Sample Out-of-Sample Inference Based on Wasserstein Distance

Jose Blanchet
Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, jose.blanchet@stanford.edu

Yang Kang
Department of Statistics, Columbia University, New York, NY 10027 yang.kang@columbia.edu

We present a novel inference approach that we call Sample Out-of-Sample (or SOS) inference. The approach can be used widely, ranging from semi-supervised learning to stress testing, and it is fundamental in the application of data-driven Distributionally Robust Optimization (DRO). Our method enables measuring the impact of plausible out-of-sample scenarios in 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 remain controlled as the dimension increases.

Key words: Non-parametric Statistics, Probability, Distributionally Robust Optimization, Optimal Transport

History: This paper was first submitted on May 25, 2016 and has been with the authors for 3 years for 3 revisions.

1. Introduction

The goal of this paper is to introduce a novel methodology for non-parametric inference allows incorporating the adverse impact of out-of-sample scenarios. We call the procedure Sample Out-of-Sample (SOS)) inference. Our method is general, and we discuss several applications, including Distributionally Robust Optimization (DRO), semi-supervised learning, and a novel stress-testing framework. We use the DRO framework in the introduction to put our contributions in perspective. We elaborate on semi-supervised learning and stress-testing applications in Section 2.

A data-driven DRO problem takes the form

$$\min_{\theta\in R^d}\max_{P\in\mathcal{U}_\delta(P_n)} E_P[\mathcal{L}(\theta,X)], \tag{1}$$

where $\mathcal{L}:\mathbb{R}^{d\times l}\to[0,\infty)$ is a cost (or loss) function, $X\in\mathbb{R}^l$ is a random element, and $\theta\in\mathbb{R}^d$ is a decision. Often, $\mathcal{L}(\cdot,x)$ is assumed to be strictly convex and smooth (e.g. twice differentiable) and we will assume this throughout our motivating discussion. The notation $E_P(\cdot)$ denotes the expectation operator associated to the probability measure P. We use P_n to denote the empirical measure corresponding to $\{X_i\}_{i=1}^n$ independent identical distributed (i.i.d.) observations that follow the distribution P_*. The set $\mathcal{U}_\delta(P_n)$ is the distributional uncertainty set. The parameter $\delta>0$ is called the "size of the distributional uncertainty" so that the family of sets $(\mathcal{U}_\delta(P_n):\delta\geq 0)$ is increasing (in the sense of inclusion) as a $\delta>0$ increases and so that for $\delta=0$, $\mathcal{U}_0(P_n)=\{P_n\}$. Therefore, intuitively, P_n is the "center" of the distributional uncertainty region and $\delta>0$ can be thought of as its "radius."

Ideally, one would like to compute $\theta_*=\arg\min E_{P_*}[\mathcal{L}(\theta,X)]$, but P_* is unknown. Therefore, the intuition behind formulation (1) is that one is interested in choosing a decision θ, which performs well uniformly over a range of models that constitute reasonable (or plausible) variations of the data (encoded by P_n).

We are interested in variations of the empirical distribution P_n (the elements in $\mathcal{U}_\delta(P_n)$) that systematically explore the impact of *out-of-sample scenarios* in the loss function $\mathcal{L}(\cdot)$. Therefore,

$P \in \mathcal{U}_\delta(P_n)$ should not be supported only on the underlying data set. Instead, we are interested in a framework that admits models in $P \in \mathcal{U}_\delta(P_n)$ that may be supported outside the sample $\{X_i\}_{i=1}^n$. Because of this out-of-sample exploration feature, we choose $\mathcal{U}_\delta(P_n)$ based on the Wasserstein distance of order 2, which is explained in Section 3. We shall also discuss different alternative norms that are supported by our analysis and discuss how these can be calibrated in a data-driven way.

Distributionally robust optimization formulations such as (1) based on the Wasserstein distances have been studied recently in a wide range of settings, especially in applications to machine learning and artificial intelligence, see for example, Shafieezadeh-Abadeh et al. (2015), Mohajerin Esfahani and Kuhn (2018), Zhao and Guan (2018), Blanchet and Murthy (2019), Gao and Kleywegt (2016), Blanchet et al. (2019b), Yang (2017), Sinha et al. (2018), Gao et al. (2018), Volpi et al. (2018), Chen et al. (2018), Blanchet et al. (2019e,c).

All of these studies focus on the setting in which the support of the distributions inside $\mathcal{U}_\delta(P_n)$ is $\mathbb{R}^d$. Moreover, within the current literature, only Blanchet et al. (2019b) studies the optimal selection of the parameter δ by defining a natural optimization criterion. The work of Blanchet et al. (2019b) also shows that such criterion recovers choices that have been argued to be effective for recovery in machine learning settings for which a DRO representation can be posed.

In contrast, compared to Blanchet et al. (2019b), *our work is the first one that studies the statistical implications of choosing the support of the members of the distributional uncertainty* $P \in U_\delta(P_n)$ *in a data-driven way.* One of our main contributions of this paper consists in providing a comprehensive study of an optimal data-driven choice of uncertainty size, δ, when the support of the members in $\mathcal{U}_\delta(P_n)$ is obtained from an arbitrary random sample whose size is increasing with n.

More generally, our contributions can be viewed in the lens of a novel inference framework that we call SOS inference, based on the analysis of the so-called SOS profile function for estimating equations.

In the DRO framework, we consider enriching the empirical data set $\mathcal{X}_n = \{X_i\}_{i=1}^n$ (which is assumed to be i.i.d.) by including a set of scenarios $\{Y_i\}_{i=1}^m$ (which is also assumed to be i.i.d.), with

$m = [\kappa n]$ for some $\kappa \in [0,\infty)$. The Y_i's and the X_i's are not assumed share the same distribution. In order to unify the notation we write $Z_i = X_i$ for $i = 1,...,n$, $Z_{n+k} = Y_k$ for $k = 1,...,m$ and set $\mathcal{Z}_{n+m} = \{Z_j\}_{j=1}^{n+m}$. (We use $\mathbf{P}$ to denote the probability measure supporting the infinite sequences $\{X_i\}_{i\geq 1}$ and $\{Y_i\}_{i\geq 1}$, where the support of $\mathbf{P}$ is dense in the support of the underlying sampling distribution.)

In order to emphasize the difference between the analysis in Blanchet et al. (2019b) and our analysis here, we write $\mathcal{U}_\delta(P_n;\mathbb{R}^{l+1})$ to denote the full support case (studied in Blanchet et al. (2019b)) and $\mathcal{U}_\delta(P_n;\mathcal{Z}_{n+m})$ for the uncertainty set considered in our current setting.

Let us describe the optimality criterion introduced in Blanchet et al. (2019b) for choosing δ. Here we restrict the support on the observed sequence and we would expect larger δ due to the extra constraint. Since the set $\mathcal{U}_\delta(P_n)$ is interpreted as the set of plausible variations of the data, then the set

$$\Lambda_\delta(P_n) = \{\theta : \theta = \arg\min E_P[\mathcal{L}(\theta, X)] \text{ for } P \in \mathcal{U}_\delta(P_n;\mathcal{Z}_{n+m})\} \tag{2}$$

corresponds to the set of plausible decisions, those that are compatible with the distributional uncertainty region. Note that $\Lambda_\delta(P_n)$ is a random set that can be interpreted as a confidence region. The criterion that we utilize is the following

$$\min\{\delta : \mathbf{P}(\theta_* \in \Lambda_\delta(P_n)) \geq \alpha\}, \tag{3}$$

where α is a desired confidence level.

To analyze (3), we first argue that

$$\{\theta_* \in \Lambda_\delta(P_n)\} = \{R_n^W(\theta_*) \leq \delta\}, \tag{4}$$

for a suitable function, $R_n^W(\cdot)$, which we call the Sample-out-of-Sample (SOS) profile function. In simple words, $R_n^W(\theta_*)$ can be computed directly in terms of the shortest Wasserstein distance between P_n and the set of probability models $P \in \mathcal{U}_\delta(P_n;\mathcal{Z}_{n+m})$ for which $E_P[\nabla_\theta \mathcal{L}(\theta_*, x)] = 0$.

As a consequence of (4), the optimal δ solving (3) is simply the α-quantile of $R_n^W(\theta_*)$.

In general, we can use our methodology to test the hypothesis that θ_* satisfies $E_{P_*}(h(\theta_*, x)) = 0$, simply replacing $\nabla_\theta \mathcal{L}(\theta, x)$ by $h(\theta, x)$ in the definition of the SOS profile function. The hypothesis is rejected for high values of the statistics $R_n^W(\theta_*)$. Thus, it is important to compute the asymptotic distribution $R_n^W(\theta_*)$.

Our contributions are then stated at this level of generality (i.e., asymptotic analysis of $R_n^W(\theta_*)$ for the purpose of hypothesis testing). In the end, this paper involves two main methodological contributions:

A) First, we characterize the asymptotic distribution of $R_n^W(\theta_*)$ as $n \to \infty$; see Theorem 1, Theorem 2, and Theorem 3. We explain how to compute the asymptotic limiting distributions in Section 4.1.2.

B) Second, we discuss various extensions that we believe are natural to study in order to define DRO optimal transport cost functions. These include implicit DRO formulations and plug-in estimators. We illustrate the extensions in the empirical result section (Section 5). For example, writing $\theta_* = (\gamma_*, v_*)$ we develop the asymptotic distribution of $R_n^W(\gamma_*, \bar{v}_n)$, where $\bar{v}_n$ is a suitable consistent plug-in estimator for v_* as $n \to \infty$; see Corollary 2. The construction of $\bar{v}_n$ may be based on standard empirical estimators. This extension may be used in the context of stochastic optimization with constraints, as illustrated in Section 5.

The theory that we develop in this paper parallels the main fundamental results obtained in the context of Empirical Likelihood (EL), introduced by Art Owen in Owen (1988, 1990, 2001). In fact, the construction of the function $R_n^W(\cdot)$ borrows a great deal of inspiration from the empirical likelihood profile function and its extensions based on divergence criteria, rather than the likelihood function (see Owen (2001)), and also see Bayraksan and Love (2015) for a comprehensive review of divergence-based distributional uncertainty sets in optimization, many of which are amenable to EL-based analysis. There are, however, several important features of our framework that, we believe, add significant value to the non-parametric inference literature.

Before we discuss these features, we want to emphasize that our motivation is not to disprove the appropriateness of divergence approaches. The DRO community is actively investigating the advantages of various choices of uncertainty sets. Our discussion should be seen as a step in this direction.

The most likely picture to eventually emerge is that divergence and Wasserstein approaches complement each other depending on issues such as convenience and tractability. For the purpose of using out-of-sample scenarios to inform the uncertainty set, we believe the Wasserstein distance is a natural choice, as we shall explain.

First, using divergence-based criteria (as it is typically done in standard EL settings) carries implicit support assumptions that seem unnatural in our setting as the sample size increases. For example, it is not difficult to see that a divergence-based distance between the empirical measure based on n i.i.d. samples and that of $m = [\kappa n]$ i.i.d. samples (both from the same distribution) may not converge to zero. In our setting, this suggests that under divergence-type constructions, it requires a large uncertainty set to include distributions that one may reasonably and intuitively see as relatively small perturbation of the data. So, choosing a large-sized uncertainty to accommodate these small perturbations may inflate the estimates artificially, just because the populations are large but unbalanced. Alternatively, if the size of uncertainty is small (which is expected under the null hypothesis as the sample size increases), the proportion of mass allocated outside the support of the empirical measure decreases to zero, so the overwhelming proportion of the mass in the models contained in the uncertainty set is concentrated in the support of the baseline model. Hence, we believe that the direct use of the EL framework may not be suitable in our setting. Additional out-of-sample issues that 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 (2018)), and see also 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 $R_n^W(\theta_*)$ 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 $\mathcal{L}(\cdot)$),

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

as $n \to \infty$. Note that the scaling depends on the dimension of the random vector X 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. When $l = 1$, we obtain a similar limiting distribution as in the EL case. The intuition here is that a sample of order $O(n)$ provides enough coverage of the space since the optimal transport plan will displace points at distance $O(1/n^{1/2})$. 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, some of the qualitative distinctions in the methodology 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). The high level intuition of the scaling is associated with the interplay between the linear programming formulation and the coverage of a sample of size n in a space in l dimension. A high-level intuition is given in more detail in Section 7.1. In contrast to the analysis of the SOS function, in the EL analysis of the profile function, the optimal solutions are amenable to a smooth perturbation analysis as $n \to \infty$ using a Taylor expansion of second (and 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 as in Wilks (1938)), which lies at the core of EL analysis.

The high-level intuition developed in Section 7.1 also underscores the distinction between our development here and the analysis in Blanchet et al. (2019b). In contrast to our development here, the scaling in Blanchet et al. (2019b) is always dimension independent. This is because the issue involving the coverage of the random scenarios in the support of the alternative distributions is not a feature that needs to be considered. Moreover, the current setting introduces a correlation structure in the optimal transportation map, which is not present in the analysis of Blanchet et al. (2019b). This is because the feasible transportation locations are now given by a random sample. To this end, we take advantage of recent sample-path martingale inequalities. The use of these inequalities is showcased in the technical Section 7.2.7 and we believe that these techniques may be applicable more broadly in non-parametric statistical analysis.

The rest of the paper is organized as follows. In Section 2 we discuss semi-supervised learning and stress-testing applications that motivate the formulation in which the support of $P \in \mathcal{U}_\delta(P_n; \mathcal{Z}_{n+m})$ is data-driven. Basic definitions, including a review of the Wasserstein distance, are given in Section 3. Our main technical results are described in Section 4. We include applications of our results to settings such as stochastic optimization, risk analysis, and semi-supervised learning in Section 5. A short section including conclusions and additional discussions is given in Section 6. Finally, our technical development is given in Section 7, starting with a high-level intuition of the nature of our results and scaling in Section 7.1.

2. Motivating Settings

2.1. Semi-supervised Learning Applications

The setting of semi-supervised learning can be used to illustrate our framework. Consider a classification problem that takes the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ and $Y_i \in \{-1, 1\}$ is the i-th response variable and $X_i \in R^l$ is the i-th predictor. For concreteness, let us consider the logistic regression setting in which

$$P(Y_i = 1|X_i) = \frac{\exp(Y_i \beta_*^T X_i)}{1 + \exp(Y_i \beta_*^T X_i)} = 1 - P(Y_i = -1|X_i).$$

Suppose that we have access to an unlabeled data set $\{X_i'\}_{i=1}^m$ and we are interested in using this data in a meaningful way for estimating β_*. This is the semi-supervised learning setting arising in cases in which obtaining responses or labels for every individual may be costly.

If the predictive variables are contained inside a lower-dimensional manifold embedded in the underlying ambient space, our intuition is that unlabeled data can be used as a proxy to profile

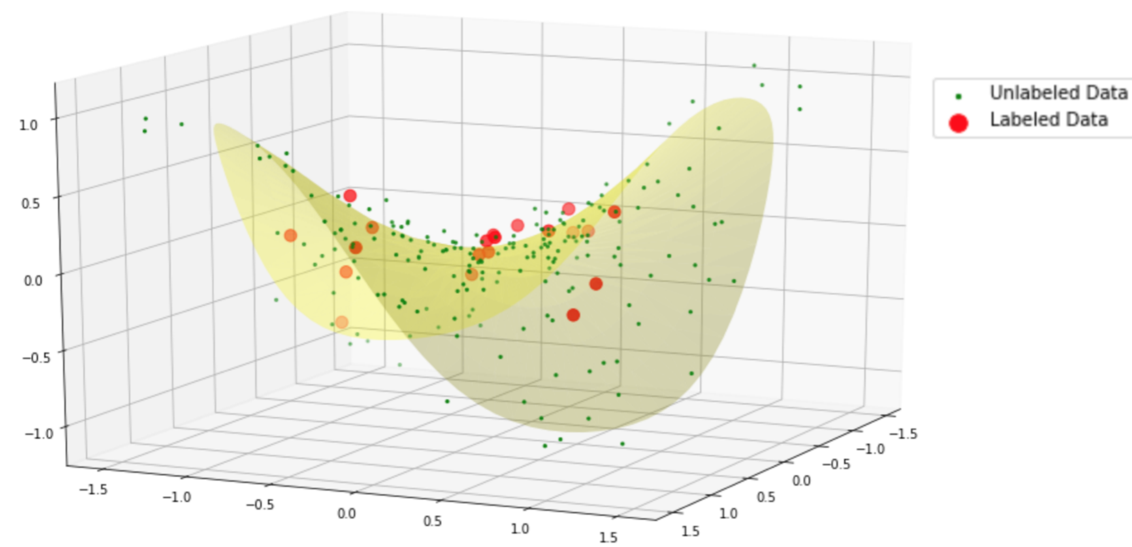

Figure 1 An illustrative example showing that the unlabeled observations (with green dots) can be used to provide a proxy for the underlying manifold (the yellow surface) in which the the predictive variables lie; whereas the labeled data points (red dots) are not sufficient to provide such information.

precisely such a lower dimensional manifold. Thus it is natural to impose a DRO formulation that enhances statistical performance by quantifying the impact of out-of-sample scenarios that lay in the relevant lower-dimensional manifold. This intuition is illustrated in Figure 1. The work of Blanchet and Kang (2018) proposes combining both the labeled and unlabeled data by forming the set $\mathcal{X}_{n,m} = \mathcal{D}_n \cup (\{X_i'\}_{i=1}^m \times \{-1,1\}_{i=1}^n)$ (i.e., the original data set is enriched by considering the unlabeled data with all the possible responses recorded by the labeled data).

