Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach

John C. Duchi[1] Peter W. Glynn[2] Hongseok Namkoong[2]

[1]Department of Statistics

[2]Department of Management Science and Engineering

Stanford University {jduchi,glynn,hnamk}@stanford.edu

Stanford University

Abstract

We study statistical inference and robust solution methods for stochastic optimization problems. We first develop an empirical likelihood framework for stochastic optimization. We show an empirical likelihood theory for Hadamard differentiable functionals with general f-divergences and give conditions under which $T(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ is Hadamard differentiable. Noting that the right endpoint of the generalized empirical likelihood confidence interval is a distributionally robust optimization problem with uncertainty regions given by f-divergences, we show various statistical properties of robust optimization. First, we give a statistically principled method of choosing the size of the uncertainty set to obtain a *calibrated* one-sided confidence interval. Next, we give general conditions under which the robust solutions are consistent. Finally, we prove an asymptotic expansion for the robust formulation, showing how robustification regularizes the problem.

1 Introduction

In this paper, we study the properties of robust solution methods, in particular statistical and inferential guarantees, for the stochastic optimization problem

$$\underset{x\in\mathcal{X}}{\text{minimize}} \ \mathbb{E}_{P_0}[\ell(x;\xi)] = \int_\Xi \ell(x;\xi) dP_0(\xi). \tag{1}$$

In the formulation (1), the feasible region $\mathcal{X} \subset \mathbb{R}^d$ is a nonempty closed set, ξ is a random vector on the probability space $(\Xi, \mathcal{A}, P_0)$, where the domain Ξ is a (subset of) a separable metric space, and the function $\ell : \mathcal{X} \times \Xi \to \mathbb{R}$ is a lower semi-continuous (loss) functional. In most data-based decision making scenarios, the underlying distribution P_0 is unknown. Even in scenarios (*e.g.*, simulation optimization) where P_0 is known, the integral $\mathbb{E}_{P_0}[\ell(x;\xi)]$ may be high-dimensional and intractable to compute. Consequently, it is standard [e.g. 56] to approximate the population objective (1) using the standard sample average approximation (SAA) based on a (Monte Carlo) sample $\xi_1, \ldots, \xi_n$ drawn from P_0, giving the empirical problem

$$\underset{x\in\mathcal{X}}{\text{minimize}} \ \mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)] = \frac{1}{n}\sum_{i=1}^n \ell(x;\xi_i). \tag{2}$$

Here $\widehat{P}_n := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\xi_i}$ denotes the usual empirical measure.

In this paper, we are concerned with inference—the construction of confidence intervals—and consistency of approximate solutions for the problem (1) based on variants of the SAA objective (2). We develop a family of optimization programs, based on the distributionally robust optimization framework [23, 6, 10, 7], which allow us to provide asymptotically calibrated confidence intervals,

along with certificates based on convex programming, for optimal values of the problem (1) (as well as approximate solutions $\widehat{x}$ achieving a guaranteed level of performance). More concretely, if we define the optimal value functional T by

$$T(P) := \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x; \xi)],$$

we show how to construct estimates l_n and u_n based on the sample $\xi_1, \ldots, \xi_n$, so that for a given confidence level α, $[l_n, u_n]$ is a *calibrated* confidence interval

$$\lim_{n \to \infty} \mathbb{P}\left(T(P) \in [l_n, u_n]\right) = 1 - \alpha \tag{3}$$

for all distributions P on ξ. Statistically speaking, calibratedness is a highly desirable property as it achieves the desired confidence level exactly in the large sample limit. We also give sharper statements than the asymptotic guarantee (3), providing expansions for l_n and u_n and giving rates at which $u_n - l_n \to 0$.

We describe our approach briefly before summarizing our main contributions, discussing related approaches simultaneously. To begin, we recall the f-divergence [3, 21]. Let $f : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ be a closed convex function satisfying $f(1) = 0$. Then for any distributions P and Q on a space Ξ, the *f-divergence* between P and Q is

$$D_f\left(P \| Q\right) = \int f\left(\frac{dP}{dQ}\right) dQ = \int_{\Xi} f\left(\frac{p(\xi)}{q(\xi)}\right) q(\xi) d\mu(\xi),$$

where μ is any σ-finite measure for which $P, Q \ll \mu$, and $p = dP/d\mu$ and $q = dQ/d\mu$. With this definition, we define the upper and lower confidence bounds

$$u_n := \inf_{x \in \mathcal{X}} \sup_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(x; \xi)] : D_f(P \| \widehat{P}_n) \le \frac{\rho}{n} \right\} \tag{4a}$$

$$l_n := \inf_{x \in \mathcal{X}} \inf_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(x; \xi)] : D_f(P \| \widehat{P}_n) \le \frac{\rho}{n} \right\}. \tag{4b}$$

In the formulation (4), the parameter $\rho = \chi^2_{1,1-\alpha}$ is chosen as the $(1-\alpha)$-quantile of the χ^2_1 distribution, that is, $\mathbb{P}(Z^2 \le \rho) = 1 - \alpha$ for $Z \sim \mathsf{N}(0, 1)$, and for the exact result (3) to hold, we assume that the convex function f is C^2 in a neighborhood of 1 with the normalization $f''(1) = 2$.

The upper endpoint (4a) is a natural robust formulation for the sample average approximation (2), proposed by [7] for distributions P with finite support. The approach in the current paper applies to essentially arbitrary distributions, and we explicitly link robust optimization formulations with stochastic optimization—approaches often considered somewhat dichotomous [6]. That is, we show how a robust optimization approach to dealing with parameter uncertainy yields solutions with a number of desirable statistical properties, and we address the questions that naturally arise in viewing robust optimization from a statistical perspective.

We now summarize our contributions, which unify the approach to uncertainy based on robust optimization with classical inferential goals.

(i) We develop an empirical likelihood framework for the value function $T(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x; \xi)]$. We show how the construction (4a)–(4b) of $[l_n, u_n]$ gives a calibrated (3) confidence interval for $T(P_0)$. To do so, we extend Owen's empirical likelihood theory [46, 45] to a general collection of nonparametric functionals (the most general that we know in the literature) with general divergence measures. Our proof is different to the classical result of Owen's and gives a novel justification for the inference framework when applied to the simple setting $T(P) = E_P[X]$.

(ii) We also show that the upper confidence set $(-\infty, u_n]$ is a calibrated one-sided confidence interval when $\rho = \chi^2_{1,1-2\alpha} = \inf\{\rho' : \mathbb{P}(Z^2 \le \rho') \ge 1 - 2\alpha, Z \sim \mathsf{N}(0,1)\}$. That is, under suitable conditions on ℓ and P_0,

$$\lim_{n\to\infty} \mathbb{P}\left(\inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \in (-\infty, u_n]\right) = 1 - \alpha.$$

This shows that the robust optimization problem (4a) provides a sharp statistical certificate for the attainable performance under the population loss (1), which is (often) efficiently computable when ℓ is convex.

(iii) We show that the robust formulation (4a) has the (almost sure) asymptotic expansion

$$\sup_{P \ll \widehat{P}_n} \left\{\mathbb{E}_P[\ell(x;\xi)] : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\right\} = \mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)] + (1 + o(1))\sqrt{\frac{\rho}{n}\mathrm{Var}_P(\ell(x;\xi))},$$

and that this expansion is uniform in x under mild restrictions. Viewing the second term in the expansion as a regularizer for the SAA problem (2) makes concrete the intuition that robust optimization provides regularization; the regularizer accounts for the variance of the objective function (which is generally non-convex in x even if ℓ is convex), reducing uncertainty. We give weak conditions under which the expansion is uniform in x, showing that the regularization interpretation is valid when we choose $\widehat{x}_n$ to minimize the robust formulation (4a).

(iv) Lastly, we prove consistency of estimators $\widehat{x}_n$ attaining the infimum in the problem (4a) under essentially the same conditions required for that of the SAA. More precisely, for the sets of optima defined by

$$S^\star := \operatorname*{argmin} \mathbb{E}_{P_0}[\ell(x;\xi)] \qquad S_n^\star := \operatorname*{argmin}_{x\in\mathcal{X}} \sup_{P\ll\widehat{P}_n} \left\{\mathbb{E}_P[\ell(x;\xi)] : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\right\},$$

we show that the Hausdorff distance between $S_n^\star$ and $S^\star$ tends to zero so long as ℓ has more than one moment under P_0 and is lower semicontinuous.

Background and prior work

The nonparametric inference framework for stochastic optimization we develop in this paper can be viewed as the empirical likelihood counterpart of the normality theory that Shapiro develops [52, 54]. While an extensive literature exists on statistical inference for stochastic optimization problems (see, for example, the line of work developed by Dupacová, Wets, King, Shapiro, Rockaffelar, and others [26, 52, 31, 53, 33, 54, 32, 55], as well as the book of Shapiro et al. [56]), Owen's empirical likelihood framework [47] has received little attention in the stochastic optimization literature. The only exception we know of is that by Lam and Zhou [38]. In its classical form, empirical likelihood provides a confidence set for a d-dimensional mean $\mathbb{E}_{P_0}[Y]$ based on an empirical distribution $\widehat{P}_n$ supported on a sample by using the set $C_{\rho,n} := \{\mathbb{E}_P[Y] : D_{\mathrm{kl}}(\widehat{P}_n\|P) \le \rho/n\}$. Here, $D_{\mathrm{kl}}(P\|Q)$ is the Kullback-Leibler divergence (with $f(t) = t\log t$). Empirical likelihood is asymptotically pivotal, meaning that if we choose

$$\rho = \chi^2_{d,1-\alpha} := \inf\left\{\rho' : \mathbb{P}(\|Z\|_2^2 \le \rho') \ge 1 - \alpha \ \text{ for } Z \sim \mathsf{N}(0, I_{d\times d})\right\},$$

then $\mathbb{P}(\mathbb{E}_{P_0}[Y] \in C_{\rho,n}) \to 1 - \alpha$. This convergence requires no knowledge or estimation of unknown quantities, such as variance. We show how such asymptotically pivotal results also apply for the robust optimization formulation (4).

Using confidence sets to robustify optimization problems involving random parameters is a common technique (see, for example, Chapter 2 of the book of Ben-Tal et al. [6]). A number of researchers have extended such techniques to situations in which one observes a sample $\xi_1, \ldots, \xi_n$ and constructs an uncertainty set over the data directly, including the papers [23, 60, 7, 10, 11]. The duality of confidence regions and hypothesis tests [39] gives a natural connection between robust optimization, uncertainty sets, and statistical tests. Delage and Ye [23] took initial steps in this direction by constructing confidence regions based on mean and covariance matrices from the data, and Jiang and Guan [30] expanded this line of research to other moment constraints. Bertsimas, Gupta, and Kallus [10, 11] developed uncertainty sets based on various linear and higher-order moment conditions; they also propose a robust SAA formulation based on goodness of fit tests, showing tractability as well as some consistency results based on Scarsini's linear convex orderings [51] so long as the underlying distribution is bounded; they also give uncalibrated confidence sets. The formulation (4) has similar motivation to the preceding works, as the uncertainty set

$$\left\{\mathbb{E}_P[\ell(x;\xi)] : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\right\}$$

is a confidence region for $\mathbb{E}_{P_0}[\ell(x;\xi)]$ for each fixed $x \in \mathcal{X}$ (as we show in the sequel). Our results take a step further in this direction by showing that, under mild conditions, the values (4a) and (4b) simultaneously provide exactly calibrated upper and lower confidence bounds.

Ben-Tal et al. [7] explore a similar scenario to ours, focusing on the robust formulation (4a), and they show that when P_0 is *finitely* supported, the robust program (4a) gives an *uncalibrated* one-sided confidence interval. In the unconstrained setting, Lam and Zhou [38] also show that standard empirical likelihood theory can give confidence bounds for stochastic optimization problems via estimating equations, but their confidence regions may be larger than ours. The result (i) generalizes these works, as we show how the robust formulations (4) yield calibrated confidence intervals for general distributions P under standard moment conditions on the loss ℓ for general constrained stochastic optimization problems.

Ben-Tal et al.'s robust sample approximation [7], and Bertsimas et al.'s goodness of fit testing-based procedures [11] provide natural motivation for formulations similar to ours (4). By considering completely nonparametric measures of fit, however—as in empirical likelihood theory [47]—we can depart from assumptions on the structure of Ξ (i.e. that it is finite or a compact subset of $\mathbb{R}^d$). The f-divergence formulation (4) allows a more nuanced understanding of the underlying structure of the population problem (1), and it also allows the precise confidence statements, expansions, and consistency guarantees outlined in (i)–(iii). Lam [37] derives an expansion similar in spirit to our contribution (iii), though it is restricted to the KL-divergence and applies pointwise in x, focusing on a simulation setting where the goal is to evaluate model sensitivity. His expansion requires finite moment generating functions of the base measure P_0 and does not immediately apply to sample-based approximations. The richness of the family of f-divergences offers significant modeling flexibility, which allows a modeler to select appropriate divergences for her problem (for example, to allow easier computation), as well as specifying a desired confidence level for appropriate robustness.

Notation We collect our mostly standard notation here. For a sequence of random variables $X_1, X_2, \ldots$ in a metric space $\mathcal{X}$, we say $X_n \overset{d}{\rightsquigarrow} X$ if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded continuous functions f. We write $X_n \overset{P^*}{\to} X$ for random variables X_n converging to a random variable X in outer probability. Given sets $A, B \subset \mathbb{R}^d$ and a norm $\|\cdot\|$, the Hausdorff distance between A and B

is

$$d_{\rm haus}(A,B) := \max\left\{\sup_{x\in A}\operatorname{dist}(x,B), \sup_{y\in B}\operatorname{dist}(y,A)\right\} \text{ where } \operatorname{dist}(x,B) = \inf_{y\in B}\|x-y\|.$$

For a measure μ on a measurable space $(\Xi, \mathcal{A})$ and $p \geq 1$, we let $L^p(\mu)$ be the usual L^p space, that is, $L^p(\mu) := \{f : \Xi \to \mathbb{R} \mid \int |f|^p d\mu < \infty\}$. For a deterministic or random sequence $a_n \in \mathbb{R}$, we say that a sequence of random variables X_n is $O_P(a_n)$ if $\lim_{c\to\infty} \limsup_n P(|X_n| \geq c \cdot a_n) = 0$. Similarly, we say that $X_n = o_P(a_n)$ if $\limsup P(|X_n| \geq c \cdot a_n) = 0$ for all $c > 0$. For a function f, we let $f^*(y) = \sup_x\{y^T x - f(x)\}$ denote its conjugate. For a convex function $f : \mathbb{R} \to \mathbb{R}$, we define the right derivative $f'_+(x) = \lim_{\delta\downarrow 0} \frac{f(x+\delta)-f(x)}{\delta}$, which must exist.

Outline

The rest of the paper is organized as follows. In Section 2, we give a generalized empirical likelihood theory for general nonparametric statistical functionals. In Section 3, we apply this theory to stochastic optimization problems and give computationally tractabile procedures. In Section 4, we give a statistical interpretation to robust optimization and show a principled way of choosing the size of the uncertainty set ρ to get a calibrated upper confidence bound on the population optimal value (1). We also state basic properties of the formulation (4a) from a robust optimization perspective and see how robust optimization regularizes the problem. In Section 5, consistency of the empirical robust optimizers are shown, verifying that the robust formulation (4a) is a competitive approach compared to the SAA. In Section 6, we present numerical experiments and conclude the paper in Section 7.

2 Generalized Empirical Likelihood and Asymptotic Expansions

In this section, we abstract away from the stochastic optimization setting that motivates us to develop our basic theoretical results. We begin by briefly reviewing generalized empirical likelihood theory [44, 29], giving a presentation dovetailing with our goals. Let $Z_1, \ldots, Z_n$ be independent random vectors—formally, measurable functions $Z : \Xi \to \mathbb{B}$ for some Banach space $\mathbb{B}$—with common distribution P_0, and let $P \mapsto T(P) \in \Theta$ (for some space Θ) be a function of interest. In our setting, this mapping generally corresponds to $T(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$, so that $Z(\xi)$ is the function $\ell(\cdot;\xi)$. Then the *generalized empirical likelihood confidence region* for $T(P_0)$ is

$$C_{n,\rho} := \left\{T(P) : D_f(P\|\widehat{P}_n) \leq \frac{\rho}{n}\right\},$$

where $\widehat{P}_n$ is the empirical distribution of $Z_1, \ldots, Z_n$. As $C_{n,\rho}$ is the image of T on the neighborhood of the empirical distribution $\widehat{P}_n$ given by the f-divergence, it forms a natural confidence region for $T(P_0)$ (for T suitably smooth with respect to distributions). We may define a dual quantity, the profile divergence $R_n : \Theta \to \mathbb{R}_+$, by

$$R_n(\theta) := \inf_{P \ll \widehat{P}_n}\left\{D_f(P\|\widehat{P}_n) : T(P) = \theta\right\}.$$

By inspection, we have $T(P_0) \in C_{n,\rho}$ if and only if $R_n(T(P_0)) \leq \frac{\rho}{n}$. Classical empirical likelihood, developed by [46, 45, 47], considers the above setting for $f(t) = -2\log t$, that is, $D_f(P\|\widehat{P}_n) = 2D_{\rm kl}(\widehat{P}_n\|P)$, so that the divergence is the nonparametric log-likelihood ratio. The goal is then to show that for appropriately smooth functionals T that

$$\mathbb{P}(T(P_0) \in C_{n,\rho}) = \mathbb{P}\left(R_n(T(P_0)) \leq \frac{\rho}{n}\right) \to 1 - \alpha(\rho) \text{ as } n \to \infty,$$

where $\alpha(\rho)$ is a desired confidence level (based on ρ) for the inclusion $T(P_0) \in C_{n,\rho}$.

2.1 Generalized Empirical likelihood for means of finite-dimensional random vectors

In the classical case in which the vectors $Z_i \in \mathbb{R}^d$, Owen [45] shows that empirical likelihood applied to the mean $\mathbb{E}_{P_0}[Z]$ guarantees elegant asymptotic properties. More precisely, one obtains the convergence $R_n(\mathbb{E}_{P_0}[Z]) \stackrel{d}{\rightsquigarrow} \chi^2_{d_0}$, where $\chi^2_{d_0}$ denotes the χ^2-distribution with d_0 degrees of freedom, whenever $\mathrm{Cov}(Z)$ has rank $d_0 \le d$. That is,, $C_{n,\rho(\alpha)}$ is a calibrated $(1-\alpha)$-confidence interval for $T(P_0) = \mathbb{E}_{P_0}[Z]$ if we set $\rho(\alpha) = \inf\{\rho' : \mathbb{P}(\chi^2_{d_0} \le \rho') \ge 1-\alpha\}$. We will extend these results to somewhat more general functions T and to a variety of f-divergences satisfying the following condition, which we henceforth assume without mention.

Assumption A (Smoothness of f-divergence)**.** *The function $f : \mathbb{R}_+ \to \overline{\mathbb{R}}$ is convex, three times differentiable on a neighborhood of 1, and satisfies $f(1) = f'(1) = 0$ and $f''(1) = 2$.*

The assumption that $f(1) = f'(1) = 0$ is no loss of generality, as the function $t \mapsto f(t)+c(t-1)$ yields identical divergence measures to f, and the assumption that $f''(1) = 2$ is simply a normalization for easier calculation. We make no restrictions on the behavior of f at 0, as a number of popular divergence measures, such as empirical likelihood with $f(t) = -2\log t + 2t - 2$, approach infinity as $t \downarrow 0$.

The following proposition is a generalization of Owen's results [45] to smooth f-divergences. We provide a novel proof of the result, which is essentially known [9], using the asymptotic expansion we develop in the sequel.

