Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

STATISTICAL LIMIT THEOREMS IN DISTRIBUTIONALLY ROBUST OPTIMIZATION

Jose Blanchet

Stanford University
975 Via Ortega, 3rd Floor
Stanford, CA 94305, USA

Alexander Shapiro

Georgia Institute of Technology
765 Ferst Dr Atlanta, 4th Floor
Atlanta, GA 30332-0205, USA

ABSTRACT

The goal of this paper is to develop a methodology for the systematic analysis of asymptotic statistical properties of data-driven DRO formulations based on their corresponding non-DRO counterparts. We illustrate our approach in various settings, including both phi-divergence and Wasserstein uncertainty sets. Different types of asymptotic behaviors are obtained depending on the rate at which the uncertainty radius decreases to zero as a function of the sample size and the geometry of the uncertainty sets.

1 INTRODUCTION

The statistical analysis of Empirical Risk Minimization (ERM) estimators is a well-investigated topic both in statistics (e.g., (van der Vaart 1998)) and stochastic optimization (e.g., (Shapiro et al. 2009)). In recent years, there has been significant interest in the investigation of distributionally robust optimization (DRO) estimators (e.g., (Rahimian and Mehrotra 2019)). The goal of this paper is to develop a methodology for the study of asymptotic statistical properties of data-driven DRO formulations based on their corresponding non-DRO counterparts.

Our objective is to illustrate the main conceptual strategies for the statistical development, emphasizing qualitative features, for instance, the different types of behavior arising from the interaction between the distributional uncertainty size and the sample size, while keeping the discussion easily accessible. Consequently, in order to keep the discussion easily accessible, we do not necessarily focus on the most general assumptions to apply our results.

To set the stage, let us introduce some notation. We use $\mathfrak{P}(\mathscr{S})$ to denote the set of Borel probability measures supported on a closed (nonempty) set $\mathscr{S} \subset \mathbb{R}^d$. Let $X_1, ..., X_n$ be a sequence of independent identically distributed (i.i.d.) random vectors viewed as realizations (or i.i.d. copies) of random vector X having distribution $P_* \in \mathfrak{P}(\mathscr{S})$. Consider the corresponding empirical measure $P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$, where δ_x denotes the Dirac measure of mass one at the point $x \in \mathbb{R}^d$. The sample mean of a function $\psi : \mathscr{S} \to \mathbb{R}$ is $\mathbb{E}_{P_n}[\psi(X)] = n^{-1}\sum_{i=1}^n \psi(X_i)$. By the Strong Law of Large Numbers, we have that $\mathbb{E}_{P_n}[\psi(X)]$ converges with probability one (w.p.1) to $\mathbb{E}_{P_*}[\psi(X)]$, provided the expectation $\mathbb{E}_{P_*}[\psi(X)]$ is well defined.

By the Central Limit Theorem, $n^{1/2}(\mathbb{E}_{P_n}[\psi(X)] - \mathbb{E}_{P_*}[\psi(X)]) \rightsquigarrow N\left(0, \sigma^2\right)$, where "$\rightsquigarrow$" denotes the weak convergence (converges in distribution) and $N\left(0, \sigma^2\right)$ represents the normal distribution with mean zero and variance $\sigma^2 = \mathrm{Var}_{P_*}[\psi(X)]$, provided this variance is finite.

We consider a loss function of the form $l : \mathbb{R}^d \times \Theta \to \mathbb{R}$, with $\Theta \subset \mathbb{R}^m$ being the parameter space. Unless stated otherwise, we assume that the set Θ is *compact* and $l(x, \theta)$ is *continuous* on $\mathscr{S} \times \Theta$. We define

$$f_n(\theta) := \mathbb{E}_{P_n}[l(X, \theta)] \quad \text{and} \quad f(\theta) := \mathbb{E}_{P_*}[l(X, \theta)]. \tag{1}$$

So, the standard ERM formulation takes the form

$$\min_{\theta \in \Theta} f_n(\theta), \tag{2}$$

and viewed as an empirical counterpart of the "true" (or limiting) form

$$\min_{\theta\in\Theta} f(\theta). \tag{3}$$

The statistical properties such as consistency and asymptotic normality of the ERM estimates have been widely studied in significant generality as the sample size $n \to \infty$. These types of results hold under structural properties of the function $f(\cdot)$ and natural stability assumptions (to be reviewed) which guarantee a functional Central Limit Theorems for $f_n(\cdot)$. Our goal is to present a development that is largely parallel to the associated distributionally robust counterpart to (2).

More precisely, (2) can be endowed with distributional robustness by defining a set of probability measures, called the *ambiguity set*, $\mathfrak{M}_\delta(P_n) \subset \mathfrak{P}(\mathscr{S})$, which are seen as "reasonable" (according to some criterion) perturbations of the empirical measure. The parameter $\delta \geq 0$ is the uncertainty size and the family of sets $\{\mathfrak{M}_\delta(P_n) : \delta \geq 0\}$ is typically nondecreasing in δ (in the inclusion partial order sense). The ambiguity set can be defined around any reference probability measure, but unless otherwise stated, we will center the ambiguity set around P_n. In this paper we deal with ambiguity sets of the form

$$\mathfrak{M}_\delta(P_n) := \{P \in \mathfrak{P}(\mathscr{S}) : D(P,P_n) \leq \delta\},$$

where $D(Q,P)$ is a divergence between $Q,P \in \mathfrak{P}(\mathscr{S})$. Specifically, we consider the phi-divergence and Wasserstein distance cases.

In order to state the DRO version of (2) we define

$$\mathscr{F}_n(\theta,\delta_n) := \sup_{P\in\mathfrak{M}_{\delta_n}(P_n)} \mathbb{E}_P[l(X,\theta)],$$

where δ_n is a monotonically decreasing sequence tending to zero as $n \to \infty$. The DRO version of (2) takes the form

$$\min_{\theta\in\Theta} \mathscr{F}_n(\theta,\delta_n). \tag{4}$$

The aim of this paper is to investigate asymptotic statistical properties of the optimal value and optimal solutions of the DRO problem (4). There are *typically* (but not always) *three types of cases* involving the limiting asymptotic statistics depending on the rate of convergence of δ_n to zero. These can be seen both in terms of the value function error

$$\min_{\theta\in\Theta} \mathscr{F}_n(\theta,\delta_n) - \min_{\theta\in\Theta} f(\theta),$$

and the optimal solution error (assuming it is unique for the limiting version of the problem and sufficient regularity conditions are in place).

Intuitively, if δ_n is smaller than a certain (to be characterized) *critical rate* relative to the canonical parametric statistical error rate $n^{-1/2}$, then the DRO effect is negligible compared to the statistical error implicit in a sample of size n. If δ_n decreases to zero right at the critical rate, the DRO effect is comparable with this statistical error and can be quantified in the form of an asymptotic bias. If δ_n is greater than the critical rate, the DRO effect overwhelms the statistical noise. These critical rates depend on the sensitivity of the optimal value function with respect to a small change in the size of uncertainty.

Our objective is to provide accessible principles that can be used to obtain explicit limiting distributions for the errors, both for value functions and optimizers, when $\delta_n \to 0$ in these *three cases*; see Theorems 1 and 2 for general principles and Theorems 3 and 4 for the application of these principles to value functions of phi-divergence and Wasserstein DRO, respectively; and Theorem 5 for the corresponding application to phi-divergence optimal solutions (we omit the Wasserstein case due to space constraints).

It is important to note that it is common in the data-driven DRO literature to suggest choosing δ_n to enforce that P_* is inside $\mathfrak{M}_{\delta_n}(P_n)$ with high probability. This selection typically *will fall in the third*

case, that is, this choice will induce estimates that are substantially larger than standard statistical noise. Therefore, prescriptions corresponding to the third case should be assigned only if the optimizer perceives that the out-of-sample environment is substantially different from the observed (empirical) environment due to errors or fluctuations that fall outside of standard statistical noise.

The remainder of the paper is organized as follows. In Section 2 we will quickly review the elements of the statistical analysis of Empirical Risk Minimization (ERM) – also known as Empirical Optimization or Sample Average Approximation – which corresponds to case $\delta_n = 0$. Then, in Section 3, we will follow a parallel discussion to that of Section 2 and discuss assumptions for the data-driven DRO version of the problem. The objective is to use these assumptions so that we can obtain a flexible and disciplined approach that can be systematically applied to various DRO formulations. Then, in Section 4 we will discuss the application of this approach to the explicit development of asymptotics for the optimal value in phi-divergence and Wasserstein DRO and, finally, in Section 5, we also develop these explicit results for associated optimal solutions.

