Ann Inst Stat Math (2019) 71:771–810
https://doi.org/10.1007/s10463-018-0662-0

Asymptotic properties of parallel Bayesian kernel density estimators

Alexey Miroshnikov[1] · **Evgeny Savelev**[2]

Received: 28 March 2017 / Revised: 27 November 2017 / Published online: 18 April 2018

Abstract In this article, we perform an asymptotic analysis of Bayesian parallel kernel density estimators introduced by Neiswanger et al. (in: Proceedings of the thirtieth conference on uncertainty in artificial intelligence, AUAI Press, pp 623–632, 2014). We derive the asymptotic expansion of the mean integrated squared error for the full data posterior estimator and investigate the properties of asymptotically optimal bandwidth parameters. Our analysis demonstrates that partitioning data into subsets requires a non-trivial choice of bandwidth parameters that optimizes the estimation error.

Keywords Density estimation · Asymptotic properties · Parametric optimization · Parallel algorithms

1 Introduction

Recent developments in data science and analytics research have produced an abundance of large data sets that are too large to be analyzed in their entirety. As the size of data sets increases, the time required for processing rises significantly. An effective solution to this problem is to perform statistical analysis of large data sets with the

✉ Evgeny Savelev
savelev@vt.edu

Alexey Miroshnikov
amiroshn@math.ucla.edu

[1] Department of Mathematics, University of California, 520 Portola Plaza, Los Angeles, CA 90095, USA

[2] Department of Mathematics, Virginia Polytechnic Institute and State University, 460 McBryde Hall, 225 Stanger Street, Blacksburg, VA 24061, USA

use of parallel computing. The prevalence of parallel processing of large data sets motivated a surge in research on parallel statistical algorithms.

One approach is to divide data sets into smaller subsets and analyze the subsets on separate machines using parallel Markov chain Monte Carlo (MCMC) methods (Langford et al. 2009; Newman et al. 2009; Smola and Narayanamurthy 2010). These methods, however, require communication between machines for generation of each sample. Communication costs in modern computer networks dwarf the speedup achieved by parallel processing, and therefore, algorithms that require extensive communications between machines are ineffective (see Scott 2017).

To address these issues, numerous alternative communication-free parallel MCMC methods have been developed for Bayesian analysis of big data. These methods partition data into subsets, perform independent Bayesian MCMC analysis on each subset, and combine the subset posterior samples to estimate the full data posterior (see Scott et al. 2016; Neiswanger et al. 2014; Miroshnikov et al. 2015).

Formally, the task at hand is to estimate the full data posterior $p(\mathbf{x}|\mathbf{y})$, by estimating posterior densities $p_m(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}_m)$, which are subject to the following relation

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x}) \prod_{m=1}^{M} p(\mathbf{y}_m|\mathbf{x}) = \prod_{m=1}^{M} p_m(\mathbf{x}), \quad p_m(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}_m) = p(\mathbf{x})^{1/M} p(\mathbf{y}_m|\mathbf{x}).$$

Neiswanger et al. (2014) introduce a parallel kernel density estimator that first approximates each subset posterior density; the full data posterior is then estimated by multiplying the subset posterior estimators together.

$$\widehat{p}(\mathbf{x}|\mathbf{y}) \ \propto \ \widehat{p}^{*}(\mathbf{x}|\mathbf{y}) := \widehat{p}_1(\mathbf{x}|\mathbf{y}_1) \cdot \widehat{p}_2(\mathbf{x}|\mathbf{y}_2) \cdot \cdots \cdot \widehat{p}_M(\mathbf{x}|\mathbf{y}_M). \tag{1}$$

Here, $\mathbf{x} \in \mathbb{R}^d$ is the model parameter, $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M\}$ is the full data set partitioned into M disjoint independent subsets, and

$$\widehat{p}_m(\mathbf{x}|\mathbf{y}_m) = \sum_{i=1}^{N_m} \frac{1}{h_m} K\Big(\frac{\mathbf{x} - \mathbf{X}_i^m}{h_m}\Big) \tag{2}$$

is the subset posterior kernel density estimator, with $h_m \in \mathbb{R}_+$ a kernel bandwidth parameter.

Neiswanger et al. (2014) show that the estimator (1) is asymptotically exact and develop a sampling algorithm that generates samples from the distribution approximating the full data estimator. Similar sampling algorithms were presented and investigated in Wang and Dunson (2013), Scott et al. (2016), and Scott (2017). It has been noted that these algorithms do not perform well for posteriors that have non-Gaussian shape and are sensitive to the choice of the kernel parameters (see Miroshnikov et al. 2015; Scott et al. 2016; Wang and Dunson 2013).

The highlighted issues indicate that the proper choice of the bandwidth can greatly benefit the accuracy of the estimation as well as sampling algorithms. Moreover, properly chosen bandwidth parameters will improve accuracy of the estimation without incurring additional computational cost.

In the present article, we are concerned with an asymptotic analysis of the parallel Bayesian kernel density estimators of form (1). In particular, we are interested in the asymptotic representation of the mean integrated squared error (MISE) for the non-normalized estimator $\widehat{p}^*$ and the density estimator $\widehat{p}$ as well as the properties of the optimal kernel bandwidth vector parameter $\mathbf{h} = (h_m)_{m=1}^M$ as $\mathbf{N} = (N_1, N_2, \ldots, N_M) \to \infty$; the issues left open in Neiswanger et al. (2014).

We also discuss a universal iterative algorithm based on the derived asymptotic expansions that locates optimal parameters without adopting any assumptions on the underlying probability densities.

The kernel density estimators for the case $M = 1$ have been studied extensively in the past five decades. Asymptotic properties of the mean integrated squared error for the estimator (1) with $M = 1$ and $d = 1$, which takes form (2), were studied by Rosenblatt (1956), Parzen (1962), and Epanechnikov (1969). In particular, for sufficiently smooth probability densities, Parzen (1962) derived the asymptotic expansion for the mean integrated squared error,

$$\mathrm{MISE}[p, \widehat{p}, \mathbf{N}, \mathbf{h}] = \frac{h^4 k_2^2}{4} \int_{\mathbb{R}} (p''(x))^2 \mathrm{d}x + \frac{1}{nh} \int_{\mathbb{R}} K^2(t)\, \mathrm{d}t + o\left(\frac{1}{nh} + h^4\right),$$

with $\mathbf{N} = n$ and $\mathbf{h} = h$, and obtained a formula for the asymptotically optimal bandwidth parameter,

$$h_{M=1}^{\mathrm{opt}} = n^{-1/5} k_2^{-2/5} \left(\int_{\mathbb{R}} K^2(t)\, \mathrm{d}t\right)^{1/5} \left(\int_{\mathbb{R}} (p''(x))^2\, \mathrm{d}x\right)^{-1/5}, \tag{3}$$

which minimizes the leading terms in the expansion.

The case of non-differentiable or discontinuous probability density functions has been shown to possess different asymptotic estimates for MISE. It has been shown by van Eeden (1985) that the optimal bandwidth parameter $h_{M=1}^{\mathrm{opt}} \in \mathbb{R}$ and the rate of convergence of the mean integrated squared error depend directly on the regularity of the probability density p.

In the case of multivariate distributions, $d \geq 1$, the complexity of the asymptotic analysis depends on the form of the bandwidth matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$. In the simplest case, one can assume that $\mathbf{H} = h\mathbf{I}$, where h is a scalar (see Silverman 1986; Simonoff 1996; Epanechnikov 1969). Another approach is to consider the bandwidth matrix of the form $\mathbf{H} = \mathrm{diag}(h_1, h_2, \ldots, h_d)$, with h_i being a bandwidth parameter for each dimension $i \in \{1, \ldots, d\}$. The most general formulation assumes that $\mathbf{H}$ is a $d \times d$ matrix, which allows one to encode correlations between components of $\mathbf{x}$ (see Duong and Hazelton 2005; Wand and Jones 1994).

In the present work, motivated by the ideas of Parzen (1962), Duong and Hazelton (2005), Wand and Jones (1994), and Rosenblatt (1956), we focus on the case $M > 1$ and $d = 1$ and do the asymptotic analysis of the mean integrated squared error for both the parallel non-normalized estimator

$$\mathrm{MISE}\big[\widehat{p}^*, p^*; \mathbf{N}, \mathbf{h}\big] = \mathbb{E} \int_{\mathbb{R}} \Big\{p^*(x|\mathbf{y}) - \widehat{p}^*(x|\mathbf{y})\Big\}^2 \mathrm{d}x$$

and the full data set posterior density estimator

$$\mathrm{MISE}\big[\widehat{p},\, p\,;\mathbf{N},\mathbf{h}\big] = \mathbb{E}\int_{\mathbb{R}}\Big\{p(x|\mathbf{y}) - \widehat{p}(x|\mathbf{y})\Big\}^2\mathrm{d}x,$$

as

$$\mathbf{N} = (N_1, N_2, \ldots, N_M) \to \infty, \quad \mathbf{h} = (h_1, h_2, \ldots, h_M) \to 0, \quad \text{and} \quad (\mathbf{N}\cdot\mathbf{h})^{-1} \to 0.$$

In Theorem 1, under appropriate condition on the regularity of the probability density, we derive the expression for AMISE$[p^*, \widehat{p}^*]$, the asymptotically leading part of MISE for the estimator $\widehat{p}^*$. The leading part turns out to be in agreement with the leading part for the case $M = 1$, but in the multi-subset case, $M > 1$, the leading part contains novel terms that take into account the relationship between M subset posterior densities p_m.

We then perform a similar analysis for the mean square error of the full data set posterior density estimator $\widehat{p}$. The presence of the normalizing constant

$$\widehat{c} = \widehat{\lambda}^{-1} = \left(\int \widehat{p}_1(x|\mathbf{y})\cdot\widehat{p}_2(x|\mathbf{y})\ldots\widehat{p}_M(x|\mathbf{y})\mathrm{d}x\right)^{-1} = \left(\int \widehat{p}^*(x|\mathbf{y})\,\mathrm{d}x\right)^{-1}$$

introduces major difficulties in the analysis of MISE because $\widehat{c}$ may in general have an infinite second moment in which case MISE$[\widehat{p}, p]$ is not defined. This may occur when the estimators $\widehat{p}_i^*$ (on some events) decay too quickly in x variable and the sets of x with the most "mass" for each $\widehat{p}_i^*$ have little common intersection, which potentially leads to large values of $\widehat{c}$. To make sure that $\mathbb{E}\widehat{c}^2 < \infty$, one must impose appropriate conditions on the density p and kernel K. In this article, however, we take another approach. Instead, we replace the mean integrated squared error by an asymptotically equivalent distance functional denoted by

$$\overline{\mathrm{MISE}}\big[\widehat{p},\, p\,;\mathbf{N},\mathbf{h}\big] = \mathbb{E}\left[\left(\frac{\widehat{\lambda}}{\lambda}\right)^2\int_{\mathbb{R}}\Big\{p(x|\mathbf{y}) - \widehat{p}(x|\mathbf{y})\Big\}^2\mathrm{d}x\right].$$

We show that the new functional is always well defined and that it is asymptotically equivalent to MISE when restricted to events $\Omega_{\mathbf{N}} \subset \Omega$ whose probability tends to one as the total number of samples $\|\mathbf{N}\| \to \infty$.

We then do the analysis of the functional $\overline{\mathrm{MISE}}$ by carrying out the same program as for the MISE of the estimator $\widehat{p}^*$. In Theorem 2, we derive the expression for $\overline{\mathrm{AMISE}}[p, \widehat{p}]$, the asymptotically leading part of the $\overline{\mathrm{MISE}}$ for the full data set posterior density estimator $\widehat{p}$. The asymptotically optimal bandwidth parameter for the full data set posterior is then defined to be a minimizer

$$\mathbf{h}^{\mathrm{opt}} = \mathrm{argmin}_{\mathbf{h}\in\mathbb{R}_+^M}\,\overline{\mathrm{AMISE}}[p, \widehat{p}; \mathbf{N}, \mathbf{h}].$$

We then compute minimizing bandwidth $\mathbf{h}^{\mathrm{opt}}$ in explicit form for two special cases. In the examples presented here, we consider subset posterior densities of normal and gamma distributions; see (28), (30), and (32). In the two examples, the optimizing

bandwidth vectors differ significantly and depend, as expected, directly on the full data set density which is typically unknown. For that reason, we discuss an iterative algorithm for locating optimal bandwidth parameters based on asymptotic expansion we derived; see Sect. 4.4.

Our analysis demonstrates that partitioning data into $M > 1$ sets affects the optimality condition of parameter $\mathbf{h}$. In addition, it indicates that the bandwidth vector

$$\mathbf{h}_0^{\text{opt}} = \left(h_{1,M=1}^{\text{opt}}, h_{2,M=1}^{\text{opt}}, \ldots, h_{M,M=1}^{\text{opt}}\right),$$

which minimizes the "componentwise" mean integrated squared error

$$\sum_{i=1}^{M} \text{MISE}[\widehat{p}_i,\, p_i,\, N_i,\, h_i],$$

where $h_{m,M=1}^{\text{opt}}$ is the optimal bandwidth parameter for the estimator $\widehat{p}_m(x|\mathbf{y_m})$ given by (3), is suboptimal for both estimators $\widehat{p}^*$ and $\widehat{p}$ whenever $M > 1$.

This observation highlights the fact that the choice of optimal parameters for parallel kernel density estimators (suitable for parallelizing data analysis) must differ from the theoretical choice suggested in case of processing on a single machine. We must also note that the increased values of MISE resulted from choosing a suboptimal bandwidth parameter get compounded in case of parallel processing. This further necessitates the importance of a proper choice of bandwidth, especially if it comes at no additional computational costs.

In the present work, we perform the analysis of kernel density estimators under the assumption that the samples $\mathbf{X}_m$ are i.i.d. However, it is well known that samples produced with MCMC methods are, in general, not independent, which introduces additional layer of complexity in the analysis. For a more detailed discussion of this issue, we refer the reader to Remark 1 in Sect. 2.

The paper is arranged as follows. In Sect. 2, we set notation and hypotheses that form the foundation of the analysis. In Sect. 3, we derive an asymptotic expansion for MISE of the non-normalized estimator as well as derive formulas for leading parts of bias$[\widehat{p}^*]$ and $\mathbb{V}[\widehat{p}^*]$, which are central to the analysis performed in subsequent sections. In Sect. 4, we perform the analysis of $\overline{\text{MISE}}$ for the full data set posterior density. In Sect. 5, we compute explicit expressions for optimal bandwidth parameters for several special cases and conduct numerical experiments. Finally, in "Appendix," we provide supplementary lemmas and theorems employed in Sects. 3 and 4.

2 Notation and hypotheses

For the convenience of the reader, we collect in this section all hypotheses and results relevant to our analysis and present the notation that is utilized throughout the article.

