SAMPLE OUT-OF-SAMPLE INFERENCE BASED ON WASSERSTEIN DISTANCE

Jose Blanchet  Yang Kang
jose.blanchet@columbia.edu  yangkang@stat.columbia.edu

Columbia University

Abstract. We present a novel inference approach which we call Sample Out-of-Sample (or SOS) inference. Our motivation is to propose a method which is well suited for data-driven stress testing, in which emphasis is placed on measuring the impact of (plausible) out-of-sample scenarios on a given performance measure of interest (such as a financial loss). The methodology is inspired by Empirical Likelihood (EL), but we optimize the empirical Wasserstein distance (instead of the empirical likelihood) induced by observations. From a methodological standpoint, our analysis of the asymptotic behavior of the induced Wasserstein-distance profile function shows dramatic qualitative differences relative to EL. For instance, in contrast to EL, which typically yields chi-squared weak convergence limits, our asymptotic distributions are often not chi-squared. Also, the rates of convergence that we obtain have some dependence on the dimension in a non-trivial way but which remains controlled as the dimension increases.

1. Introduction

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

In order to motivate our goal and the mathematical development that follows, consider the following stress-testing exercise. An insurance company wishes to estimate a certain expectation of interest, \( \mathbb{E}(L(X)) \), where \( X \) might represent a risk factor and \( L(X) \) the corresponding financial loss. The insurance company may estimate \( \mathbb{E}(L(X)) \) based on \( n \) i.i.d. (independent and identically distributed) empirical samples \( X_1, \ldots, X_n \in \mathbb{R}^l \). However, the regulator (or auditor) is also...
interested in quantifying the potential financial loss based on stress scenarios, say an i.i.d. sample \( Y_1, ..., Y_m \in \mathbb{R}^l \) (for simplicity we let \( m = n \)). The scenarios provided by the regulator may or may not come from the same distribution as the \( X_i \)’s.

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

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

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

Observe that \( R_n^W (\theta) \) is obtained by solving a linear programming problem. There is a strong connection between the SOS function and the Wasserstein’s distance of order two, this is discussed in the next section.

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

\[
R_n^W (\mathbb{E} (L(X))) \Rightarrow \nu R,
\]

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

\[
\{ \theta : R_n (\theta) \leq \delta \}
\]

(which is easily seen to be an interval) is an approximate 95% confidence interval which uses the stress scenarios in a meaningful way.
It is important to stress that the confidence interval designed via (2) contains estimates corresponding to all probability distributions which recognize the possibility of the stress scenarios, but which are also plausible given the available empirical evidence.

Let us provide additional motivation for the study of $R^W_n (\theta)$ by establishing a connection to distributional robust performance analysis of stochastic systems (see, for example, Lam (2013), Ben-Tal et al. (2013) and Goh and Sim (2010)). To illustrate such connection we continue working with the stress-testing situation introduced earlier. A distributional robust estimate of $E(L(X))$ is obtained by evaluating

\begin{equation}
\mathcal{U}_n (\Delta) = \max \left\{ \sum_{i,k} L(Z_k) \pi(i,k) : \right. \\
\left. \sum_k \pi(i,k) = 1/n \quad \forall i, \pi(i,k) \geq 0 \quad \forall i,k, \quad \sum_{i,k} ||X_i - Z_k||^2 \pi(i,k) \leq \Delta \right\}.
\end{equation}

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

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

\[ \mathcal{U}_n (\delta_n) = \max \{ \theta : R^W_n (\theta) \leq \delta_n \}. \]

To see this equality, let $\theta_n^+ = \sup \{ \theta : R^W_n (\theta) \leq \delta_n \}$ and let $\pi^R (\theta_n^+)$ be the optimizer of (1) (taking $\theta = \theta_n^+$) then, because $\pi^R (\theta_n^+)$ is feasible for (4), we have that $\mathcal{U}_n (\delta_n) \geq \theta_n^+$. Likewise, let $\pi^U (\delta_n)$ be the optimizer of (4) (taking $\Delta = \delta_n$) then, since $\pi^U (\delta_n)$ is feasible for (1) we obtain that $R^W_n (\mathcal{U}_n (\delta_n)) \leq \delta_n$ and therefore, by definition of $\theta_n^+$ we must have $\mathcal{U}_n (\delta_n) \leq \theta_n^+$. Therefore, our study of confidence intervals such as (3), and the asymptotic analysis of $R^W_n (\theta)$, as we indicate in (2) provide the means for optimally choosing $\delta_n$ in the context of distributional robust performance analysis. Similar connections to Empirical Likelihood had been noted in the literature (see Lam and Zhou (2015), Lam and Zhou (2016) and Blanchet et al. (2016a)). Additional connections to distributional robust optimization are discussed in Section 4.
The main methodological objective of this paper is to study the asymptotic behavior of general SOS functions for estimating equations (which we define in subsequent sections in the paper). That is, we wish to estimate $\theta^*$ such that

$$E(h(\theta^*, X)) = 0,$$

where $h(\theta, X) = (h_1(\theta, X), \ldots, h_q(\theta, X))^T$ (a column vector of functions) and $\theta \in \mathbb{R}^d$ (for $q \leq d$), under standard assumptions which make the inference problem of finding $\theta^*$ well posed using suitable SOS functions. Note that the particular case leading to (2) is obtained by letting $q = 1 = d$ and $h(\theta, x) = L(x) - \theta$.

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), Owen (1990) and Owen (2001)). In fact, as the reader might appreciate, we borrow a great deal of inspiration from the EL inference paradigm (and its extensions based on divergence criteria, rather than the likelihood function, Owen (2001)). There are, however, several important characteristics of our framework that, we believe, add significant value to the non-parametric inference literature.

First, from a conceptual standpoint, the EL framework restricts the support of the outcomes only to the observed empirical sample and, therefore, there is no reason to expect particularly good out of sample performance of estimates based on EL, for example, in settings similar to the stress testing exercise discussed earlier. In fact, the potentially out-of-sample problems which arise from using divergence criteria for data-driven distributional robust optimization (closely related to EL) are noted in the stochastic optimization literature, see Esfahani and Kuhn (2015); see also Wang et al. (2009) and Ben-Tal et al. (2013), for related work.

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

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

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

At a high level, these qualitative distinctions in the form of the asymptotic arise because of the linear programming formulation underlying the SOS function, which will typically lead to corner solutions (i.e. basic feasible solutions in the language of linear programming). In contrast, in the EL analysis of the profile function, the optimal solutions are amenable to a perturbation analysis as $n \to \infty$ using a Taylor expansion of higher order terms. The lack of a continuously differentiable derivative (of the optimal solution as a function of $\theta$) requires a different type of analysis relative to the approach (traced back to the classical Wilks theorem, Wilks (1938)) which lies at the core of EL analysis. We believe that the proof techniques that we develop here might have wider applicability.

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

a) We characterize the asymptotic distribution of $R^W_n(\theta_*)$ defined in (5) as $n \to \infty$ (see Theorem 1).

b) We introduce two forms of the SOS inference framework for estimating equations. We call these the implicit and the explicit SOS formulations, respectively. These formulations, as we shall discuss, are motivated by different types of applications (see Theorem 2 and Theorem 3).

c) Writing $\theta_* = (\gamma_*, v_*)$ we develop the asymptotic distribution of $R^W_n(\gamma_*, \bar{v}_n)$, where $\bar{v}_n$ is a suitable consistent plug-in estimator for $v_*$ as $n \to \infty$. This extension is particularly useful to reduce the computational burden involved in solving the optimization problem underlying the use of the SOS function for inference (see Corollary 1 and Corollary 2).
d) We apply our SOS inference framework in the context of stochastic optimization and stress testing (see Section 4).

e) Possible extensions and applications of our framework are given in our conclusions section, namely, Section 5. We also discuss results in Blanchet et al. (2016b, which include connections to machine learning, extensions beyond the Wasserstein distance of order two, and more general distributions for out-of-sample evaluation (beyond those supported on finitely many scenarios as discussed here).

We have discussed the qualitative features of our contributions in a) and b). About item c), its analysis parallels, in a way, the extensions developed by Hjort et al. (2009) in the context of EL. The applications to stochastic optimization, in particular, highlight the need for the general form of SOS function.

Regarding item d). A recent paper of Esfahani and Kuhn (2015) proposes Wasserstein’s distance in the context of distributional robust stochastic optimization. In Esfahani and Kuhn (2015), the authors take advantage of recently developed concentration inequalities for the Wasserstein distance (see Fournier and Guillin (2015)) to guarantee an asymptotically correct confidence level for the obtained stochastic programming bounds. In particular, given a certain degree of confidence (say 95%), if one wishes to estimate a plausible distributional robust feasible region within $\varepsilon$ error, their bound implies $O(\varepsilon^{-l})$ number of samples. In contrast, applying our results to the problems in Esfahani and Kuhn (2015) we can see that $O(\varepsilon^{-\min(l,2)})$ samples suffice. In simple words, the bounds obtained in Esfahani and Kuhn (2015) appear to be rather pessimistic; while the bounds in Esfahani and Kuhn (2015) suggest that estimating the distributional uncertain region suffers from the curse of dimensionality, our results show that this is not the case. We believe that our results here might be helpful when estimating Wasserstein’s distances in high dimensions.

The rest of the paper is organized as follows. In Section 2 we present and discuss our methodological results, in particular the contributions related to items a) to c) above. In Section 3 we provide the proofs of our results. Section 4 contains applications to stochastic optimization and stress testing (corresponding to item d) above), and including an empirical example. As mentioned earlier in item e), Section 5 contains final considerations and further applications.
2. Basic Definitions and Main Results

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 $\mathbb{E}(\|W\|_2^2) < \infty$, we use $\text{Var}(W) = \mathbb{E}(WW^T)$ to denote the covariance matrix of $X$.

In this section we present our results for the analysis of the SOS profile function for means first and later we move to estimating equations.

2.1. SOS Function for Means. We state the following underlying assumption throughout this subsection.

A1): Let us write $X_n = \{X_1, ..., X_n\} \subset \mathbb{R}^l$ to denote an i.i.d. sample from a continuous distribution. So, the cardinality of the set $X_n$ is $n$.

A2): We also consider an independent i.i.d. sample $Y_m = \{Y_1, ..., Y_m\} \subset \mathbb{R}^l$ from a continuous distribution. Throughout our discussion we shall assume that $m = \lceil \kappa n \rceil$ 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 $Z_{n+m} = \{Z_1, ..., Z_{n+m}\} = X_n \cup Y_m$, with $Z_k = X_k$ for $k = 1, ..., n$, and $Z_{n+j} = Y_j$ for $j = 1, ..., m$. For any closed set $C$ let us write $\mathcal{P}(C)$ to denote the set of probability measures supported on $C$. So, in particular, a typical element $\nu_n \in \mathcal{P}(Z_{n+m})$ takes the form

$$
\nu_n(dz) = \sum_{k=1}^{n+m} v(k) \delta_{Z_k}(dz),
$$

where $\delta_{Z_k}(dz)$ is a Dirac measure centered at $Z_k$. Now, we shall use $\mu_n \in \mathcal{P}(X_n)$ to denote the empirical measure associated to $X_n$, that is,

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

Given any $\pi \in \mathcal{P}(X_n \times Z_{n+m})$ we write $\pi_X \in \mathcal{P}(X_n)$ to denote the marginal distribution with respect to the first coordinate, namely $\pi_X(dx) = \int_{z \in Z_{n+m}} \pi(dx,dz)$ and, likewise, we define $\pi_Z \in \mathcal{P}(Z_n)$ as $\pi_Z(dz) = \int_{z \in X_n} \pi(dx,dz)$.

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

\[
R_n^W(\theta^*) = \inf \{ \int \int \|x - z\|^2_2 \pi(dx,dz) : \\
s.t. \pi \in \mathcal{P}(X_n \times Z_{(n+m)}), \pi_X = \mu_n, \pi_Z = v_n, \int z\pi_n(dz) = \theta^* \}.
\]

(Here and throughout the paper s.t. abbreviates “subject to”.)

Remark 1. The connection to the Wasserstein distance (of order 2), $d_2(\mu_n, v_n)$, can be directly appreciated by recalling that

\[
d_2(\mu_n, v_n)^2 = \inf \{ \int \int \|x - z\|^2_2 \pi(dx,dz) : \pi \in \mathcal{P}(X_n \times Z_{(n+m)}), \pi_X = \mu_n, \pi_Z = v_n \}.
\]

In simple words, $R_n^W(\theta^*)$ is obtained by minimizing the (squared) Wasserstein distance to the empirical measure among all distributions $v_n$ supported on $Z_{(n+m)}$ with expected value equal to $\theta^*$ (i.e. $\mathbb{E}_{v_n}(Z) = \int zv_n(dz) = \theta^*$).