Proposition 1. *Let Assumption A hold. Let $Z_i \in \mathbb{R}^d$ be i.i.d. P_0 with finite covariance of rank $d_0 \le d$. Then*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{E}_{P_0}[Z] \in \left\{\mathbb{E}_P[Z] : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\right\}\right) = \mathbb{P}\left(\chi^2_{d_0} \le \rho\right). \tag{5}$$

When $d = 1$, the proposition is a direct consequence of Lemma 1 to come; for more general dimensions d, we present the proof in Appendix C.1. If we consider the random variable $Z_x(\xi) := \ell(x;\xi)$, defined for each $x \in \mathcal{X}$, Proposition 1 allows us to construct pointwise confidence intervals for the distributionally robust problems (4). However, we require somewhat stronger results than the pointwise guarantee (5), for which we now develop an asymptotic expansion that essentially gives all of the major distributional convergence results in this paper.

2.2 An asymptotic expansion

As mentioned previously, our results on convergence and calibration build on two asymptotic expansions, which we now present. The first we may state without any further explanation.

Lemma 1. *Let $Z : \Xi \to \mathbb{R}$ be a random variable with $\mathrm{Var}(Z) < \infty$, and let Assumption A hold. Let $s_n^2 = \mathbb{E}_{\widehat{P}_n}[Z^2] - \mathbb{E}_{\widehat{P}_n}[Z]^2$ denote the sample variance of Z. Then there exists a sequence ε_n with $\varepsilon_n \to 0$ as $n \to \infty$ such that for any $\epsilon > 0$*

$$\left| \sup_{P : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[Z] - \mathbb{E}_{\widehat{P}_n}[Z] - \sqrt{\frac{\rho}{n} s_n^2} \right| \le c_f \mathbb{E}[Z^2]^{\frac{1}{2}} \sqrt{\frac{\epsilon}{n} + \frac{\rho^2}{n^2}} + \frac{\varepsilon_n}{\sqrt{n}} \tag{6}$$

eventually with probability 1.

See Appendix A for the proof. For intuition on the importance of the expansion (6), note that we may rewrite it as

$$\sup_{P:D_f(P\|\widehat{P}_n)\le\frac{\rho}{n}} \mathbb{E}_P[Z] = \mathbb{E}_{\widehat{P}_n}[Z] + \sqrt{\frac{\rho}{n}\mathrm{Var}_{\widehat{P}_n}(Z)} + o_{P_0}(1/\sqrt{n}) \;\; \text{when } \xi_i \overset{\text{iid}}{\sim} P_0.$$

But then the classical central limit theorem implies that the latter term satisfies

$$\begin{aligned}\mathbb{P}\left(\mathbb{E}_{P_0}[Z] \le \mathbb{E}_{\widehat{P}_n}[Z] + \sqrt{\frac{\rho}{n}\mathrm{Var}_{\widehat{P}_n}(Z)}\right) &= \mathbb{P}\left(0 \le \sqrt{n}(\mathbb{E}_{\widehat{P}_n}[Z] - \mathbb{E}_{P_0}[Z]) + \sqrt{\rho\mathrm{Var}_{\widehat{P}_n}(Z)}\right) \\ &\underset{n\uparrow\infty}{\to} \mathbb{P}(W \le \sqrt{\rho}) = \mathbb{P}(W^2 \le \rho)\end{aligned}$$

where $W \sim \mathsf{N}(0,1)$, yielding Proposition 1 in the case that $d = 1$.

A more general story requires somewhat more notation and background in empirical processes, which we now provide (though see, for example, the book of van der Vaart and Wellner [59] for a much fuller treatment). Let P_0 be a fixed probability distribution on the measurable space $(\Xi, \mathcal{A})$, and recall the usual L^2 space $L^2(P_0)$, where we equip functions with the $L^2(P_0)$ norm $\|h\|_{L^2(P_0)} = \mathbb{E}_{P_0}[h(\xi)^2]^{\frac{1}{2}}$. For any measure μ on Ξ and $h : \Xi \to \mathbb{R}$, we use the functional shorthand $\mu h := \int h(\xi) d\mu(\xi)$ so that for any probability measure we have $Ph = \mathbb{E}_P[h(\xi)]$. Now, for a set $\mathcal{H} \subset L^2(P_0)$, let $\mathcal{L}^\infty(\mathcal{H})$ be the space of linear functionals on $\mathcal{H}$ bounded with the uniform norm $\|L_1 - L_2\|_{\mathcal{H}} = \sup_{h\in\mathcal{H}} |L_1 h - L_2 h|$ for $L_1, L_2 \in \mathcal{L}^\infty(\mathcal{H})$. To avoid measurability issues, we use outer probability and expectation with the corresponding convergence notions as necessary [e.g. 59, Section 1.2]. We then have the following definition [cf. 59, Eq. (2.1.1)].

Definition 1. *A class of functions $\mathcal{H}$ is P_0-*Donsker *if $\sqrt{n}(\widehat{P}_n - P_0) \overset{d}{\rightsquigarrow} G$ in the space $\mathcal{L}^\infty(\mathcal{H})$, where G is a tight Borel measurable element of $\mathcal{L}^\infty(\mathcal{H})$, and $\widehat{P}_n$ is the empirical distribution of $\xi_i \overset{\text{iid}}{\sim} P_0$.*

With these preliminaries in place, we can state a more general result than Lemma 1. For this theorem, we require a space $\mathcal{X}$, which we treat abstractly for now, and for each $x \in \mathcal{X}$ we let $Z(x, \cdot) : \Xi \to \mathbb{R}$ be a random variable. We assume also that the collection $\mathcal{Z} = \{Z(x, \cdot) \mid x \in \mathcal{X}\}$ is P_0-Donsker (Definition 1) when viewed as mappings $\Xi \to \mathbb{R}$. Additionally, we assume that $\mathcal{Z}$ has L^2-integrable envelope, that is, that there is some function $M_2 : \Xi \to \mathbb{R}$ such that $Z(x, \xi) \le M_2(\xi)$ for all $x \in \mathcal{X}$ and $\mathbb{E}_{P_0}[M_2(\xi)^2] < \infty$.

Theorem 2. *Let the conditions of the preceding paragraph hold. Then*

$$\sup_{P:D_f(P\|\widehat{P}_n)\le\frac{\rho}{n}} \mathbb{E}_P[Z(x,\xi)] = \mathbb{E}_{\widehat{P}_n}[Z(x,\xi)] + \sqrt{\frac{\rho}{n}\mathrm{Var}_{P_0}(Z(x,\xi))} + \varepsilon_n(x),$$

where $\sup_{x\in\mathcal{X}} \sqrt{n}|\varepsilon_n(x)| \overset{P^*}{\to} 0$.

This theorem, whose proof we provide in Appendix B.3, is the main technical result off of which the majority of the results in this paper build; we state it here for convenience later, applying it to prove all of our distributional convergence results (including Proposition 1).

2.3 Hadamard Differentiable Functionals

In this section, we present an analogue of Proposition 1 for smooth functionals of probability distributions, which we use in the optimization context subsequently. Let $(\Xi, \mathcal{A})$ be a measurable space, and $\mathcal{H}$ be a collection of functions $h : \Xi \to \mathbb{R}$, where we assume that $\mathcal{H}$ is P_0-Donsker with envelope $M_2 \in L^2(P_0)$ (Definition 1). Let $\mathcal{P}$ be the space of probability measures on $(\Xi, \mathcal{A})$ bounded with respect to the supremum norm $\|\cdot\|_{\mathcal{H}}$ where we view measures as functionals on $\mathcal{H}$. Then, for $T : \mathcal{P} \to \mathbb{R}$, the following definition captures a form of differentiability sufficient for applying the delta method to show that T is asymptotically normal [59, Chapter 3.9]. In the definition, we let $\mathcal{M}$ denote the space of signed measures on Ξ bounded with respect to $\|\cdot\|_{\mathcal{H}}$, noting that $\mathcal{M} \subset \mathcal{L}^\infty(\mathcal{H})$ via the mapping $\mu h = \int h(\xi) d\mu(\xi)$.

Definition 2. *The functional $T : \mathcal{P} \to \mathbb{R}$ is* Hadamard directionally differentiable at $P \in \mathcal{P}$ tangentially to $B \subset \mathcal{M}$ *if for all $H \in B$, there exists $dT_P(H) \in \mathbb{R}$ such that for all convergent sequences $t_n \to 0$ and $H_n \to H$ in $\mathcal{L}^\infty(\mathcal{H})$, that is, $\|H_n - H\|_{\mathcal{H}} \to 0$, such that $P + t_n H_n \in \mathcal{P}$,*

$$\frac{T(P + t_n H_n) - T(P)}{t_n} \to dT_P(H) \quad \text{as } n \to \infty.$$

Equivalently, T is Hadamard directionally differentiable at $P \in \mathcal{P}$ tangentially to $B \subset \mathcal{M}$ *if for every compact $K \subset B$,*

$$\lim_{t \to 0} \sup_{H \in K, P + tH \in \mathcal{P}} \left| \frac{T(P + tH) - T(P)}{t} - dT_P(H) \right| = 0. \tag{7}$$

Moreover, $T : \mathcal{P} \to \mathbb{R}$ is Hadamard differentiable at $P \in \mathcal{P}$ tangentially to $B \subset \mathcal{L}^\infty(\mathcal{H})$ *if $dT_P : B \to \mathbb{R}$ is linear and continuous on B.*

By restricting ourselves very slightly to a nicer class of Hadamard differentiable functionals, we may present a result on asymptotically pivotal confidence sets provided by f-divergences. To that end, we say that $T : \mathcal{P} \to \mathbb{R}$ has *canonical gradient* $T^{(1)} : \Xi \times \mathcal{P} \to \mathbb{R}$ if

$$dT_P(Q - P) = \int_\Xi T^{(1)}(\xi; P) d(Q - P)(\xi) \tag{8}$$

and $T^{(1)}$ has normalization $\mathbb{E}_P[T^{(1)}(\xi; P)] = 0$. In this case, $T^{(1)}$ is the usual *influence function* of T (cf. [28]). [1]

We now extend Proposition 1 to Hadamard differentiable functionals $T : \mathcal{P} \to \mathbb{R}$. Owen [46] shows a similar result for empirical likelihood (i.e. with $f(t) = -2 \log t + 2t - 2$) for the smaller class of Frechét differentiable functionals. Bertail et al. [8, 9] also claim a stronger result under certain uniform entropy conditions, but their proofs [8, pg. 308] show that confidence sets converge to one another in Hausdorff distance, which we do not believe is sufficient for our claim.[2] Recall that $\mathcal{M}$ is the (vector) space of signed measures in $\mathcal{L}^\infty(\mathcal{H})$.

[1] A sufficient condition for $T^{(1)}(\cdot; P)$ to exist is that T be Hadamard differentiable at P tangentially to any set B including the measures $\mathbb{1}_\xi - P$ for each $\xi \in P$: indeed, let $H_\xi := \mathbb{1}_\xi - P$, then the $\int H_\xi dP(\xi) = 0$, and the linearity of $dT_P : B \to \mathbb{R}$ guarantees that $\int dT_P(H_\xi) dP(\xi) = \int dT_P(\mathbb{1}_\xi - P) dP(\xi) = dT_P(P - P) = 0$, and we define $T^{(1)}(\xi; P) = dT_P(\mathbb{1}_\xi - P)$. Another sufficient condition is that the tangent set B in the definition of Hadamard differentiability be a set of linear functionals continuous with respect to the $\|\cdot\|_{L^2(P)}$-norm; the derivative dT_P is then isomorphic to an element $T^{(1)}(\cdot; P)$ of the Hilbert space $L^2(P)$ by the Riesz Representation Theorem (see, for example, [58, Chapter 25.5] or [34, Chapter 18]).

[2] The sets $A_n := \{v/n : v \in \mathbb{Z}^d\}$ and $B = \mathbb{R}^d$ satisfy $d_{\text{haus}}(A_n, B) = \frac{1}{2n}$, but for any random variable Z with Lebesgue density, we certainly have $\mathbb{P}(Z \in A_n) = 0$ while $\mathbb{P}(Z \in B) = 1$.

Theorem 3. *Let Assumption A hold and let $\mathcal{H}$ be a P_0-Donsker class of functions with an L^2-envelope M. Let $B \subset \mathcal{M}$ be such that G takes values in B where G is the limit $\sqrt{n}(\widehat{P}_n - P_0) \overset{d}{\rightsquigarrow} G$ given in Definition 1. Assume that $T : \mathcal{P} \subset \mathcal{M} \to \mathbb{R}$ is Hadamard differentiable at P_0 tangentially to B with canonical gradient $T^{(1)}(\cdot; P_0)$ as defined in (8) and that dT_P is defined and continuous on the whole of $\mathcal{M}$. If $0 < \mathrm{Var}(T^{(1)}(\xi; P_0)) < \infty$, then*

$$\lim_{n\to\infty} \mathbb{P}\left(T(P_0) \in \left\{T(P) : D_f\left(P \| P_n\right) \le \frac{\rho}{n}\right\}\right) = \mathbb{P}\left(\chi_1^2 \le \rho\right), \tag{9}$$

We use Lemma 1 to show the result in Appendix C.2.

If we let $B = B(\mathcal{H}, P_0) \subset \mathcal{L}^\infty(\mathcal{H})$ be the set of linear functionals on $\mathcal{H}$ that are $\|\cdot\|_{L^2(P_0)}$-uniformly continuous and bounded, then this is sufficient for the existence of the canonical derivative $T^{(1)}$. Note that $B(\mathcal{H}, P_0)$ is the smallest natural space containing the Gaussian process on $\mathcal{L}^\infty(\mathcal{H})$.

3 Statistical Inference for Stochastic Optimization

With our asymptotic expansion and convergence results in place, we now consider application of our results to stochastic optimimization problems, which is our main goal. In particular, we consider the mapping

$$T(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)].$$

Although the functional $T(P)$ is not linear, we show shortly that mild regularity conditions guarantee its Hadamard differentiability, allowing us to apply the asymptotic results of the previous section. Danskin [22] gave the first results on smoothness of the infimum functional (see also the book by Bonnans and Shapiro [13]), and we provide a general statement giving the first-order expansion of $T(P)$ as a function of the (infinite dimensional) distribution P.

3.1 Generalized Empirical Likelihood for Stochastic Optimization

In the stochastic optimization setting under consideration, we define the function class

$$\mathcal{H} := \{\ell(x;\cdot) : x \in \mathcal{X}\}, \tag{10}$$

implicitly assuming that $\xi \mapsto \ell(x;\xi)$ is $\mathcal{A}$-measurable for each x. We also assume that $\mathrm{Var}_{P_0}(\ell(x;\xi)) < \infty$ for all $x \in \mathcal{X}$ to allow application of our previous results, as well as making the following lower semi-continuity assumption.

Assumption B. *The function $\ell(\cdot;\xi)$ is lower semi-continuous for P_0-almost all $\xi \in \Xi$. Either $\mathcal{X}$ is compact, or $|\ell(x;\xi)| \to \infty$ as $\|x\| \to \infty$.*

Assumption B is the standard assumption guaranteeing existence of minimizers of $\mathbb{E}_P[\ell(x;\xi)]$ for any $P \ll P_0$ [e.g. 49, Theorem 1.9]. As in Section 2.1, we assume that central limit theorem is valid in $\mathcal{L}^\infty(\mathcal{H})$.

Assumption C. *$\{\ell(x;\cdot) : x \in \mathcal{X}\}$ is P_0-Donsker with $\sup_{x\in\mathcal{X}} |\ell(x;\xi) - \mathbb{E}_P\ell(x;\xi)| < \infty$ for all $\xi \in \Xi$ and $\mathbb{E}_{P_0}\left[\sup_{x\in\mathcal{X}} \ell(x;\xi)^2\right] < \infty$.*

Below is an example of such a function class.

Example 1 (Lipchitz Functions)**:** In addition to the latter two conditions in Assumption C, let $\mathcal{X}$ be compact, $|\ell(x_1;\xi) - \ell(x_2;\xi)| \le C(\xi)\|x_1 - x_2\|$ for P_0 almost surely for some $C(\xi)$ with $\mathbb{E}_{P_0}C(\xi)^2 < \infty$. Then, this class of functions is P_0-Donsker (*e.g.*, [59, Theorem 2.5.6]). ♣

From Danskin's theorem [22, 50], it is well known that the value function $T(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ is Hadamard differentiable. Combining this fact with Theorem 3, we obtain a generalized empirical likelihood theory for stochastic optimization problems.

Theorem 4. *Let Assumptions A, B, C hold, and let* $x^* := \operatorname{argmin}_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$ *be the unique minimizer. Then for* $T(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$*, we have that* (9) *holds.*

See Appendix D for the proof.

In the unconstrained setting, this result was noted by Lam and Zhou [38] for the EL divergence $f(t) = -\log t$ under more restrictive assumptions. Lam and Zhou [38] used the first order optimality conditions as an estimating equation for which the standard empirical likelihood theory [47] applied. On the other hand, Theorem 4 gives a much more general result that applies to constrained problems as well as general objective functions and divergences by using the smoothness of the functional $T(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$.

When P is finitely supported on n points, $T(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ is continuous with respect to the usual topology on $\mathbb{R}^n$ since it is Hadamard differentiable. Hence, the confidence region

$$\left\{ \inf_{x\in\mathcal{X}} \sum_{i=1}^n p_i \ell(x;\xi_i) : p^\top \mathbb{1} = 1,\ p \geq 0,\ \sum_{i=1}^n f(np_i) \leq \rho \right\}. \tag{11}$$

is a confidence interval as a continuous image of a connected set is connected. To compute this confidence interval, it suffices to compute the upper and lower endpoints. We discuss how to compute the endpoints efficiently in Section 3.2.

Normality theory for stochastic optimization problem was developed in [54]. The normal approximation analogue of Theorem 4 is

$$\sqrt{n}\left(\inf_{x\in\mathcal{X}} \mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)] - \inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \right) \overset{d}{\rightsquigarrow} N(0, \operatorname{Var}_{P_0} \ell(x^*;\xi)) \tag{12}$$

which holds under the conditions of Proposition 4 (the assumptions in [54] can be relaxed by applying Lemma 17). Note that the asymptotic distribution depends on an unknown parameter $\operatorname{Var}_{P_0} \ell(x^*;\xi)$ and is not asymptotically pivotal. This quantity needs to be estimated, usually by $\operatorname{Var}_{\widehat{P}_n} \ell(x_n^*;\xi)$ where x_n^* is the optimizer for the SAA (2). In this regard, the empirical divergence setup may be desirable from a statistical perspective.

Remark 1: Note that for both the normal approximation and the empirical likelihood, uniqueness of population optimum is required to obtain asymptotic normality. By extending the empirical divergence theory to processes, this assumption can be relaxed under sufficient regularity conditions on the f-divergence. In this case, for a given confidence level $1-\alpha$, ρ should be set such that

$$P\left(\sup_{x\in\mathcal{X}} \chi^2(x) \leq \rho \right) = 1 - \alpha$$

where $\chi^2(\cdot)$ is a Chi-squared process following $\chi^2(\cdot) \overset{d}{=} G(\cdot)^2$. Here, $G(\cdot)$ is a centered Gaussian random process on $\mathcal{X}$ with the covariance function $R(x_1, x_2) := \operatorname{Corr}(\ell(x_1;\xi), \ell(x_2;\xi))$. Hence, even without the assumption that the optimum is unique, we still retain the statistical interpretation that the robust problem (4a) gives an upper bound. The threshold value ρ can be set using an approximation formula developed in [1] or by Monte Carlo simulations [2]. $\diamondsuit$

Before ending the section, we mention that the inference framework extends to the case where $\mathcal{X}$ is also estimated with random data. See Appendix E for details.

3.2 Computing the Confidence Interval

To compute the confidence interval (11), we need to solve for the two endpoints u_n and l_n given by the expressions (4a)–(4b).

