# Statistics of Robust Optimization:
# A Generalized Empirical Likelihood Approach

John C. Duchi

Department of Electrical Engineering and Statistics, Stanford University, jduchi@stanford.edu,

Peter W. Glynn, Hongseok Namkoong

Department of Management Science and Engineering, Stanford University, glynn, hnamk@stanford.edu,

We study statistical inference and distributionally robust solution methods for stochastic optimization problems, focusing on confidence intervals for optimal values and solutions that achieve exact coverage asymptotically. We develop a generalized empirical likelihood framework—based on distributional uncertainty sets constructed from nonparametric $f$-divergence balls—for Hadamard differentiable functionals, and in particular, stochastic optimization problems. As consequences of this theory, we provide a principled method for choosing the size of distributional uncertainty regions to provide one- and two-sided confidence intervals that achieve exact coverage. We also give an asymptotic expansion for our distributionally robust formulation, showing how robustification regularizes problems by their variance. Finally, we show that optimizers of the distributionally robust formulations we study enjoy (essentially) the same consistency properties as those in classical sample average approximations. Our general approach applies to quickly mixing stationary sequences, including geometrically ergodic Harris recurrent Markov chains.

*Key words*: Stochastic Optimization, Robust Optimization, Empirical Likelihood
*MSC2000 subject classification*:
*OR/MS subject classification*: Primary: ; secondary:
*History*:

**1. Introduction**  We study statistical properties of distributionally robust solution methods for the stochastic optimization problem

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

In the formulation (1), the feasible region $\mathcal{X} \subset \mathbb{R}^d$ is a nonempty closed set, $\xi$ is a random vector on the probability space $(\Xi, \mathcal{A}, P_0)$, where the domain $\Xi$ 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) function. In most data-based decision making scenarios, the underlying distribution $P_0$ is unknown, and even in scenarios, such as 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, one typically [78] approximates the population objective (1) using the sample average approximation (SAA) based on a sample $\xi_1, \ldots, \xi_n \stackrel{\text{iid}}{\sim} P_0$

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

where $\widehat{P}_n$ denotes the usual empirical measure over the sample $\{\xi_i\}_{i=1}^n$.

We study approaches to constructing confidence intervals for problem (1) and demonstrating consistency of its approximate solutions. We develop a family of convex optimization programs, based on the distributionally robust optimization framework [26, 5, 12, 6], which allow us to provide

confidence intervals with asymptotically exact coverage for optimal values of the problem (1). Our approach further yields approximate solutions $\widehat{x}_n$ that achieve an asymptotically guaranteed level of performance as measured by the population objective $\mathbb{E}_{P_0}[\ell(x;\xi)]$. More concretely, define the optimal value functional $T_{\mathrm{opt}}$ that acts on probability distributions on $\Xi$ by

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

For a fixed confidence level $\alpha$, we show how to construct an interval $[l_n, u_n]$ based on the sample $\xi_1, \ldots, \xi_n$ with (asymptotically) *exact coverage*

$$\lim_{n \to \infty} \mathbb{P}\left(T_{\mathrm{opt}}(P_0) \in [l_n, u_n]\right) = 1 - \alpha. \tag{3}$$

This exact coverage indicates the interval $[l_n, u_n]$ has correct size as the sample size $n$ tends to infinity. We also give sharper statements than the asymptotic guarantee (3), providing expansions for $l_n$ and $u_n$ and giving rates at which $u_n - l_n \to 0$.

Before summarizing our main contributions, we describe our approach and discuss related methods. We begin by recalling divergence measures for probability distributions [1, 23]. For a lower semi-continuous convex function $f : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ satisfying $f(1) = 0$, the $f$-*divergence* between probability distributions $P$ and $Q$ on $\Xi$ 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 $\mu$ is a $\sigma$-finite measure with $P, Q \ll \mu$, and $p := dP/d\mu$ and $q := dQ/d\mu$. With this definition, we will show that for $f \in \mathcal{C}^3$ near 1 with $f''(1) = 2$, the upper and lower confidence bounds

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

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

yield asymptotically exact coverage (3). In the formulation (4), the parameter $\rho = \chi^2_{1,1-\alpha}$ is chosen as the $(1-\alpha)$-quantile of the $\chi^2_1$ distribution.

The upper endpoint (4a) is a natural distributionally robust formulation for the sample average approximation (2), proposed by Ben-Tal et al. [6] for distributions $P$ with finite support. The approach in the current paper applies to arbitrary distributions, and we are therefore able to explicitly link (typically dichotomous [5]) robust optimization formulations with stochastic optimization. We show how a robust optimization approach for dealing with parameter uncertainy yields solutions with a number of desirable statistical properties, even in situations with dependent sequences $\{\xi_i\}$. The exact coverage guarantees (3) give a principled method for choosing the size $\rho$ of distributional uncertainty regions to provide one- and two-sided confidence intervals.

We now summarize our contributions, unifying the approach to uncertainty based on robust optimization with classical statistical goals.

(i) We develop an empirical likelihood framework for general smooth functionals $T(P)$, applying it in particular to the optimization functional $T_{\mathrm{opt}}(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$. We show how the construction (4a)–(4b) of $[l_n, u_n]$ gives a confidence interval with exact coverage (3) for $T_{\mathrm{opt}}(P_0)$ when the minimizer of $\mathbb{E}_{P_0}[\ell(x;\xi)]$ is unique. To do so, we extend Owen's empirical likelihood theory [63, 62] to suitably smooth (Hadamard differentiable) nonparametric functionals $T(P)$ with general $f$-divergence measures (the most general that we know in the literature); our proof is different from Owen's classical result even when $T(P) = \mathbb{E}_P[X]$ and extends to stationary sequences $\{\xi_i\}$.

(ii) We show that the upper confidence set $(-\infty, u_n]$ is a one-sided confidence interval with exact coverage 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 $\ell$ 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), which is efficiently computable when $\ell$ is convex, provides an upper confidence bound for the optimal population objective (1).

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

$$\sup_{P \ll \widehat{P}_n} \left\{\mathbb{E}_P[\ell(x;\xi)] : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\right\} = \mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)] + (1 + o(1))\sqrt{\frac{\rho}{n}\mathrm{Var}_P(\ell(x;\xi))}, \qquad (5)$$

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 $\ell$ 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 for consistency of SAA (see Assumption 5). More precisely, for the sets of optima defined by

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

the distance from any point in $S_n^\star$ to $S^\star$ tends to zero so long as $\ell$ has more than one moment under $P_0$ and is lower semi-continuous.

**Background and prior work** The nonparametric inference framework for stochastic optimization we develop in this paper is the empirical likelihood counterpart of the normality theory that Shapiro develops [74, 76]. While an extensive literature exists on statistical inference for stochastic optimization problems (see, for example, the line of work [32, 74, 44, 75, 46, 76, 45, 77, 78]), Owen's empirical likelihood framework [64] has received little attention in the stochastic optimization literature save for notable recent exceptions [51, 50]. In its classical form, empirical likelihood provides a confidence set for a $d$-dimensional mean $\mathbb{E}_{P_0}[Y]$ (with a full-rank covariance) by using the set $C_{\rho,n} := \{\mathbb{E}_P[Y] : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\}$ where $f(t) = -2\log t$. Empirical likelihood theory shows that if we set $\rho = \chi^2_{d,1-\alpha} := \inf\left\{\rho' : \mathbb{P}(\|Z\|_2^2 \le \rho') \ge 1 - \alpha \text{ for } Z \sim \mathsf{N}(0, I_{d\times d})\right\}$, then $C_{\rho,n}$ is an asymptotically exact $(1-\alpha)$-confidence region, i.e. $\mathbb{P}(\mathbb{E}_{P_0}[Y] \in C_{\rho,n}) \to 1 - \alpha$. Through a self-normalization property, empirical likelihood requires no knowledge or estimation of unknown quantities, such as variance. We show such asymptotically pivotal results also apply for the robust optimization formulation (4). The empirical likelihood confidence interval $[l_n, u_n]$ has the desirable characteristic that when $\ell(x;\xi) \ge 0$, then $l_n \ge 0$ (and similarly for $u_n$), which is not necessarily true for confidence intervals based on the normal distribution.

Using confidence sets to robustify optimization problems involving randomness is common (see Ben-Tal et al. [5, Chapter 2]). A number of researchers extend 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 [26, 83, 6, 12, 11]. The duality of confidence regions and hypothesis tests [52] gives a natural connection between robust optimization, uncertainty sets, and statistical tests. Delage and Ye [26] made initial progress in this direction by constructing confidence regions based

on mean and covariance matrices from the data, and Jiang and Guan [42] expand this line of research to other moment constraints. Bertsimas, Gupta, and Kallus [12, 11] develop 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 [71] so long as the underlying distribution is bounded; they further give confidence regions that do not have exact coverage. The formulation (4) has similar motivation to the preceding works, as the uncertainty set

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

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 extend this by showing that, under mild conditions, the values (4a) and (4b) provide upper and lower confidence bounds for $T(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ with (asymptotically) exact coverage.

Ben-Tal et al. [6] 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 a one-sided confidence interval with (asymptotically) inexact coverage (that is, they only give a bound on the coverage probability). In the unconstrained setting $\mathcal{X} = \mathbb{R}^d$, Lam and Zhou [51] used estimating equations to show that standard empirical likelihood theory gives confidence bounds for stochastic optimization problems. Their confidence bounds have asymptotically inexact confidence regions, although they do not require unique solutions of the optimization problem as our results sometimes do. The result (i) generalizes these works, as we show how the robust formulation (4) yields asymptotically exact confidence intervals for general distributions $P_0$, and general constrained ($\mathcal{X} \subset \mathbb{R}^d$) stochastic optimization problems.

Ben-Tal et al.'s robust sample approximation [6] and Bertsimas et al.'s goodness of fit testing-based procedures [11] provide natural motivation for formulations similar to ours (4). However, by considering completely nonparametric measures of fit we can depart from assumptions on the structure of $\Xi$ (i.e. that it is finite or a compact subset of $\mathbb{R}^m$). The $f$-divergence formulation (4) allows for 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). Concurrent with the initial `arXiv` version of this work, Lam [49, 50] develops variance expansions similar to ours (5), focusing on the KL-divergence and empirical likelihood cases (i.e. $f(t) = -2\log t$ with i.i.d. data). Our methods of proof are different, and our expansions hold almost-surely (as opposed to in probability), apply to general $f$-divergences, and generalize to dependent sequences under standard ergodicity conditions.

The recent line of work on distributionally robust optimization using Wasserstein distances [65, 84, 33, 72, 15, 79] is similar in spirit to the formulation considered here. Unlike $f$-divergences, uncertainty regions formed by Wasserstein distances contain distributions that have support different to that of the empirical distribution. Using concentration results for Wasserstein distances with light-tailed random variables [35], Esfahani and Kuhn [33] gave finite sample guarantees with nonparametric rates $O(n^{-1/d})$. The $f$-divergence formulation we consider yields different statistical guarantees; for random variables with only second moments, we give confidence bounds that achieve (asymptotically) *exact* coverage at the parametric rate $O(n^{-1/2})$. Further, the robustification approach via Wasserstein distances is often computationally challenging (with current techology), as tractable convex formulations are available [72, 33] only under stringent conditions on the functional $\xi \mapsto \ell(x;\xi)$. On the other hand, efficient solution methods [6, 56] for the robust problem (4a) are obtainable without restriction on the objective function $\ell(x;\xi)$ other than convexity in $x$.

**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 $X_n \Rightarrow X$ if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded continuous functions $f$. We write $X_n \xrightarrow{P^*} X$ for random variables $X_n$ converging to a random variable $X$ in outer probability [82, Section 1.2]. Given a set $A \subset \mathbb{R}^d$, norm $\|\cdot\|$, and point $x$, the distance $\mathrm{dist}(x, A) = \inf_y\{\|x - y\| : y \in A\}$. The *inclusion distance*, or the *deviation*, from a set $A$ to $B$ is

$$d_\subset(A, B) := \sup_{x \in A} \mathrm{dist}(x, B) = \inf\left\{\epsilon \geq 0 : A \subset \{y : \mathrm{dist}(y, B) \leq \epsilon\}\right\}. \tag{6}$$

For a measure $\mu$ 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 $\alpha \in [0, 1]$, we define $\chi^2_{d,\alpha}$ to be the $\alpha$-quantile of a $\chi^2_d$ random variable, that is, the value such that $\mathbb{P}(\|Z\|_2^2 \leq \chi^2_{d,\alpha}) = \alpha$ for $Z \sim \mathsf{N}(0, I_{d \times d})$. The Fenchel conjugate of a function $f$ is $f^*(y) = \sup_x\{y^T x - f(x)\}$. 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 [39]. We let $\mathbf{I}_A(x)$ be the $\{0, \infty\}$-valued membership function, so $\mathbf{I}_A(x) = \infty$ if $x \notin A$, 0 otherwise. To address measurability issues, we use outer measures and corresponding convergence notions [82, Section 1.2-5]. Throughout the paper, the sequence $\{\xi_i\}$ is i.i.d. unless explicitly stated.

**Outline** In order to highlight applications of our general results to stochastic optimization problems, we first present results for the optimal value functional $T_{\mathrm{opt}}(P) := \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x; \xi)]$, before presenting their most general forms. In Section 2, we first describe the necessary background on generalized empirical likelihood and establish our basic variance expansions. We apply these results in Section 3 to stochastic optimization problems, including those involving dependent data, and give computationally tractable procedures for solving the robust formulation (4a). In Section 4, we develop the connections between distributional robustness and principled choices of the size $\rho$ in the uncertainty sets $\{P : D_f(P\|\widehat{P}_n) \leq \rho/n\}$, choosing $\rho$ to obtain asymptotically exact bounds on the population optimal value (1). To understand that the cost of the types of robustness we consider is reasonably small, in Section 5 we show consistency of the empirical robust optimizers under (essentially) the same conditions guaranteeing consistency of SAA. We conclude the "applications" of the paper to optimization and modeling with numerical investigation in Section 6, demonstrating benefits and drawbacks of the robustness approach over classical stochastic approximations. To conclude the paper, we present the full generalization of empirical likelihood theory to $f$-divergences, Hadamard differentiable functionals, and uniform (Donsker) classes of random variables in Section 7.

**2. Generalized Empirical Likelihood and Asymptotic Expansions** We begin by briefly reviewing generalized empirical likelihood theory [64, 59, 41], showing classical results in Section 2.1 and then turning to our new expansions in Section 2.2. 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 $P_0$. Let $\mathcal{P}$ be the set of probability distributions on $\Xi$ and let $T : \mathcal{P} \to \mathbb{R}$ be the statistical quantity of interest. We typically consider $T_{\mathrm{opt}}(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x; \xi)]$ with $Z(\xi) := \ell(\cdot; \xi)$, although our theory applies in more generality (see Section 7). 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 $Z_1, \ldots, Z_n$. The set $C_{n,\rho}$ is the image of $T$ on an $f$-divergence neighborhood of the empirical distribution $\widehat{P}_n$. We may define a dual quantity, the profile divergence $R_n : \mathbb{R} \to \mathbb{R}_+$ (called the profile likelihood [64] when $f(t) = -2\log t$) by

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

Then for any $P \in \mathcal{P}$, we have $T(P) \in C_{n,\rho}$ if and only if $R_n(T(P)) \leq \frac{\rho}{n}$. Classical empirical likelihood [63, 62, 64] considers $f(t) = -2\log t$ so that $D_f(P\|\widehat{P}_n) = 2D_{\mathrm{kl}}(\widehat{P}_n\|P)$, in which case the divergence is the nonparametric log-likelihood ratio. To show that $C_{n,\rho}$ is actually a meaningful confidence set, the goal is then to demonstrate that (for appropriately smooth functionals $T$)

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

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

**2.1. Generalized Empirical Likelihood for Means** In the classical case in which the vectors $Z_i \in \mathbb{R}^d$ and are i.i.d., Owen [62] shows that empirical likelihood applied to the mean $T(P_0) := \mathbb{E}_{P_0}[Z]$ guarantees elegant asymptotic properties: when $\mathrm{Cov}(Z)$ has rank $d_0 \leq d$, as $n \to \infty$ one has $R_n(\mathbb{E}_{P_0}[Z]) \Rightarrow \chi^2_{d_0}$, where $\chi^2_{d_0}$ denotes the $\chi^2$-distribution with $d_0$ degrees of freedom. Then $C_{n,\rho(\alpha)}$ is an asymptotically exact $(1-\alpha)$-confidence interval for $T(P_0) = \mathbb{E}_{P_0}[Z]$ if we set $\rho(\alpha) = \inf\{\rho' : \mathbb{P}(\chi^2_{d_0} \leq \rho') \geq 1 - \alpha\}$. We extend these results to more general functions $T$ and to a variety of $f$-divergences satisfying the following condition, which we henceforth assume without mention (each of our theorems requires this assumption).

ASSUMPTION 1 (**Smoothness of $f$-divergence**). *The function $f : \mathbb{R}_+ \to \overline{\mathbb{R}}$ is convex, three times differentiable in 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 a normalization for easier calculation. We make no restrictions on the behavior of $f$ at $0$, as a number of divergence measures, such as KL with $f(t) = -2\log t + 2t - 2$, approach infinity as $t \downarrow 0$.

The following proposition is a generalization of Owen's results [62] to smooth $f$-divergences. While the result is essentially known [4, 21, 9], it is also an immediate consequence of our uniform variance expansions to come.

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

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{E}_{P_0}[Z] \in \left\{\mathbb{E}_P[Z] : D_f(P\|\widehat{P}_n) \leq \frac{\rho}{n}\right\}\right) = \mathbb{P}\left(\chi^2_{d_0} \leq \rho\right). \tag{7}$$

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 B.5. 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).

**2.2. Asymptotic Expansions** To obtain inferential guarantees on $T(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x; \xi)]$, we require stronger results than the pointwise guarantee (7). We now develop an asymptotic expansion that essentially gives all of the major distributional convergence results in this paper. Our results on convergence and exact coverage build on two asymptotic expansions, which we now present. In the statement of the lemma, we recall that a sequence $\{Z_i\}$ of random variables is ergodic and stationary if for all bounded functions $f : \mathbb{R}^k \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ we have

$$\lim_{n\to\infty} \mathbb{E}[f(Z_t, \ldots, Z_{t+k-1})g(Z_{t+n}, \ldots, Z_{t+n+m-1})] = \mathbb{E}[f(Z_1, \ldots, Z_k)]\mathbb{E}[g(Z_1, \ldots, Z_m)].$$

We then have the following lemma.

LEMMA 1. *Let $Z_1, Z_2, \ldots$ be a strictly stationary ergodic sequence of random variables with $\mathbb{E}[Z_1^2] < \infty$, and let Assumption 1 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*

$$\left| \sup_{P : D_f(P \| \widehat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[Z] - \mathbb{E}_{\widehat{P}_n}[Z] - \sqrt{\frac{\rho}{n} s_n^2} \right| \leq \frac{\varepsilon_n}{\sqrt{n}} \tag{8}$$

*where $\varepsilon_n \overset{a.s.}{\to} 0$.*

See Appendix A for the proof. For intuition, we may rewrite the expansion (8) as

$$\sup_{P : D_f(P \| \widehat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[Z] = \mathbb{E}_{\widehat{P}_n}[Z] + \sqrt{\frac{\rho}{n} \mathrm{Var}_{\widehat{P}_n}(Z)} + \frac{\varepsilon_n^+}{\sqrt{n}} \tag{9a}$$

$$\inf_{P : D_f(P \| \widehat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[Z] = \mathbb{E}_{\widehat{P}_n}[Z] - \sqrt{\frac{\rho}{n} \mathrm{Var}_{\widehat{P}_n}(Z)} + \frac{\varepsilon_n^-}{\sqrt{n}} \tag{9b}$$

with $\varepsilon_n^{\pm} \overset{a.s.}{\to} 0$, where the second equality follows from symmetry. Applying the classical central limit theorem and Slutsky's lemma, we then obtain

$$\mathbb{P}\left( \sqrt{n} \left| \mathbb{E}_{P_0}[Z] - \mathbb{E}_{\widehat{P}_n}[Z] \right| \leq \sqrt{\rho \mathrm{Var}_{\widehat{P}_n}(Z)} \right) \underset{n \uparrow \infty}{\to} \mathbb{P}(|N(0,1)| \leq \sqrt{\rho}) = \mathbb{P}(\chi_1^2 \leq \rho),$$

yielding Proposition 1 in the case that $d = 1$. Concurrently with the original version of this paper, Lam [50] gives an in-probability version of the result (9) (rather than almost sure) for the case $f(t) = -2 \log t$, corresponding to empirical likelihood. Our proof is new, gives a probability 1 result, and generalizes to ergodic stationary sequences.