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)-A3), suppose that the covariance matrix of $X$, $\text{Var}(X)$, has full rank $l$. The following asymptotic result follows

- When $l = 1$,
  \[
  nR_n^W(\theta^*) \Rightarrow \sigma^2 \chi^2_1
  \]
  where $\sigma^2 = \text{Var}(X)$.

- When $l = 2$,
  \[
  nR_n^W(\theta^*) \Rightarrow \rho \left( \|\tilde{Z}\|^2_2 \right) \left[ 2 - \eta \left( \|\tilde{Z}\|^2_2 \right) \rho \left( \|\tilde{Z}\|^2_2 \right) \right] \|\tilde{Z}\|^2
  \]
  where $\rho \left( \|\tilde{Z}\|^2_2 \right)$ is the unique solution to
  \[
  \frac{1}{\rho} = \mathbb{P} \left[ \rho^2 \|\tilde{Z}\|^2_2 \geq \tau \left| \|\tilde{Z}\|^2_2 \right. \right].
  \]
and
\[
\eta \left( \| \tilde{Z} \|_2^2 \right) = E \left[ \max \left( 1 - \frac{\tau}{\rho \left( \| \tilde{Z} \|_2^2 \right)^2 \| \tilde{Z} \|_2^2}, 0 \right), \| \tilde{Z} \|_2^2 \right],
\]

with \( \tilde{Z} \sim N (0, \text{Var} (X)) \in \mathbb{R}^l \) and \( \tau \) is independent of \( \tilde{Z} \) satisfying
\[
P (\tau > t) = E [\exp (- (f_X (X_1) + \kappa f_Y (X_1)) \pi t)].
\]

- When \( l \geq 3 \),
\[
n^{1/2 + \frac{3}{2l + 2}} R_n^W (\theta_*) \Rightarrow \frac{2l + 2}{l + 2} \frac{\| \tilde{Z} \|_{2+1}^{1+1}}{\left( E \left[ \frac{q^{l/2}}{\Gamma (l/2 + 1)} (f_X (X_1) + \kappa f_Y (X_1)) \right] \right)^{1/\tau}}
\]

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

### 2.2. SOS Function for Estimating Equations

Throughout this subsection we assume that\( A1 \) and \( A2 \) are in force. Let us assume that \( h : \mathbb{R}^d \times \mathbb{R}^l \rightarrow \mathbb{R}^q \), we assume that \( q \leq d \). We impose the following assumptions.

**B1** Assume \( \theta_\ast \in \mathbb{R}^d \) satisfies
\[
E (h (\theta_\ast, X)) = 0.
\]

**B2** Furthermore, suppose that
\[
E \| h (\theta_\ast, X) \|^2_2 < \infty.
\]

Our goal is to estimate \( \theta_\ast \) 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.

#### 2.2.1. Implicit SOS Formulation for Estimating Equations

The first SOS function form for estimating equations is the following, we call it Implicit SOS or Indirect SOS function because the Wasserstein distance is applied to \( h (\theta, X_i) \) and \( h (\theta, Z_k) \) and thus it implicitly or indirectly induces a notion of proximity among the samples.
The Implicit SOS formulation might lead to dimension reductions if \( l \) (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 use denote \( g_X(\cdot) \) is the density for \( h(\theta_*, X_i) \in \mathbb{R}^q \) and \( g_Y(\cdot) \) for the density of \( h(\theta_*, Y_i) \in \mathbb{R}^{q_2} \). Then, the Wasserstein profile function defined in Equation (7) have following asymptotic results:

- When \( q = 1 \),

\[
\frac{n}{R_n^W(\theta_*)} \Rightarrow \text{Var} \left( h(\theta_*, X_1) \right) \chi_1^2
\]

- When \( q = 2 \),

\[
\frac{n}{R_n^W(\theta_*)} \Rightarrow \rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \left[ 2 - \eta \left( \left\| \tilde{Z} \right\|_2^2 \right) \rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \right] \left\| \tilde{Z} \right\|_2^2
\]

where \( \rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \) is the unique solution to

\[
\frac{1}{\rho} = \mathbb{P} \left[ \rho^2 \left\| \tilde{Z} \right\|_2^2 \geq \tau \middle| \left\| \tilde{Z} \right\|_2^2 \right]
\]

and

\[
\eta \left( \left\| \tilde{Z} \right\|_2^2 \right) = \mathbb{E} \left[ \max \left( 1 - \frac{\tau}{\rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \left\| \tilde{Z} \right\|_2^2}, 0 \right) \middle| \left\| \tilde{Z} \right\|_2^2 \right]
\]
with $\tilde{Z} \sim N(0, \text{Var}(h(\theta, X))) \in \mathbb{R}^q$ and $\tau$ is independent of $\tilde{Z}$ satisfying

$$
P[\tau > t] = E[\exp(-[g_X(h(\theta, X_1)) + \kappa g_Y(h(\theta, X_1))] \pi t)].$$

- When $q \geq 3$,

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

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

2.2.2. Explicit SOS Formulation for Estimating Equations. The second SOS function form we call Explicit SOS function because the Wasserstein distance is explicitly or directly applied to the samples and the scenarios.

**Definition 3** (Explicit SOS Profile Function for Estimating Equations).

$$R_n^W(\theta_*) = \inf \left\{ \int \int \|x - z\|^2 \pi(dx, dz) : \right.$$

$$\left. \quad \text{s.t. } \pi \in \mathcal{P}(X_n \times Z_{(n+m)}), \pi_X = \mu_n, \int h(\theta, z) \pi_Z(dz) = 0 \right\}.$$  

Both the implicit and explicit SOS have their merits. We have discussed the merit of the implicit SOS formulation. For the Explicit SOS formulation, consider the stress testing application discussed in the Introduction. The interest of an auditor or a regulator might be on the impact of scenarios on a specific performance measure of interest. One might think that the regulator applies the same stress scenarios to different insurance companies or banks, and therefore the function $h(\cdot)$ is unique to each insurance company. The regulator is interested in the impact of stress testing scenarios on the structure of the bank (modeled by $h(\cdot)$). In this setting, the Explicit SOS formulation appears more appropriate.

While the analysis of the Explicit SOS formulation is also largely based on the techniques developed for Theorem 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.
Assume that the derivative of \( h(\theta_*, x) \) with respect to (w.r.t.) \( x \), \( D_x h(\theta_*, \cdot) : R^l \to R^q \), is a continuous function of \( x \) and the second derivative w.r.t. \( x \) is bounded, i.e. \( ||D_x^2 h(\theta_*, \cdot)|| < \bar{K} \) for all \( x \).

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 \( \Upsilon = E(V_i) \) is strictly positive definite.

We provide the proof of the next result in our technical Section 3.3.

**Theorem 3** (Explicit SOS Profile Function Analysis). Under assumptions A1)-A2), B1)-B2) and BE1)-BE2), we have that (8) satisfies

- When \( l = 1 \),
  \[
  nR_n^W(\theta_*) \Rightarrow \tilde{Z}^T \Upsilon^{-1} \tilde{Z}
  \]
  where \( \tilde{Z} \sim N(0, \text{Var}(h(\theta_*, X))) \in R^q \).

- Assume that \( l = 2 \). It is possible to uniquely define \( \tilde{\zeta} : R^q \to R^q \) continuous such that
  \[
  z = -E \left[ V_1 I \left( \tau \leq \tilde{\zeta}^T(z) V_1 \tilde{\zeta}(z) \right) \right] \tilde{\zeta}(z),
  \]
  where \( \tau \) is independent of \( \tilde{Z} \) satisfying
  \[
  \mathbb{P}(\tau > t) = E(\exp(-[f_X(X_1) + \kappa f_Y(X_1)] \pi t)).
  \]
  Moreover, we have that,
  \[
  nR_n^W(\theta_*) \Rightarrow -2\tilde{Z}^T \tilde{\zeta}(\tilde{Z}) - \tilde{\zeta}^T(\tilde{Z})E \left[ V_1 \max \left( 1 - \frac{\tau}{\tilde{\zeta}^T(\tilde{Z}) V_1 \tilde{\zeta}(\tilde{Z})}, 0 \right) \right] \tilde{Z} \tilde{\zeta}(\tilde{Z}).
  \]
  where \( \tilde{Z} \sim N(0, \text{Var}(h(\theta_*, X))) \in R^q \).

- Suppose that \( l \geq 3 \). It is possible to uniquely define \( \tilde{\zeta} : R^q \to R^q \) continuous such that
  \[
  z = -E \left[ \frac{\pi^{l/2}}{\Gamma(l/2 + 1)} \left( f_X(X_1) + \kappa f_Y(X_1) \right) V_1 \cdot \left( \tilde{\zeta}^T(z) V_1 \tilde{\zeta}(z) \right)^l \right] \tilde{\zeta}(z),
  \]
(note that $\tilde{V}_1$ is a function of $X_1$). Moreover,

$$n^{1/2 + \frac{3}{2l+2}} R^W_n(\theta_*) \Rightarrow -2\bar{Z}^T \tilde{\zeta} \left( \bar{Z} \right) - \frac{2}{l+2} \mathbb{E} \left[ \frac{\pi^{l/2}}{\Gamma(l/2+1)} (f_X(X_1) + \kappa f_Y(X_1)) \left( \tilde{\zeta}^T \left( \bar{Z} \right) V_1 \tilde{\zeta} \left( \bar{Z} \right) \right)^{l/2+1} \right],$$

where $\tilde{Z} \sim N(0, \text{Var}(h(\theta_*, X))) \in \mathbb{R}^q$ independent of $X_1$.

We should observe that unlike implicit formulation, the rate of convergence will only depend on the dimension of data $X_i \in \mathbb{R}^l$, but the shape of asymptotic distribution is determined by the estimating functions $h(\theta_*, X_i) \in \mathbb{R}^q$.

2.3. Plug-in Estimators for SOS Functions. In many situations, for example in the context of stochastic optimization, we are interested in a specific parameter $\theta_* = (\gamma_*^*, \nu_*^*) \in \mathbb{R}^{d+p}$ such that $\mathbb{E}[h(\gamma_*^*, \nu_*, X)] = 0$, where $\nu_* \in \mathbb{R}^p$ is the nuisance parameter. We shall discuss a method that allows us to deal with the nuisance parameter using a plug-in estimator, while taking advantage of the SOS framework for the estimation of $\gamma_*$. After we state our assumptions we will provide the results in this section and the proofs, which follow closely those of Theorems 3 and 2 will be given in Section 3.

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

**C1)** Given $\gamma_*$ there is a unique $\nu_* \in \mathbb{R}^p$ such that

$$\mathbb{E}[h(\gamma_*, \nu_*, X)] = 0$$

and, given $\nu_*$, we also assume that $\gamma_*$ satisfies

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

**C2)** We have access to a suitable estimator $\nu_n$ such that the sequence

$$\left\{ n^{1/2} (\nu_n - \nu_*) \right\}_{n=1}^{\infty} \text{ is tight},$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(\gamma_*, \nu_n, X_i) \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 $V$ around $v_*$. 

C4) Suppose that there is a function $M(\cdot): \mathbb{R}^l \to (0, \infty)$ satisfying that

$$
||h(\gamma^*, \nu, x)||_2^2 \leq M(x) \text{ for a.e. } \nu \in V,
$$

and $E(M(X_1)) < \infty$.

2.3.1. **Plug-in Estimators for Implicit SOS Functions.** We are interested in studying the plug-in implicit SOS function (or implicit pseudo-SOS profile function) given by

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

where,

$$
\mathcal{X}_n^{h}(\gamma^*, \nu_n) = \{h(\gamma^*, \nu_n, x) : x \in \mathcal{X}_n\}, \mathcal{Z}_{(n+m)}^{h}(\gamma^*, \nu_n) = \{h(\gamma^*, \nu_n, z) : z \in \mathcal{Z}_{(n+m)}\}.
$$

We typically will use (9) to find a plug-in estimator $\nu_n$. Under suitable assumptions on the consistency and convergence rate of the plug-in estimator we have an asymptotic result for (11), as we indicate next.