Interchanging the sup and the inf in (4a), we get an upper bound on u_n. This bound is tight if the objective function $\ell(\cdot;\xi)$ is convex. The result follows directly from [27, Theorem VII.4.3.1].

Lemma 2. *Let Assumption B hold, $\mathcal{X}$ be a convex set and $\ell(\cdot;\xi):\mathcal{X}\to\mathbb{R}$ be a convex function P_0-a.s.. Then,*

$$u_n = \inf_{x\in\mathcal{X}} \sup_{p\in\mathbb{R}^n} \left\{ \sum_{i=1}^n p_i \ell(x;\xi_i) : p^\top \mathbb{1} = 1,\ p \geq 0,\ \sum_{i=1}^n f(np_i) \leq \rho \right\}. \tag{13}$$

By a dual reformulation argument, we can write the minimax problem (13) as a minimization problem. Taking the dual of the inner problem above, we obtain the following reformulation as shown in [7]. The result is a direct consequence of strong duality.

Proposition 5 (Ben-Tal et al. [7])**.**

$$\sup_{P\ll\widehat{P}_n} \left\{ \mathbb{E}_P[\ell(x;\xi)] : D_f\left(P\|\widehat{P}_n\right) \leq \frac{\rho}{n} \right\} = \inf_{\lambda\geq 0,\eta\in\mathbb{R}} \mathbb{E}_{\widehat{P}_n}\left[\lambda f^*\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\right] + \frac{\rho}{n}\lambda + \eta. \tag{14}$$

The reformulation shows that when the objective is convex, the dual of the inner problem yields a convex optimization problem. To see this, observe that f^* is nondecreasing since its subgradient is nonnegative everywhere. As f^* is also convex, the result follows if the objective $\ell(\cdot;\xi)$ is convex ([14, Section 3.2.4]). In this subsection, we will henceforth assume that $\ell(\cdot;\xi)$ is convex P_0-a.s.. Hence, when conditions of Lemma 2 hold, we can compute u_n by solving the convex optimization problem

$$u_n = \inf_{x\in\mathcal{X},\lambda\geq 0,\eta\in\mathbb{R}} \mathbb{E}_{\widehat{P}_n}\left[\lambda f^*\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\right] + \frac{\rho}{n}\lambda + \eta. \tag{15}$$

Note that (4b) is in general not convex. However, we can exploit the special structure to construct efficient solution methods. Noting that the objective is linear in p, we can construct an alternating minimization algorithm as follows.

$$x^{k+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \sum_{i=1}^n p_i^k \ell(x^k;\xi_i)$$

$$p^{k+1} = \operatorname*{argmin}_{p\in\mathbb{R}^n} \left\{ \sum_{i=1}^n p_i \ell(x^{k+1};\xi_i) : p^\top \mathbb{1} = 1,\ p \geq 0,\ \sum_{i=1}^n f(np_i) \leq \rho \right\}$$

The above scheme will converge as the objective is monotonically decreasing over each iteration although convergence to global optima is not guaranteed.

4 Connections with Robust Optimization

From Theorem 4 and Lemma 2, the robust formulation (4a) is an upper confidence bound on the optimal value of the population problem (1). Hence, the robust finite sample solution is guaranteed to have loss at most u_n with probability at least $P(\chi_1^2 \leq \rho)$ asymptotically.

4.1 One-sided Confidence Interval

In this section, we discuss how we can set ρ to give a *calibrated* upper confidence bound and see how different choices of the f-divergence correspond to different levels of robustness. From Theorem 4, we have

$$\liminf_{n\to\infty} P\left(\inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \le u_n\right) \ge P\left(\chi_1^2 \le \rho\right).$$

But without the lower endpoint, the one-sided confidence interval $(-\infty, u_n]$ is no longer calibrated. That is, the inequality in the preceding display is not tight. For a robust optimizer only using the upper endpoint for her decision making, this will result in loss of statistical power—*i.e.*, the probability of Type II error is larger. The following theorem shows that asymptotic tail probabilities are symmetric and thus gives an easy way to obtain a calibrated one-sided confidence interval.

Theorem 6. *Let Assumptions A, B, C hold. Then, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \le u_n\right) = \mathbb{P}\left(\inf_{x\in S}\left\{\sqrt{\rho \mathrm{Var}_{P_0}\ell(x;\xi)} + N(0, \mathrm{Var}_{P_0}\ell(x;\xi))\right\} \ge 0\right)$$

where $S = \mathrm{argmin}_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$. *In particular, if* S *is unique, the preceding display is equal to* $1 - \frac{1}{2}P\left(\chi_1^2 \ge \rho\right)$.

We defer the proof to Appendix B.4. From Theorem 6, the one-sided confidence interval can be shortened by a simple correction to the threshold ρ. For a given confidence level $1-\alpha$, setting $\rho = \chi^2_{1,1-2\alpha}$ will give an one-sided confidence interval $(-\infty, u_n]$ has asymptotic coverage $1-\alpha$.

Remark 2: Here, we note an interesting phenomenon when the population optimum is not unique. In this case, the empirical divergence upper bound u_n given in (4a) always *undercovers* and the lower bound l_n given in (4b) always *overcovers* when we use Theorem 6 to calibrate our one-sided confidence intervals. To see this, note that

$$\begin{aligned}
&\mathbb{P}\left(\inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \le u_n\right) \to \mathbb{P}\left(\inf_{x\in S_P}\left\{\sqrt{\rho \mathrm{Var}_{P_0}\ell(x;\xi)} + N(0, \mathrm{Var}_{P_0}\ell(x;\xi))\right\} \ge 0\right) \\
&\le \mathbb{P}\left(\sqrt{\rho \mathrm{Var}_{P_0}}\ell(x^*;\xi) + N(0, \mathrm{Var}_{P_0}\ell(x^*;\xi)) \ge 0\right) = \mathbb{P}\left(N(0,1) \ge -\sqrt{\rho}\right) = 1 - \frac{1}{2}P\left(\chi_1^2 \ge \rho\right).
\end{aligned}$$

for any $x^* \in S_P$ and similarly,

$$\begin{aligned}
&\mathbb{P}\left(\inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \ge l_n\right) \to \mathbb{P}\left(\inf_{x\in S_P}\left\{-\sqrt{\rho \mathrm{Var}_{P_0}\ell(x;\xi)} + N(0, \mathrm{Var}_{P_0}\ell(x;\xi))\right\} \le 0\right) \\
&\ge \mathbb{P}\left(-\sqrt{\rho \mathrm{Var}_{P_0}}\ell(x^*;\xi) + N(0, \mathrm{Var}_{P_0}\ell(x^*;\xi)) \le 0\right) = \mathbb{P}\left(N(0,1) \ge -\sqrt{\rho}\right) = 1 - \frac{1}{2}P\left(\chi_1^2 \ge \rho\right).
\end{aligned}$$

$\diamondsuit$

4.2 Robust Formulation

As noted in [40], the SAA (2) has an intrinsic optimistic bias

$$\begin{aligned}
\inf_{x\in\mathcal{X}} \mathbb{E}[\ell(x;\xi)] &\ge \mathbb{E}\left[\inf_{x\in\mathcal{X}} \frac{1}{n+1}\sum_{i=1}^{n+1} \ell(x;\xi_i)\right] = \mathbb{E}\left[\inf_{x\in\mathcal{X}} \frac{1}{n+1}\sum_{i=1}^{n+1} \frac{1}{n}\sum_{j=1,j\neq i}^{n} \ell(x;\xi_j)\right] \\
&\ge \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{E}\left[\inf_{x\in\mathcal{X}} \frac{1}{n}\sum_{j=1,j\neq i}^{n+1} \ell(x;\xi_i)\right] = \mathbb{E}\left[\inf_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^{n} \ell(x;\xi_i)\right].
\end{aligned}$$

Thus, a more conservative approach may be appropriate. The robust formulation (4a) remedies for this inherent optimism by modeling the empirical problem in a conservative fashion around the data. In particular, it uses the confidence region formed by the f-divergence as an uncertainty set.

As is usual for robust formulations, we have from Proposition 5 that the distributionally robust problem (4a) has a convex dual reformulation. To see concretely what this reformulation gives us, consider the Cressie-Read family (16) suggested by [20]. Parameterized by k, the Cressie-Read family of divergences and their *Fenchel conjugates* $f^*(s) := \sup_t\{\langle s,t\rangle - f(t)\}$ are given by

$$f_k(t) = \frac{t^k - kt + k - 1}{2k(k-1)}, \quad f_k^*(s) = \frac{2}{k}\left[\left(\frac{k-1}{2}s + 1\right)_+^{k_*} - 1\right] \tag{16}$$

where $1/k + 1/k_* = 1$.

We implicitly take $f_k(t) = \infty$ for $t < 0$. The above expression is not defined when $k = 0, 1$ and is defined as the continuous limits as $k \to 0, 1$ respectively. The family (16) includes many commonly used statistics. If $k = 0$, we have the empirical likelihood $f_0(t) = -2\log t + 2t - 2$ and if $k = 1$, we have the KL maximum entropy $f_1(t) = 2t\log t - 2t + 2$. Note that f_k satisfies Assumption A for all k.

Then, the robust formulation (4a) for the Cressie-Read family becomes

$$\begin{aligned}&\inf_{x\in\mathcal{X},\lambda\geq 0,\eta} \mathbb{E}_{\widehat{P}_n}\left[\frac{2\lambda}{k}\left(\left[(k-1)\frac{\ell(x;\xi)-\eta}{2\lambda} + 1\right]_+^{k_*} - 1\right)\right] + \frac{\rho}{n}\lambda + \eta \\ &= \inf_{x\in\mathcal{X},\eta}\left(1 + k(k-1)\frac{\rho}{2n}\right)^{1/k}\left(\mathbb{E}_{\widehat{P}_n}\left(\ell(x;\xi)-\eta\right)_+^{k_*}\right)^{1/k_*} + \frac{\eta}{2}\end{aligned} \tag{17}$$

where $1/k + 1/k_* = 1$. Intuitively, we are penalizing large upward deviations of the objective function $\ell(x;\xi)$ from the mean in a certain way. Note that this penalization factor grows as k approaches 1 as we are imposing a more stringent notion of robustness. Thus, the modeler can choose an appropriate f-divergence depending on the level of robustness she desires.

A desirable property of the robust formulation (4a) is that the inner problem is a coherent risk measure of $\ell(x;\xi)$ (see [4] or [56, Ch 6.3] for the definition). Here, the risk preference is embodied in the f-divergence and the degree of risk aversion depends on the sample size as well as the confidence threshold ρ. In particular, the risk measure (17) induced by the Cressie-Read family of divergences is the higher order generalization of Conditional Value-at-Risk proposed by [35].

4.3 Robust Optimization as Variance Regularization

We now turn our attention to characterizing the asymptotic behavior of the robust formulation (4a). From inequality (23) and the fact that $C_{\rho,n}$ has radius $O(n^{-\frac{1}{2}})$, the finite sample problem (4a) should converge at the canonical $O_p(n^{-\frac{1}{2}})$ rate. We show an asymptotic expansion that will serve to further our understanding of what robust optimization really does. In particular, we see how robustifying regularizes the problem.

From Lemma 1, if Assumption A holds and $\sup_{x\in\mathcal{X}} \mathrm{Var}\ \ell(x;\xi) < \infty$, then

$$\sup_{P: D_f(P\|P_n)\leq \frac{\rho}{n}} \mathbb{E}_P[\ell(x;\xi)] = \mathbb{E}_{P_n}[\ell(x;\xi)] + \sqrt{\frac{\rho}{n}\mathrm{Var}_{P_n}(\ell(x;\xi))} + \varepsilon_n(x) \tag{18}$$

where $\mathbb{P}\left(\sqrt{n}|\varepsilon_n(x)| \geq c\right) \to 0$ for all $c > 0$. Further, from Theorem 2 we have that the expansion is uniform in $x \in \mathcal{X}$ under Assumption C *i.e.*, $\mathbb{P}\left(\sqrt{n}\sup_{x\in\mathcal{X}}|\varepsilon_n(x)| \geq c\right) \to 0$ for all $c > 0$.

Hence, we have that

$$\sqrt{n}\left(\sup_{P:D_f\left(P\|\widehat{P}_n\right)\leq\frac{\rho}{n}}\mathbb{E}_P[\ell(\cdot;\xi)]-\mathbb{E}_{P_0}[\ell(\cdot;\xi)]\right)\overset{d}{\rightsquigarrow}H\ \text{ in }\ \mathcal{L}^\infty(\mathcal{H})$$

where $H(\cdot)=\sqrt{\rho\mathrm{Var}_{P_0}\ell(\cdot;\xi)}+G$ and G is a mean zero Gaussian process with covariance $\mathbb{E}_{P_0}G(x_1)G(x_2)=\mathrm{Cov}_{P_0}(\ell(x_1;\xi),\ell(x_2;\xi))$. Applying the delta method as in the proof of Theorem 4, we obtain

$$\sqrt{n}\left(u_n-\inf_{x\in\mathcal{X}}\mathbb{E}_{P_0}[\ell(x;\xi)]\right)\overset{d}{\rightsquigarrow}\inf_{x\in S}H(x)\tag{19}$$

where $S=\mathrm{argmin}_{x\in\mathcal{X}}\mathbb{E}_{P_0}[\ell(x;\xi)]$. That is, we have obtained the asymptotic distribution of the robust formulation (4a).

Expansion (18) can be interpreted in many interesting ways. First, we see that in addition to remedying for the optimism bias, the robust formulation also takes into account the variance of the objective function. It is therefore expected that robustifying a SAA will reduce variance. While the optimism bias is of order $O(n^{-1})$, the robust formulation induces a conservatism bias of order $O(n^{-\frac{1}{2}})$. Although this is seemingly insensible, the bias term can be easily corrected for when the goal is to estimate the objective, thereby obtaining a error rate of $O(n^{-1})$.

Secondly, the expansion elucidates how robust optimization regularizes the problem. The connection between robust optimization and regularization has been a folk theorem where for support vector machines and lasso regression, [61, 62] noted a formal equivalence between robustifying and regularization for certain uncertainty sets. From expansion (18), robustifying can also be taken to mean explicitly setting aside a certain amount of safety stock and minimizing the overall cost. While enjoying these intuitive interpretations of what robustness entails, the formulation incorporates the conservatism in a way that is also economically sensible, namely, in a coherent fashion.

Lastly, note that the leading terms in expansion (18) is simply the upper endpoint of the confidence interval obtained from the normal approximation (12) at the threshold $\sqrt{\rho}$. However, the robust formulation

$$\inf_{x\in\mathcal{X}}\mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)]+\sqrt{\frac{\rho}{n}\mathrm{Var}_{\widehat{P}_n}\ell(x;\xi)}\tag{20}$$

is neither convex (in general) nor coherent. The robust formulation (4a) obtained from the generalized empirical likelihood framework mimics the normal approximation to the first order but incorporates higher order information in a way that attains tractability (convexity) and coherence. But in some special cases, the normal approximation upper confidence bound (20) is convex in x—in particular, when $\ell(x;\xi)$ is linear in x.

Example 2: Portfolio Optimization

Consider the portfolio optimization problem with weight constraint

$$\begin{array}{ll}\max_{x\in\mathbb{R}^d} & \mathbb{E}_{P_0}\left[\xi^\top x\right]\\ \text{s.t.} & x^\top\mathbb{1}=1\\ & x\in[l,u]\end{array}$$

where $\xi\in\mathbb{R}^d$ are the stock returns. Let $\mu_n:=\mathbb{E}_{\widehat{P}_n}\xi\in\mathbb{R}^d$ and $\Sigma_n:=\mathrm{Cov}_{\widehat{P}_n}\xi\in\mathbb{R}^{d\times d}$ be the sample mean and covariance of ξ respectively. The robust formulation corresponding to (4a) becomes

$$\max_{x\in\mathbb{R}^d}\min_{p\in R^n}\left\{\sum_{i=1}^n p_i\left(\xi_i^\top x\right):x^\top\mathbb{1}_d=1,\ x\in[l,u],\ p^\top\mathbb{1}_n=1,\ p\geq 0,\ \sum_{i=1}^n f(np_i)\leq\rho\right\}\tag{21}$$

$$=\max_{x^\top\mathbb{1}=1,\ x\in[l,u]}\ \mu_n x-\sqrt{\rho}\sqrt{x^\top\Sigma_n x}+o_p(n^{-\frac{1}{2}})\tag{22}$$

where the second line follows from the expansion (18). Note that (22) is the Lagrangian version of the Markowitz problem [41]. Since the Markowitz problem penalizes upward deviations as well as the downside, the risk function (negative utility) used here is not coherent. On the other hand, the robust formulation (21) matches the Markowitz problem to the first order and provides a optimization problem minimizing a coherent risk measure. ♣

5 Consistency

In this section, we give weak conditions under which the robust solutions are consistent. This verifies that the robust formulation (4a) is competitive compared to the SAA. In fact, we will show that robust solutions are consistent under (essentially) the same conditions required for that of the SAA.

First, we show uniform convergence of the inner problem of (4a) to the population expectation in Section 5.1. Uniform convergence is a desirable property in that in addition to guaranteeing consistency, it gives asymptotic properties of robust solutions as we saw in Section 4.3. We show that uniform convergence holds under essentially identical conditions required for the SAA to converge uniformly. We give general bracketing/entropy conditions. In particular, we see that uniform convergence holds when the objective function is continuous in $x \in \mathcal{X}$ and $\mathcal{X}$ is compact.

Optimization problems in the decision science domain often have unbounded feasible regions. While uniform convergence gives strong guarantees, it can be more stringent than what's necessary for consistency when the objective function is convex. Using the concept of epi-convergence, we relax the de facto compactness of the feasible region $\mathcal{X}$ required by uniformity to unbounded sets in Section 5.2. This is due to the fact that for lower semi-continuous convex functions, pointwise convergence is equivalent uniform convergence on compacta [33] so that we can restrict attention to a compact region around the population optimum.

Consistency of robust solutions of this type was first shown in [11] for a uniformly consistent goodness-of-fit test under equicontinuity of the objective function. Among others, they confirmed that tests induced by the KL divergence and χ^2 (Respectively $k = 1, 2$ in the Cressie-Read family (16)) are uniformly consistent. The equicontinuity of the random objective is a strong condition, often requiring uniform boundedness of the objective and compactness of the feasible region $\mathcal{X}$. In what follows, we relax the equicontinuity condition on the objective and show consistency for general f-divergences.

5.1 Uniform Convergence

In this section, we will show the uniform convergence of the robust objective $\sup_{P \ll \widehat{P}_n} \mathbb{E}_P[\ell(x;\xi)]$. We first state some well-known results in the empirical process literature. Let $\mathcal{H}$ be as in (10) and let the correspoding uniform topology as defined in Section 2.3

Definition 3. *A collection of functions $\mathcal{H}$ is called Glivenko-Cantelli if*

$$\sup_{f \in \mathcal{H}} \left| \mathbb{E}_{\widehat{P}_n} f - \mathbb{E}_{P_0} f \right| \overset{P^*}{\to} 0.$$

The definition says that the weak law of large numbers holds uniformly over $x \in \mathcal{X}$. Often, this can be shown using appropriate covering argument on $\mathcal{H}$.

Definition 4. *Given two functions l, u, we define the ϵ-bracket $[l, u]$ as the set of functions such that $l \leq f \leq u$ and $\|l - u\| < \epsilon$. The bracketing number $N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|)$ for a function class $\mathcal{H}$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$.*

A simple sufficient condition for the uniform law of large numbers can be given in terms of bracketing numbers.

Lemma 3. *van der Vaart and Wellner [59, Theorem 2.4.1]*
Let $\mathcal{H}$ be a class of measurable functions such that $N_{[\ ]}(\epsilon, \mathcal{H}, L_1(P_0)) < \infty$ for all $\epsilon > 0$. Then, $\mathcal{H}$ is Glivenko-Cantelli.