Next, we show a uniform variant of the asymptotic expansion (9) which relies on local Lipschitzness of our loss. While our results apply in significantly more generality (see Section 7), the following assumption covers many practical instances of stochastic optimization problems.

ASSUMPTION 2. *The set $\mathcal{X} \subset \mathbb{R}^d$ is compact, and there exists a measurable function $M : \Xi \to \mathbb{R}_+$ such that for all $\xi \in \Xi$, $\ell(\cdot; \xi)$ is $M(\xi)$-Lipschitz with respect to some norm $\|\cdot\|$ on $\mathcal{X}$.*

THEOREM 1. *Let Assumption 2 hold with $\mathbb{E}_{P_0}[M(\xi)^2] < \infty$ and assume that $\mathbb{E}_{P_0}[|\ell(x_0; \xi)|^2] < \infty$ for some $x_0 \in \mathcal{X}$. If $\xi_i \overset{\mathrm{iid}}{\sim} P_0$, then*

$$\sup_{P : D_f(P \| \widehat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[\ell(x, \xi)] = \mathbb{E}_{\widehat{P}_n}[\ell(x, \xi)] + \sqrt{\frac{\rho}{n} \mathrm{Var}_{\widehat{P}_n}(\ell(x, \xi))} + \varepsilon_n(x), \tag{10}$$

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

This theorem is a consequence of the more general uniform expansions we develop in Section 7, in particular Theorem 7. In addition to generalizing classical empirical likelihood theory, these results also allow a novel proof of the classical empirical likelihood result for means (Proposition 1).

**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 and study the mapping

$$T_{\mathrm{opt}} : \mathcal{P} \to \mathbb{R}, \quad P \mapsto T_{\mathrm{opt}}(P) := \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x; \xi)].$$

Although the functional $T_{\mathrm{opt}}(P)$ is nonlinear, we can provide regularity conditions guaranteeing its smoothness (Hadamard differentiability), so that the generalized empirical likelihood approach provides asymptotically exact confidence bounds on $T_{\mathrm{opt}}(P)$. Throughout this section, we make a standard assumption guaranteeing existence of minimizers [e.g. 69, Theorem 1.9].

ASSUMPTION 3. *The function $\ell(\cdot;\xi)$ is proper and lower semi-continuous for $P_0$-almost all $\xi \in \Xi$. Either $\mathcal{X}$ is compact or $x \mapsto \mathbb{E}_{P_0}[\ell(x;\xi)]$ is coercive, meaning $\mathbb{E}_{P_0}[\ell(x;\xi)] \to \infty$ as $\|x\| \to \infty$.*

In the remainder of this section, we explore the generalized empirical likelihood approach to confidence intervals on the optimal value for both i.i.d. data and dependent sequences (Sections 3.1 and 3.1, respectively), returning in Section 3.3 to discuss a few computational issues, examples, and generalizations.

**3.1. Generalized Empirical Likelihood for Stochastic Optimization** The first result we present applies in the case that the data is i.i.d.

THEOREM 2. *Let Assumptions 1, 2 hold with $\mathbb{E}_{P_0}[M(\xi)^2] < \infty$ and $\mathbb{E}_{P_0}[|\ell(x_0;\xi)|^2] < \infty$ for some $x_0 \in \mathcal{X}$. If $\xi_i \overset{\text{iid}}{\sim} P_0$ and the optimizer $x^\star := \arg\min_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$ is unique, then*

$$\lim_{n \to \infty} \mathbb{P}\left(T_{\text{opt}}(P_0) \in \left\{ T_{\text{opt}}(P) : D_f\left(P\|\widehat{P}_n\right) \le \frac{\rho}{n} \right\}\right) = \mathbb{P}\left(\chi_1^2 \le \rho\right).$$

This result follows from a combination of two steps: the generalized empirical likelihood theory for smooth functionals we give in Section 7, and a proof that the conditions of the theorem are sufficient to guarantee smoothness of $T_{\text{opt}}$. See Appendix C for the full derivation.

Setting $\mathcal{X} = \mathbb{R}^d$, meaning that the problem is unconstrained, and assuming that the loss $x \mapsto \ell(x;\xi)$ is differentiable for all $\xi \in \Xi$, Lam and Zhou [51] give a similar (but different) result to Theorem 2 for the special case that $f(t) = -2\log t$, which is the classical empirical likelihood setting. They use first order optimality conditions as an estimating equation and apply standard empirical likelihood theory [64]. This approach gives a non-pivotal asymptotic distribution; the limiting law is a $\chi_r^2$-distribution with $r = \text{rank}(\text{Cov}_{P_0}(\nabla \ell(x^\star;\xi))$ degrees of freedom, though $x^\star$ need not be unique in this approach. The resulting confidence intervals are too conservative and fail to have (asymptotically) exact coverage. The estimating equations approach also requires the loss $\ell(\cdot;\xi)$ to be differentiable and the covariance matrix of $(\ell(x^\star;\xi), \nabla_x \ell(x^\star;\xi))$ to be positive definite for some $x^* \in \arg\min_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$. In contrast, Theorem 2 applies to problems with general constraints, as well as more general objective functions $\ell$ and $f$-divergences, by leveraging smoothness properties (over the space of probability measures) of the functional $T_{\text{opt}}(P) := \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$. A consequence of the more general losses, divergences, and exact coverage is that the theorem requires the minimizer of $\mathbb{E}_{P_0}[\ell(x;\xi)]$ to be unique.

Shapiro [74, 76] develops a number of normal approximations and asymptotic normality theory for stochastic optimization problems. The normal analogue of Theorem 2 is that

$$\sqrt{n}\left(\inf_{x \in \mathcal{X}} \mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)] - \inf_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]\right) \Rightarrow N\left(0, \text{Var}_{P_0}(\ell(x^\star;\xi))\right) \tag{11}$$

which holds under the conditions of Theorem 2. The normal approximation (11) depends on the unknown parameter $\text{Var}_{P_0}(\ell(x^\star;\xi))$ and is not asymptotically pivotal. The generalized empirical likelihood approach, however, is pivotal, meaning that there are no hidden quantities we must estimate; generally speaking, the normal approximation (11) requires estimation of $\text{Var}_{P_0}(\ell(x^\star;\xi))$, for which one usually uses $\text{Var}_{\widehat{P}_n}(\ell(\widehat{x}_n;\xi))$ where $\widehat{x}_n$ minimizes the sample average (2).

When the optimum is not unique, we can still provide an exact asymptotic characterization of the limiting probabilities that $l_n \le T_{\text{opt}}(P_0) \le u_n$, where we recall the definitions (4) of $l_n = \inf_P\{T_{\text{opt}}(P) : D_f(P\|\widehat{P}_n) \le \rho/n\}$ and $u_n = \sup_P\{T_{\text{opt}}(P) : D_f(P\|\widehat{P}_n) \le \rho/n\}$, which also shows a useful symmetry between the upper and lower bounds. The characterization depends on the excursions of a non-centered Gaussian process when $x^\star$ is non-unique, which unfortunately makes it hard to evaluate. To state the result, we require the definition of a few additional processes. Let $G$ be the mean-zero Gaussian process with covariance

$$\text{Cov}(x_1, x_2) = \mathbb{E}[G(x_1)G(x_2)] = \text{Cov}(\ell(x_1;\xi), \ell(x_2;\xi))$$

for $x_1, x_2 \in \mathcal{X}$, and define the non-centered processes $H_+$ and $H_-$ by

$$H_+(x) := G(x) + \sqrt{\rho \mathrm{Var}_{P_0}(\ell(x;\xi))} \quad \text{and} \quad H_-(x) := G(x) - \sqrt{\rho \mathrm{Var}_{P_0}(\ell(x;\xi))}. \tag{12}$$

Letting $S_{P_0}^\star := \arg\min_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$ be the set of optimal solutions for the population problem (1), we obtain the following theorem. (It is possible to extend this result to mixing sequences, but we focus for simplicity on the i.i.d. case.)

THEOREM 3. *Let Assumptions 1, 2, and 3 hold, where the Lipschitz constant $M$ satisfies $\mathbb{E}_{P_0}[M(\xi)^2] < \infty$. Assume there exists $x_0 \in \mathcal{X}$ such that $\mathbb{E}_{P_0}[|\ell(x_0;\xi)|^2] < \infty$. If $\xi_i \overset{\text{iid}}{\sim} P_0$, then*

$$\lim_{n \to \infty} \mathbb{P}\left( \inf_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \leq u_n \right) = \mathbb{P}\left( \inf_{x \in S_{P_0}^\star} H_+(x) \geq 0 \right)$$

*and*

$$\lim_{n \to \infty} \mathbb{P}\left( \inf_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \geq l_n \right) = \mathbb{P}\left( \inf_{x \in S_{P_0}^\star} H_-(x) \leq 0 \right).$$

*If $S_{P_0}^\star$ is a singleton, both limits are equal to $1 - \frac{1}{2} P\left(\chi_1^2 \geq \rho\right)$.*

We defer the proof of the theorem to Appendix C.3, noting that it is essentially an immediate consequence of the uniform results in Section 7 (in particular, the uniform variance expansion of Theorem 7 and the Hadamard differentiability result of Theorem 8).

Theorem 3 provides us with a few benefits. First, if all one requires is a one-sided confidence interval (say an upper interval), we may shorten the confidence set via a simple correction to the threshold $\rho$. Indeed, for a given desired confidence level $1 - \alpha$, setting $\rho = \chi_{1,1-2\alpha}^2$ (which is smaller than $\chi_{1,1-\alpha}^2$) gives a one-sided confidence interval $(-\infty, u_n]$ with asymptotic coverage $1 - \alpha$.

**3.2. Extensions to Dependent Sequences** We now give an extension of Theorem 2 to dependent sequences, including Harris recurrent Markov chains mixing suitably quickly. To present our results, we first recall $\beta$-mixing sequences [16, 34, Chs. 7.2–3] (also called absolute regularity in the literature).

DEFINITION 1. The $\beta$-mixing coefficient between two sigma algebras $\mathcal{B}_1$ and $\mathcal{B}_2$ on $\Xi$ is

$$\beta(\mathcal{B}_1, \mathcal{B}_2) = \frac{1}{2} \sup \sum_{\mathcal{I} \times \mathcal{J}} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|$$

where the supremum is over finite partitions $\{A_i\}_{i \in \mathcal{I}}$, $\{B_j\}_{j \in \mathcal{J}}$ of $\Xi$ such that $A_i \in \mathcal{B}_1$ and $B_j \in \mathcal{B}_2$.

Let $\{\xi\}_{i \in \mathbb{Z}}$ be a sequence of strictly stationary random vectors. Given the $\sigma$-algebras

$$\mathcal{G}_0 := \sigma(\xi_i : i \leq 0) \quad \text{and} \quad \mathcal{G}_n := \sigma(\xi_i : i \geq n) \quad \text{for} \quad n \in \mathbb{N},$$

the $\beta$-mixing coefficients of $\{\xi_i\}_{i \in \mathbb{Z}}$ are defined via Definition 1 by

$$\beta_n := \beta(\mathcal{G}_0, \mathcal{G}_n). \tag{13}$$

A stationary sequence $\{\xi_i\}_{i \in \mathbb{Z}}$ is $\beta$-mixing if $\beta_n \to 0$ as $n \to \infty$. For Markov chains, $\beta$-mixing has a particularly nice interpretation: a strictly stationary Markov chain is $\beta$-mixing if and only if it is Harris recurrent and aperiodic [16, Thm. 3.5].

With these preliminaries, we may state our asymptotic convergence result, which is based on a uniform central limit theorem that requires fast enough mixing [29].

THEOREM 4.    *Let $\{\xi_n\}_{n=0}^{\infty}$ be an aperiodic, positive Harris recurrent Markov chain taking values on $\Xi$ with stationary distribution $\pi$. Let Assumptions 1 and 2 hold and assume that there exists $r > 1$ and $x_0 \in \mathcal{X}$ satisfying $\sum_{n=1}^{\infty} n^{\frac{1}{r-1}}\beta_n < \infty$, the Lipschitz moment bound $\mathbb{E}_\pi[|M(\xi)|^{2r}] < \infty$, and $\mathbb{E}_\pi[|\ell(x_0;\xi)|^{2r}] < \infty$. If the optimizer $x^\star := \arg\min_{x \in \mathcal{X}} \mathbb{E}_\pi[\ell(x;\xi)]$ is unique then for any $\xi_0 \sim \nu$*

$$\lim_{n \to \infty} \mathbb{P}_\nu \left( T_{\mathrm{opt}}(\pi) \in \left\{ T_{\mathrm{opt}}(P) : D_f \left( P \| \widehat{P}_n \right) \leq \frac{\rho}{n} \right\} \right) = \mathbb{P}\left( \chi_1^2 \leq \frac{\rho \mathrm{Var}_\pi \ell(x^\star;\xi)}{\sigma_\pi^2(x^\star)} \right) \qquad (14)$$

*where $\sigma_\pi^2(x^\star) = \mathrm{Var}_\pi \ell(x^\star;\xi) + 2\sum_{n=1}^{\infty} \mathrm{Cov}_\pi(\ell(x^\star;\xi_0), \ell(x^\star;\xi_n))$.*

Theorem 4 is more or less a consequence of the general results we prove in Section 7.3 on ergodic sequences, and we show how it follows from these results in Appendix D.3.

We give a few examples of Markov chains satisfying the mixing condition $\sum_{n=1}^{\infty} n^{\frac{1}{r-1}}\beta_n < \infty$ for some $r > 1$.

EXAMPLE 1 (UNIFORM ERGODICITY).    If an aperiodic, positive Harris recurrent Markov chain is uniformly ergodic then it is geometrically $\beta$-mixing [54, Theorem 16.0.2], meaning that $\beta_n = O(c^n)$ for some constant $c \in (0,1)$ In this case, the Lipschitzian assumption in Theorem 4 holds whenever $\mathbb{E}_\pi[M(\xi)^2 \log_+ M(\xi)] < \infty$.

As our next example, we consider geometrically $\beta$-mixing processses that are not necessarily uniformly mixing. The following result is due to Mokkadem [55].

EXAMPLE 2 (GEOMETRIC $\beta$-MIXING).    Let $\Xi = \mathbb{R}^p$ and consider the affine auto-regressive process

$$\xi_{n+1} = A(\epsilon_{n+1})\xi_n + b(\epsilon_{n+1})$$

where $A$ is a polynomial $p \times p$ matrix-valued function and $b$ is a $\mathbb{R}^p$-valued polynomial function. We assume that the noise sequence $\{\epsilon_n\}_{n \geq 1} \overset{\mathrm{iid}}{\sim} F$ where $F$ has a density with respect to the Lebesgue measure. If ($i$) eigenvalues of $A(0)$ are inside the open unit disk and ($ii$) there exists $a > 0$ such that $\mathbb{E}\|A(\epsilon_n)\|^a + \mathbb{E}\|b(\epsilon_n)\|^a < \infty$, then $\{\xi_n\}_{n \geq 0}$ is geometrically $\beta$-mixing. That is, there exists $c \in (0,1)$ such that $\beta_n = O(s^n)$.

See Doukhan [28, Section 2.4.1] for more examples of $\beta$-mixing processes.

Using the equivalence of geometric $\beta$-mixing and geometric ergodicity for Markov chains [61, 54, Chapter 15], we can give a Lyapunov criterion.

EXAMPLE 3 (LYAPUNOV CRITERION).    Let $\{\xi_n\}_{n \geq 0}$ be an aperiodic Markov chain. For shorthand, denote the regular conditional distribution of $\xi_m$ given $\xi_0 = z$ by $P^m(z,\cdot) := \mathbb{P}_z(\xi_m \in \cdot) = \mathbb{P}(\xi_m \in \cdot | \xi_0 = z)$. Assume that there exists a measurable set $C \in \mathcal{A}$, a probability measure $\nu$ on $(\Xi, \mathcal{A})$, a potential function $w : \Xi \to [1,\infty)$, and constants $m \geq 1, \lambda > 0, \gamma \in (0,1)$ such that (i) $P^m(z,B) \geq \lambda\nu(B)$ for all $z \in C, B \in \mathcal{A}$, (ii) $\mathbb{E}_z w(\xi_1) \leq \gamma w(z)$ for all $z \in C^c$, and (iii) $\sup_{z \in C} \mathbb{E}_z w(\xi_1) < \infty$. (The set $C$ is a *small set* [54, Chapter 5.2].) Then $\{\xi_n\}_{n \geq 0}$ is aperiodic, positive Harris recurrent, and geometrically ergodic [54, Theorem 15.0.1]. Further, we can show that $\{\xi_n\}_{n \geq 0}$ is geometrically $\beta$-mixing: there exists $c \in (0,1)$ with $\beta_n = O(c^n)$. For completeness, we include a proof of this in Appendix D.1.

For dependent sequences, the asymptotic distribution in the limit (14) contains unknown terms such as $\sigma_\pi^2$ and $\mathrm{Var}_\pi(\ell(x^\star;\xi))$; such quantities need to be estimated to obtain exact coverage. This loss of a pivotal limit occurs because $\sqrt{n}(\widehat{P}_n - P_0)$ converges to a Gaussian process $G$ on $\mathcal{X}$ with covariance function

$$\mathrm{Cov}(x_1, x_2) := \mathrm{Cov}(G(x_1), G(x_2)) = \sum_{n \geq 1} \mathrm{Cov}_\pi(\ell(x_1;\xi_0), \ell(x_2;\xi_n)),$$

while empirical likelihood self-normalizes based on $\mathrm{Cov}_\pi\left(\ell(x_1;\xi_0),\ell(x_2;\xi_0)\right)$. (These covariances are identical if $\xi_i$ are i.i.d.) As a result, in Theorem 4, we no longer have the self-normalizing behavior of Theorem 2 for i.i.d. sequences. To remedy this, we now give a sectioning method that yields pivotal asymptotics, even for dependent sequences.

Let $m \in \mathbb{N}$ be a fixed integer. Letting $b := \lfloor n/m \rfloor$, partition the $n$ samples into $m$ sections

$$\{\xi_1,\ldots,\xi_b\},\ \{\xi_{b+1},\ldots,\xi_{2b}\},\ \cdots,\ \{\xi_{(m-1)b+1},\ldots,\xi_{mb}\}$$

and denote by $\widehat{P}_b^j$ the empirical distribution on each of the blocks for $j = 1,\ldots,m$. Let

$$U_b^i := \sup_{P \ll \widehat{P}_b^j} \left\{ T_{\mathrm{opt}}(P) : D_f\left(P\|\widehat{P}_b^j\right) \le \frac{\rho}{n} \right\}$$

and define

$$\overline{U}_b := \frac{1}{m}\sum_{j=1}^m U_b^j \quad \text{and} \quad s_m^2(U_b) := \frac{1}{m}\sum_{j=1}^m \left(U_b^j - \overline{U}_b\right)^2.$$

Letting $\widehat{x}_n^* \in \arg\min_{x \in \mathcal{X}} \mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)]$, we obtain the following result.