**Corollary 1** (Plug-in for Implicit SOS Formulation). **Assuming A1)-A2), and C1)-C4) hold.** Moreover, suppose denote $g_X(\cdot)$ as the density for $h(\gamma^*, v_*, X_1) \in \mathbb{R}^q$ and $g_Y(\cdot)$ for the density of $h(\gamma^*, v_*, Y_1) \in \mathbb{R}^q$. We notice $\tilde{Z}' \in \mathbb{R}^q$ is defined in C2). We obtain that (11) has following asymptotic behavior

- When $q = 1$,

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

- When $q = 2$,

$$
nR_n^W(\gamma^*) \Rightarrow \rho \left( ||\tilde{Z}'||_2^2 \right) \left[ 2 - \eta \left( ||\tilde{Z}'||_2^2 \right) \rho \left( ||\tilde{Z}'||_2^2 \right) \right] ||\tilde{Z}'||^2
$$
where \( \rho \left( \left\| \tilde{Z}' \right\|_2^2 \right) \) is the unique solution to

\[
\frac{1}{\rho} = \mathbb{P} \left[ \rho^2 \left\| \tilde{Z}' \right\|_2^2 \geq \tau \left\| \tilde{Z}' \right\|_2^2 \right],
\]

and

\[
\eta \left( \left\| \tilde{Z}' \right\|_2^2 \right) = \mathbb{E} \left[ \max \left( 1 - \frac{\tau}{\rho} \left( \left\| \tilde{Z}' \right\|_2^2 \right) \, 0 \right) \left\| \tilde{Z}' \right\|_2^2 \right],
\]

with \( \tau \) is independent of \( \tilde{Z}' \) satisfying

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

- When \( q \geq 3 \),

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

2.3.2. **Plug-in Estimators for Explicit SOS Functions.** We can also analyze plug-in estimators for Explicit SOS profile functions. We now define the explicit plug-in (or pseudo) SOS function based on (8) as simply plugging-in the nuisance parameter:

\[
R_n^W (\gamma_*) = \inf \int \int \| x - z \|_2^2 \, \pi(dx, dz) : \\
\text{s.t.} \quad \pi \in \mathcal{P} \left( X_n^h (\gamma_*, \nu_n) \times \mathcal{Z}^h_{(n+\mu)} (\gamma_*, \nu_n) \right), \pi_X = \mu_n, \int h (\gamma_*, \nu_n, z) \, \pi_Z(dz) = 0 \}
\]

In addition to **C1)** to **C4)** introduced at the beginning of this subsection, we shall impose the following additional assumptions:

**C5)** Define \( \tilde{V}_i (\nu_*) = D_x h (\gamma_*, \nu_*, X_i) \cdot D_x h (\gamma_*, \nu_*, X_i)^T \) and assume that \( \tilde{Y} = E \left( \tilde{V}_i \right) \) is strictly positive definite.

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

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

\[
\left\| D_{\nu} D_x h (\gamma_*, \nu, x) \right\|_2^2 \leq M(x) \text{ for a.e. } \nu \in \mathcal{V}.
\]
The second derivative w.r.t. $x$ exist and bounded, i.e. $\|D_x^2h(\gamma, \nu, x)\| < \bar{K}$ for a.e. $\nu \in \mathcal{V}$ and all $x$.

**Corollary 2** (Plug-in for Explicit SOS Formulation). $X_i \in \mathbb{R}^l$, $h(\gamma, \nu, x) \in \mathbb{R}^q$. Assume that A1)-A2) and C1)-C7) hold. We notice $\tilde{Z}'$ is defined in C2). Then, the SOS profile function defined in Equation (12) has the following asymptotic properties.

- When $l = 1$,
  \[ nR_n^W(\gamma^*) \Rightarrow \tilde{Z}'^T \tilde{Y}^{-1} \tilde{Z}'. \]

- Suppose that $l = 2$. It is possible to uniquely define $\tilde{\zeta} : \mathbb{R}^l \to \mathbb{R}^l$ continuous such that
  \[ z = -E \left[ \tilde{V}_1 I \left( \tau \leq \tilde{\zeta}^T (z) \tilde{V}_1 \tilde{\zeta} (z) \right) \tilde{\zeta} (z) \right], \]
  where $\tau$ is independent of $\tilde{Z}'$ satisfying
  \[ \mathbb{P}(\tau > t) = E(\exp(-[f_X(X_1) + \kappa f_Y(X_1)]\pi t)). \]
  Furthermore,
  \[ nR_n^W(\theta^*) \Rightarrow -2\tilde{\zeta}^T (\tilde{Z}') \tilde{Z}' - \tilde{\zeta}^T (\tilde{Z}') \mathbb{E} \left[ \tilde{V}_1 \max \left( 1 - \frac{\tau}{\tilde{\zeta}^T (\tilde{Z}') \tilde{V}_1 \tilde{\zeta} (\tilde{Z}')}, 0 \right) \right] \tilde{\zeta} (\tilde{Z}'), \]
  and $\tilde{Z}'$ is independent with $\tilde{V}_1$ and $\tau$.

- Assume that $l \geq 3$. A continuous function $\tilde{\zeta} : \mathbb{R}^l \to \mathbb{R}^l$ can be defined uniquely so that
  \[ z = -E \left[ \frac{\pi^{l/2}(f_X(X_1) + \kappa f_Y(X_1))}{\Gamma(l/2 + 1)} \tilde{V}_1 \left( \tilde{\zeta}^T (z) \tilde{V}_1 \tilde{\zeta} (z) \right)^{l/2+1} \right] \tilde{\zeta} (z), \]
  (note that $\tilde{V}_1$ is a function of $X_1$). Moreover,
  \[ n^{1/2+\frac{1}{2(l+2)}} R_n^W(\theta^*) \Rightarrow -2\tilde{\zeta}^T (\tilde{Z}') \tilde{Z}' - \frac{2}{l+2} \mathbb{E} \left[ \frac{\pi^{l/2}(f_X(X_1) + \kappa f_Y(X_1))}{\Gamma(l/2 + 1)} \left( \tilde{\zeta}^T (\tilde{Z}') \tilde{V}_1 \tilde{\zeta} (\tilde{Z}') \right)^{l/2+1} \right] \tilde{Z}', \]
  where $\tilde{Z}'$ and $X_1$ are independent.
3. Methodological Development

We shall analyze the limiting distribution of the SOS profile function for means first. In order to gain some intuition let us perform some basic manipulations. First, without loss of generality we assume \( \theta_\ast = 0 \), otherwise, we can let \( \tilde{X}_i = X_i - \theta_\ast \) and apply the analysis to the \( \tilde{X}_i \)'s.

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

The Dual Problem. Let us define \( \tilde{\pi}(i,j) = \delta_{i,j}/n \) (where \( \delta_{i,j} = I(i = j) \) is Kronecker’s delta), and we shall also define \( \mu(i,j) = \pi(i,j) - \tilde{\pi}(i,j) \). Observe that \( n \sum_{i,j=1}^{n} \tilde{\pi}(i,j) ||X_i - X_j||_2^2 = 0 \), then we can rewrite the definition of (6) via the following equivalent linear programming problem

\[
R^W_n(\theta_\ast) = \min \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m+n} \mu(i,j) ||X_i - Z_j||_2^2 \right\}
\]

\[
\begin{align*}
\{ & \mu(i,j) \geq 0, \text{ for } i \neq j \\
& \mu(i,i) \leq 0, \text{ for all } i \\
& \sum_{j=1}^{m+n} \mu(i,j) = 0, \text{ for all } i \\
& \left( \sum_{j=1}^{m+n} \mu(i,j) \right) Z_j = -\bar{X}_n \\
& \mu(i,i) \geq -1/n
\end{align*}
\]

where \( \bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i \).

Here, we should notice that, in order to satisfies the condition \( \sum_{j=1}^{m+n} \mu(i,j) = 0 \) for all \( i \in \{1, \ldots, n\} \) and with the definition of \( \mu(i,j) \), we need to have \( \mu(i,i) \leq 0 \) for all \( i \in \{1, \ldots, n\} \) and \( \mu(i,j) \geq 0 \) for \( i \neq j \).

We know when \( n \to \infty \) with probability 1, \( \bar{0} \) 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 (13) in the dual formulation as

\[
R^W_n(\theta_\ast) = \max_{\beta, \lambda, \gamma} \left\{ -\lambda^T \bar{X}_n - \frac{1}{n} \sum_{i=1}^{n} \gamma_i \right\}
\]

\[
\text{s.t. } \begin{cases} 
\beta_i + \lambda^T Z_j \leq ||X_i - Z_j||_2^2, \text{ for } i \neq j \\
\beta_i + \lambda^T X_i + \gamma_i \geq 0, \text{ for all } i
\end{cases}
\]
For each fixed $i$, by observing the two constraints in the previous display, while noticing $\beta_i$’s are free parameters we have that

$$\beta_i \leq -\lambda^T Z_j + ||X_i - Z_j||_2^2, \text{ for } i \neq j$$

$$\beta_i \geq -\lambda^T X_i - \gamma_i.$$ 

So, we can cancel the $\beta_i$’s and combine the two optimization constrains as:

$$-\lambda^T X_i - \gamma_i \leq -\lambda^T Z_j + ||X_i - Z_j||_2^2, \text{ for } i \neq j$$

The above condition is for $i \neq j$, while for the case that $i = j$ the above condition becomes $\gamma_i \geq 0$. So, let us rewrite the dual problem as:

$$R_n^W(\theta^*) = \max_{\lambda, \gamma \geq 0} \left\{ -\lambda^T \bar{X}_n - \frac{1}{n} \sum_{i=1}^{n} \gamma_i \right\}$$

$$\text{s.t. } -\lambda^T X_i - \gamma_i \leq -\lambda^T Z_j + ||X_i - Z_j||_2^2, \text{ for all } i, j$$

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

$$R_n^W(\theta^*) = \max_{\lambda, \gamma \geq 0} \left\{ -\lambda^T \bar{X}_n - \frac{1}{n} \sum_{i=1}^{n} \gamma_i \right\}$$

$$\text{s.t. } -\lambda^T X_i - \gamma_i \leq \inf_j \left\{ -\lambda^T Z_j + ||X_i - Z_j||_2^2 \right\}, \text{ for all } i.$$ 

**High-Level Intuitive Analysis.** At this point we can perform a high-level and intuitive analysis which can help us guide our intuition about our result. First, consider an approximation performed by freeing the $Z_j$ in the constraints of (16), in this portion the reader can appreciate that the assumption that $X_j$ has a density yields

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

where error $\epsilon_n (i)$ is small and it will be discussed momentarily. Now, observe that the optimal $a^*_i = X_i + \lambda/2$, therefore

$$\inf_j \left\{ -\lambda^T Z_j + ||X_i - Z_j||_2^2 \right\} = -\lambda^T X_i - ||\lambda||_2^2 / 4 + \epsilon_n (i).$$
Hence, the $i$-th constraint in (16) takes the form
\[-\lambda^T X_i - \gamma_i \leq -\lambda^T X_i - \|\lambda\|^2_2 / 4 + \epsilon_n (i),\]
and thus (16) can ultimately be written as
\[
R_n^W (\theta_*) = - \min_{\lambda, \gamma_i \geq 0} \left\{ \lambda^T \bar{X}_n + \frac{1}{n} \sum_{i=1}^{n} \gamma_i \right\}
\text{ s.t. } \gamma_i \geq \|\lambda\|^2_2 / 4 - \epsilon_n (i) \text{ for all } i.
\]

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 $\{Z_1, \ldots, Z_{m+n}\} \setminus X_i$ which is within $O_p (1/n)$ distance of $a_* (i)$). Then it is intuitive to expect the approximation
\[
R_n^W (\theta_*) = - \min_{\lambda} \left\{ \lambda \bar{X}_n + \lambda^2 / 4 + O_p (1/n) \right\},
\]
which formally yields an optimal selection $\lambda_* (n) = -\bar{X}_n + O_p (1/n)$ and therefore we expect, due to the Central Limit Theorem (CLT), that
\[
nR_n^W (\theta_*) = n \bar{X}_n^2 + nO_p \left( 1/n^{3/2} \right) \Rightarrow Var (X) \chi^2_1
\]
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 $\epsilon_n (i) = O_p (1/n^{1/2})$; this time, with positive probability (uniformly as $n \to \infty$) we must have that a point in the sample $\{Z_1, \ldots, Z_{m+n}\} \setminus X_i$ is within $O_p (1/n^{1/2})$ 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 $O (1/n^{1/2})$). 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 section. A key technical device introduced in our proof technique is a Poisson point process which approximates the number of points in $\{Z_1, ..., Z_{m+n}\} \setminus X_i$ which are within a distance of size $O\left(n^{-1/l}\right)$ from the free optimizer $a_*(i)$ arising in (17).

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

3.2. 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.

3.2.1. Step 1 (Dual Formulation and Lower Bound): Using the same transformations introduced in (13) we can obtain the dual formulation of the SOS profile function (6), which is a natural adaptation of (16), namely

$$R^W_n(\theta_*) = \max_{\lambda, \gamma_n \geq 0} \left\{ -\lambda \bar{X}_n - \frac{1}{n} \sum_{i=1}^{n} \gamma_i \right\}$$

subject to

$$-\lambda^T X_i - \gamma_i \leq \inf_j \left\{ -\lambda^T Z_j + \|X_i - Z_j\|_2^2 \right\}, \text{ for all } i.$$

Observe that the following lower bound applies by optimizing over $a \in R^l$ instead of $a = Z_j \in Z_n$, 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 $a_*(X_i, \lambda) = X_i + \lambda/2$.