The following is a well-known example of a Glivenko-Cantelli class ([58, Example 19.8]).
Example 3 (Pointwise Compact Class)**:** When $\mathcal{X}$ is compact and $\ell(\cdot;\xi)$ is continuous in x for P_0-a.s. $\xi \in \Xi$, $\mathcal{F}$ is Glivenko-Cantelli if there exists a measurable Z such that $|\ell(x;\xi)| \leq Z$ P_0-a.s. and $\mathbb{E}_{P_0}[Z] < \infty$. From Lemma 3, it suffices to show that the bracketing number of the set is finite. Let $h(x,\delta;\xi)$ be the maximum gap

$$h(x,\delta;\xi) := \sup_{x':\|x-x'\|<\delta} \left|\ell(x;\xi) - \ell(x';\xi)\right|.$$

By continuity of $\ell(\cdot;\xi)$, we have $h(x,\delta;\xi) \to 0$ as $\delta \to 0$ P_0-a.s. for all $x \in \mathcal{X}$. Then, $\lim_{\delta\to 0} \mathbb{E}_{P_0}[h(x,\delta;\xi)] = \mathbb{E}_{P_0}[\lim_{\delta\to 0} h(x,\delta;\xi)] = 0$ by the Dominated Convergence Theorem. Fix $\epsilon > 0$ and choose $\delta(x)$ so that $\mathbb{E}[h(x,\delta;\xi)] < \epsilon$. Since $\cup_{x\in\mathcal{X}} B(x;\delta(x))$ is an open cover of $\mathcal{X}$, compactness implies there exists $\{x_1, ..., x_m\}$ such that $\cup_{j=1}^m B(x_j;\delta(x_j))$ is an open cover. Letting $u_j(x;\xi) = \ell(x_j;\xi) + h(x_j,\delta(x_j);\xi)$, $l_j(x;\xi) = \ell(x_j;\xi) - h(x_j,\delta(x_j);\xi)$, we see that $[l_j, u_j]$ is a 2ϵ-cover of $\mathcal{H}$. ♣

Letting $L := \frac{dP}{d\widehat{P}_n}$, we have

$$\begin{aligned}
&\sup_{P:D_f\left(P\|\widehat{P}_n\right)\leq\frac{\rho}{n}} \left|\mathbb{E}_P[\ell(x;\xi)] - \mathbb{E}_{P_0}[\ell(x;\xi)]\right| \\
&\leq \sup_{P:D_f\left(P\|\widehat{P}_n\right)\leq\frac{\rho}{n}} \mathbb{E}_{\widehat{P}_n}\left|L(\xi)-1\right|\ell(x;\xi) + \left|\mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)] - \mathbb{E}_{P_0}[\ell(x;\xi)]\right| \\
&\leq \sup_{P:D_f\left(P\|\widehat{P}_n\right)\leq\frac{\rho}{n}} \left(\mathbb{E}_{\widehat{P}_n}\left|L(\xi)-1\right|^p\right)^{1/p} \left(\mathbb{E}_{\widehat{P}_n}|\ell(x;\xi)|^q\right)^{1/q} + \left|\mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)] - \mathbb{E}_{P_0}[\ell(x;\xi)]\right| \qquad (23)
\end{aligned}$$

for any $p, q \geq 1$, $\frac{1}{p} + \frac{1}{q} = 1$ where the last inequality follows from Hölder, assuming appropriate moments exist.

The first term in (23) depicts the tradeoff between robustness (p) and the fatness of the tails of the objective function (q). Control on $\mathbb{E}_{\widehat{P}_n}[|L(\xi)-1|^p]^{1/p}$ depends on how fast the function f increases around 1. If f is steep, the ambiguity set for our worst-case measure will be small and it will be easier to get a handle on the likelihood ratio. We will use the bound (23) to obtain uniform convergence under the following moment condition.

Assumption D. $\sup_{x\in\mathcal{X}} \mathbb{E}_{P_0}|\ell(x;\xi)|^{1+\epsilon} < \infty$ *for some* $\epsilon > 0$.

Theorem 7. *If Assumptions A, D hold and $\{\ell(x;\cdot) : x \in \mathcal{X}\}$ is Glivenko-Cantelli, then*

$$\sup_{x\in\mathcal{X}} \sup_{P\ll\widehat{P}_n} \left\{\left|\mathbb{E}_P[\ell(x;\xi)] - \mathbb{E}_{P_0}[\ell(x;\xi)]\right| : D_f\left(P\|\widehat{P}_n\right) \leq \frac{\rho}{n}\right\} \overset{P^*}{\to} 0$$

Proof Let $q = \min(2, 1+\epsilon)$ and $p = \max(2, 1+\frac{1}{\epsilon})$. From Lemma 13, we have that $\|L(\xi)-1\|_{p,\widehat{P}_n} \leq n^{-1/p}\sqrt{\frac{\rho}{\gamma}}$ where $\gamma > 0$ only depends on f. Combining this with Assumption D, we have that the first term in the upper bound (23) goes to 0. Since the second term goes to 0 in outer probability

from the Glivenko-Cantelli property, desired result follows. □

Note that the rate of convergence of the robust objective depends on $1+\epsilon$, the order of finite moments. In the previous sections, we have taken $\epsilon = 2$ to obtain the canonical rate of convergence $n^{-\frac{1}{2}}$.

Define the sets of optima as below

$$S := \operatorname{argmin} \mathbb{E}_{P_0}[\ell(x;\xi)], \quad S_n := \operatorname*{argmin}_{x\in\mathcal{X}} \sup_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(x;\xi)] : D_f\left(P \| \widehat{P}_n\right) \leq \frac{\rho}{n} \right\}.$$

When uniform convergence holds, we can show consistency of robust solutions.

Theorem 8. *If Assumptions A, B, D hold and $\{\ell(x;\cdot) : x \in C\}$ is Glivenko-Cantelli for some compact set $C \subset \mathcal{X}$ containing S in its interior, then the robust formulation* (4a) *is consistent. That is,*

$$\left| \inf_{x\in\mathcal{X}} \sup_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(x;\xi)] : D_f\left(P \| \widehat{P}_n\right) \leq \frac{\rho}{n} \right\} - \inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \right| \to 0, \tag{24}$$

and the set of optima converges in Hausdorff distance in outer probability, $d_{\text{haus}}(S_n, S) \xrightarrow{P^*} 0$.

See Appendix F for the proof. If the objective function is convex, an identical conclusion can be shown without the Glivenko-Cantelli condition as we see in the next section.

5.2 Epi-convergence

Until this point, we have considered classes of objective functions that yield uniform convergence of the inner problem of (4a). We saw that the second moment has to be converge uniformly in order for this to happen. Checking this condition usually relies on a covering argument which depends heavily on the size of $\mathcal{X}$. If the goal is to just establish consistency, lower semi-continuity is sufficient for convex objectives. The notion of epi-convergence readily provides the tools needed for this purpose. In this section, we leverage the results of [33] along with concepts well outlined in [49]. This enables dealing with unbounded feasibility regions. The result is certainly true for the SAA (*e.g.*, [56, Chapter 5.1.1]) and holds for robust formulations too as we see next.

Theorem 9. *Let Assumptions A, B, D hold. If $\mathcal{X}$ is convex and $\ell(\cdot;\xi)$ is convex P_0-a.s., then* (24) *and $d_{\text{haus}}(S_n, S) \xrightarrow{P^*} 0$ holds.*

The above theorem establishes consistency for potentially unbounded feasible regions $\mathcal{X}$ when the objective function is convex. Modulo measurability issues, the proof hinges on a localization to a compact set and exploiting epi-convergence theory. See Appendix F for the proof.

6 Simulations

We present three simulation experiments in this section: portfolio optimization, conditional value-at-risk and the multi-item newsvendor model. For these different objective functions $\ell(x;\xi)$, we will compute the actual coverage probability of the generalized empirical likelihood confidence region $[l_n, u_n]$ given in (4) and compare it to the nominal level $\mathbb{P}\left(\chi^2_1 \leq \rho\right)$ given by our asymptotic theory in Section 3. As a benchmark, we also compute the coverage levels of the normal approximation (12). We take independent draws of n samples and compute $[l_n, u_n]$ using the methods outlined

in Section 3.2. The convex optimization problem (14) is solved to compute u_n and alternating minimization is used for l_n. We note that all three examples satisfy conditions of Lemma 2The package convex.jl [57] was used for the respective convex optimization steps. We report our numerical results in Figure 1, 2.

Figure 1: Coverage Rates (nominal = 95%)

%	Portfolio		Newsvendor		CVaR Normal		CVaR Tail=3		CVaR Tail=5	
sample size	EL	Normal	EL	Normal	EL	Normal	EL	Normal	EL	Normal
20	75.16	89.2	30.1	91.38	91.78	95.02	29	100	35.4	100
40	86.96	93.02	55.24	90.32	93.3	94.62	48.4	100	59.73	100
60	89.4	93.58	69.5	88.26	93.8	94.56	42.67	100	51.13	100
80	90.46	93.38	74.44	86.74	93.48	93.94	47.73	100	57.73	100
100	91	93.8	77.74	85.64	94.22	94.38	46.33	100	55.67	99.87
200	92.96	93.68	86.73	95.27	94.64	95.26	48.4	99.8	56.73	98.93
400	94.28	94.52	91	94.49	94.92	95.06	48.67	98.93	55.27	97.93
600	94.48	94.7	92.73	94.29	94.8	94.78	51.13	98.53	56.73	97.67
800	94.36	94.36	93.02	93.73	94.64	94.64	51.67	97.93	57.47	97.6
1000	95.25	95.15	92.84	94.31	94.62	94.7	53.07	98.47	58.6	97.33
2000	95.48	95.25	93.73	95.25	94.92	95.04	54.07	96.8	59.07	96.53
4000	96.36	95.81	95.1	95.78	95.3	95.3	58.6	96	62.07	96.6
6000	96.33	95.87	94.61	95	94.43	94.51	61.8	95.8	66.07	95.73
8000	96.46	95.9	94.56	94.71	94.85	94.85	64.67	95.67	69	95.33
10000	96.43	95.51	94.71	94.85	94.43	94.43	66.87	94.73	69.4	96.13
20000							74.27	95.8	76.8	96.13
40000							81.8	94.2	84.87	94.87
60000							86.87	93.93	89.47	94.47
80000							91.4	93.67	92.33	95
100000							94.2	94.33	95.07	95.2

First, consider the portfolio optimization problem given in Example 2. In particular, we will take $\mathcal{X} = [l, u] = [-10, 10]^{20}$ so that we have leverage constraints and $\ell(x; \xi) = \xi^\top x$. Here, $\xi \overset{\text{iid}}{\sim} N(\mu, \Sigma)$ is a 20-dimensional normal vector. Note that this example satisfies conditions of Theorem 4 (we confirm that the population optimum is unique). For the linear objective function, we plot the the objective of the Markowitz formulation as well. This is different from the normal upper confidence bound given by (12) in that the infimum over x is taken *after* taking the upper confidence bound. See Figure 2a for a plot of the intervals. It is worth noting that generalized empirical likelihood undercovers in small sample settings, which is consistent with previous empirical observations for other statistics (*e.g.*, [47, Sec 2.8]). Coverage levels for the 95% nominal level is given in Figure 1.

Example 4 (Conditional Value-at-Risk)**:** The conditional value-at-risk at α is the average above the $1 - \alpha$ quantile, $q_{1-\alpha}$,

$$\mathbb{E}_{P_0}[\xi | \xi \geq q_{1-\alpha}] = \min_x \left\{ \frac{1}{1-\alpha} \mathbb{E}_{P_0} (\xi - x)_+ + x \right\}$$

where the second expression was given by [48]. Conditional Value-at-Risk is a a quantity of interest in many financial applications. ♣

For our second simulation experiment, we use three different distributions; a mixture of normal

distribution and a mixture of heavy-tailed distributions with $\mathbb{P}(|\xi| \geq t) \propto t^{-\beta}$ where $\beta = 3, 5$. We draw ξ from an equal weight mixture distribution with parameters $\mu = [-6, -4, -2, 0, 2, 4, 6]$, $\sigma^2 = [2, 4, 6, 8, 10, 12, 14]$ respectively. Here, μ is interpreted as negative returns. Note that σ^2 increase with μ, reminiscent of volatility in bear markets, a well-observed tendency called the leverage effect [12, 17]. Since the cumulative distribution function of mixture of normals is strictly increasing, the optimal solution is unique. Although the feasible region $\mathcal{X} = \mathbb{R}$ is not compact, we proceed to compute the generalized empirical likelihood interval (4) and compare coverage rates with the nominal level 95%. Note that the actual coverage converges to the nominal level much more slowly in the heavy-tailed case. Further, even for the normal distribution we observe in Figure 2b a starker contrast with the normal approximation in small samples settings in that generalized empirical likelihood undercovers significantly.

Example 5 (Multi-item Newsvendor)**:** Let ξ_j be the uncertain demand for item $j = 1, \ldots, d$. Unsatisfied demand has backorder cost b_j per unit and extraneous orders have inventory cost h_j per unit. Then, the multi-item newsvendor loss is $\ell(x; \xi) = b^\top (x - \xi)_+ + h^\top (\xi - x)_+$ where $(\cdot)_+$ is the elementwise max operation with 0. ♣

For the last experiment, we take $\mathcal{X} = \left\{x \in \mathbb{R}^{20} : \|x\|_1 \leq 10\right\}$ and $\xi \overset{\text{iid}}{\sim} N(0, \Sigma)$ (we allow negative demand). Undercoverage is less pronounced for this example but nontheless still extant. See Figure 2c.

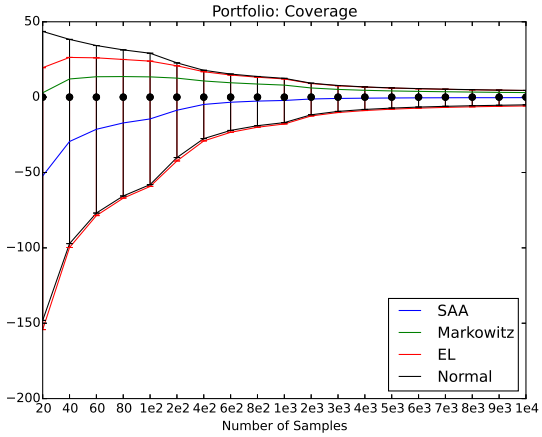
7 Conclusion

In this paper, we gave a generalized empirical likelihood theory for stochastic optimization problems. By noting that the resulting upper confidence bound is a robust formulation with desirable properties, we showed that in addition to being computationally tractable (convex) and coherent, the robust solutions are consistent under general conditions. Finally, we showed robustification can be interpreted as a form of regularization in that it also takes into account the variance of the estimator as well as the objective.

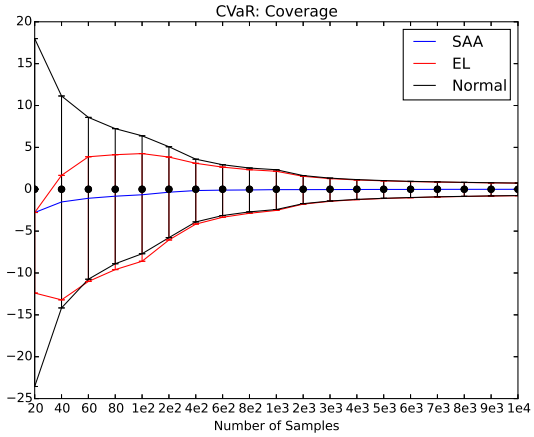
We hope that this paper takes a step towards understanding robust optimization from a statistical perspective. There are many interesting topics for further research, and we list a few of them below. On the technical side, the identifiability condition imposed in Theorem 4 is quite stringent. When the optima is not unique, there is an inherent problem in that the relevant asymptotics are no longer normal. As the normal approximation also fails in this case, it would be desirable to have a inference procedure that gives (at least) a conservative confidence interval even in the non-indentifiable case.

In large sample settings, interior point algorithms are no longer a viable option as function evaluations take operations linear in the sample size. Although there is a vast literature on online methods for the SAA (*e.g.*, [43, 25]), the literature for minimax problems of the form (4a) is still nascent [18]. Efficient solution methods need to be developed to scale up robust optimization.

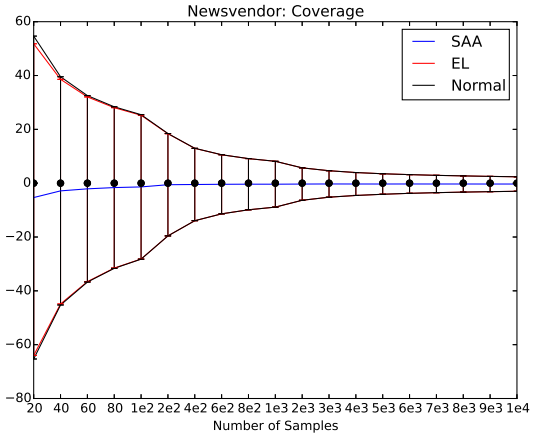
There are two ways of injecting robustness in the formulation (4a): increasing ρ and choosing a f that is more gradual near 1. This paper characterized a statistically principled way of choosing ρ to obtain calibrated confidence bounds and showed that smooth f-divergences have same first order asymptotics. Higher order characteristics of different f-divergences through the lens of robustness would be interesting. While the literature on higher order corrections for generalized empirical likelihood offer some answers for inference problems regarding mean of a distribution [24, 5, 19, 15, 16], many questions remain to be answered in the stochastic optimization setting.



(a) Portfolio Optimization



(b) Conditional Value-at-Risk



(c) Multi-item Newsvendor

Figure 2: Confidence Intervals

References

[1] R. J. Adler and J. E. Taylor. *Random fields and geometry*, volume 115. Springer, 2009.

[2] R. J. Adler, J. H. Blanchet, J. Liu, et al. Efficient monte carlo for high excursions of gaussian random fields. *The Annals of Applied Probability*, 22(3):1167–1214, 2012.

[3] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

[4] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[5] K. A. Baggerly. Empirical likelihood as a goodness-of-fit measure. *Biometrika*, 85(3):535–547, 1998.

[6] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

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

[8] P. Bertail. Empirical likelihood in some semiparametric models. *Bernoulli*, 12(2):299–331, 2006.

[9] P. Bertail, E. Gautherat, and H. Harari-Kermadec. Empirical φ^*-divergence minimizers for hadamard differentiable functionals. In *Topics in Nonparametric Statistics*, pages 21–32. Springer, 2014.

[10] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *arXiv:1401.0212 [math.OC]*, 2013. URL `http://arxiv.org/abs/1401.0212`.

[11] D. Bertsimas, V. Gupta, and N. Kallus. Robust SAA. *arXiv:1408.4445 [math.OC]*, 2014. URL `http://arxiv.org/abs/1408.4445`.

[12] F. Black. Stuedies of stock price volatility changes. *In Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economics Statistics Section, pp. 177-181*, 1976.

[13] J. F. Bonnans and A. Shapiro. Optimization problems with perturbations: A guided tour. *SIAM review*, 40(2):228–264, 1998.

[14] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[15] F. Bravo. Second-order power comparisons for a class of nonparametric likelihood-based tests. *Biometrika*, 90(4):881–890, 2003.

[16] F. Bravo. Bartlett-type adjustments for empirical discrepancy test statistics. *Journal of statistical planning and inference*, 136(3):537–554, 2006.

[17] A. A. Christie. The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of financial Economics*, 10(4):407–432, 1982.

[18] K. Clarkson, E. Hazan, and D. Woodruff. Sublinear optimization for machine learning. *Journal of the Association for Computing Machinery*, 59(5), 2012.

[19] S. A. Corcoran. Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, pages 967–972, 1998.

