Asymptotic properties and approximation of Bayesian logspline density estimators for communication-free parallel computing methods

A. Miroshnikov, K. Kotsiopoulos, E. M. Conlon

Abstract

In this article we perform an asymptotic analysis of Bayesian parallel density estimators which are based on logspline density estimation. The parallel estimator we introduce is in the spirit of a kernel density estimator introduced in recent studies. We provide a numerical procedure that produces the density estimator itself in place of the sampling algorithm. We then derive an error bound for the mean integrated squared error for the full data posterior density estimator. We also investigate the parameters that arise from logspline density estimation and the numerical approximation procedure. Our investigation identifies specific choices of parameters for logspline density estimation that result in the error bound scaling appropriately in relation to these choices.

Keywords: logspline density estimation, asymptotic properties, error analysis, parallel algorithms.

Mathematics Subject Classification (2000): 62G07, 62G20, 68W10.

Contents

1 Introduction

The recent advances in data science and big data research have brought challenges in analyzing large data sets in full. These massive data sets may be too large to read into a computers memory in full, and data sets may be located on different machines. In addition, there is a lengthy time needed to process these data sets. To alleviate these difficulties, many parallel computing methods have recently been developed. One such approach partitions large data sets into subsets, where each subset is analyzed on a separate machine using parallel Markov chain Monte Carlo (MCMC) methods [8, 9, 10]; here, communication between machines is required for each MCMC iteration, increasing computation time.

Due to the limitations of methods requiring communication between machines, a number of alternative communication- free parallel MCMC methods have been developed for Bayesian analysis of big data [5, 6]. For these approaches, Bayesian MCMC analysis is performed on each subset independently, and the subset posterior samples are combined to estimate the full data posterior distributions. Neiswanger, Wang and Xing [5] introduced a parallel kernel density estimator that first approximates each subset posterior density and then estimates the full data posterior by multiplying together the subset posterior estimators. The authors of [5] show that the estimator they use is asymptotically exact; they then develop an algorithm that generates samples from the posterior distribution approximating the full data posterior estimator. Though the estimator is asymptotically exact, the algorithm of [5] does not perform well for posteriors that have non-Gaussian shape. This under-performance is attributed to the method of construction of the subset posterior densities; this method produces near-Gaussian posteriors even if the true underlying distribution is non-Gaussian. Another limitation of the method of Neiswanger, Wang and Xing is its use in high-dimensional parameter spaces, since it becomes impractical to carry out this method when the number of model parameters increases.

Miroshnikov and Conlon [6] introduced a new approach for parallel MCMC that addresses the limitations of [5]. Their method performs well for non-Gaussian posterior distributions and only analyzes densities marginally for each parameter, so that the size of the parameter space is not a limitation. The authors use logspline density estimation for each subset posterior, and the subsets are combined by a direct numeric product of the subset posterior estimates. However, note that this technique does not produce joint posterior estimates, as in [5].

The estimator introduced in [6] follows the ideas of Neiswanger et al. [5]. Specifically, let $p(\mathbf{x}|\theta)$ be the likelihood of the full data given the parameter $\theta \in \mathbb{R}$. We partition $\mathbf{x}$ into M disjoint subsets $\mathbf{x}_m$, with $m \in \{1, 2, ..., M\}$. For each subset we draw N samples $\theta_1^m, \theta_2^m, ..., \theta_N^m$ whose distribution is given by the subset posterior density $p(\theta|\mathbf{x}_m)$. Given prior $p(\theta)$, the datasets $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M$ and assuming that they are independent from each other, then the posterior density, see [5], is expressed by

$$p(\theta|\mathbf{x}) \propto p(\theta) \prod_{m=1}^{M} p(\mathbf{x}_m|\theta) = \prod_{m=1}^{M} p_m(\theta) =: p^*(\theta)\,, \quad \text{where} \quad p(\theta|\mathbf{x}_m) := p_m(\theta) = p(\mathbf{x}_m|\theta)p(\theta)^{1/M}. \quad (1)$$

In our work, we investigate the properties of the estimator $\hat{p}(\theta|\mathbf{x})$, defined in [6], that has the form

$$\hat{p}(\theta|\mathbf{x}) \propto \prod_{m=1}^{M} \hat{p}_m(\theta) =: \hat{p}^*(\theta)\,, \quad (2)$$

where $\hat{p}_m(\theta)$ is the logspline density estimator of $p_m(\theta)$ and where we suppressed the information about the data x.

The estimated product $\hat{p}^*$ of the subset posterior densities is, in general, unnormalized. This motivates us to define the normalization constant $\hat{c}$ for the estimated product $\hat{p}^*$. Thus, the normalized density $\hat{p}$, one of the main points of interest in our work, is given by

$$\hat{p}(\theta) = \hat{c}^{-1}\hat{p}^*(\theta), \quad \text{where} \quad \hat{c} = \int \hat{p}^*(\theta)\, d\theta.$$

Computing the normalization constant analytically is a difficult task since the subset posterior densities are not explicitly calculated, with the exception of a finite number of points $\big(\theta_i, \hat{p}^*_m(\theta_i)\big)$, where $i \in \{1, \dots, n\}$. By taking the product of these values for each i we obtain the value of $\hat{p}^*(\theta_i)$. This allows us to numerically approximate the unnormalized product $\hat{p}^*$ by using a Lagrange interpolation polynomials. This approximation is denoted by $\tilde{p}^*$. Then we approximate the constant $\hat{c}$ by numerically integrating $\tilde{p}^*$. The approximation of the normalization constant $\hat{c}$ is denoted by $\tilde{c}$, given by

$$\tilde{c} = \int \tilde{p}^*(\theta)\, d\theta, \text{ and we set } \tilde{p}(\theta) := \tilde{c}^{-1}\tilde{p}^*(\theta).$$

The newly defined density $\tilde{p}$ acts as the estimator for the full-data posterior p.

In this paper, we establish error estimates between the three densities via the mean integrated squared error or MISE, defined for two functions f, g as

$$\text{MISE}(f, g) := \mathbb{E} \int \big(f(\theta) - g(\theta)\big)^2 d\theta. \tag{3}$$

Thus, our work involves two types of approximations: 1) the construction of $\hat{p}^*$ using logspline density estimators and 2) the construction of the interpolation polynomial $\tilde{p}^*$. The methodology of logspline density estimation was introduced in [2] and corresponding error estimates between the estimator and the density it is approximating are presented in [3, 4]. These error estimates depend on three factors: i) the N_m number of samples drawn from the subset posterior density, ii) the $K_m + 1$ number of knots used to create the k-order B-splines, and iii) the step-size of those knots, which we denote by h_m.

In our work we estimate the MISE between the functions $\hat{p}^*$ and p^* by adapting the estimation techniques introduced in [3, 4]. We then utilize this analysis to establish a similar estimate for the normalized densities $\hat{p}$ and p,

$$\text{MISE}(p^*, \hat{p}^*) = O\left[\left(\exp\left\{\sum_{m=1}^{M} \frac{K_m + 1 - k}{N_m^{1/2}} + h_{max}^{j+1} \sum_{m=1}^{M} \left\| \frac{d^{j+1}\log(p_m)}{d\theta^{j+1}} \right\|_\infty \right\} - 1\right)^2\right],$$

where $h_{max} = \max_m\{h_m\}$ and $j + 1$ is the number of continuous derivatives of p. Notice that the exponential contains two terms, where the first depends on the number of samples and the number of knots and the other depends on the placement of the spline knots. Both terms converge to zero and for MISE to scale optimally both terms must converge at the same rate. To this end, we choose h_{max} and each K_m to be functions of the vector $\mathbf{N} = \big\{N_1, \dots, N_M\big\}$ and scale appropriately with the norm $\|\mathbf{N}\|$. This simplifies the above estimate to

$$\text{MISE}(p^*, \hat{p}^*) = O\left(M^{2-2\beta}\|\mathbf{N}\|^{-2\beta}\right)$$

where the parameter $\beta \in (0, 1/2)$ is related to the convergence of the logspline density estimators.

The estimate for MISE between $\tilde{p}^*$ and $\hat{p}^*$ is obtained in a similar way by utilizing Lagrange interpolation error bounds, as described in [7]. This error depends on two factors: i) the step-size Δx of the grid points chosen to construct the polynomial, where the grid points correspond to the coordinates $\big(\theta_i, \hat{p}^*_m(\theta_i)\big)$ discussed earlier, and ii) the degree l of the Lagrange polynomial. The estimate obtained is also shown to hold for the normalized densities $\tilde{p}$ and $\hat{p}$.

$$\text{MISE}(\hat{p}^*, \tilde{p}^*) = O\left[\left(\frac{\Delta x}{h_{min}(\mathbf{N})} M\right)^{2(l+1)}\right]$$

where $h_{min}(\mathbf{N})$ is the minimal distance between the spline knots and is chosen to asymptotically scale with the norm of the vector of samples $\mathbf{N}$, see Section **??**.

We then combine both estimates to obtain a bound for MISE for the densities p and $\tilde{p}$. We obtain

$$\text{MISE}(p\,,\tilde{p}) = O\left[M^{2-2\beta}\|\mathbf{N}\|^{-2\beta} + \left(\frac{\Delta x}{h_{min}(\mathbf{N})}M\right)^{2(l+1)}\right].$$

In order for MISE to scale optimally the two terms in the sum must converge to zero at the same rate. As before with the distance between $\hat{p}^*$ and p^*, we choose Δx to scale appropriately with the norm of the vector $\mathbf{N}$. This leads to the optimum error bound for the distance between the estimator $\tilde{p}$ and the density p,

$$\text{MISE}(p\,,\tilde{p}) = O\left(\|\mathbf{N}\|^{-2\beta}\right) \quad \text{where we choose} \quad \Delta x = O\left(\|\mathbf{N}\|^{-\beta\left(\frac{1}{l+1}+\frac{1}{j+1}\right)}\right). \tag{4}$$

The paper is arranged as follows. In Section 2 we set notation and hypotheses that form the foundation of the analysis. In Section 3 we derive an asymptotic expansion for MISE of the non-normalized estimator, which are central to the analysis performed in subsequent sections. We also perform there the analysis of MISE for the full data set posterior density estimator $\hat{p}$. In Section 4, we perform the analysis for the numerical estimator $\tilde{p}$. In Section 5 we showcase our simulated experiments and discuss the results. Finally, in the appendix we provide supplementary lemmas and theorems employed in Section 3 and Section 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. [5] we consider the probability density function of the form

$$p(\theta) \propto p^*(\theta) \quad \text{where} \quad p^*(\theta) := \prod_{m=1}^{M} p_m(\theta) \tag{5}$$

where we assume that $p_m(\theta),\ m \in \{1,\dots,M\}$ have compact support on the interval $[a,b]$.

(H2) For each $m \in \{1,\dots,M\}$ $p_m(\theta)$ is a probability density function. We consider the estimator of p in the form

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

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

$$\hat{p}_m : \mathbb{R}\times\Omega^m_{n_m} \quad \text{defined by} \quad \hat{p}_m(\theta,\omega) = f_m(\theta, \hat{y}(\theta^m_1,\dots,\theta^m_{n_m})),\ \omega\in\Omega^m_{n_m} \tag{H2-b}$$

We also consider the additional estimators $\bar{p}_m$ of p_m as defined in (71) and

$$\bar{p}^*(\theta) := \prod_{m=1}^{M} \bar{p}_m(\theta).$$

Here $\theta^m_1, \theta^m_2, \dots, \theta^m_{n_m} \sim p_m(x)$ are independent identically distributed random variables and f_m is the logspline density estimate introduced in Definition (37) with N_m number of knots and the order of the B-splines is k_m.

$$\Omega^m_{n_m} = \left\{\omega\in\Omega : \hat{y} = \hat{y}(\theta^m_1,\dots,\theta^m_{n_m}) \in \mathbb{R}^{L_m+1} \ \text{ exists}\right\}. \tag{6}$$

where $L_m := N_m - k_m$.

The mean integrated square error of the estimator $\hat{p}^*$ of the product p^* is defined by

$$\mathrm{MISE}_{[\mathbf{N}]} := \mathrm{MISE}(p^*, \hat{p}^*) = \mathbb{E} \int \left(\hat{p}^*(\theta;\omega) - p^*(\theta)\right)^2 d\theta \tag{7}$$

where we use the notation $\mathbf{N} = (N_m)_{m=1}^N$.

We assume that the probability densities functions $p_1, \dots, p_M$ satisfy the following hypotheses:

(H3) The number of samples for each subset are parameterized by a governing parameter n as follows:

$$\mathbf{N}(n) = \{N_1(n), N_2(n), N_3(n), \dots, N_M(n)\} : \mathbb{N} \to \mathbb{N}^M$$

such that for all $m \in \{1, 2, \dots, M\}$

$$\begin{aligned} D_1 \le \frac{N_m}{n} \le D_2 \\ \lim_{n\to\infty} N_m(n) = \infty\,. \end{aligned} \tag{8}$$

Note that $C_1\|\mathbf{N}(n)\| \le N_m(n) \le C_2\|\mathbf{N}(n)\|$.

(H4) For each $m \in \{1, \dots, M\}$, $k_1 = k_2 = \cdots = k_M = k$ for some fixed k in $\mathbb{N}$. For the number of knots for each m are parameterized by n as follows:

$$\mathbf{K}(n) = \{K_1(n), K_2(n), K_3(n), \dots, K_M(n)\} : \mathbb{N} \to \mathbb{N}^M \tag{9}$$

where $K_m(n) + 1$ is the number of knots for B-splines on the interval $[a, b]$ and thus

$$\mathbf{L}(n) = \{L_1(n), L_2(n), L_3(n), \dots, L_M(n)\} : \mathbb{N} \to \mathbb{N}^M \quad \text{with} \quad L_m(n) = K_m(n) - k$$

and we require

$$\lim_{n\to\infty} L_m(n) = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{L_m(n)}{N_m(n)^{1/2-\beta}} = 0,\ 0 < \beta < \frac{1}{2}.$$

(H5) For the knots $T_{K_m(n)} = (t_i^m)_{i=0}^{K_m(n)}$, we write

$$\bar{h}_m = \max_{k-1\le i\le K_m(n)-k} (t_{i+1}^m - t_i^m) \quad \text{and} \quad \underline{h}_m = \min_{k-1\le i\le K_m(n)-k} (t_{i+1}^m - t_i^m). \tag{10}$$

(H6) For each $m \in \{1, \dots, M\}$, $j \in \{0, \dots, k-1\}$ and density $p_m \in C^{j+1}([a, b])$ there exists $C_{m,s} \ge 0$ such that

$$\left| \frac{d^{j+1} \log\left(p_m(\theta)\right)}{d\theta^{j+1}} \right| < C_{m,s} \quad \text{for all} \quad x. \tag{11}$$

(H7) Let $\|\cdot\|_2$ denote the L^2-norm on $[a, b]$. For p^* defined as in H1, there exists $C^* \ge 0$ such that

$$\|p^*\|_2^2 = \int \left(p^*(\theta)\right)^2 d\theta < C^*\,. \tag{12}$$

(H8) For each subset $\mathbf{x}_m$, the B-splines are created by choosing a uniform knot sequence. Thus,

$$\bar{h}_m = \underline{h}_m = h_m,\ \text{for } m \in \{1, \dots, M\}. \tag{13}$$

Let

$$h_{min} = \min_{1\le m\le M} \{h_m\} \quad \text{and} \quad h_{max} = \max_{1\le m\le M} \{h_m\}. \tag{14}$$

We assume that h_{min}, h_{max} scale in a similar way to the number of samples, i.e

$$c_1\|N(n)\|^{-\beta} \le h_{min}^{j+1}(n) \le h_{max}^{j+1}(n) \le c_2\|N(n)\|^{-\beta}$$

where $j \in \{0, \dots, k-1\}$ is the same as in hypothesis (H6).

3 Analysis of MISE for $\hat{p}$

3.1 Error analysis for unnormalized estimator

Suppose we are given a data set $\mathbf{x}$ and it is partitioned into $M \geq 1$ disjoint subsets $\mathbf{x}_m,\ m \in \{1, \dots, M\}$. We are interested in the subset posterior densities $p_m(\theta) = p(\theta|\mathbf{x}_m)$. For each such density we apply the analysis from before. Let $\hat{p}_m$ and $\bar{p}_m,\ m \in \{1, \dots, M\}$ be the corresponding logspline estimators as defined in (70) and (71) respectively. By definition of $\hat{p}_m$, that is equal to the logspline density estimate on $\Omega^m_{n_m} \subset \Omega$, where $\Omega^m_{n_m}$ is the set defined in (69) for $\hat{p}_m$.

