Large Deviation Bounds for Markov Chain Sample Average Approximation via Weak Convergence

Arnab Sur*

Abstract

A common approach to solve stochastic optimization problems with expectations is to replace the expectations by its sample averages. Large sample asymptotic properties of this approximation are well studied when the sample is i.i.d. In many cases, however, i.i.d. samples are not readily available. On the contrary, one can generate a Harris recurrent Markov chain with stationary distribution using Markov chain Monte Carlo (MCMC). We call it Markov chain sample average approximation (MCSAA) when the true average is replaced by the sample averages associated to a general state space Markov chain with stationary distribution. In this article, we study the large sample properties of MCSAA estimators associated to a random convex function and also construct probabilistic (exponential) error bounds using large deviation principle via weak convergence.

Key Words: Epigraphical and uniform convergence, Markov chain Sample average approximation, Consistency, Large deviation principle, Donsker-Varadhan Entropy.

*The University of Chicago. IL – 60637, USA. Email: arnabsur2002@gmail.com.

1 Introduction

This paper considers optimization problems involving expectation functions of the form $E[f(x,\xi)]$ where $f : \mathbb{R}^n \times \Xi \to \mathbb{R}$. $\xi : (\Omega, \mathcal{B}, P) \to \Xi \subset \mathbb{R}^m$ is a random vector with induced probability distribution π on Ξ and $\xi(\omega) = \boldsymbol{\xi} \in \Xi$ denotes the realizations of ξ. The following unconstrained stochastic optimization problem we consider involves minimization of :

$$E[f(x,\xi)], \;\; x \in \mathbb{R}^n \tag{1.1}$$

However, in the literature instead of solving the above problem a proxy problem was considered where expectation was replaced by the sample average of an independent and identically distributed sample $\xi_i, \; i = 1, \ldots, N$ drawn from π. At each stage $N \in \mathbb{N}$, stochastic optimizers studied the following sample average objective:

$$\frac{1}{N} \sum_{i=1}^{N} f(x, \xi_i). \tag{1.2}$$

The *sample average approximation* (SAA) objective (1.2) is studied in place of (1.1) when the integral in (1.1) is difficult to evaluate but samples from Ξ under π are readily available.

Consistency of statistical estimators is the key concept in stochastic optimization since its emergence. In stochastic programming models the expectation functionals of random functions are approximated through sample averages when the underlying probability distribution is realized partially. In such a situation epigraphical and uniform convergence of a sequence of sample averages associated to a random function are the primary ingredients in proving the consistency of the statistical estimators in stochastic optimization, especially for stochastic programming models involving expectation functional.

Literature on SAA approach in stochastic optimization had mainly developed for independent and identically distributed sample drawn from the probability distribution π. The SAA approach was introduced in the literature under different names: in Robinson [26] and Gürkan, Özge and Robinson [16], the authors named it sample path optimization, whereas in Kleywegt, Shapiro and Homem-de-Mello [19] it is named as sample average approximation method. The name SAA had been coined by the authors in [19]. SAA analysis typically assumes that the sequence of approximated problems could be solved exactly and the corresponding optimal values will approximate the optimal value of the original stochastic optimization problem. This kind of analysis had appeared in the literature even prior to the above mentioned articles, see, for instances, Wets [37], Birge and Wets [7] and Wets and Dupačová [13]. The authors had considered a sequence of probability measures which weakly converges to the probability distribution associated to the random function instead of empirical distributions. Under this condition pointwise convergence of a sequence of expectation functionals associated to a random function was achieved under the assumptions of continuity of the random function in ξ and tightness property (see Theorem 2.8 in [7]). Then they derived epigraphical convergence for random convex and locally Lipschitz function in x using pointwise convergence. On the other hand, pointwise convergence of a sequence of sample averages associated to an i.i.d. sample is obtained through strong law of large numbers. We refer "Lectures on stochastic programming: modeling and theory" [32] for further reading on SAA analysis related to i.i.d. sample.

Generating independent and identically distributed samples ξ_i, however, is often impossible. For example, in many service applications, such as queueing systems, ξ under π corresponds to a long-run distribution that is only available from discrete-event simulation. In other examples, such as complex systems including

atmospheric conditions, ξ can only be generated as a sequence given previous climatic conditions. In such cases, i.i.d. samples are not available. Moreover, in Bayesian statistics it is difficult to generate i.i.d. sample as probability distribution is known up to a certain constant which is hard to evaluate, for instance see [1], [2], [25] and the references therein. However, in such cases generation of recurrent Markov chain with stationary distribution as π is possible using Markov chain Monte Carlo (MCMC, in short) due to the advancement of computational tools. Apart from the instances mentioned above, in practice, a stochastic optimization problem could be formulated in such a way where the random variable is unobservable (hidden) but follows Markov property with stationary distribution. MCMC could then be used to simulate the sample and the sample average could be used to replace the true average. The following two examples demonstrate such situations where the samples are drawn from a Markov chain with stationary distribution, instead of drawing i.i.d. sample from the underlying distribution.

Example 1.1. Let us consider the news vendor problem where a company has to decide about the order quantity x of a certain product to satisfy the demand ξ, which is a random variable. If the demand ξ is higher than the order amount x, then there is a backorder penalty cost which is greater than the ordering cost and if the order amount x is higher than the demand ξ, then there is a holding cost of the excess amount. The objective is to minimize the total expected cost $E[f(x,\xi)]$. The distribution of ξ is estimated from the historical data of the ordering process. We assume that the distribution ξ is known upto a certain constant (for e.g., normalized exponential distribution where the normalization factor is unknown or difficult to evaluate). In such a case, i.i.d. samples are not readily available; however, MCMC algorithms can be used to generate a Markov chain with distribution of ξ as the stationary distribution.

Example 1.2. Let us consider stochastic economic dispatch problem under multiple random failure of transmission lines (cascading failure). The objective is to minimize the expected operational cost $E[f(x,\xi)]$ to serve the load under potential cascading failure of the lines, where x is the decision vector and ξ denotes the network topology of the grid. ξ is a random variable as the cascading failure of the lines is random. We assume that the exponential failure rate $\lambda > 0$ and the repair rate $\mu > 0$ of the lines are known, however, the distribution of ξ is unknown. With the help of λ and μ, we can construct a Markov chain (jump process) whose stationary distribution is the underlying distribution of ξ. We can then replace the expected operational cost $E[f(x,\xi)]$ by its sample average (cost associated to the Markov chain) if the estimators are consistent.

This paper considers the formulation of (1.2) in which ξ_i's are generated from a general state space Markov chain with stationary distribution π as in the service and complex system examples mentioned above or in Bayesian statistics. Our study assumes general state space Markov chain with stationary distribution instead of i.i.d. sample and we term it as *Markov chain sample average approximation* (MCSAA, in short).

In the stochastic optimization literature, non-i.i.d. cases have been considered in few papers, for example, in [20], the authors derived Birkhoff type ergodic theorem for random lower semi-continuous functions to obtain consistency of optimal solutions which encompasses the i.i.d. case. However, MCSAA can not be considered as a special case of that study as ergodicity implies convergence from "π-almost all" starting points in case of a general state space Markov chain. Starting point and the transition kernel define the trajectory of a Markov chain, hence, we have to include all starting points in the state space when we study consistency of the MCSAA estimators; otherwise, the convergence can not be guaranteed if the general state space Markov chain starts within a measure zero set (with respect to the stationary distribution π). This is a major difference with the result regarding consistency of SAA estimators in [20] as it does not imply

convergence from "all" starting points, i.e., the convergence can not be established for all sample paths drawn from a general state space Markov chain as starting point describes the trajectory of the sample path. Finite state-space Markov chain is, however, included in that study; for more details, we refer [34]. Moreover, in [17], the author established the convergence rates of the SAA estimators under non-independent and identically distributed sample generated by applying Latin hypercube sampling (LHS) and randomized quasi-Monte Carlo (QMC) sampling techniques.

In Section 2, we discuss epigraphical and uniform convergence of a sequence of sample averages of a random convex function associated to a general state space Markov chain with stationary distribution π. The strong law of large numbers for general state space Markov chain will be used to derive the convergence results. In fact, we need to impose more structure on the stationary Markov chain (namely, Harris recurrence) to employ strong law and include all starting point. All the results in this section are true for any starting point in the state space. Starting point is of immense importance while generating a Markov chain using MCMC to approximate expectation functional of a random function by its sample averages. In Section 3, epigraphical and uniform convergence of a sequence of the sample averages is applied to derive consistency of optimal and stationary points of stochastic constrained and unconstrained optimization problems involving expectation functionals. An approximation result on convex subdifferential (see, Corollary 2.1) will be used to establish the consistency of stationary points. The consistency of MCSAA estimators of stationary points of constrained optimization problem is a key contribution of this section.

In the last section, we construct a probabilistic bound (exponential) for the MCSAA estimators of the optimal value using large deviation theory. This section will be an important addition to the literature on SAA, in general, stochastic optimization. In the context of SAA, the large deviation theory deals with the situation where the probability that the sample averages deviate form the expected value by a fixed amount $\epsilon > 0$ approaches to zero exponentially fast as N tends to ∞. Formally speaking, it estimates the bound for the quantity

$$\frac{1}{N} \log P\{ \mid \frac{1}{N} \sum_{k=1}^{N} f(x, \xi_k) - \int f(x, \xi) d\pi \mid < \epsilon\},$$

as N tends to ∞.

Large deviation has found applications in numerous areas including information theory, risk management, statistical mechanics, thermodynamics and many more. However, it has been underused in the literature of stochastic optimization. Optimizers mostly used Cramér's theorem to derive large deviation bounds for i.i.d. sample, for example, see [18], [30], [31]. Large deviation upper bound for the SAA estimators of the optimal value had been established in [18], however, the paper does not say anything about the lower bound. The paper [31] applied Cramér's theorem to show, under further assumptions that a sequence of optimal solution of the SAA subproblems approaches to an exact optimal solution of the true problem exponentially fast for sufficiently large sample size. Cramér's theorem is applicable only for the i.i.d sequences to determine the rate function. In [9], the authors relaxed the i.i.d. assumption to estimate the large deviation upper bound for the SAA estimators of the optimal value via Gärtner-Ellis theorem. The paper does not focus on any kind of sampling technique; rather, they assume certain conditions that allow for the application of Gärtner-Ellis theorem in the large deviation. The approach employed in this paper can be used for a non-i.i.d. sequence, however, Theorem 3.7 in [9] is not applicable for the Markov chain, as later we will see that the convergence is established with respect to the probability measure $Q_{\boldsymbol{\xi}}$, induced by the probability transition function $Q(\cdot, \cdot)$ and the starting point of the Markov chain $\boldsymbol{\xi}$.Moreover, the authors in [12] applied Gärtner-Ellis theorem to

develop large deviation bounds for the estimators of the optimal value when the sample is generated using Latin hypercube sampling technique. Both Cramér's theorem and Gärtner-Ellis theorem heavily depend on the convergence of the log moment generating function to obtain the large deviation principle (LDP) with some rate. They are similar in flavour, in fact, Cramér's theorem is a special case of the Gärtner-Ellis theorem, however, the latter includes the non-i.i.d cases as well.

In this paper, we follow the weak convergence approach to the large deviation theory to construct bounds for the MCSAA estimators. This approach is well suited for the general spaces, say Polish space or even more general spaces like topological vector space. The weak convergence approach exploits the space of probability measures defined on Ξ endowed with the weak topology, rather than relying on the convergence of the log moment generating function. If Ξ is separable and complete, weak topology can be defined by Prohorov metric. This metric plays a pivotal role in establishing the large deviation bounds of the estimators. We develop a large deviation bound for the MCSAA estimators of the optimal value using Donsker-Varadhan entropy [24] when the sample is drawn form a general state space Markov chain. The bounds are valid for countable state space Markov chain as well. Moreover, we construct independent (with respect to the starting point) large deviation bounds for the MCSAA estimators using Ellis' theorem [14]. In the process of obtaining the large deviation bounds for the MCSAA estimators, we develop pointwise and uniform large deviation bounds of the functional values of the objective by applying Donsker-Varadhan entropy and Ellis' theorem. The pointwise and the uniform errors converge to zero exponentially fast in the sample size N and the exponential rate of convergence for the pointwise and uniform MCSAA estimators is carried over to the MCSAA estimators of the optimal value.