[20] N. Cressie and T. R. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 440–464, 1984.

[21] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientifica Mathematica Hungary*, 2:299–318, 1967.

[22] J. M. Danskin. *The theory of max-min and its application to weapons allocation problems*, volume 5. Springer Science & Business Media, 1967.

[23] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[24] T. DiCiccio, P. Hall, and J. Romano. Empirical likelihood is bartlett-correctable. *The Annals of Statistics*, pages 1053–1061, 1991.

[25] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[26] J. Dupacová and R. Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Annals of Statistics*, pages 1517–1549, 1988.

[27] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, New York, 1993.

[28] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley and Sons, second edition, 2009.

[29] G. Imbens. Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20(4):493–506, 2002.

[30] R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Optimization Online*, 2013. URL `http://www.optimization-online.org/DB_FILE/2013/09/4044.pdf`.

[31] A. J. King. Generalized delta theorems for multivalued mappings and measurable selections. *Mathematics of Operations Research*, 14(4):720–736, 1989.

[32] A. J. King and R. T. Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18(1):148–162, 1993.

[33] A. J. King and R. J. Wets. Epi-consistency of convex stochastic programs. *Stochastics and Stochastic Reports*, 34(1-2):83–92, 1991.

[34] M. R. Kosorok. Introduction to empirical processes. *Introduction to Empirical Processes and Semiparametric Inference*, pages 77–79, 2008.

[35] P. A. Krokhmal. Higher moment coherent risk measures. *Quantitative Finance*, 7(4):373–387, 2007.

[36] J. Kyparisis. On uniqueness of kuhn-tucker multipliers in nonlinear programming. *Mathematical Programming*, 32(2):242–246, 1985.

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

[38] H. Lam and E. Zhou. Quantifying input uncertainty in stochastic optimization. In *Proceedings of the 2015 Winter Simulation Conference*. IEEE, 2015.

[39] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses, Third Edition*. Springer, 2005.

[40] W.-K. Mak, D. P. Morton, and R. K. Wood. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24(1):47–56, 1999.

[41] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[42] S. Mendelson. Learning without concentration. In *Proceedings of the Twenty Seventh Annual Conference on Computational Learning Theory*, 2014.

[43] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

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

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

[47] A. B. Owen. *Empirical likelihood*. CRC press, 2001.

[48] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[49] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, New York, 1998.

[50] W. Römisch. Delta method, infinite dimensional. *Encyclopedia of Statistical Sciences*, 2005.

[51] M. Scarsini. Multivariate convex orderings, dependence, and stochastic equality. *Journal of Applied Probability*, 35:93–103, 1999.

[52] A. Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *The Annals of Statistics*, pages 841–858, 1989.

[53] A. Shapiro. On differential stability in stochastic programming. *Mathematical Programming*, 47(1-3):107–116, 1990.

[54] A. Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30 (1):169–186, 1991.

[55] A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993.

[56] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Programming Society, 2009.

[57] M. Udell, K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd. Convex optimization in julia. In *High Performance Technical Computing in Dynamic Languages (HPTCDL), 2014 First Workshop for*, pages 18–28. IEEE, 2014.

[58] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

[59] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

[60] Z. Wang, P. Glynn, and Y. Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, pages 1–21, 2015.

[61] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *The Journal of Machine Learning Research*, 10:1485–1510, 2009.

[62] H. Xu, C. Caramanis, and S. Mannor. Robust regression and lasso. In *Advances in Neural Information Processing Systems 21*, pages 1801–1808, 2009.

A Proof of Lemma 1

Assume without loss of generality that Z is mean-zero. If $\mathrm{Var}(Z) = 0$, then the lemma follows trivially as $s_n = 0$. In what follows, we will assume that $\mathrm{Var}(Z) > 0$. Define

$$Z_\star := \mathbb{E}[(Z)_+] \wedge \mathbb{E}[(-Z)_+],$$

which must be strictly positive, because $\mathrm{Var}(Z) > 0$.

Now consider the maximization problem

$$\begin{aligned} &\text{maximize } \mathbb{E}_P[Z] \\ &\text{subject to } D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}. \end{aligned} \tag{25}$$

Let $p_n \in \mathbb{R}^n, \lambda_n \ge 0, \eta_n \in \mathbb{R}$ denote the primal and dual variables for the problem (25); λ_n for the f-divergence constraint and η_n for the constraint $p_n^\top \mathbb{1} = 1$. We introduce no dual variable for the constraint that $p_n \ge 0$, as $f(t) = +\infty$ for all $t \le 0$. Writing down the Lagrangian, we have

$$\mathcal{L}(p_n, \lambda_n, \eta_n, \theta_n) = \sum_{i=1}^n p_{i,n} Z_i + \lambda_n \left(\rho - \sum_{i=1}^n f(n p_{i,n}) \right) + \eta_n (1 - \mathbb{1}^\top p_n).$$

We now state a lemma useful both for computing the dual to problem (25) and for calculations involving the dual. This result is a specialization of several standard results; see, for example, Hiriart-Urruty and Lemaréchal [27, Chapter I.6.2].

Lemma 4. *Let $f : (0, \infty) \to \mathbb{R}_+$ be a strictly convex C^2 convex function with $f''(1) = 2$ and $f(1) = f'(1) = 0$. Then $f'_+(0) < 0$, and*

(i) $\mathrm{Im}\, f^{*\prime} \subset [0, \infty)$ and $f^{*\prime}(0) = 1$, and

(ii) f^* is C^1 on $\mathbb{R}$, and f^* is C^2 on $(f'_+(0), \infty)$ with second derivative

$$f^{*\prime\prime}(s) = \frac{1}{f''(f^{*\prime}(s))},$$

(iii) and if f is three times differentiable in a neighborhood of 1, then for s in some neighborhood of 0 (whose size depends on f),

$$f^{*\prime\prime\prime}(s) = \frac{-1}{(f''(f^{*\prime}(s)))^2} \cdot f'''(f^{*\prime}(s)) \frac{1}{f''(f^{*\prime}(s))}.$$

Proof The first two results are standard [27, Chapter I.6.2]. The final result follows by the chain rule. Expanding formally (assuming all notated derivatives exist), we have that

$$f^{*\prime\prime\prime}(s) = \frac{-1}{(f''(f^{*\prime}(s)))^2} \cdot f'''(f^{*\prime}(s)) f^{*\prime\prime}(s).$$

Using the value of $f^{*\prime\prime}(s)$ from part (ii) and noting that $f^{*\prime}(s)$ is continuous and satisfies $f^{*\prime}(0) = 1$ by parts (i) and (ii), we obtain the result. □

Note that strong duality obtains for the problem (25) (by Slater's condition), so the KKT conditions coupled with Lemma 4 yield

$$p_{i,n} = \frac{1}{n} f^{*\prime}\left(\frac{Z_i - \eta_n}{n\lambda_n}\right)$$
$$0 = \lambda_n \left(\rho - \sum_{i=1}^n f(np_{i,n})\right), \quad p_n^\top \mathbb{1} = 1, \ \lambda_n \geq 0, \ \eta_n(p_n^\top \mathbb{1} - 1) = 0, \ \sum_{i=1}^n f(np_{i,n}) \leq \rho.$$

Let

$$\alpha_n := \frac{1}{n\lambda_n} \ \text{ and } \ \beta_n := \frac{\eta_n}{n\lambda_n} \tag{26}$$

so that the optimality conditions become equivalent to $p_{i,n} = \frac{1}{n} f^{*\prime}(\alpha_n Z_i - \beta_n)$.

First, we note the following fact, which follows from the Paley-Zygmund inequality.

Lemma 5. *Let $c_0 \in (0,1)$ and define*

$$q_\star := \mathbb{P}\left(\frac{Z}{\mathbb{E}[(Z)_+]} \geq c_0\right) \wedge \mathbb{P}\left(\frac{Z}{\mathbb{E}[(-Z)_+]} \leq -c_0\right). \tag{27}$$

Then

$$q_\star > (1 - c_0)^2 \frac{Z_\star^2}{\mathbb{E}[Z^2]} > 0.$$

See Section A.1 for a proof of this lemma. We will let $c_0 = \frac{1}{4}$ in the definition of $q_\star$ for the remainder of the proof of the theorem. We then have the following result, which shows that there is likely to be substantial variance of the Z vectors. (This technique is reminiscent of Mendelson's arguments on the variance of regression estimators [cf. 42, Theorem 5.3]).

Lemma 6. *Define the event*

$$\mathcal{E}_n := \left\{\text{card}\left\{i : \frac{Z_i}{\mathbb{E}[(Z)_+]} \geq c_0\right\} \geq \frac{q_\star}{4} n \text{ and } \text{card}\left\{i : \frac{Z_i}{\mathbb{E}[(-Z)_+]} \leq -c_0\right\} \geq \frac{q_\star}{4} n\right\}.$$

Then we have

$$\mathbb{P}\left(\mathcal{E}_n\right) \geq 1 - 2\exp\left(-nq_\star^2\right).$$

See Section A.2 for a proof of this lemma.

Now, we show that the quantities α_n and β_n are generally small—as long as $\mathcal{E}_n$ holds. Let us suppress dependence on x for simplicity. Indeed, let $\mathcal{E}_n$ hold. Then we find that

$$\rho \geq \sum_{i=1}^n f(np_{i,n}) = \sum_{i=1}^n f(f^{*\prime}(\alpha_n Z_i - \beta_n)).$$

There are four cases to consider. Let I^+ denote those indices $i \in [n]$ such that $Z_i \geq c_0 Z_\star$ and I^- those indices $i \in [n]$ such that $Z_i \leq -c_0 Z_\star$. By noting that $f^{*\prime}$ is non-decreasing and f is non-decreasing (resp. non-increasing) on $[1, \infty)$ (resp. $(0, 1]$), we have the following.

Case 1 $\alpha_n \geq 0$, $\beta_n \leq 0$. For $i \in I^+$, we have $\alpha_n Z_i - \beta_n \geq |\alpha_n| c_0 Z_\star + |\beta_n| \geq 0$, $f^{*\prime}(\alpha_n Z_i - \beta_n) \geq f^{*\prime}(|\alpha_n| c_0 Z_\star + |\beta_n|) \geq 1$ and $f(f^{*\prime}(\alpha_n Z_i - \beta_n)) \geq f(f^{*\prime}(|\alpha_n| c_0 Z_\star + |\beta_n|)) \geq 0$.

Case 2 $\alpha_n \geq 0$, $\beta_n \geq 0$. For $i \in I^-$, we have $\alpha_n Z_i - \beta_n \leq -|\alpha_n| c_0 Z_\star - |\beta_n| \leq 0$, $f^{*\prime}(\alpha_n Z_i - \beta_n) \leq f^{*\prime}(-|\alpha_n| c_0 Z_\star - |\beta_n|) \leq 1$ and $f(f^{*\prime}(\alpha_n Z_i - \beta_n)) \geq f(f^{*\prime}(-|\alpha_n| c_0 Z_\star - |\beta_n|)) \geq 0$.

Case 3 $\alpha_n \le 0$, $\beta_n \ge 0$. For $i \in I^+$, we have $\alpha_n Z_i - \beta_n \le -|\alpha_n| c_0 Z_\star - |\beta_n| \le 0$, $f^{*\prime}(\alpha_n Z_i - \beta_n) \le f^{*\prime}(-|\alpha_n| c_0 Z_\star - |\beta_n|) \le 1$ and $f(f^{*\prime}(\alpha_n Z_i - \beta_n)) \ge f(f^{*\prime}(-|\alpha_n| c_0 Z_\star - |\beta_n|)) \ge 0$.

Case 4 $\alpha_n \le 0$, $\beta_n \le 0$. For $i \in I^-$, we have $\alpha_n Z_i - \beta_n \ge |\alpha_n| c_0 Z_\star + |\beta_n| \ge 0$, $f^{*\prime}(\alpha_n Z_i - \beta_n) \ge f^{*\prime}(|\alpha_n| c_0 Z_\star + |\beta_n|) \ge 1$ and $f(f^{*\prime}(\alpha_n Z_i - \beta_n)) \ge f(f^{*\prime}(|\alpha_n| c_0 Z_\star + |\beta_n|)) \ge 0$.

Thus, in any of the preceding four cases, by noting that $\min\{|I^+|, |I^-|\} \ge \frac{q_\star n}{2}$, we have on the event $\mathcal{E}_n$,

$$\begin{aligned}\rho \ge \sum_{i=1}^n f(f^{*\prime}(\alpha_n Z_i - \beta_n)) &\ge \max\left\{\sum_{i \in I^+} f(f^{*\prime}(\alpha_n Z_i - \beta_n)), \sum_{i \in I^-} f(f^{*\prime}(\alpha_n Z_i - \beta_n))\right\} \\ &\ge \frac{n q_\star}{4} \max\left\{f(f^{*\prime}(|\alpha_n| c_0 Z_\star + |\beta_n|)), f(f^{*\prime}(-|\alpha_n| c_0 Z_\star - |\beta_n|))\right\} \quad (28)\end{aligned}$$

It follows that $f(f^{*\prime}(|\alpha_n| c_0 Z_\star + |\beta_n|)) \to 0$ which implies $f^{*\prime}(|\alpha_n| c_0 Z_\star + |\beta_n|) \to 1$ since f is strictly convex at 1 with $f(1) = f'(1) = 0$. Since $f^{*\prime}$ is strictly increasing at 0 and $f^{*\prime}(0) = 1$, we conclude $\alpha_n, \beta_n \to 0$ on $\mathcal{E}_n$.

Using Lemma 4, there exists a constant $c > 0$ (depending on f) such that $|f^{*\prime}(s) - 1 - \frac{s}{2}| \le \frac{s}{4}$ for $s \in [-c, c]$, and $f(t) \ge \frac{1}{2}(t-1)^2$ for $t \in [1-c, 1+c]$. Note that for $s \in [0, c]$, we then have

$$f(f^{*\prime}(s)) \ge \frac{1}{2}(f^{*\prime}(s) - 1)^2 \ge \frac{1}{32} s^2.$$

Now, define the event

$$\mathcal{E}_n^0 = \left\{|\alpha_n| \le \frac{1}{2} \frac{c}{c_0 Z_\star},\ |\beta_n| \le \frac{c}{2}\right\}.$$

From the definition, we have on $\mathcal{E}_n \cap \mathcal{E}_n^0$ $|f^{*\prime}(|\alpha_n| c_0 Z_\star + |\beta_n|) - 1| \le c$ and $|\alpha_n| c_0 Z_\star + |\beta_n| \le c$. Then, inequality (28) yields that

$$\rho \ge \frac{n q_\star}{8}(f^{*\prime}(|\alpha_n| c_0 Z_\star + |\beta_n|) - 1)^2 \ge \frac{n q_\star}{128}(|\alpha_n| c_0 Z_\star + |\beta_n|)^2.$$

It follows that on $\mathcal{E}_n \cap \mathcal{E}_n^0$,

$$\left\| \begin{matrix} \alpha_n Z_\star \\ \beta_n \end{matrix} \right\|_\infty \le \frac{8\sqrt{2}\rho}{\sqrt{n q_\star}} \max\{1, c_0\}. \quad (29)$$

We remark that $\alpha_n, \beta_n \to 0$ conditional on $\mathcal{E}_n$ and hence we have that $\mathcal{E}_n \cap \mathcal{E}_n^0$ holds almost surely as $n \to \infty$.

Now we turn to showing a few somewhat finer results relating α_n and β_n, providing asymptotic expansions of α_n and β_n in terms of one another, which allow us to evaluate the value of the solution of the supremum problem (25).

Before we begin, we define a few Taylor remainder values. For $k \in \{1, 2, \ldots\}$ and the functions f and f^*, let

$$R_{f,k}(t) := f(t) - \left[f(1) + \sum_{i=1}^k \frac{1}{i!} f^{(i)}(1) t^i\right] \quad \text{and} \quad R_{f^*,k}(s) := f^*(s) - \left[f^*(0) + \sum_{i=1}^k \frac{1}{i!} (f^*)^{(i)}(0) s^i\right],$$

assuming that these derivatives exist. Now for $g = f$ or $g = f^*$, define

$$\mathsf{r}_{g,k} := \limsup_{|\delta| \to 0} \frac{|R_{g,k}(\delta)|}{|\delta|^{k+1}}, \quad (30)$$

where we note that $\mathsf{r}_{f^*,k} < \infty$ and $\mathsf{r}_{f,k} < \infty$ for $k \in \{0,1,2\}$ by Lemma 4. Using the definition (30), define the constants

$$\begin{aligned}
\epsilon_{f^*} &:= \sup\Big\{s : \text{for all } |\epsilon| \le s,\ |f^{*\prime}(\epsilon) - f^{*\prime}(0) - f^{*\prime\prime}(0)\epsilon| \le 2\mathsf{r}_{f^*,2}\epsilon^2 \\
&\qquad\qquad \text{and } |f^{*\prime}(\epsilon) - f^{*\prime}(0)| \le 2\mathsf{r}_{f^*,1}\epsilon\Big\}, \\
\epsilon_f &:= \sup\Big\{t : \text{for all } |\epsilon| \le t,\ |f(1+\epsilon) - f(1) - f'(1)\epsilon - \frac{1}{2}f''(1)\epsilon^2| \le 2\mathsf{r}_{f,2}\epsilon^3 \\
&\qquad\qquad \text{and } |f(1+\epsilon) - f(1) - f'(1)\epsilon| \le 2\mathsf{r}_{f,1}\epsilon^2\Big\},
\end{aligned} \tag{31}$$

which are both finite and positive. These ϵ_f and ϵ_f^* give bounds on intervals in which the functions f and f^* have sufficiently nice Taylor approximations. Both are finite finite by the definitions (30). Using the $\epsilon > 0$ defined in the statement of the theorem, we also define the event

$$\mathcal{E}_{n,f} := \left\{\max_{i \le n} |\alpha_n Z_i - \beta_n| \le \epsilon_f \wedge \epsilon_{f^*} \wedge \epsilon\right\}, \tag{32}$$

making Taylor approximations of f and $f^{*\prime}$ around $\alpha_n Z_i - \beta_n$ valid. With these definitions, we have the following two lemmas.

Lemma 7. *Let the event $\mathcal{E}_{n,f}$ hold and define $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$. Then*

$$\beta_n = \alpha_n \overline{Z}_n + R_n, \quad \textit{where } \ |R_n| \le \frac{C_f \rho}{n}$$

for a constant C_f depending only on f.

See Section A.3 for a proof of Lemma 7.

Lemma 8. *Let the event $\mathcal{E}_{n,f}$ hold. Then*

$$\sum_{i=1}^n (\alpha_n Z_i - \beta_n)^2 \in 4\rho\left[1 - c_f(\epsilon \wedge \epsilon_f \wedge \epsilon_{f^*}), 1 + c_f(\epsilon \wedge \epsilon_f \wedge \epsilon_{f^*})\right],$$

where $c_f < \infty$ is a constant that depends only on f.

See Section A.4 for a proof of this lemma.

