Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

UNIFORM CONVERGENCE OF SAMPLE AVERAGE APPROXIMATION WITH ADAPTIVE MULTIPLE IMPORTANCE SAMPLING

M. Ben Feng

Deparment of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario, CANADA

Alvaro Maggiar

Supply Chain Optimization Technologies
Amazon.com
Seattle, WA 98109, USA

Jeremy Staum
Andreas Wächter

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA

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 function being approximated. In the optimization context, we obtain convergence of the optimal value and optimal solutions of the sample average approximation.

1 INTRODUCTION

We are interested in approximating or optimizing a function $g : \mathscr{X} \to \mathbb{R}$ 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 $\mathscr{X}$ 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 $h(x,\cdot)$. Defining $G(x,\xi) = F(x,\xi)h(x,\xi)$, we can rewrite (1) compactly as

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

For fixed x, if the integral (1) cannot be computed explicitly, simple Monte Carlo method can be applied to estimate $g(x)$ by the sample average $\widehat{g}_N^{MC}(x) = 1/N\sum_{i=1}^N F(x,\xi_i)$, where the random samples $\xi_1,\ldots,\xi_N$ are drawn from $h(x,\xi)$. In the optimization context where one wants to minimize $g(x)$, sample average approximation (SAA) provides a way to obtain an approximation of the minimizer of $g(x)$. In the simplest setting, when the probability distribution does not depend on x, i.e., $h(x,\xi) = h(\xi)$, SAA minimizes $\widehat{g}_N^{MC}(x)$ instead. In this case, the set of minimizers of $\widehat{g}_N^{MC}$ converges to the set of minimizers of $g(x)$ as $N \to \infty$, if $\widehat{g}_N^{MC}$ converges uniformly to g (Shapiro et al. 2009). In statistics, the function class for which the sample average converges uniformly is called Glivenko-Cantelli (Van der Vaart 2000).

Other Monte Carlo methods, such as importance sampling (Rubinstein and Kroese 2017), are applicable in this setting. Let ϕ be a sampling distribution on Ξ so that $\phi(\xi) > 0$ for any ξ such that there exists

978-1-5386-6572-5/18/$31.00 ©2018 IEEE

an $x \in \mathscr{X}$ with $h(x,\xi) \neq 0$. Then, when ξ is sampled from ϕ, for all $x \in \mathscr{X}$, the importance sampling estimator $G(x,\xi)/\phi(\xi)$ has expectation $g(x)$. Written in the notation of (1), this estimator has the form $\widehat{g}_N(x) = \frac{1}{N}\sum_{i=1}^{N} F(x,\xi_i)\frac{h(x,\xi_i)}{\phi(\xi_i)}$. 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.

One main contribution of this paper is to provide convergence results without assuming that the samples ξ_i are independent and identically distributed. Instead, we study the convergence of the SAA given by

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

where, for each $i = 1,\ldots,N$, ξ_i is sampled from a different density ϕ_i. A sampling density ϕ_i might 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 (Cornuet et al. 2012; Marin et al. 2014); see Section 1.1 for elaborated discussions.

The pointwise convergence of $\widehat{g}_N(x)$ 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 $\widehat{g}_N$ to g, i.e., convergence in the L_∞ sense. This in turn allows us to establish the convergence of the optimal solutions of the SAA problem

$$\min_{x\in\mathscr{X}} \ \widehat{g}_N(x) \tag{4}$$

to the optimal solutions of the original optimization problem

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

In Section 3 we extend this to the case when $\widehat{g}_N$ 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.

The convergence of the SAA under non-iid sampling, i.e., when $\xi_1,\ldots,\xi_N$ are not independent and identically distributed, has been addressed by, for example, Dai et al. (2000). 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 $\widehat{g}_N$ to g. Dupačová and Wets (1988) proved epi-convergence of $\widehat{g}_N$ 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 (2001). 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 (2008) established results on uniform convergence of $\widehat{g}_N$ to g, and of solutions to SAA problems, under non-iid sampling. His results were generalized by Xu (2010). While these papers consider non-iid sampling, our results are more general since they permit distributions that are adaptively chosen based on the previous samples.