Then, Blanchet and Kang (2018) considers a DRO formulation for estimating β_* in which the distributional uncertainty region is defined in terms of the Wasserstein distance. The DRO formulation proposed in Blanchet and Kang (2018) is equivalent to the problem

$$\min_{\beta} \max_{\mathcal{U}_\delta\left(P_n;\mathcal{X}_{n,m}\right)} E_P\left[\mathcal{L}\left(X,Y,\beta\right)\right], \tag{5}$$

which corresponds to (1).

The formulation of Blanchet and Kang (2018) (i.e. (5)) is of significant interest because it is a natural semi-supervised learning extension version of regularized linear regression, which is a highly popular supervised machine learning estimator (see Hastie et al. (2005)). In particular, it is shown in

Blanchet et al. (2019b), see also Shafieezadeh-Abadeh et al. (2015), that replacing $\mathcal{U}_\delta\left(P_n;\mathcal{X}_{n,m}\right)$ by $\mathcal{U}_\delta\left(P_n;\mathbb{R}^l\right)$ in (5) one recovers exactly regularized logistic regression and δ corresponds exactly to the regularization parameter. This connection between Wasserstein DRO and mainstream supervised learning estimators has been established for a large class of methods, including square-root Lasso (Blanchet et al. (2019b)), support vector machines (Blanchet et al. (2019b)), group Lasso (Blanchet and Kang (2017)), adaptive Lasso (Blanchet et al. (2019e)), etc.

The methods developed in this paper provide the theoretical underpinning for the choice of the uncertainty size δ in the context of (5), which yields regularized estimators that are informed by the unlabeled data in a meaningful way.

2.2. Novel Stress-testing Framework

Consider the following stress-testing exercise. An insurance company wishes to estimate a certain expectation of interest, say $\mathbb{E}_{\mathbb{P}^*}(L(X))$, where X might represent one or several risk factors, $L(X)$ is the corresponding financial loss and $\mathbb{P}^*(\cdot)$ is the underlying probability measure which may be unknown.

The insurance company may estimate $\mathbb{E}^*(L(X))$ based on n i.i.d. 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 $Y_1,...,Y_m \in \mathbb{R}^l$, where $m = [\kappa n]$ with $\kappa \in [0,\infty)$. It may be natural to choose $\kappa = 1$ so that the amount of information provided by the regulator and the company is balanced, but this is not necessary.

The scenarios provided by the regulator may or may not come from the same distribution as the X_i's. In fact, typically they will come from a different distribution. The regulator's beliefs are captured by the distribution of the Y_i's. These beliefs may, in turn, be informed by the knowledge that is accessible only by the regulator and not by the insurance company. The regulator may not necessarily question the fact that the historical data from the X_i's follows distribution $\mathbb{P}_*(\cdot)$, but the regulator might be concerned that the insurance company lacks additional information to assess the overall risk exposure better.

On the one hand, the insurance company clearly knows well its idiosyncratic risk exposures, so the data represented by the X_n's, arising from a model with such idiosyncratic information is meaningful and should be considered carefully. On the other hand, it is also correct that the regulator possesses additional information that should be considered in evaluating the potential impact of scenarios that may not be appropriately captured by the data of the insurance company.

How does one incorporate both the X_i's and the Y_i's in a meaningful way for the purposes of evaluating the risk of the company?

The methodology developed in this paper allows incorporating both the empirical data of the insurance company and the stress scenarios provided by the regulator into a Distributionally Robust Performance Analysis (DRPA) formulation (closely related to Distributionally Robust Optimization – DRO) as we describe next.

Define $Z_k = X_k$ for $k = 1, \ldots, n$ and $Z_{n+k} = Y_k$ for $k = 1, \ldots, m$ (i.e., merge both the empirical samples and the stress scenarios into a set $\mathcal{Z}_{n+m} = \{Z_1, ..., Z_{n+m}\}$). We let

$$P_n(dx) = n^{-1} \sum_{k=1}^{n} \delta_{\{X_k\}}(dx)$$

be the empirical distribution of the data generated by the insurance company. A natural estimate for $\mathbb{E}^*(L(X))$ based on the insurance company's data is given by

$$\mathbb{E}_{P_n}(L(X)) = n^{-1} \sum_{k=1}^{n} L(X_k).$$

Now, let $\mathcal{P}(\mathcal{Z}_{n+m})$ be the set of all probability distributions with support on $\mathcal{Z}_{n+m}$. Our DRPA approach consists in providing estimates for $\mathbb{E}_P(L(X))$ via

$$\theta_-(\delta), \theta_+(\delta) = \min_{P \in \mathcal{U}_\delta(P_n; \mathcal{Z}_{n,m})}, \max \mathbb{E}_P(L(X)). \tag{6}$$

We believe that the DRPA formulation (6) provides a reasonable approach for combining both the insurance company's information and the regulator's beliefs. We do not disregard the data coming from the insurance company (in fact, the empirical distribution P_n is placed at the center of the uncertainty set), but we also capture the potential impact of out-of-sample scenarios based on the regulator's beliefs.

Formulation (6) is closely related to (1) and the methodology that we present in this paper can be used to find an optimal choice for δ. In particular, an equivalent way of representing the range $[\theta_-(\delta), \theta_+(\delta)]$ is in terms of a suitably defined SOS profile function (or "SOS function"), $R_n^W(\cdot)$, as we shall see, so that

$$[\min\{\theta : R_n^W(\theta) \leq \delta\}, \max\{\theta : R_n^W(\theta) \leq \delta\}] = [\theta_-(\delta), \theta_+(\delta)]. \tag{7}$$

Therefore, the study of the function $R_n^W(\cdot)$ is a key in the analysis of (6) and the selection of δ based on statistical principles, and this leads us to our contributions A)-B) described in the Introduction.

We emphasize, however, that our choice of δ is purely statistical. That is, we operate under the blanket assumption that the risk is correctly computed solely with the bank's internal data as the sample size grows to infinity. Under this assumption there is less and less need for scenarios as the sample size of the internal data increases. In practice, the sample size is always finite and, in the end, the choice of regulatory capital is the result of an informed negotiation between the regulator and the bank. We provide a tool that helps to inform this discussion because it statistically combines both elements (internal data and external scenarios) in a way that is consistent with the guidelines described in of Governors Federal Reserve System (2019) for generating stress scenarios. However, non-statistical criteria (e.g., social cost based) may also be used to choose δ, leading to, for instance, hybrid methods that would build on our current development. However, these types of hybrid choices would require additional modeling elements that are beyond the scope of our statistical treatment.

3. Basic Definitions

Throughout our development we adopt the convention that all vectors we consider are expressed as columns, so, for example, $x^T = (x_1, ..., x_l)$ is a row vector in $\mathbb{R}^l$ (here we use x^T to denote the transpose of x). Also, given a random variable $W \in \mathbb{R}^d$ so that $\mathbb{E}(W) = 0$ and $\mathbb{E}\left(\|W\|_2^2\right) < \infty$, we use $Var(W) = \mathbb{E}(WW^T)$ to denote the covariance matrix of W.

3.1. On Wasserstein Distance and Distributional Uncertainty

As we mentioned in the introduction, we utilize the Wasserstein distance of order 2 to describe the distributional uncertainty region. We consider two closed subsets of $\mathbb{R}^l$, namely $\mathcal{X}$ and $\mathcal{Z}$. We use the notation $\mathcal{P}(\mathcal{X}\times\mathcal{Z})$ to denote all the Borel probability measures π with support on $\mathcal{X}\times\mathcal{Z}$. Any $\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{Z})$ can be thought of as the joint distribution of a pair of random vectors (X,Z). We use the notation π_X to denote the marginal distribution of X under π; similarly, π_Z is the marginal distribution of Z under π.

The Wasserstein distance (of order 2) between the Borel probability measures μ and υ, supported on $\mathcal{X}$ and $\mathcal{Z}$, respectively, is defined as $\sqrt{D(\mu,\upsilon)}$, where

$$D(\mu,\upsilon)=\inf\{\int\int\|x-z\|_2^2\,\pi(dx,dz):\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{Z}),\pi_X=\mu,\pi_Z=\upsilon\}.$$

In simple words, the square of the Wasserstein distance of order 2 (under the Euclidean metric) is defined as the minimum cost of transporting the mass encoded by μ into the mass encoded by υ; computing the unitary-cost-per-transportation of a unit of mass from x to y as the square of the Euclidean distance between the source (x) and destination (y).

Our results can be directly adapted to the situation in which the Euclidean metric is replaced by the so-called Mahalanobis distance, namely, $\|x-y\|_A^2=(x-y)^T A(x-y)$ for any positive definite matrix A. The use of this distance and procedures to fit A for classification tasks based on manifold learning tools are studied in Blanchet et al. (2019e). In order to simplify the notation and the exposition we continue with the standard Euclidean metric throughout our development, corresponding to $A=I$.

In the sequel, $\mathcal{X}$ and $\mathcal{Z}$ are finite cardinality sets. Therefore, in this case, the evaluation of $D(\mu,\upsilon)$ is a finite dimensional linear programming problem and so, conceptually, computing $D(\mu,\upsilon)$ is straightforward. The Wasserstein distance is defined in great generality (for arbitrary metric spaces) as the solution of the Monge-Kantorovich problem with the cost-per-transportation defined in terms of the underlying metric. We refer the reader to Villani (2008) for more information on

Wasserstein distances. Because we focus on the finite-cardinality case, it is enough with elementary notions of finite dimensional linear programming to understand the definition we use in this paper.

The distributional uncertainty set, $\mathcal{U}_\delta(P_n)$, mentioned in the Introduction to motivate our contributions can then be defined by choosing $\mathcal{X} = \mathcal{X}_n$ and $\mathcal{Z} = \mathcal{Z}_{n+m}$ and letting

$$\mathcal{U}_\delta(P_n) = \mathcal{U}_\delta(P_n; \mathcal{Z}_{n+m}) = \{P : D(P_n, P) \le \delta\}.$$

3.2. The SOS Profile Function

To motivate the definition of the SOS Profile function, once again, we return to the DRO framework defined in the Introduction. We note from (2) that

$$\Lambda_\delta(P_n) = \{\theta : E_P[h(X,\theta)] = 0 \text{ for } P \in \mathcal{U}_\delta(P_n)\},$$

where $h(X,\theta) = \nabla_\theta \mathcal{L}(X,\theta)$. So (by convexity) we have that $\theta_* \in \Lambda_\delta(P_n)$ if and only if there exists $P \in \mathcal{U}_\delta(P_n)$ such that

$$E_P[h(X,\theta_*)] = 0. \tag{8}$$

Let $R_n^W(\theta_*)$ be the smallest transportation cost (measured by $D(P_n,P)$) between P_n and any member $P \in \mathcal{P}(\mathcal{Z}_{n+m})$ for which (8) is true. It is easy to reason that $R_n^W(\theta_*) \le \delta$ if and only if $\theta_* \in \Lambda_\delta(P_n)$. Formally, we have the following definition for the SOS profile function $R_n^W(\theta)$, namely

$$R_n^W(\theta) = \min\{D(P_n,P) : E_P[h(X,\theta)] = 0\}. \tag{9}$$

The goal of this paper is to study the behavior of $R_n^W(\theta_*)$ under the estimating equation assumption

$$E_{P_*}[h(X,\theta_*)] = 0, \tag{10}$$

and the $\{X_i\}_{i=1}^n$ being an i.i.d. sample from P_*. We will formulate our results in terms of the estimating equation (10) for general $h(\cdot)$ (not necessarily arising from an optimization problem).

We consider this more general framework because we believe that our results may be applicable to inference settings other than DRO, for instance, the stress-testing framework described earlier. In fact, we now return to such setting to explain how to use the SOS profile function in this case.

3.2.1. The SOS Profile function for stress-testing setting In the stress-testing setting described earlier, we wish to select δ just as large to guarantee that $\theta_* := \mathbb{E}_{\mathbb{P}^*}(L(X)) \in [\theta_-(\delta), \theta_+(\delta)]$ with a certain degree of confidence, which we shall denote by α.

Therefore, because of equation (7), we are interested in choosing the smallest δ so that

$$P\{\theta_* \in [\theta_-(\delta), \theta_+(\delta)]\} = \mathbf{P}\{R_n^W(\theta_*) \leq \delta\} = \alpha. \tag{11}$$

In other words, δ is chosen to be the α-quantile of the random variable

$$R_n^W(\theta_*) = \min\{D(P_n, P) : E_P[L(X) - \theta_*] = 0\}.$$

Note that this formulation is a particular case of the one introduced in (10) by letting $h(\theta, x) = L(x) - \theta$. For pedagogical reasons, we will present our results first for the SOS profile function for means (i.e., assuming that $L(x) = x$) and later we move to more general estimating equations.

4. Main Results

4.1. SOS Function for Means

We state the following underlying assumptions 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. Therefore, the cardinality of the set $\mathcal{X}_n$ is n.

A2): We also consider an independent i.i.d. sample $\mathcal{Y}_m = \{Y_1, ..., Y_m\} \subset \mathbb{R}^l$ from a continuous distribution. Throughout our discussion we shall assume that $m = [\kappa n]$ with $\kappa \in [0, \infty)$.

A3): Assume that $\mathbb{E}\|X_1\|_2^2 + \mathbb{E}\|Y_1\|_2^2 < \infty$.

A4): If $l = 1$ we assume that X_i and Y_i have positive densities $f_X(\cdot)$ and $f_Y(\cdot)$. If $l \geq 2$ we assume that X_i and Y_i have differentiable positive densities $f_X(\cdot)$ and $f_Y(\cdot)$, 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 $Z_{n+j} = Y_j$ 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 $\mathcal{C}$. Therefore, 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 $\mathcal{X}_n$, that is,

$$\mu_n(dx) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(dx).$$

Given any $\pi \in \mathcal{P}(\mathcal{X}_n \times \mathcal{Z}_{n+m})$ 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, $R_n^W(\cdot)$, to estimate $\theta_* = E(X)$ is defined as

$$\begin{aligned} R_n^W(\theta_*) = \inf\{ & \int\int \|x-z\|_2^2 \pi(dx, dz) : \\ & \text{s.t. } \pi \in \mathcal{P}(\mathcal{X}_n \times \mathcal{Z}_{n+m}), \pi_X = \mu_n, \pi_Z = v_n, \int z v_n(dz) = \theta_* \}, \\ = \inf\{ & \int\int \|x-z\|_2^2 \pi(dx, dz) : \\ & \text{s.t. } \pi \in \mathcal{P}(\mathcal{X}_n \times \mathcal{Z}_{n+m}), \pi_X = \mu_n, \int z \pi_Z(dz) = \theta_* \}. \end{aligned} \tag{12}$$

(Here and throughout the paper, s.t. is an abbreviation for "subject to.")

We now state the following asymptotic distributional result for the SOS function.

Theorem 1 **(SOS Profile Function Analysis for Means)**. *In addition to Assumptions A1)-A4), suppose that the covariance matrix of X, $Var(X)$, exists. The following asymptotic result follows*

- *When $l = 1$,*

$$nR_n^W(\theta_*) \Rightarrow \sigma^2 \chi_1^2$$

where $\sigma^2 = Var(X)$.