Definition 1. *For $m \in \{1, \dots, M\}$, let $\Omega^m_{n_m}$ be the set defined in* (6)*. We then set*

$$\underline{\Omega}^{M,\boldsymbol{N}} := \bigcap_{m=1}^{M} \Omega^m_{n_m} \quad \textit{where} \quad \boldsymbol{N} = (n_1, \dots, n_m)$$

which is the set where the maximizer for the log-likelihood exists given each data subset and thus all logspline density estimators $\hat{p}_m$ exist.

Lemma 2. *Suppose the conditions in* (H3) *and* (H4) *hold. Given the previous definition, we have that*

$$\lim_{n\to\infty} \mathbb{P}\left(\underline{\Omega}^{M,\boldsymbol{N}(n)}\right) = 1$$

Proof. By Theorem 53 we have that

$$\mathbb{P}\left(\Omega \setminus \underline{\Omega}^{M,\mathbf{N}(n)}\right) = \mathbb{P}\left(\bigcup_{m=1}^{M} (\Omega^m_{N_m(n)})^c\right) \leq \sum_{m=1}^{M} \mathbb{P}\left((\Omega^m_{N_m(n)})^c\right) \leq \sum_{m=1}^{M} 2e^{-N_m(n)^{2\epsilon}(L_m(n)+1)\delta_m(D)}$$

and the result follows by taking n to infinity. □

Since the probability of the set where the estimators $\hat{p}_m$ exist for all $m \in \{1, \dots, M\}$ tends to 1, it makes sense to do our analysis for a conditional MISE on the set $\underline{\Omega}^{M,\mathbf{N}(n)}$. Considering the practical aspect, we will never encounter the set where the maximizer of the log-likelihood doesn't exist.
At this point, let's state a bound for $|\hat{p}^*(\theta;\omega) - p^*(\theta)|$ which will be essential in our analysis of MISE.

Lemma 3. *Suppose the hypotheses* (H1)-(H7) *hold and that we are restricted to the sample subspace $\underline{\Omega}^{M,\boldsymbol{N}(n)}$. We then have the following:*

(a) There exists a positive constant $R_1 = R_1(M)$ such that

$$\|\log(\hat{p}^*(\cdot,\omega)) - \log(\bar{p}^*(\cdot))\|_\infty \leq R_1 \sum_{m=1}^{M} \frac{L_m(n)+1}{\sqrt{N_m(n)}}.$$

(b) There exists a positive constant $R_2 = R_2(M, k, j, \mathcal{F}_p, \gamma(T_{K_1(n)}), \dots, \gamma(T_{K_M(n)}))$ such that

$$\|\log(p^*) - \log(\bar{p}^*)\|_\infty \leq R_2\ \bar{h}^{j+1}_{max} \sum_{m=1}^{M} \left\| \frac{d^{j+1}\log(p_m)}{d\theta^{j+1}} \right\|_\infty \quad \textit{where} \quad \bar{h}_{max} = \max_m\{\bar{h}_m\}.$$

(c) Using the bounds from (a) and (b) we have

$$|\hat{p}^*(\theta;\omega) - p^*(\theta)| \leq \left(\exp\left\{R_1 \sum_{m=1}^{M} \frac{L_m(n)+1}{\sqrt{N_m(n)}} + R_2\ \bar{h}^{j+1}_{max} \sum_{m=1}^{M} \left\|\frac{d^{j+1}\log(p_m)}{d\theta^{j+1}}\right\|_\infty\right\} - 1\right) p^*(\theta).$$

Proof. (a) The bound can be shown by writing

$$\begin{aligned}\|\log(\hat{p}^*(\cdot,\omega)) - \log(\bar{p}^*(\cdot))\|_\infty &= \|\log(\prod_{m=1}^{M} \hat{p}_m(\cdot;\omega)) - \log(\prod_{m=1}^{M} \bar{p}_m(\cdot))\|_\infty \\ &\leq \sum_{m=1}^{M} \|\log(\hat{p}_m(\cdot;\omega)) - \log(\bar{p}_m(\cdot))\|_\infty\end{aligned}$$

and then applying Theorem 56. For each $m \in \{1, \dots, M\}$ there will be an M_3^m appearing in the bound and we can take $R_1 = \max_m\{M_3^m\}$.

(b) Similar to part (a) we can write

$$\begin{aligned}\|\log(p^*(\cdot)) - \log(\bar{p}^*(\cdot))\|_\infty &= \|\log(\prod_{m=1}^{M} p_m(\cdot)) - \log(\prod_{m=1}^{M} \bar{p}_m(\cdot))\|_\infty \\ &\leq \sum_{m=1}^{M} \|\log(p_m(\cdot)) - \log(\bar{p}_m(\cdot))\|_\infty\end{aligned}$$

and then we apply Lemma 47. For each $m \in \{1, \dots, M\}$ there will be constants M'_m and $C_m(k,j)$ appearing and we can take $R_2 = \max_m\{M'_m C_m(k,j)\}$.

(c) To see why this is true, we write

$$|\hat{p}^*(\theta;\omega) - p^*(\theta)| = p^*(\theta)\left|\frac{\hat{p}^*(\theta;\omega)}{p^*(\theta)} - 1\right| = p^*(\theta)\left|\exp\{\log(\hat{p}^*(\theta;\omega)) - \log(p^*(\theta))\} - 1\right|.$$

If $\hat{p}^*(\theta;\omega) \geq p^*(\theta)$ then

$$|\exp\{\log(\hat{p}^*(\theta;\omega)) - \log(p^*(\theta))\} - 1| = \exp\{\log(\hat{p}^*(\theta;\omega)) - \log(p^*(\theta))\} - 1.$$

If $\hat{p}^*(\theta;\omega) < p^*(\theta)$ then

$$\begin{aligned}|\exp\{\log(\hat{p}^*(\theta;\omega)) - \log(p^*(\theta))\} - 1| &= 1 - \exp\{\log(\hat{p}^*(\theta;\omega)) - \log(p^*(\theta))\} \\ &\leq \exp\{\log(p^*(\theta)) - \log(\hat{p}^*(\theta;\omega))\} - 1\end{aligned}$$

where the last step is justified by the fact that $1 - e^{-x} \leq e^x - 1,$ for any $x \geq 0$. This implies

$$\begin{aligned}|\hat{p}^*(\theta;\omega) - p^*(\theta)| &\leq p^*(\theta)\left(\exp\{|\log(\hat{p}^*(\theta;\omega)) - \log(p^*(\theta))|\} - 1\right) \\ &\leq p^*(\theta)\left(\exp\{|\log(\hat{p}^*(\theta;\omega)) - \log(\bar{p}^*(\theta))| + |\log(\bar{p}^*(\theta)) - \log(p^*(\theta))|\} - 1\right)\end{aligned}$$

and then we apply the bounds from the previous two parts.

□

This leads us directly to the theorem for the conditional MISE of the unnormalized densities p^* and $\hat{p}^*$.

Theorem 4. *(Conditional* MISE *for unnormalized $p^*, \hat{p}^*$ and $M \geq 1$)*
Assume the conditions (H1)-(H7) *hold. Given $M \geq 1$ we have*

$$\begin{aligned}&\text{MISE}(p^*, \hat{p}^* \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}) \\ &\leq \left(\exp\left\{R_1 \sum_{m=1}^{M} \frac{L_m(n)+1}{\sqrt{N_m(n)}} + R_2\,\bar{h}_{max}^{j+1} \sum_{m=1}^{M} \left\|\frac{d^{j+1}\log(p_m)}{d\theta^{j+1}}\right\|_\infty\right\} - 1\right)^2 \|p^*\|_2^2\end{aligned} \tag{15}$$

where R_1, R_2 are as in Lemma 3.
In addition, if (H8) *holds, then* MISE *scales optimally in regards to the number of samples,*

$$\sqrt{\text{MISE}(p^*, \hat{p}^*)} = O(Mn^{-\beta}) = O(M^{1-\beta}\|\boldsymbol{N}(n)\|^{-\beta}) \tag{16}$$

Proof. By definition of the conditional MISE and Lemma 3, we have

$$\begin{aligned}\text{MISE}(p^*, \hat{p}^* \mid \underline{\Omega}^{M,\mathbf{N}(n)}) &= \mathbb{E}_{\underline{\Omega}^{M,\mathbf{N}(n)}} \int (\hat{p}^*(\theta;\omega) - p^*(\theta))^2\, d\theta \\ &\leq \mathbb{E}_{\underline{\Omega}^{M,\mathbf{N}(n)}} \int \left[\left(\exp\left\{R_1 \sum_{m=1}^{M} \frac{L_m(n)+1}{\sqrt{N_m(n)}} + R_2\,\bar{h}_{max}^{j+1} \sum_{m=1}^{M} \left\|\frac{d^{j+1}\log(p_m)}{d\theta^{j+1}}\right\|_\infty\right\} - 1\right) p^*(\theta)\right]^2 d\theta\end{aligned}$$

which implies (15). Next, if (H8) holds, then (16) follows directly. □

Remark 5. *It's interesting to note how the number of knots, their placement and the number of samples all play a role in the above bound. If we want to be accurate, all of the parameters $L_m(n), N_m(n)$ and $\bar{h}_{max}$ must be chosen appropriately.*

3.2 Analysis for renormalization constant

We will now consider the error that arises for MISE when one renormalizes the product of the estimators so it can be a probability density. The renormalization can affect the error since p^* and $\hat{p}^*$ are rescaled. We define the renormalization constant and its estimator to be

$$\lambda = \int p^*(\theta)\, d\theta \quad \text{and} \quad \hat{\lambda} = \hat{\lambda}(\omega) = \int \hat{p}^*(\theta;\omega)\, d\theta \tag{17}$$

Therefore, we are interested in analyzing

$$\text{MISE}(p,\hat{p}) = \text{MISE}(cp^*, \hat{c}\hat{p}^*), \quad \text{where} \quad c = \lambda^{-1},\ \hat{c} = \hat{\lambda}^{-1}.$$

We first state the following lemma for λ and $\hat{\lambda}(\omega)$.

Lemma 6. *Let λ and $\hat{\lambda}(\omega)$ be defined as in* (6)*. Suppose that* (H8) *holds and we are restricted to the sample subspace $\underline{\Omega}^{M,\boldsymbol{N}(n)}$. Then we have*

$$\left| \frac{\hat{\lambda}(\omega)}{\lambda} - 1 \right| = O(M^{1-\beta} \|\boldsymbol{N}(n)\|^{-\beta}) \tag{18}$$

Proof. By definition of λ and $\hat{\lambda}(\omega)$, we have

$$\begin{aligned} |\lambda - \hat{\lambda}(\omega)| &= \left| \int p^*(\theta)\, d\theta - \int \hat{p}^*(\theta;\omega)\, d\theta \right| \\ &\leq \left(\exp\left\{ R_1 \sum_{m=1}^{M} \frac{L_m(n)+1}{\sqrt{N_m(n)}} + R_2\ \bar{h}_{max}^{j+1} \sum_{m=1}^{M} \left\| \frac{d^{j+1}\log(p_m)}{d\theta^{j+1}} \right\|_\infty \right\} - 1 \right) \lambda \end{aligned}$$

where the inequality is justified by Lemma 3(c). Dividing by λ the result then follows by hypothesis (H8). ☐

So what the above lemma suggests is that when restricted to the sample subspace $\underline{\Omega}^{M,\mathbf{N}(n)}$, the space where the logspline density estimators $\hat{p}_m$, $m \in \{1,\dots,M\}$ are all defined, the renormalization constant $\hat{c}$ of the product of the estimators approximates the true renormalization constant c.

Knowing now how $\hat{\lambda}(\omega)$ scales we can start analyzing MISE$(p,\hat{p})$ on the sample subspace. However, to make the analysis slightly easier we introduce a new functional, called $\overline{\text{MISE}}$. This new functional is asymptotically equivalent to MISE as we will show, thus providing us with the means to view how MISE scales without having to directly analyze it.

Definition 7. *Suppose $M \geq 1$ and hypotheses* (H1)-(H2) *hold. Given the sample subspace $\underline{\Omega}^{M,\boldsymbol{N}(n)}$ we define the functional*

$$\overline{\text{MISE}}\left(p, \hat{p} \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}\right) = \mathbb{E}_{\underline{\Omega}^{M,\boldsymbol{N}(n)}} \left[\left(\frac{\hat{\lambda}(\omega)}{\lambda} \right)^2 \int (\hat{p}(\theta;\omega) - p(\theta))^2\, d\theta \right] \tag{19}$$

Proposition 8. *The functional $\overline{\text{MISE}}$ is asymptotically equivalent to* MISE *on $\underline{\Omega}^{M,\boldsymbol{N}(n)}$, in the sense that*

$$\lim_{\|\boldsymbol{N}(n)\|\to\infty} \frac{\overline{\text{MISE}}\left(p,\hat{p} \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}\right)}{\text{MISE}\left(p,\hat{p} \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}\right)} = 1 \tag{20}$$

Proof. Notice that $\overline{\text{MISE}}$ can be written as

$$\begin{aligned} \overline{\text{MISE}}\left(p,\hat{p} \mid \underline{\Omega}^{M,\mathbf{N}(n)}\right) &= \mathbb{E}_{\underline{\Omega}^{M,\mathbf{N}(n)}} \left[\left(\frac{\hat{\lambda}}{\lambda} - 1 + 1 \right)^2 \int (\hat{p}(\theta;\omega) - p(\theta))^2\, d\theta \right] \\ &= \mathbb{E}_{\underline{\Omega}^{M,\mathbf{N}(n)}} \left[\left[\left(\frac{\hat{\lambda}}{\lambda} - 1 \right)^2 + 2\left(\frac{\hat{\lambda}}{\lambda} - 1 \right) + 1 \right] \int (\hat{p}(\theta;\omega) - p(\theta))^2\, d\theta \right] \end{aligned}$$

and thus by Lemma 6

$$\overline{\text{MISE}}\left(p, \hat{p} \mid \underline{\Omega}^{M,\mathbf{N}(n)}\right) = (1 + \mathcal{E}(n))\text{MISE}\left(p, \hat{p} \mid \underline{\Omega}^{M,\mathbf{N}(n)}\right)$$

$$\text{where} \quad \mathcal{E}(n) = O(M^{1-\beta}\|\mathbf{N}(n)\|^{-\beta})$$

which then implies the result. □

We conclude our analysis with the next theorem, which states how MISE scales for the renormalized estimators.

Theorem 9. *Let $M \geq 1$. Assume the conditions* (H1)-(H8) *hold. Then*

$$\text{MISE}\left(p, \hat{p} \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}\right) = O(M^{2-2\beta}\|\boldsymbol{N}(n)\|^{-2\beta}). \tag{21}$$

Proof. We will do the work for $\overline{\text{MISE}}$ and the result will follow from Proposition 8. Notice that $\overline{\text{MISE}}$ can be written as below. Also, let $\mathbb{E}_n(\cdot) = \mathbb{E}(\cdot|\underline{\Omega}^{M,\mathbf{N}(n)})$

$$\begin{aligned}
\overline{\text{MISE}}\left(p, \hat{p} \mid \underline{\Omega}^{M,\mathbf{N}(n)}\right) &= \mathbb{E}_n\left[\left(\frac{\hat{\lambda}}{\lambda}\right)^2 \int (p - \hat{p})^2 \, d\theta\right] \\
&= \|p\|_2^2 \; \mathbb{E}_n\left[\left(\frac{\hat{\lambda}}{\lambda} - 1\right)^2\right] + \lambda^{-2} \; \text{MISE}_n(p^*, \hat{p}^*) \\
&\quad - 2\lambda^{-1} \; \mathbb{E}_n \int \left(\frac{\hat{\lambda}}{\lambda} - 1\right)(\hat{p}^* - p^*)p \, d\theta \\
&= J_1 + J_2 + J_3
\end{aligned}$$

We now determine how each of the J_i, $i \in \{1, 2, 3\}$ scale. For J_1 by Lemma 6 we have

$$J_1 = O(M^{2-2\beta}\|\mathbf{N}(n)\|^{-2\beta}),$$

for J_2 we have from (H8)

$$J_2 = O(M^{2-2\beta}\|\mathbf{N}(n)\|^{-2\beta})$$

and for J_3 we have from Lemmas 3(c) and 6

$$\begin{aligned}
|J_3|^2 &\leq 4\lambda^{-2} \; \left(\mathbb{E}_n \int \left|\frac{\hat{\lambda}}{\lambda} - 1\right| |\hat{p}^* - p^*| p \, d\theta\right)^2 \\
&\leq 4\lambda^{-2} \; \mathbb{E}_n\left[\left(\frac{\hat{\lambda}}{\lambda} - 1\right)^2 \int p^2 \, d\theta\right] \cdot \text{MISE}_n(p^*, \hat{p}^*).
\end{aligned}$$

Thus, by hypotheses (H7)-(H8)

$$|J_3| = O(M^{2-2\beta}\|\mathbf{N}(n)\|^{-2\beta}).$$

□

4 Numerical Error

So far we have estimated the error that arises between the unknown density p and the full-data estimator $\hat{p}$. However, in practice it is difficult to evaluate the renormalization constant

$$\hat{\lambda}(\omega) = \int \hat{p}^*(\theta) \, d\theta = \int \prod_{m=1}^{M} \hat{p}_m(\theta) \, d\theta$$

defined in (17). The difficulty is due to the process of generating MCMC samples and thus $\hat{p}^*$ is not explicitly known. In order to circumvent this issue, our idea is to approximate the integral above numerically. To accomplish this, we interpolate $\hat{p}^*$ using Lagrange polynomials. This procedure leads to the construction of an interpolant estimator $\tilde{p}^*$ which we then integrate numerically. We then normalize $\tilde{p}^*$ and use that as a density estimator for p. Unfortunately, to estimate the error by considering that kind of approximation given an arbitrary grid of points for Lagrange polynomials, independent of the set of knots (t_i) for B-splines gives a stringent condition on the smoothness of B-splines we incorporate. It turns out that we have to utilize B-splines of order at least $k = 4$. For this reason we consider using Lagrange polynomials of order $l+1$ which satisfy $l < k-2$.

