Consistent Efficient Estimators via MCMC

Supervisor : Prof. Dootika Vats

April 13, 2024

Siddharth Pathak
Indian Institute of Technology Kanpur
siddharthp21@iitk.ac.in

Overview

Notations & Definitions[3]

Let $(\mathbb{S}, \Sigma)$ is a measurable space and $A \subseteq \mathbb{S}$. Formally the discrete-time continuous state space Markov chain $\{X_i\}_{i \geqslant 0}$ follows:

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

Definition (Markov 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.

Definition (Transition Density)

Let $P(x, .)$ 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

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

Notations & Definitions

Definition (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 $P^n(x, A) > 0$.

Definition (Harris Recurrence)

Let $A \in \Sigma$ and define $\tau_A = \inf\{n \geqslant 1 : X_n \in A\}$, τ_A is called the first return time to A. If $X_n \notin A$ for all $n \geqslant 1$, $\tau_A = \infty$. 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 $x \in \mathbb{S}$.

$$\mathbb{P}(\tau_A < \infty | X_0 = x) = 1$$

Rate of Convergence

Let $M : \mathbb{S} \mapsto \mathbb{R}^+$ and $\psi : \mathbb{N} \rightarrow [0, 1]$ be such that
$||P^n(x, .) - F(.)|| \leqslant M(x)\psi(n)$ for all x, n.
1. **Geometric Ergodicity:** $\psi(n) = t^n$ for some $0 \leqslant t < 1$.
2. **Uniform Ergodicity:** $\sup_x M(x) < \infty$ and $\psi(n) = t^n$ for some $0 \leqslant t < 1$.

Notations & Definitions

MCMC Accept-Reject Algorithm

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. 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

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**

Notations & Definitions

Unadjusted Langevin Algorithm

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:

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

Algorithm 2 Unadjusted Langevin Algorithm (ULA)

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

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

Importance Sampling Estimators

For $h : \mathcal{X} \mapsto R$, we want to estimate $\theta = \mathbb{E}_F[h(X)]$. Let G be a distribution with density g defined on $\mathcal{X}$ so that,

$$\theta = \mathbb{E}_F[h(X)] = \int_{\mathcal{X}} h(x)f(x)dx = \int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)} g(x)dx = \mathbb{E}_G\left[\frac{h(x)f(x)}{g(x)}\right]$$

If $Z_1, \ldots, Z_N \overset{IID}{\sim} G$, then an estimator of θ is

$$\hat{\theta}_g = \sum_{i=1}^{N} \frac{h(Z_i)f(Z_i)}{g(Z_i)}$$

Theorem

The importance sampling estimator $\hat{\theta}_g$ is unbiased for θ. Also it is consistent for θ. That is, as $N \to \infty$,

$$\hat{\theta}_g \xrightarrow{p} \theta.$$

Markov Chain Importance Sampling Estimator[2]

Let $\{Y_k\}_{k\in\mathbb{N}}$ be the proposed samples and the corresponding density function be ρ_Y. An asymptotically unbiased estimator of ρ_Y can be defined as

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

As the K increases, the size of chain $\{X_k\}_{k\geqslant 0}$increases, it converges to some distribution ρ_X (say). So,

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

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

Markov Chain Importance Sampling Estimator

MCIS Estimators

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

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

Augmented Chain $\{Z_k\}_{k\in\mathbb{N}}$ & It's Kernel

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 Markov chain. So, the kernel of $\{Z_k\}_{k\in\mathbb{N}}$ is a function $K_Z : \mathbb{R}^{2d} \times \mathcal{B}^{2d} \mapsto [0, 1]$ define as follows:

$$K_Z((x, y), A \times B) = (1 - \alpha(x, y))\mathbb{1}_A(x)q(B|x) + \alpha(x, y)\mathbb{1}_A(y)q(B|y)$$

Inference about Augmented Chain $\{Z_k\}_{k\in\mathbb{N}}$

Theorem ([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 $q(.|.)$ be globally supported and continuous in both arguments. 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 $\{Z_k\}_{k\in\mathbb{N}}$.*

For further analysis let us define $\phi(y)$ and $w(y)$ as following:

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

Law of Large Number for $S_K^{IS}(f)$ [2]

Theorem (LLN for $S_K^{IS}(f)$)

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

Proof

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

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

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

Central Limit Theorem for $S_K^{IS}(f)$ [2]

