New Insights Into Approximate Bayesian Computation

Gérard Biau[1]
Université Pierre et Marie Curie[2] *& Ecole Normale Supérieure*[3]*, France*
gerard.biau@upmc.fr

Frédéric Cérou
INRIA Rennes – Bretagne Atlantique, France
Frederic.Cerou@inria.fr

Arnaud Guyader
Université Rennes 2 & INRIA Rennes – Bretagne Atlantique, France
arnaud.guyader@uhb.fr

Abstract

Approximate Bayesian Computation (ABC for short) is a family of computational techniques which offer an almost automated solution in situations where evaluation of the posterior likelihood is computationally prohibitive, or whenever suitable likelihoods are not available. In the present paper, we analyze the procedure from the point of view of k-nearest neighbor theory and explore the statistical properties of its outputs. We discuss in particular some asymptotic features of the genuine conditional density estimate associated with ABC, which is an interesting hybrid between a k-nearest neighbor and a kernel method.

Index Terms — Approximate Bayesian Computation, Nonparametric estimation, Conditional density estimation, Nearest neighbor methods, Mathematical statistics.

2010 Mathematics Subject Classification: 62C10, 62F15, 62G20.

[1] Corresponding author.

[2] Research partially supported by the French National Research Agency (grant ANR-09-BLAN-0051-02 "CLARA") and by the Institut universitaire de France.

[3] Research carried out within the INRIA project "CLASSIC" hosted by Ecole Normale Supérieure and CNRS.

Résumé

Le terme anglais "Approximate Bayesian Computation" (ABC en abrégé) désigne une famille de techniques bayésiennes ayant pour objet la simulation selon une loi de probabilité lorsque la vraisemblance a posteriori n'est pas disponible ou s'avère impossible à évaluer numériquement. Dans le présent article, nous envisageons cette procédure du point de vue de la théorie des k-plus proches voisins, en nous attachant plus particulièrement à examiner les propriétés statistiques des sorties de l'algorithme. Cela nous conduit à analyser le comportement asymptotique d'un estimateur de la densité conditionnelle naturellement associé à ABC, utilisé en pratique et possédant à la fois les caractéristiques d'un estimateur des k-plus proches voisins et celles d'une méthode à noyau.

Mots-clés — Approximate Bayesian Computation, Estimation non paramétrique, Estimation de la densité conditionnelle, Méthodes de plus proches voisins, Statistique mathématique.

Classification par Sujets Mathématiques 2010 : 62C10, 62F15, 62G20.

1 Introduction

Let $\boldsymbol{Y}$ be a generic random observation which may, for example, take the form of a sample of independent and identically distributed (i.i.d.) random variables. More generally, it may also be the first observations of a time series or a more complex random object, such as a DNA sequence. We denote by $\ell(\boldsymbol{y}|\boldsymbol{\theta})$ the distribution (likelihood) of $\boldsymbol{Y}$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ is an unknown parameter that we wish to estimate. In the Bayesian paradigm, the parameter itself is seen as a random variable $\boldsymbol{\Theta}$, and the likelihood $\ell(\boldsymbol{y}|\boldsymbol{\theta})$ becomes the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{\Theta} = \boldsymbol{\theta}$. The distribution $\pi(\boldsymbol{\theta})$ of $\boldsymbol{\Theta}$ is called the prior distribution, while the distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ of $\boldsymbol{\Theta}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is termed posterior.

When taking a Bayesian perspective, inference about the parameter $\boldsymbol{\Theta}$ typically proceeds via calculation or simulation of the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. A variety of methods exist for inference in this context, such as rejection algorithms (Ripley, 1982), Markov Chain Monte Carlo (MCMC) methods (e.g., the Metropolis-Hastings algorithm, Metropolis et al., 1953; Hastings, 1970), and Importance Sampling (Ripley, 1982). For a comprehensive introduction to the domain, the reader is referred to the monographs by Robert and Casella (2004) and Marin and Robert (2007). However, in some contexts, computation of the posterior is problematic, either because the size of the data makes the calculation computationally intractable, or because calculation is impossible when using realistic models for how the data arises. Thus, despite their power and flexibility, MCMC procedures and their variants may prove irrelevant in a growing number of contemporary applications involving very large dimensions or complicated models. This computational burden typically arises in fields such as ecology, population genetics and image analysis, just to name a few.

This difficulty has motivated a drive to more approximate approaches, in particular the field of Approximate Bayesian Computation (ABC for short). In a nutshell, ABC is a family of computational techniques which offer an almost automated solution in situations where evaluation of the likelihood is computationally prohibitive, or whenever suitable likelihoods are not available. The approach was originally mentioned, but not analyzed, by Rubin (1984). It was further developed in population genetics by Fu and Li (1997); Tavaré et al. (1997); Pritchard et al. (1999); Beaumont et al. (2002), who gave the name of Approximate Bayesian Computation to a family of likelihood-free inference methods. Since its original developments, the ABC paradigm has successfully been applied to various scientific areas, ranging from archaeological science and ecology to epidemiology, stereology and protein network analysis. There are too many references to be included here, but the recent survey by Marin et al. (2012) offers both a historical and technical review of the domain.