With Lemmas 7 and 8 in place, we can more rigourously give several asymptotic expansions. Let $\overline{Z_n^2} = \frac{1}{n}\sum_{i=1}^n Z_i^2$ for shorthand, and $\overline{Z_n} = \frac{1}{n}\sum_{i=1}^n Z_i$. Also, let $4\rho_n$ be the value of $\sum_{i=1}^n (\alpha_n Z_i - \beta_n)^2$, which by Lemma 8 may be random but on $\mathcal{E}_{n,f}$ we have $\rho_n \in \rho[1 \pm c\epsilon]$ for a constant c depending only on f. Combining the two lemmas, then, we see that

$$\begin{aligned}
4\rho_n \stackrel{(i)}{=} \sum_{i=1}^n (\alpha_n Z_i - \beta_n)^2 &= n\alpha_n^2 \overline{Z_n^2} - 2n\alpha_n\beta_n\overline{Z_n} + n\beta_n^2 \\
&\stackrel{(ii)}{=} n\alpha_n^2\overline{Z_n^2} - 2n\alpha_n^2\overline{Z_n}^2 - 2n\alpha_n\overline{Z_n}R_n + n\alpha_n^2\overline{Z_n}^2 + 2n\alpha_n R_n\overline{Z_n} + nR_n^2 \\
&= n\alpha_n^2\overline{Z_n^2} - n\alpha_n^2\overline{Z_n}^2 + nR_n^2,
\end{aligned}$$

where step (i) is a consequence of Lemma 8 and step (ii) substitutes $\beta_n = \alpha_n\overline{Z_n} + R_n$ as in Lemma 7. Dividing both sides by n and solving for α, we find

$$\alpha_n^2 = 4\left[\frac{\rho_n}{n} - R_n^2\right] \cdot \frac{1}{\overline{Z_n^2} - \overline{Z_n}^2} \in \frac{4\rho}{ns_n^2} + \left[-\frac{c_f\epsilon}{ns_n^2} - \frac{c_f\rho^2}{n^2 s_n^2}, \frac{c_f\epsilon}{ns_n^2} + \frac{c_f\rho^2}{n^2 s_n^2}\right], \tag{33}$$

where $s_n^2 = \frac{1}{n}\sum_{i=1}^n Z_i^2 - (\frac{1}{n}\sum_{i=1}^n Z_i)^2$ is the sample variance.

Returning to the supremum problem, we have on the event $\mathcal{E}_{n,f}$ that

$$f^{*\prime}(\alpha_n Z_i - \beta_n) = f^{*\prime}(0) + f^{*\prime\prime}(0)(\alpha_n Z_i - \beta_n) + R(\alpha_n Z_i - \beta_n) = 1 + \frac{1}{2}(\alpha_n Z_i - \beta_n) + R(\alpha_n Z_i - \beta_n)$$

where by assumption the remainder R satisfies $\limsup_{|\delta|\to 0} \delta^{-2}|R(\delta)| < \infty$. Thus, recalling the remainder R_n from Lemma 7, we have

$$\begin{aligned}
\sum_{i=1}^n p_{i,n} Z_i &= \sum_{i=1}^n \frac{1}{n} f^{*\prime}(\alpha_n Z_i - \beta_n) Z_i \\
&= \frac{1}{n}\sum_{i=1}^n Z_i + \frac{1}{2n}\alpha_n \sum_{i=1}^n Z_i^2 - \frac{1}{2n}\beta_n \sum_{i=1}^n Z_i + \frac{1}{n}\sum_{i=1}^n R(\alpha_n Z_i - \beta_n) Z_i \\
&= \overline{Z_n} + \frac{1}{2}\alpha_n s_n^2 - R_n \overline{Z_n} + \frac{1}{n}\sum_{i=1}^n R(\alpha_n Z_i - \beta_n) Z_i.
\end{aligned}$$

We expand $\sqrt{\alpha_n}$ using the value (33). Noting that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $\sqrt{a-b} \ge \sqrt{a} - \sqrt{b}$, we have

$$\alpha_n s_n^2 \in 2 s_n \sqrt{\frac{\rho}{n} \pm c_f \left(\frac{\epsilon}{n} + \frac{\rho^2}{n^2}\right)} \in 2 s_n \sqrt{\frac{\rho}{n}} \pm c_f^{\frac{1}{2}} s_n \sqrt{\frac{\epsilon}{n} + \frac{\rho^2}{n^2}}.$$

Thus on the event $\mathcal{E}_{n,f}$, we have

$$\begin{aligned}
\sup_{P : D_f(P\|P_n) \le \frac{\rho}{n}} \mathbb{E}_P[Z] &= \sum_{i=1}^n p_{i,n} Z_i \\
&\in \overline{Z_n} + \sqrt{\frac{\rho}{n} s_n^2} \\
&\quad - R_n \overline{Z_n} + \frac{1}{n}\sum_{i=1}^n R(\alpha_n Z_i - \beta_n) Z_i \pm c_f s_n \sqrt{\frac{\epsilon}{n} + \frac{\rho^2}{n^2}}.
\end{aligned} \tag{34}$$

On the event $\mathcal{E}_{n,f}$, we have $|R_n| \le c_f \rho / n$ for a constant c_f depending only on f. From the strong law of large numbers $|\overline{Z_n}| \overset{a.s.}{\to} 0$, we have

$$n \mathbf{1}\left\{\mathcal{E}_{n,f}\right\} R_n \overline{Z_n} \overset{a.s.}{\to} 0, \tag{35}$$

Additionally, on the event $\mathcal{E}_{n,f}$, we have

$$\sum_{i=1}^n |R(\alpha_n Z_i - \beta_n)| \overset{(i)}{\le} \sum_{i=1}^n c_f (\alpha_n Z_i - \beta_n)^2 \overset{(ii)}{\le} c_f' \rho,$$

where the constants c_f, c_f' depend only on f, inequality (i) follows from the definitions (31) and (32) of $\mathcal{E}_{n,f}$, and inequality (ii) follows from Lemma 8. Thus on the event $\mathcal{E}_{n,f}$, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^n R(\alpha_n Z_i - \beta_n) Z_i &\le \sum_{i=1}^n |R(\alpha_n Z_i - \beta_n)| \max_{i \le n} \frac{|Z_i|}{\sqrt{n}} \\
&\le c_f \rho \max_{i \le n} \frac{|Z_i|}{\sqrt{n}}.
\end{aligned} \tag{36}$$

Lastly, we show that the events $\mathcal{E}_{n,f}$ are suitably likely to hold:

Lemma 9. *Under the conditions of the theorem, we have*

$$\max_{i \le n} \left\{ |\alpha_n Z_i|, \frac{|Z_i|}{\sqrt{n}}, \frac{|Z_i|}{s_n \sqrt{n}} \right\} \stackrel{a.s.}{\to} 0.$$

See Section A.5 for a proof of Lemma 9.

We may now complete the proof of the theorem. Recall the definitions of $\mathcal{E}_n$ and $\mathcal{E}_{n,f}$ from Lemma 6 and Eq. (32), respectively. We prove the pointwise result first. Let $\epsilon > 0$ be arbitrary in the definition (32) of $\mathcal{E}_{n,f}$. Lemma 6 implies that $\mathbb{P}(\mathcal{E}_n \text{ eventually}) = 1$, and Lemma 9 implies that $\mathbb{P}(\mathcal{E}_{n,f} \text{ eventually}) = 1$. Thus with probability one there exists some (potentially random) N such that $n \ge N$ implies $|\alpha_n Z_n| \le \epsilon/2$ (by Lemma 9) and $|\beta_n| \le \epsilon/2$ (by inequality (29) and the following remark). In particular, expression (35) imples that $nR_n\overline{Z_n} \stackrel{a.s.}{\to} 0$, while expression (36) and Lemma 9 imply that

$$\sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^n R(\alpha_n Z_i - \beta_n) Z_i \stackrel{a.s.}{\to} 0.$$

Thus, asymptotically almost surely, we have by expression (34) that

$$\left| \sup_{P: D_f(P\|P_n) \le \frac{\rho}{n}} \mathbb{E}_P[Z] - \overline{Z_n} - \sqrt{\frac{\rho}{n} s_n^2} \right| \le c_f s_n \sqrt{\frac{\epsilon}{n} + \frac{\rho^2}{n^2}} + o(n^{-\frac{1}{2}}).$$

We also have that $s_n \stackrel{a.s.}{\to} \mathbb{E}[Z^2]^{\frac{1}{2}}$, which gives the result.

A.1 Proof of Lemma 5

Recall from the Paley-Zygmund inequality that for any $c_0 \in [0, 1]$, we have

$$\mathbb{P}(Z \ge c_0 \mathbb{E}[(Z)_+]) \ge (1 - c_0)^2 \frac{\mathbb{E}[(Z)_+]^2}{\mathbb{E}[(Z)_+^2]}$$

and

$$\mathbb{P}(-Z \ge c_0 \mathbb{E}[(-Z)_+]) \ge (1 - c_0)^2 \frac{\mathbb{E}[(-Z)_+]^2}{\mathbb{E}[(-Z)_+^2]}.$$

Noting that $\mathbb{E}[Z^2] = \mathbb{E}[(Z)_+^2 + \mathbb{E}[(-Z)_+^2]$, we thus obtain

$$\mathbb{P}\left(\frac{Z}{\mathbb{E}[(Z)_+]} \ge c_0 \right) \wedge \mathbb{P}\left(\frac{Z}{\mathbb{E}[(-Z)_+]} \le -c_0 \right) \ge (1 - c_0)^2 \frac{Z_\star^2}{\mathbb{E}[Z^2]} > 0.$$

This gives the lemma.

A.2 Proof of Lemma 6

We prove the first result. Let Let $B_i^+ = \mathbf{1}\left\{Z_i \ge c_0 \mathbb{E}[(Z)_+]\right\}$, so the B_i^+ are i.i.d. Bernoulli random variables with $\mathbb{P}(B_i^+ = 1) \ge q_\star$. Let $B_i^- = \mathbf{1}\left\{Z_i \le -c_0 \mathbb{E}[(-Z)_+]\right\}$, noting that similarly $\mathbb{P}(B_i^- =$

1) $\geq q_\star$. Then, we have

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n B_i^+ \leq \frac{nq_\star}{4}\right) &= \mathbb{P}\left(\sum_{i=1}^n (1-B_i^+) \geq n - \frac{nq_\star}{4}\right) \\ &= \mathbb{P}\left(\sum_{i=1}^n \left((1-B_i^+) - (1-\mathbb{E}B_i^+)\right) \geq n - \frac{nq_\star}{4} - n(1-\mathbb{E}B_i^+)\right) \\ &\overset{(a)}{\leq} \mathbb{P}\left(\sum_{i=1}^n \left((1-B_i^+) - (1-\mathbb{E}B_i^+)\right) \geq \frac{3nq_\star}{4}\right) \overset{(b)}{\leq} \exp\left(-\frac{9nq_\star^2}{8}\right).\end{aligned}$$

where (a) follows from $\mathbb{P}(B_i^+ = 1) \geq q_\star$ and (b) from Hoeffding's inequality. A similar bound holds for B_i^-'s by a symmetric argument. Hence,

$$\mathbb{P}(\mathcal{E}_n^c) \leq \mathbb{P}\left(\sum_{i=1}^n B_i^+ \leq \frac{nq_\star}{2}\right) + \mathbb{P}\left(\sum_{i=1}^n B_i^- \leq \frac{nq_\star}{2}\right) \leq 2\exp\left(-nq_\star^2\right) \leq 2\exp\left(-nq_\star^2\right).$$

A.3 Proof of Lemma 7

On the event (32), we may perform a Taylor expansion of $f^{*\prime}$ around $\alpha_n Z_i - \beta_n$ for all i, and we have

$$\begin{aligned}f^{*\prime}(\alpha_n Z_i - \beta_n) &= f^{*\prime}(0) + f^{*\prime\prime}(0)(\alpha_n Z_i - \beta_n) + R_{f^*}(\alpha_n Z_i - \beta_n) \\ &= 1 + 2(\alpha_n Z_i - \beta_n) + R_{f^*}(\alpha_n Z_i - \beta_n),\end{aligned}$$

where $|R_{f^*}(\alpha_n Z_i - \beta_n)| \leq 2c_f|\alpha_n Z_i - \beta_n|^2$. Because $1 = \sum_{i=1}^n p_{i,n}$, we thus find that on the event $\mathcal{E}_{n,f}$ of definition (32), we have

$$\beta_n = \alpha_n\left(\frac{1}{n}\sum_{i=1}^n Z_i\right) - \frac{1}{2n}\sum_{i=1}^n R_{f^*}(\alpha_n Z_i - \beta_n).$$

We now show that the remainder term is small on the event $\mathcal{E}_{n,f}$. We have by the smoothness of f near 1 (and f^* near 0) that there exists a constant $c_f' > 0$, dependent only on f, such that

$$f(f^{*\prime}(s)) = f(\underbrace{f^{*\prime}(0)}_{=1} + \underbrace{f^{*\prime\prime}(0)}_{=\frac{1}{2}}s + O(s^2)) \geq c_f' s^2$$

for $|s| \leq \epsilon_f$, where ϵ_f is the valid Taylor expansion threshold (31). Consequently, we find that on the event $\mathcal{E}_{n,f}$, we have

$$\rho \geq \sum_{i=1}^n f(f^{*\prime}(\alpha_n Z_i - \beta_n)) \geq \sum_{i=1}^n c_f'(\alpha_n Z_i - \beta_n)^2,$$

while for another constant C_f (dependent on f) we thus have

$$\sum_{i=1}^n |R_{f^*}(\alpha_n Z_i - \beta_n)| \leq C_f \sum_{i=1}^n (\alpha_n Z_i - \beta_n)^2 \leq \frac{C_f}{c_f'}\sum_{i=1}^n f(f^{*\prime}(\alpha_n Z_i - \beta_n)) \leq C_f\rho,$$

where the value of C_f changes from line to line but depends only on f. In particular, we find that on the event $\mathcal{E}_{n,f}$, the conclusions of the lemma hold.

A.4 Proof of Lemma 8

We first claim that on the event $\mathcal{E}_{n,f}$, we have

$$\rho \in [1 \pm c_f(\epsilon \wedge \epsilon_f \wedge \epsilon_{f^*})] \sum_{i=1}^n (f^{*\prime}(\alpha_n Z_i - \beta_n) - 1)^2, \tag{37}$$

where the constant c_f depends only on the function f. Deferring the proof of the claim (37) to the end of this section, we proceed with the proof of the lemma. Indeed, we note that

$$f^{*\prime}(\alpha_n Z_i - \beta_n) - 1 = f^{*\prime\prime}(\delta_i)(\alpha_n Z_i - \beta_n)$$

where $|\delta_i| \le |\alpha_n Z_i - \beta_n|$. As we have $|\delta_i| \le |\alpha_n Z_i - \beta_n| \le \epsilon \wedge \epsilon_f \wedge \epsilon_{f^*}$, a Taylor expansion implies that for a constant c_f only dependent on f, we have $f^{*\prime\prime}(\delta_i) \in [\frac{1}{2} \pm c_f(\epsilon \wedge \epsilon_f \wedge \epsilon_{f^*})]$. This gives the result of the lemma.

Proof of claim (37)**:** When the variance of $\{Z_i\}_{i=1}^n$ is non-zero, we have equality in the constraint $\sum_{i=1}^n f(np_{i,n}) \le \rho$. As $|\alpha_n Z_i - \beta_n| \le \epsilon_f$ on the event $\mathcal{E}_{n,f}$, we may thus perform a Taylor approximation to find

$$\begin{aligned} \rho &= \sum_{i=1}^n f(f^{*\prime}(\alpha_n Z_i - \beta_n)) \qquad (38) \\ &= \sum_{i=1}^n \left[f(1) + f'(1)(f^{*\prime}(\alpha_n Z_i - \beta_n) - 1) + \frac{1}{2} f''(1)(f^{*\prime}(\alpha_n Z_i - \beta_n) - 1)^2 + R_f(\alpha_n Z_i - \beta_n) \right], \end{aligned}$$

with remainder

$$R_f(\delta) = f(f^{*\prime}(\delta)) - \left[f(1) + f'(1)(f^{*\prime}(\delta) - 1) + \frac{1}{2} f''(1)(f^{*\prime}(\delta) - 1)^2 \right].$$

Noting that $|f^{*\prime}(\delta) - 1| \le 2\mathsf{r}_{f^*,1}|\delta|$ for $|\delta| \le \epsilon_{f^*}$, this remainder satisfies

$$\begin{aligned} |R_f(\delta)| \le \sup_{|\delta'| \le |\delta|} |f'''(f^{*\prime}(\delta'))| |f^{*\prime}(\delta) - 1|^3 &\le \sup_{|\delta'| \le |\delta|} |f'''(1 + 2\mathsf{r}_{f^*,1}\delta')| |f^{*\prime}(\delta) - 1|^3 \\ &\le 2\mathsf{r}_{f,3} |f^{*\prime}(\delta) - 1|^3 \end{aligned}$$

whenever $|\delta| \le \epsilon_{f^*}$. Thus, we have $|R_f(\alpha_n Z_i - \beta_n)| \le c_f |f^{*\prime}(\alpha_n Z_i - \beta_n) - 1|^3$ for a constant c_f dependent only on f as long as the event $\mathcal{E}_{n,f}$ holds. By expression (38) and that $f(1) = f'(1) = 0$ and $f''(1) = 2$, we obtain

$$\rho \in \sum_{i=1}^n (f^{*\prime}(\alpha_n Z_i - \beta_n) - 1)^2 + [-c_f, c_f] \sum_{i=1}^n |f^{*\prime}(\alpha_n Z_i - \beta_n) - 1|^3. \tag{39}$$

Now, we apply a Taylor expansion to $f^{*\prime}(\alpha_n Z_i - \beta_n)$, which yields that on the event $\mathcal{E}_{n,f}$, we have $|f^{*\prime}(\alpha_n Z_i - \beta_n) - 1| \le c_f |\alpha_n Z_i - \beta_n| \le c_f(\epsilon \wedge \epsilon_f \wedge \epsilon_{f^*})$ for a constant c_f depending on f only.

A.5 Proof of Lemma 9

The first result is a nearly immediate consequence of expression (29) and the following remark. Indeed, as noted by Owen [45, Lemma 3], we have

$$\sum_{n=1}^{\infty} \mathbb{P}(|Z_n|^2 \geq nZ_\star^2) \lesssim \frac{1}{Z_\star^2} \mathbb{E}[|Z_1|^2] < \infty,$$

so that $\mathbb{P}(|Z_n| \geq \sqrt{n} Z_\star \text{ i.o.}) = 0$. An identical argument with any positive constant $a > 0$ multiplying $Z_\star$ gives that $\mathbb{P}(|Z_n| \geq a\sqrt{n} Z_\star \text{ i.o.}) = 0$, so that $\max_{i \leq n} \frac{|Z_i|}{\sqrt{n}} \stackrel{a.s.}{\to} 0$. Recall the event $\mathcal{E}_n$ defined in Lemma 6. By Lemma 6, we have with probability 1 that there exists some (potentially random) N such that $\mathcal{E}_n$ holds for all $n \geq N$. On the event, inequality (29) shows that $\alpha_n \lesssim n^{-\frac{1}{2}} / Z_\star$, which shows that $\max_{i \leq n} \alpha_n |Z_i| \stackrel{a.s.}{\to} 0$. For the final term in the maximum, we note that on the event $\mathcal{E}_n$, we have

$$s_n^2 \geq \frac{q_\star}{2} \max\left\{(c_0 Z_\star - \overline{Z_n})^2, (-c_0 Z_\star - \overline{Z_n})^2\right\} \geq \frac{q_\star c_0^2}{2} Z_\star^2. \tag{40}$$

Thus we have $\max_{i \leq n} \frac{|Z_i|}{s_n \sqrt{n}} \lesssim \max_{i \leq n} \frac{|Z_i|}{Z_\star \sqrt{n}}$ on the event $\mathcal{E}_n$, which gives the result.

B Proof of uniform expansion

In this section, we give the proofs of theorems related to our uniform expansions. We begin by collecting important technical definitions and results, as well as a few preliminary lemmas, before providing the proof in Section B.3.

B.1 Preliminary results and definitions

We begin with several definitions and assorted standard lemmas important for our results, focusing on results on convergence in distribution in general metric spaces. See, for example, the first section of the book by van der Vaart and Wellner [59] for an overview.

