3
Sample Average Approximation

We have seen in Chapter 2 that chance constrained problems are usually hard to solve and explicit solutions are only available in very particular cases. The main idea of SAA is to replace the original problem by an *approximate problem* obtained via sampling from the distribution of the original problem. We claim that good candidate solutions and bounds on the true optimal value can be obtained by solving such approximations. We start with the theoretical background for the method, stating and proving consistency results. The discussion follows [PAS].

3.1
Theoretical background for SAA

As stated in Chapter 1, we consider chance constrained problems

$$\begin{array}{ll} \underset{x\in X}{\text{Min}} & f(x) \\ \text{s.t.} & \mathsf{Prob}\Big\{G(x,\xi)\le 0\Big\}\ge 1-\varepsilon. \end{array} \tag{3-1}$$

In order to simplify the presentation we assume in this section that the constraint function $G:\mathbb{R}^n\times\Xi\to\mathbb{R}$ is real valued. Of course, a number of constraints $G_i(x,\xi)\le 0$, $i=1,\dots,m$, can be equivalently replaced by one constraint with the max-function as discussed in (2-6). We assume that the set X is *closed*, the function $f(x)$ is *continuous* and the function $G(x,\xi)$ is a *Carathéodory function*, i.e., $G(x,\cdot)$ is measurable for every $x\in\mathbb{R}^n$ and $G(\cdot,\xi)$ continuous for a.e. $\xi\in\Xi$.

Problem (3-1) can be written in the following equivalent form

$$\underset{x\in X}{\text{Min}}\, f(x)\ \text{ s.t. } p(x)\le\varepsilon, \tag{3-2}$$

where

$$p(x):=P\{G(x,\xi)>0\}.$$

Now let $\xi^1,\dots,\xi^N$ be an *independent identically distributed* (iid) sample of N realizations of random vector ξ and $P_N:=N^{-1}\sum_{j=1}^N\Delta(\xi^j)$ be the respective empirical measure. Here $\Delta(\xi)$ denotes measure of mass one at point ξ, and hence P_N is a discrete measure assigning probability $1/N$ to each point ξ^j,

$j = 1, \ldots, N$. The sample average approximation $\hat{p}_N(x)$ of function $p(x)$ is obtained by replacing the 'true' distribution P by the empirical measure P_N. That is, $\hat{p}_N(x) := P_N\{G(x,\xi) > 0\}$. Let $\mathbb{1}_{(0,\infty)} : \mathbb{R} \to \mathbb{R}$ be the indicator function of $(0,\infty)$, i.e.,

$$\mathbb{1}_{(0,\infty)}(t) := \begin{cases} 1, & \text{if } t > 0, \\ 0, & \text{if } t \le 0. \end{cases}$$

Then we can write that $p(x) = \mathbb{E}_P[\mathbb{1}_{(0,\infty)}(G(x,\xi))]$ and

$$\hat{p}_N(x) = \mathbb{E}_{P_N}[\mathbb{1}_{(0,\infty)}(G(x,\xi))] = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{(0,\infty)}\left(G(x,\xi^j)\right).$$

That is, $\hat{p}_N(x)$ is equal to the proportion of times that $G(x,\xi^j) > 0$. The problem, associated with the generated sample $\xi^1, \ldots, \xi^N$, is

$$\operatorname*{Min}_{x \in X} f(x) \ \text{ s.t. } \hat{p}_N(x) \le \gamma. \tag{3-3}$$

We refer to problems (3-2) and (3-3) as *the true and SAA problems*, respectively, at the respective significance levels ε and γ. Note that, following [LA] and [PAS], we allow the significance level $\gamma \ge 0$ of the approximate problem to be different from the significance level ε of the true problem. Next we discuss the convergence of a solution of the SAA problem (3-3) to that of the true problem (3-2) with respect to the sample size N and the significance level γ. A convergence analysis of problem (3-3) has been given in [LA]. Here we present complementary results under slightly different assumptions.

