Noname manuscript No.
(will be inserted by the editor)

Sample Average Approximation with Adaptive Importance Sampling

Andreas Wächter · Jeremy Staum · Alvaro Maggiar · Mingbin Feng

October 9, 2017

Abstract We study sample average approximations under adaptive importance sampling in which the sample densities may depend on previous random samples. Based on a generic uniform law of large numbers, we establish uniform convergence of the sample average approximation to the true function. We obtain convergence of the optimal value and optimal solutions of the sample average approximation. The relevance of this result is demonstrated in the context of the convergence analysis of a randomized optimization algorithm.

Keywords sample average approximation, adaptive importance sampling, likelihood ratio, parametric integration, uniform convergence

1 Introduction

We are interested in minimizing a function $g : \mathcal{X} \to \mathbb{R}$ given by

$$g(x) = \int_{\Xi} F(x, \xi) h(x, \xi) \, \mathrm{d}\xi \tag{1}$$

where $F(x, \cdot)$ is measurable for all x, and $h(x, \cdot)$ is a probability density function that might depend on x. We assume that $\mathcal{X}$ is a compact subset of $\mathbb{R}^n$. The integral $g(x)$ can be interpreted as an expectation $\mathrm{E}_x[F(x, \xi)]$ taken under the assumption that ξ is a random vector with density $h(x, \cdot)$.

When the integral (1) cannot be computed or is too expensive to evaluate, sample average approximation (SAA) provides a way to obtain an approximation of the minimizer of $g(x)$. In the most simple setting, when the probability distribution does not depend on x, that is, $h(x, \xi) = h(\xi)$, this approach consists of minimizing the sample average approximation $\hat{g}_N(x) = 1/N \sum_{i=1}^{N} F(x, \xi_i)$, where the realizations $\xi_1, \ldots, \xi_N$ of the random variable

Address(es) of author(s) should be given

are drawn from $h(\xi)$. In this case, the set of minimizers of $\hat{g}_N$ converges to the set of minimizers of $g(x)$ as $N \to \infty$, if $\hat{g}_N$ converges uniformly to g [19].

To extend this approach, consider the parametric integral

$$g(x) = \int_{\Xi} G(x, \xi) \, \mathrm{d}\xi. \tag{2}$$

Let ϕ be a sampling distribution so that $\phi(\xi) > 0$ for any ξ such that there exists an $x \in \mathcal{X}$ with $G(x, \xi) > 0$. Then, when $\{\xi_i\}_1^\infty$ is sampled i.i.d. from ϕ, the Monte Carlo estimator $1/N \sum_{i=1}^N G(x, \xi_i)/\phi(\xi_i)$ converges a.s. to $g(x)$ for all $x \in \mathcal{X}$. In the context of problem (1), define $G(x, \xi) = F(x, \xi)h(x, \xi)$. Then the estimator has the form

$$\hat{g}_N(x) = \frac{1}{N} \sum_{i=1}^{N} F(x, \xi_i) \frac{h(x, \xi_i)}{\phi(\xi_i)}.$$

This approach is known as importance sampling [18]. The sampling density ϕ may be different from the target density h. Usually, ϕ is chosen to reduce the variance of estimating the expectation of F.

The key contribution of this paper is that we provide convergence results without assuming that the samples ξ_i are independent and identically distributed. Instead, we study the convergence of the sample average approximation given by

$$\hat{g}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{G(x, \xi_i)}{\phi_i(\xi_i)}, \tag{3}$$

where, for each $i = 1, \ldots, N$, ξ_i is sampled from a different importance sampling density ϕ_i. A sampling density ϕ_i might even depend on the previous samples $\xi_1, \ldots, \xi_{i-1}$ and is therefore by itself a random variable. This setting is similar to that of adaptive multiple importance sampling [5,16]. There, however, the estimator uses mixture distributions, a case not considered here.

The pointwise convergence of $\hat{g}_N(x)$ to $g(x)$ for a single fixed x as the sample size N goes to infinity is by itself of interest and, depending on the choice ϕ_i, might be relatively elementary (see Section 4 for two examples). In Section 2, we give conditions under which pointwise convergence leads to uniform convergence of the functions $\hat{g}_N$ to g. This in turn allows us to establish the convergence of the optimal solutions of the sample average approximation

$$\min_{x \in \mathcal{X}} \ \hat{g}_N(x) \tag{4}$$

to the optimal solutions of the original optimization problem

$$\min_{x \in \mathcal{X}} \ g(x). \tag{5}$$

In Section 3 we extend this to the case when $\hat{g}_N$ depends on additional random nuisance parameters z_N that converge to a random limit point z^*. Section 5 gives simplified conditions for uniform convergence for the case that all probability distributions are normal. Finally, in Section 6 we apply our results to

prove convergence of the parameters in a quadratic regression model that approximates a stochastic function in the context of a randomized optimization algorithm.

In stochastic optimization, importance sampling has been used, for example, in the context of Benders decomposition [7,10,12]. Royset and Polak [17] presented a result on uniform convergence of the sample average approximation when $\xi_1, \ldots, \xi_N$ are independently sampled from an identical importance sampling distribution. In their work, both the target and the sampling distributions are assumed to be normal. The convergence of the sample average approximation under non-iid sampling has been addressed, for example, by Dai et al. [6]. They proved results about convergence of solutions to SAA problems when $\xi_1, \ldots, \xi_N$ are neither identically distributed nor independent, but did not discuss uniform convergence of $\hat{g}_N$ to g. Dupačkocá and Wets [9] proved epi-convergence of $\hat{g}_N$ to g, from which convergence of solutions to SAA problems follows. Their analysis assumes that $\{\phi_i\}_{i=1}^{\infty}$ converges in distribution. A similar result was obtained by Korf and Wets [14]. One of their assumptions is that $\{\xi_i\}_{i=1}^{\infty}$ forms an ergodic process, which may not be easy to verify in many applications. Homem-de-Mello [11] established results on uniform convergence of $\hat{g}_N$ to g, and of solutions to SAA problems, under non-iid sampling. His results were generalized by Xu [21]. While the these papers consider non-iid sampling, our results are more general since they permit distributions that are adaptively chosen based on the previous samples.

