Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach

John C. Duchi , Peter W. Glynn , Hongseok Namkoong

To cite this article:
John C. Duchi , Peter W. Glynn , Hongseok Namkoong (2021) Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. Mathematics of Operations Research 46(3):946-969. https://doi.org/10.1287/moor.2020.1085

http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 46, No. 3, August 2021, pp. 946–969
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach

John C. Duchi,[a] Peter W. Glynn,[b] Hongseok Namkoong[c]

[a] Department of Electrical Engineering and Statistics, Stanford University, Stanford, California 94305; [b] Department of Management Science and Engineering, Stanford University, Stanford, California 94305; [c] Decision, Risk, and Operations Division, Columbia Business School, New York City, NY 10027

Contact: jduchi@stanford.edu (JCD); glynn@stanford.edu (PWG); namkoong@gsb.columbia.edu, https://orcid.org/0000-0002-5708-4044 (HN)

Received: July 13, 2018
Revised: October 3, 2019
Accepted: July 29, 2020
Published Online in Articles in Advance: January 27, 2021

MSC2010 Subject Classification: Primary: 62Mxx inference from stochastic processes
OR/MS Subject Classification: Primary: Programming/stochastic; secondary: statistics

https://doi.org/10.1287/moor.2020.1085

Abstract. 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.

Funding: All authors were supported by SAIL-Toyota Center for AI Research. J. C. Duchi and H. Namkoong were supported by a Sloan Fellowship. J. C. Duchi was supported by National Science Foundation award NSF-CAREER-1553086 and the Office of Naval Research Young Investigator Program award N00014-19-2288. H. Namkoong was supported by a Samsung Fellowship.
Supplemental Material: The online appendices are available at https://doi.org/10.1287/moor.2020.1085.

Keywords: stochastic optimization • robust optimization • empirical likelihood

1. Introduction

We study statistical properties of distributionally robust solution methods for the stochastic optimization problem

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

In the formulation (1), the feasible region $\mathcal{X} \subset \mathbb{R}^d$ is a nonempty closed set, ξ is a random vector on the probability space $(\Xi, \mathcal{A}, P_0)$, where the domain Ξ is a (subset of) a separable metric space, and the function $\ell : \mathcal{X} \times \Xi \to \mathbb{R}$ is a lower-semicontinuous (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 (Shapiro et al. [94]) approximates the population objective (1) using the sample average approximation (SAA) based on a sample $\xi_1, \ldots, \xi_n \overset{\text{iid}}{\sim} P_0$,

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

where $\hat{P}_n$ 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 (Ben-Tal et al. [7, 9], Bertsimas et al. [14], Delage and Ye [33]), which allow us to provide confidence intervals with asymptotically exact coverage for optimal values of the problem (1). Our approach further yields approximate solutions $\hat{x}_n$ 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_{opt} that acts on probability distributions on Ξ by

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

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

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

This exact coverage indicates that 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 (Ali and Silvey [1], Csiszár [30]). For a lower-semicontinuous convex function $f : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ satisfying $f(1) = 0$, the *f-divergence* between probability distributions P and Q on Ξ is

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

where μ is a σ-finite measure such that P and Q are absolutely continuous with respect to μ ($P, Q \ll \mu$), 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

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

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

yield the 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 χ^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. [9] 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; Ben-Tal et al. [7]) 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 in (3) give a principled method for choosing the size ρ 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 $T_{\text{opt}}(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$. We show how the construction (4a)–(4b) of $[l_n, u_n]$ gives a confidence interval with exact coverage (3) for $T_{\text{opt}}(P_0)$ when the minimizer of $\mathbb{E}_{P_0}[\ell(x;\xi)]$ is unique. To do so, we extend Owen's empirical likelihood theory (Owen [78, 79]) 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 ℓ and P_0,

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

This shows that the robust optimization problem (4a), which is efficiently computable when ℓ 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

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

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 nonconvex in x, even if ℓ is convex), reducing uncertainty. We give weak conditions under which the expansion is uniform in x, showing that the regularization interpretation is valid when we choose $\hat{x}_n$ to minimize the robust formulation (4a).

iv. Lastly, we prove consistency of estimators $\hat{x}_n$ 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

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

the distance from any point in $S^\star_{\hat{P}_n}$ to $S^\star_{P_0}$ tends to zero so long as ℓ has more than one moment under P_0 and is lower semicontinuous.

As we show in Section 3.1, the generalized empirical likelihood confidence interval $[l_n, u_n]$ is tighter than the confidence interval generated from the central limit approximation based on the SAA [see inequality (12) and its discussion]. This tightening comes at the cost of undercoverage in small samples, as we observe our simulation experiments (Section 6). To address poor coverage in small-sample or high-dimensional scenarios, two of the authors have extended the results of this paper (see Duchi and Namkoong [40]) to provide finite sample guarantees for the upper bound u_n. Letting $u_n(x) = \sup_{P \ll \hat{P}_n} \{\mathbb{E}_P[\ell(x;\xi)] : D_f\left(P \| \hat{P}_n\right) \le \rho/n\}$, we have for problem-dependent constants C_1, C_2, with probability at least $1 - \exp(-C_1\rho)$,

$$\mathbb{E}_{P_0}[\ell(x;\xi)] \le u_n(x) + \frac{C_2\rho}{n} \quad \text{simultaneously for all } x \in \mathcal{X}.$$

The uniformity in these guarantees (Duchi and Namkoong [40, theorem 3]) means that the constants C_1, C_2 must grow with some notion of the complexity of ℓ and $\mathcal{X}$, such as localized Rademacher complexities (Bartlett et al. [5]); these finite sample bounds thus lack the asymptotically exact coverage (i.e. pivotal, problem-independent behavior) that we develop for u_n. Although developing a deeper understanding of finite sample behavior of robust estimators is important, such results are complementary to our asymptotic guarantees.

1.1. Background and Prior Work

The nonparametric inference framework for stochastic optimization that we develop in this paper is the empirical likelihood counterpart of the normality theory that Shapiro develops in [90] and [92]. Although an extensive literature exists on statistical inference for stochastic optimization problems (see, e.g., the line of work of Dupacová and Wets [42], King [56], King and Wets [58], King and Rockafellar [57], Shapiro [90, 91, 92, 93], Shapiro et al. [94]), Owen's empirical likelihood framework in [80] has received little attention in the stochastic optimization literature, save for notable recent exceptions (Lam [62], Lam and Zhou [63]). 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 $C_{\rho,n} := \{\mathbb{E}_P[Y] : D_f\left(P \| \hat{P}_n\right) \le \frac{\rho}{n}\}$, where $f(t) = -2\log t$. Empirical likelihood theory shows that if we set $\rho = \chi^2_{d,1-\alpha} := \inf\{\rho' : \mathbb{P}(\|Z\|_2^2 \le \rho') \ge 1 - \alpha \text{ for } Z \sim \mathsf{N}(0, I_{d\times d})\}$, then $C_{\rho,n}$ is an asymptotically exact $(1-\alpha)$-confidence region; that is, $\mathbb{P}(\mathbb{E}_{P_0}[Y] \in C_{\rho,n}) \to 1 - \alpha$. Through a self-normalization property, empirical likelihood requires no knowledge or estimation of unknown quantities, such as variance. We show that 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 $\ell(x;\xi) \ge 0$, $l_n \ge 0$ (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. [7, 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 by Ben-Tal et al. [9], Bertsimas et al. [13, 14], Delage and Ye [33], Gupta [47], and Wang et al. [99]. The duality of confidence regions and hypothesis tests (Lehmann and Romano [65]) gives a natural connection between robust optimization, uncertainty sets, and statistical tests. Delage and Ye [33] made initial progress in this

direction by constructing confidence regions based on mean and covariance matrices from the data, and Jiang and Guan [54] expanded this line of research to other moment constraints. Bertsimas et al. [14, 13] developed uncertainty sets based on various linear and higher-order moment conditions. They also propose a robust SAA formulation based on goodness of fit tests, showing tractability as well as some consistency results based on Scarsini's linear convex orderings in [87], so long as the underlying distribution is bounded; they further give confidence regions that do not have exact coverage. Gupta [47] provided a Bayesian perspective for choosing uncertainty sets for stochastic constraints, proposing a particular notion of (Bayesian) asymptotic optimality and tractable approximations. The formulation (4) has similar motivation to the preceding works, as the uncertainty set