1.1 Importance Sampling and the Likelihood Ratio Method

Estimating $g(x)$ by sampling from a density other than h or $h(x,\cdot)$ is called "importance sampling" (IS) or the "likelihood ratio" (LR) method. IS and the LR method are distinguished by the purpose for which they are employed, not by their mathematics. The purpose of the LR method (Rubinstein and Shapiro 1993) is to allow samples from a density ϕ to be used to estimate the function g. The purpose of IS is to reduce variance in estimating $g(x)$ by choosing an appropriate sampling density ϕ or sampling densities $\phi_1,\ldots,\phi_N$ from which to sample. "Multiple importance sampling" (MIS) is the specific name for using multiple sampling densities in IS. In standard forms of MIS, the sampling densities $\phi_1,\dots,\phi_N$ are regarded as fixed; ϕ_i cannot depend on $\xi_1,\dots,\xi_{i-1}$. Cornuet et al. (2012) proposed adaptive multiple importance sample (AMIS), which allows for such dependence: $\xi_1,\dots,\xi_N$ are sampled from different sampling densities $\phi_1,\dots,\phi_N$, respectively, and the sampling distribution ϕ_i for ξ_i can depend on $\xi_1,\dots,\xi_{i-1}$. This is the setting within which we study the convergence of the SAA $\widehat{g}_N$ given in (3). However, we adopt the perspective of the LR method rather than the perspective of IS: we suppose that there is a sequence of random densities $\phi_1,\dots,\phi_N$ each depending on previous sampling, but we focus on using the sample $\xi_1,\dots,\xi_N$ to estimate the function g and not on how the densities $\phi_1,\dots,\phi_N$ should be chosen to minimize variance. That question is investigated in the literature on AMIS when the goal is to estimate $g(x)$ for fixed x. In the LR method, when the goal is to estimate the function g, the choice of a single sampling density ϕ was investigated by Rubinstein and Shapiro (1993).

When multiple sampling densities are used, an alternative estimator to (3) using the "balance heuristic weights" by Veach and Guibas (1995) has been employed. Regarding the sampling of $\xi_1,\dots,\xi_N$ as stratified sampling from the mixture distribution $\sum_{j=1}^N \phi_j/N$, justifies the mixture likelihood ratio (MLR) estimator $\sum_{i=1}^N \frac{G(x,\xi_i)}{\sum_{j=1}^N \phi_j(\xi_i)}$ instead of (3). In some situations, the MLR estimator performs much better than (3) (Feng and Staum 2015; Feng and Staum 2017; Owen and Zhou 2000). The convergence of AMIS, i.e., consistency in the statistical sense, is addressed by Marin et al. (2014), using a weak law of large numbers for triangular arrays. AMIS procedures vary in the way that they learn from function values and reuse them. Consistency results for some variants of AMIS are yet to be obtained (Marin et al. 2014). We are not aware of any previous uniform convergence results for AMIS. We hope that the present work for the specific approximation (3) will further research on strong and uniform convergence of different AMIS schemes, such as MLR.

In stochastic optimization, IS has been used, for example, in the context of Benders decomposition (Dantzig and Glynn 1990; Glynn 2013; Infanger 1992). Royset and Polak (2004) presented a result on uniform convergence of the SAA when $\xi_1,\dots,\xi_N$ are independently sampled from an identical sampling distribution. In their work, both the target and the sampling distributions are assumed to be normal.

2 UNIFORM CONVERGENCE

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

We are concerned with uniform convergence as $N\to\infty$ of the SAA $\widehat{g}_N$ 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 $x\in\mathscr{X}$, w.p. 1, $\lim_{N\to\infty}|\widehat{g}_N(x)-g(x)|=0$.

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 Andrews (1992).

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

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

Lipschitz-type conditions similar to (8) are common in uniform convergence results (see, for example, Duffie and Singleton (1993), Jenish and Prucha (2009), Shapiro and Xu (2007)). 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 Andrews (1992). It establishes uniform convergence of the estimator $\widehat{g}_N$ to g.

