithm 6.3.

Algorithm 6.3 Stochastic Gradient Descent with Continuous State

Initialize $\lambda = 0$, and β to be empirical risk minimizer, $\epsilon = 0.5$, tracking error $Error = 100$.

while $Error > 10^{-3}$ **do**

Sample a mini-batch uniformly from observations $\left\{X_{(j)}, Y_{(j)}\right\}_{j=1}^{M}$, with $M \leq n$.

For each $j = 1, \ldots, M$, sample i.i.d. $\{U_k^{(j)}\}_{k=1}^{L}$ from $\mathcal{N}(0, \sigma^2 I_{d\times d})$.

We denote f_L^j as empirical distribution for $U_k^{(j)}$'s, and estimate the batched as

$$\nabla_\beta \phi_{\epsilon,f} = \frac{1}{M}\sum_{j=1}^{M} \nabla_\beta \phi_{\epsilon,f_L^j}\left(X_{(j)}, Y_{(j)}, \beta, \lambda\right), \ \nabla_\lambda \phi_{\epsilon,f} = \frac{1}{M}\sum_{j=1}^{M} \nabla_\lambda \phi_{\epsilon,f_L^j}\left(X_{(j)}, Y_{(j)}, \beta, \lambda\right).$$

Update β and λ using

$$\beta = \beta + \alpha_\beta \nabla_\beta(L)\phi_{\epsilon,f}(X,Y,\beta,\lambda), \ \lambda = \left(\lambda + \alpha_\lambda \nabla_\lambda(L)\phi_{\epsilon,f}(X,Y,\beta,\lambda)\right)_+.$$

Update tracking error $Error$ as the norm of difference between latest parameter and average of last 50 iterations.

end while

Output β.

6.7 Numerical Experiments

We validate our data-driven cost function based DRO using six real data examples from the UCI machine learning database Lichman [2013]. We focus on a DRO formulation based on the log-exponential loss for a linear model. We consider logistic regression (LR), regularized logistic regression (LRL1), DRO with cost function learned using absolute constraints (DRO (absolute)) and its $\alpha = 50\%, 90\%$ level of doubly robust DRO (DD-R-DRO (absolute)); DRO with cost function learned using absolute constraints with linear and quadratic polynomial transformation of the data (DRO-NL (absolute)), and its $\alpha = 50\%, 90\%$ level of doubly robust DRO (DD-R-DRO (absolute)); DRO with cost function learned using relative constraints (DRO (relative)) and its $\alpha = 50\%, 90\%$ level of doubly robust DRO (DD-R-DRO (relative)).

For each iteration and each data set, the data is split randomly into training and test sets. We fit the models on the training and evaluate the performance on test set. The regularization parameter is chosen via $5-$fold cross-validation for LRL1, DRO-L and DRO-NL. We report the mean and standard deviation for training and testing log-exponential error and testing accuracy for 200 independent experiments for each data set. The details of the numerical results and basic information of the data is summarized in Table 6.1.

We observe that DRO with the data-driven cost function could improve the generalization performance comparing to the empirical risk minimization problem (logistic regression) and its DRO counterpart with regular Euclidean norm as cost function (regularized logistic regression). The doubly robust DRO framework, in general, get robust improvement comparing to its non-robust counterpart with $\alpha = 90\%$. More importantly, the robust methods tend to enjoy the variance reduction property due to robust optimization. Also, as the robust level increases, i.e. $\alpha = 50\%$, where we

