The Empirical Likelihood Approach to Quantifying Uncertainty in Sample Average Approximation*

Henry Lam† Enlu Zhou‡

Abstract

We study the empirical likelihood approach to construct confidence intervals for the optimal value and the optimality gap of a given solution, henceforth quantify the statistical uncertainty of sample average approximation, for optimization problems with expected value objectives and constraints where the underlying probability distributions are observed via limited data. This approach relies on two distributionally robust optimization problems posited over the uncertain distribution, with a divergence-based uncertainty set that is suitably calibrated to provide asymptotic statistical guarantees.

Keywords: empirical likelihood; sample average approximation; confidence interval; constrained optimization; stochastic program; statistical uncertainty

1 Introduction

We consider a stochastic optimization problem in the form

$$\min_{x\in\Theta}\{h(x) := E[H(x;\xi)]\}, \tag{1}$$

where $x = (x_1, \ldots, x_p)$ is a continuous decision variable in the deterministic feasible region $\Theta \subseteq \mathbb{R}^p$, and ξ is a random vector on $\mathbb{R}^d$. We are interested in situations where the underlying probability distribution that controls the expectation $E[\cdot]$ is not fully known and can only be accessed via limited data $\xi_1, \ldots, \xi_n$. It is customary in this setting to work on an empirical counterpart of the problem, namely by solving the sample average approximation (SAA) (e.g., [16]):

$$\min_{x\in\Theta} \frac{1}{n}\sum_{i=1}^{n} H(x;\xi_i). \tag{2}$$

We further consider problems with expected value constraints, in the form

$$\begin{array}{ll} \min & h(x) = E[H(x;\xi)] \\ \text{subject to} & f_k(x) = E[F_k(x;\xi)] \le 0,\ k = 1,\ldots,m \\ & g_k(x) \le 0,\ k = 1,\ldots,s \end{array} \tag{3}$$

*A preliminary conference version of this work has appeared in [11]. Research of the first author was partially supported by the National Science Foundation under Grants CMMI-1400391/1542020 and CMMI-1436247/1523453. Research of the second author was partially supported by the National Science Foundation under Grant CAREER CMMI-1453934, and Air Force Office of Scientific Research under Grant YIP FA- 9550-14-1-0059.

†Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI. Email: khlam@umich.edu

‡H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA. Email: enlu.zhou@isye.gatech.edu

where $g_k(\cdot)$'s are deterministic functions. Thus (3) can include both stochastic and deterministic constraints. Again, under limited data $\xi_1, \ldots, \xi_n$, an SAA version of (3) is (e.g., [17])

$$\begin{array}{ll} \min & \frac{1}{n}\sum_{i=1}^n H(x;\xi_i) \\ \text{subject to} & \frac{1}{n}\sum_{i=1}^n F_k(x;\xi_i) \le 0,\ k=1,\ldots,m \\ & g_k(x) \le 0,\ k=1,\ldots,s \end{array} \tag{4}$$

Our premise is that beyond the n observations, new samples are not easily accessible because of either a lack of data or limited computational capacity in running further Monte Carlo simulation. The optimal value and solution obtained from (2) or (4) thus deviate from those under the genuine distribution in (1) or (3). Moreover, the error of the solution implies a non-zero optimality gap with the true optimal value, resulting in suboptimal decisions. Estimating these errors is important and has been studied over the years (e.g., [10], [12], Chapter 5 in [16]).

Our main contribution is to bring in a new approach to rigorously quantify the uncertainty in (2) and (4) through constructing confidence intervals (CIs) for the true optimal value and the optimality gap for a given solution. The machinery underlying our framework uses the so-called empirical likelihood (EL) method in statistics, and culminates at a reformulation of the problem of finding the upper and lower bounds of a CI into solving two optimization problems that closely resemble distributionally robust optimization (DRO). The uncertainty set in the DRO is a divergence-based ball cast over an uncertain probability distribution, where the size of the ball is suitably calibrated so that it provides asymptotic guarantees for the coverage probability of the resulting CI.

We study the theory giving rise to such guarantees. We demonstrate through several numerical examples that our method compares favorably with some existing methods, such as bounds using the central limit theorem (CLT) and the delta method, in terms of finite-sample performance. In the remainder of this paper, Sections 2 and 3 study the theory of our approach applied to the optimal value and the optimality gap, and our online Supplemental Material shows the numerical results and comparison with previous methods.

2 The Empirical Likelihood Method for Constructing Confidence Bounds for Optimal Values

This section studies in detail the EL method in constructing CIs for the optimal values. Section 2.1 focuses on (1) that only has deterministic constraints, and Section 2.2 generalizes to the stochastically constrained case (3).

2.1 Deterministically Constrained Optimization

Let us first fix some notations. Given the set of i.i.d. data $\xi_1, \xi_2, \ldots, \xi_n$, we denote a probability vector over $\{\xi_1, \ldots, \xi_n\}$ as $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$, where $\sum_{i=1}^n w_i = 1$ and $w_i \ge 0$ for all $i = 1, \ldots, n$. We denote $\chi^2_{q,\beta}$ as the $1-\beta$ quantile of a χ^2 distribution with degree of freedom q. We use "$\Rightarrow$" to denote convergence in distribution, and "a.s." to denote "almost surely".

Our method utilizes the optimization problems

$$\begin{array}{ll} \max / \min_w & \min_{x\in\Theta} \sum_{i=1}^n w_i H(x;\xi_i) \\ \text{subject to} & -2\sum_{i=1}^n \log(nw_i) \le \chi^2_{p+1,\beta} \\ & \sum_{i=1}^n w_i = 1 \\ & w_i \ge 0 \text{ for all } i = 1, \ldots, n \end{array} \tag{5}$$

where "max / min" denotes a pair of maximization and minimization. Note that the optimal value of the SAA problem (2) lies between those of (5).

The quantity $-(1/n)\sum_{i=1}^{n}\log(nw_i)$ can be interpreted as the Burg-entropy divergence ([15], [3]) between the probability distributions represented by the weights w and by the uniform weights $(1/n)_{i=1,\ldots,n}$ on the support $\{\xi_1,\ldots,\xi_n\}$. Thus, the first constraint in (5) is a Burg-entropy divergence ball centered at the uniform weights, with radius $\chi^2_{p+1,\beta}/(2n)$. From the viewpoint of DRO (e.g., [6, 3, 18]), the optimization problems in (5) output the worst-case estimates of $\min_{x\in\Theta}\{h(x) = E[H(x;\xi)]\}$ when $E[\cdot]$ is uncertain and its underlying distribution is believed to lie inside the divergence ball. We should point out, however, that this DRO interpretation differs from those in the existing literature (e.g., [4]), as our divergence ball (i.e. the "uncertainty set" in the terminology of robust optimization) may have low coverage of the true distribution P. This can be seen particularly when P is a continuous distribution, in which case the coverage of the divergence ball is zero because of the violation of the absolute continuity requirement needed in properly defining the divergence.

The EL method is a mechanism to endow statistical meaning to (5). In particular, it asserts that using the ball size $\chi^2_{p+1,\beta}/(2n)$ in (5) gives rise to statistically valid $1-\beta$ confidence bounds for the optimal value of (1) (despite that the ball may under-cover the true distribution). This method originates as a nonparametric analog of maximum likelihood estimation first proposed by [13]. On the data set $\{\xi_1,\ldots,\xi_n\}$, we first define a "nonparametric likelihood" $\prod_{i=1}^{n} w_i$, where w_i is a probability weight applied to each datum. It is straightforward to see that the maximum value of $\prod_{i=1}^{n} w_i$, among all w in the probability simplex, is $\prod_{i=1}^{n}(1/n)$. In fact, the same conclusion holds even if one allows putting weights outside the support of the data, which could only make the likelihood $\prod_{i=1}^{n} w_i$ smaller. In this sense, $\prod_{i=1}^{n}(1/n)$ can be viewed as a maximum likelihood in the nonparametric space. Correspondingly, we define the nonparametric likelihood ratio between the weights w and the maximum likelihood weights as $\prod_{i=1}^{n} w_i / \prod_{i=1}^{n}(1/n) = \prod_{i=1}^{n}(nw_i)$.