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

is a confidence region for $\mathbb{E}_{P_0}[\ell(x;\xi)]$ for each fixed $x \in \mathcal{X}$ (as we show in the sequel). Our results 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. [9] explored a similar scenario to ours, focusing on the robust formulation (4a), and they showed that, when P_0 is finitely supported, the robust program (4a) gives a one-sided confidence interval with (asymptotically) inexact coverage (i.e., they only give a bound on the coverage probability). In the unconstrained setting $\mathcal{X} = \mathbb{R}^d$, Lam and Zhou [63] 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 in [9] and Bertsimas et al.'s goodness of fit testing-based procedures in [13] provide natural motivation for formulations similar to ours in (4). However, by considering completely nonparametric measures of fit, we can depart from assumptions on the structure of Ξ (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 [61, 62] developed variance expansions similar to ours in (5), focusing on the Kullback–Leibler (KL) divergence and empirical likelihood cases (i.e., $f(t) = -2\log t$ with independent and identically distributed [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 (Blanchet and Murthy [17], Esfahani and Kuhn [43], Pflug and Wozabal [81], Shafieezadeh-Abadeh et al. [88], Sinha et al. [95], Wozabal [100]) 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 (Fournier and Guillin [45]), Esfahani and Kuhn [43] gave finite sample guarantees with nonparametric rates $O(n^{-1/d})$, in particular, showing consistency guarantees for Wasserstein-based robust formulations. The f-divergence formulation that 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 (Esfahani and Kuhn [43], Shafieezadeh-Abadeh [88]) only under stringent conditions on the functional $\xi \mapsto \ell(x;\xi)$. On the other hand, efficient solution methods (Ben-Tal et al. [9], Namkoong and Duchi [72]) for the robust problem (4a) are obtainable without restriction on the objective function $\ell(x;\xi)$ other than convexity in x.

A literature somewhat orthogonal to the distributionally robust approach that we take considers confidence regions for either the optimal population value or its solution. In the first vein, Lan et al. [64] and Mak et al. [67] provided upper and lower bounds on the optimal value (1) using finite sample concentration results; both results are conservative, thus allowing strong finite sample coverage guarantees but necessarily failing to achieve asymptotically exact coverage. In the second vein of providing confidence regions for the solution $x^\star$ itself, several researchers propose using the iterates of a stochastic approximation algorithm (Bubeck et al. [22],

Chen et al. [25], Li et al. [66], Mandt et al. [68]). These results apply both in Bayesian (Bubeck et al. [22], Mandt et al. [68]) and frequentist inference (Chen et al. [25], Li et al. [66]).

1.2. 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 that $X_n \rightrightarrows X$ 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 (van der Vaart and Wellner [98, section 1.2]). Given a set $A \subset \mathbb{R}^d$, norm $\|\cdot\|$, and point x, the distance $\operatorname{dist}(x, A) = \inf_y\{\|x - y\| : y \in A\}$. The *inclusion distance*, or the *deviation*, from a set A to B is

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

For a measure μ on a measurable space $(\Xi, \mathcal{A})$ and $p \geq 1$, we let $L^p(\mu)$ be the usual L^p space; that is, $L^p(\mu) := \{f : \Xi \to \mathbb{R} \mid \int |f|^p d\mu < \infty\}$. For a deterministic or random sequence $a_n \in \mathbb{R}$, we say that a sequence of random variables X_n is $O_P(a_n)$ if $\lim_{c\to\infty} \limsup_n P(|X_n| \geq c \cdot a_n) = 0$. Similarly, we say that $X_n = o_P(a_n)$ if $\limsup P(|X_n| \geq c \cdot a_n) = 0$ for all $c > 0$. For $\alpha \in [0, 1]$, we define $\chi^2_{d,\alpha}$ to be the α-quantile of a χ^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 (Hiriart-Urruty and Lemaréchal [50]). We let $\mathbf{I}_A(x)$ be the $\{0, \infty\}$-valued membership function, so $\mathbf{I}_A(x) = \infty$ if $x \notin A$, and 0 otherwise. To address measurability issues, we use outer measures and corresponding convergence notions (van der Vaart and Wellner [98, section 1.2–1.5]). Throughout the paper, the sequence $\{\xi_i\}$ is i.i.d. unless explicitly stated.

1.3. Outline

In order to highlight applications of our general results to stochastic optimization problems, we first present results for the optimal value functional $T_{\text{opt}}(P) := \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x; \xi)]$, 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 ρ in the uncertainty sets $\{P : D_f(P\|\hat{P}_n) \leq \rho/n\}$, choosing ρ 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 (Imbens [53], Newey and Smith [75], Owen [80]), 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 Ξ, and let $T : \mathcal{P} \to \mathbb{R}$ be the statistical quantity of interest. We typically consider $T_{\text{opt}}(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x; \xi)]$ 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

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

where $\hat{P}_n$ 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 $\hat{P}_n$. We may define a dual quantity, the profile divergence $R_n : \mathbb{R} \to \mathbb{R}_+$ (called the *profile likelihood* in Owen [80] when $f(t) = -2\log t$), by

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

Then, for any $P \in \mathcal{P}$, we have $T(P) \in C_{n,\rho}$ if and only if $R_n(T(P)) \le \frac{\rho}{n}$. Classical empirical likelihood (Owen [78, 79, 80]) considers $f(t) = -2\log t$ so that $D_f(P\|\hat{P}_n) = 2D_{\mathrm{kl}}(\hat{P}_n\|P)$, 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)

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

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

2.1. Generalized Empirical Likelihood for Means

In the classical case in which the vectors $Z_i \in \mathbb{R}^d$ and are i.i.d., Owen [78] showed 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 $d_0 \le d$, 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 χ^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 $\rho(\alpha) = \inf\{\rho' : \mathbb{P}(\chi^2_{d_0} \le \rho') \ge 1 - \alpha\}$. 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 in [78] to smooth f-divergences. Whereas the result is essentially known (Baggerly [4], Bertail et al. [11], Corcoran [28]), 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 $d_0 \le d$. Then,*

$$\lim_{n\to\infty} \mathbb{P}\Big(\mathbb{E}_{P_0}[Z] \in \Big\{\mathbb{E}_P[Z] \; : \; D_f\big(P\|\hat{P}_n\big) \le \frac{\rho}{n}\Big\}\Big) = \mathbb{P}\Big(\chi^2_{d_0} \le \rho\Big). \tag{7}$$

When $d = 1$, the proposition is a direct consequence of Lemma 1 to come; for more general dimensions d, we present the proof in Online Appendix B.5. If we consider the random variable $Z_x(\xi) := \ell(x;\xi)$, defined for each $x \in \mathcal{X}$, then 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

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

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 $s_n^2 = \mathbb{E}_{\hat{P}_n}[Z^2] - \mathbb{E}_{\hat{P}_n}[Z]^2$ denote the sample variance of Z. Then,*

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

where $\varepsilon_n \xrightarrow{a.s.} 0$.

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

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

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

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

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

yielding Proposition 1 in the case that $d = 1$. Concurrently with the original version of this paper, Lam [62] gave 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) that relies on the local Lipschitzness of our loss. Although 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 2. *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*

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

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

This theorem is a consequence of the more general uniform expansions that we develop in Section 7, in particular Theorem 9. 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 the 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., Rockafellar and Wets [85, theorem 1.9]).

Assumption 3. *The function $\ell(\cdot;\xi)$ is proper and lower-semicontinuous 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 that $\mathbb{E}_{P_0}[\ell(x;\xi)] \to \infty$ 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 3.2, 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 are i.i.d.

Theorem 3. *Let Assumptions 1 and 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* $\xi_i \stackrel{\text{iid}}{\sim} P_0$ *and the optimizer* $x^\star := \arg\min_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$ *is unique, then*

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

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_{opt}. See Online Appendix C for the full derivation.

