Nonparametric Kernel Density Estimation for the Metropolis- Hastings Algorithm

Arkajyoti Bhattacharjee* Nitin Garg* Suchismita Roy*

*Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

April 16, 2022

Contents

The Metropolis-Hastings (M-H) Algorithm

- One of the most widely used Markov Chain Monte Carlo (MCMC) algorithms is the Metropolis-Hastings (Metropolis et al. (1953), Hastings (1970)) algorithm.
- The Markov transition kernel of the M-H chain, $P_x(A)$ is:

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

  where $r(x) := \int (1 - \alpha(x,y)q(x,y))dy =: 1 - a(x)$.
- The i^{th} step transition kernel is given by,

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

Algorithm 1 Metropolis-Hastings algorithm

1: **Input:** $X_n = x$.
2: Draw $Y \sim Q(x,.)$ and independently $U \sim \mathscr{U}(0,1)$.
3: **if** $U < \alpha(x,y) = min\{1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\}$, **then**
4: set $X_{n+1} = y$.
5: **else**
6: set $X_{n+1} = x$.
7: **Output:** X_{n+1}.

The Metropolis-Hastings (M-H) Algorithm

- One of the most widely used Markov Chain Monte Carlo (MCMC) algorithms is the Metropolis-Hastings (Metropolis et al. (1953), Hastings (1970)) algorithm.
- The Markov transition kernel of the M-H chain, $P_x(A)$ is:

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

where $r(x) := \int (1 - \alpha(x,y)q(x,y))dy =: 1 - a(x)$.

- The i^{th} step transition kernel is given by,

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

Algorithm 1 Metropolis-Hastings algorithm

1: **Input:** $X_n = x$.
2: Draw $Y \sim Q(x,.)$ and independently $U \sim \mathscr{U}(0,1)$.
3: **if** $U < \alpha(x,y) = min\{1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\}$, **then**
4: set $X_{n+1} = y$.
5: **else**
6: set $X_{n+1} = x$.
7: **Output:** X_{n+1}.

The Metropolis-Hastings (M-H) Algorithm

- One of the most widely used Markov Chain Monte Carlo (MCMC) algorithms is the Metropolis-Hastings (Metropolis et al. (1953), Hastings (1970)) algorithm.
- The Markov transition kernel of the M-H chain, $P_x(A)$ is:

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

where $r(x) := \int(1 - \alpha(x,y)q(x,y))dy =: 1 - a(x)$.

- The i^{th} step transition kernel is given by,

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

Algorithm 1 Metropolis-Hastings algorithm

1: **Input:** $X_n = x$.
2: Draw $Y \sim Q(x,.)$ and independently $U \sim \mathscr{U}(0,1)$.
3: **if** $U < \alpha(x,y) = min\{1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\}$, **then**
4: set $X_{n+1} = y$.
5: **else**
6: set $X_{n+1} = x$.
7: **Output:** X_{n+1}.

The Metropolis-Hastings (M-H) Algorithm

- One of the most widely used Markov Chain Monte Carlo (MCMC) algorithms is the Metropolis-Hastings (Metropolis et al. (1953), Hastings (1970)) algorithm.
- The Markov transition kernel of the M-H chain, $P_x(A)$ is:

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

where $r(x) := \int (1 - \alpha(x,y)q(x,y))dy =: 1 - a(x)$.

- The i^{th} step transition kernel is given by,

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

Algorithm 1 Metropolis-Hastings algorithm

1: **Input:** $X_n = x$.
2: Draw $Y \sim Q(x,.)$ and independently $U \sim \mathscr{U}(0,1)$.
3: **if** $U < \alpha(x,y) = min\{1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\}$, **then**
4: set $X_{n+1} = y$.
5: **else**
6: set $X_{n+1} = x$.
7: **Output:** X_{n+1}.

Definitions and Properties

- Suppose $X_1, \ldots, X_n \sim f$.
- **KDE**:

$$\hat{f}(x) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} K_h(x, X_i) \stackrel{sym}{=} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

- **Properties**:

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

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

- The properties ensure $\hat{f}(x)$ is a valid p.d.f.