Definition 5 (Tightness)**.** *A random variable X on a metric space $(\mathcal{X}, \mathsf{d})$ is* tight *if for all $\epsilon > 0$, there exists a compact set K_ϵ such that $\mathbb{P}(X \in K_\epsilon) \geq 1 - \epsilon$. A sequence of random variables $X_n \in \mathcal{X}$ is* asymptotically tight *if for every $\epsilon > 0$ there exists a compact set K such that*

$$\liminf_n P_*(X_n \in K^\delta) \geq 1 - \epsilon \quad \text{for all } \delta > 0,$$

where $K^\delta = \{x \in \mathcal{X} : \mathsf{d}(x, K) < \delta\}$ is the δ-enlargement of K and P_* denotes inner measure.

Lemma 10 (Prohorov's theorem [59], Theorem 1.3.9)**.** *Let $X_n \in \mathcal{X}$ be a sequence of random variables in the metric space $\mathcal{X}$. Then*

1. *If $X_n \stackrel{d}{\rightsquigarrow} X$ for some random variable X where X is tight, then X_n is asymptotically tight and measurable.*
2. *If X_n is asymptotically tight, then there is a subsequence $n(m)$ such that $X_{n(m)} \stackrel{d}{\rightsquigarrow} X$ for some tight random variable X.*

Thus, to show that a sequence of random vectors converges in distribution, one necessary step is to show that the sequence is tight. We now present two technical lemmas on this fact. In each lemma, $\mathcal{H}$ is some set (generally a collection of functions in our applications), and Ω_n is a sample space defined for each n. (In our applications, we take $\Omega_n = \Xi^n$.) We let $X_n(h) \in \mathbb{R}$ denote the random realization of X_n evaluated at $h \in \mathcal{H}$.

Lemma 11 (Van der Vaart and Wellner [59], Theorem 1.5.4)**.** *Let $X_n : \Omega_n \to \mathcal{L}^\infty(\mathcal{H})$. Then X_n converges weakly to a tight limit if and only if X_n is asymptotically tight and the marginals $(X_n(h_1), \ldots, X_n(h_k))$ converge weakly to a limit for every finite subset $\{h_1, \ldots, h_k\}$ of $\mathcal{H}$. If X_n is asymptotically tight and its marginals converge weakly to the marginals of $(X(h_1), \ldots, X(h_k))$ of X, then there is a version of X with uniformly bounded sample paths and $X_n \overset{d}{\rightsquigarrow} X$.*

Lemma 12 (Van der Vaart and Wellner [59], Theorem 1.5.7)**.** *A sequence of mappings $X_n : \Omega_n \to \mathcal{L}^\infty(\mathcal{H})$ is asymptotically tight if and only if (i) $X_n(h)$ is asymptotically tight in $\mathbb{R}$ for all $h \in \mathcal{H}$, (ii) there exists a semi-metric $\|\cdot\|$ on $\mathcal{H}$ such that $(\mathcal{H}, \|\cdot\|)$ is totally bounded, and (iii) X_n is asymptotically uniformly equicontinuous in probability, i.e., for every $\epsilon, \eta > 0$, there exists $\delta > 0$ such that $\limsup_{n\to\infty} \mathbb{P}\left(\sup_{\|h-h'\|<\delta} |X_n(h) - X_n(h')| > \epsilon\right) < \eta$.*

B.2 Technical lemmas

With these preliminary results stated, we now give a few technical lemmas necessary for the proof to come. Let $\mathcal{P}_{n,\rho} := \{P : D_f(P\|\widehat{P}_n) \leq \frac{\rho}{n}\}$ be the collection of distributions near $\widehat{P}_n$. We first show that $\mathcal{P}_{n,\rho}$ shrinks around the empirical measure as $n \to \infty$.

Lemma 13. *Let Assumption A hold. Then*

$$\sup_{n\in\mathbb{N}} \sup_{p\in\mathbb{R}^n} \left\{ \|np - \mathbb{1}\|_2 : p^\top \mathbb{1} = 1,\ p \geq 0,\ \sum_{i=1}^n f(np_i) \leq \rho \right\} \leq \sqrt{\frac{\rho}{\gamma_f}}$$

for some $\gamma_f > 0$ depending only on f.

Proof By performing a Taylor expansion of f around 1 for the point np_i and using $f(1) = f'(1) = 0$, we obtain

$$f(np_i) = \frac{1}{2} f''(s_i)(np_i - 1)^2$$

for some s_i between np_i and 1. As f is convex with $f''(1) > 0$, it is strictly increasing on $[1, \infty)$. Thus there exists a unique $M > 1$ such that $f(M) = \rho$. If $f(0) = \infty$, there is similarly a unique $m \in (0, 1)$ such that $f(m) = \rho$ (if no such m exists, because $f(0) < \rho$, define $m = 0$). Any $p \in \{p \in \mathbb{R}^n : p^\top \mathbb{1} = 1,\ p \geq 0,\ \sum_{i=1}^n f(np_i) \leq \rho\}$ must thus satisfy $np_i \in [m, M]$. Because f is C^2 and strictly convex, $\gamma := \inf_{s\in[m,M]} f''(s)$ exists, is attained, and is strictly positive. Using the Taylor expansion of the f-divergence, we have $(np_i - 1)^2 \leq 2f(np_i)/f''(s_i)$ for each i, and thus

$$\sum_{i=1}^n (np_i - 1)^2 \leq \sum_{i=1}^n \frac{2f(np_i)}{f''(s_i)} \leq \frac{2}{\gamma} \sum_{i=1}^n f(np_i) \leq \frac{2}{\gamma}\rho.$$

Taking the square root of each side gives the lemma. □

We also give a general result on tightness for large classes of functions $h : \Xi \to \mathbb{R}$.

Lemma 14. *If $\mathcal{H}$ is P_0-Donsker with L^2-integrable envelope M_2, that is, $h(\xi) \leq M_2(\xi)$ for all $h \in \mathcal{H}$ and $\mathbb{E}_{P_0}[M_2^2(\xi)] < \infty$, then $\sqrt{n}(Q_n - P_0)$ is asymptotically tight when viewed as a mapping $\mathcal{L}^\infty(\mathcal{H}) \to \mathbb{R}$ for any arbitrary sequence $Q_n \in \mathcal{P}_{n,\rho}$.*

Proof We use the charactization of asymptotic tightness in Lemma 12. With that in mind, consider an arbitrary sequence $Q_n \in \mathcal{P}_{n,\rho}$. We have

$$\begin{aligned}\mathbb{P}\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{n}(Q_n - \widehat{P}_n)(h-h')\right| \geq \epsilon\right) &\overset{(a)}{\leq} \mathbb{P}\left(\sup_{\|h-h'\|<\delta}\|nq - \mathbb{1}\|_2 \left\|h-h'\right\|_{L^2(\widehat{P}_n)} \geq \epsilon\right) \\ &\overset{(b)}{\leq} \mathbb{P}\left(\sqrt{\frac{\rho}{\gamma_f}}\sup_{\|h-h'\|<\delta}\left\|h-h'\right\|_{L^2(\widehat{P}_n)} \geq \epsilon\right)\end{aligned}$$

where inequality (a) follows from the Cauchy-Schwarz inequality and inequality (b) follows from Lemma 13. Since $\mathcal{H}$ is P_0-Donsker, the last term goes to 0 as $n \to \infty$ and $\delta \to 0$. Note that

$$\begin{aligned}&\limsup_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{n}(Q_n - P)(h-h')\right| \geq \epsilon\right) \\ &\leq \limsup_{\delta,n}\left\{\mathbb{P}\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{n}(Q_n - \widehat{P}_n)(h-h')\right| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{n}(\widehat{P}_n - P)(h-h')\right| \geq \frac{\epsilon}{2}\right)\right\}.\end{aligned}$$

As $\sqrt{n}(\widehat{P}_n - P_0)$ is asymptotically tight in $\mathcal{L}^\infty(\mathcal{H})$ (*e.g.*, [59, Theorem 1.5.4]), the second term vanishes by Lemma 12. Applying Lemma 12 again, we conclude that $\sqrt{n}(Q_n - P_0)$ is asymptotically tight. □

B.3 Proof of Theorem 2

The proof of the theorem uses Lemma 1 and standard tools of empirical process theory to make the expansion uniform. Without loss of generality, we assume that Z is a mean-zero random variable for all $x \in \mathcal{X}$ (as we may replace $Z(x,\xi)$ with $Z(x,\xi) - \mathbb{E}[Z(x,\xi)]$). We use the standard characterization of asymptotic tightness given by Lemma 11, so we show the finite dimensional convergence to zero of our process. It is clear that there is *some* random function ε_n such that

$$\sup_{P\in\mathcal{P}_{\rho,n}}\mathbb{E}_P[Z(x;\xi)] = \mathbb{E}_{\widehat{P}_n}[Z(x;\xi)] + \sqrt{\frac{\rho}{n}\mathrm{Var}_{P_0}(Z(x;\xi))} + \varepsilon_n(x),$$

but we must establish its uniform convergence to zero at a rate $o(n^{-\frac{1}{2}})$.

To establish asymptotic tightness of $\sup_{P\in\mathcal{P}_{\rho,n}}\mathbb{E}_P[Z(\cdot,\xi)]$, first note that we have finite dimensional marginal convergence. Indeed, we have $\sqrt{n}\varepsilon_n(x) \overset{P^*}{\to} 0$ for all $x \in \mathcal{X}$ by Lemma 1. By the definition of convergence in probability, we obtain for any finite k and any $x_1,\ldots,x_k \in \mathcal{X}$ that $\sqrt{n}\left(\varepsilon_n(x_1),\ldots,\varepsilon_n(x_k)\right) \overset{P^*}{\to} 0$. Further, by our Donsker assumption on $Z(x,\cdot)$ we have that $\{Z(x,\cdot)^2, x \in \mathcal{X}\}$ is a Glivenko-Cantelli class [59, Lemma 2.10.14]. That is,

$$\sup_{x\in\mathcal{X}}\left|\mathrm{Var}_{\widehat{P}_n}(Z(x,\xi)) - \mathrm{Var}_{P_0}(Z(x,\xi))\right| \overset{P^*}{\to} 0. \tag{41}$$

Now, we write the error term ε_n as

$$\sqrt{n}\varepsilon_n(\cdot) = \underbrace{\sqrt{n} \sup_{P: D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[Z(\cdot;\xi)]}_{(a)} - \underbrace{\sqrt{n}\,\mathbb{E}_{\widehat{P}_n}[Z(\cdot;\xi)]}_{(b)} - \underbrace{\sqrt{\rho \mathrm{Var}_{\widehat{P}_n}(Z(\cdot;\xi))}}_{(c)}.$$

Then term (a) is asymptotically tight in $\mathcal{L}^\infty(\mathcal{Z})$ by Lemma 14. The term (b) is tight because $\mathcal{Z}$ is P_0-Donsker by assumption, and term (c) is tight by the uniform Glivenko-Cantelli result (41). In particular, $\sqrt{n}\varepsilon_n(\cdot)$ is an asymptotically tight sequence in $\mathcal{L}^\infty(\mathcal{Z})$. As the finite dimensional distributions all converge to 0 in probability, Lemma 11 implies that $\sqrt{n}\varepsilon_n \overset{d}{\rightsquigarrow} 0$ in $\mathcal{L}^\infty(\mathcal{H})$ as desired. Of course, convergence in distribution to a constant implies convergence in probability to the constant.

B.4 Proof of Theorem 6

From Theorem 2, we have that

$$\sqrt{n}\left(\sup_{P: D_f\left(P\|\widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\cdot;\xi)] - \mathbb{E}_{P_0}[\ell(\cdot;\xi)]\right) \overset{d}{\rightsquigarrow} H \quad \text{in} \quad \mathcal{L}^\infty(\mathcal{H})$$

where $H(\cdot) = \sqrt{\rho \mathrm{Var}_{P_0}\ell(\cdot;\xi)} + G$ and G is a mean zero Gaussian process with covariance $\mathbb{E}_{P_0}G(x_1)G(x_2) = \mathrm{Cov}_{P_0}(\ell(x_1;\xi), \ell(x_2;\xi))$. Applying the delta method as in the proof of Theorem 4, we obtain

$$\sqrt{n}\left(u_n - \inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]\right) \overset{d}{\rightsquigarrow} \inf_{x\in S} H(x) \tag{42}$$

where $S = \mathrm{argmin}_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$. That is, we have

From the weak convergence (42), we obtain

$$\begin{aligned}\mathbb{P}\left(\inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \le u_n\right) &= \mathbb{P}\left(\sqrt{n}(u_n - \inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]) \ge 0\right) \\ &\to \mathbb{P}\left(\inf_{x\in S}\left\{\sqrt{\rho \mathrm{Var}_{P_0}\ell(x^*;\xi)} + N(0, \mathrm{Var}_{P_0}\ell(x^*;\xi))\right\} \ge 0\right).\end{aligned}$$

For S unique, the last term is equal to $\mathbb{P}\left(N(0,1) \ge -\sqrt{\rho}\right) = 1 - \frac{1}{2}P\left(\chi_1^2 \ge \rho\right)$.

C Proofs of generalized empirical likelihood convergence

In this appendix, we collect the proofs related to our distributional results on generalized empirical likelihood in abstract settings.

C.1 Proof of Proposition 1

Let $Z \in \mathbb{R}^d$ be random vectors with covariance Σ, where $\mathrm{rank}(\Sigma) = d_0$. From Theorem 3, we have that if we define

$$T_{s,n}(\lambda) := s \sup_{P: D_f(P\|P_n) \le \rho/n} \left\{s\mathbb{E}_P[Z^T\lambda]\right\}, \quad s \in \{-1, 1\},$$

then

$$\begin{bmatrix} \sqrt{n}(T_{1,n}(\lambda) - \mathbb{E}_{P_0}[Z^T\lambda]) \\ \sqrt{n}(T_{-1,n}(\lambda) - \mathbb{E}_{P_0}[Z^T\lambda]) \end{bmatrix} = \begin{bmatrix} \sqrt{n}(\mathbb{E}_{P_n}[Z]^T\lambda - \mathbb{E}_{P_0}[Z]^T\lambda) + \sqrt{\rho\lambda^T\Sigma\lambda} \\ \sqrt{n}(\mathbb{E}_{P_n}[Z]^T\lambda - \mathbb{E}_{P_0}[Z]^T\lambda) - \sqrt{\rho\lambda^T\Sigma\lambda} \end{bmatrix} + o_P(1)$$

uniformly in λ such that $\|\lambda\|_2 = 1$. (This class of functions is trivially P_0-Donsker.) The latter quantity converges (uniformly in λ) to

$$\begin{bmatrix} \lambda^T W + \sqrt{\rho\lambda^T\Sigma\lambda} \\ \lambda^T W - \sqrt{\rho\lambda^T\Sigma\lambda} \end{bmatrix}$$

for $W \sim \mathsf{N}(0, \Sigma)$ by the central limit theorem. Now, we have that

$$\mathbb{E}_{P_0}[Z] \in \underbrace{\{\mathbb{E}_P[Z] : D_f(P\|P_n) \le \rho/n\}}_{=:C_{\rho,n}}$$

if and only if

$$\inf_{\lambda:\|\lambda\|_2\le 1}\{T_{1,n}(\lambda) - \mathbb{E}_{P_0}[Z^T\lambda]\} \le 0 \text{ and } \sup_{\lambda:\|\lambda\|_2\le 1}\{T_{-1,n}(\lambda) - \mathbb{E}_{P_0}[Z^T\lambda]\} \ge 0$$

by convexity of the set $C_{\rho,n}$. But of course, by convergence in distribution and the homogeneity of $\lambda \mapsto \lambda^T W + \sqrt{\rho\lambda^T\Sigma\lambda}$, the probabilities of this event converge to

$$\mathbb{P}\left(\inf_\lambda\{\lambda^T W + \sqrt{\rho\lambda^T\Sigma\lambda}\} \ge 0, \sup_\lambda\{\lambda^T W - \sqrt{\rho\lambda^T\Sigma\lambda}\} \le 0\right) = \mathbb{P}\left(\|W\|_{\Sigma^\dagger} \ge \sqrt{\rho}\right) = \mathbb{P}(\chi^2_{d_0} \ge \rho)$$

by the continuous mapping theorem.

C.2 Proof of Theorem 3

We first state a standard result that the delta method applies for Hadamard differentiable functionals, as given by van der Vaart and Wellner [59, Section 3.9]. In the lemma, the sets Ω_n denote the implicit sample spaces defined for each n. For a proof, see [59, Theorem 3.9.4].

Lemma 15 (Delta method)**.** *Let $T : \mathcal{P} \subset \mathcal{M} \to \mathbb{R}$ be Hadamard differentiable at P tangentially to B with dT_P linear and continuous on the whole of $\mathcal{M}$. Let $P_n : \Omega_n \to \mathbb{R}$ be maps (treated as random elements of $\mathcal{M} \subset \mathcal{L}^\infty(\mathcal{H})$) with $r_n(P_n - P) \overset{d}{\rightsquigarrow} Z$ in $\mathcal{L}^\infty(\mathcal{H})$, where $r_n \to \infty$ and Z is a separable, Borel-measurable map. Then $r_n(T(P_n) - T(P)) - dT_P(r_n(P_n - P)) \overset{P^*}{\to} 0$.*

Since $\mathcal{H}$ was assumed to be P_0-Donsker, we have $\sqrt{n}(\widehat{P}_n - P_0) \overset{d}{\rightsquigarrow} G$ in $\mathcal{L}^\infty(\mathcal{H})$. Using the canonical derivative guaranteed by the Hadamard differentiability assumption (recall the limit (7)), we have from Lemma 15 that

$$T(\widehat{P}_n) = T(P_0) + \mathbb{E}_{\widehat{P}_n}[T^{(1)}(\xi, P_0)] + \kappa_n(\widehat{P}_n) \tag{43}$$

where $\kappa_n(\widehat{P}_n) = o_P(n^{-\frac{1}{2}})$. Next, we show that this is true uniformly over $\{P : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\}$. See Section C.3 for the proof.

Lemma 16. *Under assumptions of Theorem 3, we have*

$$\limsup_n \mathbb{P}\left(\sup\left\{|\kappa_n(P)| : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\right\} \ge \frac{\epsilon}{\sqrt{n}}\right) = 0 \tag{44}$$

where $\kappa_n(P) := T(P) - T(P_0) - \mathbb{E}_Q[T^{(1)}(\xi; P_0)]$.

We now see how the theorem is a direct consequence of Lemma 1 and Lemma 16. Taking sup over $\{P : D_f(P\|\widehat{P}_n) \leq \frac{\rho}{n}\}$ in (43), we have

$$\left| \sup_{P:D_f(P\|\widehat{P}_n)\leq \frac{\rho}{n}} T(P) - T(P_0) - \sup_{P:D_f(P\|\widehat{P}_n)\leq \frac{\rho}{n}} \mathbb{E}_P[T^{(1)}(\xi;P_0)] \right| \leq \sup_{P:D_f\left(P\|\widehat{P}_n\right)\leq \frac{\rho}{n}} \kappa_n(P).$$

Now, multiply both sides by $\sqrt{n}$ and apply Lemmas 1, 16 to obtain

$$\left| \sqrt{n}\left(\sup_{P:D_f\left(P\|\widehat{P}_n\right)\leq \frac{\rho}{n}} T(P) - T(P_0) \right) - \sqrt{n}\mathbb{E}_{\widehat{P}_n}[T^{(1)}(\xi;P_0)] - \sqrt{\rho \mathrm{Var} T^{(1)}(\xi;P_0)} \right| = o_p(1).$$

That is,

$$\sqrt{n}\left(\sup_{P:D_f\left(P\|\widehat{P}_n\right)\leq \frac{\rho}{n}} T(P) - T(P_0) \right) \overset{d}{\rightsquigarrow} \sqrt{\rho \mathrm{Var} T^{(1)}(\xi;P_0)} + N\left(0, \mathrm{Var} T^{(1)}(\xi;P_0)\right).$$

Hence, we have $\mathbb{P}\left(T(P_0) \leq \sup_{P:D_f\left(P\|\widehat{P}_n\right)\leq \frac{\rho}{n}} T(P)\right) \to P(N(0,1) \geq -\sqrt{\rho})$. By an exactly symmetric argument on $-T(P_0)$, we similarly have $\mathbb{P}\left(T(P_0) \geq \inf_{P:D_f\left(P\|\widehat{P}_n\right)\leq \frac{\rho}{n}} T(P)\right) \to P(N(0,1) \leq \sqrt{\rho})$. We conclude that

$$\mathbb{P}\left(T(P_0) \in \left\{T(P) : D_f\left(P\|\widehat{P}_n\right) \leq \frac{\rho}{n}\right\}\right) \to P(\chi_1^2 \leq \rho).$$

C.3 Proof of Lemma 16

Let $\mathcal{P}_{n,\rho} := \{P : D_f(P\|\widehat{P}_n) \leq \frac{\rho}{n}\}$. Recall that $\{X_n\} \subset \mathcal{L}^\infty(\mathcal{H})$ is asymptotically tight if for every $\epsilon > 0$, there exists a compact K such that $\liminf_{n\to\infty} \mathbb{P}\left(X_n \in K^\delta\right) \geq 1-\epsilon$ for all $\delta > 0$ where $K^\delta := \{y \in \mathcal{L}^\infty(\mathcal{H}) : d(y,K) < \delta\}$ (*e.g.*, [59, Def 1.3.7]).