(H1) Motivated by the form of the posterior density at Neiswanger et al. (2014), we consider the probability density function of the form

$$p(x) \propto p^*(x) \quad \text{where} \quad p^*(x) := \prod_{m=1}^{M} p_m(x).$$

Here, $p_m(x)$ is a probability density function for each $m \in \{1, \dots, M\}$.

(H2) We consider the estimator of p in the form

$$\hat{p}(x) \propto \hat{p}^*(x) \quad \text{where} \quad \hat{p}^*(x) := \prod_{m=1}^{M} \hat{p}_m(x), \tag{H2-a}$$

and for each $m \in \{1, \dots, M\}$ $\hat{p}_m(x)$ is the kernel density estimator of the probability density $p_m(x)$ that has the form

$$\hat{p}_m(x) = \frac{1}{N_m h_m} \sum_{i=1}^{N_m} K\left(\frac{x - X_i^m}{h_m}\right).$$

Here, $X_1^m, X_2^m, \dots, X_{N_m}^m \sim p_m(x)$ are independent identically distributed random variables, K is a kernel density function, and $h_m > 0$ is a bandwidth parameter.

The mean integrated squared error of the estimator $\widehat{p}^*$ of the non-normalized product p^* as well as for the estimator $\widehat{p}(x)$ of the full posterior density $p(x)$ is defined by

$$\mathrm{MISE}[p^*, \widehat{p}^*, \mathbf{N}, \mathbf{h}] = \mathrm{MISE}[p^*, \widehat{p}^*(x)] := \mathbb{E} \int_{\mathbb{R}} (\widehat{p}^*(x) - p^*(x))^2 \, \mathrm{d}x,$$
$$\mathrm{MISE}[p, \widehat{p}, \mathbf{N}, \mathbf{h}] = \mathrm{MISE}[p, \widehat{p}(x)] := \mathbb{E} \int_{\mathbb{R}} (\widehat{p}(x) - p(x))^2 \, \mathrm{d}x,$$

where we use the notation $\mathbf{h} = (h_m)_{m=1}^M$ and $\mathbf{N} = (N_m)_{m=1}^N$. We also use the following convention for the bias and variance of estimators $\widehat{p}(x)$, $\widehat{p}^*(x)$, $\widehat{p}_m(x)$

$$\begin{aligned}
\mathrm{bias}[\hat{p}(x)] &= \mathbb{E}\big[\hat{p}(x)\big] - p(x), \\
\mathrm{bias}[\hat{p}^*(x)] &= \mathbb{E}\big[\hat{p}^*(x)\big] - p^*(x), \\
\mathrm{bias}[\hat{p}_m(x)] &= \mathbb{E}\big[\hat{p}_m(x)\big] - p_m(x), \quad m \in \{1, \dots, M\}.
\end{aligned}$$

We assume that the kernel density function K and probability densities functions $p_1, \dots, p_M$ satisfy the following hypotheses:

(H3) K is positive, bounded, normalized, and its first moment is zero, that is

$$0 \le K(t) \le C, \quad \int_{\mathbb{R}} K(t) \, \mathrm{d}t = 1, \quad \int_{\mathbb{R}} t \, K(t) \, \mathrm{d}t = 0, \quad \int_{\mathbb{R}} K^2(t) \, \mathrm{d}t < \infty.$$

(H4) For each $s \in \{0, 1, 2, 3\}$

$$k_s = \int_{\mathbb{R}} |t|^s K(t)\, \mathrm{d}t < \infty.$$

(H5) For each $m \in \{1, \ldots, M\}$, $s \in \{0, 1, 2, 3\}$, and density $p_m \in C^3(\mathbb{R})$, there exists a constant $C \geq 0$ such that

$$\left| p_m^{(s)}(x) \right| < C \quad \text{for all } x \in \mathbb{R}.$$

(H6) For each $m \in \{1, \ldots, M\}$ and $s \in \{0, 1, 2, 3\}$, the density $p_m(x)$ and its derivatives are integrable, that is, there is a constant C so that

$$\int_{\mathbb{R}} \left| p_m^{(s)}(x) \right| \mathrm{d}x = C < \infty.$$

(H7) Functions

$$\mathbf{N}(n) = \{N_1(n), N_2(n), N_3(n), \ldots, N_M(n)\} : \mathbb{N} \to \mathbb{N}^M,$$
$$\mathbf{h}(n) = \{h_1(n), h_2(n), \ldots, h_M(n)\} : \mathbb{N} \to \mathbb{R}_{++}^M,$$

satisfy for all $i \in \{1, 2, \ldots, M\}$

$$D_1 \leq \frac{N_i}{n} \leq D_2 \quad \text{for some } 0 < D_1 < D_2,$$
$$A_1 N_i(n)^{-\alpha_0} \leq h_i(n) \leq A_2 N_i(n)^{-\alpha_0} \quad \text{for some} \quad \alpha_0 \in (0, 1),$$
$$\lim_{n \to \infty} h_i(n) N_i(n) = \infty.$$

We also define $\underline{N}(n) = \min_i N_i(n)$ and note that $C_1 \|\mathbf{N}\| \leq \underline{N}(n) \leq C_2 \|\mathbf{N}(n)\|$.

Remark 1 In our work, asymptotic error analysis of kernel density estimators is performed under assumption that the samples drawn from each subset posterior distribution are i.i.d. However, it is well known that samples produced with MCMC methods always have a degree of autocorrelation present in them. This dependence does not break the convergence of KDE estimators to the posterior density, and such estimators are shown to be asymptotically exact under certain conditions (De Valpine 2004; West 1993; Sköld and Roberts 2003). Despite the difficulty of choosing optimal parameters for kernel density estimators, the method has been successfully applied to obtain an approximation of posterior density in many projects (De Valpine 2004; West 1993; Neiswanger et al. 2014), many of which are not concerned with the choice of optimal parameters. One particular study, Sköld and Roberts (2003) stand out, as it shows that the optimal bandwidth parameter, h_{M-H}, for KDE based on MCMC samples is a scalar multiple of the optimal bandwidth $h_{i.i.d.}$ for i.i.d samples from the

same posterior density. The scaling is shown to depend on the data set and is linked to the rejection rate of the Metropolis-Hastings sampler:

$$h_{M-H} = A^{1/5} h_{i.i.d.}, \quad A = \mathbb{E}(2/a(X)) - 1,$$

where X is the random variable whose samples are used to estimate the probability density function and $a(X)$ is the acceptance probability rate. Note that if $a(X) \equiv 1$, then $A = 1$. This result highlights the fact that using MCMC samples in general changes the optimal bandwidth and directly depends on the MCMC algorithm.

We must also note that there are several techniques available that can reduce the dependence between samples obtained with MCMC methods. One can run independent Markov chains for each sample, discard a number of intermediate samples between the recorded samples, or employ so-called perfect sampling (Propp and Wilson 1996), which guarantees to produce i.i.d samples.

3 Asymptotic analysis of MISE for $\widehat{p}^*$

We start with the observation that MISE can be expressed via the combination of bias and variance

$$\begin{aligned} \mathrm{MISE}[p^*, \widehat{p}^*] &= \mathbb{E} \int_{\mathbb{R}} \left(\widehat{p}^*(x) - p^*(x)\right)^2 \mathrm{d}x \\ &= \int_{\mathbb{R}} \left(\mathrm{bias}\big[\widehat{p}^*(x)\big]\right)^2 \mathrm{d}x + \int_{\mathbb{R}} \mathbb{V}[\widehat{p}^*(x)]\, \mathrm{d}x. \end{aligned} \tag{4}$$

In what follows, we do the analysis of the bias, then that of variance and conclude with the section where we derive the formula for the optimal bandwidth vector.

3.1 Bias expansion

Using the fact that $\widehat{p}_i(x)$, $i = 1, \ldots, M$ are independent, we obtain

$$\begin{aligned} \mathrm{bias}[\widehat{p}^*(x)] &= \mathbb{E}[\widehat{p}^*(x)] - p^*(x) \\ &= \prod_{m=1}^{M} \mathbb{E}[\widehat{p}_m](x) - \prod_{m=1}^{M} p_m(x) \\ &= \prod_{m=1}^{M} \big(\mathrm{bias}[p_m(x)] + p_m(x)\big) - \prod_{m=1}^{M} p_m(x). \end{aligned} \tag{5}$$

To simplify notation in (5), we shall employ the multiindex notation. Let α be the multiindex with

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_M) \qquad \alpha_m \in \{0, 1\},$$

Then, the above formula can rewritten as follows:

$$\begin{aligned}\text{bias}[\widehat{p}^*(x)] &= \sum_{1\leqslant|\alpha|\leqslant M}\prod_{m=1}^{M}\text{bias}^{\alpha_m}[\widehat{p}_m(x)]\big(p_m(x)\big)^{(1-\alpha_m)}\\ &= \sum_{m=1}^{M}\left[\text{bias}[\widehat{p}_m(x)]\prod_{\substack{k=1\\k\neq m}}^{M}p_k(x)\right]\\ &\quad+\sum_{2\leqslant|\alpha|\leqslant M}\prod_{m=1}^{M}\big(\text{bias}[\widehat{p}_m(x)]\big)^{\alpha_m}(p_m(x))^{(1-\alpha_m)}.\end{aligned}\tag{6}$$

Using this decomposition, we prove the following lemma.

Lemma 1 *Suppose Hypotheses (H3)–(H6) hold. Then,*

(i) *The bias can be expressed as*

$$bias[\widehat{p}^*(x)] = \frac{k_2}{2}\sum_{m=1}^{M}\left[h_m^2\,p_m''(x)\prod_{\substack{k=1\\k\neq m}}^{M}p_k(x)\right]+E_b(x;\mathbf{h}),$$

where the error term $E_b(x;\mathbf{h})$ satisfies the bounds

$$\begin{aligned}|E_b(x;\mathbf{h})| &\leq E_\infty||\mathbf{h}||^3,\quad \forall x\in\mathbb{R}\\ \int_{\mathbb{R}}|E_b(x;\mathbf{h})|\,\mathrm{d}x &\leq E_1||\mathbf{h}||^3,\\ \int_{\mathbb{R}}|E_b(x;\mathbf{h})|^2\,\mathrm{d}x &\leq E_2||\mathbf{h}||^6.\end{aligned}\tag{7}$$

(ii) *The square-integrated bias satisfies*

$$\int_{\mathbb{R}}bias^2[\widehat{p}^*(x)]\,\mathrm{d}x = \frac{k_2^2}{4}\int_{\mathbb{R}}\left[\sum_{m=1}^{M}h_m^2\,p_m''(x)\prod_{\substack{k=1\\k\neq m}}^{M}p_k(x)\right]^2\mathrm{d}x+\mathcal{E}_b(\mathbf{h})\;<\;\infty\tag{8}$$

with the error term satisfying

$$|\mathcal{E}_b(\mathbf{h})|\leq C||\mathbf{h}||^5,\tag{9}$$

where the constant C is independent of $\mathbf{N}$ and $\mathbf{h}\in\mathbb{R}_+^M$.

Proof According to (6) and (34), we have

$$\begin{aligned}
&\mathrm{bias}[\widehat{p}^{*}(x)] \\
&= \frac{k_2}{2}\sum_{m=1}^{M}\left[h_m^2\, p_m''(x)\prod_{\substack{k=1\\k\neq m}}^{M} p_k(x)\right] + \sum_{m=1}^{M}\left[E_{b,m}\prod_{\substack{k=1\\k\neq m}}^{M} p_k(x)\right] \\
&\quad + \sum_{2\leqslant|\alpha|\leqslant M}\prod_{m=1}^{M}\left(\frac{h_m^2 k_2}{2}p_m''(x) + E_{b,m}\right)^{\alpha_m}(p_m(x))^{(1-\alpha_m)}.
\end{aligned}$$

Here, $E_{b,m}$ is the error in bias approximation for each $\widehat{p}_m$ from (34). We are computing bounds for

$$\begin{aligned}
E_b(x;\mathbf{h}) &= \sum_{m=1}^{M}\left[E_{b,m}\prod_{\substack{k=1\\k\neq m}}^{M} p_k(x)\right] \\
&\quad + \sum_{2\leqslant|\alpha|\leqslant M}\prod_{m=1}^{M}\left(\frac{h_m^2 k_2}{2}p_m''(x) + E_{b,m}\right)^{\alpha_m}(p_m(x))^{(1-\alpha_m)}.
\end{aligned} \tag{10}$$

To simplify the derivations, we separate the terms in (10) into two groups: terms with at least one multiple of $E_{b,m}$ and terms free of $E_{b,m}$. We define the sets

$$A_m = \left\{\alpha = (\alpha_j)_{j=1}^{M} \;:\; \alpha_m = 0 \text{ and } 1 \leq |\alpha| \leq (M-1)\right\},$$

and functions

$$P_m(x) = \prod_{\substack{k=1\\k\neq m}}^{M} p_k(x) + \sum_{\alpha\in A_m}\left[\prod_{\substack{j=1\\j\neq m}}^{M}\left(\frac{h_j^2 k_2}{2}p_j''(x) + \mathbb{1}_{\{j>m\}}E_{b,j}\right)^{\alpha_j}(p_j(x))^{(1-\alpha_j)}\right].$$

Here, $\mathbb{1}$ is the characteristic function. Consequently, the error term can be written as follows:

$$\begin{aligned}
E_b(x;\mathbf{h}) &= \sum_{m=1}^{M}\left[E_{b,m}P_m(x)\right] \\
&\quad + \sum_{2\leqslant|\alpha|\leqslant M}\prod_{m=1}^{M}\left(\frac{h_m^2 k_2}{2}p_m''(x)\right)^{\alpha_m}(p_m(x))^{(1-\alpha_m)}.
\end{aligned} \tag{11}$$

Assuming that $||\mathbf{h}||$ is bounded, (H5) and (34), we can conclude that there is a constant C_P so that

$$|P_m(x)| \leq C_P \text{ for any } x \in \mathbb{R} \text{ and } 1 \leq m \leq M.$$

Using (H5) and (34), we conclude that the first term is bounded, and there is a constant C so that

$$\sum_{m=1}^{M} \left| E_{b,m} P_m(x) \right| \leq C \sum_{m=1}^{M} \left(\frac{k_3 h_m^3}{6} \right) \leq CM \frac{||\mathbf{h}||^3 k_3}{6}. \tag{12}$$

The next sum in (11) contains terms that are bounded due to (H5):

$$\left| \frac{h_m^2 k_2}{2} p_m''(x) \right| \leq \frac{||\mathbf{h}||^2 C k_2}{2} \qquad \text{and} \qquad |p_m(x)| \leq C$$

for some appropriate constants C. Since each one of the products below has at least two terms with $p_m''(x)$ for some m, a constant C_Q must exist, so that

$$\left| \sum_{2 \leqslant |\alpha| \leqslant M} \prod_{m=1}^{M} \left(\frac{h_m^2 k_2}{2} p_m''(x) \right)^{\alpha_m} (p_m(x))^{(1-\alpha_m)} \right| \leq C_Q \frac{||\mathbf{h}||^4 k_2^2}{4}. \tag{13}$$