We use the following notation throughout the paper. For a sequence Y_n of random variables, by writing $Y_n = o_p(n^{-\gamma})$ we mean that $n^{\gamma}Y_n$ tends in probability to zero as $n \to \infty$. In particular $Y_n = o_p(1)$ means that Y_n tends in probability to zero. The notation $Q \ll P$ means that $Q \in \mathfrak{P}(\mathscr{S})$ is absolutely continuous with respect to $P \in \mathfrak{P}(\mathscr{S})$. Unless stated otherwise, probabilistic statements like "almost every" (a.e.), are made with respect to the probability measure P_*. By saying that a function $h : \mathscr{S} \to \mathbb{R}$ is integrable we mean that $\mathbb{E}_{P_*}|h(X)| < \infty$. It is said that a mapping $\phi : \mathbb{R}^m \to \mathbb{R}^k$ is directionally differentiable at a point $\theta \in \mathbb{R}^m$ if the directional derivative

$$\phi'(\theta, d) := \lim_{t \downarrow 0} \frac{\phi(\theta + td) - \phi(\theta)}{t}$$

exists for every $d \in \mathbb{R}^m$. We will use the term $\varepsilon_n(\theta)$, $\theta \in \Theta$, to denote a random field such that

$$\sup_{\theta \in \Theta} |\varepsilon_n(\theta)| = o_p(1). \tag{5}$$

2 STATISTICS OF ERM: REVIEW

In addition to the population objective function $f(\theta) := \mathbb{E}_{P_*}[l(X, \theta)]$, introduced in (1), we also let

$$\vartheta := \inf_{\theta \in \Theta} f(\theta) \text{ and } \Theta^* := \arg\min_{\theta \in \Theta} f(\theta),$$

be the optimal value and the set of optimal solutions of the population version of the optimization problem, respectively.

As defined in (1), $f_n(\theta) = \mathbb{E}_{P_n}[l(X, \theta)]$ is the objective function of the ERM version of the problem and

$$\vartheta_n := \inf_{\theta \in \Theta} f_n(\theta) \text{ and } \theta_n \in \arg\min_{\theta \in \Theta} f_n(\theta)$$

are the respective optimal value and an optimal solutions of the ERM problem. We will now quickly review the development of the asymptotic statistics of the optimal value in ERM and then we will discuss the corresponding results for optimal solutions.

2.1 Asymptotics of the Optimal Value

In order to analyze the statistical error in the difference between the optimal values $\vartheta_n - \vartheta$, we start by enforcing a functional Central Limit Theorem (CLT) for $f_n(\cdot)$. In particular, one imposes assumptions that guarantee an expansion of the form (recall that $\varepsilon_n(\cdot)$ satisfies (5))

$$f_n(\theta) = f(\theta) + n^{-1/2} r_n(\theta) + n^{-1/2} \varepsilon_n(\theta), \tag{6}$$

where we have functional weak convergence

$$r_n(\cdot) \rightsquigarrow \mathfrak{g}(\cdot)$$

in the uniform topology on compact sets, with $\mathfrak{g}(\cdot)$ being a mean zero Gaussian random field with covariance function

$$\mathrm{Cov}\left(\mathfrak{g}(\theta), \mathfrak{g}\left(\theta'\right)\right) = \mathrm{Cov}_{P_*}\left(l(X,\theta), l\left(X,\theta'\right)\right).$$

There are several ways to enforce (6); a simple set of sufficient conditions satisfying this is given next (cf., (van der Vaart 1998, example 19.7)).

Assumption 1 (i) For some $\bar{\theta} \in \Theta$ the expectation $\mathbb{E}_{P_*}[l(X,\bar{\theta})^2]$ is finite. (ii) There is a measurable function $\psi : \mathscr{S} \to \mathbb{R}_+$ such that $\mathbb{E}_{P_*}[\psi(X)^2]$ is finite and

$$|l(X,\theta) - l(X,\theta')| \leq \psi(X)\|\theta - \theta'\|$$

for all $\theta, \theta' \in \Theta$ and a.e. $X \in \mathscr{S}$.

In particular, under this assumption, it follows that the expectation function $f(\theta)$ and variance

$$\sigma^2(\theta) := \mathrm{Var}_{P_*}(l(X,\theta))$$

are finite valued and continuous on Θ. Furthermore, since the set Θ is compact, it follows that the optimal value, ϑ_n, of the ERM problem converges to ϑ in probability (in fact, almost surely). Moreover, it is not difficult to show from (6) that the distance from θ_n to Θ_* converges in probability to zero (actually, convergence occurs almost surely) as $n \to \infty$. Finally, since the functional $V(\phi) := \inf_{\theta\in\Theta} \phi(\theta)$, mapping continuous functions $\phi : \Theta \to \mathbb{R}$ to the real line, is directionally differentiable, the following classical result is a direct consequence of the (functional) Delta Theorem (cf., (Shapiro 1991)).

Proposition 1 Under Assumption 1,

$$n^{1/2}(\vartheta_n - \vartheta) \rightsquigarrow \inf_{\theta\in\Theta^*} \mathfrak{g}(\theta) \tag{7}$$

as $n \to \infty$. In particular, if $\Theta^* = \{\theta^*\}$ is a singleton, i.e. θ^* is the unique optimal solution of the true problem, then $n^{1/2}(\vartheta_n - \vartheta^*)$ converges in distribution to normal $N(0, \sigma^2(\theta^*))$.

2.2 Asymptotics of Optimal Solutions

We assume now that $\Theta^* = \{\theta^*\}$ is a singleton, i.e., θ^* is the unique optimal solution of the true (population) problem (3). We also assume that for a.e. X, the function $l(X,\cdot)$ is continuously differentiable. As was argued in the previous section, the asymptotics of the optimal value is governed by the asymptotics of the objective function. On the other hand, the asymptotics of optimal solutions can be derived from the asymptotics of the gradients of the objective function.

Let us consider the following parametrization of problem (3):

$$\min_{\theta\in\Theta} f(\theta) + v^T\theta, \tag{8}$$

with parameter vector $v \in \mathbb{R}^m$. Denote by $\theta_*(v)$ an optimal solution of the above problem (8) viewed as a function of vector v. Of course, we have $\theta_*(0) = \theta^*$.

Assumption 2 (uniform second order growth) There is a neighborhood $\mathscr{V}$ of θ^* and a positive constant κ such that for every v in a neighborhood of $0 \in \mathbb{R}^m$, problem (8) has an optimal solution $\theta_*(v) \in \mathscr{V}$ and

$$f(\theta) + v^T(\theta - \theta_*(v)) \geq f(\theta_*(v)) + \kappa\|\theta - \theta_*(v)\|^2,$$

for all $\theta \in \Theta \cap \mathscr{V}$.

The following assumption can be viewed as a counterpart of Assumption 1 applied to the gradients of the objective function.

Assumption 3 (i) For some $\bar{\theta} \in \Theta$ the expectation $\mathbb{E}_{P_*}\left[\|\nabla l(X,\bar{\theta})\|^2\right]$ is finite. (ii) There is a measurable function $\Psi : \mathscr{S} \to \mathbb{R}_+$ such that $\mathbb{E}_{P_*}[\Psi(X)^2]$ is finite and

$$\|\nabla l(X,\theta) - \nabla l(X,\theta')\| \leq \Psi(X)\|\theta - \theta'\|, \tag{9}$$

for all $\theta, \theta' \in \Theta$ and a.e. $X \in \mathscr{S}$.

By the functional CLT it follows that

$$\nabla f_n(\theta) = \nabla f(\theta) + n^{-1/2} d_n(\theta) + n^{-1/2}\varepsilon_n(\theta), \tag{10}$$

where we have a functional weak convergence $d_n(\cdot) \rightsquigarrow \mathfrak{G}(\cdot)$ in the uniform topology on a closed neighborhood of θ^*, with $\mathfrak{G}(\cdot)$ being a continuous mean zero Gaussian random field with covariance function

$$\mathrm{Cov}\left[\mathfrak{G}(\theta), \mathfrak{G}(\theta')\right] = \mathbb{E}_{P_*}[(\nabla l(X,\theta) - \nabla f(\theta))(\nabla l(X,\theta') - \nabla f(\theta'))^T].$$

It follows from (10) that

$$[\nabla f_n(\theta) - \nabla f(\theta)] - [\nabla f_n(\theta^*) - \nabla f(\theta^*)] = n^{-1/2}\left[d_n(\theta) - d_n(\theta^*) + \varepsilon_n(\theta) - \varepsilon_n(\theta^*)\right].$$

Also since $\rho_n := \|\theta_n - \theta^*\|$ tends in probability to zero, we have

$$\sup_{\theta:\|\theta-\theta^*\|\leq\rho_n} [d_n(\theta) - d_n(\theta^*) + \varepsilon_n(\theta) - \varepsilon_n(\theta^*)] = o_p(1). \tag{11}$$

Thus we have the following result from (Shapiro 1993, Theorem 2.1), where the respective regularity conditions are ensured by the above property (11).