Recall that a sequence $f_k(x)$ of extended real valued functions is said to *epiconverge* to a function $f(x)$, written $f_k \xrightarrow{e} f$, if for any point x the following two conditions hold: (i) for any sequence x_k converging to x one has

$$\liminf_{k \to \infty} f_k(x_k) \ge f(x), \tag{3-4}$$

(ii) there exists a sequence x_k converging to x such that

$$\limsup_{k \to \infty} f_k(x_k) \le f(x). \tag{3-5}$$

Note that by the (strong) Law of Large Numbers (LLN) we have that for any x, $\hat{p}_N(x)$ converges w.p.1 to $p(x)$.

Proposition 5 *Let $G(x,\xi)$ be a Carathéodory function. Then the functions $p(x)$ and $\hat{p}_N(x)$ are lower semicontinuous, and $\hat{p}_N \xrightarrow{e} p$ w.p.1. Moreover, suppose that for every $\bar{x} \in X$ the set $\{\xi \in \Xi : G(\bar{x},\xi) = 0\}$ has P-measure zero, i.e., $G(\bar{x},\xi) \ne 0$ w.p.1. Then the function $p(x)$ is continuous at every $x \in X$ and $\hat{p}_N(x)$ converges to $p(x)$ w.p.1 uniformly on any compact set $C \subset X$, i.e.,*

$$\sup_{x \in C} |\hat{p}_N(x) - p(x)| \to 0 \text{ w.p.1 as } N \to \infty. \tag{3-6}$$

Proof. Consider function $\psi(x,\xi) := \mathbb{1}_{(0,\infty)}\big(G(x,\xi)\big)$. Recall that $p(x) = \mathbb{E}_P[\psi(x,\xi)]$ and $\hat{p}_N(x) = \mathbb{E}_{P_N}[\psi(x,\xi)]$. Since the function $\mathbb{1}_{(0,\infty)}(\cdot)$ is lower semicontinuous and $G(\cdot,\xi)$ is a Carathéodory function, it follows that the function $\psi(x,\xi)$ is random lower semicontinuous[1] (see, e.g., [RW, Proposition 14.45]). Then by Fatou's lemma we have for any $\bar{x} \in \mathbb{R}^n$,

$$\begin{aligned} \liminf_{x\to\bar{x}} p(x) &= \liminf_{x\to\bar{x}} \int_\Xi \psi(x,\xi)dP(\xi) \\ &\geq \int_\Xi \liminf_{x\to\bar{x}} \psi(x,\xi)dP(\xi) \geq \int_\Xi \psi(\bar{x},\xi)dP(\xi) = p(\bar{x}). \end{aligned}$$

This shows lower semicontinuity of $p(x)$. Lower semicontinuity of $\hat{p}_N(x)$ can be shown in the same way.

The epiconvergence $\hat{p}_N \xrightarrow{\text{e}} p$ w.p.1 is a direct implication of Artstein and Wets [AW, Theorem 2.3]. Note that, of course, $|\psi(x,\xi)|$ is dominated by an integrable function since $|\psi(x,\xi)| \leq 1$.

Suppose, further, that for every $\bar{x} \in X$, $G(\bar{x},\xi) \neq 0$ w.p.1, which implies that $\psi(\cdot,\xi)$ is continuous at $\bar{x}$ w.p.1. Then by the Lebesgue Dominated Convergence Theorem we have for any $\bar{x} \in X$,

$$\begin{aligned} \lim_{x\to\bar{x}} p(x) &= \lim_{x\to\bar{x}} \int_\Xi \psi(x,\xi)dP(\xi) \\ &= \int_\Xi \lim_{x\to\bar{x}} \psi(x,\xi)dP(\xi) = \int_\Xi \psi(\bar{x},\xi)dP(\xi) = p(\bar{x}). \end{aligned}$$