The inequalities (12) and (13) imply the first inequality in (7):

$$|E_b(x\,;\mathbf{h})| \leq CM \frac{||\mathbf{h}||^3 k_3}{6} + \frac{||\mathbf{h}||^4 k_2^2}{4} C_Q.$$

L_1 integrability follows from conditions (H5), (H6), the expansion (11) and the second formula in (35)

$$\int_{\mathbb{R}} |E_b(x\,;\mathbf{h})| \, \mathrm{d}x \leq C \left(\frac{k_3 ||\mathbf{h}||^3}{6} + \frac{||\mathbf{h}||^4 k_2^2}{4} \right),$$

which proves the second estimate in (7).
Using the estimates obtained above, we conclude

$$\int_{\mathbb{R}} |E_b(x\,;\mathbf{h})|^2 \, \mathrm{d}x \leq \sup_{\mathbb{R}} |E_b(x\,;\mathbf{h})| \cdot \int_{\mathbb{R}} |E_b(x\,;\mathbf{h})| \, \mathrm{d}x \leq E_\infty \cdot E_1 ||\mathbf{h}||^6.$$

Finally, (ii) follows from Cauchy–Schwarz inequality applied to

$$\mathrm{bias}^2[\widehat{p}^*(x)] = \frac{k_2^2}{4}\left[\sum_{m=1}^{M} h_m^2 p_m''(x) \prod_{\substack{k=1\\k\neq m}}^{M} p_k(x)\right]^2 + E_b(x;\mathbf{h})k_2\left[\sum_{m=1}^{M} h_m^2 p_m''(x) \prod_{\substack{k=1\\k\neq m}}^{M} p_k(x)\right] + E_b^2(x;\mathbf{h}),$$

which leads directly to (8) and (9). □

3.2 Variance expansion

We next obtain an asymptotic formula for the variance of $\hat{p}^*$. For the proof of the lemma, we perform the following preliminary calculation:

$$\begin{aligned}\mathbb{V}[\widehat{p}^*(x)] &= \mathbb{E}[(\widehat{p}^*(x))^2] - \left(\mathbb{E}[\widehat{p}^*(x)]\right)^2 = \prod_{m=1}^{M}\mathbb{E}\left[\widehat{p}_m^2\right] - \prod_{m=1}^{M}\mathbb{E}^2[\widehat{p}_m]\\ &= \prod_{m=1}^{M}\left(\mathbb{V}[\widehat{p}_m] + \left(p_m + \mathrm{bias}[\widehat{p}_m]\right)^2\right) - \prod_{m=1}^{M}\left(p_m + \mathrm{bias}[\widehat{p}_m]\right)^2 \qquad (14)\\ &= \sum_{1\leqslant|\alpha|\leqslant M}\prod_{m=1}^{M}\left(\mathbb{V}[\widehat{p}_m]\right)^{\alpha_m}\left(p_m + \mathrm{bias}[\widehat{p}_m]\right)^{2(1-\alpha_m)}.\end{aligned}$$

Lemma 2 *Let Hypotheses (H3)–(H7) hold. Then,*

(i) *The variation of $\hat{p}^*$ is given by*

$$\mathbb{V}[\hat{p}^*(x)] = \left(\sum_{m=1}^{M}\left[\frac{p_m}{N_m h_m}\prod_{\substack{k=1\\k\neq m}}^{M} p_k^2(x)\right]\right)\int_{\mathbb{R}} K^2(t)\,\mathrm{d}t + E_V(x;\mathbf{N},\mathbf{h}), \quad x\in\mathbb{R},$$

where the error term $E_V(x;n,\mathbf{h})$ satisfies the bounds

$$|\mathcal{E}_V(N,h)| := \left|\int_{\mathbb{R}} E_V(x)\,\mathrm{d}x\right| = o\left(\frac{1}{\|\mathbf{N}\|}\right). \qquad (15)$$

Proof According to (14), we have

$$\begin{aligned}\mathbb{V}(\hat{p}^*(x)) &= \sum_{1\leqslant|\alpha|\leqslant M}\prod_{m=1}^{M}\left(\frac{p_m(x)}{N_m h_m}\int_{\mathbb{R}}K^2(t)\,\mathrm{d}t + E_{V,m}\right)^{\alpha_m} \\ &\quad \left(p_m + \mathrm{bias}[\widehat{p}_m]\right)^{2(1-\alpha_m)} \\ &= \sum_{1\leqslant|\alpha|\leqslant M}\prod_{m=1}^{M}\left(\frac{p_m(x)}{N_m h_m}\int_{\mathbb{R}}K^2(t)\,\mathrm{d}t + E_{V,m}\right)^{\alpha_m} \\ &\quad \left(p_m^2 + 2p_m\,\mathrm{bias}[\widehat{p}_m] + \mathrm{bias}^2[\widehat{p}_m]\right)^{(1-\alpha_m)}.\end{aligned} \tag{16}$$

Here, $E_{V,m}$ is the approximation error of variance of each $p_m(x)$ from (41). In a fashion similar to the previous proof, we separate the terms in (16). We single out the leading-order terms, the terms with at least one multiple of $E_{V,m}$, the terms with multiples of $\mathrm{bias}[\widehat{p}_m]$, and the terms of the order $o\left(\frac{1}{\|\mathbf{N}\|\|\mathbf{h}\|}\right)$.

We define sets

$$\begin{aligned}A_m^0 &= \left\{\alpha = (\alpha_j)_{j=1}^M \;:\; \alpha_m = 0 \text{ and } 0 \leq |\alpha| \leq (M-1)\right\}, \\ B_m^1 &= \left\{\alpha = (\alpha_j)_{j=1}^M \;:\; \alpha_m = 0 \text{ and } |\alpha| = 1\right\},\end{aligned}$$

and functions

$$P_m^0(x) = \sum_{\alpha\in A_m^0}\left[\prod_{\substack{j=1\\ j\neq m}}^{M}\left(\frac{p_m(x)}{N_m h_m}\int_{\mathbb{R}}K^2(t)\,\mathrm{d}t + \mathbb{1}_{\{j>m\}}E_{V,m}\right)^{\alpha_m}\left(\mathbb{E}^2[\widehat{p}_m]\right)^{(1-\alpha_m)}\right],$$

$$Q_m^1(x) = \sum_{\alpha\in B_m^1}\left[\prod_{\substack{j=1\\ j\neq m}}^{M}\left(\frac{p_m(x)}{N_m h_m}\int_{\mathbb{R}}K^2(t)\,\mathrm{d}t\right)^{\alpha_m}\left(\mathbb{E}^2[\widehat{p}_m]\right)^{(1-\alpha_m)}\right].$$

The variance expansion can be rewritten as

$$\begin{aligned}\mathbb{V}(\hat{p}^*(x)) &= \sum_{1\leqslant|\alpha|\leqslant M}\prod_{m=1}^{M}\left(\frac{p_m(x)}{N_m h_m}\int_{\mathbb{R}}K^2(t)\,\mathrm{d}t\right)^{\alpha_m} \\ &\quad \left(p_m^2 + 2p_m\,\mathrm{bias}[\widehat{p}_m] + \mathrm{bias}^2[\widehat{p}_m]\right)^{(1-\alpha_m)} \\ &\quad + \sum_{m=1}^{M}E_{V,m}P_m^0(x)\end{aligned}$$

$$
\begin{aligned}
&= \sum_{m=1}^{M} \left(\frac{p_m(x)}{N_m h_m} \int_{\mathbb{R}} K^2(t)\,\mathrm{d}t \right) \prod_{\substack{j=1 \\ j\neq m}}^{M} p_m^2(x) \\
&\quad + \sum_{m=1}^{M} \mathrm{bias}[\widehat{p}_m](2p_m(x) + \mathrm{bias}[\widehat{p}_m])\, Q_m^1(x) \\
&\quad + \sum_{2\leqslant|\alpha|\leqslant M} \prod_{m=1}^{M} \left(\frac{p_m(x)}{N_m h_m} \int_{\mathbb{R}} K^2(t)\,\mathrm{d}t \right)^{\alpha_m} \left(\mathbb{E}^2[\widehat{p}_m]\right)^{(1-\alpha_m)} \\
&\quad + \sum_{m=1}^{M} E_{V,m} P_m^0(x).
\end{aligned}
$$

Based on definitions of functions $P_m^0(x)$ and $Q_m^1(x)$, Hypotheses (H5), (H6), and (H7), we can conclude that there are constants $C_{\mathbb{E}}$, C_P, C_Q so that

$$
\begin{aligned}
\mathbb{E}[\widehat{p}_m] &\leq C_{\mathbb{E}}, \\
|P_m^0(x)| &\leq C_P, \\
|Q_m^1(x)| &\leq C_Q \frac{1}{\|\mathbf{N}\|\|\mathbf{h}\|}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\int_{\mathbb{R}} |E_V(x)|\mathrm{d}x \leq \sum_{m=1}^{M} C \left(2 + \frac{||\mathbf{h}||^2 k_2}{2}\right) \frac{C_Q}{\|\mathbf{N}\|\|\mathbf{h}\|} \int_{\mathbb{R}} |\mathrm{bias}[\widehat{p}_m]|\mathrm{d}x \\
+ \frac{1}{\|\mathbf{N}\|^2\|\mathbf{h}\|^2} \sum_{2\leqslant|\alpha|\leqslant M} \left(\frac{1}{\|\mathbf{N}\|^2\|\mathbf{h}\|^2}\right)^{(|\alpha|-2)} C_{\mathbb{E}}^{(M-|\alpha|)} \\
+ \frac{M \cdot C_P}{\|\mathbf{N}\|}.
\end{aligned}
$$

This leads directly to (15). □

3.3 AMISE formula and optimal bandwidth vector

With the lemmas above, we can derive the decomposition of MISE$[p^*, \widehat{p}^*]$ into leading-order terms and higher order terms.

Theorem 1 *Let Hypotheses (H3)–(H7) hold. Then,* MISE *can be represented as*

$$
\mathrm{MISE}[p^*, \widehat{p}^*, \mathbf{N}, \mathbf{h}] = \mathrm{AMISE}[p^*, \widehat{p}^*, \mathbf{N}, \mathbf{h}] + \mathcal{E}(\mathbf{N}, \mathbf{h}),
$$

where the leading term

$$\mathrm{AMISE}[p^*, \widehat{p}^*; \mathbf{N}, \mathbf{h}] = \frac{k_2^2}{4} \int_{\mathbb{R}} \left(\sum_{m=1}^{M} \left[h_m^2 p_m''(x) \prod_{\substack{k=1 \\ k \neq m}}^{M} p_k(x) \right] \right)^2 \mathrm{d}x + \\ + \int_{\mathbb{R}} \left(\sum_{m=1}^{M} \left[\frac{p_m(x)}{N_m h_m} \prod_{\substack{k=1 \\ k \neq m}}^{M} (p_k(x))^2 \right] \right) \mathrm{d}x \int_{\mathbb{R}} K^2(t)\, \mathrm{d}t,$$

and the error term $\mathcal{E}$ satisfies

$$\mathcal{E}(\mathbf{N}, \mathbf{h}) = \mathcal{E}_b(\mathbf{N}, \mathbf{h}) + \mathcal{E}_V(\mathbf{N}, \mathbf{h}) = o\Big(||\mathbf{h}||^4 + \frac{1}{\|\mathbf{N}\| \|\mathbf{h}\|} \Big),$$

as $\mathbf{h} \to 0$, $\mathbf{N} \to \infty$, *and* $(\|\mathbf{N}\| \|\mathbf{h}\|)^{-1} \to 0$.

Proof The result follows from Lemma 1, Lemma 2, and formula (4). □

Remark 2 We would like to note that the analysis we perform here is in spirit of the asymptotic analysis performed for multivariate kernel density estimators by Epanechnikov (1969). However, the full data set density p under consideration is a univariate density expressed as a product and cannot be viewed as a special case of the expansion obtained in Epanechnikov (1969).

Remark 3 The asymptotically leading part derived here is the first step of our analysis. It serves as a stepping stone for the analysis of full MISE carried out in the next section. We would like to note that one can find optimal bandwidth that minimizes AMISE for the non-normalized estimator. One has to be aware, however, that these optimal parameters would not take into account a normalization constant and, as a consequence, would be suboptimal for MISE of the normalized full data set density $\widehat{p}$.

4 Asymptotic analysis of MISE for $\widehat{p}$

4.1 Normalizing constant

In this section, we consider the error that arises when one takes into account the normalizing constant. Recall that by assumption

$$p(x) \propto p^*(x) \quad \text{where} \quad p^*(x) := \prod_{m=1}^{M} p_m(x),$$

where $p_m(x)$, $m \in \{1, \ldots, M\}$ is a probability density function. Then, we define

$$\lambda := \int p^*(x)\, \mathrm{d}x > 0 \quad \text{and} \quad c := \lambda^{-1}$$

and obtain $p(x) = cp^*(x)$. For the estimator

$$\hat{p}(x) \propto \hat{p}^*(x) \quad \text{with} \quad \hat{p}^*(x) := \prod_{m=1}^{M} \hat{p}_m(x),$$

we similarly define

$$\widehat{\lambda} := \int \widehat{p}^*(x)\,\mathrm{d}x > 0 \quad \text{and} \quad \widehat{c} := \widehat{\lambda}^{-1}$$

and hence $\widehat{p}(x) = \widehat{c}\widehat{p}^*(x)$.

We are interested in the optimal bandwidth vector $\mathbf{h} = (h)_{m=1}^{M}$ that optimizes the leading term of the mean integrated squared error

$$\mathrm{MISE}(\hat{p},\ p) = \mathrm{MISE}(\widehat{c}\widehat{p}^*,\ cp^*) = \mathbb{E}\int_{\mathbb{R}} \left(cp^*(x) - \widehat{c}\widehat{p}^*(x)\right)^2 \mathrm{d}x.$$

Observe that $\widehat{c}$ and $\widehat{p}^*$ are not independent and the previously performed analysis is not directly applicable. Moreover, we observe that the estimator of the normalizing constant

$$\widehat{c} = \left(\int \prod_{i=1}^{M} \widehat{p}_i(x)\,\mathrm{d}x\right)^{-1} < \infty \tag{17}$$

may in general have an infinite expectation. This may happen because the estimators in the above product may decay too quickly in x variable and the sets of x with the most "mass" for each p_i may have no common intersection. This potentially may lead to small values of $\widehat{\lambda}$ and hence large $\widehat{c}$. To avoid this situation, one would need to choose the kernel K in appropriate way and establish the finiteness of the expectation of $\widehat{c}$.