2 Uniform Convergence

To recapitulate with more mathematical detail: let $\mathcal{X}$ be a compact subset of $\mathbb{R}^n$, Ξ be a subset of $\mathbb{R}^d$ and G be a function from $\mathcal{X} \times \mathbb{R}^d$ to $\mathbb{R}$ whose support is contained in $\mathcal{X} \times \Xi$. Let $(\Omega, \mathcal{G}, \mathbb{Q})$ be a probability space on which there is an infinite sequence of random vectors $\{\xi_i\}_{i=1}^{\infty}$, each ξ_i being a $\mathcal{G}$-measurable function from Ω to $\mathbb{R}^d$. Define $\{\mathcal{F}_i\}_{i=1}^{\infty}$ as the natural filtration of this sequence, i.e., $\mathcal{F}_i$ contains the information in $\xi_1, \ldots, \xi_i$. Suppose that under $\mathbb{Q}$, for every $i \in \mathbb{N}$, the conditional distribution of ξ_i given $\mathcal{F}_{i-1}$ has a density ϕ_i. Let Ξ_i represent the support of ϕ_i; this subset of $\mathbb{R}^d$ can be random. Suppose that $G : \mathcal{X} \times \Xi \to \mathbb{R}$ be a real-valued function so that, for all $x \in \mathcal{X}$, (2) exists and is finite.

We are concerned with uniform convergence as $N \to \infty$ of the sample average approximation $\hat{g}_N$ defined by (3) to the function g defined by (2). The following assumption ensures that the ratios in (3) are finite.

Assumption 1 *With probability one, for every $i \in \mathbb{N}$, $\Xi \subseteq \Xi_i$.*

Our strategy is to assume that a pointwise strong law of large numbers applies (Assumption 2), and then to specify a Lipschitz-type condition (Assumption 3) that guarantees that the convergence is uniform.

Assumption 2 *For all $x \in \mathcal{X}$, w.p. 1, $\lim_{N\to\infty} |\hat{g}_N(x) - g(x)| = 0$.*

In Section 4 we discuss two pointwise laws of large numbers, including one in which $\{\xi_i\}_{i=1}^{\infty}$ is neither independently nor identically distributed. The following Lipschitz assumption corresponds to Assumption S-LIP in [1].

Assumption 3 *There exists a function $\gamma : \mathbb{R}_+ \to \mathbb{R}$ such that $\lim_{\delta\to 0} \gamma(\delta) = 0$ and, for every $i \in \mathbb{N}$, there exists a (random) measurable function $\gamma_i : \Xi_i \to \mathbb{R}$, such that*

$$\sup_{N\in\mathbb{N}} \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}[\gamma_i(\xi_i)] < \infty, \tag{6}$$

and, with probability one,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (\gamma_i(\xi_i) - \mathrm{E}[\gamma_i(\xi_i)]) = 0, \tag{7}$$

and, for all $x, x' \in \mathcal{X}$ and $i \in \mathbb{N}$, with probability one,

$$\left| \frac{G(x,\xi_i)}{\phi_i(\xi_i)} - \frac{G(x',\xi_i)}{\phi_i(\xi_i)} \right| \le \gamma_i(\xi_i)\gamma(\|x - x'\|_2). \tag{8}$$

Lipschitz-type conditions similar to (8) are common in uniform convergence results (see, for example, [8,13,20]). Together with the compactness of the parameters, it allows for the extension of pointwise results to uniform ones. The Lipschitz constants are allowed to vary from sample to sample to accommodate a greater variety of sampling distributions, so long as they satisfy the regularity conditions given by (6) and (7). For the case of normal distributions, Section 5 presents conditions that are easier to verify than those above.

The next theorem follows from Theorem 3(b) in [1]. It establishes uniform convergence of the estimator $\hat{g}_N$ to g.

Theorem 1 *If Assumptions 1, 2, and 3 hold, then, with probability one, $\lim_{N\to\infty} \|\hat{g}_N - g\|_\infty = 0$.*

Next we consider the convergence of the optimal solutions of the sample average approximation (4) to the optimal solution of the original problem (5). Let $\hat{\vartheta}_N$ and ϑ_* denote the optimal objective values of (4) and (5), respectively. Similarly, let $\hat{S}_N$ and S_* denote the set of optimal solutions of (4) and (5), respectively. Finally, we define the distance of a point $x \in \mathcal{X}$ to a set $B \subseteq \mathcal{X}$ as $\mathrm{dist}(x, B) = \inf_{x'\in B} \|x - x'\|_2$ and the deviation of a set $A \subseteq \mathcal{X}$ from the set B as $\mathbb{D}(A, B) = \sup_{x\in A} \mathrm{dist}(x, B)$.

Theorem 2 *Suppose that Assumptions 1, 2, and 3 hold, that (i) $G(\cdot, \xi)$ is lower semi-continuous for all $\xi \in \mathbb{R}^d$, and (ii) that there exists an integrable function $Z(\xi)$ such that $G(x, \xi) \ge Z(\xi)$ for all $x \in \mathcal{X}$ and almost all $\xi \in \Xi$. Further assume that there exists a compact set $C \subseteq \mathcal{X}$ such that S_* is non-empty and contained in C, and with probability one, for N large enough, $\hat{S}_N$ is non-empty and contained in C. Then, with probability one, $\lim_{N\to\infty} \hat{\vartheta}_N = \vartheta_*$ and $\lim_{N\to\infty} \mathbb{D}(\hat{S}_N, S_*) = 0$.*

Having established the uniform convergence in Theorem 1, the proof of Theorem 2 follows closely the proof of Theorem 5.3 in [19]. (See Appendix A.)

3 Results When Some Parameters Converge

In this section we consider the situation in which the vector x in the parametric integral (2) is partitioned into optimization variables y and nuisance parameters z, writing $x = (y, z)$. We provide results relevant to sample average approximation and optimization over y alone, where the sample average approximations are constructed using a convergent sequence of random values of the z parameters. For example, z may represent estimators of statistical parameters, decisions that are updated and converge over time, etc. Section 6 describes an example in which z corresponds to the iterates of a randomized optimization algorithm.