This shows that $p(x)$ is continuous at $x = \bar{x}$. Finally, the uniform convergence (3-6) follows by a version of the uniform Law of Large Numbers (see, e.g., [SHA, Proposition 7, p.363]). ∎

By lower semicontinuity of $p(x)$ and $\hat{p}_N(x)$ we have that the feasible sets of the 'true' problem (3-2) and its SAA counterpart (3-3) are closed sets. Therefore, if the set X is bounded (i.e., compact), then problems (3-2) and (3-3) have nonempty sets of optimal solutions denoted, respectively, as S and $\hat{S}_N$, provided that these problems have nonempty feasible sets. We also denote by ϑ^* and $\hat{\vartheta}_N$ the optimal values of the true and the SAA problems, respectively. The following result shows that for $\gamma = \varepsilon$, under mild regularity conditions, $\hat{\vartheta}_N$ and $\hat{S}_N$ converge w.p.1 to their counterparts of the true problem.

We make the following assumption.

(A) There is an optimal solution $\bar{x}$ of the true problem (3-2) such that for any $\varepsilon > 0$ there is $x \in X$ with $\|x - \bar{x}\| \leq \varepsilon$ and $p(x) < \varepsilon$.

[1] Random lower semicontinuous functions are called normal integrands in [RW].

In other words the above condition (A) assumes existence of a sequence $\{x_k\} \subset X$ converging to an optimal solution $\bar{x} \in S$ such that $p(x_k) < \varepsilon$ for all k, i.e., $\bar{x}$ is an accumulation point of the set $\{x \in X : p(x) < \varepsilon\}$.

Proposition 6 *Suppose that the significance levels of the true and SAA problems are the same, i.e.,* $\gamma = \varepsilon$, *the set* X *is compact, the function* $f(x)$ *is continuous,* $G(x, \xi)$ *is a Carathéodory function, and condition* (A) *holds. Then* $\hat{\vartheta}_N \to \vartheta^*$ *and* $\mathbb{D}(\hat{S}_N, S) \to 0$ *w.p.1 as* $N \to \infty$.

Proof. By the condition (A), the set S is nonempty and there is $x \in X$ such that $p(x) < \varepsilon$. We have that $\hat{p}_N(x)$ converges to $p(x)$ w.p.1. Consequently $\hat{p}_N(x) < \varepsilon$, and hence the SAA problem has a feasible solution, w.p.1 for N large enough. Since $\hat{p}_N(\cdot)$ is lower semicontinuous, the feasible set of the SAA problem is compact, and hence $\hat{S}_N$ is nonempty w.p.1 for N large enough. Of course, if x is a feasible solution of an SAA problem, then $f(x) \geq \hat{\vartheta}_N$. Since we can take such point x arbitrary close to $\bar{x}$ and $f(\cdot)$ is continuous, we obtain that

$$\limsup_{N\to\infty} \hat{\vartheta}_N \leq f(\bar{x}) = \vartheta^* \ \text{ w.p.1.} \tag{3-7}$$

Now let $\hat{x}_N \in \hat{S}_N$, i.e., $\hat{x}_N \in X$, $\hat{p}_N(\hat{x}_N) \leq \varepsilon$ and $\hat{\vartheta}_N = f(\hat{x}_N)$. Since the set X is compact, we can assume by passing to a subsequence if necessary that $\hat{x}_N$ converges to a point $\bar{x} \in X$ w.p.1. Also we have that $\hat{p}_N \xrightarrow{\text{e}} p$ w.p.1, and hence

$$\liminf_{N\to\infty} \hat{p}_N(\hat{x}_N) \geq p(\bar{x}) \ \text{ w.p.1.}$$

It follows that $p(\bar{x}) \leq \varepsilon$ and hence $\bar{x}$ is a feasible point of the true problem, and thus $f(\bar{x}) \geq \vartheta^*$. Also $f(\hat{x}_N) \to f(\bar{x})$ w.p.1, and hence

$$\liminf_{N\to\infty} \hat{\vartheta}_N \geq \vartheta^* \ \text{ w.p.1.} \tag{3-8}$$