- *When $l = 2$, define $\tilde{Z} \sim N(0, Var(X)) \in \mathbb{R}^l$, then*

$$nR_n^W(\theta_*) \Rightarrow \rho\left(\tilde{Z}\right)\left(2 - \tilde{\eta}\left(\tilde{Z}\right)\rho\left(\tilde{Z}\right)\right)\left\|\tilde{Z}\right\|_2^2$$

where $\rho := \rho\left(\tilde{Z}\right)$ *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*

$$\tilde{g}(x) = \mathbb{P}\left(\tau(0) \leq \|x\|_2^2\right),$$

where τ *is a random variable satisfying*

$$\mathbb{P}(\tau > t) = \mathbb{E}\left[\exp\left(-\left(f_X(X_1) + \kappa f_Y(X_1)\right)\pi t\right)\right].$$

And the function $\tilde{\eta} : \mathbb{R}^l \to \mathbb{R}$ *is a deterministic function given as*

$$\tilde{\eta}(x) = \mathbb{E}\left[\max\left(1 - \tau(0)/\|x\|_2^2, 0\right)\right].$$

- *When* $l \geq 3$,

$$n^{1/2 + \frac{3}{2l+2}} R_n^W(\theta_*) \Rightarrow \frac{2l+2}{l+2} \frac{\left\|\tilde{Z}\right\|_2^{1+\frac{1}{l+1}}}{\left(\mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left(f_X(X_1) + \kappa f_Y(X_1)\right)\right]\right)^{\frac{1}{l+1}}}$$

where $\tilde{Z} \sim N(0, Var(X)) \in \mathbb{R}^l$.

4.1.1. More on the limiting distribution The limiting distributions that we obtain are explicitly characterized. They depend on parameters that are meaningful in the application settings that we shall discuss. For example, the distribution from which stress scenarios are generated or the distribution of the predictors of the unlabeled data clearly play a key role in the limiting distribution. These parameters dictate the "spread"of the distribution and, consequently, the size of quantiles. So, the parameters that appear in our limit theorems readily affect the uncertainty size in a quantifiable way.

In order to make this point relatively more tangible, consider the following example based on simulated data.

Our asymptotic theorem gives different asymptotic distributions for different degrees of freedom (d.f.) in the Student-t distribution. If we select the 95% quantile for the construction of our robust risk valuation interval, we can see that the higher the d.f., the smaller the quantile, as we show in Figure 4.1.1. So, an increase in the likelihood of more extreme scenarios provided by the regulator translates directly into a larger confidence region for the risk or a larger size in the uncertainty region, in a precisely quantifiable way thanks to our results. The SOS profile function is the distance

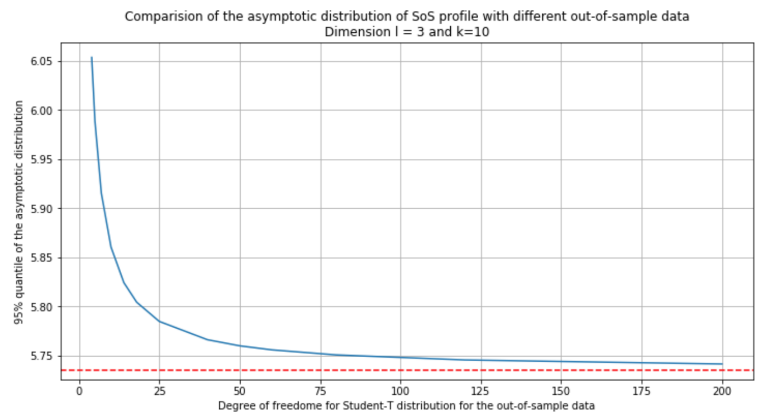

Figure 2 95% quantile for the SOS profile function asymptotic distribution (dimension being 3 and the $\kappa = 10$) with different degree of freedom for the Student-t distribution in stress scenarios. The in-sample data is standard Gaussian. The red dashed line illustrates the situation in which stress scenarios are also chosen to be standard Gaussian.

between the empirical distribution and the manifold determined by the estimating equation(s). If the in-sample data and the stress-scenario data are more similar, we would expect smaller quantiles (this corresponds to the setting in which the d.f. is large for the Student-t distribution), and we will observe larger quantiles when the two distributions are different from each other (this is the setting in which the d.f. is small for Student-t).

4.1.2. Evaluating the Limiting Distribution In Theorem 1 and in the rest of our results, the limiting distribution depends on parameters that might be unknown. For example, take the case $l \geq 3$ in Theorem 1. We obtain that

$$n^{1/2+\frac{3}{2l+2}} R_n^W(\theta_*) \Rightarrow \frac{2l+2}{l+2} \frac{\left\|\tilde{Z}\right\|_2^{1+\frac{1}{l+1}}}{(c_0)^{1/(l+1)}}, \tag{13}$$

where

$$c_0 := \mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left(f_X\left(X_1\right)+\kappa f_Y\left(X_1\right)\right)\right]$$

and $\tilde{Z} \sim N\left(0, Var\left(X\right)\right)$. This situation is quite standard when developing asymptotic distributions for hypothesis testing and the remedy is to simply use any consistent plug-in estimator to estimate the unknown quantities. For instance, we can use

$$\Sigma_n = \frac{1}{n}\sum_{j=1}^{n}\left(X_j - E_{P_n}\left(X_j\right)\right)\left(X_j - E_{P_n}\left(X_j\right)\right)^T$$

instead of $\Sigma = Var\left(X\right)$. We can also use any consistent estimator (converging on compact sets and with rapid decay at infinity) for the densities of $f_X\left(\cdot\right)$ and $f_Y\left(\cdot\right)$, say $f_X^{(n)}\left(\cdot\right)$ and $f_Y^{(n)}\left(\cdot\right)$, respectively, and estimate c_0 via

$$c_0\left(n\right) = E_{P_n}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left(f_X^{(n)}\left(X\right)+\kappa f_Y^{(n)}\left(X_1\right)\right)\right],$$

which is consistent as $n \to \infty$. Because the asymptotic distribution in (13) is continuous in c_0 and Σ, it follows that estimating quantiles based on the plug-in estimators $c_0\left(n\right)$ and Σ_n in place of c_0 and Σ leads to asymptotically equivalent specifications for the asymptotic quantiles of $R_n^W\left(\theta_*\right)$. These quantiles, in turn, can be estimated by Monte Carlo using the asymptotic limits, with the plug-in estimators in place. A completely analogous approach can be followed for the asymptotic distributions obtained in the developments that we discuss next.

4.2. SOS Function for Estimating Equations

Throughout this subsection we assume that **A1)** and **A2)** are in force. Let us assume that $h : R^d \times R^l \to R^q$ and $q \leq d$. We also impose the following assumptions.

B1) Assume $\theta_* \in R^d$ satisfies

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

B2) Furthermore, suppose that

$$\mathbb{E}\left\|h\left(\theta_*, X\right)\right\|_2^2 < \infty, \text{ and } \mathbb{E}\left\|h\left(\theta_*, Y\right)\right\|_2^2 < \infty.$$

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.

4.2.1. 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 $\mathcal{Z}_{n+m}^h(\theta_*) = \{h(\theta_*, Z_k) : Z_k \in \mathcal{Z}_{n+m}\}$ then

$$R_n^W(\theta_*) = \inf\{\int\int \|h(\theta_*, x) - h(\theta_*, z)\|_2^2 \pi(dx, dz) : \quad (14)$$
$$\text{s.t. } \pi \in \mathcal{P}\left(\mathcal{X}_n^h(\theta_*) \times \mathcal{Z}_{n+m}^h(\theta_*)\right), \pi_X = \mu_n, \int h(\theta_*, z) \pi_Z(dz) = 0\}.$$

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 1; just redefine $X_i \leftarrow h(\theta_*, X_i)$, $Z_k \leftarrow h(\theta_*, Z_k)$, and apply Theorem 1 directly. Thus the proof of the next result is omitted.

Theorem 2 **(Implicit SOS Profile Function Analysis)**. *Let us denote $g_X(\cdot)$ as 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 (14) has the following asymptotic results:*

- *When $q = 1$,*

$$nR_n^W(\theta_*) \Rightarrow Var(h(\theta_*, X_1)) \chi_1^2$$

- *When $q = 2$, if $\tilde{Z} \sim N(0, Var(h(\theta_*, X))) \in R^q$ then*

$$nR_n^W(\theta_*) \Rightarrow \rho\left(\tilde{Z}\right)\left[2 - \eta\left(\tilde{Z}\right)\rho\left(\tilde{Z}\right)\right]\left\|\tilde{Z}\right\|_2^2,$$

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 function defined as

$$\tilde{g}\left(x\right)=\mathbb{P}\left(\|x\|_2^2\geq\tau(0)\right),$$

where τ is a random variable satisfying

$$\mathbb{P}\left[\tau>t\right]=\mathbb{E}\left[\exp\left(-\left[g_X\left(h\left(\theta_*,X_1\right)\right)+\kappa g_Y\left(h\left(\theta_*,X_1\right)\right)\right]\pi t\right)\right].$$

And the function $\tilde{\eta}:\mathbb{R}^q\to\mathbb{R}$ is a deterministic continuous function given as

$$\tilde{\eta}\left(x\right)=\mathbb{E}\left[\max\left(1-\tau(0)/\|x\|_2^2,0\right)\right].$$

- *When $q\geq 3$,*

$$n^{1/2+\frac{3}{2q+2}}R_n^W\left(\theta_*\right)\Rightarrow\frac{2q+2}{q+2}\frac{\left\|\tilde{Z}\right\|_2^{1+\frac{1}{q+1}}}{\left(\mathbb{E}\left[\frac{\pi^{q/2}}{\Gamma(q/2+1)}\left(g_X\left(h\left(\theta_*,X_1\right)\right)+\kappa g_Y\left(h\left(\theta_*,X_1\right)\right)\right)\right]\right)^{\frac{1}{q+1}}}$$

where $\tilde{Z}\sim N\left(0,Var\left(h\left(\theta_*,X\right)\right)\right)\in\mathbb{R}^q$.

4.2.2. 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).

$$R_n^W\left(\theta_*\right)=\inf\{\int\int\|x-z\|_2^2\,\pi\left(dx,dz\right): \quad \text{s.t.}\;\; \pi\in\mathcal{P}\left(\mathcal{X}_n\times\mathcal{Z}_{(n+m)}\right),\pi_X=\mu_n,\int h\left(\theta_*,z\right)\pi_Z\left(dz\right)=0\}. \tag{15}$$

Both the implicit and explicit SOS formulations 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 Section 2.2. 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 1, it does require some additional assumptions that are not immediately clear without examining the proof of Theorem 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 7.3.

Theorem 3 **(Explicit SOS Profile Function Analysis)**. *Under assumptions A1)-A2), B1)-B2) and BE1)-BE2), we have that (15) satisfies*

- *When $l = 1$,*

$$nR_n^W(\theta_*) \Rightarrow \tilde{Z}^T \Upsilon^{-1} \tilde{Z}$$

where $\tilde{Z} \sim N(0, Var(h(\theta_*, X))) \in \mathbb{R}^q$.

- *Assume that $l = 2$. Let $\tilde{Z} \sim N(0, Var(h(\theta_*, X))) \in \mathbb{R}^q$. It is possible to uniquely define deterministic continuous mapping, $\tilde{\zeta} : \mathbb{R}^q \to \mathbb{R}^q$, such that $\tilde{\zeta}(z)$ is defined via*

$$z = -\mathbb{E}\left[V_1 I\left(\tau \le \tilde{\zeta}^T(z) V_1 \tilde{\zeta}(z)\right)\right] \tilde{\zeta}(z),$$

where τ is independent of V_1 satisfying

$$\mathbb{P}(\tau > t) = \mathbb{E}(\exp(-[f_X(X_1) + \kappa f_Y(X_1)]\pi t)).$$

Then, we have that,

$$nR_n^W(\theta_*) \Rightarrow -2\tilde{Z}^T\tilde{\zeta}\left(\tilde{Z}\right) - \tilde{\zeta}^T\left(\tilde{Z}\right)\tilde{G}\left(\tilde{\zeta}\left(\tilde{Z}\right)\right)\tilde{\zeta}\left(\tilde{Z}\right),$$

where $\tilde{G}:\mathbb{R}^q \to \mathbb{R}^{q\times q}$ *is a deterministic continuous mapping defined as*

$$\tilde{G}(\zeta) = \mathbb{E}\left[V_1 \max\left(1-\tau/\left(\zeta^T V_1 \zeta\right), 0\right)\right].$$

- *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*

$$z = -\mathbb{E}\left[\frac{\pi^{l/2}\left(f_X\left(X_1\right)+\kappa f_Y\left(X_1\right)\right)}{\Gamma(l/2+1)}V_1\cdot\left(\tilde{\zeta}^T(z)V_1\tilde{\zeta}(z)\right)^l\right]\tilde{\zeta}(z),$$

(note that V_1 *is a function of* X_1, *so these are correlated). Moreover,*

$$n^{1/2+\frac{3}{2l+2}}R_n^W(\theta_*) \Rightarrow -2\tilde{Z}^T\tilde{\zeta}\left(\tilde{Z}\right) - \frac{2}{l+2}\tilde{G}\left(\tilde{Z}\right),$$

where $\tilde{Z} \sim N\left(0, Var\left(h\left(\theta_*, X\right)\right)\right) \in \mathbb{R}^q$ *and* $\tilde{G}:\mathbb{R}^q \to \mathbb{R}$ *is a deterministic continuous function defined as*

$$\tilde{G}(\zeta) = \mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left(f_X\left(X_1\right)+\kappa f_Y\left(X_1\right)\right)\left(\zeta^T V_1 \zeta\right)^{l/2+1}\right].$$

We should observe that unlike the 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 $h(\theta_*, X_i) \in \mathbb{R}^q$.

4.3. 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 (for example Lagrange multipliers in the setting of constrained optimization).

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 Theorem 2 and Theorem 3, will be given in Section 7.

Throughout this subsection, let us suppose that $h(\gamma,\nu,x)\in\mathbb{R}^q$. In addition, we impose the following assumptions.

C1) Given γ_* there is a unique $\nu_*\in R^p$ such that

$$\mathbb{E}\left[h\left(\gamma_*,\nu,X\right)\right]=0 \tag{16}$$

and, given ν_*, we also assume that γ_* satisfies

$$\mathbb{E}\left[h\left(\gamma,\nu_*,X\right)\right]=0. \tag{17}$$

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 $h(\gamma,\cdot,x)$ 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 $M(\cdot):R^l\to(0,\infty)$ satisfying that

$$\left\|h\left(\gamma_*,\nu,x\right)\right\|_2^2\le M(x) \text{ for a.e. } \nu\in\mathcal{V},$$

$$\left\|D_\nu h\left(\gamma_*,\nu,x\right)\right\|_2^2\le M(x) \text{ for a.e. } \nu\in\mathcal{V},$$

and $E\left(M\left(X_1\right)\right)<\infty$ and $E\left(M\left(Y_1\right)\right)<\infty$.

4.3.1. 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

$$R_n^W\left(\gamma_*\right)=\inf\{\int\int\left\|h\left(\gamma_*,v_n,x\right)-h\left(\gamma_*,v_n,z\right)\right\|_2^2\pi\left(dx,dz\right): \tag{18}$$
$$\text{s.t. }\ \pi\in\mathcal{P}\left(\mathcal{X}_n^h\left(\gamma_*,v_n\right)\times\mathcal{Z}_{n+m}^h\left(\gamma_*,v_n\right)\right),\pi_X=\mu_n,\int h\left(\gamma_*,v_n,z\right)\pi_Z\left(dz\right)=0\},$$

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 (16) 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 (18), as we indicate next.

COROLLARY 1 **(Plug-in for Implicit SOS Formulation)**. *Assume* ***A1)-A2)*** *and* ***C1)-C4)*** *hold. Moreover, suppose we 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 (18) has following asymptotic behavior:*

- *When $q = 1$,*

$$nR_n^W(\gamma_*) \Rightarrow \left(\tilde{Z}'\right)^2.$$

- *When $q = 2$,*

$$nR_n^W(\gamma_*) \Rightarrow \rho\left(\tilde{Z}'\right)\left[2 - \tilde{\eta}\left(\tilde{Z}'\right)\rho\left(\tilde{Z}'\right)\right]\left\|\tilde{Z}'\right\|_2^2$$

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, τ satisfies

$$\mathbb{P}[\tau > t] = \mathbb{E}\left[\exp\left(-\left[g_X(h(\gamma_*, \nu_*, X_1)) + \kappa g_Y(h(\gamma_*, \nu_*, X_1))\right]\pi t\right)\right].$$

- *When $q \geq 3$,*

$$n^{1/2 + \frac{3}{2q+2}} R_n^W(\gamma_*) \Rightarrow \frac{2q+2}{q+2} \frac{\left\|\tilde{Z}'\right\|_2^{1+\frac{1}{q+1}}}{\left(\mathbb{E}\left[\frac{\pi^{q/2}}{\Gamma(q/2+1)}\left(g_X(h(\gamma_*, \nu_*, X_1)) + \kappa g_Y(h(\gamma_*, \nu_*, X_1))\right)\right]\right)^{\frac{1}{q+1}}}.$$

4.3.2. 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 (15) as simply plugging in the nuisance parameter:

$$R_n^W(\gamma_*) = \inf \Big\{ \int\int \|x-z\|_2^2 \, \pi(dx, dz) : \tag{19}$$
$$\text{s.t. } \pi \in \mathcal{P}\left(\mathcal{X}_n \times \mathcal{Z}_{(n+m)}\right), \pi_X = \mu_n, \int h(\gamma_*, v_n, z)\, \pi_Z(dz) = 0 \Big\}.$$

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 $\bar{\Upsilon} = \mathbb{E}\left(\bar{V}_i\right)$ is strictly positive definite.