There are of course many results on confidence sets for optimal values in stochastic programming. Shapiro [90, 92] developed a number of normal approximations and asymptotic normality theory for stochastic optimization problems. The normal analogue of Theorem 2 is that

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

which holds under the conditions of Theorem 2. The normal approximation (11) requires estimation of the unknown parameter $\text{Var}_{P_0}(\ell(x^\star;\xi))$; the plug-in estimator $\text{Var}_{\hat{P}_n}(\ell(\hat{x}_n;\xi))$ is a frequent choice, where $\hat{x}_n$ minimizes the sample average (2). The generalized empirical likelihood approach in Theorem 2 is asymptotically pivotal, so there are no hidden quantities that we must estimate.

In an approach more directly using empirical likelihood, Lam and Zhou [63] gave a result related to Theorem 2 for the special case that $f(t) = -2\log t$ when the domain $\mathcal{X} = \mathbb{R}^d$ (so the problem is unconstrained) and the loss $x \mapsto \ell(x;\xi)$ is differentiable for all $\xi \in \Xi$. They used first-order optimality conditions as an estimating equation and applied standard empirical likelihood theory (Owen [80]). This approach gives a nonpivotal asymptotic distribution; the limiting law is a χ_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 $x^\star \in \arg\min_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)]$. In contrast, Theorem 2 applies to problems with general constraints, as well as more general objective functions ℓ 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 Theorem 2 requires the minimizer of $\mathbb{E}_{P_0}[\ell(x;\xi)]$ to be unique.

We now argue that the generalized empirical likelihood confidence interval $[l_n, u_n]$ is typically tighter than its normal approximation counterpart (11). Let $\hat{x}_n$ be the solution to the sample average approximation (2). From the definition (4a) and the asymptotic expansion (10),

$$\begin{aligned} u_n &\le \sup_{P \ll \hat{P}_n} \left\{\mathbb{E}_P[\ell(\hat{x}_n;\xi)] \, : \, D_f\left(P\|\hat{P}_n\right) \le \frac{\rho}{n}\right\} \\ &= \mathbb{E}_{\hat{P}_n}[\ell(\hat{x}_n,\xi)] + \sqrt{\frac{\rho}{n}\text{Var}_{\hat{P}_n}(\ell(\hat{x}_n,\xi))} + \varepsilon_n(\hat{x}_n), \end{aligned} \tag{12}$$

where $\sqrt{n}\varepsilon_n(\hat{x}_n) \xrightarrow{P^*} 0$ under the conditions of Theorem 1. The term $\mathbb{E}_{\hat{P}_n}[\ell(\hat{x}_n,\xi)] + \sqrt{\frac{\rho}{n}\text{Var}_{\hat{P}_n}(\ell(\hat{x}_n,\xi))}$ is the normal upper confidence bound (11)—with the same confidence level as u_n—with the plug-in estimator $\text{Var}_{\hat{P}_n}(\ell(\hat{x}_n;\xi))$ for the asymptotic variance. As u_n minimizes over all x, it is tighter than its normal approximation counterpart when the inequality (12) is looser: $\varepsilon_n(\hat{x}_n) = o_p(n^{-1/2})$.

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 $l_n = \inf_P\{T_{\text{opt}}(P) \, : \, D_f(P\|\hat{P}_n) \le \rho/n\}$ and $u_n = \sup_P\{T_{\text{opt}}(P) \, : \, D_f(P\|\hat{P}_n) \le \rho/n\}$, which also shows a useful symmetry between the upper and lower bounds. The characterization depends on the excursions of a noncentered Gaussian process when $x^\star$ is nonunique, 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 noncentered processes H_+ and H_- by

$$H_+(x) := G(x) + \sqrt{\rho\text{Var}_{P_0}(\ell(x;\xi))} \quad \text{and} \quad H_-(x) := G(x) - \sqrt{\rho\text{Var}_{P_0}(\ell(x;\xi))}. \tag{13}$$

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 2. *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*

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

and

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

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

We defer the proof of the theorem to Online 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),then we may shorten the confidence set via a simple correction to the threshold ρ. Indeed, for a given desired confidence level $1-\alpha$, setting $\rho = \chi^2_{1,1-2\alpha}$ (which is smaller than $\chi^2_{1,1-\alpha}$) 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 3 to dependent sequences, including Harris-recurrent Markov chains mixing suitably quickly. To present our results, we first recall β-mixing sequences (Bradley [19], Ethier and Kurtz [44, sections 7.2–7.3]; also called *absolute regularity* in the literature).

Definition 1. The β-mixing coefficient between two sigma algbras $\mathcal{B}_1$ and $\mathcal{B}_2$ on Ξ is

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

where the supremum is over finite partitions $\{A_i\}_{i\in\mathcal{I}}$, $\{B_j\}_{j\in\mathcal{J}}$ of Ξ 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 σ-algebras

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

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

$$\beta_n := \beta(\mathcal{G}_0, \mathcal{G}_n). \tag{14}$$

A stationary sequence $\{\xi_i\}_{i\in\mathbb{Z}}$ is β-mixing if $\beta_n \to 0$ as $n \to \infty$. For Markov chains, β-mixing has a particularly nice interpretation: a strictly stationary Markov chain is β-mixing if and only if it is Harris-recurrent and aperiodic (Bradley [19, theorem 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 (Doukhan [38]).

Theorem 4. *Let $\{\xi_n\}_{n=0}^\infty$ be an aperiodic, positive Harris-recurrent Markov chain taking values on Ξ with stationary distribution π. 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 $\xi_0 \sim \nu$,*

$$\lim_{n\to\infty} \mathbb{P}_\nu\left(T_{\text{opt}}(\pi) \in \left\{T_{\text{opt}}(P) \,:\, D_f(P\|\hat{P}_n) \le \frac{\rho}{n}\right\}\right) = \mathbb{P}\left(\chi_1^2 \le \frac{\rho \text{Var}_\pi \ell(x^\star;\xi)}{\sigma_\pi^2(x^\star)}\right), \tag{15}$$

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

Theorem 4 is more or less a consequence of the general results that we prove in Section 7.3 on ergodic sequences, and we show how it follows from these results in Online 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 β-mixing (Meyn and Tweedie [70, 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 β-mixing processses that are not necessarily uniformly mixing. The following result is due to Mokkadem [71].

Example 2 (Geometric β-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 $\{\epsilon_n\}_{n\geq 1} \overset{\text{iid}}{\sim} F$, 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 β-mixing. That is, there exists $c \in (0,1)$ such that $\beta_n = O(s^n)$.

See Doukhan [37, section 2.4.1] for more examples of β-mixing processes.

Using the equivalence of geometric β-mixing and geometric ergodicity for Markov chains (Nummelin and Tweedie [77], Meyn and Tweedie [70, 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 ξ_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 ν 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*; Meyn and Tweedie [70, section 5.2].) Then, $\{\xi_n\}_{n\geq 0}$ is aperiodic, positive Harris-recurrent, and geometrically ergodic (Meyn and Tweedie [70, theorem 15.0.1]). Further, we can show that $\{\xi_n\}_{n\geq 0}$ is geometrically β-mixing: there exists $c \in (0,1)$ with $\beta_n = O(c^n)$. For completeness, we include a proof of this in Online Appendix D.1.

For dependent sequences, the asymptotic distribution in the limit (15) contains unknown terms such as σ_π^2 and $\text{Var}_\pi(\ell(x^\star;\xi))$; such quantities need to be estimated to obtain exact coverage. This loss of a pivotal limit occurs because $\sqrt{n}(\hat{P}_n - P_0)$ converges to a Gaussian process G on $\mathcal{X}$ with covariance function

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

whereas empirical likelihood self-normalizes based on $\text{Cov}_\pi(\ell(x_1;\xi_0), \ell(x_2;\xi_0))$. (These covariances are identical if ξ_i are i.i.d.) As a result, in Theorem 5, we no longer have the self-normalizing behavior of Theorem 3 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 $\hat{P}_b^j$ the empirical distribution on each of the blocks for $j = 1, \ldots, m$. Let

$$U_b^i := \sup_{P \ll \hat{P}_b^j} \left\{ T_{\text{opt}}(P) : D_f(P \| \hat{P}_b^j) \leq \frac{\rho}{n} \right\},$$

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

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

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

where T_{m-1} is the Student's t-distribution with $(m-1)$-degress of freedom.

