Epi-convergence of Sample Averages of a Random Lower Semi-continuous Functional Generated by a Markov Chain and Application to Stochastic Optimization

Arnab Sur* and John R. Birge†

Abstract

The purpose of this article is to establish epigraphical convergence of the sample averages of a random lower semi-continuous functional associated with a Harris recurrent Markov chain with stationary distribution π. Sample averages associated with an ergodic Markov chain with stationary probability distribution will epigraphically converge from π-almost all starting points. The property of Harris recurrence allows us to replace "almost all" by "all", which is potentially important when running Markov chain Monte Carlo algorithms. That result on epi-convergence is then applied to establish the consistency of the optimal solutions and optimal value of a stochastic optimization problem involving expectation functional of the form $E_\pi[f(x,\xi)]$. Moreover, we develop asymptotic normality of the statistical estimator of the optimal value using a Markov chain central limit theorem.

Key Words: Sample average approximation method, Epigraphical convergence, Random lower semi-continuous function, Measure-preserving ergodic transformation, Harris recurrent Markov chain, Consistency, Asymptotic normality.

*The University of Chicago Booth School of Business, Chicago, IL 60637, USA.
email: arnabsur2002@gmail.com.

†The University of Chicago Booth School of Business, Chicago, IL 60637, USA.
email: jbirge@chicagobooth.edu.

This research is supported by the University of Chicago Booth School of Business.

1 Introduction

In this article we study the epigraphical convergence of the sample averages of a random lower semi-continuous (LSC) function $f : \mathbb{R}^n \times \Xi \to \overline{\mathbb{R}} = \mathbb{R} \cup \infty$, where the random argument is $\xi : (\Omega, \mathcal{A}, P) \to \Xi \subseteq \mathbb{R}^m$. The epigraphical convergence plays a pivotal role to establish the consistency of the sample average estimators of the following optimization problem (P):

$$\min_{x \in X \subset \mathbb{R}^n} \; E_\pi[f(x, \xi)], \tag{1.1}$$

where the underlying distribution π is partially known. We consider the following sample average approximation problem (SAA)

$$\min_{x \in X \subset \mathbb{R}^n} \quad \frac{1}{N} \sum_{k=1}^{N} f(x, \xi_k)\,, \tag{1.2}$$

where $\xi_1, ...\xi_N$ is typically an an independent and identically distributed (IID) sample drawn from π. This technique is known as the sample average approximation method or sample path optimization; see [32], for example. The literature on SAA is well developed. At times, to generate an IID sample becomes difficult, for example, in Bayesian statistics and the study of many stochastic models arising in mathematical physics and image processing; see [2], [3], [11], [30] and the references therein. In most cases in practice, the density of the distribution π is known up to a certain constant which is difficult to evaluate; in such a case, f is the density of π and $f = h/c$, where $c = \int h d\pi$ is not known explicitly, while the unnormalized density h is known. In these situations IID samples are not directly available; however, Markov chain Monte Carlo (MCMC) methods (see [12]) may be used to generate samples using dependent samples generated from a Markov chain whose invariant (stationary) distribution is π. Useful MCMC algorithms are Gibbs sampler, Metropolis-Hastings algorithm and their variants and extensions (see [29] and the references therein). This article focuses on defining explicit conditions under which the corresponding Markov chain Monte Carlo version of SAA (1.2) produces consistent solutions for the original problem (1.1) and asymptotic distributions defined by the Markov chain.

In particular, we study the epigraphical convergence of the sample averages of a random lower semi-continuous (LSC) function f associated with a Markov chain which is generated using an MCMC algorithm or otherwise. As epigraphical convergence is the fundamental tool to establish the consistency of the optimal solutions of the SAA problems, i.e., the sequence of the solutions of the SAA problem converges to the solution of the true problem (P) almost surely. The consistency of the optimal solutions is well known in the literature of stochastic optimization when the sample is IID and the convergence is established π-almost surely. On the contrary, a dependent sample drawn from a Markov chain with stationary distribution as π does not necessarily guarantee π-almost sure convergence. Rather the almost sure convergence will be with respect to the distribution of the entire chain, i.e., $Q_{\boldsymbol{\xi_1}}$, where Q is the transition kernel and $\boldsymbol{\xi_1}$ is the initial (starting) point of the chain, as it depends on the law of large numbers for stationary Markov chains.

Therefore, the starting point of the Markov chain is a critical consideration in the context of epigraphical convergence and consistency of optimal solutions.

To achieve the consistency of optimal solutions, we need to establish the epigraphical convergence of the SAA problems to the true problem (P) for any starting point of the Markov chain. The connection between epigraphical convergence and consistency of the SAA estimators has been studied extensively in the literature; they have appeared in [1], [7], [17], [36] and [37] and the references therein. Korf and Wets [23] developed an ergodic theorem on epi-convergence for random LSC function. All the aforementioned articles cover the IID sample; however, they do not include the Markov chain framework. Moreover, prior results in [23] provide the epigraphical convergence obtained from the ergodic theorem for stationary Markov chains of the random LSC function holds for π-almost all starting points, but, in general, the distribution of initial points need not intersect positive probability sets under π. For those results, there exists a π-null B from which we cannot guarantee epi-convergence if the chain starts from B. This creates an issue if the chain is initialized in this null set, for example, see [10], [31] and [35]. Since all initial points could be in B (i.e., $P(\boldsymbol{\xi_1} \in B) = 1$ as, for example, observed in [8]), such results are meaningless for practical situations where the stationary distribution is unknown.

To address this issue, all initial points should be included, i.e., we need to consider results with initial distributions that are not defined with respect to π. We use Harris recurrence for a Markov chain with stationary distribution to resolve this matter with a strong law of large numbers that holds for all starting points. The use of samples from a Harris recurrent Markov chain with stationary distribution has become a fundamental numerical tool. As background, the connection between Harris recurrence and MCMC algorithms is investigated in [10], [31] and [35]. As noted, results using IID samples to establish the convergence of SAA solutions to an optimal solution to (P) do not apply directly when samples are drawn from a Markov chain since, in particular, the initial state of the chain is inherently a singular event. This paper provides conditions for convergence (almost surely among process realizations) that apply for any initial state.

The results in this paper are related to Markov chain Monte Carlo (MCMC) methods for maximum likelihood estimation (MLE) (see [13] and [15]). In that case, the objective represents a log-likelihood such that the optimization problem has the form:

$$\max_{x \in X} \sum_{k=1}^{N} \log\left(\frac{h(x, \xi_k)}{h(z, \xi_k)}\right) - \log \int h(x, \xi) d\mu + \log \int h(z, \xi) d\mu, \tag{1.3}$$

where μ is a measure over Ξ and z is an arbitrary fixed parameter vector such that the non-negative functions, $h(x, \xi)$ and $h(z, \xi)$, have the same positive support regions. As noted above, in these cases, the difficulty is that the integrals are not analytically available but may be estimated with the use of MCMC to select ξ_k sequentially with a measure proportional to $h(z, \xi)$. The MCMC estimate of the integral terms in (1.3) is

then:

$$\log \sum_{k=1}^{N} \frac{h(x,\xi_k)}{h(z,\xi_k)}, \tag{1.4}$$

which provides an overall objective similar to that of (1.2). (An alternative form in which the integral term is estimated with a distinct sample from IID observations governing the likelihood appears in [26]. A similar second-level IID sampling procedure is also used in [14] for MLE in models with missing variables.) [15] establishes that, for exponential families (to ensure concavity of the objective), optimal solutions of the sample-average estimate of (1.3) with the empirical estimate from (1.4) achieve almost sure convergence to a solution of (1.3). [13] provides a generalization for any $h(x,\xi)$ that is lower semi-continuous for all x and all ξ except for a null set (possibly depending on x) under the probability measure implied by $h(z,\xi)$ and that is upper semi-continuous for all x, all ξ_k observations, and at all ξ except for a null set (not depending on x) under the probability measure implied by $h(z,\xi)$. We highlight the differences between these results and our result here in a following section.