The key of the EL method is a nonparametric counterpart of the celebrated Wilks' Theorem [19] in parametric likelihood inference. The latter states that the ratio between the maximum likelihood and the true likelihood (the parametric likelihood ratio) converges to a χ^2-distribution in a suitable logarithmic scale. To develop this analog, we first incorporate a target parameter of interest, i.e. the quantity whose statistical uncertainty is to be assessed (or to be "estimated"). Say this parameter is $\theta\in\mathbb{R}^p$. Suppose the true parameter is known to satisfy the set of equations $E[t(\theta;\xi)] = 0$ where $E[\cdot]$ is the expectation for the random object $\xi\in\mathbb{R}^d$, and $t(\theta;\xi), 0\in\mathbb{R}^b$. We define the nonparametric profile likelihood ratio as

$$\mathcal{R}(\theta) = \max\left\{\prod_{i=1}^{n} nw_i : \sum_{i=1}^{n} w_i t(\theta;\xi_i) = 0,\ \sum_{i=1}^{n} w_i = 1,\ w_i \geq 0 \text{ for all } i = 1,\ldots,n\right\} \tag{6}$$

where profiling refers to the categorization of all weights that respect the set of equations $E[t(\theta;\xi)] = 0$.

With the above definitions, the crux is the empirical likelihood theorem (ELT):

Theorem 1 (Theorem 3.4 in [14])**.** *Let $\xi_1,\ldots,\xi_n\in\mathbb{R}^d$ be i.i.d. data. Let $\theta_0\in\mathbb{R}^p$ be a value of the parameter that satisfies $E[t(\theta;\xi)] = 0$, where $t(\theta;\xi), 0\in\mathbb{R}^b$. Assume the covariance matrix $Var(t(\theta_0;\xi))$ is finite and has rank $q > 0$. Then $-2\log\mathcal{R}(\theta_0) \Rightarrow \chi^2_q$, where $\mathcal{R}(\theta)$ is defined in* (6)*.*

The quantity $-2\log\mathcal{R}(\theta)$ is defined as ∞ if the optimization in (6) is infeasible.

We now explain how (5) provides confidence bounds for optimization problem (1). We make the following assumptions:

Assumption 1. *1. $h(x)$ is differentiable in x with $\nabla_x h(x) = E[\nabla_x H(x;\xi)]$ for all $x \in \Theta$.*

2. $x^* \in argmin_{x\in\Theta} h(x)$ if and only if $\nabla_x h(x^*) = 0$. Moreover, this relation is distributionally stable*, meaning that $\tilde{x}^* \in argmin_{x\in\Theta} \tilde{h}(x)$ if and only if $\nabla_x \tilde{h}(\tilde{x}^*) = 0$ for any $\tilde{h}(x) = \tilde{E}[H(x;\xi)]$ that has the expectation $\tilde{E}[\cdot]$ generated under an arbitrary distribution $\tilde{P}$ such that*

$$\sup_{x\in\Theta} |\tilde{h}(x) - h(x)| < \epsilon$$

for small enough $\epsilon > 0$.

3. There exists an $x^* \in argmin_{x\in\Theta} h(x)$ such that the covariance matrix of the random vector $(\nabla_x H(x^*;\xi), H(x^*;\xi)) \in \mathbb{R}^{p+1}$ is finite and has positive rank.

4. $\frac{1}{n}\sum_{i=1}^n H(x;\xi_i) \to h(x)$ uniformly over $x \in \Theta$ a.s..

5. $E\left[\sup_{x\in\Theta} H(x;\xi)^2\right] < \infty$

Assuming the existence of $\nabla_x H(x;\xi)$ a.s., the interchangeability of derivative and expectation in Assumption 1.1 can generally be justified by the pathwise Lipschitz continuity condition

$$|H((x_1,\ldots,x_{j-1},u,x_{j+1},\ldots,x_p);\xi) - H((x_1,\ldots,x_{j-1},v,x_{j+1},\ldots,x_p);\xi)| \le M_j|u-v| \text{ a.s.}$$

for any u, v in a nonrandom neighborhood around the point x_j to be differentiated and M_j measurable with $EM_j < \infty$ (e.g., [1]). Another sufficient condition is that $H(x;\xi)$ is a.s. continuous and piecewise differentiable in x_j and $\sup_{u\in D} |(\partial/\partial x_j)H((x_1,\ldots,x_{j-1},u,x_{j+1},\ldots,x_p);\xi)|$ is integrable where D is a neighborhood around x_j [8]. Assumption 1.2 states that the first order condition for optimality is both sufficient and necessary. Assumptions 1.2 and 1.4 together ensure that this first order condition is unchanged when the true distribution is replaced by a (weighted) empirical version as the sample size gets large. Assumption 1.5 is a technical condition required to bound the error between the empirical distribution and its weighted version within the divergence ball. Assumption 1.3 is used to invoke Theorem 1. Note that x^* is not necessarily unique.

As our subsequent development will reveal, both the necessity and the sufficiency of the first order condition in Assumption 1.2 are required; in particular, we need the necessity of $\nabla_x h(x^*) = 0$ for $x^* \in \text{argmin}_{x\in\Theta} h(x)$ and the sufficiency of $\nabla_x \tilde{h}(\tilde{x}^*) = 0$ for $\tilde{x}^* \in \text{argmin}_{x\in\Theta} \tilde{h}(x)$ in order for our argument on statistical guarantee to go through. Assumptions 1.2, 1.4 and 1.5 can be replaced by a single condition

Assumption 2. *$x^* \in argmin_{x\in\Theta} h(x)$ if and only if $\nabla_x h(x^*) = 0$, and $\tilde{x}^* \in argmin_{x\in\Theta} \sum_{i=1}^n w_i H(x;\xi_i)$ if and only if $\sum_{i=1}^n w_i \nabla_x H(\tilde{x}^*;\xi_i) = 0$ for any support set $\{\xi_1,\ldots,\xi_n\} \subset \Theta$ and arbitrary probability vector w.*

Assumption 2 is satisfied by, for instance, $H(\cdot;\xi)$ that is coersive and convex for any ξ and $\Theta = \mathbb{R}^p$.

We have the following statistical guarantee:

Theorem 2. *Suppose $\xi_1,\ldots,\xi_n \in \mathbb{R}^d$ are i.i.d. data. Let z^* be the optimal value of* (1)*, and $\overline{z}$ and $\underline{z}$ be the maximum and minimum values of* (5) *respectively. Then, under Assumption 1, we have*

$$\liminf_{n\to\infty} P\left(z^* \in [\underline{z}, \overline{z}]\right) \ge 1 - \beta.$$

Proof. By Assumption 1.3, there exists an $x^* \in \operatorname{argmin}_{x\in\Theta} h(x)$ such that the covariance matrix of the random vector $(\nabla_x H(x^*;\xi), H(x^*;\xi))$ has a positive rank, call it r. Also, by Assumptions 1.1 and 1.2, x^* satisfies $E[\nabla_x H(x^*;\xi)] = 0$.

We define the nonparametric profile likelihood ratio as

$$\mathcal{R}(x,z) = \max\left\{\prod_{i=1}^n nw_i : \sum_{i=1}^n w_i \nabla_x H(x;\xi_i) = 0,\ \sum_{i=1}^n w_i H(x;\xi_i) = z,\ \sum_{i=1}^n w_i = 1,\ w_i \geq 0 \text{ for all } i = 1,\ldots,n\right\}. \tag{7}$$

Let $z^* = \min_{x\in\Theta} h(x) = h(x^*)$. From Theorem 1, the nonparametric profile likelihood ratio (7) satisfies $-2\log\mathcal{R}(x^*,z^*) \Rightarrow \chi^2_r$ as $n\to\infty$. This implies $P(-2\log\mathcal{R}(x^*,z^*) \leq \chi^2_{r,\beta}) \to 1-\beta$.