		BC	BN	QSAR	Magic	MB	SB
LR	Train	0 ± 0	$.008 \pm .003$	$.026 \pm .008$	$.213 \pm .153$	0 ± 0	0 ± 0
	Test	8.75 ± 4.75	2.80 ± 1.44	35.5 ± 12.8	17.8 ± 6.77	18.2 ± 10.0	14.5 ± 9.04
	Accur	$.762 \pm .061$	$.926 \pm .048$	$.701 \pm .040$	$.668 \pm .042$	$.678 \pm .059$	$.789 \pm .035$
LRL1	Train	$.185 \pm .123$	$.080 \pm .030$	$.614 \pm .038$	$.548 \pm .087$	$.401 \pm .167$	$.470 \pm .040$
	Test	$.428 \pm .338$	$.340 \pm .228$	$.755 \pm .019$	$.610 \pm .050$	$.910 \pm .131$	$.588 \pm .140$
	Accur	$.929 \pm .023$	$.930 \pm .042$	$.646 \pm .036$	$.665 \pm .045$	$.717 \pm .041$	$.811 \pm .034$
DRO (absolute)	Train	$.022 \pm .019$	$.197 \pm .112$	$.402 \pm .039$	$.469 \pm .064$	$.294 \pm .046$	$.166 \pm .031$
	Test	$.126 \pm .034$	$.275 \pm .093$	$.557 \pm .023$	$.571 \pm .043$	$.613 \pm .053$	$.333 \pm .023$
	Accur	$.954 \pm .015$	$.919 \pm .050$	$.733 \pm .0.026$	$.727 \pm .039$	$.714 \pm .032$	$.887 \pm .011$
DD-R-DRO (absolute) $\alpha = 90\%$	Train	$.029 \pm .013$	$.078 \pm .031$	$.397 \pm .036$	$.420 \pm .063$	$.249 \pm .055$	$.194 \pm .031$
	Test	$.126 \pm .023$	$.259 \pm .086$	$.554 \pm .019$	$.561 \pm .035$	$.609 \pm .044$	$.331 \pm .018$
	Accur	$.954 \pm .012$	$.910 \pm .042$	$.736 \pm .025$	$.729 \pm .032$	$.709 \pm .025$	$.890 \pm .008$
DD-R-DRO (absolute) $\alpha = 50\%$	Train	$.040 \pm .055$	$.137 \pm .030$	$.448 \pm .032$	$.504 \pm .041$	$.351 \pm .048$	$.166 \pm .030$
	Test	$.132 \pm .015$	$.288 \pm .059$	$.579 \pm .017$	$.590 \pm .029$	$.623 \pm .029$	$.337 \pm .013$
	Accur	$.952 \pm .012$	$.918 \pm .037$	$.733 \pm .025$	$.710 \pm .033$	$.715 \pm .021$	$.888 \pm .008$
DRO-NL (absolute)	Train	$.032 \pm .015$	$.113 \pm .035$	$.339 \pm .044$	$.381 \pm .084$	$.287 \pm .049$	$.195 \pm .034$
	Test	$.119 \pm .044$	$.194 \pm .067$	$.557 \pm .032$	$.577 \pm ..049$	$.607 \pm .060$	$.332 \pm .015$
	Accur	$.955 \pm .016$	$.931 \pm .036$	$.736 \pm .027$	$.730 \pm .043$	$.716 \pm .054$	$.889 \pm .009$
DD-R-DRO-NL (absolute) $\alpha = 90\%$	Train	$.018 \pm .008$	$.049 \pm .030$	$.367 \pm .043$	$.420 \pm .080$	$.269 \pm .056$	$.196 \pm .031$
	Test	$.113 \pm .030$	$.209 \pm .053$	$.551 \pm .022$	$.567 \pm .033$	$.603 \pm .052$	$.332 \pm .013$
	Accur	$.954 \pm .011$	$.926 \pm .037$	$.740 \pm .026$	$.731 \pm .032$	$.716 \pm .027$	$.889 \pm .008$
DD-R-DRO-NL (absolute) $\alpha = 50\%$	Train	$.045 \pm .005$	$.283 \pm .029$	$.397 \pm .044$	$.383 \pm .079$	$.201 \pm .054$	$.157 \pm .030$
	Test	$.136 \pm .023$	$.266 \pm .044$	$.559 \pm .019$	$.580 \pm .030$	$.614 \pm .041$	$.341 \pm .010$
	Accur	$.950 \pm .010$	$.915 \pm .033$	$.733 \pm .026$	$.726 \pm .021$	$.709 \pm .026$	$.888 \pm .009$
DRO (relative)	Train	$.086 \pm .038$	$.436 \pm .138$	$.392 \pm .040$	$.457 \pm .071$	$.322 \pm .061$	$.181 \pm, 036$
	Test	$.153 \pm .060$	$.329 \pm .124$	$.559 \pm .025$	$582 \pm .033$	$.613 \pm .031$	$.332 \pm .016$
	Accur	$.946 \pm .018$	$.916 \pm .075$	$.714 \pm .029$	$.710 \pm, 027$	$.704 \pm .021$	$.890 \pm .008$
DD-R-DRO (relative) $\alpha = 90\%$	Train	$.030 \pm .014$	$.244 \pm .121$	$.375 \pm .038$	$.452 \pm .067$	$.402 \pm .058$	$.234 \pm .032$
	Test	$.141 \pm .054$	$.300 \pm .108$	$.556 \pm .022$	$.577 \pm .032$	$.610 \pm .024$	$.332 \pm .011$
	Accur	$.949 \pm .019$	$.921 \pm .070$	$.729 \pm .023$	$.717 \pm .025$	$.710 \pm .020$	$.892 \pm .007$
DD-R-DRO (relative) $\alpha = 90\%$	Train	$.031 \pm .016$	$.232 \pm .094$	$.445 \pm .032$	$.544 \pm .057$	$.365 \pm .054$	$.288 \pm .029$
	Test	$.154 \pm .049$	$.319 \pm .078$	$.570 \pm .019$	$.594 \pm .018$	$.624 \pm .018$	$.357 \pm .008$
	Accur	$.948 \pm .019$	$.918 \pm .081$	$.705 \pm .023$	$.699 \pm .028$	$.698 \pm .018$	$.881 \pm .005$
Num Predictors		30	4	30	10	20	56
Train Size		40	20	80	30	30	150
Test Size		329	752	475	9990	125034	2951

Table 6.1: Numerical results of data-driven optimal transportation cost selection DRO with real data sets.

believe in higher noise in cost function learning, we can observe, the doubly robust based approach seems to shrink towards to regularized logistic regression, and benefits less from the data-driven cost structure.

6.8 Conclusion and Discussion

Our fully data-driven DRO procedure combines a semi-parametric approach (i.e. the metric learning procedure) with a parametric procedure (expected loss minimization) to enhance the generalization performance of the underlying parametric model. We emphasize that our approach is applicable to any data-driven DRO formulation and is not restricted to classification tasks. An interesting research avenue that might be considered include the development of a semi-supervised framework as in Chapter 4, in which unlabeled data is used to inform the support of the elements in $\mathcal{U}_\delta(P_n)$.