In the next section, we assume a stationary Markov chain to establish the well-known consistency of the SAA estimators of the optimal solutions and optimal values applying the epigraphical convergence of the sample averages to the associated expectation functional. Section 3 extends these results to positive Harris recurrent Markov chains. Section 4 establishes asymptotic properties of the SAA estimators associated with given Markov chain properties.

2 Ergodic Theorem and Epigraphical Convergence for Almost All Starting Points

The purpose of this section is to establish epigraphical convergence of sample averages of random LSC functions assuming a stationary Markov chain. Let $\xi_1, \xi_2, \ldots$ be such a stationary Markov chain on $(\Xi, \mathcal{B})$ with transition probability function Q and let π be the stationary (invariant) distribution of the chain, i.e., $\pi Q = Q$. The Markov chain is stationary if the invariant probability measure π is its initial distribution. That is, π is the marginal distribution of ξ_n, for all n. Since π and Q determine the finite dimensional distributions of the Markov chain, the joint distribution of ξ_{n+1}, ξ_{n+2}, $\ldots \xi_{n+k}$ does not depend on n. $(\Xi^\infty, \mathcal{B}^\infty)$ is the associated canonical sample space and Q_π denotes the corresponding probability measure on $(\Xi^\infty, \mathcal{B}^\infty)$ given that the initial distribution of the Markov chain is π.

We next introduce the unilateral shift operator $T : \Xi^\infty \to \Xi^\infty$ on Ξ^∞, defined by $\omega = (\boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \ldots) \mapsto T(\omega) = (\boldsymbol{\xi_2}, \boldsymbol{\xi_3}, \ldots)$ for all $\omega \in \Xi^\infty$. Recall that the Borel sigma-algebra on $\mathbb{R}^n$ is the smallest sigma-algebra containing all open sets, also given as the smallest sigma-algebra making all the coordinate projections measurable. In analogy with this, $\mathcal{B}^\infty$ is the smallest sigma-algebra making all the coordinate projections measurable. The following family of sets $\{\omega \in \Xi^\infty : \boldsymbol{\xi_1} \in B_1, \boldsymbol{\xi_2} \in B_2, \ldots \boldsymbol{\xi_n} \in B_n\}$, where $n \geq 1$ and B_i's

belong to the generating family of $\mathcal{B}$, generate the sigma-algebra $\mathcal{B}^\infty$ which is stable under finite intersection. Therefore, to show the measurability of any transformation we need to analyze its behaviour over the finite dimensional sets (cylinders). Let us consider a measurable finite dimensional cylinder $B = B_1 \times B_2 \times \dots B_n \in \mathcal{B}^\infty$. Then,

$$\begin{aligned} T^{-1}(B) &= \{\omega \in \Xi^\infty : (T\omega)_1 \in B_1, (T\omega)_2 \in B_2, \dots (T\omega)_n \in B_n\} \\ &= \{\omega \in \Xi^\infty : \boldsymbol{\xi_2} \in B_1, \boldsymbol{\xi_3} \in B_2, \dots \boldsymbol{\xi_{n+1}} \in B_n\} \in \mathcal{B}^\infty \end{aligned}$$

where $(T\omega)_k$ is the k^{th} coordinate of $T\omega$. Hence, T is measurable. T is also measure-preserving, i.e., $Q_\pi(T^{-1}(B)) = Q_\pi(B)$ for all $B \in \mathcal{B}^\infty$. For all B that are invariant events, i.e., such that $T^{-1}(B) = B$, if $Q_\pi(B) \in \{0, 1\}$, then we say that T and the associated Markov chain are ergodic.

We are now ready to obtain the epigraphical convergence (law of large numbers) for random LSC functions using the ergodic theorem stated in [23]. For that result, we employ the definition of a random LSC function.

Definition 2.1. $f : X \times \Xi \to \overline{\mathbb{R}}$, where $X \subset \mathbb{R}^n$, is a random lower semi-continuous function at $\overline{x} \in X$, if the following conditions hold:
(i) the function $(x, \boldsymbol{\xi}) \mapsto f(x, \boldsymbol{\xi})$ is $(\mathcal{B}_X \times \mathcal{B})$-measurable, and
(ii) for every $\boldsymbol{\xi} \in \Xi$, the function $x \mapsto f(x, \boldsymbol{\xi})$ is LSC, i.e., for each $\epsilon > 0$, there exists $\delta(\boldsymbol{\xi}) > 0$ such that

$$f(x, \boldsymbol{\xi}) \geq f(\overline{x}, \boldsymbol{\xi}) - \epsilon \text{ for all } x \in B(\overline{x}, \delta(\boldsymbol{\xi})) \cap X,$$

where $B(\overline{x}, \delta(\boldsymbol{\xi}))$ is a ball around $\overline{x}$ with radius δ which depends on $\boldsymbol{\xi}$.

Theorem 2.1. Let $\{\xi_k : k = 1, 2, ...\}$ be an ergodic Markov chain with transition probability function Q on $(\Xi, \mathcal{B})$. Suppose that π is the stationary distribution of the chain. Let f be a random LSC function on $X \times \Xi$ in $x \in X$ such that $f \in L_1(\Xi, \mathcal{B}, \pi)$. Then, $\dfrac{1}{N}\displaystyle\sum_{k=1}^{N} f(\cdot, \xi_k)$ converges epigraphically on X to $E_\pi[f(\cdot, \xi)]$, Q_π-a.s.

Proof. Let us define $g : X \times \mathcal{B}^\infty \to \mathbb{R}$, by $g(x, \omega) = f(x, \xi_1(\omega))$. By definition g is measurable and LSC in x. Moreover, we recall the unilateral shift operator $T : \mathcal{B}^\infty \to \mathcal{B}^\infty$, which is ergodic and measure preserving. Using Theorem 8.2 in [23] we have $\dfrac{1}{N}\displaystyle\sum_{k=0}^{N} g(\cdot, T^k(\omega))$ converges epigraphically on X to $E_{Q_\pi}[g(\cdot, \omega)]$, Q_π-a.s., i.e., $\dfrac{1}{N}\displaystyle\sum_{k=1}^{N} f(\cdot, \xi_k)$ converges epigraphically on X to $E_\pi[f(\cdot, \xi)]$, Q_π-a.s. ∎

We transferred from the probability space $(\Xi, \mathcal{B}, \pi)$ to the probability space $(\Xi^\infty, \mathcal{B}^\infty, Q_\pi)$ and then used the shift operator defined on the representation space to obtain the result using the ergodic theorem for random LSC functions. Almost all materials related to the invariance and ergodicity can be formulated in terms of the Markov chain $\{\xi_k : k = 1, 2, ...\}$ rather than going into the representation space. First, we define the concept of irreducibility for a general state space Markov chain.

We say a set $A \in \mathcal{B}$ in the state space is ϕ-positive when $\phi(A) > 0$, where ϕ is an arbitrary measure on the state space. A non-negative kernel Q on the state space is called ϕ-irreducible if for every $\boldsymbol{\xi} \in \Xi$ and and a ϕ-positive set $A \in \mathcal{B}$, there exists a positive integer n such that $Q^n(\boldsymbol{\xi}, A) > 0$. A Markov chain is irreducible if its transition kernel is ϕ-irreducible for some ϕ. The definition seems quite arbitrary since ϕ is arbitrary. However, note that ϕ is only used to specify a family of null sets. Moreover, if a transition kernel is ϕ-irreducible for some ϕ, then there always exists a maximal irreducible measure Φ that specifies the minimal family of null sets, i.e., $\Phi(A) = 0$ implies $\phi(A) = 0$ for any irreducible measure ϕ.

Let Q be a ϕ-irreducible transition kernel. If a transition kernel is irreducible and has an invariant probability measure, then the invariant measure is unique up to multiplication by positive constant. Moreover, the invariant measure is a maximal irreducibility measure (see, [25]). The next theorem relates ergodicity to the irreducibility of a stationary Markov chain.

Theorem 2.2. Every irreducible stationary Markov chain is ergodic.