Proposition 2 Suppose that Assumptions 2 and 3 hold. Then it follows that

$$\theta_n = \theta_*(Z_n) + o_p(n^{-1/2}), \tag{12}$$

where $Z_n := \nabla f_n(\theta^*) - \nabla f(\theta^*)$.

The above result reduces the analysis of asymptotic properties of optimal solutions to the investigation of the asymptotic behavior of optimal solutions of the finite dimensional problem (8). By the (finite dimensional) Central Limit Theorem, $n^{1/2}Z_n$ converges in distribution to normal $N(0,\Sigma)$ with covariance matrix $\Sigma = \mathrm{Cov}(\nabla l(X,\theta^*))$. Moreover, if the mapping $\theta_*(v)$ is directionally differentiable at $v = 0$ (in the Hadamard sense), then by the finite dimensional Delta Theorem it follows from (12) that

$$n^{1/2}(\theta_n - \theta^*) \rightsquigarrow \theta'_*(0,Z),$$

where $Z \sim N(0,\Sigma)$. In particular, if $\theta'_*(0,w) = Aw$ is linear (i.e., $\theta_*(v)$ is differentiable at $v = 0$ with Jacobian matrix A), then $n^{1/2}(\theta_n - \theta^*)$ converges in distribution to normal with null mean vector and covariance matrix $A\Sigma A^T$.

Directional differentiability of optimal solutions of parameterized problems is well investigated. For example, if θ^* is an interior point of Θ, $f(\theta)$ is twice continuously differentiable at θ^* and the Hessian matrix $H := \nabla^2 f(\theta^*)$ is nonsingular, then the uniform second order growth (Assumption 2) holds, and $\theta_*(v)$ is differentiable at $v = 0$ with $\theta'_*(0,w) = H^{-1}w$. When θ^* is on the boundary of the set Θ, the sensitivity analysis of the parameterized problem (8) is more delicate and involves a certain measure of the curvature of the set Θ at the point θ^*. This is discussed extensively in (Bonnans and Shapiro 2000). We also refer to (Shapiro, Dentcheva, and Ruszczyński 2009, sections 5.1.3 and 7.1.5) for a basic summary of these results.

It is worthwhile to note at this point that the regularity conditions of Assumptions 2 and 3 address different properties of the considered setting. Assumption 2 deals with the limiting optimization problem and is of deterministic nature. The *uniform* second order growth condition was introduced in (Shapiro 1993), and in a more general form was discussed in (Bonnans and Shapiro 2000, section 5.1.3). On the other hand Assumption 3 is related to the stochastic behavior of the ERM problem (2).

3 STATISTICS OF DRO: GENERAL PRINCIPLES

We now provide sufficient conditions for the development of DRO statistical principles based on assumptions which are parallel to those imposed in the ERM section. Define

$$\bar{\vartheta}_n := \inf_{\theta\in\Theta} \mathscr{F}_n(\theta,\delta_n) \text{ and } \bar{\theta}_n \in \arg\min_{\theta\in\Theta} \mathscr{F}_n(\theta,\delta_n),$$

the optimal value and an optimal solution of the DRO problem (4).

3.1 DRO Asymptotics of the Optimal Value

Similar to the ERM case, in the DRO setting, we will typically have an expansion of the form

$$\mathscr{F}_n(\theta,\delta_n) = f_n(\theta) + \delta_n^\gamma R_n(\theta) + \delta_n^\gamma \varepsilon_n(\theta), \tag{13}$$

for some $\gamma > 0$, where $R_n(\cdot)$ converges in probability in the uniform topology over Θ to a continuous deterministic process $\rho(\cdot)$,

$$R_n(\theta) = \rho(\theta) + \varepsilon_n(\theta). \tag{14}$$

Since $\mathscr{F}_n(\cdot,\delta_n) \geq f_n(\cdot)$, it follows then that $\rho(\cdot) \geq 0$. We will characterize $\gamma > 0$ and $\rho(\cdot)$ explicitly in the next sections in the context of phi-divergence and Wasserstein DRO formulations under suitable conditions.The following result, summarizes the type of behavior that we expect in DRO formulations depending on the rate of decay to zero of the uncertainty size δ_n.

Theorem 1 Suppose that Assumption 1 and conditions (13) - (14) hold. Then there are three types of asymptotic behavior of the DRO optimal value:
(a) If $\delta_n^\gamma = o\left(n^{-1/2}\right)$, then

$$\bar{\vartheta}_n = \vartheta_n + o_p\left(n^{-1/2}\right), \tag{15}$$

and hence

$$n^{1/2}\left(\bar{\vartheta}_n - \vartheta\right) \rightsquigarrow \inf_{\theta\in\Theta^*} \mathfrak{g}(\theta),$$

which coincides with (7) and thus the DRO formulation has no asymptotic impact.
(b) If $\delta_n^\gamma = n^{-1/2}$, then

$$n^{1/2}\left(\bar{\vartheta}_n - \vartheta\right) \rightsquigarrow \inf_{\theta\in\Theta^*} \{\mathfrak{g}(\theta) + \rho(\theta)\}, \tag{16}$$

so the DRO formulation introduces an explicit and quantifiable asymptotic bias which can be interpreted as a regularization term.
(c) If $o\left(\delta_n^\gamma\right) = n^{-1/2}$, then

$$\delta_n^{-\gamma}\left(\bar{\vartheta}_n - \vartheta\right) \rightsquigarrow \inf_{\theta\in\Theta^*} \rho(\theta),$$

so the bias term induced by the DRO formulation is larger than the statistical error.

Proof. Part (a). By (13) and (14) we have that in the considered case

$$\mathscr{F}_n(\theta,\delta_n) = f_n(\theta) + o(n^{-1/2})\varepsilon_n(\theta),$$

where $\varepsilon_n(\theta)$ is the generic term satisfying (5). Thus (15) follows.

Part (b). By (13) and (14) in the considered case, we can write

$$n^{1/2}\left(\mathscr{F}_n(\theta,\delta_n) - f(\theta)\right) = n^{1/2}(f_n(\theta) - f(\theta)) + \rho(\theta) + \varepsilon_n(\theta).$$

Under Assumption 1, by the functional CLT we see that $n^{1/2}(f_n(\theta)-f(\theta))+\rho(\theta)$ converges in distribution to $\mathfrak{g}(\theta)+\rho(\theta)$, and hence (16) follows by the Delta Theorem.

Part (c) may appear somewhat different because the right hand side is deterministic but, under case (c) note that we can simply write

$$\mathscr{F}_n(\theta,\delta_n) = f(\theta)+\delta_n^{\gamma}R_n(\theta)+\delta_n^{\gamma}\varepsilon_n(\theta),$$

so case (c) also follows from the standard analysis since $R_n(\cdot)$ converges uniformly to $\rho(\cdot)$ in probability (thus it converges weakly in the uniform topology). □

3.2 DRO Asymptotics of the Optimal Solutions

As in the ERM development, in addition to Assumptions 2, it is convenient to guarantee that for all n large enough, $\mathscr{F}_n(\theta,\delta_n)$ is differentiable in a neighborhood $\mathscr{V}$ of θ^* and

$$\nabla\mathscr{F}_n(\theta,\delta_n) = \nabla f_n(\theta)+\delta_n^{\gamma}D_n(\theta)+\delta_n^{\gamma}\varepsilon_n(\theta), \tag{17}$$

for some $\gamma>0$, where $D_n(\theta)$ converges in probability to $\nabla\rho(\theta)$ uniformly around a closed neighborhood $\bar{\mathscr{V}}$ of θ^*. As a consequence, we obtain the following analog of Theorem 1, which follows from the *finite dimensional* Delta Theorem. Recall that $\theta_*(v)$ is an optimal solution to problem (8) and $\theta_*'(0,\cdot)$ is its directional derivative at $v=0$.