2 Epigraphical and Uniform Convergence

This section is dedicated to establish the epigraphical and uniform convergence of a sequence of sample averages associated to a random convex function. The sample averages are generated by a Harris recurrent (general state space) Markov chain with a stationary distribution π. We start this section by recalling the following definitions.

Definition 2.1. $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is a measurable space. A probability transition kernel is a function $\mathcal{K} : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to [0, 1]$ with the following properties:
i) $x \mapsto \mathcal{K}(x, B)$ is $\mathcal{B}(\mathcal{X})$-measurable for each $B \in \mathcal{B}(\mathcal{X})$,
ii) $B \mapsto \mathcal{K}(x, B)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ for each $x \in \mathcal{X}$.

A Markov chain can be constructed if the initial distribution (starting point) and probability transition kernel are known. The probability transition kernel will serve as a one step transition laws for such a chain. Measurability of the kernel in the first variable plays a pivotal role to define the distribution of the Markov chain. We assume a Markov chain with state space $(\Xi, \mathcal{B}(\Xi))$ and transition kernel $Q(\cdot, \cdot)$. For each starting point $\boldsymbol{\xi} \in \Xi$, the transition kernel induces a probability measure $Q_{\boldsymbol{\xi}}$ on Ω. Therefore, the induced probability measures $Q_{\boldsymbol{\xi}}$ will be different for different starting points $\boldsymbol{\xi}$. A general state space Markov chain is Harris recurrent with respect to a reference measure ϕ if $A \in \mathcal{B}(\Xi)$, $\phi(A) > 0$ implies $Q_{\boldsymbol{\xi}}(T_A < \infty) = 1$ for all $\boldsymbol{\xi} \in \Xi$, where T_A is the first entrance time or hitting time to A. Details can be found in Meyn and Tweedie [21].

Next we state the *strong law of large numbers* for Markov chain which is the main foundation of every result in this section. The strong law of large numbers for Harris recurrent Markov chain is the following :

Theorem 2.1 ([3])**.** Let $\{\xi_k\}_{k\geq 0}$ be a Harris recurrent Markov chain with state space $(\Xi, \mathcal{B}(\Xi))$ and transition function $Q(\cdot,\cdot)$. Suppose there exists a stationary distribution π. Then

i) π is unique and

ii) for all $f \in L^1(\Xi, \mathcal{B}(\Xi), \pi)$ and for all $\boldsymbol{\xi} \in \Xi$, we have

$$\frac{1}{N}\sum_{k=0}^{N-1} f(\xi_k) \to \int f d\pi\,,$$

$Q_{\boldsymbol{\xi}}$-a.s.

Throughout this article the following notions will be used for the sake of consistency. Let $\xi : (\Omega, \mathcal{B}, P) \to \mathbb{R}^m$ be a random variable with support $\Xi \subseteq \mathbb{R}^m$ and $E[\cdot]$ denotes the expected value with respect to the probability distribution π of ξ. $\{\xi_k : k = 1, 2\ldots\}$ is a Harris recurrent Markov chain with state space $(\Xi, \mathcal{B}(\Xi))$ and transition kernel $Q(\cdot,\cdot)$. Moreover, π is the stationary distribution of the Markov chain $\{\xi_k : k = 1, 2\ldots\}$. Birkhoff type ergodicity of a general state space Markov chain implies convergence from almost all starting points, however, to derive convergence from all starting points Harris recurrence is necessary.

The next theorem is on epigraphical convergence of the sample averages of a random convex function to its true average. Strong law of large numbers will play a pivotal role to establish epigraphical convergence.

Theorem 2.2. Suppose that $f : \mathbb{R}^n \times \Xi \to \overline{\mathbb{R}}$ be a convex function in x, π-a.s.; and for all x, $f(x,\cdot)$ is dominated by an integrable function, π-a.s. Then $E[f(\cdot,\xi)]$ is finite valued and convex, and $\frac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k)$ converges epigraphically to $E[f(\cdot,\xi)]$, $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi}$. Moreover, $\frac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k)$ converges uniformly to $E[f(\cdot,\xi)]$, $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi}$, on compact subsets of $\mathbb{R}^n$.

Proof. Let us assume that $f(x,\xi) \leq g(\xi)$, π-a.s. Hence, it follows from the assumption that $\mid E[f(\cdot,\xi)] \mid \leq E[g(\xi)]$ and consequently $\mid E[f(\cdot,\xi)] \mid < +\infty$. The convexity of $E[f(\cdot,\xi)]$ and $\frac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k)$ follows from the convexity of $f(\cdot,\xi)$.
Let us choose a countable dense subset D in $\mathbb{R}^n$. The strong law of large numbers for Markov chains (ref. Theorem 2.1) asserts that $\frac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k)$ converges pointwise to $E[f(\cdot,\xi)]$, $(Q_{\boldsymbol{\xi}})$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Hence, for $x \in D$, $\frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k)$ converges pointwise to $E[f(x,\xi)]$ for all $\omega \in \Omega \setminus B_x$, where $(Q_{\boldsymbol{\xi}})(B_x) = 0$ for any $\boldsymbol{\xi} \in \Xi$. We have, $(Q_{\boldsymbol{\xi}})(B = \bigcup_{x\in D} B_x) = 0$ as D is countable. Thus, $\frac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k)$ converges pointwise to $E[f(\cdot,\xi)]$ on D for all $\omega \in \Omega \setminus B$, where $(Q_{\boldsymbol{\xi}})(B) = 0$ for any $\boldsymbol{\xi} \in \Xi$. Consequently by using Theorem 7.17 in [27], we conclude that $\frac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k)$ epigraphically converges to $E[f(\cdot,\xi)]$ for all $\omega \in \Omega \setminus B$, where $(Q_{\boldsymbol{\xi}})(B) = 0$ for any $\boldsymbol{\xi} \in \Xi$. Again using Theorem 7.17 in [27], we have uniform convergence on compact subsets as well. This completes the proof. ∎

We will now deduce an important corollary of the previous theorem. Applying Corollary 2.4 in [6] with

Theorem 2.2 we obtain the following result which will play a crucial role to obtain consistency of the stationary points in the next section.

Corollary 2.1. Suppose that all assumption in Theorem 2.2 be true. Then for each x,

$$\partial E[f(x,\xi)] = \{\lim_{N\to\infty} u_N : u_N \in \partial E_N[f(x_N,\xi)] \text{ and } x_N \to x, \text{ as } N \to \infty\}$$

($Q_{\boldsymbol{\xi}}$)-a.s. for any $\boldsymbol{\xi} \in \Xi$, where $E_N[f(.,\xi)] := \dfrac{1}{N}\sum_{k=1}^{N} f(.,\xi_k)$.

Proof. The result follows from Corollary 2.4 in [6] after combining it with Theorem 2.2. ∎

Next theorem drops the convexity assumption still establishes uniform convergence under continuity of $f(\cdot,\xi)$.

Theorem 2.3. Let us assume C be a non-empty compact subset of $\mathbb{R}^n$. Suppose that the function $f(\cdot,\xi)$ is continuous and is dominated by an integrable function on C, π-a.s. Then the expected value function $E[f(x,\xi)]$ is finite valued and continuous on C, and $\dfrac{1}{N}\sum_{k=1}^{N} f(x,\xi_k)$ converges to $E[f(x,\xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi}$ uniformly on C.

Proof. $E[f(x,\xi)]$ is finite valued as it is dominated by an integrable function $G(\xi)$ π-a.s. and the continuity follows from the continuity of $f(\cdot,\xi)$.

To establish uniform continuity we need to show the following.

$$\sup_{x\in C} \mid E[f(x,\xi)] - \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) \mid < \epsilon \; Q_{\boldsymbol{\xi}}\text{-a.s. for any } \boldsymbol{\xi}, \text{ for sufficiently large } N.$$

Let us denote $F(x) := E[f(x,\xi)]$ and $F_N(x) := \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k)$. Hence, we have to prove that

$$\sup_{x\in C} \mid F(x) - F_N(x) \mid < \epsilon \; Q_{\boldsymbol{\xi}}\text{-a.s. for any } \boldsymbol{\xi}, \text{ for sufficiently large } N.$$

Let us choose a point $\bar{x} \in C$, and using the triangle inequality we have

$$\begin{aligned}
&\sup_{x\in C} \mid F(x) - F_N(x) \mid = \sup_{x\in C} \mid F(x) - F(\bar{x}) + F(\bar{x}) - F_N(\bar{x}) + F_N(\bar{x}) - F_N(x) \mid \\
&\leq \sup_{x\in C} \left(\mid F(x) - F(\bar{x}) \mid + \mid F(\bar{x}) - F_N(\bar{x}) \mid + \mid F_N(\bar{x}) - F_N(x) \mid \right) \\
&\leq \sup_{x\in C} \mid F(x) - F(\bar{x}) \mid + \mid F(\bar{x}) - F_N(\bar{x}) \mid + \sup_{x\in C} \mid F_N(\bar{x}) - F_N(x) \mid .
\end{aligned}$$

Choose a sequence α_j such that $\alpha_j > 0$ and $\alpha_j \to 0$, as $j \to \infty$ and define neighbourhoods around $\bar{x}$ as

$$V_j := \{x \in C :\parallel x - \bar{x} \parallel \leq \alpha_j\} \quad \text{and} \quad \delta_j(\xi) := \sup_{x\in V_j} \mid f(x,\xi) - f(\bar{x},\xi) \mid.$$

Continuity implies that for any $\xi \in \Xi$, $\delta_j(\xi) \to 0$ as $j \to \infty$. Moreover, $\delta_j(\xi)$ is dominated by an integrable function as $\mid \delta_j(\xi) \mid \leq 2G(\xi)$. Hence, using Lebesgue Dominated Convergence Theorem we conclude that

$$\lim_{j\to\infty} E[\delta_j(\xi)] = E[\lim_{j\to\infty} \delta_j(\xi)] = 0, \quad \text{as } \delta_j(\xi) \to 0 \text{ when } j \to \infty. \tag{2.1}$$

We also have that

$$\mid F_N(\bar{x}) - F_N(x) \mid = \mid \tfrac{1}{N} \sum_{k=1}^{N} (f(\bar{x}, \xi_k) - f(x, \xi_k)) \mid \leq \frac{1}{N} \sum_{k=1}^{N} \mid (f(\bar{x}, \xi_k) - f(x, \xi_k)) \mid .$$

Hence,

$$\sup_{x \in V_j} \mid F_N(\bar{x}) - F_N(x) \mid \leq \sup_{x \in V_j} \frac{1}{N} \sum_{k=1}^{N} \mid (f(\bar{x}, \xi_k) - f(x, \xi_k)) \mid$$

$$\leq \frac{1}{N} \sum_{k=1}^{N} \sup_{x \in V_j} \mid (f(\bar{x}, \xi_k) - f(x, \xi_k)) \mid = \frac{1}{N} \sum_{k=1}^{N} \delta_j(\xi_k). \tag{2.2}$$

$\frac{1}{N} \sum_{k=1}^{N} \delta_j(\xi_k) \to E[\delta_j(\xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi}$, as $N \to \infty$ follows from the strong law of large numbers of Markov chain

Consequently, for sufficiently large N we obtain from equation (2.2) that

$$\sup_{x \in V_j} \mid F_N(\bar{x}) - F_N(x) \mid \leq E[\delta_j(\xi)] \quad Q_{\boldsymbol{\xi}}\text{-a.s. for any } \boldsymbol{\xi}.$$

And from equation (2.1) we obtain $\lim_{j \to \infty} E[\delta_j(\xi)] = 0$. Hence, for given $\epsilon > 0$, there exists $n_1 \in \mathbb{N}$ such that $\mid E[\delta_j(\xi)] \mid < \frac{\epsilon}{3}$, for all $j \geq n_1$. Which implies that for sufficiently large N,

$$\sup_{x \in V_j} \mid F_N(\bar{x}) - F_N(x) \mid \leq E[\delta_j(\xi)] < \frac{\epsilon}{3} \ Q_{\boldsymbol{\xi}}\text{-a.s. for any } \boldsymbol{\xi}, \text{ for all } j \geq n_1.$$

So, there exists a neighbourhood $W := \bigcup_{j \geq n_1} V_j$ of $\bar{x}$ such that for sufficiently large N,

$$\sup_{x \in W} \mid F_N(\bar{x}) - F_N(x) \mid < \frac{\epsilon}{3} \quad Q_{\boldsymbol{\xi}}\text{-a.s. for any } \boldsymbol{\xi}.$$

Due to continuity we have, for a given $\epsilon > 0$ there exists a neighbourhood B of $\bar{x}$ such that $\mid F(\bar{x}) - F(x) \mid < \frac{\epsilon}{3}$, for all $x \in B \bigcap C$.