This result can be proved using Theorem 7.16 in [9] combined with the Proposition 2.3 in [27]. Moreover, ergodicity of the Markov chain guarantees the uniqueness of the invariant probability measure. The shift operator allows us to express stationarity and ergodicity in terms of the probability space $(\Xi^\infty, \mathcal{B}^\infty, Q_\pi)$ and then use the well known epigraphical ergodic theorems to derive the epigraphical law of large numbers. However, in practice it is convenient to be able to express stationarity and ergodicity in terms of the probability space $(\Xi, \mathcal{B}, \pi)$ rather than going into the representation space. Therefore, we can reproduce Theorem 2.1 if $\{\xi_k : k = 1, 2, ...\}$ is an irreducible stationary Markov chain.

The law of large numbers on epigraphical convergence stated in Theorem 2.1 says that the sample average associated with an irreducible stationary Markov chain with stationary (invariant) distribution π converges epigraphically, Q_π-a.s., to the expectation of f calculated with respect to the unique invariant distribution of the Markov chain.

Let B denote the subset of Ξ^∞ consisting of the points $\omega \in \Xi^\infty$ such that

$$B = \{\omega : \frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k(\omega)) \text{ converges epigraphically to } E_\pi[f(\cdot, \xi)]\}.$$

That is, B^c is a null set with respect to the measure Q_π for which the epiconvergence derived from the ergodic theorem in [23] fails. Let us write

$$C_y = \{\omega \in B : \boldsymbol{\xi_1} = y\}.$$

The above set contains all those sequences in B which start at y. The initial distribution and the marginal distributions are same as the stationary distribution π. Therefore, Fubini's theorem states that

$$Q_\pi(B) = \int_\Xi Q_y(C_y)\pi(dy),$$

and this is equal to one only if

$$Q_y(C_y) = \Pr\{(\xi_2, \xi_3, \dots) \in C_y | \xi_1 = y\} = 1, \quad \pi\text{-a.s.}$$

So, this gives us that the epigraphical law of large numbers in Theorem 2.1 derived from the ergodic theorem is conditional on the initial value. We see that it holds for all initial values except possibly a null set with respect to the invariant measure π. Therefore, the epigraphical convergence holds $Q_{\boldsymbol{\xi_1}}$-a.s. for all initial $\boldsymbol{\xi_1}$ except a π-null set. That is, for π-a.a. $\boldsymbol{\xi_1} \in \Xi$, $\frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k)$ converges epigraphically to $E_\pi[f(\cdot, \xi)]$, $Q_{\boldsymbol{\xi_1}}$-a.s.

Allowing for arbitrary starting points to avoid inherent issues from unknown stationary distributions (particularly in the context of learning models) is the motivation for Breiman [8]'s classic analysis of the Strong Law of Large Numbers for Markov chains. In the MCMC literature, Tierney [35] mentioned this exceptional null set of starting points from which convergence fails for ϕ-irreducible Markov chains as "nuisance" and Chan and Geyer [10] referred to this null set as "measure-theoretic pathology". However, the null set can arise quite naturally, even for MCMC algorithms. Example 4 in [31] affirms that the null set can indeed arise for general state space Markov chain and Example 9 in [31] demonstrates that a simple two-dimensional Metropolis-within-Gibbs algorithm with continuous target and proposal densities fails to converge to stationarity from all starting points. Hence, in practice this null set of points from which the convergence fails could cause problems for MCMC algorithms if the user happens to choose an initial point, to construct the Markov chain, in the null set as the convergence fails then; see [10], [31] and [35]. Thus, understanding the nature of the null set is important to realize the point of initialization of the Markov chain and to guarantee the convergence. However, since the use of a Markov chain approximation is generally because the underlying distribution π, which is the stationary distribution of the Markov chain, is not known, it is not practically possible to identify the π-null set and to verify sampling from points with positive probability with respect to the stationary distribution, i.e., to initialize a Markov chain outside of the π-null set. Therefore, to minimize a stochastic optimization problem with expectation functional, we would not be able to replace the expectation by its sample average associated to a stationary ergodic Markov chain if the chain starts from the null set, because the epigraphical convergence cannot be guaranteed for this instance.

This situation suggests, as emphasized in [8], that we require a stronger property to resolve this problem to include all starting points and to replace "π-a.a." by "all" starting points. In other words, we have to ensure epigraphical convergence for any starting point. For instance, in the Metropolis-Hastings algorithm, the restriction on the transition matrix Q was imposed in [35] to ensure that the Markov chain should enter the subset $E^+ = \{x : \pi(x) > 0\}$ of the sample space after atmost one step from any initial state, where π is the invariant distribution, i.e., the Markov chain should reach a π-positive state after at most one step from any initial point. Thus, the convergence of the Metropolis-Hastings algorithm has been established from all starting points. Similarly, we require more restriction (structure) on the transition matrix of the Markov chain to include all starting points. In the next section, we define the concept of Harris recurrence and then

use it to obtain the epigraphical convergence in Theorem 2.1 for all starting points.

3 Harris Recurrence and Epigraphical Convergence for All Starting Points

The ϕ-irreducibility of a Markov chain implies that $E_{\boldsymbol{\xi}}(\nu_A) = \sum_{n=1}^{\infty} Q^n(\boldsymbol{\xi}, A) > 0$, for all $\boldsymbol{\xi} \in \Xi$ whenever $\phi(A) > 0$, where ν_A denotes the number of visits of the chain to A. In other words, we have for any initial state $\boldsymbol{\xi} \in \Xi$, the expected number of visits of the Markov chain is positive whenever $\phi(A) > 0$. That is, the important sets A (i.e., $\phi(A) > 0$) are always reached with positive probability from every initial state $\boldsymbol{\xi} \in \Xi$. ϕ-irreducibility gives us epigraphical convergence of the sample averages to the expected value calculated with respect to the stationary distribution for almost all starting points (ref. Theorem 2.1). However, as we have noted above, we require the inclusion of all starting points to use MCMC algorithms. Harris recurrence provides the platform to include all starting points to establish the relevant law of large numbers. First, we define Harris recurrence and then use it to establish epigraphical convergence for all starting points.

A Markov chain is said to be Harris recurrent with respect to a sigma-finite measure ϕ if $A \in \mathcal{B}$, $\phi(A) > 0$ 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, Q is the transition probability function, and $Q_{\boldsymbol{\xi}}$ is the probability measure of the entire chain under Q given initial state $\boldsymbol{\xi}$, for references, see [4], [25], [24]. This implies that $Q_{\boldsymbol{\xi}}(\nu_A = \infty) = 1$ for all $\boldsymbol{\xi} \in \Xi$ whenever $\phi(A) > 0$; that is, from every initial state $\boldsymbol{\xi} \in \Xi$, the chain visits A infinitely many times $Q_{\boldsymbol{\xi}}$-a.s. Harris recurrence states that any important set A can be reached in finite steps with probability one from any initial state $\boldsymbol{\xi} \in \Xi$. Therefore, it is clear that Harris recurrence implies ϕ-irreducibility. The restriction on Q described in [35] for Metropolis-Hastings algorithm as described at the end of the previous section is a special case of Harris recurrence as it is natural to assume $\pi = \phi$. When such chains have an invariant probability measure π, they are then positive Harris recurrent chains. Hence, a stationary positive Harris recurrent chain is ergodic. Positive Harris recurrent chains enjoy the strongest properties. Harris recurrent chains are recurrent but the converse is not true. However, the difference between them is only by a ϕ-null set. In our discussion below, we use these three descriptors (positive, Harris, and recurrent) to emphasize the properties even though they are sometimes redundant in the contexts below.

Consider a sequence of functions $f_N : \mathbb{R}^n \to \overline{\mathbb{R}}$, a set $X \subseteq \mathbb{R}^n$ and a point $\overline{x} \in X$. Moreover, assume that $\{f_N, N = 1, 2....\}$ is *locally bounded* at $\overline{x}$ in the sense that, for some $\rho \in \mathbb{R}_+$ and a neighbourhood V of $\overline{x}$, $\mid f_N(\overline{x}) \mid \leq \rho$ for all $N \in \mathbb{N}$ and $x \in V$. Then, the family $\{f_N : N = 1, 2....\}$ is asymptotically equi-lower semi-continuous at $\overline{x}$ relative to X if, for every $\epsilon > 0$, there exists $\delta > 0$ and an index set $\mathbf{N} \in \mathcal{N}_\infty$ (all subsequences of $\mathbb{N}$ formed by eliminating a finite number of elements of $\mathbb{N}$) such that $f_N(x) \geq f_N(\overline{x}) - \epsilon$ for all $N \in \mathbf{N}$ when $x \in X$ and $\mid x - \overline{x} \mid \leq \delta$. If this property holds at every $\bar{x} \in X$, then the sequence $\{f_N\}$

is asymptotically equi-lower semi-continuous relative to X.