Before we go into more details on ABC, some more notation is required. We assume to be given a statistic $\mathbf{S}$, taking values in $\mathbb{R}^m$. It is a function of the original observation $\boldsymbol{Y}$, with a dimension m typically much smaller than the dimension of $\boldsymbol{Y}$. The statistic $\mathbf{S}$ is supposed to admit a conditional density $f(\mathbf{s}|\boldsymbol{\theta})$ with respect to the Lebesgue measure on $\mathbb{R}^m$. Note that, strictly speaking, we should write $\mathbf{S}(\boldsymbol{Y})$ instead of $\mathbf{S}$. However, since there is no ambiguity, we continue to use the latter notation. As such, the statistic $\mathbf{S}$ should be understood as a low-dimensional summary of $\boldsymbol{Y}$. It can be, for example, a sufficient statistic for the parameter $\boldsymbol{\Theta}$, but not necessarily. Assuming that $\boldsymbol{\Theta}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$, the conditional distribution of $\boldsymbol{\Theta}$ given $\mathbf{S} = \mathbf{s}$ has a density $g(\boldsymbol{\theta}|\mathbf{s})$ which, according to Bayes' rule, takes the form

$$g(\boldsymbol{\theta}|\mathbf{s}) = \frac{f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\bar{f}(\mathbf{s})}, \quad \text{where } \bar{f}(\mathbf{s}) = \int_{\mathbb{R}^p} f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

is the marginal density of $\mathbf{S}$. Finally, we denote by $\boldsymbol{y}_0$ the observed realization of $\boldsymbol{Y}$ (i.e., the data set), and let $\mathbf{s}_0(= \mathbf{s}(\boldsymbol{y}_0))$ be the corresponding realization of $\mathbf{S}$. Throughout the document, both $\boldsymbol{y}_0$ and $\mathbf{s}_0$ should be considered as fixed quantities.

In its most common form, the generic ABC algorithm is framed as follows:

Algorithm 1 Pseudo-code 1 of a generic ABC algorithm

```
Require: A positive integer $N$ and a tolerance level $\varepsilon$.
  for $i = 1$ to $N$ do
    Generate $\boldsymbol{\theta}_i$ from the prior $\pi(\boldsymbol{\theta})$;
    Generate $\boldsymbol{y}_i$ from the likelihood $\ell(.|\boldsymbol{\theta}_i)$.
  end for
  return The $\boldsymbol{\theta}_i$'s such that $\|\mathbf{s}(\boldsymbol{y}_i) - \mathbf{s}_0\| \leq \varepsilon$.
```

The basic idea behind this formulation is that using a representative enough summary statistic $\mathbf{S}$ coupled with a small enough tolerance level ε should produce a good approximation of the posterior distribution. A moment's thought reveals that pseudo-code 1 has the flavor of a nonparametric kernel conditional density estimation procedure, for which ε plays the role of a bandwidth. This is, for example, the point of view that prevails in the analysis of Blum (2010), who explores the asymptotic bias and variance of kernel-type estimates of the posterior density $g(.|\mathbf{s}_0)$ evaluated over the code outputs.

However, as made transparent by Marin et al. (2012), pseudo-code 1, despite its widespread diffusion, does not exactly match what people do in practice. A more accurate formulation is the following one:

Algorithm 2 Pseudo-code 2 of a generic ABC algorithm

```
Require: A positive integer $N$ and an integer $k_N$ between 1 and $N$.
  for $i = 1$ to $N$ do
    Generate $\boldsymbol{\theta}_i$ from the prior $\pi(\boldsymbol{\theta})$;
    Generate $\boldsymbol{y}_i$ from the likelihood $\ell(.|\boldsymbol{\theta}_i)$.
  end for
  return The $\boldsymbol{\theta}_i$'s such that $\mathbf{s}(\boldsymbol{y}_i)$ is among the $k_N$-nearest neighbors of $\mathbf{s}_0$.
```

Algorithm 1 and Algorithm 2 are dual, in the sense that the number of accepted points is fixed in the second and random in the first, while their range is random in the second and fixed in the first. In practice, the parameter N is chosen to be very large (typically of the order of 10^6), while k_N is most commonly expressed as a percentile. Thus, for example, the choice $N = 10^6$ and a percentile $k_N/N = 0.1\%$ allow to retain 1000 simulated $\boldsymbol{\theta}_i$'s.

From a nonparametric perspective, pseudo-code 2 falls within the broad family of nearest neighbor-type procedures (Fix and Hodges, 1951; Loftsgaarden and Quesenberry, 1965; Cover, 1968). Such procedures have the favor of

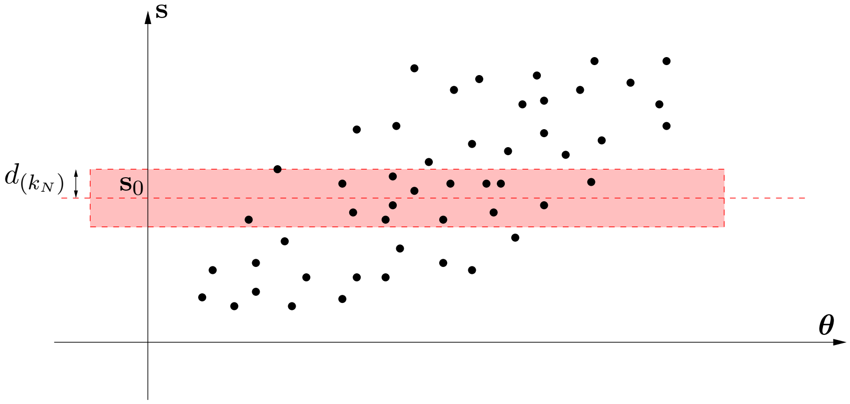

Figure 1: Illustration of ABC in dimension $m = p = 1$ ($d_{(k_N)} = \|\mathbf{S}_{(k_N)} - \mathbf{s}_0\|$).

practitioners, because they are fast, easy to compute and flexible. For implementation, they require only a measure of distance in the sample space, hence their popularity as a starting-point for refinement, improvement and adaptation to new settings (see, e.g., Devroye et al., 1996, Chapter 19). In any case, it is our belief that ABC should be analyzed in this context, and this is the point of view that is taken in the present article.