See Online Appendix 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 $T_{\rm opt}(P_0)$. 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 $P \mapsto T_{\rm opt}(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ is continuous (on $\mathbb{R}^n$) because it is concave and finite-valued; as a consequence, the set

$$\left\{T_{\rm opt}(P) \,:\, D_f\left(P\|\hat{P}_n\right) \le \rho/n\right\} = \left\{\inf_{x\in\mathcal{X}} \sum_{i=1}^n p_i\ell(x;\xi_i) \,:\, p^\top\mathbb{1} = 1,\ p \ge 0,\ \sum_{i=1}^n f\left(np_i\right) \le \rho\right\}, \tag{16}$$

is an interval, and, in this section, we discuss a few methods for computing it. To compute the interval (16), 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 (Hiriart-Urruty and Lemaréchal [49, theorem VII.4.3.1]).

Lemma 2. *Let Assumptions 3 and 4 hold. Then,*

$$u_n = \inf_{x\in\mathcal{X}} \sup_{p\in\mathbb{R}^n}\left\{\sum_{i=1}^n p_i\ell(x;\xi_i) : p^\top\mathbb{1} = 1,\ p \ge 0,\ \sum_{i=1}^n f\left(np_i\right) \le \rho\right\}. \tag{17}$$

By strong duality, we can write the minimax problem (17) as a joint minimization problem. Recall that f^* denotes the Fenchel conjugate of f $f^*(s) := \sup_t\{st - f(t)\}$.

Lemma 3 (Ben-Tal et al. [9])**.** *The following duality holds:*

$$\sup_{P\ll\hat{P}_n}\left\{\mathbb{E}_P[\ell(x;\xi)] \,:\, D_f(P\|\hat{P}_n) \le \frac{\rho}{n}\right\} = \inf_{\lambda\ge0,\eta\in\mathbb{R}}\left\{\mathbb{E}_{\hat{P}_n}\left[\lambda f^*\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\right] + \frac{\rho}{n}\lambda + \eta\right\}. \tag{18}$$

When $x \mapsto \ell(x;\xi)$ is convex in x, the minimization (17) is a convex problem because it is the supremum of convex functions. The reformulation (18) shows that we can compute u_n by solving a problem jointly convex in η, λ, 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 $\hat{l}_n$ 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 (Shapiro et al. [94, chapter 5]) guarantee that, under our conditions, the empirical minimizer $\hat{x}_n = \arg\min_{x\in\mathcal{X}}\mathbb{E}_{\hat{P}_n}[\ell(x;\xi)]$ satisfies $\hat{x}_n \xrightarrow{a.s.} x^\star$, where $x^\star = \arg\min_{x\in\mathcal{X}}\mathbb{E}_{P_0}[\ell(x;\xi)]$. Now, consider the one-step estimator

$$\hat{l}_n := \inf_{P\,:\,D_f\left(P\|\hat{P}_n\right)\le\rho/n} \mathbb{E}_P[\ell(\hat{x}_n;\xi)]. \tag{19}$$

Then, by Theorem 1, we have

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

because $\hat{x}_n$ is eventually in any open set (or set open relative to $\mathcal{X}$) containing $x^\star$. Standard limit results (van der Vaart and Wellner [98]) guarantee that $\mathrm{Var}_{\hat{P}_n}(\ell(\hat{x}_n;\xi)) \xrightarrow{a.s.} \mathrm{Var}_{P_0}(\ell(x^\star;\xi))$, because $x \mapsto \ell(x;\xi)$ is Lipschitzian by Assumption 2. Noting that $\mathbb{E}_{\hat{P}_n}[\ell(\hat{x}_n;\xi)] \le \mathbb{E}_{\hat{P}_n}[\ell(x^\star;\xi)]$, we thus obtain

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

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

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

Combining these results, we have 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

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

Summarizing, the one-step estimator (19) 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 $\hat{l}_n$ defined by expression (19) guarantees that

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

giving a computationally feasible and asymptotically pivotal statistic. We remark in passing that alternating by minimizing over $P : D_f(P\|\hat{P}_n) \le \rho/n$ and x (i.e., more than the single-step minimizer) simply gives a lower bound $\tilde{l}_n$ satisfying $l_n \le \tilde{l}_n \le \hat{l}_n$, 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 (Mak et al. [67], Shapiro et al. [94]), because $\inf_{x\in\mathcal{X}} \mathbb{E}[\ell(x;\xi)] \ge \mathbb{E}[\inf_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \ell(x;\xi_i)]$. The robust formulation (4a) addresses this optimism by looking at a worst-case objective based on the confidence region $\{P : D_f(P\|\hat{P}_n) \le \rho/n\}$. It is clear that the robust formulation (4a) is a coherent risk measure (Shapiro et al. [94, section 6.3]) of $\ell(x;\xi)$: it is convex, monotonic in the loss ℓ, equivariant to translations $\ell \mapsto \ell + a$, and positively homogeneous in ℓ. A number of authors have studied coherent risk measures (Artzner et al. [3], Krokhmal [60], Rockafellar and Uryasev [84], Shapiro et al. [94]), and we next study their connections to statistical confidence regions for the optimal population objective (1).

As a concrete example, we consider Krokhmal's higher-order generalizations in [60] of conditional value at risk, where, for $k_* \ge 1$ and a constant $c > 0$, the risk functional has the form

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

The risk R_{k_*} penalizes upward deviations of the objective $\ell(x;\xi)$ from a fixed value η, where the parameter $k_* \ge 1$ 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 (Krokhmal [60]).

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 [29] family 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

$$f_k(t) = \frac{2\left(t^k - kt + k - 1\right)}{k(k-1)} \quad \text{so} \quad f_k^*(s) = \frac{2}{k}\left[\left(\frac{k-1}{2}s + 1\right)_+^{k_*} - 1\right], \tag{20}$$

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 (20) includes divergences such as the χ^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 (18) 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 Online Appendix C.4.

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

$$\sup_{P \in \mathcal{P}_n} \mathbb{E}_P[\ell(x;\xi)] = \inf_{\eta \in \mathbb{R}}\left\{\left(1 + \frac{2k(k-1)\rho}{n}\right)^{\frac{1}{k}} \mathbb{E}_{\hat{P}_n}\left[\left(\ell(x;\xi) - \eta\right)_+^{k_*}\right]^{\frac{1}{k_*}} + \eta\right\}. \tag{21}$$

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 (21) depends only on ρ, 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

$$\sup_{P : D_f(P\|P_n) \leq \frac{\rho}{n}} \mathbb{E}_P[\ell(x;\xi)] = \mathbb{E}_{P_n}[\ell(x;\xi)] + \sqrt{\frac{\rho}{n}\mathrm{Var}_{P_n}(\ell(x;\xi))} + \varepsilon_n(x), \tag{22}$$

where $\sqrt{n}\sup_{x \in \mathcal{X}} |\varepsilon_n(x)| \xrightarrow{P^*} 0$. In a companion to this paper, Duchi and Namkoong [39, 73] explore the expansion (22) in substantial depth for the special case $f(t) = \frac{1}{2}(t-1)^2$. Equation (22) 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 (22) 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. [101] for standard supervised machine learning tasks (see also Ben-Tal et al. [7, chapter 12]), though these results consider uncertainty on the data vectors ξ themselves, rather than the distribution. Our approach yields a qualitatively different type of (approximate) regularization by variance. In our follow-up work (Duchi and Namkoong [39, 73]), we analyze finite sample performance of the robust solutions. The naive variance-regularized objective

$$\mathbb{E}_{\hat{P}_n}[\ell(x;\xi)] + \sqrt{\frac{\rho}{n}\mathrm{Var}_{\hat{P}_n}\ell(x;\xi)} \tag{23}$$

is neither convex (in general) nor coherent, so that the expansion (22) allows us to solve a convex optimization problem that approximately regularizes variance.