To be mathematically precise, let us assume that in the framework established in Section 2, $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{Y} \subseteq \mathbb{R}^{n_y}$ and $\mathcal{Z} \subseteq \mathbb{R}^{n_z}$ for some n_y and n_z that sum to n. Further suppose there is a sequence of random vectors $\{Z_N\}_{N=1}^{\infty}$, each Z_N being a $\mathcal{G}$-measurable function from Ω to $\mathbb{R}^{n_z}$. This sequence need *not* be adapted to the filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$. We analyze problems in which this sequence converges to a limiting random variable Z_*.

Assumption 4 *There exists a random variable Z_* such that $\lim_{N\to\infty} \|Z_N - Z_*\|_2 = 0$ with probability one.*

We study the convergence of sample average approximations $\hat{g}_N^Z : \Omega \to L_\infty(\mathcal{Y})$ given by $\hat{g}_N^Z(y) = \frac{1}{N}\sum_{i=1}^N \frac{G(y,Z_N,\xi_i)}{\phi_i(\xi_i)}$ to the function $g^Z : \Omega \to L_\infty(\mathcal{Y})$ given by $g^Z(y) = g(y, Z_*)$.

The following result is a generalization of Theorem 1 in this context. Here, Assumptions 1, 2, and 3 refer to $G : \mathcal{X} \times \Xi \to \mathbb{R}$ with $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ and $x = (y, z) \in \mathcal{Y} \times \mathcal{Z}$.

Theorem 3 *If Assumptions 1, 2, 3, and 4 hold, then with probability one, $\lim_{N\to\infty} \|\hat{g}_N^Z - g^Z\|_\infty = 0$.*

Proof We have

$$\begin{aligned}
&\|\hat{g}_N^Z - g^Z\|_\infty = \sup_{y\in\mathcal{Y}} \left| \frac{1}{N}\sum_{i=1}^N \frac{G(y, Z_N, \xi_i)}{\phi_i(\xi_i)} - g(y, Z_*) \right| \\
&\leq \sup_{y\in\mathcal{Y}} \frac{1}{N}\sum_{i=1}^N \left| \frac{G(y, Z_N, \xi_i) - G(y, Z_*, \xi_i)}{\phi_i(\xi_i)} \right| + \sup_{y\in\mathcal{Y}} \left| \frac{1}{N}\sum_{i=1}^N \frac{G(y, Z_*, \xi_i)}{\phi_i(\xi_i)} - g(y, Z_*) \right| \\
&\overset{(8)}{\leq} \sup_{y\in\mathcal{Y}} \frac{1}{N}\sum_{i=1}^N \gamma_i(\xi_i)\gamma(\|Z_N - Z_*\|_2) + \sup_{y\in\mathcal{Y}} \left| \frac{1}{N}\sum_{i=1}^N \frac{G(y, Z_*, \xi_i)}{\phi_i(\xi_i)} - g(y, Z_*) \right|.
\end{aligned} \tag{9}$$

By Theorem 1, the second term converges to zero. For the first term, we see that

$$\frac{1}{N}\sum_{i=1}^N \gamma_i(\xi_i) = \frac{1}{N}\sum_{i=1}^N (\gamma_i(\xi_i) - \mathrm{E}[\gamma_i(\xi_i)]) + \frac{1}{N}\sum_{i=1}^N \mathrm{E}[\gamma_i(\xi_i)]$$

where, by Assumption 3, the first term converges to zero and the second term is bounded. Since Z_N converges to Z_*, we have from the continuity of γ at 0 that $\gamma(\|Z_N - Z_*\|_2) \to 0$. Hence, also the first term in (9) converges to zero.

Finally, in analogy to (5) and (4), we consider the optimization problem $\vartheta_*^Z := \min_{y\in\mathcal{Y}} g^Z(y)$ and its sample average approximation $\hat{\vartheta}_N^Z := \min_{y\in\mathcal{Y}} \hat{g}_N^Z(y)$. Let S_*^Z and $\hat{S}_N^Z$ denote the set of optimal minimizers of g^Z and $\hat{g}_N^Z$, respectively. Theorem 4 follows from Theorem 3 in the same way that Theorem 2 follows from Theorem 1.

Theorem 4 *Suppose that Assumptions 1, 2, 3, and 4 hold, that (i) $G(\cdot,\xi)$ is lower semi-continuous for all $\xi \in \mathbb{R}^d$, and (ii) that there exists an integrable function $Z(\xi)$ such that $G(y,z,\xi) \geq Z(\xi)$ for all $(y,z) \in \mathcal{Y}\times\mathcal{Z}$ and almost all $\xi \in \Xi$. Further assume that there exists a compact set $C \subseteq \mathcal{Y}$ such that, with probability one, S_*^Z is non-empty and contained in C and for N large enough, $\hat{S}_N^Z$ is non-empty and contained in C. Then, with probability one, $\lim_{N\to\infty} \hat{\vartheta}_N^Z = \vartheta_*^Z$ and $\lim_{N\to\infty} \mathbb{D}(\hat{S}_N^Z, S_*^Z) = 0$.*

4 Pointwise Strong Laws of Large Numbers

In this section, we give two examples of theorems that imply the pointwise convergence required in Assumption 2. The first is the well-known strong law of large numbers for independent and identically distributed random variables. It follows, for example, from Theorem 6.1 in [2], using the fact that ϕ_i is the density for ξ_i, and therefore $\mathrm{E}\left[\frac{G(x,\xi_i)}{\phi_i(\xi_i)}\right] = g(x)$. We however need the following assumption on the measurability of $G(x,\cdot)$.

Assumption 5 *For all $x \in \mathcal{X}$, $G(x,\cdot)$ is a measurable function on $\mathbb{R}^d$ and $g(x) < \infty$.*

Theorem 5 *Suppose Assumption 1 and 5 hold. If $\{\xi_i\}_{i=1}^\infty$ are independent and identically distributed (i.e., $\phi_i = \phi_1$ for all i), then for all $x \in \mathcal{X}$, with probability one, $\lim_{N\to\infty} |\hat{g}_N(x) - g(x)| = 0$.*

Next we establish a pointwise strong law of large numbers for the case in which $\{\xi_i\}_{i=1}^\infty$ are neither independently nor identically distributed.