Theorem 2 Suppose that: Assumptions 2 and 3 hold, conditions (13) - (14) are satisfied, identity (17) holds with $D_n(\cdot)$ converging in probability to $\nabla\rho(\cdot)$ uniformly around a closed neighborhood $\bar{\mathscr{V}}$ of θ^*, and that $\theta_*(v)$ is directionally differentiable at $v=0$ (in the Hadamard sense). Let $Z\sim N(0,\Sigma)$ with covariance matrix $\Sigma=\mathrm{Cov}(\nabla l(X,\theta^*))$. Then the DRO optimal solutions can have three types of asymptotic behavior:
(A) If $\delta_n^{\gamma}=o\left(n^{-1/2}\right)$, then

$$\bar{\theta}_n = \theta_n+o_p\left(n^{-1/2}\right),$$

thus

$$n^{1/2}\left(\bar{\theta}_n-\theta^*\right)\rightsquigarrow\theta_*'(0,Z).$$

(B) If $\delta_n^{\gamma}=n^{-1/2}$, then

$$n^{1/2}\left(\bar{\theta}_n-\theta^*\right)\rightsquigarrow\theta_*'(0,Z+\nabla\rho(\theta^*)).$$

(C) If $o\left(\delta_n^{\gamma}\right)=n^{-1/2}$, then

$$\delta_n^{-\gamma}\left(\bar{\theta}_n-\theta^*\right)\rightsquigarrow\theta_*'(0,\nabla\rho(\theta^*)).$$

4 GENERAL PRINCIPLE IN ACTION: OPTIMAL VALUES

4.1 The Phi-Divergence Case

We recall the definition of the distributional uncertainty set for the phi-divergence case. Consider a convex lower semi-continuous function $\phi:\mathbb{R}\to\mathbb{R}_+\cup\{+\infty\}$ such that $\phi(1)=0$ and $\phi(t)=+\infty$ for $t<0$. For probability measures $Q,P\in\mathfrak{P}$ such that Q is absolutely continuous with respect to P with the corresponding density dQ/dP, the ϕ-divergence is defined as (cf., (Csiszár 1963),(Morimoto 1963))

$$D_\phi(Q\|P) := \mathbb{E}_P[\phi(dQ/dP)] = \int\phi(dQ/dP)dP.$$

In particular, for $\phi(t):=t\log(t)-t+1$, $t\geq 0$, this becomes the Kullback–Leibler (KL) divergence of Q from P. The ambiguity set $\mathfrak{M}_\delta(P)$ associated with $D_\phi(\cdot\|P)$ is defined as

$$\mathfrak{M}_\delta(P) := \{Q\ll P : D_\phi(Q\|P)\leq\delta\}.$$

By duality arguments the corresponding distributionally robust functional can be written in the form (cf., (Bayraksan and Love 2015), (Ben-Tal and Teboulle 1987), (Shapiro 2017))

$$\sup_{Q\in\mathfrak{M}_\delta(P)} \mathbb{E}_Q[Y] = \inf_{\mu,\lambda>0} \{\lambda\delta + \mu + \lambda \mathbb{E}_P[\phi^*((Y-\mu)/\lambda)]\}, \tag{18}$$

where $\phi^*(y) = \sup_{t\in\mathbb{R}}\{yt - \phi(t)\}$ is the convex conjugate of ϕ. Using this representation, we can obtain an asymptotic expansion for (18) as a function of δ. This expansion can be helpful to suggest the form of the expansion in (13) and (17). For this, we need to assume certain regularity properties of $\phi(t)$ around $t = 1$.

Assumption 4 Assume that $\phi(t)$ is two times continuously differentiable in a neighborhood of $t = 1$ with $\kappa := 2/\phi''(1) > 0$.

Under this condition, we have the following expansion, which is obtained, in order to simplify our exposition, under the assumption that the probability measure P has compact support. See also the results in (Lam 2016), which provide additional correction terms under a fixed P. The uniform feature of the statement below is helpful in the statistical analysis. Our development here will also be used in the expansion of optimal solutions.

Proposition 3 Suppose that Assumption 4 holds, that $P(|Y| \le \nu) = 1$ for some $\nu \in (0,\infty)$. Then, for any $b_0 > 0$,

$$\sup_{Q\in\mathfrak{M}_\delta(P)} \mathbb{E}_Q[Y] - \mathbb{E}_P(Y) - \delta^{1/2}\kappa^{1/2}\sqrt{\mathrm{Var}_P[Y]} = o\left(\delta^{1/2}\right), \tag{19}$$

uniformly over Borel probability measures P supported on $[-\nu,\nu]$ such that $\mathrm{Var}_P[Y] \ge b_0$. Moreover, there is $\bar\delta > 0$ such that for all $\delta < \bar\delta$

$$\arg\max\{\mathbb{E}_Q[Y] : Q \in \mathfrak{M}_\delta(P)\}$$

is unique.

Proof. Note that we can write

$$\sup_{Q\in\mathfrak{M}_\delta(P)} \mathbb{E}_Q[Y] = \sup_{\mathbb{E}_P(Z)=1,\mathbb{E}_P(\phi(Z))\le\delta} \mathbb{E}_P[YZ],$$

where the sup is taken over the set of positive random variables Z satisfying the specified moment constraints. We may assume that $\mathbb{E}_P[Y] = 0$ for simplicity since we can always center the objective function around $\mathbb{E}_P[Y)]$. In turn, by letting $\bar\Delta = (Z-1)/\delta^{1/2}$, the previous optimization problem is equivalent to

$$\delta^{1/2} \sup_{\mathbb{E}_P(\bar\Delta)=0,\bar\Delta\ge-\delta^{-1/2},\mathbb{E}_P(\phi(1+\delta^{1/2}\bar\Delta))\le\delta} \mathbb{E}_P[Y\bar\Delta]. \tag{20}$$

Since $|Y| \le \nu$ and $\mathbb{E}_P[Y] = 0$, then $\bar\Delta = aY$ is feasible for any $a > 0$ provided that $a\nu \le \delta^{-1/2}$ and

$$\mathbb{E}_P[\phi(1+\delta^{1/2}\bar\Delta)] \le \delta.$$

In turn, since $\phi(t)$ is two times continuously differentiable at $t = 1$, we have that

$$\delta^{-1}\phi(1+\delta^{1/2}ay)) \to a^2y^2\phi''(1)/2$$

as $\delta \to 0$ uniformly in compact sets. Therefore, we conclude that there exists $\delta_0 > 0$ such that for any $\delta < \delta_0$

$$\sup_{\mathbb{E}_P(\bar\Delta)=0,\bar\Delta\ge-\delta^{-1/2},\mathbb{E}_P(\phi(1+\delta^{1/2}\bar\Delta))\le\delta} \mathbb{E}_P[Y\bar\Delta]$$

$$\ge \sup_{a>0,a^2\mathbb{E}_P(Y^2)/2\le(1-\delta_0)/\phi''(1)} \mathbb{E}_P[aY^2] = \sqrt{\kappa(1-\delta_0)}\cdot\sqrt{\mathbb{E}_P[Y^2]}.$$

Since $\delta_0 > 0$ can be chosen to be arbitrarily small, we conclude an asymptotic lower bound which retrieves (19). For the upper bound, we apply the duality result (18) in the form corresponding to (20), we obtain

$$\begin{aligned}
&\sup_{\mathbb{E}_P(\bar{\Delta})=0,\bar{\Delta}\geq-\delta^{-1/2},\delta^{-1}\mathbb{E}_P(\phi(1+\delta^{1/2}\bar{\Delta}))\leq 1} \mathbb{E}_P[Y\bar{\Delta}] \\
&= \min_{\bar{\lambda}>0,\bar{\mu}} \{\bar{\lambda} + \mathbb{E}_P[\sup_{\bar{\Delta}\geq-\delta^{-1/2}} \{(Y+\bar{\mu})\bar{\Delta} - \bar{\lambda}\delta^{-1/2}\phi(1+\delta^{1/2}\bar{\Delta})\}]\} \\
&\leq \min_{\bar{\lambda}>0} \{\bar{\lambda} + \mathbb{E}_P[\sup_{\bar{\Delta}\geq-\delta^{-1/2}} \{Y\bar{\Delta} - \bar{\lambda}\delta^{-1/2}\phi(1+\delta^{1/2}\bar{\Delta})\}]\}. \qquad (21)
\end{aligned}$$