In order to better understand the rationale behind Algorithm 2, denote by $(\boldsymbol{\Theta}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{\Theta}_N, \boldsymbol{Y}_N)$ an i.i.d. sample, with common joint distribution $\ell(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. This sample is naturally associated with the i.i.d. sequence $(\boldsymbol{\Theta}_1, \mathbf{S}_1), \ldots, (\boldsymbol{\Theta}_N, \mathbf{S}_N)$, where each pair has density $f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Finally, let $\mathbf{S}_{(1)}, \ldots, \mathbf{S}_{(k_N)}$ be the k_N-nearest neighbors of $\mathbf{s}_0$ among $\mathbf{S}_1, \ldots, \mathbf{S}_N$, and let $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k_N)}$ be the corresponding $\boldsymbol{\Theta}_i$'s (see Figure 1 for an illustration in dimension $m = p = 1$).

With this notation, we see that the generic ABC Algorithm 2 proceeds in two steps:

1. First, simulate (realizations of) an N-sample $(\boldsymbol{\Theta}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{\Theta}_N, \boldsymbol{Y}_N)$;

2. Seconds, return (realizations of) the variables $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k_N)}$.

This simple observation opens the way to a mathematical analysis of ABC via techniques based on nearest neighbors. In fact, despite a growing number of practical applications, theoretical results guaranteeing the validity of the approach are still lacking (see Wilkinson, 2008; Blum, 2010; Fearnhead and Prangle, 2012, for results in this direction). Our present contribution is twofold:

(i) We offer in Section 2 an explicit result regarding the distribution of the algorithm outputs $(\mathbf{\Theta}_{(1)}, \mathbf{S}_{(1)}), \ldots, (\mathbf{\Theta}_{(k_N)}, \mathbf{S}_{(k_N)})$. Let $\mathcal{B}_m(\mathbf{s}_0, \delta)$ denote the closed ball in $\mathbb{R}^m$ centered at $\mathbf{s}_0$ with nonnegative radius δ, i.e., $\mathcal{B}_m(\mathbf{s}_0, \delta) = \{\mathbf{s} \in \mathbb{R}^m : \|\mathbf{s} - \mathbf{s}_0\| \leq \delta\}$. In a nutshell, Proposition 2.1 reveals that, conditionally on the distance $d_{(k_N+1)} = \|\mathbf{S}_{(k_N+1)} - \mathbf{s}_0\|$, the simulated data set may be regarded as k_N i.i.d. realizations of the joint density of $(\mathbf{\Theta}, \mathbf{S})$ restricted to the cylinder $\mathbb{R}^p \times \mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})$. This result is important since it gives a precise description of the output distribution of ABC Algorithm 2.

(ii) For a fixed $\mathbf{s}_0 \in \mathbb{R}^m$, the estimate practitioners use most to infer the posterior density $g(.|\mathbf{s}_0)$ at some point $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is

$$\hat{g}_{N,\mathbf{s}_0}(\boldsymbol{\theta}_0) = \frac{1}{k_N h_N^p} \sum_{j=1}^{k_N} K\left(\frac{\boldsymbol{\theta}_0 - \mathbf{\Theta}_{(j)}}{h_N}\right), \qquad (1.1)$$

where $\{h_N\}$ is a sequence of positive real numbers (bandwidth) and K is a nonnegative Borel measurable function (kernel) on $\mathbb{R}^p$. The idea is simple: In order to estimate the posterior, just look at the k_N-nearest neighbors of $\mathbf{s}_0$ and smooth the corresponding $\mathbf{\Theta}_j$'s around $\boldsymbol{\theta}_0$. It should be noted that (1.1) is a smart hybrid between a k-nearest neighbor and a kernel density estimation procedure. It is different from the Rosenblatt-type (Rosenblatt, 1969) kernel conditional density estimates proposed in Beaumont et al. (2002) and further explored by Blum (2010). In Section 3 and Section 4, we establish some consistency properties of this genuine estimate and discuss its rates of convergence.

For the sake of clarity, proofs are postponed to Section 5 and Section 6. An appendix at the end of the paper offers some new results on convolution and approximation of the identity.

To conclude this introduction, we would like to make a few comments on the topics that will **not** be addressed in the present document. An important part of the performance of the ABC approach, especially for high-dimensional data sets, relies upon a good choice of the summary statistic $\mathbf{S}$. In many practical applications, this statistic is picked by an expert in the field, without any particular guarantee of success. A systematic approach to choosing such a statistic, based upon a sound theoretical framework, is currently under active investigation in the Bayesian community. This important issue will not be pursued further here. As a good starting point, the interested reader is referred to Joyce and Marjoran (2008), who develop a sequential scheme for scoring statistics according to whether their inclusion in the analysis will substantially improve the quality of inference. Similarly, we will not address issues regarding how to enhance efficiency of ABC and its variants, as for example with the sequential techniques of Sisson et al. (2007) and Beaumont et al. (2009). Nor won't we explore the important question of ABC model choice, for which theoretical arguments are still missing (Robert et al., 2011; Marin et al., 2011).

