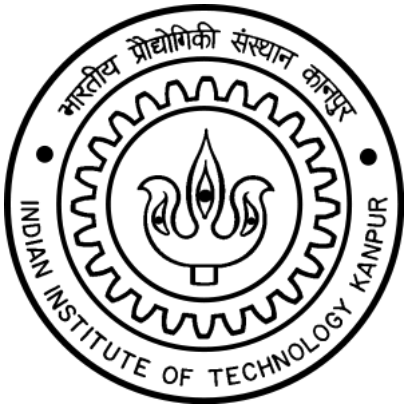
UNDERGRADUATE PROJECT REPORT

Efficient Estimators from MCMC Samples



Supervisor: Prof. Dootika Vats
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur

Submitted by:

Siddharth Pathak

Spring Semester 2023-24
May 1, 2024

Abstract

This is the report for an undergraduate project undertaken during the spring semester of 2024. This project is based on a research paper from Journal of Computational and Graphical Statistics. This work is devoted to developing a highly efficient MCMC estimator to utilize the rejected proposed samples in MCMC sampling algorithms. The main references for this project are Markov Chain Importance Sampling—A Highly Efficient Estimator for MCMC and Monte Carlo Estimation of Bayesian Credible and HPD Intervals. First, we discussed the *MCIS* estimator proposed in [2]. We will investigate its theoretical properties and computational efficiency. Later we will discuss the noble approach of having accurate *HPD* quantiles from the cheap samples (i.e. form samples come from efficient MCMC algorithms like ULA).

Acknowledgement

First and foremost, I would like to express my sincere gratitude to Prof. Dootika Vats for her guidance throughout this project. I am extremely thankful for all the discussions where she always had invaluable insights to offer. Her research methods and ways of tackling a problem have always amazed me. I hope to emulate her enthusiasm and dedication towards research in my future endeavours.

I would like to extend my heartfelt acknowledgement to all my esteemed mathematics and statistics professors at IIT Kanpur. Your expertise, guidance, and commitment to nurturing an understanding of these subjects have been invaluable throughout this project. Your passion for teaching and dedication to our growth as students have left an indelible mark on my academic journey.

Contents

1. Introduction and Preliminaries

A Markov Chain is a mathematical process that undergoes transitions from one state to another. Key properties of a Markov process are that it is random and that each step in the process is "memoryless" in other words, the future state depends only on the current state of the process and not the past. Many real-world phenomena possess this Markov property. Let us consider any continuous state space $\mathbb{S}$ and $A \subseteq \mathbb{S}$. Let $(\mathbb{S}, \Sigma)$ is a measurable space. Formally the discrete-time continuous state space Markov chain $\{X_i\}_{i\geq 0}$ follows:

$$\mathbb{P}(X_{n+1} \in A | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} \in A | X_n = i).$$

1.1. Markov Chain Transition Kernel

A Markov transition kernel is a map $P : \mathbb{S} \times \Sigma \mapsto [0,1]$ such that

- For all $A \in \Sigma$, $P(\cdot, A)$ is a measurable function on $\mathbb{S}$.
- For all $x \in \mathbb{S}$, $P(x, \cdot)$ is a probability measure on Σ.

Informally, this is just like a conditional probability. For $x \in \mathbb{S}$ and $A \in \Sigma$

$$P(x, A) = \mathbb{P}(X_{i+1} \in A | X_i = x).$$

1.2. Transition Density

The continuous state analogue of the one-step transition probability p_{ij} is the one-step transition density. Let $P(x, \cdot)$ be absolutely continuous with respect to a measure μ. Denote $p : \mathbb{S} \times \mathbb{S} \mapsto [0, \infty)$ as the Markov transition density defined as

$$p(x, y)\mu(dy) = P(x, dy)$$

This is not the probability that the chain makes a move from state x to state y. Instead, it is a probability density function in y which describes a curve under which area represents probability. Here, x can be thought of as a parameter of this density. The analogue of the n-step transition probability $p_{ij}^{(n)}$is the n-step transition density denoted by $p^{(n)}(x, y)$. We must have,

$$\int_{\mathbb{S}} p(x, y)dy = 1 \quad \& \quad \int_{\mathbb{S}} p^{(n)}(x, y)dy = 1, \quad \forall n \geq 1$$

1.3. Stationary Distribution

Let $\{X_n\}_{n\geq 0}$ be a Markov chain living on a continuous state space $\mathbb{S}$ with transition probability density $p(x, y)$. A stationary distribution for $\{X_n\}$ on $\mathbb{S}$ is a probability density function $\pi(x)$ on $\mathbb{S}$ satisfying

$$\pi(y) = \int_{\mathbb{S}} \pi(x)p(x, y)dx$$

Given a continuous state Markov chain with n-step transition densities $p^{(n)}(x,y)$, suppose that the $\lim_{n\to\infty} p^{(n)}(x,y)$ exists and is independent of x. Then the limit is a probability density function in y that is also a stationary distribution.

1.4. ***Reversibility***

A Markov chain on a continuous state space $\mathbb{S}$ with transition probability density $p(x,y)$ is said to be reversible with respect to a density $\pi(x)$ if

$$\pi(x)p(x,y) = \pi(y)p(y,x), \quad \forall x,y \in \mathbb{S}$$

One can see that reversibility implies stationarity of the distribution π but the converse is not true.

1.5. ***F-irreducible***

A Markov Chain Transition Kernel P is $F-$irreducible if $\forall x \in \mathbb{S}$ and $A \in \Sigma$ such that $F(A) > 0$ there exists n such that $P^n(x,A) > 0$. Otherwise, P is reducible.

1.6. ***Aperiodicity***

For $d \geq 2$, consider disjoint sets $A_1, \ldots, A_d$ such that for $N = (A_1 \cup \ldots \cup A_d)^c$ then $F(N) = 0$. Further, let the sets satisfy $F(A_i) > 0$ and $P(x,A_{i+1}) = 1$ for all $x \in A_i, 1 \leq i \leq d-1$ and $P(x,A_1) = 1$ for all $x \in A_d$. If such sets do not exist, then P is aperiodic. Otherwise, P is periodic with period d.

1.7. ***Markov Chain Monte Carlo (MCMC)***

Markov chain Monte Carlo (MCMC) is a large class of algorithms that one might turn to where one creates a Markov chain that converges, in the limit, to a distribution of interest. In other words, if one wanted to draw/simulate values from a particular posterior density $\pi(\theta|\mathbf{x})$, an MCMC algorithm might give us a recipe for a transition density $p(\cdot,\cdot)$ that walks around on the support of $\pi(\theta|\mathbf{x})$ so that

$$\lim_{n\to\infty} p^{(n)}(\cdot,\theta) = \pi(\theta|\mathbf{x})$$

Several MCMC algorithms are available, such as Metropolis-Hastings, Gibbs sampling, and Hamiltonian Monte Carlo. The choice of algorithm depends on the nature of our problem and the structure of our model.

