Nonparametric Kernel Density Estimation for the Metropolis-Hastings Algorithm[*]

Submitted by:

Arkajyoti Bhattacharjee [†‡]
Nitin Garg [§†]
Suchismita Roy [¶†]

Supervised by:

Dr. Dootika Vats [†]

Submitted on:

20th April, 2022

Abstract

In this report, we discuss how the rejection step of the Metropolis-Hastings algorithm affects kernel density estimation. We elaborate on the theory developed by Sköld and Roberts (2003) by providing extensive proofs and explore applications exhibiting their efficiency in various problems.

Contents

[*] This report has been prepared towards the partial fulfillment of the requirements of the course *MTH516A: Non-Parametric Inference.*

[†] Department of Mathematics & Statistics, Indian Institute of Kanpur, India.

[‡] 201277, M.Sc. Statistics (Final year).

[§] 180490, B.S. Statistics (Final year).

[¶] 201440, M.Sc. Statistics (Final year).

1 Introduction

Over the course of the 21^{st} century, the use of Markov Chain Monte Carlo (MCMC) algorithms has grown exponentially, with manifold applications in astronomy (Thrane and Talbot (2019), Sharma (2017)), health science (Sorensen et al. (2002), Vajargah et al. (2021)), cognitive science (Kim et al. (2003)), image compression, optimization (Mahendran et al. (2012), Ma et al. (2015)) and machine learning (Andrieu et al. (2003), Hensman et al. (2015)). It is a sampling methodology widely used in estimating expected values under complicated and high-dimensional distributions, which are known up to a normalizing constant (see Brooks et al. (2011), Liu and Liu (2001), Gilks et al. (1995)).

MCMC consists of two parts - *Markov chain* and *Monte Carlo*. The *Monte Carlo* methods are a broad class of computational algorithms used to compute closed form analytical solutions of complicated numerical integrals. For example, we may be interested in obtaining an analytical solution of

$$\int_{\pi}^{2\pi} e^{sin(log(x))cos(e^x)} dx.$$

Clearly, finding a closed form solution to this integral is difficult as no standard anti-derivative exists of the integrand. A Monte Carlo approach to this problem is to sample a large number of $\mathscr{U}(\pi, 2\pi)$ variables and compute the above integral as an expectation under the uniform distribution. Mathematically,

$$\int_{\pi}^{2\pi} e^{sin(log(x))cos(e^x)} dx = \pi \mathbb{E}_{\mathscr{U}}(e^{sin(log(X))cos(e^X)}) \hat{=} \frac{1}{N}\sum_{i=1}^{N} e^{sin(log(X_i))cos(e^{X_i})},$$

where $X_1, \ldots, X_N$ is a random sample drawn from $\mathscr{U}(\pi, 2\pi)$, $\mathbb{E}_{\mathscr{U}}$ is expectation under $\mathscr{U}(\pi, 2\pi)$ and "$\hat{=}$" means 'is estimated by'. The right hand side of the above equation holds due to the weak law of large numbers, assuming N is large enough. The Monte Carlo approach easily provides 3.16 as an 'estimated' solution of the otherwise intractable integral. Although other numerical integration techniques may be able to provide an 'approximate' solution, difficulty increases as the dimensionality increases, and MCMC is often the better alternative.

The *Markov chain* part of MCMC uses the Markovian property, which means that the proposed random value depends on the current value and not on the previous values of the sequential process (hence, 'chain').

MCMC is extensively used in Bayesian inference as it involves generating samples from complicated posterior distributions where the exact form of the likelihood is unknown or difficult to derive analytically. One of most popular MCMC algorithms is the *Metropolis-Hastings* (M-H) algorithm (Metropolis et al. (1953), Hastings (1970), Chib and Greenberg (1995), Robert and Casella (1999)). A survey (Beichl and Sullivan (2000)) placed the MH algorithm among the ten algorithms that have had the greatest influence on the development and practice of science and engineering in the 20^{th} century.

After generating the samples via MCMC algorithms, one is naturally interested in visualizations, like the sample path of the Markov chains, and posterior density estimation. There are two approaches to density estimation. In the *parametric* approach, we assume that the data has been drawn from a known parametric family of distributions, for example the normal distribution with mean μ and variance σ^2. The underlying density f can then be estimated by estimating the associated parameters from the observed data and substituting the estimated parameters in the assumed distribution. On the other hand, *nonparametric* density estimation is a much more flexible approach where no assumption is made about the underlying distribution of the observed data. There is a rich literature on density estimation (see Silverman (2018), Scott (2015), Sheather (2004)), of which kernel based methods are the most popular (Nadaraya (1964), Watson (1964), Priestley and Chao (1972), Gasser and Müller (1979)). For a review of nonparametric kernel density estimation (KDE) and its applications, the reader can look into Weglarczyk (2018), Gramacki (2018), Zambom and Ronaldo (2013).

The literature on KDE is majorly developed on independent data. Kernel estimators designed for independent data can lead to overly rough estimators on dependent data (Sköld and Roberts (2003)). However, generally for time series data and particularly, for Markov chain output, where transition densities with respect to Lebesgue measure are known to exist, the optimal bandwidths are asymptotically equal to those for independent data from the target distribution (see Yakowitz (1989), Hall et al. (1995)). But for the M-H algorithm the transition density does not exist, generally, because of the presence of a spike/point mass at the current state owing to the rejection of a proposed move. It is to be noted that the Gibbs sampler, which is a particular case of the M-H algorithm, has no rejection step and hence, it has a transition density. The point masses do affect the optimal smoothing parameter of the kernel estimator (Sköld and Roberts (2003)), as can be seen in Figure (1).

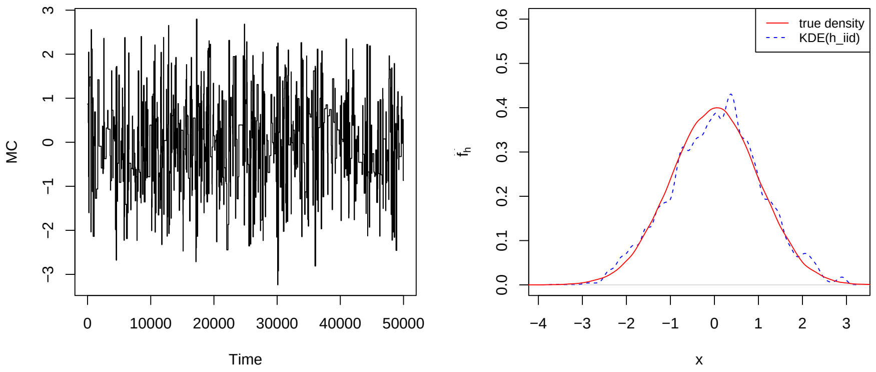

Figure 1: Plot showing that $h_{i.i.d.}$ indeed fails and hence, samples obtained via Metropolis-Hastings algorithm require special attention for KDE.

In this report, we review this effect and the modifications of the existing i.i.d. bandwidth selection procedures that are optimal for the M-H output. The rest of the report is organized as follows. In Section (2), we present an overview of KDE, measures of discrepancy for the estimators, and bandwidth selection. we present the Metropolis-Hastings algorithm in Section (3). In Section (4), we present the theory developed for KDE for samples generated by the Metropolis-Hastings algorithm. we also present two bandwidth selection methods, viz., plug-in method and bump-killing in Section (4.1). We finally present examples where we explore the practical aspects of the theory presented in the report.

1.1 Notations

Throughout the report, we will be using a few notations that we define here. If μ_1 and μ_2 are two probability measures, the *total variation* distance between them is defined as $||\mu_1 - \mu_2||_{TV} \overset{def}{=} \underset{A\in\mathscr{B}}{sup}|\mu_1(A) - \mu_2(A)|$, where $\mathscr{B}$ is the Borel σ-field on $\mathbb{R}$. If the measures admit densities π_1 and π_2 with respect to Lebesgue measure $||\mu_1 - \mu_2||_{TV} = \frac{1}{2}\int|\pi_1(x) - \pi_2(x)|dx$.

2 An overview of Kernel Density Estimation

Suppose we have n real data points $X_1, \ldots, X_n$ drawn from a population with underlying density f, which we are interested to estimate. The kernel density estimator of f, denoted by $\hat{f}$, with *kernel* function K and *bandwidth*, *window* or *smoothing parameter* $h > 0$ is defined by:

$$\hat{f}(x) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} K_h(x, X_i) \tag{1}$$

We assume $K_h(x,u)$ satisfies the following properties:

$$0 \leq K_h(x,u) < \infty \; \forall \; x, u \in \mathbb{R} \tag{2}$$

$$\int_{-\infty}^{\infty} K_h(x,u)dx = 1 \tag{3}$$

Property (2) ensures that

$$0 \leq \hat{f}(x) < \infty, \tag{4}$$

and Property (3) ensures that

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1} K_h(x,X_i)dx = \frac{1}{n}\sum_{i=1} \int_{-\infty}^{\infty} K_h(x,X_i)dx = 1. \tag{5}$$

Properties (4) and (5) imply that $\hat{f}$ defined in (1) is a probability density function (PDF).

The kernel K_h transforms the "bump" X_i into an interval centred, symmetrically or asymmetrically corresponding to symmetric or asymmetric kernel functions respectively, around X_i. The kernel estimator $\hat{f}$ is a sum of these "bumps", as can be seen in Figure 2. Some commonly used symmetric and asymmetric (Chen (2000), Scaillet (2004), Jin et al. (2003)) kernel functions are given in Table 1.
Symmetric kernel allows us to write equation (1) in the form:

$$\hat{f}_{sym}(x) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \tag{6}$$

The kernel function K governs the shape of the bumps, whereas the bandwidth h determines their width. Throughout the rest of the report, we will be using kernel estimators of the form defined in (6) and, abusing notation, denote it as $\hat{f}$.