In addition, we introduced the doubly robust approach, DD-R-DRO, which calibrates a transportation cost function by using a data-driven approach based on robust optimization. The overall methodology is doubly robust. On one hand, data driven DRO, which fully uses the training data to estimate the underlying transportation cost enhances out-of-sample performance by allowing an adversary to perturb the data (represented by the empirical distribution) in order to obtain bounds on the testing risk which are tight. On the other hand, the tightness of bounds might come at the cost of potentially introducing noise in the testing error performance. The second layer of robustification, as shown in the numerical examples, mitigates precisely the presence of this noise.

APPENDIX 6.A: Proof of Theorem 6.1

We first state and prove Lemma 6.1 which will be useful in proving Theorem 6.1.

Lemma 6.1. If Λ is a is positive definite matrix and we define $\|x\|_\Lambda = \left(x^T \Lambda x\right)^{1/2}$, then $\|\cdot\|_{\Lambda^{-1}}$ is the dual norm of $\|\cdot\|_\Lambda$. Furthermore, we have

$$u^T w \leq \|u\|_\Lambda \|w\|_{\Lambda^{-1}},$$

where the equality holds if and only if, there exists non-negative constant τ, s.t $\tau\Lambda u = \Lambda^{-1} w$ or $\tau\Lambda^{-1} w = \Lambda u$.

Proof for Lemma 6.1. This result is a direct generalization of l_2 norm in Euclidean space. Note that

$$u^T w = (\Lambda u)^T \left(\Lambda^{-1} w\right) \leq \|\Lambda u\|_2 \left\|\Lambda^{-1} w\right\|_2 = \|u\|_\Lambda \|w\|_{\Lambda^{-1}}. \tag{6.17}$$

The inequality in the above is Cauchy-Schwartz inequality for $\mathbb{R}^d$ applies to Λu and $\Lambda^{-1} w$, and the equality holds if and only if there exists nonnegative τ, s.t. $\tau\Lambda u = \Lambda^{-1} w$ or $\tau\Lambda^{-1} w = \Lambda u$. Now, by the definition of the dual norm,

$$\|w\|_\Lambda^* = \sup_{u:\|u\|_\Lambda \leq 1} u^T w = \sup_{u:\|u\|_\Lambda \leq 1} \|u\|_\Lambda \|w\|_{\Lambda^{-1}} = \|w\|_{\Lambda^{-1}}.$$

While the first equality follows from the definition of dual norm, the second equality is due to Cauchy-Schwartz inequality (6.17), and the equality condition therein, and the last equality are immediate after maximizing. □

Proof for Theorem 6.1. The technique is a generalization of the method used in proving Theorem 2.2 in Chapter 2. We can apply the strong duality result, see Proposition

2.1 in Appendix of Chapter 2, for worst-case expected loss function, which is a semi-infinite linear programming problem, to obtain

$$\sup_{P:D_{c_\Lambda}(P,P_n)\leq\delta} \mathbb{E}_P\left[\left(Y-X^T\beta\right)^2\right] = \min_{\gamma\geq 0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^{n}\sup_u\left\{\left(y_i-u^T\beta\right)^2-\gamma\left\|x_i-u\right\|_\Lambda^2\right\}\right\}.$$

For the inner suprema , let us denote $\Delta = u - X_i$ and $e_i = Y_i - X_i^T\beta$ for notation simplicity. The inner optimization problem associated with (X_i, Y_i) becomes,

$$\begin{aligned}
&\sup_u\left\{\left(y_i-u^T\beta\right)^2-\gamma\left\|x_i-u\right\|_\Lambda^2\right\}\\
&\quad= e_i^2+\sup_\Delta\left\{\left(\Delta^T\beta\right)^2-2e_i\Delta^T\beta-\gamma\left\|\Delta\right\|_\Lambda^2\right\},\\
&\quad= e_i^2+\sup_\Delta\left\{\left(\sum_j\left|\Delta_j\right|\left|\beta_j\right|\right)^2+2\left|e_i\right|\sum_j\left|\Delta_j\right|\left|\beta_j\right|-\gamma\left\|\Delta\right\|_\Lambda^2\right\},\\
&\quad= e_i^2+\sup_{\left\|\Delta\right\|_\Lambda}\left\{\left\|\Delta\right\|_\Lambda^2\left\|\beta\right\|_{\Lambda^{-1}}^2+2\left|e_i\right|\left\|\Delta\right\|_\Lambda\left\|\beta\right\|_{\Lambda^{-1}}-\gamma\left\|\Delta\right\|_\Lambda^2\right\},\\
&\quad= \begin{cases} e_i^2\frac{\gamma}{\gamma-\left\|\beta\right\|_{\Lambda^{-1}}^2} & \text{if } \gamma>\left\|\beta\right\|_{\Lambda^{-1}}^2,\\ +\infty & \text{if } \gamma\leq\left\|\beta\right\|_{\Lambda^{-1}}^2. \end{cases}
\end{aligned}$$

While the first equality is due to the change of variable, the second equality is because we are working on a maximization problem, and the last term only depends on the magnitude rather than sign of Δ, thus the optimization problem will always pick Δ that satisfying the equality. Considering the third equality, for the optimization problem, we can first apply the Cauchy-Schwartz inequality in Lemma 6.1 and we know that the maximization problem is to take Δ satisfying the equality constraint. For the last equality, if $\gamma \leq \left\|\beta\right\|_{\Lambda^{-1}}^2$, the optimization problem is unbounded, while $\gamma > \left\|\beta\right\|_{\Lambda^{-1}}^2$, we can solve the quadratic optimization problem and it leads to the final equality.