The rest of the proof focuses on the event $-2\log\mathcal{R}(x^*,z^*) \leq \chi^2_{r,\beta}$. Write

$$\begin{aligned} & -2\log\mathcal{R}(x^*,z^*) \\ = \ & \min\left\{ -2\sum_{i=1}^n \log(nw_i) : \sum_{i=1}^n w_i \nabla_x H(x^*;\xi_i) = 0,\ \sum_{i=1}^n w_i H(x^*;\xi_i) = z^*,\ \sum_{i=1}^n w_i = 1,\ w_i \geq 0 \right. \\ & \left. \text{for all } i = 1,\ldots,n \right\} \end{aligned} \tag{8}$$

We argue that $-2\log\mathcal{R}(x^*,z^*) \leq \chi^2_{r,\beta}$ implies the existence of a probability vector w such that

$$\sum_{i=1}^n w_i \nabla_x H(x^*;\xi_i) = 0,\ \sum_{i=1}^n w_i H(x^*;\xi_i) = z^*,\ -2\sum_{i=1}^n \log(nw_i) \leq \chi^2_{r,\beta} \tag{9}$$

Notice that $-2\sum_{i=1}^n \log(nw_i) = \infty$ if $w_i = 0$ for any i. Hence it suffices to replace, in (8), $w_i \geq 0$ with $w_i \geq \epsilon$ for all i, for some small enough $\epsilon > 0$. In this modified, compact, feasible set, $-2\sum_{i=1}^n \log(nw_i)$ is bounded and hence must possess an optimal solution w, which is a probability vector that satisfies (9).

This further implies that z^* is bounded from above and below by the optimization problems

$$\begin{array}{ll} \max / \min_w & \sum_{i=1}^n w_i H(x^*;\xi_i) \\ \text{subject to} & \sum_{i=1}^n w_i \nabla_x H(x^*;\xi_i) = 0 \\ & -2\sum_{i=1}^n \log(nw_i) \leq \chi^2_{r,\beta} \\ & \sum_{i=1}^n w_i = 1 \\ & w_i \geq 0 \text{ for all } i = 1,\ldots,n \end{array} \tag{10}$$

We argue that as $n\to\infty$, (10) is equivalent to

$$\begin{array}{ll} \max / \min_w & \sum_{i=1}^n w_i H(x^*;\xi_i) \\ \text{subject to} & w \in \{(w_1,\ldots,w_n) : x^* \in \operatorname{argmin}_{x\in\Theta} \sum_{i=1}^n w_i H(x;\xi_i)\} \\ & -2\sum_{i=1}^n \log(nw_i) \leq \chi^2_{r,\beta} \\ & \sum_{i=1}^n w_i = 1 \\ & w_i \geq 0 \text{ for all } i = 1,\ldots,n \end{array} \tag{11}$$

eventually (i.e. with probability 1). Note that the first constraint in (11) states that the probability vector w must be chosen such that x^*, a minimizer of $h(x)$ picked at the beginning of this proof, also minimizes $\sum_{i=1}^n w_i H(x;\xi_i)$.

To develop the argument for the asymptotic equivalence, let us denote P^w as the distribution represented by the probability weights w on the support $\{\xi_1, \ldots, \xi_n\}$. Denote $E^w[\cdot]$ as the associated expectation and $h^w(x) = E^w[H(x;\xi)]$. We will show that

$$\sup_{x\in\Theta, w\in\mathcal{W}_r} |h^w(x) - h(x)| \to 0 \ \text{ a.s.} \tag{12}$$

where

$$\mathcal{W}_r = \left\{ (w_1, \ldots, w_n) \in \mathbb{R}^n : -2\sum_{i=1}^n \log(nw_i) \le \chi^2_{r,\beta},\ \sum_{i=1}^n w_i = 1,\ w_i \ge 0 \text{ for all } i = 1, \ldots, n \right\} \tag{13}$$

Assumption 1.2 then implies that with probability 1, for sufficiently large n, $\sum_{i=1}^n w_i \nabla_x H(x^*;\xi_i) = 0$ if and only if $x^* \in \operatorname{argmin}_{x\in\Theta} \sum_{i=1}^n w_i H(x;\xi_i)$ for any $w \in \mathcal{W}_r$, leading to the equivalence.

We now show (12). Consider

$$\begin{aligned}
\sup_{x\in\Theta, w\in\mathcal{W}_r} |h^w(x) - h(x)| &= \sup_{x\in\Theta, w\in\mathcal{W}_r} \left| \int H(x;\xi) d(P^w - P)(\xi) \right| \\
&\le \sup_{x\in\Theta, w\in\mathcal{W}_r} \left| \int H(x;\xi) d(P^w - \hat{P})(\xi) \right| + \sup_{x\in\Theta} \left| \int H(x;\xi) d(\hat{P} - P)(\xi) \right| \\
&\qquad \text{where } \hat{P} \text{ denotes the empirical distribution generated from } \{\xi_1, \ldots, \xi_n\} \\
&\le \sup_{x\in\Theta, 1\le i\le n, w\in\mathcal{W}_r} |H(x;\xi_i)| d_{TV}(P^w, \hat{P}) + \sup_{x\in\Theta} \left| \int H(x;\xi) d(\hat{P} - P)(\xi) \right| \qquad (14) \\
&\qquad \text{where } d_{TV} \text{ denotes the total variation distance}
\end{aligned}$$

Now by Lemma 11.5 in [14] (restated in the Appendix) and Assumption 1.5, we have

$$\sup_{x\in\Theta, 1\le i\le n} |H(x;\xi_i)| = \max_{1\le i\le n} \sup_{x\in\Theta} |H(x;\xi_i)| = o(n^{1/2}) \ \text{ a.s.} \tag{15}$$

On the other hand, by Pinsker's inequality, for any $w \in \mathcal{W}_r$,

$$d_{TV}(P^w, \hat{P}) \le \sqrt{\frac{d_{KL}(P^w, \hat{P})}{2}} = \sqrt{\frac{-\sum_{i=1}^n \log(nw_i)}{2n}} \le \sqrt{\frac{\chi^2_r}{4n}} \tag{16}$$

where d_{KL} denotes the Kullback-Leibler divergence. Combining (15) and (16), the first term in (14) goes to 0 a.s.. The second term in (14) converges to 0 a.s. by Assumption 1.4. Hence $\sup_{x\in\Theta} |h^w(x) - h(x)| \to 0$ a.s.. Therefore (10) is equivalent to (11) eventually as $n \to \infty$.

Consider (11). With the first constraint, the objective function must be equal to $\min_{x\in\Theta} \sum_{i=1}^n w_i H(x;\xi_i)$. Thus (11) is equivalent to

$$\begin{array}{ll}
\max / \min_w & \min_{x\in\Theta} \sum_{i=1}^n w_i H(x;\xi_i) \\
\text{subject to} & w \in \{(w_1, \ldots, w_n) : x^* \in \operatorname{argmin}_{x\in\Theta} \sum_{i=1}^n w_i H(x;\xi_i)\} \\
& -2\sum_{i=1}^n \log(nw_i) \le \chi^2_{r,\beta} \\
& \sum_{i=1}^n w_i = 1 \\
& w_i \ge 0 \text{ for all } i = 1, \ldots, n
\end{array} \tag{17}$$

Let $\overline{v}$ and $\underline{v}$ be the maximum and minimum values of (17). Note that $r \le p+1$ since r is the rank of a $\mathbb{R}^{(p+1)\times(p+1)}$ matrix. This implies $\chi^2_{r,\beta} \le \chi^2_{p+1,\beta}$. Together with a relaxation by removing the

first constraint in (17), we have $\underline{v} \geq \underline{z}$ and $\overline{v} \leq \overline{z}$ where $\overline{z}$ and $\underline{z}$ are the maximum and minimum values of (5). From this we conclude that

$$\liminf_{n\to\infty} P(\underline{z} \leq z^* \leq \overline{z}) \geq \liminf_{n\to\infty} P(\underline{v} \leq z^* \leq \overline{v}) \geq \liminf_{n\to\infty} P(-2\log \mathcal{R}(x^*, z^*) \leq \chi^2_{r,\beta}) = 1 - \beta$$

□

We also have the following result based on Assumption 2 that can be more natural to verify in some cases:

Corollary 1. *Suppose $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ are i.i.d. data. Let z^* be the optimal value of (1), and $\overline{z}$ and $\underline{z}$ be the maximum and minimum values of (5) respectively. Then, under Assumptions 1.1, 1.3 and 2, we have*

$$\liminf_{n\to\infty} P\left(z^* \in [\underline{z}, \overline{z}]\right) \geq 1 - \beta.$$

Proof. The proof follows similarly as that of Theorem 2, with the observation that the equivalence of (10) and (11) holds for all n if Assumption 2 replaces Assumptions 1.2, 1.4 and 1.5. □

Note that we have obtained bounds by relaxing the constraints in (17), and the degree of freedom in the χ^2-distribution may not be optimally chosen. Nevertheless, our numerical examples show that, at least for small p, the EL method provides reasonably tight CIs. There exist techniques (e.g., bootstrap calibration or Bartlett correction; [13, 7]) that can improve the coverage of the EL method in estimation problems. Investigation of these techniques in the optimization context is delegated to future work.