There are several instances where drawing samples directly from the distribution or finding a given expectation analytically is impossible. MCMC is a saviour in these situations, but it can only give us samples to estimate the given quantity. How we use these samples to estimate the desired quantity depends on us. This work gives an efficient estimator based on the utilization of rejected proposals. For the unadjusted Langevin algorithm (ULA), it introduces a unique approach to rectify discretization errors.

1.8. *Generalized form of MCMC Accept-Reject Algorithm*

Inside MCMC, a subclass of algorithms follows accept-reject methods to provide samples from a given density. Let ρ be the target density function on $\mathbb{R}^d$ without normalization constant, $Q = q(.|x) : \mathbb{R}^d \mapsto [0,1]$ is the proposal density function. Note that Q is not a fixed pdf; it depends upon the samples that were previously drawn, but the family of probability density remains the same. In each accept-reject algorithm, there is a computable function $\alpha : \mathbb{R}^d \times \mathbb{R}^d \mapsto [0,1]$ called the acceptance probability function. This function is the only way to differentiate between different accept-reject algorithms. Generally α depends upon ρ and Q. The generic structure of an arbitrary k^{th} iteration of accept reject algorithm starting from some initial point X_1 is as follows:

Algorithm 1 Generic MCMC Accept-Reject Algorithm

1: Draw $Y_k \sim q(.|X_k)$ independently from $X_{k-1}, \ldots, X_1$.
2: Compute $\alpha_k = \alpha(X_k, Y_k)$.
3: Draw $U \sim Uniform(0,1)$.
4: **if** $U < \alpha_k$ **then**
5: Set $X_{k+1} = Y_k$.
6: **else**
7: Set $X_{k+1} = X_k$.
8: **end if**

As said before, α is the soul of this algorithm. Different methods of calculating alpha give us different algorithms. For example in famous **Metropolis-Hastings Algorithm** we define

$$\alpha(x,y) = \min\left\{1, \frac{q(x|y)\rho(y)}{q(y|x)\rho(x)}\right\}$$

The Markov chain generated by the MH algorithm is reversible and stationary concerning the target distribution.

1.9. *Langevin algorithm*

The Langevin algorithm, also known as Langevin Monte Carlo and Langevin MCMC, generates samples from a probability distribution for which we have access to the gradient of its log probability. It's a key ingredient of machine learning methods, such as diffusion models and differentially private learning.

Given a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ for which we have access to its gradient Δf, and where $\int \exp(-f(x))dx$ is finite, the Langevin algorithm produces a sequence of random iterates x_0, x_1, with associated density function $x_0 \sim p_0, x_1 \sim p_1, ..$ increasingly approximates the following target distribution:

$$q(x) := \frac{1}{Z}\exp(-f(x)), \quad \text{with} \quad Z = \int_{\mathbb{R}^d} \exp(-f(x))dx.$$

And it does so in a remarkably simple way:

The Langevin algorithm admits different variants, depending on whether the step size is constant

Algorithm 2 Unadjusted Langevin Algorithm (ULA)

Require: starting guess $x_0 \in \mathbb{R}^d$ and step-size $\gamma > 0$

1: **for** t=0,1,... **do**
2: sample $\varepsilon_t \sim N(0,I)$
3: $x_{t+1} = x_t - \gamma \Delta f(x_t) + \sqrt{2\gamma\varepsilon_t}$
4: **end for**
5: **return** $x_1, x_2, \ldots$

or decreasing and whether there's a rejection step or not. The variant above with a constant step-size γ and no rejection step is the most commonly used in practice and is often referred to as the Unadjusted Langevin Algorithm (ULA).

1.10. Harris Recurrence

[3] Let $A \in \Sigma$ and define $\tau_A = \inf\{n \geq 1 : X_n \in A\}$, τ_A is called the first return time to A. If $X_n \notin A$ for all $n \geq 1$, $\tau_A = \infty$. If $FP = F$ and P is $F-$irreducible, then P is Harris Recurrent if for all $A \in \Sigma$ with $F(A) > 0$ and all $x \in \mathbb{S}$.

Harris recurrence, stronger than irreducibility, ensures that a Markov chain will visit set A with probability 1 in finite time, contrasting with irreducibility, which only guarantees a positive probability of visiting A. This distinction is crucial for addressing edge cases and ensuring convergence even in scenarios where initial conditions may pose challenges.

1.11. Ergodicity

[3] If $FP = F$, P is $F-$irreducible, aperiodic, and Harris recurrent, then for every initial distribution λ

$$\lim_{n\to\infty} ||\lambda P^n - F|| = 0$$

Consequently, for all $x \in \mathbb{S}$

$$\lim_{n\to\infty} ||P^n(x,.) - F(.)|| = 0$$

Moreover, for any two initial distributions λ_1 and λ_2

$$\lim_{n\to\infty} ||\lambda_1 P^n - \lambda_2 P^n|| = 0$$

The Markov chain is then said to be ergodic.

1.12. Rate of Convergence

Three rates of convergence are given particular interest in Markov chains. Let $M : X \mapsto R^+$ and $\psi : N \to [0,1]$ be such that

$$||P^n(x,.) - F(.)|| \leq M(x)\psi(n) \text{ for all } x,n.$$

1. Polynomial Ergodicity of order k: $\psi(n) = n^{-k}$ for some $k > 0$.
2. Geometric Ergodicity: $\psi(n) = t^n$ for some $0 \leq t < 1$.

3. Uniform Ergodicity: $\sup_x M(x) < \infty$ and $\psi(n) = t^n$ for some $0 \leq t < 1$.

Polynomial ergodicity is the weaker rate of convergence, followed by geometric ergodicity. If the rate of convergence is a bounded function of the starting values, then we have uniform ergodicity.

2. Problem Statement

Markov chain algorithms are pivotal in diverse fields such as machine learning, statistics, and various other disciplines. Typically, these algorithms are formulated as acceptance-rejection methods. This study includes a novel estimator known as Markov Chain Importance Sampling, specifically designed for these algorithms. Notably, this estimator efficiently utilizes rejected proposals, offering a unique approach for addressing discretization errors in the unadjusted Langevin algorithm. The proposed estimator demonstrates the satisfaction of a central limit theorem and significantly enhances accuracy per CPU cycle. Additionally, it can estimate the normalizing constant, a crucial parameter in Bayesian machine learning and statistics. The primary objective of this study is to investigate the efficiency and effectiveness of the Markov Chain Importance Sampling estimator, contributing valuable insights to improving Markov chain algorithms and their applications in various domains.