Definition 3.1. Let $f_N(\cdot,\cdot)$ has a second random argument defined on a probability space $(\Omega, \mathcal{A}, P)$, with $f_N : X \times \Xi \to \overline{\mathbb{R}}$, f_N is a random asymptotically equi-lower semi-continuous family relative to X π-almost surely if the property holds for $f_N(x, \boldsymbol{\xi})$ for all $\boldsymbol{\xi}$ except a π-null set, i.e., for any $\bar{x} \in X$ and $\epsilon > 0$, there exists $\delta(\boldsymbol{\xi}) > 0$ and an index set $\mathbf{N}(\boldsymbol{\xi}) \in \mathcal{N}_\infty$ such that $f_N(x, \boldsymbol{\xi}) \geq f_N(\overline{x}, \boldsymbol{\xi}) - \epsilon$ for all $N \in \mathbf{N}(\boldsymbol{\xi})$ when $x \in X$ and $\mid x - \overline{x} \mid \leq \delta(\boldsymbol{\xi})$, for all $\boldsymbol{\xi}$ except a π-null set. The sequence $\{f_N\}$ is equi-lower semi-continuous relative to X if the inequality condition holds for all N.

Next, we introduce pathwise independence of a random LSC function which we require to prove the epigraphical convergence. First, we recall the Definition 2.1 to define the following.

Definition 3.2. A random LSC function is pathwise independent at $\overline{x}$ with respect to a sample path $\{\boldsymbol{\xi_k} : k = 1, 2, \ldots\}$, a countable collection of realizations of the random variable, if, for $B(\overline{x}, \delta(\boldsymbol{\xi_k}))$ defined as in Definition 2.1 for a given ϵ:

$$\bigcap_k B(\overline{x}, \delta(\boldsymbol{\xi_k})) \setminus \{\overline{x}\} \neq \emptyset, \quad \text{i.e.,} \;\; \delta = \inf_k \delta(\boldsymbol{\xi_k}) > 0.$$

The above property asserts that the radius δ of the ball B is independent with respect to the realizations along a sample path, i.e., particular value of δ will work for each realization along the sample path, although different sample paths may have different values of δ.

We present a discussion on comparative relationship of pathwise independence with the existing assumptions in the literature. In [1], the authors established the epigraphical convergence of the sample averages of a random lower semi-continuous function corresponding to an IID sample. To prove the result, they assumed the following: for each $x_0 \in X$, there exists an open set N_0 in $\mathbb{R}^n$ and an integrable function $\alpha_0(\cdot) : \Xi \to (-\infty, \infty)$ such that $x_0 \in N_0$ and for almost all $\boldsymbol{\xi} \in \Xi$ the inequality

$$f(x, \boldsymbol{\xi}) \geq \alpha_0(\boldsymbol{\xi}) \tag{3.1}$$

holds for all $x \in N_0$. Zervos in [37] considered the stochastic optimization problem which consists of the following performance criterion:

$$r(x) := \int f(x, \xi)\pi(d\xi).$$

The above problem is approximated by the following sequence of performance criteria:

$$r_n(x, \omega) := \int f(x, \xi)\pi_n(d\xi)(\omega),$$

where $\pi_n(\omega)$ converges weakly to π a.s. (π is the distribution of the random variable). The epigraphical convergence of the sequence $\{r_n(\cdot,\cdot)\}$ to $r(\cdot)$ has been established for π-almost all $\boldsymbol{\xi} \in \Xi$. The result obtained in [1] can be viewed as a special case where μ_n's are empirical distributions associated with an IID. Zervos

established the result for a more general decision space and state space. The minor difference in assumption is that the integrability condition (3.1) holds for π-almost all $\boldsymbol{\xi}$ in [1], whereas it holds for all $\boldsymbol{\xi}$ in [37]. In the literature, the condition (3.1) or more restrictive assumptions have been imposed in all references to obtain epigraphical convergence, see [1], [5], [17], [21] and [37]. The authors in [1] stated that (3.1) is the most relaxed assumption (among all the assumptions available in the literature) under which the epigraphical convergence has been established and Zervos [37] followed that to establish epigraphical convergence for more general spaces than $\mathbb{R}^n$.

For $x_0 \in X$, if we fix $\epsilon > 0$ and set $\alpha_0(\boldsymbol{\xi}) := f(x_0, \boldsymbol{\xi}) - \epsilon$, then Definitions 2.1 and 3.2 give us a neighborhood $N_0 = B(x_0, \delta)$ around x_0 which ensures the integrability property (3.1) for the sequence $\{\boldsymbol{\xi_k} : k = 1, 2, \ldots\}$, where $\delta = \inf_k \delta(\boldsymbol{\xi_k})$. Hence, pathwise independence of LSC functions and the assumption (3.1) are equivalent for a sequence of realizations, i.e., for a sample path. Pathwise independence confirms a neighborhood B around x_0 like assumption (3.1) for each sample path. Moreover, pathwise independence further relax the assumption (3.1) as the neighborhood around x_0 may differ for different sample paths (i.e., δ may be different) in case of pathwise independence, whereas the neighbourhood remains same for π-almost all $\boldsymbol{\xi}$ in [1]. In our framework, we have partial information about π, hence, it would be difficult to validate any condition which holds π-a.s. We next prove the epigraphical convergence under the assumption of pathwise independence.

Theorem 3.1. Suppose (i) $\Xi \subseteq \mathbb{R}^m$ is the support set and $f : \mathbb{R}^n \times \Xi \to \overline{\mathbb{R}}$ is a lower semi-continuous function in x for each $\boldsymbol{\xi} \in \Xi$, where $f(\cdot, \xi)$ is locally bounded for every $x \in X$, (ii) $E_\pi[f(\cdot, \xi)]$ is lower semi-continuous, and (iii) $\{\xi_k : k = 1, 2, ...\}$ is a positive Harris recurrent Markov chain on Ξ as defined above and $f(\cdot, \cdot)$ is pathwise independent for every $x \in X$ almost surely with respect to $Q_{\boldsymbol{\xi_1}}$, for any $\boldsymbol{\xi_1} \in \Xi$. Then, $\{\frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k) : N = 1, 2....\}$ is a random equi-lower semi-continuous family relative to X almost surely with respect to $Q_{\boldsymbol{\xi_1}}$ and $\frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k)$ converges epigraphically on X to $E_\pi[f(\cdot, \xi)]$, $Q_{\boldsymbol{\xi_1}}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$.

Proof. Given $f : \mathbb{R}^n \times \Xi \to \overline{\mathbb{R}}$ is a lower semi-continuous function, let us choose an arbitrary and fixed $\overline{x} \in \mathbb{R}^n$; then, by definition of lower semi-continuity, we have, for each $\epsilon > 0$, there exists $\delta > 0$ such that

$$f(x, \boldsymbol{\xi}) \geq f(\overline{x}, \boldsymbol{\xi}) - \epsilon \text{ for all } x \in B(\overline{x}, \delta(\boldsymbol{\xi})) \cap X,$$

where $B(\overline{x}, \delta(\boldsymbol{\xi}))$ is a ball around $\overline{x}$ with radius δ which depends on $\boldsymbol{\xi}$. Assumption (ii) says that the random sample $\{\xi_k : k = 1, 2.....\}$ is a Harris recurrent Markov chain with state space $(\Xi, \mathcal{B})$ and stationary distribution π. For each k, we have that

$$f(x, \boldsymbol{\xi_k}) \geq f(\overline{x}, \boldsymbol{\xi_k}) - \epsilon \text{ for all } x \in B(\overline{x}, \delta(\boldsymbol{\xi_k})) \cap X.$$

Hence, we obtain for each $N \in \mathbb{N}$,

$$\frac{1}{N}\sum_{k=1}^{N} f(x, \boldsymbol{\xi_k}) \geq \frac{1}{N}\sum_{i=1}^{N} f(\overline{x}, \boldsymbol{\xi_k}) - \epsilon \text{ for all } x \in B(\overline{x}, \delta_N) \cap X,$$

where δ_N is the minimum value of $\delta(\boldsymbol{\xi}_1)$, $\delta(\boldsymbol{\xi}_2), \ldots$ $\delta(\boldsymbol{\xi}_N)$. This implies, by the virtue of pathwise independence that the sequence $\{\frac{1}{N}\sum_{k=1}^{N} f(., \xi_k) : N = 1, 2, \ldots\}$ is equi-lower semi-continuous at $\overline{x}$, almost surely with respect to $Q_{\boldsymbol{\xi_1}}$ for any $\boldsymbol{\xi_1} \in \Xi$, i.e., an equi-lower semi-continuous family $Q_{\boldsymbol{\xi_1}}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$.