2 Distribution of ABC outputs

We continue to use the notation of Section 1 and recall in particular that $(\mathbf{\Theta}_1, \mathbf{S}_1), \ldots, (\mathbf{\Theta}_N, \mathbf{S}_N)$ are i.i.d. $\mathbb{R}^p \times \mathbb{R}^m$-valued random variables, with common probability density $f(\boldsymbol{\theta}, \mathbf{s}) = f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Both $\mathbb{R}^p$ (the space of $\mathbf{\Theta}_i$'s) and $\mathbb{R}^m$ (the space of $\mathbf{S}_i$'s) are equipped with the Euclidean norm $\|.\|$. In this section, attention is focused on analyzing the distribution of the algorithm outputs $(\mathbf{\Theta}_{(1)}, \mathbf{S}_{(1)}), \ldots, (\mathbf{\Theta}_{(k_N)}, \mathbf{S}_{(k_N)})$.

In what follows, we keep $\mathbf{s}_0$ fixed and denote by d_i the (random) distance between $\mathbf{s}_0$ and $\mathbf{S}_i$. (To be rigorous, we should write $d_i(\mathbf{s}_0)$, but since no confusion can arise we write it simply d_i.) Similarly, we let $d_{(i)}$ be the distance between $\mathbf{s}_0$ and its i-th nearest neighbor among $\mathbf{S}_1, \ldots, \mathbf{S}_N$, that is

$$d_{(i)} = \|\mathbf{S}_{(i)} - \mathbf{s}_0\|.$$

(If distance ties occur, a tie-breaking strategy must be defined. For example, if $\|\mathbf{S}_i - \mathbf{s}_0\| = \|\mathbf{S}_j - \mathbf{s}_0\|$, $\mathbf{S}_i$ may be declared "closer" if $i < j$, i.e., the tie-breaking is done by indices. Note however that ties occur with probability 0 since all random variables are absolutely continuous.) It is assumed throughout the paper that $N \geq 2$ and $1 \leq k_N \leq N - 1$.

Rearranging the k_N (ordered) statistics $(\mathbf{\Theta}_{(1)}, \mathbf{S}_{(1)}), \ldots, (\mathbf{\Theta}_{(k_N)}, \mathbf{S}_{(k_N)})$ in the original order of their outcome, one obtains the k_N (non-ordered) random variables $(\mathbf{\Theta}_1^\star, \mathbf{S}_1^\star), \ldots, (\mathbf{\Theta}_{k_N}^\star, \mathbf{S}_{k_N}^\star)$. Our first result is concerned with the conditional distributions

$$\mathcal{L}\left\{(\mathbf{\Theta}_1^\star, \mathbf{S}_1^\star), \ldots, (\mathbf{\Theta}_{k_N}^\star, \mathbf{S}_{k_N}^\star) \,|\, d_{(k_N+1)}\right\}$$

and

$$\mathcal{L}\left\{(\mathbf{\Theta}_{(1)}, \mathbf{S}_{(1)}), \ldots, (\mathbf{\Theta}_{(k_N)}, \mathbf{S}_{(k_N)}) \,|\, d_{(k_N+1)}\right\}.$$

Recall that the collection of all $\mathbf{s}_0 \in \mathbb{R}^m$ with $\int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} > 0$ for all $\delta > 0$ is called the support of $\bar{f}$.

Proposition 2.1 (Distribution of ABC outputs) *Assume that* $\mathbf{s}_0$ *belongs to the support of* $\bar{f}$. *Let* $(\tilde{\mathbf{\Theta}}_1, \tilde{\mathbf{S}}_1), \ldots, (\tilde{\mathbf{\Theta}}_{k_N}, \tilde{\mathbf{S}}_{k_N})$ *be i.i.d. random variables, with common probability density (conditional on* $d_{(k_N+1)}$*)*

$$\frac{\mathbf{1}_{[\|\mathbf{s}-\mathbf{s}_0\| \leq d_{(k_N+1)}]} f(\boldsymbol{\theta}, \mathbf{s})}{\displaystyle\int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{s}}. \tag{2.1}$$

Then

$$\mathcal{L}\left\{(\mathbf{\Theta}_1^\star, \mathbf{S}_1^\star), \ldots, (\mathbf{\Theta}_{k_N}^\star, \mathbf{S}_{k_N}^\star) \mid d_{(k_N+1)}\right\} = \mathcal{L}\left\{(\tilde{\mathbf{\Theta}}_1, \tilde{\mathbf{S}}_1), \ldots, (\tilde{\mathbf{\Theta}}_{k_N}, \tilde{\mathbf{S}}_{k_N})\right\}.$$

Moreover

$$\begin{aligned}&\mathcal{L}\left\{(\mathbf{\Theta}_{(1)}, \mathbf{S}_{(1)}), \ldots, (\mathbf{\Theta}_{(k_N)}, \mathbf{S}_{(k_N)}) \mid d_{(k_N+1)}\right\} \\ &\quad = \mathcal{L}\left\{(\tilde{\mathbf{\Theta}}_{(1)}, \tilde{\mathbf{S}}_{(1)}), \ldots, (\tilde{\mathbf{\Theta}}_{(k_N)}, \tilde{\mathbf{S}}_{(k_N)})\right\}.\end{aligned}$$

Note, since $\mathbf{s}_0$ belongs by assumption to the support of $\bar{f}$, that the normalizing constant in the denominator of (2.1) is positive. This theorem may be regarded as an extension of a result of Kaufmann and Reiss (1992), who provide explicit representations of the conditional distribution of an empirical point process given some order statistics. However, the present Bayesian setting is not covered by the conclusions of Kaufmann and Reiss (1992), and our proof actually relies on simpler arguments.