We will plug in

$$\bar{\lambda}_0 = \arg\min\{\bar{\lambda} + \kappa\mathbb{E}_P[Y^2]/4\bar{\lambda} : \bar{\lambda} > 0\} = 2^{-1}\sqrt{\kappa\mathbb{E}_P[Y^2]} > 0$$

into (21) to obtain our upper bound. Using that $\bar{\lambda}_0 > 0$ and that ϕ is convex with $\phi''(1) > 0$, we have that the family of (continuous) functions

$$s_\delta(y) := \sup_{\bar{\Delta}\geq-\delta^{-1/2}} \{y\bar{\Delta} - \bar{\lambda}\delta^{-1/2}\phi(1+\delta^{1/2}\bar{\Delta})\}$$

converges uniformly on compact sets to

$$s_0(y) = \sup_{\bar{\Delta}}\{y\bar{\Delta} - \bar{\lambda}\bar{\Delta}^2/\kappa\} = \frac{\kappa y^2}{4\bar{\lambda}}.$$

Therefore we obtain that

$$\begin{aligned}
&\min_{\bar{\lambda}>0}\{\bar{\lambda} + \mathbb{E}_P[\sup_{\bar{\Delta}\geq-\delta^{-1/2}} \{Y\bar{\Delta} - \bar{\lambda}\delta^{-1/2}\phi(1+\delta^{1/2}\bar{\Delta})\}]\} \\
&\leq \bar{\lambda}_0 + \mathbb{E}_P[\sup_{\bar{\Delta}\geq-\delta^{-1/2}} \{Y\bar{\Delta} - \bar{\lambda}_0\delta^{-1/2}\phi(1+\delta^{1/2}\bar{\Delta})\}] \to \sqrt{\kappa}\cdot\sqrt{\mathbb{E}_P[Y^2]}.
\end{aligned}$$

These estimates, which are uniform given that $|Y| \leq \nu$, yield the estimate in the proposition. The uniqueness is standard; it follows from the local strong convexity of $\phi(\cdot)$ at the origin. □

Recall that $\sigma^2(\theta) := \mathrm{Var}_{P_*}(l(X,\theta))$, and that $\mathfrak{g}(\cdot)$ is a mean zero Gaussian random field. Expansion (19) immediately yields, at least when $\sup_{\theta\in\Theta}|l(X,\theta)|$ is P_*-bounded, that

$$\mathscr{F}_n(\theta,\delta_n) = f_n(\theta) + \delta_n^{1/2}\kappa^{1/2}\sigma(\theta) + \delta_n^{1/2}\varepsilon_n(\theta). \qquad (22)$$

Consequently, we obtain the following result.

Theorem 3 Suppose that $\sup_{\theta\in\Theta}|l(X,\theta)|$ is P_*-essentially bounded, that Assumption 1 and Assumption 4 hold, and that $\sigma^2(\theta) > 0$ for all $\theta \in \Theta^*$. Then, we have the following types of asymptotic behavior of the DRO optimal values.
(a-phi) If $\delta_n = o\left(n^{-1}\right)$, then

$$n^{1/2}\left(\bar{\vartheta}_n - \vartheta\right) \rightsquigarrow \inf_{\theta\in\Theta^*}\mathfrak{g}(\theta).$$

(b-phi) If $\delta_n = n^{-1}$, then

$$n^{1/2}\left(\bar{\vartheta}_n - \vartheta\right) \rightsquigarrow \inf_{\theta\in\Theta^*}\left\{\mathfrak{g}(\theta) + \kappa^{1/2}\sigma(\theta)\right\}. \qquad (23)$$

(c-phi) If $o(\delta_n) = n^{-1}$, then

$$\delta_n^{-1/2}\left(\bar{\vartheta}_n - \vartheta\right) \rightsquigarrow \kappa^{1/2}\inf_{\theta\in\Theta^*}\sigma(\theta),$$

so the bias term induced by the DRO formulation dominates the statistical error.

Proof. Proof of this theorem follows standard techniques (cf., (Shapiro, Dentcheva, and Ruszczyński 2009, proof of Theorem 5.7)). For the sake of completeness we briefly outline proof of case (b-phi). Note that our assumptions imply Assumption 1, and hence $\sigma^2(\theta)$ is a continuous function of θ. Therefore there is a compact neighborhood $\bar{\Theta}$ of Θ^* such that $\sigma^2(\theta) > 0$ for all $\theta \in \bar{\Theta}$. We can restrict the minimization to $\bar{\Theta}$ for which the expansion (22) holds.

Consider the space $C(\bar{\Theta})$ of continuous functions $g : \bar{\Theta} \to \mathbb{R}$ equipped with the sup-norm, and functional $V(g) := \inf_{\theta\in\bar{\Theta}} g(\theta)$, mapping $C(\bar{\Theta})$ into the real line. This functional is directionally differentiable in the Hadamard sense with the directional derivative at a point $\mu \in C(\bar{\Theta})$ given by $V'(\mu, h) = \inf_{\theta\in\bar{\Theta}(\mu)} h(\theta)$, where $\bar{\Theta}(\mu) := \arg\min_{\theta\in\bar{\Theta}} \mu(\theta)$. We have that $\bar{\vartheta}_n = V(\mathscr{F}_n)$ and $\vartheta = V(f)$, where $\mathscr{F}_n(\cdot) := \mathscr{F}_n(\cdot, \delta_n)$. By the functional CLT and (22) it follows that $n^{1/2}(\mathscr{F}_n - f)$ converges in distribution (weakly) to $\mathfrak{g}(\theta) + \kappa^{1/2}\sigma(\theta)$. We can apply now the functional Delta Theorem to conclude (23). □

Given that $\phi(\cdot)$ is only assumed to satisfy Assumption (4), without imposing any growth condition, situations such as the (c-phi) case require imposing stronger moment conditions than just assuming $\mathrm{Var}_{P_*}[l(X,\theta)] < \infty$. This can be seen in the KL-divergence case in which $\phi(t) = t\log(t) - t + 1$. For fixed $\delta > 0$, the population version of the DRO problem requires that $l(X,\theta)$ has a finite moment-generating function in a neighborhood of the origin. Therefore, if δ_n converges to zero sufficiently slowly and $l(X,\theta)$ has infinite moments of order $2+\varepsilon$, an expansion such as (22) may not hold. However, if $\phi(t) = (t-1)^2$, it follows that expansion (22) holds exactly with $\varepsilon_n(\theta) = 0$.

On the other hand, (Duchi et al. 2021, Theorem 2) provides a more general result for the case (b-phi) since it does not require compact support (although it requires ϕ to be three times continuously differentiable). The following example shows that the smoothness of $\phi(\cdot)$ is important in deriving the asymptotics in the previous result with $\delta_n = n^{-1/2}$.

Example 1 Consider $\phi(t) := |t-1|$, $t \geq 0$. In that case (e.g., (Shapiro 2017, Example 3.12)), for $\delta \in (0,2)$ and essentially bounded Y,

$$\sup_{Q\in\mathfrak{M}_\delta(P)} \mathbb{E}_Q[Y] = (\delta/2)\mathrm{ess\,sup}(Y) + (1-\delta/2)\mathsf{AV@R}_{P,1-\delta/2}(Y),$$

where

$$\mathsf{AV@R}_{P,\alpha}(Y) := \inf_{\tau\in\mathbb{R}} \left\{\tau + \alpha^{-1}\mathbb{E}_P[Y-\tau]_+\right\},\ \alpha \in (0,1].$$

Note that $\mathsf{AV@R}_{P,1}(Y) = \mathbb{E}_P[Y]$ and as α tends to one,

$$\left|\mathsf{AV@R}_{P,\alpha}(Y) - \mathbb{E}_P[Y]\right| = O(1-\alpha), \tag{24}$$

provided Y is essentially bounded.

Suppose that $l(x,\theta)$ is bounded on $\mathscr{S}\times\Theta$, and hence

$$\mathscr{F}_n(\theta,\delta_n) = (\delta_n/2)\max_{1\leq i\leq N} l(X_i,\theta) + (1-\delta_n/2)\mathsf{AV@R}_{P_n,1-\delta_n/2}(l(X,\theta)). \tag{25}$$

Consider $\delta_n = n^{-1}$. Then the first term in (25) is of order $O(n^{-1})$, and by (24) the second term is $\mathbb{E}_{P_n}[l(X,\theta)] + O\left(n^{-1}\right)$. Consequently in that case $\bar{\vartheta}_n = \vartheta_n + o_p(n^{-1/2})$, and hence this corresponds to case (a) in Theorem 1. This shows that the assumption of smoothness (differentiability) of $\phi(\cdot)$ is essential for deriving the asymptotics of $\bar{\vartheta}_n$. Here, some additional terms in the asymptotics of $\bar{\vartheta}_n$ appear when δ_n is of order $O(n^{-1/2})$, rather than $O(n^{-1})$. □

4.2 The Wasserstein Distance Case

We use $\mathfrak{P}(\mathscr{S}\times\mathscr{S})$ to denote the set of Borel probability measures on the product space $\mathscr{S}\times\mathscr{S}$. Let $c : \mathscr{S}\times\mathscr{S} \to \mathbb{R}_+ \cup \{+\infty\}$ be a lower semi-continuous function such that $c(x,y) = 0$ if and only if $x = y$.