Moreover, due to Assumption (ii) and the strong law of large numbers for positive Harris recurrent Markov chains with finite function values (for example, see [4] and [2]), we obtain that $\frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k)$ converges pointwise to $E_\pi[f(\cdot, \xi)]$, $Q_{\boldsymbol{\xi_1}}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$. We then have that $\frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k)$ converges pointwise to $E_\pi[f(\cdot, \xi)]$ and $\{\frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k) : N = 1, 2....\}$ is random equi-lower semi-continuous at $\overline{x}$, $Q_{\boldsymbol{\xi_1}}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$. Applying Theorem 7.10 in [33] we can conclude that $\frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k)$ converges epigraphically to $E_\pi[f(\cdot, \xi)]$ at $\overline{x}$, $Q_{\boldsymbol{\xi_1}}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$. This completes the proof as $\overline{x}$ was chosen arbitrarily. ∎

As observed above, the authors require assumption (3.1) in [1] to hold π-almost surely to establish the epigraphical convergence. On the contrary, we assumed pathwise independence to obtain the result corresponding to a Markov chain, which may be more straightforward to validate. Moreover, in the Markov chain case, the epigraphical convergence is obtained $Q_{\boldsymbol{\xi_1}}$-a.s. Therefore, we do not need a condition which holds for π-almost all $\boldsymbol{\xi}$. Nevertheless, we need an assumption to hold for $Q_{\boldsymbol{\xi_1}}$-almost all sequences. Pathwise independence is the minimal (as discussed before Theorem 3.1) sufficient condition to establish epigraphical convergence in case of Markovian dependence among the data.

We note that the epigraphical convergence obtained in Theorem 3.1 cannot be viewed as a special case of Theorem 1 or Theorem 2 in [37]. The epigraphical convergence in [37] holds for P-a.s., i.e., with respect to the probability measure associated with the random variable, whereas the convergence in Theorem 3.1 holds for $Q_{\boldsymbol{\xi_1}}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$, i.e., with respect to the distribution of the entire chain. Moreover, the result introduces a new methodological approach using pathwise independence to establish epigraphical convergence.

We already proved the almost sure epigraphical convergence of a sequence of sample average of LSC functions to the function's expectation with respect to the stationary probability measure in Theorem 2.1 assuming an ergodic measure-preserving transformation to generate samples. While this result is similar to that of Theorem 3.1, however, the convergence in Theorem 3.1 applies for any initial point $\boldsymbol{\xi}_1$ while the result (Theorem 2.1) obtained from [23] does not apply to cases in which epi-convergence does not occur from the

initial point, which may result either from an infinite function value or weaker chain properties than those of positive Harris chains. The result in Theorem 3.1 is particularly relevant for Markov chain sampling since initial points in Markov chains are inherently singular events and the stationary distribution is not known a priori, but the properties for Harris recurrence can be confirmed (as demonstrated, for example, for typical Markov Chain Monte Carlo implementations in [2] and the general conditions in [16]).

The result in Theorem 3.1 is also similar to the result in [13] as noted in the introduction. That result requires both upper semi-continuity of $h(x, \xi_i)$ (equivalent to lower semi-continuity of $f(x, \xi)$ in the minimization context here) and lower semi-continuity (except on a possibly x-dependent null set) because of the opposite sign on the integration constant term in (1.4). The conditions in Theorem 3.1 only require lower semi-continuity (and local boundedness) since the additional term with negative coefficient does not appear in (P). The MLE result in [13] also avoids difficulties in which pointwise- and epi-convergence may not agree by the assumptions of non-negative densities, continuity, and integrability (in the case of missing data models). The local boundedness and equi-lower semi-continuity assumptions in Theorem 3.1 replace those assumptions here. The result in [13] also assumes that the observations are generated by an irreducible Metropolis-Hastings algorithm, which is generalized in Theorem 3.1 to Harris recurrent Markov chains. In particular, the result in Theorem 3.1 does not require initialization on the support of the stationary distribution, which, as we have emphasized, in many cases in practice may not be known.

Definition 3.3. Consider a sequence of functions $f_N : \mathbb{R}^n \to \overline{R}$, a set $X \subseteq \mathbb{R}^n$ and a point $\overline{x} \in X$. Moreover, assume that $\{f_N : N = 1, 2....\}$ is locally bounded at $\overline{x}$. Then the family $\{f_N : N = 1, 2....\}$ is equi-upper semi-continuous at $\overline{x}$ relative to X if for every $\epsilon > 0$ there exists $\delta > 0$ such that $f_N(x) \leq f_N(\overline{x}) + \epsilon$ for all $N \in \mathbb{N}$, when $x \in X$ and $\mid x - \overline{x} \mid \leq \delta$. We define asymptotically equi-upper semi-continuous at $\bar{x}$ relative to X, and random asymptotically equi-upper semicontinuous at $\bar{x}$ relative to X almost surely with respect to a measure π analogously to the definitions for LSC functions.

The following lemma on upper-semicontinuity follows directly from the proof of Theorem 3.1.

Lemma 3.1. Suppose that (i) $\Xi \subseteq \mathbb{R}^m$ is the support set and $f : \mathbb{R}^n \times \Xi \to \overline{\mathbb{R}}$ defines an upper semi-continuous function in x for each $\boldsymbol{\xi} \in \Xi$, where $f(\cdot, \xi)$ is locally bounded for every $x \in X$, (ii) $E_\pi[f(\cdot, \xi)]$ is upper semi-continuous, and $iii)$ $\{\xi_k : k = 1, 2, ...\}$ is a positive Harris recurrent Markov chain on Ξ as defined above and $f(\cdot, \cdot)$ is pathwise independent for every $x \in X$ and almost surely with respect to $Q_{\boldsymbol{\xi_1}}$, for any $\boldsymbol{\xi_1} \in \Xi$. Then $\{\frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k) : N = 1, 2....\}$ is a random equi-upper semi-continuous family relative to X almost surely with respect to $Q_{\boldsymbol{\xi}_1}$, for any $\boldsymbol{\xi_1} \in \Xi$.

The next theorem establishes the continuous convergence of the sequence of sample averages. The result is based on Theorem 3.1 and Lemma 3.1

Theorem 3.2. Suppose that (i) $\Xi \subseteq \mathbb{R}^m$ is the support set and $f : \mathbb{R}^n \times \Xi \to \overline{\mathbb{R}}$ is continuous in x, and

$E_\pi[f(\cdot,\xi)]$ is locally bounded for every $x \in X \subset \mathbb{R}^n$, and (ii) that $\{\xi_k : k = 1, 2.....\}$ is a positive Harris recurrent Markov chain with state space $(\Xi, \mathfrak{B})$, transition probability $Q(\cdot,\cdot)$, and stationary distribution π. Moreover, $f(\cdot,\cdot)$ is pathwise independent for every $x \in X$ and almost surely with respect to $Q_{\boldsymbol{\xi_1}}$, for any $\boldsymbol{\xi_1} \in \Xi$. Then $\dfrac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k)$ converges continuously to $E_\pi[f(\cdot,\xi)]$ relative to X, $Q_{\boldsymbol{\xi}_1}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$.