Assumption 6 *There exist non-negative constants k and b such that, with probability one, for all $i \in \mathbb{N}$, $x \in \mathcal{X}$, and $\xi \in \Xi_i$, $\frac{|G(x,\xi)|}{\phi_i(\xi)} \leq k\exp(b\|\xi\|_2)$.*

Assumptions on the unconditional moment generating function of $F(x,\xi)$ in (1), for each $x \in \mathcal{X}$, are common in this type of analysis [6,11,21]. In Assumption 7, we focus instead on the moment generating function M_i of the conditional distribution of $\|\xi_i\|_2$ given $\mathcal{F}_{i-1}$, defined as $M_i(s) = \mathrm{E}[\exp(s\|\xi_i\|_2)|\mathcal{F}_{i-1}] = \int_{\Xi_i} \exp(s\|\xi\|_2)\phi_i(\xi)\,\mathrm{d}\xi$. Note that M_i is a random function.

Assumption 7 *There exists $\alpha \geq 1$ such that $\sum_{i=1}^\infty i^{-2\alpha}\mathrm{E}[M_i(2\alpha b)] < \infty$, where b is as in Assumption 6.*

In Section 5 we show that Assumption 7 is satisfied when the densities ϕ_i are normal distributions with bounded means.

Theorem 6 *Suppose Assumption 1, 5, 6, and 7 hold. Then for all $x \in \mathcal{X}$, with probability one, $\lim_{N\to\infty} |\hat{g}_N(x) - g(x)| = 0$.*

The proof requires a simple relationship that is easy to show.

Lemma 1 *Given $a, c \in \mathbb{R}$ and $r \geq 1$, we have $|a + c|^r \leq (1 + |c|)^r (1 + |a|^r)$.*

Proof (Proof of Theorem 6) For a given fixed $x \in \mathcal{X}$ and all $i, N \in \mathbb{N}$, define $U_i = \frac{G(x,\xi_i)}{\phi_i(\xi_i)} - g(x)$ and $V_N = \sum_{i=1}^N U_i$, so that $\hat{g}_N(x) - g(x) = V_N/N$. The claim of the theorem follows from Chow's strong law of large numbers for martingales (see [3]) which that states that $V_N/N \to 0$ with probability one.

The remainder of this proof verifies that our setting satisfies the conditions for the theorem in [3]. The conditions are that V_N be a martingale whose increments satisfy Chung's condition (Equation (3.2) in [4]). That is, there exists $\alpha \geq 1$ such that $\sum_{i=1}^\infty i^{-(1+\alpha)} \mathrm{E}[|U_i|^{2\alpha}] < \infty$.

To see that V_N is a martingale, recall that ϕ_i is the density of ξ_i, and therefore $\mathrm{E}[U_i] = 0$ for all $i \in \mathbb{N}$ with probability one. Letting $a = U_i + g(x) = \frac{|G(x,\xi_i)|}{\phi_i(\xi_i)}$, $c = -g(x)$ and $r = 2\alpha$ in Lemma 1, we find $\mathrm{E}\left[|U_i|^{2\alpha}\right] \leq C\left(1 + \mathrm{E}\left[\left(\frac{|G(x,\xi_i)|}{\phi_i(\xi_i)}\right)^{2\alpha}\right]\right)$, where $C = (1 + |g(x)|)^{2\alpha}$. Assumption 6 then yields

$$\begin{aligned}\mathrm{E}\left[\left(\frac{|G(x,\xi_i)|}{\phi_i(\xi_i)}\right)^{2\alpha}\right] &= \mathrm{E}\left[\mathrm{E}\left[\left(\frac{|G(x,\xi_i)|}{\phi_i(\xi_i)}\right)^{2\alpha}\middle|\mathcal{F}_{i-1}\right]\right] \\ &\leq \mathrm{E}\left[k^{2\alpha}\exp(2\alpha b\|\xi_i\|_2)|\mathcal{F}_{i-1}\right] = k^{2\alpha}\mathrm{E}[M_i(2\alpha b)].\end{aligned}$$

Since $\alpha + 1 > 1$, we have $\sum_{i=1}^\infty i^{-(1+\alpha)} < \infty$, and with Assumption 7

$$\sum_{i=1}^\infty i^{-(1+\alpha)}\mathrm{E}[|U_i|^{2\alpha}] \leq C\left(\sum_{i=1}^\infty i^{-(1+\alpha)} + k^{2\alpha}\sum_{i=1}^\infty i^{-(1+\alpha)}\mathrm{E}[M_i(2\alpha b)]\right) < \infty.$$

Hence, Chung's condition holds.

5 Normal Distributions and Smooth Functions

Assumption 3 is stated in very general terms. Now we present specific conditions that are easier to verify. We consider the case in which all density functions correspond to normal distributions with different means μ and variances σ^2, so they are of the form

$$\varphi(\mu, \sigma, \xi) = \frac{1}{(\sqrt{2\pi}\sigma)^d}\exp\left(-\frac{\|\xi - \mu\|_2^2}{2\sigma^2}\right). \tag{10}$$

Assumption 8 *Let $\Xi = \mathbb{R}^d$, and for all $x \in \mathcal{X}$ and $\xi \in \mathbb{R}^d$ we have $h(x,\xi) = \varphi(x,\bar\sigma,\xi)$ for some $\bar\sigma > 0$. Furthermore, for all $i \in \mathbb{N}$, and $\xi \in \mathbb{R}^d$, we have $\Xi_i = \mathbb{R}^d$ and $\phi_i(\xi) = \varphi(\mu_i,\sigma_i,\xi)$ for some random variables $\mu_i \in \mathbb{R}$ and $\sigma_i \geq \bar\sigma$. The sequence $\{\mu_i\}$ is uniformly bounded with probability one.*

Under this assumption, the moment generating functions

$$M_i(s) = \int \exp(s\|\xi\|_2)\phi_i(\xi)\,\mathrm{d}\xi = \frac{1}{(\sqrt{2\pi\sigma_i})^d}\int \exp\left(s\|\xi\|_2 - \frac{\|\xi_i-\mu_i\|_2^2}{2\sigma_i^2}\right)\mathrm{d}\xi,$$

are uniformly bounded for fixed s, and Assumption 7 holds (for any values of $\alpha \geq 1$ and $b > 0$). Furthermore, the following lemma establishes that the likelihood ratio has subexponential growth.

