Operations Research Letters 45 (2017) 301–307

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

The empirical likelihood approach to quantifying uncertainty in sample average approximation

Henry Lam [a,*], Enlu Zhou [b]

[a] Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, United States

[b] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, United States

ARTICLE INFO

Article history:
Received 8 April 2016
Received in revised form
8 April 2017
Accepted 8 April 2017
Available online 17 April 2017

Keywords:
Empirical likelihood
Sample average approximation
Confidence interval
Statistical uncertainty

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.

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., [13]):

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

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

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

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., [14])

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

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., [7,9], Chapter 5 in [13]).

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

* Corresponding author.
E-mail addresses: khlam@umich.edu (H. Lam), enlu.zhou@isye.gatech.edu (E. Zhou).

http://dx.doi.org/10.1016/j.orl.2017.04.003

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 (see Appendix B) 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 $w_i \geq 0$ 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

$$\begin{array}{lll} \max\limits_{w} / \min\limits_{w} & \min\limits_{x\in\Theta} \sum\limits_{i=1}^n w_i H(x;\xi_i) & \\ \text{subject to} & -2\sum\limits_{i=1}^n \log(nw_i) \leq \chi^2_{p+1,\beta} & (5) \\ & \sum\limits_{i=1}^n w_i = 1 & \\ & w_i \geq 0 \quad \text{for all } i = 1, \ldots, n & \end{array}$$

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 [12,2] 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., [4,2,15]), 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., [3]), 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 [10]. 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 [16] 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

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

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 [11]*). *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. $x^* \in \operatorname{argmin}_{x\in\Theta} h(x)$ if and only if $\nabla_x h(x^*) = 0$. Moreover, this relation is *distributionally stable*, meaning that $\tilde{x}^* \in \operatorname{argmin}_{x\in\Theta}\tilde{h}(x)$ 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 $x^* \in \operatorname{argmin}_{x\in\Theta} h(x)$ 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 a 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. $E\left[\sup_{x\in\Theta} H(x;\xi)^2\right] < \infty$.