Let us assume that $U(\bar{x}) := W \cap B$, which is a neighbourhood of $\bar{x}$. Collection of all such neighbourhoods $U(x)$ will form an open cover of C, i.e. $\bigcup_{x \in C} U(x) = C$. Since C is compact, there exists a finite number of points $x_1, \ldots x_t$ in C such that $\bigcup_{i=1}^{t} U_i = C$, where U_i is a neighbourhood of x_i.

Hence, for sufficiently large N and for $i = 1, \ldots t$ we have

$$\sup_{x \in U_i} \mid F_N(x_i) - F_N(x) \mid + \sup_{x \in U_i} \mid F(x_i) - F(x) \mid < \frac{\epsilon}{3} + \frac{\epsilon}{3} = \frac{2\epsilon}{3} \quad Q_{\boldsymbol{\xi}}\text{-a.s. for any } \boldsymbol{\xi}.$$

Moreover, using strong law of large numbers for Markov chain we obtain that for sufficiently large N and for given $\epsilon > 0$, $\mid F_N(x) - F(x) \mid < \frac{\epsilon}{3}$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi}$ and for all $x \in C$.

Consequently, for $i = 1, \ldots t$ and for sufficiently large N, we have $\mid F_N(x_i) - F(x_i) \mid < \frac{\epsilon}{3}$, $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi}$.

Combining them we deduce that for $i = 1, \ldots t$ and for sufficiently large N,

$\sup_{x \in U_i} \mid F(x) - F_N(x) \mid \leq \sup_{x \in U_i} \mid F(x) - F(x_i) \mid + \mid F(x_i) - F_N(x_i) \mid + \sup_{x \in U_i} \mid F_N(x_i) - F_N(x) \mid < \epsilon$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi}$.

As $\bigcup_{i=1}^{t} U_i = C$, we conclude the proof. ∎

One can observe that minimal information about π is required in comparison to i.i.d set up. The only requirement is a Harris recurrent Markov chain with π as its stationary distribution. The induced probability measure $Q_{\boldsymbol{\xi}}$ on Ω depends on the starting point $\boldsymbol{\xi}$ of the Markov chain. Consequently, the null set on which the convergence does not hold depends on the starting point. The whole texture of the work on convergence depends on the starting point of the Markov chain. Therefore, we have to assume Harris recurrence to obtain convergence from all starting points; whereas, ergodicity of the Markov chain implies convergence from all starting points except for a π-null set A. The convergence can not be guaranteed if we consider an ergodic Markov chain which starts from set A, i.e., the convergence is not guaranteed for all the sample paths. To read more on why the ergodic theorem developed in [20] is not suitable to establish convergence for a general state space Markov chain from all starting points, we refer [34].

3 Asymptotic Properties of the MCSAA Estimators

SAA is a familiar technique to solve optimization problems involving expectation functionals. Instead of evaluating the expectations in a closed form they are replaced by sample averages. The researchers predominantly used sample averages corresponding to an i.i.d. sample. However, Dai et. al. [9] relaxed the i.i.d. assumption to derive probabilistic bounds using large deviation principle and Homem-de-Mello in [17] employed non-i.i.d. samples generated through quasi-Monte Carlo and Latin hypercube sampling schemes. Large sample properties like consistency of the optimal solutions, convergence rates are well studied in the literature. In this section, we present the most relevant extension of the existing literature. We assume a Harris recurrent Markov chain with stationary distribution π as the sample and call it as Markov chain sample average approximation (MCSAA). MCMC algorithms guarantee existence of such a sample. Connections between Harris recurrence and MCMC algorithms were well studied in [8] and [35]. Applying results in the previous section we can deduce the consistency of optimal solutions and stationary points of the stochastic optimization problems.

3.1 Unconstrained Optimization Problem

Let us consider the following unconstrained minimization problem:

$$\min_{x\in\mathbb{R}^n} \quad E[f(x,\xi)]\,, \tag{3.1}$$

where $\xi : (\Omega, \mathcal{B}, P) \to \mathbb{R}^m$, is a random variable with support $\Xi \subseteq \mathbb{R}^m$, and $f : \mathbb{R}^n \times \Xi \to \mathbb{R}$ is a convex function with finite expectation for all $x \in \mathbb{R}^n$. $E[\cdot]$ is evaluated with respect to π.

The corresponding sample average approximation problem, for a fixed $N \in \mathbb{N}$, is the following:

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k)\,, \tag{3.2}$$

where $\xi_1, ...\xi_N$ is a Harris recurrent Markov chain with stationary distribution π and state space $(\Xi, \mathcal{B}(\Xi))$. This assumption is appropriate due to the advancement of the computational mechanism. If π is partially

realized, for instance, in Bayesian statistics it is known up to a certain constant, oftentimes researchers face difficulties to generate i.i.d. sample. Nevertheless, this situation does not restrict us from generating a Harris recurrent Markov chain with π as its stationary distribution using MCMC. Moreover, we can formulate a system which has a underlying irreducible Markov structure with finite states but the states are unobservable. We then can simulate the states using MCMC and minimize the problem with respect to its stationary distribution. These situations describe the importance of MCSAA.
Next theorem is on consistency of the optimal solutions of unconstrained problems.

Theorem 3.1. Let us consider the optimization problem (3.1) and the corresponding sample average approximation problems (3.2), for each $N \in \mathbb{N}$.
i) If $\{\overline{x}_N : N = 1, 2.....\}$ is a sequence of global minimizers of the sample average approximation problems, i.e. $\overline{x}_N$ is a global minimizer of the problem (3.2), for each $N \in \mathbb{N}$ and $\overline{x}$ is an accumulation point of this sequence $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, then $\overline{x}$ is a global minimizer of the problem (3.1) and $\frac{1}{N}\sum_{k=1}^{N} f(\overline{x}_N, \xi_k)$ converges to $E[f(\overline{x}, \xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$.
ii) If $\{\overline{x}_N : N = 1, 2.....\}$ is a sequence of local minimizers of the sample average approximation problems sharing a common radius of attraction $\rho > 0$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$ (i.e. for each $N \in \mathbb{N}$, $\frac{1}{N}\sum_{k=1}^{N} f(\overline{x}_N, \xi_k) \leq \frac{1}{N}\sum_{k=1}^{N} f(x, \xi_k)$ for all $x \in \mathbb{R}^n$ such that $\| \overline{x}_N - x \| < \rho$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$) and $\overline{x}$ is an accumulation point of this sequence $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, then $\overline{x}$ is a local minimizer of the problem (3.1) and $\frac{1}{N}\sum_{k=1}^{N} f(\overline{x}_N, \xi_k)$ converges to $E[f(\overline{x}, \xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$.

Proof. Applying Theorem 2.2 we have $\frac{1}{N}\sum_{k=1}^{N} f(., \xi_k)$ epi-converges to $E[f(., \xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Hence, using the functional definition of epi-convergence we can conclude that for every infinite sequence $\{x_N : N = 1, 2....\}$ such that $x_N \in \mathbb{R}^n$ and $x_N \to x$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, as $N \to \infty$; $\liminf_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(x_N, \xi_k) \geq E[f(x, \xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Moreover, using strong law of large numbers for Markov chain we know that for every $x \in \mathbb{R}^n$, there exists a sequence $\{x_N = x : N = 1, 2....\}$, such that x_N converges to x and $\limsup_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(x_N, \xi_k) \leq E[f(x, \xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Therefore, Theorem 3.3.2 in [23] implies that the epigraphs of the sample average approximation problems (3.2) converge to epigraph of the optimization problem (3.1). Consequently using Theorem 3.3.3 in [23], we accomplish the above result. ∎

3.2 Constrained Optimization Problem

We consider the following stochastic constrained optimization problem.

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & E[f(x, \xi)] \\ \text{subject to} \quad & E[G(x, \xi)] = 0, \\ & E[H(x, \xi)] \leq 0; \end{aligned} \tag{3.3}$$

where $\xi : (\Omega, \mathfrak{B}, P) \to \mathbb{R}^m$, is a random variable with support $\Xi \subseteq \mathbb{R}^m$. $f : \mathbb{R}^n \times \Xi \to \mathbb{R}$ is a convex function with finite expectation for all $x \in \mathbb{R}^n$ and $G : \mathbb{R}^n \times \Xi \to \mathbb{R}^l$, $H : \mathbb{R}^n \times \Xi \to \mathbb{R}^p$ are continuously differentiable with finite expectation evaluated with respect to π.

The corresponding sample average approximation problem for a fixed $N \in \mathbb{N}$, is:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{N}\sum_{k=1}^{N} f(x, \xi_k) \\ \text{subject to} \quad & \frac{1}{N}\sum_{k=1}^{N} G(x, \xi_k) = 0, \\ & \frac{1}{N}\sum_{k=1}^{N} H(x, \xi_k) \leq 0; \end{aligned} \tag{3.4}$$

where $\xi_1, ...\xi_N$ is a Harris recurrent Markov chain with stationary distribution π and state space $(\Xi, \mathcal{B}(\Xi))$.

Theorem 3.2. Let us consider the stochastic constrained optimization problem (3.3) and the corresponding sample average approximation problems (3.4), for each $N \in \mathbb{N}$. Moreover assume that for every $x \in X$, there exists a sequence $\{x_N : N = 1, 2.....\}$ with $x_N \in X_N$ such that x_N converges to x as $N \to \infty$; where X is feasible set of the optimization problem (3.3) and X_N, $N = 1, 2.....$ are feasible sets of the sample average approximation problems (3.4). Then the conclusions $i)$ and $ii)$ of Theorem 3.1 will be true for constrained optimization problem.

Proof. Due to the assumption in Theorem 3.2 and applying Theorem 2.2 (uniform convergence on compact subsets) we have, for every $x \in X$ there exists a sequence $\{x_N : N = 1, 2.....\}$ with $x_N \in X_N$ such that x_N converges to x as $N \to \infty$ and $\limsup_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(x_N, \xi_k) = E[f(x, \xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Moreover, consider an infinite sequence $\{x_N : N = 1, 2....\}$ such that $x_N \in X_N$ and $x_N \to x$. This implies that $\frac{1}{N}\sum_{k=1}^{N} G(x_N, \xi_k) = 0$ and $\frac{1}{N}\sum_{k=1}^{N} H(x_N, \xi_k) \leq 0$, for each $N \in \mathbb{N}$. Applying Theorem 2.3 we obtain $E[G(x, \xi)] = 0$ and $E[H(x, \xi)] \leq 0$; hence, $x \in X$. Further, applying Theorem 2.2 we can conclude that, $\liminf_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(x_N, \xi_k) = E[f(x, \xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Consequently, the result follows from Theorem 3.3.2 and Theorem 3.3.3 in [23]. ∎

Remark 1. The notable difference between constrained and unconstrained optimization problem is that feasible sets of the sample average approximation problems are different for each $N \in \mathbb{N}$ for the constrained case, whereas they are same for each $N \in \mathbb{N}$ in case of unconstrained one. So we need the extra assumption in Theorem 3.2.