Lemma 2 *Suppose Assumption 8 holds. Then there exist constants $k_h, b_h \geq 0$ so that $\frac{h(x,\xi)}{\phi_i(\xi)} \leq k_h \exp(b_h\|\xi\|_2)$ for all $i \in \mathbb{N}$, $x \in \mathcal{X}$, and $\xi \in \Xi$.*

Proof Choose any $i \in \mathbb{N}$, $x \in \mathcal{X}$, and $\xi \in \Xi$. Then

$$\begin{aligned}
&\log\left(\frac{h(x,\xi)}{\phi_i(\xi)}\right) = \log\left(\varphi(x,\bar\sigma,\xi)\right) - \log\left(\varphi(\mu_i,\sigma_i,\xi)\right)\\
\overset{(10)}{=}& -\frac{\|x-\xi\|_2^2}{2\bar\sigma^2} + \frac{\|\xi-\mu_i\|_2^2}{2\sigma_i^2}\\
=&\frac{1}{2\bar\sigma^2}\left(-\|x\|_2^2 + 2\langle x,\xi\rangle - \|\xi\|_2^2 + \frac{\bar\sigma^2}{\sigma_i^2}\|\mu_i\|_2^2 - 2\frac{\bar\sigma^2}{\sigma_i^2}\langle\mu_i,\xi\rangle + \frac{\bar\sigma^2}{\sigma_i^2}\|\xi\|_2^2\right)\\
=&\frac{1}{2\bar\sigma^2}\left(\frac{\bar\sigma^2}{\sigma_i^2}\|\mu_i\|_2^2 - \|x\|_2^2 + 2\langle x - \frac{\bar\sigma^2}{\sigma_i^2}\mu_i,\xi\rangle + \left[\frac{\bar\sigma^2}{\sigma_i^2} - 1\right]\|\xi_i\|_2^2\right). \qquad (11)
\end{aligned}$$

By Assumption 8, $\sigma_i \geq \bar\sigma$, and the term in the square brackets is non-positive. Because $\mathcal{X}$ is compact and μ_i is bounded by Assumption 8, there exist positive constants $\tilde k$ and b_h so that for all $i \in \mathbb{N}$, $x \in \mathcal{X}$, and $\xi \in \Xi$, we have $\log\left(\frac{h(x,\xi)}{\phi_i(\xi)}\right) \leq \tilde k + b_h\|\xi\|_2$. The claim of Lemma 2 follows with $k_h = \exp(\tilde k)$.

We also require some differentiability properties for F.

Assumption 9 *Suppose, that F in (1) is continuously differentiable in x for any $\xi \in \Xi$, and that there exist $k_F, b_F > 0$ so that for any $x \in \mathcal{X}$ and $\xi \in \Xi$*

$$|F(x,\xi)| \leq k_F\exp(b_F\|\xi\|_2) \quad \text{and} \qquad (12a)$$

$$\|\nabla_x F(x,\xi)\|_2 \leq k_F\exp(b_F\|\xi\|_2). \qquad (12b)$$

Here, ∇_x denotes the gradient with respect to x.

A consequence of the final proposition is that the claims of Theorems 1, 2, 3, and 4 hold under Assumptions 8 and 9.

Proposition 1 *If Assumptions 8 and 9 hold, then Assumptions 1, 2, and 3 hold for $G(x,\xi) = F(x,\xi)h(x,\xi)$.*

Proof Suppose the assumptions of Proposition 1 hold. Assumption 8 implies Assumption 1, and Assumption 9 implies Assumption 5. We already argued above that Assumption 7 holds because of Assumption 8. Assumption 6 holds, since for any $i \in \mathbb{N}$, $x \in \mathcal{X}$, and $\xi \in \Xi$,

$$\frac{|G(x,\xi)|}{\phi_i(\xi)} = |F(x,\xi)|\frac{h(x,\xi)}{\phi_i(\xi)} \overset{(12a)}{\leq} k_F \exp(b_F\|\xi\|_2) \cdot k_h \exp(b_h\|\xi\|_2),$$

where we used Lemma 2. Therefore, Theorem 6 implies that Assumption 2 holds. It remains to prove that Assumption 3 is also implied.

Note that $\nabla_x h(x,\xi) = \frac{1}{\bar{\sigma}^2} h(x,\xi)(x-\xi)$ for all $x, \xi \in \mathbb{R}^d$. Using this and the mean value theorem, we have for all $i \in \mathbb{N}$ and $x, x' \in \mathcal{X}$ that

$$\begin{aligned}
&\left|\frac{G(x,\xi_i)}{\phi_i(\xi_i)} - \frac{G(x',\xi_i)}{\phi_i(\xi_i)}\right| = \frac{1}{\phi_i(\xi_i)}\langle \nabla_x G(\bar{x}_i,\xi_i), x - x'\rangle \\
=&\frac{1}{\phi_i(\xi_i)}\langle \nabla_x F(\bar{x}_i,\xi_i)h(\bar{x}_i,\xi_i) + F(\bar{x}_i,\xi_i)\nabla_x h(\bar{x}_i,\xi_i), x - x'\rangle \\
=&\frac{h(\bar{x}_i,\xi_i)}{\phi_i(\xi_i)}\langle \nabla_x F(\bar{x}_i,\xi_i) + \frac{1}{\bar{\sigma}^2}F(\bar{x}_i,\xi_i)(\bar{x}_i - \xi_i), x - x'\rangle \qquad (13)
\end{aligned}$$

for some $\bar{x}_i \in \{\lambda_i x + (1-\lambda_i)x' : \lambda_i \in (0,1)\}$. With $M_x = \max\{\|x\|_2 : x \in \mathcal{X}\} < \infty$, we find

$$\left\|\nabla_x F(\bar{x}_i,\xi_i) + \frac{1}{\bar{\sigma}^2}F(\bar{x}_i,\xi_i)(\bar{x}_i - \xi_i)\right\|_2 \leq k_F \frac{\bar{\sigma}^2 + M_x + 1}{\bar{\sigma}^2} \exp\left((b_F+1)\|\xi_i\|_2\right) \qquad (14)$$

where we used Assumption 9 and $\|\xi_i\|_2 \leq \exp(\|\xi_i\|_2)$.