PROPOSITION 2. *Under the conditions of Theorem 4, for any initial distribution $\xi_0 \sim \nu$*

$$\lim_{n\to\infty} \mathbb{P}_\nu\left( T_{\mathrm{opt}}(\pi) \le \overline{U}_b - \sqrt{\frac{\rho}{b}\mathrm{Var}_{\widehat{P}_n}\ell(\widehat{x}_n^*;\xi)} + s_m(U_b)t \right) = \mathbb{P}(T_{m-1} \ge -t)$$

*where $T_{m-1}$ is the Student-t distribution with $(m-1)$-degress of freedom.*

See Section D.4 for the proof of Proposition 2. Thus, we recover an estimable quantity guaranteeing an exact confidence limit.

**3.3. Computing the Confidence Interval and its Properties** For convex objectives, we can provide efficient procedures for computing our desired confidence intervals on the optimal value $T_{\mathrm{opt}}(P_0)$. We begin by making the following assumption.

ASSUMPTION 4. *The set $\mathcal{X} \subset \mathbb{R}^d$ is convex and $\ell(\cdot;\xi):\mathcal{X} \to \mathbb{R}$ is a proper closed convex function for $P_0$-almost all $\xi \in \Xi$.*

For $P$ finitely supported on $n$ points, the functional $P \mapsto T_{\mathrm{opt}}(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ is continuous (on $\mathbb{R}^n$) because it is concave and finite-valued; as a consequence, the set

$$\left\{ T_{\mathrm{opt}}(P) : D_f(P\|\widehat{P}_n) \le \rho/n \right\} = \left\{ \inf_{x \in \mathcal{X}} \sum_{i=1}^n p_i \ell(x;\xi_i) : p^\top \mathbb{1} = 1,\ p \ge 0,\ \sum_{i=1}^n f(np_i) \le \rho \right\}, \qquad (15)$$

is an interval, and in this section we discuss a few methods for computing it. To compute the interval (15), we solve for the two endpoints $u_n$ and $l_n$ of expressions (4a)–(4b).

The upper bound is computable using convex optimization methods under Assumption 4, which follows from the coming results. The first is a minimax theorem [38, Theorem VII.4.3.1].

LEMMA 2. *Let Assumptions 3 and 4 hold. Then*

$$u_n = \inf_{x \in \mathcal{X}} \sup_{p \in \mathbb{R}^n} \left\{ \sum_{i=1}^n p_i \ell(x;\xi_i) : p^\top \mathbb{1} = 1,\ p \ge 0,\ \sum_{i=1}^n f(np_i) \le \rho \right\}. \qquad (16)$$

By strong duality, we can write the minimax problem (16) as a joint minimization problem.

LEMMA 3 **(Ben-Tal et al. [6])**. *The following duality holds:*

$$\sup_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(x;\xi)] : D_f\left(P\|\widehat{P}_n\right) \leq \frac{\rho}{n} \right\} = \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ \mathbb{E}_{\widehat{P}_n}\left[\lambda f^*\left(\frac{\ell(x;\xi) - \eta}{\lambda}\right)\right] + \frac{\rho}{n}\lambda + \eta \right\}. \quad (17)$$

When $x \mapsto \ell(x;\xi)$ is convex in $x$, the minimization (16) is a convex problem because it is the supremum of convex functions. The reformulation (17) shows that we can compute $u_n$ by solving a problem jointly convex in $\eta$, $\lambda$, and $x$.

Finding the lower confidence bound (4b) is in general not a convex problem even when the loss $\ell(\cdot;\xi)$ is convex in its first argument. With that said, the conditions of Theorem 3, coupled with convexity, allow us to give an efficient two-step minimization procedure that yields an estimated lower confidence bound $\widehat{l}_n$ that achieves the asymptotic pivotal behavior of $l_n$ whenever the population optimizer for problem (1) is unique. Indeed, let us assume the conditions of Theorem 3 and Assumption 4, additionally assuming that the set $S^\star_{P_0}$ is a singleton. Then standard consistency results [78, Chapter 5] guarantee that under our conditions, the empirical minimizer $\widehat{x}_n = \arg\min_{x \in \mathcal{X}} \mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)]$ satisfies $\widehat{x}_n \stackrel{a.s.}{\to} x^\star$, where $x^\star = \arg\min_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$. Now, consider the one-step estimator

$$\widehat{l}_n := \inf_{P:D_f(P\|\widehat{P}_n) \leq \rho/n} \mathbb{E}_P[\ell(\widehat{x}_n;\xi)]. \quad (18)$$

Then by Theorem 1, we have

$$\widehat{l}_n = \frac{1}{n}\sum_{i=1}^n \ell(\widehat{x}_n;\xi_i) - \sqrt{\frac{\rho}{n}\mathrm{Var}_{\widehat{P}_n}(\ell(\widehat{x}_n;\xi))} + o_{P_0}(n^{-\frac{1}{2}})$$

because $\widehat{x}_n$ is eventually in any open set (or set open relative to $\mathcal{X}$) containing $x^\star$. Standard limit results [82] guarantee that $\mathrm{Var}_{\widehat{P}_n}(\ell(\widehat{x}_n;\xi)) \stackrel{a.s.}{\to} \mathrm{Var}_{P_0}(\ell(x^\star;\xi))$, because $x \mapsto \ell(x;\xi)$ is Lipschitzian by Assumption 2. Noting that $\mathbb{E}_{\widehat{P}_n}[\ell(\widehat{x}_n;\xi)] \leq \mathbb{E}_{\widehat{P}_n}[\ell(x^\star;\xi)]$, we thus obtain

$$\inf_{P:D_f(P\|\widehat{P}_n) \leq \rho/n} \mathbb{E}_P[\ell(\widehat{x}_n;\xi)] \leq \mathbb{E}_{\widehat{P}_n}[\ell(x^\star;\xi)] - \sqrt{\frac{\rho}{n}\mathrm{Var}_{P_0}(\ell(x^\star;\xi))} + o_{P_0}(n^{-\frac{1}{2}})$$

Defining $\sigma^2(x^\star) = \mathrm{Var}_{P_0}(\ell(x^\star;\xi))$ for notational convenience and rescaling by $\sqrt{n}$, we have

$$\sqrt{n}\left(\mathbb{E}_{\widehat{P}_n}[\ell(x^\star;\xi)] - \mathbb{E}_{P_0}[\ell(x^\star;\xi)] - \sqrt{\frac{\rho}{n}\sigma^2(x^\star)} + o_{P_0}(n^{-\frac{1}{2}})\right) \Rightarrow \mathsf{N}\left(-\sqrt{\rho\sigma^2(x^\star)}, \sigma^2(x^\star)\right).$$

Combining these results, we have that that $\sqrt{n}(l_n - \mathbb{E}_{P_0}[\ell(x^\star;\xi)]) \Rightarrow \mathsf{N}(-\sqrt{\rho\sigma^2(x^\star)}, \sigma^2(x^\star))$ (looking forward to Theorem 7 and using Theorem 2), and

$$l_n \leq \widehat{l}_n \leq \mathbb{E}_{\widehat{P}_n}[\ell(x^\star;\xi)] - \sqrt{\frac{\rho}{n}\sigma^2(x^\star)} + o_{P_0}(n^{-\frac{1}{2}}).$$

Summarizing, the one-step estimator (18) is upper and lower bounded by quantities that, when shifted by $-\mathbb{E}_{P_0}[\ell(x^\star;\xi)]$ and rescaled by $\sqrt{n}$, are asymptotically $\mathsf{N}(-\sqrt{\rho\sigma^2(x^\star)}, \sigma^2(x^\star))$. Thus under the conditions of Theorem 2 and Assumption 2, the one-step estimator $\widehat{l}_n$ defined by expression (18) guarantees that

$$\lim_{n \to \infty} \mathbb{P}\left(\widehat{l}_n \leq \mathbb{E}_{P_0}[\ell(x^\star;\xi)] \leq u_n\right) = \mathbb{P}\left(\chi_1^2 \leq \rho\right),$$

giving a computationally feasible and asymptotically pivotal statistic. We remark in passing that alternating by minimizing over $P : D_f(P\|\widehat{P}_n) \leq \rho/n$ and $x$ (i.e. more than the single-step minimizer) simply gives a lower bound $\widetilde{l}_n$ satisfying $l_n \leq \widetilde{l}_n \leq \widehat{l}_n$, which will evidently have the same convergence properties.

**4. Connections to Robust Optimization and Examples**    To this point, we have studied the statistical properties of generalized empirical likelihood estimators, with particular application to estimating the population objective $\inf_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$. We now make connections between our approach of minimizing worst-case risk over $f$-divergence balls and approaches from the robust optimization and risk minimization literatures. We first relate our approach to classical work on coherent risk measures for optimization problems, after which we briefly discuss regularization properties of the formulation.

**4.1. Upper Confidence Bounds as a Risk Measure**    Sample average approximation is optimistic [78], because $\inf_{x \in \mathcal{X}} \mathbb{E}[\ell(x;\xi)] \geq \mathbb{E}[\inf_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} \ell(x;\xi_i)]$. The robust formulation (4a) addresses this optimism by looking at a worst case objective based on the confidence region $\{P : D_f(P \| \widehat{P}_n) \leq \rho/n\}$. It is clear that the robust formulation (4a) is a coherent risk measure [78, Ch. 6.3] of $\ell(x;\xi)$: it is convex, monotonic in the loss $\ell$, equivariant to translations $\ell \mapsto \ell + a$, and positively homogeneous in $\ell$. A number of authors have studied coherent risk measures measures [3, 68, 48, 78], and we study their connections to statistical confidence regions for the optimal population objective (1) below.

As a concrete example, we consider Krokhmal's higher-order generalizations [48] of conditional value at risk, where for $k_* \geq 1$ and a constant $c > 0$ the risk functional has the form

$$R_{k_*}(x; P) := \inf_{\eta \in \mathbb{R}} \left\{ (1 + c)\mathbb{E}_P \left[ (\ell(x;\xi) - \eta)_+^{k_*} \right]^{\frac{1}{k_*}} + \eta \right\}.$$

The risk $R_{k_*}$ penalizes upward deviations of the objective $\ell(x;\xi)$ from a fixed value $\eta$, where the parameter $k_* \geq 1$ determines the degree of penalization (so $k_* \uparrow \infty$ implies substantial penalties for upward deviations). These risk measures capture a natural type of risk aversion [48].

We can recover such formulations, thus providing asymptotic guarantees for their empirical minimizers, via the robust formulation (4a). To see this, we consider the classical Cressie-Read family [22] of $f$-divergences. Recalling that $f^*$ denotes the Fenchel conjugate $f^*(s) := \sup_t \{st - f(t)\}$, for $k \in (-\infty, \infty)$ with $k \notin \{0, 1\}$, one defines

$$f_k(t) = \frac{t^k - kt + k - 1}{2k(k-1)} \quad \text{so} \quad f_k^*(s) = \frac{2}{k} \left[ \left( \frac{k-1}{2}s + 1 \right)_+^{k_*} - 1 \right] \tag{19}$$

where we define $f_k(t) = +\infty$ for $t < 0$, and $k_*$ is given by $1/k + 1/k_* = 1$. We define $f_1$ and $f_0$ as their respective limits as $k \to 0, 1$. (We provide the dual calculation $f_k^*$ in the proof of Lemma 4.) The family (19) includes divergences such as the $\chi^2$-divergence ($k = 2$), empirical likelihood $f_0(t) = -2\log t + 2t - 2$, and KL-divergence $f_1(t) = 2t\log t - 2t + 2$. All such $f_k$ satisfy Assumption 1.

For the Cressie-Read family, we may compute the dual (17) more carefully by infimizing over $\lambda \geq 0$, which yields the following duality result. As the lemma is a straightforward consequence of Lemma 3, we defer its proof to Appendix C.4.

LEMMA 4.    *Let $k \in (1, \infty)$ and define $\mathcal{P}_n := \{P : D_{f_k}(P \| \widehat{P}_n) \leq \rho/n\}$. Then*

$$\sup_{P \in \mathcal{P}_n} \mathbb{E}_P[\ell(x;\xi)] = \inf_{\eta \in \mathbb{R}} \left\{ \left( 1 + \frac{k(k-1)\rho}{2n} \right)^{\frac{1}{k}} \mathbb{E}_{\widehat{P}_n} \left[ (\ell(x;\xi) - \eta)_+^{k_*} \right]^{\frac{1}{k_*}} + \eta \right\} \tag{20}$$

The lemma shows that we indeed recover a variant of the risk $R_{k_*}$, where taking $\rho \uparrow \infty$ and $k \downarrow 1$ (so that $k_* \uparrow \infty$) increases robustness—penalties for upward deviations of the loss $\ell(x;\xi)$—in a natural way. The confidence guarantees of Theorem 3, on the other hand, show how (to within first order) the asymptotic behavior of the risk (20) depends only on $\rho$, as each value of $k$ allows upper confidence bounds on the optimal population objective (1) with asymptotically exact coverage.

**4.2. Variance Regularization** We now consider the asymptotic variance expansions of Theorem 1, which is that

$$\sup_{P:D_f(P\|P_n)\le\frac{\rho}{n}} \mathbb{E}_P[\ell(x;\xi)] = \mathbb{E}_{P_n}[\ell(x;\xi)] + \sqrt{\frac{\rho}{n}\mathrm{Var}_{P_n}(\ell(x;\xi))} + \varepsilon_n(x) \tag{21}$$

where $\sqrt{n}\sup_{x\in\mathcal{X}}|\varepsilon_n(x)| \overset{P^*}{\to} 0$. In a companion to this paper, Duchi and Namkoong [30, 57] explore the expansion (21) in substantial depth for the special case $f(t) = \frac{1}{2}(t-1)^2$. Eq. (21) shows that in an asymptotic sense, we expect similar results to theirs to extend to general $f$-divergences, and we discuss this idea briefly.

The expansion (21) shows that the robust formulation (4a) ameliorates the optimism bias of standard sample average approximation by penalizing the variance of the loss. Researchers have investigated connections between regularization and robustness, including Xu et al. [85] for standard supervised machine learning tasks (see also [5, Chapter 12]), though these results consider uncertainty on the data vectors $\xi$ themselves, rather than the distribution. Our approach yields a qualitatively different type of (approximate) regularization by variance. In our follow-up work [30, 57], we analyze finite-sample performance of the robust solutions. The naive variance-regularized objective

$$\mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)] + \sqrt{\frac{\rho}{n}\mathrm{Var}_{\widehat{P}_n}\ell(x;\xi)} \tag{22}$$

is neither convex (in general) nor coherent, so that the expansion (21) allows us to solve a convex optimization problem that approximately regularizes variance.

In some restricted situations, the variance-penalized objective (22) is convex—namely, when $\ell(x;\xi)$ is linear in $x$. A classical example of this is the sample version of the Markowitz portfolio problem [53].

EXAMPLE 4 (PORTFOLIO OPTIMIZATION). Let $x \in \mathbb{R}^d$ denote investment allocations and $\xi \in \mathbb{R}^d$ returns on investiment, and consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\mathrm{maximize}}\ \mathbb{E}_{P_0}\left[\xi^\top x\right] \quad \text{subject to}\quad x^\top \mathbb{1} = 1, x \in [a,b]^d.$$

Given a sample $\{\xi_1,\ldots,\xi_n\}$ of returns, we define $\mu_n := \mathbb{E}_{\widehat{P}_n}[\xi]$ and $\Sigma_n := \mathrm{Cov}_{\widehat{P}_n}(\xi)$ to be the sample mean and covariance. Then the Lagrangian form of the Markowitz problem is to solve

$$\underset{x\in\mathbb{R}^d}{\mathrm{maximize}}\ \mu_n^\top x - \sqrt{\frac{\rho}{n}x^\top\Sigma_n x} \quad \text{subject to}\quad x^\top \mathbb{1} = 1, x \in [a,b]^d.$$

The robust approximation of Theorem 7 (and Eq. (21)) shows that

$$\inf\left\{\mathbb{E}_P[\xi^\top x]\ :\ D_f(P\|\widehat{P}_n) \le \rho/n\right\} = \mu_n^\top x - \sqrt{\frac{\rho}{n}x^\top\Sigma_n x} + o_p(n^{-\frac{1}{2}}),$$

so that distributionally robust formulation approximates the Markowitz objective to $o_p(n^{-\frac{1}{2}})$. There are minor differences, however, in that the Markowitz problem penalizes both upward deviations (via the variance) as well as the downside counterpart. The robust formulation, on the other hand, penalizes downside risk only and is a coherent risk measure.

**5. Consistency** In addition to the inferential guarantees—confidence intervals and variance expansions—we have thus far discussed, we can also give a number of guarantees on the asymptotic consistency of minimizers of the robust upper bound (4a). We show that robust solutions are consistent under (essentially) the same conditions required for that of sample average approximation, which are more general than that required for the uniform variance expansions of Theorem 1. We show this in two ways: first, by considering uniform convergence of the robust objective (4a) to the population risk $\mathbb{E}_{P_0}[\ell(x;\xi)]$ over compacta (Section 5.1), and second by leveraging epigraphical convergence [69] to allow unbounded feasible region $\mathcal{X}$ when $\ell(\cdot;\xi)$ is convex (Section 5.2). In the latter case, we require no assumptions on the magnitude of the noise in estimating $\mathbb{E}_{P_0}[\ell(x;\xi)]$ as a function of $x \in \mathcal{X}$; convexity forces the objective to be large far from the minimizers, so the noise cannot create minimizers far from the solution set.

Bertsimas et al. [11] also provide consistency results for robust variants of sample average approximation based on goodness-of-fit tests, though they require a number of conditions on the domain $\Xi$ of the random variables for their results (in addition to certain continuity conditions on $\ell$). In our context, we abstract away from this by parameterizing our problems by the $n$-vectors $\{P : D_f(P\|\widehat{P}_n) \le \rho/n\}$ and give more direct consistency results that generalize to mixing sequences.

**5.1. Uniform Convergence** For our first set of consistency results, we focus on uniform convergence of the robust objective to the population (1). We begin by recapitulating a few standard statistical results abstractly. Let $\mathcal{H}$ be a collection of functions $h : \Xi \to \mathbb{R}$. We have the following definition on uniform strong laws of large numbers.

DEFINITION 2. A collection of functions $\mathcal{H}$, $h : \Xi \to \mathbb{R}$ for $h \in \mathcal{H}$, is *Glivenko-Cantelli* if

$$\sup_{h \in \mathcal{H}} \left| \mathbb{E}_{\widehat{P}_n}[h] - \mathbb{E}_{P_0}[h] \right| \overset{a.s.,*}{\to} 0.$$

There are many conditions sufficient to guarantee Glivenko-Cantelli properties. Typical approaches include covering number bounds on $\mathcal{H}$ [82, Chapter 2.4]; for example, Lipschitz functions form a Glivenko-Cantelli class, as do continuous functions that are uniformly dominated by an integrable function in the next example.

EXAMPLE 5 (POINTWISE COMPACT CLASS [81], EXAMPLE 19.8). Let $\mathcal{X}$ be compact and $\ell(\cdot;\xi)$ be continuous in $x$ for $P_0$-almost all $\xi \in \Xi$. Then $\mathcal{H} = \{\ell(x;\cdot) : x \in \mathcal{X}\}$ is Glivenko-Cantelli if there exists a measurable envelope function $Z : \Xi \to \mathbb{R}_+$ such that $|\ell(x;\xi)| \le Z(\xi)$ for all $x \in \mathcal{X}$ and $\mathbb{E}_{P_0}[Z] < \infty$.