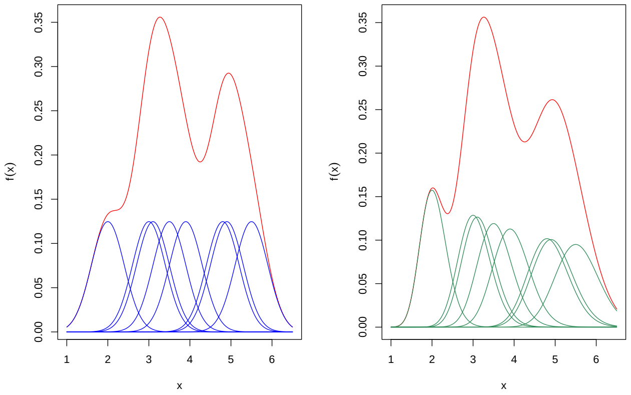

Figure 2: Kernel estimate showing the contributions of the symmetric Gaussian (left; blue) and the asymmetric Gamma (right; green) kernels evaluated for the individual observations with bandwidths $h = 0.4, 0.05$ respectively.

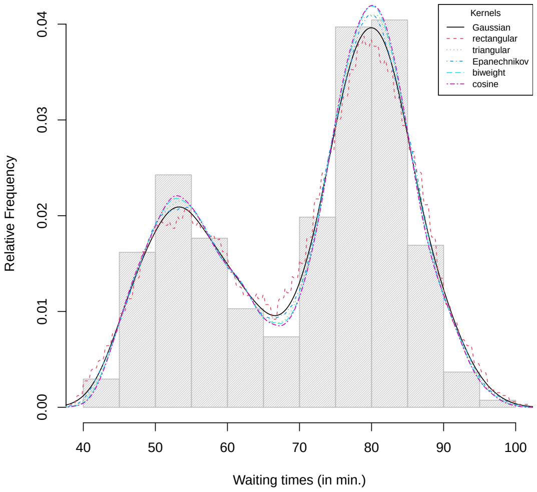

Figure 3: Density estimates of the Old Faithful Geyser eruption data based on common symmetric kernels imposed on a histogram of the data.

Type	Name	Definition
Symmetric	Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$
	rectangular	$\frac{1}{2}\mathbb{I}(\lvert x\rvert \le 1)$
	triangular	$(1-\lvert x\rvert)\mathbb{I}(\lvert x\rvert \le 1)$
	Epanechnikov	$\frac{3}{4}(1-x^2)\mathbb{I}(\lvert x\rvert \le 1)$
	biweight	$\frac{15}{16}(1-x^2)^2\mathbb{I}(\lvert x\rvert \le 1)$
	cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right)\mathbb{I}(\lvert x\rvert \le 1)$
Asymmetric	Gamma	$\frac{x^{u/h}e^{-x/h}}{h^{u/h+1}\Gamma(x/h+1)}\mathbb{I}(x>0)$
	Lognormal	$\frac{1}{x\sqrt{8\pi\ln(1+h)}}e^{-\frac{(\ln x-\ln u)^2}{8\ln(1+h)}}\mathbb{I}(x>0)$
	Inverse Gaussian	$\frac{1}{\sqrt{2\pi hx^3}}e^{-\frac{1}{2hu}(\frac{x}{u}-2+\frac{u}{x})}\mathbb{I}(x>0)$
	Reciprocal Inverse Gaussian	$\frac{1}{\sqrt{2\pi hx}}e^{-\frac{u-h}{2h}(\frac{x}{u-h}-2+\frac{u-h}{x})}\mathbb{I}(x>0)$

Table 1: Examples of common symmetric and asymmetric kernels. $\mathbb{I}(.)$ indicates the indicator function.

2.1 Measures of Discrepancy

Let $X_1, X_2, \ldots, X_n$ be the set of random samples of size n. We have defined the kernel density estimator in the equation (6). Now our aim is to calculate bias, Mean Square Error (MSE) and variance of the estimator.

2.1.1 Independent Data

Bias: We first find the bias in the estimator.

$$\begin{aligned}\mathbb{E}(\hat{f}(x)) &= \mathbb{E}\left(\frac{1}{nh}\sum_{i=1}^{n}K(\frac{X_i-x}{h})\right)\\ &= \frac{1}{nh}\times n\times\int K(\frac{u-x}{h})f(u)du \qquad \text{[Since } X_i\text{'s are identically distributed.]}\\ &= \int k(y)f(x+hy)dy. \qquad \text{[Taking } \frac{u-x}{h}=y.\text{]} \qquad (7)\end{aligned}$$

Using Taylor series expansion we have,

$$f(x+hy) = f(x) + yhf'(x) + \frac{h^2y^2}{2}f''(x) + o(h^2). \tag{8}$$

Thus using (8) in (7) we get,

$$\mathbb{E}(\hat{f}(x)) = f(x)\int k(y)dy + f'(x)h\int yk(y)dy + f''(x)\frac{h^2}{2}\int y^2k(y)dy + o(h^2)$$

$$= f(x) + f''(x)\frac{h^2}{2}\mu_{2,1} + o(h^2).$$

As $k(y)$ is a symmetric function $yk(y)$ is an odd function and the second term becomes zero. Thus, the bias is given by

$$\begin{aligned}\mathbb{B}(\hat{f}(x)) &= \mathbb{E}(\hat{f}(x)) - f(x)\\ &= f''(x)\frac{h^2}{2}\mu_{2,1} + o(h^2).\end{aligned}$$

Mean Square Error: Now, we will calculate the Mean Square Error of the KDE estimate. To do that, we first find the variance.

$$\begin{aligned}\mathbb{V}(\hat{f}(x)) &= \mathbb{V}\left(\frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h})\right)\\ &\leq \frac{1}{n^2h^2} \times n \times \mathbb{E}\left(K^2(\frac{X-x}{h})\right) \qquad \text{[Since } X_i\text{'s are identically distributed.]}\\ &= \frac{1}{nh^2}\int K^2(\frac{u-x}{h})f(u)du\\ &= \frac{1}{nh}\int K^2(y)f(x+hy)dy\\ &= \frac{1}{nh}\int K^2(y)[f(x) + yhf'(x) + o(h)]dy \qquad \text{[Using Taylor series expansion]}\\ &= \frac{1}{nh}f(x)\int K^2(y)dy + o(\frac{1}{nh})\\ &= \frac{1}{nh}f(x)\mu_{0,2} + o(\frac{1}{nh}).\end{aligned}$$

Using this, we can find the MSE

$$\begin{aligned}\mathbb{MSE}(\hat{f}(x)) &= V(\hat{f}(x)) + \mathbb{B}(\hat{f}(x))^2\\ &= \big(f''(x)\frac{h^2}{2}\mu_{2,1} + o(h^2)\big)^2 + \frac{1}{nh}f(x)\mu_{0,2} + o(\frac{1}{nh}).\end{aligned}$$

Now, let $o(h^2) = t(x)$ for some function $t(\cdot)$, such that, $\lim_{h\to 0}\frac{t(x)}{h^2} = 0$. Then,

$$\mathbb{B}\big(\hat{f}(x)\big)^2 = \frac{h^4}{4}\big(f''(x)\big)^2\mu_{2,1}^2 + t^2(x) + h^2f''(x)\mu_{2,1}^2t(x).$$

Now, $\lim\limits_{\substack{nh\to\infty\\ h\to 0}} \frac{t^{(}x)}{h^4} = \left(\lim\limits_{\substack{nh\to\infty\\ h\to 0}} \frac{t(x)}{h^2}\right)^2 = 0$ and $\lim\limits_{\substack{nh\to\infty\\ h\to 0}} \frac{h^2f''(x)\mu_{2,1}^2t(x)}{h^4} = f''(x)\mu_{2,1}^2 \times \lim\limits_{\substack{nh\to\infty\\ h\to 0}} \frac{f''(x)\mu_{2,1}^2t(x)}{h^2} = 0$. So, we get

$$\mathbb{B}\big(\hat{f}(x)\big)^2 = \frac{h^4}{4}\big(f''(x)\big)^2\mu_{2,1}^2 + o(h^4).$$

Hence, we get the final expression of MSE as

$$\mathbb{MSE}(\hat{f}(x)) = \frac{\mu_{0,2} f(x)}{nh} + \frac{1}{4}\big(f''(x)\big)^2 \mu_{2,1}^2 + o(h^4) + o(\frac{1}{nh}). \tag{9}$$

Additionally, the mean integrated square error is given by,

$$\begin{aligned} \mathbb{MISE}(\hat{f}(x)) &= \int \mathbb{MSE}\big(\hat{f}(x)\big)dx \\ &= \frac{\mu_{0,2}}{nh}\int f(x)dx + \frac{1}{4}h^4\mu_{2,1}^2 \int \big(f''(x)\big)^2 dx + o(h^4) + o\big(\frac{1}{nh}\big) \\ &= \frac{\mu_{0,2}}{nh} + \frac{1}{4}h^4\mu_{2,1}^2 ||f''||_2^2 + o(h^4) + o\big(\frac{1}{nh}\big) \end{aligned} \tag{10}$$

2.1.2 Time Series Data

The case of time series data i.e. dependent data has received considerable attention as well. Observe that the bias term is not affected due to the stationary assumption. To analyze variance, we consider the following additional assumptions:

Assumption 1. *Restricting the local dependence: It is assumed that* (X_i, X_{i+j}) *has a bounded bivariate density for all* $j > 0$.

Assumption 2. *Restricting the long-range dependence: It is assumed that the process satisfies a certain mixing condition and that the mixing coefficients decay at a sufficiently fast rate.*

Assumption 3. *The sequence is stationary and ergodic.*

The last assumption is to make sure that we can construct a consistent estimate. The general result provided in the literature is that, given assumptions 1–3, the asymptotic expansions of MSE and MISE agree with (9)–(10) for independent data and the bandwidth sequence in (11) remains optimal for dependent data. This might seem surprising at first as it is certainly not true for the empirical distribution function, for example.