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

Next we consider the convergence of the optimal solutions of the SAA (4) to the optimal solution of the original problem (5). Let $\widehat{\vartheta}_N$ and ϑ_* denote the optimal objective values of (4) and (5), respectively. Similarly, let $\widehat{S}_N$ and S_* denote the set of optimal solutions of (4) and (5), respectively. Finally, we define the distance of a point $x \in \mathscr{X}$ to a set $B \subseteq \mathscr{X}$ as $\mathrm{dist}(x,B) = \inf_{x' \in B} \|x - x'\|_2$ and the deviation of a set $A \subseteq \mathscr{X}$ 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 $G(x,\xi) \geq Z(\xi)$ for all $x \in \mathscr{X}$ and almost all $\xi \in \Xi$. Further assume that there exists a compact set $C \subseteq \mathscr{X}$ such that S_* is non-empty and contained in C, and with probability one, for N large enough, $\widehat{S}_N$ is non-empty and contained in C. Then, with probability one, $\lim_{N \to \infty} \widehat{\vartheta}_N = \vartheta_*$ and $\lim_{N \to \infty} \mathbb{D}(\widehat{S}_N, S_*) = 0$.

The proof of this theorem follows very closely the proof of Theorem 5.3 in Shapiro et al. (2009), but we include it here for completeness because the conditions are slightly different. We establish the result in two lemmas.

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

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

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

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

Lemma 2 Suppose the assumptions of Theorem 2 hold. Then, with probability one, $\lim_{N\to\infty}\mathbb{D}(\widehat{S}_N, S_*) = 0$.

Proof. We prove $\lim_{N\to\infty}\mathbb{D}(\widehat{S}_N, S_*) = 0$ in the event that $\lim_{N\to\infty}\|\widehat{g}_N - g\|_\infty = 0$, $\lim_{N\to\infty}\widehat{\vartheta}_N = \vartheta_*$, and $\widehat{S}_N$ is non-empty and contained in C for all N sufficiently large. This event has probability one by Theorem 1, by Lemma 1, 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 $\mathscr{X}$ such that for every $i = 1,\ldots,\infty$, $x_{N_i} \in \widehat{S}_{N_i}$. 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, $\widehat{\vartheta}_{N_i'} - g(x^*) = \widehat{g}_{N_i'}(x_{N_i'}) - g(x^*) = \left(\widehat{g}_{N_i'}(x_{N_i'}) - g(x_{N_i'})\right) + \left(g(x_{N_i'}) - g(x^*)\right)$. It follows from assumptions (i) and (ii) in Theorem 2 and Theorem 7.47 in Shapiro et al. (2009) 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 $\lim_{i\to\infty}(\widehat{g}_{N_i'}(x_{N_i'}) - g(x_{N_i'})) = 0$ since $\lim_{N\to\infty}\|\widehat{g}_N - g\|_\infty = 0$. Therefore $\lim_{i\to\infty}\widehat{\vartheta}_{N_i'} \geq g(x^*)$. We also have $\lim_{N\to\infty}\widehat{\vartheta}_N = \vartheta_*$. Thus, $g(x^*) \leq \vartheta_*$, which implies $x^* \in S_*$. Overall, we show that 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 SAA problems given by (4). Therefore x^* is in S_* so $\limsup_{N\to\infty}\mathbb{D}(\widehat{S}_N, S_*) = \limsup_{N\to\infty}\sup_{x\in\widehat{S}_N}\operatorname{dist}(x, S_*) = 0$. □

3 RESULTS WHEN SOME PARAMETERS CONVERGE

In this section we consider the situation in which the vector x in the parametric integral (2) may include some parameters that are of primary interest and other parameters that are of auxiliary interest or are mere nuisance parameters. We write $x = (y,z)$ where y is of primary interest and z is not. We provide results relevant to SAA and optimization over y alone, where the SAAs are constructed using a convergent sequence of random values of the z parameters. For example, these values of the z parameters may represent estimators of statistical parameters, stochastic processes that converge to a limiting random variable, 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, $\mathscr{X} = \mathscr{Y} \times \mathscr{Z}$, where $\mathscr{Y} \subseteq \mathbb{R}^{n_y}$ and $\mathscr{Z} \subseteq \mathbb{R}^{n_z}$ 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 $\mathscr{G}$-measurable function from Ω to $\mathbb{R}^{n_z}$. This sequence need *not* be adapted to the filtration $\{\mathscr{F}_i\}_{i=1}^\infty$. 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 SAAs $\widehat{g}_N^Z : \Omega \to L_\infty(\mathscr{Y})$ given by $\widehat{g}_N^Z(y) = \frac{1}{N}\sum_{i=1}^N \frac{G(y,Z_N,\xi_i)}{\phi_i(\xi_i)}$ to the function $g^Z : \Omega \to L_\infty(\mathscr{Y})$ 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 $G : \mathscr{X} \times \Xi \to \mathbb{R}$ with $\mathscr{X} = \mathscr{Y} \times \mathscr{Z}$ and $x = (y,z) \in \mathscr{Y} \times \mathscr{Z}$.