If $\mathcal{H}$ is Glivenko-Cantelli for $\xi \overset{\text{iid}}{\sim} P_0$, then it is Glivenko-Cantelli for $\beta$-mixing sequences [60] (those with coefficients (13) $\beta_n \to 0$). Our subsequent results thus apply to $\beta$-mixing sequences $\{\xi_i\}$.

If there is an envelope function for objective $\ell(x;\xi)$ that has more than one moment under $P_0$, we can show that the robust risk converges uniformly to the population risk (compared to just the first moment for SAA).

ASSUMPTION 5. *There exists $Z : \Xi \to \mathbb{R}_+$ with $|\ell(x;\xi)| \le Z(\xi)$ for all $x \in \mathcal{X}$ and $\epsilon > 0$ such that* $\mathbb{E}_{P_0}[Z(\xi)^{1+\epsilon}] < \infty$.

Under this assumption, we have the following theorem.

THEOREM 5. *Let Assumptions 1 and 5 hold, and assume the class $\{\ell(x;\cdot) : x \in \mathcal{X}\}$ is Glivenko-Cantelli. Then*

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

See Appendix E.1 for a proof of the result.

When uniform convergence holds, the consistency of robust solutions follows. As in the previous section, we define the sets of optima

$$
S_{P_0}^\star := \underset{x \in \mathcal{X}}{\arg\min}\, \mathbb{E}_{P_0}[\ell(x;\xi)] \quad \text{and} \quad S_{\widehat{P}_n}^\star := \underset{x \in \mathcal{X}}{\arg\min}\, \sup_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(x;\xi)] : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n} \right\}. \tag{23}
$$

Then we immediately attain the following corollary to Theorem 5. In the corollary, we recall the definition of the inclusion distance, or deviation, between sets (6).

COROLLARY 1. *Let Assumptions 1 and 5 hold, let $\mathcal{X}$ be compact, and assume $\ell(\cdot;\xi)$ is continuous on $\mathcal{X}$. Then*

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

*and $d_\subset(S_{\widehat{P}_n}^\star, S_{P_0}^\star) \overset{P^*}{\to} 0$.*

*Proof.* The first conclusion is immediate by Theorem 5 and Example 5. To show convergence of the optimal sets, we denote by $A^\epsilon = \{x : \mathrm{dist}(x, A) \le \epsilon\}$ the $\epsilon$-enlargement of $A$. By the uniform convergence given in Theorem 5 and continuous mapping theorem [82, Theorem 1.3.6], for all $\epsilon > 0$

$$
\begin{aligned}
\limsup_{n \to \infty} \mathbb{P}^* \left( d_\subset(S_{\widehat{P}_n}^\star, S_{P_0}^\star) \ge \epsilon \right) &\le \limsup_{n \to \infty} \mathbb{P}^* \left( \inf_{x \in S_{P_0}^{\star\epsilon}} \widehat{F}_n(x) > \inf_{x \in \mathcal{X}} \widehat{F}_n(x) \right) \\
&= \mathbb{P}^* \left( \inf_{x \in S_{P_0}^{\star\epsilon}} F(x) > \inf_{x \in \mathcal{X}} F(x) \right) = 0
\end{aligned}
$$

where $\widehat{F}_n(x) := \sup_{P \ll \widehat{P}_n} \{ \mathbb{E}_P[\ell(x;\xi)] : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n} \}$ and $F(x) := \mathbb{E}_{P_0}[\ell(x;\xi)]$. $\square$

**5.2. Consistency for convex problems**    When the function $\ell(\cdot;\xi)$ is convex, we can give consistency guarantees for minimizers of the robust upper bound (4a) by leveraging epigraphical convergence theory [46, 69], bypassing the uniform convergence and compactness conditions above. Analogous results hold for sample average approximation [78, Chapter 5.1.1].

In the theorem, we let $S_{P_0}^\star$ and $S_{\widehat{P}_n}^\star$ be the solution sets (23) as before. We require a much weaker variant of Assumption 5: we assume that for some $\epsilon > 0$, we have $\mathbb{E}[|\ell(x;\xi)|^{1+\epsilon}] < \infty$ for all $x \in \mathcal{X}$. We also assume there exists a function $g : \mathcal{X} \times \Xi \to \mathbb{R}$ such that for each $x \in \mathcal{X}$, there is a neighborhood $U$ of $x$ with $\inf_{z \in U} \ell(z;\xi) \ge g(x,\xi)$ and $\mathbb{E}[|g(x,\xi)|] < \infty$. Then we have the following result, whose proof we provide in Appendix E.2.

THEOREM 6. *In addition to the conditions of the previous paragraph, let Assumptions 1, 3, and 4 hold. Assume that $\mathbb{E}_{\widehat{P}_n}[|\ell(x;\xi)|^{1+\epsilon}] \overset{a.s.}{\to} \mathbb{E}_{P_0}[|\ell(x;\xi)|^{1+\epsilon}]$ for $x \in \mathcal{X}$. Then*

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

*and $d_\subset(S_{\widehat{P}_n}^\star, S_{P_0}^\star) \overset{P^*}{\to} 0$.*

By comparison with Theorem 5 and Corollary 1, we see that Theorem 6 requires weaker conditions on the boundedness of the domain $\mathcal{X}$, instead relying on the compactness of the solution set $S_{P_0}^\star$ and the growth of $\mathbb{E}_{P_0}[\ell(x;\xi)]$ off of this set, which means that eventually $S_{\widehat{P}_n}^\star$ is nearly contained in $S_{P_0}^\star$. When $\{\xi_i\}$ are not i.i.d., the pointwise strong law for $|\ell(x;\xi)|^{1+\epsilon}$ holds if $\{\xi_i\}$ is strongly mixing ($\alpha$-mixing) [40], so the theorem immediately generalizes to dependent sequences.

**6. Simulations** We present three simulation experiments in this section: portfolio optimization, conditional value-at-risk optimization, and optimization in the multi-item newsvendor model. In each of our three simulations, we compute and compare the following approaches to estimation and inference:

(i) We compute the generalized empirical likelihood confidence interval $[l_n, u_n]$ as in expression (4), but we use the (computable) estimate $\widehat{l}_n$ of Eq. (18) in Section 3.3. With these, we simulate the true coverage probability $\mathbb{P}(\inf_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x; \xi)] \in [\widehat{l}_n, u_n])$ (because we control the distribution $P_0$ and $\ell(x; \xi)$) of our confidence intervals, and we compare it to the nominal $\chi^2$-confidence level $\mathbb{P}(\chi_1^2 \leq \rho)$ our asymptotic theory in Section 3 suggests.

(ii) We compute the coverage rates of the normal confidence intervals (11) at the same level as our $\chi^2$ confidence level.

Throughout our simulations (and for both the normal and generalized empirical likelihood/robust approximations), we use the nominal 95% confidence level, setting $\rho = \chi^2_{1,0.95}$, so that we attain the asymptotic coverage $\mathbb{P}(\chi_1^2 \leq \rho) = 0.95$. We focus on i.i.d. sequences and assume that $\xi_i \overset{\text{iid}}{\sim} P_0$ in the rest of the section.

To solve the convex optimization problems (17) and (18) to compute $u_n$ and $\widehat{l}_n$, respectively, we use the `Julia` package `convex.jl` [80]. Each experiment consists of 1250 independent replications for each of the sample sizes $n$ we report, and we vary the sample size $n$ to explore its effects on coverage probabilities. In all of our experiments, because of its computational advantages, we use the $\chi^2$-squared divergence $f_2(t) = \frac{1}{2}(t-1)^2$. We summarize our numerical results in Table 1, where we simulate runs of sample size up to $n = 10,000$ for light-tailed distributions, and $n = 100,000$ for heavy-tailed distributions; in both cases, we see that actual coverage very closely approximates the nominal coverage 95% at the largest value of sample size ($n$) reported. In Figure 1, we plot upper/lower confidence bounds and mean estimates, all of which are averaged over the 1250 independent runs.

**6.1. Portfolio Optimization** Our first simulation investigates the standard portfolio optimization problem (recall Example 4, though we *minimize* to be consistent with our development). We consider problems in dimension $d = 20$ (i.e. there are 20 assets). For this problem, the objective is $\ell(x; \xi) = x^\top \xi$, we set $\mathcal{X} = \{x \in \mathbb{R}^d \mid \mathbb{1}^\top x = 1, -10 \leq x \leq 10\}$ as our feasible region (allowing leveraged investments), and we simulate returns $\xi \overset{\text{iid}}{\sim} N(\mu, \Sigma)$. Within each simulation, the vector $\mu$ and covariance $\Sigma$ are chosen as $\mu \sim N(0, I_d)$ and $\Sigma$ is standard Wishart distributed with $d$ degrees of freedom. The population optimal value is $\inf_{x \in \mathcal{X}} \mu^\top x$. As $\mu \in \mathbb{R}^d$ has distinct entries, the conditions of Theorem 2 are satisfied because the population optimizer is unique. We also consider the (negative) Markowitz problem

$$\underset{x \in \mathcal{X}}{\text{minimize}} \ \mathbb{E}_{\widehat{P}_n}[x^\top \xi] + \sqrt{\frac{\rho}{n} \text{Var}_{\widehat{P}_n}(x^\top \xi)},$$

as the variance-regularized expression is efficiently minimizable (it is convex) in the special case of linear objectives. In Figure 1a, we plot the results of our simulations. The vertical axis is the estimated confidence interval for the optimal solution value for each of the methods, shifted so that $0 = \mu^\top x^\star$, while the horizontal axis is the sample size $n$. We also plot the estimated value of the objective returned by the Markowitz optimization (which is somewhat conservative) and the estimated value given by sample average approximation (which is optimistic), averaging the confidence intervals over 1250 independent simulations. Concretely, we see that the robust/empirical likelihood-based confidence interval at $n = 20$ is approximately $[-150, 40]$, and the Markowitz portfolio is the line slightly above 0, but below each of the other estimates of expected returns. In Table 1, we give the actual coverage rates— the fraction of time the estimated confidence interval

TABLE 1. Coverage Rates (nominal = 95%)

| % | Portfolio | | Newsvendor | | CVaR Normal | | CVaR tail $a=3$ | | CVaR tail $a=5$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

contains the true value $\mu^\top x^\star$. In comparison with the normal confidence interval, generalized empirical likelihood undercovers in small sample settings, which is consistent with previous observations for empirical likelihood (*e.g.*, [64, Sec 2.8]).
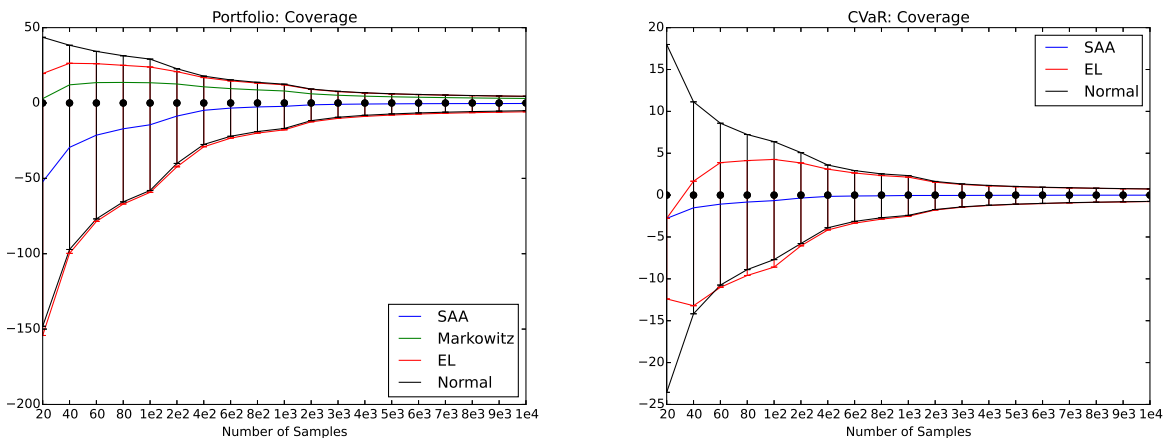
**6.2. Conditional Value-at-Risk** For a real-valued random variable $\xi$, the *conditional value-at-risk* $\alpha$ (CVaR) is the expectation of $\xi$ conditional on it taking values above its $1-\alpha$ quantile [68]. More concisely, the CVaR (at $\alpha$) is

$$\mathbb{E}[\xi \mid \xi \geq q_{1-\alpha}] \overset{(i)}{=} \inf_x \left\{ \frac{1}{1-\alpha} \mathbb{E}[(\xi-x)_+] + x \right\} \quad \text{where } q_{1-\alpha} = \inf\{x \in \mathbb{R} : 1-\alpha \leq \mathbb{P}(\xi \leq x)\}.$$

Conditional Value-at-Risk is of interest in many financial applications [68, 78].

For our second simulation experiment, we investigate three different distributions: a mixture of normal distributions and two different mixtures of heavy-tailed distributions. For our normal experiments, we draw $\xi$ from an equal weight mixture of normal distributions with means $\mu \in \{-6, -4, -2, 0, 2, 4, 6\}$ and variances $\sigma^2 \in \{2, 4, 6, 8, 10, 12, 14\}$, respectively. In keeping with our financial motivation, we interpret $\mu$ as negative returns, where $\sigma^2$ increases as $\mu$ increases, reminiscent of the oft-observed tendency in bear markets (the leverage effect) [14, 19]. For the heavy-tailed experiments, we take $\xi = \mu + Z$ for $Z$ symmetric with $\mathbb{P}(|Z| \geq t) \propto \min\{1, t^{-a}\}$, and we take an equal weight mixture of distributions centered at $\mu \in \{-6, -4, -2, 0, 2, 4, 6\}$.

Our optimization problem is thus to minimize the loss $\ell(x; \xi) = \frac{1}{1-\alpha}(\xi-x)_+ + x$, and we compare the performance of the generalized empirical likelihood confidence regions we describe and normal approximations. For all three mixture distributions, the cumulative distribution function is increasing, so there is a unique population minimizer. To approximate the population optimal value, we take $n = 1,000,000$ to obtain a close sample-based approximation to the CVaR $\mathbb{E}_{P_0}[\xi \mid \xi \geq q_{1-\alpha}]$.

(a) Portfolio Optimization



(b) Conditional Value-at-Risk
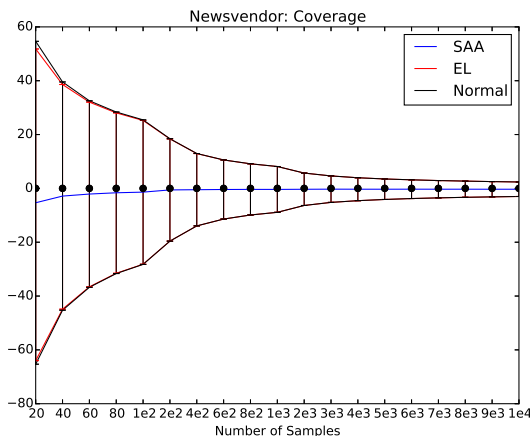


(c) Multi-item Newsvendor

FIGURE 1. Average confidence sets for $\inf_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x; \xi)]$ for normal approximations (11) ("Normal") and the generalized empirical likelihood confidence set (4) ("EL"). The true value being approximated in each plot is centered at 0. The optimal objective computed from the sample average approximation ("SAA") has a negative bias.

Although the feasible region $\mathcal{X} = \mathbb{R}$ is not compact, we compute the generalized empirical likelihood interval (4) and compare coverage rates for confidence regions that asymptotically have the nominal level 95%. In Table 1, we see that the empirical likelihood coverage rates are generally smaller than the normal coverage rates, which is evidently (see Figure 1b) a consequence of still remaining negative bias (optimism) in the robust estimator (4a). In addition, the true coverage rate converges to the nominal level (95%) more slowly for heavy-tailed data (with $\beta \in \{3, 5\}$).

**6.3. Multi-item Newsvendor**  Our final simulation investigates the performance of the generalized empirical likelihood integral (4) for the multi-item newsvendor problem. In this problem, the random variables $\xi \in \mathbb{R}^d$ denote demands for items $j = 1, \dots, d$, and for each item $j$, there is a backorder cost $b_j$ per unit and inventory cost $h_j$ per unit. For a given allocation $x \in \mathbb{R}^d$ of items, then, the loss upon receiving demand $\xi$ is $\ell(x; \xi) = b^\top (x - \xi)_+ + h^\top (\xi - x)_+$, where $(\cdot)_+$ denotes the elementwise positive-part of its argument.

For this experiment, we take $d = 20$ and set $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_1 \leq 10\}$, letting $\xi \overset{\text{iid}}{\sim} \mathsf{N}(0, \Sigma)$ (there may be negative demand), where $\Sigma$ is again standard Wishart distributed with $d$ degrees of freedom. We choose $b, h$ to have i.i.d. entries distributed as $\mathsf{Exp}(\frac{1}{10})$. For each individual simulation, we

approximate the population optimum using a sample average approximation based on a sample of size $n = 10^5$. As Table 1 shows, the proportion of simulations in which $[\widehat{l}_n, u_n]$ covers the true optimal value is still lower than the nominal 95%, though it is less pronounced than other cases. Figure 1c shows average confidence intervals for the optimal value for both generalized empirical likelihood-based and normal-based confidence sets.

**7. General Results**    In this section, we abstract away from the stochastic optimization setting that motivates us. By leveraging empirical process theory, we give general results that apply to suitably smooth functionals (Hadamard differentiable) and classes of functions $\{\ell(x; \cdot) : x \in \mathcal{X}\}$ for which a uniform central limit theorem holds ($P_0$-Donsker). Our subsequent development implies the results presented in previous sections as corollaries. We begin by showing results for i.i.d. sequences and defer extensions to dependent sequences to Section 7.3. Let $Z_1, \ldots, Z_n$ be independent random vectors with common distribution $P_0$. Let $\mathcal{P}$ be the set of probability distributions on $\Xi$ and let $T : \mathcal{P} \to \mathbb{R}$ be a functional of interest.

First, we show a general version of the uniform asymptotic expansion (10) that applies to $P_0$-Donsker classes in Section 7.1. In Section 7.2 we give a generalized empirical likelihood theory for Hadamard differentiable functionals $T(P)$, which in particular applies to $T_{\mathrm{opt}}(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x; \xi)]$ (cf. Theorem 2). The general treatment for Hadamard differentiable functionals is necessary as Frechét differentiability is too stringent for studying constrained stochastic optimization [76]. Finally, we present extensions of the above results to (quickly-mixing) dependent sequences in Section 7.3.

**7.1. Uniform Asymptotic Expansion**    A more general story requires some background on empirical processes, which we now briefly summarize (see van der Vaart and Wellner [82] for a full treatment). Let $P_0$ be a fixed probability distribution on the measurable space $(\Xi, \mathcal{A})$, and recall the space $L^2(P_0)$ of functions square integrable with respect to $P_0$, 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 signed measure $\mu$ on $\Xi$ 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 bounded linear functionals on $\mathcal{H}$ equipped 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. 82, Section 1.2]. We then have the following definition [82, Eq. (2.1.1)] that describes sets of functions on which the central limit theorem holds uniformly.

DEFINITION 3.   A class of functions $\mathcal{H}$ is $P_0$-*Donsker* if $\sqrt{n}(\widehat{P}_n - P_0) \Rightarrow G$ 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 $\xi_i \overset{\text{iid}}{\sim} P_0$.

In Definition 3, the measures $\widehat{P}_n$, $P_0$ are considered as elements in $\mathcal{L}^\infty(\mathcal{H})$ with $\widehat{P}_n f = \mathbb{E}_{\widehat{P}_n} f$, $P_0 f = \mathbb{E}_{P_0} f$ for $f \in \mathcal{H}$.