This function measures the marginal cost of transporting a unit of mass from a source location to a target location, respectively. The domain of of $c(\cdot,\cdot)$ is

$$\mathrm{dom}(c) = \{(x,y) \in \mathscr{S} \times \mathscr{S} : c(x,y) < \infty\}.$$

The optimal transport cost between $P, Q \in \mathfrak{P}(\mathscr{S})$ is given by

$$D_c(P,Q) := \min\{\mathbb{E}_\pi[c(X,Y)] : \pi \in \mathfrak{P}(\mathscr{S} \times \mathscr{S}),\ \pi_X = P,\ \pi_Y = Q\},$$

where $\mathbb{E}_\pi[\cdot]$ is the expectation under a joint distribution $\pi \in \mathfrak{P}(\mathscr{S} \times \mathscr{S})$ and π_X and π_Y denote the marginal distributions of X and Y, respectively. It turns out that the optimizer is always achieved, thus we write 'min' instead of 'inf'. Let $\|\cdot\|$ be a norm in the space $\mathbb{R}^d$. An important special case corresponds to the choice $c(x,y) := \|x-y\|^p$ for some $p > 0$, in which case $D_c(P,Q)^{1/p}$ is the so-called p-Wasserstein distance. The reader is referred to the text of Villani (Villani 2003) for more background on optimal transport.

For any given $P \in \mathfrak{P}(\mathscr{S})$ and $\delta \geq 0$ we have the following dual result (cf., (Esfahani and Kuhn 2018), (Blanchet and Murthy 2019), (Gao and Kleywegt 2016)) assuming that $\mathfrak{h}(\cdot)$ is upper semi-continuous and $\mathfrak{h}(X)$ is P-integrable,

$$\sup_{Q: D_c(P,Q) \leq \delta} \mathbb{E}_Q[\mathfrak{h}(Y)] = \min_{\lambda \geq 0} \left\{ \lambda \delta + \mathbb{E}_P\left[\bar{\mathfrak{h}}_\lambda(X)\right] \right\}, \tag{26}$$

where

$$\bar{\mathfrak{h}}_\lambda(x) := \sup_{y \in \mathscr{S}} \{\mathfrak{h}(y) - \lambda c(x,y)\},\ \lambda \geq 0.$$

Throughout the rest of our discussion, we will choose $c(x,y) := \|x-y\|^p$ for $p \in (1,\infty)$ and therefore write $D_p(P,Q)$ for this choice of cost function. Further, we use $\|\cdot\|_*$ to denote the dual norm, namely,

$$\|y\|_* = \sup\{x^T y : \|x\| = 1\}.$$

As in the case of phi-divergence, assuming that P is fixed and has compact support, for example, we can obtain an asymptotic expansion for (26) as a function of δ. By writing $\mathbb{E}_P^{(p-1)/p}[\cdot]$ we mean $(\mathbb{E}_P[\cdot])^{(p-1)/p}$.

Proposition 4 Suppose that $\mathfrak{h}(\cdot)$ is continuously differentiable and the mapping

$$x \mapsto \sup\{\|\nabla \mathfrak{h}(x+\Delta) - \nabla \mathfrak{h}(x)\| / (1 + \|\Delta\|^{p-1}) : \Delta \in \mathbb{R}^d\}$$

is bounded on compact sets. Then, for any $b_0 > 0$,

$$\sup_{Q: D_p(P,Q) \leq \delta} \mathbb{E}_Q[\mathfrak{h}(Y)] - \mathbb{E}_P[\mathfrak{h}(X)] - \delta^{1/p} \mathbb{E}_P^{(p-1)/p}[\|\nabla \mathfrak{h}(X)\|_*^{p/(p-1)}] = o\left(\delta^{1/p}\right),$$

uniformly over $P \in \mathfrak{P}([-\nu,\nu]^d)$ such that $\mathbb{E}_P \|\nabla \mathfrak{h}(Y)\| \geq b_0$.

Proof. The proof of this result is similar to the one given in the phi-divergence case. Therefore, we only provide a sketch. We start by observing that

$$\sup_{Q: D_p(P,Q) \leq \delta} \mathbb{E}_Q[\mathfrak{h}(Y)] = \mathbb{E}_P[\mathfrak{h}(X)] + \sup_{\mathbb{E}_P \|\Delta\|^p \leq \delta} \mathbb{E}_P[\mathfrak{h}(X+\Delta) - \mathfrak{h}(X)],$$

where the optimization in the right hand side is taken over random variables Δ. We let $\delta^{1/p}\bar{\Delta} = \Delta$ and note that

$$\begin{aligned}\sup_{\mathbb{E}_P \|\Delta\|^p \leq \delta} \mathbb{E}_P[\mathfrak{h}(X+\Delta) - \mathfrak{h}(X)] &= \delta^{1/p} \sup_{\mathbb{E}_P \|\bar{\Delta}\|^p \leq 1} \mathbb{E}_P[\left(\mathfrak{h}(X+\delta^{1/p}\bar{\Delta}) - \mathfrak{h}(X)\right)/\delta^{1/p}] \\ &= \delta^{1/p} \sup_{\mathbb{E}_P \|\bar{\Delta}\|^p \leq 1} \mathbb{E}_P\left[\int_0^1 \nabla \mathfrak{h}(X + t\delta^{1/p}\bar{\Delta}) \cdot \bar{\Delta} dt\right].\end{aligned}$$

Next, we can obtain a lower bound by considering a specific form of $\bar{\Delta}$ suggested by the formal asymptotic limit as $\delta \to 0$. Note that

$$\mathbb{E}_P[\nabla \mathfrak{h}(X) \cdot \bar{\Delta}] \leq \mathbb{E}_P[\|\nabla \mathfrak{h}(X)\|_* \left\|\bar{\Delta}\right\|],$$

and the equality is achieved if we choose any $\bar{\Delta}_*$ which is a constant multiple of

$$\bar{\Delta}_1(X) \in \arg\max\{\nabla \mathfrak{h}(X) \cdot \bar{\Delta} : \left\|\bar{\Delta}\right\| = 1\},$$

(The function $\bar{\Delta}_1(\cdot)$ can be selected in a measurable way using the uniformization technique of Jankov-von Neumann.) Next, if $\left\|\bar{\Delta}^*\right\| = a\|\nabla \mathfrak{h}(X)\|_*^{\gamma}$, then

$$\mathbb{E}_P[\|\nabla \mathfrak{h}(X)\|_* \left\|\bar{\Delta}^*\right\|] = a\mathbb{E}_P[\|\nabla \mathfrak{h}(X)\|_*^{\gamma+1}]$$

and

$$\mathbb{E}_P\left(\left\|\bar{\Delta}^*\right\|^p\right) = a^p \mathbb{E}_P \|\nabla \mathfrak{h}(X)\|_*^{\gamma p} = 1.$$

Letting $\gamma p = \gamma + 1$ we have that $\gamma = 1/(p-1)$ and therefore

$$\sup_{\mathbb{E}_P\left\|\bar{\Delta}\right\|^p \leq 1} \mathbb{E}_P[\nabla \mathfrak{h}(X) \cdot \bar{\Delta}^*] = \mathbb{E}_P^{(p-1)/p}[\|\nabla \mathfrak{h}(X)\|_*^{p/(p-1)}],$$

with

$$\bar{\Delta}^*(X) = \bar{\Delta}_1(X) \|\nabla \mathfrak{h}(X)\|_*^{1/(p-1)} \mathbb{E}_P^{-1/p} \|\nabla \mathfrak{h}(X)\|_*^{p/(p-1)}.$$