2.2 Bandwidth Selection

The kernel function does not affect density estimation as much as the bandwidth does, as can be seen from Figures (3) and (4). So, there is an extensive literature present on bandwidth selection. The interested reader can see Chapter 3 of Wand and Jones (1994), Jones et al. (1996) for a review of the different bandwidth selection methods. In this report, we consider bandwidth selection by minimizing the asymptotic MISE (AMISE). For the i.i.d case, by minimizing the MISE (10) after taking $n \to \infty$ and $h \to 0$ such that $nh \to \infty$, we find the asymptotically optimal bandwidth as:

$$h_{\text{i.i.d}} := \left[\frac{\mu_{0,2}}{\mu_{2,1}^2 ||\pi''||_2^2 n} \right]^{1/5}. \tag{11}$$

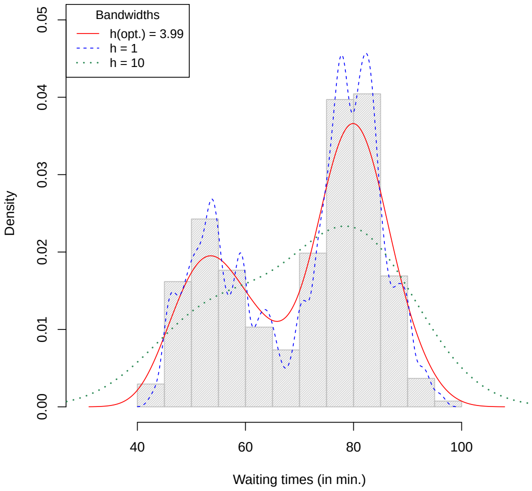

Figure 4: Density estimates of the Old Faithful Geyser eruption data based on different bandwidths imposed on a histogram of the data.

Most of the available methods have been verified to retain optimality criteria when data are dependent, but only under the essential assumption of smooth transition densities. That optimal bandwidth for samples obtained via the Metropolis-Hastings algorithm requires separate attention is already motivated and evidenced by Figure (1).

3 The Metropolis-Hastings algorithm

Let $f(\cdot)$ be the target density function. The aim of the M-H algorithm is to propose values from a proposal distribution $Q(x,.)$, x being the current state of the Markov chain, and store them sequentially in a Markov chain if they are accepted with acceptance probability $\alpha(x,y) = min\{1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\}$. $Q(x,.)$ is chosen such that it is the invariant distribution of f. The M-H algorithm is given by in Algorithm 1.

Algorithm 1 Metropolis-Hastings algorithm

```
Let $X_n = x$. To obtain $X_{n+1}$:

    1. $Y \sim Q(x,.)$ and independently $U \sim \mathscr{U}(0,1)$.
    2. If $U < \alpha(x,y) = min\left\{1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\right\}$,
           set $X_{n+1} = y$.
    3. Else
           set $X_{n+1} = x$.
```

It is to be noted that $q(x,.)$ in Algorithm 1 is the proposal density corresponding to $Q(x,.)$. This algorithm produces $X_0, X_1, \ldots$ that from a Markov chain with stationary distribution f and the transition kernel,

$$P_x(A) := P(X_{i+1} \in A | X_i = x) = \int_A \alpha(x,y)q(x,y)dy + r(x)\mathbf{1}_{\{x \in A\}}$$

where $r(x)$ is the probability of rejecting a move which is given by, $r(x) = 1 - a(x) := \int [1-\alpha(x,y)]q(x,y)dy$. Further, the i^{th} step transition kernel is given by,

$$P_x^{(i)}(A) := P(X_{j+i} \in A | X_i = x) = \int_A \tilde{p}_x^{(i)}(y)dy + r(x)^i \mathbf{1}_{\{x \in A\}}$$

Here $\tilde{p}_x^i$ is the absolutely continuous part of the transition kernel, $P_x^{(i)}$.

4 KDE for the M-H algorithm

In this section, we will derive the expressions for the bias and variance of the estimator in the case of samples from M-H algorithm. Sköld and Roberts (2003) provide two theorem (one local and the other global) to show that the optimal bandwidth in this case can be written as a multiple of the expression in (11), where the multiplier can be expressed in terms of the acceptance probabilities of the algorithm. The results of theorems 1 & 2 can also be applied to the Gibbs sampling algorithm by taking $r(x) \equiv 0$.

Theorem 1. *Sköld and Roberts (2003) Fix $u \in \mathbb{R}$ and suppose there are functions $V : \mathbb{R} \to \mathbb{R}^+$, $R : \mathbb{N} \to (0,1)$ and constants $\varepsilon > 0$, $M < \infty$ such that uniformly for $x \in [u-\varepsilon, u+\varepsilon]$ and for $i = 0, 1 \ldots$.*

Assumption L1. $|\tilde{p}_y^{(i)} - f(x)| < V(y)R(i)$ *and* $\sum_{i=0}^{\infty} R(i) < M$,

Assumption L2. *$f(x)$, $\frac{1}{a(x)}$, $p^{(i)}(x)$, $V(x)$, $\mathbb{E}[V(X_0)]$ all are bounded by M,*

Assumption L3. *a(x) and f(x) are uniformly continuous.*

Then

$$\mathbb{V}\big(\hat{f}(u)\big) = \left(\frac{2}{a(u)} - 1\right)\frac{\mu_{0,2} f(u)}{nh} + o\left(\frac{1}{nh}\right) \qquad \text{as } n \to \infty \text{ and } h \to 0.$$

If we further assume,

Assumption L4. *$f(x)$ has a bounded third derivative in $x \in [u-\varepsilon, u+\varepsilon]$,*

we get the asymptotic bias,

$$\mathbb{E}(\hat{f}(x)) - f(u)) = \frac{1}{2}h^2\mu_{2,1} f''(u) + o\left(\frac{1}{n}\right) + o(h^2), \qquad \text{as } n \to \infty \text{ and } h \to 0.$$

To show that a M-H chain satisfies these assumptions, we need to discuss a little bit on convergence rate. If μ_1 and μ_2 are two probability measures, we can define the total variation distance between them as,

$$||\mu_1 - \mu_2|| := \sup_{A \in \mathscr{B}} |\mu_1(A) - \mu_2(A)| \tag{12}$$

where, $\mathscr{B}$ denotes the Borel σ algebra on $\mathbb{R}$. If μ_1 and μ_2 admit densities π_1 and π_2 respectively with respect to the Lebesgue measure, we will have,

$$||\mu_1 - \mu_2|| = \frac{1}{2}\int |\pi_1(x) - \pi_2(x)| dx.$$

Due to rejection step, transition density of the M-H chain with respect to Lebesgue measure does not exist but it exists for the smooth part. Thus, we can write

$$D(x,i) := ||P_x^{(i)} - f|| = \frac{1}{2}\int |\tilde{p}_x^{(i)}(y) - f(y)| dy + \frac{1}{2} r(x)^i.$$

We now define three different rates of convergence seen in Markov chains. Let $M : \mathscr{X} \to \mathbb{R}^+$ and $\chi : \mathbb{N} \to [0,1]$ be such that,

$$D(x,n) \le M(x)\chi(n) \text{ for all } x, n.$$

Then

1. If $\chi(n) = n^{-k}$ for some $k > 0$, the chain is *polynomially ergodic* at rate k.

2. If $\chi(n) = t^n$ for some $0 \le t < 1$, the chain is *geometrically ergodic*.

3. If $\sup_x M(x) < \infty$, then we say the chain is *uniformly ergodic*.

We can argue using these convergence rates that the assumptions are satisfied. Using ergodicity and uniform continuity of $f(x)$ and $\tilde{p}_y^{(i)}(x)$, we can prove that $R(i) \to 0$ and the conditions of ergodicity ensure that $\sum_i R(i) < M$. To get a bound on $p^{(i)}$, we need to start the chain from a density, $p^{(0)}$, that is bounded in the interval, $[u-\varepsilon, u+\varepsilon]$. We need the global assumption that $\mathbb{E}[V(X_0)] < M$ for the theorem to be valid. We can make it local by, for example, selecting X_0 to be supported on $[u-\varepsilon, u+\varepsilon]$, and the simulation algorithm can perform arbitrarily badly outside this interval without affecting the density estimator at u.

Theorem 2. *Sköld and Roberts (2003)*
Suppose there are functions $V : \mathbb{R} \mapsto \mathbb{R}^+$, $R : \mathbb{N} \mapsto (0,1)$ and constants $\varepsilon > 0$, $M < \infty$ such that uniformly for $(x,y) \in \mathbb{R}^2$ and for $i = 0,1,\ldots$

Assumption G1. *$\int \frac{|\tilde{p}_y^{(i)}(x) - f(x)|}{a(x)} dx \le V(y)R(i)$ and $\sum_{i=0} R^{1-\varepsilon} < M$.*

Assumption G2. *$p^{(i)}(x)$, $f(x)$, $\tilde{p}_y^{(i)}(x)$, $E[V(X_i)]$ and $E[\frac{1}{a^2(X_i)}]$ are bounded by M for $x \in \mathbb{R}$ and $\frac{1}{a(x)} < M$ on the support of $p^{(0)}$.*

Assumption G3. *$f^{(3)}(x)^2$ is bounded by an integrable function which is monotone for large enough $|x|$.*

Under Assumptions G1 and G2,

$$\begin{aligned}\int Var[\hat{f}(u)]du &= A\frac{\mu_{0,2}}{nh} + o\left(\frac{1}{nh}\right) \\ &= A\int Var[\hat{f}_{iid}(u)]du \quad \text{as } n \to \infty \quad \text{and} \quad h \to 0.\end{aligned}$$

where, $A = \left(\mathbb{E}\left[\frac{2}{a(X)}\right] - 1\right)$.
Here X is a random variable with density f.
Additionally, under Assumption G3, we get the asymptotic integrated squared bias

$$\int \mathbb{E}[\hat{f}(x) - f(x)]^2 dx = \frac{1}{4}h^4\mu_{2,1}^2 ||f''(u)||_2^2 + O\left(\frac{1}{n}\right) + o(h^2).$$

Mean Integrated Square Error is given by,

$$MISE = A\frac{\mu_{0,2}}{nh} + o\left(\frac{1}{nh}\right) + \frac{1}{4}h^4\mu_{2,1}^2 ||f''(u)||_2^2 + O\left(\frac{1}{n}\right) + o(h^2).$$

Our concern is to show that the assumptions are satisfied by samples obtained from the M-H algorithm. Roberts and Tweedie (1996) show that if a is not bounded away from zero, the resulting chain cannot be geometric ergodic. Thus for the term a^{-1}, the factor V is going to be multiplied by a constant factor if R decreases to zero at a geometric rate. It is proved in literature that $\mathbb{E}[a(X_i)^{-1}] < \infty$ is required to achieve convergence in CLTs as the usual rate of $(n^{-1/2})$. If these assumptions are not satisfied, then the algorithm is inefficient, so we need to change the proposal then.

4.1 Bandwidth Selection for MH

The expression for Bias is unchanged in case of samples from the MH algorithm whereas the Variance term has an extra A term multiplied to it. This leads to a change in the expression for the optimal bandwidth. We discuss two methods for bandwidth selection as proposed by Sköld and Roberts (2003).