4.1 Interpolation of an estimator: preliminaries

We remind the reader the model we deal with throughout our work. We recall that the (marginal) posterior of the parameter $\theta \in \mathbb{R}$ (which is a component of a multidimensional parameter $\boldsymbol{\theta} \in \mathbb{R}^d$) given the data

$$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$$

partitioned into M disjoint sets $\mathbf{x}_m$, $m = 1, \ldots, M$ is assumed to have the form

$$p(\theta|\mathbf{x}) \propto \prod_{m=1}^{M} p_m(\theta) \tag{22}$$

with $p(\theta|\mathbf{x}_m)$ denoting the (marginal) posterior density of θ given data $\mathbf{x}_m$.

The estimator $\hat{p}(\theta|\mathbf{x})$ of the posterior $p(\theta|\mathbf{x})$ is taken to be

$$\hat{p}(\theta|\mathbf{x}) \propto \prod_{m=1}^{M} \hat{p}_m(\theta) \tag{23}$$

where $\hat{p}_m(\theta)$ stands for the logspline density estimator of the sub-posterior density $p_m(\theta)$. Recall from Definition 37 and hypotheses (H1)-(H5) that for each $m \in \{1, \ldots, M\}$, the estimator $\hat{p}_m$ has the form

$$\hat{p}_m(\theta) = \exp\left(B_m(\theta; \hat{y}^m) - c(\hat{y}^m)\right) \tag{24}$$

where

$$B_m(\theta; \hat{y}^m) = \sum_{j=0}^{L_m(n)} \hat{y}_j^m B_{j,k,T_{K_m(n)}}(\theta)$$

$$\text{and} \quad c(\hat{y}^m) = \log\left(\int \exp\left(B_m(\theta; \hat{y}^m)\, d\theta\right)\right)$$

The vector $\hat{y}^m = (\hat{y}_1^m, \ldots, \hat{y}_{L_m(n)}^m)$ is the argument that maximizes the log-likelihood, as described in equation (65) and we also remind the reader that this maximizer exists for all $m \in \{1, \ldots, M\}$ as we carry out our analysis on the sample subspace $\underline{\Omega}^{M,\mathbf{N}(n)}$.

Together with the hypotheses stated in section 3, we now add the next proposition which will be necessary for our work later on.

Proposition 10. *Suppose hypotheses* (H1)-(H8) *hold. Given the space $\underline{\Omega}^{M,\boldsymbol{N}(n)}$, we have that the estimator $\hat{p}_m$ is bounded and its derivatives of all orders satisfy*

$$\left|\hat{p}_m^{(\alpha)}(\theta)\right| \leq C(\alpha, k, p_m) \|\boldsymbol{N}(n)\|^{\alpha\beta/(j+1)} \quad \text{for} \quad \theta \in (a,b) \text{ and } \alpha < k-1$$

where the constant $C(\alpha, k, p_m)$ depends on the order k of the B-splines, the order α of the derivative and the density p_m.

Proof. Observe that the estimator $\hat{p}_m$ can be expressed as

$$\hat{p}_m(\theta) = \exp\Big[\sum_{j=0}^{L_m(n)} \hat{y}_j^m B_{j,k}(\theta) - c(\hat{y}^m)\Big] = \exp\Big[\sum_{j=0}^{L_m(n)} (\hat{y}_j^m - c(\hat{y}^m)) B_{j,k}(\theta)\Big]$$

Then, applying Faa di Bruno's formula, we obtain

$$|\hat{p}_m^{(\alpha)}(\theta)| \leq \hat{p}_m(\theta) \sum_{k_1+2k_2+\cdots+\alpha k_\alpha=\alpha} \frac{\alpha!}{k_1!k_2!\dots k_\alpha!} \prod_{i=1}^{\alpha} \left(\frac{\left| \frac{d^i}{d\theta^i} \sum_{j=0}^{L_m(n)} (\hat{y}_j^m - c(\hat{y}^m)) B_{j,k}(\theta) \right|}{i!} \right)^{k_i}, \text{ for } \theta \in [t_i, t_{i+1}].$$

where $k_1, \dots, k_\alpha$ are nonnegative integers and if $k_i > 0$ with $i \geq k$ then that term in the sum above will be zero since almost everywhere $B_{j,k}^{(i)}(\theta) = 0$. By De Boor's formula [1, p.132], we can estimate the derivative of a spline as follows

$$\left| \frac{d^i}{d\theta^i} \sum_{j=0}^{L_m(n)} (\hat{y}_j^m - c(\hat{y}^m)) B_{j,k}(\theta) \right| = \left| \frac{d^i}{d\theta^i} \log \hat{p}_m(\theta) \right| \leq C \frac{\| \log \hat{p} \|_\infty}{\underline{h}_m^i}.$$

where the constant C depends only on the order k of the B-splines. Therefore, we can bound $|\hat{p}_m^{(\alpha)}(\theta)|$ as follows

$$\begin{aligned} |\hat{p}_m^{(\alpha)}(\theta)| &\leq \hat{p}_m(\theta) \sum_{k_1+2k_2+\cdots+\alpha k_\alpha=\alpha} \frac{\alpha!}{k_1!k_2!\dots k_\alpha!} \prod_{i=1}^{\alpha} \left(C \frac{\| \log \hat{p}_m \|_\infty}{i!\, \underline{h}_m^i} \right)^{k_i} \\ &\leq \hat{p}_m(\theta) \left(\frac{1 + C^\alpha \| \log \hat{p}_m \|_\infty^\alpha}{\underline{h}_m^\alpha} \right) \sum_{k_1+2k_2+\cdots+\alpha k_\alpha=\alpha} \frac{\alpha!}{k_1!k_2!\dots k_\alpha!}. \end{aligned}$$

The above leads to the following bound:

$$\begin{aligned} \left| \hat{p}_m^{(\alpha)}(\theta) \right| &\leq \hat{p}_m(\theta) \frac{1 + C^\alpha \| \log \hat{p}_m \|_\infty^\alpha}{\underline{h}_m^\alpha} \sum_{\zeta=1}^{\alpha} \frac{\alpha!}{\zeta!} (\alpha - \zeta + 1)^\zeta \\ &\leq C(k, \alpha)\, \hat{p}_m(\theta) \frac{1 + \| \log \hat{p}_m \|_\infty^\alpha}{\underline{h}_m^\alpha} \end{aligned}$$

where $C(k, \alpha)$ is a constant that depends on the order k and the α. Next, recalling the hypotheses (H3), (H4),(H6) and (H8), we obtain

$$\hat{p}_m(\theta) \leq |\hat{p}_m(\theta) - p_m(\theta)| + p_m(\theta) \leq \|p_m\|_\infty (1 + c\|\mathbf{N}(n)\|^{-\beta})$$

and

$$\begin{aligned} \| \log \hat{p}_m \|_\infty &\leq \| \log \hat{p}_m - \log \bar{p}_m \|_\infty + \| \log \bar{p}_m - \log p_m \|_\infty + \| \log p_m \|_\infty \\ &\leq c\|\mathbf{N}(n)\|^{-\beta} + \| \log p_m \|_\infty \end{aligned}$$

where we also used Lemma 47, Theorem 56, Lemma 3. Therefore,

$$\begin{aligned} \left| \hat{p}_m^{(\alpha)}(\theta) \right| &\leq C(k, \alpha)\, \|p_m\|_\infty (1 + \|\mathbf{N}(n)\|^{-\beta}) \frac{1 + \|\mathbf{N}(n)\|^{-\alpha\beta} + \| \log p_m \|_\infty^\alpha}{\underline{h}_m^\alpha} \\ &\leq C(\alpha, k, p_m) \frac{1}{\underline{h}_m^\alpha} \\ &= C(\alpha, k, p_m) (\underline{h}_m^{j+1})^{-\alpha/(j+1)} \;\sim\; C(\alpha, k, p_m) \|\mathbf{N}(n)\|^{\alpha\beta/(j+1)} \end{aligned}$$

The final result follows immediately and since the index i was chosen arbitrarily and that all interior knots are simple, this concludes the proof. □

Remark 11. *Remark 30, in Section 6, allowed us to extend the bound for all $\theta \in (a, b)$ in the proof above. In reality, we can also extend the bound to the closed interval $[a, b]$. Since $a = t_0$ and $b = t_{K_m(n)}$ are knots with multiplicity k, any B-spline that isn't continuous at those knots will just be a polynomial that has been cut off, which means there is no blow-up. Thus, we can extend the bound by considering right-hand and left-hand limits of derivatives at a and b, respectively. From this point on we consider the bound in Proposition 10 holds for all $\theta \in [a, b]$.*

Lemma 12. *Assume hypotheses* (H1)-(H8) *hold. Suppose that for each* $m = 1, \dots, M$ *the sub-posterior estimator* $\hat{p}_m(\theta)$ *is* α*-times differentiable on* $[a, b]$ *for some positive integer* $\alpha < k - 1$.
Then, the estimator $\hat{p}^*$ *satisfies*

$$\Big|\frac{d^\alpha}{d\theta^\alpha}\hat{p}^*(\theta)\Big| = \big|(\hat{p}_1 \ldots \hat{p}_M)^{(\alpha)}(\theta)\big| \le C(\alpha, k, p_1, \dots, p_M)\|\boldsymbol{N}(n)\|^{\alpha\beta/(j+1)}M^\alpha \quad \textit{for} \quad \theta \in [a, b], \tag{25}$$

where $C(\alpha, k, p_1, \dots, p_M)$ *depends on the order* k *of the B-splines, the order* α *of the derivative and the densities* $p_1, \dots, p_M$.

Proof. Let $\theta \in [a, b]$. By Proposition (10) we have

$$|\hat{p}_m^{(\alpha)}(\theta)| \le C(\alpha, k, p_m)\|\mathbf{N}(n)\|^{\alpha\beta/(j+1)}.$$

Then, using the general Leibnitz rule and employing the above inequality we obtain

$$\begin{aligned}
\Big|\frac{d^\alpha}{d\theta^\alpha}\hat{p}^*(\theta)\Big| &= \big|(\hat{p}_1 \ldots \hat{p}_M)^{(\alpha)}(\theta)\big| = \\
&= \Big|\sum_{i_1+\dots+i_M=\alpha}\frac{\alpha!}{i_1!\dots i_M!}\hat{p}_1^{(i_1)} \ldots \hat{p}_M^{(i_M)}\Big| \\
&\le \sum_{i_1+\ldots+i_M=\alpha}\frac{\alpha!}{i_1! \ldots i_M!}C(i_1, k, p_1)\|\mathbf{N}(n)\|^{i_1\beta/(j+1)} \ldots C(i_M, k, p_M)\|\mathbf{N}(n)\|^{i_M\beta/(j+1)} \\
&= \|\mathbf{N}(n)\|^{\alpha\beta/(j+1)}\sum_{i_1+\ldots+i_M=\alpha}\frac{\alpha!}{i_1! \ldots i_M!}C(i_1, k, p_1) \ldots C(i_M, k, p_M)
\end{aligned}$$

From the proof of Proposition 10, notice that $C(i, k, p_m) \le C(j, k, p_m)$ for positive integers $i \le j$. Therefore, we have

$$|\hat{p}_m^{(\alpha)}(\theta)| \le C(\alpha, k, p_1, \dots, p_M)\|\mathbf{N}(n)\|^{\alpha\beta/(j+1)}\sum_{i_1+\ldots+i_M=\alpha}\frac{\alpha!}{i_1! \ldots i_M!}$$

where $C(\alpha, k, p_1, \dots, p_M) = C(\alpha, k, p_1) \dots C(\alpha, k, p_M)$ and the result follows from the multinomial theorem. This concludes the proof. ☐

4.2 Numerical approximation of the renormalization constant $\hat{c} = \hat{\lambda}^{-1}$

By Remark 30, in Section 6, we have that B-splines of order k, and therefore any splines that arise from these, will have $k-2$ continuous derivatives on (a, b). Thus, in order to utilize Lemma 59, we must have that the order of the Lagrange polynomials be at most $k - 2$, i.e. $l \le k - 3$. Since $l \ge 1$ this implies that the B-splines used in the construction of the logspline estimators be at least cubic. Thus, assume $k \ge 4$ and let $1 \le l \le k - 3$ be a positive integer that denotes the degree of the interpolating polynomials. Let $N \in \mathbb{N}$ be the number of sub-intervals of $[a, b]$ on each of which we will interpolate the product of estimators by the polynomial of degree l. Thus each sub-interval has to be further subdivided into l intervals. Define the partition $\mathcal{X}$ of $[a, b]$ such that

$$\mathcal{X} = \{a = x_0 < x_1 < x_2 < \dots < x_{Nl} = b\} \quad \text{and} \quad x_{i+1} - x_i = \frac{b-a}{Nl} = \Delta x. \tag{26}$$

For each $i = 0, \dots, N-1$, recalling the formula (87), we define the (random) Lagrange polynomial

$$\hat{q}_i(\theta) := \sum_{\tau=0}^{l}\hat{p}^*(x_{il+\tau})l_{\tau,i}(\theta) \quad \text{with} \qquad l_{\tau,i}(\theta) := \prod_{j\in\{0,\dots,l\}\setminus\{\tau\}}\left(\frac{\theta - x_{il+j}}{x_{il+\tau} - x_{il+j}}\right), \tag{27}$$

which is a polynomial that interpolates the estimator $\hat{p}^*(\theta)$ on the interval $[x_{il}, x_{(i+1)l}]$. We next define an interpolant estimator $\tilde{p}^*$ to be a *random* composite polynomial given by

$$\tilde{p}^*(\theta) := \begin{cases} 0, & \theta \in \mathbb{R}\backslash[a, b] \\ \hat{q}_i(\theta), & \theta \in [x_{il}, x_{(i+1)l}] \end{cases} \tag{28}$$

which approximates the estimator $\hat{p}^*$ on the whole interval $[a,b]$.
We are now ready to estimate the mean integrated squared error given by

$$\text{MISE}\big(p^*, \tilde{p}^* \mid \underline{\Omega}^{M,\mathbf{N}(n)}\big) = \mathbb{E}\int \big(p^*(\theta) - \tilde{p}^*(\theta)\big)^2 d\theta \tag{29}$$

Lemma 13. *Assume that hypotheses* (H1)-(H8) *hold and $\tilde{p}^*$ is the estimator of $\hat{p}^*$ as defined in* (28) *given the partition $\mathcal{X}$ from* (26) *respectively. The following estimate holds provided $1 \le l \le k-3$.*

$$\begin{aligned}\text{MISE}(\hat{p}^*, \tilde{p}^* \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}) &= \mathbb{E}\int_a^b \big(\hat{p}^*(\theta) - \tilde{p}^*(\theta)\big)^2 d\theta \\ &\le \left(\frac{(\Delta x)^{l+1}}{4(l+1)}\|\boldsymbol{N}(n)\|^{(l+1)\beta/(j+1)} M^{l+1}\right)^2 \mathcal{C}(l+1,k,p_1,\dots,p_M,(a,b))\end{aligned} \tag{30}$$

where the constant $\mathcal{C}(l+1,k,p_1,\dots,p_M,(a,b))$ depends on the order $l+1$ of the Lagrange polynomials, the order k of the B-splines, the densities $p_1,\dots,p_M$ and the length of the interval (a,b).