Proof. $\{\dfrac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k) : N = 1, 2....\}$ is a random equi-semi-continuous family, which follows from Theorem 3.1 and Lemma 3.1 by noting that local boundedness of the integral and continuity of the integrand imply continuity of $E_\pi[f(\cdot,\xi)]$. Further applying Theorem 3.1 with Theorem 7.11 in [33], we have the desired result. ∎

Corollary 3.1. Suppose all the assumptions of Theorem 3.2 hold, then $E_\pi[f(\cdot,\xi)]$ is continuous and $\dfrac{1}{N}\sum_{k=1}^{N} f(\cdot,\xi_k)$ converges uniformly to $E_\pi[f(\cdot,\xi)]$, $Q_{\boldsymbol{\xi}_1}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$, on all compact subsets of X.

Proof. The result follows immediately from Theorem 3.2 by applying Theorem 7.14 in [33]. ∎

The result in Corollary 3.1 is actually the statement of the uniform law of large numbers when the sample is a Harris recurrent Markov chain with stationary distribution. In this paper, we present an approach to achieve the result using Theorem 3.1.

4 Application in Stochastic Optimization

4.1 Consistency of Minimizers

As an application of Theorem 3.1 in stochastic optimization, we consider the stochastic optimization problem (P). Sample average approximation is an 'exterior' approach where an IID sample is generated from the underlying distribution π and then the SAA subproblems are solved by an approximate deterministic algorithm. As mentioned in the introduction, generation of an IID sample is difficult (or impossible) in general. However, one can often use MCMC algorithms to generate a stationary Markov chain even if we have limited information about the underlying distribution. We can formulate SAA subproblems using a Harris recurrent Markov chain with π as a stationary distribution, instead of IID samples. Epigraphical convergence of the SAA subproblems play the central role to establish the consistency of the statistical estimators, see [1]. Therefore, Theorem 3.1 will be applied to establish the consistency of the minimizers of (P) when the sample is a Markov chain. The theorem on consistency is the following:

Theorem 4.1. For optimization problem (P) and the corresponding sample average approximation problems of type (SAA) with f satisfying the assumptions of Theorem 3.1, where $\{\xi_k : k = 1, 2, \ldots\}$, forms a Harris

recurrent Markov chain with stationary distribution π; then:
i) if $\{\tilde{x}_N : N = 1, 2 \ldots\}$ is a sequence of global minimizers of sample average approximation problems (SAA) for given $\boldsymbol{\xi}_1 \in \Xi$, i.e., $\tilde{x}_N$ is a global minimizer of the problem (SAA), for each $N \in \mathbb{N}$, and $\tilde{x}$ is an accumulation point of this sequence, $Q_{\boldsymbol{\xi}_1}$-a.s., then $\tilde{x}$ is a global minimizer of problem (P) and $\frac{1}{N}\sum_{k=1}^{N} f(\tilde{x}_N, \xi_k)$ converges to $E_\pi[f(\tilde{x}, \xi)]$, $Q_{\boldsymbol{\xi}_1}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$;
ii) if $\{\tilde{x}_N : N = 1, 2.....\}$ is a sequence of local minimizers of the sample average approximation problems (SAA) for given $\boldsymbol{\xi}_1 \in \Xi$ sharing a common radius of attraction $\rho > 0$, $Q_{\boldsymbol{\xi}_1}$-a.s. (i.e., for each $N \in \mathbb{N}$, $\frac{1}{N}\sum_{k=1}^{N} f(\tilde{x}_N, \xi_k) \leq \frac{1}{N}\sum_{k=1}^{N} f(x, \xi_k)$ for all $x \in X$ such that $\| \tilde{x}_N - x \| < \rho$, $Q_{\boldsymbol{\xi}_1}$-a.s.) and $\tilde{x}$ is an accumulation point of this sequence, $Q_{\boldsymbol{\xi}_1}$-a.s., then $\tilde{x}$ is a local minimizer of the problem (P) and $\frac{1}{N}\sum_{k=1}^{N} f(\tilde{x}_N, \xi_k)$ converges to $E_\pi[f(\tilde{x}, \xi)]$, $Q_{\boldsymbol{\xi}_1}$-a.s. for any $\boldsymbol{\xi_1} \in \Xi$.

Proof. Applying Theorem 3.1 we have $\frac{1}{N}\sum_{k=1}^{N} f(\cdot, \xi_k)$ epi-converges to $E_\pi[f(\cdot, \xi)]$, $Q_{\boldsymbol{\xi}_1}$-a.s. Hence, using the functional definition of epi-convergence, we can conclude that for every infinite sequence $\{x_N : N = 1, 2....\}$ such that $x_N \in X$ and $x_N \to x$, as $N \to \infty$; $\liminf_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(x_N, \xi_k) \geq E_\pi[f(x, \xi)]$, $Q_{\boldsymbol{\xi}_1}$-a.s. Moreover, using the strong law of large numbers for positive Harris recurrent Markov chains with finite function values, we know that, for every $x \in X$, there exists a sequence $\{x_N = x : N = 1, 2....\}$, such that x_N converges to x and $\limsup_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(x_N, \xi_k) \leq E_\pi[f(x, \xi)]$, $Q_{\boldsymbol{\xi}_1}$-a.s. Thus, Theorem 3.3.2 in [28] implies that the epigraphs of the sample average approximation problems of type (SAA) converge to the epigraph of the optimization problem (P) almost surely. Then using Theorem 3.3.3 in [28], we conclude with the above result. ∎

Now consider the following equality and inequality constrained stochastic optimization problem:

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & E_\pi[f(x, \xi)] \\ \text{subject to} & E_\pi[G(x, \xi)] = 0, \\ & E_\pi[H(x, \xi)] \leq 0; \end{array} \tag{4.1}$$

where ξ is a random variable as before with support $\Xi \subseteq \mathbb{R}^m$, $f : \mathbb{R}^n \times \Xi \to \mathbb{R}$ defines a convex function $f(\cdot, \xi)$ for any $\xi \in \Xi$ 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 in x for all $\xi \in \Xi$ with finite expectation functionals.

We can proceed in a similar fashion to solve the above stochastic optimization problem through the sample average approximation method. Applying Theorem 3.1, one can again have the conclusions of Theorem 4.1 for the constrained optimization problem under one more additional assumption 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 the feasible set of the optimization problem (4.1) and X_N, $N = 1, 2.....$ are feasible sets of the corresponding sample average approximation problems.

4.2 Asymptotic Normality

In the previous section we derived the consistency of SAA estimators of optimal points of the problem (P) when the sample is a stationary Markov chain. The main goal of this section is to describe conditions on Ξ and f under which a Central Limit Theorem type result holds to assess the error of such an approximation of the optimal points. First, we shall define the drift condition. The book of Meyn and Tweedie [25] made the use of this computational methodology popular as it helps one to deal with complicated situations that arise in MCMC. The drift operator Δ is defined for any non-negative measurable function V on Ξ by

$$\Delta V(\boldsymbol{\xi}) := \int V(\boldsymbol{\zeta})Q(\boldsymbol{\xi}, d\boldsymbol{\zeta}) - V(\boldsymbol{\xi}), \;\; \boldsymbol{\xi} \in \Xi.$$

To define the drift conditions, first we need to define the concept of small set. The theory of small sets is discussed in detail in the book of Meyn and Tweedie [25].

Definition 4.1. A set $C \in \mathfrak{B}(\Xi)$ is called a small set if there exists a non-trivial probability measure μ on $\mathfrak{B}(\Xi)$ and a positive integer m, such that for all $x \in C$ and $B \in \mathfrak{B}(\Xi)$,

$$Q^m(x, B) \geq \mu(B). \tag{4.2}$$

When this condition holds, we say that C is μ-small.

The following result on asymptotic normality can be obtained at every $x \in X$ using the Markov chain central limit theorem.