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

Proof. We have

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

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 $\vartheta_*^Z := \min_{y\in\mathscr{Y}}\ g^Z(y)$ and its sample average approximation $\widehat{\vartheta}_N^Z := \min_{y\in\mathscr{Y}}\ \widehat{g}_N^Z(y)$. Let S_*^Z and $\widehat{S}_N^Z$ denote the set of optimal minimizers of g^Z and $\widehat{g}_N^Z$, 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 $(y,z)\in\mathscr{Y}\times\mathscr{Z}$ and almost all $\xi\in\Xi$. Further assume that there exists a compact set $C\subseteq\mathscr{Y}$ such that, with probability one, S_*^Z is non-empty and contained in C and for N large enough, $\widehat{S}_N^Z$ is non-empty and contained in C. Then, with probability one, $\lim_{N\to\infty}\widehat{\vartheta}_N^Z=\vartheta_*^Z$ and $\lim_{N\to\infty}\mathbb{D}(\widehat{S}_N^Z,S_*^Z)=0$.

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 Billingsley (1995), 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 $x\in\mathscr{X}$, $G(x,\cdot)$ 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 $x\in\mathscr{X}$, with probability one, $\lim_{N\to\infty}|\widehat{g}_N(x)-g(x)|=0$.

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 $i\in\mathbb{N}$, $x\in\mathscr{X}$, 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 $x\in\mathscr{X}$, are common in this type of analysis (Dai et al. 2000; Homem-de-Mello 2008; Xu 2010). In Assumption 7, we focus instead on the moment generating function M_i of the conditional distribution of $\|\xi_i\|_2$ given $\mathscr{F}_{i-1}$, defined as $M_i(s)=\mathrm{E}[\exp(s\|\xi_i\|_2)|\mathscr{F}_{i-1}]=\int_{\Xi_i}\exp(s\|\xi\|_2)\phi_i(\xi)\,\mathrm{d}\xi$. 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 $x\in\mathscr{X}$, with probability one, $\lim_{N\to\infty}|\widehat{g}_N(x)-g(x)|=0$.

For the proof of Theorem 6 we require the following relationship.

Lemma 3 Given $a,c\in\mathbb{R}$ and $r\geq 1$, the following inequality holds: $|a+c|^r\leq(1+|c|)^r(1+|a|^r)$.

Proof. Clearly, $|a+c|^r\leq(|a|+|c|)^r$. We now consider the two cases $|a|\geq 1$ and $|a|<1$.

If $|a| \geq 1$, then $(|a|+|c|)^r = |a|^r \left(1+\frac{|c|}{|a|}\right)^r \leq |a|^r (1+|c|)^r \leq (1+|a|^r)(1+|c|)^r$.

If $|a| < 1$, then $(|a|+|c|)^r \leq (1+|c|)^r \leq (1+|a|^r)(1+|c|)^r$. □

Proof of Theorem 6. For a given fixed $x \in \mathscr{X}$ 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 $\widehat{g}_N(x) - g(x) = V_N/N$. The claim of the theorem follows from Chow's strong law of large numbers for martingales (see Chow (1967)) 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 Chow (1967). The conditions are that V_N be a martingale whose increments satisfy Chung's condition (Equation (3.2) in Chung (1951)). 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 3, 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

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

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 $x \in \mathscr{X}$ 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

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

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 4 Suppose Assumption 8 holds. The there exist non-negative constants k_h and b_h so that

$$\frac{h(x,\xi)}{\phi_i(\xi)} \leq k_h\exp(b_h\|\xi\|_2) \tag{11}$$

for all $i \in \mathbb{N}$, $x \in \mathscr{X}$, and $\xi \in \Xi$.