Definition 3.1. A feasible solution x^* of the optimization problem (3.3) is said to satisfy Fritz John necessary optimality conditions if there exist scalars $\bar{\lambda}_0 \geq 0$, $\bar{\lambda}_i^G \in \mathbb{R}$, $i = 1, ...l$ and $\bar{\lambda}_i^H \geq 0$, $i = 1, ...p$ such that

$i)$ $\quad 0 \in \bar{\lambda}_0 E[\partial f(x^*, \xi)] + \sum_{i=1}^{l} \bar{\lambda}_i^G E[\nabla_x G_i(x^*, \xi)] + \sum_{i=1}^{p} \bar{\lambda}_i^H E[\nabla_x H_i(x^*, \xi)].$

$ii)$ $\quad (\bar{\lambda}_0, \bar{\lambda}^G, \bar{\lambda}^H) \neq 0,$

$iii)$ $\quad \bar{\lambda}_i^H E[H_i(x^*, \xi)] = 0,$ for $i = 1, ...p.$

Then x^* is called Fritz John point. If $\bar{\lambda}_0 > 0$, then x^* is known as a stationary point and the necessary optimality condition is known as Karush-Kuhn-Tucker (KKT, for short) condition.

Theorem 3.3. Let us consider the constrained optimization problem (3.3) and the corresponding sample average approximation problems of the type (3.4), for each $N \in \mathbb{N}$. Suppose that x^N be a Fritz John point of the sample average approximation problem (3.4) and x^* is a cluster point of the sequence $\{x^N : N = 1, 2.....\}$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Further assume that there exist positive valued random variables C_i^G and C_j^H such that $E[C_i^G]$ and $E[C_j^H]$ for all $i = 1, \dots l$ and $j = 1, \dots p$ are finite and let there exists a neighbourhood U of x^* such that for every $x_1, x_2 \in U$, we have $|G_i(x_1, \xi) - G_i(x_2, \xi)| \leq C_i^G ||x_1 - x_2||$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$ and $|H_j(x_1, \xi) - H_j(x_2, \xi)| \leq C_j^H ||x_1 - x_2||$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Then x^* is a stationary point of the constrained optimization problem (3.3) provided $E[\nabla_x G_i(x^*, \xi)]$, $i = 1, \dots l$ and $E[\nabla_x H_j(x^*, \xi)]$, $j = 1, \dots p$ are linearly independent.

Proof. x^N is a Fritz John point of the sample average approximation problem (3.4). Then by the definition of Fritz John necessary optimality conditions there exist scalars $(\bar{\lambda}_0)^N \geq 0$, $(\bar{\lambda}_i^G)^N \in \mathbb{R}$, $i = 1, \dots l$ and $(\bar{\lambda}_i^H)^N \in \mathbb{R}$, $i = 1, \dots p$ such that

$$i) \quad 0 \in \frac{1}{N} \sum_{k=1}^{N} (\bar{\lambda}_0)^N \partial f(x^N, \xi^k)] + \sum_{i=1}^{l} [\frac{1}{N} \sum_{k=1}^{N} (\bar{\lambda}_i^G)^N \nabla_x G_i(x^N, \xi_k)] + \sum_{i=1}^{p} [\frac{1}{N} \sum_{k=1}^{N} (\bar{\lambda}_i^H)^N \nabla_x H_i(x^N, \xi_k)].$$

$$ii) \quad ((\bar{\lambda}_0)^N, (\bar{\lambda}^G)^N, (\bar{\lambda}^H)^N) \neq 0,$$

$$iii) \quad \frac{1}{N} \sum_{k=1}^{N} (\bar{\lambda}_i^H)^N H_i(x^N, \xi_k) = 0, \text{ for } i = 1, \dots p.$$

Let

$$\lambda^N = \frac{((\bar{\lambda}_0)^N, (\bar{\lambda}_1^G)^N, ..., (\bar{\lambda}_l^G)^N, (\bar{\lambda}_1^H)^N, ...(\bar{\lambda}_l^H)^N)}{\sqrt{\{(\bar{\lambda}_0)^N\}^2 + \{(\bar{\lambda}_1^G)^N\}^2 + ...\{(\bar{\lambda}_l^G)^N\}^2 + \{(\bar{\lambda}_1^H)^N\}^2 + ...\{(\bar{\lambda}_l^H)^N\}^2}}.$$

So $\| \lambda^N \| = 1$ for all $N \in \mathbb{N}$, boundedness of the sequence implies that there exists a convergent subsequence $\{\lambda^N\}$ such that $\lambda^N \to \lambda^*$ and by the continuity of the norm $\| \lambda^* \| = 1$. Moreover, given that $x^N \to x^*$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Let us denote

$$\lambda^N = (\lambda_0^N, (\lambda_1^G)^N, ...(\lambda_l^G)^N, (\lambda_1^H)^N, ...(\lambda_l^H)^N).$$

$$\lambda^* = (\lambda_0^*, (\lambda_1^G)^*, ...(\lambda_l^G)^*, (\lambda_1^H)^*, ...(\lambda_l^H)^*).$$

Dividing both sides of $i)$ by $\sqrt{\{(\bar{\lambda}_0)^N\}^2 + \{(\bar{\lambda}_1^G)^N\}^2 + ...\{(\bar{\lambda}_i^G)^N\}^2 + \{(\bar{\lambda}_1^H)^N\}^2 + ...\{(\bar{\lambda}_l^H)^N\}^2}$, we get

$$0 \in \frac{1}{N} \sum_{k=1}^{N} (\lambda_0^N) \partial f(x^N, \xi_k) + \sum_{i=1}^{l} [\frac{1}{N} \sum_{k=1}^{N} (\lambda_i^G)^N \nabla_x G_i(x^N, \xi_k)] + \sum_{i=1}^{p} [\frac{1}{N} \sum_{k=1}^{N} (\lambda_i^H)^N \nabla_x H_i(x^N, \xi_k)]. \tag{3.5}$$

Further from $ii)$ and $iii)$ we obtain

$$iv) \quad ((\lambda_0)^N, (\lambda^G)^N, (\lambda^H)^N) \neq 0.$$

$$v) \quad \frac{1}{N} \sum_{k=1}^{N} (\lambda_i^H)^N H_i(x^N, \xi_k) = 0, \text{ for } i = 1, \dots p.$$

Then the above equation (3.5) implies that

$$0 \in (\lambda_0^N)[\frac{1}{N}\sum_{k=1}^{N}\partial f(x^N,\xi^k)] + \sum_{i=1}^{l}(\lambda_i^G)^N[\frac{1}{N}\sum_{k=1}^{N}\nabla_x G_i(x^N,\xi_k)] + \sum_{i=1}^{p}(\lambda_i^H)^N[\frac{1}{N}\sum_{k=1}^{N}\nabla_x H_i(x^N,\xi^k)]. \quad (3.6)$$

Given that, $\nabla_x G_i(\cdot,\xi_k)$ and $\nabla_x H_j(\cdot,\xi_k)$ for $i = 1,\dots l$ and $j = 1,\dots p$ are continuous on a closed unit ball B_1^N around x^N, for $k = 1,\dots N$. Hence, $\frac{1}{N}\sum_{k=1}^{N}\nabla_x G_i(\cdot,\xi_k)$ and $\frac{1}{N}\sum_{k=1}^{N}\nabla_x H_j(\cdot,\xi_k)$ are continuous on the closed unit ball B_1^N, for $i = 1,\dots l.$ and $j = 1,\dots p$.

We know that $x^N \to x^*$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. Therefore, for some given $\varepsilon \in (0, \frac{1}{2})$ there exists $m \in \mathbb{N}$, such that $\| x^N - x^* \| < \varepsilon$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, $\forall\ N > m$. Consider any $N > m$ and as we have seen above $\| x^N - x^* \| < \frac{1}{2} < 1$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, hence $x^* \in B_1^N$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, as B_1^N is the closed unit ball with center x^N. So, if we take the closed unit ball B_1^M around x^M for some $M > m$, then $\{x^*\} \subseteq B_1^M$, $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$; and also note that for any $N > m$ we have $\| x^N - x^M \| \leq \| x^N - x^* \| + \| x^M - x^* \| < \frac{1}{2} + \frac{1}{2} = 1$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$. This implies $x^N \in B_1^M$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, for all $N > m$ and for some $M > m$. This is true for any arbitrary $M > m$, hence the sequence $\{x^N\}_{N>m} \subset B_1^M$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$ for any $M > m$. Thus $x^* \in B$ and $\{x^N\}_{N>m} \subset B$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, where $B = \bigcap_{N>m} B_1^N$.

Due to the assumption in Theorem 3.3 we have that $x \in B$ and for $i = 1,\dots l$

$$\| \nabla_x G_i(x,\xi) \| \leq C_i^G \qquad Q_{\boldsymbol{\xi}}\text{-a.s. for any } \boldsymbol{\xi} \in \Xi.$$

Further noting that $\nabla_x G_i(.,\xi(\omega))$ is continuous on B $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, for $i = 1,\dots l$. Applying Theorem 2.3, we can conclude that $E[\nabla_x G_i(x,\xi)]$ is finite valued and continuous; moreover, $\frac{1}{N}\sum_{k=1}^{N}\nabla_x G_i(.,\xi_k)$ converges to $E[\nabla_x G_i(.,\xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$ uniformly on B, for $i = 1,....l$. Hence, we obtain $\frac{1}{N}\sum_{k=1}^{N}\nabla_x G_i(x^N,\xi_k)$ converges to $E[\nabla_x G_i(x^*,\xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, for $i = 1,\dots l$. Similarly, we conclude that $\frac{1}{N}\sum_{k=1}^{N}\nabla_x H_j(x^N,\xi_k)$ converges to $E[\nabla_x H_j(x^*,\xi)]$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi} \in \Xi$, for $j = 1,\dots p$.

Further, passing the limit $N \to \infty$ to equation (3.6) and using Corollary 2.1 we obtain

$$0 \in \lambda_0^*\partial E[f(x^*,\xi)] + \sum_{i=1}^{l}(\lambda_i^G)^* E[\nabla_x G_i(x^*,\xi)] + \sum_{i=1}^{p}(\lambda_i^H)^* E[\nabla_x H_i(x^*,\xi)]. \quad (3.7)$$

Since $f(\cdot,\xi)$ is convex with finite expectation functional we exchange the Aumann integral and sub-differential operation (for details see [4]) in the above equation to obtain

$$0 \in \lambda_0^* E[\partial f(x^*,\xi)] + \sum_{i=1}^{l}(\lambda_i^G)^* E[\nabla_x G_i(x^*,\xi)] + \sum_{i=1}^{p}(\lambda_i^H)^* E[\nabla_x H_i(x^*,\xi)]. \quad (3.8)$$

Moreover, using Theorem 2.2 we conclude that $(\lambda_i^H)^* H_i(x^*,\xi) = 0$, for $i = 1,\dots p.$.

This concludes that x^* is a stationary point of the constrained optimization problem (3.3) as $E[\nabla_x G_i(x^*,\xi)]$, $i = 1,\dots l$ and $E[\nabla_x H_j(x^*,\xi)]$, $j = 1,\dots p$ are linearly independent. ∎

Epigraphical and uniform convergence of the sample averages play the pivotal role to establish large sample properties of MCSAA. One prominent dissimilarity with the existing literature is that the results on consistency and the assumptions to obtain those results hold almost surely with respect to the probability induced

by the Markov chain for any starting point. Another point worth mentioning is that to achieve consistency of optimal solutions and stationary points of the constrained problem uniform convergence is required as opposed to epigraphical convergence for unconstrained problem.