Proof. Let $i \in \{0,\dots,N-1\}$. By Lemma 59, Lemma 12, and (28) for any $\theta \in [x_{il}, x_{(i+1)l}]$ we have

$$\begin{aligned}\big|\hat{p}^*(\theta) - \tilde{p}^*(\theta)\big| &= \big|\hat{p}^*(\theta) - \hat{q}_i(\theta)\big| \\ &\le \left(\sup_{\theta\in[x_{il},x_{(i+1)l}]}\Big|\frac{d}{d\theta}^{(l+1)}\hat{p}^*(\theta)\Big|\right)\frac{(\Delta x)^{l+1}}{4(l+1)} \\ &\le \frac{(\Delta x)^{l+1}}{4(l+1)} C(l+1,k,p_1,\dots,p_M)\|\mathbf{N}(n)\|^{(l+1)\beta/(j+1)} M^{l+1}.\end{aligned} \tag{31}$$

Thus we conclude that

$$\begin{aligned}\mathbb{E}\int_a^b \big(\hat{p}^*(\theta) - \tilde{p}^*(\theta)\big)^2 d\theta &= \sum_{i=0}^{N-1}\mathbb{E}\int_{x_{il}}^{x_{(i+1)l}} \big(\hat{p}^*(\theta) - \hat{q}_i(\theta)\big)^2 d\theta \\ &\le \left(\frac{(\Delta x)^{l+1}}{4(l+1)}\|\mathbf{N}(n)\|^{(l+1)\beta/(j+1)} M^{l+1}\right)^2 \mathcal{C}(l+1,k,p_1,\dots,p_M,(a,b)).\end{aligned}$$

where $\mathcal{C}(l+1,k,p_1,\dots,p_M,(a,b)) = C^2(l+1,k,p_1,\dots,p_M)(b-a)$. □

Now that we have bounded the error between $\hat{p}^*$ and $\tilde{p}^*$, we define the renormalization constant $\tilde{c}$ and the density estimator $\tilde{p}$ of $\hat{p}$.

$$\frac{1}{\tilde{c}} = \tilde{\lambda} = \int_a^b \tilde{p}^*(\theta)\, d\theta \quad \text{and} \quad \tilde{p} := \tilde{c}\tilde{p}^* \tag{32}$$

Now the question is, how close is $\tilde{\lambda}$ to $\hat{\lambda}$. This is answered in the following lemma.

Lemma 14. *Given the definitions of $\hat{\lambda}$ and $\tilde{\lambda}$ in* (17) *and* (32) *respectively, we have that the distance between the two renormalization constants is bounded by*

$$|\hat{\lambda} - \tilde{\lambda}| \le \left(\frac{(\Delta x)^{l+1}}{4(l+1)}\|\boldsymbol{N}(n)\|^{(l+1)\beta/(j+1)} M^{l+1}\right)\mathcal{R}(l+1,k,p_1,\dots,p_M,(a,b)) \tag{33}$$

where the constant $\mathcal{R}(l+1,k,p_1,\dots,p_M,(a,b)) = C(l+1,k,p_1,\dots,p_M)(b-a)$.

Proof. We write

$$|\hat{\lambda} - \tilde{\lambda}| \le \int_a^b |\hat{p}^*(\theta) - \tilde{p}^*(\theta)|\, d\theta$$

and then we just apply the Lagrange interpolation error from Lemma 59. □

We will continue by following the same steps as in subsection 3.2. The idea is to introduce a functional that will scale the same as $\text{MISE}(\hat{p}, \tilde{p} \mid \underline{\Omega}^{M,\mathbf{N}(n)})$.

Definition 15. *Suppose $M \geq 1$ and hypotheses* (H1),(H2) *and* (H8) *hold. Given the sample subspace $\underline{\Omega}^{M,\boldsymbol{N}(n)}$ we define the functional*

$$\underline{\text{MISE}}\left(\hat{p}, \tilde{p} \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}\right) = \mathbb{E}_{\underline{\Omega}^{M,\boldsymbol{N}(n)}}\left[\left(\frac{\tilde{\lambda}}{\hat{\lambda}(\omega)}\right)^2 \int (\hat{p}(\theta;\omega) - \tilde{p}(\theta))^2\, d\theta\right] \tag{34}$$

Proposition 16. *The functional $\underline{\text{MISE}}$ is asymptotically equivalent to* MISE *on $\underline{\Omega}^{M,\boldsymbol{N}(n)}$, in the sense that*

$$\lim_{\Delta x \to 0} \frac{\underline{\text{MISE}}\left(\hat{p}, \tilde{p} \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}\right)}{\text{MISE}\left(\hat{p}, \tilde{p} \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}\right)} = 1 \tag{35}$$

Proof. Notice that $\underline{\text{MISE}}$ can be written as

$$\begin{aligned}
\underline{\text{MISE}}\left(\hat{p}, \tilde{p} \mid \underline{\Omega}^{M,\mathbf{N}(n)}\right) &= \mathbb{E}_{\underline{\Omega}^{M,\mathbf{N}(n)}}\left[\left(\frac{\tilde{\lambda}}{\hat{\lambda}} - 1 + 1\right)^2 \int (\hat{p}(\theta;\omega) - \tilde{p}(\theta))^2\, d\theta\right] \\
&= \mathbb{E}_{\underline{\Omega}^{M,\mathbf{N}(n)}}\left[\left(\lambda^{-2}\left(\frac{\lambda}{\hat{\lambda}}\right)^2 \left(\tilde{\lambda} - \hat{\lambda}\right)^2 + 2\lambda^{-1}\frac{\lambda}{\hat{\lambda}}\left(\tilde{\lambda} - \hat{\lambda}\right) + 1\right) \int (\hat{p}(\theta;\omega) - \tilde{p}(\theta))^2\, d\theta\right].
\end{aligned}$$

Thus, by Lemmas 6 and 14, where the former implies

$$\frac{\lambda}{\hat{\lambda}} \leq \frac{1}{1 - C\, M^{1-\beta} \|\mathbf{N}(n)\|^{-\beta}}\,,$$

and for large enough n for which $1 - C\, M^{1-\beta}\|\mathbf{N}(n)\|^{-\beta} > 0$, we have

$$\underline{\text{MISE}}\left(\hat{p}, \tilde{p} \mid \underline{\Omega}^{M,\mathbf{N}(n)}\right) = (1 + \mathcal{E}(n))\text{MISE}\left(\hat{p}, \tilde{p} \mid \underline{\Omega}^{M,\mathbf{N}(n)}\right)$$

with $\mathcal{E}(n) = O(M^{l+1}(\Delta x)^{l+1})$. This then implies the result. ◻

Theorem 17. *Let $M \geq 1$. Assume the conditions* (H1)-(H8) *hold. Then*

$$\text{MISE}\left(\hat{p}, \tilde{p} \mid \underline{\Omega}^{M,\boldsymbol{N}(n)}\right) = O\left[\left(\|\boldsymbol{N}(n)\|^{\beta/(j+1)}(\Delta x) M\right)^{2(l+1)}\right]. \tag{36}$$

Proof. We will do the work for $\underline{\text{MISE}}$ and the result will follow from Proposition 16. Notice that $\underline{\text{MISE}}$ can be written as below. Also, let $\mathbb{E}_n(\cdot) = \mathbb{E}(\cdot|\underline{\Omega}^{M,\mathbf{N}(n)})$

$$\begin{aligned}
\underline{\text{MISE}}\left(\hat{p}, \tilde{p} \mid \underline{\Omega}^{M,\mathbf{N}(n)}\right) &= \mathbb{E}_n\left[\left(\frac{\tilde{\lambda}}{\hat{\lambda}}\right)^2 \int (\hat{p}(\theta;\omega) - \tilde{p}(\theta))^2\, d\theta\right] \\
&= \mathbb{E}_n \int \left(\frac{\tilde{\lambda}}{\hat{\lambda}}\hat{p} - \frac{1}{\hat{\lambda}}\tilde{p}^* - \hat{p} + \hat{p}\right)^2 d\theta \\
&\leq \frac{\lambda^{-1}}{1 - C\, M^{1-\beta}\|\mathbf{N}(n)\|^{-\beta}} \mathbb{E}_n \int \left((\tilde{\lambda} - \hat{\lambda})(\hat{p} - p) + (\tilde{\lambda} - \hat{\lambda})p + (\hat{p}^* - \tilde{p}^*)\right)^2 d\theta \\
&\leq \frac{\lambda^{-1}}{1 - C\, M^{1-\beta}\|\mathbf{N}(n)\|^{-\beta}} (J_1 + J_2 + J_3 + J_4 + J_5 + J_6)
\end{aligned}$$

where

$$\begin{aligned}
J_1 &= \mathbb{E}_n \int (\tilde{\lambda} - \hat{\lambda})^2 (\hat{p} - p)^2\, d\theta, & J_2 &= \mathbb{E}_n \int (\tilde{\lambda} - \hat{\lambda})^2 p^2\, d\theta, \\
J_3 &= \mathbb{E}_n \int (\hat{p}^* - \tilde{p}^*)^2\, d\theta, & J_4 &= 2\,\mathbb{E}_n \int (\tilde{\lambda} - \hat{\lambda})^2 (\hat{p} - p) p\, d\theta, \\
J_5 &= 2\,\mathbb{E}_n \int (\tilde{\lambda} - \hat{\lambda})(\hat{p} - p)(\hat{p}^* - \tilde{p}^*)\, d\theta, & J_6 &= 2\,\mathbb{E}_n \int (\tilde{\lambda} - \hat{\lambda})(\hat{p}^* - \tilde{p}^*) p\, d\theta.
\end{aligned}$$

and by hypotheses (H1)-(H8) and Lemmas 9, 13 and 14, we obtain

$$\begin{aligned}|J_1| &\le C_1\Big(\|\mathbf{N}(n)\|^{\beta/(j+1)}(\Delta x)M\Big)^{2(l+1)}\cdot M^{2-2\beta}\|\mathbf{N}(n)\|^{-2\beta}\\ |J_2|+|J_3|+|J_6| &\le C_2\Big(\|\mathbf{N}(n)\|^{\beta/(j+1)}(\Delta x)M\Big)^{2(l+1)}\\ |J_4|+|J_5| &\le C_3\Big(\|\mathbf{N}(n)\|^{\beta/(j+1)}(\Delta x)M\Big)^{2(l+1)}\cdot M^{1-\beta}\|\mathbf{N}(n)\|^{-\beta}\end{aligned}$$

which for large n implies the result. □

Theorem 18. *Assume that hypotheses* (H1)-(H8) *hold. Let $\tilde{p}$ be the polynomial that interpolates $\hat{p}$ as defined in* (28)*, given the partition $\mathcal{X}$. We then have the estimate*

$$\begin{aligned}\mathrm{MISE}(p\,,\tilde{p}\mid\underline{\Omega}^{M,\boldsymbol{N}(n)}) &= \mathbb{E}\int_a^b\big(p(\theta)-\tilde{p}(\theta)\big)^2d\theta\\ &\le \mathcal{C}\left[M^{2-2\beta}\|\boldsymbol{N}(n)\|^{-2\beta}+\left((\Delta x)\|\boldsymbol{N}(n)\|^{\beta/(j+1)}M\right)^{2(l+1)}\right]\end{aligned} \tag{37}$$

where the constant $\mathcal{C}$ depends on the order k of the B-splines, the degree l of the interpolating polynomial, the densities $p_1,\dots,p_M$ and the length of the interval (a,b). Furthermore, assuming that Δx is a function of the vector of samples $\boldsymbol{N}(n)$, then MISE *scales optimally with respect to $\boldsymbol{N}(n)$ such that*

$$\mathrm{MISE}(p\,,\tilde{p}\mid\underline{\Omega}^{M,\boldsymbol{N}(n)}) \quad\le C\|\boldsymbol{N}(n)\|^{-2\beta}\quad \textit{when}\quad \Delta x = O\bigg(\|\boldsymbol{N}(n)\|^{-\beta\left(\frac{1}{l+1}+\frac{1}{j+1}\right)}\bigg). \tag{38}$$

Proof. Observe that

$$\begin{aligned}\mathrm{MISE}(p\,,\tilde{p}\mid\underline{\Omega}^{M,\mathbf{N}(n)}) &\le \mathbb{E}\int_a^b\big(p(\theta)-\hat{p}(\theta)\big)^2d\theta+\mathbb{E}\int_a^b\big(\hat{p}(\theta)-\tilde{p}(\theta)\big)^2d\theta\\ &=: I_1+I_2.\end{aligned}$$

(37) then follows from Theorem 9 and Theorem 17. Using that estimate we can ask the following question. Suppose that we chose Δx to be a function of the number of samples so that

$$c_1\|\mathbf{N}(n)\|^{-\alpha}\le \Delta x(n)\le c_2\|\mathbf{N}(n)\|^{-\alpha} \tag{39}$$

for some constants c_1, c_2 and α. Clearly, one would not like Δx to be excessively small in order to avoid difficulties that appear with round-off error when computing. On the other hand one would like the error to converge to zero as fast as possible. Thus let us find the smallest rate α for which the asymptotic rate achieves its maximum. To this end we define the function

$$R(\alpha) := -\lim_{\|\mathbf{N}(n)\|\to\infty}\log_{\|\mathbf{N}(n)\|}\mathrm{MISE}(p,\tilde{p}\mid\underline{\Omega}^{M,\mathbf{N}(n)})$$

that describes the asymptotic rate of convergence of the mean integrated squared error. By (37) we have

$$R(\alpha)=\begin{cases}2\beta, & \alpha\ge\beta\left(\dfrac{1}{l+1}+\dfrac{1}{j+1}\right)\\ \left(\alpha-\dfrac{\beta}{j+1}\right)2(l+1), & \alpha<\beta\left(\dfrac{1}{l+1}+\dfrac{1}{j+1}\right)\end{cases}$$

It is obvious that the smallest rate for which the function $R(\alpha)$ achieves its maximum value of 2β is given by $\alpha=\beta\left(\frac{1}{l+1}+\frac{1}{j+1}\right)$. This concludes the proof. □

5 Numerical Experiments

5.1 Numerical experiment with normal subset posterior densities

5.1.1 Description of experiment

This numerical experiment, as well as the following, is designed to investigate the relationship between the approximated value of $\mathrm{MISE}(p\,,\tilde{p}\mid \underline{\Omega}^{M,\mathbf{N}(n)})$ and the bound given by (38). One iteration of the experiment generates $M=3$ subsets of a predetermined number of MCMC samples with $\hat{p}_m \sim \mathcal{N}(2,1),\ m=1,2,3$. Then for each iteration the Lagrange polynomial $\tilde{p}$ is computed a hundred times by re-sampling in order to obtain an approximation to MISE and its standard deviation. For this specific example, we perform ten iterations starting with $20,000$ samples and increasing that number by $10,000$ for each experiment. In the experiments we ran, we chose the parameters so that the optimal rate of convergence for MISE was obtained. Thus, $\beta=1/2$ was chosen. The logspline density estimation that was implemented utilized cubic B-splines (thus, order $k=4$), which implies $l=1$ in (38). Furthermore, we chose $j=1$. This yields the rate $C\|\mathbf{N}\|^{-1}$ as the upper bound for the convergence rate of MISE.

5.1.2 Numerical results

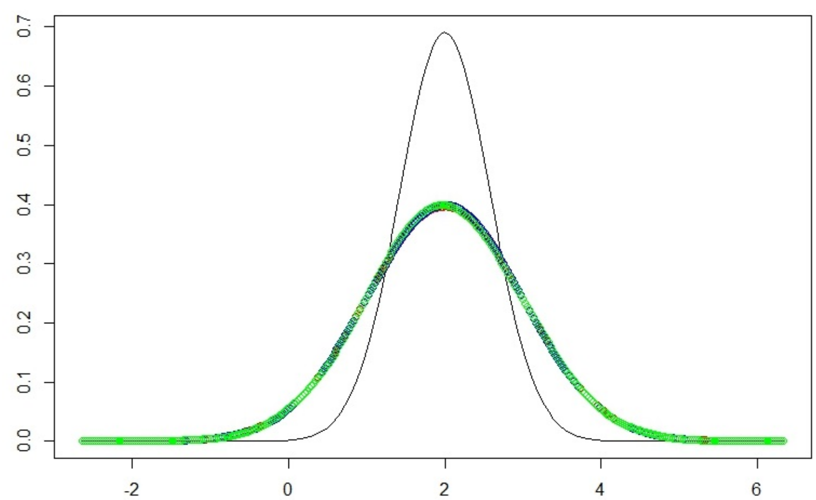


Figure 1: The full data posterior (black line) is shown with the 3 subset posterior densities (red, blue, green) for one iteration of 110,000 samples.