C6) The function $M(\cdot)$ from condition **C4)** also satisfies

$$\|D_x h(\gamma_*, \nu, x)\|_2^2 \le M(x) \text{ for a.e. } \nu \in \mathcal{V}.$$
$$\|D_\nu D_x h(\gamma_*, \nu, x)\|_2^2 \le M(x) \text{ for a.e. } \nu \in \mathcal{V}.$$

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 2 **(Plug-in for Explicit SOS Formulation)**. *Let $X_i \in \mathbb{R}^l$, $h(\gamma, \nu, x) \in \mathbb{R}^q$, and assume that A1)-A2) and C1)-C7) hold. We notice $\tilde{Z}'$ is defined in C2). Then, the SOS profile function defined in Equation (19) has the following asymptotic properties:*

- *When $l = 1$,*

$$nR_n^W(\gamma_*) \Rightarrow \tilde{Z}'^T \bar{\Upsilon}^{-1} \tilde{Z}'.$$

- *Suppose that $l = 2$. It is possible to uniquely define deterministic continuous mapping $\tilde{\zeta} : \mathbb{R}^q \to \mathbb{R}^q$, such that*

$$z = -\mathbb{E}\left[\bar{V}_1 I\left(\tau \le \tilde{\zeta}^T(z)\, \bar{V}_1 \tilde{\zeta}(z)\right)\right] \tilde{\zeta}(z),$$

where τ is independent of $\bar{V}_1$ and it satisfies

$$\mathbb{P}(\tau > t) = \mathbb{E}\left(\exp\left(-\left[f_X(X_1) + \kappa f_Y(X_1)\right] \pi t\right)\right).$$

Furthermore,

$$nR_n^W(\gamma_*) \Rightarrow -2\tilde{\zeta}^T\left(\tilde{Z}'\right)\tilde{Z}' - \tilde{\zeta}^T\left(\tilde{Z}'\right)\tilde{G}\left(\tilde{\zeta}\left(\tilde{Z}'\right)\right)\tilde{\zeta}\left(\tilde{Z}'\right),$$

where $\tilde{G}:\mathbb{R}^q \to \mathbb{R}^{q\times q}$ *is a deterministic continuous mapping defined as*

$$\tilde{G}(\zeta) = \mathbb{E}\left[\bar{V}_1 \max\left(1 - \tau/\left(\zeta^T \bar{V}_1 \zeta\right), 0\right)\right].$$

- *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*

$$z = -\mathbb{E}\left[\frac{\pi^{l/2}\left(f_X(X_1) + \kappa f_Y(X_1)\right)}{\Gamma(l/2+1)}\bar{V}_1\left(\tilde{\zeta}^T(z)\bar{V}_1\tilde{\zeta}(z)\right)^l\right]\tilde{\zeta}(z)$$

(note that $\bar{V}_1$ *is a function of* X_1*). Moreover,*

$$n^{1/2+\frac{3}{2l+2}}R_n^W(\gamma_*) \Rightarrow -2\tilde{\zeta}^T\left(\tilde{Z}'\right)\tilde{Z}' - \frac{2}{l+2}\tilde{G}\left(\tilde{\zeta}\left(\tilde{Z}'\right)\right),$$

where $\tilde{G}:\mathbb{R}^q \to \mathbb{R}$ *is a deterministic continuous function defined as*

$$\tilde{G}(\zeta) = \mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left(f_X(X_1) + \kappa f_Y(X_1)\right)\left(\zeta^T\bar{V}_1\zeta\right)^{l/2+1}\right].$$

5. Application to Stochastic Optimization and Stress Testing

We will provide an application of the SOS inference framework to quantify model uncertainty in the context of stochastic programming. Motivating applications include the evaluation of Conditional Value at Risk (C-VaR) and semi-supervised learning settings, as we shall discuss in the examples below.

We are interested in the value function of a stochastic programming problem formulation via

$$\begin{aligned} C_* = \min_{\theta} \quad & \mathbb{E}\left[m(\theta, X)\right] \\ s.t.\ & \mathbb{E}[\phi(\theta, X)] \leq 0. \end{aligned} \tag{20}$$

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 shall assume that $\phi(\cdot, x)$ is a convex function for any x.

Following Blanchet and Murthy (2019), 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 (20) 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 $C_* = \min_\theta \{\mathbb{E}[m(\theta, X)]\}$.

The authors in Lam and Zhou (2015, 2017) provide a discussion for some potential approaches to derive nonparametric confidence interval (including Empirical Likelihood, a Bayesian approach, bootstrap and the delta method). In Lam and Zhou (2015, 2017) it is argued that the Empirical Likelihood method tends to have superior finite sample performance, and Blanchet et al. (2019a) provides an optimal (in certain sense) specification for the Empirical Likelihood approach. More importantly, in Blanchet et al. (2019a) an approach combining Empirical Likelihood and a plug-in estimator for the 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 4. The approach that we introduce next is the analog of the plug-in strategy discussed in Blanchet et al. (2019a) in order to find a robustified confidence interval for C_*.

The following corollary plays the key role in specifying confidence interval for C_*. The result is a direct extension of Corollary 1 and Corollary 2, provided the following assumptions are in place.

We define $M(\theta) = \mathbb{E}[m(\theta, X)]$, and the assumptions are

D1): Assume $m(\cdot)$ is convex differentiable in θ, then $M(\theta)$ is also convex differentiable. We assume there is a unique optimizer θ_*n for $M(\theta)$.

D2): Assume that $m(\cdot)$ is strongly convex at θ_*, that is, there exist $\delta > 0$, such that for every θ

$$M(\theta) \geq M(\theta_*) + \delta \|\theta - \theta_*\|_2^2.$$

Corollary 3. *Let us consider stochastic programming problem* $C_* = \min_\theta M(\theta) = \min_\theta \mathbb{E}[m(\theta, X)]$. *Assume that D1)-D2) hold. We consider the estimating equations to be the derivative condition and value function condition*

$$\mathbb{E}[m(\theta_*, X) - C_*] = 0, \text{ and } \mathbb{E}[D_\theta m(\theta_*, X)] = 0.$$

For simplicity, let us denote $h(\theta_*, C_*, x) = \left(m(\theta_*, x) - C_*, D_\theta m(\theta_*, x)^T\right)^T$. *We are interested in* C_* *only and consider a sample average approximation (SAA) estimator for* θ_* *to be* $\hat{\theta}_{SAA}$. *For* $h(\cdot, C_*, x)$ *we assume C1)-C7) hold. Let us denote* $U \sim N(0, Var(m(\theta_*, X))) \in \mathbb{R}$ *and* $U(0) = \left(U, \vec{0}\right)^T \in \mathbb{R}^{d+1}$. *Recalling 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* $d+1 = 2$*)*

$$nR_n^W(C_*) \Rightarrow \rho(U)\left[2 - \tilde{\eta}(U)\rho(U)\right]U^{2,}$$

where $\rho(U)$ *is the unique solution to*

$$\frac{1}{\rho} = \tilde{g}(\rho U),$$

and $\tilde{g}: \mathbb{R} \to \mathbb{R}$ *is a deterministic continuous function defined as*

$$\tilde{g}(x) = \mathbb{P}\left[x^2 \geq \tau\right].$$

$\tilde{\eta}(x)$ *is also a deterministic function, defined as*

$$\tilde{\eta}(x) = \mathbb{E}\left[\max\left(1 - \tau/x^2, 0\right)\right],$$

and τ *satisfies*

$$\mathbb{P}[\tau > t] = E\left(\exp\left(-\left(g_X\left(h\left(\theta_*, C_*, X_1\right) + \kappa g_Y\left(h\left(\theta_*, C_*, X_1\right)\right)\right)\pi t\right)\right)\right).$$

- *When* $d \geq 2$,

$$n^{1/2+\frac{3}{2d+4}} R_n^W(C_*) \Rightarrow \frac{2d+4}{d+3} \frac{||U||_2^{1+\frac{1}{d+2}}}{\mathbb{E}\left[\frac{\pi^{(d+1)/2}}{\Gamma((d+3)/2)}\left(g_X\left(h\left(\theta_*, C_*, X_1\right)\right) + g_Y\left(h\left(\theta_*, C_*, X_1\right)\right)\right)\right]^{\frac{1}{d+2}}}.$$

For the explicit formulation, we have the following asymptotic results (we use $\zeta_{[1]}$ *to denote the 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* $\tilde{\zeta}:\mathbb{R}^q\rightarrow\mathbb{R}^q$, *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*

$$\mathbb{P}(\tau>t)=\mathbb{E}\left(\exp\left(-\left[f_X(X_1)+\kappa f_Y(X_1)\right]\pi t\right)\right).$$

Furthermore,

$$nR_n^W(C_*)\Rightarrow -2U\tilde{\zeta}_{[1]}-\tilde{\zeta}^T(U(0))\tilde{G}\left(\tilde{\zeta}(U(0))\right)\tilde{\zeta}(U(0)),$$

where $\tilde{G}:\mathbb{R}^q\rightarrow\mathbb{R}^{q\times q}$ *is a deterministic continuous mapping defined as*

$$\tilde{G}(\zeta)=\mathbb{E}\left[\bar{V}_1\max\left(1-\frac{\tau}{\zeta^T\bar{V}_1\zeta},0\right)\right],$$

and U *is independent with* $\bar{V}_1$ *and* τ.

- *Assume that* $l\geq 3$. *A continuous function* $\tilde{\zeta}:\mathbb{R}^q\rightarrow\mathbb{R}^q$ *can be defined uniquely so that*

$$z=-\mathbb{E}\left[\frac{\pi^{l/2}\left(f_X(X_1)+\kappa f_Y(X_1)\right)}{\Gamma(l/2+1)}\bar{V}_1\left(\tilde{\zeta}^T(z)\bar{V}_1\tilde{\zeta}(z)\right)^l\right]\tilde{\zeta}(z)$$

(note that $\bar{V}_1$ *is a function of* X_1*). Moreover,*

$$n^{1/2+\frac{3}{2l+2}}R_n^W(C_*)\Rightarrow -2U\tilde{\zeta}_{[1]}-\frac{2}{l+2}\tilde{G}\left(\tilde{\zeta}(U(0))\right),$$

where $\tilde{G}:\mathbb{R}^q\rightarrow\mathbb{R}$ *is a deterministic function given as*

$$\tilde{G}(\zeta)=\mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left(f_X(X_1)+\kappa f_Y(X_1)\right)\left(\zeta^T\bar{V}_1\zeta\right)^{l/2+1}\right],$$

and U *and* X_1 *are independent.*

As indicated earlier, the corollary is a special case of Corollary 1 and Corollary 2, so the proof is omitted. The estimating equations correspond to the first order optimality condition (i.e., the first derivative equal to zero) and the corresponding optimal value equation. We use sample average approximation estimator as the underlying plug-in estimator.

We notice that for sample average approximation, under assumptions D1)-D2), it has been shown in Ruszczynski and Shapiro (2003), Shapiro and Dentcheva (2014) that the optimizer $\hat{\theta}_{SAA}$ and the optimal value function $\frac{1}{n}\sum_{i=1}^{n} m\left(\hat{\theta}_{SAA}, X_i\right)$ satisfy

$$\begin{aligned}
&\hat{\theta}_{SAA} - \theta_* = O\left(1/n^{1/2}\right)\\
&\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta m\left(\hat{\theta}_{SAA}, X_i\right) = 0,\\
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m\left(\hat{\theta}_{SAA}, X_i\right) - C_*\right) \Rightarrow N\left(0, \mathrm{Var}\left(m\left(\theta_*, X\right)\right)\right).
\end{aligned}$$

Therefore, Corollary 2 and Corollary 1 apply.

Similar to the derivation in Blanchet et al. (2019a) in the setting of Empirical Likelihood, for the plug-in estimator derived from sample average approximation, if we denote $n^{1/2+3/(2d+4)} R_n^{W(implicit)}(C_*) \Rightarrow R_0^{(implicit)}$ 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 and $R_0^{(\cdot)}(95\%)$ is the upper 95% quantile for $R_0^{(explicit)}$ (or $R_0^{(implicit)}$). The upper/lower bound of confidence interval ($C_{up}^{(\cdot)}/C_{l0}^{(\cdot)}$) can be found by solving the linear programming problem

$$\begin{aligned}
C_{up}^{(\cdot)}/C_{lo}^{(\cdot)} = &\max_{\pi(i,j)} \ / \ \min_{\pi(i,j)} \{\sum_{i,j=1}^{n} \pi(i,j) m(\hat{\theta}_{SAA}, X_i)\\
&s.t.\ \pi(i,j) \geq 0 \ \sum_{j=1}^{n} \pi(i,j) = 1/n; \sum_{i,j=1}^{n} \pi(i,j) \left\|X_i - X_j\right\|_2^2 \leq \frac{R_0^{(\cdot)}\left(95\%\right)}{n^\alpha}\}.
\end{aligned}$$

Next, we are going to provide a numerical example in quantifying C-VaR using the methodology we developed above.

EXAMPLE 1 (QUANTIFY THE UNCERTAINTY OF CONDITIONAL VALUE AT RISK (C-VAR)). In this example we would like to find an 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 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$(0.9) = 5.066$.

In this example, we have three approaches in which our SOS procedure can be applied: 1) implicit SOS formulation (ISOS); 2) explicit SOS formulation assuming data being of dimension l (ESOS-O), i.e. $X_i = \left(X_i^{(1)}, \ldots, X_i^{(l)}\right)^T \in \mathbb{R}^l$; 3) explicit formulation assuming data being of dimension 1 (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) in Blanchet et al. (2019a), nonparametric bootstrap method (BT), and central limit theorem-based Delta method (CLT) discussed in Theorem 5.7 Shapiro and Dentcheva (2014). 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 1 for Laplace distribution below.

We can observe that the three SOS-based approaches seem to have comparable coverage probabilities in most cases, for both generating distributions, in comparison to the EL, bootstrap, and delta method. However, for small sample situations ($n = 20$) EL and all of the SOS-based approaches appear to perform better than the rest. It is discussed in Lam and Zhou (2015) that EL has better finite sample performance compared to 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 compared 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 combined the data into 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 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.

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
	IISOS	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 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.

In addition, we report the computational time for our calculation in Table 5. The different formulations of SOS-based methods share the same computation cost, thus we only report the case for implicit SOS. We report the average calculating time in seconds with thousands of experiments, where the experiments are implemented in Python with Scipy optimizers and our machine is equipped with an Intel i7 3.5Ghz processor and 16GB memory. Our SOS based method requires solving the C-VaR optimization problem once, then solve the linear programming. The EL based method is similar, with solving the C-VaR optimization problem once, it then solves a convex optimization problem. Finally, the bootstrap based method requires solving the C-VaR optimization repetitively. We can observe that for the example we consider, our SOS-based method does not face computational challenges compared with other methods.

	20	50	100	500
ISOS	0.042	0.108	0.613	14.069
EL	0.018	0.069	0.401	7.272
BT	0.099	1.038	2.085	18.023

Table 3 Computational Cost for Our C-VaR examples. The average computational time in seconds for different algorithms with different sample sizes.