In this article, we do not investigate this. Instead, we will show that one can replace MISE by an equivalent functional which is well defined and finite on the whole sample space Ω and that there exists a sequence of smaller sample subspaces Ω_n with $\mathbb{P}(\Omega_n) \to 1$, on which the new functional is asymptotically *equivalent* to MISE$[p, \widehat{p}]$ restricted to Ω_n. We then analyze the new functional and investigate its optimal parameters.

4.2 Preliminary estimates

Lemma 3 (Covariance) *Let $\widehat{p}^*(x)$ be an estimator of form* (H2-a) *where the vector of sample sizes* $\mathbf{N}(n)$ *and bandwidth vector* $\mathbf{h}(n)$ *satisfy (H7). Then,*

$$\mathrm{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)] = \mathbb{E}[\widehat{p}^*(x)\widehat{p}^*(y)] - \mathbb{E}[\widehat{p}^*(x)]\mathbb{E}[\widehat{p}^*(y)]$$

satisfies the estimates

$$\begin{aligned} |\mathrm{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)]| &\leq \frac{C_{abs}}{\|\mathbf{N}\|\|\mathbf{h}\|}, \\ \left| \iint \mathrm{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)] \, \mathrm{d}x\mathrm{d}y \right| &\leq \frac{C_{int}}{\|\mathbf{N}\|}, \end{aligned} \tag{18}$$

for some constants $C_{abs}, C_{int} > 0$ *independent of* n.

Proof We can expand the product as follows:

$$\begin{aligned} &\prod_{i=1}^{M} \mathbb{E}[\widehat{p}_i(x)\widehat{p}_i(y)] - \prod_{i=1}^{M} \mathbb{E}[\widehat{p}_i(x)]\mathbb{E}[\widehat{p}_i(y)] \\ &= \sum_{j=1}^{M} \left(\mathbb{E}[\widehat{p}_j(x)\widehat{p}_j(y)] - \mathbb{E}[\widehat{p}_j(x)]\mathbb{E}[\widehat{p}_j(y)] \right) \left(\prod_{i=1}^{j-1} \mathbb{E}[\widehat{p}_i(x)\widehat{p}_i(y)] \right) \\ &\quad \left(\prod_{i=j+1}^{M} \mathbb{E}[\widehat{p}_i(x)]\mathbb{E}[\widehat{p}_i(y)] \right), \end{aligned}$$

where the products with the top index smaller than the bottom index should be taken as having the value one.

We next observe that, according to (34), for each $i \in \{1, \ldots, M\}$

$$|\mathbb{E}[\widehat{p}_i(x)]\mathbb{E}[\widehat{p}_i(y)]| \leq C \left(1 + \frac{k_2 h_i^2}{2} + \frac{k_3 h_i^3}{6} \right)^2.$$

Also Lemma 9 implies that

$$\begin{aligned} \left| \mathbb{E}[\widehat{p}_i(x)\widehat{p}_i(y)] \right| &\leq C \left(1 + \frac{k_2 h_i^2}{2} + \frac{k_3 h_i^3}{6} \right)^2 + \frac{C}{N_i h_i} \\ &\quad + \frac{C}{N_i} \left(1 + \left(1 + \frac{k_2 h_i^2}{2} + \frac{k_3 h_i^3}{6} \right)^2 \right). \end{aligned}$$

Then, we conclude that for some $C_{\mathbb{E}} \geq 0$

$$|\mathbb{E}[\widehat{p}_i(x)\widehat{p}_i(y)]|, |\mathbb{E}[\widehat{p}_i(x)]\mathbb{E}[\widehat{p}_i(y)]| \leq C_{\mathbb{E}} < \infty, \quad \text{for all} \quad x, y \in \mathbb{R}.$$

Therefore, by Lemma 9, we obtain the estimate

$$|\mathrm{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)]| \leq M\, C \left(\frac{1}{\|\mathbf{N}\|\|\mathbf{h}\|} + \frac{1}{\|\mathbf{N}\|} \left(1 + \left(1 + \frac{k_2 \|\mathbf{h}\|^2}{2} + \frac{k_3 \|\mathbf{h}\|^3}{6} \right)^2 \right) \right)$$

for some appropriate constant C, which gives $(18)_1$.

The integral of $\mathrm{Cov}[\widehat{p}^*(x), \widehat{p}^*(y)]$ is also finite. Using the result of Lemma 9 and the Hypothesis (H6),

$$\begin{aligned}
&\iint \left|\mathrm{Cov}\big[\widehat{p}^*(x), \widehat{p}^*(y)\big]\right| \mathrm{d}x\mathrm{d}y \\
&\le C_{\mathbb{E}}^{M-1} \sum_{i=1}^{M} \iint \left|\mathbb{E}[\widehat{p}_j(x)\widehat{p}_j(y)] - \mathbb{E}[\widehat{p}_j(x)]\mathbb{E}[\widehat{p}_j(y)]\right| \mathrm{d}x\mathrm{d}y \\
&\le C_{\mathbb{E}}^{M-1} \sum_{i=1}^{M} \iint \left(\frac{1}{N_i} p_i(x) \frac{1}{h_i} K_2\left(\frac{x-y}{h_i}\right) + |E_{\Pi,i}(x,y)|\right) \mathrm{d}x\mathrm{d}y \\
&\le C_{\mathbb{E}}^{M-1} \frac{M}{\|\mathbf{N}\|} \left(2 + C\left(k_1 + \frac{k_2\|\mathbf{h}\|^2}{2} + \frac{k_3\|\mathbf{h}\|^3}{6}\right)\right),
\end{aligned}$$

where at the last step we used the facts that $\frac{1}{h} K_2\left(\frac{x-y}{h}\right)$ is a probability density function in y for any fixed x and $p_i(x)$ is also a probability density function. □

Lemma 4 *Let $\widehat{p}^*(x)$ be an estimator of form (H2-a) where the vector of sample sizes $\mathbf{N}(n)$ and bandwidth vector $\mathbf{h}(n)$ satisfy (H7). Then, following identity and the estimate holds*

$$\mathbb{V}[\widehat{\lambda} - \lambda] = \mathbb{V}\left[\int \widehat{p}^*(x)\,\mathrm{d}x - \int p^*(x)\,\mathrm{d}x\right] \le \frac{C_{int}}{\|\mathbf{N}\|} < \infty,$$

where $C_{int} > 0$ is defined in (18).

Proof Since λ is constant, we have

$$\begin{aligned}
\mathbb{V}[\widehat{\lambda} - \lambda] &= \mathbb{E}\big[\widehat{\lambda} - \mathbb{E}[\widehat{\lambda}]\big]^2 \\
&= \mathbb{E}\left[\int_{\mathbb{R}} \widehat{p}^*(x) - \mathbb{E}[\widehat{p}^*(x)]\,\mathrm{d}x\right]^2 \\
&= \mathbb{E}\left[\int \big(\widehat{p}^*(x) - \mathbb{E}[\widehat{p}^*(x)]\big)\,\mathrm{d}x \cdot \int \big(\widehat{p}^*(y) - \mathbb{E}[\widehat{p}^*(y)]\big)\,\mathrm{d}y\right] \\
&= \iint \Big(\mathbb{E}\big[\widehat{p}^*(x)\widehat{p}^*(y)\big] - \mathbb{E}[\widehat{p}^*(x)]\mathbb{E}[\widehat{p}^*(y)]\Big)\mathrm{d}x\,\mathrm{d}y \le \frac{C_{int}}{\|\mathbf{N}\|},
\end{aligned}$$

where the last inequality is from Lemma 3. □

Lemma 5 *Let $\widehat{p}^*(x)$ be an estimator of form (H2-a) where the vector of sample sizes $\mathbf{N}(n)$ and bandwidth vector $\mathbf{h}(n)$ satisfy (H7). Then, for any $\alpha \in (0, 1]$,*

$$\mathbb{P}\left(\left\{\omega : |\mathbb{E}\widehat{\lambda} - \widehat{\lambda}(\omega; \mathbf{N}(n), \mathbf{h}(n))| > \frac{\lambda}{\sqrt{2}\|\mathbf{N}\|^{\frac{1-\alpha}{2}}}\right\}\right) \le \frac{2C_{int}}{\lambda^2\|\mathbf{N}\|^{\alpha}}.$$

Moreover, for any α satisfying

$$\max(0, 1 - 4\alpha_0) < \alpha < 1,$$

where α_0 is defined in (H7), we have

$$\mathbb{P}\left\{\left|\frac{\widehat{\lambda}}{\lambda} - 1\right| > \frac{1}{\|\mathbf{N}\|^{\frac{1-\alpha}{2}}}\right\} \leq \frac{2C_{int}}{\lambda^2\|\mathbf{N}\|^{\alpha}},$$

for all sufficiently large n.

Proof By Lemma 4 and Chebyshev inequality, we obtain

$$\begin{aligned}&\mathbb{P}\left\{\left|\widehat{\lambda} - \mathbb{E}[\widehat{\lambda}]\right|^2 > \frac{\lambda^2}{2\|\mathbf{N}\|^{1-\alpha}}\right\}\\ &\quad\leq \mathbb{P}\left\{\left|\widehat{\lambda} - \mathbb{E}[\widehat{\lambda}]\right|^2 > \mathbb{V}(\widehat{\lambda})\frac{\lambda^2\|\mathbf{N}\|^{\alpha}}{2C_{int}}\right\} \leq \frac{2C_{int}}{\lambda^2\|\mathbf{N}\|^{\alpha}}.\end{aligned}$$

Recall next that

$$|\mathbb{E}(\widehat{\lambda}) - \lambda| = \left|\int \Big(\mathbb{E}[\widehat{p}^*(x)] - p^*(x)\Big)\,\mathrm{d}x\right| \leq \int |\mathrm{bias}[\widehat{p}^*]|\,\mathrm{d}x \leq C\|\mathbf{h}\|^2,$$

where C is independent of $\mathbf{h}$. According to (H7), we have $\|\mathbf{h}(n)\| \leq A\|\mathbf{N}\|^{-\alpha_0}$ for some $\alpha_0 \in (0, 1)$. Fix an arbitrary α that satisfies

$$\max(0, 1 - 4\alpha_0) < \alpha < 1 \quad \text{so that} \quad 4\alpha_0 > 1 - \alpha.$$

Then,

$$\|\mathbf{h}\|^2\|\mathbf{N}\|^{\frac{1-\alpha}{2}} \leq A\|\mathbf{N}\|^{-2\alpha_0}\|\mathbf{N}\|^{\frac{1-\alpha}{2}} = A\|\mathbf{N}\|^{\frac{-4\alpha_0+(1-\alpha)}{2}} \to 0 \quad \text{as}\, n \to \infty.$$

Thus there exists n_0 such that

$$C\|\mathbf{h}(n)\|^2 < \frac{\lambda}{4}\|\mathbf{N}(n)\|^{-\frac{(1-\alpha)}{2}} \quad \text{for all}\, n > n_0.$$

By the triangle inequality, we have

$$\left|\widehat{\lambda} - \mathbb{E}\widehat{\lambda}\right| > \left|\widehat{\lambda} - \lambda\right| - \left|\lambda - \mathbb{E}\widehat{\lambda}\right| > \left|\widehat{\lambda} - \lambda\right| - \frac{\lambda}{4}\|\mathbf{N}\|^{-\frac{(1-\alpha)}{2}},$$

and hence for every

$$\omega_0 \in \left\{\omega :\ |\widehat{\lambda}(\omega) - \lambda| > \frac{\lambda}{\|\mathbf{N}\|^{\frac{1-\alpha}{2}}}\right\} \tag{19}$$

we have

$$\left|\widehat{\lambda}(\omega_0) - \mathbb{E}\widehat{\lambda}\right| > \left|\widehat{\lambda}(\omega_0) - \lambda\right| - \frac{\lambda}{4}\|\mathbf{N}\|^{-\frac{(1-\alpha)}{2}} > \frac{3\lambda}{4}\|\mathbf{N}\|^{-\frac{(1-\alpha)}{2}} > \frac{\lambda}{\sqrt{2}}\|\mathbf{N}\|^{-\frac{(1-\alpha)}{2}}. \tag{20}$$

Then, (19) and (20), we obtain

$$\left\{\omega : |\widehat{\lambda}(\omega) - \lambda| > \frac{\lambda}{\|\mathbf{N}\|^{\frac{1-\alpha}{2}}}\right\} \subset \left\{\omega : |\widehat{\lambda}(\omega) - \mathbb{E}\widehat{\lambda}| > \frac{\lambda}{\sqrt{2}\|\mathbf{N}\|^{\frac{1-\alpha}{2}}}\right\},$$

and hence

$$\mathbb{P}\left\{\omega : |\widehat{\lambda}(\omega) - \lambda| > \frac{\lambda}{\|\mathbf{N}\|^{\frac{1-\alpha}{2}}}\right\} \leq \mathbb{P}\left\{\omega : |\widehat{\lambda}(\omega) - \mathbb{E}\widehat{\lambda}| > \frac{\lambda}{\sqrt{2}\|\mathbf{N}\|^{\frac{1-\alpha}{2}}}\right\} \leq \frac{2C_{int}}{\lambda^2\|\mathbf{N}\|^{\alpha}}.$$

□

4.3 Functional equivalent to MISE

As it was pointed, the functional MISE defined in (17) is not defined in the whole space Ω because the reciprocal of the renormalization random variable $(\widehat{\lambda})^{-1}$ may in general have en infinite expectation.

The event space Ω_n ensures that the constant $\widehat{c}$ has a finite expectation and stays close to the true normalization constant c. However, even on this smaller and safer space, the functional MISE$[\widehat{p}, p]$ is rather difficult to analyze. To help resolve this issue, we introduce a functional that is asymptotically equivalent to MISE on the space Ω_n

Definition 1

$$\overline{\text{MISE}} = \mathbb{E}\left[\left(\frac{\widehat{\lambda}}{\lambda}\right)^2 \int_{\mathbb{R}} (\widehat{p}(x) - p(x))^2 \mathrm{d}x\right]. \tag{21}$$

The equivalence follows from the definition of the space Ω_n

Proposition 1 *The functional* $\overline{\text{MISE}}$ *is asymptotically equivalent to* MISE *on smaller events* Ω_n *uniformly in n, that is*

$$\lim_{\|\mathbf{N}(n)\|\to\infty} \frac{\overline{\text{MISE}}\left[p(x), \widehat{p}(x;\omega)|\omega \in \Omega_n\right]}{\text{MISE}\left[p(x), \widehat{p}(x;\omega)|\omega \in \Omega_n\right]} = 1, \tag{22}$$

where

$$\Omega_n = \left\{\omega \in \Omega : |\widehat{\lambda} - \lambda| \leq \frac{\lambda}{\|\mathbf{N}\|^{\frac{1-\alpha}{2}}}\right\} \text{ with } \mathbb{P}(\Omega_n) \geq 1 - \frac{C}{\|\mathbf{N}\|^{\alpha}}, \tag{23}$$

and α *is a fixed constant satisfying* $1 > \alpha > \min(1 - 4\alpha_0, 0)$.