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

$$N_n^{(i)}(t, \lambda) = \# \left\{ Z_j : \|Z_j - a_*(X_i, \lambda)\|_2 \leq t^{2/l}/n^{2/l}, Z_j \neq X_i \right\}.$$
(recall that $Z_j \in \mathbb{R}^l$). Observe that, actually, we have

$$N_n^{(i)}(t, \lambda) = N_n^{(i)}(t, \lambda, 1) + N_n^{(i)}(t, \lambda, 2),$$

where

$$N_n^{(i)}(t, \lambda, 1) = \# \left\{ X_j : \| X_j - a_*(X_i, \lambda) \|_2^2 \leq t^{2/l}/n^{2/l}, X_j \neq X_i \right\},$$

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

For any $X_j$ with $j \neq i$, conditional on $X_i$, we have,

$$P \left[ \| X_j - a_*(X_i, \lambda) \|_2^2 \leq t^{2/l}/n^{2/l} \mid X_i \right] = f_X \left( X_i + \lambda/2 \right) \frac{\pi^{l/2}}{\Gamma(l/2 + 1)} \frac{t}{n} + o_p(t/n).$$

Similarly,

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

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,

$$N_n^{(i)}(t, \lambda, 1) \mid X_i \sim \text{Bin} \left( f_X \left( X_i + \lambda/2 \right) \frac{\pi^{l/2}}{\Gamma(l/2 + 1)} \frac{t}{n} + o_p(t/n), n - 1 \right),$$

$$N_n^{(i)}(t, \lambda, 2) \mid X_i \sim \text{Bin} \left( f_Y \left( X_i + \lambda/2 \right) \frac{\pi^{l/2}}{\Gamma(l/2 + 1)} \frac{t}{n} + o_p(t/n), [\kappa n] \right),$$

$$N_n^{(i)}(t, \lambda) = N_n^{(i)}(t, \lambda, 1) + N_n^{(i)}(t, \lambda, 2).$$

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

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

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

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

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 $f_X(X_i + \lambda/2) \pi^{d/2}/\Gamma(d/2 + 1)$. Similar considerations apply to $N_n^{(i)}(\cdot, \lambda, 2) | X_i$ which yield that

$$N_n^{(i)}(\cdot, \lambda) | X_i \Rightarrow \text{Poi} \left( \Lambda(X_i, \lambda) \right),$$

where

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

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

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

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

(20)

$$\mathbb{P}[T_i(n) > t | X_i] = \mathbb{P} \left[ N_n^{(1)}(t, \lambda) + N_n^{(2)}(t, \lambda) = 0 \bigg| X_i \right] = \exp \left( -\Lambda(X_i, \lambda) t \right) \left( 1 + O \left( 1/n^{1/l} \right) \right),$$

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 (20) we define $\tau_i(X_i)$ to be a random variable such that

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

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

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

We finish Step 2 with the statement of two technical lemmas. The first provides a rate of convergence for the Glivenko-Cantelli theorem associated to the sequence $\{T_i(n)\}_{i=1}^n$.

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

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

and

$$\lim_{n \to \infty} \mathbb{E} \left( \sup_{t \in [0, T]} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n (\max(t^2 - T_i(n)^\alpha, 0) - \mathbb{E}[\max(t^2 - T_i(n)^\alpha, 0)]) \right| \right) < \infty.$$
The second technical lemma deals with local properties of the distribution of \( T_i(n) \) at the origin. The proofs of both of these technical results are given at the end of the proof of Theorem 1, in Section 3.2.7.

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

\[
\mathbb{P}[T_i(n) \leq t] - \mathbb{P}[\tau \leq t] = O(t^{1+1/l}/n^{1/l})
\]

### 3.2.3. Step 3 (Closest Point and SOS Function Simplification):

Note that the \( i \)-th constraint, namely,

\[
-\gamma_i \leq \lambda^T X_i + \inf_j \left\{ -\lambda^T Z_j + ||X_i - Z_j||_2^2 \right\},
\]

can be written as

\[
-\gamma_i \leq \inf_j \left\{ -\lambda^T(Z_j - X_i) + ||X_i - Z_j||_2^2 \right\}
\]

\[
= - ||\lambda||_2^2 / 4 + \inf_j \left\{ ||Z_j - (\lambda/2 + X_i)||_2^2 \right\}
\]

\[
= - ||\lambda||_2^2 / 4 + T_i^{2/l}(n)/n^{2/l}.
\]

However, since \( \gamma_i \geq 0 \) we must have that

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

Therefore, the SOS profile function takes the form

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

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

\[
(21) \quad R_n^W(\theta_*) = \max_{\lambda} \left\{ -2\lambda^T X_n - \frac{1}{n} \sum_{i=1}^{n} \max \left( ||\lambda||_2^2 - \frac{T_i^{2/l}(n)}{n^{2/l}}, 0 \right) \right\}.
\]
3.2.4. **Step 4 (Case \( l = 1 \)):** When \( l = 1 \), let’s denote \( \sqrt{n} \bar{X}_n = Z_n \) and \( \sqrt{n} \lambda = \zeta \), where by CLT we can show \( Z_n \Rightarrow N(0, \sigma^2) \). Then we have:

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

The second equation follows the estimate in (Lemma 1). We know the objective function as a function of \( \theta \) is a strictly convex function. Since as \( \zeta = b |Z_n| \) with \( b \rightarrow \pm\infty \) implies that the objective function will tend to \(-\infty\), we conclude that the sequence of global optimizers is tight and each optimizer (i.e. for each \( n \)) could be characterized by the first order optimality condition almost surely.

We can take the derivative with respect to \( \zeta \) in

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

and set it to zero. We need to notice that while taking the derivative we require exchanging the derivative and expectation, this can be done true here by the dominating convergence theorem since

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

for all \( \delta > 0 \). So we obtain

\[
Z_n = -\zeta P(T_i(n) \leq n\zeta^2) = -\zeta P(\tau \leq n\zeta^2) + o_p(1) = -\zeta + o_p(1).
\]

This estimate follows (Lemma 2). Therefore, the optimizer \( \zeta_n^* \), satisfies \( \zeta_n^* = -Z_n + o_p(1) \). Then, we plug it into the objective function to obtain that the scaled SOS profile function satisfies

\[
\begin{align*}
nR_n^W(\theta_*) &= 2Z_n^2 - \mathbb{E} \left[ \max \left( Z_n^2 - T_i^2 (n) /n, 0 \right) \right] + o_p(1) \\
&= 2Z_n^2 - \int_{-Z_n^2}^{Z_n^2} \mathbb{P} \left[ T_i^2 (n) \leq n \left( Z_n^2 - t \right) \right] dt + o_p(1) \\
&= 2Z_n^2 - \int_{0}^{Z_n^2} \mathbb{P} \left[ \tau^2 \leq n \left( Z_n^2 - t \right) \right] dt + o_p(1) \\
&= Z_n^2 + o_p(1).
\end{align*}
\]
For the above claim, in the second equation we are using that \( E \|X\| = \int_0^\infty P \|X\| \geq t \, dt \), the third equality follows due to an application of Lemma 2, and the last equality is made by simple algebra, thus applying the continuous mapping theorem and convergence together results we obtain

\[
nR_n^W(\theta_*) \Rightarrow \sigma^2 \chi^2_1.
\]

3.2.5. **Step 5 (Case \( l = 2 \)):** Once again we introduce the substitution \( \zeta = \sqrt{n}\lambda \) and \( \sqrt{n}X_n = Z_n \) into \((21)\). Then, scaling the profile function by \( n \), we have

\[
nR_n^W(\theta_*) = \max_{\zeta} \left\{ -2\zeta^T Z_n - \frac{1}{n} \sum_{i=1}^n \max \left( \|\zeta\|_2^2 - T_i(n), 0 \right) \right\}
\]

\[
= \max_{\zeta} \left\{ -2\zeta^T Z_n - \mathbb{E} \left[ \max \left( \|\zeta\|_2^2 - T_i(n), 0 \right) \right] \right\} + o_p(1),
\]

where the previous estimate follows by applying Lemma 1 (the error is obtained by localizing \( \zeta \) on a compact set, which is valid because the sequence of global optimizers is easily seen to be tight). The objective function is strictly convex as a function of \( \zeta \) 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. By taking the derivative \( \zeta \) and set it to zero, applying dominating convergence theorem as for \( l = 1 \), we have

\[
Z_n = -\zeta \mathbb{P} \left[ \|\zeta\|_2^2 \geq T_i(n) \right] = -\zeta \mathbb{P} \left[ \|\zeta\|_2^2 \geq \tau \right] + o_p(1).
\]

The previous estimate follows by applying Lemma 2. Using equation \((23)\), we conclude that the optimizer \( \zeta^*_n \), satisfies \( \zeta^*_n = -\rho Z_n + o_p(1) \), for some \( \rho \). In turn, plugging in this representation into equation \((23)\), we have

\[
\|\zeta^*_n\|_2 \mathbb{P} \left[ \|\zeta^*_n\|_2^2 \geq \tau \right] + o_p(1) = \|Z_n\|_2.
\]

Sending \( n \to \infty \), we conclude that \( \rho \) is the unique solution to

\[
\frac{1}{\rho} = \mathbb{P} \left[ \rho^2 \left\| \tilde{Z} \right\|_2^2 \geq \tau \right] \left\| \tilde{Z} \right\|_2^2.
\]

Since the objective function is strict convex and the above equation is derived from first order optimality condition, we know the solution exists and is unique (alternatively we can use the continuity and monotonicity of left and right hand side of \((24)\), 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
\[ nR^W_n(\theta_s) = 2\rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \left\| Z_n \right\|_2^2 - \mathbb{E} \left[ \max \left( \left\| \zeta_n^* \right\|_2^2 - T_i(n), 0 \right) \right] \left\| Z_n \right\|_2^2 + o_p(1). \]

Let us denote
\[ A_2 = \mathbb{E} \left[ \max \left( \left\| \zeta_n^* \right\|_2^2 - T_i(n), 0 \right) \right] \left\| Z_n \right\|_2^2. \]

For a positive random variable \( Y \), we have: \( \mathbb{E}[Y] = \int_0^\infty \mathbb{P}[Y \geq t] \, dt \). Therefore,
\[
A_2 = \int_0^{\left\| \zeta_n^* \right\|_2^2} \mathbb{P} \left[ \left\| \zeta_n^* \right\|_2^2 - T_i(n) \geq t \right| \left\| Z_n \right\|_2^2 \right] \, dt
\]
\[
= \int_0^{\left\| \zeta_n^* \right\|_2^2} \mathbb{P} \left[ \left\| \zeta_n^* \right\|_2^2 - \tau \geq t \right| \left\| Z_n \right\|_2^2 \right] \, dt + o_p(1)
\]
\[
= \left\| \zeta_n^* \right\|_2^2 \int_0^1 \mathbb{P} \left[ 1 - \frac{\tau}{\rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \left\| Z_n \right\|_2^2} \geq s \right| \left\| Z_n \right\|_2^2 \right] \, ds + o_p(1)
\]
\[
= \rho \left( \left\| \tilde{Z} \right\|_2^2 \right)^2 \left\| Z_n \right\|_2^2 \mathbb{E} \left[ \max \left( 1 - \frac{\tau}{\rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \left\| Z_n \right\|_2^2}, 0 \right) \right] \left\| Z_n \right\|_2^2 + o_p(1)
\]
\[
= \rho \left( \left\| \tilde{Z} \right\|_2^2 \right)^2 \eta \left( \left\| Z_n \right\|_2^2 \right) \left\| Z_n \right\|_2^2 + o_p(1)
\]

The second equation follows (Lemma 2). Finally combine \( A_2 \) term and the first term, using the CLT and continuous mapping theorem, where we denote \( Z_n \Rightarrow \tilde{Z} \sim N(0, \text{Var}(X)) \), we have:
\[
nR^W_n(\theta_s) = \rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \left[ 2 - \eta \left( \left\| Z_n \right\|_2^2 \right) \rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \right] \left\| Z_n \right\|_2^2 + o_p(1)
\]
\[
\Rightarrow \rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \left[ 2 - \eta \left( \left\| \tilde{Z} \right\|_2^2 \right) \rho \left( \left\| \tilde{Z} \right\|_2^2 \right) \right] \left\| \tilde{Z} \right\|_2^2.
\]

3.2.6. **Step 6 (Case \( l \geq 3 \):** For simplicity, let us write \( \sqrt{n}X_n = Z_n \) and \( n^{\frac{2}{l+2}} \lambda = \zeta \), then we have
where we denote

\[ \frac{n^{1/2+\frac{3}{2(l+2)}}}{P_n(\theta_s)} \]

\[ = \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{(1/2+\frac{3}{2(l+2)}-\frac{3}{l})} \frac{1}{n} \sum_{i=1}^{n} \max \left( \left\| \frac{\zeta}{n^{(\frac{3}{2(l+2)}-\frac{3}{l})}} \right\|^2 - T_i^{2/l}(n), 0 \right) \right\} \]

\[ = \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{(1/2+\frac{3}{2(l+2)}-\frac{3}{l})} \mathbb{E} \left[ \max \left( \left\| \frac{\zeta}{n^{(\frac{3}{2(l+2)}-\frac{3}{l})}} \right\|^2 - T_i^{2/l}(n), 0 \right) \right] \right\} + o_p(1). \]