The main message of Proposition 2.1 is that, **conditionally on** $d_{(k_N+1)}$, one can consider the k_N-tuple $(\mathbf{\Theta}_{(1)}, \mathbf{S}_{(1)}), \ldots, (\mathbf{\Theta}_{(k_N)}, \mathbf{S}_{(k_N)})$ as an ordered sample drawn according to the probability density (2.1). Alternatively, the (unordered) simulated values may be treated like i.i.d. realizations of variables with common density proportional to $\mathbf{1}_{[\|\mathbf{s}-\mathbf{s}_0\| \leq d_{(k_N+1)}]} f(\boldsymbol{\theta}, \mathbf{s})$. Conditionally on $d_{(k_N+1)}$, the accepted $\boldsymbol{\theta}_j$'s are nothing but i.i.d. realizations of the probability density

$$\frac{\displaystyle\int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\mathbf{s}}{\displaystyle\int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{s}}.$$

Although this conclusion is intuitively clear, its proof requires a careful mathematical analysis.

As will be made transparent in the next section, Proposition 2.1 plays a key role in the mathematical analysis of the natural conditional density estimate associated with ABC methodology. In fact, investigating ABC in terms of nearest neighbors has other important consequences. Suppose, for example, that we are interested in estimating some finite conditional expectation $\mathbb{E}[\varphi(\boldsymbol{\Theta})|\mathbf{S} = \mathbf{s}_0]$, where the random variable $\varphi(\boldsymbol{\Theta})$ is bounded. This includes in particular the important setting where φ is polynomial and one wishes to estimate the conditional moments of $\boldsymbol{\Theta}$. Then, provided $k_N/\log\log N \to \infty$ and $k_N/N \to 0$ as $N \to \infty$, it can be shown that for almost all $\mathbf{s}_0$ (with respect to the distribution of $\mathbf{S}$), with probability 1,

$$\frac{1}{k_N}\sum_{j=1}^{k_N} \varphi\left(\boldsymbol{\Theta}_{(j)}\right) \to \mathbb{E}[\varphi(\boldsymbol{\Theta})|\mathbf{S} = \mathbf{s}_0]. \tag{2.2}$$

Proof of such a result uses the full power of the vast and rich nearest neighbor estimation theory. To be more precise, let us make a quick detour through this theory and consider an i.i.d. sample $(\mathbf{X}_1, Z_1), \ldots, (\mathbf{X}_N, Z_N)$ taking values in $\mathbb{R}^m \times \mathbb{R}$, where the output variables Z_i's are bounded. Assume, to keep things simple, that the $\mathbf{X}_i$'s have a probability density and that our goal is to assess the regression function $r(\mathbf{x}) = \mathbb{E}[Z \mid \mathbf{X} = \mathbf{x}]$, $\mathbf{x} \in \mathbb{R}^m$. In this context, the k-nearest neighbor regression function estimate of r (Royall, 1966; Cover, 1968; Stone, 1977) takes the form

$$\hat{r}_N(\mathbf{x}) = \frac{1}{k_N}\sum_{j=1}^{k_N} Z_{(j)}, \quad \mathbf{x} \in \mathbb{R}^m,$$

where $Z_{(j)}$ is the Z-observation corresponding to $\mathbf{X}_{(j)}$, the j-th-closest point to $\mathbf{x}$ among $\mathbf{X}_1, \ldots, \mathbf{X}_N$. Denoting by μ the distribution of $\mathbf{X}_1$, it is proved in Theorem 3 of Devroye (1982) that provided $k_N/\log\log N \to \infty$ and $k_N/N \to 0$, for μ-almost all $\mathbf{x}$,

$$\hat{r}_N(\mathbf{x}) \to r(\mathbf{x}) \quad \text{with probability 1 as } N \to \infty.$$

This result can be transposed without further effort to our ABC setting via the correspondence $\varphi(\boldsymbol{\Theta}) \leftrightarrow Z$ and $\mathbf{S} \leftrightarrow \mathbf{X}$, thereby establishing validity of (2.2). The decisive step towards that conclusion is accomplished by making a connection between ABC and nearest neighbor methodology. We leave it to the reader to draw his own conclusions as to further possible utilizations of this correspondence.

3 Mean square error consistency

As in Section 2, we keep the conditioning vector $\mathbf{s}_0$ fixed and consider the i.i.d. sample $(\boldsymbol{\Theta}_1, \mathbf{S}_1), \ldots, (\boldsymbol{\Theta}_N, \mathbf{S}_N)$, where each pair is distributed according to the probability density $f(\boldsymbol{\theta}, \mathbf{s}) = f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ on $\mathbb{R}^p \times \mathbb{R}^m$. Based on this sample, our new objective is to estimate the posterior density $g(\boldsymbol{\theta}_0|\mathbf{s}_0)$, $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. This estimation step is an important ingredient of the Bayesian analysis, whether this may be for visualization purposes or more involved mathematical achievements.