2.2 Stochastically Constrained Optimization

We generalize the EL method to the stochastically constrained problem (3). In this setting, we construct CI via the following optimization problems

$$\begin{array}{ll} \max / \min_w & \left\{ \begin{array}{ll} \min_x & \sum_{i=1}^n w_i H(x;\xi_i) \\ \text{subject to} & \sum_{i=1}^n w_i F_k(x;\xi_i) \leq 0,\ k = 1,\ldots,m \\ & g_k(x) \leq 0,\ k = 1,\ldots,s \end{array} \right\} \\ \text{subject to} & -2\sum_{i=1}^n \log(nw_i) \leq \chi^2_{p+m+1,\beta} \\ & \sum_{i=1}^n w_i = 1 \\ & w_i \geq 0 \text{ for all } i = 1,\ldots,n \end{array} \tag{18}$$

While resembling (5), we note that the degree of freedom in the χ^2-distribution is now $p + m + 1$, which includes the number of stochastic constraints compared to (5).

For convenience, we denote

$$\Lambda = \{x \in \mathbb{R}^p : g_k(x) \leq 0,\ k = 1,\ldots,s\}$$

as the set of x satisfying the deterministic constraints in (3).

We make the following assumptions in parallel to Assumption 1:

Assumption 3. *We assume:*

1. $h(x) = E[H(x;\xi)]$, $f_k(x) = E[F_k(x;\xi)]$, $k = 1,\dots,m$ *and* $g_k(x)$, $k = 1,\dots,s$ *are all differentiable in* $x \in \Lambda$, *and*

$$\nabla_x h(x) = E[\nabla_x H(x;\xi)],\ \nabla_x f_k(x) = E[\nabla_x F_k(x;\xi)]$$

2. *Let* S^* *be the set of all optimal solutions for* (3). $x^* \in S^*$ *if and only if* x^* *satisfies the KKT condition, where the active set of the KKT condition (i.e. equalities) is unique among all* $x^* \in S^*$ *and is sufficient for determining* S^*. *This relation is* distributionally stable, *meaning that* $\tilde{x}^* \in \tilde{S}^*$, *where* $\tilde{S}^*$ *is the set of optimal solutions for*

$$\begin{array}{ll} \min & \tilde{h}(x) \\ \textit{subject to} & \tilde{f}_k(x) \le 0,\ k = 1,\dots,m \\ & \tilde{g}_k(x) \le 0,\ k = 1,\dots,s \end{array} \tag{19}$$

if and only if $\tilde{x}^*$ *satisfies the corresponding KKT condition, where* $\tilde{h}(x) = \tilde{E}[H(x;\xi)]$ *and* $\tilde{f}_k(x) = \tilde{E}[F_k(x;\xi)]$, *with* $\tilde{E}$ *denoting the expectation under an arbitrary distribution* $\tilde{P}$ *such that*

$$\sup_{x\in\Lambda} |\tilde{h}(x) - h(x)| < \epsilon$$

$$\sup_{x\in\Lambda} |\tilde{f}_k(x) - f_k(x)| < \epsilon \ \ \textit{for all } k = 1,\dots,m$$

for small enough $\epsilon > 0$. *Moreover, for any such* $\epsilon > 0$, *the active set of the KKT condition at any* $\tilde{x}^* \in \tilde{S}^*$ *for* (19) *is the same as that at any* $x^* \in S^*$ *for* (3) *and is sufficient for determining* $\tilde{S}^*$.

3. *There exists an optimal solution* x^* *for* (3), *with associated Lagrange multipliers for the stochastic constraints in* (3) *given by* $\lambda^* = (\lambda_1^*,\dots,\lambda_m^*)$, *such that the covariance matrix of the variables* $H(x^*;\xi)$, $\frac{\partial}{\partial x_j}H(x^*;\xi) + \sum_{k=1}^m \lambda_k^* \frac{\partial}{\partial x_j}F_k(x^*;\xi)$, *and* $F_k(x^*;\xi)$, *for all indices* j *and* k *corresponding to the active set of the KKT condition, is finite and has positive rank.*

4. $$\frac{1}{n}\sum_{i=1}^n H(x;\xi_i) \to h(x) \ \ \textit{and} \ \ \frac{1}{n}\sum_{i=1}^n F_k(x;\xi_i) \to f_k(x), \quad k = 1,\dots,m$$

uniformly over $x \in \Lambda$ *a.s.*.

5. $E\left[\sup_{x\in\Theta} H(x;\xi)^2\right] < \infty$ *and* $E\left[\sup_{x\in\Theta} F_k(x;\xi)^2\right] < \infty$ *for* $k = 1,\dots,m$.

Denote $\nu^* = (\nu_1^*,\dots,\nu_s^*)$ as the Lagrange multiplier for the deterministic constraints in (3). In Assumptions 3.2 and 3.3 above, the active set of the KKT condition satisfied by (x^*,λ^*,ν^*) is in the form

$$\frac{\partial}{\partial x_j}h(x^*) + \sum_{k=1}^m \lambda_k^* \frac{\partial}{\partial x_j} f_k(x^*) + \sum_{k=1}^s \nu_k^* \frac{\partial}{\partial x_j} g_k(x^*) = 0,\ j \in \mathcal{A}_1^* \equiv \{1,\dots,p\}$$

$$f_k(x^*) = 0,\ k \in \mathcal{A}_2^* \subset \{1,\dots,m\}$$

$$g_k(x^*) = 0,\ k \in \mathcal{A}_3^* \subset \{1,\dots,s\}$$

$$\lambda_k^* = 0\ k \in \{1,\dots,m\} \setminus \mathcal{A}_2^*$$

$$\nu_k^* = 0, \quad k \in \{1, \ldots, s\} \setminus \mathcal{A}_3^*$$

where $\mathcal{A}_1^*$, $\mathcal{A}_2^*$ and $\mathcal{A}_3^*$ denote the sets of indices that correspond to the equalities in the optimality condition, which are unique among any optimal solutions of (3) by Assumption 3.2. The j and k described in Assumption 3.3 refer to the indices in $\mathcal{A}_1^*$ and $\mathcal{A}_2^*$. Assumption 3.2 further enforces the sets $\mathcal{A}_1^*$, $\mathcal{A}_2^*$ and $\mathcal{A}_3^*$ to remain as the active sets under a perturbation to $\tilde{P}$ described therein, and the equalities indexed via these sets are enough to determine S^* and $\tilde{S}^*$. Assumptions 3.2 and 3.3 generalize Assumptions 1.2 and 1.3 from the simple zero-derivative optimality condition to the KKT condition. Similar to Section 2.1, we require the necessity of the KKT and the active set conditions regarding (3) and the sufficiency regarding (19) for our development to go through. Constraint qualification for the validity of the KKT condition is implicitly assumed in Assumption 3.2.