Proof Observe that

$$\begin{aligned}\overline{\mathrm{MISE}}\Big[p(x),\widehat{p}(x)|\Omega_n\Big] &= \frac{1}{\mathbb{P}(\Omega_n)}\int_{\Omega_n}\left(\frac{\widehat{\lambda}}{\lambda}\right)^2\int_{\mathbb{R}}(\widehat{p}(x,\omega)-p(x))^2\mathrm{d}x\,\mathbb{P}(d\omega)\\ &= \frac{1}{\mathbb{P}(\Omega_n)}\left(\int_{\Omega_n}\left[\left(\frac{\widehat{\lambda}}{\lambda}-1\right)^2+2\left(\frac{\widehat{\lambda}}{\lambda}-1\right)+1\right]\right.\\ &\qquad\left.\int_{\mathbb{R}}(\widehat{p}(x,\omega)-p(x))^2\mathrm{d}x\right)\mathbb{P}(d\omega).\end{aligned}$$

Then, by (23), we obtain that

$$\overline{\mathrm{MISE}}\Big[p(x),\widehat{p}(x)|\Omega_n\Big] = (1+\varepsilon(n))\mathrm{MISE}\Big[p(x),\widehat{p}(x)|\Omega_n\Big],$$

where

$$|\varepsilon(n)| \le \frac{C}{\|\mathbf{N}\|^{\frac{1-\alpha}{2}}},$$

for some constant $C > 0$ independent of n. This implies (22). □

One of the positive side effects we must mention is that the functional defined in (21) is not only easier to analyze but also it is defined throughout the whole space Ω. We take advantage of this fact and continue the discussion with expectations taken over the whole unrestricted space.

With the slight modification of the functional, we can now extract the leading-order part

Theorem 2 *The distance functional* $\overline{\mathrm{MISE}}$ *can be represented as*

$$\overline{\mathrm{MISE}}[p,\widehat{p},\mathbf{N},\mathbf{h}] = \overline{\mathrm{AMISE}}[p,\widehat{p},\mathbf{N},\mathbf{h}] + \mathcal{E}(\mathbf{N},\mathbf{h}), \tag{24}$$

where the leading term

$$\begin{aligned}\overline{\mathrm{AMISE}}[p,\widehat{p},\mathbf{N},\mathbf{h}] &:= \left(\int B(x)\,dx\right)^2\int\left(cp^*(x)\right)^2\mathrm{d}x\\ &\quad+\int(B(x))^2\,\mathrm{d}x+\int_{\mathbb{R}}(V(x))\,\mathrm{d}x\int_{\mathbb{R}}K^2(t)\,\mathrm{d}t\\ &\quad-2\iint B(y)\,B(x)\,cp^*(x)\,\mathrm{d}x\,\mathrm{d}y\\ B(x) &= \frac{ck_2}{2}\sum_{m=1}^{M}\left[h_m^2\,p_m''(x)\prod_{\substack{k=1\\k\neq m}}^{M}p_k(x)\right]\\ V(x) &= \sum_{m=1}^{M}\left[\frac{p_m}{N_m h_m}\prod_{\substack{k=1\\k\neq m}}^{M}p_k^2(x)\right],\end{aligned} \tag{25}$$

and the error term $\mathcal{E}$ satisfies

$$\mathcal{E}(\mathbf{N}, \mathbf{h}) = o\Big(||\mathbf{h}||^4 + \frac{1}{\|\mathbf{N}\|\|\mathbf{h}\|}\Big),$$

as $\mathbf{h} \to 0$, $\mathbf{N} \to \infty$, *and* $(\|\mathbf{N}\|\|\mathbf{h}\|)^{-1} \to 0$.

Proof We can divide the functional $\overline{\text{MISE}}$ into three components

$$\begin{aligned}\overline{\text{MISE}}[p, \widehat{p}] &= J_1 + J_2 + J_3 \\ &= c^2\mathbb{E}[(\lambda - \widehat{\lambda})^2] \int_{\mathbb{R}} (p(x))^2 \, \mathrm{d}x \\ &\quad + c^2\mathbb{E} \int_{\mathbb{R}} (\widehat{p}^* - p^*)^2 \mathrm{d}x \\ &\quad - 2c^2\mathbb{E} \int_{\mathbb{R}} (\widehat{\lambda} - \lambda)(\widehat{p}^* - p^*) p(x) \, \mathrm{d}x.\end{aligned}$$

Our first step will be to express each term J_i, i=1,…,3 as a sum of a higher order term and the term containing a bias, variance, or their combination. We then will use the results of the previous section and "Appendix" to obtain a leading part of each term.

First, observe that

$$\begin{aligned}\mathbb{E}[(\lambda - \widehat{\lambda})^2] &= \mathbb{E}[(\lambda - \mathbb{E}[\widehat{\lambda}])^2] + \mathbb{E}[(\mathbb{E}[\widehat{\lambda}] - \widehat{\lambda})^2] \\ &= \left(\mathbb{E}\left[\int p(x) - \widehat{p}(x) \, \mathrm{d}x\right]\right)^2 + \mathbb{E}[(\mathbb{E}[\widehat{\lambda}] - \widehat{\lambda})^2].\end{aligned}$$

The second term turns out to be of higher order. This can be seen from the following estimate:

$$\begin{aligned}\mathbb{E}[(\widehat{\lambda} - \mathbb{E}[\widehat{\lambda}])^2] &= \mathbb{E}\left[\int \big(\widehat{p}^* - \mathbb{E}\widehat{p}^*\big) \, \mathrm{d}x\right]^2 \\ &= \mathbb{E}\left[\int \big(\widehat{p}^* - \mathbb{E}\widehat{p}^*\big) \, \mathrm{d}x \cdot \int \big(\widehat{p}^* - \mathbb{E}\widehat{p}^*\big) \, \mathrm{d}x\right] \\ &= \mathbb{E}\left[\int \big(\widehat{p}^*(x) - \mathbb{E}\widehat{p}^*(x)\big) \, \mathrm{d}x \cdot \int \big(\widehat{p}^*(y) - \mathbb{E}\widehat{p}^*(y)\big) \, \mathrm{d}y\right] \\ &= \iint \big(\mathbb{E}\left[\widehat{p}^*(x)\widehat{p}^*(y)\right] - \mathbb{E}[\widehat{p}^*(x)]\mathbb{E}[\widehat{p}^*(y)]\big) \, \mathrm{d}x \, \mathrm{d}y \le \frac{C_1}{\|\mathbf{N}\|},\end{aligned}$$

where the last inequality follows from Lemma 3.

Thus, we conclude

$$J_1 = c^2 \left(\int \text{bias}[p^*, \widehat{p}^*] \, \mathrm{d}x\right)^2 \int \big(p(x)\big)^2 \, \mathrm{d}x + E_1 \quad \text{where } |E_1| \le \frac{C}{\|\mathbf{N}\|}.$$

From (4), we have that

$$J_2 = c^2 \int \Big(\mathrm{bias}^2[p^*, \widehat{p}^*] + \mathbb{V}[\widehat{p}^*] \Big) \, \mathrm{d}x.$$

The term J_3 can be expressed as

$$\begin{aligned} J_3 &= c^2 \mathbb{E}_n \iint \big(\widehat{p}^*(y) - p^*(y)\big)\Big(\big(\widehat{p}^*(x) - p^*(x)\big)p(x)\Big) \, \mathrm{d}y\mathrm{d}x \\ &= c^2 \iint \mathrm{bias}[\widehat{p}^*(y)]\mathrm{bias}[\widehat{p}^*(x)] \, p(x) \, \mathrm{d}x \, \mathrm{d}y \\ &\quad + c^2 \mathbb{E} \iint \big(\mathbb{E}[\widehat{p}^*(y)] - \widehat{p}^*(y)\big)\big(\mathbb{E}[\widehat{p}^*(x)] - \widehat{p}^*(x)\big)p(x) \, \mathrm{d}y\mathrm{d}x. \end{aligned}$$

Since $p^*(x)$ is uniformly bounded, Lemma 3 implies that the last term in the above identity satisfies

$$\left| c^2 \mathbb{E} \iint \big(\mathbb{E}[\widehat{p}^*(y)] - \widehat{p}^*(y)\big)\big(\mathbb{E}[\widehat{p}^*(x)] - \widehat{p}^*(x)\big) c p^*(x) \, \mathrm{d}y\mathrm{d}x \right| \leq \frac{C}{\|\mathbf{N}\|}.$$

This gives

$$J_3 = c^2 \iint \mathrm{bias}[\widehat{p}^*(y)]\mathrm{bias}[\widehat{p}^*(x)] \, p(x) \, \mathrm{d}x \, \mathrm{d}y + E_3, \quad |E_3| < \frac{1}{\|\mathbf{N}\|}.$$

Combining the above estimates gives

$$\begin{aligned} \overline{\mathrm{MISE}}\big[p, \widehat{p}\big] &= c^2 \left(\int \mathrm{bias}[p^*, \widehat{p}^*] \, \mathrm{d}x \right)^2 \int \big(p(x)\big)^2 \, \mathrm{d}x \\ &\quad + c^2 \int \mathrm{bias}^2[p^*, \widehat{p}^*] + \mathbb{V}[\widehat{p}^*] \, \mathrm{d}x \\ &\quad - 2c^2 \iint \mathrm{bias}[\widehat{p}^*(y)]\mathrm{bias}[\widehat{p}^*(x)] \, p(x) \, \mathrm{d}x \, \mathrm{d}y + E_M, \\ |E_M| &\leq \frac{C}{\|\mathbf{N}\|}. \end{aligned} \tag{26}$$

Applying the results of Lemma 1 and Lemma 2 to the identity (26) leads to (24) and (25), and this finishes the proof. □

4.4 Numerical optimization scheme for optimal bandwidth

In the absence of knowledge of probability density functions $p(x)$ and $p_m(x)$, it may seem that formula (25) has little practical use. However, this formula may be tweaked to produce an approximation of $\overline{\mathrm{AMISE}}$, which may be used to compute the

approximate optimal bandwidth. One can replace the densities $p(x)$, $p_m(x)$ with their approximations $\widehat{p}(x)$ and $\widehat{p}_m(x)$ in (25). This defines a function

$$\mathbf{h} \to \widehat{\overline{\text{AMISE}}}(\mathbf{h}), \tag{27}$$

which then can be minimized over $\mathbb{R}_+^M$.

There is a variety of iterative numerical optimization methods that can be employed to minimize (27). In cases where the function (27) is convex, algorithms such as gradient descent or conjugate gradient descent will be guaranteed to converge to the global minimizer (see Boyd and Vandenberghe 2004). Without the assumption of convexity, one may use stochastic gradient descent (Ge et al. 2015), which is guaranteed to converge to a local minimizer, or Nelder-Mead algorithm (Nelder and Mead 1965) for which convergence is not guaranteed, but it does not require the knowledge of the gradient.

The asymptotically leading part $\overline{\text{AMISE}}(\mathbf{h})$ is continuous in $\mathbf{h}$ and blows up as $\mathbf{h} \to \partial\mathbb{R}_+^M$ or $\mathbf{h} \to \infty$ and therefore must posses a minimizer. However, the conditions under which one can guarantee that a minimizer of (27) has an asymptotic behavior equivalent to the corresponding local minimizer of $\overline{\text{AMISE}}(\mathbf{h})$ as the number of samples increases is an open question, and it is the subject of an ongoing investigation.

5 Examples

In a general setting, finding a bandwidth vector $\mathbf{h}$ that minimizes (25) would require solving a system of nonlinear equations, which would probably not have a closed-form solution and require application of numerical methods. In this section, we discuss several special cases, for which closed-form solutions can be obtained with relative ease.

5.1 $\overline{\text{AMISE}}$ optimization for a symmetric case

In this case, we assume that all posterior densities for each subset of samples are the same and that all subsets contain the same number of samples. In other words, we employ the following assumptions:

1. $p_1(x) = p_2(x) = \cdots = p_M(x) = f(x)$,
2. $N_1 = N_2 = \cdots = N_M$, that is, $\mathbf{N} = (n, n, \ldots, n)$, for some $n \in \mathbb{N}$.

In view of the symmetry, all components of the optimal bandwidth vector should be the same, that is $\mathbf{h} = (h, h, \ldots, h)$. Under these assumptions, the expression for $\overline{\text{AMISE}}$ simplifies into

$$\overline{\text{AMISE}}[p(x), \widehat{p}(x)|\mathbf{N}, \mathbf{h}]$$
$$:= M^2 \frac{c^2 k_2^2 h^4}{4} \left(\int p_1''(x) p_1^{M-1}(x)\, \mathrm{d}x \right)^2 \int \left(c p_1^M(x)\right)^2 \mathrm{d}x$$

$$+ M^2 \frac{c^2 k_2^2 h^4}{4} \int \left(p_1''(x) p_1^{M-1}(x) \right)^2 \mathrm{d}x + M \int_{\mathbb{R}} \left(\frac{p_1^{2M-1}}{nh} \right) \mathrm{d}x \int_{\mathbb{R}} K^2(t)\, \mathrm{d}t$$
$$- M^2 \frac{c^3 k_2^2 h^4}{2} \iint \left(p_1''(y) p_1^{M-1}(y) \right) \left(p_1''(x) p_1^{2M-1}(x) \right) \mathrm{d}x\, \mathrm{d}y.$$

This expression achieves its minimum when $h = h^{\mathrm{opt}}$ where

$$h^{\mathrm{opt}} = (4n)^{-1/5} \left(\frac{B(M)}{A(M)} \right)^{1/5}, \tag{28}$$

and the constants A and B are given by

$$\begin{aligned} A(M) = M \frac{c^2 k_2^2}{4} \Bigg[& \left(\int_{\mathbb{R}} p_1''(x) p_1^{M-1}(x)\, \mathrm{d}x \right)^2 \int_{\mathbb{R}} \left(c p_1^M(x) \right)^2 \mathrm{d}x \\ & + \int_{\mathbb{R}} \left(p_1''(x) p_1^{M-1}(x) \right)^2 \mathrm{d}x \\ & - 2c \iint_{\mathbb{R}^2} \left(p_1''(y) p_1^{M-1}(y) \right) \left(p_1''(x) p_1^{2M-1}(x) \right) \mathrm{d}x\, \mathrm{d}y \Bigg] \\ B(M) = c^2 & \int_{\mathbb{R}} \left(p_1^{2M-1} \right) \mathrm{d}x \int_{\mathbb{R}} K^2(t)\, \mathrm{d}t. \end{aligned} \tag{29}$$

Forming the bandwidth vector $\mathbf{h}^{\mathrm{opt}} = (h^{\mathrm{opt}}, h^{\mathrm{opt}}, \ldots, h^{\mathrm{opt}})$ should yield a smaller value for $\overline{\mathrm{AMISE}}$ than the one achieved with the conventional choice given in (3).