3. Markov Chain Importance Sampling (MCIS) Estimator

[2] As discussed above, our estimator uses all proposed samples, letting μ_Y be the asymptotic distribution of the proposed samples $\{Y_k\}_{k\geq 0}$. Let the corresponding density function be ρ_Y. An asymptotically unbiased estimator of ρ_Y can be defined as

$$\hat{\rho}_Y(y) = \frac{1}{K}\sum_{k=0}^{K} q(y|X_k) \xrightarrow{K\to\infty} \rho_Y = \int \rho_X(x)q(y|x)dx$$

The intuition behind the above definition of an estimator of ρ_Y is that, as the K increases, the size of chain $\{X_k\}_{k\geq 0}$ increases, it converges to some distribution ρ_X (say). So,

$$\frac{1}{K}\sum_{k=0}^{K} q(y|X_k) = \sum_{x\in S_X}\left[q(y|x)\sum_{k=1}^{K}\frac{\mathbb{1}_x(x_k)}{K}\right]$$

As, $K \to \infty$, it is obvious to see that $\sum_{k=1}^{K}\frac{\mathbb{1}_x(x_k)}{K} \to \rho_X(x)$. Hence,

$$\frac{1}{K}\sum_{k=0}^{K} q(y|X_k) = \sum_{x\in S_X}\left[q(y|x)\sum_{k=1}^{K}\frac{\mathbb{1}_x(x_k)}{K}\right] \xrightarrow{K\to\infty} \int \rho_X(x)q(y|x)dx$$

Now, ρ_Y and $\hat{\rho}_Y$ give their respective estimates of the expected value of any function f over density ρ. So, the MCIS estimators for $\mathbb{E}_\mu(f)$ can be defined as

$$S_K^{IS}(f) = \frac{\sum_{k=1}^{K} w(Y_k)f(Y_k)}{\sum_{k=1}^{K} w(Y_k)}, \qquad w = \frac{\rho}{\rho_Y}$$

$$\hat{S}_K^{IS}(f) = \frac{\sum_{k=1}^K \hat{w}(Y_k) f(Y_k)}{\sum_{k=1}^K \hat{w}(Y_k)}, \qquad \hat{w} = \frac{\rho}{\hat{\rho}_Y}$$

These estimators are very useful because they make use of rejected samples in *MH* algorithms. Also, for ULA they provide a consistent estimator. This estimator reduces the error of discretization when used with samples of ULA.

One can observe that $\hat{\varepsilon} = K^{-1}\sum_{k=1}^K w(Y_k)$ is the consistent estimator of the normalization constant for the density ρ. This can be seen as

$$\varepsilon = \int \rho(x)dx = \int \frac{\rho(x)}{\rho_Y(x)}\rho_Y(x)dx = \mathbb{E}_Y\left[\frac{\rho(x)}{\rho_Y(x)}\right]$$

and we have $\mathbb{E}_Y[\hat{\varepsilon}] = \mathbb{E}_Y[K^{-1}\sum_{k=1}^K w(Y_k)] = \mathbb{E}_Y\left[\frac{\rho(x)}{\rho_Y(x)}\right] = \varepsilon$

4. Convergence Properties of the MCIS Estimator

In this section, we will investigate the convergence properties of our estimator S_K^{IS}. We will establish uniform ergodicity, geometric ergodicity for $\{(X_k,Y_k)\}_{k\in\mathbb{N}}$, followed by the Law of Large Numbers, and finally CLT for the estimator S_K^{IS}. After this much, we will establish results for our practical estimator $\hat{S}_K^{IS}$.

4.1. Augmented Chain $\{Z_k\}_{k\in\mathbb{N}}$

Let $\{Z_k\}_{k\in\mathbb{N}}$ be an augmented chain where, $Z_k = (X_k,Y_k)$. Note that $\{Z_k\}_{k\in\mathbb{N}}$ is not necessarily a markov chain, if $\{X_k\}_{k\in\mathbb{N}}, \{Y_k\}_{k\in\mathbb{N}}$ are generated by MH algorithm. In case, the samples generated by ULA, $\{Z_k\}_{k\in\mathbb{N}}$ will also be a Makrov chain.

Let $\{X_k\}_{k\in\mathbb{N}}, \{Y_k\}_{k\in\mathbb{N}}$ are obtained from general accept-reject algorithm discussed before. Let $\alpha(x,y)$ be the acceptance probability function of the algorithm. At each stage $Z_k = (X_k,Y_k)$ will be transformed to either (X_k,Y_{k+1}) or (Y_k,Y_{k+1}). So, the kernel of $\{Z_k\}_{k\in\mathbb{N}}$ is a function $K_Z : \mathbb{R}^{2d} \times \mathscr{B}(\mathbb{R}^{2d}) \mapsto [0,1]$ define as follows:

$$\begin{aligned} K_Z((x,y),A\times B) &= (1-\alpha(x,y))\mathbb{P}((x_k,y_{k+1}) \in A\times B) + \alpha(x,y)\mathbb{P}((y_k,y_{k+1}) \in A\times B) \\ &= (1-\alpha(x,y))\mathbb{1}_A(x)q(B|x) + \alpha(x,y)\mathbb{1}_A(y)q(B|y) \end{aligned}$$

It can be observed that K_Z is degenerate, but K_Z^m will be globally supported for $m > 1$. Now, we will explore some properties of the augmented chain $Z_{kk\in\mathbb{N}}$ that will be useful for further analysis.

Theorem 4.1. *[2] Let $(Z_k)_{k\in\mathbb{N}} = (X_k,Y_k)_{k\in\mathbb{N}}$ is an augmented chain, where $\{X_k\}_{k\in\mathbb{N}}$ and $\{Y_k\}_{k\in\mathbb{N}}$ are the chains generated by Markov chain accept reject algorithm. Then $\{Z_k\}_{k\in\mathbb{N}}$ has following properties:*

- *If $\{X_k\}_{k\in\mathbb{N}}$ has a stationary distribution μ_X with density ρ_X, then $\{Z_k\}_{k\in\mathbb{N}}$ has the stationary*

distribution μ_Z with density $\rho_Z(x,y) = \rho_X(x)q(y|x)$.

- *Let the proposal densities $q(\cdot|\cdot)$ be globally supported and continuous. If $\{X_k\}_{k\in\mathbb{N}}$ is irreducible, aperiodic and/or Harris positive, so is $\{Z_k\}_{k\in\mathbb{N}}$.*
- *If $\{X_k\}_{k\in\mathbb{N}}$ is geometrically ergodic, so is $\{Z_k\}_{k\in\mathbb{N}}$.*
- *If $\{X_k\}_{k\in\mathbb{N}}$ is uniformly ergodic, so is $\{Z_k\}_{k\in\mathbb{N}}$.*

Proof. • Density of chain $\{Z_k\}_{k\in\mathbb{N}}$ is found by using the hierarchical structure of the two variables as follows:

$$\rho_Z(x,y) = \rho_{(X,Y)}(x,y) = \rho_X(x)\rho_{Y|X}(y|x) = \rho_X(x)q(y|x)$$

- Aperiodicity and irreducibility of $\{Z_k\}$ in follows directly since the proposal densities q are assumed to be globally supported. For Harris recurrence, let $A \subseteq \mathscr{B}(\mathbb{R}^{2d})$ with $\mu_Z(A) > 0$. By continuity of q there exist subsets $A_X, A_Y \in \mathscr{B}(\mathbb{R}^d)$ and $\varepsilon > 0$, such that

$$A_X \times A_Y \subseteq A, \quad \mu_X(A_X) > 0 \text{ and } \int_{A_Y} q(y|x)dy > \varepsilon \ \forall x \in A_X.$$

Since $\{X_k\}$ is Harris positive by assumption, we have $\mathbb{P}[X_k \in A_X \text{ infinitely often }] = 1$, regardless of the initial value X_1. This implies that, almost surely, Z_k will enter $A_X \times A_Y \subseteq A$ in finite time,

$$\mathbb{P}[min_{k>1}(Z_k \in A) < \infty] = 1,$$

regardless of the initial value X_1. This completes the proof.