Definitions and Properties

- Suppose $X_1, \ldots, X_n \sim f$.
- **KDE**:

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

- **Properties**:

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

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

- The properties ensure $\hat{f}(x)$ is a valid p.d.f.

Definitions and Properties

- Suppose $X_1, \ldots, X_n \sim f$.
- **KDE**:

$$\hat{f}(x) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} K_h(x, X_i) \stackrel{sym}{=} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

- **Properties**:

$$0 \leq K_h(x, u) < \infty \; \forall \; x, u \in \mathbb{R},$$
$$\int_{-\infty}^{\infty} K_h(x, u) dx = 1.$$

- The properties ensure $\hat{f}(x)$ is a valid p.d.f.

Definitions and Properties

- Suppose $X_1, \ldots, X_n \sim f$.
- **KDE**:

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

- **Properties**:

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

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

- The properties ensure $\hat{f}(x)$ is a valid p.d.f.

How KDE works?

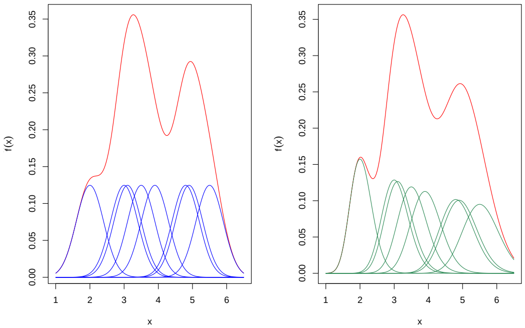

Figure: Kernel estimate showing the contributions of the symmetric Gaussian (left) and the asymmetric Gamma (right) kernels evaluated for the individual observations with bandwidths $h = 0.4, 0.05$ respectively.

Some common symmetric kernels

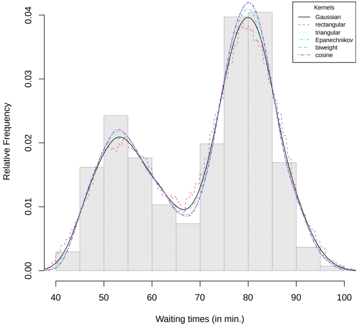

Figure: Density estimates of the Old Faithful Geyser eruption data based on common symmetric kernels imposed on a histogram of the data.

Bandwidth Selection

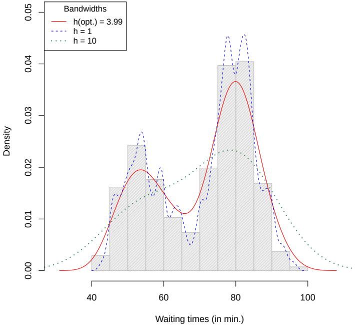

Figure: Density estimates of the Old Faithful Geyser eruption data based on different bandwidths imposed on a histogram of the data.

Measures of Discrepancy and Asymptotic Expansions: Independent Data

- **MSE** (local measure):

$$MSE(\hat{f}(u)) := \mathbb{E}(f(u) - \hat{f}(u))^2 = Var(\hat{f}(u)) + Bias(\hat{f}(u))^2$$

$$= \frac{\mu_{0,2} f(u)}{nh} + \frac{1}{4} h^4 \mu_{2,1} f''(u)^2 + o\left(\frac{1}{nh}\right) + o(h^4),$$

valid as $h \to 0, nh \to \infty$ and where $\mu_{i,j} = \int x^i K(x)^j dx$.

- **MISE** (global measure):

$$MISE(\hat{f}) =: \mathbb{E} \int (f(u) - \hat{f}(u))^2 du = \int Var(\hat{f}(u)) du + \int Bias(\hat{f}(u))^2 du$$

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

as $h \to 0, nh \to \infty$.

- We minimize MISE with respect to h, to get the optimal bandwidth.

Measures of Discrepancy and Asymptotic Expansions: Independent Data

- **MSE** (local measure):

$$MSE(\hat{f}(u)) := \mathbb{E}(f(u)-\hat{f}(u))^2 = Var(\hat{f}(u)) + Bias(\hat{f}(u))^2$$

$$= \frac{\mu_{0,2} f(u)}{nh} + \frac{1}{4} h^4 \mu_{2,1} f''(u)^2 + o\left(\frac{1}{nh}\right) + o(h^4),$$

valid as $h \to 0, nh \to \infty$ and where $\mu_{i,j} = \int x^i K(x)^j dx$.