As exposed in the introduction, the natural ABC-companion estimate of $g(\boldsymbol{\theta}_0|\mathbf{s}_0)$ takes the form

$$\hat{g}_N(\boldsymbol{\theta}_0) = \frac{1}{k_N h_N^p} \sum_{j=1}^{k_N} K\left(\frac{\boldsymbol{\theta}_0 - \boldsymbol{\Theta}_{(j)}}{h_N}\right), \quad \boldsymbol{\theta}_0 \in \mathbb{R}^p, \tag{3.1}$$

where $\{h_N\}$ is a sequence of positive real numbers (bandwidth) and K is a nonnegative Borel measurable function (kernel) on $\mathbb{R}^p$. (To reduce the notational burden, we dropped the dependency of the estimate upon $\mathbf{s}_0$, keeping in mind that $\mathbf{s}_0$ is held fixed.) Kernel estimates were originally studied in density estimation by Rosenblatt (1969) and Parzen (1962), and were latter introduced in regression estimation by Nadaraya (1964, 1965) and Watson (1964). The origins of k-nearest neighbor density estimation go back to Fix and Hodges (1951) and Loftsgaarden and Quesenberry (1965). Kernel estimates have been extended to the conditional density setting by Rosenblatt (1969), who proceeds by separately inferring the bivariate density $f(\boldsymbol{\theta}, \mathbf{s})$ of $(\boldsymbol{\Theta}, \mathbf{S})$ and the marginal density of $\mathbf{S}$. Rosenblatt's estimate reads

$$\tilde{g}_N(\boldsymbol{\theta}_0) = \frac{\sum_{i=1}^N L\left(\frac{\mathbf{s}_0 - \mathbf{S}_i}{\delta_N}\right) K\left(\frac{\boldsymbol{\theta}_0 - \boldsymbol{\Theta}_i}{h_N}\right)}{h_N^p \sum_{i=1}^N L\left(\frac{\mathbf{s}_0 - \mathbf{S}_i}{\delta_N}\right)},$$

where L is a kernel in $\mathbb{R}^m$, and δ_N is the corresponding bandwidth. ABC-compatible estimates of this type have been discussed in Beaumont et al. (2002) and further explored by Blum (2010) (additional references for the conditional density estimation problem are Hyndman et al., 1996; Györfi and Kohler, 2007; Faugeras, 2009, and the survey of Hansen, 2004).

The conditional density estimate we are interested in is different, in the sense that it has both the flavor of a k-nearest neighbor approach (it retains only the k_N-nearest neighbors of $\mathbf{s}_0$ among $\mathbf{S}_1, \ldots, \mathbf{S}_N$) and a kernel method (it smoothes the corresponding $\boldsymbol{\Theta}_j$'s). Obviously, the main advantage of (3.1) over its kernel-type competitors is its simplicity (it does not involve evaluation of a ratio, with a denominator that can be small), which makes it easy to implement.

A related procedure to density estimation has been originally proposed by Breiman et al. (1977), who suggested varying the kernel bandwidth with respect to the sample points. Various extensions and modifications of the Breiman et al. (1977) estimate have been later proposed in the literature. The rationale behind the approach is to combine the desirable smoothness properties of kernel estimates with the data-adaptive character of nearest neighbor procedures. Particularly influential papers in the study of variable kernel estimates were those of Abramson (1982) and Hall and Marron (1988), who showed how variable bandwidths with positive kernels can nevertheless induce convergence rates usually attainable with fixed bandwidths and fourth order kernels. For a complete and comprehensive description of variable kernel estimates and their properties, we refer the reader to Jones (1990).

Our goal in this section is to investigate some consistency properties of the ABC-companion estimate (3.1). Pointwise mean square error consistency is proved in Theorem 3.3 and mean integrated square error consistency is established in Theorem 3.4. We stress that this part of the document is concerned with minimal conditions of convergence. We did indeed try to reduce as much as possible the assumptions on the various unknown probability densities by resorting to real analysis arguments.

The following assumptions on the kernel will be needed throughout the paper:

Assumption [K1] The kernel K is nonnegative and belongs to $L^1(\mathbb{R}^p)$, with $\int_{\mathbb{R}^p} K(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1$. Moreover, the function $\sup_{\|\mathbf{y}\| \geq \|\boldsymbol{\theta}\|} |K(\mathbf{y})|$, $\boldsymbol{\theta} \in \mathbb{R}^p$, is in $L^1(\mathbb{R}^p)$.

Assumption set [**K1**] is in no way restrictive and is satisfied by all standard kernels such as, for example, the naive kernel

$$K(\boldsymbol{\theta}) = \frac{1}{V_p} \mathbf{1}_{\mathcal{B}_p(\mathbf{0},1)}(\boldsymbol{\theta}),$$

where V_p is the volume of the closed unit ball $\mathcal{B}_p(\mathbf{0},1)$ in $\mathbb{R}^p$, or the Gaussian kernel

$$K(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\|\boldsymbol{\theta}\|^2/2\right).$$

We recall for further references that, in the p-dimensional Euclidean space,

$$V_p = \frac{\pi^{p/2}}{\Gamma\left(1+\frac{p}{2}\right)}, \tag{3.2}$$

where $\Gamma(.)$ is the gamma function. Everywhere in the document, we denote by λ_p (respectively, λ_m) the Lebesgue measure on $\mathbb{R}^p$ (respectively, $\mathbb{R}^m$) and set, for any positive h,

$$K_h(\boldsymbol{\theta}) = \frac{1}{h^p} K(\boldsymbol{\theta}/h), \quad \boldsymbol{\theta} \in \mathbb{R}^p.$$

We note once and for all that, under Assumption [**K1**], $\int_{\mathbb{R}^p} K_h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1$.