For the outer minimization problem over γ, as the inner suprema equal infinity if $\gamma \leq \|\beta\|^2_{\Lambda^{-1}}$, the worst-case expected loss becomes,

$$\begin{aligned} &\sup_{P:D_{c_{\mathcal{D}_n}}(P,P_n)\leq\delta} \mathbb{E}_P\left[\left(Y - X^T\beta\right)^2\right] \\ &= \min_{\gamma>\|\beta\|^2_{\alpha\text{-}(p,s)}} \left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^T\beta\right)\frac{\gamma}{\gamma - \|\beta\|^2_{\Lambda^{-1}}}\right\}, \\ &= \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)} + \sqrt{\delta}\,\|\beta\|_{\Lambda^{-1}}\right)^2. \end{aligned} \tag{6.18}$$

The first equality follows the discussion above for restricting $\gamma > \|\beta\|^2_{\Lambda^{-1}}$. We can observe that the objective function in the right hand side of (6.18) is convex and differentiable and as $\gamma \to \infty$ and $\gamma \to \|\beta\|^2_{\Lambda}$, the value function will be infinity. We know the optimizer could be uniquely characterized via first order optimality condition. Solving for γ in this way (through first order optimality), it is straightforward to obtain the last equality in (6.18). If we take square root on both sides, we prove the claim for linear regression.

For the log-exponential loss function, the proof follows a similar strategy. By applying strong duality results for semi-infinity linear programming problem in Proposition 2.1 in Appendix of Chapter 2, we can write the worst case expected loss function as,

$$\begin{aligned} &\sup_{P:D_{c_{\mathcal{D}_n}}(P,P_n)\leq\delta} \mathbb{E}_P\left[\log\left(1+\exp\left(-Y\beta^T X\right)\right)\right] \\ &= \min_{\gamma\geq 0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^{n}\sup_u\left\{\log\left(1+\exp\left(-Y_i\beta^T u\right)\right) - \gamma\,\|X_i - u\|_\Lambda\right\}\right\}. \end{aligned}$$

For each i, we can apply Lemma 1 in Shafieezadeh-Abadeh *et al.* [2015] and dual-norm

result in Lemma 6.1 to deal with the inner optimization problem. It gives us,

$$\begin{aligned}&\sup_u \left\{\log\left(1+\exp\left(-Y_i\beta^T u\right)\right) - \gamma \left\|X_i - u\right\|_\Lambda\right\} \\ =&\begin{cases} \log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right) & \text{if} \quad \|\beta\|_{\Lambda^{-1}} \leq \gamma, \\ \infty & \text{if} \quad \|\beta\|_{\Lambda^{-1}} > \gamma. \end{cases}\end{aligned}$$

Moreover, since the outer optimization is trying to minimize, following the same discussion for the proof for linear regression case, we can plug-in the result above and it leads the first equality below,

$$\begin{aligned}&\min_{\gamma\geq 0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^n \sup_u \left\{\log\left(1+\exp\left(-Y_i\beta^T u\right)\right) - \gamma\left\|X_i-u\right\|_\Lambda\right\}\right\} \\ &= \min_{\gamma\geq\|\beta\|_{\Lambda^{-1}}}\left\{\delta\gamma + \frac{1}{n}\sum_{i=1}^n \log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right)\right\} \\ &= \frac{1}{n}\sum_{i=1}^n \log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right) + \delta\|\beta\|_{\Lambda^{-1}}.\end{aligned}$$

We know that the target function is continuous and monotone increasing in γ, thus we can notice it is optimized by taking $\gamma = \|\beta\|_{\Lambda^{-1}}$, which leads to second equality above. This proves the claim for logistic regression in the statement of the theorem. □

APPENDIX 6.B: Proof of Theorem 6.2

Let us denote, for any set S, the r-neighborhood of S is defined as the set of all points in $\mathbb{R}^d$ that are at distance less than r from S, i.e. $S_r = \cup_{u\in S}\{\bar{u} : \|\bar{u}-u\|_2 \leq r\}$.

Proof of Theorem 6.2. The first part of the inequality is easy to derive. For the second part, we proceed as follows: Under Assumptions A3 and A4, we can define

the compact set

$$\mathcal{C} = \mathcal{C}(X, Y, \beta, \lambda) = \{u : c(u, X) \leq l(X, Y, \beta) - K + \lambda_0/(\lambda - \lambda_0)\}.$$

It is easy to check that $\arg\max\{\psi(u, X, Y, \lambda)\} \subset \mathcal{C}$. Owing to optimality of u^* and from Assumption A2 that $K \geq \phi(X, Y, \beta, \lambda_0)$, we can see that

$$\begin{aligned} l(X, Y) &\leq l(u^*, Y)) - \lambda c(u, X) \\ &= l(u^*, Y) - \lambda_0 c(u^*, X) - (\lambda - \lambda_0) c(u^*, X) \\ &\leq K - \lambda_0 - (\lambda - \lambda_0) c(u^*, X). \end{aligned}$$