Using similar arguments as in (11), we have with an arbitrary but fixed $\hat{x} \in \mathcal{X}$ and all $i \in \mathbb{N}$ that $\log\left(\frac{h(\bar{x}_i,\xi_i)}{h(\hat{x},\xi_i)}\right) = \frac{1}{2\bar{\sigma}^2}\left(\|\bar{x}_i\|_2^2 - \|\hat{x}\|_2^2 + 2\langle \bar{x}_i - \hat{x}, \xi\rangle\right) \leq \frac{1}{2\bar{\sigma}^2}\left(2M_x^2 + 2M_x\|\xi_i\|_2\right)$, so $h(\bar{x}_i,\xi_i) \leq h(\hat{x},\xi_i) \cdot \exp\left(\frac{M_x^2}{\bar{\sigma}^2}\right) \cdot \exp\left(\frac{M_x\|\xi_i\|_2}{\bar{\sigma}^2}\right)$. Combining this with (13) and (14) we have

$$\left|\frac{G(x,\xi_i)}{\phi_i(\xi_i)} - \frac{G(x',\xi_i)}{\phi_i(\xi_i)}\right| \leq \frac{h(\hat{x},\xi_i)}{\phi_i(\xi_i)} \cdot k_G \exp(b_G\|\xi_i\|_2) \cdot \|x - x'\|_2$$

with $k_G = k_F\left(1 + \frac{M_x+1}{\bar{\sigma}^2}\right)\exp\left(\frac{M_x^2}{\bar{\sigma}^2}\right)$ and $b_G = b_F + 1 + \frac{M_x}{\bar{\sigma}^2}$. Defining $\gamma_i(\xi_i) = \frac{k_G \exp(b_G\|\xi_i\|_2)h(\hat{x},\xi_i)}{\phi_i(\xi_i)}$, it remains to show that (6) and (7) hold.

We are now going to apply Theorem 6 to the function $G_\gamma(\hat{x},\xi) = k_G \exp(b_G\|\xi\|_2)h(\hat{x},\xi)$ with $\mathcal{X}_\gamma = \{\hat{x}\}$. For this, note that $g_\gamma(\hat{x})$ defined as

$$g_\gamma(\hat{x}) := \int_\Xi G_\gamma(\hat{x},\xi)\,\mathrm{d}\xi = \int_\Xi \frac{G_\gamma(\hat{x},\xi)}{\phi_i(\xi_i)}\phi_i(\xi_i)\,\mathrm{d}\xi = \int_\Xi \gamma_i(\xi_i)\phi_i(\xi_i)\,\mathrm{d}\xi = \mathrm{E}[\gamma_i(\xi_i)]$$

is finite. The last equality follows because ξ_i is sampled from density ϕ_i. Therefore, (6) holds, and Assumption 5 holds for $G = G_\gamma$. Further consider $\hat{g}_{\gamma,N}(\hat{x}) := \frac{1}{N}\sum_{i=1}^N \frac{G_\gamma(\hat{x},\xi_i)}{\phi_i(\xi_i)} = \frac{1}{N}\sum_{i=1}^N \gamma_i(\xi_i)$. From the definition

of G_γ and Lemma 2, we have for any $\xi \in \Xi$ that $\frac{|G_\gamma(\hat{x},\xi)|}{\phi_i(\xi)} = \frac{h(\hat{x},\xi)}{\phi_i(\xi)} \cdot k_G \exp(b_G\|\xi\|_2) \le k_h\, k_G \exp((b_h + b_G)\|\xi\|_2)$. Therefore, Assumption 6 holds for $G = G_\gamma$, and using Theorem 6 we obtain $0 = \lim_{N\to\infty} (\hat{g}_{\gamma,N}(\hat{x}) - g_\gamma(\hat{x})) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N (\gamma_i(\xi_i) - \mathrm{E}[\gamma_i(\xi_i)])$, which is (7).

6 Example: Regression Models for Step Computation in an Optimization Algorithm

As an illustration in which the importance sampling is adaptive and nuisance parameters are present, we consider the randomized optimization algorithm proposed by Maggiar et al. [15] in which a local model of the objective is constructed via a SAA regression problem in every iteration.

The algorithm in [15] addresses the minimization of the function $\overline{L} : \mathcal{Z} \to \mathbb{R}$ given by $\overline{L}(z) = \int_\Xi L(\xi)h(z,\xi)\,\mathrm{d}\xi$, where $\mathcal{Z} \subset \mathbb{R}^d$ is a compact set, $\Xi = \mathbb{R}^d$, and $h(y,\xi) = \varphi(y,\sigma,\xi)$ is the normal density with mean y and variance σ^2. The integral is finite because $L : \mathbb{R}^d \to \mathbb{R}$ is assumed to exhibit subexponential growth. $L(\xi)$ is the output of a deterministic computer simulation with input ξ and the "original" objective function one would like to minimizer. However, since L is subject to numerical noise and therefore discontinuous, the task of minimizing L is ill-defined. To overcome this difficulty, [15] proposes to minimize the convolution $\overline{L}$ as a smooth approximation of L.

The derivative-free trust-region optimization algorithm proposed in [15] utilizes an SAA approximation $\overline{L}_N(z) = \frac{1}{N}\sum_{i=1}^N L(\xi_i)\frac{\varphi(z,\sigma,\xi_i)}{\varphi(t_i,\sigma,\xi_i)}$ of $\overline{L}$. The points ξ_i are sampled randomly according to the normal pdf $\varphi(t_i,\sigma,\cdot)$, where its mean t_i is either an iterate or a trial point encountered by the algorithm up to iteration N. Note that the likelihood ratio in the definition of $\overline{L}(z)$ has the form of that in (3) and therefore falls into our framework.