In some restricted situations, the variance-penalized objective (23) 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 in [69].

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

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

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

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

The robust approximation of Theorem 7 [and Equation (22)] shows that

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

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—that we have discussed thus far, 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 2. 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 (Rockafellar and Wets [85]) 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. [13] also provided consistency results for robust variants of sample average approximation based on goodness-of-fit tests, though they required a number of conditions on the domain Ξ of the random variables for their results (in addition to certain continuity conditions on ℓ). In our context, we abstract away from this by parameterizing our problems by the n-vectors $\{P : D_f(P \| \hat{P}_n) \le \rho/n\}$ 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

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

There are many conditions sufficient to guarantee Glivenko–Cantelli properties. Typical approaches include covering number bounds on $\mathcal{H}$ (van der Vaart and Wellner [98, section 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; [97, example 19.8])**.** Let $\mathcal{X}$ be compact, and let $\ell(\cdot; \xi)$ 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 β-mixing sequences (Nobel and Dembo [76]) [those with coefficients (14) $\beta_n \to 0$]. Our subsequent results thus apply to β-mixing sequences $\{\xi_i\}$.

If there is an envelope function for objective $\ell(x; \xi)$ that has more than one moment under P_0, then we can show that the robust risk converges uniformly to the population risk (compared with 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 that the class $\{\ell(x;\cdot) : x \in \mathcal{X}\}$ is Glivenko–Cantelli. Then,*

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

See Online 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:

$$S^\star_{P_0} := \arg\min_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \quad \text{and} \quad S^\star_{\hat{P}_n} := \arg\min_{x\in\mathcal{X}} \sup_{P \ll \hat{P}_n} \left\{ \mathbb{E}_P[\ell(x;\xi)] : D_f\left(P\|\hat{P}_n\right) \le \frac{\rho}{n} \right\}. \tag{24}$$

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 that $\ell(\cdot;\xi)$ is continuous on $\mathcal{X}$. Then,*

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

and $d_C(S^\star_{\hat{P}_n}, S^\star_{P_0}) \xrightarrow{P^*} 0$.

Proof. The first conclusion is immediate by Theorem 5 and Example 5. To show convergence of the optimal sets, we denote by $A^\epsilon = \{x : \operatorname{dist}(x, A) \le \epsilon\}$ the ϵ-enlargement of A. By the uniform convergence given in Theorem 5 and continuous mapping theorem (van der Vaart and Wellner [98, theorem 1.3.6]), for all $\epsilon > 0$,

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

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

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 (King and Wets [58], Rockafellar and Wets [85]), bypassing the aforementioned uniform convergence and compactness conditions. Analogous results hold for sample average approximation (Shapiro et al. [94, section 5.1.1]).

In the theorem, we let $S^\star_{P_0}$ and $S^\star_{\hat{P}_n}$ be the solution sets (24) 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 that 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 Online Appendix E.2.

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

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

and $d_C(S^\star_{\hat{P}_n}, S^\star_{P_0}) \xrightarrow{P^*} 0$.

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 $S^\star_{P_0}$ and the growth of $\mathbb{E}_{P_0}[\ell(x;\xi)]$ off of this set, which means that, eventually, $S^\star_{\hat{P}_n}$ is nearly contained in $S^\star_{P_0}$. When the $\{\xi_i\}$ are

not i.i.d., the pointwise strong law for $|\ell(x;\xi)|^{1+\epsilon}$ holds if the $\{\xi_i\}$ are strongly mixing (α-mixing; Ibragimov [52]), and 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 $\hat{l}_n$ of Equation (19) in Section 3.3. With these, we simulate the true coverage probability $\mathbb{P}(\inf_{x\in\mathcal{X}} \mathbb{E}_{P_0}[\ell(x;\xi)] \in [\hat{l}_n, u_n])$ (because we control the distribution P_0 and $\ell(x;\xi)$) of our confidence intervals, and we compare it to the nominal χ^2-confidence level $\mathbb{P}(\chi_1^2 \le \rho)$ that 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 χ^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 $\mathbb{P}(\chi_1^2 \le \rho) = 0.95$. 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 (18) and (19) to compute u_n and $\hat{l}_n$, respectively, we use the Julia package convex.jl (Udell et al. [96]). Each experiment consists of 1,250 independent replications for each of the sample sizes n that 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 χ^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 1,250 independent runs. We observed undercoverage in small sample regimes, and enumerate possible approaches to this issue in Section 8.

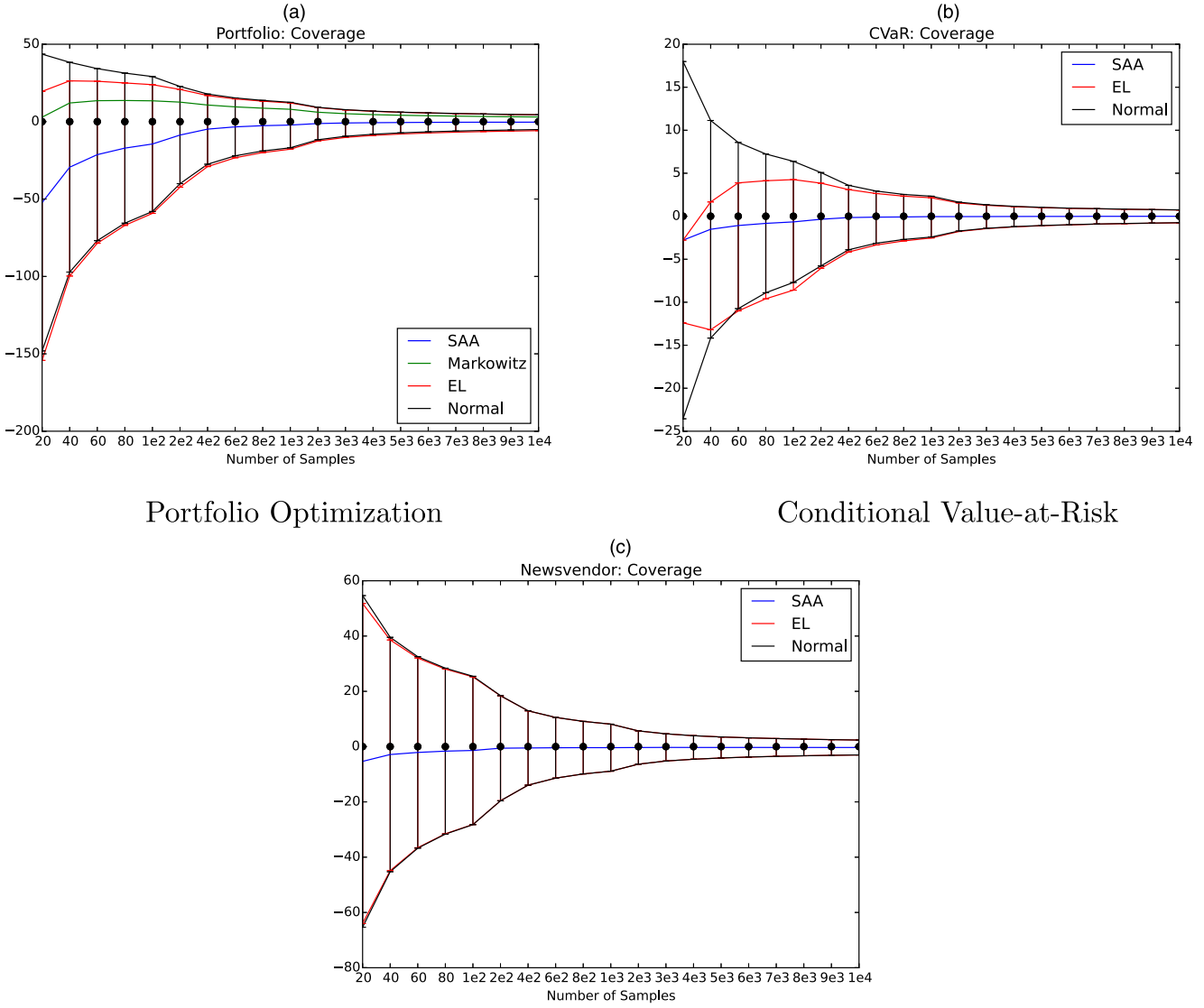
6.1. Portfolio Optimization

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 $\mathcal{X} = \{x \in \mathbb{R}^d \mid \mathbb{1}^\top x = 1, -10 \le x \le 10\}$ as our feasible region (allowing leveraged investments), and we simulate returns $\xi \overset{\text{iid}}{\sim} N(\mu, \Sigma)$. Within each simulation,