We have the following result:

Theorem 3. *Suppose $\xi_1, \ldots, \xi_n$ are i.i.d. data. Under Assumption 3, we have*

$$\liminf_{n\to\infty} P(z^* \in [\underline{z}, \overline{z}]) \geq 1 - \beta$$

where z^* is the optimal value of (3)*, and $\underline{z}$ and $\overline{z}$ are the minimum and maximum values of* (18)*.*

Proof. Consider the nonparametric profile likelihood ratio

$$\mathcal{R}(x, \lambda, \nu, z) = \max \left\{ \prod_{i=1}^n n w_i : \begin{array}{l} \sum_{i=1}^n w_i H(x; \xi_i) = z \\ \sum_{i=1}^n w_i \left(\frac{\partial}{\partial x_j} H(x; \xi_i) + \sum_{k=1}^m \lambda_k \frac{\partial}{\partial x_j} F_k(x; \xi_i) \right) + \sum_{k=1}^s \nu_k \frac{\partial}{\partial x_j} g_k(x) = 0, \ j \in \mathcal{A}_1^* \\ \sum_{i=1}^n w_i F_k(x; \xi_i) = 0, \ k \in \mathcal{A}_2^* \\ g_k(x) = 0, \ k \in \mathcal{A}_3^* \\ \lambda_k = 0 \ k \in \{1, \ldots, m\} \setminus \mathcal{A}_2^* \\ \nu_k = 0, \quad k \in \{1, \ldots, s\} \setminus \mathcal{A}_3^* \\ \sum_{i=1}^n w_i = 1 \\ w_i \geq 0 \text{ for all } i = 1, \ldots, n \end{array} \right\} \tag{20}$$

Let x^* be an optimal solution for (3) satisfying Assumption 3.3, and $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*), \nu^* = (\nu_1^*, \ldots, \nu_s^*)$ be its associated Lagrange multipliers. By Assumption 3.3, the covariance of the random vector concatenated by

$$\begin{array}{l} H(x^*; \xi) \\ \dfrac{\partial}{\partial x_j} H(x^*; \xi) + \displaystyle\sum_{k=1}^m \lambda_k^* \frac{\partial}{\partial x_j} F_k(x^*; \xi) + \sum_{k=1}^s \nu_k^* \frac{\partial}{\partial x_j} g_k(x^*) \ \text{ for } j \in \mathcal{A}_1^* \\ F_k(x^*; \xi) \ \text{ for } k \in \mathcal{A}_2^* \end{array}$$

has rank r for some $r > 0$. Let z^* be the optimal value of (3) equal to $h(x^*)$. Since the other active KKT conditions are deterministic, Theorem 1 implies that $-2 \log \mathcal{R}(x^*, \lambda^*, \nu^*, z^*) \Rightarrow \chi_r^2$, which further implies $P(-2 \log \mathcal{R}(x^*, \lambda^*, \nu^*, z^*) \leq \chi_{r,\beta}^2) \to 1 - \beta$.

Similar to the proof of Theorem 2, $-2 \log \mathcal{R}(x^*, \lambda^*, \nu^*, z^*) \leq \chi_{r,\beta}^2$ implies the existence of a w that satisfies $-2 \sum_{i=1}^n \log(n w_i) \leq \chi_{r,\beta}^2$ and all constraints in (20) evaluated at $x^*, \lambda^*, \nu^*, z^*$. This in

turn implies that z^* is bounded by

$$\begin{array}{ll}
\max / \min_w & \sum_{i=1}^n w_i H(x^*;\xi_i) \\
\text{subject to} & \sum_{i=1}^n w_i \left(\frac{\partial}{\partial x_j} H(x^*;\xi_i) + \sum_{k=1}^m \lambda_k^* \frac{\partial}{\partial x_j} F_k(x^*;\xi_i) \right) + \sum_{k=1}^s \nu_k^* \frac{\partial}{\partial x_j} g_k(x^*) = 0,\ j \in \mathcal{A}_1^* \\
& \sum_{i=1}^n w_i F_k(x^*;\xi_i) = 0,\ k \in \mathcal{A}_2^* \\
& g_k(x^*) = 0,\ k \in \mathcal{A}_3^* \\
& \lambda_k^* = 0\ k \in \{1,\dots,m\} \setminus \mathcal{A}_2^* \\
& \nu_k^* = 0, \quad k \in \{1,\dots,s\} \setminus \mathcal{A}_3^* \\
& -2\sum_{i=1}^n \log(n w_i) \le \chi^2_{r,\beta} \\
& \sum_{i=1}^n w_i = 1 \\
& w_i \ge 0 \text{ for all } i = 1,\dots,n
\end{array} \tag{21}$$

Using the same argument as in the proof of Theorem 2, we obtain from Assumptions 3.4 and 3.5 that

$$\sup_{x\in\Lambda, w\in\mathcal{W}_r} |h^w(x) - h(x)| \to 0 \ \text{ a.s.}$$

$$\sup_{x\in\Lambda, w\in\mathcal{W}_r} |f_k^w(x) - f_k(x)| \to 0 \ \text{ a.s. for all } k = 1,\dots,m$$

where $\mathcal{W}_r$ is defined in (13), and $h^w(x) = E^w[H(x;\xi)]$, $f_k^w(x) = E^w[F_k(x;\xi)]$ with E^w denoting the expectation with respect to P^w, the probability distribution represented by the weights w on the support $\{\xi_1,\dots,\xi_n\}$. Thus, by Assumption 3.2, the set of active KKT conditions for an optimal solution of the weighted sample problem

$$\begin{array}{ll}
\min_x & \sum_{i=1}^n w_i H(x;\xi_i) \\
\text{subject to} & \sum_{i=1}^n w_i F_k(x;\xi_i) \le 0,\ k = 1,\dots,m \\
& g_k(x) \le 0,\ k = 1,\dots,s
\end{array}$$

for any $w \in \mathcal{W}_r$ is identical to that for x^* for (3) eventually as $n \to \infty$, and Assumption 3.2 further implies that (21) is equivalent to

$$\begin{array}{ll}
\max / \min_w & \sum_{i=1}^n w_i H(x^*;\xi_i) \\
\text{subject to} & w \in \left\{ (w_1,\dots,w_n) : x^* \in \left\{ \begin{array}{ll} \text{argmin}_x & \sum_{i=1}^n w_i H(x;\xi_i) \\ \text{subject to} & \sum_{i=1}^n w_i F_k(x;\xi_i) \le 0,\ k = 1,\dots,m \\ & g_k(x) \le 0,\ k = 1,\dots,s \end{array} \right\} \right\} \\
& -2\sum_{i=1}^n \log(n w_i) \le \chi^2_{r,\beta} \\
& \sum_{i=1}^n w_i = 1 \\
& w_i \ge 0 \text{ for all } i = 1,\dots,n
\end{array} \tag{22}$$

eventually as $n \to \infty$. With the first constraint, the objective function in (22) must be equal to $\min_x\{\sum_{i=1}^n w_i H(x;\xi_i) : \sum_{i=1}^n w_i F_k(x;\xi_i) \le 0, k = 1,\dots,m,\ g_k(x) \le 0, k = 1,\dots,s\}$. Note that

$$r \le 1 + |\mathcal{A}_1^*| + |\mathcal{A}_2^*| \le 1 + p + m \tag{23}$$

where $|\cdot|$ denotes cardinality. This implies that $\chi^2_{r,\beta} \le \chi^2_{p+m+1,\beta}$. Thus, together with a relaxation of the first constraint in (22), the same argument as in the proof of Theorem 2 stipulates that the maximum and minimum values of (22) are bounded from above and below respectively by those of (18) and concludes the theorem.

□

Note that, much like the proof of Theorem 2, we have relaxed constraints and placed a conservative bound on the degree of freedom of the χ^2-distribution in (23), which could potentially be improved with more refined analysis.

3 The Empirical Likelihood Method for Constructing Confidence Bounds for Optimality Gaps

We study the construction of CI for the optimality gap of a given solution using the EL method. We focus on optimization problem (1) and suppose $\hat{x}$ is a feasible solution obtained from some procedure independently of the data $\xi_1, \ldots, \xi_n$. The optimality gap of $\hat{x}$ is given by $\mathcal{G}(\hat{x}) = h(\hat{x}) - z^*$ where z^* is the optimal value of (1). We will show how we can apply the results in Section 2 to find the CI for $\mathcal{G}(\hat{x})$.