- If $\{X_k\}$ is geometrically ergodic, there exists $r > 1$ such that

$$\sum_{r=1}^{\infty} r^m ||K_X^m(x,\cdot) - \mu||_{TV} < \infty$$

for all $x \in \mathbb{R}^d$. For any signed measure ν on $\mathbb{R}^d$ we define the signed measure $\nu \odot q$ on $\mathbb{R}^{2d}$ given by

$$\nu \odot q = \int_A \int_B q(y|x)dyd\nu(x)$$

Since q is non-negative, its Jordan decomposition $\nu \odot q = (\nu \odot q)_+ - (\nu \odot q)_-$ is given by

$$(\nu \odot q)_+ = \nu_+ \odot q \quad \& \quad (\nu \odot q)_- = \nu_- \odot q$$

and therefore

$$||\nu \odot q||_{TV} = (\nu \odot q)_+(\mathbb{R}^{2d}) + (\nu \odot q)_-(\mathbb{R}^{2d}) = \nu_+(\mathbb{R}^{2d}) + \nu_-(\mathbb{R}^{2d}) = ||\nu||_{TV}$$

Here, we used the connection between the total variation norm and the Jordan decomposition

of a measure Since X_2 is either equal to X_1 (with probability $1-\alpha_1$) or to Y_1 (with probability α_1), this implies for each $(x,y) \in \mathbb{R}^{2d}$, $\alpha := \alpha(x,y)$ and $m \geq 1$.

$$K_Z^m((x,y),\cdot) = [(1-\alpha)K_X^{m-1}(x,\cdot) + \alpha K_Y^{m-1}(y,\cdot)] \odot q,$$

$$||K_Z^m((x,y),\cdot) - \mu \odot q||_{TV} = (1-\alpha)||K_X^m(x,\cdot) - \mu||_{TV} + \alpha||K_Y^m(y,\cdot) - \mu||_{TV}$$

and thus

$$\sum_{r=1}^{\infty} r^m ||K_Z^m((x,y),\cdot) - \mu \odot q||_{TV} < \infty$$

which proves this property.

- Proof for Uniform ergodicity will follow similarly.

□

4.2. *Central Limit Theorem & Law of Large Number*

[3] The LLN is important because it guarantees stable long-term results for the averages of random events. We are now set to establish the law of large numbers for MCIS estimator S_K^{IS}. Let μ be a probability measure on $\mathbb{R}^d, \rho : \mathbb{R}^d \to \mathbb{R}$ be proportional to the probability density function of μ and $f \in L^1(\mu)$. Let the processes $\{X_k\}_{k\in\mathbb{N}}$ and $\{Y_k\}_{k\in\mathbb{N}}$ be given by a Markov chain acceptance-rejection algorithm. Let us define some random variables to make our calculations easier.

$$\phi(y) = \frac{f(y)\rho(y)}{\rho_Y(y)}, \quad w(y) = \frac{\rho(y)}{\rho_Y(y)}$$

Claim: $S_K^{IS}(f) = \frac{\sum_{i=1}^K \phi(Y_i)}{\sum_{i=1}^K w(Y_i)} \xrightarrow{a.s.} \mathbb{E}_\mu(f)$

Proof.

$$\bar{\phi}_K = \frac{1}{K}\sum_{i=1}^K \phi(Y_i) \xrightarrow{a.s.} \mathbb{E}_{Y\sim\rho_Y}\left(\frac{f\rho}{\rho_Y}\right) = \int f(y)\rho(y)dy = \varepsilon\mathbb{E}_\rho(f)$$

$$\bar{w}_K = \frac{1}{K}\sum_{i=1}^K w(Y_i) \xrightarrow{a.s.} \mathbb{E}_{Y\sim\rho_Y}\left(\frac{\rho}{\rho_Y}\right) = \int \rho(y)dy = \varepsilon$$

Above two, statements imply $S_K^{IS}(f) = \frac{\sum_{i=1}^K \phi(Y_i)}{\sum_{i=1}^K w(Y_i)} \xrightarrow{a.s.} \mathbb{E}_\mu(f)$. This proves the claim. □

Now to prove CLT of our theoretical estimator $S_K^{IS}(f)$, The statement for CLT for MCMC is as follows:

Theorem 4.2. *[2] The sequence $X_1, X_2, X_3, \ldots$ of random elements of some set is a Markov chain that has a stationary probability distribution and the initial distribution of the process, i.e. the distribution of X_1, is the stationary distribution, so that $X_1, X_2, X_3, \ldots$ are identically distributed. In the classic central limit theorem these random variables would be assumed to be independent, but here we have only the weaker assumption that the process has the Markov property, and g is some (measurable)*

real-valued function for which $var(g(X_1)) < \infty$. *Now let*

$$\mu = \mathrm{E}(g(X_1)),$$
$$\widehat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} g(X_k)$$
$$\sigma^2 := \lim_{n\to\infty} \mathrm{var}(\sqrt{n}\widehat{\mu}_n) = \lim_{n\to\infty} n\,\mathrm{var}(\widehat{\mu}_n) = \mathrm{var}(g(X_1)) + 2\sum_{k=1}^{\infty} \mathrm{cov}(g(X_1), g(X_{1+k})).$$

Then as $n \to \infty$, *we have*

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathscr{D}} Normal(0, \sigma^2)$$

Let us consider that $\{Z_k\}_{k\in\mathbb{N}} = (\{(X_k, Y_k)\}_{k\in\mathbb{N}}$ is the augmented chain, we discussed before. Let us define a function $h : \mathbb{R}^{2d} \mapsto \mathbb{R}^2$ such that

$$h(z) = h(x,y) = (\phi(y), w(y))^t$$

Now, we will proceed further to establish CLT for S_K^{IS}. let us have the following assumptions:

- $\{X_k\}_{k\geq 1}$ is geometrically ergodic.
- for some $\psi > 0$, $\mathbb{E}_{Y\sim\rho_Y}(|\phi(Y)|^{2+\psi}) < \infty$ and $\mathbb{E}_{Y\sim\rho_Y}(|w(Y)|^{2+\psi}) < \infty$