With these preliminaries in place, we can state a general form of Theorem 1. We let $\mathcal{H}$ be a $P_0$-Donsker collection of functions $h : \Xi \to \mathbb{R}$ with $L^2$-integrable envelope, that is, $M_2 : \Xi \to \mathbb{R}_+$ with $h(\xi) \le M_2(\xi)$ for all $h \in \mathcal{H}$ with $\mathbb{E}_{P_0}[M_2(\xi)^2] < \infty$. Assume the data $\xi_i \overset{\text{iid}}{\sim} P_0$. Then we have

THEOREM 7.   *Let the conditions of the preceding paragraph hold. Then*

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

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

See Appendix B, in particular Appendix B.3, for the proof.

Theorem 7 is useful, and in particular, we can derive Theorem 1 as a corollary:

EXAMPLE 6 (FUNCTIONS LIPSCHITZ IN $x$).   Suppose that for each $\xi \in \Xi$, the function $x \mapsto \ell(x; \xi)$ is $L(\xi)$-Lipschitz, where $\mathbb{E}[L(\xi)^2] < \infty$. If in addition the set $\mathcal{X}$ is compact, then functions $\mathcal{H} := \{\ell(x; \cdot)\}_{x \in \mathcal{X}}$ satisfy all the conditions of Theorem 7. (See also [82, Chs. 2.7.4 and 3.2].)

**7.2. Hadamard Differentiable Functionals**   In this section, we present an analogue of the asymptotic calibration in Proposition 1 for smooth functionals of probability distributions, which when specialized to the optimization context yield the results in Section 3. 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 3). 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 [82, Chapter 3.9]. In the definition, we let $\mathcal{M}$ denote the space of signed measures on $\Xi$ 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 4.   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 $\mathcal{L}^{\infty}(\mathcal{H})$ (i.e. $\|H_n - H\|_{\mathcal{H}} \to 0$) for which $P + t_n H_n \in \mathcal{P}$, and

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

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

$$\lim_{t \to 0} \sup_{H \in K, P + tH \in \mathcal{P}} \left| \frac{T(P + tH) - T(P)}{t} - dT_P(H) \right| = 0. \tag{24}$$

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 *influence function* $T^{(1)} : \Xi \times \mathcal{P} \to \mathbb{R}$ if

$$dT_P(Q - P) = \int_{\Xi} T^{(1)}(\xi; P) d(Q - P)(\xi) \tag{25}$$

and $T^{(1)}$ satisfies $\mathbb{E}_P[T^{(1)}(\xi; P)] = 0$.[1] If we let $B = B(\mathcal{H}, P) \subset \mathcal{L}^{\infty}(\mathcal{H})$ be the set of linear functionals on $\mathcal{H}$ that are $\|\cdot\|_{L^2(P)}$-uniformly continuous and bounded, then this is sufficient for the existence of the canonical derivative $T^{(1)}$, by the Riesz Representation Theorem for $L^2$ spaces (see [81, Chapter 25.5] or [47, Chapter 18]).

We now extend Proposition 1 to Hadamard differentiable functionals $T : \mathcal{P} \to \mathbb{R}$. Owen [63] 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 similar result under certain uniform entropy conditions, but their proofs are incomplete.[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 $T^{(1)}(\xi; P) = dT_P(\mathbb{1}_{\xi} - P)$.

[2] Their proofs [8, pg. 308] show that confidence sets converge to one another in Hausdorff distance, which is not sufficient for their claim. The sets $A_n := \{v/n : v \in \mathbb{Z}^d\}$ and $B = \mathbb{R}^d$ have Hausdorff distance $\frac{1}{2n}$, 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 8. *Let Assumption 1 hold and let $\mathcal{H}$ be a $P_0$-Donsker class of functions with an $L^2$-envelope $M$. Let $\xi_i \overset{\text{iid}}{\sim} P_0$ and let $B \subset \mathcal{M}$ be such that $G$ takes values in $B$ where $G$ is the limit $\sqrt{n}(\widehat{P}_n - P_0) \Rightarrow G$ in $\mathcal{L}^\infty(\mathcal{H})$ given in Definition 3. Assume that $T : \mathcal{P} \subset \mathcal{M} \to \mathbb{R}$ is Hadamard differentiable at $P_0$ tangentially to $B$ with influence function $T^{(1)}(\cdot; P_0)$ 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*

$$\lim_{n \to \infty} \mathbb{P}\left(T(P_0) \in \left\{T(P) : D_f\left(P \| P_n\right) \le \frac{\rho}{n}\right\}\right) = \mathbb{P}\left(\chi_1^2 \le \rho\right), \qquad (26)$$

We use Theorem 7 to show the result in Appendix B.4.

**7.3. Extensions to Dependent Sequences**   In this subsection, we show an extension of the empirical likelihood theory for smooth functionals (Theorem 8) to $\beta$-mixing sequence of random variables. Let $\{\xi\}_{i \in \mathbb{Z}}$ be a sequence of strictly stationary random variables taking values in the Polish space $\Xi$. We follow the approach of Doukhan et al. [29] to prove our results, giving bracketing number conditions sufficient for our convergence guarantees (alternative approaches are possible [60, 2, 86, 67]).

We first define bracketing numbers.

DEFINITION 5.   Let $\|\cdot\|$ be a (semi)norm on $\mathcal{H}$. For functions $l, u : \Xi \to \mathbb{R}$ with $l \le u$, the *bracket* $[l, u]$ is the set of functions $h : \Xi \to \mathbb{R}$ such that $l \le h \le u$, and $[l, u]$ is an $\epsilon$-*bracket* if $\|l - u\| \le \epsilon$. Brackets $\{[l_i, u_i]\}_{i=1}^m$ *cover* $\mathcal{H}$ if for all $h \in \mathcal{H}$, there exists $i$ such that $h \in [l_i, u_i]$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|)$ is the minimum number of $\epsilon$-brackets needed to cover $\mathcal{H}$.

For i.i.d. sequences, if the bracketing integral is finite,

$$\int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|_{L^2(P_0)})} \, d\epsilon < \infty,$$

then $\mathcal{H}$ is $P_0$-Donsker [82, Theorem 2.5.6]. For $\beta$-mixing sequences, a modification of the $L^2(P_0)$-norm yields similar result. To state the required bracketing condition in full, we first provide requisite notation. For any $h \in L^1(P_0)$, we let

$$Q_h(u) = \inf\{t : \mathbb{P}(|h(\xi_0)| > t) \le u\}.$$

be the quantile function of $|h(\xi_0)|$. Define $\beta(t) := \beta_{\lfloor t \rfloor}$ where $\beta_n$ are the mixing coefficients (13), and define the norm

$$\|h\|_{L^{2,\beta}(P_0)} = \sqrt{\int_0^1 \beta^{-1}(u) Q_h(u)^2 du}, \qquad (27)$$

where $\beta^{-1}(u) = \inf\{t : \beta(t) \le u\}$. When $\{\xi_i\}_{i \in \mathbb{Z}}$ are i.i.d., the $(2, \beta)$-norm $\|\cdot\|_{L^{2,\beta}(P_0)}$ is the $L^2(P_0)$-norm as $\beta^{-1}(u) = 1$ for $u > 0$. Lastly, we let $\Gamma$ be the covariance function

$$\Gamma(h_1, h_2) := \sum_{i \in \mathbb{Z}} \mathrm{Cov}(h_1(\xi_0), h_2(\xi_i)). \qquad (28)$$

We then have the following result, which extends bracketing entropy conditions to $\beta$-mixing sequences.

LEMMA 5 (**Doukhan et al. [29, Theorem 1]**).   *Let $\{\xi\}_{i \in \mathbb{Z}}$ be a strictly stationary sequence of random vectors taking values in the Polish space $\Xi$ with common distribution $P_0$ satisfying $\sum_{n=1}^\infty \beta_n < \infty$. Let $\mathcal{H}$ be a class of functions $h : \Xi \to \mathbb{R}$ with envelope $M(\cdot)$ such that $\|M\|_{L^{2,\beta}(P_0)} < \infty$. If*

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|_{L^{2,\beta}(P_0)})} \, d\epsilon < \infty,$$

*then the series* $\sum_i \text{Cov}(h(\xi_0), h(\xi_i))$ *is absolutely convergent to* $\Gamma(h,h) < \infty$ *uniformly in h, and*

$$\sqrt{n}(\widehat{P}_n - P_0) \Rightarrow G \quad in \quad \mathcal{L}^\infty(\mathcal{H})$$

*where G is a Gaussian process with covariance function* $\Gamma$ *and almost surely uniformly continuous sample paths.*

The discussion following [29, Theorem 1] provides connections between $\|\cdot\|_{L^{2,\beta}(P_0)}$ and other norms, as well as sufficient conditions for Lemma 5 to hold. For example, if the bracketing integral with respect to the norm $\|\cdot\|_{L^{2r}(P_0)}$ is finite with $\sum_{n \geq 1} n^{\frac{1}{r-1}} \beta_n < \infty$, the conditions of Lemma 5 are satisfied.

We now give an extension of Theorem 8 for dependent sequences. Recall that $\mathcal{M}$ is the (vector) space of signed measures in $\mathcal{L}^\infty(\mathcal{H})$. Let $B \subset \mathcal{M}$ be such that $G$ takes values in $B$.

THEOREM 9. *Let Assumption 1 and the hypotheses of Lemma 5 hold. Let $B \subset \mathcal{M}$ be such that $G$ takes values in $B$, where $\sqrt{n}(\widehat{P}_n - P_0) \Rightarrow G$ in $\mathcal{L}^\infty(\mathcal{H})$ as in Lemma 5. Assume that $T : \mathcal{P} \subset \mathcal{M} \to \mathbb{R}$ is Hadamard differentiable at $P_0$ tangentially to $B$ with influence function $T^{(1)}(\cdot; P_0)$ as (Eq. (25)) and that $dT_P$ is defined and continuous on the whole of $\mathcal{M}$. If $0 < \text{Var}(T^{(1)}(\xi; P_0)) < \infty$, then*

$$\lim_{n \to \infty} \mathbb{P}\left(T(P_0) \in \left\{T(P) : D_f(P\|P_n) \leq \frac{\rho}{n}\right\}\right) = \mathbb{P}\left(\chi_1^2 \leq \frac{\rho \text{Var}_{P_\xi} T^{(1)}(\xi; P_0)}{\Gamma(T^{(1)}, T^{(1)})}\right). \tag{29}$$

See Section D.2 for the proof. We show in Section D.3 that Theorem 4 follows from Theorem 9.

**8. Conclusion** We have extended generalized empirical likelihood theory in a number of directions, showing how it provides inferential guarantees for stochastic optimization problems. The upper confidence bound (4a) is a natural robust optimization problem [5, 6], and our results show that this robust formulation gives exact asymptotic coverage. The robust formulation implements a type of regularization by variance, while maintaining convexity and risk coherence (Theorem 7). This variance expansion explains the coverage properties of (generalized) empirical likelihood, and we believe it is likely to be effective in a number of optimization problems [30].

There are a number of interesting topics for further research, and we list a few of them. On the statistical and inferential side, the uniqueness conditions imposed in Theorem 2 are stringent, so it is of interest to develop procedures that are (asymptotically) adaptive to the size of the solution set $S_{P_0}^\star$ without being too conservative; this is likely to be challenging, as we no longer have normality of the asymptotic distribution of solutions. On the computational side, interior point algorithms are often too expensive for large scale optimization problems (i.e. when $n$ is very large)—just evaluating the objective or its gradient requires time at least linear in the sample size. While there is a substantial and developed literature on efficient methods for sample average approximation and stochastic gradient methods [66, 58, 31, 25, 43, 37], there are fewer established and computationally efficient solution methods for minimax problems of the form (4a) (though see the papers [58, 20, 7, 73, 56] for work in this direction). Efficient solution methods need to be developed to scale up robust optimization.

There are two ways of injecting robustness in the formulation (4a): increasing $\rho$ and choosing a function $f$ defining the $f$-divergence $D_f(\cdot\|\cdot)$ that grows slowly in a neighborhood of 1 (recall the Cressie-Read family (19) and associated dual problems). We characterize a statistically principled way of choosing $\rho$ to obtain calibrated confidence bounds, and we show that all smooth $f$-divergences have the same asymptotic ($n \to \infty$) behavior to first-order. We do not know, however, the extent to which different choices of the divergence measure $f$ impact higher order or finite-sample behavior of the estimators we study. While the literature on higher order corrections for empirical likelihood offers some answers for inference problems regarding the mean of a distribution [27, 4, 21, 17, 18], the more complex settings arising in large-scale optimization problems leave a number of open questions.

## References

[1] 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.

[2] M. A. Arcones and B. Yu. Central limit theorems for empirical and U-processes of stationary mixing sequences. *Journal of Theoretical Probability*, 7(1):47–71, 1994.

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

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

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

[6] 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.

[7] A. Ben-Tal, E. Hazan, T. Koren, and S. Mannor. Oracle-based robust optimization via online learning. *Operations Research*, 63(3):628–638, 2015.

[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 $\varphi^*$-divergence minimizers for hadamard differentiable functionals. In *Topics in Nonparametric Statistics*, pages 21–32. Springer, 2014.

[10] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

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

[12] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming, Series A*, 167(2):235–292, 2018. URL http://arxiv.org/abs/1401.0212.

[13] P. Billingsley. *Probability and Measure*. Wiley, Second edition, 1986.

[14] 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.

[15] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *arXiv:1604.01446 [math.PR]*, 2016. URL https://arxiv.org/abs/1604.01446.

[16] R. C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

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

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

[19] 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.

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

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

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

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

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

[25] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27*, 2014.

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

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

[28] P. Doukhan. *Mixing, Properties and Examples*. Number 85 in Lecture Notes in Statistics. Springer, 1994.

[29] P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 393–427. Elsevier, 1995.

[30] J. C. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *arXiv:1610.02581 [stat.ML]*, 2016.

[31] 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.

[32] 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.

[33] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming, Series A*, to appear, 2017.

[34] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

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

[36] P. W. Glynn and A. Zeevi. Bounding stationary expectations of markov processes. In *Markov processes and related topics: a Festschrift for Thomas G. Kurtz*, pages 195–214. Institute of Mathematical Statistics, 2008.

[37] E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2016.

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

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

[40] I. A. Ibragimov. Some limit theorems for stationary processes. *Theory of Probability & Its Applications*, 7(4):349–382, 1962.

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

[42] 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.

[43] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, 2013.

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

[45] 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.

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

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

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

[49] H. Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

[50] H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 2018. URL http://arXiv.org/abs/1605.09349.

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

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

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

[54] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Second edition, 2009.

[55] A. Mokkadem. Propriétés de mélange des processus autorégressifs polynomiaux. *Ann. Inst. H. Poincaré Probab. Statist*, 26(2):219–260, 1990.

[56] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with $f$-divergences. In *Advances in Neural Information Processing Systems 29*, 2016.

[57] H. Namkoong and J. C. Duchi. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems 30*, 2017.

[58] 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.

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

[60] A. Nobel and A. Dembo. A note on uniform laws of averages for dependent processes. *Statistics & Probability Letters*, 17(3):169–172, 1993.

[61] E. Nummelin and R. L. Tweedie. Geometric ergodicity and r-positivity for general markov chains. *The Annals of Probability*, pages 404–420, 1978.

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

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

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

[65] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[66] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[67] E. Rio. *Asymptotic Theory of Weakly Dependent Random Processes*. Springer, 2017.

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

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

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

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

[72] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[73] S. Shalev-Shwartz and Y. Wexler. Minimizing the maximal loss: How and why? In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

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

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

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

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

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

[79] A. Sinha, H. Namkoong, and J. C. Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv:1710.10571 [stat.ML]*, 2017.

[80] M. Udell, K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd. Convex optimization in Julia. In *First Workshop on High Performance Technical Computing in Dynamic Languages*, pages 18–28. IEEE, 2014.

[81] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

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

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

[84] D. Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research*, 193(1): 21–47, 2012.

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

[86] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994.

**Appendix A: Proof of Lemma 1** We assume without loss of generality for both that $Z$ is mean-zero and that $\mathrm{Var}(Z) > 0$, as if $\mathrm{Var}(Z) = 0$ then $s_n = 0$ and the lemma is trivial. We prove the result by approximating the function $f$ with simpler functions, which allows a fairly immediate proof.

The starting point of the proof of each of Lemma 1 is the following lemma, which gives sufficient conditions for the robust expectation to be well approximated by the variance.

LEMMA 6. *Let $0 \le \epsilon < 1$ and define the event*

$$\mathcal{E}_n := \left\{ \max_{i \le n} \frac{|Z_i - \overline{Z}_n|}{\sqrt{n}} \le \epsilon s_n \sqrt{\frac{1 - C\epsilon}{\rho}} \right\}.$$

*Then on $\mathcal{E}_n$,*

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

This result gives a nearly immediate proof of Lemma 1, as we require showing only that $\mathcal{E}_n$ holds eventually (i.e. for all large enough $n$). We defer its proof to Section A.1.

To that end, we state the following result, due essentially to Owen [62, Lemma 3].

LEMMA 7. *Let $Z_i$ be (potentially dependent) identically distributed random variables with $\mathbb{E}[|Z_1|^k] < \infty$ for some $k > 0$. Then for all $\epsilon > 0$, $\mathbb{P}(|Z_n| \ge \epsilon n^{1/k} \ i.o.) = 0$ and $\max_{i \le n} |Z_i|/n^{1/k} \overset{a.s.}{\to} 0$.*

*Proof.* A standard change of variables gives $\mathbb{E}[|Z_1|^k] = \int_0^\infty \mathbb{P}(|Z_1|^k \ge t) dt \gtrsim \sum_{n=1}^\infty \mathbb{P}(|Z_n|^k \ge n)$. Thus for any $\epsilon > 0$ we obtain

$$\sum_{n=1}^\infty \mathbb{P}(|Z_n|^k \ge \epsilon n) \lesssim \frac{1}{\epsilon} \mathbb{E}[|Z_1|^k] < \infty,$$

and the Borel-Cantelli lemma gives the result.

Birkhoff's Ergodic Theorem and that the sequence $\{Z_n\}$ is strictly stationary and ergodic with $\mathbb{E}[Z_1^2] < \infty$ implies that

$$\overline{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i \overset{a.s.}{\to} \mathbb{E}[Z_1] \quad \text{and} \quad \overline{Z_n^2} = \frac{1}{n} \sum_{i=1}^{n} Z_i^2 \overset{a.s.}{\to} \mathbb{E}[Z_1^2].$$

Thus we have that $s_n^2 = \overline{Z_n^2} - \overline{Z}_n^2 \overset{a.s.}{\to} \mathrm{Var}(Z) > 0$ and by Lemma 7, the event $\mathcal{E}_n$ holds eventually. Lemma 6 thus gives Lemma 1.

**A.1. Proof of Lemma 6** We require a few auxiliary functions before continuing with the arguments. First, for $\epsilon > 0$ define the Huber function

$$h_\epsilon(t) = \inf_y \left\{ \frac{1}{2}(t-y)^2 + \epsilon|y| \right\} = \begin{cases} \frac{1}{2}t^2 & \text{if } |t| \le \epsilon \\ \epsilon|t| - \frac{1}{2}\epsilon^2 & \text{if } |t| > \epsilon. \end{cases}$$

Now, because the function $f$ is $\mathcal{C}^3$ in a neighborhood of 1 with $f''(1) = 2$, there exist constants $0 < c, C < \infty$, depending only on $f$, such that

$$2(1 - C\epsilon)h_\epsilon(t) \le f(t+1) \le (1 + C\epsilon)t^2 + \mathbf{I}_{[-\epsilon,\epsilon]}(t) \quad \text{for all } t \in \mathbb{R}, \text{ and } |\epsilon| \le c, \tag{30}$$

where the first inequality follows because $t \mapsto f'(t)$ is non-decreasing and $f'(1) = 0$ and $f''(1) = 2$, and the second similarly because $f$ is $\mathcal{C}^3$ near 1.