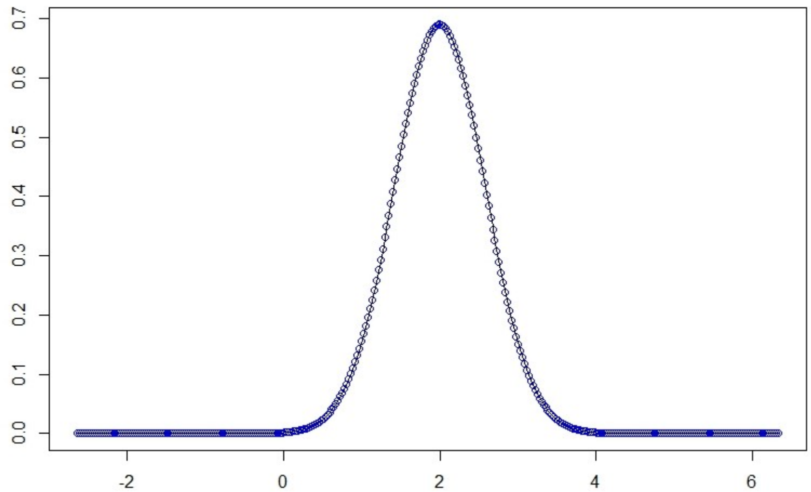


Figure 2: The full data posterior (black line) is shown with the combined subset posterior density (blue points) for one iteration of 110,000 samples.

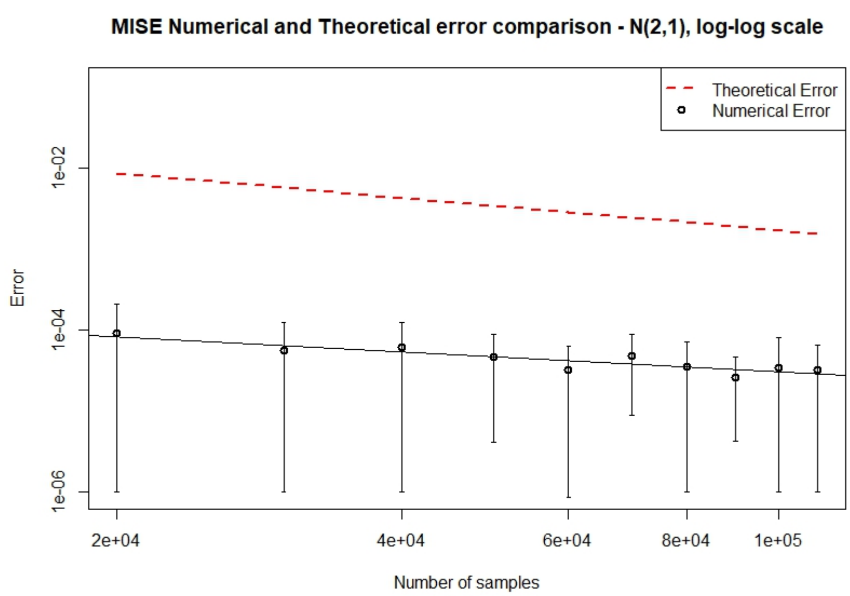


Figure 3: The average MISE estimate is depicted for the ten experiments along with standard deviation bars (black) plotted on a log-log scale with a regression line added. The red line is the upper bound of (38) as calculated for the different number of samples.

Notice in Figure 3 how the regression line and the theoretical error line seem parallel. This implies that the rate obtained from (38) is numerically satisfied.

5.2 Numerical experiment with gamma subset posterior densities

5.2.1 Description of experiment

This experiment mimics the previous one with the normally distributed generated MCMC samples, with the difference now that they are generated by a $Gamma(1, 1)$. The number of samples again increases from $20,000$ to $110,000$ by an increment of $10,000$ for each iteration. Furthermore, $M = 5$ subsets are now created.

5.2.2 Numerical results

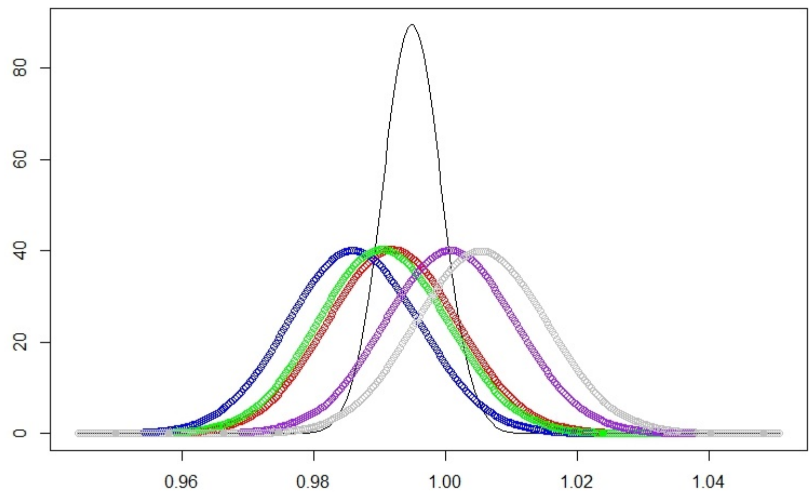


Figure 4: The full data posterior (black line) is shown with the 5 subset posterior densities (red, blue, green, purple, gray) for one iteration of 110,000 samples.

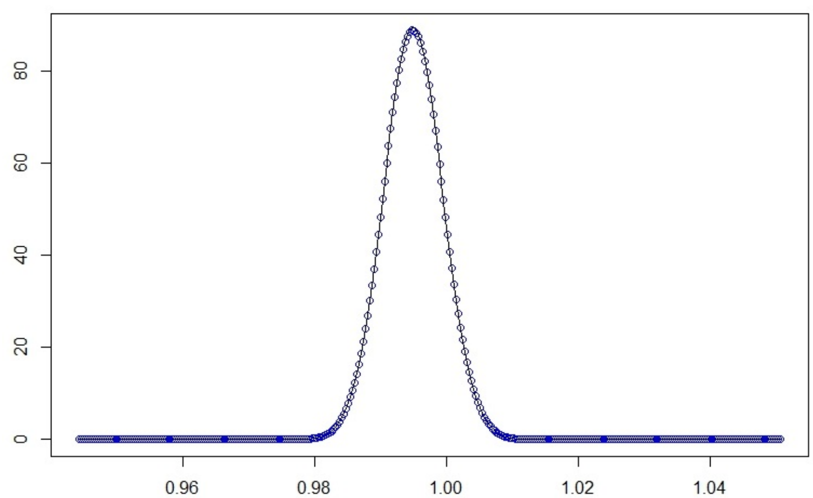


Figure 5: The full data posterior (black line) is shown with the combined subset posterior density (blue points) for one iteration of 110,000 samples.

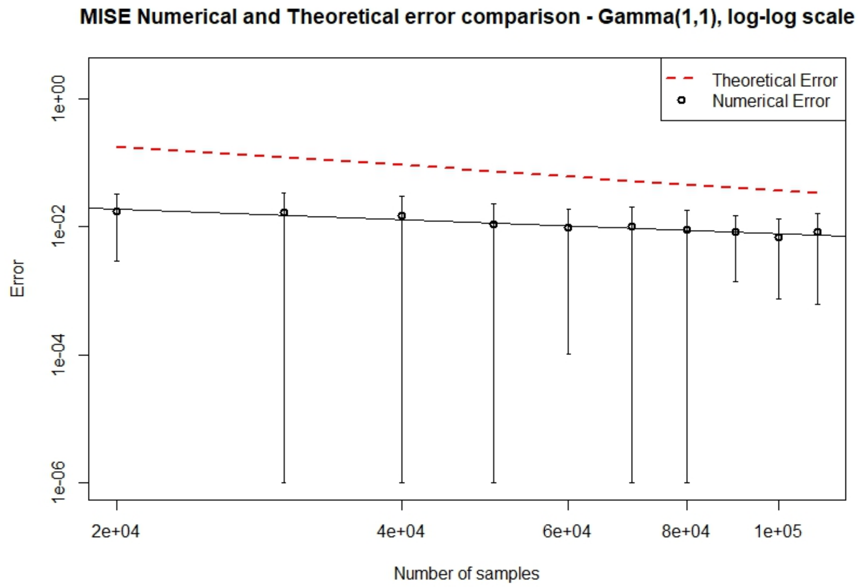


Figure 6: The averaged MISE is depicted for the ten experiments along with standard deviation bars (black) plotted on a log-log scale with a regression line added. The red line is the upper bound of (38) as calculated for the different number of samples.

The result we obtain from Figure 6 is similar to the previous example. The rate from the bound (38) is again numerically satisfied.

5.3 Numerical experiment conducted on flights from/to New York in January 2018

5.3.1 Description of experiment

In this series of experiments we employ the data of US flights that are from or to the state of New York for the month of January 2018. The data was obtained from the Bureau of Transportation Statistics [12]. We were specifically interested in the delayed arrival times, thus flights that arrived 15 minutes or later from the scheduled time. There were a total of 12,100 such flights, which in turn were divided into 5 data subsets of 2,420 each. We assumed that the delayed arrival times are distributed according to a Gamma distribution with some shape parameter and rate parameter. In what follows, we will be doing inference for the shape parameter, denoted by α. Using the JAGS sampling package [11], we generated samples from the marginal full data posterior distribution and the subset posterior distributions for α. The data were shuffled beforehand to ensure that the condition of independence between subsets is satisfied. Ten iterations were performed, starting with 20,000 samples and increasing that number by 10,000. In each iteration, the values were then re-sampled 100 times in order to obtain an approximation to MISE and its standard deviation. Similar to the first example in this section, the parameters were chosen in a manner to achieve optimal convergence for MISE. Therefore, $\beta = 1/2$, cubic B-splines were implemented, which implies $l = 1$, and we chose $j = 1$. These yield the rate of $C\|\mathbf{N}\|^{-1}$ for MISE, as given in (38).

5.3.2 Numerical results

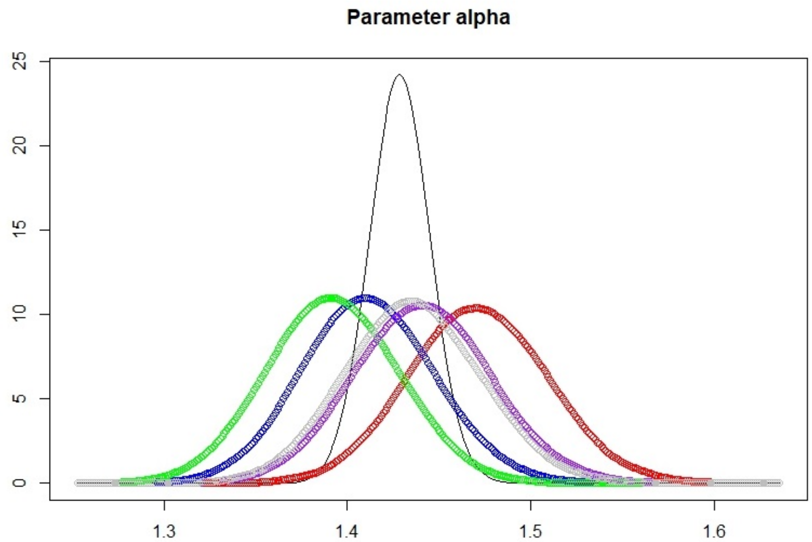


Figure 7: The full data posterior (black line) is shown with the 5 subset posterior densities (red, blue, green, purple, gray) for one iteration of 110,000 samples.

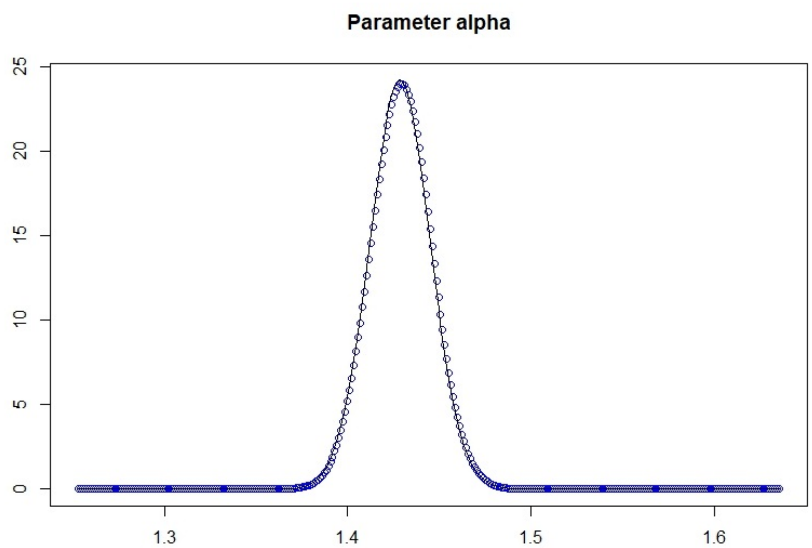


Figure 8: The full data posterior (black line) is shown with the combined subset posterior density (blue points) for one iteration of 110,000 samples.

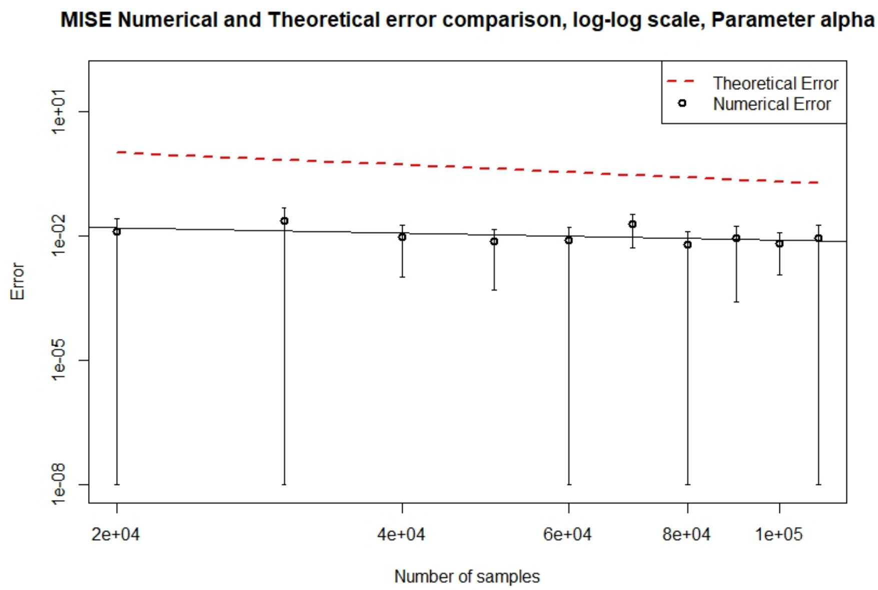


Figure 9: The averaged MISE is depicted for the ten experiments along with standard deviation bars (black) plotted on a log-log scale with a regression line added. The red line is the upper bound of (38) as calculated for the different number of samples.

From Figure 9, the conclusion is similar to the previous examples with the simulated data. The regression line shows that the rate given by (38), with the choice of parameters as mentioned in the description, is again numerically satisfied.

6 Appendix

Here we provide all the relevant results related to B-splines and logspline density estimators based on the works of [1, 2, 3, 4].

6.1 B-Splines

In this section we will define the logspline family of densities and present an overview of how the logspline density estimator is chosen for the density p. The idea behind logspline density estimation of an unknown density p is that the logarithm of p is estimated by a spline function, a piecewise polynomial that interpolates the function to be estimated. Therefore, the family of estimators constructed for the unknown density is a family of functions that are exponentials of splines that are suitably normalized so that they can be densities. Thus, to build up the estimation method, we need to start the theory with the building blocks of splines themselves, the functions we call basis splines or B-splines for short whose linear combination generates the set of splines of a given order.