4.1.1 Plug-in method

On minimising MISE obtained in the case of MH samples, we get the the optimal bandwidth as

$$h_{\text{M-H}} := \left[\frac{A\mu_{0,2}}{\mu_{2,1}^2 ||f''||_2^2 n} \right]^{1/5} = A^{1/5} h_{\text{i.i.d}}.$$

Using the above bandwidth $h_{\text{M-H}}$, we get the asymptotic MISE

$$\inf_{h>0} \text{AMISE}(\hat{p}_h) = \text{AMISE}(\hat{p}_{h_{\text{M-H}}}) = \frac{5}{4} \frac{(\mu_{2,1}^2 ||f''||_2^2 A^4 \mu_{0,2}^4)^{1/5}}{n^{4/5}}.$$

We can observe that this AMISE is the smallest possible using some kernel K. Also, if $h_{\text{i.i.d}}$ is used as our bandwidth then the asymptotic MISE increases by a factor of

$$(4A+1)/(5A^{4/5}) \geq 1.$$

So, when we use an optimal acceptance rate i.e., $E[a(X)] \approx 1/4$ there is at least a 22 percent increase of AMISE if $h_{\text{i.i.d}}$ in place of $h_{\text{M-H}}$. We use the following steps to implement a kernel density estimator based on $h_{\text{M-H}}$. A and $||\pi''||_2^2$ need to be estimated in the expression for $h_{\text{M-H}}$. First, we estimate A by defining T_i as the time the process spends at a point X_i after having visited X_{i-1}, i.e.

$$T_i := \sum_{j=i}^{n-1} \mathbb{I}_{\{X_i = X_j\}}.$$

Now, $T_i|X_i$ has a truncated geometric distribution which gives us $P(T_i = k|X_i) = r(X_i)^{k-1}a(X_i)$ for $k = 1,\ldots,n-i-1$ and $P(T_i = n-i|X_i) = r(X_i)^{n-i-1}$. We can then write

$$E[T_i|X_i] = \frac{1}{a(X_i)} + O_P[r(X_i)^{n-i}]$$

Using this, an unbiased estimator of A can be given by

$$\hat{A} = \frac{1}{n}\sum_{i=0}^{n-1}(2T_i - 1) \tag{13}$$

This is a consistent estimator of A as if all moves are accepted, then $\hat{A} \equiv 1$.
For estimating $||\pi^{''}||_2^2$ quickly, we can use a normal scale rule which just finds the value of $||\pi^{''}||_2^2$ for the normal density with the same variance as the sample, i.e. we use $||\phi_{\hat{\sigma}}^{''}||_2^2$ as the estimate where $\phi_\sigma \equiv N(0,\sigma^2)$. We can use the result that

$$||\phi_\sigma^{(k)}||_2^2 = \frac{(2k)!}{(2\sigma)^{2k+1}k!\pi^{1/2}} \tag{14}$$

to get the final estimate. This gives the normal-scale rule bandwidth

$$\hat{h}_{\text{ns}} := 2\hat{\sigma}\left[\frac{\pi^{1/2}\hat{A}\mu_{0,2}}{12\mu_{2,1}^2 n}\right]^{1/5}.$$

This gives us a quick and *ad hoc* way of estimating the bandwidth parameter but this does not work well for densities that do not have a shape close to that of the normal distribution. Hence, a non-parametric estimate of $||\pi^{''}||_2^2$ is used instead. We make the following additional assumption.

Assumption 1. *The kernel K has six non-zero derivatives and that* $(-1)^k K^{(2k)}(0) > 0$ *for* $k = 2,3$.

These assumptions are satisfied by most kernels. If π has $2k$ bounded derivatives then

$$||\pi^{(k)}||_2^2 = (-1)^k \int \pi^{(2k)}(x)\pi(x)dx = (-1)^k E\left[\pi^{(2k)}(X)\right] =: I_k$$

An estimator of I_k can be given by

$$\hat{I}_k := \frac{(-1)^k}{n}\sum_{i=0}^{n-1}\hat{p}_{g_k}^{(2k)}(X_i) = \frac{(-1)^k}{n^2 g_k^{2k+1}}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}K^{(2k)}\left[\frac{X_i - X_j}{g_k}\right]. \tag{15}$$

We then find the bias of this estimator, first define $S_i = \sum_{j=0}^{n-1} \mathbf{1}_{\{X_i = X_j\}}$ then $E(S_i) = E(2T_i - 1) = A + O(E[r^{n-i}(X_i)])$. The bias can now be calculated as

$$\begin{aligned}
E(\hat{I}_k) - I_k = E\left[\sum_{i=0}^{n-1} S_i\right] \frac{(-1)^k K^{(2k)}(0)}{n^2 g_k^{2k+1}} + \frac{1}{n^2 g_k^{2k+1}} \\
\times E\left\{\sum_{i,j;X_i \neq X_j} K^{(2k)}\left[\frac{X_i - X_j}{g_k}\right]\right\} - I_k \\
= \frac{(-1)^k A K^{(2k)}(0)}{n^2 g_k^{2k+1}} - \frac{g_k^2}{2}\mu_{2,1} I_{k+1} + o(g_k^2) + O\left(\frac{1}{n}\right)
\end{aligned}$$

The main terms of the bias can be made to cancel by choosing

$$g_k := \left|\frac{2AK^{(2k)}(0)}{\mu_{2,1} I_{k+1} n}\right|^{1/(2k+3)}. \tag{16}$$

The following step-wise process is used to implement $h_{\text{M-H}}$:

Step 1. Estimate A by $\hat{A}$ as in (13).

Step 2. Estimate σ by the sample standard deviation $\hat{\sigma}$ and use a normal scale rule to estimate I_4 using (14), i.e.

$$\tilde{I}_4 := \frac{8!}{(2\hat{\sigma})^9 4! \pi^{1/2}}.$$

Step 3. Estimate I_3 using $\hat{I}_k$ as in (15) using bandwidth $\hat{g}_3$ as in (16).

Step 4. Estimate I_4 using $\hat{I}_k$ as in (15) using bandwidth $\hat{g}_2$ as in (16).

Step 5. Estimate $h_{\text{M-H}}$ using

$$\hat{h}_{\text{M-H}} := \left[\frac{\hat{A}\mu_{0,2}}{\mu_{2,1}^2 \hat{I}_2 n}\right]^{1/5}. \tag{17}$$

Note: Computing $\hat{I}_k$ involves $O(n^2)$ calculations making the process computationally intensive for large n. So for large n, we use an approximation based on linear *binning* given in appendix D of Wand and Jones (1994) to speed up the computations.

4.1.2 Bump Killing

Often the rejection step in the MH algorithm can produce false modes in the KDE estimate, especially if the acceptance probability goes to zero in the tail of the target distribution π. This problem isn't corrected by using the correction factor A like in $h_{\text{M-H}}$ as A is generally too large in the main support of the density and too small in the tails. This hints towards the requirement of a local correction factor. To do this, a simple and *ad hoc* method is used that effectively kills bumps that are introduced by long rejection periods. We use a different bandwidth for each observation. The modified kernel density estimator is defined as

$$\tilde{p}(u) := \frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{h_{\text{bk}}(i)}K\left[\frac{X_i-u}{h_{\text{bk}}(i)}\right],$$

where

$$h_{\text{bk}}(i) := (2T_i-1)^{1/5}h_{\text{i.i.d}}. \tag{18}$$

The motivation for using this estimator can be understood by observing that $h_{\text{M-H}} = A^{1/5}h_{\text{i.i.d}} = [2E(T_i)-1]^{1/5}h_{\text{i.i.d}}$ is the optimal global value, and $h_{\text{bk}}(\cdot)$ spreads the mass of the replicates.
Note: By Jensen's inequality, we get that

$$E[(2T_i-1)^{1/5}]h_{\text{i.i.d}} \geq [2E(T_i)-1]^{1/5}h_{\text{i.i.d}}$$

and using $h_{\text{bk}}(\cdot)$ will over-smooth on average. We can implement this by replacing step 5 in the implementation of $\hat{h}_{\text{M-H}}$ with

Step 5. Estimate $h_{\text{bk}}(i)$ using

$$\hat{h}_{\text{bk}}(i) := \left[\frac{(2T_i-1)\mu_{0,2}}{\mu_{2,1}^2\hat{I}_2 n}\right]^{1/5}. \tag{19}$$

5 Applications

In this section, we look at three examples showing the effectiveness of h_{mh} and h_{bk} discussed in this report.

5.1 Problem 1

In this example, we revisit Figure (1). Here, our target distribution is the standard Gaussian distribution and we use a Gaussian random walk proposal with mean

10 and standard deviation 100. We obtain a Markov chain of length 10^5 using Metropolis-Hastings algorithm. We have used the `density` function in R to obtain $\hat{h}_{iid} = 0.09$. We note that the `density` function uses a standard Gaussian distribution as the kernel function and calculates h_{iid} using the expression obtained by minimizing the AMSE for a standard Gaussian kernel. We then use $\hat{A}$ in (13) to find $\hat{h}_{MH} = 0.25$ using (17). To calculate h_{bk}, we use equation (19).

From Figure (5), it is evident that h_{iid} fails to efficiently estimate the true target density and produces a false peak near the mode of the target distribution. h_{MH}, on the other hand, successfully estimates the underlying density and produces a smoother curve, even towards the tails, where h_{iid} produces small bumps.

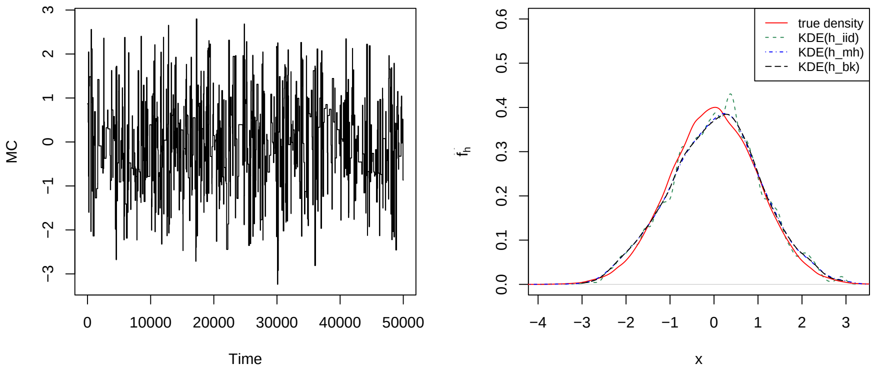

Figure 5: Trace plot (left) and KDEs (right) based on MH samples generated using $\mathscr{N}(10, 100)$ for the target $\mathscr{N}(0, 1)$. Clearly, h_{iid} is a poor smoothing parameter and h_{mh} is more effective.