The denominator is well defined, since $\mathbb{E}_P \|\nabla \mathfrak{h}(Y)\| > 0$ and the random variable $\bar{\Delta}^*(X)$ is essentially bounded uniformly over the family $P \in \mathfrak{P}([-\nu, \nu]^d)$ and $\mathbb{E}_P \|\nabla \mathfrak{h}(Y)\| \geq b_0$. Since the gradient $\nabla \mathfrak{h}(\cdot)$ is continuous, then it is uniformly continuous over compact sets and, consequently, uniformly over $\bar{\Delta}$ in compact sets,

$$\int_0^1 \left\|\nabla \mathfrak{h}(x + t\delta^{1/p}\bar{\Delta}) - \nabla \mathfrak{h}(x)\right\| \bar{\Delta} dt = o(1)$$

as $\delta \to 0$. This yields that

$$\sup_{\mathbb{E}_P\left\|\bar{\Delta}\right\|^p \leq 1} \mathbb{E}_P\left[\int_0^1 \nabla \mathfrak{h}(X + t\delta^{1/p}\bar{\Delta}) \cdot \bar{\Delta} dt] \geq \mathbb{E}_P^{(p-1)/p}[\|\nabla \mathfrak{h}(X)\|_*^{p/(p-1)}\right] + o(1)$$

uniformly over $P \in \mathfrak{P}([-\nu, \nu]^d)$ and $\mathbb{E}_P \|\nabla \mathfrak{h}(Y)\| \geq b_0$. For the upper bound, we can apply the duality representation, just as we did in the phi-divergence case. Using duality, we have that

$$\sup_{\mathbb{E}_P\left\|\bar{\Delta}\right\|^p \leq 1} \mathbb{E}_P\left[\int_0^1 \nabla \mathfrak{h}(X + t\delta^{1/p}\bar{\Delta}) \cdot \bar{\Delta} dt\right] = \min_{\bar{\lambda} > 0}\left\{\bar{\lambda} + \mathbb{E}_P\left[\sup_{\bar{\Delta}} \int_0^1 \nabla \mathfrak{h}(X + t\delta^{1/p}\bar{\Delta}) \cdot \bar{\Delta} dt - \bar{\lambda}\left\|\bar{\Delta}\right\|^p\right]\right\}.$$

Once again, we select a specific choice $\bar{\lambda}_0$ given by

$$0 < \bar{\lambda}_0 = \arg\min\left\{\bar{\lambda} + \mathbb{E}_P[\sup_{\bar{\Delta}}\{\|\nabla \mathfrak{h}(X)\|_* \cdot \left\|\bar{\Delta}\right\| - \bar{\lambda}\left\|\bar{\Delta}\right\|^p\}] : \bar{\lambda} \geq 0\right\}.$$

The fact that $\bar{\lambda}_0 > 0$ follows because $\mathbb{E}_P \|\nabla \mathfrak{h}(X)\|_* > 0$. We then obtain

$$\sup_{\mathbb{E}_P\left\|\bar{\Delta}\right\|^p \leq 1} \mathbb{E}_P\left[\int_0^1 \nabla \mathfrak{h}(X + t\delta^{1/p}\bar{\Delta}) \cdot \bar{\Delta}\right] \quad \leq \bar{\lambda}_0 + \mathbb{E}_P\left[\sup_{\bar{\Delta}}\{\int_0^1 \nabla \mathfrak{h}(X + t\delta^{1/p}\bar{\Delta}) \cdot \bar{\Delta} dt - \bar{\lambda}_0\left\|\bar{\Delta}\right\|^p\}\right].$$

The rest of the proof is similar to the phi-divergence case. We omit the details due to space constraints. □

Similar results have appeared in the literature (cf., (Bartl et al. 2021)). An important difference which is useful in our analysis is that the above result is uniform over a class $P \in \mathfrak{P}([-v,v]^d)$ such that $\mathbb{E}_P \|\nabla \mathfrak{h}(Y)\| \geq b_0$. In order to write the expansion of $\mathscr{F}_n(\theta,\delta_n)$ we clarify that here we use $\nabla_x l(x,\theta)$ to denote the gradient with respect to x. Under suitable boundedness and smoothness assumptions, the previous result yields

$$\mathscr{F}_n(\theta,\delta_n) = f_n(\theta) + \delta_n^{1/p}\mathbb{E}_{P_n}^{(p-1)/p}[\|\nabla_x l(X,\theta)\|_*^{p/(p-1)}] + \delta_n^{1/p}\varepsilon_n(\theta).$$

We collect the precise statement of our result next. The proof is similar to that of Theorem 3 and thus omitted. Related results are given in (Blanchet et al. 2019; Blanchet et al. 2022).

Theorem 4 Suppose $l(\cdot,\theta)$ is continuously differentiable, that

$$(x,\theta) \mapsto \sup\{\|\nabla l(x+\Delta,\theta) - \nabla l(x,\theta)\|/(1+\|\Delta\|^{p-1}) : \|\Delta\| \geq 0\}$$

is locally bounded, that P_* has compact support, $l(x,\cdot)$ is Lipschitz continuous and

$$\inf_{\theta\in\Theta^*} \mathbb{E}_{P_*}[\|\nabla_x l(X,\theta)\|] > 0.$$

Then, we have the following types of behavior of optimal values .
(a-W) If $\delta_n^{1/p} = o\left(n^{-1/2}\right)$, then

$$n^{1/2}\left(\bar{\vartheta}_n - \vartheta\right) \rightsquigarrow \min_{\theta\in\Theta^*} \mathfrak{g}(\theta).$$

(b-W) If $\delta_n^{1/p} = n^{-1/2}$, then

$$n^{1/2}\left(\bar{\vartheta}_n - \vartheta\right) \rightsquigarrow \min_{\theta\in\Theta^*} \left\{\mathfrak{g}(\theta) + \mathbb{E}_{P_*}^{(p-1)/p}\left[\|\nabla_x l(X,\theta)\|_*^{p/(p-1)}\right]\right\}.$$

(c-W) If $o\left(\delta_n^{1/p}\right) = n^{-1/2}$, then

$$\delta_n^{-1/p}\left(\bar{\vartheta}_n - \vartheta\right) \rightsquigarrow \min_{\theta\in\Theta^*} \mathbb{E}_{P_*}^{(p-1)/p}[\|\nabla_x l(X,\theta)\|_*^{p/(p-1)}].$$

5 GENERAL PRINCIPLE IN ACTION: OPTIMAL SOLUTIONS

We complete our discussion in this section, considering optimal solutions. Due to space constraints, we focus only on the phi-divergence case. A key observation is that the uncertainty set is compact in the weak topology and therefore, if Assumption 3 holds, the function $\mathscr{F}_n(\cdot,\delta_n)$ is differentiable and its gradient has expansion (17). In fact, the derivative can be shown to exist if we are able to argue that, for δ sufficiently small, the worst-case measure is unique. This is precisely the strategy we will pursue in this section. Throughout the section, we impose the condition that $\Theta^* = \{\theta^*\}$. Recall that $\sigma^2(\theta) := \mathrm{Var}_{P_*}(l(X,\theta))$.

Theorem 5 Suppose that Assumptions 2, 3 and 4 hold, that $l(x,\cdot)$ is essentially bounded under P_* and $\sigma^2(\theta^*) > 0$, and that $\theta_*(v)$ is directionally differentiable at $v = 0$ (in the Hadamard sense). Let $Z \sim N(0,\Sigma)$, where Σ is the covariance matrix of $\nabla l(X,\theta^*)$. Then we have the following.
(A-phi) If $\delta_n = o\left(n^{-1}\right)$, then

$$n^{1/2}\left(\bar{\theta}_n - \theta_*\right) \rightsquigarrow \theta_*'(0,Z).$$

(B-phi) If $\delta_n = n^{-1}$, then

$$n^{1/2}\left(\bar{\theta}_n - \theta_*\right) \rightsquigarrow \theta_*'\left(0, Z + \kappa^{1/2}\nabla\sigma(\theta^*)\right).$$

(C-phi) If $o(\delta_n) = n^{-1}$, then

$$\delta_n^{-1/2}\left(\bar{\theta}_n - \theta_*\right) \rightsquigarrow \theta_*'\left(0, \kappa^{1/2}\nabla\sigma(\theta^*)\right).$$

Proof. Applying the centering and scaling used to obtain (20) we obtain

$$\mathscr{F}_n(\theta,\delta_n) = f_n(\theta) + \delta_n^{1/2} D_n(\theta,\delta_n),$$

where

$$\mathscr{D}_n(\theta,\delta_n) = \sup_{\mathbb{E}_{P_n}(\Delta)=0, \Delta \geq -\delta_n^{-1/2}, \delta_n^{-1/2}\mathbb{E}_{P_n}(\phi(1+\delta_n^{1/2}\Delta))\leq 1} \mathbb{E}_{P_n}[l_n(X,\theta)\Delta], \tag{27}$$

and

$$\bar{l}_n(X,\theta) = l(X,\theta) - f_n(\theta).$$

It suffices to show that

$$\nabla \mathscr{D}_n(\theta,\delta_n) \to \nabla \rho(\theta)$$

uniformly over some region $\|\theta-\theta_*\| \leq \delta_0$ for some δ_0. Note that the optimization region in (27) is compact in the weak topology and therefore, by Danskin's Theorem (see (Shapiro et al. 2009, sections 5.1.3 and 7.1.5), Section 7), we have that $\mathscr{D}_n(\cdot,\delta_n)$ is directionally differentiable and by the uniqueness of the optimal $\bar{\Delta}_n$ for δ_n sufficiently small we have that

$$\nabla \mathscr{D}_n(\theta,\delta_n) = \mathbb{E}_{P_n}[\nabla l_n(X,\theta)\bar{\Delta}_n(\theta)].$$