The first crucial result from real analysis that is needed here is the so-called Lebesgue's differentiation theorem (see, e.g., Theorem 7.16 in Wheeden and Zygmund, 1977), which asserts that if φ is a locally integrable function in $\mathbb{R}^n$, then

$$\frac{1}{V_n \delta^n} \int_{\mathcal{B}_n(\mathbf{x}_0,\delta)} |\varphi(\mathbf{x}) - \varphi(\mathbf{x}_0)| \, \mathrm{d}\mathbf{x} \to 0 \quad \text{as } \delta \to 0$$

for λ_n-almost all $\mathbf{x}_0 \in \mathbb{R}^n$. A point $\mathbf{x}_0$ at which this statement is valid is called a Lebesgue point of φ. In the proofs, we shall in fact need some convolution-type variations around the Lebesgue's theorem regarding the prior density π. These important results are gathered in the next theorem, whose proof can be found in Theorem 1, page 5 and Theorem 2, pages 62-63 of Stein (1970).

Theorem 3.1 *Let K be a kernel satisfying Assumption* [**K1**]*, and let the function $\pi^\star$ be defined on $\mathbb{R}^p$ by*

$$\boldsymbol{\theta}_0 \mapsto \pi^\star(\boldsymbol{\theta}_0) = \sup_{h>0} \left[\int_{\mathbb{R}^p} K_h(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \right].$$

(*i*) *For λ_p-almost all $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, one has*

$$\int_{\mathbb{R}^p} K_h(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \to \pi(\boldsymbol{\theta}_0) \quad \text{as } h \to 0.$$

(*ii*) *The quantity $\pi^\star(\boldsymbol{\theta}_0)$ is finite for λ_p-almost all $\boldsymbol{\theta}_0 \in \mathbb{R}^p$.*

(*iii*) *For any $q > 1$, the function $\pi^\star$ is in $L^q(\mathbb{R}^p)$ whenever π is in $L^q(\mathbb{R}^p)$.*

When K is chosen to be the naive kernel, the function $\pi^\star$ of Theorem 3.1 is called the Hardy-Littlewood maximal function of π. It should be understood as a gauge of the size of the averages of π around $\boldsymbol{\theta}_0$.

We shall also need an equivalent of Theorem 3.1 for the joint density f, which this time is defined on $\mathbb{R}^p \times \mathbb{R}^m$. Things turn out to be slightly more complicated in this case if one is willing pairs of points $(\boldsymbol{\theta}_0, \mathbf{s}_0)$ to be approached as $(h, \delta) \to (0, 0)$ by general product kernels over $\mathbb{R}^p \times \mathbb{R}^m$. These kernels take the form $K_h(.) \otimes L_\delta(.)$, without any restriction on the joint behavior of h and δ (in particular, we do not impose that $h = \delta$). The so-called Jessen-Marcinkiewicz-Zygmund theorem (Jessen et al., 1935, see also Zygmund, 1959, Chapter 17, pages 305-309) answers the question for naive kernels, at the price of a slight integrability assumption on f. On the other hand, the literature offers surprisingly little help for general kernels, with the exception of arguments presented in Devroye and Krzyżak (2002). This is astonishing since this real analysis issue is at the basis of pointwise convergence properties of multivariate kernel estimates and indeed most density estimates. To fill the gap, we begin with the following theorem, which is tailored to our ABC context (that is, when the second kernel L is restricted to be the naive one). A more general result (that is, for both K and L general kernels) together with interesting new results on convolution and approximation of the identity are given in the Appendix section, at the end of the paper (Theorem 3.2 is thus a consequence of Theorem A.1). In the sequel, notation u^+ means $\max(u, 0)$.

Theorem 3.2 *Let K be a kernel satisfying Assumption* [**K1**]*, and let the function $f^\star$ be defined on $\mathbb{R}^p \times \mathbb{R}^m$ by*

$$(\boldsymbol{\theta}_0, \mathrm{s}_0) \mapsto f^\star(\boldsymbol{\theta}_0, \mathrm{s}_0) = \sup_{h>0,\delta>0} \left[\frac{1}{V_m \delta^m} \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathrm{s}_0,\delta)} K_h(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) f(\boldsymbol{\theta}, \mathrm{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathrm{s} \right].$$

(*i*) *If*

$$\int_{\mathbb{R}^p} \int_{\mathbb{R}^m} f(\boldsymbol{\theta}, \mathrm{s}) \log^+ f(\boldsymbol{\theta}, \mathrm{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathrm{s} < \infty \qquad (3.3)$$

then, for $\lambda_p \otimes \lambda_m$-almost all $(\boldsymbol{\theta}_0, \mathrm{s}_0) \in \mathbb{R}^p \times \mathbb{R}^m$,

$$\lim_{(h,\delta)\to(0,0)} \frac{1}{V_m \delta^m} \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathrm{s}_0,\delta)} K_h(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) f(\boldsymbol{\theta}, \mathrm{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathrm{s} = f(\boldsymbol{\theta}_0, \mathrm{s}_0).$$

(*ii*) *If condition (3.3) is satisfied, then $f^\star(\boldsymbol{\theta}_0, \mathrm{s}_0)$ is finite for $\lambda_p \otimes \lambda_m$-almost all $(\boldsymbol{\theta}_0, \mathrm{s}_0) \in \mathbb{R}^p \times \mathbb{R}^m$.*

(*iii*) *For any $q > 1$, the function $f^\star$ is in $L^q(\mathbb{R}^p \times \mathbb{R}^m)$ whenever f is in $L^q(\mathbb{R}^p \times \mathbb{R}^m)$.*

A remarkable feature of Theorem 3.2 (*i*) is that the result is true as soon as $(h, \delta) \to (0, 0)$, without any restriction on these parameters. This comes however at the price of the mild integrability assumption (3.3), which is true, in particular, if f is in any $L^q(\mathbb{R}^p \times \mathbb{R}^m)$, $q > 1$.