The estimate in the previous display is due to an application of Lemma 1. As we discussed for the case \( l = 2 \) case, the objective function is strictly convex in \( \zeta \), the global optimizers are not only tight, but each is also characterized by first order optimality conditions almost surely. Then by taking the derivative w.r.t \( \zeta \) and set it equal to zero, and applying the dominating convergence theorem as we discussed for \( l = 1 \), we have

\[ Z_n = -n^{(1/2-\frac{3}{2(l+2)})} \zeta \mathbb{P} \left[ T_i(n) \leq \left\| \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right\|^l \right] \]

\[ = -n^{(1/2-\frac{3}{2(l+2)})} \zeta \mathbb{P} \left[ \tau \leq \left\| \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right\|^l \right] + o_p(1), \]

where the second equation follows by Lemma 2. For fixed \( \zeta \), we know \( \left\| \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right\|^l \to 0 \), we can write

\[ \mathbb{P} \left[ \tau \leq \left\| \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right\|^l \right] = 1 - \mathbb{P} \left[ \tau > \left\| \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right\|^l \right] \]

\[ = 1 - \mathbb{E} \left[ \mathbb{P} \left[ \tau > \left\| \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right\|^l \right| X_1 \right] \]

\[ = \mathbb{E} \left[ 1 - \exp \left( -\frac{\pi^{l/2}}{\Gamma(l/2+1)} \frac{f_X(X_1 + \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)}) + f_Y(X_1 + \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)})}{\Gamma(l/2+1)} \left\| \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right\|^l \right) \right] \]

\[ = \mathbb{E} \left[ \frac{\pi^{l/2}}{\Gamma(l/2+1)} \left| f_X(X_1) + f_Y(X_1) \right| \left\| \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right\|^l \right] + o_p \left( n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right) \]

\[ = C \left\| \zeta n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right\|^l + o_p \left( n^{-\left(\frac{3}{2(l+2)}-\frac{1}{l}\right)} \right), \]

where we denote

\[ C = \frac{\pi^{l/2}}{\Gamma(l/2+1)} \mathbb{E} \left| f_X(X_1) + f_Y(X_1) \right|. \]
Plug it back into the optimizer, and we have:

\[
Z_n = -C n^{(1/2 - 3/2^2/2 + 1)} \zeta ||\zeta||^l + o_p(1) = -C \zeta ||\zeta||^l + o_p(1)
\]

We know that within the object function, the second term is only based on the \(L_2\) norm of \(\zeta\), thus to maximize the objective function we will asymptotically select \(\zeta^* = -c_*Z_n (1 + o(1))\), where \(c_* > 0\) is suitably chosen, thus, we conclude that the optimizer takes the form,

\[
\zeta^* = -Z_n \frac{||Z_n||^l}{C} + o_p(1).
\]

Plugging-in the optimizer back into the objective function we have:

\[
n^{1/2 + 3/2^2/2} R_n^W (\theta_s) = -2\zeta^T Z_n n^{(1/2 + 3/2^2/2 - 2^2/2 + 1)} \mathbb{E} \left[ \max \left( \left| \left| \frac{\zeta^*}{n^{1/2 + 2^2/2 - 1}} \right| - T_1^{2/l} (n), 0 \right| \right| ||Z_n||_2 \right] + o_p(1).
\]

Let us focus on the second term in the right hand side, which we call \(B2\), that is

\[
B2 = n^{(1/2 + 3/2^2/2 - 2^2/2 + 1)} \mathbb{E} \left[ \max \left( \left| \left| \frac{\zeta^*}{n^{1/2 + 2^2/2 - 1}} \right| - T_1^{2/l} (n), 0 \right| \right| ||Z_n||_2 \right].
\]
We can notice that inside the previous expectation there is a positive random variable bounded by
\[ \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^2, \]
thus we can write the expectation term as
\[
\mathbb{E} \left[ \max \left( \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^2 - T_1^{2/l} (n), 0 \right) \right] ||Z_n||_2
\]
\[
= \mathbb{E} \left[ \mathbb{E} \left[ \max \left( \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^2 - T_1^{2/l} (n), 0 \right) \right] \right] \mathbb{P} \left[ T_1 (n) \leq \left( \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^2 - t \right] \right] \frac{1}{l/2} X_1 \right) \right] dt ||Z_n||_2
\]
\[
= \mathbb{E} \left[ \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^2 \left( 1 - e^{-\frac{\pi^{1/2} (f_X (X_1 + \frac{\zeta}{n^{(2/1+\frac{1}{l})}}) + f_Y (X_1 + \frac{\zeta}{n^{(2/1+\frac{1}{l})}})) \left( \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^2 - t \right) \right] \right) \right] dt ||Z_n||_2
\]
\[
+ O \left( 1/n^{-1/2+3/l-6/2+1/2} \right).
\]

The estimate in third equation follows by applying Lemma 2. Since we know that \( \left\| \zeta_n^{-(3/2+1/2)} \right\|^2 \ll 1 \) as \( n \to \infty \), we have
\[
\mathbb{E} \left[ \max \left( \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^2 - T_1^{2/l} (n), 0 \right) \right] ||Z_n||_2
\]
\[
= \mathbb{E} \left[ \left( \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^2 \right) \left( \frac{\pi^{1/2}}{\Gamma(l/2+1)} \left( f_X (X_1 + \frac{\zeta}{n^{(2/1+\frac{1}{l})}}) + f_Y (X_1 + \frac{\zeta}{n^{(2/1+\frac{1}{l})}}) \right) \left( \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^2 - t \right) \right) \right] dt \right] ||Z_n||_2
\]
\[
+ O \left( 1/n^{-1/2+3/l-6/2+1/2} \right)
\]
\[
= C \frac{2}{l+2} \left\| \frac{\zeta}{n^{(2/1+\frac{1}{l})}} \right\|^{l+2} + O \left( 1/n^{-1/2+3/l-6/2+1/2} \right).
\]
Then, owing to the previous estimates, $B2$ becomes

$$B2 = -\frac{2C}{l+2} n^{(1/2+\frac{9}{2l+2}-\frac{4}{l})} \left( -\frac{9}{2l+2} \frac{4}{l} \right) ||\zeta||^{l+2} + o_p(1)$$

$$= -\frac{2C}{l+2} \frac{||Z_n||^{l+1}}{C_1^{l+1}} + o_p(1) = -\frac{2}{l+2} \frac{||Z_n||^{l+1}}{C_1^{l+1}} + o_p(1).$$

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

$$n^{1/2+\frac{7}{2l+2}} R_n^W(\theta) = \frac{2l + 2}{l+2} \frac{||Z_n||^{l+1}}{C_1^{l+1}} \Rightarrow \frac{2l + 2}{l+2} \frac{||\tilde{Z}||^{l+1}}{C_1^{l+1}}.$$

3.2.7. 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 $t \geq 0$,

$$F_n(t) = P(T_i(n) \leq t)$$

$$D_i(t) = I(T_i(n) \leq t), \quad \tilde{D}_i(t) = I(T_i(n) \leq t) - F_n(t),$$

$$\bar{F}_n(t) = 1 + n^{-1/2} \sum_{i=1}^{n} \tilde{D}_i(t).$$

Therefore, we are interested in studying

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

We will start by studying

$$\mathbb{E}[\sup\{F_n(t) : t \in [0,T]\}].$$

First, we define

$$h_n(t) = \frac{\bar{F}_n(t_-)}{\left( \bar{F}_n^*(t_-)^2 + [\bar{F}_n(t_-)] \right)^{1/2}},$$

where, for a given function $\{g(t) : t \in [0,T]\}$, we define

$$g^*(t) = \sup\{g(s) : s \in [0,t]\},$$

$$[g](t) = \int_{0}^{t} (dg(s))^2.$$
In particular,
\[ [\tilde{F}_n](t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i(n) \leq t). \]

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

\[ h_n(t)^2 \leq 1. \]

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

\[ \sup_{0 \leq t \leq T} \tilde{F}_n(t) \leq 6\sqrt{[\tilde{F}_n](T)} + 2 \int_{0}^{T} h_n(t) d\tilde{F}_n(t). \]

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

\[ \mathbb{E} \left( \int_{0}^{T} h_n(t) d\tilde{F}_n(t) \right) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \mathbb{E} \left( \int_{0}^{T} h_n(t) d\tilde{D}_i(t) \right), \]

\[ = n^{1/2} \mathbb{E} \left( \int_{0}^{T} h_n(t) d\tilde{D}_1(t) \right). \]

(25)

Let us write

\[ 1\tilde{F}_n(t) = \bar{F}_n(t) - \bar{D}_1(t) / n^{1/2}, \]

that is, we simply remove the last term in the sum defining \( \bar{F}_n(t) \). We have that

\[ h_n(t) = \frac{1\tilde{F}_n(t_\cdot) + \bar{D}_1(t_\cdot) / n^{1/2}}{\left( \bar{F}_n^*(t_\cdot) + [1\tilde{F}_n](t_\cdot) + [\bar{D}_1](t_\cdot) / n \right)^{1/2}}, \]

moreover,

\[ |1\tilde{F}_n^*(t) - \bar{F}_n^*(t)| \leq 1/n^{1/2}. \]

We then can write

(26) \[ h_n(t) = \frac{1\tilde{F}_n(t_\cdot) + \bar{D}_1(t_\cdot) / n^{1/2}}{\left( \bar{F}_n^*(t_\cdot) + [1\tilde{F}_n](t_\cdot) + [\bar{D}_1](t_\cdot) / n \right)^{1/2}} = \frac{1\tilde{F}_n(t_\cdot)}{\left( 1\tilde{F}_n^*(t_\cdot) + [1\tilde{F}_n](t_\cdot) \right)^{1/2}} \left( 1 + \frac{L_n(t_\cdot)}{n^{1/2}} \right), \]

where we can select a deterministic constant \( c \in (0, \infty) \) such that \( |L_n(t)| \leq c \) for \( j = 0 \) and 1 assuming \( n \geq 4 \) (this constrain in \( n \) is imposed so that a Taylor expansion for the function \( 1/(1-x) \)
can be developed for $x \in (0, 1)$. We now insert (26) into (25) and conclude that if we define

$$
\tilde{h}_n (t) = \frac{1_{F_n} (t_\omega)}{\left(1_{F_n}^2 (t_\omega)^2 + [1_{F_n} (t_\omega)]^2\right)^{1/2}},
$$

it suffices to verify that

$$
n^{1/2}E \left( \int_0^T \tilde{h}_n (t) d\bar{D}_1 (t) \right) < \infty.
$$

Define $\tilde{h}_n (t)$ to be a copy of $\bar{h}_n (t)$, independent of $X_1$ and $T_1 (n)$. In particular, $\tilde{h}_n (t)$ 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 $\{\tilde{D}_i (t) : t \leq T\}$ that depend on $T_1 (n)$ and $X_1$ is smaller than $N_n (T, \lambda, 1)$. Therefore, similarly as we obtained from the analysis leading to the definition of $\bar{h}_n (\cdot)$, we have that a random variable $\bar{L}_{N_n (T, \lambda, 1)}$ can be defined so that $|\bar{L}_{N_n (T, \lambda, 1)}| \leq c (1 + N_n (T, \lambda, 1))$ for some (deterministic) $c > 0$ and $n \geq 4$ and satisfying

$$
E \left( \int_0^T \tilde{h}_n (t) d\bar{D}_1 (t) \right)
= E \left( \bar{h}_n (T_1 (n)) I (T_1 (n) \leq T) \right) - E \left( \bar{h}_n (T_1 (n)) I (T_1 (n) \leq T) \right)
= E \left( \bar{h}_n (T_1 (n)) I (T_1 (n) \leq T) \right) - E \left( \bar{h}_n (\tau_i (X_i)) I (\tau_i (X_i) \leq T) \right)
+ E \left( \bar{L}_{N_n (T, \lambda, 1)} / n^{1/2} \right)
= E \left( \bar{L}_{N_n (T, \lambda, 1)} / n^{1/2} \right).
$$

We have that

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

Consequently, we conclude that

$$
n^{1/2}E \left( \int_0^T h_n (t) d\bar{D}_1 (t) \right) = O (1)
$$
as \( n \to \infty \), as required. Thus we proved that the first part of the lemma holds. For the second part, we observe that
\[
\lim_{n \to \infty} E \left( \sup_{t \in [0,T]} \left( \frac{1}{n^{1/2}} \sum_{i=1}^{n} (\max (t^2 - T_i(n)\alpha, 0) - E \left[ \max (t^2 - T_i(n)\alpha, 0) \right]) \right) \right) \\
= \lim_{n \to \infty} E \left( \sup_{t \in [0,T]} \left( \int_{0}^{t} \frac{1}{n^{1/2}} \sum_{i=1}^{n} (2sI (T_i\alpha(n) \leq s^2) - 2sP[T_i\alpha(n) \leq s^2]) ds \right) \right) \\
\leq \lim_{n \to \infty} \int_{0}^{T} E \left( \sup_{t \in [0,T]} \left( \frac{1}{n^{1/2}} \sum_{i=1}^{n} (2tI (T_i\alpha(n) \leq t^2) - 2tP[T_i\alpha(n) \leq t^2]) \right) \right) dt \\
\leq 2T^2 \lim_{n \to \infty} E \left( \sup_{t \in [0,T]} \left( \frac{1}{n^{1/2}} \sum_{i=1}^{n} (I (T_i(n) \leq t) - P[T_i(n) \leq t]) \right) \right) < \infty.
\]
Hence, applying the result for the first part of the lemma, we conclude the second part as well. \( \Box \)