5.2 Problem 2

In this example, we exhibit the effectiveness of bump-killing. Here, our target distribution is the gamma distribution with shape parameter 3 and scale parameter 1. We use a gamma proposal with shape parameter 3 and scale parameter 1.7. We obtain a Markov chain of length 10^5 using Metropolis-Hastings algorithm. As done in the previous example, we have used the `density` function in R to obtain $\hat{h}_{iid} = 0.15$. We then use $\hat{A}$ in (13) to find $\hat{h}_{MH} = 0.28$ using (17). To calculate h_{bk}, we use equation (19).

From Figure (6), it is evident that h_{iid} fails to efficiently estimate the tails of true target density and produces a false mode near the tail. Although h_{MH} smoothens the bump to some extent, there is still a non-existent mode present in the estimate. This is due to long rejection periods in the MH chain, as can be seen from the tails of the trace plot. However, h_{bk} effectively kills the bump and produces a much smoother plot that is much closer to the true density.

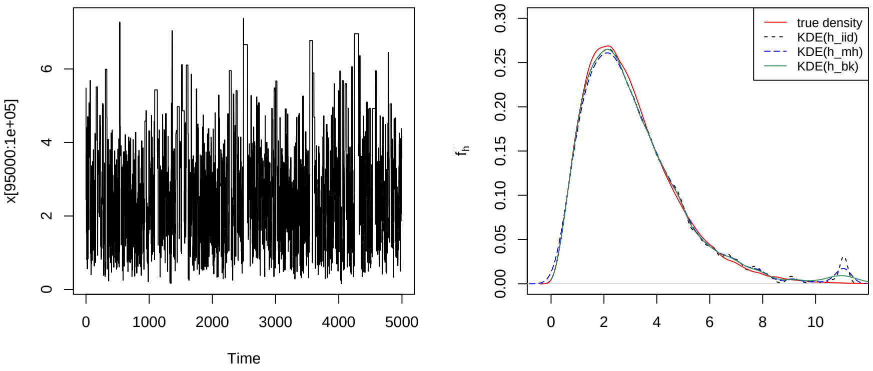

Figure 6: Plot of true density and KDE of M-H samples with proposal $q(y,x) \propto x^2 e^{-1.7x}$ for target $f(x) = x^2 e^{-x}/2$ based on h_{iid}, h_{mh} and h_{bk}.

5.3 Problem 3

In this final example, we take our target distribution to be the standard Gaussian distribution and we use a skew-normal distribution with location parameter 0, scale parameter 0.54 and shape parameter 10 as our proposal. The pdf of the skew-normal distribution, with location parameter ξ, scale parameter ω, and shape parameter α is:

$$f(x;\xi,\omega,\alpha) = \left(\frac{2}{\omega\sqrt{2\pi}} e^{-\frac{(x-\xi)^2}{2\omega^2}} \int_{-\infty}^{\alpha\left(\frac{x-\xi}{\omega}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt \right) \mathbb{I}(x \in \mathbb{R}).$$

We obtain a Markov chain of length $10^5 5$ using Metropolis-Hastings algorithm. We then proceed similarly as in the previous examples to obtain $\hat{h}_{iid} = 0.05$ and $\hat{h}_{MH} = 0.17$.

It can be seen from Figure (7) that h_{iid} fails to efficiently estimate the true target density density and produces a false peak near the tail of the target distribution.

h_{MH} provides a smoother curve, but yet produces a false mode at the tail. h_{bk} effectively smoothens outs the bump, thus, again validating its effectiveness when there are long rejection periods near the tails of the target distribution, which is evident from the trace plot. However, we observe that there is still room for improvements as the proposal, having a high density at the mode of the target, results in over estimating the density near the mode of the target distribution. we state that our motive here was solely showing the effectiveness of bump-killing and not producing the best possible estimate of the true density.

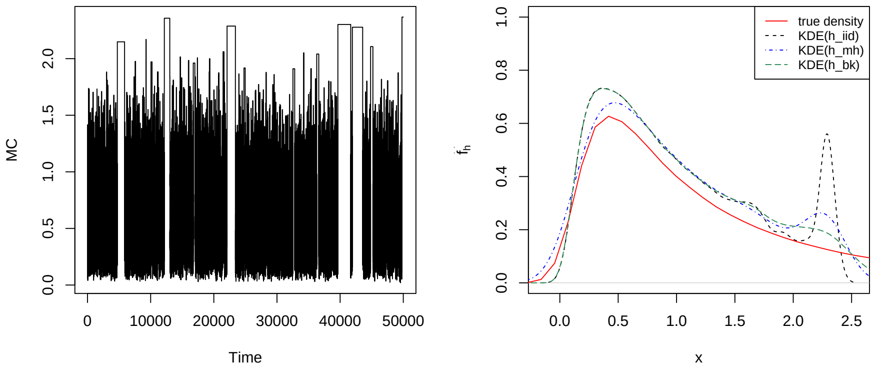

Figure 7: Trace plot (left) and KDEs (right) based on MH samples generated using skew-$\mathcal{N}(0, 0.54, 10)$ for the target log-$\mathcal{N}(0, 1)$. Bump-killing effectively kills the bumps in h_{mh} and is quite smoother than h_{iid}.

6 Supplementary Material

The interested reader is directed to https://github.com/ArkaB-DS/NDE4MH which contains all the figures present here in the directory `images` and the corresponding codes to generate them in the `R` directory.

7 Acknowledgements

We take this opportunity to heartily thank our supervisor Prof. Dootika Vats for her valuable feedback and constant guidance on this project.

References

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, 50(1):5–43.

Beichl, I. and Sullivan, F. (2000). The metropolis algorithm. *Computing in Science & Engineering*, 2(1):65–69.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Chen, S. X. (2000). Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52(3):471–480.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation*, pages 23–68. Springer.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Gramacki, A. (2018). *Nonparametric kernel density estimation and its computational aspects*, volume 37. Springer.

Hall, P., Lahiri, S. N., and Truong, Y. K. (1995). On bandwidth choice for density estimation with dependent data. *The Annals of Statistics*, 23(6):2241–2263.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hensman, J., Matthews, A. G. d. G., Filippone, M., and Ghahramani, Z. (2015). Mcmc for variationally sparse Gaussian processes. *arXiv preprint arXiv:1506.04000*.

Jin, X., Kawczak, J., et al. (2003). Birnbaum-saunders and lognormal kernel estimators for modelling durations in high frequency financial data. *Annals of Economics and Finance*, 4:103–124.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association*, 91(433):401–407.

Kim, W., Navarro, D. J., Pitt, M. A., Myung, I. J., Thrun, S., and Saul, L. (2003). An MCMC-Based Method of Comparing Connectionist Models in Cognitive Science. In *NIPS*, pages 937–944. Citeseer.

Liu, J. S. and Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*, volume 10. Springer.

Ma, Y.-A., Chen, T., and Fox, E. B. (2015). A complete recipe for stochastic gradient mcmc. *arXiv preprint arXiv:1506.04696*.

Mahendran, N., Wang, Z., Hamze, F., and De Freitas, N. (2012). Adaptive MCMC with Bayesian optimization. In *Artificial Intelligence and Statistics*, pages 751–760. PMLR.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Priestley, M. B. and Chao, M. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(3):385–392.

Robert, C. P. and Casella, G. (1999). The metropolis—hastings algorithm. In *Monte Carlo Statistical Methods*, pages 231–283. Springer.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83:95–110.

Scaillet, O. (2004). Density estimation using inverse and reciprocal inverse gaussian kernels. *Nonparametric statistics*, 16(1-2):217–226.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Sharma, S. (2017). Markov chain Monte Carlo methods for Bayesian data analysis in astronomy. *Annual Review of Astronomy and Astrophysics*, 55:213–259.

Sheather, S. J. (2004). Density estimation. *Statistical science*, pages 588–597.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Sköld, M. and Roberts, G. O. (2003). Density estimation for the metropolis–hastings algorithm. *Scandinavian journal of statistics*, 30(4):699–718.

Sorensen, D., Gianola, D., et al. (2002). Likelihood, bayesian, and mcmc methods in quantitative genetics. *Likelihood, Bayesian, and MCMC methods in quantitative genetics.*

Thrane, E. and Talbot, C. (2019). An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models. *Publications of the Astronomical Society of Australia*, 36.

Vajargah, K. F., Benis, S. G., and Golshan, H. M. (2021). Detection of the quality of vital signals by the Monte Carlo Markov Chain (mcmc) method and noise deleting. *Health Information Science and Systems*, 9(1):1–10.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. CRC press.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Weglarczyk, S. (2018). Kernel density estimation and its application. In *ITM Web of Conferences*, volume 23, page 00037. EDP Sciences.

Yakowitz, S. (1989). Nonparametric density and regression estimation for markov sequences without mixing assumptions. *Journal of Multivariate Analysis*, 30(1):124–136.

Zambom, A. Z. and Ronaldo, D. (2013). A review of kernel density estimation with applications to econometrics. *International Econometric Review*, 5(1):20–42.