Example 2 (Semi-supervised Learning). We consider the DRO formulation for Semi-supervised Learning (SSL) as suggested in Blanchet and Kang (2018). We formulate the data-driven DRO problem and compare the results for choosing the distributional uncertainty size with the above asymptotic results of SOS function as suggested in Corollary 3. We consider the MiniBooNE data set from UCI machine learning data base Blake and Merz (1998). We consider logistic regression as our baseline model and form SSL-DRO formulation. For each iteration, we randomly split the data into labeled training set with size $n = 30$, unlabeled training set with size $N - n = 5000$, and testing set with size $n = 125034$. We compare the choice of the uncertainty size using 5-fold cross-validation and SoS asymptotic results. We also include the results for logistic regression and regularized logistic regression as reference. We report the average training error and testing error as log-exponential loss and testing accuracy as accurate classification rate. The mean and standard deviation of the training error, testing error, and testing accuracy are evaluated via 500 independent experiments. The details are included in Table 4.

	Training Error	Testing Error	Testing Accuracy
Logistic Regression	0 ± 0	18.2 ± 10.0	$.678 \pm .059$
LRL1 with CV	$.401 \pm .167$	$.910 \pm .131$	$.717 \pm .041$
DRO-SSL with CV	$.287 \pm .047$	$.609 \pm .054$	$.710 \pm .032$
DRO-SSL with SoS	$.304 \pm .045$	$.682 \pm .048$	$.709 \pm .028$

Table 4 Numerical Results for Semi-supervised Learning.

6. Conclusions and Discussion

This paper introduces a methodology inspired by Empirical Likelihood, but in which the likelihood ratio function is replaced by a Wasserstein distance. The method that we propose is motivated by the problem of systematically finding estimators that incorporate out-of-sample performance in their design.

In turn, as a motivation for the need to find these types of estimators we discussed applications to stress testing and semi-supervised learning, which have been discussed in the body of this paper. Another way in which we can justify our framework is as an approximation approach to solving the problem

$$\min_{\theta\in R^l} \max_{P\in\mathcal{U}_\delta\left(P_n;\mathbb{R}^d\right)} E_P[\mathcal{L}\left(X,\theta\right)].$$

It turns out that in great generality (see Esfahani and Kuhn (2018))

$$\max_{P\in\mathcal{U}_\delta\left(P_n;\mathbb{R}^d\right)} E_P[\mathcal{L}\left(X,\theta\right)] = \min_{\lambda\geq 0}\{\lambda\delta + E_{P_n}[f\left(X,\theta;\lambda\right)]\},$$

where $f\left(x,\theta;\lambda\right)$ is defined as the solution of an optimization problem involving a parameter $y\in\mathbb{R}^d$ which we refer to as the "inner optimization problem."The inner optimization problem is typically not convex and therefore it is challenging to solve. There are cases in which the inner optimization problem can be solved in closed form, however, and many of those cases have been documented in the literature in Esfahani and Kuhn (2018). Our results can be used to suitably calibrate an alternative formulation that may be more tractable given that $y\in\mathbb{R}^d$ is replaced by $y\in\mathcal{Z}_{n,m}$.

There are a number of structural properties in our procedure that are worth investigating and that we plan to explore in future work. For instance, we believe the choice of a particular cost in optimal transport distance deserves substantial analysis. In this paper we have chosen the L_2 Wasserstein metric to illustrate our results. The methodology that we propose can be extended to cover other Wasserstein metrics, so on the technical side our work provides the foundations for such extensions. However, it is the impact of such selection that appears to also bring about interesting connections. This already is made evident from our work Blanchet et al. (2019b) in which we see that the connections that we mentioned earlier in this discussion (to LASSO and SVM) are made after carefully choosing a natural Wasserstein metric.

In addition, given the parallel philosophy underpinning the method that we proposed (based on Empirical Likelihood), the results described in this paper open up a significant amount of research opportunities that 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), Wang and Rao (2001), Zhao and Wang (2008), Chen and Keilegom (2009), Murphy (1995), Li et al. (1996), Hollander and McKeague (1997), Li et al. (1997), Einmahl and McKeague (1999), Wang et al. (2009), Zhou (2015)), machine learning (see Duchi et al. (2016), Hu et al. (2018), Duchi and Namkoong (2018), Blanchet et al. (2019d)), 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, 2017), Blanchet et al. (2019a)). 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 paper (or natural extensions).

7. 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 $\tilde{X}_i = X_i - \theta_*$ and apply the analysis to the $\tilde{X}_i$'s.

7.1. The Dual Problem and High-Level Understanding of Results

The Dual Problem Let us revisit the definition of (12) and write it as a linear programming problem,

$$R_n^W(\theta_*) = \min_{\pi(i,j)\geq 0} \quad \sum_{i=1}^{n}\sum_{j=1}^{m+n} \pi(i,j)\left\|X_i - Z_j\right\|_2^2 \tag{21}$$

$$\text{s.t.} \begin{cases} \sum_{j=1}^{(m+n)} \pi(i,j) = 1/n, \text{ for all } i \\ \sum_{j=1}^{(m+n)} \left(\sum_{i=1}^{n} \pi(i,j)\right) Z_j = 0 \end{cases}.$$

We know with probability one 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 (1973), we can write (21) in the dual formulation as

$$R_n^W(\theta_*) = \max_{\lambda, \tilde{\gamma}_i} \left\{ -\frac{1}{n} \sum_{i=1}^n \tilde{\gamma}_i \right\}$$
$$s.t. \quad \tilde{\gamma}_i + \|X_i - Z_j\|_2^2 - \lambda^T Z_j \geq 0 \text{ for all } i, j.$$

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

$$R_n^W(\theta_*) = \max_{\lambda, \gamma_i \geq 0} \left\{ -\lambda^T \bar{X}_n - \frac{1}{n} \sum_{i=1}^n \gamma_i \right\}$$
$$\text{s.t.} \ -\lambda^T X_i - \gamma_i \leq -\lambda^T Z_j + \|X_i - Z_j\|_2^2, \text{ for all } i, j.$$

We can further simplify the constraints by minimizing over j, while keeping i fixed, therefore arriving to the simplified dual formulation

$$R_n^W(\theta_*) = \max_{\lambda, \gamma_i \geq 0} \left\{ -\lambda^T \bar{X}_n - \frac{1}{n} \sum_{i=1}^n \gamma_i \right\} \tag{22}$$
$$\text{s.t.} \ -\lambda^T X_i - \gamma_i \leq \inf_j \left\{ -\lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\}, \text{ for all } i.$$

High-Level Intuitive Analysis At this point we can perform a high-level 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 (22), in this portion the reader can appreciate that the assumption that X_j has a density yields

$$\inf_j \left\{ \|Z_j - (X_i + \lambda/2)\|_2^2 \right\} = \epsilon_n(i), \tag{23}$$

where error $\epsilon_n(i)$ is small as $n \to \infty$ and it will be discussed momentarily. Equation (23) is equivalent to

$$\inf_j \left\{ -\lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\} = -\lambda^T X_i - \|\lambda\|_2^2/4 + \epsilon_n(i).$$

Hence, the i-th constraint in (22) takes the form

$$-\lambda^T X_i - \gamma_i \leq -\lambda^T X_i - \|\lambda\|_2^2/4 + \epsilon_n(i),$$

and thus (22) can ultimately be written as

$$R_n^W(\theta_*) = -\min_{\lambda,\gamma_i \geq 0} \left\{ \lambda^T \bar{X}_n + \frac{1}{n}\sum_{i=1}^{n} \gamma_i \right\} \tag{24}$$

$$\text{s.t. } \gamma_i \geq (1-\epsilon_n(i)) \|\lambda\|_2^2/4 \text{ for all } i.$$

Now, observe that if Z_j was free, then the optimal choice in (23) would be $a_*(i) = X_i + \lambda/2$.

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

$$R_n^W(\theta_*) = -\min_{\lambda} \left\{ \lambda \bar{X}_n + (1 + O_p(1/n)) \lambda^2/4 \right\},$$

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

$$nR_n^W(\theta_*) = n\bar{X}_n^2 + nO_p\left(1/n^{3/2}\right) \Rightarrow Var(X)\chi_1^2, \tag{25}$$

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 $R_n^W(\theta_*)$ 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

$\lambda_*(n) = -\bar{X}_n + O_p(n^{-1/2})$, but we know from the CLT that $\bar{X}_n = O_p(n^{-1/2})$ 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 $\epsilon_i(n) = O(n^{-1/l})$ 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 sub-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 $O(n^{-1/l})$ from the free optimizer $a_*(i)$ arising in (23).

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, 1990), Qin and Lawless (1994)), which builds on Wilks (1938).

7.2. Proof of Theorem 1

The proof of Theorem 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 paper.

7.2.1. Step 1 (Dual Formulation and Lower Bound): Using the same transformations introduced in (21) we can obtain the dual formulation of the SOS profile function (12), which is a natural adaptation of (22), namely

$$R_n^W(\theta_*) = \max_{\lambda, \gamma_i \geq 0} \left\{ -\lambda \bar{X}_n - \frac{1}{n} \sum_{i=1}^{n} \gamma_i \right\}$$
$$\text{s.t. } -\lambda^T X_i - \gamma_i \leq \inf_j \left\{ -\lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\}, \text{ for all } i.$$

Observe that the following lower bound applies by optimizing over $a \in R^l$ instead of $a = Z_j \in \mathcal{Z}_n$, therefore obtaining the lower bound

$$\inf_j \left\{ -\lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\} \geq \inf_a \left\{ -\lambda^T a + \|X_i - a\|_2^2 \right\}$$
$$= -\lambda^T X_i - \|\lambda\|_2^2 / 4,$$

with the optimizer $a_*(X_i, \lambda) = X_i + \lambda/2$.

7.2.2. Step 2 (Auxiliary Poisson Point Processes): Then, for each i let us define a point process,

$$N_n^{(i)}(t,\lambda) = \#\left\{Z_j : \|Z_j - a_*(X_i,\lambda)\|_2^2 \le t^{2/l}/n^{2/l}, Z_j \ne X_i\right\},$$

(recall that $Z_j \in R^l$). 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,

$$\mathbb{P}\left[\|Y_j - a_*(X_i,\lambda)\|_2^2 \le t^{2/l}/n^{2/l} \middle| X_i\right] = f_Y(X_i + \lambda/2) \frac{\pi^{l/2}}{\Gamma(l/2+1)} \frac{t}{n} + o_p(t/n).$$

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,

$$\begin{aligned}
N_n^{(i)}(t,\lambda,1)|X_i &\sim \text{Bin}\left(f_X(X_i+\lambda/2)\frac{\pi^{l/2}}{\Gamma(d/2+1)}\frac{t}{n} + o_p(t/n), n-1\right),\\
N_n^{(i)}(t,\lambda,2)|X_i &\sim \text{Bin}\left(f_Y(X_i+\lambda/2)\frac{\pi^{l/2}}{\Gamma(l/2+1)}\frac{t}{n} + o_p(t/n), [\kappa n]\right),\\
N_n^{(i)}(t,\lambda) &= N_n^{(i)}(t,\lambda,1) + N_n^{(i)}(t,\lambda,2).
\end{aligned}$$

Moreover, we have as $n \to \infty$,

$$f_X(X_i+\lambda/2)\frac{\pi^{l/2}}{\Gamma(l/2+1)}\frac{t}{n} \times (n-1) \to f_X(X_i+\lambda/2)\frac{\pi^{l/2}}{\Gamma(l/2+1)} t.$$

Thus, by Poisson approximation to the binomial distribution, we have the weak convergence result

$$N_n^{(i)}(\cdot,\lambda,1)|X_i \Rightarrow \text{Poisson}\left(f_X\left(X_i+\lambda/2\right)\frac{\pi^{l/2}}{\Gamma(l/2+1)}\cdot\right),$$

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

$$\Lambda\left(X_i,\lambda\right) = \left[f_X\left(X_i+\lambda/2\right)+\kappa f_Y\left(X_i+\lambda/2\right)\right]\frac{\pi^{l/2}}{\Gamma(l/2+1)}.$$

Let us write $T_i\left(n,\lambda\right)$ to denote the first arrival time of $N_n^{(i)}(\cdot,\lambda)$, that is,

$$T_i\left(n,\lambda\right) = \inf\left\{t\geq 0 : N_n^{(i)}(t,\lambda)\geq 1\right\}$$

Then, we can specify the survival function for $T_i\left(n\right)$ to be:

$$\mathbb{P}\left[T_i\left(n,\lambda\right)>t \mid X_i\right] = \mathbb{P}\left[N_n^{(i)}(t,\lambda)=0 \;\middle|\; X_i\right] = \exp\left(-\Lambda\left(X_i,\lambda\right)t\right)\left(1+O\left(1/n^{1/l}\right)\right), \tag{26}$$

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 previous string of equalities. Motivated by the form in the right hand side of (26) we define $\tau_i\left(X_i,\lambda\right)$ to be a random variable such that

$$\mathbb{P}\left[\tau_i\left(X_i,\lambda\right)>t|X_i\right] = \exp\left(-\Lambda\left(X_i,\lambda\right)t\right),$$

and we drop the dependence on X_i and the subindex i when we refer to the unconditional version of $\tau_i\left(X_i,\lambda\right)$, namely

$$\mathbb{P}\left[\tau\left(\lambda\right)>t\right] = \mathbb{E}\left[\exp\left(-\Lambda\left(X_1,\lambda\right)t\right)\right].$$

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 $\left\{T_i\left(n,\lambda\right)\right\}_{i=1}^n$.

LEMMA 1. *For any $T \in (0,\infty)$ (deterministic) and $\alpha \in (0,2]$, we have that*

$$\overline{\lim}_{n\to\infty}\mathbb{E}\left(\sup_{t\in[0,T]}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(I\left(T_i\left(n,\lambda\right)\le t\right)-\mathbb{P}[T_i\left(n,\lambda\right)\le t]\right)\right|\right)<\infty,$$

and

$$\overline{\lim}_{n\to\infty}\mathbb{E}\left(\sup_{t\in[0,T]}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(\max\left(t^2-T_i(n,\lambda)^{\alpha},0\right)-\mathbb{E}\left[\max\left(t^2-T_i(n,\lambda)^{\alpha},0\right)\right]\right)\right|\right)<\infty.$$

The second technical lemma deals with local properties of the distribution of $T_i\left(n,\lambda\right)$. The proofs of both of these technical results are given at the end of the proof of Theorem 1, in Section 7.2.7.

LEMMA 2. *For $X_i \in \mathbb{R}^l$ and any finite t, we have the Poisson approximation to binomial as:*

$$\mathbb{P}\left[T_i\left(n,\lambda\right)\le t\right]-\mathbb{P}\left[\tau(\lambda)\le t\right]=O(t^{1+1/l}/n^{1/l}),$$

and

$$\mathbb{P}\left[T_i\left(n,\lambda\right)\le t\right]-\mathbb{P}\left[\tau(\lambda)\le t\right]=\mathbb{P}\left[\tau>t\right]O\left(1/n^{l}\right).$$

7.2.3. Step 3 (Closest Point and SOS Function Simplification): Note that the i-th constraint, namely,

$$-\gamma_i\le\lambda^TX_i+\inf_j\left\{-\lambda^TZ_j+\left\|X_i-Z_j\right\|_2^2\right\},$$

can be written as

$$\begin{aligned}-\gamma_i&\le\inf_j\left\{-\lambda^T(Z_j-X_i)+\left\|X_i-Z_j\right\|_2^2\right\}\\&=-\left\|\lambda\right\|_2^2/4+\inf_j\left\{\left\|Z_j-(\lambda/2+X_i)\right\|_2^2\right\}\\&=-\left\|\lambda\right\|_2^2/4+T_i^{2/l}\left(n,\lambda\right)/n^{2/l}.\end{aligned}$$

However, since $\gamma_i\ge 0$ we must have that

$$-\gamma_i\le-\left\|\lambda\right\|_2^2/4+\min\left(T_i^{2/l}\left(n,\lambda\right)/n^{2/l},\left\|\lambda\right\|_2^2/4\right).$$

Therefore, the SOS profile function takes the form

$$R_n^W\left(\theta_*\right)=\max_{\lambda}\left\{-\lambda^T\bar{X}_n-\left\|\lambda\right\|_2^2/4+\frac{1}{n}\sum_{i=1}^{n}\min\left(\frac{T_i^{2/l}\left(n,\lambda\right)}{n^{2/l}},\left\|\lambda\right\|_2^2/4\right)\right\}.$$

To simplify the notation, let us redefine $\lambda \longleftarrow 2\lambda$ then we have that the simplified SOS profile function becomes:

$$R_n^W(\theta_*) = \max_\lambda \left\{ -2\lambda^T \bar{X}_n - \frac{1}{n} \sum_{i=1}^n \max\left(\|\lambda\|_2^2 - \frac{T_i^{2/l}(n,\lambda)}{n^{2/l}}, 0 \right) \right\}. \tag{27}$$