**Proof of Lemma 2.**
\[
P[T_i(n) \leq t] = P \left( Bin \left( n - 1, P \left( \|X_i - a (X_i, \lambda)\|_2 \leq t^{1/l}/n^{1/l} \right) \right) \geq 1 \right) \\
= 1 - \left( 1 - P \left( \|X_i - a (X_i, \lambda)\|_2 \leq t^{1/l}/n^{1/l} \right) \right)^n.
\]
Then,
\[
P \left( \|X_j - a (X_i, \lambda)\|_2 \leq t^{1/l}/n^{1/l} \right) = c_0 t/n + c_1 t/n \cdot t^{1/l}/n^{1/l} + o \left( t^{1+1/l}/n^{1+1/l} \right).
\]
Therefore by the Poisson approximation to the Binomial distribution we know:
\[
P[T_i(n) \leq t] = 1 - \exp (-c_0 t) + O \left( t^{1+1/l}/n^{1/l} \right) \\
P[\tau \leq t] = 1 - \exp (-c_0 t).
\]
Thus we proved the claim:
\[
P[T_i(n) \leq t] - P[\tau \leq t] = O \left( t^{1+1/l}/n^{1/l} \right).
\]
\( \Box \)

### 3.3. 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.
3.3.1. **Proofs of SOS Theorems for General Estimation.** We will first prove the corresponding theorems for general estimating equations. As we discussed before, Theorem 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 \( \bar{h}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} h(\theta, X_i) \). The analogue of Step 1, namely, the dual formulation takes the form

\[
R_W^W(\theta^*) = \max_{\lambda} \left\{ -\lambda^T \bar{h}_n(\theta^*) - \frac{1}{n} \sum_{i=1}^{n} \max_j \left\{ \lambda^T h(\theta^*, Z_j) - \lambda^T h(\theta^*, X_i) - ||X_i - Z_j||^2_2 \right\}^+ \right\}
\]

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

\[
n R_W^W(\theta^*) = \max_{\zeta} \left\{ -2\zeta Z_n - \frac{1}{n} \sum_{i=1}^{n} n \max_j \left\{ 2\zeta^T h(\theta^*, Z_j) - 2\zeta^T h(\theta^*, X_i) - ||X_i - Z_j||^2_2 \right\}^+ \right\}.
\]

For each \( i \), let us consider the maximization problem

\[
(27) \quad \max_j \left\{ 2\zeta^T h(\theta^*, Z_j) - 2\zeta^T h(\theta^*, X_i) - ||X_i - Z_j||^2_2 \right\}.
\]

Similar as Step 1 of the proof for Theorem 1, we would like to solve the maximization problem (27) by first minimizing over \( z \) (as a free variable), instead of over \( j \) and then quantify the gap. Observe that the uniform bound \( ||D^2_x h(\theta^*, \cdot)|| < \tilde{K} \) stated in BE1) implies that for all \( n \) large enough (in particular, \( n^{1/2} > 2\tilde{K} ||\zeta|| \)) implies that

\[
(28) \quad \max_z \left\{ 2\zeta^T h(\theta^*, z) - 2\zeta^T h(\theta^*, X_i) - ||X_i - z||^2_2 \right\},
\]

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 (28) satisfies

\[
\bar{a}_*(X_i, \zeta) = X_i + D_x h(\theta^*, \bar{a}_*(X_i, \zeta))^T \cdot \frac{\zeta}{n^{1/2}}
\]

\[
= X_i + D_x h(\theta^*, X_i)^T \cdot \frac{\zeta}{n^{1/2}} + O \left( \frac{||\zeta||^2_2}{n} \right). \quad (29)
\]

Moreover, owing to BE1), we obtain that

\[
(30) \quad ||D_x h(\theta^*, \bar{a}_*(X_i, \zeta)) - D_x h(\theta^*, X_i)|| \leq \tilde{K} \frac{||\zeta||}{n^{1/2}}.
\]
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 (29) and (30) that
\[ \|a_*(X_i, \zeta) - \bar{a}_*(X_i, \zeta)\| = O\left(\frac{\|\zeta\|^2}{n} \left(\|D_x h(\theta_*, X_i)\| + \frac{\|\zeta\|}{n^{1/2}}\right)\right). \]

Then, using after performing a Taylor expansion and applying inequality (30) we obtain that
\[
\begin{align*}
2 \zeta^T \frac{\zeta}{\sqrt{n}} h(\theta_*, X_i) - 2 \zeta^T \frac{\zeta}{\sqrt{n}} h(\theta_*, a_*(X_i, \zeta)) + \|X_i - a_*(X_i, \zeta)\|^2 \ + \ O\left(\frac{\|D_x h(\theta_*, X_i)\|^2 \|\zeta\|^3}{n^{3/2}}\right).
\end{align*}
\]

In turn, a direct calculation gives that
\[
-\zeta^T V_i \zeta/n = 2 \zeta^T \frac{\zeta}{\sqrt{n}} h(\theta_*, X_i) - 2 \zeta^T \frac{\zeta}{\sqrt{n}} h(\theta_*, a_*(X_i, \zeta)) + \|X_i - a_*(X_i, \zeta)\|^2 + O\left(\frac{\|D_x h(\theta_*, X_i)\|^2 \|\zeta\|^3}{n^{3/2}}\right).
\]

Similarly as in Step 2 of the proof of Theorem 1 we can define the point process \( N_l^i(t, \zeta) \) and \( T_l(i) \). We know the gap between freeing the variable \( z \) and restricting the maximization over the \( Z_j \)'s (i.e. the difference between (28) and (27)) is
\[
\max_j \left\{ \frac{1}{n} \zeta^T V_i \zeta - \left(2 \zeta^T \frac{\zeta}{\sqrt{n}} h(\theta_*, Z_j) - 2 \zeta^T \frac{\zeta}{\sqrt{n}} h(\theta_*, X_i) - \|X_i - Z_j\|^2 \right) \right\} + O\left(\frac{\|D_x h(\theta_*, X_i)\|^2 \|\zeta\|^3}{n^{3/2}}\right).
\]

By the definition of \( T_l(i) \), we can write the profile function for \( l = 1 \) as
\[
nR_l^n(\theta_*) = \max_{\zeta} \left\{ -2 \zeta^T Z_n - \frac{1}{n} \sum_{i=1}^n \max \left( \zeta^T V_i \zeta - \frac{T_l^2(n)}{n}, 0 \right) + O\left(\frac{\|D_x h(\theta_*, X_i)\|^2 \|\zeta\|^3}{n^{1/2}}\right) \right\}.
\]

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 $\zeta$ can be restricted to compact sets, that

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

Then, for $l = 2$, the profile function is

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

When $l \geq 3$, let us denote $\sqrt{n} h_n(\theta_*) = Z_n$ 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

$$n^{\frac{1}{2} + \frac{3}{2l+2}} R_n^W(\theta_*)$$

$$= \max_{\zeta} \left\{ -2\zeta^T Z_n - \frac{1}{n} \sum_{i=1}^{n} n^{\frac{1}{2} + \frac{3}{2l+2}} \max_{j} \left\{ 2 \frac{\zeta^T}{n^{\frac{3}{2l+2}}} h(\theta_*, Z_j) - 2 \frac{\zeta^T}{n^{\frac{3}{2l+2}}} h(\theta_*, X_i) - ||X_i - Z_j||_2^2 \right\} \right\} + o_p(1).$$

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. We have the profile function becomes

$$n^{\frac{1}{2} + \frac{3}{2l+2}} R_n^W(\theta_*)$$

$$= \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{\frac{1}{2} + \frac{3}{2l+2}} \frac{1}{n} \sum_{i=1}^{n} \max \left( n^{-(\frac{6}{2l+2} - \frac{3}{2})} \zeta^T V_i \zeta - T_i^{2/l}(n), 0 \right) \right\} + o_p(1)$$

$$= \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{\frac{1}{2} + \frac{3}{2l+2}} \frac{2}{7} \mathbb{E} \left[ \max \left( n^{-(\frac{6}{2l+2} - \frac{3}{2})} \zeta^T V_i \zeta - T_i^{2/l}(n), 0 \right) \right] \right\} + o_p(1).$$

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

In **Step 4** for $l = 1$, the objective function is

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

We know $\Upsilon = \mathbb{E} [V_i]$ is symmetric strictly positive definite matrix, then the objective function is strictly convex and differentiable in $\zeta$. Thus the (unique) global maximizer is characterized by the first order optimality condition almost surely. We take derivative w.r.t. $\zeta$ and set it to be 0, applying the same estimation in the original proof the first order optimality condition becomes

$$Z_n = -\Upsilon \zeta + o_p(1).$$
Since $\Upsilon$ is invertible, for any $n$ we can solve optimal $\zeta_n^* = -\Upsilon^{-1}Z_n$. Plugging $\zeta_n^*$ in the objective function we have

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

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 $l = 2$, the objective function is

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

Same as discussed in for $l = 1$, the objective function is strictly convex and differentiable in $\zeta$, thus the (unique) global maximizer could be characterized via first order optimality condition almost surely. We take derivative w.r.t. $\zeta$ and set it to be 0, applying same estimation in the proof of Theorem 1 the first order optimality condition becomes

$$(32) \quad Z_n = -\mathbb{E} \left[ V_1 1_{\tau \leq \zeta^T V_1 \zeta} \right] \zeta + o_p(1)$$

We know the objective function is strictly convex differentiable, then for fixed $Z_n$ there is a unique $\zeta_n^*$ that satisfies the first order optimality condition (32). We plug in the optimizer and the objective function becomes

$$nR_n^W(\theta^*) = -2Z_n^T \zeta_n^* - \mathbb{E} \left[ \max \left( \zeta_n^*^T V_1 \zeta_n^* - T_1(n) \right) \right] + o_p(1)$$

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

$$nR_n^W(\theta^*) \Rightarrow -2\tilde{Z}^T \tilde{\zeta} - \tilde{\zeta}^T \mathbb{E} \left[ V_1 \max \left( 1 - \frac{\tau}{\zeta^T V_1 \zeta}, 0 \right) \right] \tilde{\zeta}$$

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

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

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

Finally, in Step 6 for \( l \geq 3 \), the objective function is

\[
n^{1/2 + \frac{3}{2 + \sqrt{2}} - \frac{3}{2} + \frac{3}{2 + \sqrt{2}} + 2R_W n(\theta_*)} = \max_{\zeta} \left\{ -2\zeta^T Z_n - n^{(1/2 + \frac{3}{2 + \sqrt{2}} - \frac{3}{2} + \frac{3}{2 + \sqrt{2}} + 2R_W n(\theta_*)} \right\}
\]

Follows the same discussion above for \( l = 1 \) and 2, we know the objective function is strictly convex differentiable in \( \zeta \) and the global maximizer is characterized by first order optimality condition almost surely. We take derivative of the objective function w.r.t. \( \zeta \) and set it to be 0. We apply the same technique as in the proof of Theorem 1, the first order optimality condition becomes

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

The objective condition is strictly convex differentiable and for fixed \( Z_n \) there is a unique \( \zeta_n^* \) satisfying the first optimality condition (33). We plug \( \zeta_n^* \) into the objective function and it becomes

\[
n^{1/2 + \frac{3}{2 + \sqrt{2}} - \frac{3}{2} + \frac{3}{2 + \sqrt{2}} + 2R_W n(\theta_*)} = -2Z_n^{l/2} \zeta_n^* - n^{(1/2 + \frac{3}{2 + \sqrt{2}} - \frac{3}{2} + \frac{3}{2 + \sqrt{2}} + 2R_W n(\theta_*)} \mathbb{E} \left[ \max \left( n^{-\frac{3}{2 + \sqrt{2}} - \frac{3}{2} + \frac{3}{2 + \sqrt{2}} + 2R_W n(\theta_*)} \zeta_n^* V_1 \zeta_n^* - T^{2/l} (n) , 0 \right) \right] + o_p(1)
\]