Given an iterate $z_N \in \mathcal{Y}$, the optimization algorithm generates a trial point $\overline{z}_N$ as the minimizer of a quadratic model within a ball around z_N. The model has the form $q_N(\xi; z_N) = b + \langle g, \xi - z_N\rangle_2 + \frac{1}{2}\langle \xi - z_N, Q(\xi - z_N)\rangle$, with coefficients $b \in \mathbb{R}$, $g \in \mathbb{R}^d$. The matrix $Q \in \mathbb{R}^{d\times d}$ is symmetric, and $q_N(\xi; z_N)$ should approximate the simulation output $L(\xi)$ for ξ close to z_N. Convergence of the optimization algorithm would follow if the model parameters are computed by a weighted local regression of L; that is, if $y = (b, g, Q)$ are the minimizers of

$$\min_{y\in\mathcal{Y}} \int_\Xi F(y,z,\xi)h(z,\xi)\,\mathrm{d}\xi, \tag{15}$$

where $F(y,z,\xi) = (b + \langle g, \xi - z\rangle_2 + \frac{1}{2}\langle \xi - z, Q(\xi - z)\rangle_2 - L(\xi))^2$. This objective function has the form of (2). (In abuse of notation, we collect the model parameters b, g, and Q in the vector y.)

To get an approximate solution of (15), at an iterate Z_N (using an upper case letter to emphasize its stochastic nature), the optimization algorithm computes the quadratic model from the stochastic average approximation of

(15); that is

$$\min_{y\in\mathcal{Y}} \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(Z_N,\sigma,\xi_i)}{\varphi(T_i,\sigma,\xi_i)} \left(b + \langle g, \xi_i - Z_N\rangle_2 + \frac{1}{2}\langle \xi_i - Z_N, Q(\xi_i - z_N)\rangle_2 - L(\xi) \right)^2 . \tag{16}$$

The analysis of the algorithm in [15] requires that the model $q_N(\xi; Z_N)$ converges to the optimal solution of (15) at any limit point Z_* of the iterates Z_N. This can be proved using the results in Section 3.

For any $\omega \in \Omega$, let $\{Z_N(\omega)\}_{N=1}^{\infty}$ be a subsequence of iterates such that $\{Z_N(\omega)\}_{N=1}^{\infty}$ converges to a limit point $Z_*(\omega)$. Such a subsequence exists, due to compactness of $\mathcal{Z}$; thus Assumption 4 holds. Furthermore, since $F(y,z,\xi)$ is a polynomial in (y,z) and L exhibits subexponential growth, Assumption 9 holds. Also, because all iterates and trial points are contained in $\mathcal{Z}$, the sequence $\{T_i\}$, consisting of such points, is uniformly bounded. Finally, the algorithm in [15] ensures that the optimal solutions of (15) and (16) are unique and uniformly bounded, by monitoring the condition number of matrices involved in the computation of the optimal solution of (16). In summary, Assumptions 4 and 9 hold, and Proposition 1 together with Theorem 2 yields $\lim_{N\to\infty} \mathbb{D}(\hat{S}_N^Z, S_*^Z) = 0$. So, the approximate model parameters in $\hat{S}_N^Z$ in iteration N converge to the optimal parameters in S_*^Z.

7 Conclusion

We considered the sample average approximation of stochastic optimization problems whose objective function is expressed as a parametric integral. The key contribution is that we permit non-independent, non-identical, and adaptive sampling, where the importance sampling distribution may depend on previous samples. Under the assumption of pointwise convergence and a stochastic Lipschitz condition, we proved uniform convergence of the sample average approximation of the parametric integral over a compact set as well as convergence of the optimal values and optimal solution sets of the sample average approximation problems as the number of samples goes to infinity.

Acknowledgments

We thank Tito Homem-de-Mello, David Morton, Imry Rosenbaum, and Johannes Royset for discussions.

References

1. Andrews, D.W.: Generic uniform convergence. Econometric theory **8**(02), 241–257 (1992)
2. Billingsley, P.: Probability and Measure, 3rd edn. John Wiley & Sons (1995)
3. Chow, Y.S.: On a strong law of large numbers for martingales. The Annals of Mathematical Statistics **38**(2), 610–610 (1967)

4. Chung, K.L.: The strong law of large numbers. In: Processdings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950, pp. 341–352. University of California Press, Berkeley and Los Angeles (1951)
5. Cornuet, J.M., Marin, J.M., Mira, A., Robert, C.P.: Adaptive multiple importance sampling. Scandinavian Journal of Statistics **39**, 798–812 (2012)
6. Dai, L., Chen, C.H., Birge, J.R.: Convergence properties of two-stage stochastic programming. Journal of Optimization Theory and Applications **106**(3), 489–509 (2000)
7. Dantzig, G.B., Glynn, P.W.: Parallel processors for planning under uncertainty. Annals of Operations Research **22**(1), 1–21 (1990)
8. Duffie, D., Singleton, K.J.: Simulated moments estimation of markov models of asset prices (1990)
9. Dupačová, J., Wets, R.: Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. The Annals of Statistics pp. 1517–1549 (1988)
10. Glynn, P.W., Infanger, G.: Simulation-based confidence bounds for two-stage stochastic programs. Mathematical Programming **138**(1), 15–42 (2013)
11. Homem-de-Mello, T.: On rates of convergence for stochastic optimization problems under non-independent and identically distributed sampling. SIAM Journal on Optimization **19**(2), 524–551 (2008)
12. Infanger, G.: Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. Annals of Operations Research **39**(1), 69–95 (1992)
13. Jenish, N., Prucha, I.R.: Central limit theorems and uniform laws of large numbers for arrays of random fields. Journal of econometrics **150**(1), 86–98 (2009)
14. Korf, L., Wets, R.J.B.: Random LSC functions: An ergodic theorem. Mathematics of Operations Research **26**(2), 421–445 (2001)
15. Maggiar, A., Wächter, A., Dolinskaya, I.S., Staum, J.: A derivative-free trust-region algorithm for the optimization of functions smoothed via Gaussian convolution using multiple importance sampling (2015). `http://www.optimization-online.org/DB_HTML/2015/07/5017.html`
16. Marin, J.M., Pudlo, P., Sedki, M.: Consistency of the adaptive multiple importance sampling (2014). ArXiv:1211.2548v2
17. Royset, J.O., Polak, E.: Implementable algorithm for stochastic optimization using sample average approximations. Journal of Optimization Theory and Applications **122**(1), 157–184 (2004)
18. Rubinstein, R.Y., Kroese, D.P.: Simulation and the Monte Carlo method, 3rd edn. John Wiley & Sons (2017)
19. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM, Philadelphia (2009)
20. Shapiro, A., Xu, H.: Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. Journal of mathematical analysis and applications **325**(2), 1390–1399 (2007)
21. Xu, H.: Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming. Journal of Mathematical Analysis and Applications **368**, 692–710 (2010)