It follows from (3-7) and (3-8) that $\hat{\vartheta}_N \to \vartheta^*$ w.p.1. It also follows that the point $\bar{x}$ is an optimal solution of the true problem and consequently we obtain that $\mathbb{D}(\hat{S}_N, S) \to 0$ w.p.1. ∎

Condition (A) is essential for the consistency of $\hat{\vartheta}_N$ and $\hat{S}_N$. Think, for example, about a situation where the constraint $p(x) \leq \varepsilon$ defines just one feasible point $\bar{x}$ such that $p(\bar{x}) = \varepsilon$. Then arbitrary small changes in the constraint $\hat{p}_N(x) \leq \varepsilon$ may result in that the feasible set of the corresponding SAA problem becomes empty. Note also that condition (A) was not used in the proof of inequality (3-8). Verification of condition (A) can be done by ad hoc methods.

Suppose now that $\gamma > \varepsilon$. Then by Proposition 6 we may expect that with increase of the sample size N, an optimal solution of the SAA problem will approach an optimal solution of the true problem with the significance level γ rather than ε. Of course, increasing the significance level leads to enlarging the feasible set of the true problem, which in turn may result in decreasing of the optimal value of the true problem. For a point $\bar{x} \in X$ we have that $\hat{p}_N(\bar{x}) \leq \gamma$, i.e., $\bar{x}$ is a feasible point of the SAA problem, iff no more than γN times the event "$G(\bar{x}, \xi^j) > 0$" happens in N trials. Since probability of the event "$G(\bar{x}, \xi^j) > 0$" is $p(\bar{x})$, it follows that

$$\mathsf{Prob}\{\hat{p}_N(\bar{x}) \leq \gamma\} = B\big(\lfloor \gamma N \rfloor; p(\bar{x}), N\big). \tag{3-9}$$

Recall that by Chernoff inequality [CHE], for $k > Np$,

$$B(k; p, N) \geq 1 - \exp\left\{-N(k/N - p)^2/(2p)\right\}.$$

It follows that if $p(\bar{x}) \leq \varepsilon$ and $\gamma > \varepsilon$, then $1 - \mathsf{Prob}\{\hat{p}_N(\bar{x}) \leq \gamma\}$ approaches zero at a rate of $\exp(-\kappa N)$, where $\kappa := (\gamma - \varepsilon)^2/(2\varepsilon)$. Of course, if $\bar{x}$ is an optimal solution of the true problem and $\bar{x}$ is a feasible point of the SAA problem, then $\hat{\vartheta}_N \leq \vartheta^*$. That is, if $\gamma > \varepsilon$, then the probability of the event "$\hat{\vartheta}_N \leq \vartheta^*$" approaches one exponentially fast. By similar analysis we have that if $p(\bar{x}) = \varepsilon$ and $\gamma < \varepsilon$, then probability that $\bar{x}$ is a feasible point of the corresponding SAA problem approaches zero exponentially fast (cf., [LA]).

The above is a qualitative analysis. For a given candidate point $\bar{x} \in X$, say obtained as a solution of a SAA problem, we would like to validate its quality as a solution of the true problem. This involves two questions, namely whether $\bar{x}$ is a feasible point of the true problem, and if so, then what is the optimality gap $f(\bar{x}) - \vartheta^*$. Of course, if $\bar{x}$ is a feasible point of the true problem, then $f(\bar{x}) - \vartheta^*$ is nonnegative and is zero iff $\bar{x}$ is an optimal solution of the true problem.