7.2.4. Step 4 (Case $l = 1$): When $l = 1$, let us 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:

$$\begin{aligned} nR_n^W(\theta_*) &= \max_\zeta \left\{ -2\zeta Z_n - \frac{1}{n} \sum_{i=1}^n \max\left(\zeta^2 - T_i^2\left(n, \zeta/\sqrt{n}\right) n^{-1}, 0 \right) \right\} \\ &= \max_\zeta \left\{ -2\zeta Z_n - \mathbb{E}\left[\max\left(\zeta^2 - T_i^2\left(n, \zeta/\sqrt{n}\right) n^{-1}, 0 \right) \right] \right\} + o_p(1) \\ &= \max_\zeta \left\{ -2\zeta Z_n - \mathbb{E}\left[\max\left(\zeta^2 - T_i^2(n,0)\, n^{-1}, 0 \right) \right] \right\} + o_p(1) \end{aligned}$$

The second equation follows the estimate in (Lemma 1). Using the bonded derivative for the density condition and first order Taylor expansion, we can prove that $\mathbb{E}[T_i^2(n,0)] - \mathbb{E}[T_i^2(n, \zeta/\sqrt{n})] \to 0$ as $n \to \infty$ for any fixed ζ. Since max function is Lipschitz continuous function with constant 1, and using the Dominating Convergence Theorem, the third equation above could be derived as

$$\begin{aligned} &\mathbb{E}\left[\max\left(\zeta^2 - T_i^2(n,0)\, n^{-1}, 0 \right) \right] - \mathbb{E}\left[\max\left(\zeta^2 - T_i^2\left(n, \zeta/\sqrt{n}\right) n^{-1}, 0 \right) \right] \\ \le &\mathbb{E}\left[\left| T_i^2\left(n, \zeta/\sqrt{n}\right) n^{-1} - T_i^2(n,0)\, n^{-1} \right| \right] = o_p(1). \end{aligned} \tag{28}$$

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.

$$G(\zeta, n) = \mathbb{E}\left[\max\left(\zeta^2 - T_i^2(n,0)\, n^{-1}, 0 \right) \right],$$

which is a deterministic function of ζ and for any n it is convex. Moreover, the derivative of $G(\zeta, n)$ is,

$$g(\zeta, n) = \nabla_\zeta G(\zeta, n) = 2\zeta \mathbb{P}\left(T_i^2(n,0) \le n\zeta^2 \right).$$

We need to notice that while taking the derivative we require exchanging the derivative and expectation, this can be done true hereby the Dominated Convergence Theorem since

$$\delta^{-1}\left|\max\left(\left(\zeta+\delta\right)^2 - T_i^2\left(n,0\right)n^{-1},0\right) - \max\left(\zeta^2 - T_i^2\left(n,0\right)n^{-1},0\right)\right| \leq 2|\zeta|,$$

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

$$Z_n = -\zeta P\left(T_i^2(n,0)\leq n\zeta^2\right) = -\zeta P\left(\tau^2(0)\leq n\zeta^2\right) + o_p(1) = -\zeta + o_p(1).$$

This estimate follows the second result of Lemma 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

$$nR_n^W\left(\theta_*\right) = 2Z_n^2 - G\left(Z_n,n\right) + o_p\left(1\right) \text{ as } n\to\infty.$$

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 that 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 $n\to\infty$,

$$\begin{aligned}
G\left(\zeta,n\right) &= \int_0^{\zeta^2} \mathbb{P}\left[T_i^2\left(n,0\right)\leq n\left(\zeta^2-t\right)\right]dt \\
&= \int_0^{\zeta^2} \mathbb{P}\left[\tau^2\left(0\right)\leq n\left(\zeta^2-t\right)\right]dt + o(1) \\
&= \int_0^{\zeta^2} 1dt + o(1) = \zeta^2 + o(1),
\end{aligned}$$

where the second equality is derived from the second argument of Lemma 2. Then for the SOS profile function, it becomes,

$$nR_n^W\left(\theta_*\right) = 2Z_n^2 - Z_n^2 + o_p(1) = Z_n^2 + o_p(1) \text{ as } n\to\infty.$$

Applying the continuous mapping theorem and the Central Limit Theorem for Z_n, we have

$$nR_n^W\left(\theta_*\right) \Rightarrow \sigma^2\chi_1^2.$$

7.2.5. Step 5 (Case $l=2$): Once again we introduce the substitution $\zeta=\sqrt{n}\lambda$ and $\sqrt{n}\bar{X}_n = Z_n$ into (27). Then, scaling the profile function by n, as $n\to\infty$ we have

$$\begin{aligned}
nR_n^W(\theta_*) &= \max_{\zeta}\left\{-2\zeta^T Z_n - \frac{1}{n}\sum_{i=1}^{n}\max\left(\|\zeta\|_2^2 - T_i\left(n,\zeta/\sqrt{n}\right),0\right)\right\}\\
&= \max_{\zeta}\left\{-2\zeta^T Z_n - \mathbb{E}\left[\max\left(\|\zeta\|_2^2 - T_i\left(n,\zeta/\sqrt{n}\right),0\right)\right]\right\} + o_p(1)\\
&= \max_{\zeta}\left\{-2\zeta^T Z_n - \mathbb{E}\left[\max\left(\|\zeta\|_2^2 - T_i\left(n,0\right),0\right)\right]\right\} + o_p(1), \qquad (29)
\end{aligned}$$

where the second equality is by applying Lemma 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), and the third equality is applying similar derivation as in (28),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

$$G(\zeta,n) = \mathbb{E}\left[\max\left(\|\zeta\|_2^2 - T_i(n,0),0\right)\right].$$

It is a continuous differentiable and convex function in ζ and with derivative equals

$$g(\zeta,n) = \nabla_\zeta G(\zeta,n) = 2\zeta\mathbb{P}\left[\|\zeta\|_2^2 \geq T_i(n,0)\right] = 2\zeta\mathbb{P}\left[\|\zeta\|_2^2\geq\tau(0)\right] + o(1) \text{ as } n\to\infty,$$

where the first equality requires applying the Dominated Convergence Theorem, as in the case $l=1$ and the second estimate follows the first argument in Lemma 2. Combining the above estimation, we have the first order optimality condition becomes

$$Z_n = -\zeta\mathbb{P}\left[\|\zeta\|_2^2\geq\tau(0)\right] + o_p(1) = -\zeta\tilde{g}(\zeta) + o_p(1) \text{ as } n\to\infty, \qquad (30)$$

where $\tilde{g}(\zeta(0)) = \mathbb{P}\left[\|\zeta\|_2^2\geq\tau\right]$ is a deterministic function of ζ. Using equation (30), 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 (30), 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

$$\frac{1}{\rho} = \tilde{g}\left(\rho \tilde{Z}\right). \tag{31}$$

Since the objective function is strictly 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 (31), 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

$$nR_n^W(\theta_*) = 2\rho\left(\left\|\tilde{Z}\right\|_2^2\right)\|Z_n\|_2^2 - G(\zeta_n^*, n) + o_p(1).$$

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

$$\begin{aligned}
G(\zeta, n) &= \int_0^{\|\zeta\|_2^2} \mathbb{P}\left[\|\zeta\|_2^2 - T_i(n,0) \geq t\right] dt \\
&= \int_0^{\|\zeta\|_2^2} \mathbb{P}\left[\|\zeta\|_2^2 - \tau(0) \geq t\right] dt + o(1) \\
&= \|\zeta\|_2^2 \int_0^1 \mathbb{P}\left[1 - \tau(0)/\|\zeta\|_2^2 \geq s\right] ds + o(1) \\
&= \|\zeta\|_2^2\, \mathbb{E}\left[\max\left(1 - \tau(0)/\|\zeta\|_2^2, 0\right)\right] + o(1) \\
&= \|\zeta\|_2^2\, \tilde{\eta}(\zeta) + o(1),
\end{aligned}$$

where we define $\tilde{\eta}(\zeta) = \mathbb{E}\left[\max\left(1 - \tau(0)/\|\zeta\|_2^2, 0\right)\right]$ is a deterministic continuous function of ζ. The second equation follows the first result of Lemma 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:

$$\begin{aligned}
nR_n^W(\theta_*) &= 2\rho\left(\tilde{Z}\right)\|Z_n\|_2^2 - \rho\left(\tilde{Z}\right)^2 \tilde{\eta}(Z_n)\|Z_n\|_2^2 + o_p(1) \\
&\Rightarrow 2\rho\left(\tilde{Z}\right)\left\|\tilde{Z}\right\|_2^2 - \rho\left(\tilde{Z}\right)^2 \tilde{\eta}\left(\tilde{Z}\right)\left\|\tilde{Z}\right\|_2^2.
\end{aligned}$$

7.2.6. 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

$$\begin{aligned}
&n^{1/2+\frac{3}{2l+2}} R_n^W(\theta_*) \\
&= \max_{\zeta}\left\{-2\zeta^T Z_n - n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)}\frac{1}{n}\sum_{i=1}^{n}\max\left(\left\|\frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}}\right\|_2^2 - T_i^{2/l}\left(n, \zeta/n^{\frac{3}{2l+2}}\right), 0\right)\right\} \\
&= \max_{\zeta}\left\{-2\zeta^T Z_n - n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)}\mathbb{E}\left[\max\left(\left\|\frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}}\right\|_2^2 - T_1^{2/l}\left(n, \zeta/n^{\frac{3}{2l+2}}\right), 0\right)\right]\right\} + o_p(1) \\
&= \max_{\zeta}\left\{-2\zeta^T Z_n - n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)}\mathbb{E}\left[\max\left(\left\|\frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}}\right\|_2^2 - T_1^{2/l}(n,0), 0\right)\right]\right\} + o_p(1).
\end{aligned}$$

The estimate in second equation the previous display is due to an application of Lemma 1, and the third equation follows the similar derivation as in (28). Similar as for the lower dimensional case, let us denote

$$G(\zeta, n) = n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)}\mathbb{E}\left[\max\left(\left\|\frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}}\right\|_2^2 - T_1^{2/l}(n,0), 0\right)\right],$$

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 Dominated Convergence Theorem, as we discussed for $l = 1$ and the gradient of $G(\zeta, n)$ has the following estimate as $n \to \infty$,

$$\begin{aligned}
g(\zeta, n) = \nabla_{\zeta} G(\zeta, n) &= 2n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)}\zeta\mathbb{P}\left[T_i(n,0) \leq \left\|\zeta n^{-\left(\frac{3}{2l+2}-\frac{1}{l}\right)}\right\|_2^l\right] \\
&= 2n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)}\zeta\mathbb{P}\left[\tau(0) \leq \left\|\zeta n^{-\left(\frac{3}{2l+2}-\frac{1}{l}\right)}\right\|_2^l\right] + o(1).
\end{aligned}$$

The second equality estimate is considering ζ within a compact set and the derivation follows the first argument in Lemma 2. Then the first order optimality condition for the SOS profile function becomes,

$$Z_n = -n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)}\zeta\mathbb{P}\left[\tau(0) \leq \left\|\zeta n^{-\left(\frac{3}{2l+2}-\frac{1}{l}\right)}\right\|_2^l\right] + o(1) \text{ as } n \to \infty.$$

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

$$\begin{aligned}
&\mathbb{P}\left[\tau(0) \le \|\kappa_n\|_2^l\right] = 1-\mathbb{P}\left[\tau(0) > \|\kappa_n\|_2^l\right] = 1-\mathbb{E}\left[\mathbb{P}\left[\tau(0) > \|\kappa_n\|_2^l \;\middle|\; X_1\right]\right]\\
&= \mathbb{E}\left[1-\exp\left(-\frac{\pi^{l/2}\left(f_X(X_1)+\kappa f_Y(X_1)\right)}{\Gamma(l/2+1)}\|\kappa_n\|_2^l\right)\right]\\
&= \mathbb{E}\left[\frac{\pi^{l/2}}{\Gamma(l/2+1)}\left[f_X(X_1)+\kappa f_Y(X_1)\right]\|\kappa_n\|_2^l\right] + o_p\left(n^{-\left(\frac{3l}{2l+2}-1\right)}\right)\\
&= C\|\kappa_n\|_2^l + o_p\left(n^{-\left(\frac{3l}{2l+2}-1\right)}\right),
\end{aligned}$$

where we denote

$$C = \frac{\pi^{l/2}}{\Gamma(l/2+1)}\mathbb{E}\left[f_X(X_1)+\kappa f_Y(X_1)\right].$$

Plug it back into the optimizer, and as $n\to\infty$ we have:

$$Z_n = -Cn^{(1/2-\frac{3}{2l+2})}n^{\left(-\frac{3l}{2l+2}+1\right)}\zeta\|\zeta\|_2^l + o_p(1) = -C\zeta\|\zeta\|_2^l + o_p(1).$$

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 $\zeta_n^* = -c_* Z_n(1+o(1))$, 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:

$$n^{1/2+\frac{3}{2l+2}}R_n^W(\theta_*) = -2\zeta_n^{*\,T}Z_n - G(\zeta_n^*, n) + o_p(1).$$

Let us focus on the analysis of $G(\zeta,n)$ 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.

$$\mathbb{E}\left[\max\left(\|\kappa_n\|_2^2 - T_1^{2/l}(n,0),0\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\max\left(\|\kappa_n\|_2^2 - T_1^{2/l}(n,0),0\right) \;\middle|\; X_1\right]\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\int_0^{\kappa_n} \mathbb{P}\left[T_1(n,0) \le (\kappa_n - t)^{l/2} \;\middle|\; X_1\right] dt\right] \\
&= \mathbb{E}\left[\int_0^{\|\kappa_n\|_2^2} \mathbb{P}\left[\tau(0) \le \left(\|\kappa_n\|_2^2 - t\right)^{l/2} \;\middle|\; X_1\right] + O\left(1/n^{-1/2+1/l}\right) dt\right] \\
&= \mathbb{E}\left[\int_0^{\|\kappa_n\|_2^2} \left(1 - \exp\left(-\frac{\pi^{l/2}\left(f_X(X_1) + \kappa f_Y(X_1)\right)}{\Gamma(l/2+1)} \left(\|\kappa_n\|_2^2 - t\right)^{l/2}\right)\right) dt\right] \\
&\qquad + O\left(1/n^{-1/2+3/l-\frac{6}{2l+2}}\right) \\
&= C\frac{2}{l+2}\left\|\frac{\zeta}{n^{\left(\frac{3}{2l+2}-\frac{1}{l}\right)}}\right\|^{l+2} + O\left(1/n^{-1/2+3/l-\frac{6}{2l+2}}\right)
\end{aligned}
$$

The estimate in third equation follows by applying the first argument in Lemma 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

$$
\begin{aligned}
& n^{1/2+\frac{5}{4l+2}} R_n^W(\theta_*) \\
&= 2\|Z_n\|_2^2 \frac{\|Z_n\|_2^{\left(\frac{1}{l+1}-1\right)}}{C^{\frac{1}{l+1}}} - \frac{2}{l+2}\frac{\|Z_n\|_2^{1+\frac{1}{l+1}}}{C^{\frac{1}{l+1}}} + o_p(1) \\
&= \frac{2l+2}{l+2}\frac{\|Z_n\|_2^{1+\frac{1}{l+1}}}{C^{\frac{1}{l+1}}} + o_p(1) \Rightarrow \frac{2l+2}{l+2}\frac{\left\|\tilde{Z}\right\|_2^{1+\frac{1}{l+1}}}{C^{\frac{1}{l+1}}}.
\end{aligned}
$$

7.2.7. Proofs of Technical Lemmas in Step 2

[**Proof of Lemma 1**]We shall introduce some notation which will be convenient throughout our development. Define for $t \ge 0$,

$$
\begin{aligned}
&F_n(t) = P\left(T_i(n,\lambda) \le t\right), \\
&D_i(t) = I\left(T_i(n,\lambda) \le t\right), \quad \bar{D}_i(t) = I\left(T_i(n,\lambda) \le t\right) - F_n(t), \\
&\bar{F}_n(t) = 1 + n^{-1/2}\sum_{i=1}^{n} \bar{D}_i(t).
\end{aligned}
$$

Therefore, we are interested in studying

$$\bar{F}_n(t) - 1 = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left(I\left(T_i(n,\lambda) \le t\right) - F_n(t) \right).$$

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

$$\begin{aligned} g^*(t) &= \sup\{g(s) : s \in [0,t]\}, \\ [g](t) &= \int_0^t (dg(s))^2. \end{aligned}$$