5.2 $\overline{\mathrm{AMISE}}$ optimization for normal subset posterior densities

Let us assume that all subsets of samples of x satisfy

- $p_m = \mathcal{N}(x, \mu, \sigma)$ is a normal distribution with the same mean and standard deviation for each $m = 1, \ldots, M$,
- $N_1 = N_2 = \cdots = N_M$, that is, $\mathbf{N} = (n, n, \ldots, n)$, for some $n \in \mathbb{N}$.

Again, using symmetry argument, we look for the minimizer on the set of positive vectors $\mathbf{h} = (h, h, \ldots, h)$. In that case, the optimal $h = h^{\mathrm{opt}}$ is computed by (28), where constants A and B are computed by (29) with $p_1(x)$ replaced by $\mathcal{N}(x, \mu, \sigma)$. This gives

$$A(M) = \frac{3}{32 \pi^{1/2} M^{1/2} \sigma^5},$$

and

$$B(M) = \frac{M}{2 \pi^{1/2} \sqrt{2M - 1}},$$

and hence the minimizer of the leading part is given by

$$\mathbf{h}^{\text{opt}} = (1, 1, \ldots, 1)h^{\text{opt}} \quad \text{with} \quad h^{\text{opt}} = \left(\frac{16}{9}\frac{M^3}{(2M-1)}\right)^{1/10} \sigma n^{-1/5}. \tag{30}$$

Recall that n is the number of samples that each subset contains and hence the total number of samples for all subsets is given by $\|\mathbf{N}\|_1 = n \cdot M$. Thus, letting $M \to \infty$, we obtain

$$h^{\text{opt}} = \left((8/9)^{1/10} + O(M^{-1})\right)(n\,M)^{-1/5}\sigma \qquad \text{as} \quad M \to \infty.$$

Setting $M = 1$ in (30), we once again obtain the bandwidth vector

$$\mathbf{h}_0^{\text{opt}} = (1, 1, \ldots)h_{M=1}^{\text{opt}} \quad \text{with} \quad h_{M=1}^{\text{opt}} = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5}, \tag{31}$$

where each component $h_{M=1}^{\text{opt}}$ is the optimal bandwidth parameter for the individual subset posterior density estimator. Thus, the "intuitive" choice of the bandwidth vector as $\mathbf{h}_0^{\text{opt}}$ leads to a suboptimal approximation of $\widehat{p}(x)$.

5.3 $\overline{\text{AMISE}}$ optimization for gamma-distributed subset posterior densities

Let us assume that all subsets of samples of x satisfy

- $p_m = \Gamma(x, \alpha, \beta)$ is a gamma distribution where α and β are the same for each $m = 1, \ldots, M$,
- $N_1 = N_2 = \cdots = N_M$, that is, $\mathbf{N} = (n, n, \ldots, n)$, for some $n \in \mathbb{N}$.

By symmetry argument, we look for the minimizer on the set of positive vectors $\mathbf{h} = (h, h, \ldots, h)$. By substituting $p_1(x)$ by $\Gamma(x, \alpha, \beta)$ in (28) and (29), we can obtain formulas similar to the ones derived in the previous section. Evaluating the integrals is not very challenging; however, the integration results in very bulky expressions.

$$\begin{aligned}
h(n, M, \alpha) &= \frac{1}{(4n^2\pi)^{1/10}}\left(\frac{A}{B + C + D}\right)^{1/5}, \\
A &= 2^{2(\alpha-1)M}(2M-1)^{-2\alpha M+\alpha+2M-2}\Gamma(\alpha)\left(\frac{M}{\theta}\right)^{3(\alpha-1)M-1} \\
&\quad \times \theta^{3\alpha M-2M+4}\Gamma(2M\alpha - \alpha - 2M + 2), \\
B &= \frac{(\alpha-1)^2(M-1)^2M^2\left(\frac{M}{\theta}\right)^{(\alpha-1)M}\theta^{\alpha M}\Gamma(2(\alpha-1)M)\Gamma((\alpha-1)M-1)^2}{\Gamma((\alpha-1)M+1)^2}, \\
C &= 2\Big(M\big(\alpha\big(4(M-1)M+3\big) - 4(M-1)M - 15\big) + 9\Big)\left(\frac{M}{\theta}\right)^{(\alpha-1)M} \\
&\quad \times \theta^{\alpha M}\Gamma(2(\alpha-1)M - 3),
\end{aligned}$$

$$D = \frac{2(\alpha-1)(M-1)(2M-1)M^{(\alpha-1)M+1}\theta^M\Gamma((\alpha-1)M-1)\Gamma(2(\alpha-1)M-1)}{\Gamma((\alpha-1)M+1)}. \tag{32}$$

It must be noted that this result is very different from the normal distribution one ,and the suggested values of h are approximately thirty percent smaller than those in case of normal distribution even if the standard deviation of the samples is the same. This further necessitates the need for an easy-to-apply method for numerical approximation of the bandwidth vector $\mathbf{h}$, as the KDE method even for very similar families of distributions (such as normal and gamma ones) achieves best performance for very different bandwidth values. We discussed one such possible numerical scheme in Sect. 4.4.

5.4 Numerical experiments with normal subset posterior densities

5.4.1 Description of the experiment

The numerical experiment is designed to investigate the location of the optimal bandwidth parameter by approximating the true value of MISE$[p, \widehat{p}]$ by repeated simulation. One iteration of the experiment generates M subsets of a predetermined number of samples with $p_m = \mathcal{N}(x, 0, 1)$, $m = 1, \ldots, M$. Then, the approximation $\widehat{p}(x)$ is computed several times with varied bandwidth parameters h, and integrated square error ISE$[p(x), \widehat{p}(x), h]$ is then computed via numerical integration. The iteration is repeated a thousand times to obtain an approximation of MISE$[p(x), \widehat{p}(x), h]$ and its standard deviation. This process is repeated for varying sample sizes and numbers of subsets.

Once the data are collected, the minimum of MISE$[p(x), \widehat{p}(x), h]$ is located and the bandwidth parameter h for which the minimum is obtained is recorded. Since h computed this way is a random variable, the whole experiment is repeated a hundred times to compute the approximation of the expected value of h that minimizes MISE$[p(x), \widehat{p}(x), h]$ and its variance.

5.4.2 Numerical results

The experiments we ran allow us to compare the behavior of MISE$[p(x), \widehat{p}(x), \mathbf{h}]$ when we select $\mathbf{h} = \mathbf{h}_0^{\text{opt}}$ from (31) and when we select $\mathbf{h} = \mathbf{h}^{\text{opt}}$ from (30). Figure 1a, b demonstrates that the latter choice is clearly a superior one. The rate of decay of the error is very close to $O(\|\mathbf{N}\|^{-4/5})$, which is consistent with our calculations.

It must be noted that the graphs are plotted at the theoretically optimal values of $\mathbf{h}$, and the question of whether or not the error can be improved must be addressed. Our experiment computes the values of MISE for a variety of values of $\mathbf{h}$, and the bandwidth that produces the smallest error is indeed slightly different from our theoretical predictions. However, the discrepancy between them is negligible and it does become smaller as sample sizes increase.

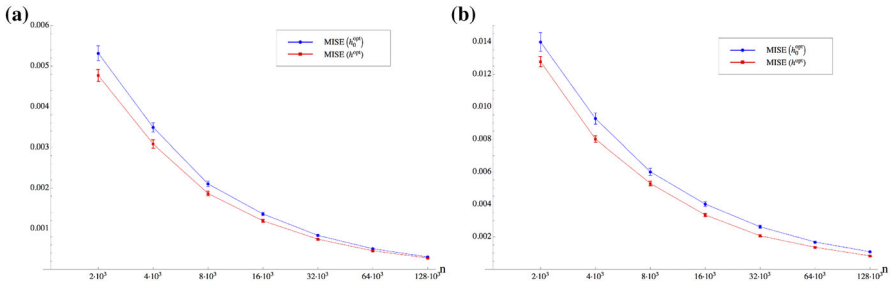

Fig. 1 MISE$[p, \widehat{p}, \mathbf{N}(n), \mathbf{h}]$ for $\mathbf{h}^{\text{opt}}$ and $\mathbf{h}_0^{\text{opt}}$. **a** $M = 4$, **b** $M = 8$

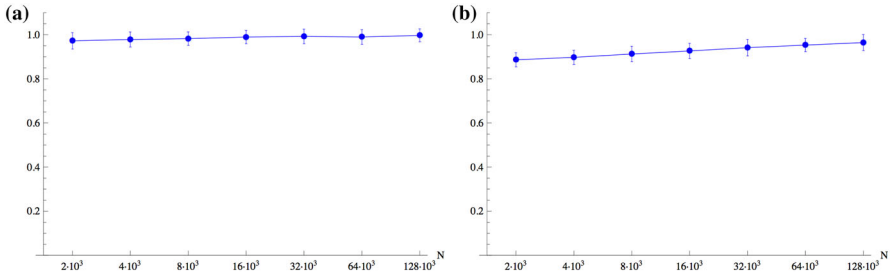

Fig. 2 Ratio $h^{\text{opt}}/h^{\text{opt}}_{\text{MISE}}$ for different subset configurations. **a** $M = 4$, **b** $M = 8$

Let us define

$$\begin{aligned}\mathbf{h}^{\text{opt}}_{\text{MISE}} &= \text{argmin}_{\mathbf{h}\in\mathbb{R}_+^M}\text{MISE}[p^*, \widehat{p}^*; \mathbf{N}, \mathbf{h}]\\ &= \text{argmin}_{h\in\mathbb{R}_+}\text{MISE}[p^*, \widehat{p}^*; \mathbf{N}, (h, h, \ldots, h)],\\ &= h^{\text{opt}}_{\text{MISE}} \cdot (1, 1, \ldots, 1),\end{aligned}$$

where the last two equalities hold in view of the symmetry assumption on p^*.

Figure 2 shows that the ratio of the numerically computed approximation of $\mathbf{h}^{\text{opt}}_{\text{MISE}}$ to the theoretically predicted value $\mathbf{h}^{\text{opt}}$ stays very close to one, which confirms the validity of our approach.

5.5 Numerical experiments with gamma-distributed subset posterior densities

5.5.1 Description of the experiment

The numerical experiment mimics the one with normally distributed samples, with the only difference that this experiment generates samples distributed with $\Gamma(x, \alpha = 3, \beta = 3)$.

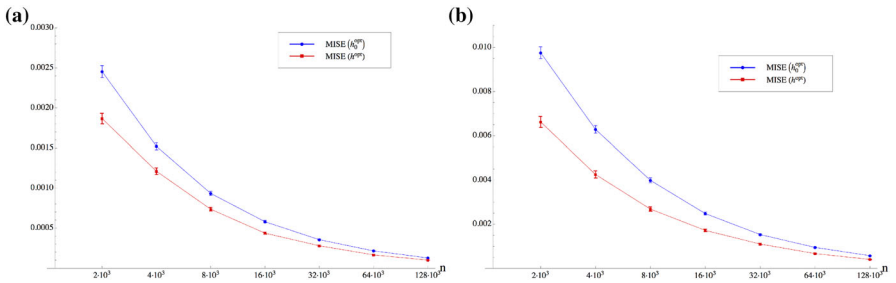

Fig. 3 MISE$[p, \widehat{p}, \mathbf{N}(n), \mathbf{h}]$ for $\mathbf{h}^{\text{opt}}$ and $\mathbf{h}_0^{\text{opt}}$. **a** $M = 4$, **b** $M = 8$

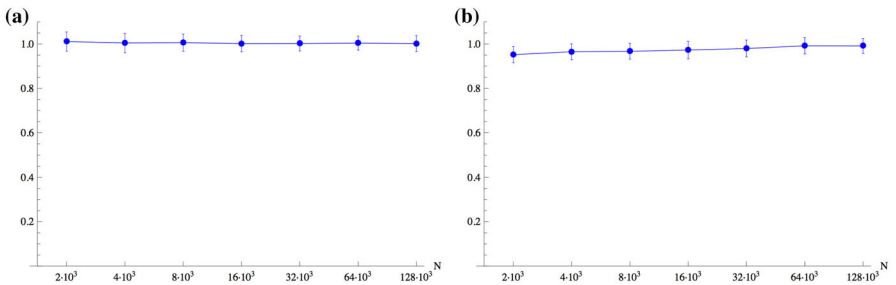

Fig. 4 Ratio of $h^{\text{opt}}/h_{\text{MISE}}^{\text{opt}}$ for different subset configurations. **a** $M = 4$, **b** $M = 8$

5.5.2 Numerical results

The results of the experiments replicate the same behavior for gamma-distributed samples. We must note that the location of the optimal bandwidth parameter is significantly different from that in the case of normally distributed samples. Nevertheless, the results clearly show the advantage of our choice of **h**, which is demonstrated in Fig. 3a, b.

Just as before, our experiment verifies that formula (32) yields near optimum values of MISE, see Fig. 4.

5.6 Numerical experiments with eruption data of the "Old Faithful" geyser

5.6.1 Description of the experiment

In this experiment, we employ the data of the waiting times between eruptions of the "Old Faithful" geyser in Yellowstone National Park. The data for the geyser eruption, between September 2009 and August 2011, were obtained from Geyser Observation and Study Association (2017).

We are interested in computing the posterior density estimator of the mode of the waiting time distribution given the data. To generate the samples of the mode, we assume that the waiting times are distributed according to a gamma distribution with the shape parameter α and rate parameter β. To generate samples from subset

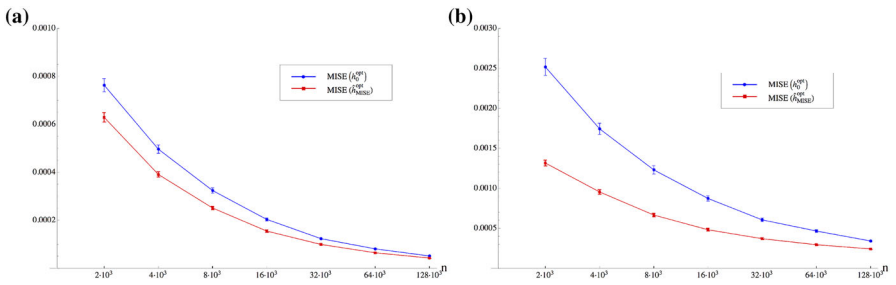

Fig. 5 MISE$[p, \widehat{p}, \mathbf{N}(n), \mathbf{h}]$ for $\widehat{\mathbf{h}}^{\text{opt}}_{\text{MISE}}$ and $\mathbf{h}^{\text{opt}}_{0}$. **a** $M = 4$, **b** $M = 8$

posterior distributions and full set posterior distribution of α and β, we use JAGS sampling package. We compute the mode samples given by the formula

$$\text{mode} = \frac{\alpha - 1}{\beta}.$$

The true posterior distribution of the mode is unknown. For this reason, we approximate the true full set posterior density of the mode, by computing a KDE estimate based on 10^7 samples generated with JAGS and bootstrapping using these samples to estimate the expectation of the estimator to reduce variance.