8 Appendix: Proofs

Proof of Theorem 1:

Proof. Let K be a compactly supported and bounded Borel function, and $g(x)$ be integrable in a neighbourhood of x. Let us define the following:

$$K_h(\cdot) = \frac{K(\cdot/h)}{h},$$

$$\text{And} \qquad K_h * g(u) \to g(u) \int K(x)dx, \qquad \text{as } h \to 0 \text{ for all continuity points } u \text{ of } g.$$

We assume g is bounded, so is $K_h * g$ by standard properties of the convolution operator. Furthermore, if $\int K_h(x)dx \leq 1$ the bounding constant is the same. We have assumed that the difference $|f(x) - p^{(i)}(x)|$. We can write,

$$\begin{aligned} P^{(i)}(A) = \mathbb{E}[P^{(i)}_{X_0}(A)] &= \mathbb{E}\int_A \tilde{p}^{(i)}_{X_0}(y)dy + \mathbb{E}[r(X_0)^i \mathbb{I}_{\{x \in A\}}] \\ &= \int_A \mathbb{E}\big(\tilde{p}^{(i)}_{X_0}(x)dx\big) + \int_A r(x^i)p^{(0)}(x)dx \\ &= \int_A [\mathbb{E}\big(\tilde{p}^{(i)}_{X_0}(x)dx\big) + r(x^i)p^{(0)}(x)]dx \\ &= \int_A p^{(i)}(x)dx. \end{aligned}$$

Thus we can write the density a $p^{(i)}$ as,

$$p^{(i)}(x) = \mathbb{E}\big(\tilde{p}^{(i)}_{X_0}(x)dx\big) + r(x^i)p^{(0)}(x) \tag{20}$$

Thus we can find a bound as follows,

$$\begin{aligned} |f(x) - p^{(i)}(x)| &= |f(x) - \mathbb{E}\big(\tilde{p}^{(i)}_{X_0}(x)dx\big) - r(x^i)p^{(0)}(x)| \\ &\leq |f(x) - \mathbb{E}\big(\tilde{p}^{(i)}_{X_0}(x)dx\big)| + |r(x^i)p^{(0)}(x)| \\ &= \mathbb{E}|f(x) - \tilde{p}^{(i)}_{X_0}(x)dx| + r(x^i)p^{(0)}(x) \end{aligned} \tag{21}$$

In the last line, we have used that x being fixed and $f(x)$ is constant as we have taken the expectation w.r.t. the variable X_0.
Without loss of generality, let us assume u = 0. We have,

$$Var\big(\hat{f}(0)\big) = \frac{1}{n^2}\sum_{i=0}^{n-1} \mathbb{V}[K_h(X_i)] + \frac{2}{n^2}\sum_{i=0}^{n-1}\sum_{i=i+1}^{n-1} \mathbb{C}ov[K_h(X_i), K_h(X_j)] \tag{22}$$

We define a function $g : \mathbb{N} \times \mathbb{N} \times \mathbb{R}^2 \to \mathbb{R}$ as follows,

$$g_{i,j}(x,y) \coloneqq p^{(i)}(x)\tilde{p}^{(j-i)}_x(y) - p^{(i)}(x)p^{(j)}(y) \tag{23}$$

This is the difference of the probability of reaching x at i^{th} step and at y in next $(j-i)$ using smooth part of MH when we accepted the draw at each step and the product of the probabilities of reaching x at i^{th} step and y at j^{th} step using MH chain. Thus the function describes the long-range dependence and we have assumed time series assumption $\tilde{p}^{(j-i)}_x(y) - p^{(j)}(y) \to 0$ sufficiently fast which is also ensured by L1. Now consider,

$$\mathbb{C}ov[K_h(X_i), K_h(X_j)] = \mathbb{E}[K_h(X_i)K_h(X_j)] - \mathbb{E}[K_h(X_i)]\mathbb{E}[K_h(X_j)]$$

$$
\begin{aligned}
&= \mathbb{E}[K_h(X_i)K_h(X_j)\mathbb{I}(X_i \neq X_j)] - \mathbb{E}[K_h(X_i)]\mathbb{E}[K_h(X_j)] \\
&+ \mathbb{E}[K_h(X_i)K_h(X_j)\mathbb{I}(X_i = X_j)] \\
&= \int\int_{x\neq y} K_h(x)K_h(y)p^{(i)}\tilde{p}_x^{(j-i)}(y)dxdy - \mathbb{E}[K_h(X_i)]\mathbb{E}[K_h(X_j)] \\
&+ \mathbb{E}\bigg(K_h^2(X_i)\mathbb{E}[\mathbb{I}(X_i = X_j)|X_i]\bigg) \\
&= \int\int_{x\neq y} K_h(x)K_h(y)\bigg(g_{i,j}(x,y) + p^{(i)}(x)p^{(j)}(x)\bigg) - \mathbb{E}[K_h(X_i)]\mathbb{E}[K_h(X_j)] \\
&+ \mathbb{E}\bigg(K_h^2(X_i)\mathbb{E}[\mathbb{I}(X_i = X_j)|X_i]\bigg) \\
&= \int\int_{x\neq y} K_h(x)K_h(y)g_{i,j}(x,y)dxdy + \int K_h(x)p^{(i)}(x)dy \int K_h(y)p^{(j)}(y)dy \\
&- \mathbb{E}[K_h(X_i)]\mathbb{E}[K_h(X_j)] + \mathbb{E}[K_h^2(X_i)r(X_i)^{j-i}] \\
&= \int\int_{x\neq y} K_h(x)K_h(y)g_{i,j}(x,y)dxdy + \mathbb{E}[K_h^2(X_i)r(X_i)^{j-i}] \\
&:= \eta_{i,j} + \theta_{i,j}
\end{aligned}
$$

Now, we will try to show that contribution of $\eta_{i,j}$ is of order $\frac{1}{n}$. We have assumed $\frac{1}{a(x)} < M$. So,

$$
\frac{1}{a(x)} < M \implies a(x) > \frac{1}{M} \implies 1 - a(x) < 1 - \frac{1}{M} \implies r(x) < 1 - \frac{1}{M} \tag{24}
$$

Thus using (21) and L1 we have, we can get a bound on $g_{i,j}$. If $(x,y) \in [-\varepsilon, \varepsilon]^2$,

$$
\begin{aligned}
|g_{i,j}(x,y)| &\leq |p^{(i)}(x)\bigg(\tilde{p}_x^{(j-i)}(y) - p^{(j)}(y)\bigg)| \\
&= |p^{(i)}(x)\bigg(\tilde{p}_x^{(j-i)}(y) - f(y) + f(y)p^{(j)}(y)\bigg)| \\
&\leq p^{(i)}(x)\big(|\tilde{p}_x^{(j-i)}(y) - f(y)| + |f(y)p^{(j)}(y)|\big) \\
&\leq M[V(x)R(j-i) + \mathbb{E}\big(V(X_0)\big)R(j) + r(y)^j p^{(0)}(y)] \\
&\leq M^2[R(j-i) + R(j) + (1 - \frac{1}{M})^j]
\end{aligned} \tag{25}
$$

As by assumption $p^{(i)}$, V and $1/a$ are bounded by M. Now consider the term,

$$
\sum_{i=1}^{n}\sum_{j=i+1}^{n}[R(j-i) + R(j)] \leq \sum_{i=1}^{n} 2\sum_{j=1}^{n} R(j)
$$

$$
\begin{aligned}
&\leq n \times 2 \sum_{j=1}^{\infty} R(j) \\
&\leq 2nM
\end{aligned}
$$

Again,

$$
\begin{aligned}
\sum_{i=1}^{n} \sum_{j=i+1}^{n} (1-\frac{1}{M})^{j} &= \sum_{i=1}^{n} M\left(1-\frac{1}{M}\right)^{i+1}\left[1-\big(1-\frac{1}{M}\big)^{n-1-i}\right] \\
&\leq M
\end{aligned}
$$

Using the fact that $\frac{1}{a(x)} < M \implies M > 1 \implies \frac{1}{M} < 1$, the remaining terms are less than 1.
For sufficiently small h, $\sup K_h \subset [-\varepsilon, \varepsilon]$ and as the bound of $g_{i,j}$ does not depend on (x,y) and $\int K_h(x)dx \leq 1$, we can find,

$$
\begin{aligned}
\frac{2}{n^2} \sum_{i=0}^{n-1} \sum_{j=i+1}^{n-1} \eta_{i,j} &= \frac{2M^2}{n^2} \sum_{i=0}^{n-1} \sum_{j=i+1}^{n-1} [R(j-i) + R(j) + (1-\frac{1}{M})^{j}] \\
&\leq \frac{6M^3}{n}
\end{aligned}
$$

This agrees with the standard results on density data for dependent data. Now we will concentrate on the term $\theta_{i,j}$.

$$
\begin{aligned}
\frac{2}{n^2} \sum_{i=0}^{n-1} \sum_{j=i+1}^{n-1} \theta_{i,j} &= \frac{2}{n^2} \sum_{i=0}^{n-1} \sum_{j=i+1}^{n-1} \mathbb{E}[K_h^2(X_i) r(X_i)^{j-i} \\
&= \frac{2}{n^2} \sum_{i=0}^{n-1} \mathbb{E}[K_h^2(X_i) \sum_{j=i+1}^{n-1} r(X_i)^{j-i}] \\
&= \frac{2}{n^2} \sum_{i=0}^{n-1} \mathbb{E}[K_h^2(X_i) r(X_i) \frac{1-r(X_i)^{n-1-i}}{1-r(X_i)}] \\
&= \frac{2}{n^2} \sum_{i=0}^{n-1} \int K_h^2(x) \frac{r(x)}{a(x)}) \big(1-r(x)^{n-1-i}\big) p^{(i)}(0) dx \\
&= \frac{2}{n^2} \sum_{i=0}^{n-1} \frac{r(0)}{a(0)}) \big(1-r(0)^{n-1-i}\big) p^{(i)}(0) \times \int \frac{K^2(x/h)}{h^2} dx \\
&= \frac{2}{n^2 h} \sum_{i=0}^{n-1} (K^2)_h * \left[\frac{p^{(i)} r(1-r^{n-1-i})}{a}\right](0) \\
&= \frac{2}{n^2 h} \sum_{i=0}^{n-1} \left[\frac{r(0)}{a(0)}) \big(1-r(0)^{n-1-i}\big) p^{(i)}(0) + \frac{fr}{a}(0) - \frac{fr}{a}(0)\right] \times \int K^2(x) dx
\end{aligned}
$$

$$= \frac{2}{n^2 h} \times n \times \frac{fr}{a}(0) \int K^2(x)dx + \frac{2}{n^2 h} \sum_{i=0}^{n-1} \sum_{i=0}^{n-1} (K^2)_h * G_i(0) \tag{26}$$

where, $G_i(x) = \frac{r(x)\left[p^{(i)}(x)\left(1-r(x)^{n-1-i}\right)-f(x)\right]}{a(x)}$. Now we try to find a bound of $G_i(x)$.

$$\begin{aligned} G_i(x) &= \frac{r(x)\left[p^{(i)}(x)\left(1-r(x)^{n-1-i}\right)-f(x)\right]}{a(x)} \\ &= \frac{r(x)}{a(x)}\left[p^{(i)}(x) - f(x) - p^{(i)}(x)r(x)^{n-1-i}\right)\right] \end{aligned}$$