So, the first question we will answer is how we construct B-splines. There are several ways to do this, some less intuitive than others. The approach we will take will be through the use of **divided differences**. It is a recursive division process that is used to calculate the coefficients of interpolating polynomials written in a specific form called the Newton form.

Definition 19. *The kth divided difference of a function g at the knots $t_0, \ldots, t_k$ is the leading coefficient (meaning the coefficient of x^k) of the interpolating polynomial q of order k+1 that agrees with g at those knots. We denote this number as*

$$[t_0, \ldots, t_k]g \tag{40}$$

Here we use the terminology found in De Boor [1], where a polynomial of order k+1 is a polynomial of degree less than or equal to k. It's better to work with the "order" of a polynomial since all polynomials of a certain order form a vector space, whereas polynomials of a certain degree do not. The term "agree" in the definition means that for the sequence of knots $(t_i)_{i=0}^k$, if ζ appears in the sequence m times, then for the interpolating polynomial we have

$$q^{(i-1)}(\zeta) = g^{(i-1)}(\zeta), \qquad i = 1, \dots, m \tag{41}$$

Since the interpolating polynomial depends only on the data points, the order in which the values of $t_0, \dots, t_1$ appear in the notation in (19) does not matter. Also, if all the knots are distinct, then the interpolating polynomial is unique.

At this point let's write down some examples to see how the recursion algorithm pops up. If we want to interpolate a function g using only one knot, say t_0, then we will of course have the constant polynomial $q(x) = g(t_0)$. Thus, since $g(t_0)$ is the only coefficient, we have

$$[t_0]g = g(t_0) \tag{42}$$

Now suppose we have two knots, t_0, t_1.
If $t_0 \neq t_1$, then q is the secant line defined by the two points $(t_0, g(t_0))$ and $(t_1, g(t_1))$. Thus, the interpolating polynomial will be given by

$$q(x) = g(t_0) + (x - t_0)\frac{g(t_1) - g(t_0)}{t_1 - t_0} \tag{43}$$

Therefore,

$$[t_0, t_1]g = \frac{g(t_1) - g(t_0)}{t_1 - t_0} = \frac{[t_1]g - [t_0]g}{t_1 - t_0} \tag{44}$$

To see what happens when $t_0 = t_1$, we can take the limit $t_1 \to t_0$ above and thus $[t_0, t_1]g = g'(t_0)$.
By continuing these calculations for more knots yields the following result:

Lemma 20. *Given a function g and a sequence of knots $(t_i)_{i=0}^k$, the kth divided difference of g is given by*

(a) $[t_0, \dots, t_k]g = \dfrac{g^{(k)}(t_0)}{k!}$ *when $t_0 = \dots = t_k, g \in C^k$, therefore yielding the leading coefficient of the Taylor approximation of order k+1 to g.*

(b) $[t_0, \dots, t_k]g = \dfrac{[t_0, \dots, t_{r-1}, t_{r+1}, \dots, t_k]g - [t_0, \dots, t_{s-1}, t_{s+1}, \dots, t_k]g}{t_s - t_r}$*, where t_r and t_s are any two distinct knots in the sequence $(t_i)_{i=0}^k$.*

Now that we have defined the kth divided difference of a function, we can easily state what B-splines are. B-splines arise as appropriately scaled divided differences of the positive part of a certain power function and it can be shown that B-splines form a basis of the linear space of splines of some order. Let's start with the definition.

Definition 21. *Let $t = (t_i)_{i=0}^N$ be a nondecreasing sequence of knots. Let $1 \leq k \leq N$. The j-th B-spline of order k, with $j \in \{0, 1, \dots, N-k\}$, for the knot sequence $(t_i)_{i=0}^N$ is denoted by $B_{j,k,t}$ and is defined by the rule*

$$B_{j,k,t}(x) = (t_{j+k} - t_j)[t_j, \dots, t_{j+k}](\cdot - x)_+^{k-1} \tag{45}$$

where $(\cdot)_+$ defines the positive part of a function, i.e. $(f(x))_+ = \max_x\{f(x), 0\}$.

The "placeholder" notation in the above definition says that the kth divided difference of $(\cdot - x)_+^{k-1}$ is to be considered for the function $(t - x)_+^{k-1}$ as a function of t and have x fixed. Of course, in the end the number will vary as x varies, giving rise to the function $B_{j,k,t}$. If either k or t can be inferred from context then we will usually drop them from the notation and write B_j instead of $B_{j,k,t}$. A direct consequence we receive from the above definition is the support of $B_{j,k,t}$.

Lemma 22. *Let $B_{j,k,t}$ be defined as in 21. Then the support of the function is contained in the interval $[t_j, t_{j+k})$.*

Proof. All we need to do is show that if $x \notin [t_j, t_{j+k})$, then $B_{j,k,t}(x) = 0$.
Suppose first that $x \geq t_{j+k}$. Then we will have that $t_i - x \leq 0$ for $i = j, \dots, j+k$ which in turn implies $(t_i - x)_+ = 0$ and finally $[t_j, \dots, t_{j+k}](\cdot - x)_+^{k-1} = 0$.
On the other hand, if $x < t_j$, then since $(t - x)_+^{k-1}$ as a function of t is a polynomial of order k and we have $k+1$ sites where it agrees with its interpolating polynomial, necessarily they are both the same. This implies $[t_j, \dots, t_{j+k}](\cdot - x)_+^{k-1} = 0$ since the coefficient of t^k is zero. ☐

6.2 Recurrence relation and various properties

Since we stated the definition of B-splines using divided differences, we can use that to state the **recurrence relation** for B-splines which will be useful when we will later prove various properties of these functions. We start by stating and proving the Leibniz formula which will be needed in the proof of the recurrence relation

Lemma 23. *Suppose f, g, h are functions such that $f = g \cdot h$, meaning $f(x) = g(x)h(x)$ for all x and let (t_i) be a sequence of knots. Then we have the following formula*

$$[t_j, \dots, t_{j+k}]f = \sum_{r=j}^{j+k}([t_j, \dots, t_r]g)([t_r, \dots, t_{j+k}]h), \quad \textit{for some } j, k \in \mathbb{N}. \tag{46}$$

Proof. First of all, observe that the function

$$\left(g(t_j) + \sum_{r=j+1}^{j+k}(x - t_j)\dots(x - t_{r-1})[t_j, \dots, t_r]g\right) \cdot \left(h(t_{j+k}) + \sum_{s=j}^{j+k-1}(x - t_{s+1})\dots(x - t_{j+k})[t_s, \dots, t_{j+k}]h\right)$$

agrees with f at the knots $t_j, \dots, t_{j+k}$ since the first and second factor agree with g and h respectively at those values. Now, observe that if $r > s$ then the above product vanishes at all the knots since the term $(x - t_i)$ for $i = j, \dots, j+k$ will appear in at least one of the two factors. Thus, the above agrees with f at $t_j, \dots, t_{j+k}$ when $r \leq s$. But then the product turns into a polynomial of order $k+1$ whose leading coefficient is

$$\sum_{r=s}([t_j, \dots, t_r]g)([t_s, \dots, t_{j+k}]h)$$

and that of course must be equal to

$$[t_j, \dots, t_{j+k}]f$$

☐

Now we can state and prove the recurrence relation for B-splines.

Lemma 24. *Let $t = (t_i)_{i=0}^N$ be a sequence of knots and let $1 \leq k \leq N$. For $j \in \{0, 1, \dots, N-k\}$ we can construct the j-th B-spline $B_{j,k}$ of order k associated with the knots $t = (t_i)_{i=0}^N$ as follows:*

(1) First we have $B_{j,1}$ be the characteristic function on the interval $[t_j, t_{j+1})$

$$B_{j,1}(x) = \begin{cases} 1, & x \in [t_j, t_{j+1}) \\ 0, & x \notin [t_j, t_{j+1}) \end{cases} \tag{47}$$

(2) The B-splines of order k for $k > 1$ on $[t_j, t_{j+k})$ are given by

$$B_{j,k}(x) = \frac{x - t_j}{t_{j+k-1} - t_j} B_{j,k-1}(x) + \frac{t_{j+k} - x}{t_{j+k} - t_{j+1}} B_{j+1,k-1}(x) \tag{48}$$

Proof. (1) easily follows from the definition we gave for B-splines using divided differences in Definition 21. (2) can be proven using Lemma 23. Since B-splines were defined using the function $(t - x)_+^{k-1}$ for fixed x, we apply the Leibniz formula for the kth divided difference to the product

$$(t - x)_+^{k-1} = (t - x)(t - x)_+^{k-2}$$

This yields

$$[t_j, \dots, t_{j+k}](\cdot - x)_+^{k-1} = (t_j - x)[t_j, \dots, t_{j+k}](\cdot - x)_+^{k-2} + 1 \cdot [t_{j+1}, \dots, t_{j+k}](\cdot - x)_+^{k-2} \tag{49}$$

since $[t_j](\cdot - x) = (t_j - x)$, $[t_j, t_{j+1}](\cdot - x) = 1$ and $[t_j, \dots, t_r](\cdot - x) = 0$ for $r > j+1$. Now, from Lemma 20 (b), we have that $(t_j - x)[t_j, \dots, t_{j+k}](\cdot - x)_+^{k-2}$ can be written as

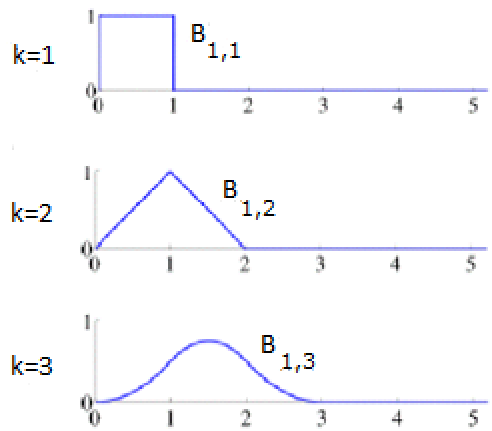
$$(t_j - x)[t_j, \dots, t_{j+k}](\cdot - x)_+^{k-2} = \frac{t_j - x}{t_{j+k} - t_j}([t_{j+1}, \dots, t_{j+k}] - [t_j, \dots, t_{j+k-1}]) \tag{50}$$

Thus, by replacing that term in the result (49) we obtained by Leibniz, we get

$$[t_j, \dots, t_{j+k}](\cdot - x)_+^{k-1} = \frac{x - t_j}{t_{j+k} - t_j}[t_j, \dots, t_{j+k-1}](\cdot - x)_+^{k-2} + \frac{t_{j+k} - x}{t_{j+k} - t_j}[t_{j+1}, \dots, t_{j+k}](\cdot - x)_+^{k-2} \tag{51}$$

The result in (2) follows immediately once we multiply both sides by $(t_{j+k} - t_j)$ and then multiply and divide the first term in the sum on the right hand side by $(t_{j+k-1} - t_j)$ and then multiply and divide the second term by $(t_{j+k} - t_{j+1})$. □

From the recurrence relation we acquire information about B-splines that was not clear from the first definition we gave using divided differences. $B_{j,1}$ is a characteristic function, or otherwise piecewise constant. By Lemma 24 (b), since the coefficients of $B_{j,k-1}$ are linear functions of x, we have $B_{j,2}$ is a piecewise linear function on $[t_j, t_{j+2})$. Therefore, inductively we have $B_{j,3}$ is a piecewise parabolic function on $[t_j, t_{j+3})$, $B_{j,4}$ is a piecewise polynomial of degree 3 on $[t_j, t_{j+4})$ and so on. Below there is a visual representation of B-splines showing how the graph changes as the order increases.



Since we now have defined what a B-spline is as a function, the next step is to ask what set is generated when considering linear combinations of these functions. Since B-splines are piecewise polynomials themselves, we have that this set is a subset of the set of piecewise polynomials with breaks at the knots (t_i). Something that can be proven though, is that it is exactly the set of piecewise polynomials with certain break and continuity conditions at the knots and this equality occurs on a smaller interval, which we call the basic interval, denoted by $I_{k,t}$.

Definition 25. Suppose $t = (t_0, \dots, t_N)$ is a nondecreasing sequence of knots. Then for the B-splines of order k, with $2k < N + 2$, that arise from these knots, we define $I_{k,t} = [t_{k-1}, t_{N-k+1}]$ and call it the **basic interval**.

Remark 26. In order for this definition to be correct, we need to extend the B-splines and have them be left continuous at the right endpoint of the basic interval since we are defining it as a closed interval.

Remark 27. The basic interval for the $N - k + 1$ B-splines of order $k > 1$ is defined in such a way so that at least two of them are always supported on any subinterval of $I_{k,t}$ and later we will see that the B-splines form a partition of unity on the basic interval. For $k = 1$, by construction the B-splines already form a partition of unity on $I_{1,t} = [t_0, t_N]$.

For example, let $t = (t_i)_{i=0}^6$ be disjoint and $k = 3$. Then there are 4 B-splines, $B_{j,3}$, $j = 0, 1, 2, 3$, of order 3 that arise in this framework. Their supports are $[t_0, t_3)$, $[t_1, t_4)$, $[t_2, t_5)$, $[t_3, t_6)$ respectively. Clearly, on $[t_0, t_1)$ only $B_{0,3}$ is supported and since as a function is non-constant we cannot have $\sum_{j=0}^{3} B_{j,3} = B_{0,3}$ on $[t_0, t_1)$ be equal to 1.

The partition of unity is stated and proved in the next lemma together with other properties of the B-splines. The recurrence relation makes the proofs fairly easy compared to using the divided difference definition of the B-splines.

Lemma 28. *Let $B_{j,k,t}$ be the function as given in Definition 21 for the knot sequence $t = (t_i)_{i=0}^{N}$. Then the following hold:*

(a) $B_{j,k,t}(x) > 0$ for $x \in (t_j, t_{j+k})$.

(b) (Marsden's Identity) For any $\alpha \in \mathbb{R}$, we have $(x-\alpha)^{k-1} = \sum_j \psi_{j,k}(\alpha) B_{j,k,t}(x)$, where $\psi_{j,k}(\alpha) = (t_{j+1} - \alpha)\dots(t_{j+k-1} - \alpha)$ and $\psi_{j,1}(\alpha) = 1$.

(c) $\sum_j B_{j,k,t} = 1$ on the basic interval $I_{k,t}$.

Proof. (a) This is a simple induction. For $k = 1$ the hypothesis holds since the B-splines are just characteristic functions on $[t_j, t_{j+1})$ and thus strictly positive in the interior.
For $k = 2$ by the recurrence relation, $B_{j,2,t}$ is a linear combination of $B_{j,1},\ B_{j+1,1}$ with coefficients the linear functions $\frac{x-t_j}{t_{j+1}-t_j},\ \frac{t_{j+2}-x}{t_{j+2}-t_{j+1}}$ which is positive on (t_j, t_{j+2}).
Assuming the hypothesis holds for $k = r$, we can show it is true for $k = r+1$ by using the same argument as in the previous case.

(b) Let $\omega_{j,k}(x) = \frac{x-t_j}{t_{j+k-1}-t_j}$. Thus, $\frac{t_{j+k}-x}{t_{j+k}-t_{j+1}} = 1 - \omega_{j+1,k}(x)$. This way we can write the recurrence relation as

$$B_{j,k}(x) = \omega_{j,k}(x) B_{j,k-1}(x) + (1 - \omega_{j+1,k}(x)) B_{j+1,k-1} \tag{52}$$

Using this we can write $\sum_j \psi_{j,k}(\alpha) B_{j,k,t}(x)$ as

$$\begin{aligned}
\sum_j \psi_{j,k}(\alpha) B_{j,k,t}(x) &= \sum_j [\omega_{j,k}(x)\psi_{j,k}(\alpha) + (1-\omega_{j,k}(x))\psi_{j-1,k}(\alpha)] B_{j,k-1,t}(x) \\
&= \sum_j \psi_{j,k-1}(\alpha)[\omega_{j,k}(x)(t_{j+k-1} - \alpha) + (1-\omega_{j,k}(x))(t_j - \alpha)] B_{j,k-1,t}(x) \\
&= \sum_j \psi_{j,k-1}(\alpha)(x-\alpha) B_{j,k-1,t}(x)
\end{aligned} \tag{53}$$

since $\omega_{j,k}(x) f(t_{j+k-1}) + (1-\omega_{j,k}(x)) f(t_j)$ is the unique straight line that intersects f at $x = t_j$ and $x = t_{j+k-1}$. Thus,

$$\omega_{j,k}(x)(t_{j+k-1} - \alpha) + (1-\omega_{j,k}(x))(t_j - \alpha) = x - \alpha$$

Therefore, by induction we have

$$\begin{aligned}
\sum_j \psi_{j,k}(\alpha) B_{j,k,t}(x) &= \sum_j \psi_{j,1}(\alpha)(x-\alpha)^{k-1} B_{j,1,t}(x) \\
&= (x-\alpha)^{k-1} \sum_j \psi_{j,1}(\alpha) B_{j,1,t}(x) \\
&= (x-\alpha)^{k-1}
\end{aligned}$$

since $\psi_{j,1}(\alpha) = 1$ and $B_{j,1,t}$ are just characteristic functions.