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

$$\begin{aligned}
\log\left(\frac{h(x,\xi)}{\phi_i(\xi)}\right) &= \log\left(\varphi(x,\bar{\sigma},\xi)\right) - \log\left(\varphi(\mu_i,\sigma_i,\xi)\right) \overset{(10)}{=} -\frac{\|x-\xi\|_2^2}{2\bar{\sigma}^2} + \frac{\|\xi-\mu_i\|_2^2}{2\sigma_i^2} \\
&= \frac{1}{2\bar{\sigma}^2}\left(-\|x\|_2^2 + 2\langle x,\xi\rangle - \|\xi\|_2^2 + \frac{\bar{\sigma}^2}{\sigma_i^2}\|\mu_i\|_2^2 - 2\frac{\bar{\sigma}^2}{\sigma_i^2}\langle \mu_i,\xi\rangle + \frac{\bar{\sigma}^2}{\sigma_i^2}\|\xi\|_2^2\right) \\
&= \frac{1}{2\bar{\sigma}^2}\left(\frac{\bar{\sigma}^2}{\sigma_i^2}\|\mu_i\|_2^2 - \|x\|_2^2 + 2\langle x - \frac{\bar{\sigma}^2}{\sigma_i^2}\mu_i,\xi\rangle + \left[\frac{\bar{\sigma}^2}{\sigma_i^2} - 1\right]\|\xi\|_2^2\right). \qquad (12)
\end{aligned}$$

By Assumption 8, $\sigma_i \geq \bar{\sigma}$, and the term in the square brackets is non-positive. Because $\mathscr{X}$ is compact and μ_i is bounded by Assumption 8, there exist positive constants $\tilde{k}$ and b_h so that for all $i \in \mathbb{N}$, $x \in \mathscr{X}$, and $\xi \in \Xi$, $\log\left(\frac{h(x,\xi)}{\phi_i(\xi)}\right) \leq \tilde{k} + b_h\|\xi\|_2$. The claim of Lemma 4 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 $x \in \mathscr{X}$ and any $\xi \in \Xi$

$$|F(x,\xi)| \leq k_F \exp(b_F\|\xi\|_2) \quad \text{and} \qquad (13a)$$

$$\|\nabla_x F(x,\xi)\|_2 \leq k_F \exp(b_F\|\xi\|_2). \qquad (13b)$$

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 $i \in \mathbb{N}$, $x \in \mathscr{X}$, and $\xi \in \Xi$,

$$\frac{|G(x,\xi)|}{\phi_i(\xi)} \overset{(10)}{=} |F(x,\xi)|\frac{h(x,\xi)}{\phi_i(\xi)} \overset{(11),(13a)}{\leq} k_F \exp(b_F\|\xi\|_2) \cdot k_h \exp(b_h\|\xi\|_2).$$

where we used Lemma 4. 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 $x,x' \in \mathscr{X}$ that

$$\begin{aligned}
\left|\frac{G(x,\xi_i)}{\phi_i(\xi_i)} - \frac{G(x',\xi_i)}{\phi_i(\xi_i)}\right| &= \frac{1}{\phi_i(\xi_i)}\langle \nabla_x G(\bar{x}_i,\xi_i), x - x'\rangle \\
&= \frac{1}{\phi_i(\xi_i)}\langle \nabla_x F(\bar{x}_i,\xi_i)h(\bar{x}_i,\xi_i) + F(\bar{x}_i,\xi_i)\nabla_x h(\bar{x}_i,\xi_i), x - x'\rangle \\
&= \frac{h(\bar{x}_i,\xi_i)}{\phi_i(\xi_i)}\langle \nabla_x F(\bar{x}_i,\xi_i) + \frac{1}{\bar{\sigma}^2}F(\bar{x}_i,\xi_i)(\bar{x}_i - \xi_i), x - x'\rangle \qquad (14)
\end{aligned}$$