If the above two conditions hold then $\|\Sigma_h\| < \infty$, where

$$\Sigma_h := 0.5\Sigma_h^{(1)} + \sum_{k=2}^{\infty} \Sigma_h^{(k)}$$

$$\Sigma_h^{(k)} := \mathbb{C}ov_{X_1\sim\rho}[h(Z_1), h(Z_k)] + \mathbb{C}ov_{X_1\sim\rho}[h(Z_k), h(Z_1)]$$

and, MCIS estimator S_K^{IS} follows the CLT, stating below

$$\sqrt{K}(S_K^{IS} - \mathbb{E}_\rho(f)) \xrightarrow{\mathscr{D}} \mathrm{Normal}(0, \Sigma_{CLT})$$

With

$$\Sigma_{CLT} = \varepsilon^{-2}\begin{bmatrix} 1 & -\mathbb{E}_\mu(f) \end{bmatrix} \Sigma_h \begin{bmatrix} 1 \\ -\mathbb{E}_\mu(f) \end{bmatrix}$$

To prove the above theorem, we need some results as follows:

Proposition 4.3. *[2] Let* $\{Z_k\}_{k\in\mathbb{N}}$ *be an aperiodic, irreducible, Harris positive and geometrically ergodic Markov chain in* R^n *with stationary distribution* π *and* $h : R^n \to \mathbb{R}^m$ *be a measurable function such that*

$$E_{Z\sim\pi}(||h(Z)||^{2+\varepsilon}) < \infty$$

for some $\varepsilon > 0$. *Then* $||\gamma_h|| < \infty$, *where* $\gamma_h \in \mathbb{R}^{m\times m}$ *is given by*

$$\gamma_h := 0.5\gamma_h^{(1)} + \sum_{k=2}^{\infty} \gamma_h^{(k)}, \quad \gamma_h^{(k)} := \mathbb{C}ov_{Z_1\sim\pi}[h(Z_1), h(Z_k)] + \mathbb{C}ov_{Z_1\sim\pi}[h(Z_k), h(Z_1)]$$

and

$$G_K = \sqrt{K}\left(\frac{1}{K}\sum_{i=1}^{K} h(Z_i) + \mathbb{E}_{Z\sim\pi}(h(Z))\right) \xrightarrow[K\to\infty]{d} N(0,\gamma_h)$$

Proof. Let $v \in \mathbb{R}^m$ and $h_v = v^T h : \mathbb{R}^n \mapsto \mathbb{R}$. The above conditions imply $\mathbb{E}_{Z\sim\pi}[|h_v(Z)|^{2+\varepsilon}] < \infty$, hence we yields $0 < \gamma_{h_v} < \infty$ and

$$v^T G_K = \sqrt{K}\left(\frac{1}{K}\sum_{i=1}^{K} h_v(Z_i) + \mathbb{E}_{Z\sim\pi}(h_v(Z))\right) \xrightarrow[K\to\infty]{d} N(0,\gamma_{h_v})$$

Note that $\gamma_{h_v}^{(k)} = v^T \gamma_h^{(k)} v$ and thereby $\gamma h_v = v^T \gamma_h v$. Since $v \in \mathbb{R}^m$ was chosen arbitrarily and h is symmetric (for each partial sum), this implies $||\gamma_h|| < \infty$. Since G_K converges in distribution to $H \sim N(0,\gamma_h)$ if and only if

$$v^T G_K \xrightarrow[K\to\infty]{d} v^T H \sim N(0, v^T \gamma_h v)$$

for each $v \in \mathbb{R}^m$, (8) proves the claim. □

Proof (CLT for MCIS). Since $\{Z_k\}$ is aperiodic, irreducible, Harris positive and geometrically ergodic by Theorem 4.1, proposition 4.3 yields

$$\sqrt{K}\left[\begin{bmatrix}\bar{\phi}_k \\ \bar{w}_k\end{bmatrix} - \begin{bmatrix}\varepsilon\mathbb{E}_\mu[f] \\ \varepsilon\end{bmatrix}\right] \xrightarrow{d} N(0,\Sigma_h)$$

By applying the delta method with

$$g(u,v) = \frac{u}{v}, \quad \nabla(\varepsilon\mathbb{E}_\mu[f],\varepsilon) = \varepsilon^{-1}\begin{bmatrix}1 \\ -\mathbb{E}_\mu(f)\end{bmatrix}$$

This establishes the CLT for $S_K^{IS}(f) = \frac{\bar{\phi}}{\bar{w}}$ □

5. Applications to Sampling Algorithm & Covergence Properties of $\hat{S}_K^{IS}$

We conducted an analysis of the convergence properties of the Monte Carlo Importance Sampling (MCIS) estimator S_K^{IS}. The estimator $\hat{S}_K^{IS}$, which is practically relevant, presented challenges due to the intricate nature arising from the interdependence of the points Y_k. In this section, we undertake a comparative assessment of the approximation properties of $\hat{S}_K^{IS}$ concerning stratified sampling and layered adaptive importance sampling (LAIS). This exploration aims to elucidate the operational mechanisms of $\hat{S}_K^{IS}$ and lay the groundwork for rigorous mathematical scrutiny of $\hat{S}_K^{IS}$. Regrettably, our endeavours to establish such a theoretical foundation have thus far proven unsuccessful.

$\hat{S}_K^{IS}$ has been exclusively regarded as an approximation to S_K^{IS}. Nevertheless, empirical evidence demonstrates that in numerous experiments, $\hat{S}_K^{IS}$ outperforms S_K^{IS} substantially and performs comparably to LAIS while evaluating the target density only half as frequently as LAIS. Consequently, it is prudent to scrutinize the estimation properties of $\hat{S}_K^{IS} \approx \mathbb{E}_\mu(f)$ directly, rather than

construing $\hat{S}_K^{IS}$ solely as an approximation to S_K^{IS}. For this purpose, a meaningful comparison between the samples Y_k and random variables $Z_k \sim q(\cdot|X_k), k = 1,\ldots,K$, sampled independently after the establishment of the Markov chain $\{X_k\}_{k\geq 0}$, is undertaken. Specifically, if $X_1,\ldots,X_K$ are held constant, $Z_1,\ldots,Z_K$ become conditionally independent stratified samples drawn from the mixture distribution $\hat{\rho}_Y = (1/K)\sum_{k=1}^{K} q(\cdot|X_k)$.

This approach offers three advantages:

- The samples are drawn from $\hat{\rho}_Y$ rather than ρ_Y, facilitating a straightforward application of importance sampling without the need for an approximation of the importance density.
- Stratified samples exhibit provably superior variance properties compared to independent samples from $\hat{\rho}_Y$.
- The samples are drawn directly from $\hat{\rho}_Y$ without reliance on asymptotic augmenteds or Markov chains. The utilization of the underlying Markov chain $\{X_k\}_{k\in\mathbb{N}}$ is solely to obtain a good importance sampling density $\hat{\rho}_Y$.