Thus by definition of $\mathcal{C} = \mathcal{C}(X, Y, \beta, \lambda)$, it follows easily that $u^* \in \mathcal{C}$, which further implies $\{u : \|u - u^*\|_2 \leq \eta\} \subset \mathcal{C}_\eta$. Then we combine the strongly convexity assumption in Assumption A2 and the definition of $\phi_{\epsilon,f}(u, X, Y, \beta, \lambda)$, which yields

$$\begin{aligned} \phi_{\epsilon,f}(X, Y, \beta, \lambda) &\geq \epsilon \log \left(\int_{\|u-u^*\|_2 \leq \eta} \exp\left(\left[\phi(X, Y, \beta, \lambda) - \frac{\theta}{2} \|u - u^*\|_2^2 \right] / \epsilon \right) f(u) du \right) \\ &= \epsilon \log \left(\exp\left(\phi(X, Y, \beta, \lambda) / \epsilon \right) \right) \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta}{2} \|u - u^*\|_2^2 / \epsilon \right) f(u) du \\ &= \phi(X, Y, \beta, \lambda) + \epsilon \log \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta \|u - u^*\|_2^2}{2\epsilon} \right) f(u) du. \end{aligned}$$

As $\{u : \|u - u^*\|_2 \leq \eta\} \subset \mathcal{C}_\eta$, we can use the lower bound of $f(\cdot)$ to deduce that

$$\begin{aligned} &\int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta \|u - u^*\|_2^2}{2\epsilon} \right) f(u) du \\ &\geq \inf_{u \in \mathcal{C}_\eta} f(u) \times \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta \|u - u^*\|_2^2}{2\epsilon} \right) du \\ &= \inf_{u \in \mathcal{C}_\eta} f(u) \times (2\pi\epsilon/\theta)^{d/2} P(Z_d \leq \eta^2 \theta / \epsilon), \end{aligned}$$

where Z_d is a chi-squared random variable of d degrees of freedom. To conclude, recall that $\epsilon \in (0, \eta^2\theta\chi_\alpha)$, the lower bound of $\phi_{\epsilon,f}(\cdot)$ can be written as

$$\phi_{\epsilon,f}(X, Y, \beta, \lambda) \geq \phi(X, Y, \beta, \lambda) - \frac{d}{2}\epsilon \log(1/\epsilon) + \frac{d}{2}\epsilon \log\left((2\pi\alpha/\theta) \inf_{u \in \mathcal{C}_\eta} f(u)\right).$$

This completes the proof of Theorem 6.2. □

Chapter 7

Discussion and Conclusion

To close this dissertation, we are going to discuss further the potential applications of our data-driven DRO formulation to improve generalization in statistical learning. We will focus on the example of multi-task training in Section 7.1. In Section 7.2, we discuss the different state-of-the-art for robustness in classical statistics and robustness we discussed for our data-driven DRO formulation. In addition, we will propose a conclusion to this dissertation in Section 7.3.

7.1 Distributionally Robust Multi-task training

In this section, in addition to the connections of data-driven DRO formulation we made for square-root Lasso, regularized logistic regression (in Chapter 2); semi-supervised learning (in Chapter 4); groupwise regularization method, square-root group Lasso and group Lasso logistic regression (in Chapter 5); and adaptive regularized regression (in Chapter 6), we shall argue discuss that data-driven DRO is a formulation which improves generalization performance. Moreover, other well-known methods address overfitting also can be approximately interpreted using the data-

driven DRO representation. Next we are going to use multi-task training as an example to illustrate how to formulate DRO problem.

Multi-task training, originally developed in Caruana [1993], is an approach to improve the generalization error by considering pooling multiple training goals and example, into modeling. Intuitively, multi-task training is trying to utilize the information for other related tasks, by sharing part of the model, to put pressures on the parameters towards the direction with better generalization performance. For a brief overview of multi-task training we refer to Section 7.7 of Goodfellow *et al.* [2016] and more systematical details in Chapter 5 of Thrun and Pratt [2012]. We believe, we can utilize the data-driven DRO formulation to implement multi-task training in a meaningful way. Intuitively speaking, let us assume we have data $\mathcal{D}_n = \{X_i, Y_i\}_{i=1}^n$, where predictors $X_i \in \mathbb{R}^d$ and response variable $Y_i = \left(Y_i^{(1)}, Y_i^{(2)}\right)^T \in \mathbb{R}^2$, where we have two tasks of learning: one for $Y_i^{(1)}$ and the other for $Y_i^{(2)}$. A direct data-driven DRO formulation for multi-task training would be consider encode the multi-task information into optimal transport cost function, that is we consider

$$c\left((x,y),(x',y')\right) = c_x\left(x,x'\right) + c_y\left(y,y'\right),$$

where $c_x\left(\cdot\right)$ is the cost function considering variability in x and $c_y\left(\cdot\right)$ is the transport cost in y. Different from the cost function $N_q\left(\cdot\right)$ defined in Equation (2.21) of Chapter 2, where we put infinity transport cost in y, we are considering allowing variability in y to encode the multi-task information. Let us consider the example with $Y_i \in \mathbb{R}^2$. For simplicity, let us consider our main goal is to train $Y_i^{(1)}$ and the task $Y_i^{(2)}$ is trying to help us. Let us consider $c_y\left(\cdot\right)$ to be:

$$c_y\left(y,y'\right) = \kappa|y^{(2)} - y'^{(2)}|I_{y^{(1)}=y'^{(1)}} + \infty I_{y^{(1)}\neq y'^{(1)}},$$

where κ is a non-negative constant encoding the belief in second training task.