- **MISE** (global measure):

$$MISE(\hat{f}) =: \mathbb{E} \int (f(u)-\hat{f}(u))^2 du = \int Var(\hat{f}(u)) du + \int Bias(\hat{f}(u))^2 du$$

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

as $h \to 0, nh \to \infty$.

- We minimize MISE with respect to h, to get the optimal bandwidth.

Measures of Discrepancy and Asymptotic Expansions: Independent Data

- **MSE** (local measure):

$$MSE(\hat{f}(u)) := \mathbb{E}(f(u) - \hat{f}(u))^2 = Var(\hat{f}(u)) + Bias(\hat{f}(u))^2$$

$$= \frac{\mu_{0,2} f(u)}{nh} + \frac{1}{4} h^4 \mu_{2,1} f''(u)^2 + o\left(\frac{1}{nh}\right) + o(h^4),$$

valid as $h \to 0, nh \to \infty$ and where $\mu_{i,j} = \int x^i K(x)^j dx$.

- **MISE** (global measure):

$$MISE(\hat{f}) =: \mathbb{E} \int (f(u) - \hat{f}(u))^2 du = \int Var(\hat{f}(u)) du + \int Bias(\hat{f}(u))^2 du$$

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

as $h \to 0, nh \to \infty$.

- We minimize MISE with respect to h, to get the optimal bandwidth.

Measures of Discrepancy and Asymptotic Expansions: Dependent/ Time Series Data

- By stationarity, the bias is not affected by the dependence in the data.
- For results on the variance, two assumptions on the dependence structure of the sequence are applied in the vast majority of this literature:

Assumption (Restricting the local dependence)

Here, it is assumed that (X_i, X_{i+j}) has a bounded bivariate density for all $j > 0$.

Assumption (Restricting the long-range dependence)

Here, it is assumed that the process satisfies a certain mixing condition and that the mixing coefficients decay at a sufficiently fast rate.

- To further ensure that we can construct a consistent estimate, we need:

Assumption

The sequence is stationary and ergodic.

Problem: Motivation

- **Assumption 1 fails**: Based on the form of P_x in (1), due to the rejection step of M-H, (X_i, X_{i+1}) will not a have bounded bivariate density. Infact, the transition density does not exist w.r.t the Lebesgue measure.
- So, **Density Estimation for MH** requires some special attention.
- We present the theory relating KDE of MH samples with that of i.i.d. samples and provide expressions for plug-in bandwidth, h_{mh} and variable KDE based bandwidth, h_{bk}.

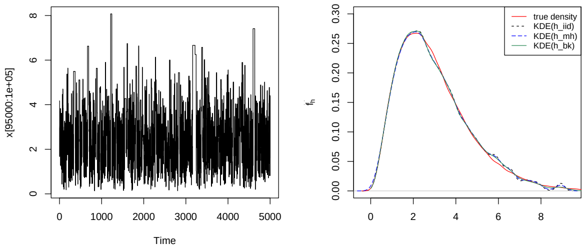

Figure: Plot of true density and KDE of M-H samples with proposal $q(y,x) \propto x^2 e^{-1.7x}$ for target $f(x) = x^2 e^{-x}/2$ based on h_{iid}, h_{mh} and h_{bk}.

Problem: Motivation

- **Assumption 1 fails**: Based on the form of P_x in (1), due to the rejection step of M-H, (X_i, X_{i+1}) will not a have bounded bivariate density. Infact, the transition density does not exist w.r.t the Lebesgue measure.
- So, **Density Estimation for MH** requires some special attention.
- We present the theory relating KDE of MH samples with that of i.i.d. samples and provide expressions for plug-in bandwidth, h_{mh} and variable KDE based bandwidth, h_{bk}.

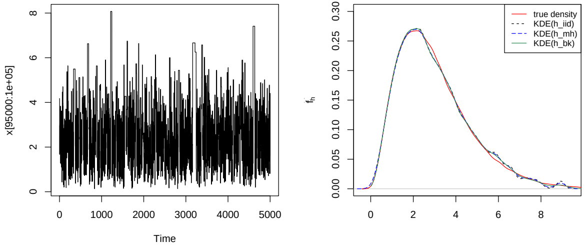

Figure: Plot of true density and KDE of M-H samples with proposal $q(y,x) \propto x^2 e^{-1.7x}$ for target $f(x) = x^2 e^{-x}/2$ based on h_{iid}, h_{mh} and h_{bk}.