Now, for an arbitrary $\delta > 0$, let $Q_n \in \mathcal{P}_{n,\rho}$ such that $|\kappa(Q_n)| \geq (1-\delta)\sup_{Q\in\mathcal{P}_{n,\rho}} |\kappa(Q)|$. Since the sequence $\sqrt{n}(Q_n - P_0)$ is asymptotically tight by Lemma 14, every subsequence has a further subsequence $n(m)$ such that $\sqrt{n(m)}(Q_{n(m)} - P_0) \overset{d}{\rightsquigarrow} X$ for some tight and Borel-measurable map X. It then follows from Lemma 15 that $\sqrt{n(m)}\kappa_{n(m)}(Q_{n(m)}) \to 0$ as $m \to \infty$. The desired result follows since

$$\mathbb{P}\left((1-\epsilon)\sqrt{n} \sup_{Q\in\mathcal{P}_{n,\rho}} |\kappa_{n(m)}(Q)| \geq \epsilon \right) \leq \mathbb{P}\left(\sqrt{n}|\kappa_{n(m)}(Q_n)| \geq \epsilon\right) \to 0.$$

D Proof of Theorem 4

We first show a result explicitly guaranteeing smoothness of $T(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$. To this end, we will show Hadamard differentiability of T tangentially to the set B where B is defined as the set of signed measures $H \in \mathcal{M} \subset \mathcal{L}^\infty(\mathcal{H})$ such that $\int \ell(x_n;\xi)dH(\xi) \to \ell(x;\xi)dH(\xi)$ when $\|x_n - x\| \to 0$. Note that since $\mathcal{X} \subset \mathbb{R}^d$ is a finite dimensional closed set, it does not matter which norm we endow $\mathcal{X}$ with. Here, $\mathcal{M} \subset \mathcal{L}^\infty(\mathcal{H})$ denotes the set of signed measures in $\mathcal{L}^\infty(\mathcal{H})$.

We can prove the following variant of Danskin's theorem. Our proof is motivated from [50].

Lemma 17. *If Assumption B holds and $\mathcal{X}$ is compact, then $T : \mathcal{P} \subset \mathcal{M} \to \mathbb{R}, P \mapsto \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ is Hadamard directionally differentiable on $\mathcal{P}$ tangentially to B with the derivative*

$$dT_P(H) := \inf_{x\in S_P} \int \ell(x;\xi) dH(\xi)$$

where $S_P = \operatorname{argmin}_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$.

Proof To ease notation, let $H(x) = \int \ell(x;\xi)dH(\xi)$ and denote the set of ϵ-optimal solutions for P as

$$S_P(\epsilon) := \left\{ x \in \mathcal{X} : \mathbb{E}_P[\ell(x;\xi)] \le \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)] + \epsilon \right\}.$$

By Assumption B, the level sets of $\mathbb{E}_P[\ell(\cdot;\xi)]$ are compact so that any sequence $x(\epsilon) \in S_P(\epsilon)$ has a further subsequence that converges to some $x \in S_P$ as $\epsilon \to 0$ (we use the same notation for the subsequence). Let $x_n \in S_P(t_n^2)$ for some fixed $t_n \to 0$. Then, for $H_n \to H$ in $\mathcal{L}^\infty(\mathcal{H})$ such that $P_0 + t_n H_n \in \mathcal{P}$, we have

$$T(P + t_n H_n) - T(P) \le (P + t_n H_n)(x_n) - P(x_n) + t_n^2 = t_n H_n(x_n) + t_n^2.$$

For some $x \in S_P$ is such that $x_n \to x$, we have $|H_n(x_n) - H(x)| \le \|H_n - H\|_{\mathcal{H}} + |H(x_n) - H(x)| \to 0$ by continuity of $H \in B$. Dividing the preceeding display by t_n and sending n to ∞, we hence obtain

$$\limsup_{n\to\infty} \frac{1}{t_n}(T(P + t_n H_n) - T(P)) \le \inf_{x\in S_P} H(x).$$

To see the other direction, let $y_n \in S_n(P + t_n H_n, t_n^2)$. By noting that $(P + t_n H_n)(y_n) \le \inf_{x\in\mathcal{X}}(P + t_n H_n)(x) + t_n^2 \le \inf_{x\in\mathcal{X}} P(x) + |t_n| \, \|H_n\|_{\mathcal{H}} + t_n^2$, we have

$$T(P + t_n H_n) - T(P) \ge (P + t_n H_n)(y_n) - t_n^2 - P(y_n) = t_n H_n(y_n) - t_n^2.$$

Again, dividing both sides by t_n and sending n to ∞, we obtain

$$\liminf_{n\to\infty} \frac{1}{t_n}(T(P + t_n H_n) - T(P)) \ge inf_{x\in S_P} H(x)$$

which shows the desired result. □

Using Lemma 17, it is straightforward to check that we satisfy all the requirements of Theorem 3. First, the derivative $dT_P(H)$ can be extended to the whole of $\mathcal{M}$ using the exact same definition. It is then clear that when the set of P-optima S_P is unique, dT_P is a linear functional on $\mathcal{M}$. Further, dT_P is continuous with respect to the supremum norm $\|H\|_{\mathcal{H}} = \sup_{h\in\mathcal{H}} |Hh|$ and has canonical gradient $T^{(1)}(\xi;P) = \ell(x^*;\xi) - \mathbb{E}_{P_0}[\ell(x;\xi)]$.

Now, to apply Theorem 3, it remains to see that G takes values on B. Since G is a tight element in $\mathcal{L}^\infty(\mathcal{H})$, there exists a metric $\|\cdot\|$ on $\mathcal{H}$ such that $h \mapsto Gh$ is uniformly continuous in $\|\cdot\|$ (see, for example, [59, Theorem 1.5.8]). Noting that $\mathcal{H}$ can be identified with $\mathcal{X} \subset \mathbb{R}^d$ which is a finite dimensional set, it follows that there exists a version of $G : \Xi \to \mathcal{L}^\infty(\mathcal{X})$ that has continuous sample paths with respect to the norm on $\mathcal{X}$. Hence, we have shown that there is version of G such that $G \in B$ P_0-a.s.. Applying Theorem 3 to T at $P_0 \in \mathcal{P}$, we obtain the result.

E Extension to Stochastic Constraints

We can further extend Theorem 4 to the case where the feasible region $\mathcal{X}$ is estimated via random samples as well. Let $\mathcal{X}$ be given by the set of constraints

$$\mathcal{X} = \{x : \mathbb{E}_{P_0}[c_k(x;\xi)] \leq 0 \ k = 1,\dots,r,\ \mathbb{E}_{P_0}[c_k(x;\xi)] = 0 \ k = r+1,\dots,q\}\,.$$

The space under consideration will be restricted by setting

$$\mathcal{H} = \{\ell(x;\cdot) : x \in \mathcal{X}\} \cup \cup_{k=1}^{q} \{c_k(x;\cdot) : x \in \mathcal{X}\}\,.$$

The statistical functional in question can be written as

$$\begin{array}{rlll} T(P) = & \text{minimize} & \mathbb{E}_P[\ell(x;\xi)] & \\ & \text{subject to} & \mathbb{E}_P[c_k(x;\xi)] \leq 0, & k = 1,\dots,r \\ & & \mathbb{E}_P[c_k(x;\xi)] = 0, & k = r+1,\dots,q \end{array} \tag{45}$$

We will assume that $\ell(\cdot;\xi)$, $c_k(\cdot;\xi)$ $k = 1,\dots,m$ are convex P_0-a.s.. Consider the Lagrangian

$$L(P;x,\lambda) := \mathbb{E}_P[\ell(x;\xi)] + \sum_{k=1}^{q} \lambda_k \mathbb{E}_P[c_k(x;\xi)]$$

and denote by $S_P = \operatorname{argmin}_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ as before. If Slater's condition holds for P, then for any $\bar{x} \in S_P$, there is a $\bar{\lambda} \in \mathbb{R}^q$ such that

$$\begin{aligned} &L(P;\bar{x},\bar{\lambda}) = \min_{x\in\mathcal{X}} L(P;x,\bar{\lambda}) \\ &\bar{\lambda}_k \geq 0,\ \bar{\lambda}_k \mathbb{E}_P[c_k(x;\xi)] = 0, \quad k = 1,\dots,r \end{aligned}$$

If we denote by Λ_P the set of dual variables satisfying the above KKT conditions (it does not depend on $\bar{x}$), Slater's condition guarantees that Λ_P is bounded.

[54, Theorem 3.4] showed that $T(P)$ Hadamard directionally differentiable. The proof proceeds similarly as in Lemma 17.

Proposition 10. *Shapiro [54]*
Assume that $\ell(\cdot;\xi)$, $c_k(\cdot;\xi)$ $k = 1,\dots,q$ are convex P-a.s. and $\mathcal{X}$ compact. If Slater's condition holds for the optimization problem (45), $T(\cdot)$ is Hadamard directionally differentiable at P tangentially to the whole of $\mathcal{L}^\infty(\mathcal{H})$ with

$$dT_P(Q) = \min_{x\in S_P} \max_{\lambda\in\Lambda_P} L(P;x,\lambda)$$

As before, we need $T(P)$ to be Hadamard differentiable. For this, we will require Λ_P to be singleton. The following necessary and sufficient condition for an unique dual optimum was proved in [36].

Assumption E. *For a $\bar{x} \in S_P$ and $\bar{\lambda} \in \Lambda_P$,*

$$\nabla\mathbb{E}_P[\ell(\bar{x};\xi)], \quad \nabla\mathbb{E}_P[c_k(\bar{x};\xi)], \quad k = 1,\dots,q$$

are linearly independent. Further, there exists a $z \in \mathbb{R}^d$ such that

$$\nabla\mathbb{E}_P[c_k(\bar{x};\xi)]^\top z < 0 \text{ for } k \in I, \quad \nabla\mathbb{E}_P[c_k(\bar{x};\xi)]^\top z = 0 \text{ for } k \in J \cup \{k : r+1 \leq k \leq q\}$$

where $I = \{k : \bar{\lambda}_k > 0\}$, $J = \{k : \bar{\lambda}_k = 0\}$.

Using Proposition 10 in conjunction with Theorem 3, we obtain the following extension of Theorem 4.

Theorem 11. *Let Assumptions A, B, C, E hold with S_{P_0} and Λ_{P_0} singleton. Then for $T(P)$ given by (45), we have that (9) holds.*

F Proof of Theorems 8, 9

We first show Theorem 9. Below we define some concepts used in the proof.

Definition 6. *Let $\{A_n\}$ be a sequence of sets in $\mathbb{R}^d$. The* limit supremum *(or* limit exterior *or* outer limit*) and* limit infimum *(*limit interior *or* inner limit*) of the sequence $\{A_n\}$ are*

$$\limsup_n A_n := \left\{ x \in \mathbb{R}^d \mid \liminf_{n\to\infty} \operatorname{dist}(x, A_n) = 0 \right\}$$
$$\liminf_n A_n := \left\{ x \in \mathbb{R}^d \mid \limsup_{n\to\infty} \operatorname{dist}(x, A_n) = 0 \right\}.$$

Equivalently, the limit supremum is the set of all cluster points of all selections of the pertinent multifunction and the limit infimum the set of limits of all convergence selections.

Building on this definition, we define a notion of convergence in terms of epigraphs.

Definition 7. *We say that the given sequence of functions g_n* epi-converges *to a function g if*

$$\operatorname{epi} g = \liminf_{n\to\infty} \operatorname{epi} g_n = \limsup_{n\to\infty} \operatorname{epi} g_n. \tag{46}$$

where $\operatorname{epi} g = \{(x, r) : g(x) \le r\}$.

We use $g_n \overset{\text{epi}}{\to} g$ to denote the epigraphical convergence of g_n to g. If g is proper ($\operatorname{dom} g \neq \varnothing$), we can characterize epigraphical convergence (46) for convex lower semi-continuous functions. See Rockafellar and Wets [49, Theorem 7.17] for a proof.

(i) There exists a dense set $A \subset \mathbb{R}^d$ such that $g_n(x) \to g(x)$ for all $x \in A$.

(ii) For all compact $C \subset \operatorname{dom} g$ not containing a boundary point of $\operatorname{dom} g$,

$$\lim_n \sup_{x\in C} |g_n(x) - g(x)| = 0.$$

The last characterization says that epigraphical convergence is equivalent to uniform convergence on compacts. This property gives us a more general consistency result.

Proof of Theorem 9 We want to use the last characterization of epiconvergence given above. To this end, we first confirm the hypothesis for the equivalence holds. The conjugate function f^* is proper since it is the conjugate of a real-valued convex function on $[0, \infty)$. From the dual representation in Proposition 5,

$$F(x) := \mathbb{E}_{P_0}[\ell(x;\xi)]$$
$$\widehat{F_n}(x) := \sup_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_{P_0}[\ell(x;\xi)] : D_f\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n} \right\} = \inf_{\lambda\ge 0,\eta} \frac{1}{n} \sum_{i=1}^n \lambda f^*\left(\frac{\ell(x;\xi_i) - \eta}{\lambda}\right) + \frac{\rho}{n}\lambda + \eta$$

and hence both functions are proper and have a minorizing affine function. Since it is a conjugate function, f^* is lower semi-continuous. Hence, $\widehat{F_n}$ is lower semi-continuous P_0-a.s. from the hypothesis. To see that F is lower semi-continuous, we use Fatou's Lemma:

$$\mathbb{E}_{P_0}[\ell(x;\xi)] \le \mathbb{E}_{P_0}[\liminf_{z\to x} \ell(z;\xi)] \le \liminf_{z\to x} \mathbb{E}_{P_0}[\ell(z;\xi)].$$

From Lemma 13 and (23), the right hand side of (23) tends to 0 on a dense subset of $\mathcal{X}$. Then, as noted in Section 5.2, $\widehat{F_n} \overset{\text{epi}}{\to} F$ P_0-a.s. in view of Rockafellar and Wets [49, Theorem 7.17].

From Assumption B, the level sets of $\mathbb{E}_{P_0}[\ell(x;\xi)]$ are bounded and $S = \operatorname{argmin}_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$ is nonempty and compact (Rockafellar and Wets [49, Theorem 1.9]). Let C be a compact subset such that it contains S in its interior. From the third characterization of epi-convergence for convex functions, we have that

$$\sup_{x\in\mathcal{C}} \left|\widehat{F_n}(x) - F(x)\right| \overset{P^*}{\to} 0.$$

Or equivalently, $\widehat{F_n} \overset{P^*}{\to} F$ in $\mathcal{L}^\infty(\mathcal{C})$ where the function class is taken over C instead of $\mathcal{X}$ in (10). Note that the convergence is in outer probability since the supremum over $x \in \mathcal{X}$ may not be measurable. By the a.s. representation theorem, (van der Vaart and Wellner [59, Theorem 1.10.4]), there exists $\overline{F_n}$ defined on some probability space $(\overline{\Xi}^\infty, \overline{\mathcal{E}}^\infty, \overline{P_0}^\infty)$ such that $\overline{F_n} \overset{a.u.}{\to} F$ in $\mathcal{L}^\infty(C)$ and $\overline{\mathbb{E}}^*_{P_0} h(\overline{F_n}) = \mathbb{E}^*_{P_0} h(F_n)$ for all bounded $h : \mathcal{L}^\infty(C) \to \mathbb{R}$. By definition of almost uniform convergence, the first claim is equivalent to, for a fixed $\epsilon > 0$, there exists $A \subset \overline{\mathcal{E}}^\infty$ such that

$$\sup_{x\in\mathcal{C}} \left|\overline{F_n}(x) - F(x)\right| \to 0$$

uniformly over A (note that the preceding display is random and the uniformity is over $\omega \in A$).

Denote by $\overline{S_n}$ the set of optimizers of $\overline{F_n}$ and $\overline{S_n}(C)$ the set of optimizers of $\overline{F_n}$ over C. We first claim that $d_{\text{haus}}(\overline{S_n}(C), S) \overset{P^*}{\to} 0$. If this is true, since S is contained in the interior of C, it will follow that $d_{\text{haus}}(\overline{S_n}, S) \overset{P^*}{\to} 0$. Further, $S_n(C)$ and S are both contained in C and hence the function $h : \overline{\mathcal{L}}^\infty(C) \to \mathbb{R}, Q \mapsto d_{\text{haus}}(\operatorname{argmin}_{x\in C} \int \ell(x;\xi) dQ, S)$ is bounded. Here the space of bounded functions $\overline{\mathcal{L}}^\infty$ is taken to be appropriately redefined for $(\overline{\Xi}, \overline{\mathcal{E}}, \overline{P_0})$. Note that the argmin function on $\overline{\mathcal{L}}^\infty(C)$ is well defined since $\int \ell(x;\xi) dQ$ also has a compact set of minima since C is compact. It follows that $d_{\text{haus}}(\overline{S_n}(C), S)$ and $d_{\text{haus}}(S_n(C), S)$ has the same law. Hence, $d_{\text{haus}}(S_n(C), S) \overset{P^*}{\to} 0$ and as before we can conclude $d_{\text{haus}}(S_n, S) \overset{P^*}{\to} 0$.

We now remains to show the claim $d_{\text{haus}}(\overline{S_n}(C), S) \overset{P^*}{\to} 0$. Let us assume otherwise for contradiction. Then, there exists $x_n \in \overline{S_n}(C)$, such that $\operatorname{dist}(x_n, S) \geq \delta$ for some $\delta > 0$ when n is large. For each fixed $\omega \in \overline{\Xi}^\infty$, $x_n(\omega) \in C$ so that there exists a subsequence (which we also denote by x_n) that converges to a value x^*. By assumption, $\operatorname{dist}(x^*, S) \geq \epsilon$ and hence $\overline{F_n}(x_n)(\omega) \leq \overline{F_n}(x)(\omega)$ for all $x \in C$. For $\omega \in A$, let $n \to \infty$ to obtain $F(x^*)(\omega) \leq F(x)(\omega)$ for all $x \in C$. But then, this contradicts $x^* \notin S$. Hence, $d_{\text{haus}}(\overline{S_n}(C), S) \to 0$ on A. Since $P(A) \geq 1 - \epsilon$ where ϵ was arbitrarily chosen, we obtain the desired result. □

Proof of Theorem 8 The proof of Theorem 8 proceeds identically as above, except now, the uniform convergence is established by the Glivenko-Cantelli hypothesis. □