Theorem 4.2. Consider the stochastic optimization problem (P) and the corresponding sample average approximation problems of type (SAA), for each $N \in \mathbb{N}$ where $\{\xi_k : k = 1, 2, \ldots\}$, forms a Harris recurrent Markov chain with stationary distribution π. Assume that one of the following drift conditions holds at any x:

1. suppose that for a function $V : \Xi \to [1, \infty)$ there exist constants $d > 0$, $b < \infty$ such that

$$\Delta V(\boldsymbol{\xi}) \leq -dV(\boldsymbol{\xi}) + bI(\boldsymbol{\xi} \in C), \;\; \boldsymbol{\xi} \in \Xi;$$

where C is a small set and I is the indicator function, holds and $f^2(x, \boldsymbol{\xi}) \leq V(\boldsymbol{\xi})$, for all $\boldsymbol{\xi} \in \Xi$;

2. suppose that for a function $V : \Xi \to [1, \infty)$ there exist constants $d > 0$, $b < \infty$ and $0 \leq r < 1$ such that

$$\Delta V(\boldsymbol{\xi}) \leq -d[V(\boldsymbol{\xi})]^r + bI(\boldsymbol{\xi} \in C), \;\; \boldsymbol{\xi} \in \Xi;$$

where C is a small set and I is the indicator function, holds and $\| \; f(x, \boldsymbol{\xi}) \; \| \leq [V(\boldsymbol{\xi})]^{r+\eta-1}$, for all $\boldsymbol{\xi} \in \Xi$ where $1 - r \leq \eta < 1$ is such that $E_\pi[V^2\eta] < \infty$. Then for any x, $\sigma_x^2 \in [0, \infty)$ and, if $\sigma_x^2 > 0$, then for any initial state $\boldsymbol{\xi_1} \in \Xi$:

$$\lim_{N\to\infty} Q_{\boldsymbol{\xi_1}}\{\frac{1}{\sqrt{N}\sigma_x}(\frac{1}{N}\sum_{k=1}^{N} f(x, \xi_k) - E_\pi[f(x, \xi)]) \leq t\} = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} d\xi \, ;$$

where $\sigma_x^2 := var_\pi\{f(x,\xi_1)\}+2\sum_{i=2}^{\infty} cov_\pi\{f(x,\xi_1), f(x,\xi_i)\}$, and var_π and cov_π denote variance and covariance respectively with respect to Q_π.

Remark 1. The above theorem can be applied to assess the error at any optimal point of (P) . Proof of the first part can be found in Chapter 17 of [25] and the second part is proved in [18]. The rates of convergence in the CLT at x can be calculated for these drift conditions, see [18] and [22]. A detailed discussion on CLT, drift conditions and various ergodicity of Markov chain can be found in [19]. We can apply Theorem 4.2 at $\tilde{x}$, an optimal value of (P), to obtain convergence rates of the solution values at any optimal point $\tilde{x}$ for (P).

We can also obtain a result on convergence of the finite-sample minimum values in (SAA) to the optimal value of (P). Let $\tilde{z}$ be the optimal value in (P) and $\tilde{X} = \{x|E_f(x) = \tilde{z}\}$, the optimal solution set in (P), and let z^N be the (random) optimal value of (SAA). We further define $Z(x) = \mathcal{N}(0, \sigma_x^2)$, i.e., a normally distributed random function on X with variance given by σ_x^2. The following result then follows from Theorem 4.2 and Theorem 3.2 in [34]. As a preliminary step, we define a continuous approximation of lower semi-continuous as follows:

$$f_\epsilon(x,\xi) = \min_{y\in B(x,\epsilon)\cap X} f(y,\xi) + \|x-y\|.$$

and let

$$M_\epsilon(\xi) = \max_{x\in X} f_\epsilon(x,\xi) - \min_{x\in X} f_\epsilon(x,\xi).$$

Lemma 4.1. For compact X, $\epsilon > 0$, $\epsilon < M_\epsilon(\xi) < \infty$, $f_\epsilon(x,\xi)$ is Lipschitz continuous on X with Lipschitz constant $\frac{M(\epsilon)}{\epsilon}$ on X.

Proof. Consider $x_1, x_2 \in X$. First, if $\|x_1 - x_2\| \geq \epsilon$, then:

$$|f_\epsilon(x_1,\xi) - f_\epsilon(x_2,\xi)| \leq M_\epsilon(\xi) \leq \frac{M_\epsilon(\xi)}{\epsilon}\|x_1 - x_2\|,$$

which establishes the result under this condition.

For $\|x_1 - x_2\| \leq \epsilon$, suppose without loss of generality that $f_\epsilon(x_2,\xi) \leq f_\epsilon(x_1,\xi)$. Let $f(y_2,\xi) = f_\epsilon(x_2,\xi)$ (where by the definition $y_2 \in X$). Note that if $y_2 \in B(x_1,\epsilon)$, then $|f_\epsilon(x_1,\xi) - f_\epsilon(x_2,\xi)| \leq |\|x_1 - y_2\| - \|x_2 - y_2\|| \leq \|x_1 - x_2\|$, which establishes the condition in this case since $M_\epsilon(\xi) \geq \epsilon$. For the remaining condition (i.e., $y_2 \notin B(x_1,\epsilon)$), note that $f_\epsilon(x_2,\xi) = f(y_2) + \|x_2 - y_2\|$ and $\|x_1 - y_2\| > \epsilon \geq \|x_1 - x_2\|$. From these conditions,

$$\begin{aligned}\frac{|f_\epsilon(x_1,\xi) - f_\epsilon(x_2,\xi)|}{\|x_1 - x_2\|} &= \frac{f_\epsilon(x_1,\xi) - f_\epsilon(x_2,\xi)}{\|x_1 - x_2\|} \\ &= \frac{f_\epsilon(x_1,\xi) - f(y_2,\xi) - \|x_2 - y_2\|}{\|x_1 - x_2\|} \leq \frac{f_\epsilon(x_1,\xi) - f(y_2,\xi) - \|x_2 - y_2\|}{\|x_1 - y_2\|} \\ &\leq \frac{f_\epsilon(x_1,\xi) - f(y_2,\xi)}{\|x_1 - y_2\|} = \frac{|f_\epsilon(x_1,\xi) - f_\epsilon(y_2,\xi)|}{\|x_1 - y_2\|} \leq \frac{M_\epsilon(\xi)}{\epsilon},\end{aligned}$$

as shown earlier for any x_1, y_2 with $\|x_1 - y_2\| \geq \epsilon$. This then establishes the Lipschitz continuity with constant $\frac{M_\epsilon(\xi)}{\epsilon}$. ∎

This approximation allows the use of Theorem 4.2 and Theorem 3.2 in [34] for the following.

Theorem 4.3. Suppose the conditions of Theorem 4.2 and Lemma 4.1, and that $\mathbb{E}_\pi[(M_\epsilon(\xi))^2] < \infty$ for all $\epsilon > 0$; then,

$$(z^N - \tilde{z})N^{1/2} \to \min_{x \in \bar{X}} Z(x) \; Q_{\boldsymbol{\xi_1}}\text{-a.s., for all starting points } \boldsymbol{\xi_1}. \tag{4.3}$$

Proof. Theorem 3.2 of [34] states that the minimum of a scaled sample average of continuous functions that converges to a continuous function on a compact set converges to the minimum of the distributional limits over the optimal solution set (i.e., the conclusion in (4.3) for a random continuous function Z). To invoke this result, we use that the sequence of continuous approximations $f_\epsilon(x, \xi_k)$ from a Markov chain $\{\xi_k, k = 1, 2, \ldots\}$ satisfying the properties in Theorem 4.3, which, with the scaling of $N^{1/2}$, of the difference between the expectation $E_\pi[f_\epsilon(x, \xi)]$ and the sample averages of $f_\epsilon(x, \xi_k)$, $k = 1, \ldots, N$, along the Markov chain with the properties of Theorem 4.3 and the expectation $E_\pi[f_\epsilon(x, \xi)]$, has the property of converging in distribution to $(Z(x))$ at any x. We first need to extend this pointwise convergence to convergence in distribution for a continuous function on x.