From game-theoretic interpretation of DRO problem, we are allowing the adversary player also exploring the variability of second task $Y_{\cdot}^{(2)}$, where the shape of the distributional uncertainty neighborhood $\mathcal{U}_\delta(P_n)$ will be affected by the measure of closeness in the label of second task, $Y_{\cdot}^{(2)}$. If the two learning tasks are related, intuitively we would expect the side information we gain on the distributional uncertainty neighborhood $\mathcal{U}_\delta(P_n)$ will regularize the model towards the direction with better generalization performance. We plan to report this line of work in the future.

7.2 Distributionally Robustness and Robustness in Statistics

In statistics, the terminology “robustness" is mainly reserved for the purpose of considering the outliers or data-contaminations in of observed sample, which has been studied in Huber [1964]; Donoho and Huber [1983]; Huber [1996, 2011]. For example, let us assume we have i.i.d. samples $\mathcal{D}_n = \{W_i\}_{i=1}^n$, where $W_i \in \mathbb{R}$. We assume that the distribution of W_i is symmetric around θ and that $Var\left(W_i\right) < \infty$.

We are interested in estimating the location parameter θ. However, we know that during the data collection or recording procedure, some of the samples may be contaminated. It is not difficult to convince ourselves that the sample mean estimator, $\bar{\theta}_n = n^{-1}\sum_{i=1}^n W_i$, which is the minimizer to the squared loss $\mathbb{E}_{P_n}\left[\left(W-\theta\right)^2\right]$ might perform poorly due to those contaminated samples.

An intuitive approach to address this contamination issue and propose a more robust estimator is to consider the median, $\hat{\theta}_{med}$, of the sample $\mathcal{D}_n$. In turn, this is equivalent to minimizing the empirical absolute loss, $\mathbb{E}_{P_n}\left[|W-\theta|\right]$, instead of the

squared loss.

This example illustrates the spirit of robustness underlying much of the work of Huber Huber [1964, 2011]. In contrast, the sense of robustness in our *data-driven* DRO formulation, is focused on improving out-of-sample performance out of finite-sample information. For example, let us consider the linear regression, where $Y = \beta_*^T X + e$, with β_* being the true regression parameter and e being an independent random error. To address the robustness in Huber's sense, researchers normally consider the absolute loss or Huber's loss Huber [1964] for empirical risk minimization, which is known as the robust regression in the literature Huber [1973]; Rousseeuw and Leroy [2005]. We can easily impose an optimal transport cost uncertainty set and formulate a data-driven DRO version of Huber's empirical loss minimization problem. This formulation might, which considers two layers of robustness, emphasizes that we are studying two different phenomena.

7.3 Conclusion

In this dissertation we study the data-driven DRO with optimal transport cost discrepancy. We show that our data-driven DRO formulation unifies many successful machine learning algorithms which have been studied and which are well known from practice to exhibit good generalization properties.

In addition, we develop a statistical methodology to study data-driven DRO with optimal transport costs. Using the theory, we provide a sharp characterization of the optimal selection of the uncertainty size for DRO problems. Furthermore, we explore multiple ways of choosing the uncertainty region in a data driven way. For example, we studied how to inform the uncertainty region using side information to form novel machine learning algorithms to improve generalization performance.

As we have illustrated in Chapter 4 and Chapter 6, our DRO formulation is considered for a general learning problem, rather than linear and logistic regression settings as we mainly considered in this dissertation for the sake of concreteness. Our discussion on the DRO formulation and its connections to model regularization and multi-task training strongly suggest that there are many applications to be discovered. One such application which we did not explore, but we believe is particularly interesting is that of enhancing generalization error in the setting of training deep-learning algorithms.

Bibliography

Pierre Alquier. LASSO, iterative feature selection and the correlation selector: Oracle inequalities and numerical performances. *Electronic Journal of Statistics*, 2:1129–1152, 2008.

Bertille Antoine, Helene Bonnal, and Eric Renault. On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood. *Journal of Econometrics*, 138(2):461–487, 2007.

Akshay Balsubramani and Yoav Freund. Scalable semi-supervised aggregation of classifiers. In *Advances in Neural Information Processing Systems*, pages 1351–1359, 2015.

Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems 27*, pages 1556–1564. 2014.

Adrian Bernard Druke Becker. *Decomposition methods for large scale stochastic and robust optimization problems*. PhD thesis, Massachusetts Institute of Technology, 2011.

Mathias Beiglböck and Pietro Siorpaes. Pathwise versions of the Burkholder-Davis-Gundy inequality. *Bernoulli*, 21(1):360–373, February 2015.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of operations research*, 23(4):769–805, 1998.

Aharon Ben-Tal and Arkadi Nemirovski. Robust optimization–methodology and applications. *Mathematical Programming*, 92(3):453–480, 2002.

Aharon Ben-Tal, Dimitris Bertsimas, and David B Brown. A soft robust model for optimization under ambiguity. *Operations research*, 58(4-part-2):1220–1234, 2010.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Dimitris Bertsimas, Xuan Vinh Doan, Karthik Natarajan, and Chung-Piaw Teo. Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research*, 35(3):580–602, 2010.

Dimitris Bertsimas, David Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *arXiv preprint arXiv:1401.0212*, 2013.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Christopher M Bishop. Pattern recognition. *Machine Learning*, 128:1–58, 2006.