A Proof of Theorem 4

We establish the result in two lemmas.

Lemma 3 *Suppose Assumptions 1, 2, and 3 hold. Further assume that S_* is not empty and that, with probability one, $\hat{S}_N$ is non-empty for all N sufficiently large. Then $\lim_{N\to\infty} \hat{\vartheta}_N = \vartheta_*$ with probability one.*

Proof We prove $\lim_{N\to\infty} \hat{\vartheta}_N = \vartheta_*$ in the event that $\hat{S}_N$ is non-empty for all N sufficiently large and that $\lim_{N\to\infty} \|\hat{g}_N - g\|_\infty = 0$. This event has probability one by assumption and by Theorem 1.

Let x_* be an optimal solution of (5). Because $\lim_{N\to\infty} \|\hat{g}_N - g\|_\infty = 0$, $\lim_{N\to\infty} \hat{g}_N(x_*) = g(x_*) = \vartheta_*$. Since $\hat{\vartheta}_N$ is the optimal value of (4), $\hat{\vartheta}_N \leq \hat{g}_N(x_*)$ for all N. As a consequence, $\limsup_{N\to\infty} \hat{\vartheta}_N \leq \vartheta_*$.

Define $\hat{\vartheta}_{\inf} = \liminf_{N\to\infty} \hat{\vartheta}_N$. There exist a subsequence $\{N_i\}_{i=1}^\infty$ of the natural numbers and a sequence $\{x_N\}_{N=1}^\infty$ of points in $\mathcal{X}$ such that for every $i = 1, \ldots, \infty$, $x_{N_i} \in \hat{S}_{N_i}$, and $\lim_{i\to\infty} \hat{g}_{N_i}(x_{N_i}) = \hat{\vartheta}_{\inf}$. Because $\lim_{N\to\infty} \|\hat{g}_N - g\|_\infty = 0$, we also have $\lim_{i\to\infty} g(x_{N_i}) = \hat{\vartheta}_{\inf}$. Since ϑ_* is the optimal value of (5), $\vartheta_* \leq g(x_{N_i})$ for all N_i. Therefore $\vartheta_* \leq \hat{\vartheta}_{\inf}$.

Overall, we have obtained $\limsup_{N\to\infty} \hat{\vartheta}_N \leq \vartheta_* \leq \liminf_{N\to\infty} \hat{\vartheta}_N$.

Lemma 4 *Suppose the assumptions of Theorem 2 hold. Then, w.p.1,* $\lim_{N\to\infty} \mathbb{D}(\hat{S}_N, S_*) = 0$.

Proof We prove $\lim_{N\to\infty} \mathbb{D}(\hat{S}_N, S_*) = 0$ in the event that $\lim_{N\to\infty} \|\hat{g}_N - g\|_\infty = 0$, $\lim_{N\to\infty} \hat{\vartheta}_N = \vartheta_*$, and $\hat{S}_N$ is non-empty and contained in C for all N sufficiently large. This event has probability one by Theorem 1, by Lemma 3, and by assumption.

Consider any subsequence $\{N_i\}_{i=1}^\infty$ of the natural numbers and sequence $\{x_N\}_{N=1}^\infty$ of points in $\mathcal{X}$ such that for every $i = 1, \ldots, \infty$, $x_{N_i} \in \hat{S}_{N_i}$. Because C is compact, the sequence $\{x_{N_i}\}_{i=1}^\infty$ has a limit point. Consider any such limit point, and denote it as x^*. Consider any subsequence $\{N_i'\}_{i=1}^\infty$ of $\{N_i\}_{i=1}^\infty$ such that $\lim_{i\to\infty} x_{N_i'} = x^*$. For any i,

$$\hat{\vartheta}_{N_i'} - g(x^*) = \hat{g}_{N_i'}(x_{N_i'}) - g(x^*) = \left(\hat{g}_{N_i'}(x_{N_i'}) - g(x_{N_i'})\right) + \left(g(x_{N_i'}) - g(x^*)\right).$$

It follows from assumptions (i) and (ii) in Theorem 2 and Theorem 7.47 in [19] that g is lower semi-continuous, which in turn implies that $\liminf_{i\to\infty}(g(x_{N_i'}) - g(x^*)) \geq 0$. We also have $\lim_{i\to\infty}(\hat{g}_{N_i'}(x_{N_i'}) - g(x_{N_i'})) = 0$ since $\lim_{N\to\infty} \|\hat{g}_N - g\|_\infty = 0$. Therefore $\lim_{i\to\infty} \hat{\vartheta}_{N_i'} \geq g(x^*)$. We also have $\lim_{N\to\infty} \hat{\vartheta}_N = \vartheta_*$. Thus, $g(x^*) \leq \vartheta_*$, which implies $x^* \in S_*$.

In words: if x^* is a limit point of a sequence $\{x_{N_i}\}_{i=1}^\infty$ of points that are optimal solutions of a sequence of sample average approximation problems given by (4), then x^* is in S_*. Therefore,

$$\limsup_{N\to\infty} \mathbb{D}(\hat{S}_N, S_*) = \limsup_{N\to\infty} \sup_{x\in\hat{S}_N} \operatorname{dist}(x, S_*) = 0.$$