4 Large Deviation Bounds for the MCSAA Estimators

The large deviation theory deals with the methods of computing asymptotics of probabilities of rare events. It provides an estimation if the probability of visiting a non-typical state is exponentially small. A precise formula for the exponential rate of convergence can be established as the size of the system goes to infinity. Application of large deviation principle in the SAA literature is not very well studied. However, there are few articles where the optimizers used large deviation principle to create exponential bounds. They predominantly applied Cramér's Theorem for i.i.d. sample to establish exponential rate of convergence, for example see [18], [30] and [31]. As mentioned in the introduction, the authors in [9] and [12] applied Gärtner-Ellis theorem to obtain large deviation bounds. Both the theorems heavily depend on the convergence of the log moment generating function to obtain the large deviation principle. However, we follow the weak convergence approach to the large deviation theory to construct bounds for the MCSAA estimators. Weak convergence approach enables us to include general state space Markov chain.

Let us start the section by recalling the concept of weak topology in functional analysis. Let B be a Banach space and B^* be its adjoint space, i.e., B^* is the collection of all real-valued bounded linear functionals on B. The weak topology on B^* are characterized by the following family of neighbourhoods which for the basis of the weak topology: at any $g_0 \in B^*$, the family is defined as the sets $\{N(g_0; x_1, x_2, \dots x_n; \epsilon)\}$ for all possible choices of $\epsilon > 0$, integer n and points $x_1, x_2, \dots x_n$ in B; where $\{N(g_0; x_1, x_2, \dots x_n; \epsilon)\} := \{g : \mid g(x_i) - g_0(x_i) \mid < \epsilon, \text{ for } i = 1, 2, \dots n\}$. A sequence $\{g_n\}$ in B^* converges to g in B^* in weak topology if and only if $g_n(x) \to g(x)$, $\forall\ x \in B$ and we say g_n converges weakly to g.

We now recapitulate the functional analysis approach to define the underlying topological structure of weak convergence on the space of finite signed measures. It is important to understand the structure thoroughly to derive large deviation bounds using weak convergence. Let (X, d) be a metric space and $\mathcal{M}(X)$ be the space of all finite signed measures defined on $\mathcal{B}_X$. Moreover, $\mathcal{C}(X)$ stands for the space of all bounded real-valued continuous functions on X. $\mathcal{C}(X)$ is a Banach space under sup-norm. If X is compact metric space, it is well known that $\mathcal{C}(X)$ is a separable Banach space. Define a map from $(\mathcal{C}(X), \mathcal{M}(X))$ to $\mathbb{R}$ by

$$(f, \alpha) \mapsto < f, \alpha > := \int_X f d\alpha.$$

This is a bilinear form. Recall the following theorem in Varadarajan [36] which is fundamental in defining the weak topology on $\mathcal{M}(X)$.

Theorem 4.1. For any two finite measures α_1 and α_2 on a metrizable space X, the following conditions are equivalent:
(i) $\alpha_1 = \alpha_2$
(ii) $\int_X f d\alpha_1 = \int_X f d\alpha_2$, for every $f \in \mathcal{C}_{\mathcal{U}}(X)$,
where $\mathcal{C}_{\mathcal{U}}(X) \subset \mathcal{C}(X)$ is the subset of all bounded uniformly continuous functions on X and $\mathcal{C}_{\mathcal{U}}(X)$ is dense in $\mathcal{C}(X)$ under sup-norm.

Therefor, for a fixed $\alpha \in \mathcal{M}(X)$, $f \mapsto < f, \alpha > := \int_X f d\alpha$ defines a functional from $\mathcal{M}(X)$ to $\mathbb{R}$. Hence, we may identify $\mathcal{M}(X)$ as a subspace of $\mathbb{R}^{\mathcal{C}(X)}$, equipped with the product topology (due to the above theorem). As we mentioned earlier, for fixed $\alpha \in \mathcal{M}(X)$, $\phi_\alpha(\cdot) := < \cdot, \alpha >$ is linear, i.e., $\phi_\alpha(\cdot) \in \mathcal{C}(X)^*$. Hence, $\mathcal{M}(X) \subset \mathcal{C}(X)^*$. Let us now define a map from $\mathcal{C}(X)$ to $\mathcal{C}(X)^{**}$ as: $f \mapsto T_f$ such that $T_f(\phi_\alpha) = \phi_\alpha(f) =< f, \alpha >$, where $\mathcal{C}(X)^{**}$ is the double dual of $\mathcal{C}(X)$. The weak* topology on $\mathcal{M}(X)(\subset \mathcal{C}(X)^*)$ is the weak topology induced by the image of T: $T(\mathcal{C}(X)) \subset \mathcal{C}(X)^{**}$ as we discussed at the beginning of the section (in functional analysis). Since $\mathcal{M}_p(X) \subset \mathcal{M}(X)$, the weak* topology induces a topology on $\mathcal{M}_p(X)$, where $\mathcal{M}_p(X)$ is the space of all probability measures defined on X. In the literature of probability, this weak* topology is often called the weak topology, but precisely it is the weak* topology.

Similarly as in functional analysis, the base of the weak topology on $\mathcal{M}_p(X)$ can be characterized as follows: at any point α, the family of sets of the form $V(\alpha; f_1, f_2, \dots f_n; \epsilon)$ constitute the basis of the weak topology, where n is a positive integer, $f_1, f_2, \dots f_n \in \mathcal{C}(X)$ and $\epsilon > 0$. It is clear from the construction that a sequence $\{\alpha_n\}$ in $\mathcal{M}_p(X)$ converges in the weak topology to a probability measure α if and only if $\int_X f d\alpha_n \to \int_X f d\alpha$, for all $f \in \mathcal{C}(X)$. In such a case we say that α_n converges weakly to α. Hence, weak convergence of probability measures corresponds to weak topology on $\mathcal{M}_p(X)$. The detailed discussion about the properties of $\mathcal{M}_p(X)$ equipped with weak topology can be found in [22]. If X is separable and complete, weak topology can be defined by Prohorov metric. The Prohorov metric $d_p(\alpha_1, \alpha_2)$ between elements α_1 and α_2 in $\mathcal{M}_p(X)$ is defined as the infimum of those positive ϵ for which the two inequalities

$$\alpha_1(A) \leq \alpha_2(A^\epsilon) + \epsilon \quad \text{and} \quad \alpha_2(A) \leq \alpha_1(A^\epsilon) + \epsilon,$$

hold for all $A \in \mathcal{B}_X$, where $A^\epsilon = \{x \in X : \exists\, y \in A, d(x, y) < \epsilon\}$. For more details on Prohorov metric and weak convergence, we refer [5].

Suppose that (X, d) is a Polish space, and $\mathcal{M}_p(X)$ denotes the space of all probability measures on $(X, \mathcal{B}_X)$ endowed with weak topology. The precise definition of the large deviation principle is the following:

Definition 4.1. Let $I : X \to [0, 1]$ be a lower semicontinuous function and $r_n \to \infty$ be a sequence of positive real constants. A sequence of probability measures $\{\alpha_n\} \in \mathcal{M}_p(X)$ is said to satisfy a large deviation principle with rate function I and normalization factors r_n if the following inequalities hold:

$$\limsup_{n\to\infty} \quad \frac{1}{r_n} \log \alpha_n(F) \leq - \inf_{x \in F} I(x) \qquad \forall \text{ closed } F \subset X, \tag{4.1}$$

$$\liminf_{n\to\infty} \quad \frac{1}{r_n} \log \alpha_n(G) \geq - \inf_{x \in G} I(x) \qquad \forall \text{ open } G \subset X. \tag{4.2}$$

We abbreviate it as LDP(α_n, r_n, I) and call I as a rate function only if I is lower semicontinuous. When the sets $\{I \leq c\}$ are compact for all $c \in \mathbb{R}$, we say I is a tight rate function.

4.1 Pointwise Large Deviation Bounds

Let us consider our problem framework where Ξ is a closed subspace of $\mathbb{R}^n$, i.e., Ξ is a Polish space. This implies that $\mathcal{M}_p(\Xi)$ is a Polish space as well. Hence, weak convergence of probability measures in $\mathcal{M}_p(\Xi)$ and Prohorov metric on $\mathcal{M}_p(\Xi)$ are equivalent. The weak topology on $\mathcal{M}_p(\Xi)$ can be defined by Prohorov metric.

$Q_{\boldsymbol{\xi}}$ is the induced probability measure by the Markov chain $\{\xi_k : k = 1, 2, \ldots\}$ on Ω, for a fixed starting point $\boldsymbol{\xi} \in \Xi$. For each $\omega \in \Omega$, positive integer N and Borel subset B in $\mathcal{B}(\Xi)$, define the empirical measure

$$L_N(\omega, B) = \frac{1}{N} \sum_{k=1}^{N} \delta_{\xi_k(\omega)} B.$$

For each positive integer N, L_N maps Ω into $\mathcal{M}_p(\Xi)$. L_N's can be considered as the $\mathcal{M}_p(\Xi)$-valued random variables defined on $(\Omega, \mathcal{B}, Q_{\boldsymbol{\xi}})$. Therefore, for a fixed starting point $\boldsymbol{\xi}$, the sequence of empirical measures $\{L_N\}$ induces a sequence of probability measures $\{\Gamma_{N,\boldsymbol{\xi}}\}$ on $\mathcal{M}_p(\Xi)$ such that $\Gamma_{N,\boldsymbol{\xi}} = Q_{\boldsymbol{\xi}} L_N^{-1}$, i.e., $\Gamma_{N,\boldsymbol{\xi}}(A) = Q_{\boldsymbol{\xi}}\{\omega : L_N(\omega, \cdot) \in A\}$, where A is a Borel subset of $\mathcal{M}_p(\Xi)$ endowed with weak topology. For different starting points of the Markov chain, the induced probability measures $\{\Gamma_{N,\boldsymbol{\xi}}\}$ on $\mathcal{M}_p(\Xi)$ will be different. In case of the i.i.d. random variables, induced probability measures $\{\Gamma_N\} = PL_N^{-1}$ are independent of the starting point, where P is the underlying probability on Ω. Sanov in [28] developed a large deviation bound using Kullback-Leibler divergence (relative entropy) for the sequence $\{\Gamma_N\}$. Detailed discussion with historical notes and references can be found in [10].

The work of Sanov has been extended by Donsker and Varadhan in [11]. They developed a large deviation bound for the sequence $\{\Gamma_{N,\boldsymbol{\xi}}\}$ associated to a Ξ-valued Markov chain $\{\xi_k : k = 1, 2, \ldots\}$ with $\boldsymbol{\xi}$ as the starting point and whose transition kernel $Q(\boldsymbol{\xi}, d\boldsymbol{\xi'})$ satisfies the following property:

Assumption 4.1. (1) $Q(\boldsymbol{\xi}, d\boldsymbol{\xi'}) = Q(\boldsymbol{\xi}, \boldsymbol{\xi'})\lambda(d\boldsymbol{\xi'})$, where λ is the reference measure on Ξ,
(2) there exist constants a and b such that $0 < a \leq Q(\boldsymbol{\xi}, \boldsymbol{\xi'}) \leq b < +\infty$, for all $\boldsymbol{\xi} \in \Xi$ and λ-almost all $\boldsymbol{\xi'} \in \Xi$,
(3) for any function $u(\cdot) \in L_1(\lambda)$,

$$\int_{\Xi} Q(\boldsymbol{\xi}, \boldsymbol{\xi'}) u(\boldsymbol{\xi'}) \lambda(d\boldsymbol{\xi'})$$

is continuous in $\boldsymbol{\xi}$. That means, transition kernel have a density $Q(\boldsymbol{\xi}, \boldsymbol{\xi'})$ with respect to the reference measure $\lambda(d\boldsymbol{\xi'})$ which satisfies (2) and (3).

Let $\mathcal{U}_1$ be the set of functions $u \in \mathcal{C}(\Xi)$ such that $u > 0$ on Ξ. With $Qu(\boldsymbol{\xi'}) := \int_{\Xi} u(\boldsymbol{\xi'}) Q(\boldsymbol{\xi}, d\boldsymbol{\xi'})$, the authors defined that

$$I(\mu) = -\inf_{u \in \mathcal{U}_1} \int_{\Xi} \log\left(\frac{Qu}{u}\right)(\boldsymbol{\xi}) \mu(d\boldsymbol{\xi}),$$

for any probability measure μ on Ξ. This is known as the Donsker-Varadhan entropy [24]. The following version of the large deviation principle was developed by Donsker and Varadhan under the assumptions $(1) - (3)$ with $I(\cdot)$ as the rate function.