Let us start with verification of feasibility of $\bar{x}$. For that we need to estimate the probability $p(\bar{x})$. We proceed by employing again the Monte Carlo sampling techniques. Generate an iid sample $\xi^1, ..., \xi^N$ and estimate $p(\bar{x})$ by $\hat{p}_N(\bar{x})$. Note that this random sample should be generated independently of a random procedure which produced the candidate solution $\bar{x}$, and that we can use a very large sample since we do not need to solve any optimization problem here. The estimator $\hat{p}_N(\bar{x})$ of $p(\bar{x})$ is unbiased and for large N and not "too small" $p(\bar{x})$ its distribution can be approximated reasonably well by the normal distribution with mean $p(\bar{x})$ and variance $p(\bar{x})(1 - p(\bar{x}))/N$. This leads to the following approximate $(1 - \beta)$-confidence upper bound on $p(\bar{x})$:

$$U_{\beta,N}(\bar{x}) := \hat{p}_N(\bar{x}) + z_\beta \sqrt{\hat{p}_N(\bar{x})(1 - \hat{p}_N(\bar{x}))/N}. \tag{3-10}$$

A more accurate $(1-\beta)$-confidence upper bound is given by (cf., [NS]):

$$U^*_{\beta,N}(\bar{x}) := \sup_{\rho\in[0,1]} \{\rho : B(k;\rho,N) \geq \beta\}, \tag{3-11}$$

where $k := N\hat{p}_N(\bar{x}) = \sum_{j=1}^N \mathbb{1}_{(0,\infty)}(G(\bar{x},\xi^j))$.

In order to get a lower bound for the optimal value ϑ^* we proceed as follows. Let us choose two positive integers M and N, and let

$$\theta_N := B\big(\lfloor\gamma N\rfloor;\varepsilon,N\big)$$

and L be the *largest* integer such that

$$B(L-1;\theta_N,M) \leq \beta. \tag{3-12}$$

Next generate M independent samples $\xi^{1,m},\dots,\xi^{N,m}$, $m=1,\dots,M$, each of size N, of random vector ξ. For each sample solve the associated optimization problem

$$\begin{array}{ll} \underset{x\in X}{\text{Min}} & f(x) \\ \text{s.t.} & \sum_{j=1}^N \mathbb{1}_{(0,\infty)}(G(x,\xi^{j,m})) \leq \gamma N, \end{array} \tag{3-13}$$

and hence calculate its optimal value $\hat{\vartheta}^m_N$, $m=1,\dots,M$. That is, solve M times the corresponding SAA problem at the significance level γ. It may happen that problem (3-13) is either infeasible or unbounded from below, in which case we assign its optimal value as $+\infty$ or $-\infty$, respectively. We can view $\hat{\vartheta}^m_N$, $m=1,\dots,M$, as an iid sample of the random variable $\hat{\vartheta}_N$, where $\hat{\vartheta}_N$ is the optimal value of the respective SAA problem at significance level γ. Next we rearrange the calculated optimal values in the nondecreasing order as follows $\hat{\vartheta}^{(1)}_N \leq \dots \leq \hat{\vartheta}^{(M)}_N$, i.e., $\hat{\vartheta}^{(1)}_N$ is the smallest, $\hat{\vartheta}^{(2)}_N$ is the second smallest etc, among the values $\hat{\vartheta}^m_N$, $m=1,\dots,M$. We use the random quantity $\hat{\vartheta}^{(L)}_N$ as a lower bound of the true optimal value ϑ^*. It is possible to show that with probability at least $1-\beta$, the random quantity $\hat{\vartheta}^{(L)}_N$ is below the true optimal value ϑ^*, i.e., $\hat{\vartheta}^{(L)}_N$ is indeed a lower bound of the true optimal value with confidence at least $1-\beta$ (see[2] [NS]). We will discuss later how to choose the constants M,N and γ based on numerical experiments.

[2]In [NS] this lower bound was derived for $\gamma=0$. It is straightforward to extend the derivations to the case of $\gamma>0$.