With the upper and lower bounds (30) in place, let us rewrite the supremum problem slightly. For $\epsilon < 1$, define the sets

$$\begin{aligned} \mathcal{U}_{\mathrm{sm}} &:= \left\{ u \in \mathbb{R}^n \mid \mathbb{1}^T u = 0, \|nu\|_\infty \le \epsilon, (1 + C\epsilon)\|nu\|_2^2 \le \rho \right\} \subset \dots \\ \mathcal{U} &:= \left\{ u \in \mathbb{R}^n \mid \mathbb{1}^T u = 0, u \ge -1/n, \sum_{i=1}^{n} f(nu_i + 1) \le \rho \right\} \subset \dots \\ \mathcal{U}_{\mathrm{big}} &:= \left\{ u \in \mathbb{R}^n \mid \mathbb{1}^T u = 0, \sum_{i=1}^{n} h_\epsilon(nu_i) \le \frac{\rho}{2(1 - C\epsilon)} \right\}. \end{aligned} \tag{31}$$

Then for any vector $z \in \mathbb{R}^n$, by inspection (replacing $p \in \mathbb{R}_+^n$ with $\mathbb{1}^T p = 1$ with $u \in \mathbb{R}^n$ with $\mathbb{1}^T u = 0$ and $u \ge -(1/n)$), we have

$$\sup_{u \in \mathcal{U}_{\mathrm{sm}}} u^T z \le \sup_p \left\{ p^T z \mid D_f(p \| (1/n)\mathbb{1}) \le \rho/n \right\} - \frac{1}{n}\mathbb{1}^T z = \sup_{u \in \mathcal{U}} u^T z \le \sup_{u \in \mathcal{U}_{\mathrm{big}}} u^T z. \tag{32}$$

To show the lemma, then, it suffices to lower bound $\sup_{u \in \mathcal{U}_{\mathrm{sm}}} u^T z$ and upper bound $\sup_{u \in \mathcal{U}_{\mathrm{big}}} u^T z$.

To that end, the next two lemmas control both the upper and lower bounds in expression (32). In the lemmas, we let $\overline{z_n^2} = \frac{1}{n} \sum_{i=1}^{n} z_i^2$ and $\overline{z}_n = \frac{1}{n} \sum_{i=1}^{n} z_i$.

LEMMA 8. Let $s_n(z)^2 = \overline{z_n^2} - \overline{z}_n^2$. If $\|z - \overline{z}_n\|_\infty / \sqrt{n} \le \epsilon s_n(z)\sqrt{(1 + C\epsilon)/\rho}$, then

$$\sup_{u \in \mathcal{U}_{\mathrm{sm}}} u^T z = \sqrt{\frac{\rho}{n} s_n(z)^2} \frac{1}{\sqrt{1 + C\epsilon}}.$$

*Proof.* We can without loss of generality replace $z$ with $z - \overline{z}_n \mathbb{1}$ in the supremum, as $\mathbb{1}^T u = 0$ so $u^T z = u^T(z - \overline{z}_n \mathbb{1})$, and so we simply assume that $\mathbb{1}^T z = 0$ and thus $\|z\|_2 = \sqrt{n} s_n(z)$. By the Cauchy-Schwarz inequality, $\sup_{u \in \mathcal{U}_{\mathrm{sm}}} u^T z \le \sqrt{\rho} \|z\|_2 / (n\sqrt{1 + C\epsilon})$. We claim that under the conditions of the lemma, it is achieved. Indeed, let

$$u = \frac{\sqrt{\rho}}{n\sqrt{1 + C\epsilon} \|z\|_2} z = \frac{\sqrt{\rho}}{n^{3/2}\sqrt{(1 + C\epsilon)\overline{z_n^2}}} z,$$

so $\|nu\|_2 = \rho$ and $u^T z = \rho \|z\|_2 / (n\sqrt{1+C\epsilon})$. Then $u$ satisfies $u^T \mathbb{1} = 0$, $\|u\|_2^2 \le \rho/(n^2(1+C\epsilon))$, and because $\|z\|_\infty / \sqrt{n} \le \epsilon \sqrt{1+C\epsilon} s_n(z)/\sqrt{\rho}$ by assumption, we have $u \in \mathcal{U}_{\mathrm{sm}}$, which gives the result.

LEMMA 9.    Let $s_n(z)^2 = \overline{z_n^2} - \overline{z}_n^2$. If $\|z - \overline{z}_n\|_\infty / \sqrt{n} \le \epsilon s_n(z)\sqrt{(1-C\epsilon)/\rho}$, then

$$\sup_{u \in \mathcal{U}_{\mathrm{big}}} u^T z \le \sqrt{\frac{\rho}{n} s_n(z)^2} \frac{1}{\sqrt{1-C\epsilon}}.$$

*Proof.* We have $u^T z = u^T(z - \mathbb{1}\overline{z}_n)$ for $u^T \mathbb{1} = 0$. Thus, we always have upper bound that

$$\sup_{u \in \mathcal{U}_{\mathrm{big}}} u^T z \le \sup_{u \in \mathbb{R}^n} \left\{ u^T(z - \mathbb{1}\overline{z}_n) \mid \sum_{i=1}^n h_\epsilon(nu_i) \le \frac{\rho}{2(1-C\epsilon)} \right\}.$$

Let us assume w.l.o.g. (as in the proof of Lemma 8) that $\overline{z}_n = 0$ so that $\|z\|_2 = \sqrt{n} s_n(z)$. Introducing multiplier $\lambda \ge 0$, the Lagrangian for the above maximization problem is

$$L(u, \lambda) = u^T z - \lambda \sum_{i=1}^n h_\epsilon(nu_i) + \lambda \frac{\rho}{2(1-C\epsilon)}.$$

Let us supremize over $u$. In one dimension, we have

$$
\begin{aligned}
\sup_{u_i} \{u_i z_i - \lambda h_\epsilon(nu_i)\} &= \lambda \sup_{v_i} \sup_y \left\{ v_i \frac{z_i}{\lambda n} - \frac{1}{2}(v_i - y)^2 - \epsilon|y| \right\} \\
&= \lambda \sup_y \left\{ \frac{z_i^2}{2\lambda^2 n^2} + y \frac{z_i}{\lambda n} - \epsilon|y| \right\} = \frac{z_i^2}{2\lambda n^2} + \mathbf{I}_{[-\epsilon, \epsilon]}\left( \frac{z_i}{\lambda n} \right).
\end{aligned}
$$

We obtain

$$\sup_{u \in \mathcal{U}_{\mathrm{big}}} u^T z \le \inf_{\lambda \ge 0} \sup_u L(u, \lambda) = \inf_\lambda \left\{ \frac{\|z\|_2^2}{2\lambda n^2} + \frac{\rho}{2(1-C\epsilon)}\lambda \mid \lambda \ge \frac{\|z\|_\infty}{\epsilon n} \right\}. \tag{33}$$

Now, by taking

$$\lambda = \sqrt{\frac{1-C\epsilon}{\rho}} \frac{\|z\|_2}{n},$$

we see that under the conditions of the lemma, we have $\lambda \ge \|z\|_\infty / (\epsilon n)$ and substituting into the Lagrangian dual (33), the lemma follows.

Combining Lemmas 8 and 9 gives Lemma 6.

**Appendix B: Uniform convergence results**    In this section, we give the proofs of theorems related to uniform convergence guarantees and the uniform variance expansions. The order of proofs is not completely reflective of that we present in the main body of the paper, but we order the proofs so that the dependencies among the results are linear. We begin by collecting important technical definitions, results, and a few preliminary lemmas. In Section B.3, we then provide the proof of Theorem 7, after which we prove Theorem 8 in Section B.4. Based on Theorem 8, we are then able to give a nearly immediate proof (in Section B.5) we of Proposition 1.

**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 [82] for an overview.

DEFINITION 6 (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 $K_\epsilon$ 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

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

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

LEMMA 10 (**Prohorov's theorem [82], 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 $X_n \Rightarrow X$ 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 $X_{n(m)} \Rightarrow X$ 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 for random vectors in $\mathcal{L}^\infty(\mathcal{H})$. In each, $\mathcal{H}$ is some set (generally a collection of functions in our applications), and $\Omega_n$ 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 [82], 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 $X_n \Rightarrow X$.*

Although the convergence in distribution in outer probability does not require measurability of the pre-limit quantities, the above lemma guarantees it.

LEMMA 12 (**Van der Vaart and Wellner [82], 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 provide two technical lemmas. Recall the definition

$$\mathcal{P}_{n,\rho} = \{P : D_f(P\|\widehat{P}_n) \leq \frac{\rho}{n}\} \tag{34}$$

The first, Lemma 13, shows that the vector $np$ is close to the all-ones vector for all vectors $p \in \mathcal{P}_{n,\rho}$. The second (Lemma 14) gives conditions for tightness of classes of functions $h : \Xi \to \mathbb{R}$.

LEMMA 13. *Let Assumption 1 hold. Then*

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

*for some $C_f \geq c_f > 0$ 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, $C_f^{-1} := \inf_{s \in [m,M]} f''(s)$ and $c_f^{-1} := \sup_{s \in [m,M]} f''(s)$ exists, are attained, and are strictly positive. Using the Taylor expansion of the $f$-divergence, we have $(np_i - 1)^2 = 2f(np_i)/f''(s_i)$ for each $i$, and thus

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

and similarly $\sum_{i=1}^n (np_i - 1)^2 \geq 2c_f \rho$. Taking the square root of each sides gives the lemma.

LEMMA 14. *Let $\mathcal{H}$ be $P_0$-Donsker with $L^2$-integrable envelope $M_2$, i.e. $|h(\xi)| \leq M_2(\xi)$ for all $h \in \mathcal{H}$ with $\mathbb{E}_{P_0}[M_2^2(\xi)] < \infty$. Then for any sequence $Q_n \in \mathcal{P}_{n,\rho}$, the mapping $\sqrt{n}(Q_n - P_0) : \mathcal{L}^\infty(\mathcal{H}) \to \mathbb{R}$ is asymptotically tight.*

$W$ e 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

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

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

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

As $\sqrt{n}(\widehat{P}_n - P_0)$ is asymptotically tight in $\mathcal{L}^\infty(\mathcal{H})$ [82, 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 7**   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 each $h \in \mathcal{H}$ mean-zero (as we may replace $h(\xi)$ with $h(\xi) - \mathbb{E}[h(\xi)]$). 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 $\varepsilon_n$ such that

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

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

To establish asymptotic tightness of the collection $\{\sup_{P \in \mathcal{P}_{\rho,n}} \mathbb{E}_P[h(\xi)]\}_{h \in \mathcal{H}}$, first note that we have finite dimensional marginal convergence. Indeed, we have $\sqrt{n}\varepsilon_n(h) \overset{p}{\to} 0$ for all $h \in \mathcal{H}$ by Lemma 1, and so for any finite $k$ and any $h_1, \ldots, h_k \in \mathcal{H}$, $\sqrt{n}(\varepsilon_n(h_1), \ldots, \varepsilon_n(h_k)) \overset{p}{\to} 0$. Further, by

our Donsker assumption on $\mathcal{H}$ we have that $\{h(\cdot)^2, h \in \mathcal{H}\}$ is a Glivenko-Cantelli class [82, Lemma 2.10.14], and

$$\sup_{h \in \mathcal{H}} \left| \mathrm{Var}_{\widehat{P}_n}(h(\xi)) - \mathrm{Var}_{P_0}(h(\xi)) \right| \overset{P^*}{\to} 0. \tag{35}$$

Now, we write the error term $\varepsilon_n$ as

$$\sqrt{n}\varepsilon_n(h) = \underbrace{\sqrt{n} \sup \left\{ \mathbb{E}_P[h(\xi)] - \mathbb{E}_{P_0}[h(\xi)] \mid D_f(P\|\widehat{P}_n) \le \rho/n \right\}}_{(a)}$$
$$- \underbrace{\sqrt{n} \left( \mathbb{E}_{\widehat{P}_n}[h(\xi)] - \mathbb{E}_{P_0}[h(\xi)] \right)}_{(b)} - \underbrace{\sqrt{\rho \mathrm{Var}_{\widehat{P}_n}(h(\xi))}}_{(c)}.$$

Then term $(a)$ is asymptotically tight (as a process on $h \in \mathcal{H}$) in $\mathcal{L}^\infty(\mathcal{H})$ by Lemma 14. The term $(b)$ is similarly tight because $\mathcal{H}$ is $P_0$-Donsker by assumption, and term $(c)$ is tight by the uniform Glivenko-Cantelli result (35). In particular, $\sqrt{n}\varepsilon_n(\cdot)$ is an asymptotically tight sequence in $\mathcal{L}^\infty(\mathcal{H})$. As the finite dimensional distributions all converge to 0 in probability, Lemma 11 implies that $\sqrt{n}\varepsilon_n \Rightarrow 0$ 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 8**    We first state a standard result that the delta method applies for Hadamard differentiable functionals, as given by van der Vaart and Wellner [82, Section 3.9]. In the lemma, the sets $\Omega_n$ denote the implicit sample spaces defined for each $n$. For a proof, see [82, Theorem 3.9.4].

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

For a probability measure $P$, define $\kappa(P) := T(P) - T(P_0) - \mathbb{E}_P[T^{(1)}(\xi; P_0)]$. Since $\mathcal{H}$ was assumed to be $P_0$-Donsker, we have $\sqrt{n}(\widehat{P}_n - P_0) \Rightarrow G$ in $\mathcal{L}^\infty(\mathcal{H})$. Recalling the canonical derivative $T^{(1)}$, we have from Lemma 15 that

$$T(\widehat{P}_n) = T(P_0) + \mathbb{E}_{\widehat{P}_n}[T^{(1)}(\xi, P_0)] + \kappa(\widehat{P}_n) \tag{36}$$

where $\kappa(\widehat{P}_n) = o_P(n^{-\frac{1}{2}})$. Next, we show that this is true uniformly over $\{P : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\}$. We return to prove the lemma in Section B.4.1 for the proof.

LEMMA 16.    *Under the assumptions of Theorem 8, for any $\epsilon > 0$*

$$\limsup_n \mathbb{P}\left( \sup_P \left\{ |\kappa(P)| : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n} \right\} \ge \frac{\epsilon}{\sqrt{n}} \right) = 0 \tag{37}$$

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

We now see how the theorem is a direct consequence of Theorem 7 and Lemma 16. Taking sup over $\{P : D_f(P\|\widehat{P}_n) \le \frac{\rho}{n}\}$ in the definition of $\kappa(\cdot)$, we have

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

Now, multiply both sides by $\sqrt{n}$ and apply Theorem 7 and Lemma 16 to obtain

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

Since $\mathbb{E}_{P_0}[T^{(1)}(\xi;P_0)] = 0$ by assumption, the central limit theorem then implies that

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

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

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

**B.4.1. 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 $\liminf_{n\to\infty} \mathbb{P}(X_n \in K^\delta) \geq 1 - \epsilon$ for all $\delta > 0$ where $K^\delta := \{y \in \mathcal{L}^\infty(\mathcal{H}) : d(y,K) < \delta\}$ (*e.g.*, [82, 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 $\sqrt{n(m)}(Q_{n(m)} - P_0) \Rightarrow X$ 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

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

**B.5. Proof of Proposition 1** Let $Z \in \mathbb{R}^d$ be random vectors with covariance $\Sigma$, where $\mathrm{rank}(\Sigma) = d_0$. From Theorem 8, we have that if we define

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

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

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

if and only if

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

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

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

by the continuous mapping theorem.

**Appendix C: Proofs of Statistical Inference for Stochastic Optimization**    In this appendix, we collect the proofs of the results in Sections 3 and 4 on statistical inference for the stochastic optimization problem (1). We first give a result explicitly guaranteeing smoothness of $T_{\text{opt}}(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$. The following variant of Danskin's theorem [24] gives Hadamard differentiability of $T_{\text{opt}}$ tangentially to the space $B(\mathcal{H}, P_0) \subset \mathcal{L}^\infty(\mathcal{H})$ of bounded linear functionals continuous w.r.t. $L^2(P_0)$ (which we may identify with measures, following the discussion after Definition 4). The proof of Lemma 17—which we include in Appendix C.2 for completeness—essentially follows that of Römisch [70].

LEMMA 17.   *Let Assumption 3 hold and assume $x \mapsto \ell(x;\xi)$ is continuous for $P_0$-almost all $\xi \in \Xi$. Then the functional $T_{\text{opt}} : \mathcal{P} \to \mathbb{R}$ defined by $T_{\text{opt}}(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ is Hadamard directionally differentiable on $\mathcal{P}$ tangentially to $B(\mathcal{H}, P_0)$ with derivative*

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

*where $S_P^\star = \arg\min_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$.*

**C.1. Proof of Theorem 2**    By Example 6, we have that $\{\ell(x;\cdot : x \in \mathcal{X}\}$ is $P_0$-Donsker with envelope function $M_2(\xi) = |\ell(x_0;\xi)| + M(\xi)\operatorname{diam}(\mathcal{X})$. Further, Lemma 17 implies that the hypotheses of Theorem 8 are satisfied. Indeed, when the set of $P$-optima $S_P^\star$ is a singleton, Lemma 17 gives that $dT_P$ is a linear functional on the space of bounded measures $\mathcal{M}$,

$$dT_{P_0}(H) = \int \ell(x^\star;\xi)dH(\xi)$$

where $x^\star = \arg\min_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$ and the canonical gradient of $T_{\text{opt}}$ is given by $T^{(1)}(\xi; P_0) = \ell(x^\star;\xi) - \mathbb{E}_{P_0}[\ell(x^\star;\xi)]$.

**C.2. Proof of Lemma 17**    For notational convenience, we identify the set $\mathcal{H} := \{\ell(x;\cdot) : x \in \mathcal{X}\}$ as a subset of $L^2(P_0)$, viewed as functions mapping $\Xi \to \mathbb{R}$ indexed by $x$. Let $H \in B(\mathcal{H}, P_0)$, where for convenience we use the notational shorthand

$$H(x) := H(\ell(x;\cdot)) = \int \ell(x;\xi)dH(\xi),$$

where we have identified $H$ with a measure in $\mathcal{M}$, as in the discussion following Definition 4. We have the norm $\|H\| := \sup_{x \in \mathcal{X}} |H(x)|$, where $\|H\| < \infty$ for $H \in B(\mathcal{H}, P_0)$. In addition, we denote the set of $\epsilon$-minimal points for problem (1) with distribution $P$ by

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

where we let $S_P^\star = S_P^\star(0)$.

We first show that for $H_n \in B(\mathcal{H}, P_0)$ with $\|H - H_n\| \to 0$, we have for any sequence $t_n \to 0$ that

$$\limsup_n \frac{1}{t_n}\left(T_{\text{opt}}(P_0 + t_n H_n) - T_{\text{opt}}(P_0)\right) \leq \inf_{x^\star \in S_{P_0}^\star} H(x^\star). \tag{38}$$

Indeed, let $x^\star \in S^\star_{P_0}$. Then

$$T_{\mathrm{opt}}(P_0 + t_n H_n) - T_{\mathrm{opt}}(P_0) \leq \mathbb{E}_{P_0}[\ell(x^\star; \xi)] + t_n H_n(x^\star) - \mathbb{E}_{P_0}[\ell(x^\star; \xi)] \leq t_n H_n(x^\star).$$

By definition, we have $|H_n(x^\star) - H(x^\star)| \leq \|H_n - H\| \to 0$ as $n \to \infty$, whence

$$\limsup_n \frac{1}{t_n} \left( T_{\mathrm{opt}}(P_0 + t_n H_n) - T_{\mathrm{opt}}(P_0) \right) \leq \limsup_n \frac{1}{t_n} t_n H_n(x^\star) = H(x^\star).$$

As $x^\star \in S^\star_{P_0}$ is otherwise arbitrary, this yields expression (38).