Problem: Motivation

- **Assumption 1 fails**: Based on the form of P_x in (1), due to the rejection step of M-H, (X_i, X_{i+1}) will not a have bounded bivariate density. Infact, the transition density does not exist w.r.t the Lebesgue measure.
- So, **Density Estimation for MH** requires some special attention.
- We present the theory relating KDE of MH samples with that of i.i.d. samples and provide expressions for plug-in bandwidth, h_{mh} and variable KDE based bandwidth, h_{bk}.

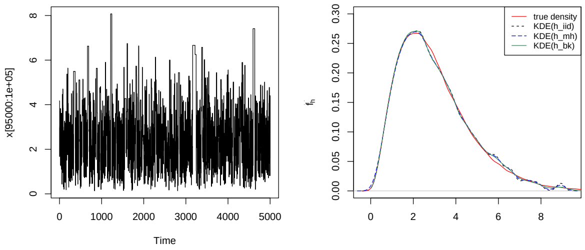

Figure: Plot of true density and KDE of M-H samples with proposal $q(y,x) \propto x^2 e^{-1.7x}$ for target $f(x) = x^2 e^{-x}/2$ based on h_{iid}, h_{mh} and h_{bk}.

Problem: Motivation

- **Assumption 1 fails**: Based on the form of P_x in (1), due to the rejection step of M-H, (X_i, X_{i+1}) will not a have bounded bivariate density. Infact, the transition density does not exist w.r.t the Lebesgue measure.
- So, **Density Estimation for MH** requires some special attention.
- We present the theory relating KDE of MH samples with that of i.i.d. samples and provide expressions for plug-in bandwidth, h_{mh} and variable KDE based bandwidth, h_{bk}.

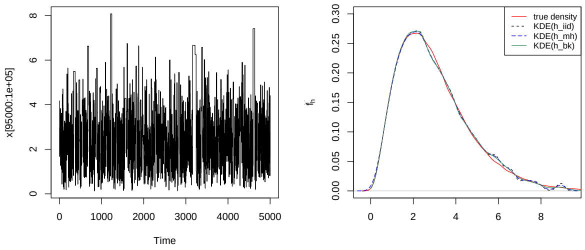

Figure: Plot of true density and KDE of M-H samples with proposal $q(y,x) \propto x^2 e^{-1.7x}$ for target $f(x) = x^2 e^{-x}/2$ based on h_{iid}, h_{mh} and h_{bk}.

Local Assumptions

- 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)

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

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.

Assumption (L4)

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

Local Asymptotic Variance and Bias Expansion for the M-H algorithm

Theorem (Sköld and Roberts (2003))

Under Assumptions L1-L3,

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

where, $A(u) := \left(\frac{2}{a(u)} - 1\right)$ *and* $a(u)$ *denotes probability of accepting a move from* u.

Additionally, under Assumption L4, we get the asymptotic bias,

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

- **Comment:** Variance of kernel density estimator based on i.i.d. samples is multiplied by the factor $A(u)$, which is always greater than or equal to 1 and inversely proportional with the acceptance probability.

Local Asymptotic Variance and Bias Expansion for the M-H algorithm

Theorem (Sköld and Roberts (2003))

Under Assumptions L1-L3,

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

where, $A(u) := \left(\frac{2}{a(u)} - 1\right)$ *and* $a(u)$ *denotes probability of accepting a move from* u.
Additionally, under Assumption L4, we get the asymptotic bias,

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

- **Comment:** Variance of kernel density estimator based on i.i.d. samples is multiplied by the factor $A(u)$, which is always greater than or equal to 1 and inversely proportional with the acceptance probability.