Now, from (21)

$$\begin{aligned} |p^{(i)}(x) - f(x)| &\le \mathbb{E}|f(x) - \tilde{p}^{(i)}_{X_0}| + r(x)^i p^0(x) \le V(X_0)R(i) + M \times (1 - \frac{1}{M})^i \\ &\le MR(i) + M \times (1 - \frac{1}{M})^i \end{aligned}$$

Also, $\frac{r(x)}{a(x)} \le \frac{1-a(x)}{a(x)} \le \frac{1}{a(x)} - 1 \le M - 1 \le M$. Thus we can conclude,

$$\begin{aligned} G_i(x) &= \frac{r(x)\left[p^{(i)}(x)\left(1-r(x)^{n-1-i}\right)-f(x)\right]}{a(x)} \\ &\le M \times \left[MR(i) + M \times (1 - \frac{1}{M})^i + M \times (1 - \frac{1}{M})^{n-1-i}\right] \\ &= M^2\left[R(i) + (1 - \frac{1}{M})^i + (1 - \frac{1}{M})^{n-1-i}\right] \end{aligned}$$

As $\sum_{i=0}^{\infty} R(i)$ is assumed to be finite, $\frac{2}{n^2 h} \sum_{i=0}^{n-1} \sum_{i=0}^{n-1} (K^2)_h * G_i(0) < \infty$. Thus the sum in (26) is finite. Hence, we finally have,

$$\frac{2}{n^2} \sum_{i=0}^{n-1} \sum_{j=i+1}^{n-1} \theta_{i,j} = \frac{2(K^2)_h * (fr/a)(0)}{nh} + O\left(\frac{1}{n^2 h}\right)$$

$$\begin{aligned} \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(K_h(X_i)) &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\mathbb{E}\left(K_h^2(X_i)\right) - \mathbb{E}^2\left(K_h^2(X_i)\right)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\int \frac{K^2(u/h)}{h^2} p^{(i)}(0)du - \left(\int \frac{K(u/h)}{h} p^{(i)}(0)du\right)^2\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\frac{1}{h}(K^2)_h * p^{(i)}(0) - \left(K_h * p^{(i)}(0)\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\int \frac{K^2(u/h)}{h^2}[p^{(i)}(0)+f(0)-f(0)]du - \left(\int \frac{K(u/h)}{h}p^{(i)}(0)du\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\int \frac{K^2(u/h)}{h^2}f(0)du + \int \frac{K^2(u/h)}{h^2}\left(p^{(i)}(0)-f(0)\right)du - \left(K_h * p^{(i)}(0)\right)^2\right] \\
&= \frac{1}{n^2 h} \times n \times K_h^2 * f(0) + o\left(\frac{1}{n}\right) \\
&= \frac{1}{nh}K_h^2 * f(0) + o\left(\frac{1}{n}\right)
\end{aligned}
$$

Using (21), we can bound the second term in the third last line and as $\sum_{i=0}^{\infty} R(i) < \infty$, the sum of the bound will be finite, as we have discussed before. Further, $\left(K_h * p^{(i)}(0)\right)^2 = \left(p^{(i)}(0)\int K(u)du\right)^2 \leq (M \times 1)^2$. So the third sum of the third last line is finite. Thus we can write the sum as $o\left(\frac{1}{n}\right)$. Thus from (22) we have,

$$
\begin{aligned}
\mathbb{V}(\hat{f}(0)) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(K_h(X_i)) + \frac{2}{n^2}\sum_{i=0}^{n}\sum_{j=i+1}^{n}(\eta_{i,j}+\theta_{i,j}) \\
&= \frac{1}{nh}K_h^2 * f(0) + \frac{2(K^2)_h * (fr/a)(0)}{nh} + O\left(\frac{1}{n^2 h}\right) + o\left(\frac{1}{n}\right) \\
&= \frac{1}{nh}(\frac{2}{a(0)} - 1)f(0)\mu_{0,2} + o\left(\frac{1}{nh}\right)
\end{aligned}
$$

Now our aim is to find the bias. We have,

$$
\begin{aligned}
\mathbb{E}\big(\hat{f}(0)\big) - f(0) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}K_h(X_i)\right) - f(0) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big(K_h(X_i)\big) - f(0) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int K_h(x)p^{(i)}(0)dx - f(0) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int \frac{K\left(\frac{u}{h}\right)}{h}p^{(i)}(0)du - f(0) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int K(x)p^{(i)}(0)dx - f(0)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} K_h * p^{(i)}(0) - f(0)\\
&= \frac{1}{n}\sum_{i=1}^{n}\int K(x)\Big[p^{(i)}(0) + f(0) - f(0)\Big]dx - f(0)\\
&= \frac{1}{n}\times n \times \int K(x) f(0) dx - f(0) + \frac{1}{n}\sum_{i=1}^{n}\int K(x)\Big[p^{(i)}(0) - f(0)\Big]dx\\
&= K_h * f(0) - f(0) + \frac{1}{n}\sum_{i=1}^{n}\int K(x)\Big[p^{(i)}(0) - f(0)\Big]dx
\end{aligned}
$$

As we have discussed before using (21), we can show the last sum $\sum_{i=1}^{n}\int K(x)\Big[p^{(i)}(0) - f(0)\Big]$ is bounded. Furthermore, we have assumed K is symmetric and f has a bounded third derivative. Thus we have got,

$$
\mathbb{E}\big(\hat{f}(0)\big) - f(0) = f''(0)\frac{h^2}{2}\mu_{2,1} + o(h^2) + O\left(\frac{1}{n}\right) \tag{27}
$$

□

Proof of Theorem 2:

Proof. Let us first define two quantities.

$$
\begin{aligned}
K_h(X_i, x) &= K_h(X_i - x) - K_h * p^{(i)}(x)\\
d_{i,j}(x) &= \tilde{p}_{X_i}^{j-i}(x) - p^{j}(x)
\end{aligned}
$$

We get these quantities for $j > i$ by conditioning on X_i.

$$
\begin{aligned}
\mathbb{E}\big(K_h(X_i, x)\big) &= \mathbb{E}\big(K_h(X_i - x)\big) - \mathbb{E}\big(K_h * p^{(i)}(x)\big)\\
&= \int K_h(v - x) p^{(i)}(x) dv - \mathbb{E}\big(K_h * p^{(i)}(x)\big)\\
&= \int \frac{K(\frac{v-x}{h})}{h} p^{(i)}(x) dv - \mathbb{E}\big(K_h * p^{(i)}(x)\big)\\
&= \big(K_h * p^{(i)}(x)\big) - \big(K_h * p^{(i)}(x)\big)\\
&= 0.
\end{aligned}
$$

$$
\mathbb{C}ov\big(K_h(X_i - x), K_h(X_j - x)\big) = \mathbb{E}\big[K_h(X_i - x), K_h(X_j - x)\big] - K_h * p^{(i)}(x) K_h * p^{(j)}(x)
$$

$$
\begin{aligned}
&= \mathbb{E}\big[K_h(X_i,x) - K_h * p^{(i)}(x)\big]\big[K_h(X_j,x) - K_h * p^{(j)}(x)\big] \\
&\quad - K_h * p^{(i)}(x) K_h * p^{(j)}(x) \\
&= \mathbb{E}\big[K_h(X_i,x)K_h(X_j,x)\big] - K_h * p^{(i)}(x)\mathbb{E}K_h(X_j,x) \\
&\quad - K_h * p^{(j)}(x)\mathbb{E}K_h(X_i,x) \\
&\quad + K_h * p^{(i)}(x)K_h * p^{(j)}(x) - K_h * p^{(i)}(x)K_h * p^{(j)}(x) \\
&= \mathbb{E}\big[K_h(X_i,x)K_h(X_j,x)\big] \\
&= \mathbb{E}\{K_h(X_i,x)\mathbb{E}[K_h(X_i,x)|X_i]\}
\end{aligned}
$$

Given X_i, we can reach X_j using smooth part with probability $\tilde{p}_{X_i}^{j-i}(x)$, where X_j is different from X_i and can take any value in $\mathscr{X}$. On the other hand if it is rejected for $(j-i)$ steps, it will stuck at the same place. Thus we get the following.

$$
\begin{aligned}
\mathbb{C}ov\big(K_h(X_i-x), K_h(X_j-x)\big) &= \mathbb{E}[K_h(X_i,x)K_h * \tilde{p}_{X_i}^{j-i}(x) - K_h * p^{(j)}(x)] \\
&= \mathbb{E}[K_h(X_i,x)K_h * d_{i,j}(x)] + \mathbb{E}[K_h(X_i,x)K_h(X_i-x)r(X_i)^{j-i}] \\
&:= \eta_{i,j}(x) + \theta_{i,j}(x)
\end{aligned}
$$

The integrated variance of the estimator can be written as,

$$
\int \mathbb{V}(\hat{f}_h(x))dx = \frac{1}{n^2}\sum_{i=0}^{n-1}\int Var[K_h(X_i-x)]dx + \frac{2}{n^2}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\int[\eta_{i,j}(x) + \theta_{i,j}(x)]dx \tag{28}
$$

Now, consider the terms.

$$
\begin{aligned}
\int \eta_{i,j}(x)dx &= \int \mathbb{E}[K_h(X_i,x)K_h * d_{i,j}(x)]dx \\
&\le \int \mathbb{E}|K_h(X_i,x)K_h * d_{i,j}(x)|dx \\
&\le \int [\mathbb{E}|K_h(X_I,x)|^q]^{\frac{1}{q}}[\mathbb{E}|K_h * d_{i,j}(x)|^p]^{\frac{1}{p}} \text{ , by Holder's inequality} \\
&\le [\int \mathbb{E}|K_h(X_I,x)|^q]^{\frac{1}{q}}[\int \mathbb{E}|K_h * d_{i,j}(x)|^p]^{\frac{1}{p}} \text{ , by Jensen's inequality}
\end{aligned} \tag{29}
$$

Now we will use,

$$
\big(K_h(X_i-x)\big)^q = \left(\frac{K(\frac{X_i-x}{h})}{h}\right)^q = \frac{K^q\big(\frac{X_i-x}{h}\big)}{h^q} = \frac{K_h^q(\frac{X_i-x}{h})}{h^{q-1}}
$$