We now turn to the corresponding lower bound that

$$\liminf_n \frac{1}{t_n} \left( T_{\mathrm{opt}}(P_0 + t_n H_n) - T_{\mathrm{opt}}(P_0) \right) \geq \inf_{x^\star \in S^\star_{P_0}} H(x^\star). \tag{39}$$

Because $\|H\| < \infty$ and $\|H_n - H\| \to 0$, we see that

$$\begin{aligned} T_{\mathrm{opt}}(P_0 + t_n H_n) = \inf_{x \in \mathcal{X}} \left\{ \mathbb{E}_{P_0}[\ell(x; \xi)] + t_n H_n(x) \right\} &\leq \inf_{x \in \mathcal{X}} \left\{ \mathbb{E}_{P_0}[\ell(x; \xi)] + t_n \|H_n - H\| + t_n \|H\| \right\} \\ &\leq \inf_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x; \xi)] + O(1) \cdot t_n. \end{aligned}$$

Thus, for any $y_n \in S^\star_{P_0 + t_n H_n}$ we have $y_n \in S_{P_0}(ct_n)$ for a constant $c < \infty$. Thus each subsequence of $y_n$ has a further subsequence converging to some $x^\star \in S^\star_{P_0}$ by the assumed compactness of $S_{P_0}(\epsilon)$, and the dominated convergence theorem implies that $\|\ell(y_n; \cdot) - \ell(x^\star; \cdot)\|_{L^2(P_0)} \to 0$ if $y_n \to x^\star$. In particular, we find that

$$\liminf_n \mathbb{E}_{P_0}[\ell(y_n; \xi)] = \mathbb{E}_{P_0}[\ell(x^\star; \xi)]$$

for any $x^\star \in S^\star_{P_0}$. Letting $y_n \in S^\star_{P_0 + t_n H_n}$, then, we have

$$T_{\mathrm{opt}}(P_0 + t_n H_n) - T_{\mathrm{opt}}(P_0) \geq \mathbb{E}_{P_0}[\ell(y_n; \xi)] + t_n H_n(y_n) - \mathbb{E}_{P_0}[\ell(y_n; \xi)] = t_n H_n(y_n).$$

Moving to a subsequence if necessary along which $y_n \to x^\star$, we have $H_n(y_n) - H(x^\star) \leq \|H_n - H\| + |H(y_n) - H(x^\star)| \to 0$, where we have used that $H$ is continuous with respect to $L^2(P_0)$. Thus $t_n H(y_n) \geq t_n H(x^\star) - o(t_n)$, which gives the lower bound (39).

**C.3. Proof of Theorem 3** We prove only the asymptotic result for the upper confidence bound $u_n$, as the proof of the lower bound is completely parallel. By Theorem 7, we have that

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

where we recall the definition (12) of the Gaussian processes $H_+$ and $H_-$. Applying the delta method as in the proof of Theorem 2 (see Section C.1 and Lemma 17, noting that this is essentially equivalent to the continuity of the infimum operator in the sup-norm topology) we obtain

$$\sqrt{n} \left( u_n - \inf_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x; \xi)] \right) \Rightarrow \inf_{x \in S^\star_{P_0}} H_+(x)$$

where $S^\star_{P_0} = \arg\min_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x; \xi)]$. This is equivalent to the first claim of the theorem, and the result when $S^\star_{P_0}$ is a singleton is immediate.

**C.4. Proof of Lemma 4** We first show the calculation to derive expression (19) of the conjugate $f_k^*$. For $k > 1$, we when $t \geq 0$ we have

$$\frac{\partial}{\partial t}\left[st - f_k(t)\right] = s - \frac{1}{2(k-1)}(t^{k-1} - 1).$$

If $s < 0$, then the supremum is attained at $t = 0$, as the derivative above is $< 0$ at $t = 0$. If $s \geq -\frac{1}{2(k-1)}$, then we solve $\frac{\partial}{\partial t}\left[st - f_k(t)\right] = 0$ to find $t = ((k-1)s/2 + 1)^{1/(k-1)}$, and substituting gives

$$st - f(t) = \frac{2}{k}\left(\frac{k-1}{2}s + 1\right)^{\frac{k}{k-1}} - \frac{2}{k}$$

which is our desired result as $1 - 1/k = 1/k_*$. When $k < 1$, a completely similar proof gives the result.

We now turn to computing the supremum in the lemma. For shorthand, let $Z = \ell(x; \xi)$. By the duality result of Lemma 3, for any $P_0$ and $\rho \geq 0$ we have

$$
\begin{aligned}
\sup_{P \in D_f(P\|P_0) \leq \rho} \mathbb{E}_P[Z] &= \inf_{\lambda \geq 0, \eta}\left\{\lambda\mathbb{E}_{P_0}\left[f_k^*\left(\frac{Z - \eta}{\lambda}\right)\right] + \lambda\rho + \eta\right\}\\
&= \inf_{\lambda \geq 0, \eta}\left\{2^{1-k_*}\frac{(k-1)^{k_*}}{k}\lambda^{1-k_*}\mathbb{E}_{P_0}\left[\left(Z - \eta + \frac{2\lambda}{k-1}\right)_+^{k_*}\right] + \lambda\left(\rho - \frac{2}{k}\right) + \eta\right\}\\
&= \inf_{\lambda \geq 0, \widetilde{\eta}}\left\{\frac{2^{1-k_*}(k-1)^{k_*}}{k}\lambda^{1-k_*}\mathbb{E}_{P_0}\left[(Z - \widetilde{\eta})_+^{k_*}\right] + \lambda\left(\rho + \frac{2}{k(k-1)}\right) + \widetilde{\eta}\right\}
\end{aligned}
$$

where in the final equality we set $\widetilde{\eta} = \eta - \frac{2\lambda}{k-1}$, because $\eta$ is unconstrained. Taking derivatives with respect to $\lambda$ to infimize the preceding expression, we have (noting that $\frac{k_*-1}{k_*} = \frac{1}{k}$)

$$2\left(\frac{k-1}{2\lambda}\right)^{k_*}\frac{1-k_*}{k}\mathbb{E}_{P_0}\left[(Z - \widetilde{\eta})_+^{k_*}\right] + \left(\rho + \frac{2}{k(k-1)}\right) = 0$$

$$\text{or}\quad \lambda = 2^{\frac{1}{k}}(k-1)(2 + \rho k(k-1))^{-\frac{1}{k_*}}\mathbb{E}_{P_0}\left[(Z - \widetilde{\eta})_+^{k_*}\right]^{\frac{1}{k_*}}.$$

Substituting $\lambda$ into the preceding display and mapping $\rho \mapsto \rho/n$ gives the claim of the lemma.

**Appendix D: Proofs for Dependent Sequences** In this section, we present proofs of our results on dependent sequences (Example 3, Theorem 4, Proposition 2, and Theorem 9). We begin by giving a proof of claims in Example 3 in Section D.1. Then, for logical consistency, we first present the proof of the general result Theorem 9 in Section D.2, which is an extension of Theorem 8 to $\beta$-mixing sequences. We apply this general result for Hadamard differentiable functionals to stochastic optimization problems $T_{\mathrm{opt}}(P) = \inf_{x \in \mathcal{X}}\mathbb{E}_P[\ell(x; \xi)]$ and prove Theorem 4 in Section D.2. Finally, we prove Proposition 2 in Section D.4, a sectioning result that provides exact coverages even for dependent sequences.

**D.1. Proof of Example 3** First, we note from Meyn and Tweedie [54, Theorem 15.0.1] that $\{\xi_n\}_{n\geq 0}$ is aperiodic, positive Harris recurrent and geometrically ergodic. Letting $\pi$ be the stationary distribution of $\{\xi_n\}$, it follows that for some $s \in (0, 1)$ and $R \in (0, \infty)$, we have

$$\sum_{n=1}^{\infty} s^n \|P^n(z, \cdot) - \pi(\cdot)\|_w \leq Rw(z) \quad \text{for all} \quad z \in \Xi \tag{40}$$

where the distance $\|P - Q\|_w$ between two probabilities $P$ and $Q$ is given by

$$\|P(\cdot) - Q(\cdot)\|_w := \sup \left\{ \left| \int_\Xi f(y) P(dy) - \int_\Xi f(y) Q(dy) \right| : f \text{ measurable}, \ |f| \le w \right\}.$$

Now, let $\{A_i\}_{i \in \mathcal{I}}$ and $\{B_j\}_{j \in \mathcal{J}}$ be finite partitions of $\Xi$ such that $A_i, B_j \in \mathcal{A}$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$. By definition, the $\beta$-mixing coefficient can be written as

$$
\begin{aligned}
\beta_n &= \frac{1}{2} \sup \sum_{i \in \mathcal{I}, j \in \mathcal{J}} |\mathbb{P}_\pi \left( X_0 \in A_i, \xi_n \in B_j \right) - \pi \left( A_i \right) \pi \left( B_j \right)| \\
&= \frac{1}{2} \sup \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \left| \int_{A_i} \left( \mathbb{P}_z(\xi_n \in B_j) - \pi(B_j) \right) \pi(dz) \right| = \frac{1}{2} \sup \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \left| \int_{A_i} \nu_n(z, B_j) \pi(dz) \right| \qquad (41)
\end{aligned}
$$

where $\nu_n(z, \cdot)$ is the signed measure on $(\Xi, \mathcal{A})$ given by $\nu_n(z, B) := \mathbb{P}_z(\xi_n \in B) - \pi(B)$. From the Hahn-Jordan decomposition theorem, there exists a positive and negative set, $P_{n,z}$ and $N_{n,z}$, for the signed measure $\nu_n(z, \cdot)$ so that we can write

$$\nu_n(z, B) = \nu_n(z, B \cap P_{n,z}) + \nu_n(z, B \cap N_{n,z}) \ \text{ for all } \ B \in \mathcal{A}.$$

Now, note that

$$
\begin{aligned}
\sum_{j \in \mathcal{J}} |\nu_n(z, B_j)| &= \sum_{j \in \mathcal{J}} \nu_n(z, B_j \cap P_{n,z}) - \sum_{j \in \mathcal{J}} \nu_n(z, B_j \cap N_{n,z}) \\
&= \nu_n(z, P_{n,z}) - \nu_n(z, N_{n,z}) = 2\nu_n(z, P_{n,z})
\end{aligned}
$$

where second equality follows since $\{B_j\}_{j \in \mathcal{J}}$ is a partition of $\Xi$ and the last inequality follows from definition of $\nu_n(z, \cdot)$.

Since $w \ge 1$, we further have that $2\nu_n(z, P_{n,z}) \le 2 \|P^n(x, \cdot) - \pi(\cdot)\|_w$. Collecting these bounds, we have from inequality (40) that $\sum_{j \in \mathcal{J}} |\nu_n(z, B_j)| \le s^n R w(z)$. From the representation (41), we then obtain $\beta_n \le s^{-n} R \mathbb{E}_\pi w(\xi_0)$. Now, from the Lyapunov conditions

$$\mathbb{E}_z w(\xi_1) \le \gamma w(z) + b \ \text{ for all } \ z \in \Xi$$

where we let $b := \sup_{z' \in C} \mathbb{E}_{z'} w(\xi_1)$. By taking expectations over $\xi_0 \sim \pi$, note that $\mathbb{E}_\pi w(\xi_0) \le b/(1 - \gamma) < \infty$ (see, for example, Glynn and Zeevi [36]). This yields our final claim $\beta_n = O(s^n)$.

**D.2. Proof of Theorem 9** Armed with Lemma 1 and its uniform counterpart given in Theorem 7 (the proof goes through, *mutatis mutandis*, under the hypotheses of Theorem 9), we proceed similarly as in the proof of Theorem 8. Only now, Lemma 5 implies that

$$\sqrt{n} \left( \mathbb{E}_{\widehat{P}_n}[T^{(1)}(\xi; P_0)] - \mathbb{E}_{P_\xi}[T^{(1)}(\xi; P_0)] \right) \Rightarrow N \left( 0, \Gamma(T^{(1)}, T^{(1)}) \right),$$

we have

$$\sqrt{n} \left( \sup_{P : D_f(P \| \widehat{P}_n) \le \rho/n} T(P) - T(P_0) \right) \Rightarrow \sqrt{\rho \mathrm{Var} \left( T^{(1)}(\xi; P_0) \right)} + N \left( 0, \Gamma(T^{(1)}, T^{(1)}) \right) \qquad (42)$$

and

$$\mathbb{P} \left( T(P_0) \le \sup_P \{ T(P) \mid D_f(P \| \widehat{P}_n) \le \rho/n \} \right) \to \mathbb{P} \left( W \ge -\sqrt{\frac{\rho \mathrm{Var}_{P_\xi} \left( T^{(1)}(\xi; P_0) \right)}{\Gamma(T^{(1)}, T^{(1)})}} \right),$$

where $W \sim \mathsf{N}(0, 1)$. From a symmetric argument for inf, we obtain the desired result.

### D.3. Proof of Theorem 4

***Case 1:*** $\xi_0 \sim \pi$    We first show the result for $\xi_0 \sim \pi$. Since $P \mapsto T_{\mathrm{opt}}(P) = \inf_{x \in \mathcal{X}} \mathbb{E}_P[\ell(x; \xi)]$ is Hadamard differentiable by Lemma 17, it suffices to verify the hypothesis of Theorem 9. Note from the discussion following Theorem 1 in Doukhan et al. [29] that if $\mathcal{H} \subset L^{2r}(\pi)$ satisfies $\sum_{n \geq 1} n^{\frac{1}{r-1}} \beta_n < \infty$ and

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|_{L^{2r}(\pi)})} \, d\epsilon < \infty,$$

then the hypotheses of Lemma 5 holds. Since the other assumptions of Theorem 9 hold from Lemma 17, we need only show that this bracketing integral is finite. Conveniently, that $\ell(\cdot; \xi)$ is $M(\xi)$-Lipschitz by Assumption 2 implies [82, Chs. 2.7.4 & 3.2]

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|_{L^{2r}(\pi)})} \leq C \|M\|_{L^{2r}(\pi)} \int_0^1 \sqrt{d \log \epsilon^{-1}} \, d\epsilon < \infty$$

for a compact set $\mathcal{X} \subset \mathbb{R}^d$.

***Case 2:*** $\xi_0 \sim \nu$ ***for general measures*** $\nu$.    As $\ell(\cdot; \xi)$ is continuous, we can ignore issues of outer measure and treat convergence in the space $\mathcal{C}(\mathcal{X})$ of continuous functions on $\mathcal{X}$. We will show

$$\sqrt{n} \left( \sup_{P : D_f(P\|\widehat{P}_n) \leq \frac{\rho}{n}} T(P) - T(P_0) \right) \Rightarrow W := \mathsf{N} \left( \sqrt{\rho \mathrm{Var}_\pi T^{(1)}(\xi; P_0)}, \Gamma(T^{(1)}, T^{(1)}) \right) \quad (43)$$

under any initial distribution $\xi_0 \sim \nu$. The result for infima of $T(P)$ over $\{P : D_f(P\|\widehat{P}_n) \leq \rho/n\}$ is analogous, so that this implies the theorem.

We abuse notation and let $W_n(\xi_j^{j+n}) = \sqrt{n}(\sup_{P : D_f(P\|\widehat{P}_n) \leq \frac{\rho}{n}} T(P) - T(P_0))$, except that we replace the empirical $\widehat{P}_n$ with the empirical distribution over $\xi_j, \ldots, \xi_{j+n}$. To show the limit (43), it suffices to show $W_n(\xi_{m_n}^{n+m_n}) \Rightarrow W$ for appropriate increasing sequences $m_n$:

**LEMMA 18.** *For any initial distribution* $\xi_0 \sim \nu$, $W_n(\xi_{m_n}^{n+m_n}) - W_n(\xi_0^n) \overset{a.s.}{\to} 0$ *whenever* $m_n \to \infty$ *and* $m_n/\sqrt{n} \to 0$.

*Proof.* By Lemma 13, there exists $C > 0$ depending only on the choice of $f$ and $\rho$ such that

$$\left| W_n(\xi_{m_n}^{n+m_n}) - W_n(\xi_0^n) \right| \leq \sqrt{n} \frac{C}{n} \sup_{x \in \mathcal{X}} \sum_{i=0}^{m_n} |\ell(x; \xi_i) - \ell(x; \xi_{n+i})|$$

$$\leq \frac{C m_n}{\sqrt{n}} \frac{1}{m_n} \sum_{i=0}^{m_n} \left( \sup_{x \in \mathcal{X}} |\ell(x; \xi_i)| + \sup_{x \in \mathcal{X}} |\ell(x; \xi_{n+i})| \right).$$

By hypothesis, we have $\sup_{x \in \mathcal{X}} |\ell(x; \xi)| \leq |\ell(x_0; \xi)| + M(\xi) \mathrm{diam}(\mathcal{X})$ where $\mathbb{E}|\ell(x_0; \xi)| + \mathbb{E}[M(\xi)] < \infty$, and as $\{M(\xi_i)\}_{i=1}^\infty$ are $\beta$-mixing, the law of large numbers holds for any initial distribution [54, Proposition 17.1.6]. Then $\frac{1}{\sqrt{n}} \sum_{i=0}^{m_n} \sup_{x \in \mathcal{X}} |\ell(x; \xi_i)| \overset{a.s.}{\to} 0$ so long as $m_n/\sqrt{n} \to 0$.

Fix an arbitrary initial distribution $\xi_0 \sim \nu$. Letting $m_n = n^{1/4}$, Case 1 and Lemma 18 yield

$$W_n(\xi_{m_n}^{n+m_n}) \Rightarrow W \quad \text{when} \quad \xi_0 \sim \pi.$$

Let $\mathcal{L}_\nu(\xi_{m_n}^{n+m_n})$ denote the law of $(\xi_{m_n}, \ldots, \xi_{n+m_n})$ when $\xi_0 \sim \nu$, and let $Q^n(\xi, \cdot)$ be the distribution of $\xi_n$ conditional on $\xi_0 = \xi$ and $\nu \circ Q^m = \int Q^m(\xi, \cdot) d\nu(\cdot)$. The Markov property then implies

$$\left\| \mathcal{L}_\nu(\xi_{m_n}^{n+m_n}) - \mathcal{L}_\pi(\xi_0^n) \right\|_{\mathrm{TV}} = \| \nu \circ Q^{m_n} - \pi \circ Q^{m_n} \|_{\mathrm{TV}}.$$

By positive Harris recurrence and aperiodicity [54, Theorem 13.0.1], $\| \nu \circ Q^{m_n} - \pi \circ Q^{m_n} \|_{\mathrm{TV}} \to 0$ for any $m_n \to \infty$. We conclude that $W_n(\xi_{m_n}^{n+m_n}) \Rightarrow W$ for any $\nu$; Lemma 18 gives the final result.