Theorem 4.2. For any closed set $C \subset \mathcal{M}_p(\Xi)$,

$$\limsup_{N \to \infty} \quad \frac{1}{N} \log \Gamma_{N,\boldsymbol{\xi}}(C) \leq -\inf_{\alpha \in C} I(\alpha)$$

For any open set $G \subset \mathcal{M}_p(\Xi)$,

$$\liminf_{N \to \infty} \quad \frac{1}{N} \log \Gamma_{N,\boldsymbol{\xi}}(G) \geq -\inf_{\alpha \in G} I(\alpha)$$

The open and the closed sets are considered with respect to the weak topology on $\mathcal{M}_p(\Xi)$.

We are now all set to construct our large deviation bounds. We apply the preceding theorem to construct our pointwise large deviation bounds for the functional values of the objective corresponding to the constrained and unconstrained stochastic optimization problems formulated in the previous section.

Theorem 4.3. Let $f(x,\cdot)$ be bounded and continuous in $\boldsymbol{\xi}$, and Ξ be closed. We also assume that $\{\xi_k : k = 1,2,\ldots\}$ with transition kernel $Q(\cdot,\cdot)$ satisfies Assumption 4.1. Then the following exponential bounds hold for all x, for any starting point $\boldsymbol{\xi} \in \Xi$ of the Markov chain and given $\epsilon > 0$.

$$\limsup_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{ \mid \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \int f(x,\xi)d\pi \mid < \epsilon\} \leq -\inf_{\alpha\in\overline{B(\pi)}} I(\alpha),$$

$$\text{and,} \quad \liminf_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{ \mid \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \int f(x,\xi)d\pi \mid < \epsilon\} \geq -\inf_{\alpha\in B(\pi)} I(\alpha);$$

where $B(\pi)$ is the ϵ-neighbourhood around π with respect to Prohorov metric and $\overline{B(\pi)}$ is the closure.

Proof. Ξ is a Polish space as wee assumed it to be closed. Hence, $\mathcal{M}_p(\Xi)$ is also a Polish space with respect to the weak topology. From the previous discussion, we know that the weak topology can be defined by Prohorov metric if Ξ is a Polish space. Consider an ϵ-neighbourhood around π in $\mathcal{M}_p(\Xi)$ with respect to the Prohorov metric. Therefore, by using the large deviation principle developed in Theorem 4.2, we have

$$\limsup_{N\to\infty} \quad \frac{1}{N}\log \Gamma_{N,\boldsymbol{\xi}}(\overline{B(\pi)}) \leq -\inf_{\alpha\in\overline{B(\pi)}} I(\alpha).$$

This implies,

$$\limsup_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{\omega : L_N(\omega,\cdot) \in (\overline{B(\pi)}) \leq -\inf_{\alpha\in\overline{B(\pi)}} I(\alpha),$$

where $\overline{B(\pi)}$ is the closure. Moreover, we assume that $f(x,\cdot)$ is bounded and real-valued continuous in $\boldsymbol{\xi}$, i.e., $f(x,\cdot) \in \mathcal{C}(\Xi)$, for all $x \in \mathbb{R}^n$. Hence, exploiting the definition of convergence in weak topology we conclude that

$$\limsup_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{ \mid \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \int f(x,\xi)d\pi \mid \leq \epsilon\} \leq -\inf_{\alpha\in\overline{B(\pi)}} I(\alpha).$$

Hence,

$$\limsup_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{ \mid \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \int f(x,\xi)d\pi \mid < \epsilon\} \leq -\inf_{\alpha\in\overline{B(\pi)}} I(\alpha).$$

The upper bound holds for all x and for any $\epsilon > 0$. Similarly, we can show that

$$\liminf_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{ \mid \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \int f(x,\xi)d\pi \mid < \epsilon\} \geq -\inf_{\alpha\in B(\pi)} I(\alpha).$$

∎

The countable state space irreducible Markov chain follows Donsker-Varadhan large deviation principle. Hence, we can achieve the pointwise large deviation bounds for the functional values of the objective using Donsker-Varadhan entropy, see [24].

Donsker and Varadhan [11] developed a large deviation principle for the probability measures associated to the sequence of the empirical measures. Whereas, R. S. Ellis [14] in 1988 had established a large deviation

principle for the sequence of empirical measures itself. Instead of assuming a density $Q(\boldsymbol{\xi}, \boldsymbol{\xi'})$ of the transition kernel $Q(\boldsymbol{\xi}, d\boldsymbol{\xi'})$ with respect to the reference measure $\lambda(d\boldsymbol{\xi'})$, the author assumed the following uniformity hypothesis.

Assumption 4.2. For some β and N in $\mathbb{Z}^+$, some $M \in [1, \infty)$, all $\boldsymbol{\xi}$, $\boldsymbol{\xi'}$ in Ξ and all Borel subsets A of Ξ,

$$Q^{\beta}(\boldsymbol{\xi}, A) \leq \frac{M}{N} \sum_{i=1}^{N} Q^{i}(\boldsymbol{\xi'}, A),$$

where $Q^i(\boldsymbol{\xi'}, \cdot)$ denotes the i^{th}-step transition probability of the Markov chain for initial condition $\boldsymbol{\xi'}$ which can be obtained by applying Chapman-Kolmogorov equation.

This assumption was actually employed by Stroock [33] to formulate its preceding result. This assumption clearly holds true for every finite state irreducible Markov chain, see [10].

Theorem 4.4. Let Ξ be a complete separable metric space space. We assume that $Q(\cdot, \cdot)$ satisfies Assumption 4.2. Define for $\alpha \in \mathcal{M}_p(\Xi)$

$$J_Q(\alpha) = \sup_{u \in \mathcal{U}(\Xi)} \int_{\Xi} \frac{u(\boldsymbol{\xi})}{(Qu)(\boldsymbol{\xi})} \alpha(d\boldsymbol{\xi}),$$

where $\mathcal{U}(\Xi)$ denotes the set of $u \in \mathcal{C}(\Xi)$ such that $u \geq \delta$ on Ξ for some $\delta = \delta(u) > 0$ and $(Qu)(\boldsymbol{\xi}) := \int_{\Xi} u(\boldsymbol{\xi'}) Q(\boldsymbol{\xi}, d\boldsymbol{\xi'})$. The following large deviation principle LDP (L_N, N, J_Q) holds with the convex rate function J_Q.
For each closed set F in $\mathcal{M}_p(\Xi)$

$$\limsup_{N \to \infty} \quad \frac{1}{N} \log\Big(\sup_{\boldsymbol{\xi} \in \Xi} \; Q_{\boldsymbol{\xi}}\{\omega : L_N(\omega, \cdot) \in F\} \Big) \leq - \inf_{\alpha \in F} J_Q(\alpha),$$

and for each open set G in $\mathcal{M}_p(\Xi)$

$$\liminf_{N \to \infty} \quad \frac{1}{N} \log\Big(\inf_{\boldsymbol{\xi} \in \Xi} \; Q_{\boldsymbol{\xi}}\{\omega : L_N(\omega, \cdot) \in G\} \Big) \geq - \inf_{\alpha \in G} J_Q(\alpha).$$

$\mathcal{M}_p(\Xi)$ is equipped with weak topology.

Using the above large deviation principle [14], we can obtain an independent pointwise large deviation bounds for the functional values of the objective, i.e., independent of the starting point. The theorem on independent pointwise large deviation bounds is the following.

Theorem 4.5. Let $f(x, \cdot)$ be bounded and continuous in $\boldsymbol{\xi}$, and Ξ be closed. We also assume that $\{\xi_k : k = 1, 2, \ldots\}$ with transition kernel $Q(\cdot, \cdot)$ satisfies Assumption 4.2. Then the following exponential bounds hold for all x and given $\epsilon > 0$.

$$\limsup_{N \to \infty} \quad \frac{1}{N} \log \left(\sup_{\boldsymbol{\xi} \in \Xi} Q_{\boldsymbol{\xi}}\{ \mid \frac{1}{N} \sum_{k=1}^{N} f(x, \xi_k) - \int f(x, \xi) d\pi \mid < \epsilon \} \right) \leq - \inf_{\alpha \in \overline{B(\pi)}} J_Q(\alpha),$$

$$\text{and,} \quad \liminf_{N \to \infty} \quad \frac{1}{N} \log\Big(\inf_{\boldsymbol{\xi} \in \Xi} \; Q_{\boldsymbol{\xi}}\{ \mid \frac{1}{N} \sum_{k=1}^{N} f(x, \xi_k) - \int f(x, \xi) d\pi \mid < \epsilon \} \Big) \geq - \inf_{\alpha \in B(\pi)} J_Q(\alpha);$$

where $B(\pi)$ is the ϵ-neighbourhood around π with respect to Prohorov metric and $\overline{B(\pi)}$ is the closure.

Proof. The proof follows the same path as Theorem 4.3. ∎

The above large deviation bounds are independent with respect to the starting point of the Markov chain and pointwise with respect to the decision variable. These bounds also hold for finite state space irreducible Markov chain as it satisfies the Assumption 4.2. Hence, both large deviation bounds are satisfied by the finite state space irreducible Markov chain.

4.2 Uniform Large Deviation Bounds

Let us assume that the constraints of the stochastic optimization problem (3.3) and the MCSAA subproblems (3.4) are included in a fixed set $\mathcal{S}$. We assume that $\mathcal{S}$ is compact and $\mathcal{C}(\mathcal{S})$ is the space of all real-valued continuos functions defined on $\mathcal{S}$ and endowed with the sup-norm. This forms a Banach space under the sup-norm. Let us denote

$$F_N(\cdot,\omega) := \frac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k(\omega)) \quad \text{and} \quad F(\cdot) := \int_{\Xi} f(\cdot,\xi)d\pi.$$

$F_N(\cdot,\omega)$ and $F(\cdot)$ are defined on $\mathcal{S}$ and $F_N(\cdot,\omega), F_N(\cdot) \in \mathcal{C}(\mathcal{S})$, for all $\omega \in \Omega$. Hence, for each N, $F_N(\cdot,\omega)$ is a $\mathcal{C}(\mathcal{S})$-valued random variable. The next theorem develops uniform bounds with the help of the pointwise large deviation bounds.

Theorem 4.6. Suppose that the set $\mathcal{S}$ is compact . Let $f(x,\cdot)$ be bounded and continuous in $\boldsymbol{\xi}$, and Ξ be closed. We also assume that $\{\xi_k : k = 1,2,\ldots\}$ with transition kernel $Q(\cdot,\cdot)$ satisfies Assumption 4.1. Then the following exponential bounds hold for any starting point $\boldsymbol{\xi} \in \Xi$ of the Markov chain and given $\epsilon > 0$.

$$\limsup_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{\ ||F_N(\cdot,\omega) - F(\cdot)|| < \epsilon\} \leq -\inf_{\alpha\in\overline{B(\pi)}} I(\alpha),$$

$$\text{and,} \quad \liminf_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{\ ||F_N(\cdot,\omega) - F(\cdot)|| < \epsilon\} \geq -\inf_{\alpha\in B(\pi)} I(\alpha),$$

where $||\cdot||$ denotes the sup-norm on $\mathcal{C}(\mathcal{S})$; $B(\pi)$ is the ϵ-neighbourhood around π with respect to Prohorov metric and $\overline{B(\pi)}$ is the closure.