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,

$$[\bar{F}_n](t) = \frac{1}{n} \sum_{i=1}^{n} I\left(T_i(n,\lambda) \le t\right).$$

We observe that $\bar{F}_n^*(t) \ge 1$, therefore $h_n(t)$ is well defined; moreover, note that

$$h_n(t)^2 \le 1.$$

We invoke Theorem 1.2 of Beiglbck and Siorpaes (2015) and conclude that

$$\sup_{0 \le t \le T} \bar{F}_n(t) \le 6\sqrt{[\bar{F}_n](T)} + 2\int_0^T h_n(t)\, d\bar{F}_n(t).$$

Now we analyze the integral in the right hand side of the previous display. Observe that

$$\begin{aligned} \mathbb{E}\left(\int_0^T h_n(t)\, d\bar{F}_n(t) \right) &= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \mathbb{E}\left(\int_0^T h_n(t)\, d\bar{D}_i(t) \right) \\ &= n^{1/2} \mathbb{E}\left(\int_0^T h_n(t)\, d\bar{D}_1(t) \right). \end{aligned} \tag{32}$$

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,

$$ \left|{}_1\bar{F}_n^*(t) - \bar{F}_n^*(t)\right| \le 1/n^{1/2}. $$

We then can write

$$ \begin{aligned} h_n(t) &= \frac{{}_1\bar{F}_n(t_-) + \bar{D}_1(t_-)/n^{1/2}}{\left(\bar{F}_n^*(t_-)^2 + [{}_1\bar{F}_n](t_-) + [D_1](t_-)/n\right)^{1/2}} \\ &= \frac{{}_1\bar{F}_n(t_-)}{\left({}_1\bar{F}_n^*(t_-)^2 + [{}_1\bar{F}_n](t_-)\right)^{1/2}}\left(1 + \frac{L_n(t_-)}{n^{1/2}}\right), \end{aligned} \tag{33} $$

where we can select a deterministic constant $c \in (0,\infty)$ such that $|L_n(t)| \le c$ for $j = 0$ and 1 assuming $n \ge 4$ (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 (33) into (32) 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 $\{\bar{D}_i(t) : t \le T\}$ 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 $\left|\bar{L}_{N_n(T,\lambda,1)}\right| \le c(1 + N_n(T,\lambda,1))$ for some (deterministic) $c > 0$ and $n \ge 4$ and satisfying

$$ \mathbb{E}\left(\int_0^T \bar{h}_n(t)\, d\bar{D}_1(t)\right) $$

$$
\begin{aligned}
&= \mathbb{E}\left(\bar{h}_n\left(T_1\left(n\right)\right) I\left(T_1\left(n\right) \leq T\right)\right) - \mathbb{E}\left(\widetilde{h}_n\left(T_1\left(n\right)\right) I\left(T_1\left(n\right) \leq T\right)\right) \\
&= \mathbb{E}\left(\widetilde{h}_n\left(T_1\left(n\right)\right) I\left(T_1\left(n\right) \leq T\right)\right) - \mathbb{E}\left(\widetilde{h}_n\left(\tau_i\left(X_i\right)\right) I\left(\tau_i\left(X_i\right) \leq T\right)\right) \\
&+ \mathbb{E}\left(\bar{L}_{N_n(T,\lambda,1)}/n^{1/2}\right) \\
&= \mathbb{E}\left(\bar{L}_{N_n(T,\lambda,1)}/n^{1/2}\right).
\end{aligned}
$$

We have that

$$
\left|\mathbb{E}\left(\bar{L}_{N_n(T,\lambda,1)}/n^{1/2}\right)\right| \leq \left|\mathbb{E}\left(c(1+N_n\left(T,\lambda,1\right))\right)\right|/n^{1/2} = O\left(1/n^{1/2}\right).
$$

Consequently, we conclude that

$$
n^{1/2}\mathbb{E}\left(\int_0^T h_n\left(t\right) d\bar{D}_1\left(t\right)\right) = O\left(1\right),
$$

as $n \to \infty$, as required. Thus we proved that the first part of the lemma holds. For the second part, we observe that

$$
\begin{aligned}
&\overline{\lim}_{n\to\infty}\mathbb{E}\left(\sup_{t\in[0,T]}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(\max\left(t^2 - T_i(n,\lambda)^{\alpha}, 0\right) - \mathbb{E}\left[\max\left(t^2 - T_i(n,\lambda)^{\alpha}, 0\right)\right]\right)\right|\right) \\
&= \overline{\lim}_{n\to\infty}\mathbb{E}\left(\sup_{t\in[0,T]}\left|\int_0^t \frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(2sI\left(T_i^{\alpha}\left(n,\lambda\right) \leq s^2\right) - 2s\mathbb{P}[T_i^{\alpha}\left(n,\lambda\right) \leq s^2]\right) ds\right|\right) \\
&\leq \overline{\lim}_{n\to\infty}\int_0^T \mathbb{E}\left(\sup_{t\in[0,T]}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(2tI\left(T_i^{\alpha}\left(n,\lambda\right) \leq t^2\right) - 2t\mathbb{P}[T_i^{\alpha}\left(n,\lambda\right) \leq t^2]\right)\right|\right) dt \\
&\leq 2T^2\overline{\lim}_{n\to\infty}\mathbb{E}\left(\sup_{t\in[0,T]}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(I\left(T_i\left(n,\lambda\right) \leq t\right) - \mathbb{P}[T_i\left(n,\lambda\right) \leq t]\right)\right|\right) < \infty.
\end{aligned}
$$

Hence, applying the result for the first part of the lemma, we conclude the second part as well.

[**Proof of Lemma 2**]

$$
\begin{aligned}
\mathbb{P}\left[T_i\left(n,\lambda\right) \leq t\right] &= \mathbb{P}\left(Bin\left(\mathbb{P}\left(\left\|X_i - a\left(X_i,\lambda\right)\right\|_2 \leq t^{1/l}/n^{1/l}\right), n-1\right) \geq 1\right) \\
&= 1 - \left(1 - \mathbb{P}\left(\left\|X_i - a\left(X_i,\lambda\right)\right\|_2 \leq t^{1/l}/n^{1/l}\right)\right)^n.
\end{aligned}
$$

Then, as $n \to \infty$ and $t \to 0^+$

$$
\mathbb{P}\left(\left\|X_j - a\left(X_i,\lambda\right)\right\|_2 \leq t^{1/l}/n^{1/l}\right) = c_0 t/n + c_1 t/n \cdot t^{1/l}/n^{1/l} + o\left(t^{1+1/l}/n^{1+1/l}\right).
$$

Therefore by the Poisson approximation to the Binomial distribution we know:

$$\mathbb{P}\left[T_i\left(n,\lambda\right)\le t\right]=1-\exp\left(-c_0t\right)+O\left(t^{1+1/l}/n^{1/l}\right),$$

$$\mathbb{P}\left[\tau(\lambda)\le t\right]=1-\exp\left(-c_0t\right).$$

Thus we proved the first claim:

$$\mathbb{P}\left[T_i\left(n,\lambda\right)\le t\right]-\mathbb{P}\left[\tau(\lambda)\le t\right]=O\left(t^{1+1/l}/n^{1/l}\right).$$

The second claim follows the definition of τ and equation (26), where as $n\to\infty$ we have

$$\begin{aligned}&\mathbb{P}\left[T_i\left(n,\lambda\right)\le t\right]-\mathbb{P}\left[\tau(\lambda)\le t\right]=\mathbb{P}\left[T_i\left(n,\lambda\right)>t\right]-\mathbb{P}\left[\tau(\lambda)>t\right]\\&=\mathbb{E}\left[\exp\left(-\Lambda\left(\lambda,X_1\right)\right)\right]\left(1+O\left(1/n^l\right)\right)-\mathbb{E}\left[\exp\left(-\Lambda\left(\lambda,X_1\right)\right)\right]\\&=\mathbb{P}\left[\tau(\lambda)>t\right]O\left(1/n^l\right).\end{aligned}$$

7.3. Proofs of Additional Theorems

In this subsection, we are going to provide the proofs of the remaining theorems and corollaries (Theorem 2, Theorem 3, Corollary 1 and Corollary 2). We are going to follow closely the proof of Theorem 1 and discuss the differences inside each of its steps.

7.3.1. Proofs of SOS Theorems for General Estimation We will first prove the corresponding theorems for general estimating equations. As we discussed before, Theorem 2 is the direct generalization of Theorem 1 and we are going to only discuss the proof of Theorem 3 in this part.

[**Proof of Theorem 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

$$R_n^W\left(\theta_*\right)=\max_{\lambda}\left\{-\lambda^T\bar{h}_n\left(\theta_*\right)-\frac{1}{n}\sum_{i=1}^{n}\max_j\left\{\lambda^Th\left(\theta_*,Z_j\right)-\lambda^Th\left(\theta_*,X_i\right)-\|X_i-Z_j\|_2^2\right\}^+\right\}.$$

Step 2 and Step 3 are given as follows, for $l=1$ and $l=2$, let us denote $\sqrt{n}\bar{h}_n\left(\theta_*\right)=Z_n$ and $\sqrt{n}\lambda=2\zeta$, we can scale the SOS profile function by n, arriving to

$$nR_n^W\left(\theta_*\right)=\max_{\zeta}\left\{-2\zeta^TZ_n-\frac{1}{n}\sum_{i=1}^{n}n\max_j\left\{2\frac{\zeta^T}{\sqrt{n}}h\left(\theta_*,Z_j\right)-2\frac{\zeta^T}{\sqrt{n}}h\left(\theta_*,X_i\right)-\|X_i-Z_j\|_2^2\right\}^+\right\}.$$

For each i, let us consider the maximization problem

$$\max_j \left\{ 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, Z_j) - 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, X_i) - \|X_i - Z_j\|_2^2 \right\}. \tag{34}$$

Similar as Step 1 of the proof for Theorem 1, we would like to solve the maximization problem (34) 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

$$\max_z \left\{ 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, z) - 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, X_i) - \|X_i - z\|_2^2 \right\}, \tag{35}$$

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 (35) satisfies

$$\begin{aligned} \bar{a}_*(X_i,\zeta) &= X_i + D_x h(\theta_*, \bar{a}_*(X_i,\zeta))^T \cdot \frac{\zeta}{n^{1/2}} \\ &= X_i + D_x h(\theta_*, X_i)^T \cdot \frac{\zeta}{n^{1/2}} + O\left(\frac{\|\zeta\|_2^2}{n} \|D_x h(\theta_*, \bar{a}_*(X_i,\zeta))\|_2 \right). \end{aligned} \tag{36}$$

Moreover, owing to BE1), we obtain that

$$\|D_x h(\theta_*, \bar{a}_*(X_i,\zeta)) - D_x h(\theta_*, X_i)\|_2 \le \tilde{K} \frac{\|\zeta\|_2}{n^{1/2}}. \tag{37}$$

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 (36) and (37) 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 (37) 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 $n \to \infty$

$$
\begin{aligned}
-\frac{\zeta^T V_i \zeta}{n} &= 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, X_i) - 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, a_*(X_i, \zeta)) + \|X_i - a_*(X_i, \zeta)\|_2^2 \\
&\quad + O\left(\frac{\|D_x h(\theta_*, X_i)\|^2 \|\zeta\|^3}{n^{3/2}}\right).
\end{aligned}
$$

Similarly as in Step 2 of the proof of Theorem 1 we can define the point process $N^{(i)}(t, \zeta)$ and $T_i(n, \lambda)$. We know the gap between freeing the variable z and restricting the maximization over the Z_j's (i.e. the difference between (35) and (34)) is

$$
\begin{aligned}
&\max_j \left\{ \frac{1}{n}\zeta^T V_i \zeta - \left(2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, Z_j) - 2\frac{\zeta^T}{\sqrt{n}} h(\theta_*, X_i) - \|X_i - Z_j\|_2^2 \right) \right\} \\
&+ O\left(\frac{\|D_x h(\theta_*, X_i)\|^2 \|\zeta\|^3}{n^{3/2}}\right).
\end{aligned}
$$

By the definition of $T_i(n, \lambda)$, we can write the profile function for $l = 1$ as

$$
\begin{aligned}
&nR_n^W(\theta_*) \\
&= \max_\zeta \left\{ -2\zeta^T Z_n - \frac{1}{n}\sum_{i=1}^n \max\left(\zeta^T V_i \zeta - \frac{T_i^2(n, \lambda)}{n} + O\left(\frac{\|D_x h(\theta_*, X_i)\|^2 \|\zeta\|^3}{n^{1/2}}\right), 0 \right) \right\}.
\end{aligned}
$$

Note that the sequence of global optimizers is tight as $n \to \infty$ because $\mathbb{E}(V_i)$ 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 1 (invoking Lemma 1) and using the fact that ζ can be restricted to compact sets, that as $n \to \infty$

$$
nR_n^W(\theta_*) = \max_\zeta \left\{ -2\zeta^T Z_n - \mathbb{E}\left[\max\left(\zeta^T V_1 \zeta - \frac{T_1^2(n, \lambda)}{n} \right) \right] \right\} + o_p(1).
$$

Then, for $l = 2$, as $n \to \infty$ we have estimate for the profile function as

$$
nR_n^W(\theta_*) = \max_\zeta \left\{ -2\zeta^T Z_n - \mathbb{E}\left[\max\left(\zeta^T V_1 \zeta - T_1^2(n, \lambda) \right) \right] \right\} + o_p(1).
$$

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

$$
\begin{aligned}
&n^{\frac{1}{2}+\frac{3}{2l+2}} R_n^W(\theta_*) \\
&= \max_\zeta \left\{ -2\zeta^T Z_n - \frac{1}{n}\sum_{i=1}^n n^{\frac{1}{2}+\frac{3}{2l+2}} \max_j \left\{ 2\frac{\zeta^T}{n^{\frac{3}{2l+2}}} h(\theta_*, Z_j) - 2\frac{\zeta^T}{n^{\frac{3}{2l+2}}} h(\theta_*, X_i) - \|X_i - Z_j\|_2^2 \right\}^+ \right\}.
\end{aligned}
$$

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 1. As $n\to\infty$, we have the estimate for profile function becomes

$$\begin{aligned} & n^{\frac{1}{2}+\frac{3}{2l+2}}R_n^W(\theta_*) \\ &= \max_{\zeta}\left\{-2\zeta^T Z_n - n^{\frac{1}{2}+\frac{3}{2l+2}-\frac{2}{l}}\frac{1}{n}\sum_{i=1}^{n}\max\left(n^{-\left(\frac{6}{2l+2}-\frac{2}{l}\right)}\zeta^T V_i\zeta - T_i^{2/l}(n,\lambda),0\right)\right\} + o_p(1) \\ &= \max_{\zeta}\left\{-2\zeta^T Z_n - n^{\frac{1}{2}+\frac{3}{2l+2}-\frac{2}{l}}\mathbb{E}\left[\max\left(n^{-\left(\frac{6}{2l+2}-\frac{2}{l}\right)}\zeta^T V_1\zeta - T_1^{2/l}(n,\lambda),0\right)\right\}\right] + o_p(1). \end{aligned}$$

The final estimation follows as in the proof for Theorem 1 (i.e. applying Lemma 1).

In **Step 4** for $l=1$, as $n\to\infty$ the objective function is

$$nR_n^W(\theta_*) = \max_{\zeta}\left\{-2\zeta^T Z_n(\theta_*) - \mathbb{E}\left[\max\left(\zeta^T V_1\zeta - \frac{T_1^2(n,\lambda)}{n},0\right)\right]\right\} + o_p(1).$$

Let us denote $G:\mathbb{R}^l\to\mathbb{R}$ to be a deterministic continuous function, defined as

$$G(\zeta,n) = \mathbb{E}\left[\max\left(\zeta^T V_1\zeta - \frac{T_1^2(n,\lambda)}{n},0\right)\right].$$

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

$$Z_n = -\Upsilon\zeta + o_p(1) \text{ as } n\to\infty. \tag{38}$$

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

$$nR_n^W(\theta_*) = 2Z_n^T\Upsilon^{-1}Z_n - G\left(-\Upsilon^{-1}Z_n,n\right) + o_p(1).$$

As $n\to\infty$, we can apply the same estimation in the proof of Theorem 1, it becomes

$$nR_n^W(\theta_*) \Rightarrow \tilde{Z}^T\Upsilon^{-1}\tilde{Z}.$$

Thus we proof the claim for $l=1$.