We construct subset posterior density estimators based on the number of samples $n \in \{2 \cdot 10^3, 4 \cdot 10^3, \ldots, 128 \cdot 10^3\}$ generated by JAGS. The data are scrambled before dividing it into subsets, to ensure that the data distributions are the same for each subset. We then construct the product posterior estimator (1), which is compared to the full set posterior approximation. We estimate MISE by averaging squared L^2 distance between the product posterior and the approximation of the true posterior over experiments repeated 10^3 times.

Due to inability to compute the minimizer of AMISE analytically, we opted to employ numerically computed estimation of the optimal bandwidth parameter

$$\widehat{\mathbf{h}}^{\text{opt}}_{\text{MISE}} = \text{argmin}_{\mathbf{h} \in \mathbb{R}^M_+} \widehat{\text{MISE}}[p^*, \widehat{p}^*; \mathbf{N}, \mathbf{h}].$$

The results of the experiments are plotted in Fig. 5a, b, and they are consistent with the behavior observed in the cases of synthetic samples.

6 Appendix

6.1 Kernel density estimators and asymptotic error analysis

In this section, we will use the following notation. The function f denotes a probability density, and its kernel density estimator is given by

$$\hat{f}(x; X_1, X_2, \ldots, X_N, h) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right). \tag{33}$$

where $X_1, X_2, \ldots X_n \sim f$ are i.i.d. samples.

Lemma 6 (Bias expansion) *Let K satisfy (H3) and (H4). Let f be a probability density function satisfying (H5) and (H6). Let $\widehat{f}_{n,h}(x)$ be an estimation of f given by (33). Then,*

(i) *$bias(\widehat{f}_{n,h})$ is given by*

$$\left[bias(\widehat{f}_{n,h})\right](x) = \mathbb{E}\left[\widehat{f}_{n,h}(x)\right] - f(x) = \frac{h^2 k_2 f''(x)}{2} + [E_b(f, K)](x; h), \tag{34}$$

where

$$E_b(x; h) := \int_{\mathbb{R}} K(t) \left(\int_x^{x-ht} \frac{f'''(z)(x - ht - z)^2}{2} \, \mathrm{d}z \right) \mathrm{d}t.$$

(ii) *For all $n \geq 1$ and $h > 0$, the term $E_b(\cdot; n, h)$ satisfies the bounds*

$$\begin{aligned} |E_b(x; h)| &\leq \frac{Ck_3}{6} h^3, \quad x \in \mathbb{R}, \\ \int_{\mathbb{R}} |E_b(x; h)| \, \mathrm{d}x &\leq C \frac{k_3}{6} h^3, \\ \int_{\mathbb{R}} |E_b(x; n, h)|^2 \, \mathrm{d}x &\leq \frac{C^2 k_3^2}{36} h^6, \end{aligned} \tag{35}$$

for some constant C.

(iii) *The square-integrated $bias(\widehat{f}_{n,k})$ satisfies*

$$\int_{\mathbb{R}} bias^2(\widehat{f}_{n,k}) \, \mathrm{d}x = \frac{h^4 k_2^2}{4} \int_{\mathbb{R}} (f''(x))^2 \, \mathrm{d}x + \mathcal{E}_b(n, h) < \infty$$

with

$$|\mathcal{E}_b(n, h)| \leq C_b \left(k_2 + \frac{k_3}{6} h\right) \frac{k_3 h^5}{6}, \tag{36}$$

for some constant C_b, and all $n \geq 1$, $h > 0$.

Proof Using (33) and the fact that X_i, $i = 1, \ldots, n$ are i.i.d., we obtain

$$\begin{aligned}\mathrm{bias}_{n,h}(x) &= \mathbb{E}\big[\widehat{f}_{n,h}(x)\big] - f(x)\\ &= \frac{1}{h}\mathbb{E}\left[K\left(\frac{x - X_1}{h}\right)\right] - f(x)\\ &= \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)\,\mathrm{d}y - f(x)\\ &= \int_{\mathbb{R}} K(t)(f(x-ht) - f(x))\,\mathrm{d}t,\end{aligned}$$

where we used the substitution $t = (x - y)/h$. Employing Taylor's theorem with an error term in integral form and using (H3), we get

$$\begin{aligned}&\mathrm{bias}_{n,h}(x)\\ &= \int_{\mathbb{R}} K(t)\left(-htf'(x) + \frac{h^2t^2}{2}f''(x) + \int_x^{x-ht} \frac{f'''(z)(x-ht-z)^2}{2}\,\mathrm{d}z\right)\mathrm{d}t\\ &= \frac{h^2 f''(x)}{2}\int_{\mathbb{R}} t^2 K(t)\,\mathrm{d}t + \int_{\mathbb{R}} K(t)\left(\int_x^{x-ht} \frac{f'''(z)(x-ht-z)^2}{2}\,\mathrm{d}z\right)\mathrm{d}t,\end{aligned}$$

which proves (i).

By (H4), we have

$$|E_b(x;n,h)| \leqslant C\left(\int_{\mathbb{R}} K(t)\left|\int_x^{x-ht} \frac{(x-ht-z)^2}{2}\,\mathrm{d}z\right|\mathrm{d}t\right) = \frac{Ck_3}{6}h^3, \tag{37}$$

and by (H6), using the substitution $\alpha = x - ht - z$ and employing Tonelli's theorem, we obtain

$$\begin{aligned}&\int_{\mathbb{R}} |E_b(x;n,h)|\,\mathrm{d}x\\ &\leq \int_{\mathbb{R}}\int_{\mathbb{R}} K(t)\int_{x-\frac{h}{2}(|t|+t)}^{x+\frac{h}{2}(|t|-t)} \frac{|f'''(z)|(x-ht-z)^2}{2}\,\mathrm{d}z\,\mathrm{d}t\,\mathrm{d}x\\ &= \int_{\mathbb{R}} K(t)\int_{-\frac{h}{2}(|t|+t)}^{\frac{h}{2}(|t|-t)}\left(\left(\int_{\mathbb{R}} |f'''(x-ht-\alpha)|\,\mathrm{d}x\right)\frac{\alpha^2}{2}\right)\mathrm{d}\alpha\,\mathrm{d}t\\ &\leq C\int_{\mathbb{R}} K(t)\left(\int_{-\frac{h}{2}(|t|-t)}^{\frac{h}{2}(|t|+t)} \frac{\alpha^2}{2}\,\mathrm{d}\alpha\right)\mathrm{d}t = \frac{h^3}{6}Ck_3.\end{aligned} \tag{38}$$

Thus, combining the two bounds above, we conclude

$$\int_{\mathbb{R}} |E_b(x;n,h)|^2\,\mathrm{d}x \leq \frac{Ck_3}{6}h^3\int_{\mathbb{R}} |E_b(x;n,h)|\,\mathrm{d}x \leq \frac{C^2k_3^2}{36}h^6.$$

Observe that

$$\text{bias}^2(\widehat{f}_{n,h})(x) = \frac{h^4 k_2^2}{4}(f''(x))^2 + h^2 k_2 f''(x) E_b(x;n,h) + E_b^2(x;n,h). \tag{39}$$

By (H5), (37), and (38)

$$\begin{aligned} \left|\mathcal{E}_b(n,h)\right| &:= \left| \int_{\mathbb{R}} \left(h^2 k_2 f''(x) E_b(x;n,h) + E_b^2(x;n,h) \right) \mathrm{d}x \right| \\ &\leq \left(h^2 k_2 C + \frac{C k_3}{6} h^3 \right) \int_{\mathbb{R}} |E_b(x;n,h)| \\ &\leq \left(h^2 k_2 C + \frac{C k_3}{6} h^3 \right) \frac{h^3}{6} C k_3. \end{aligned} \tag{40}$$

By (H5) and (H6), we have $\int_{\mathbb{R}} (f''(x))^2 \,\mathrm{d}x < \infty$. Hence, by setting $C_b = C^2$, using (39) and (40), we obtain (36). □

Lemma 7 (Variation expansion) *Let K satisfy (H3) and (H4), with $r = 2$. Let f satisfy (H5) and (H6), and $\widehat{f}_{n,h}(x)$ be the estimator of f given by (33). Then,*

(i) $\mathbb{V}(\widehat{f}_{n,h})$ *is given by*

$$\left[\mathbb{V}(\widehat{f}_{n,h})\right](x) = f(x)\frac{1}{nh}\int_{\mathbb{R}} K^2(t)\,\mathrm{d}t + E_V(x;n,h), \quad x \in \mathbb{R} \tag{41}$$

with

$$\begin{aligned} E_V(x;n,h) = -\frac{1}{n}\Bigg(&\int_{\mathbb{R}} t K^2(t) \int_0^1 f'(x - htu)\,\mathrm{d}u\,\mathrm{d}t \\ &+ \Big(f(x) + bias(\widehat{f}_{n,h})(x) \Big)^2 \Bigg). \end{aligned} \tag{42}$$

(ii) *The term $E_V(x;n,h)$ satisfies*

$$\begin{aligned} \mathcal{E}_V(n,h) &= \left| \int_{\mathbb{R}} E_V(x)\,\mathrm{d}x \right| \\ &\leq \frac{C_V}{n}\left(2 + h^2 k_2 + \left(k_2 + \frac{k_3}{3}h\right)\frac{h^5}{6}k_3 \right). \end{aligned} \tag{43}$$

Proof Using (34) and the fact that X_i, $i = 1, \ldots, n$ are i.i.d., we obtain

$$\begin{aligned} \mathbb{V}(\widehat{f}_{n,h}(x)) &= \mathbb{V}\left(\frac{1}{h} K\left(\frac{x - X_1}{h} \right) \right) \\ &= \frac{1}{n}\int_{\mathbb{R}} \frac{1}{h^2} K^2\left(\frac{x-y}{h} \right) f(y)\,\mathrm{d}y - \frac{1}{n}\left(\int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x-y}{h} \right) f(y)\,\mathrm{d}y \right)^2 \\ &= \frac{1}{nh}\int_{\mathbb{R}} K^2(t) f(x - ht)\,\mathrm{d}t - \frac{1}{n}\Big(f(x) + \text{bias}(\widehat{f}_{n,h})(x) \Big)^2 \end{aligned}$$

$$= \frac{1}{nh}\int_{\mathbb{R}} K^2(t) f(x)\,\mathrm{d}t + \frac{1}{nh}\int_{\mathbb{R}} K^2(t)\left(\int_x^{x-ht} f'(z)\,\mathrm{d}z\right)\mathrm{d}t$$
$$-\frac{1}{n}\Big(f(x) + \mathrm{bias}(\widehat{f}_{n,h})(x)\Big)^2$$
$$= \frac{1}{nh}\int_{\mathbb{R}} K^2(t) f(x)\,\mathrm{d}t - \frac{1}{n}\int_{\mathbb{R}} tK^2(t)\int_0^1 f'(x-htu)\,\mathrm{d}u\,\mathrm{d}t$$
$$-\frac{1}{n}\Big(f(x) + \mathrm{bias}(\widehat{f}_{n,h})(x)\Big)^2,$$

which proves (41) and (42).

We next estimate the terms

$$E_1(x) := \int_{\mathbb{R}} tK^2(t)\left(\int_0^1 f'(x-htu)\,\mathrm{d}u\right)\mathrm{d}t,$$
$$E_2(x) := \Big(f(x) + \mathrm{bias}(\widehat{f}_{n,h})(x)\Big)^2.$$

Observe that (H5)–(H6) imply

$$\int_{\mathbb{R}} |f'(x)|\,\mathrm{d}x = \int_{\mathbb{R}} |f'(x+\alpha)|\,\mathrm{d}x := I_1 < \infty$$

for any $\alpha \in \mathbb{R}$. Then, using Tonelli's Theorem and (H4), we obtain

$$\int_{\mathbb{R}} |E_1(x)|\,\mathrm{d}x \le \int_{\mathbb{R}} |t|K^2(t)\left(\int_{\mathbb{R}}\int_0^1 |f'(x-htu)|\,\mathrm{d}u\,\mathrm{d}x\right)\mathrm{d}t$$
$$\le \int_{\mathbb{R}} |t|K^2(t)\left(\int_0^1\left(\int_{\mathbb{R}} |f'(x-htu)|\,\mathrm{d}x\right)\mathrm{d}u\right)\mathrm{d}t \le I_1 k_1.$$

Since E_1 is integrable, we can use Fubini's theorem, and this yields

$$\int_{\mathbb{R}} E_1(x)\,\mathrm{d}x = \int_{\mathbb{R}} tK^2(t)\left(\int_{\mathbb{R}}\int_0^1 f'(x-htu)\,\mathrm{d}u\,\mathrm{d}x\right)\mathrm{d}t$$
$$= \int_{\mathbb{R}} tK^2(t)\left(\int_0^1\left(\int_{\mathbb{R}} f'(x-htu)\,\mathrm{d}x\right)\mathrm{d}u\right)\mathrm{d}t = 0,$$

where we used the fact that $\lim_{x\to\pm\infty} f(x) = 0$. Next, by (H5) and (35), we get

$$\int_{\mathbb{R}} |E_2(x)|\,\mathrm{d}x \le 2\int_{\mathbb{R}}\Big(f^2(x) + \mathrm{bias}^2(\widehat{f}_{n,h})(x)\Big)\,\mathrm{d}x$$
$$\le 2C + Ch^2k_2 + \left(k_2C + \frac{Ck_3}{6}h\right)\frac{h^5}{3}Ck_3.$$

Combining the above estimates, we obtain (43). □

Lemma 8 (Kernel autocorrelation) *Let K satisfy (H3) and (H4), then the function*

$$K_2(z) = \int_{\mathbb{R}} K(s)K(s-z)\,\mathrm{d}s \geq 0, \quad z \in \mathbb{R}$$

satisfies

$$\int_{\mathbb{R}} K_2(z)\,\mathrm{d}z = 1, \quad \int_{\mathbb{R}} z\,K_2(z)\,\mathrm{d}z \quad = 0.$$

Moreover, for any sufficiently smooth $f(x)$

$$\int \frac{1}{h} K_2\left(\frac{z-x}{h}\right) f(z)\,\mathrm{d}z = f(x) + E_{C,f} \quad \textit{with} \quad |E_{C,f}| \leq \|f''\|_\infty k_2 h^2.$$