Consider the optimization problems

$$\begin{array}{ll} \max / \min_w & \max_{x\in\Theta} \sum_{i=1}^n w_i[H(\hat{x};\xi_i) - H(x;\xi_i)] \\ \text{subject to} & -2\sum_{i=1}^n \log(nw_i) \le \chi^2_{p+1,\beta} \\ & \sum_{i=1}^n w_i = 1 \\ & w_i \ge 0 \text{ for all } i = 1, \ldots, n \end{array} \tag{24}$$

We have the following guarantee in using (24) to construct the CI for $\mathcal{G}(\hat{x})$ for (1):

Theorem 4. *Suppose $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ are i.i.d. data independent of a given solution $\hat{x}$ that is feasible for (1). Let $\mathcal{G}(\hat{x})$ be the optimality gap of $\hat{x}$ for (1), and $\overline{z}$ and $\underline{z}$ be the maximum and minimum values of the programs in (24) respectively. Suppose Assumption 1 holds except that in Condition 2, the relation holds for any $\tilde{h}(x) = \tilde{E}[H(x;\xi)]$ such that $\sup_{x\in\Theta} |(\tilde{h}(x) - h(x)) - (\tilde{h}(\hat{x}) - h(\hat{x}))| < \epsilon$ and in Condition 3, we consider the covariance matrix of $(\nabla_x H(x^*;\xi), H(x^*;\xi) - H(\hat{x};\xi))$ instead. We have*

$$\liminf_{n\to\infty} P\left(\mathcal{G}(\hat{x}) \in [\underline{z}, \overline{z}]\right) \ge 1 - \beta. \tag{25}$$

Proof. Let $\bar{H}(x;\xi) = H(x;\xi) - H(\hat{x};\xi)$, and $\bar{h}(x) = E[\bar{H}(x;\xi)] = h(x) - h(\hat{x})$. We verify that Assumption 1, with the change that $\sup_{x\in\Theta} |(\tilde{h}(x) - h(x)) - (\tilde{h}(\hat{x}) - h(\hat{x}))| < \epsilon$ is used in Condition 2 and the covariance matrix of $(\nabla_x H(x^*;\xi), H(x^*;\xi) - H(\hat{x};\xi))$ is considered instead in Condition 3, implies that $\bar{h}$ and $\bar{H}$ satisfies Assumption 1 too with h and H replaced by $\bar{h}$ and $\bar{H}$.
Condition 1: We have $\nabla_x \bar{h}(x) = \nabla_x(h(x) - h(\hat{x})) = \nabla_x h(x) = E[\nabla_x H(x;\xi)] = E[\nabla_x(H(x;\xi) - H(\hat{x};\xi))] = E[\nabla_x \bar{H}(x;\xi)]$.
Condition 2: We have $x^* \in \operatorname{argmin}_{x\in\Theta} \bar{h}(x) \Leftrightarrow x^* \in \operatorname{argmin}_{x\in\Theta} h(x) \Leftrightarrow \nabla_x h(x^*) = 0 \Leftrightarrow \nabla_x \bar{h}(x^*) = 0$. Similarly, $\tilde{x}^* \in \operatorname{argmin}_{x\in\Theta} \bar{\tilde{h}}(x) \Leftrightarrow \tilde{x}^* \in \operatorname{argmin}_{x\in\Theta} \tilde{h}(x) \Leftrightarrow \nabla_x \tilde{h}(\tilde{x}^*) = 0 \Leftrightarrow \nabla_x \bar{\tilde{h}}(\tilde{x}^*) = 0$ for any $\bar{\tilde{h}}(x) = \tilde{h}(x) - \tilde{h}(\hat{x})$ that satisfies $\sup_{x\in\Theta} |\bar{\tilde{h}}(x) - \bar{h}(x)| < \epsilon$ by our modification of this condition.
Condition 3: By our modification of this condition we have the covariance of $(\nabla_x \bar{H}(x^*;\xi), \bar{H}(x^*;\xi)) = (\nabla_x H(x^*;\xi), H(x^*;\xi) - H(\hat{x};\xi))$ finite and having a positive rank.
Condition 4: It is straightforward to show that $\frac{1}{n}\sum_{i=1}^n \bar{H}(x;\xi_i) = \frac{1}{n}\sum_{i=1}^n H(x;\xi_i) - \frac{1}{n}\sum_{i=1}^n H(\hat{x};\xi_i) \to \bar{h}(x)$ a.s. uniformly over $x \in \Theta$.
Condition 5: We have $E[\sup_{x\in\Theta} \bar{H}(x;\xi)^2] = E[\sup_{x\in\Theta}(H(x;\xi) - H(\hat{x};\xi))^2] \le 4(E[\sup_{x\in\Theta} H(x;\xi)^2] + E[H(\hat{x};\xi)^2]) < \infty$.

We have therefore verified our claim. Using Theorem 2, we get that

$$\liminf_{n\to\infty} P(\underline{v} \le \bar{h}(x^*) \le \overline{v}) \ge 1 - \beta$$

where $\overline{v}$ and $\underline{v}$ are the maximum and minimum values of

$$\begin{array}{ll} \max / \min_w & \min_{x\in\Theta} \sum_{i=1}^n w_i \bar{H}(x;\xi_i) \\ \text{subject to} & -2\sum_{i=1}^n \log(nw_i) \le \chi^2_{p+1,\beta} \\ & \sum_{i=1}^n w_i = 1 \\ & w_i \ge 0 \text{ for all } i = 1,\ldots,n \end{array} \tag{26}$$

Noting that $\mathcal{G}(x^*) = -\bar{h}(x^*)$, we get (25) immediately. □

In parallel to Corollary 1, we have the following result on optimality gap assessment based on an alternative set of assumptions:

Corollary 2. *Suppose $\xi_1,\ldots,\xi_n \in \mathbb{R}^d$ are i.i.d. data independent of a given solution $\hat{x}$ that is feasible for (1). Let $\mathcal{G}(\hat{x})$ be the optimality gap of $\hat{x}$ for (1), and $\overline{z}$ and $\underline{z}$ be the maximum and minimum values of the programs in (24) respectively. Under Assumption 1.1, Assumption 1.3 where the covariance matrix of $(\nabla_x H(x^*;\xi), H(x^*;\xi) - H(\hat{x};\xi))$ is considered instead, and Assumption 2, we have*

$$\liminf_{n\to\infty} P\left(\mathcal{G}(\hat{x}) \in [\underline{z},\overline{z}]\right) \ge 1-\beta$$

Proof. Similar to the proof of Theorem 4, we can define $\bar{H}$ and $\bar{h}$ and verify that Assumption 1.1 and the modified Assumption 1.3 hold for $\bar{H}$ and $\bar{h}$. Assumption 2 is also satisfied for $\bar{h}$ and $\bar{H}$ because $x^* \in \operatorname{argmin}_{x\in\Theta}\bar{h}(x) \Leftrightarrow x^* \in \operatorname{argmin}_{x\in\Theta} h(x) \Leftrightarrow \nabla_x h(x^*) = 0 \Leftrightarrow \nabla_x \bar{h}(x^*) = 0$ and $\tilde{x}^* \in \operatorname{argmin}_{x\in\Theta}\sum_{i=1}^n w_i\bar{H}(x) \Leftrightarrow \tilde{x}^* \in \operatorname{argmin}_{x\in\Theta}\sum_{i=1}^n w_i H(x) \Leftrightarrow \sum_{i=1}^n w_i \nabla_x H(\tilde{x}^*) = 0 \Leftrightarrow \sum_{i=1}^n w_i \nabla_x \bar{H}(\tilde{x}^*) = 0$ for any support set $\{\xi_1,\ldots,\xi_n\}$ and arbitrary probability weight vector w. The rest of the proof then follows as that of Theorem 4. □

The statistical uncertainty of the optimality gap of the stochastically constrained problem (3) can in principle be analyzed in a similar fashion. However, validating the feasibility of a given solution $\hat{x}$ under limited data is not straightforward in such a scenario (see, e.g., Section 3 in [17]), and for this reason we skip the corresponding result in this work.

4 Conclusion

We have studied the EL method to construct statistically valid CIs for the optimal value and the optimality gap of a given solution for stochastic optimization problems. The method builds on positing two optimization problems that resemble DRO problems with Burg-entropy divergence ball constraints, with the ball size suitably calibrated by a χ^2-quantile with a chosen degree of freedom. We have studied the theory leading to the statistical guarantees and numerically compared our method to approaches suggested by the CLT (in our online Supplemental Material). Built on a rigorous foundation, our method provides a competitive method for evaluating the statistical uncertainty for stochastic optimization problems under limited data. In future work, we plan to further refine the accuracy and extend the scope of our method.