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

$$\begin{aligned}
\left\|\nabla_x F(\bar{x}_i,\xi_i) + \frac{1}{\bar{\sigma}^2}F(\bar{x}_i,\xi_i)(\bar{x}_i - \xi_i)\right\|_2 &\leq k_F \exp(b_F\|\xi_i\|_2) + \frac{1}{\bar{\sigma}^2}k_F \exp(b_F\|\xi_i\|_2)(M_x + \|\xi_i\|_2) \\
&\leq k_F\left(1 + \frac{M_x+1}{\bar{\sigma}^2}\right)\exp\left((b_F+1)\|\xi_i\|_2\right), \qquad (15)
\end{aligned}$$

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

Using similar arguments as in (12), we have with an arbitrary but fixed $\widehat{x} \in \mathscr{X}$ and all $i \in \mathbb{N}$ that

$$\log\left(\frac{h(\bar{x}_i, \xi_i)}{h(\widehat{x}, \xi_i)}\right) = \frac{1}{2\bar{\sigma}^2}\left(\|\bar{x}_i\|_2^2 - \|\widehat{x}\|_2^2 + 2\langle \bar{x}_i - \widehat{x}, \xi_i\rangle\right) \leq \frac{1}{2\bar{\sigma}^2}\left(2M_x^2 + 2M_x\|\xi_i\|_2\right),$$

so $h(\bar{x}_i, \xi_i) \leq h(\widehat{x}, \xi_i) \cdot \exp\left(\frac{M_x^2}{\bar{\sigma}^2}\right) \cdot \exp\left(\frac{M_x\|\xi_i\|_2}{\bar{\sigma}^2}\right)$. Combining this with (14) and (15) we have

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

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 $\gamma_i(\xi_i) = \frac{k_G \exp(b_G\|\xi_i\|_2) h(\widehat{x}, \xi_i)}{\phi_i(\xi_i)}$, it remains to show that (6) and (7) hold.

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

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

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

$$\widehat{g}_{\gamma,N}(\widehat{x}) := \frac{1}{N}\sum_{i=1}^{N}\frac{G_\gamma(\widehat{x}, \xi_i)}{\phi_i(\xi_i)} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i(\xi_i)$$

From the definition of G_γ and Lemma 4, we have for any $\xi \in \Xi$ that

$$\frac{\left|G_\gamma(\widehat{x}, \xi)\right|}{\phi_i(\xi)} = \frac{h(\widehat{x}, \xi)}{\phi_i(\xi)} \cdot k_G \exp(b_G\|\xi\|_2) \leq k_h k_G \exp((b_h + b_G)\|\xi\|_2).$$

Therefore, Assumption 6 holds for $G = G_\gamma$, and using Theorem 6 to obtain

$$0 = \lim_{N\to\infty}\left(\widehat{g}_{\gamma,N}(\widehat{x}) - g_\gamma(\widehat{x})\right) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left(\gamma_i(\xi_i) - \mathrm{E}[\gamma_i(\xi_i)]\right),$$

which is (7). □

6 EXAMPLE: REGRESSION MODELS FOR STEP COMPUTATION IN OPTIMIZATION

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. (2018) in which a local model of the objective is constructed via a SAA regression problem in every iteration.

The algorithm in Maggiar et al. (2018) addresses the minimization of the function $\overline{L} : \mathscr{X} \to \mathbb{R}$ given by

$$\overline{L}(z) = \int_\Xi L(\xi) h(z, \xi)\,\mathrm{d}\xi,$$

where $\mathscr{X} \subset \mathbb{R}^d$ 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. In the context of Maggiar et al. (2018), $L(\xi)$ is the output of a deterministic computer simulation with input ξ and the "original" objective function one would like to minimize. However, since L is subject to numerical noise (small deterministic jumps in function value) and therefore discontinuous, the task of minimizing L ill-defined. To overcome this difficulty, Maggiar et al. (2018) proposes to minimize the convolution $\overline{L}(z)$ as a smooth approximation

The derivative-free trust-region optimization algorithm proposed in Maggiar et al. (2018) utilizes an SAA of $\overline{L}(z)$, i.e.,

$$\overline{L}_N(z) = \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(z,\sigma,\xi_i)}{\varphi(t_i,\sigma,\xi_i)} L(\xi_i).$$