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
1,000	95.25	95.15	92.84	94.31	94.62	94.7	53.07	98.47	58.6	97.33
2,000	95.48	95.25	93.73	95.25	94.92	95.04	54.07	96.8	59.07	96.53
4,000	96.36	95.81	95.1	95.78	95.3	95.3	58.6	96	62.07	96.6
6,000	96.33	95.87	94.61	95	94.43	94.51	61.8	95.8	66.07	95.73
8,000	96.46	95.9	94.56	94.71	94.85	94.85	64.67	95.67	69	95.33
10,000	96.43	95.51	94.71	94.85	94.43	94.43	66.87	94.73	69.4	96.13
20,000							74.27	95.8	76.8	96.13
40,000							81.8	94.2	84.87	94.87
60,000							86.87	93.93	89.47	94.47
80,000							91.4	93.67	92.33	95
100,000							94.2	94.33	95.07	95.2

Figure 1. (Color online) 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 zero. The optimal objective computed from the sample average approximation ("SAA") has a negative bias.



the vector μ and covariance Σ are chosen as $\mu \sim N(0, I_d)$ and Σ 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

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

as the variance-regularized expression is efficiently minimizable (it is convex) in the special case of linear objectives. In Figure 1(a), 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 $0 = \mu^\top x^\star$, whereas 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 1,250 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 contains the true value $\mu^\top x^\star$. In comparison with the normal confidence interval, generalized empirical likelihood (EL) undercovers in small sample settings, which is consistent with previous observations for empirical likelihood (e.g., Owen [80, section 2.8]).

6.2. Conditional Value-at-Risk

For a real-valued random variable ξ, the *conditional value-at-risk* α (CVaR) is the expectation of ξ conditional on it taking values above its $1-\alpha$ quantile (Rockafellar and Uryasev [84]). More concisely, the CVaR (at α) is

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

Conditional value-at-risk is of interest in many financial applications (Rockafellar and Uryasev [84], Shapiro et al. [94]).

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 ξ 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 μ as negative returns, where σ^2 increases as μ increases, reminiscent of the oft-observed tendency in bear markets (the leverage effect) (Black [16], Christie [26]). 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 that 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 $\mathbb{E}_{P_0}[\xi \mid \xi \geq q_{1-\alpha}]$. 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 1(b)] 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 $j = 1, \ldots, d$, 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 ξ 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 $\xi \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \Sigma)$ (there may be negative demand), where Σ 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 $[\hat{l}_n, u_n]$ covers the true optimal value is still lower than the nominal 95%, though it is less pronounced than other cases. Figure 1(c) 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 Ξ, 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 $T_{\text{opt}}(P) = \inf_{x\in\mathcal{X}} \mathbb{E}_P[\ell(x;\xi)]$ (cf. Theorem 2). As Shapiro [92] noted, the general treatment for Hadamard differentiable functionals is necessary as Frechét differentiability is too stringent for studying constrained stochastic optimization. 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 [98] 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 μ on Ξ and $h : \Xi \to \mathbb{R}$, we use the functional shorthand $\mu h := \int h(\xi) d\mu(\xi)$ so that, for any probability measure, we have $Ph = \mathbb{E}_P[h(\xi)]$. Now, for a set $\mathcal{H} \subset L^2(P_0)$, let $\mathcal{L}^\infty(\mathcal{H})$ be the space of 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., van der Vaart and Wellner [98, section 1.2]). We then have the following definition (van der Vaart and Wellner [98, equation (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 $\sqrt{n}(\hat{P}_n - P_0) \Rightarrow G$ in the space $\mathcal{L}^\infty(\mathcal{H})$, where G is a tight Borel-measurable element of $\mathcal{L}^\infty(\mathcal{H})$, and $\hat{P}_n$ is the empirical distribution of $\xi_i \overset{\text{iid}}{\sim} P_0$.

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

Theorem 7. *Let the conditions of the preceding paragraph hold. Then,*

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

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

See Online Appendix B, in particular, Online 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 van der Vaart and Wellner [98, sections 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 let $\mathcal{H}$ 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 (van der Vaart and Wellner [98, section 3.9]). In the definition, we let $\mathcal{M}$ denote the space of signed measures on Ξ bounded with respect to $\|\cdot\|_{\mathcal{H}}$, noting that $\mathcal{M} \subset \mathcal{L}^\infty(\mathcal{H})$ via the mapping $\mu h = \int h(\xi) d\mu(\xi)$.

Definition 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 $K \subset B$,

$$\lim_{t \to 0} \sup_{H \in K, P+tH \in \mathcal{P}} \left| \frac{T(P+tH) - T(P)}{t} - dT_P(H) \right| = 0. \tag{25}$$

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

$$dT_P(Q - P) = \int_\Xi T^{(1)}(\xi; P) d(Q - P)(\xi) \tag{26}$$

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 van der Vaart [97, section 25.5] or [59]).

We now extend Proposition 1 to Hadamard-differentiable functionals $T : \mathcal{P} \to \mathbb{R}$. Owen [79] showed 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. [10, 11] 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 $\mathcal{L}^\infty(\mathcal{H})$.

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 $\sqrt{n}(\hat{P}_n - P_0) \Rightarrow G$ 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*

$$\lim_{n \to \infty} \mathbb{P}\left(T(P_0) \in \left\{T(P) : D_f(P \| P_n) \le \frac{\rho}{n}\right\}\right) = \mathbb{P}(\chi_1^2 \le \rho). \tag{27}$$

We use Theorem 7 to show the result in Online 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 a β-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 Ξ. We follow the approach of Doukhan et al. [38] to prove our results, giving bracketing number conditions sufficient for our convergence guarantees (alternative approaches are possible; see Arcones and Yu [2], Nobel and Dembo [76], Rio [83], Yu [102]).

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 *ϵ-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 ϵ-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 (van der Vaart and Wellner [98, theorem 2.5.6]). For β-mixing sequences, a modification of the $L^2(P_0)$-norm yields a similar result. To state the required bracketing condition in full, we first provide the 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 β_n are the mixing coefficients (14), and define the norm

$$\|h\|_{L^{2,\beta}(P_0)} = \sqrt{\int_0^1 \beta^{-1}(u) Q_h(u)^2 du}, \tag{28}$$

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 Γ be the covariance function

$$\Gamma(h_1, h_2) := \sum_{i\in\mathbb{Z}} \mathrm{Cov}(h_1(\xi_0), h_2(\xi_i)). \tag{29}$$

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

Lemma 5 (Doukhan et al. [38, theorem 1]). *Let $\{\xi\}_{i\in\mathbb{Z}}$ be a strictly stationary sequence of random vectors taking values in the Polish space Ξ 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*

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

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

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

where G is a Gaussian process with covariance function Γ and almost surely uniformly continuous sample paths.

The discussion following Doukhan et al. [38, 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 $\sum_{n\ge1} n^{\frac{1}{r-1}}\beta_n < \infty$, then 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 $\sqrt{n}(\hat{P}_n - P_0) \Rightarrow G$ 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 [Equation (26)] 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*

$$\lim_{n\to\infty} \mathbb{P}\left(T(P_0) \in \left\{T(P) : D_f(P\|P_n) \le \frac{\rho}{n}\right\}\right) = \mathbb{P}\left(\chi_1^2 \le \frac{\rho \mathrm{Var}_{P_\xi} T^{(1)}(\xi; P_0)}{\Gamma(T^{(1)}, T^{(1)})}\right). \tag{30}$$

See Online Appendix D.2 for the proof. We show in Online Appendix 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 (Ben-Tal et al. [7, 9]), 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 (Duchi and Namkoong [39]).