We can precisely characterize $\bar{\Delta}_n(\theta)$ from Proposition 3 over a region $\|\theta-\theta_*\| \leq \delta_0$ for which we can guarantee $Var_{P_n}[l(X,\theta)] > 0$. Note that such $\delta_0 > 0$ can be found assuming that $n > N$ (for some random but finite almost surely N because of the Strong Law of Large Numbers and continuity since $Var_{P_*}[l(X,\theta_*)] > 0$. We have, uniformly over $\|\theta-\theta_*\| \leq \delta_0$, for $n > N$,

$$\bar{\Delta}_n(\theta) = \sqrt{\kappa}\frac{l_n(X,\theta)}{\sqrt{\phi''(1)Var_{P_n}[l(X,\theta)]}} + \varepsilon_n(\theta).$$

On the other hand, defining

$$\bar{l}(X,\theta) = l(X,\theta) - f(\theta),$$

we have that

$$\nabla \rho(\theta) = \mathbb{E}_{P_*}[\bar{l}(X,\theta)\cdot\bar{\Delta}(\theta)],$$

where

$$\bar{\Delta}(\theta) = \sqrt{\kappa}\frac{\bar{l}(X,\theta)}{\sqrt{\phi''(1)Var_{P_*}[l(X,\theta)]}}.$$

We obtain

$$\begin{aligned}
&\nabla \mathscr{D}_n(\theta,\delta_n) - \nabla\rho(\theta) \\
&= \mathbb{E}_{P_n}[\nabla l_n(X,\theta)\bar{\Delta}_n(\theta)] - \mathbb{E}_{P_*}[\nabla\bar{l}(X,\theta)\cdot\bar{\Delta}(\theta)] \\
&= \mathbb{E}_{P_n}[\left(\nabla l_n(X,\theta) - \nabla\bar{l}(X,\theta)\right)\bar{\Delta}_n(\theta)] + \mathbb{E}_{P_n}[\nabla\bar{l}(X,\theta)(\bar{\Delta}_n(\theta)-\bar{\Delta}(\theta))] \\
&+ \mathbb{E}_{P_n}[\nabla\bar{l}(X,\theta)\bar{\Delta}(\theta)] - \mathbb{E}_{P_*}[\nabla\bar{l}(X,\theta)\cdot\bar{\Delta}(\theta)].
\end{aligned}$$

It follows that $\bar{\Delta}_n(\theta) \to \bar{\Delta}(\theta)$ uniformly over $\|\theta-\theta_*\| \leq \delta_0$, and $\left(\nabla l_n(X,\theta) - \nabla\bar{l}(X,\theta)\right) \to 0$ uniformly in probability (in fact almost surely) as $n \to 0$. Uniform convergence in probability over $\|\theta-\theta_*\| \leq \delta_0$ follows from these observations. □

ACKNOWLEDGMENTS

J. Blanchet's research was partially supported by the Air Force Office of Scientific Research (AFOSR), award FA9550-20-1-0397, with additional from NSF 1915967, 2118199, 2229012, 2312204. The research of A. Shapiro was partially supported by (AFOSR) Grant FA9550-22-1-0244. We also acknowledge Dr. Yang Liu's assistance in proofreading some sections of this paper.

REFERENCES

Bartl, D., S. Drapeau, J. Obłój, and J. Wiesel. 2021. "Sensitivity Analysis of Wasserstein Distributionally Robust Optimization Problems". *Proc. of the Royal Society A.* 447:2256.

Bayraksan, G., and D. K. Love. 2015. "Data-Driven Stochastic Programming Using Phi-Divergences". *Tutorials in Operations Research, INFORMS*:1563–1581.

Ben-Tal, A., and M. Teboulle. 1987. "Penalty Functions and Duality in Stochastic Programming via Phi-Divergence Functionals". *Mathematics of Operations Research* 12:224–240.

Blanchet, J., Y. Kang, and K. Murthy. 2019. "Robust Wasserstein Profile Inference and Applications to Machine Learning". *Journal of Applied Probability* 56:830–857.

Blanchet, J., and K. Murthy. 2019. "Quantifying Distributional Model Risk via Optimal Transport". *Mathematics of Operations Research* 44:377–766.

Blanchet, J., K. Murthy, and N. Si. 2022. "Confidence Regions in Wasserstein Distributionally Robust Estimation". *Biometrika* 109:295–315.

Bonnans, J. F., and A. Shapiro. 2000. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer.

Csiszár, I. 1963. "Eine Informationstheoretische Ungleichung und Ihre Anwendung auf Den Beweis der Ergodizitat von Markoffschen Ketten". *Magyar. Tud. Akad. Mat. Kutato Int. Kozls* 8:85–108.

Duchi, J. C., P. W. Glynn, and H. Namkoong. 2021. "Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach". *Mathematics of Operations Research* 46(3):946–969.

Esfahani, P. M., and D. Kuhn. 2018. "Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations". *Mathematical Programming* 171:115–166.

Gao, R., and A. Kleywegt. 2016. "Distributionally Robust Stochastic Optimization with Wasserstein Distance". *arXiv preprint arXiv:1604.02199*.

Lam, H. 2016. "Robust Sensitivity Analysis for Stochastic Systems.". *Math. of Oper. Research* 41:1248–1275.

Morimoto, T. 1963. "Markov Processes and the H-Theorem". *J. Phys. Soc. Jap.* 18(3):328–333.

Rahimian, H., and S. Mehrotra. 2019. "Distributionally Robust Optimization: A Review". *Arxiv* https://arxiv.org/abs/1908.05659.

Shapiro, A. 1991. "Asymptotic Analysis of Stochastic Programs". *Annals of Operations Research* 30:169–186.

Shapiro, A. 1993. "Asymptotic Behavior of Optimal Solutions in Stochastic Programming". *Mathematics of Operations Research* 18:829 – 845.

Shapiro, A. 2017. "Distributionally Robust Stochastic Programming". *SIAM J. Optimization* 27:2258–2275.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. 2009. *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia: SIAM.

van der Vaart, A. 1998. *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Villani, C. 2003. *Topics in Optimal Transportation*. Graduate Studies in Mathematics, Vol. 58: American Mathematical Society.

AUTHOR BIOGRAPHIES

JOSE BLANCHET is a Professor of Management Science and Engineering (MS&E) at Stanford. Prior to joining Stanford, he was a professor at Columbia (Industrial Engineering and Operations Research, and Statistics, 2008-2017), and before that he taught at Harvard (Statistics, 2004-2008). Jose is a recipient of the 2010 Erlang Prize and several best publication awards in areas such as applied probability, simulation, operations management, and revenue management. He also received a Presidential Early Career Award for Scientists and Engineers in 2010. He is the Area Editor of Stochastic Models in Mathematics of Operations Research and has served on the editorial board of Advances in Applied Probability, Bernoulli, Extremes, Insurance: Mathematics and Economics, Journal of Applied Probability, Queueing Systems: Theory and Applications, and Stochastic Systems, among others. His email address is jose.blanchet@stanford.edu.

ALEXANDER SHAPIRO is the A. Russell Chandler III Chair and Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. Dr. Shapiro's research interests are focused on stochastic programming, risk analysis, simulation-based optimization, and multivariate statistical analysis. In 2013 he was awarded the Khachiyan Prize of INFORMS for lifetime achievements in optimization, and in 2018 he was a recipient of the Dantzig Prize awarded by the Mathematical Optimization Society and the Society for Industrial and Applied Mathematics. In 2020 he was elected to the National Academy of Engineering. In 2021 he was a recipient of the John von Neumann Theory Prize awarded by the Institute for Operations Research and the Management Sciences (INFORMS). Dr. Shapiro served on the editorial board of a number of professional journals. He was an area editor (optimization) of the Operations Research Journal and the editor-in-chief of the journal Mathematical Programming, Series A. His email address is ashapiro@isye.gatech.edu.