**D.4. Proof of Proposition 2** We first show the result for $\nu = \pi$. The general result follows by a similar argument as in the second part of the proof of Theorem 4, which we omit for conciseness. To ease notation, define $N_b^j = \sqrt{b}(U_b^j - T(\pi))$. From the proof of Theorem 9, we have the asymptotic expansion

$$N_b^j = \sqrt{b}\left(\frac{1}{b}\sum_{k=1}^{b}\ell(x^\star;\xi_{(j-1)b+k}) - \mathbb{E}_\pi[\ell(x^\star;\xi)]\right) + \sqrt{\rho \mathrm{Var}_\pi \ell(x^\star;\xi)} + \epsilon_{b,j}$$

where $\epsilon_{b,j}$ is a remainder term that satisfies $\epsilon_{b,j} \xrightarrow{p} 0$ as $b \to \infty$. From Cramer's device [13], we have that $(N_b^j)_{j=1}^m$ jointly converges in distribution to a normal distribution with marginals given by

$$N_b^j \Rightarrow \sqrt{\rho \mathrm{Var}_\pi \ell(x^*;\xi)} + N\left(0, \sigma_\pi^2\right)$$

for all $j = 1, \ldots, m$. If we can show that $N_b^j$ have asymptotic covariance equal to 0, then we have

$$\frac{\frac{1}{m}\sum_{j=1}^m N_b^j - \sqrt{\rho \mathrm{Var}_\pi \ell(x^*;\xi)}}{\sqrt{b} s_m^2(U_b)} \Rightarrow T_{m-1}$$

by the continuous mapping theorem. Since $\mathrm{Var}_{\widehat{P}_n}\ell(x_n^*;\xi) \xrightarrow{p} \mathrm{Var}_\pi \ell(x^*;\xi)$ from Corollary 1, this gives our desired result. We now show that $N_b^j$ have asymptotic covariance equal to 0.

Since $\beta$-mixing coefficients upper bound their strongly mixing counterparts, we have from Ethier and Kurtz [34, Corollary 2.5.5]

$$\mathrm{Cov}_\pi(N_b^1, N_b^j) \le 2^{2r+1} \beta_b^{1-1/r} \left(\mathbb{E}_\pi |N_b^j|^{2r}\right)^{\frac{1}{2r}}$$

for $j \ge 3$ (we deal with $j = 2$ case separately below). The below lemma controls moments of $N_b^j$.

LEMMA 19. *Let Assumption 2 hold with $\mathbb{E}_\pi[M(\xi)^{2r}] < \infty$. Then, for all $j = 1, \ldots, m$,*

$$\mathbb{E}_\pi[|N_b^j|^{2r}] \le C_{f,\rho,r,\mathcal{X}} \left(\mathbb{E}_\pi\left[M(\xi)^{2r}\right] + \mathbb{E}_\pi\left[|\ell(x_0;\xi)|^{2r}\right]\right)$$

*where* $\mathrm{diam}(\mathcal{X}) = \sup_{x,x'} \|x - x'\|$ *and* $C_{f,\rho,r,\mathcal{X}}$ *is a constant that only depends on $f$, $\rho$, $r$ and* $\mathrm{diam}(\mathcal{X})$.

We defer the proof of the lemma to Section D.4.1. We conclude that $\mathrm{Cov}_\pi(N_b^1, N_b^i) \to 0$ for $i \ge 3$.

To show that $\mathrm{Cov}_\pi(N_b^1, N_b^2) \to 0$, define $N_{b,\epsilon_b}^2$ identically as $N_b^2$, except now we leave $\lfloor \epsilon_b b \rfloor$ number of samples in the beginning. Letting $\epsilon_b = 1/\sqrt{b}$, we still obtain $\mathrm{Cov}_\pi(N_b^1, N_{b,\epsilon_b}^2) \to 0$ from an identical argument as above. Since $\mathrm{Cov}_\pi(N_b^1, N_{b,\epsilon_b}^2) - \mathrm{Cov}_\pi(N_b^1, N_b^2) \to 0$ by dominated convergence theorem, we obtain the result.

**D.4.1. Proof of Lemma 19** First, we consider the following decomposition

$$N_b^j = \sqrt{b}\left(\sup_{P \in \mathcal{P}_{n,\rho,j}} T(P) - T(\widehat{P}_b^j)\right) + \sqrt{b}\left(T(\widehat{P}_b^j) - T(\pi)\right). \tag{44}$$

To bound the moments of the first term, note that

$$0 \le \sqrt{b}\left(\sup_{P \in \mathcal{P}_{n,\rho,j}} T(P) - T(\widehat{P}_b^j)\right) = \sqrt{b}\left(\inf_{x \in \mathcal{X}} \sup_{P \in \mathcal{P}_{n,\rho,j}} \mathbb{E}_P[\ell(x;\xi)] - \inf_{x \in \mathcal{X}} \mathbb{E}_{\widehat{P}_b^j}[\ell(x;\xi)]\right)$$

$$\le \sqrt{b} \sup_{x \in \mathcal{X}}\left(\sup_{P \in \mathcal{P}_{n,\rho,j}} \mathbb{E}_P[\ell(x;\xi)] - \mathbb{E}_{\widehat{P}_b^j}[\ell(x;\xi)]\right) \tag{45}$$

From Lemma 13, for some constant $C_f$ depending only on $f$, we have

$$\mathcal{P}_{n,\rho,j} \subseteq \left\{ P \ll \widehat{P}_b^j : D_{\chi^2}(P\|\widehat{P}_b^j) \leq \frac{C_f \rho}{n} \right\} =: \mathcal{P}_{2,n,\rho,j}.$$

The right hand side of the bound (45) is then bounded by

$$\sqrt{b} \sup_{x \in \mathcal{X}} \left( \sup_{P \in \mathcal{P}_{2,n,\rho,j}} \mathbb{E}_P[\ell(x;\xi)] - \mathbb{E}_{\widehat{P}_b^j}[\ell(x;\xi)] \right).$$

Now, the following lemma bounds the error term in the variance expansion (8).

LEMMA 20 ([30, 57, **Theorem 1**]). *Let* $f(t) = \frac{1}{2}(t-1)^2$. *Then*

$$\sup_P \left\{ \mathbb{E}_P[Z] : D_f(P\|\widehat{P}_n) \leq \frac{\rho}{n} \right\} \leq \mathbb{E}_{\widehat{P}_n}[Z] + \sqrt{\frac{2\rho}{n} \mathrm{Var}_{\widehat{P}_n}(Z)}.$$

We conclude that

$$0 \leq \sqrt{b} \left( \sup_{P \in \mathcal{P}_{n,\rho,j}} T(P) - T(\widehat{P}_b^j) \right) \leq \sup_{x \in \mathcal{X}} \sqrt{2 C_f \rho \mathrm{Var}_{\widehat{P}_b^j}(\ell(x;\xi))}$$

$$\leq 2\sqrt{C_f \rho} \left( \mathbb{E}_{\widehat{P}_b^j} |\ell(x_0;\xi)| + \mathrm{diam}(\mathcal{X}) \mathbb{E}_{\widehat{P}_b^j}[M(\xi)] \right)$$

and hence

$$b^r \mathbb{E} \left| \sup_{P \in \mathcal{P}_{n,\rho,j}} T(P) - T(\widehat{P}_b^j) \right|^{2r} \leq 2^{4r-1} (C_f \rho)^r \left( \mathbb{E}_\pi |\ell(x_0;\xi)|^{2r} + \mathrm{diam}(\mathcal{X})^{2r} \mathbb{E}_\pi[M(\xi)]^{2r} \right). \tag{46}$$

To bound the second term in the decomposition (44), note that

$$\sqrt{b} \left| T(\widehat{P}_b^j) - T(\pi) \right| = \sqrt{b} \left| \inf_{x \in \mathcal{X}} \mathbb{E}_{\widehat{P}_b^j}[\ell(x;\xi)] - \inf_{x \in \mathcal{X}} \mathbb{E}_\pi[\ell(x;\xi)] \right|$$

$$\leq \sqrt{b} \sup_{x \in \mathcal{X}} \left| \mathbb{E}_{\widehat{P}_b^j}[\ell(x;\xi)] - \mathbb{E}_\pi[\ell(x;\xi)] \right|$$

Now, from a standard symmetrization argument [82, Section 2.3], we have

$$b^r \mathbb{E} \left[ \sup_{x \in \mathcal{X}} \left| \mathbb{E}_{\widehat{P}_b^j}[\ell(x;\xi)] - \mathbb{E}_\pi[\ell(x;\xi)] \right|^{2r} \right] \leq \mathbb{E} \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{\sqrt{b}} \sum_{i=1}^b \epsilon_i \ell(x;\xi_i) \right|^{2r} \right]$$

where $\epsilon_i$'s are i.i.d. random signs so that $\mathbb{P}(\epsilon_i = +1) = \mathbb{P}(\epsilon_i = -1) = \frac{1}{2}$. The following standard chaining bound controls the right hand side [82].

LEMMA 21. *Under the conditions of Proposition 2, for* $j = 1, \ldots, m$,

$$\mathbb{E}_\epsilon \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{\sqrt{b}} \sum_{i=1}^b \epsilon_i \ell(x;\xi_i) \right|^{2r} \right]$$

$$\leq C_r \left( d^r \mathrm{diam}(\mathcal{X}) \|M(\xi)\|_{L^2(\widehat{P}_b^j)}^{2r} + \left( \sqrt{d \, \mathrm{diam}(\mathcal{X})} + 1 \right)^{2r} \|\ell(x_0;\xi)\|_{L^2(\widehat{P}_b^j)}^{2r} \right)$$

*for a constant* $C_r > 0$ *depending only on* $r \geq 1$.

Taking expectations with respect to $\xi_i$'s in the preceeding display, we obtain

$$
\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{\sqrt{b}}\sum_{i=1}^{b}\epsilon_i\ell(x;\xi_i)\right|^{2r}\right]
$$
$$
\leq C_r\left(d^r\operatorname{diam}(\mathcal{X})\,\mathbb{E}[M(\xi)^{2r}]+\left(\sqrt{d\operatorname{diam}(\mathcal{X})}+1\right)^{2r}\mathbb{E}[|\ell(x_0;\xi)|^{2r}]\right).
$$

Using the bound (46) to bound the first term in the decomposition (44), and using the preceeding display to bound the second term, we obtain the final result.

**Appendix E: Proofs of Consistency Results**  In this appendix, we collect the proofs of the major theorems in Section 5 on consistency of minimizers of the robust objective (4a).

**E.1.  Proof of Theorem 5**  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 use Lemma 13 to prove the theorem. Let $\epsilon>0$ be as in Assumption 5, and define $p$ and $q$ by $q=\min\{2,1+\epsilon\}$, $p=\max\{2,1+\frac{1}{\epsilon}\}$. Then defining the likelihood ratio $L(\xi):=\frac{dP}{d\widehat{P}_n}(\xi)$ and $\mathcal{L}_{n,\rho}$ the likelihood ratio set corresponding to $\mathcal{P}_{n,\rho}$, we have

$$
\sup_{P\in\mathcal{P}_{n,\rho}}|\mathbb{E}_P[\ell(x;\xi)]-\mathbb{E}_{P_0}[\ell(x;\xi)]|
$$
$$
\leq\sup_{L\in\mathcal{L}_{n,\rho}}\mathbb{E}_{\widehat{P}_n}\left[|L(\xi)-1|\,\ell(x;\xi)\right]+\left|\mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)]-\mathbb{E}_{P_0}[\ell(x;\xi)]\right|
$$
$$
\leq\sup_{L\in\mathcal{L}_{n,\rho}}\mathbb{E}_{\widehat{P}_n}\left[|L(\xi)-1|^p\right]^{1/p}\cdot\mathbb{E}_{\widehat{P}_n}[|\ell(x;\xi)|^q]^{1/q}+\left|\mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)]-\mathbb{E}_{P_0}[\ell(x;\xi)]\right| \tag{47}
$$

where inequality (47) is a consequence of Hölder's inequality. Applying Lemma 13, we have that

$$
\mathbb{E}_{\widehat{P}_n}[|L(\xi)-1|^p]^{1/p}=n^{-1/p}\,\|np-\mathbb{1}\|_p\leq n^{-1/p}\,\|np-\mathbb{1}\|_2\leq n^{-1/p}\sqrt{\frac{\rho}{\gamma_f}}
$$

where $\gamma_f$ is as in the lemma. Combining this inequality with Assumption 5, the first term in the upper bound (47) goes to 0. Since the second term converges uniformly to 0 (in outer probability) by the Glivenko-Cantelli property, the desired result follows.

**E.2.  Proof of Theorem 6**  Before giving the proof proper, we provide a few standard definitions that are useful.

DEFINITION 7.  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

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

Moreover, we write $A_n\to A$ if $\limsup_n A_n=\liminf_n A_n=A$.
The last definition of convergence of $A_n\to A$ is *Painlevé-Kuratowski convergence*. With this definition, we may define epigraphical convergence of functions.

DEFINITION 8.  Let $g_n:\mathbb{R}^d\to\mathbb{R}$ be a sequence of functions, and $g:\mathbb{R}^d\to\mathbb{R}$. Then $g_n$ *epi-converges to a function $g$* if

$$
\operatorname{epi}g_n\to\operatorname{epi}g \tag{48}
$$

in the sense of Painlevé-Kuratowski convergence, where $\operatorname{epi}g=\{(x,r)\in\mathbb{R}^d\times\mathbb{R}:g(x)\leq r\}$.

We use $g_n \overset{\text{epi}}{\to} g$ to denote the epi-convergence of $g_n$ to $g$. If $g$ is proper (meaning that $\operatorname{dom} g \neq \varnothing$), the following lemma characterizes epi-convergence for closed convex functions.

LEMMA 22 (**Rockafellar and Wets [69], Theorem 7.17**). *Let $g_n, g$ be convex, proper, and lower semi-continuous. The following are equivalent.*

(i) $g_n \overset{\text{epi}}{\to} g$

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

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

$$\lim_{n \to \infty} \sup_{x \in C} |g_n(x) - g(x)| = 0.$$

The last characterization says that epi-convergence is equivalent to uniform convergence on compacta. Before moving to the proof of the theorem, we give one more useful result.

LEMMA 23 (**Rockafellar and Wets [69], Theorem 7.31**). *Let $g_n \overset{\text{epi}}{\to} g$, where $g_n$ and $g$ are extended real-valued functions and $\inf_x g(x) \in (-\infty, \infty)$. Then $\inf_x g_n(x) \to \inf_x g(x)$ if and only if for all $\epsilon > 0$, there exists a compact set $C$ such that*

$$\inf_{x \in C} g_n(x) \leq \inf_x g_n(x) + \epsilon \quad \text{eventually.}$$

We now show that the sample-based robust upper bound converges to the population risk. For notational convenience, based on a sample $\xi_1, \ldots, \xi_n$ (represented by the empirical distribution $\widehat{P}_n$), define the functions

$$F(x) := \mathbb{E}_{P_0}[\ell(x;\xi)] \quad \text{and} \quad \widehat{F}_n(x) := \sup_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(x;\xi)] : D_f(P \| \widehat{P}_n) \leq \frac{\rho}{n} \right\}.$$

These are both closed convex: $F$ by [10] and $\widehat{F}_n$ as it is the supremum of closed convex functions. We now show condition (ii) of Lemma 22 holds. Indeed, let $\epsilon > 0$ be such that $\mathbb{E}_{P_0}[|\ell(x;\xi)|^{1+\epsilon}] < \infty$ for all $x \in \mathcal{X}$, and define $q = \min\{2, 1+\epsilon\}$ and $p = \max\{2, 1+\epsilon^{-1}\}$ to be its conjugate. Then the bound (47) in the proof of Theorem 5 implies that for any $x \in \mathcal{X}$ we have

$$|F(x) - \widehat{F}_n(x)| \leq n^{-1/p} \sqrt{\frac{\rho}{\gamma_f}} \mathbb{E}_{\widehat{P}_n}[|\ell(x;\xi)|^q]^{\frac{1}{q}} + \left| \mathbb{E}_{\widehat{P}_n}[\ell(x;\xi)] - \mathbb{E}_{P_0}[\ell(x;\xi)] \right|.$$

The strong law of large numbers and continuous mapping theorem imply that $\mathbb{E}_{\widehat{P}_n}[|\ell(x;\xi)|^q]^{1/q} \overset{a.s.}{\to} \mathbb{E}_{P_0}[|\ell(x;\xi)|^q]^{1/q}$ for each $x$, and thus for each $x \in \mathcal{X}$, we have $\widehat{F}_n(x) \overset{a.s.}{\to} F(x)$. Letting $\widehat{\mathcal{X}}$ denote any dense but countable subset of $\mathcal{X}$, we then have

$$\widehat{F}_n(x) \to F(x) \quad \text{for all } x \in \widehat{\mathcal{X}}$$

except on a set of $P_0$-probability 0. This is condition (ii) of Lemma 22, whence we see that

$$\widehat{F}_n \overset{\text{epi}}{\to} F \quad \text{with probability 1.}$$

With these convergence guarantees, we prove the claims of the theorem. Let us assume that we are on the event that $\widehat{F}_n \overset{\text{epi}}{\to} F$, which occurs with probability 1. For simplicity of notation and with no loss of generality, we assume that $F(x) = \widehat{F}_n(x) = \infty$ for $x \notin \mathcal{X}$. By Assumption 3, the sub-level sets $\{x \in \mathcal{X} : \mathbb{E}_{P_0}[\ell(x;\xi)] \leq \alpha\}$ are compact, and $S_{P_0}^\star = \arg\min_{x \in \mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$ is non-empty ($F$ is closed), convex, and compact. Let $C \subset \mathbb{R}^d$ be a compact set containing $S_{P_0}^\star$ in its interior. We may then apply Lemma 22(iii) to see that

$$\sup_{x \in C} |\widehat{F}_n(x) - F(x)| \to 0.$$

Now, we claim that $S^\star_{\widehat{P}_n} \subset \operatorname{int} C$ eventually. Indeed, because $F$ is closed and $S^\star_{P_0} \subset \operatorname{int} C$, we know that on the compact set $\operatorname{bd} C$, we have $\inf_{x \in \operatorname{bd} C} F(x) > \inf_x F(x)$. The uniform convergence of $\widehat{F}_n$ to $F$ on $C$ then implies that eventually $\inf_{x \in \operatorname{bd} C} \widehat{F}_n(x) > \inf_{x \in C} \widehat{F}_n(x)$, and thus $S^\star_{\widehat{P}_n} \subset \operatorname{int} C$. This shows that for any sequence $x_n \in S^\star_{\widehat{P}_n}$, the points $x_n$ are eventually in the interior of any compact set $C \supset S^\star_{P_0}$ and thus $\sup_{x_n \in S^\star_{\widehat{P}_n}} \operatorname{dist}(x_n, S^\star_{P_0}) \to 0$.

The argument of the preceding paragraph shows that any compact set $C$ containing $S^\star_{P_0}$ in its interior guarantees that, on the event $\widehat{F}_n \overset{\mathrm{epi}}{\to} F$, we have $\inf_{x \in C} \widehat{F}_n(x) \leq \inf_x \widehat{F}_n(x) + \epsilon$ and $S^\star_{\widehat{P}_n} \subset \operatorname{int} C$ eventually. Applying Lemma 23 gives that $\inf_x \widehat{F}_n(x) \overset{P^*}{\to} \inf_x F(x)$ as desired. To show the second result, we note that from the continuous mapping theorem [82, Theorem 1.3.6] and $\widehat{F}_n \overset{P^*}{\to} F$ uniformly on $C$

$$\limsup_{n \to \infty} \mathbb{P}^* \left( d_{\subset}(S^\star_{\widehat{P}_n}, S^\star_{P_0}) \geq \epsilon \right) \leq \limsup_{n \to \infty} \mathbb{P}^* \left( \inf_{x \in S^{\star\epsilon}_{P_0}} \widehat{F}_n(x) > \inf_{x \in \mathcal{X}} \widehat{F}_n(x) \right)$$
$$= \mathbb{P}^* \left( \inf_{x \in S^{\star\epsilon}_{P_0}} F(x) > \inf_{x \in \mathcal{X}} F(x) \right) = 0$$

where $A^\epsilon = \{x : \operatorname{dist}(x, A) \leq \epsilon\}$ denotes the $\epsilon$-enlargement of $A$.