In **Step 5** for $l=2$, as $n\to\infty$ the objective function has estimate

$$nR_n^W(\theta_*)=\max_{\zeta}\left\{-2\zeta^T Z_n(\theta_*)-\mathbb{E}\left[\max\left(\zeta^T V_1\zeta-T_1(n,\lambda),0\right)\right]\right\}+o_p(1).$$

Still, we denote $G(\zeta,n)$ to be a deterministic function given as,

$$G(\zeta,n)=\mathbb{E}\left[\max\left(\zeta^T V_1\zeta-T_1(n,\lambda),0\right)\right].$$

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 1 the first order optimality condition becomes

$$Z_n=-\mathbb{E}\left[V_1 1_{\left(\tau(0)\le\zeta^T V_1\zeta\right)}\right]\zeta+o_p(1)\text{ as }n\to\infty. \tag{39}$$

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 (39). We plug in the optimizer and the objective function becomes

$$nR_n^W(\theta_*)=-2Z_n^T\zeta_n^*-G(\zeta_n^*,n)+o_p(1)\text{ as }n\to\infty.$$

As $n\to\infty$, we can apply the same estimation in the proof of Theorem 1, we have

$$nR_n^W(\theta_*)\Rightarrow-2\tilde{Z}^T\tilde{\zeta}-\tilde{\zeta}^T\tilde{G}\left(\tilde{\zeta}\right)\tilde{\zeta},$$

where $\tilde{G}:\mathbb{R}^q\to\mathbb{R}^q\times\mathbb{R}^q$ is a deterministic continuous mapping defined as,

$$\tilde{G}(\zeta)=\mathbb{E}\left[V_1\max\left(1-\tau(0)/\left(\zeta^T V_1\zeta\right),0\right)\right],$$

and $\tilde{\zeta}:=\tilde{\zeta}\left(\tilde{Z}\right)$ is the unique solution to

$$\tilde{Z}=-\zeta\mathbb{E}\left[V_1 1_{\left(\tau(0)\le\zeta^T V_1\zeta\right)}\right].$$

Then we proved the claim for $l=2$.

Finally, in **Step 6** for $l \geq 3$, as $n \to \infty$ the objective function is

$$\begin{aligned} & n^{1/2+\frac{3}{2l+2}} R_n^W(\theta_*) \\ & = \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} \mathbb{E}\left[\max\left(n^{-\left(\frac{6}{2l+2}-\frac{2}{l}\right)} \zeta^T V_1 \zeta - T_1^{2/l}(n,\lambda), 0 \right) \right] \right\} + o_p(1). \end{aligned}$$

We denote $G(\zeta, n)$ to be a deterministic function defined as,

$$G(\zeta, n) = n^{\left(1/2+\frac{3}{2l+2}-\frac{2}{l}\right)} \mathbb{E}\left[\max\left(n^{-\left(\frac{6}{2l+2}-\frac{2}{l}\right)} \zeta^T V_1 \zeta - T_1^{2/l}(n,\lambda), 0 \right) \right].$$

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 1, the first order optimality condition becomes

$$Z_n = -\mathbb{E}\left[V_1 \frac{\pi^{l/2} (f_X(X_1) + \kappa f_Y(X_1))}{\Gamma(l/2+1)} V_1 \left(\zeta^T V_1 \zeta\right)^l \right] \zeta + o_p(1). \text{ as } n \to \infty \tag{40}$$

The objective condition is strictly convex differentiable and for fixed Z_n there is a unique ζ_n^* satisfying the first optimality condition (40). We plug ζ_n^* into the objective function and it becomes

$$n^{1/2+\frac{3}{2l+2}} R_n^W(\theta_*) = -2Z_n^T \zeta_n^* - G(\zeta_n^*, n) + o_p(1) \text{ as } n \to \infty.$$

As $n \to \infty$, we can apply same estimate in the proof of Theorem 1, we have

$$n^{1/2+\frac{3}{2l+2}} R_n^W(\theta_*) \Rightarrow -2\tilde{Z}^T \tilde{\zeta} - \frac{2}{l+2} \tilde{G}\left(\tilde{\zeta}\right),$$

where $\tilde{G} : \mathbb{R}^q \to \mathbb{R}$ is a deterministic continuous function given as,

$$\tilde{G}(\zeta) = \mathbb{E}\left[\frac{\pi^{l/2} (f_X(X_1) + \kappa f_Y(X_1))}{\Gamma(l/2+1)} \left(\zeta^T V_1 \zeta\right)^{l/2+1} \right],$$

and $\tilde{\zeta} := \tilde{\zeta}\left(\tilde{Z}\right)$ is the unique solution to

$$\tilde{Z} = -\mathbb{E}\left[V_1 \frac{\pi^{l/2} (f_X(X_1) + \kappa f_Y(X_1))}{\Gamma(l/2+1)} V_1 \left(\zeta^T V_1 \zeta\right)^l \right] \zeta.$$

We proved the claim for $l \geq 3$ and finish the proof for Theorem 3.

7.3.2. 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 1, we are going to discuss the different steps here.

[**Proof of Corollary 1**]For implicit formulation, as we discussed for Theorem 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

$$R_n^W(\gamma_*) = \max_{\lambda,\gamma_i \geq 0} \left\{ -\lambda \bar{X}_n - \frac{1}{n}\sum_{i=1}^n \gamma_i \right\}$$
$$\text{s.t. } -\gamma_i \leq \min_j \left\{ \lambda^T X_i - \lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\}, \text{ for all } i.$$

We can apply first order Taylor expansion to $h(\gamma_*, \nu_n, X_i)$ w.r.t. ν, then we have

$$h(\gamma_*, \nu_n, X_i) = h(\gamma_*, \nu_*, X_i) + O_p\left(\frac{\|D_\nu h(\gamma_*, \bar{\nu}_n, X_i)\|}{n^{1/2}} \right),$$

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

$$Z_k = Z_k(*) + \epsilon_n(Z_k),$$

where $\epsilon_n(Z_k) = O_p\left(\|D_\nu h(\gamma_*, \bar{\nu}_n, Z_k)\| / n^{1/2} \right)$.

In **Step 2** we can define a point process $N_n^{(i)}(t,\lambda)$ and $T_i(n)$ as in the proof of Theorem 1, but the rate becomes

$$\Lambda(X_i, \lambda) = \left[f_X(X_i + \lambda/2 + \epsilon_n(X_i)) + \kappa f_Y(X_i + \lambda/2 + \epsilon_n(X_i)) \right] \frac{\pi^{l/2}}{\Gamma(l/2+1)}.$$

As $n \to \infty$, same as in the proof of Theorem 1 and Theorem 3 we can argue $\lambda \to 0$. Then we can define $\tau(0)$ same as in the proof of Theorem 1 and has the with same distribution

$$\mathbb{P}[\tau(0) \geq t] = \mathbb{E}\left[\exp\left(-\left(f_X(X_1) + \kappa f_Y(X_1)\right) \frac{\pi^{l/2}}{\Gamma(l/2+1)} \right) \right].$$

Then the rest of the proof in **Step 3, 4, 5 and 6** stay the same as that of Theorem 1, but replacing the CLT for Z_n by the asymptotic distribution given in C2).

[**Proof of Corollary 2**]For explicit formulation, the proof follows more closely the proof of Theorem 3 and we are discussing the differences as follows.

In **Step 1**, the dual formulation takes the form

$$\begin{aligned} & R_n^W(\theta_*) \\ & = \max_{\lambda} \left\{ -\lambda^T \bar{h}_n(\gamma_*, \nu_n) - \frac{1}{n} \sum_{i=1}^{n} \max_{j} \left\{ \lambda^T h(\gamma_*, \nu_n, Z_j) - \lambda^T h(\gamma_*, \nu_n, X_i) - \|X_i - Z_j\|_2^2 \right\}^+ \right\}. \end{aligned}$$

Step 2 and Step 3 Follows the same as for the proof of Theorem 3 however we need to notice that difference is the definition of $\bar{a}_*(X_i, \zeta)$, for $l = 1$ and 2 we have

$$\begin{aligned} \bar{a}_*(X_i, \zeta) &= X_i + D_x h(\gamma_*, \nu_n, \bar{a}_*(X_i, \zeta)) \cdot \frac{\zeta}{n^{1/2}} \quad (41) \\ &= X_i + D_x h(\gamma_*, \nu_n, X_i) \cdot \frac{\zeta}{n^{1/2}} + O\left(\frac{\|\zeta\|_2^2}{n} \|D_x h(\gamma_*, \nu_n, \bar{a}_*(X_i, \zeta))\|_2 \right) \\ &= X_i + D_x h(\gamma_*, \nu_*, X_i) \cdot \frac{\zeta}{n^{1/2}} + O\left(\frac{\|\zeta\|_2^2}{n} \|D_x h(\gamma_*, \nu_n, \bar{a}_*(X_i, \zeta))\|_2 \right) \\ &+ O\left(\frac{\|\zeta\|_2}{n^{1/2}} \|\nu_n - \nu_*\|_2 \|D_x h(\gamma_*, \nu_n, \bar{a}_*(X_i, \zeta))\|_2 \|D_\nu D_x h(\gamma_*, \bar{\nu}_n, \bar{a}_*(X_i, \zeta))\|_2 \right), \quad (42) \end{aligned}$$

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. 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).

Acknowledgments

The authors gratefully acknowledge support from the following NSF grants 1915967, 1820942, 1838676, as well as support from the Ford Motors Company. The authors also are grateful to the editorial team and the referees for their careful review of the paper and their constructive suggestions, which greatly helped to improve it. The authors also thank Karthyek Murthy and Henry Lam for helpful discussions.

References

Antoine B, Bonnal H, Renault E (2007) On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood. *Journal of Econometrics* 138(2):461–487.

Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. *The Operations Research Revolution*, 1–19 (INFORMS).

Beiglbck M, Siorpaes P (2015) Pathwise versions of the BurkholderDavisGundy inequality. *Bernoulli* 21(1):360–373, ISSN 1350-7265, URL `http://dx.doi.org/10.3150/13-BEJ570`.

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.

Blake C, Merz CJ (1998) Uci repository of machine learning databases.

Blanchet J, Kang Y (2017) Distributionally robust groupwise regularization estimator. *Proceedings of Machine Learning Research* 77:97112.

Blanchet J, Kang Y (2018) Semi-supervised learning based on distributionally robust optimization. *Stochastic Modeling Techniques and Data Analysis International Conference*, 87–116.

Blanchet J, Kang Y, Lam H (2019a) Quantify uncertainty for stochastic programming problem: an empirical likelihood approach. Manuscript.

Blanchet J, Kang Y, Murthy K (2019b) Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* 56(3):830–857.

Blanchet J, Kang Y, Zhang F, He F, Hu Z (2019c) Doubly robust data-driven distributionally robust optimization. *Proceedings of 18th Applied Stochastic Models and Data Analysis International Conference with Demographics Workshop*, 249–262.

Blanchet J, Kang Y, Zhang F, Hu Z (2019d) A distributionally robust boosting algorithm. *Proceedings of the 2019 Winter Simulation Conference (WSC)*, 3728–3739.

Blanchet J, Kang Y, Zhang F, Murthy K (2019e) Data-driven optimal transport cost selection for distributionally robust optimization. *Proceedings of the 2019 Winter Simulation Conference (WSC)*, 3740–3751.

Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research* 44(2):565–600.

Bravo F (2004) Empirical likelihood based inference with applications to some econometric models. *Econometric Theory* 20(02):231–264.

Chen SX (1993) On the accuracy of empirical likelihood confidence regions for linear regression model. *Annals of the Institute of Statistical Mathematics* 45(4):621–637.

Chen SX, Keilegom IV (2009) A review on empirical likelihood methods for regression. *TEST* 18(3):415–447, ISSN 1133-0686, 1863-8260, URL `http://dx.doi.org/10.1007/s11749-009-0159-5`.

Chen Z, Kuhn D, Wiesemann W (2018) Data-driven chance constrained programs over wasserstein balls. *arXiv preprint arXiv:1809.00210* .

Duchi J, Glynn P, Namkoong H (2016) Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425* .

Duchi J, Namkoong H (2018) Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750* .

Einmahl JH, McKeague IW (1999) Confidence tubes for multiple quantile plots via empirical likelihood. *The Annals of Statistics* 27(4):1348–1367.

Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171(1-2):115–166.

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199* .

Gao R, Xie L, Xie Y, Xu H (2018) Robust hypothesis testing using Wasserstein uncertainty sets. *arXiv preprint arXiv:1805.10611* .

Guggenberger P (2008) Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator. *Econometric Reviews* 27(4-6):526–541.

Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* 27(2):83–85.

Hollander M, McKeague IW (1997) Likelihood ratio-based confidence bands for survival functions. *Journal of the American Statistical Association* 92(437):215–226.

Hu W, Niu G, Sato I, Sugiyama M (2018) Does distributionally robust supervised learning give robust classifiers? *International Conference on Machine Learning*, 2034–2042.

Imbens GW (2012) Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics* .

Kitamura Y (2006) Empirical likelihood methods in econometrics: Theory and practice. *Cowles Foundation discussion paper* .

Lam H, Zhou E (2015) Quantifying uncertainty in sample average approximation. *2015 Winter Simulation Conference (WSC)*, 3846–3857, URL http://dx.doi.org/10.1109/WSC.2015.7408541.

Lam H, Zhou E (2017) The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters* 45(4):301–307.

Li G, Hollander M, McKeague IW, Yang J (1996) Nonparametric likelihood ratio confidence bands for quantile functions from incomplete survival data. *The Annals of Statistics* 24(2):628–640.

Li G, Qin J, Tiwari RC (1997) Semiparametric likelihood ratio-based inferences for truncated data. *Journal of the American Statistical Association* 92(437):236–245.

Luenberger DG (1973) *Introduction to linear and nonlinear programming*, volume 28 (Addison-Wesley Reading, MA).

Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming* 171(1):115–166, ISSN 1436-4646, URL http://dx.doi.org/10.1007/s10107-017-1172-1.

Murphy SA (1995) Likelihood ratio-based confidence intervals in survival analysis. *Journal of the American Statistical Association* 90(432):1399–1405.

Newey WK, Smith RJ (2004) Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72(1):219–255.

of Governors Federal Reserve System B (2019) Dodd-frank act stress test 2019: Supervisory stress test-methodology. *Federal Reserve System Publications* .

Owen A (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75(2):237–249.

Owen A (1990) Empirical likelihood ratio confidence regions. *The Annals of Statistics* 90–120.

Owen A (1991) Empirical likelihood for linear models. *The Annals of Statistics* 1725–1747.

Owen A (2001) *Empirical likelihood* (CRC press).

Qin J, Lawless J (1994) Empirical likelihood and general estimating equations. *The Annals of Statistics* 300–325.

Rudin W (1964) *Principles of mathematical analysis*, volume 3 (McGraw-Hill New York).

Ruszczynski AP, Shapiro A (2003) *Stochastic programming*, volume 10 (Elsevier Amsterdam).

Shafieezadeh-Abadeh S, Esfahani PM, Kuhn D (2015) Distributionally robust logistic regression. *Advances in Neural Information Processing Systems 28*, 1576–1584 (Curran Associates, Inc.).

Shapiro A, Dentcheva D (2014) *Lectures on stochastic programming: modeling and theory*, volume 16 (SIAM).

Sinha A, Namkoong H, Duchi J (2018) Certifiable distributional robustness with principled adversarial training. *International Conference on Learning Representations*, URL `https://openreview.net/forum?id=Hk6kPgZA-`.

Villani C (2008) *Optimal transport: old and new* (Springer Science & Business Media).

Volpi R, Namkoong H, Sener O, Duchi JC, Murino V, Savarese S (2018) Generalizing to unseen domains via adversarial data augmentation. *Advances in Neural Information Processing Systems*, 5334–5344.

Wang Q, Rao JNK (2001) Empirical likelihood for linear regression models under imputation for missing responses. *Canadian Journal of Statistics* 29(4):597–608.

Wang Z, Glynn PW, Ye Y (2009) Likelihood robust optimization for data-driven newsvendor problems. *Preprint* .

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 9(1):60–62.

Yang I (2017) A convex optimization approach to distributionally robust markov decision processes with wasserstein distance. *IEEE control systems letters* 1(1):164–169.

Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters* 46(2):262 – 267, ISSN 0167-6377, URL `http://dx.doi.org/https://doi.org/10.1016/j.orl.2018.01.011`.

Zhao Y, Wang H (2008) Empirical likelihood inference for the regression model of mean quality-adjusted lifetime with censored data. *Canadian Journal of Statistics* 36(3):463–478.

Zhou M (2015) *Empirical Likelihood Method in Survival Analysis*, volume 79 (CRC Press).