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 $S^{\star}_{P_0}$ 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. Whereas there is a substantial and developed literature on efficient methods for sample average approximation and stochastic gradient methods (Defazio et al. [32], Duchi et al. [41], Hazan [48], Johnson and Zhang [55], Nemirovski et al. [74], Polyak and Juditsky[82]), there are fewer established and computationally efficient solution methods for minimax problems of the form (4a) (though see the papers Ben-Tal et al. [8], Clarkson et al. [27], Namkoong and Duchi [72], Nemirovski et al. [74], and Shalev-Shwartz and Wexler [89], for work in this direction). Efficient solution methods need to be developed to scale up robust optimization.

As our results in Section 6 show, the confidence interval $[l_n, u_n]$ undercovers in small sample settings. We may use finite sample bounds to address this (Duchi and Namkoong [40]), but these are too conservative and

fail to achieve correct asymptotic coverage. Designing small sample corrections to the confidence interval $[l_n, u_n]$ to improve coverage—while maintaining asymptotical exactness—is an important open direction of research. We highlight a few possibilities here. One idea is to use Bartlett corrections; see DiCiccio et al. [34, 35] for Bartlett-correctability of empirical likelihood confidence intervals for smooth functions of means; Bartlett correctability of general Hadamard differentiable functionals remains open. Alternatively, it may be possible to generalize results on high-dimensional M-estimation, where the dimension d scales with the sample size n (e.g., Bean et al. [6], Candès and Sur [23], Donoho and Montanari [36]) to generalized empirical likelihood. In this context, extending the works by Chen et al. [24] and Hjort et al. [51,], which give limit theorems for high-dimensional estimating equations (with $d = o(\sqrt{n})$) to Hadamard-differentiable functionals may yield fruit; current analyses where $d/n \to c$ appear to require somewhat specialized data-generating distributions (Candès and Sur [23], Donoho and Montanari [36]). Approaches based on Wasserstein distances (Blanchet et al. [18]) may also address this issue, as the associated uncertainty sets need not contain the support only of the empirical distribution.

There are two ways of injecting robustness in the formulation (4a): increasing ρ 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 (20) and associated dual problems]. We characterize a statistically principled way of choosing ρ 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 that we study. Whereas the literature on higher-order corrections for empirical likelihood offers some answers for inference problems regarding the mean of a distribution (Baggerly [4], Bravo [20, 21], Corcoran [28], DiCiccio et al. [35]), the more complex settings arising in large-scale optimization problems leave a number of open questions.

Endnotes

[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 (Bertail [10, p. 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 $\mathbb{P}(Z \in A_n) = 0$, whereas $\mathbb{P}(Z \in B) = 1$.

References

[1] Ali SM, Silvey SD (1966) A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. B*. 28:131–142.
[2] Arcones MA, Yu B (1994) Central limit theorems for empirical and U-processes of stationary mixing sequences. *J. Theoret. Probab.* 7(1):47–71.
[3] Artzner P, Delbaen F, Eber J-M, Heath D (1999) Coherent measures of risk. *Math. Finance* 9(3):203–228.
[4] Baggerly KA (1998) Empirical likelihood as a goodness-of-fit measure. *Biometrika* 85(3):535–547.
[5] Bartlett PL, Bousquet O, Mendelson S (2005) Local Rademacher complexities. *Ann. Statist.* 33(4):1497–1537.
[6] Bean D, Bickel P, El Karoui N, Yu B (2013) Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* 110(36):14563–14568.
[7] Ben-Tal A, Ghaoui LE, Nemirovski A (2009) *Robust Optimization* (Princeton University Press, Princeton, NJ).
[8] Ben-Tal A, Hazan E, Koren T, Mannor S (2015) Oracle-based robust optimization via online learning. *Oper. Res.* 63(3):628–638.
[9] Ben-Tal A, den Hertog D, Waegenaere AD, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Sci.* 59(2):341–357.
[10] Bertail P (2006) Empirical likelihood in some semiparametric models. *Bernoulli* 12(2):299–331.
[11] Bertail P, Gautherat E, Harari-Kermadec H (2014) Empirical ϕ^* p-divergence minimizers for hadamard differentiable functionals. Akritas MG, Lahiri SN, Politis DN, eds. *Topics in Nonparametric Statistics* (Springer, New York), 21–32.
[12] Bertsekas DP (1973) Stochastic optimization problems with nondifferentiable cost functionals. *J. Optim. Theory Appl.* 12(2):218–231.
[13] Bertsimas D, Gupta V, Kallus N (2014) Robust sample average approximation. Preprint, submitted August 19, https://arxiv.org/abs/1408.4445.
[14] Bertsimas D, Gupta V, Kallus N (2018) Data-driven robust optimization. *Math. Programming* 167(2):235–292.
[15] Billingsley P (1986) *Probability and Measure*, 2nd ed. (Wiley, New York).
[16] Black F (1976) Studies of stock price volatility changes. *Proc. 1976 Meetings Amer. Statist. Assoc.* (American Statistical Association, Washington, DC), 177–181.
[17] Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* 44(2):565–600.
[18] Blanchet J, Kang Y, Murthy K (2019) Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.* 56(3):830–857.
[19] Bradley RC (2005) Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surveys* 2:107–144.
[20] Bravo F (2003) Second-order power comparisons for a class of nonparametric likelihood-based tests. *Biometrika* 90(4):881–890.
[21] Bravo F (2006) Bartlett-type adjustments for empirical discrepancy test statistics. *J. Statist. Planning Inference* 136(3):537–554.