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 $z_N \in \mathscr{Y}$, 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

$$\min_{y\in\mathscr{Y}} \int_{\Xi} F(y,z,\xi)h(z,\xi)\,\mathrm{d}\xi, \tag{16}$$

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 (16), 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 (16); that is

$$\min_{y\in\mathscr{Y}} \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(Z_N,\sigma,\xi_i)}{\varphi(T_i,\sigma,\xi_i)} \left(b + \langle g, \xi_i - Z_N\rangle_2 + \frac{1}{2}\langle \xi_i - Z_N, Q(\xi_i - Z_N)\rangle_2 - L(\xi_i)\right)^2. \tag{17}$$

The analysis of the algorithm in Maggiar et al. (2018) requires that the model $q_N(\xi;Z_N)$ converges to the optimal solution of (16) 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 $\mathscr{Z}$; thus Assumption 4 holds. Also, because all iterates and trial points are contained in $\mathscr{Z}$, the sequence $\{T_i\}$, consisting of such points, is uniformly bounded. Therefore Assumption 8 holds. Furthermore, since $F(y,z,\xi)$ is a polynomial in (y,z) and L exhibits subexponential growth, Assumption 9 holds. Finally, the algorithm in Maggiar et al. (2018) ensures that the optimal solutions of (16) and (17) are unique and uniformly bounded, by monitoring the condition number of matrices involved in the computation of the optimal solution of (17). In summary, Assumptions 4, 8, and 9 hold, and Proposition 1 together with Theorem 2 yields $\lim_{N\to\infty}\mathbb{D}(\widehat{S}_N^Z, S_*^Z) = 0$. So, the approximate model parameters in $\widehat{S}_N^Z$ in iteration N converge to the optimal parameters in S_*^Z.

7 CONCLUSION

We considered the SAA 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.

The differentiability assumption on the function F in Section 5 for the normal case is somewhat restrictive and presented here merely as a simple example that uses the general results in this paper. We conjecture that Assumption 9 can be relaxed considerably. In addition, in future research we plan to extend results in Section 5 to more general class of distributions such as (natural) exponential family.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation of the United States under Grant Number CMMI-1634982. Andreas Waechter was supported in part by the National Science Foundation of the United States under Grant Number DMS-1522747. We thank Tito Homem-de-Mello, David Morton, Imry Rosenbaum, and Johannes Royset for discussions.

REFERENCES

Andrews, D. W. 1992. "Generic Uniform Convergence". *Econometric Theory* 8(02):241–257.

Billingsley, P. 1995. *Probability and Measure*. 3rd ed. John Wiley & Sons.

Chow, Y. S. 1967. "On a Strong Law of Large Numbers for Martingales". *The Annals of Mathematical Statistics* 38(2):610–610.

Chung, K. L. 1951. "The Strong Law of Large Numbers". In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, 341–352: University of California Press, Berkeley and Los Angeles.

Cornuet, J.-M., J.-M. Marin, A. Mira, and C. P. Robert. 2012. "Adaptive Multiple Importance Sampling". *Scandinavian Journal of Statistics* 39:798–812.

Dai, L., C. H. Chen, and J. R. Birge. 2000. "Convergence Properties of Two-Stage Stochastic Programming". *Journal of Optimization Theory and Applications* 106(3):489–509.

Dantzig, G. B., and P. W. Glynn. 1990. "Parallel Processors for Planning Under Uncertainty". *Annals of Operations Research* 22(1):1–21.

Duffie, D., and K. J. Singleton. 1993. "Simulated Moments Estimation of Markov Models of Asset Prices". *Econometrica* 61(4):929–952.

Dupačová, J., and R. Wets. 1988. "Asymptotic Behavior of Statistical Estimators and of Optimal Solutions of Stochastic Optimization Problems". *The Annals of Statistics* 1517–1549.

Feng, M., and J. Staum. 2015. "Green Simulation Designs for Repeated Experiments". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz. et al, 403–413. Piscataway, New Jersey: IEEE.

Feng, M., and J. Staum. 2017. "Green Simulation: Reusing the Output of Repeated Experiments". *ACM Transactions on Modeling and Computer Simulation* 27(4):23:1–23:28.