Jose Blanchet and Peter Glynn. Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. In *Proceedings of the 2015 Winter Simulation Conference*, pages 3656–3667. IEEE Press, 2015.

Jose Blanchet and Yang Kang. Distributionally robust groupwise regularization estimator. *arXiv preprint arXiv:1705.04241*, 2017.

Jose Blanchet and Yang Kang. Distributionally robust semi-supervised learning. *arXiv preprint arXiv:1702.08848*, 2017.

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

Jose Blanchet, Yang Kang, and Henry Lam. Quantify Uncertainty for Stochastic Programming Problem: an Empirical Likelihood Approach, June 2016. Manuscript.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. 2001.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Francesco Bravo. Empirical likelihood based inference with applications to some econometric models. *Econometric Theory*, 20(02):231–264, 2004.

Patrick Breheny and Maintainer Patrick Breheny. Package 'grpreg'. 2016.

Florentina Bunea, Johannes Lederer, and Yiyuan She. The group square-root Lasso: Theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60(2):1313–1325, 2014.

Giuseppe Carlo Calafiore and L El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.

Giuseppe C Calafiore. Ambiguous risk measures and optimal robust portfolios. *SIAM Journal on Optimization*, 18(3):853–877, 2007.

Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 12 2007.

Richard A Caruana. Multitask connectionist learning. In *In Proceedings of the 1993 Connectionist Models Summer School*. Citeseer, 1993.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Song Xi Chen and Peter Hall. Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 21(3):1166–1181, September 1993.

Song Xi Chen and Ingrid Van Keilegom. A review on empirical likelihood methods for regression. *TEST*, 18(3):415–447, November 2009.

Song Xi Chen. On the accuracy of empirical likelihood confidence regions for linear regression model. *Annals of the Institute of Statistical Mathematics*, 45(4):621–637, 1993.

Song Xi Chen. Empirical likelihood confidence intervals for linear regression coefficients. *Journal of Multivariate Analysis*, 49(1):24–40, 1994.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

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

David L Donoho and Peter J Huber. The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157184, 1983.

Jitka Dupačová and Miloš Kopa. Robustness in stochastic programs with risk constraints. *Annals of Operations Research*, 200(1):55–74, 2012.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

John HJ Einmahl and Ian W. McKeague. Confidence tubes for multiple quantile plots via empirical likelihood. *The Annals of Statistics*, 27(4):1348–1367, 1999.

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

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

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

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

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

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group Lasso and a sparse group Lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems 28*, pages 2053–2061. 2015.

Michael Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

Michael Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259, 2015.

Paul Glasserman and Linan Yang. Bounding wrong-way risk in cva calculation. *Mathematical Finance*, 2016.

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in NIPS*, pages 529–536, 2005.

Patrik Guggenberger. Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator. *Econometric Reviews*, 27(4-6), 2008.

Nicholas G Hall, Daniel Zhuoyu Long, Jin Qi, and Melvyn Sim. Managing underperformance risk in project portfolio selection. *Operations Research*, 63(3):660–675, 2015.

Nils Lid Hjort, Ian McKeague, and Ingrid Van Keilegom. Extending the scope of empirical likelihood. *The Annals of Statistics*, pages 1079–1111, 2009.

Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

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

Myles Hollander and Ian W. McKeague. Likelihood ratio-based confidence bands for survival functions. *Journal of the American Statistical Association*, 92(437):215–226, 1997.

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

Kaizhu Huang, Rong Jin, Zenglin Xu, and Cheng-Lin Liu. Robust metric learning by smooth optimization. *arXiv preprint arXiv:1203.3461*, 2012.

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

Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, pages 799–821, 1973.

Peter J Huber. *Robust statistical procedures*. SIAM, 1996.

Peter J Huber. *Robust statistics*. Springer, 2011.

Guido W. Imbens. Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*, 2012.

Hemant Ishwaran and J Sunil Rao. Geometry and properties of generalized ridge regression in high dimensions. *Contemp. Math*, 622:81–93, 2014.

Keiiti Isii. On sharpness of Tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197, 1962.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

Yuichi Kitamura. Empirical likelihood methods in econometrics: Theory and practice. 2006.

Diego Klabjan, David Simchi-Levi, and Miao Song. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management*, 22(3):691–710, 2013.

Henry Lam and Enlu Zhou. Quantifying uncertainty in sample average approximation. In *Proceedings of the 2015 Winter Simulation Conference*, pages 3846–3857. IEEE Press, 2015.

Henry Lam and Enlu Zhou. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 2017.

Lihua Lei and Michael I Jordan. Less than a single pass: Stochastically controlled stochastic gradient method. *arXiv preprint arXiv:1609.03261*, 2016.

Gang Li, Myles Hollander, Ian W. McKeague, and Jie Yang. Nonparametric likelihood ratio confidence bands for quantile functions from incomplete survival data. *The Annals of Statistics*, 24(2):628–640, 1996.

Gang Li, Jing Qin, and Ram C. Tiwari. Semiparametric likelihood ratio-based inferences for truncated data. *Journal of the American Statistical Association*, 92(437):236–245, 1997.

Yuanqing Li, Cuntai Guan, Huiqi Li, and Zhengyang Chin. A self-training semi-supervised svm algorithm and its application in an eeg-based brain computer interface speller system. *Pattern Recognition Letters*, 29(9):1285–1294, 2008.