Proof. By the pointwise large deviation bounds developed in Theorem 4.3, for each $x \in \mathcal{S}$ and given $\epsilon > 0$, we have that

$$\limsup_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{\ |\ \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \int f(x,\xi)d\pi\ | < \epsilon\} \leq -\inf_{\alpha\in\overline{B(\pi)}} I(\alpha).$$

This implies that the bound holds uniformly in x (as $\mathcal{S}$ is compact), i.e.,

$$\limsup_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{\ \sup_{x\in\mathcal{S}}\ |\ \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \int f(x,\xi)d\pi\ | < \epsilon\} \leq -\inf_{\alpha\in\overline{B(\pi)}} I(\alpha).$$

Therefore, we have

$$\limsup_{N\to\infty} \quad \frac{1}{N}\log Q_{\boldsymbol{\xi}}\{\ ||F_N(\cdot,\omega) - F(\cdot)|| < \epsilon\} \leq -\inf_{\alpha\in\overline{B(\pi)}} I(\alpha),$$

where $||\cdot||$ denotes the sup-norm on $\mathcal{C}(\mathcal{S})$. Similarly, we have the lower bound as

$$\liminf_{N\to\infty} \quad \frac{1}{N} \log Q_{\boldsymbol{\xi}}\{ \, ||F_N(\cdot,\omega) - F(\cdot)|| < \epsilon\} \geq - \inf_{\alpha\in B(\pi)} I(\alpha).$$

■

The independent (with respect to starting point of the Markov chain) uniform large deviation bounds can be obtained from the independent pointwise large deviation bounds developed in Theorem 4.5 if $\{\xi_k : k = 1, 2, \ldots\}$ with transition kernel $Q(\cdot,\cdot)$ satisfies Assumption 4.2. The proof follows the similar path as in Theorem 4.6. Consequently, the uniform large deviation bounds hold for the finite state space Markov chain.

4.3 Large Deviation Bounds for the Optimal Value

In this section, we develop large deviation bounds for the Markov chain SAA estimators of the optimal value of the constrained and unconstrained stochastic optimization problems defined in the previous section (Section 3). We construct the bounds by applying a delta method developed for the large deviation principle.

Let us first recall the concept of Hadamard directional differentiability and the related results which are required for the delta method. Let $\mathcal{X}$ and $\mathcal{Y}$ be two metrizable topological linear spaces. A map ϕ defined on $\mathcal{X}$ with values in $\mathcal{Y}$ is called Hadamard differentiable at $x \in \mathcal{X}$ if there exists a continuous mapping $\phi'_x : \mathcal{X} \to \mathcal{Y}$ such that

$$\lim_{n\to\infty} \frac{\phi(x + t_n h_n) - \phi(x)}{t_n} = \phi'_x(h) \tag{4.3}$$

holds for all sequences $t_n \to 0+$ and $h_n \to h \in \mathcal{X}$. $\phi'_x(\cdot)$ is often nonlinear, however, by definition we can see that $\phi'_x(\cdot)$ is positively homogeneous, i.e., $\phi'_x(th) = t\phi'_x(h)$, for all $t > 0$ and $h \in \mathcal{X}$.

The notion of Hadamard directional derivative can be refined to Hadamard differentiable tangentially to a subset $\mathcal{K}$ of $\mathcal{X}$; the map ϕ is said to be Hadamard differentiable at $x \in \mathcal{X}$ tangentially to $\mathcal{K}$ if the limit (4.3) exists for all sequences $t_n \to 0+$ and $h_n \to h$ in $\mathcal{K}$ such that $x + t_n h_n \in \mathcal{X}$, for every n. In this case, $\phi'_x(\cdot)$ is a continuos mapping on $\mathcal{K}$. For more detailed discussion on Hadamard directional differentiability, we refer [29].

We state the following delta method in the large deviation theory which is the fundamental result in developing our large deviation bounds for the MCSAA estimators of the optimal value.

Theorem 4.7. Let $\mathcal{X}$ and $\mathcal{Y}$ be two metrizable topological linear spaces and let d and ρ be the compatible metrics on $\mathcal{X}$ and $\mathcal{Y}$. Let $\phi : \mathcal{X} \to \mathcal{Y}$ be Hadamard differentiable at θ tangentially to $\mathcal{K} \subset \mathcal{X}$. Let $X_n : \Omega_n \to \mathcal{X}$, $n \geq 1$ be a sequence of random maps and let r_n, $n \geq 1$ be a sequence of positive real numbers satisfying $r_n \to +\infty$.
If $\{r_n(X_n - \theta) : n \geq 1\}$ satisfies the large deviation principle with speed $\lambda(n)$ and rate function I and $\{I < \infty\} \subset \mathcal{K}$, then $\{r_n(\phi(X_n) - \phi(\theta)) : n \geq 1\}$ satisfies the large deviation principle with speed $\lambda(n)$ and rate function $I_{\phi'_\theta}$, where

$$I_{\phi'_\theta}(y) = \inf\{I(x) : \phi'_\theta(x) = y\} \quad y \in \mathcal{Y}.$$

The statement and the proof can be found in [15].

Let us assume that the constraints of the stochastic optimization problem (3.3) and the MCSAA subproblems (3.4) are included in a fixed set $\mathcal{S}$. Therefore, define the optimal value function as:

$$\overline{\phi}(g) := \inf\{g(x) : x \in \mathcal{S}\}. \tag{4.4}$$

Following the notion of Shapiro [29], we assume that $\mathcal{S}$ is compact and that g belongs to the Banach space $\mathcal{C}(\mathcal{S})$ of the real-valued continuos functions defined on $\mathcal{S}$ and endowed with the sup-norm. The following theorem stated in [29] prove the Hadamard differentiability of the optimal value function $\overline{\phi}(\cdot)$ on $\mathcal{C}(\mathcal{S})$.

Theorem 4.8. Suppose that the set $\mathcal{S}$ is compact and $\overline{\phi} : \mathcal{C}(\mathcal{S}) \to \mathbb{R}$ is the optimal value function defined in (4.4). Then for any $\eta \in \mathcal{C}(\mathcal{S})$, $\overline{\phi}$ is Hadamard directionally differentiable at η and

$$\overline{\phi'_\eta}(\zeta) = \min_{x \,\in\, \operatorname{argmin}(\eta)} \zeta(x).$$

With the help of the pointwise large deviation bounds and the Hadamard directional differentiability of the optimal value function, we establish the bounds of the optimal value of the optimization problem. Let $x^*$and x^N be the solutions of the stochastic optimization problem (3.3) and the MCSAA subproblems (3.4). We know that $x^N \to x^*$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi}$.

Theorem 4.9. Suppose that the set $\mathcal{S}$ is compact and $\overline{\phi} : \mathcal{C}(\mathcal{S}) \to \mathbb{R}$ is the optimal value function. Let $f(x, \cdot)$ be bounded and continuous in $\boldsymbol{\xi}$, and Ξ be closed. We also assume that $\{\xi_k : k = 1, 2, \ldots\}$ with transition kernel $Q(\cdot, \cdot)$ satisfies Assumption 4.1. Then the following exponential bounds hold for any starting point $\boldsymbol{\xi} \in \Xi$ of the Markov chain and given $\epsilon > 0$.

$$\limsup_{N\to\infty} \quad \frac{1}{N} \log Q_{\boldsymbol{\xi}}\{ \mid \min_{x\in\mathcal{S}} \frac{1}{N} \sum_{k=1}^{N} f(x, \xi_k) - \min_{x\in\mathcal{S}} \int f(x, \xi) d\pi \mid < \delta(\epsilon)\} \leq - \inf_{\alpha \in \overline{B(\pi)}} I(\alpha),$$

$$\text{and,} \quad \liminf_{N\to\infty} \quad \frac{1}{N} \log Q_{\boldsymbol{\xi}}\{ \mid \min_{x\in\mathcal{S}} \frac{1}{N} \sum_{k=1}^{N} f(x, \xi_k) - \min_{x\in\mathcal{S}} \int f(x, \xi) d\pi \mid < \delta(\epsilon)\} \geq - \inf_{\alpha \in B(\pi)} I(\alpha),$$

for any $\delta(\epsilon) > \epsilon$; where $B(\pi)$ is the ϵ-neighbourhood around π with respect to Prohorov metric and $\overline{B(\pi)}$ is the closure.

Proof. Let us fix some $\epsilon > 0$. By the uniform upper bound in Theorem 4.6, we have

$$\limsup_{N\to\infty} \quad \frac{1}{N} \log Q_{\boldsymbol{\xi}}\{ \, ||F_N(\cdot, \omega) - F(\cdot)|| < \epsilon\} \leq - \inf_{\alpha \in \overline{B(\pi)}} I(\alpha),$$

where $||\cdot||$ denotes the sup-norm on $\mathcal{C}(\mathcal{S})$. The upper bound holds true even if we consider the closed ϵ-ball around $F(\cdot)$. Say, $J^{upper}(\cdot)$ is the upper bound defined on $\overline{B_\epsilon(F)} \subset \mathcal{C}(\mathcal{S})$ such that $J^{upper}(\zeta) = \inf_{\alpha \in \overline{B(\pi)}} I(\alpha)$, i.e., constant rate function on $\overline{B_\epsilon(F)}$, where $\overline{B_\epsilon(F)}$ is the closed ϵ-neighbourhood of F in $\mathcal{C}(\mathcal{S})$. Hence, we obtain a large deviation upper bound locally on $\overline{B_\epsilon(F)}$ for $\{F_N(\cdot, \omega) - F(\cdot) : N \geq 1\}$.
Moreover, we have

$$|F_N(x^N, \omega) - F(x^*)| \leq |F_N(x^N, \omega) - F(x^N)| + |F(x^N) - F(x^*)|.$$

The first term on the right hand side of the above inequality is less than $\epsilon > 0$ (fixed) for all $N \geq N(\epsilon)$ and let us assume the second term is less than some fixed quantity $\epsilon' > 0$ for all $N > N(\epsilon', \omega)$. Hence,

$$|F_N(x^N, \omega) - F(x^*)| < \epsilon + \epsilon' = \delta(\epsilon), \text{for all } N \geq \max\{N(\epsilon), N(\epsilon', \omega)\}$$

By Theorem 4.8, $\overline{\phi} : \mathcal{C}(\mathcal{S}) \to \mathbb{R}$ is Hadamard directionally differentiable at $F(\cdot) \in \mathcal{C}(\mathcal{S})$. Hence, by using the delta method in large deviations stated in Theorem 4.7, we obtain a local large deviation upper bound for $\{\overline{\phi}(F_N(\cdot,\omega)) - \overline{\phi}(F(\cdot)) : N \geq 1\}$ with rate function $J^{upper}_{\overline{\phi}'_F}(\cdot)$ defined on $\overline{B_{\delta(\epsilon)}(F(x^*))}$. Then, for any starting point $\boldsymbol{\xi} \in \Xi$ of the Markov chain and given $\epsilon > 0$, the upper bound can be written as

$$\limsup_{N\to\infty} \quad \frac{1}{N} \log Q_{\boldsymbol{\xi}}\{ \mid \min_{x\in\mathcal{S}} \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \min_{x\in\mathcal{S}} \int f(x,\xi) d\pi \mid < \delta(\epsilon)\} \leq - \inf_{y\in\overline{B_{\delta(\epsilon)}(F(x^*))}} J^{upper}_{\overline{\phi}'_F}(y),$$

where $\overline{B_{\delta(\epsilon)}(F(x^*))}$ is the closure of the $\delta(\epsilon)$-radius ball around $F(x^*)$. Moreover,

$$J^{upper}_{\overline{\phi}'_F}(y) = \inf\{J^{upper}(\zeta) : \zeta \in \overline{B_\epsilon(F)} \text{ and } \min_{x\in \text{argmin}(F)} \zeta(x) = y\} = \inf_{\alpha\in\overline{B(\pi)}} I(\alpha).$$

This implies that,

$$\limsup_{N\to\infty} \quad \frac{1}{N} \log Q_{\boldsymbol{\xi}}\{ \mid \min_{x\in\mathcal{S}} \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \min_{x\in\mathcal{S}} \int f(x,\xi) d\pi \mid < \delta(\epsilon)\} \leq - \inf_{y\in\overline{B_{\delta(\epsilon)}(F(x^*))}} \inf_{\alpha\in\overline{B(\pi)}} I(\alpha).$$

$$\text{i.e.,} \quad \limsup_{N\to\infty} \quad \frac{1}{N} \log Q_{\boldsymbol{\xi}}\{ \mid \min_{x\in\mathcal{S}} \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \min_{x\in\mathcal{S}} \int f(x,\xi) d\pi \mid < \delta(\epsilon)\} \leq - \inf_{\alpha\in\overline{B(\pi)}} I(\alpha).$$

The value of $\delta(\epsilon)$ depends on the chosen values of ϵ and ϵ' and $\delta(\epsilon) > \epsilon$ as $\epsilon' > 0$. We also observe that ϵ' can assume any positive value, hence the upper bound holds for any $\delta(\epsilon) > \epsilon$. Similarly, we obtain a large deviation lower bound locally on $B_\epsilon(F)$ for $\{F_N(\cdot,\omega) - F(\cdot) : N \geq 1\}$ with constant rate function $J^{lower}(\zeta) = \inf_{\alpha\in B(\pi)} I(\alpha)$, which implies

$$\liminf_{N\to\infty} \quad \frac{1}{N} \log Q_{\boldsymbol{\xi}}\{ \mid \min_{x\in\mathcal{S}} \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \min_{x\in\mathcal{S}} \int f(x,\xi) d\pi \mid < \delta(\epsilon)\} \geq - \inf_{\alpha\in B(\pi)} I(\alpha).$$

∎

Remark 2. Let us fix an $\epsilon > 0$. The bounds hold for any subset of $\mathbb{R}$ which contains a δ-neighbourhood of $F(x^*)$, for all $\delta > \epsilon$. However, we should consider the δ-neighbourhoods around $F(x^*)$ with $\delta > \epsilon$, as $F_N(x^N,\omega) \to F(x^*)$ $Q_{\boldsymbol{\xi}}$-a.s. for any $\boldsymbol{\xi}$.