[22] Bubeck S, Eldan R, Lehec J (2015) Finite-time analysis of projected Langevin Monte Carlo. Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 28 (Neural Information Processing Systems Foundation, San Diego), 1243–1251.
[23] Candès E, Sur P (2020) The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.* 48(1):27–42.
[24] Chen SX, Peng L, Qin YL (2009) Effects of data dimension on empirical likelihood. *Biometrika* 96(3):711–722.
[25] Chen X, Lee JD, Tong XT, Zhang Y (2020) Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.* 48(1):251–273.
[26] Christie AA (1982) The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *J. Financial Econom.* 10(4):407–432.
[27] Clarkson K, Hazan E, Woodruff D (2012) Sublinear optimization for machine learning. *J. ACM* 59(5):23.
[28] Corcoran SA (1998) Bartlett adjustment of empirical discrepancy statistics. *Biometrika* 85(4):967–972.
[29] Cressie N, Read TR (1984) Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. B* 46(3):440–464.
[30] Csiszár I (1967) Information-type measures of difference of probability distributions and indirect observation. *Studia Scientifica Mathematica Hungary* 2:299–318.
[31] Danskin JM (1967) *The Theory of Max-Min and Its Application to Weapons Allocation Problems* (Springer, Berlin).
[32] Defazio A, Bach F, Lacoste-Julien S (2014) SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 27 (Neural Information Processing Systems Foundation, San Diego), 1646–1654.
[33] Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.
[34] DiCiccio T, Hall P, Romano J (1988) Bartlett adjustment for empirical likelihood. Technical Report 298. Department of Statistics, Stanford University, Stanford, CA.
[35] DiCiccio T, Hall P, Romano J (1991) Empirical likelihood is Bartlett-correctable. *Ann. Statist.* 19(2):1053–1061.
[36] Donoho D, Montanari A (2016) High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probab. Theory Related Fields* 166(3–4):935–969.
[37] Doukhan P (1994) *Mixing, Properties and Examples* (Springer, New York).
[38] Doukhan P, Massart P, Rio E (1995) Invariance principles for absolutely regular empirical processes. *Annales de l'IHP probabilités et statistiques* 31(2):393–427.
[39] Duchi JC, Namkoong H (2016) Variance-based regularization with convex objectives. Preprint, submitted October 8, https://arxiv.org/abs/1610.02581.
[40] Duchi JC, Namkoong H (2019) Variance-based regularization with convex objectives. *J. Machine Learn. Res.* 20(68):1–55.
[41] Duchi JC, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. *J. Machine Learn. Res.* 12(61):2121–2159.
[42] Dupacová J, Wets R (1988) Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann. Statist.* 16(4):1517–1549.
[43] Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming* 171(1–2):115–166.
[44] Ethier SN, Kurtz TG (2009) *Markov Processes: Characterization and Convergence* (Wiley, New York).
[45] Fournier N, Guillin A (2015) On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields* 162(3–4):707–738.
[46] Glynn PW, Zeevi A (2008) Bounding stationary expectations of markov processes. Ethier SN, Feng J, Stockbridge RH, eds. *Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz* (Institute of Mathematical Statistics, Beachwood, OH), 195–214.
[47] Gupta V (2019) Near-optimal Bayesian ambiguity sets for distributionally robust optimization. *Management Sci.* 65(9):4242–4260.
[48] Hazan E (2016) Introduction to online convex optimization. *Foundations Trends Optim.* 2(3–4):157–325.
[49] Hiriart-Urruty J, Lemaréchal C (1993) *Convex Analysis and Minimization Algorithms I* (Springer, New York).
[50] Hiriart-Urruty J, Lemaréchal C (1993) *Convex Analysis and Minimization Algorithms II* (Springer, New York).
[51] Hjort NL, McKeague IW, Van Keilegom I (2009) Extending the scope of empirical likelihood. *Ann. Statist.* 37(3):1079–1111.
[52] Ibragimov IA (1962) Some limit theorems for stationary processes. *Theory Probab. Appl.* 7(4):349–382.
[53] Imbens G (2002) Generalized method of moments and empirical likelihood. *J. Bus. Econom. Statist.* 20(4):493–506.
[54] Jiang R, Guan Y (2016) Data-driven chance constrained stochastic program. *Math. Programming* 158(1–2):291–327.
[55] Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 26 (Neural Information Processing Systems Foundation, San Diego), 315–323.
[56] King AJ (1989) Generalized delta theorems for multivalued mappings and measurable selections. *Math. Oper. Res.* 14(4):720–736.
[57] King AJ, Rockafellar RT (1993) Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res.* 18(1):148–162.
[58] King AJ, Wets RJ (1991) Epi-consistency of convex stochastic programs. *Stochastics Stochastic Rep.* 34(1–2):83–92.
[59] Kosorok MR (2008) Introduction to empirical processes. *Introduction to Empirical Processes and Semiparametric Inference* (Springer, New York), 77–79.
[60] Krokhmal PA (2007) Higher moment coherent risk measures. *Quant. Finance* 7(4):373–387.
[61] Lam H (2016) Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.* 41(4):1248–1275.
[62] Lam H (2018) Sensitivity to serial dependency of input processes: A robust approach. *Management Sci.* 64(3):1311–1327.
[63] Lam H, Zhou E (2017) The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Oper. Res. Lett.* 45(4):301–307.
[64] Lan G, Nemirovski A, Shapiro A (2012) Validation analysis of robust stochastic approximation method. *Math. Programming* 134(2):425–458.

[65] Lehmann EL, Romano JP (2005) *Testing Statistical Hypotheses*, 3rd ed. (Springer, New York).
[66] Li T, Liu L, Kyrillidis A, Caramanis C (2018) Statistical inference using SGD. *Thirty-Second AAAI Conf. Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, Menlo Park, CA), 3571–3578.
[67] Mak W-K, Morton DP, Wood RK (1999) Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Oper. Res. Lett.* 24(1):47–56.
[68] Mandt S, Hoffman M, Blei D (2017) Stochastic gradient descent as approximate Bayesian inference. *J. Machine Learn. Res.* 18(134):1–35.
[69] Markowitz H (1952) Portfolio selection. *J. Finance* 7(1):77–91.
[70] Meyn S, Tweedie RL (2009) *Markov Chains and Stochastic Stability*, 2nd ed. (Cambridge University Press, New York).
[71] Mokkadem A (1990) Propriétés de mélange des processus autorégressifs polynomiaux. *Ann. Inst. Henri Poincaré Probab. Statist.* 26(2):219–260.
[72] Namkoong H, Duchi JC (2016) Stochastic gradient methods for distributionally robust optimization with f-divergences. Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 29 (Neural Information Processing Systems Foundation, San Diego), 2208–2216.
[73] Namkoong H, Duchi JC (2017) Variance-based regularization with convex objectives. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 30 (Neural Information Processing Systems Foundation, San Diego), 2971–2980.
[74] Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.
[75] Newey W, Smith R (2004) Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica* 72(1):219–255.
[76] Nobel A, Dembo A (1993) A note on uniform laws of averages for dependent processes. *Statist. Probab. Lett.* 17(3):169–172.
[77] Nummelin E, Tweedie RL (1978) Geometric ergodicity and r-positivity for general markov chains. *Ann. Probab.* 6(3):404–420.
[78] Owen A (1990) Empirical likelihood ratio confidence regions. *Ann. Statist.* 18(1):90–120.
[79] Owen AB (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75(2):237–249.
[80] Owen AB (2001) *Empirical Likelihood* (CRC Press, Boca Raton, FL).
[81] Pflug G, Wozabal D (2007) Ambiguity in portfolio selection. *Quant. Finance* 7(4):435–442.
[82] Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* 30(4):838–855.
[83] Rio E (2017) *Asymptotic Theory of Weakly Dependent Random Processes* (Springer, New York).
[84] Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. *J. Risk* 2(3):21–42.
[85] Rockafellar RT, Wets RJB (1998) *Variational Analysis* (Springer, New York).
[86] Römisch W (2005) Delta method, infinite dimensional. Kotz S, Read CB, Balakrishnan N, Vidakovic B, eds. *Encyclopedia of Statistical Sciences* (Wiley, Hoboken, NJ).
[87] Scarsini M (1999) Multivariate convex orderings, dependence, and stochastic equality. *J. Appl. Probab.* 35(1):93–103.
[88] Shafieezadeh-Abadeh S, Esfahani PM, Kuhn D (2015) Distributionally robust logistic regression. Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 28 (Neural Information Processing Systems Foundation, San Diego), 1576–1584.
[89] Shalev-Shwartz S, Wexler Y (2016) Minimizing the maximal loss: How and why? Balcan MF, Weinberger KQ, eds. *Proc. 33rd Internat. Conf. Machine Learn.* (Association for Computing Machinery, New York), 793–801.
[90] Shapiro A (1989) Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.* 17(2):841–858.
[91] Shapiro A (1990) On differential stability in stochastic programming. *Math. Programming* 47(1–3):107–116.
[92] Shapiro A (1991) Asymptotic analysis of stochastic programs. *Ann. Oper. Res.* 30(1):169–186.
[93] Shapiro A (1993) Asymptotic behavior of optimal solutions in stochastic programming. *Math. Oper. Res.* 18(4):829–845.
[94] Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM and Mathematical Programming Society, Philadelphia).
[95] Sinha A, Namkoong H, Volpi R, Duchi JC (2017) Certifiable distributional robustness with principled adversarial training. Preprint, submitted October 29, https://arxiv.org/abs/1710.10571.
[96] Udell M, Mohan K, Zeng D, Hong J, Diamond S, Boyd S (2014) Convex optimization in Julia. *First Workshop High Performance Tech. Comput. Dynam. Languages* (IEEE, New York), 18–28.
[97] van der Vaart AW (1998) *Asymptotic Statistics* (Cambridge University Press, New York).
[98] van der Vaart AW, Wellner JA (1996) *Weak Convergence and Empirical Processes with Applications to Statistics* (Springer, New York).
[99] Wang Z, Glynn P, Ye Y (2016) Likelihood robust optimization for data-driven problems. *Comput. Management Sci.* 13:241–261.
[100] Wozabal D (2012) A framework for optimization under ambiguity. *Ann. Oper. Res.* 193(1):21–47.
[101] Xu H, Caramanis C, Mannor S (2009) Robustness and regularization of support vector machines. *J. Machine Learn. Res.* 10:1485–1510.
[102] Yu B (1994) Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Probab.* 22(1):94–116.