It is noteworthy that $\hat{S}_K^{IS}$, based on the samples Z_k instead of Y_k, aligns precisely (as a special case) with the LAIS estimator. Additionally, the behaviour of samples Y_k closely resembles that of Z_k, with each Y_k being a sample from $q(.|X_k)$, differing only in the violation of conditional independence given $\{X_k\}_{k\in\mathbb{N}}$. The challenge in analyzing the estimator $\hat{S}_K^{IS}$ stems from this lack of conditional independence, through the expected decline in conditional dependence with increasing $|k-l|$ suggests a promising direction for analysis. Also, one must note that the estimator $\hat{S}_K^{IS}$ is not efficient because it is taking $\Theta(K^2)$ in computing the estimate. In the areas of big data, this estimator has no use we have to prove this fact computationally. The results are as follows:

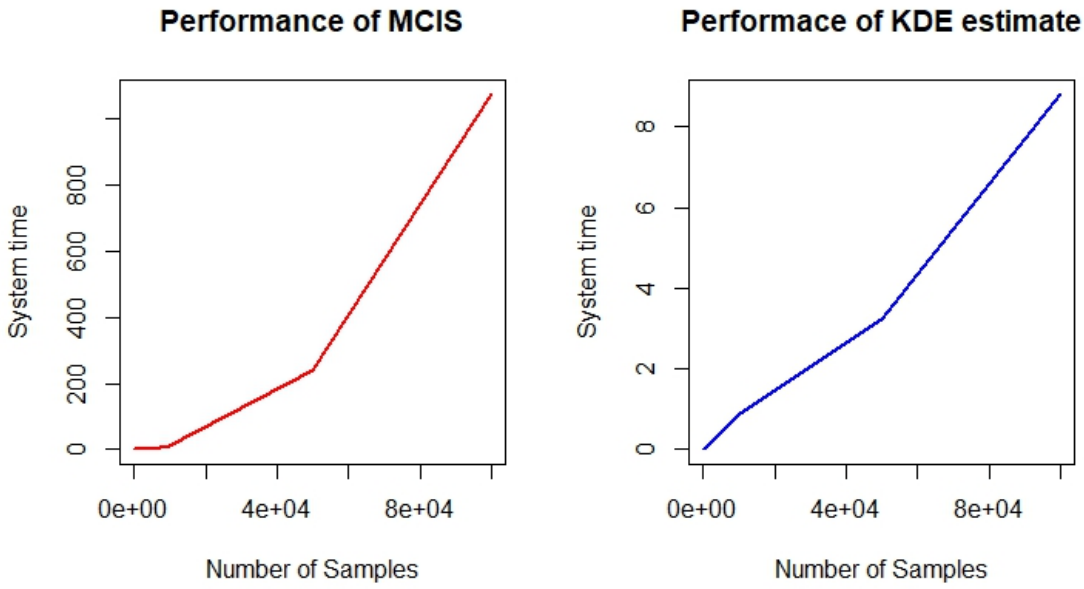


Fig. 1. Efficiency of Estimators

We consider the density function $\rho = e^{-x^2}(1 + \sin(5x) + \sin(2x))$. Note that this is an unnormalized density. In our experiment, we estimate the expectation of RV having density ρ. The first plot is for the proposed *MCIS* estimator. One can see that we can not use *MCIS* for a large sample as it is inefficient. If we replace $\hat{\rho}_Y$ with the kernel density estimation of Y we get a reasonably good estimate very efficiently. The computation of the estimate from this approach is taking $1/100$ time as compared to *MCIS*. Also, the variance of these estimates is smaller than that of *MCIS*.

Repetitions	Variance of MCIS	Variance of Estimate with KDE
1000	0.003161821	7.337105e-07

Table 1. For 10000 Samples

6. Motivation and Intuition of Estimation of Quantiles from Cheap Samples

In the previous section, we discussed a novel approach to estimating the expectation of any measurable function via cheap MCMC samples. Although quantiles did not have closed-form expression in general, we can have a method through which we can estimate quantiles from inaccurate and cheap samples. In this section, we will extend the approach to estimate CDF, quantile and *HPD* intervals via MCMC sampling. These days, it's becoming quite common for folks to summarize posterior distributions by listing out $100(1-\alpha)\%$ posterior credible interval parameters of interest in Bayesian inference. The main reason behind this is that these intervals are pretty easy to get hold of. We can obtain such credible intervals by considering a Bayesian posterior density represented by the equation:

$$\pi(\theta, \phi|D) \propto L(\theta, \phi|D)\pi(\theta, \phi)$$

where D represents the data, the parameter θ is one-dimensional, and ϕ is the multidimensional parameter. $\pi(\theta, \phi)$ represents the prior on the joint distribution of $\pi(\theta, \phi)$. In this equation, $L(\theta, \phi|D)$ is the likelihood function given the data. We aim to derive Bayesian credible and HPD (highest posterior density) intervals for θ.

Let $\pi(\theta|D)$ and $\Pi(\theta|D)$ denote the marginal posterior density function and the maximum posterior cumulative distribution function (CDF) of θ, respectively. To simplify things, let's first assume that $\pi(\theta|D)$ is unimodal. However, we'll discuss possible extensions to multimodal cases also. In this section, we'll also assume that θ can be generated from $\Pi(\theta|D)$ using a Markov chain Monte Carlo (MCMC) sampling algorithm (Especially ULA). We'll then derive exact or approximate credible and HPD intervals for θ. For exact intervals, we further assume that $\pi(\theta|D)$ and $\Pi(\theta|D)$ are analytically available, meaning we know their closed forms.

6.1. Some Important Results

Assuming we have the closed forms of $\pi(\theta|D)$ and $\Pi(\theta|D)$ available, obtaining a Bayesian credible interval for θ is straightforward. For instance, if we're interested in a $100(1-\alpha)\%$ credible interval, we calculate $\theta_{(\alpha/2)}$ and $\theta_{(1-\alpha/2)}$ such that:

$$\Pi(\theta_{(a/2)}|D) = a/2 \quad \text{and} \quad \Pi(\theta_{(1-a/2)}|D) = 1-a/2$$

Then, a $100(1-\alpha)\%$ credible interval for θ is $(\theta_{\alpha/2}, \theta(1-\alpha/2))$. When $\pi(\theta|D)$ is not symmetric, an HPD interval is more plausible. A $100(1-\alpha)\%$ HPD interval for θ is simply given by

$$R(\pi_\alpha) = \{\theta : \pi(\theta|D) > \pi_\alpha\}$$

where π_α, is the largest constant such that $\mathbb{P}(\theta \in R(\pi_\alpha)) \geq 1-\alpha$. The value π_α can be thought of as a horizontal line placed over the posterior density whose intersection(s) with the posterior defined regions with probability $1-\alpha$.