Let us define a function $h : \mathbb{R}^{2d} \mapsto \mathbb{R}^2$ such that $h(z) = h(x, y) = (\phi(y), w(y))^t$. Let us have the following assumptions:

- $\{X_k\}_{k \geqslant 1}$ 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

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

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

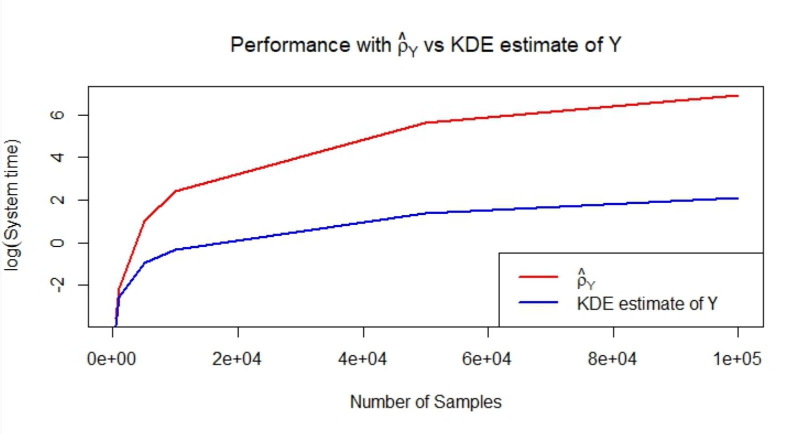
Theorem (CLT for $S_K^{IS}(f)$)

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

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

Lack of Efficiency

One must note that the estimator $\hat{S}_K^{IS}$ is not efficient because it is taking $\Theta(K^2)$ in computing the estimate. This estimator has no use in big data, and we have proved this fact computationally. One can have another estimator by replacing $\hat{\rho}(y)$ with the KDE estimate of y.



Test Results

We consider the density function $\rho = e^{-x^2}(1 + \sin(5x) + \sin(2x))$. In our experiment, we estimate the expectation of a random variable having density ρ. 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: For 10000 Samples

Posterior Inference via MCMC samples

It is expected to summarize posterior distributions by listing out $100(1-\alpha)\%$ posterior credible interval parameters of interest. 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.

Consistent Estimator for CDF[1]

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, obtaining an estimation of Bayesian credible or HPD interval. It is easy to observe that for a given θ^*

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

Consistent Estimator for CDF[1]

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

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

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 a notation. Denote

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

Estimator for CDF

for $1 \leqslant i \leqslant n$. We have

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

Estimation of Quantiles and HPD[1]

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

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

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

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

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 $j = 1, 2, \ldots, n - [(1-\alpha)n]$.

Theorem

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

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

Generalization

Corollary

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*

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

Using the above result, we can compute

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

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

Results

Example 1

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 being $N(0, 100)$ 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 $\frac{1}{(n/\sigma^2)+(1/100)}$.

Our analysis fixes $\mu = 0$ 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 mentioned above.

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

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

Example 1 Continued

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 (Proposed Estimator)	MSE (Naive Estimator)
0.25	$2.361802e-07$	$9.630869e-05$
0.50	$3.490802e-08$	$8.786842e-05$
0.75	$3.491659e-08$	$9.504091e-05$

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

Example 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 the prior to $Gamma(1, 1)$. Posterior for λ will be $Gamma(\sum_{i=1}^{n} Y_i + 1, n + 1)$.

Example 2 Continued

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: HPD Interval for $\pi(\lambda|Y)$

Results after repeating the experiments 1000 times each for $\alpha = 0.25, 0.5, 0.75$ are as following:

α	MSE (Proposed Estimator)	MSE (Naive Estimator)
0.25	$3.686399e-06$	0.001605519
0.50	$6.695139e-07$	0.001539897
0.75	$4.705776e-06$	0.001783654

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

References

- M.-H. Chen and Q.-M. Shao.

  Monte carlo estimation of bayesian credible and hpd intervals.

  Journal of Computational and Graphical Statistics, 8(1):69–92, 1999.

- I. Schuster and I. Klebanov.

  Markov chain importance sampling—a highly efficient estimator for mcmc.

  Journal of Computational and Graphical Statistics, 30(2):260–268, 2021.

- D. Vats.

  Lecture Notes for Markov Chain Monte Carlo.

  2021.

Thank You ☺!

Questions?