(c) To prove the partition of unity, we start with Marsden's Identity and divide both sides by $(k-1)!$ and differentiate $\nu - 1$ times with respect to α for some positive integer $\nu \le k-1$. We then have

$$\frac{(x-\alpha)^{k-\nu}}{(k-\nu)!} = \sum_j \frac{(-1)^{\nu-1}}{(k-1)!} \frac{d^{\nu-1}\psi_{j,k}(\alpha)}{d\alpha^{\nu-1}} B_{j,k,t}(x) \tag{54}$$

Now, for some polynomial q of order k, we can use the Taylor expansion of q

$$q = \sum_{\nu=1}^{k} \frac{(x-\alpha)^{k-\nu}}{(k-\nu)!} \frac{d^{k-\nu} q(\alpha)}{d\alpha^{k-\nu}} \tag{55}$$

Using this we see that

$$q = \sum_j \lambda_{j,k}[q] B_{j,k,t} \quad \text{where} \quad \lambda_{j,k}[q] = \sum_{\nu=1}^{k} \frac{(-1)^{\nu-1}}{(k-1)!} \frac{d^{\nu-1}\psi_{j,k}(\alpha)}{d\alpha^{\nu-1}} \frac{d^{k-\nu}q(\alpha)}{d\alpha^{k-\nu}} \tag{56}$$

which holds only on the basic interval. Now, to show that the B-splines are a partition of unity, we just use this identity for $q = 1$. □

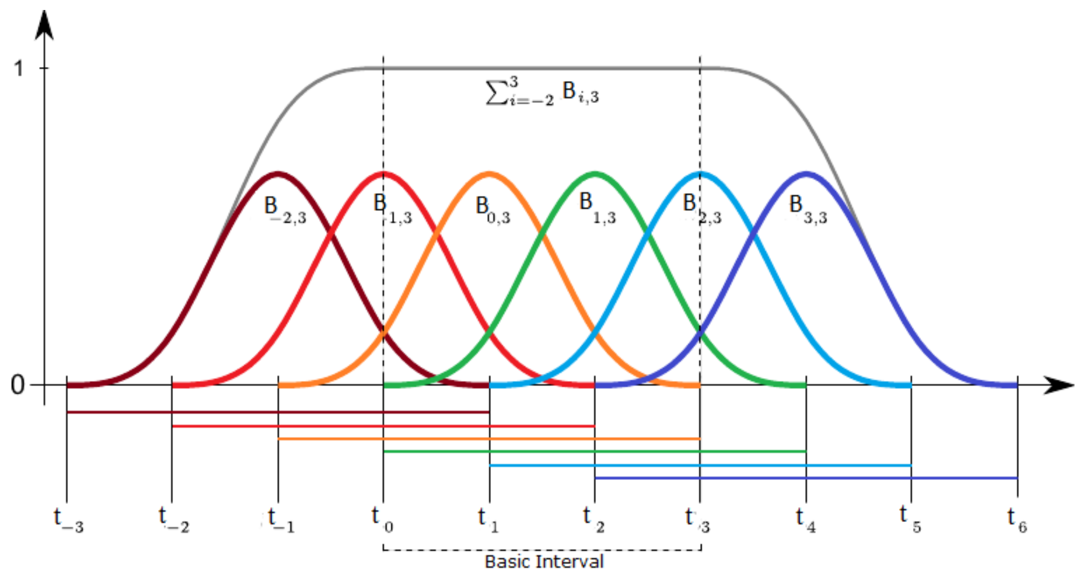
Remark 29. *Marsden's Identity says something very important. That all polynomials of order k are contained in the set generated by the B-splines $B_{j,k}$, which is also what makes the step in the proof of (c) viable. Furthermore, we can replace the $(x - \alpha)$ in the identity by $(x - \alpha)_+$ which shows that piecewise polynomials are also contained in the same set.*

Remark 30. *Another consequence of Marsden's Identity is the Curry-Schoenberg theorem. We do not explicitly state the theorem as we do not require it, rather we state a simple result from it for B-splines of order k given a sequence of knots $(t_i)_{i=0}^N$, which can be summarized as*

$$\textit{number of continuity conditions at } t_i + \textit{multiplicity of } t_i = k$$

Therefore, for a simple knot t_i, any B-spline of order k there will be continuous and also have $k-2$ continuous derivatives. On the other hand, if t_i has multiplicity k, any k-th order B-spline will have a discontinuity there.

Below there is a figure which shows the importance of the basic interval as the interval where we have partition of unity.



Remark 31. When the sequence of $t_i's$ is distinct then the sum of B-splines belongs to $C_0\big((t_0, t_N)\big)$. However, the sum of B-splines on the basic interval $I_{k,t}$ is equal to 1. To make sure that the sum equals to 1 on the whole interval (t_0, t_N), the assumption of the knots being distinct has to be dropped. It is obvious that we have to take $t_0 = \cdots = t_{k-1}$ and $t_{N-k+1} = \cdots = t_N$.

Definition 32. *Let $(t_i)_{i=0}^N$ be a sequence of knots such that $t_0 = \cdots = t_{k-1}$ and $t_{N-k+1} = \cdots = t_N$, where $1 \leq k \leq N$. Let $B_{j,k,t}$ be the B-splines as defined in 21 with knot sequence $t = (t_i)_{i=0}^N$. The set generated by the sequence $\{B_{j,k,t} : \text{all } j\}$, denoted by $\mathcal{S}_{k,t}$, is the set of splines of order k with knot sequence t. In symbols we have*

$$\mathcal{S}_{k,t} = \left\{ \sum_j a_j B_{j,k,t} : a_j \in \mathbb{R}, \text{ all } j \right\} \tag{57}$$

Remark 33. *Fix an interval $[a, b]$. Let $T_N = (t_i)_{i=0}^N$ be a sequence as in definition 32 with $t_0 = a$ and $t_N = b$, where $N \in \mathbb{N}$. The choice in definition 32 implies that*

$$\bigcup_{N \in \mathbb{N}} S_{k,T_N} \quad \text{is dense in} \quad C([a, b]) \tag{58}$$

6.3 Derivatives of B-spline functions

Later in this paper when we will be conducting our analysis on MISE, derivatives of spline functions will factor in. Since splines are just linear combinations of B-splines we just need to investigate the result of differentiating a B-spline on the interior of its support. The derivative of a k-th order B-spline is directly associated with B-splines of order $k-1$. To see this we use the recurrence relation which leads us to the following theorem:

Theorem 34. *Let $B_{j,k,t}$ be the function as defined in 21. The support of $B_{j,k,t}$ is the interval $[t_j, t_{j+k})$. Then the following equation holds on the open interval (t_j, t_{j+k})*

$$\frac{d}{d\theta} B_{j,k,t}(\theta) = \begin{cases} 0, & k = 1 \\ (k-1)\left(\dfrac{B_{j,k-1,t}(\theta)}{t_{j+k-1} - t_j} - \dfrac{B_{j+1,k-1,t}(\theta)}{t_{j+k} - t_{j+1}}\right), & k > 1 \end{cases} \tag{59}$$

Proof. The proof is done by induction on k. For $k = 1$ it is straightforward since $B_{j,1,t}$ is a constant on (t_j, t_{j+1}) and for $k > 1$ we use the recurrence relation described in lemma 24. □

Using the above formula we can easily obtain bounds for higher derivatives of B-splines. First of all, by construction of the space $\mathcal{S}_{k,t}$, the B-splines we will be working with form a partition of unity on $[t_0, t_N]$ and since they are strictly positive on the interior of their supports, we have that each B-spline is bounded by 1 for all θ.

$$B_{j,k,t}(\theta) \leq 1, \quad \forall \theta \in \mathbb{R}$$

Furthermore, by induction we can prove the following lemma:

Lemma 35. *Let $t = (t_i)_{i=0}^N$ be a sequence of knots as in definition 32 and $B_{j,k,t}$ be the function as defined in 21. Let $h_N = \min\limits_{k \leq i \leq N-k+1} (t_i - t_{i-1})$ and α be a positive integer such that $\alpha < k - 1$. Then, on the open interval (t_j, t_{j+k}) we have*

$$\sup_{\theta \in (t_j, t_{j+k})} \left| \frac{d^\alpha}{d\theta^\alpha} B_{j,k,t}(\theta) \right| \leq \frac{2^\alpha}{h_N^\alpha} \frac{(k-1)!}{(k-\alpha-1)!}, \textit{for any } j$$

Proof. We fix k and we do induction on α. Let's start with $\alpha = 1$

$$\left| \frac{d}{d\theta} B_{j,k,t}(\theta) \right| = \left| (k-1)\left(\frac{B_{j,k-1,t}(\theta)}{t_{j+k-1} - t_j} - \frac{B_{j+1,k-1,t}(\theta)}{t_{j+k} - t_{j+1}}\right) \right| \leq \frac{2}{h_N} \frac{(k-1)!}{(k-2)!}.$$

Thus the inequality holds for $\alpha = 1$.
Now we assume it holds for $\alpha = n$ and we will show it holds for $\alpha = n + 1$.

$$\left| \frac{d^{n+1}}{d\theta^{n+1}} B_{j,k,t}(\theta) \right| = \left| \frac{d^n}{d\theta^n} (k-1)\left(\frac{B_{j,k-1,t}(\theta)}{t_{j+k-1} - t_j} - \frac{B_{j+1,k-1,t}(\theta)}{t_{j+k} - t_{j+1}}\right) \right| \leq \frac{2^{n+1}}{h_N^{n+1}} \frac{(k-1)!}{[k-(n+1)-1]!}.$$

This concludes the proof. □

Remark 36. *Considering Remark 30, the bound in Lemma 35 can be extended to hold on the closed interval $[t_j, t_{j+k}]$ assuming the knots $t_j, \ldots, t_{j+k}$ are simple. Also, it is clear that we need to utilize at least parabolic B-splines in order to have a bound on a continuous derivative.*

6.4 Logspline Density Estimation

In this part we will present the method for constructing logspline density estimators using B-splines. Let p be a continuous probability density function supported on an interval $[a, b]$. Suppose p is unknown and we would like to construct density estimators for this function. The methodology is as follows

Definition 37. *Let $T_N = (t_i)_{i=0}^N$, $N \in \mathbb{N}$, be a sequence of knots such that $t_0 = \cdots = t_{k-1} = a$ and $t_{N-k+1} = \cdots = t_N = b$, where $1 \le k \le N$, k fixed. Thus, the set of splines S_{k,T_N} of order k generated by the B-splines B_{j,k,T_N} can be obtained. We suppress the parameters k, T_N and just write B_j instead of B_{j,k,T_N}. Define the spline function*

$$B(\theta; y) = \sum_{j=0}^{L} y_j B_j(\theta)\,, \quad y = (y_0, \ldots, y_L) \in \mathbb{R}^{L+1} \quad \textit{with } L := N - k. \tag{60}$$

and for each y we set the probability density function

$$\begin{aligned} f(\theta; y) &= \exp\Big(\sum_{j=0}^{L} y_j B_j(\theta) - c(y))\Big) = \exp\Big(B(\theta; y) - c(y))\Big)\,, \\ \textit{where} \quad c(y) &= \log\left(\int_a^b \exp\Big(\sum_{j=0}^{L} y_j B_j(\theta)\Big) d\theta\right) < \infty\,. \end{aligned} \tag{61}$$

The family of exponential densities $\{f(\theta; y) : \ y \in \mathbb{R}^{L+1}\}$ is not identifiable since if β is any constant, then $c((y_0 + \beta, \ldots, y_L + \beta)) = c(y) + \beta$ and thus

$$f(\theta; (y_0 + \beta, \ldots, y_L + \beta)) = f(\theta; y)$$

To make the family identifiable we restrict the vectors y to the set

$$Y_0 = \left\{ y \in \mathbb{R}^{L+1} : \sum_{i=0}^{L} y_i = 0 \right\}. \tag{62}$$

Remark 38. *Y_0 depends only on the number of knots and the order of the B-splines and not the number of samples.*

Definition 39. *We define the logspline model as the family of estimators*

$$\mathcal{L} = \big\{ f(\theta; y) \ \textit{given by} \ (61) : \ y \in Y_0 \big\}. \tag{63}$$

For any $f \in \mathcal{L}$

$$\log(f) = \sum_{j=0}^{L} y_j B_j(\theta) - c(y) \in S_{k,T_N}. \tag{64}$$

Next, let us pick a set of independent, identically distributed random variables

$$\Theta_n = \big(\theta_1, \theta_2, ..., \theta_n\big) \in \mathbb{R}^n, \ n \in \mathbb{N}$$

where each θ_i is drawn from a distribution that has density $p(\theta)$.
We next define the log-likelihood function $l_n : \mathbb{R}^{L+1+n} \to \mathbb{R}$ corresponding to the logspline model by

$$\begin{aligned} l_n(y) &= l_n(y; \theta_1, \theta, \ldots, \theta_n) = l_n(y; \Theta_n) \\ &= \sum_{i=1}^{n} \log(f(\theta_i; y)) = \sum_{i=1}^{n} \left(\sum_{j=0}^{L} y_j B_j(\theta_i) \right) - nc(y)\,, \quad y \in Y_0 \end{aligned} \tag{65}$$

and the maximizer of the log-likelihood $l_n(y)$ by

$$\hat{y}_n = \hat{y}_n(\theta_1, \ldots, \theta_n) = \arg\max_{y \in Y_0} l_n(y) \tag{66}$$

whenever this random variable exists, which will be shown on a subset of the sample space whose probability will tend to 1. The density $f(\,\cdot\,; \hat{y}_n)$ is called the *logspline density estimate* of p.

We define the expected log-likelihood function $\lambda_n(y)$ by

$$\lambda_n(y) = \mathbb{E}[l(y;\theta_1,\ldots,\theta_n)] = n\left(-c(y) + \int_a^b \left(\sum_{j=0}^{L} y_j B_j(\theta)\right)p(\theta)\,d\theta\right) < \infty\,, \quad y \in Y_0. \tag{67}$$

It follows by a convexity argument that the expected log-likelihood function has a unique maximizing value

$$\bar{y} = \arg\max_{y\in Y_0} \lambda_n(y) = \arg\max_{y\in Y_0} \frac{\lambda_n(y)}{n} \tag{68}$$

which is independent of n but depends on the knots.

Note that the function $\lambda_n(y)$ is bounded above and goes to $-\infty$ as $|y| \to \infty$ within Y_0 and therefore, due to Jensen's Inequality, the constant $\bar{y}$ is finite; see Stone [4]. The estimator $\hat{y}(\theta_1,\ldots,\theta_n)$, in general does not exist. This motivates us to define the set

$$\Omega_n = \left\{\omega \in \Omega : \hat{y} = \hat{y}(\theta_1,\ldots,\theta_n) \in \mathbb{R}^{L+1} \text{ exists}\right\}. \tag{69}$$

In what follows we will show that $\mathbb{P}(\Omega_n) \to 1$ as $n \to \infty$. We also note that due to convexity of $l_n(y)$ and $\lambda_n(y)$ the estimators $\hat{y}$ and $\bar{y}$ are unique whenever they exist.

We define the logspline estimator $\hat{p}$ of p on the space Ω_n by

$$\hat{p} : \mathbb{R}\times\Omega_n \quad \text{defined by} \quad \hat{p}(\theta,\omega) = f(\theta,\hat{y}(\theta_1,\ldots,\theta_n)),\ \omega\in\Omega_n \tag{70}$$

and define the function

$$\bar{p}(\theta) := f(\theta,\bar{y})\,. \tag{71}$$

Remark 40. *In order for the maximum likelihood estimates to be reliable, we require that the modeling error tend to 0 as $n \to \infty$, as described in hypothesis* (H4).

6.5 Notions of distance from the set of splines $S_{k,t}$

It is a well known fact that continuous functions can be approximated by polynomials. Now that we have defined the set of splines $S_{k,t}$ in Definition 32 and from what we have stated in remark 33, that $\bigcup_{N\in\mathbb{N}} S_{k,T_N}$ is dense in the space of continuous functions, there is a question that arises at this point:

Given an arbitrary continuous function g on $[a,b]$, an integer $k \geq 1$ and a set of knots $T_N = (t_i)_{i=0}^N$ as in Remark 33, how close is g to the set S_{k,T_N} of splines of order k?