For the continuous functions, from Lemma 4.1, $f_\epsilon(x, \xi)$ is Lipschitz-continuous with constant $M_\epsilon(\xi)/\epsilon$ on X. This and the assumption then implies that f_ϵ has the $\mathcal{L}_2$-Lipschitz property assumed in [34] and, for example, in Proposition A5 in [20],which establishes a functional Central Limit Theorem in the IID setting. To see how to obtain similar results in this context, we note that convergence of the sample averages of a sequence of random continuous functions to a random continuous function over a compact set follows, for example, from Theorem 7.5 in [6], which requires weak convergence of random continuous function values F_N at any vector of function values $x_1, \ldots, x_k$ to a random continuous function F at those points, and a bound on local deviations such that $\lim_{\delta \to 0} \lim_{N \to \infty} P\{\omega | \sup_{\|x-y\|<\delta} |F_N(x, \omega) - F_N(y, \omega)| \geq \epsilon\} = 0$ for any $\epsilon > 0$ where P is the distribution over random elements ω. For the first condition, Theorem 4.3 implies such convergence for a single point. The Cramér-Wold Theorem then extends this result to vectors $x_1, \ldots, x_k$ (by applying the scalar result to any linear combination of the random function values). The second condition is satisfied by the Lipschitz property of $f_\epsilon(x, \xi)$.

The result in [34] also requires that $E_\pi f(x, \xi)^2 < \infty$, which follows from the drift condition assumptions and applies equally to $f_\epsilon(x, \xi)$ from $f_\epsilon(x, \xi) \leq f(x, \xi)$. Also, note that $f_\epsilon(x, \xi) \geq \min_{x \in X} f(x, \epsilon)$. Theorem 3.2 of [34] then applies to the random selections $f_\epsilon(x, \xi_k), k = 1, 2, \ldots$ and the result in Theorem 4.2 applies for convergence at any $x \in X$ to obtain that $N^{1/2}(\sum_{i=1}^{N} f_\epsilon(x, \xi_i)/N - E_\pi[f_\epsilon(x, \xi)])$ converges in distribution to the random function $Z^\epsilon : X \times \Xi \to \bar{R}$ where $Z^\epsilon(x) \propto \mathcal{N}(0, (\sigma_x^\epsilon)^2)$, where $(\sigma_x^\epsilon)^2 = var_\pi\{f_\epsilon(x, \xi_1)\} +$

$2\sum_{i=2}^{\infty} cov_\pi\{f_\epsilon(x,\xi_1), f_\epsilon(x,\xi_i)\}$. The result in (4.3) then follows if $N^{1/2}(\sum_{i=1}^{N} f_\epsilon(x,\xi_i)/N - E_\pi[f_\epsilon(x,\xi)]) \to N^{1/2}(\sum_{k=1}^{N} f(x,\xi_k)/N - E_\pi[f(x,\xi)])$ for any sequence $\{\xi_i\}$ and $(\sigma_x^\epsilon)^2$ converges to σ_x^2 as $\epsilon \to 0$.

This result follows by showing that

$$\lim_{\epsilon \downarrow 0} f_\epsilon(x,\xi) = f(x,\xi), \tag{4.4}$$

for all $x \in X$ and any $\xi \in \Xi$, which we prove by contradiction. If (4.4) does not hold for some $x \in X$ and ξ, then, since $f_\epsilon(x,\xi) \le f(x,\xi)$, there must some sequence $x^j, j = 1,2,\ldots, x^j \in X$ such that $x^j \in B(x,1/j)$ and $\liminf_j f(x^j,\xi) < f(x,\xi)$ while $x^j \to x$, but this contradicts that f is LSC on X for all ξ. This and the assumptions of Theorem 4.2 then establish also that $(\sigma_x^\epsilon)^2$ converges to σ_x^2 at any $x \in X$, completing the proof.

∎

References

[1] Z. Artstein and R. J.-B. Wets (1995). Consistency of minimizers and the SLLN for stochastic programs. *J. Convex Anal.* Vol. 2, pp. 1-17.

[2] 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.

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

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

[5] H. Attouch and R. J.-B. Wets (1990). Epigraphical Processes: Laws of Large Numbers for Random LSC Functions. *Sém. Anal. Convexe, Montepellier* 13.1-13.29.

[6] P. Billingsley (1999). *Convergence of Probability Measures*, 2nd edition, Wiley, New York.

[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] L. Breiman (1960). The Strong Law of Large Numbers for a class of Markov chains, *The Annals of Mathematical Statistics* Vol. 31, No. 3, pp. 801–803.

[9] L. Breiman (1992). *Probability. Corrected reprint of the 1968 original.* Classics in Applied Mathematics, 7. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[10] 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.

[11] S. Geman and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis Mach. Intell.* Vol. 6, pp. 721-741.

[12] C. J. Geyer (2011). Introduction to Markov chain Monte Carlo. *Handbook of Markov chain Monte Carlo.* Edited by Steve Brooks, Andrew Gelman, Galin L. Jones and Xiao-Li Meng, pp. 3-48. Chapman and Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL.

[13] C.J. Geyer (1994). On the convergence of Monte Carlo maximum likelihood calculations, *J. R. Statist. Soc. B*, 56, 261-274.

[14] C. J. Geyer and Y.J. Sung (2007). Monte Carlo likelihood inference for missing data models, *Ann. Statist.*, 35, 990-1011.

[15] C.J. Geyer and E.A. Thompson (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J.R. Statist. Soc. B*, 54, 657-699.

[16] O. Hernández-Lerma and J. B. Lasserre (2001). Further Criteria for Positive Harris Recurrence of Markov Chains. *Proceedings of the American Mathematical Society*, Vol. 129, pp. 1521-1524.

[17] C. Hess (1996). Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator. *The Annals of Statistics* Vol. 24, pp. 1298–1315.

[18] S. F. Jarner and G. O. Roberts (2002). Polynomial convergence rates of Markov chains. *The Annals of Applied Probability.* Vol. 12, pp. 224-247.

[19] G. L. Jones (2004). On the Markov chain central limit theorem. *Probability Surveys.* Vol. 1, pp. 299-320.

[20] A. J. King (1986). Asymptotic behavior of solutions in stochastic optimization: nonsmooth analysis and the derivation of non-normal limit distributions, Dissertation, University of Washington.

[21] A. J. King and R. J. -B. Wets (1991). Epi-consistency of convex stochastic programs. *Stochastics and Stochastics Rep.* Vol. 34, pp. 83-92.

[22] I. Kontoyiannis and S. P. Meyn (2003). Spectral theory and limit theorems for geometrically ergodic Markov processes. *The Annals of Applied Probability.* Vol. 13, pp. 304-362.

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

[24] O. Hernández-Lerma and J. B. Lasserre (2003). *Markov chains and invariant probabilities.* Progress in Mathematics, 211. Birkhäuser Verlag, Basel.

[25] 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.

[26] B. Miasojedow, W. Niemiro, J. Palczewski, and W. Rejchel (2014). Asymptotics of Monte Carlo maximum likelihood estimators. ArXiv e-prints 1412.6371. Available at: {{http://adsabs.harvard.edu/abs/2014arXiv1412.6371M}}.

[27] E. Nummelin (1984). *General irreducible Markov chains and nonnegative operators.* Cambridge Tracts in Mathematics, 83. Cambridge University Press, Cambridge.

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

[29] C. Robert and G. Casella (2011). A short history of MCMC: subjective recollections from incomplete data. *Handbook of Markov chain Monte Carlo.* Edited by Steve Brooks, Andrew Gelman, Galin L. Jones and Xiao-Li Meng, pp. 49-66. Chapman and Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL.

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

[31] G. O. Roberts and J. S. Rosenthal (2006). Harris recurrence of Metropolis-within-Gibbs and transdimensional Markov chains. *Ann. Appl. Probab.* Vol. 16, pp. 2123-2139.

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

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

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

[35] L. Tierney (1998). A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.* Vol. 8, pp. 1-9.

[36] R. J. B. Wets (1984). Modeling and solution strategies for unconstrained stochastic optimisation problems. *Ann. Oper. Res.* pp. 3-22.

[37] M. Zervos (1999). On the epiconvergence of stochastic optimization problems, *Mathematics of Operations Research.* Vol. 24, pp. 495-508.