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

\[
n^{1/2 + \frac{3}{2 + \sqrt{2}} - \frac{3}{2} + \frac{3}{2 + \sqrt{2}} + 2R_W n(\theta_*)} \Rightarrow -2\bar{Z}^T \bar{\zeta} - \frac{2}{l + 2} \mathbb{E} \left[ \frac{\pi^{l/2} (f_X(X_1) + \kappa f_Y(X_1))}{\Gamma(l/2 + 1)} \left( \bar{\zeta}^T V_1 \bar{\zeta} \right)^{1/2 + 1} \right] \bar{Z},
\]

where \( \bar{\zeta} := \bar{\zeta}(\bar{Z}) \) is the unique solution to

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

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

---

3.3.2. Proofs of SOS Theorems for General Estimation with Plug-In. The proofs of the plug-in version of SOS theorems for general estimation equation also mainly follows the proof of Theorem 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

\[
X_i \leftarrow h(\gamma_s, \nu_n, X_i), \ Z_k \leftarrow h(\gamma_s, \nu_n, Z_k), \ X_i(*) \leftarrow h(\gamma_s, \nu_s, X_i) \text{ and } Z_k(*) \leftarrow h(\gamma_s, \nu_s, X_i).
\]

Then the proof for the implicit formulation with plug-in goes as follows.
In Step 1, the dual formulation is similar given as
\[
R_n^W(\gamma^*) = \max_{\lambda, \gamma \geq 0} \left\{ -\lambda \bar{X}_n - \frac{1}{n} \sum_{i=1}^{n} \gamma_i \right\}
\]
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. \( \nu \), 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 \( \nu_n \) and \( \nu^* \). By our change of notation for \( X_i, X_i(\ast), Z_k \) and \( Z_k(\ast) \) and the above Taylor expansion, we can observe
\[
Z_k = Z_k(\ast) + \epsilon_n (Z_k),
\]
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
\[
\Lambda(X_i, \lambda) = [f_X(X_i + \lambda/2 + \epsilon_n(X_i)) + \kappa f_Y(X_i + \lambda/2 + \epsilon_n(X_i))] \pi^{l/2} \Gamma(l/2 + 1).
\]
As \( n \to \infty \), same the proof of Theorem 1 and Theorem 3 we can argue \( \lambda \to 0 \). Then we can define \( \tau \) same as in the proof of Theorem 1 and has the with same distribution
\[
P[\tau \geq t] = E \left[ \exp \left( -(f_X(X_1) + \kappa f_Y(X_1)) \frac{\pi^{l/2}}{\Gamma(l/2 + 1)} \right) \right].
\]
Then the rest of the proof in Step 3, 4, 5 and 6 stays the same as that of Theorem 1, but replacing the CLT for \( Z_n \) by asymptotic distribution given in C2). \( \square \)

**Proof of Corollary 2.** For explicit formulation, the proof is more close to the proof of Theorem 3 and we are discussing the difference as follows.

In Step 1, the dual formulation takes the form
\[
R_n^W(\theta^*)
\]
\[
= \max_{\lambda} \left\{ -\lambda^T \bar{h}_n(\gamma^*, \nu_n) - \frac{1}{n} \sum_{i=1}^{n} \max_j \left\{ \lambda^T h(\gamma^*, \nu_n, Z_j) - \lambda^T h(\gamma^*, \nu_n, X_i) - ||X_i - Z_j||_2 \right\}^+ \right\}
\]
Step 2 and 3 Follows the same as for the proof of Theorem 3 however we need to notice that difference is the definition of $\bar{a}_s(X_i, \zeta)$, for $l = 1$ and 2 we have

$$
\bar{a}_s(X_i, \zeta) = X_i + D_x h(\gamma_s, \nu_n, \bar{a}_s(X_i, \zeta)) \cdot \frac{\zeta}{n^{1/2}}
$$

$$
= X_i + D_x h(\gamma_s, \nu_n, X_i) \cdot \frac{\zeta}{n^{1/2}} + O\left(\frac{\|\zeta\|^2}{n} \|D_x h(\gamma_s, \nu_n, \bar{a}_s(X_i, \zeta))\|\right)
$$

$$
= X_i + D_x h(\gamma_s, \nu_s, X_i) \cdot \frac{\zeta}{n^{1/2}} + O\left(\frac{\|\zeta\|^2}{n} \|D_x h(\gamma_s, \nu_n, \bar{a}_s(X_i, \zeta))\|\right)
$$

$$
+ O\left(\frac{\|\zeta\|^2}{n^{1/2}} \|\nu_n - \nu_s\| \|D_x h(\gamma_s, \nu_n, \bar{a}_s(X_i, \zeta))\| \|D_v D_x h(\gamma_s, \bar{\nu}_n, \bar{a}_s(X_i, \zeta))\|\right),
$$

where $\bar{\nu}_n$ is a point between $\nu_n$ and $\nu_s$. 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 $Z_n = \frac{1}{n^{1/2}} \sum_{i=1}^n h(\gamma_s, \nu_n, X_i) \Rightarrow \tilde{Z}$ given in C2) instead of CLT. \hfill \square

4. Application to Stochastic Optimization and Stress Testing

We are going to provide an application of the SOS inference framework to quantify model uncertainty in the context of stochastic programming. As a motivating application we consider the problem of evaluating Conditional Value at Risk (C-VaR).

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

$$
C_* = \min_{\theta} \{ \mathbb{E}[m(\theta, X)] \}
$$

$$
s.t. \mathbb{E}[\phi(\theta, X)] \leq 0\}.
$$

We assume that the objective function $\psi(\theta) = \mathbb{E}[m(\theta, X)]$ is a convex function in $\theta$; while the constraints $\mathbb{E}[\phi(\theta, X)] \leq 0$ specify a convex region in $\theta$, for example we can assume $\phi(\theta, X)$ is a convex function in $\theta$ for any $X$.

Following Blanchet et al. (2016a), the goal is to estimate the optimal value function using the SOS formulation and we will apply a plug-in estimator for $\theta_*$ (which is treated as a nuisance parameter). Subsequently, when introducing the Lagrangian relaxation of (37) 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) 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, Lam and Zhou (2015, 2016) it is argued that the Empirical Likelihood method tends to have best finite sample performance, and Blanchet et al. (2016a) provides an optimal (in certain sense) specification for Empirical Likelihood approach. More importantly, in Blanchet et al. (2016a) an approach combining empirical likelihood and a plug-in estimator for optimizer is introduced, which avoids solving a non-convex optimization problem introduced in the discussion of Lam and Zhou (2015).

Our goal in this section is to derive a plug-in estimator based on the SOS inference approach introduced in Section 2. The approach that we introduce next is the analog of the plug-in strategy discussed in Blanchet et al. (2016a) in order to find a robust confidence interval for $C_*$. The following result is a direct extension of Corollary 1 and Corollary 2. This corollary plays the key role in specifying confidence interval for $C_*$. To ensure the corollary hold, we need some assumptions

**D1):** Assume $\psi(\cdot)$ is convex differentiable in $\theta$ and there is a unique optimizer $\theta_*$.

**D2):** Assume that $\psi(\cdot)$ is strongly convex at $\theta_*$, that is, for every $\theta$ there exist $\delta > 0$, such that

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

**Corollary 3.** [Plug-in for Implicit/Explicit SOS Function for Stochastic Optimization] Let us consider stochastic programming problem $C_* = \min_{\theta} M(\theta) = \min_{\theta} \mathbb{E}[m(\theta, X)]$. We assume assumption D1)-D2) hold. We consider the estimating equations to be the derivative condition and value function condition

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

For simplicity, let us denote $h(\theta_*, C_*, x) = \left(m(\theta_*, x) - C_*, D_{\theta}m(\theta_*, X)\right)^T$. We are interested in $C_*$ only and consider a sample average approximation (SAA) estimator for $\theta_*$ to be $\hat{\theta}_{SAA}$. For $h(\cdot, C_*, x)$ we assume C1)-C7) hold. Let us denote $U \sim N(0, \text{Var}(m(\theta_*, X))) \in \mathbb{R}$ and $U(0) = \left(U, 0\right)^T \in \mathbb{R}^{d+1}$. Recall the implicit and explicit formulations for general estimating equation SOS function defined in Definition 2 and Definition 3, we have the following asymptotic results.

For the implicit SOS formulation, we have
• When $d = 1$ (estimating equation dimension is $d + 1 = 2$)

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

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

$$\frac{1}{\rho} = \mathbb{P} \left[ \rho^2 U^2 \geq \tau | U \right]$$

$\tau$ is independent of $U$ satisfying

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

• When $d \geq 2$,

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

For the explicit formulation, we have following asymptotic results (we use $\zeta[1]$ denote first element of vector $\zeta$)

• When $l = 1$,

$$nR_n^W(C_*) \Rightarrow v_{1,1} U^2,$$

where $v_{1,1}$ is the $(1, 1)$ element of matrix $\Upsilon^{-1}$.

• Suppose that $l = 2$. It is possible to uniquely define $\tilde{\zeta} : \mathbb{R}^l \to \mathbb{R}^l$ continuous such that

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

where $\tau$ is independent of $U$ satisfying

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

Furthermore,

$$nR_n^W(\theta_*) \Rightarrow -2U \tilde{\zeta}[1] - \tilde{\zeta}^T \mathbb{E} \left[ \tilde{V}_1 \max \left( 1 - \frac{\tau}{\tilde{\zeta}^T \tilde{V}_1 \tilde{\zeta}}, 0 \right) | U \right] \tilde{\zeta}$$

where $\tilde{Z}'$ is independent with $\tilde{V}_1$ and $\tau$. 
• Assume that \( l \geq 3 \). A continuous function \( \tilde{\zeta} : \mathbb{R}^l \rightarrow \mathbb{R}^l \) can be defined uniquely so that

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

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

\[
n^{1/2 + \frac{3}{2l+2}} R_n W(\theta_*) \Rightarrow -2U \tilde{\zeta}[l] - \mathbb{E} \left[ \frac{\pi^{l/2}}{\Gamma(l/2 + 1)} f_X(X_1) \left( \tilde{\zeta}^T \tilde{V}_1 \tilde{\zeta} \right)^{l/2+1} \right],
\]

where \( \tilde{Z}^l \) and \( X_1 \) are independent.

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

We notice for sample average approximation algorithm, guaranteed by assumptions D1)-D3, it has been shown in Ruszczyński and Shapiro (2003) and Shapiro and Dentcheva (2014) the optimizer \( \hat{\theta}_{SAA} \) and optimal value function \( \frac{1}{n} \sum_{i=1}^{n} m \left( \hat{\theta}_{SAA}, X_i \right) \) have

\[
 \hat{\theta}_{SAA} - \theta_* = O \left( 1/n^{1/2} \right)
\]

\[
 \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta} m \left( \hat{\theta}_{SAA}, X_i \right) = 0,
\]

\[
 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( m \left( \hat{\theta}_{SAA}, X_i \right) - C_* \right) \Rightarrow N \left( 0, \text{Var} \left( m(\theta_*, X_i) \right) \right).
\]

Since Corollary 3 follows as a direct application of Corollary 2 and Corollary 1, its proof is omitted.

Similar as the derivation in Blanchet et al. (2016a) for empirical likelihood, for the plug-in estimator derived from sample average approximation, if we denote \( n^{1/2 + 3/(2d+4)} R_n^{W(\text{implicit})}(C_*) \Rightarrow R_0^{(\text{implicit})} \) and \( n^{1/2 + 3/(2l+2)} R_n^{W(\text{explicit})}(C_*) \Rightarrow R_0^{(\text{explicit})} \), we can specify a robust 95% confidence interval for \( C_* \) under both explicit and implicit formulation by:

\[
 \text{CI}^{(c)}(C_*) = \left\{ C \in \mathbb{R} \mid n^{\alpha} R_n^{W(\cdot)}(C) \leq R_0^{(c)}(95\%) \right\}
\]

where \( \alpha \) depends on the formulation and dimension as in Corollary 3 and \( R_0^{(c)}(95\%) \) is the upper 95% quantile for \( R_0^{(\text{explicit})} \) (or \( R_0^{(\text{implicit})} \)). The upper/lower bound of confidence interval (\( C_{up}/C_{lo}^{(c)} \)
can be found by solving the linear programming problem

\[
\frac{C_{up}(\cdot)}{C_{lo}(\cdot)} = \max_{\pi(i,j)} \frac{\min_{\pi(i,j)} \left\{ \sum_{i,j=1}^{n} \pi(i,j) m(\hat{S}_{SA}, X_i) \right\}}{\sum_{i,j=1}^{n} \pi(i,j) \left||X_i - X_j||_2^2 \right|} \geq R_0(95\%) .
\]

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

**Example 4** (Quantify the uncertainty of Conditional Value at Risk (C-VaR)). *In this example we would like to consider find a SOS based 95% confidence interval for conditional value at risk with 90% level. The conditional value at risk with \(\alpha\)-level is given as solving the stochastic programming problem:

\[
C-VaR(\alpha) = \inf_\theta \left\{ \theta + \frac{1}{1 - \alpha} \mathbb{E} \left[ \left( \sum_{k=1}^{l} X^{(k)} - \theta \right)^+ \right] \right\}.
\]

We shall test our method using simulated data under different distributional assumptions. We a sample i.i.d. observations \(\{X_i\}_{i=1}^{n} \subset \mathbb{R}^l\). We will apply the SOS inference procedure to provide a non-parametric confidence interval for C-VaR(90%). In order to verify the coverage probability we use data simulated from normal distribution and Laplace (double exponential) distributions. We consider the case \(l = 4\). For the normal distribution setting we assume \(X_i \sim N(0, I_{4 \times 4})\), while for Laplace distribution we consider for each \(k = 1, ..., 4\), \(X_i^k \sim \text{Laplace}(0, 1)\) 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 \(\theta^* = 2.5632\) and optimal value function is C-VaR(0.9) = 3.510; for Laplace setting the optimizer is \(\theta^* = 3.497\) with optimal value function equal to C-VaR(0.9) = 5.066.