Xingguo Li, Tuo Zhao, Xiaoming Yuan, and Han Liu. The flare package for high dimensional linear regression and precision matrix estimation in R. *The Journal of Machine Learning Research*, 16(1):553–557, 2015.

Li Li, Chao Sun, Lianlei Lin, Junbao Li, and Shouda Jiang. A mahalanobis metric learning-based polynomial kernel for classification of hyperspectral images. *Neural Computing and Applications*, pages 1–11, 2016.

Moshe. Lichman. UCI machine learning repository, 2013.

Daryl Lim, Brian McFee, and Gert R Lanckriet. Robust structural metric learning. In *ICML-13*, pages 615–623, 2013.

Marco Loog. Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):462–475, 2016.

David G. Luenberger. *Introduction to linear and nonlinear programming*, volume 28. Addison-Wesley Reading, MA, 1973.

David G Luenberger. *Introduction to linear and nonlinear programming*, volume 28. Addison-Wesley Reading, MA, 1973.

Don McLeish. A general method for debiasing a Monte Carlo estimator. *Monte Carlo Meth. and Appl.*, 17(4):301–315, 2011.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70(1):53–71, 2008.

S. A. Murphy. Likelihood ratio-based confidence intervals in survival analysis. *Journal of the American Statistical Association*, 90(432):1399–1405, 1995.

Sahand Negahban, Pradeep Ravikumar, Martin Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-Estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Whitney Newey and Richard Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

Art B Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

Art Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, pages 90–120, 1990.

Art Owen. Empirical likelihood for linear models. *The Annals of Statistics*, pages 1725–1747, 1991.

Art B Owen. *Empirical likelihood*. CRC press, 2001.

Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized LASSO: A precise analysis. In *In proceedings of the 51st Annual Allerton Conference of Communication, Control, and Computing*, pages 1002–1009, October 2013.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *ICML*, 2016.

Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, pages 300–325, 1994.

Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems. Volume I: Theory*. Springer Science & Business Media, 1998.

Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems. Volume II: Applications*. Springer Science & Business Media, 1998.

Sundhar Ram, Angelia Nedić, and Venugopal Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010.

Chang-han Rhee and Peter Glynn. Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043, 2015.

Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 630–638, 2016.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

Yossi Rubner and Carlo Tomasi. The earth mover's distance. In *Perceptual Metrics for Image Database Navigation*, pages 13–28. Springer, 2001.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66. IEEE, 1998.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

Peter Ruckdeschel. Optimally (distributional-) robust kalman filtering. *arXiv preprint arXiv:1004.3393*, 2010.

Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-Hill New York, 1964.

Andrzej P Ruszczynski and Alexander Shapiro. *Stochastic programming*, volume 10. Elsevier Amsterdam, 2003.

Herbert Scarf, KJ Arrow, and S Karlin. A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production*, 10(2):201, 1958.

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *Advances in neural information processing systems*, pages 41–48, 2004.

Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems 28*, pages 3312–3320. 2015.

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

Alexander Shapiro and Darinka Dentcheva. *Lectures on stochastic programming: modeling and theory*, volume 16. SIAM, 2014.

Alexander Shapiro and Darinka Dentcheva. *Lectures on stochastic programming: modeling and theory*, volume 16. Siam, 2014.

Alexander Shapiro and Anton Kleywegt. Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17(3):523–542, 2002.

James Smith. Generalized Chebychev inequalities: Theory and applications in decision analysis. *Operations Research*, 43(5):807–825, 1995.

Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, July 2015.

Sanvesh Srivastava, Volkan Cevher, Quoc Tran-Dinh, and David B Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *AISTATS*, 2015.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *arXiv preprint arXiv:1104.4595*, 2011.

Frode Terkelsen. Some minimax theorems. *Mathematica Scandinavica*, 31(2):405–413, 1973.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Qihua Wang and J. N. K. Rao. Empirical likelihood for linear regression models under imputation for missing responses. *Canadian Journal of Statistics*, 29(4):597–608, 2001.

Zizhuo Wang, Peter W. Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven newsvendor problems. *Preprint*, 2009.

Dan Wang, Canxiang Yan, Shiguang Shan, and Xilin Chen. Unsupervised person re-identification with locality-constrained earth mover's distance. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 4289–4293. IEEE, 2016.

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

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

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

Samuel S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Changbao Wu. Weighted empirical likelihood inference. *Statistics & Probability Letters*, 66(1):67–79, January 2004.

Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, volume 15, page 12, 2002.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. In *Advances in Neural Information Processing Systems*, pages 1801–1808, 2009.

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

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

Zheng-Jun Zha, Tao Mei, Meng Wang, Zengfu Wang, and Xian-Sheng Hua. Robust distance metric learning with auxiliary knowledge. In *IJCAI*, pages 1327–1332, 2009.

Yichuan Zhao and Hongkun Wang. Empirical likelihood inference for the regression model of mean quality-adjusted lifetime with censored data. *Canadian Journal of Statistics*, 36(3):463–478, 2008.

Mai Zhou. *Empirical likelihood method in survival analysis*, volume 79. CRC Press, 2015.

Shushang Zhu and Masao Fukushima. Worst-case conditional value-at-risk with application to robust portfolio management. *Operations research*, 57(5):1155–1168, 2009.

Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-supervised learning with graphs*. Carnegie Mellon University, language technologies institute, school of computer science, 2005.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Steve Zymler. *Distributionally robust optimization with applications to risk management*. Imperial College London, 2010.