The following independent (with respect to starting points) large deviation bounds for the optimal value can be obtained by applying Theorem 4.5.

Theorem 4.10. Suppose that the set $\mathcal{S}$ is compact and $\overline{\phi} : \mathcal{C}(\mathcal{S}) \to \mathbb{R}$ is the optimal value function. Let $f(x,\cdot)$ be bounded and continuous in $\boldsymbol{\xi}$, and Ξ be closed. We also assume that $\{\xi_k : k = 1,2,\ldots\}$ with transition kernel $Q(\cdot,\cdot)$ satisfies Assumption 4.2. Then the following exponential bounds hold for any starting point $\boldsymbol{\xi} \in \Xi$ of the Markov chain and given $\epsilon > 0$.

$$\limsup_{N\to\infty} \quad \frac{1}{N} \log \left(\sup_{\boldsymbol{\xi}\in\Xi} Q_{\boldsymbol{\xi}}\{ \mid \min_{x\in\mathcal{S}} \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \min_{x\in\mathcal{S}} \int f(x,\xi) d\pi \mid < \delta(\epsilon)\} \right) \leq - \inf_{\alpha\in\overline{B(\pi)}} J_Q(\alpha),$$

$$\text{and,} \quad \liminf_{N\to\infty} \quad \frac{1}{N} \log \left(\sup_{\boldsymbol{\xi}\in\Xi} Q_{\boldsymbol{\xi}}\{ \mid \min_{x\in\mathcal{S}} \frac{1}{N}\sum_{k=1}^{N} f(x,\xi_k) - \min_{x\in\mathcal{S}} \int f(x,\xi) d\pi \mid < \delta(\epsilon)\} \right) \geq - \inf_{\alpha\in B(\pi)} J_Q(\alpha),$$

for any $\delta(\epsilon) > \epsilon$.

Proof. The proof is similar to that of Theorem 4.9. ∎

5 Conclusion

This article provides a theoretical technique to extend the literature of SAA, in general stochastic optimization, beyond i.i.d. sample. The sample is drawn from a Harris recurrent general state space Markov chain with the underlying distribution of ξ as the stationary distribution, instead of an i.i.d. sample. MCMC algorithms confirm generation of such a Markov chain even if the underlying distribution is partially realized. We then explore the asymptotic properties of the MCSAA estimators of the optimal solutions, optimal value and the optimality conditions of stochastic optimization problems (3.1) and (3.3) using epigraphical and uniform convergence of the sample averages established in this article. All properties hold almost surely with respect to the probability distribution $Q_{\boldsymbol{\xi}}$ induced by the Markov chain for any starting point $\boldsymbol{\xi}$. Harris recurrence allows us to include all starting points of a general state space Markov chain.

Last section is dedicated to develop large deviation bounds. The weak convergence approach to the large deviation theory has been employed to establish the bounds. This approach relies heavily on the space of probability measures defined on Ξ and equipped with the weak topology. For separable and complete Ξ, the weak topology can be obtained by Prohorov metric and the rate function is defined with respect to this metric. In the literature of stochastic optimization, to the best of my knowledge, this approach has never been employed. Pointwise and uniform large deviation bounds for the functionals values of the objective function are obtained by applying the large deviation principle developed by Donsker and Varadhan (Theorem 4.2). As an extension, the independent (with respect to starting point of the Markov chain) large deviation bounds are obtained using the large deviation principle developed by R. S. Ellis (Theorem 4.4). Both large deviation principles were established in terms of the weak topology defined on the space of probability measures on Ξ.

The pointwise and the uniform errors converge to zero exponentially fast in the sample size N, i.e., the rate of convergence is of order $\frac{1}{N}$. The independent large deviation errors also converge to zero exponentially fast as N tends to ∞. Consequently, the large deviation bounds for the MCSAA estimators of the optimal value are established by using the uniform large deviation bounds of the functionals values. The rate of convergence for the pointwise and uniform MCSAA estimators is carried over to the MCSAA estimators of the optimal value, i.e., the convergence of the approximating optimal values to their true counterpart has the same rate as in pointwise and uniform estimation.

References

[1] S. Asmussen and P. W. Glynn (2011). A New Proof of Convergence of MCMC via the Ergodic Theorem. *Statistics and Probability Letters*. Vol. 81, pp. 1482-1485.

[2] S. Asmussen and P. W. Glynn (2007). *Stochastic Simulation:Algorithms and Analysis*. Springer-Verlag.

[3] K. B. Athreya and S. N. Lahiri (2006). *Measure Theory and Probability Theory*. Springer Texts in Statistics, New York, USA.

[4] Robert J. Aumann (1965). Integrals of set-valued functions. *J. Math. Anal. Appl.* Vol. 12, pp. 1-12.

[5] P. Billingsley (1999). *Convergence of probability measures.* Second edition. Wiley Series in Probability and Statistics: Probability and Statistics. A Wiley-Interscience Publication. John Wiley and Sons, Inc., New York.

[6] J. R. Birge and L. Q. Qi (1995). Subdifferential convergence in stochastic programs. *SIAM J. Optimization.* Vol. 2, pp. 436-453.

[7] J. R. Birge and R. J.-B. Wets (1986). Designing approximation schemes for stochastic problems, in particular for stochastic programs with recourse. *Math. Programming Stud.* Vol. 27, pp. 54-102.

[8] K. S. Chan and C. J. Geyer (1994). Comment on "Markov chains for exploring posterior distributions" by L. Tierney. *Ann. Statist.* Vol. 22, pp. 1747-1758.

[9] L. Dai, C. H. Chen and J. R. Birge (2000). Convergence properties of two-stage stochastic programming. *J. Optim Theory Appl.* Vol. 106, pp. 489-509.

[10] A. Dembo and O. Zeitouni (1998). *Large Deviations Techniques and Applications.* 2nd ed., Springer-Verlag, New York.

[11] M. D. Donsker and S. R. S. Varadhan (1975). Asymptotic evaluation of certain Markov process expectations for large time, I. *Comm. Pure Appl. Math.* Vol. 28, pp. 1-47.

[12] S. S. Drew and T. Homem-de-Mello (2012). Some large deviations results for Latin hypercube sampling. *Methodol. Comput. Appl. Probab.* Vol. 14, pp. 203-232.

[13] J. Dupačová and R. J.-B. Wets (1988). Asymptotic behaviour of statistical estimators and of optimal solutions of stochastic optimization problems. *Annals of Statistics.* Vol. 16, pp. 1517-1549.

[14] R. S. Ellis (1988). Large deviations for the empirical measure of a Markov chain with an application to the multivariate empirical measure. *Ann. Probab.* Vol. 16, pp. 1496-1508.

[15] F. Gao and X. Zhao (2011). Delta method for large deviations and moderate deviations for estimators. *The Annals of Statistics.* Vol. 39, pp. 1211-1240.

[16] G. Gürkan, A. Y. Özge and S. M. Robinson (1999). Sample-path solutions of stochastic variational inequalities. *Math. Program.* Vol. 84, pp. 313-334.

[17] T. Homem-de-Mello (2008). On rates of convergence for stochastic optimization problems under non-independent and identically distributed sampling. *SIAM J. Optim.* Vol. 19, pp. 524-551.

[18] Y. M. Kaniovski, A. J. King and R. J.-B. Wets (1995). Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems. Stochastic programming. *Ann. Oper. Res.* Vol. 56, pp. 189-208.

[19] A. Kleywegt, A. Shapiro and T. Homem-de-Mello (2002). The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* Vol. 12, pp. 479-502.

[20] L. A. Korf and R. J.-B. Wets (2001). Random lsc functions: An ergodic theorem. *Mathematics of Operations Research.* Vol. 26, pp. 421-445.

[21] S. P. Meyn and R. L. Tweedie (1993). *Markov chains and stochastic stability.* Communications and Control Engineering Series. Springer-Verlag London, Ltd., London. xvi+ 548 pp. ISBN: 3-540-19832-6.

[22] K. R. Parthasarathy (1967). *Probability Measure on Metric Spaces.* American Mathemati- cal Society, Providence, Rhode Island.

[23] E. Polak (1997). *Optimization. Algorithms and consistent approximations.* Applied Mathematical Sciences, 124. Springer-Verlag, New York.

[24] F. Rassoul-Agha and T. Seppäläinen (2015). *A corse on large deviations with an introduction to Gibbs measures.* Graduate Studies in Mathematics, 162. American Mathematical Society, Providence, RI.

[25] C. Robert and G. Casella (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer-Verlag.

[26] S. M. Robinson (1996). Analysis of sample-path optimization. *Math. Oper. Res.* Vol. 21, pp. 513-528.

[27] R. T. Rockafellar and R. J.-B Wets (1998). *Variational analysis.* Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 317. Springer-Verlag, Berlin.

[28] I. N. Sanov (1957). On the probability of large deviations of random variables. *Mat. Sbornik.* Vol. 42(84), No. 1, pp.11-44.

[29] A. Shapiro (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research.* Vol. 30, pp. 169–186.

[30] A. Shapiro (2006). Stochastic programming with equilibrium constraints. *Journal of Optimization Theory and Applications.* Vol. 128, pp. 223-243.

[31] A. Shapiro and T. Homem-de-Mello (2000). On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. *SIAM J. Optim.* Vol. 11, pp. 70-86.

[32] A. Shapiro, D. Dentcheva, A. Ruszczyński (2014). *Lectures on stochastic programming: modeling and theory.* MPS/SIAM Series on Optimization. 9. Philadelphia: Society for Industrial and Applied Mathematics.

[33] D. W. Stroock (1984). *An introduction to the theory of large deviations.* Universitext. Springer-Verlag, New York.

[34] A. Sur and J. R. Birge (2020). *Epi-convergence of sample averages of a random lower semi-continuous functional generated by a markov chain and application to stochastic optimization.* `http://www.optimization-online.org/DB_HTML/2020/04/7762.html`

[35] L. Tierney (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* Vol. 22, pp. 1701-1762.

[36] V. S. Varadarajan (1958). Weak convergence of measures on separable metric spaces, *Sankhyā.* Vol. 19, pp. 15-22.

[37] R. J.-B. Wets (1984). Modelling and solution strategies for unconstrained stochastic optimization problems. *Ann. Oper. Res.* Vol. 1, pp. 3-22.