As for this example, we have three approaches in which our SOS procedure can be applied: 1) implicit SOS formulation (ISOS); 2) explicit SOS formulation while assume underline data is \(l\) dimension (ESOS-O), i.e. \(X_i = (X_i^{(1)}, ..., X_i^{(l)})^T \in \mathbb{R}^l\); 3) explicit formulation while assume underline data is \(1\) dimension (ESOS-C), i.e. \(X_i = X_i^{(1)} + ... + X_i^{(l)} \in \mathbb{R}\). We compare our methods with empirical likelihood method (EL) in Blanchet et al. (2016a), nonparametric bootstrap method (BT), and CLT based Delta method (CLT) discussed in Theorem 5.7 Shapiro and Dentcheva (2014). We consider four settings \(n = 20, 50, 100\) and 200. For each setting, we repeat the experiment 1000 times, and
note down the empirical coverage probability, mean of upper and lower bounds, and the mean and standard deviation of the interval width for each method. The results are summarized in Table 1 for Normal distribution and Table 2 for Laplace distribution below.

We can observe that, the three SOS-based approaches tend to have better coverage probabilities in all cases for both distributions comparing to EL, bootstrap and the Delta method. Especially for small sample situations \((n = 10, 20)\) EL and all of the SOS-based approaches appear to perform better than everything else. It is discussed in Lam and Zhou (2015) and Blanchet et al. (2016a) that EL has better finite sample performance compared to the Delta method and bootstrap. We can also notice that all empirical SOS methods tend to have smaller variance compared to others, especially for relatively large sample sizes \((n = 100, 200)\). Between the three SOS methods, we can see that explicit formulations work better comparing to implicit, which follows our discussion after Definition 3. For the two explicit-formulation methods, since we know the data affects the objective function in the form \(X_{1}^{(1)} + \ldots + X_{i}^{(l)}\), we would expect better performance if we combine the data in a single dimension. The numerical results validates our intuition.

| \(n\) | Method | Coverage Probability | Mean Lower Bound | Mean Upper Bound | Mean Interval Length | S.D. of Length |
|---|---|---|---|---|---|---|
| 20 | ESOS-C | 80.3\% | 2.60 | 4.75 | 2.15 | 0.75 |
| | ESOS-O | 77.5\% | 2.56 | 4.68 | 2.12 | 1.05 |
| | ISOS | 79.2\% | 2.75 | 5.17 | 2.42 | 0.74 |
| | EL | 71.2\% | 2.63 | 5.15 | 2.52 | 1.74 |
| | BT | 56.7\% | 1.78 | 3.87 | 2.19 | 1.15 |
| | CLT | 70.9\% | 2.03 | 4.51 | 2.48 | 1.69 |
| 50 | ESOS-C | 93.3\% | 2.66 | 4.58 | 1.92 | 0.28 |
| | ESOS-O | 90.0\% | 2.64 | 4.55 | 1.91 | 0.53 |
| | ISOS | 84.3\% | 2.94 | 4.89 | 1.95 | 0.36 |
| | EL | 84.4\% | 2.82 | 4.80 | 1.98 | 0.80 |
| | BT | 80.6\% | 2.30 | 4.27 | 1.97 | 0.75 |
| | CLT | 84.0\% | 2.47 | 4.37 | 1.90 | 0.76 |
| 100 | ESOS-C | 96.3\% | 2.78 | 4.32 | 1.54 | 0.11 |
| | ESOS-O | 95.1\% | 2.76 | 4.31 | 1.55 | 0.26 |
| | ISOS | 90.1\% | 3.05 | 4.56 | 1.51 | 0.23 |
| | EL | 91.6\% | 2.93 | 4.46 | 1.53 | 0.42 |
| | BT | 89.9\% | 2.69 | 4.24 | 1.55 | 0.40 |
| | CLT | 90.3\% | 2.71 | 4.29 | 1.51 | 0.39 |
| 200 | ESOS-C | 98.3\% | 2.97 | 4.11 | 1.14 | 0.03 |
| | ESOS-O | 97.2\% | 2.96 | 4.08 | 1.12 | 0.11 |
| | ISOS | 89.9\% | 3.22 | 4.35 | 1.13 | 0.11 |
| | EL | 94.3\% | 2.94 | 4.21 | 1.12 | 0.23 |
| | BT | 90.2\% | 2.94 | 4.05 | 1.11 | 0.22 |
| | CLT | 91.5\% | 2.95 | 4.06 | 1.11 | 0.22 |

**Table 1.** \(\alpha = 0.9\)—Conditional Value at Risk with Gaussian Data
### Table 2. $\alpha = 0.9$–Conditional Value at Risk with Laplace Data

| $n$  | Method | Coverage Probability | Mean Lower Bound | Mean Upper Bound | Mean Interval Length | S.D. of Length |
|------|--------|----------------------|------------------|------------------|----------------------|---------------|
| 20   | ESOS-C | 78.2%                | 3.37             | 6.39             | 3.02                 | 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%                | 2.57             | 6.90             | 4.37                 | 3.21          |
| 50   | ESOS-C | 89.4%                | 3.78             | 6.64             | 2.86                 | 0.42          |
|      | ESOS-O | 89.3%                | 3.69             | 6.78             | 3.09                 | 0.89          |
|      | ISOS   | 80.1%                | 4.21             | 7.17             | 2.96                 | 0.63          |
|      | EL     | 86.2%                | 3.89             | 7.43             | 3.53                 | 1.66          |
|      | BT     | 80.5%                | 3.15             | 6.58             | 3.43                 | 1.54          |
|      | CLT    | 83.6%                | 3.29             | 6.64             | 3.35                 | 1.54          |
| 100  | ESOS-C | 91.9%                | 3.93             | 6.22             | 2.29                 | 0.14          |
|      | ESOS-O | 90.8%                | 3.88             | 6.30             | 2.42                 | 0.43          |
|      | ISOS   | 86.6%                | 4.30             | 6.78             | 2.44                 | 0.36          |
|      | EL     | 89.9%                | 4.10             | 6.66             | 2.56                 | 0.86          |
|      | BT     | 86.2%                | 3.71             | 6.16             | 2.45                 | 0.81          |
|      | CLT    | 87.6%                | 3.76             | 6.17             | 2.41                 | 0.79          |
| 200  | ESOS-C | 94.4%                | 4.25             | 5.90             | 1.65                 | 0.05          |
|      | ESOS-O | 93.1%                | 4.18             | 5.93             | 1.75                 | 0.19          |
|      | ISOS   | 90.9%                | 4.58             | 6.36             | 1.78                 | 0.20          |
|      | EL     | 94.0%                | 4.21             | 6.13             | 1.91                 | 0.44          |
|      | BT     | 91.0%                | 4.11             | 5.91             | 1.80                 | 0.41          |
|      | CLT    | 91.1%                | 4.12             | 5.91             | 1.79                 | 0.40          |

5. 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 methodology that we propose is motivated by the problem of systematically finding estimators which are incorporate out-of-sample performance in their design. In turn, as a motivation for the need of finding these types of estimators we discussed applications to stress testing. We envision this paper as the first installment on this research area and we plan to explore more deeply applications not only in stress testing but also in machine learning. For example, in our work [5], we study a connection between the estimation procedure that we introduce here and statistical techniques such as LASSO and support vector machine (SVM) which are popular in machine learning.

In [5] we also explore the limiting distribution obtained for the SOS function when we compare the empirical distribution against any other distribution, as opposed to only distributions supported on a finite set of scenarios and, in this case, we show that the distribution is typically chi-squared (so this case is, in some sense, closer to the Empirical Likelihood setting).

More importantly, there are a number of structural properties in our procedure that are worth investigating and which we plan to explore in future work. For instance, the choice of a particular
Wasserstein’s metric, we believe 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 what appears to also bring about interesting connections; this already is made evident from our work [5] 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 on this paper open up a significant amount of research opportunities which are parallel to the substantial literature produced in the area of Empirical Likelihood during the last three decades. We mention, in particular, applications to regression problems (see Owen (1991), Chen (1993), Chen (1994), Wang and Rao (2001), Zhao and Wang (2008) and Chen and Keilegom (2009)), survival analysis (see Murphy (1995), Li et al. (1996), Hollander and McKeague (1997), Li et al. (1997), Einmahl and McKeague (1999), Wang et al. (2009) and Zhou (2015)), econometrics (see Newey and Smith (2004), Bravo (2004), Kitamura (2006), Antoine et al. (2007), Guggenberger (2008) and Imbens (2012)) and additional recent work on stochastic optimization (see Lam and Zhou (2015), Lam and Zhou (2016) and Blanchet et al. (2016a)). 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).

ACKNOWLEDGMENT

Support from NSF through Award 1436700 is gratefully acknowledged. The authors gratefully acknowledge discussions with Karthyek Murthy and Henry Lam.

REFERENCES

[1] Antoine, B., Bonnal, H., and 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.

[2] Beiglbek, M. and Siorpaes, P. (2015). Pathwise versions of the BurkholderDavisGundy inequality. *Bernoulli*, 21(1):360–373.
[3] Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.

[4] Blanchet, J., Kang, Y., and Lam, H. (2016a). Quantify Uncertainty for Stochastic Programming Problem: an Empirical Likelihood Approach. Manuscript.

[5] Blanchet, J., Kang, Y., and Murthy, K. (2016b). Machine Learning Based On Robust Wasserstein Profile Inference. Manuscript.

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

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

[8] Chen, S. X. (1994). Empirical likelihood confidence intervals for linear regression coefficients. *Journal of Multivariate Analysis*, 49(1):24–40.

[9] Chen, S. X. and Keilegom, I. V. (2009). A review on empirical likelihood methods for regression. *TEST*, 18(3):415–447.

[10] Einmahl, J. H. and McKeague, I. W. (1999). Confidence tubes for multiple quantile plots via empirical likelihood. *The Annals of Statistics*, 27(4):1348–1367.

[11] Esfahani, P. M. and Kuhn, D. (2015). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*.

[12] Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738.

[13] Goh, J. and Sim, M. (2010). Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917.

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

[15] Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics*, pages 1079–1111.

[16] Hollander, M. and McKeague, I. W. (1997). Likelihood ratio-based confidence bands for survival functions. *Journal of the American Statistical Association*, 92(437):215–226.
[17] Imbens, G. W. (2012). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*.

[18] Kitamura, Y. (2006). Empirical likelihood methods in econometrics: Theory and practice.

[19] Lam, H. (2013). Robust sensitivity analysis for stochastic systems. *arXiv preprint arXiv:1303.0326*.

[20] Lam, H. and Zhou, E. (2015). Quantifying uncertainty in sample average approximation. In *2015 Winter Simulation Conference (WSC)*, pages 3846–3857.

[21] Lam, H. and Zhou, E. (2016). The Empirical Likelihood Approach to Quantifying Uncertainty in Sample Average Approximation. *arXiv:1604.02573 [math, stat]*. arXiv: 1604.02573.

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

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

[24] Luenberger, D. G. (1973). *Introduction to linear and nonlinear programming*, volume 28. Addison-Wesley Reading, MA.

[25] Murphy, S. A. (1995). Likelihood ratio-based confidence intervals in survival analysis. *Journal of the American Statistical Association*, 90(432):1399–1405.

[26] Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

[27] Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, pages 90–120.

[28] Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, pages 1725–1747.

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

[30] Owen, A. B. (2001). *Empirical likelihood*. CRC press.

[31] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, pages 300–325.
[32] Ruszczynski, A. P. and Shapiro, A. (2003). *Stochastic programming*, volume 10. Elsevier Amsterdam.

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

[34] Wang, Q. and Rao, J. N. K. (2001). Empirical likelihood for linear regression models under imputation for missing responses. *Canadian Journal of Statistics*, 29(4):597–608.

[35] Wang, Z., Glynn, P. W., and Ye, Y. (2009). Likelihood robust optimization for data-driven newsvendor problems. *Preprint*.

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

[37] Zhao, Y. and 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.

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

E-mail address: jose.blanchet@columbia.edu, yangkang@stat.columbia.edu