Consider the term,

$$
\{\int \mathbb{E}|K_h(X_i,x)|^q dx\}^{\frac{1}{q}} = \{\int \mathbb{E}|K_h(X_i-x) - K_h * p^{(i)}(x)|^q dx\}^{\frac{1}{q}}
$$

$$
\begin{aligned}
&\leq [\int\int |K_h(u-x)|^q p^{(i)}(x)dudx + \int M[\frac{|K_h * p^{(i)}(x)|}{M}]^q dx]^{\frac{1}{q}} \\
&\leq \{\int \frac{(K^q)_h * p^{(i)}(x)}{h^{q-1}}dx\}^{\frac{1}{q}} + M[\int [\frac{|K_h * p^{(i)}(x)|}{M}]^q dx]^{\frac{1}{q}} \\
&\leq h^{\frac{1-q}{q}}\{\int K(x)^q dx\}^{\frac{1}{q}} + M \\
&\leq h^{\frac{1-q}{q}} C,
\end{aligned}
$$

for some constant C and sufficiently small h. Here we have used $K_h * p^{(i)}(x) = p^{(i)}(x)\int k(u)du \leq M\int K(u)du \leq M$. Now we will consider the second term of (29). Assuming p to be be an integer, we get by applying Hölder's inequality p times, we get,

$$
\begin{aligned}
\int \mathbb{E}|K_h * d_{i,j}(x)|^p dx &\leq \mathbb{E}\int_{\mathbb{R}^p}\prod_{k=1}^{p} K_h(s_k)\int\prod_{k=1}^{p}|d_{i,j}(x+s_k)|dx\mathbf{ds} \\
&\leq \int_{\mathbb{R}^p}\prod_{k=1}^{p} K_h(s_k)\int \mathbb{E}|d_{i,j}(x)|^p dx
\end{aligned}
$$

Now we will find a bound for $|d_{i,j}(x)|$. We have proved $p^{(i)}(x) = \mathbb{E}\tilde{p}_{X_0}^{(i)}(x) + r(x)^i p^{(0)}(x)$ in our previous theorem. We will again use it.

$$
\begin{aligned}
|d_{i,j}(x)| &= |\tilde{p}_{X_i}^{(j-i)}(x) - p^{(j)}(x)| \\
&= |\tilde{p}_{X_i}^{(j-i)}(x) - \mathbb{E}\tilde{p}_{X_0}^{(j)}(x) + r(x)^j p^{(0)}(x)| \\
&\leq |\tilde{p}_{X_i}^{(j-i)}(x) - f(x)| + \mathbb{E}|\tilde{p}_{X_0}^{(j)}(x) - f(x)| + r(x)^j p^{(0)}(x) \\
&\leq \tilde{p}_{X_i}^{(j-i)}(x) + f(x) + \mathbb{E}|\tilde{p}_{X_0}^{(j)}(x)| + f(x) + r(x)^j p^{(0)}(x) \\
&\leq M+M+M+M+M \text{ r(x) being the rejection probability, } r(x)^j \leq 1 \\
&= 5M
\end{aligned}
$$

Thus $\frac{|d_{i,j}(x)|}{5M} \leq 1$. By assumption of the theorem, we have,
$\int|\tilde{p}_y^{(i)}(x) - f(x)|dx \leq a(x)^{-1}\int|\tilde{p}_y^{(i)}(x) - f(x)|dx \leq V(y)R(i)$. Now we will use it.

$$
\begin{aligned}
&\{\int \mathbb{E}|d_{i,j}(x)|^p dx\}^{\frac{1}{p}} \\
&= 5M\{\int \frac{\mathbb{E}|d_{i,j}(x)|^p}{(5M)^p}dx\}^{\frac{1}{p}} \\
&\leq 5M\{\int \frac{\mathbb{E}|d_{i,j}(x)|}{5M}dx\}^{\frac{1}{p}} \text{ [As } \frac{|d_{i,j}(x)|}{5M} \leq 1 \text{ and } p > 1]
\end{aligned}
$$

$$\begin{aligned}
&\leq (5M)^{1-\frac{1}{p}}[\int \mathbb{E}|\tilde{p}^{(j-i)}_{X_i}(x) - f(x)| + \mathbb{E}|f(x) - \tilde{p}^{(j)}_{X_0}(x)|dx + \mathbb{E}(r(X_0)^j)^{\frac{1}{p}}] \\
&\leq (5M)^{1-\frac{1}{p}}[\mathbb{E}[V(X_i)]R(j-i) + \mathbb{E}[V(X_0)]R(i) + \mathbb{E}[r(X_0)^i]^{\frac{1}{p}}] \\
&\leq C'\{R(j-i)^{1-\varepsilon} + R(i)^{1-\varepsilon} + \mathbb{E}[r(X_0)^i]^{1-\varepsilon}\}
\end{aligned}$$

Hence, combining all the inequalities, we can say,

$$\begin{aligned}
\frac{2}{n^2}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n-1}\int \eta_{i,j}(x)dx &\leq \frac{2h^{1/q-1}CC'}{n^2}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n-1}\{R(j-i)^{1-\varepsilon} + R(i)^{1-\varepsilon} + \mathbb{E}[r(X_0)]^{1-\varepsilon}\} \\
&= o\left(\frac{1}{nh}\right), \text{ as } n \to \infty,\ h \to 0.
\end{aligned}$$

We can argue that $\sum\limits_{j=i+1}^{n-1}\{R(j-i)^{1-\varepsilon} + R(i)^{1-\varepsilon} + \mathbb{E}[r(X_0)]^{1-\varepsilon}\} \leq (n-1)M$. Thus we get this function as $o\left(\frac{1}{nh}\right)$. Now, let us move towards the other term, $\theta_{i,j}$.

$$\begin{aligned}
\int \theta_{i,j}(x)dx &= \int \mathbb{E}[K_h(X_i,x)K_h(X_i - x)r(X_i)^{j-i}]dx \\
&= \int \mathbb{E}[K_h^2(X_i - x)r(X_i)^{j-i}]dx - \int K_h * p^{(i)}(x)\mathbb{E}[K_h(X_i - x)r(X_i)^{j-i}]dx \\
&= \frac{\mu_{0,2}\mathbb{E}[r(X_i)^{j-i}]}{h} - \int K_h * p^{(i)}(x)K_h * (r^{j-i}p^{(i)})(x)dx \qquad (30)
\end{aligned}$$

We have $M\int K_h * (r^{j-i}p^{(i)})(x)dx = M\mathbb{E}[r(X_i)^{j-i}]$. Since $\int K_h * p^{(i)x}$ can be shown bounded by assumption, we can say that the later quantity is bounded. Hence, the terms of these type, contributes with $O\left(\frac{1}{n}\right)$ to the sum (28).

Consider the following quantity,

$$\begin{aligned}
\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\frac{r(X)^{n-1-i}}{a(X)}\right] &= \frac{1}{n}\mathbb{E}\left(\frac{1-r(X)^n}{a(X)(1-r(X))}\right) \\
&\leq \frac{\mathbb{E}(a(X)^{-2})}{n}
\end{aligned}$$

Now, let us define a quantity,

$$H_i = \mathbb{E}\left[\frac{r(X_i)[1-r(X_i)^{n-1-i}]}{a(X_i)} - \frac{r(X)[1-r(X)^{n-1-i}]}{a(X)}\right] \qquad (31)$$

$$\frac{1}{n}\sum_{i=0}^{n-1}|H_i| = \frac{1}{n}\sum_{i=0}^{n-1}\int \frac{r(x)(1-r(x)^{n-i-1})}{a(x)}|p^{(i)}(x) - f(x)|dx$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=0}^{n-1}\int \frac{1}{a(x)}|p^{(i)}(x)-f(x)|dx \\
&\leq \frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{a(x)}(p^{(i)}(x)+f(x))dx \\
&\leq \frac{1}{n}\sum_{i=0}^{n-1}\{\mathbb{E}[V(X_i)]R(i)+\mathbb{E}\left[\frac{r(X_0)\hat{\imath}}{a(X_0)}\right]\} \\
&\leq \frac{1}{n}\{M^2+\mathbb{E}[a(X_0)^{-2}]
\end{aligned}
$$

Hence, it is a quantity of order $O\left(\frac{1}{n}\right)$. Let's now, go back to the quantity (30).

$$
\begin{aligned}
\frac{2}{n^2}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n-1}\int \theta_{i,j}(x)dx &= \frac{2}{n^2}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n-1}\frac{\mu_{0,2}\mathbb{E}[r(X_i)^{j-i}]}{h}+O\left(\frac{1}{n}\right) \\
&= \frac{2\mu_{0,2}}{n^2h}\sum_{i=0}^{n-1}\mathbb{E}\left[\frac{r(X_i)[1-r(X_i)^{n-1-i}]}{a(X_i)}\right]+O\left(\frac{1}{n}\right) \\
&= \frac{2\mu_{0,2}}{n^2h}\left\{\mathbb{E}\left[\frac{r(X)}{a(X)}-\frac{1}{n}\sum_{i=0}^{n-1}\frac{r(X)^{n-1-i}}{a(X)}\right]\right\}+\frac{1}{n}\sum_{i=0}^{n-1}H_i+O\left(\frac{1}{n}\right) \\
&= \frac{2\mu_{0,2}}{nh}\mathbb{E}\left[\frac{r(X)}{a(X)}\right]+O\left(\frac{1}{n}\right)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\int \mathbb{V}[K_h(X_i-x)]dx &= \int \frac{K^2\left(\frac{X_i-x}{h}\right)}{h^2}p^{(i)}(x)dx-\int (K_h*p^{(i)}(x))^2dx \\
&= \frac{\mu_{0,2}}{h}+O(1)
\end{aligned}
$$

Thus we finally get,

$$
\begin{aligned}
\int \mathbb{V}[\hat{f}(x)]dx &= \frac{2\mu_{0,2}}{nh}\mathbb{E}\left[\frac{r(X)}{a(X)}\right]+o\left(\frac{1}{nh}\right)+\frac{\mu_{0,2}}{nh} \\
&= A\frac{\mu_{0,2}}{nh}+o\left(\frac{1}{nh}\right)
\end{aligned}
$$

The expression of integrated squared bias can be obtained similarly, as we get in theorem 1. □