Proof Since $K \geq 0$, we have $K_2 \geq 0$. Moreover, we have

$$\int_{\mathbb{R}} K_2(z)\,\mathrm{d}z = \iint_{\mathbb{R}\times\mathbb{R}} K(s)K(s-z)\,\mathrm{d}z\mathrm{d}s = 1$$

and this proves the first property. Similarly, recalling that $\int zK(z)\,\mathrm{d}z = 0$, we obtain

$$\int_{\mathbb{R}} z\,K_2(z)\,\mathrm{d}z = \int_{\mathbb{R}} K(s) \int_{\mathbb{R}} (z-s+s)K(s-z)\,\mathrm{d}z\,\mathrm{d}s = 0.$$

Next, we take any smooth function f and compute

$$\begin{aligned}\int \frac{1}{h} K_2\left(\frac{z-x}{h}\right) f(z)\,\mathrm{d}z &= \int K_2(u)\, f(x-hu)\,\mathrm{d}u \\ &= f(x) + \int K_2(u) \int_x^{x-hu} f''(t)(t-x+hu)\,\mathrm{d}t\,\mathrm{d}u.\end{aligned}$$

Finally, we estimate the last term in the above formula as follows:

$$\begin{aligned}&\left|\int K_2(u) \int_x^{x-hu} f''(t)(t-x+hu)\,\mathrm{d}t\,\mathrm{d}u\right| \\ &\quad\leq \|f''\|_\infty \int K_2(u)\frac{h^2u^2}{2}\,\mathrm{d}u \\ &\quad= \frac{\|f''\|_\infty h^2}{2}\left(\int K(s)\int (s-u)^2K(s-u)\,\mathrm{d}u\mathrm{d}s\right. \\ &\qquad\qquad\left. + \int s^2K(s)\int K(s-u)\,\mathrm{d}u\mathrm{d}s\right) \\ &\quad\leq \|f''\|_\infty k_2 h^2.\end{aligned}$$

□

Lemma 9 (Product expectation) *Let K satisfy (H3) and (H4), with $r = 2$. Let f be a probability density function that satisfies (H5) and (H6), and let $\widehat{f}_{n,h}(x)$ be an estimate of f given by (33). Then,*

$$\mathbb{E}[\widehat{f}_{n,h}(x)\widehat{f}_{n,h}(y)] - \mathbb{E}[\widehat{f}_{n,h}(x)]\mathbb{E}[\widehat{f}_{n,h}(y)] = \frac{1}{Nh} f(x) K_2\left(\frac{x-y}{h}\right) - E_{\Pi}, \quad (44)$$

where the error term

$$E_{\Pi} = \frac{1}{N}\int \left(sK(s)K\left(s - \frac{x-y}{h}\right)\left(\int_0^1 f'(x - shu)\,\mathrm{d}u\right)\right)\mathrm{d}s + \frac{1}{N}\mathbb{E}[\widehat{f}(x)]\mathbb{E}[\widehat{f}(y)]$$

satisfies

$$|E_{\Pi}(x,y)| \le \frac{C_{\Pi}}{N}, \qquad \left|\int\int E_{\Pi}(x,y)\,\mathrm{d}x\mathrm{d}y\right| \le \frac{1}{N}\left(1 + \frac{Ck_3h^3}{6}\right)^2$$
$$\int\int |E_{\Pi}(x,y)|\mathrm{d}x\mathrm{d}y \le \frac{1}{N}\left(1 + k_1\, C\frac{Ck_2h^2}{2} + \frac{Ck_3h^3}{6}\right)^2,$$

for some constant C_{Π} and constants C given in (H6) and K_2 defined in Lemma 8.

Proof By the definition of the estimator $\widehat{f}$, we have

$$\mathbb{E}\Big(\widehat{f}(x)\widehat{f}(y)\Big) = \mathbb{E}\left(\frac{1}{N^2h^2}\sum_{i,j=1}^{N} K\left(\frac{x-X_i}{h}\right)K\left(\frac{y-X_j}{h}\right)\right).$$

Since all $\{X_i\}_{i=1}^N$ are i.i.d., we can split the calculation into two parts, one for the part, where the indexes coincide and the part, where indexes are different. We then can use the independence of the samples to simplify the calculation

$$\begin{aligned}\mathbb{E}\Big(\widehat{f}(x)\widehat{f}(y)\Big) &= \frac{1}{N^2h^2}\mathbb{E}\left(\sum_{i=j} K\left(\frac{x-X_i}{h}\right)K\left(\frac{y-X_i}{h}\right)\right)\\ &\quad + \frac{1}{N^2h^2}\mathbb{E}\left(\sum_{i\neq j} K\left(\frac{x-X_i}{h}\right)K\left(\frac{y-X_j}{h}\right)\right) \qquad (45)\\ &= \frac{1}{Nh^2}\left[\mathbb{E}\left(K\left(\frac{x-X}{h}\right)K\left(\frac{y-X}{h}\right)\right)\right]\\ &\quad + \left(1 - \frac{1}{N}\right)\mathbb{E}[\widehat{f}(x)]\mathbb{E}[\widehat{f}(y)],\end{aligned}$$

where $X = X_1$. The first expectation term in (45) can be expanded as

$$\begin{aligned}
&\frac{1}{Nh^2}\mathbb{E}\left[K\left(\frac{x-X}{h}\right)K\left(\frac{y-X}{h}\right)\right] \\
&= \frac{1}{Nh^2}\int K\left(\frac{x-t}{h}\right)K\left(\frac{y-t}{h}\right) f(t)\,\mathrm{d}t \\
&= \frac{1}{Nh}\int K(s)K\left(s-\frac{x-y}{h}\right)\left(f(x)+\int_x^{x-sh} f'(z)\,\mathrm{d}z\right)\mathrm{d}s \\
&= f(x)\frac{1}{Nh}K_2\left(\frac{x-y}{h}\right) \\
&\quad -\frac{1}{N}\int sK(s)K\left(s-\frac{x-y}{h}\right)\left(\int_0^1 f'\left(x-shu\right)\,\mathrm{d}u\right)\mathrm{d}s.
\end{aligned}$$

Let us denote

$$\begin{aligned}
E_{\Pi,1} &= \frac{1}{N}\int\left(sK(s)K\left(s-\frac{x-y}{h}\right)\left(\int_0^1 f'(x-shu)\,\mathrm{d}u\right)\right)\mathrm{d}s, \\
E_{\Pi,2} &= \frac{1}{N}\mathbb{E}[\widehat{f}(x)]\mathbb{E}[\widehat{f}(y)].
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
&\mathbb{E}\left(\widehat{f}_{n,h}(x)\widehat{f}_{n,h}(y)\right) - \mathbb{E}[\widehat{f}_{n,h}(x)]\mathbb{E}[\widehat{f}_{n,h}(y)] \\
&\quad = f(x)\frac{1}{Nh}K_2\left(\frac{x-y}{h}\right)\mathrm{d}s - (E_{\Pi,1}+E_{\Pi,2}),
\end{aligned}$$

and this establishes (44).

Observe that (H3), (H4), and (H5) imply

$$|E_{\Pi,1}| \le \frac{C\,k_1}{N}.$$

Next, according to (34) and (35)

$$|\mathbb{E}[\widehat{f}(x)]| \le C + \frac{Ck_2h^2}{2} + \frac{Ck_3h^3}{6} \quad \text{for all} \quad x \in \mathbb{R},$$

where C is a maximum of constants from (H5) and hence

$$|E_{\Pi,2}| \le \frac{1}{N}\left(C + \frac{Ck_2h^2}{2} + \frac{Ck_3h^3}{6}\right)^2.$$

Combining the above estimate, we conclude that

$$|E_{\Pi}| = |E_{\Pi,1} + E_{\Pi,2}| \le \frac{1}{N}\left(Ck_1 + \left(C + \frac{Ck_2h^2}{2} + \frac{Ck_3h^3}{6}\right)^2\right).$$

To obtain bounds on the integral of the error term, let us consider each component of the error separately. The term $E_{\Pi,1}$ is integrable

$$\begin{aligned}
&\iint |E_{\Pi,1}(x,y)|\,\mathrm{d}x\mathrm{d}y \\
&\le \frac{1}{N}\iiint_{\mathbb{R}^3} |s|\,K(s)K\left(s - \frac{x-y}{h}\right)\left(\int_0^1 |f'(x-shu)|\,\mathrm{d}u\right)\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}y \qquad (46)\\
&\le \frac{1}{N}\int_{\mathbb{R}} |s|\,K(s)\left(\int_0^1\int_{\mathbb{R}} |f'(x-shu)|\,\mathrm{d}x\,\mathrm{d}u\right)\mathrm{d}s \le \frac{k_1 C}{N}.
\end{aligned}$$

Next using Fubini's theorem, we obtain

$$\begin{aligned}
&\left|\iint E_{\Pi,1}(x,y)\,\mathrm{d}x\mathrm{d}y\right| \\
&\le \frac{1}{N}\left|\iiint_{\mathbb{R}^3} s\,K(s)K\left(s - \frac{x-y}{h}\right)\left(\int_0^1 f'(x-shu)\,\mathrm{d}u\right)\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}y\right| \\
&= \frac{1}{N}\left|\int_{\mathbb{R}} s\,K(s)\left(\int_0^1\int_{\mathbb{R}} f'(x-shu)\,\mathrm{d}x\,\mathrm{d}u\right)\mathrm{d}s\right| = 0.
\end{aligned}$$

Therefore, using Lemma 6, (34), (35), and the Hypothesis (H6), we obtain

$$\left|\iint_{\mathbb{R}^2} E_{\Pi}(x,y)\,\mathrm{d}x\mathrm{d}y\right| = \frac{1}{N}\left|\int_{\mathbb{R}} \mathbb{E}[\widehat{f}(x)]\,\mathrm{d}x\right|^2 \le \frac{1}{N}\left(1 + \frac{Ck_3h^3}{6}\right)^2.$$

Finally, directly from (46), (34), and (35), we obtain

$$\begin{aligned}
\iint_{\mathbb{R}^2} |E_{\Pi}(x,y)|\,\mathrm{d}x\mathrm{d}y &\le \iint_{\mathbb{R}^2} \left|E_{\Pi,1}(x,y)\right|\mathrm{d}x\mathrm{d}y + \iint_{\mathbb{R}^2} \left|E_{\Pi,2}(x,y)\right|\mathrm{d}x\mathrm{d}y \\
&\le \frac{k_1 C}{N} + \frac{1}{N}\left(1 + \frac{Ck_2h^2}{2} + \frac{Ck_3h^3}{6}\right)^2.
\end{aligned}$$

□

Theorem 3 (MISE expansion) *Let K satisfy (H3) and (H4), with $r = 2$. Let f be a probability density function that satisfies (H5) and (H6), and let $\widehat{f}_{n,h}(x)$ be an estimate of f given by (33). Then,*

$$\mathrm{MISE}(\widehat{f}_{n,h}) = \frac{h^4k_2^2}{4}\int_{\mathbb{R}} (f''(x))^2\mathrm{d}x + \frac{1}{nh}f(x)\int_{\mathbb{R}} K^2(t)\,\mathrm{d}t + \mathcal{E}_b(n,h) + \mathcal{E}_V(n,h)$$

with $\mathcal{E}_b$ and $\mathcal{E}_V$ defined in (40) and (43), respectively. Moreover, for every $H > 0$, there exists $C_{f,K,H}$ such that

$$|\mathcal{E}_b(h,n) + \mathcal{E}_V(h,n)| \leqslant C_{f,K,H}\left(h^5 + \frac{1}{n}\right)$$

for all $n \geq 1$ and $H \geq h > 0$.

Proof It is easy to show (see Silverman 1986) that

$$\begin{aligned}\mathrm{MISE}(\widehat{f}_{n,h}) &= \int_{\mathbb{R}} \mathbb{E}[\widehat{f}_{n,h}(x) - f(x)]^2 \,\mathrm{d}x \\ &= \int_{\mathbb{R}} \left(\mathrm{bias}(\widehat{f}_{n,h})(x)\right)^2 \mathrm{d}x + \int_{\mathbb{R}} \mathbb{V}(\widehat{f}_{n,h}(x)) \,\mathrm{d}x,\end{aligned}$$

and hence the result follows from Lemma 6 and Lemma 7. □

References

Boyd, S., Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

De Valpine, P. (2004). Monte Carlo state-space likelihoods by weighted posterior kernel density estimation. *Journal of the American Statistical Association*, *99*(466), 523–536.

Duong, T., Hazelton, M. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, *32*(3), 485–506.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, *14*(1), 153–158.

Ge, R., Huang, F., Jin, C., Yuan, Y. (2015). Escaping from saddle points—Online stochastic gradient for tensor decomposition. In *Proceedings of conference on learning theory* (pp. 797–842), 3–6 July.

Geyser Observation and Study Association. (2017). Old faithful. http://www.geyserstudy.org/geyser.aspx?pGeyserNo=OLDFAITHFUL. Accessed 21 Oct 2017.

Langford, J., Smola, A. J., Zinkevich, M. (2009). Slow learners are fast. *Advances in Neural Information Processing Systems*, *22*, 2331–2339.

Miroshnikov, A., Wei, Z., Conlon, E. M. (2015). Parallel Markov chain Monte Carlo for non-Gaussian posterior distributions. *Stat*, *4*(1), 304–319.

Neiswanger, W., Wang, C., Xing, E. P. (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the thirtieth conference on uncertainty in artificial intelligence* (pp. 623–632). AUAI Press, 23–27 July.

Nelder, J. A., Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313.

Newman, D., Asuncion, A., Smyth, P., Welling, M. (2009). Distributed algorithms for topic models. *Journal of Machine Learning Research*, **10**, 1801–1828.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076.

Propp, J. G., Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, *9*(1–2), 223–252.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*(3), 832–837.

Scott, S. L. (2017). Comparing consensus Monte Carlo strategies for distributed Bayesian computation. *Brazilian Journal of Probability and Statistics*, *31*(4), 668–685.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, *11*(2), 78–88.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). New York: Chapman & Hall/CRC.

Simonoff, J. (1996). *Smoothing methods in statistics. Springer series in statistics*. New York: Springer.

Sköld, M., Roberts, G. O. (2003). Density estimation for the Metropolis-Hastings algorithm. *Scandinavian Journal of Statistics*, *30*(4), 699–718.

Smola, A., Narayanamurthy, S. (2010). An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, *3*(1–2), 703–710.

van Eeden, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous. *Annals of the Institute of Statistical Mathematics*, *37*(1), 461–472.

Wand, M. P., Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, *9*(2), 97–116.

Wang, X., Dunson, D. B. (2013). Parallelizing MCMC via Weierstrass sampler. arXiv preprint arXiv:1312.4605.

West, M. (1993). Approximating posterior distributions by mixture. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(2), 409–422.