Local Asymptotic Variance and Bias Expansion for the M-H algorithm

Theorem (Sköld and Roberts (2003))

Under Assumptions L1-L3,

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

where, $A(u) := \left(\frac{2}{a(u)} - 1\right)$ *and* $a(u)$ *denotes probability of accepting a move from* u*.*
Additionally, under Assumption L4, we get the asymptotic bias,

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

- **Comment:** Variance of kernel density estimator based on i.i.d. samples is multiplied by the factor $A(u)$, which is always greater than or equal to 1 and inversely proportional with the acceptance probability.

Global Assumptions

- 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)

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

Assumption (G2)

$p^{(i)}(x)$, $\pi(x)$, $\tilde{p}_y^{(i)}(x)$, $E[V(X_i)]$ *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)

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

Global Asymptotic Variance and Bias Expansion for the M-H algorithm

Theorem (Sköld and Roberts (2003))

Under Assumptions G1 and G2,

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

where, $A = \left(\mathbb{E}\left[\frac{2}{a(u)}\right] - 1\right)$.

Additionally, under Assumption G3, we get the asymptotic integrated squared bias

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

Mean Integrated Square Error is given by,

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

Global Asymptotic Variance and Bias Expansion for the M-H algorithm

Theorem (Sköld and Roberts (2003))

Under Assumptions G1 and G2,

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

where, $A = \left(\mathbb{E}\left[\frac{2}{a(u)}\right] - 1\right)$.

Additionally, under Assumption G3, we get the asymptotic integrated squared bias

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

Mean Integrated Square Error is given by,

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

Global Asymptotic Variance and Bias Expansion for the M-H algorithm

Theorem (Sköld and Roberts (2003))

Under Assumptions G1 and G2,

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

where, $A = \left(\mathbb{E}\left[\frac{2}{a(u)}\right] - 1\right)$.
Additionally, under Assumption G3, we get the asymptotic integrated squared bias

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

Mean Integrated Square Error is given by,

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

Optimal Bandwidth for MH Chain

- Minimising the MISE, we get the optimal bandwidth of M-H Chain as,

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

- **Comment:** Since the factor A is inversely proportional to the acceptance probability, we should smooth the region of lower acceptance probability more. It is likely to introduce a mode in the curve in the region of lower probability under the target, that is not present in the true density.

Optimal Bandwidth for MH Chain

- Minimising the MISE, we get the optimal bandwidth of M-H Chain as,

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

- **Comment:** Since the factor A is inversely proportional to the acceptance probability, we should smooth the region of lower acceptance probability more. It is likely to introduce a mode in the curve in the region of lower probability under the target, that is not present in the true density.

Plug-in Method

Based on the expression of h_{MH} in (2), we require the following estimation steps:

- **Estimating** A:

$$\hat{A} = \frac{1}{n}\sum_{i=0}^{n-1}(2T_i - 1) \text{ , where } T_i = \sum_{j=i}^{n-1}\mathbb{I}_{\{X_i = X_j\}}.$$

- **Estimating** $||\pi''||_2^2$ by $\hat{I}_k$:

$$\hat{I}_k = \frac{(-1)^k}{n^2 g_k^{2k+1}}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}K^{(2k)}\left[\frac{X_i - X_j}{g_k}\right] \text{ , where } g_k = \left|\frac{2AK^{(2k)}(0)}{\mu_{2,1} I_{k+1} n}\right|^{1/(2k+3)}.$$

- This expression involves $O(n^2)$ calculations so we use an approximation based on *binning*.

Plug-in Method

Based on the expression of h_{MH} in (2), we require the following estimation steps:

- **Estimating** A:

$$\hat{A} = \frac{1}{n}\sum_{i=0}^{n-1}(2T_i - 1) \text{ , where } T_i = \sum_{j=i}^{n-1} \mathbb{I}_{\{X_i = X_j\}}.$$

- **Estimating** $||\pi''||_2^2$ by $\hat{I}_k$:

$$\hat{I}_k = \frac{(-1)^k}{n^2 g_k^{2k+1}} \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} K^{(2k)}\left[\frac{X_i - X_j}{g_k}\right] \text{ , where } g_k = \left|\frac{2AK^{(2k)}(0)}{\mu_{2,1} I_{k+1} n}\right|^{1/(2k+3)}.$$

- This expression involves $O(n^2)$ calculations so we use an approximation based on *binning*.