Let's state this question in a slightly different way. What we would like to do is find a bound for the sup-norm distance between $g \in C[a,b]$ and S_{k,T_N}, where this distance is denoted by $dist(g,S_{k,T_N})$ and is defined as

$$\text{dist}(g,S_{k,T_N}) = \inf_{s\in S_{k,T_N}} \|g-s\|_\infty, \quad g\in C[a,b]. \tag{72}$$

The answer to our question is given by Jackson's Theorem found in de Boor [1]. To state it we first need the following definition.

Definition 41. *The modulus of continuity $\omega(g;h)$ of some function $g\in C[a,b]$ for some positive number h is defined as*

$$\omega(g;h) = \max\{|g(\theta_1)-g(\theta_2)| :\ \theta_1,\theta_2\in[a,b],\ |\theta_1-\theta_2|\leq h\}. \tag{73}$$

The bound given by Jackson's Theorem contains the modulus of continuity of the function whose sup-norm distance we want to estimate from the set of splines. The theorem is stated below.

Theorem 42. *Let $T_N = (t_i)_{i=0}^N$, $N\in\mathbb{N}$, be a sequence of knots such that $t_0 = \cdots = t_{k-1} = a$ and $b = t_{N-k+1} = \cdots = t_N$, where $1\leq k\leq N$. Let S_{k,T_N} be the set of splines as in definition 32 for the knot sequence T_N. For each $j\in\{0,\ldots,k-1\}$, there exists $C = C(k,j)$ such that for $g\in C^j[a,b]$*

$$dist(g,S_{k,T_N}) \leq C\ h^j\ \omega\left(\frac{d^j g}{d\theta^j};|t|\right) \quad \textit{where} \quad h = \max_i |t_{i+1}-t_i|. \tag{74}$$

In particular, from the Mean Value Theorem it follows

$$dist(g,S_{k,T_N}) \leq C\ h^{j+1}\ \left\|\frac{d^{j+1}g}{d\theta^{j+1}}\right\|_\infty \tag{75}$$

in the case that $g\in C^{j+1}[a,b]$.

Remark 43. *Please note that for the approximation the mesh size enters into the bound in* (75) *which dictates the placement for the knots.*

Jackson's Theorem supplies us with an estimate of how good an approximation is contained in the space of splines for a continuous function. However, in this paper we are interested in estimates for probability densities, especially since the focus is on logspline density estimates. At this point let's state results specifically for densities. The following can be found in Stone [3].
Suppose that p is a continuous probability density supported on some interval $[a, b]$, similar to the set-up when we defined the logspline density estimation method. Define the family $\mathcal{F}_p$ of densities such that

$$\mathcal{F}_p = \left\{ p_\alpha : \; p_\alpha(x) = \frac{(p(x))^\alpha}{\int (p(y))^\alpha \, dy}, \; 0 \le \alpha \le 1 \right\}. \tag{76}$$

It is easy to see that for $\alpha \in [0, 1]$ p_α is a probability density on [a,b]. An interesting consequence from this family is the following

Lemma 44. *We define the family of functions*

$$\mathcal{F}_p^{log} = \{\log(u) : \; u \in \mathcal{F}_p\}. \tag{77}$$

Then, $\mathcal{F}_p^{log}$ *defines a family of functions that is equicontinuous on the set* $\{\theta : \; p(\theta) > 0\}$.

Proof. The proof is simple enough. Pick $\epsilon > 0$. There exists $\delta > 0$ such that $|\log(p(x)) - \log(p(y))| < \epsilon$ whenever $|x - y| < \delta$. Pick any $\alpha \in [0, 1)$.
If $\alpha = 0$ then p_0 is just a constant and thus $|\log(p_0(x)) - \log(p_0(y))| = 0 < \epsilon$.
If $0 < \alpha < 1$, then $|\log(p_\alpha(x)) - \log(p_\alpha(y))| = |\alpha \log(p(x)) - \alpha \log(p(y))| < \alpha \; \epsilon < \epsilon$. □

Remark 45. *It is practical to work with* $p(x) > 0$ *on the set* $[a, b]$ *and this is what we assume until the end of the manuscript. In this case,* $\log(p) \in C[a, b]$.

Remark 46. *We will be using the notation* $\bar{h} = \max_i |t_{i+1} - t_i|$ *and* $\underline{h} = \min_i |t_{i+1} - t_i|$, *and* $\gamma(T_N) = \bar{h}/\underline{h}$.

We can apply the logspline estimation method to p. Let $\bar{p}$ be defined as in (71), the density estimate given by maximizing the expected log-likelihood. We then have the following lemma:

Lemma 47. *Suppose* p *is an unknown continuous density function supported on* $[a, b]$ *and* $\bar{p}$ *is as in* (71). *Then there exists constant* $M' = M'(\mathcal{F}_p, k, \gamma(T_N))$ *that depends on the family* $\mathcal{F}_p$, *order* k *and global mesh ratio* $\gamma(T_N)$ *of* S_{k,T_N} *such that*

$$\|\log(p) - \log(\bar{p})\|_\infty \le M' \; \operatorname{dist}(\log(p), S_{k,T_N}) \tag{78}$$

and therefore

$$\|p - \bar{p}\|_\infty \le \big(\exp\{M' \; \operatorname{dist}(\log(p), S_{k,T_N})\} - 1\big) \|p\|_\infty. \tag{79}$$

Moreover, if $\log(p) \in C^{j+1}([a, b])$ *for some* $j \in \{0, \ldots, k-1\}$ *then by Jackson's Theorem we obtain*

$$\begin{aligned} \|\log(p) - \log(\bar{p})\|_\infty &\le M' \; C(k, j) \; \bar{h}^{j+1} \; \left\| \frac{d^{j+1} \log(p)}{d\theta^{j+1}} \right\|_\infty \\ \|p - \bar{p}\|_\infty &\le \left(\exp\left\{ M' \; C(k, j) \; \bar{h}^{j+1} \; \left\| \frac{d^{j+1} \log(p)}{d\theta^{j+1}} \right\|_\infty \right\} - 1 \right) \|p\|_\infty. \end{aligned} \tag{80}$$

Remark 48. *Please note that the constant* M *does not depend on the dimension of* S_{k,T_N}. *For all practical purposes, we will be using uniformly placed knots, thus suppressing the dependence on* $\gamma(T_N)$, *which will be equal to the constant 1.*

Now we will present certain error bounds required to calculate a bound for MISE. Assume $p, \hat{p}$ and $\bar{p}$ as in the previous section. Also, assume that n is the number of random samples drawn from p.

We will state a series of definitions and theorems that encompass the results from Lemma 5, Lemma 6, Lemma 7, and Lemma 8 in the work of Stone[4][pp.728-729].

Definition 49. *Let $n \geq 1$ and $b > 0$. Let $y \in Y_0$. Let l_n and λ_n be defined by* (65) *and* (67)*, respectively. We define*

$$A_{n,b}(y) = \Bigg\{ \omega \in \Omega : |l(y;\Theta_n(\omega)) - l(\bar{y};\Theta_n(\omega)) - (\lambda_n(y) - \lambda_n(\bar{y}))| \\ < nb\bigg(\int |\log(f(\theta;y)) - \log(f(\theta;\bar{y}))|^2 d\,\theta\bigg)^{1/2}\Bigg\}. \tag{81}$$

where f is defined in (61) *as a function in the logspline family.*

Definition 50. *Given $n \geq 1$ and $0 < \epsilon$ we define $E_{\epsilon,n}$ to be the subset of $\mathcal{F} = \{f(\cdot\ ; y) : y \in Y_0\}$ such that*

$$E_{\epsilon,n} = \left\{ f(\cdot\ ; y) : y \in Y_0 \text{ and } \bigg(\int |\log(f(\theta;y)) - \log(f(\theta;\bar{y}))|^2 d\,\theta\bigg)^{1/2} \leq n^{\epsilon}\sqrt{\frac{L+1}{n}} \right\}. \tag{82}$$

Lemma 51 (Stone[4][p.728])**.** *For each $y_1, y_2 \in Y_0$ and $\omega \in \Omega$ we have*

$$|l(y_1;\Theta_n(\omega)) - l(y_2;\Theta_n(\omega)) - (\lambda_n(y_1) - \lambda_n(y_2))| \leq 2n\|\log f(\cdot\ ; y_1) - \log f(\cdot\ ; y_2)\|_\infty\,. \tag{83}$$

Lemma 52 (Stone[4][p.729])**.** *Let $n \geq 1$. Given $\epsilon > 0$ and $\delta > 0$, there exists an integer $N = N(n) > 0$ and sets $E_j \subset \mathcal{F}$, $j = 1, \ldots, N$ satisfying*

$$\sup_{f_1,f_2\in E_j} \|\log(f_1) - \log(f_2)\|_\infty \leq \delta n^{2\epsilon-1}(L+1)$$

such that $E_{\varepsilon,n} \subset \bigcup_{i=1}^N E_i$.

Combining the above lemmas it leads to the following theorem, which is a result outlined in lemmas 5 and 8 found in Stone[4].

Theorem 53. *Given $D > 0$ and $\epsilon > 0$, let $b_n = n^{\epsilon}\sqrt{\dfrac{L(n)+1}{n}}$, $n \geq 1$, and $0 < \epsilon < \frac{1}{2}$ and $\beta = \epsilon$ in* (**??**)*. There exists $N = N(D)$ such that for all $n > N$*

$$A_{n,b_n}(y) \subset \Omega_n \quad \text{for each } y \in Y_0 \tag{84}$$

and thus

$$\mathbb{P}(\Omega_n^c) \leq \mathbb{P}\big(A_{n,b_n}^c(y)\big) \leq 2e^{-n^{2\epsilon}(L+1)\delta(D)}\,. \tag{85}$$

Remark 54. *From* (85) *we can see that as number of samples goes to infinity, we have that*

$$\mathbb{P}(\Omega_n) \to 1 \text{ as } n \to \infty.$$

Remark 55. *The bound* (85) *presented in Theorem 53 is a consequence of Hoeffdings inequality which states that for any $t > 0$*

$$\mathbb{P}\bigg(\Big|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_1\Big| \geq t\bigg) \leq 2\exp\bigg(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\bigg)$$

where $X_1, \ldots, X_n$ are identically distributed independent random variables with $\mathbb{P}(X_1 \in [a_i, b_i]) = 1$. To get the bound (85) *one needs to choose*

$$t = b\bigg(\int |\log(f(\theta;y)) - \log(f(\theta;\bar{y}))|^2\, d\theta\bigg)^{\frac{1}{2}}.$$

Now that we have defined the set where $\hat{y}$ exists and showed that the probability of its complement vanishes as $n \to \infty$ with a specific exponential rate, we will now state certain rates of convergence that only apply on Ω_n. The following theorem contains results of Theorem 2 and Lemma 12 of Stone[4].

Theorem 56. *There exist constants M_1, M_2, M_3 and M_4 such that for all $\omega \in \Omega_n$*

$$\begin{aligned}
|\hat{y}(\theta_1(\omega), \ldots, \theta_n(\omega)) - \bar{y}| &\leq \frac{M_1(L+1)}{\sqrt{n}} \\
\|\hat{p}(\cdot, \omega) - \bar{p}(\cdot)\|_2 &\leq M_3 \sqrt{\frac{L+1}{n}} \\
\|\log(\hat{p}(\cdot, \omega)) - \log(\bar{p}(\cdot))\|_\infty &\leq \frac{M_4(L+1)}{\sqrt{n}}.
\end{aligned} \tag{86}$$

6.6 Lagrange interpolation

The following two theorems are well-known facts which we cite from [7, p.132, p.134].

Theorem 57. *Let $f : [a, b] \to \mathbb{R}$. Given distinct points $a = x_0 < x_1 < ... < x_l = b$ and $l + 1$ ordinates $y_i = f(x_i)$, $i = 0, \ldots, l$ there exists an interpolating polynomial $q(x)$ of degree at most l such that $f(x_i) = q(x_i)$, $i = 0, \ldots, l$. This polynomial $q(x)$ is unique among the set of all polynomials of degree at most l. Moreover, $q(x)$ is called the Lagrange interpolating polynomial of f and can be written in the explicit form*

$$q(x) = \sum_{i=0}^{l} y_i l_i(x) \quad \text{with} \quad l_i(x) = \prod_{j \neq i} \left(\frac{x - x_j}{x_i - x_j} \right), \; i = 0, 1, \ldots, l. \tag{87}$$

Theorem 58. *Suppose that $f : [a, b] \to \mathbb{R}$ has $l + 1$ continuous derivatives on (a, b). Let $a = x_0 < x_1 < ... < x_l = b$ and $y_i = f(x_i)$, $i = 0, \ldots, l$. Let $q(x)$ be the Lagrange interpolating polynomial of f given by formula (87). Then for every $x \in [a, b]$ there exists $\xi \in (a, b)$ such that*

$$f(x) - q(x) = \frac{\prod_{i=0}^{l}(x - x_i)}{(l+1)!} f^{(l+1)}(\xi). \tag{88}$$

We next prove an elementary lemma that provides the estimate of the interpolation error when information on the derivatives of f is available. This lemma is used later in Theorem 18 to compute the mean integrated squared error.

Lemma 59. *Let $f(x)$, $q(x)$, and (x_i, y_i), $i = 0, \ldots, l$, with $l \geq 1$, be as in Theorem 58. Suppose that*

$$\sup_{x \in [a,b]} |f^{(l+1)}(x)| \leq C$$

for some constant $C \geq 0$ and $x_{i+1} - x_i = \dfrac{b-a}{l} =: \Delta x$ for each $i = 0, \ldots, l-1$. Then

$$\max_{x \in [a,b]} |f(x) - q(x)| \leq C \frac{(\Delta x)^{l+1}}{4(l+1)}. \tag{89}$$

Proof. Let $x \in [a, b]$. Then $x \in [x_j, x_{j+1}]$ for some $j \in \{0, \ldots, l-1\}$. Observe that

$$|(x - x_j)(x - x_{j+1})| \leq \frac{1}{4}(\Delta x)^2$$

and for $m \in \{-j, -j+1, \ldots, -1\} \cup \{2, \ldots, l-j\}$ we have $|x - x_{j+m}| \leq (\Delta x)|m|$. From this it follows that

$$\prod_{i=0}^{l} |x - x_i| \leq \frac{(\Delta x)^{(l+1)}}{4} j!(l-j)! \leq \frac{(\Delta x)^{(l+1)} l!}{4}.$$

Then Theorem 58 together with the above estimate implies (89). ☐

References

[1] C. de Boor, *A Practical Guide to Splines (Revised Edition)*, Springer, New York, 2001.

[2] C. J.Stone and C.-Y. Koo, *Logspline Density Estimation*, in Contemporary Mathematics, Volume 59, 1986.

[3] C. J. Stone, *Uniform Error Bounds Involving Logspline Models*, in Probability, Statistics and Mathematics: Papers in honor of Samuel Karlin, 335-355, 1989.

[4] C. J. Stone, *Large-sample Inference for Log-spline Models*, in Annals of Statistics, Vol. 18, No. 2, 717-741, 1990.

[5] W. Neiswanger, C. Wang and E. P. Xing, *Asymptotically Exact, Embarrassingly Parallel MCMC*, Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence. 2014; pp. 623-632.

[6] A. Miroshnikov, Z. Wei and E. Conlon, *Parallel Markov Chain Monte Carlo for Non-Gaussian Posterior Distributions*, Stat. Accepted (2015).

[7] K. E. Atkinson, *An Introduction to Numerical Analysis, 2nd Edition*, John Wiley & Sons, 1989.

[8] J. Langford, A.J. Smola and M. Zinkevich, Slow learners are fast. In: Bengio Y, Schuurmans D, J.D. Lafferty, C.K.I. Williams, A. Culotta, *Advances in Neural Information Processing Systems* (2009), 22 (NIPS), New York: Curran Associates, Inc.

[9] D. Newman, A. Asuncion, P. Smyth and M. Welling, *Distributed algorithms for topic models.* J Machine Learn Res (2009), 10, 1801-1828.

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

[11] M. Plummer, *JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling*, Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March 20-22, Vienna, Austria (2003).

[12] United States Department of Transportation, Bureau of Transportation Statistics, URL `https://www.transtats.bts.gov/`