Consider having a Markov Chain Monte Carlo (MCMC) sample, denoted as $\{\theta_i, i = 1,2,...,n\}$, drawn from the marginal posterior distribution $\pi(\theta|D)$. It's noteworthy that if $\{(\theta_i, \phi_i), i = 1,2,\ldots,n\}$ constitutes an MCMC sample from the joint posterior distribution $\pi(\theta,\phi|D)$, then $\{\theta_i, i = 1,2,...,n\}$ forms an MCMC sample from $\pi(\theta|D)$. Now, let $\psi_i = \pi(\theta|D)$ for $i = 1,2,\ldots,n$. We select $\hat{\pi}_\alpha = \psi_{(j)}$, where $\psi_{(j)}$ represents the j^{th} smallest element of $\{\psi_i\}$, $j = [\alpha n]$, and $[\alpha n]$ denotes the integer part of αn. Consequently, $\hat{\pi}_\alpha \to \pi_\alpha$ as $n \to \infty$, and thus, $R(\pi_\alpha)$ converges to $R(\hat{\pi}_\alpha)$ as n tends to infinity. Also, a $100(1-\alpha)\%$ Bayesian credible interval is

$$(\theta_{([(a/2)n])}, \theta_{([(1-a/2)n])})$$

where $\theta_{([(\alpha/2)n)}$ and $\theta_{([(1-\alpha/2)n])}$ are the $[(\alpha/2)n]^{th}$ smallest and the $[(1-\alpha/2)n]^{th}$ smallest of θ_i, respectively. To obtain a $100(1-\alpha)\%$ HPD interval, Let us define

$$R_j(\pi) = \left(\theta_{(j)}, \theta_{(j+(1-\alpha)n])}\right)$$

for $j = 1,2,\ldots,n-[(1-c)n]$. Then we have the following result.

Theorem 6.1. *[1] Let $\{\theta_i, i = 1,2,\ldots,n\}$ be an ergodic MCMC sample from $\pi(\theta|D)$ and let $R_j^*(n) = (\theta_{(j^*)}, \theta_{(j+[(1-\alpha)n])})$, where j^* is chosen so that*

$$\theta_{(j^*+[(1-\alpha)n])} - \theta_{(j^*)} = min_{1\leq j\leq n-[(1-\alpha)n]}(\theta_{(j+(1-\alpha)n])} - \theta_{(j)})$$

That is, $R_{j^*}(n)$ has the smallest interval width among all $R_j(n)$'s. If $\pi(\theta|D)$ is unimodal and the HPD interval uniquely exists, then we have $R_{j^*}(n) \to R(\pi_\alpha)$ almost surely as $n \to \infty$. Thus, to find a $100(1-\alpha)\%$ HPD interval, we look at all the $100(1-\alpha)\%$ credible intervals in the sample and then take the one with the smallest interval width.

Note 6.2. *Proof of the above theorem is quite lengthy and involved hence, we are skipping it in our discussion.*

To obtain the HPD interval for $\eta = h(\theta,\phi)$. The result is given in the following corollary.

Corollary 6.3. *[1] Let $\{(\theta_i,\phi_i)i=1,2,\ldots,n\}$ be an ergodic MCMC sample from $\pi(\theta,\phi|D)$. Also let $\eta_i=h(\theta_i,\phi_i)$ and the $\eta_{(i)}$ be the ordered values of the η_i. Then a $100(1-\alpha)\%$ HPD interval of η can be approximate by $R_j^*(n)=(\eta_{(j^*)},\eta_{(j+[(1-\alpha)n])})$ where j^* is chosen so that*

$$\eta_{(j^*+[(1-\alpha)n])}-\eta_{(j^*)}=min_{1\le j\le n-[(1-\alpha)n]}(\eta_{(j+(1-\alpha)n])}-\eta_{(j)})$$

That is, $R_{j^*}(n)$ has the smallest interval width among all $R_j(n)$'s.

7. Importance Sampling Approach

The objective of this section is to develop a Monte Carlo method for computing posterior HPD intervals using samples from an importance sampling distribution. [1]

7.1. Consistent Estimator for CDF

Assume that $g(\theta,\phi)$ is a joint importance sampling density for $\pi(\theta,\phi)$. Also note that $\pi(\theta,\phi)$ may be evaluated only up to an unknown normalizing constant. It can be seen as

$$\pi(\theta,\phi|D)\propto p(\theta,\phi|D)=L(\theta,\phi|D)\pi(\theta,\phi)$$

Let $\Pi(\theta|D)$ be the marginal posterior cumulative distribution function of θ. We formalize the Monte Carlo approach to approximate the α^{th} quantile to obtain an estimation of Bayesian credible or HPD interval. It is easy to observe that for a given θ^*

$$\Pi(\theta*|D)=\mathbb{E}(\mathbb{1}_{\theta\le\theta^*})=\frac{\int\mathbb{1}_{\theta\le\theta^*}\frac{p(\theta,\phi|D)}{g(\theta,\phi)}g(\theta,\phi|D)d\phi d\theta}{\int\frac{p(\theta,\phi|D)}{g(\theta,\phi)}g(\theta,\phi|D)d\phi d\theta}$$

Then, a simulation consistent estimator of $\Pi(\theta|D)$ can be obtained as

$$\hat{\Pi}(\theta*|D)=\frac{\sum_{i=1}^n\mathbb{1}_{\theta\le\theta^*}\frac{p(\theta,\phi|D)}{g(\theta,\phi)}g(\theta,\phi|D)}{\sum_{i=1}^n\frac{p(\theta,\phi|D)}{g(\theta,\phi)}g(\theta,\phi|D)}$$

Let $\{\theta_{(i)}\}$ be the ordered values of $\{\theta_i\}$. Corresponding to $\{\theta_{(i)},i=1,\ldots,n\}$, we rewrite the ergodic MCMC sample $\{(\theta_{(i)},\phi_{(i)},i=1,\ldots,n\}$. Note that $\phi_{(i)}$ is just a notation. Denote

$$w_i=\frac{\frac{p(\theta,\phi|D)}{g(\theta,\phi)}}{\sum_{i=1}^n\frac{p(\theta,\phi|D)}{g(\theta,\phi)}}$$

for $1\le i\le n$. Then, we have

$$\hat{\Pi}(\theta|D)=\left\{\begin{array}{ll}0, & \text{if } \theta\le\theta_{(1)}\\ \sum_{j=1}^i w_j, & \text{if } \theta_{(i)}\le\theta<\theta_{(i+1)}\\ 1, & \text{if } \theta\ge\theta_{(n)}\end{array}\right\}$$

Note that $\hat{\Pi}(\theta|D)$ is nothing but the empirical cdf of θ. Under certain regularity conditions such as ergodicity, we can show that the central limit theorem still holds for $\hat{\Pi}(\theta|D)$.