Glynn, Peter W.and Infanger, G. 2013. "Simulation-Based Confidence Bounds for Two-Stage Stochastic Programs". *Mathematical Programming* 138(1):15–42.

Homem-de-Mello, T. 2008. "On Rates of Convergence for Stochastic Optimization Problems under Non-Independent and Identically Distributed Sampling". *SIAM Journal on Optimization* 19(2):524–551.

Infanger, G. 1992. "Monte Carlo (Importance) Sampling within a Benders Decomposition Algorithm for Stochastic Linear Programs". *Annals of Operations Research* 39(1):69–95.

Jenish, N., and I. R. Prucha. 2009. "Central Limit Theorems and Uniform Laws of Large Numbers for Arrays of Random Fields". *Journal of Econometrics* 150(1):86–98.

Korf, L., and R. J.-B. Wets. 2001. "Random LSC Functions: An Ergodic Theorem". *Mathematics of Operations Research* 26(2):421–445.

Maggiar, A., A. Wächter, I. S. Dolinskaya, and J. Staum. 2018. "A Derivative-Free Trust-Region Algorithm for the Optimization of Functions Smoothed via Gaussian Convolution Using Adaptive Multiple Importance Sampling". *SIAM Journal on Optimization* 28(2):1478–1507.

Marin, J.-M., P. Pudlo, and M. Sedki. 2014, May. "Consistency of the Adaptive Multiple Importance Sampling". arXiv:1211.2548v2.

Owen, A., and Y. Zhou. 2000. "Safe and Effective Importance Sampling". *Journal of the American Statistical Association* 95(449):135–143.

Royset, J. O., and E. Polak. 2004. "Implementable Algorithm for Stochastic Optimization Using Sample Average Approximations". *Journal of Optimization Theory and Applications* 122(1):157–184.

Rubinstein, R. Y., and A. Shapiro. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons.

Rubinstein, R. Y., and D. P. Kroese. 2017. *Simulation and the Monte Carlo Method*. 3rd ed. John Wiley & Sons.

Shapiro, A., and H. Xu. 2007. "Uniform Laws of Large Numbers for Set-Valued Mappings and Subdifferentials of Random Functions". *Journal of Mathematical Analysis and Applications* 325(2):1390–1399.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. 2009. *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia: SIAM.

Van der Vaart, A. W. 2000. *Asymptotic statistics*, Volume 3. Cambridge University Press.

Veach, E., and L. J. Guibas. 1995. "Optimally Combining Sampling Techniques for Monte Carlo Rendering". In *SIGGRAPH 1995 Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, edited by S. G. Mair and R. Cook, 419–428: ACM.

Xu, H. 2010. "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.

AUTHOR BIOGRAPHIES

M. BEN FENG is an assistant professor in actuarial science at the University of Waterloo. He earned his Ph.D. in the Department of Industrial Engineering and Management Sciences at Northwestern University. He is an Associate of the Society of Actuaries (ASA). His research interests include stochastic simulation design and analysis, optimization via simulation, nonlinear optimization, and financial and actuarial applications of simulation and optimization methodologies. His e-mail address is ben.feng@uwaterloo.ca. His website is http://www.math.uwaterloo.ca/ mbfeng/.

ALVARO MAGGIAR is a senior research scientist at Amazon.com in the Supply Chain Optimization Technologies. He holds a Ph.D in Industrial Engineering and Management Sciences from Northwestern University and a Masters in Mathematics and Finance from Imperial College London. His research interests include optimization and inventory management. His e-mail address is alvaro.maggiar@u.northwestern.edu.

JEREMY STAUM was an Associate Professor of Industrial Engineering and Management Sciences at Northwestern University. He coordinated the Risk Analysis track of the 2007 and 2011 Winter Simulation Conferences. His research interests include simulation metamodeling and optimization via simulation. His e-mail address is j-staum@northwestern.edu.

ANDREAS WÄCHTER is an Associate Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. His research centers on the design, analysis, implementation, and application of numerical algorithms for nonlinear optimization. He is a recipient of the J. H. Wilkinson Prize for Numerical Software for the Ipopt open-source optimization package, and of the INFORMS Computing Society Award. His e-mail address is andreas.waechter@northwestern.edu.