Appendix

Lemma 1 (Lemma 11.2 in [14])**.** *Let Y_i be i.i.d. random variables in $\mathbb{R}$ with $EY_i^2 < \infty$. We have $\max_{1\le i\le n}|Y_i| = o(n^{1/2})$ a.s..*

References

[1] S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*, volume 57. Springer Science & Business Media, 2007.

[2] G. Bayraksan and D. Morton. Assessing solution quality in stochastic programs. *Mathematical Programming*, 108:495–514, 2006.

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

[4] D. Bertsimas, V. Gupta, and N. Kallus. Robust sample average approximation. *available at arXiv:1408.4445*, 2016.

[5] W. Chen, M. Sim, J. Sun, and C.-P. Teo. From CVaR to uncertainty set: Implications in joint chance-constrained optimization. *Operations Research*, 58(2):470–485, 2010.

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

[7] T. DiCiccio, P. Hall, and J. Romano. Empirical likelihood is Bartlett-correctable. *The Annals of Statistics*, 19(2):1053–1061, 1991.

[8] P. Glasserman. Performance continuity and differentiability in Monte Carlo optimization. In *Proceedings of the 20th Winter Simulation Conference*, pages 518–524, New York, NY, USA, 1988. ACM.

[9] R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Mathematical Programming, Series A*, 158(1):291–327, 2012.

[10] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[11] H. Lam and E. Zhou. Quantifying uncertainty in sample average approximation. In *Proceedings of the 2015 Winter Simulation Conference*, pages 3846–3857. IEEE Press, 2015.

[12] W.-K. Mak, D. Morton, and R. Wood. Monte Carlo bounding techniques for determining solution quality. *Operations Research Letters*, 24:47–56, 1999.

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

[14] A. B. Owen. *Empirical Likelihood*. CRC press, 2001.

[15] L. Pardo. *Statistical Inference Based on Divergence Measures*. CRC Press, 2005.

[16] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*, volume 16. SIAM, 2014.

[17] W. Wang and S. Ahmed. Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters*, 36(5):515–519, 2008.

[18] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

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

[20] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming, Series A*, 137(1):167–198, 2013.

Supplemental Material

5 Numerical Examples

We test the presented method numerically on three examples. For proof of concept, the first example is a simple unconstrained quadratic optimization problem. Then we apply the proposed method to two more examples, including the problem of estimating Conditional-Value-at-Risk (CVaR) and a stochastically constrained portfolio optimization problem. The latter examples strictly speaking do not satisfy our assumptions, since the needed optimality conditions may not hold for their sample counterparts (Assumption 1.2, 2 or 3.2). However, given that the EL method does not rely on these conditions procedurally, we can still test its performance on these examples.

We compare EL with the CIs obtained from the CLT and the delta method ([16], Theorem 5.7). For deterministically constrained problems in the form (1), the $(1-\beta)$ CI on the optimal value is given by

$$\left[\hat{z}_n^* - z_{1-\beta/2}\frac{\hat{\sigma}(\hat{x}_n^*)}{\sqrt{n}}, \hat{z}_n^* + z_{1-\beta/2}\frac{\hat{\sigma}(\hat{x}_n^*)}{\sqrt{n}}\right] \tag{27}$$

where $z_{1-\beta/2}$ is the critical value of the standard normal distribution at level $1-\beta/2$, $\hat{x}_n^*$ is the empirical optimal solution obtained from (2), $\hat{z}_n^* = (1/n)\sum_{i=1}^n H(\hat{x}^*;\xi_i)$ is the empirical optimal value, and $\hat{\sigma}(\hat{x}_n^*) = \sqrt{(1/(n-1))\sum_{i=1}^n (H(\hat{x}_n^*;\xi_i) - \hat{z}_n^*)^2}$ is the empirical standard deviation of $H(\hat{x}_n^*;\xi)$. Since $\hat{z}_n^*$ is a negatively biased estimator of z^*, the CI (27) can suffer from under coverage. So we also compare with a 2-sample CLT (CLT2) method, as suggested by [12], which uses first half of the data to compute the empirical optimal value and solution, and then uses the remaining half of the data to estimate the objective value fixed at the solution to generate an upper bound. The 2-sample CLT CI is given by

$$\left[\hat{z}_{n/2}^* - z_{1-\beta/2}\frac{\hat{\sigma}(\hat{x}_{n/2}^*)}{\sqrt{n/2}}, \bar{z}_{n/2}^* + z_{1-\beta/2}\frac{\bar{\sigma}(\hat{x}_{n/2}^*)}{\sqrt{n/2}}\right] \tag{28}$$

where $\hat{z}_{n/2}^*, \hat{x}_{n/2}^*, \hat{\sigma}(\hat{x}_{n/2}^*)$ are computed as before using first half of the data $\{\xi_1, \ldots, \xi_{n/2}\}$, $\bar{z}_{n/2}^* = (2/n)\sum_{i=\frac{n}{2}+1}^n H(\hat{x}_{n/2}^*;\xi_i)$ is the evaluation of $\hat{x}_{n/2}^*$ using the remaining half of the data, and $\bar{\sigma}(\hat{x}_{n/2}^*) = \sqrt{(1/(n/2-1))\sum_{i=\frac{n}{2}+1}^n (H(\hat{x}_{n/2}^*;\xi_i) - \bar{z}_{n/2}^*)^2}$ is the empirical standard deviation at $\hat{x}_{n/2}^*$. Note that $\bar{z}_{n/2}^*$ is a positively biased estimator of z^*, and thus the CI (28) alleviates the under coverage issue; on the other hand, the effective sample size is reduced by half, and thus the estimates are less accurate especially when the data size is small, which may in turn affect the coverage probability of the CI.

Due to the limited data size, we use the single replication procedure (SRP) proposed in [2] to estimate CIs on the optimality gap. For a given solution $\hat{x}$ that is independent of the data, the SRP outputs a one-sided $(1-\beta)$ CI on the optimality gap given by

$$\left[0, \hat{\mathcal{G}}_n(\hat{x}) + z_{1-\beta}\frac{\tilde{\sigma}(\hat{x}_n^*)}{\sqrt{n}}\right], \tag{29}$$

where as before $\hat{x}_n^*$ is the empirical optimal solution, $\hat{\mathcal{G}}_n(\hat{x}) = (1/n)\sum_{i=1}^n \left(H(\hat{x},\xi_i) - H(\hat{x}_n^*,\xi_i)\right)^2$, and

$$\tilde{\sigma}^2(\hat{x}_n^*) = \frac{1}{n-1}\sum_{i=1}^n \left[(H(\hat{x},\xi_i) - H(\hat{x}_n^*,\xi_i)) - (\hat{h}(\hat{x}) - \hat{z}_n^*)\right]^2,$$

where $\hat{z}_n^*$ is the empirical optimal value and $\hat{h}(\hat{x}) = (1/n)\sum_{i=1}^n H(\hat{x}, \xi_i)$. In all the examples considered below, we set $\beta = 0.05$; since we consider cases of small data size, we replace $z_{1-\beta/2}$ and $z_{1-\beta}$ in (27)-(29) respectively with $t_{n-1,1-\beta/2}$ and $t_{n-1,1-\beta}$, which are the critical values of the Student's t-distribution with $n-1$ degree of freedom. Note that all the above discussion holds for deterministically constrained problems. Nonetheless, we also apply these methods in a stochastically constrained problem as a benchmark which, like the EL method (see the remark at the beginning of this section), are heuristic in this case without any formal validity proof.

Note that the EL method consists of solving a max-min and a min-min problem. Supposing that the original problem (1) or (3) is convex, then the max-min program is convex. In our examples we use the built-in Matlab solvers. The min-min program, on the other hand, is more challenging because the outer optimization involves minimizing the concave function $\min_{x\in\Theta}\sum_{i=1}^n w_i H(x;\xi_i)$ over w. This is not a convex problem in general. However, fixing either w or x, optimizing over the other variable becomes a convex problem. Thus one approach is to do alternating minimization, by iteratively minimizing w and x while fixing each others, until no improvement is observed. Such type of schemes has appeared in chance-constrained programming (e.g., [5, 20, 9]), and it appears to work well in our examples despite a lack of global convergence guarantee.