Plug-in Method

Based on the expression of h_{MH} in (2), we require the following estimation steps:

- **Estimating** A:

$$\hat{A} = \frac{1}{n} \sum_{i=0}^{n-1} (2T_i - 1) \text{ , where } T_i = \sum_{j=i}^{n-1} \mathbb{I}_{\{X_i = X_j\}}.$$

- **Estimating** $||\pi''||_2^2$ by $\hat{I}_k$:

$$\hat{I}_k = \frac{(-1)^k}{n^2 g_k^{2k+1}} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K^{(2k)} \left[\frac{X_i - X_j}{g_k} \right] \text{ , where } g_k = \left| \frac{2AK^{(2k)}(0)}{\mu_{2,1} I_{k+1} n} \right|^{1/(2k+3)}.$$

- This expression involves $O(n^2)$ calculations so we use an approximation based on *binning*.

Plug-in Method

Based on the expression of h_{MH} in (2), we require the following estimation steps:

- **Estimating** A:

$$\hat{A} = \frac{1}{n} \sum_{i=0}^{n-1} (2T_i - 1) \text{ , where } T_i = \sum_{j=i}^{n-1} \mathbb{I}_{\{X_i = X_j\}}.$$

- **Estimating** $||\pi''||_2^2$ by $\hat{I}_k$:

$$\hat{I}_k = \frac{(-1)^k}{n^2 g_k^{2k+1}} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K^{(2k)} \left[\frac{X_i - X_j}{g_k} \right] \text{ , where } g_k = \left| \frac{2AK^{(2k)}(0)}{\mu_{2,1} I_{k+1} n} \right|^{1/(2k+3)}.$$

- This expression involves $O(n^2)$ calculations so we use an approximation based on *binning*.

Bump-killing

- To account for long rejection periods and nullify bumps that are produced, we can use a different bandwidth for each data point:

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

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

- **Estimating** h_{bk}:

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

- Efficiently kills bumps but can over-smooth the estimate.

Bump-killing

- To account for long rejection periods and nullify bumps that are produced, we can use a different bandwidth for each data point:

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

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

- **Estimating** h_{bk}:

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

- Efficiently kills bumps but can over-smooth the estimate.

Bump-killing

- To account for long rejection periods and nullify bumps that are produced, we can use a different bandwidth for each data point:

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

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

- **Estimating** h_{bk}:

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

- Efficiently kills bumps but can over-smooth the estimate.

Example 1: h_{iid} fails, h_{mh} works

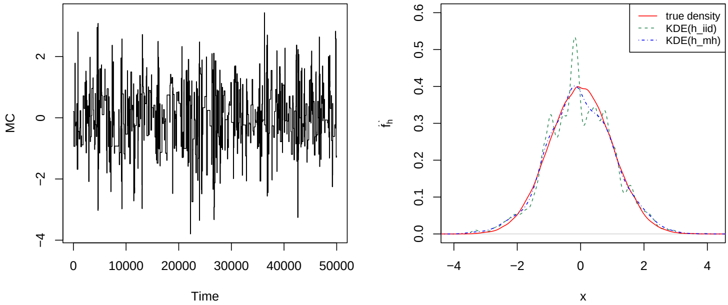

Figure: Trace plot (left) and KDEs (right) based on MH samples generated using $\mathcal{N}(10, 100)$ for the target $\mathcal{N}(0, 1)$. Clearly, h_{iid} is a poor smoothing parameter and h_{mh} is more effective.

Example 2: Bump-killing

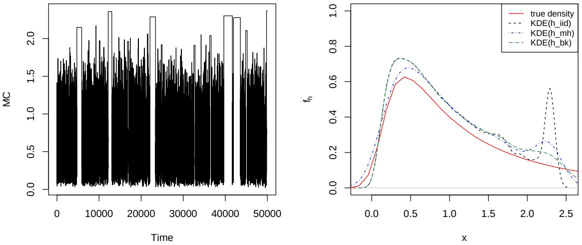

Figure: Trace plot (left) and KDEs (right) based on MH samples generated using skew-$\mathcal{N}(0, 0.54, 10)$[2]for the target log-$\mathcal{N}(0, 1)$. Bump-killing effectively kills the bumps in h_{mh} and is quite smoother than h_{iid}.

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

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

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.

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