7.2. *Estimation of Quantiles through Empirical CDF*

Let $\theta^{(\alpha)}$ be the α^{th} quantile of θ. i.e.

$$\theta^{(\alpha)} = inf\{\theta : \Pi(\theta|D) \geq \alpha\}$$

Using empirical CDF $\theta^{(\alpha)}$ can be estimated as

$$\hat{\theta}^{(\alpha)} = \left\{ \begin{array}{ll} \theta_{(1)}, & \text{if } \alpha = 0 \\ \theta_{(i)}, & \text{if } \sum_{j=1}^{i-1} w_j < \alpha \leq \sum_{j=1}^{i} w_j \end{array} \right\}$$

To obtain a $100(1-\alpha)\%$ HPD interval for θ, we let

$$R_j(n) = \left(\hat{\theta}^{(\frac{j}{n})}, \hat{\theta}^{\left(\frac{j+(1-\alpha)n}{n}\right)} \right)$$

For $j = 1, 2, \ldots, n - [(1-\alpha)n]$. Then similar to the theorem 6.1, we have the following result.

Theorem 7.1. *[1] Let $R_{j^*}(n)$ be the interval that has the smallest width among all $R_j(n)$'s. If $\pi(\theta|D)$ is unimodal and has unique HPD interval for α, then we have*

$$R_{j^*}(n) \to R(\pi_\alpha) \text{ as } n \to \infty$$

Using the result for $\hat{\theta}^{(\alpha)}$, an estimate of $100(1-\alpha)\%$ Bayesian credible interval for θ is

$$\left(\hat{\theta}^{(\alpha/2)}, \hat{\theta}^{(1-\alpha/2)} \right)$$

Similar to Corollary 6.3, we can obtain an HPD interval of $\eta = h(\theta, \phi)$ as follows.

Corollary 7.2. *[1] Let $\eta = h(\theta, \phi)$ for $i = 1, 2, \ldots, n$. Also, let the $\eta_{(i)}$ denote the ordered values of the η_i. Then, the α^{th} quantile of the marginal posterior distribution of η can be estimated by*

$$\hat{\eta}^{(\alpha)} = \left\{ \begin{array}{ll} \eta_{(1)}, & \text{if } \alpha = 0 \\ \eta_{(i)}, & \text{if } \sum_{j=1}^{i-1} w_j < \alpha \leq \sum_{j=1}^{i} w_j \end{array} \right\}$$

Using the above result we can compute

$$R_j(n) = \left(\hat{\eta}^{(\frac{j}{n})}, \hat{\eta}^{\left(\frac{j+(1-\alpha)n}{n}\right)} \right)$$

A $100(1-\alpha)\%$ HPD interval of η is $R_{j^*}(n)$ that has the smallest interval among all $R_j(n)$.

7.3. ***Results***

Let us take two examples to verify the above results computationally. Let $Y_i \overset{IID}{\sim} N(\mu, \sigma^2)$. Let us consider the case we know σ exactly. We want to find the quantiles of the posterior distribution of μ. Let us consider prior to be $N(0, 100)$ It will be easy to establish that the posterior distribution of μ with σ known is Normal distribution with mean $\frac{\sum_{i=1}^{n} Y_i/\sigma^2}{(n/\sigma^2)+(1/100)}$ and variance $\frac{1}{(n/\sigma^2)+(1/100)}$. We will apply the above method to find quantiles and verify the estimator's accuracy by mean squared error. In our analysis, we fix $\mu = 0$ and $\sigma = 1$. After summarising the posterior distribution of μ i.e. $\pi(\mu|Y)$ we obtained $n = 100000$ samples for vis Unadjusted Langevin algorithm and calculated the 95% HPD interval as mentioned above.

Actual HPD Interval	$(-0.05607639,\ 0.49822971)$
Obtained HPD Interval	$(-0.05941751,\ 0.49253528)$

Table 2. HPD Interval for $\pi(\mu|Y, \sigma^2)$

Obtained *HPD* are decently accurate. To test the consistency of the estimator we obtained mean squared error for $\alpha = 0.25, 0.5, 0.75$. The results are shown below:

α	MSE
0.25	$2.361802e-07$
0.50	$3.490802e-08$
0.75	$3.491659e-08$

Table 3. MSE $= 0.0001 \times \sum_{i=0}^{10000} (\hat{\mu}^{(\alpha)} - \mu^{(\alpha)})^2$

Consider another example where $Y_i \overset{IID}{\sim} Poisson(\lambda)$ for $i = 1, 2, \ldots, n$. We will infer about the posterior of λ. Let us consider prior to be $Gamma(1, 1)$. Posterior for λ will be $Gamma(\sum_{i=1}^{n} Y_i + 1, n + 1)$. After getting posterior samples from ULA we have the following results.

Actual HPD Interval	$(12.55808,\ 14.57834)$
Obtained HPD Interval	$(12.56114,\ 14.56607)$

Table 4. HPD Interval for $\pi(\lambda|Y)$

Results after repeating the experiments 1000 times each for $\alpha = 0.25, 0.5, 0.75$ we get

α	MSE
0.25	$3.686399e-06$
0.50	$6.695139e-07$
0.75	$4.705776e-06$

Table 5. MSE $= 0.0001 \times \sum_{i=0}^{10000} (\hat{\lambda}^{(\alpha)} - \lambda^{(\alpha)})^2$

8. Conclusion

In this undergraduate project, we embarked on an exploration inspired by research from the Journal of Computational and Graphical Statistics, aiming to enhance the efficiency of Markov Chain Monte Carlo (MCMC) estimators by leveraging rejected proposed samples within MCMC sampling algorithms. Grounded in the seminal works of "Markov Chain Importance Sampling—A Highly Efficient Estimator for MCMC" and "Monte Carlo Estimation of Bayesian Credible and HPD Intervals", our endeavour centred on understanding and implementing the MCIS estimator proposed in the former, while also delving into the novel strategy of extracting accurate Highest Posterior Density (HPD) quantiles from inexpensive samples, such as those generated by efficient MCMC algorithms like Unadjusted Langevin Algorithm (ULA).

Through our investigation, we delved into the theoretical underpinnings and computational intricacies of the MCIS estimator, assessing its potential to revolutionize MCMC-based inference techniques. Furthermore, we explored the innovative approach of deriving precise HPD quantiles from samples yielded by cost-effective MCMC methods, thereby enriching the spectrum of available techniques for Bayesian analysis.

As we conclude this project, we recognize the value of continuous exploration and innovation within computational statistics. Our journey underscores the significance of bridging theoretical advancements with practical implementations, paving the way for more robust and efficient methodologies in statistical inference.

References

1. Ming-Hui Chen and Qi-Man Shao, *Monte carlo estimation of bayesian credible and hpd intervals*, Journal of Computational and Graphical Statistics **8** (1999), no. 1, 69–92.
2. Ingmar Schuster and Ilja Klebanov, *Markov chain importance sampling—a highly efficient estimator for mcmc*, Journal of Computational and Graphical Statistics **30** (2021), no. 2, 260–268.
3. Dootika Vats, *Lecture notes for markov chain monte carlo*, 2021.