5.1 Quadratic Optimization

We consider a simple unconstrained problem of minimizing a quadratic function

$$\min_x E[(x-\xi)^2], \tag{30}$$

where ξ follows an unknown distribution F^c. It is easy to see that the optimal solution is $x^* = E[\xi]$ and the optimal value is $z^* = Var(\xi)$. We set F^c as a standard normal distribution, and thus $x^* = 0$ and $z^* = 1$.

Assuming we are given n observations from the normal distribution, we implement the different methods to obtain 95% confidence bounds for the optimal value of (30). We test on three cases where we randomly generate $n = 10, 50, 100$ data points from F^c. For each case, we repeat the experiment 1000 times, and note the empirical coverage probability, mean upper and lower bounds, and the mean and standard deviation of the interval width for each method. The results are summarized in Table 1.

		Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
$n = 10$	EL	0.74	0.40	1.57	1.17	0.59
	CLT	0.80	0.12	1.68	1.56	0.83
	CLT2	0.83	−0.05	2.54	2.59	1.77
$n = 50$	EL	0.95	0.34	2.11	1.78	1.20
	CLT	0.92	0.59	1.36	0.77	0.19
	CLT2	0.91	0.44	1.59	1.15	0.50
$n = 100$	EL	0.96	0.74	1.39	0.65	0.29
	CLT	0.92	0.72	1.26	0.54	0.10
	CLT2	0.95	0.60	1.41	0.81	0.33

Table 1: Confidence intervals on optimal value of the quadratic optimization problem

To compare EL and SRP on optimality gap, we first generate a solution $\hat{x}$ and compute its true optimality gap. Then for each of the three cases $n = 10, 50, 100$, we repeat the experiment

1000 times for each method to obtain 95% confidence bounds and estimate their empirical coverage probabilities. The results are summarized in Table 2, where the suboptimal solution $\hat{x} = 0.624$ and its corresponding optimality gap is 0.39.

		Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
$n = 10$	EL	0.95	0.05	1.89	1.84	0.92
	CLT-SRP	0.89	0	1.36	1.36	0.84
$n = 50$	EL	0.99	0.10	1.45	1.35	1.58
	CLT-SRP	0.95	0	0.77	0.77	0.27
$n = 100$	EL	0.98	0.15	0.79	0.64	0.35
	CLT-SRP	0.94	0	0.64	0.64	0.17

Table 2: Confidence intervals on optimality gap of the quadratic optimization problem

5.2 CVaR Estimation

In this example, we consider estimating $\text{CVaR}_{\alpha,F^c}(\xi)$, the α-level conditional-value-at-risk of a random variable ξ, which we assume follows an unknown distribution F^c. This can be rewritten as a stochastic optimization problem:

$$\min_{x\in\mathbb{R}} \left\{ x + \frac{1}{1-\alpha} E[(\xi - x)^+] \right\}, \tag{31}$$

where $(\cdot)^+$ is short for $\max(\cdot, 0)$. We set F^c as a standard normal distribution and $\alpha = 0.9$. As the previous example in Section 5.1, we run the experiment 1000 times for each method and each case of $n = 50, 100$ (the case of $n = 10$ is not included in this example since the data size is too small for estimating $E[(\xi - x)^+]$). To obtain confidence intervals on optimality gap, we first generate a solution $\hat{x}$ and evaluate its optimality gap using a large (10^8) sample size, and then run the experiment 1000 times for each case. The results are summarized in Table 3 and 4. Note that the true optimal value can be accurately calculated and is equal to 1.755; the suboptimal solution in this experiment is 0.71 with optimality gap 0.36.

		Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
$n = 50$	EL	0.87	1.22	2.27	1.04	0.42
	CLT	0.85	1.23	2.19	0.97	0.40
	CLT2	0.76	1.04	2.59	1.55	1.20
$n = 100$	EL	0.95	1.35	2.29	0.94	0.28
	CLT	0.90	1.37	2.09	0.72	0.21
	CLT2	0.85	1.22	2.32	1.10	0.65

Table 3: Confidence intervals on optimal value of the CVaR estimation problem

5.3 Portfolio Optimization

Our last example considers minimizing the CVaR risk associated with the loss of an investment, subject to the condition that the expected return should exceed a certain threshold. Let's denote by $x = [x^1, \ldots, x^d]'$ the vector of holding proportions in d assets, $\xi = [\xi^1, \ldots, \xi^d]'$ the random vector

		Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
$n = 50$	EL	0.99	0.04	1.51	1.47	0.51
	CLT-SRP	0.89	0	0.98	0.98	0.50
$n = 100$	EL	0.99	0.07	1.04	0.97	0.26
	CLT-SRP	0.93	0	0.76	0.76	0.29

Table 4: Confidence intervals on optimality gap of the CVaR estimation problem

of asset returns, and r_b the threshold for expected return. We assume short selling is not allowed. The problem can be written as

$$\begin{array}{ll} \min_x & CVaR_\alpha(-\xi' x) \\ \text{subject to} & E[\xi' x] \geq r_b \\ & \sum_{i=1}^d x_i = 1 \\ & x_i \geq 0, i = 1, \ldots, d \end{array} \tag{32}$$

We can rewrite the problem in the form of (3) as

$$\begin{array}{ll} \min_{x,c} & c + \frac{1}{1-\alpha} E[(-\xi' x - c)^+] \\ \text{subject to} & E[\xi' x] \geq r_b \\ & \sum_{i=1}^d x_i = 1 \\ & x_i \geq 0, i = 1, \ldots, d \end{array} \tag{33}$$

The parameter setting is as follows: ξ follows a normal distribution with mean $\mu = [0.8, 1.2]'$ and covariance $\Sigma = [1\ 0; 0\ 4]$; the minimum expected return is $r_b = 1$; the CVaR level is $\alpha = 0.9$, and the confidence level is $1 - \beta = 0.95$. It is easy to verify that the optimal solution to (33) is $x^* = [0.5, 0.5]'$, and the associated optimal value can be evaluated by Monte Carlo simulation with a large number (10^8) of samples, which yields $z^* \approx 0.96$. For comparison, we also implement the CLT and 2-sample CLT methods by computing the CIs according to (27) or (28); though the validity of these schemes has not been proved, we use them as heuristic to provide a benchmark. To compare CIs on optimality gap, we randomly generate a suboptimal solution and evaluate its optimality gap using a large (10^8) sample size. For each case of $n = 50, 100$, we repeat the experiment 1000 times, and summarize the numerical results in Table 5 and 6. In this experiment, the suboptimal solution is $[0.09, 0.91]$ with optimality gap 1.09.

		Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
$n = 50$	EL	0.63	0.11	1.22	1.11	0.60
	CLT	0.52	0.51	1.71	1.20	0.71
	CLT2	0.73	0.47	2.57	2.09	1.55
$n = 100$	EL	0.71	0.29	1.36	1.07	0.64
	CLT	0.50	0.71	1.61	0.90	0.37
	CLT2	0.69	0.62	2.01	1.39	0.96

Table 5: Confidence intervals on optimal value of the portfolio optimization problem

5.4 Summary of Numerical Results

We note in all three examples EL in general has the highest coverage probability on optimal values. Although EL in general has wider intervals than the direct CLT method, its interval widths are

		Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
$n = 50$	EL	0.90	0.25	3.24	2.99	2.41
	CLT-SRP	0.72	0	1.59	1.59	1.06
$n = 100$	EL	0.95	0.63	2.15	1.52	0.84
	CLT-SRP	0.65	0	1.13	1.13	2.51

Table 6: Confidence intervals on optimality gap of the portfolio optimization problem

often comparable to or smaller than the 2-sample CLT method, which usually has higher coverage probability than the plain CLT method. EL also has higher coverage probabilities on the optimality gap than SRP, accompanied by wider intervals than SRP. Overall speaking, EL performs competitively compared to the CLT methods.

One thing worth mentioning is that the empirical coverage probability in the last example is smaller compared to the previous two examples. A conjectured potential reason is the invalidity of both the CLT-based and the EL methods in this stochastically constrained problem. Here the EL method gives roughly comparable coverage probabilities as the CLT methods on optimal values, and higher coverage probabilities on optimality gaps.