Recall that we denote by $\bar{f}$ the marginal density of $f(\boldsymbol{\theta}, \mathrm{s})$ in s, that is

$$\bar{f}(\mathrm{s}) = \int_{\mathbb{R}^p} f(\boldsymbol{\theta}, \mathrm{s}) \mathrm{d}\boldsymbol{\theta}, \quad \mathrm{s} \in \mathbb{R}^m.$$

We are now in a position to state the two main results of this section.

Theorem 3.3 (Pointwise mean square error consistency) *Assume that the kernel K is bounded and satisfies Assumption* [**K1**]. *Assume, in addition, that the joint probability density f is such that*

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^m} f(\boldsymbol{\theta},\mathrm{s})\log^+ f(\boldsymbol{\theta},\mathrm{s})\mathrm{d}\boldsymbol{\theta}\mathrm{ds} < \infty.$$

Then, for $\lambda_p \otimes \lambda_m$-almost all $(\boldsymbol{\theta}_0, \mathrm{s}_0) \in \mathbb{R}^p \times \mathbb{R}^m$, with $\bar{f}(\mathrm{s}_0) > 0$, if $k_N \to \infty$, $k_N/N \to 0$, $h_N \to 0$ and $k_N h_N^p \to \infty$,

$$\mathbb{E}\left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0|\mathrm{s}_0)\right]^2 \to 0 \quad \textit{as } N \to \infty.$$

It is stressed that the integral assumption required on f is mild. It is for example satisfied whenever f is bounded from above or whenever f belongs to $L^q(\mathbb{R}^p \times \mathbb{R}^m)$ with $q > 1$. There are, however, situations where this assumption is not satisfied. As an illustration, take $p = m = 1$ and let

$$\mathcal{T} = \left\{(\boldsymbol{\theta}, \mathrm{s}) \in \mathbb{R} \times \mathbb{R} : \boldsymbol{\theta} > 0, \mathrm{s} > 0, \boldsymbol{\theta} + \mathrm{s} \leq \frac{1}{2}\right\}.$$

Clearly,

$$\iint_{\mathcal{T}} \frac{1}{(\boldsymbol{\theta}+\mathrm{s})^2 \log^2(\boldsymbol{\theta}+\mathrm{s})}\mathrm{d}\boldsymbol{\theta}\mathrm{ds} < \infty.$$

Choose

$$f(\boldsymbol{\theta},\mathrm{s}) = \frac{C}{(\boldsymbol{\theta}+\mathrm{s})^2 \log^2(\boldsymbol{\theta}+\mathrm{s})}\mathbf{1}_{[(\boldsymbol{\theta},\mathrm{s})\in\mathcal{T}]},$$

where C is a normalizing constant ensuring that f is a probability density. Then

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^m} f(\boldsymbol{\theta},\mathrm{s})\mathrm{d}\boldsymbol{\theta}\mathrm{ds} = 1$$

whereas

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^m} f(\boldsymbol{\theta},\mathrm{s})\log^+ f(\boldsymbol{\theta},\mathrm{s})\mathrm{d}\boldsymbol{\theta}\mathrm{ds} = \infty.$$

Theorem 3.4 below states that the estimate $\hat{g}_N$ is also consistent with respect to the mean integrated square error criterion.

Theorem 3.4 (Mean integrated square error consistency) *Assume that the kernel K belongs to $L^2(\mathbb{R}^p)$ and satisfies Assumption* [**K1**]. *Assume, in addition, that the joint probability density f and the prior π are in $L^2(\mathbb{R}^p \times \mathbb{R}^m)$ and $L^2(\mathbb{R}^p)$, respectively. Then, for λ_m-almost all $\mathrm{s}_0 \in \mathbb{R}^m$, with $\bar{f}(\mathrm{s}_0) > 0$, if $k_N \to \infty$, $k_N/N \to 0$, $h_N \to 0$ and $k_N h_N^p \to \infty$,*

$$\mathbb{E}\left[\int_{\mathbb{R}^p}\left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0|\mathrm{s}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right] \to 0 \quad \textit{as } N \to \infty.$$

Here again, the regularity assumptions required on f and π are minimal. One could envisage an additional degree of smoothing in the estimate (3.1) by observing that taking the k_N nearest neighbors of $\mathbf{s}_0$ can be viewed as the uniform kernel case of the more general quantity

$$\sum_{i=1}^{N} L\left(\frac{\mathbf{s}_0 - \mathbf{S}_i}{\|\mathbf{S}_{(k_N)} - \mathbf{s}_0\|}\right),$$

which allows unequal weights to be given to the $\mathbf{S}_i$'s. The corresponding smoothed conditional density estimate is defined by

$$\tilde{g}_N(\boldsymbol{\theta}_0) = \frac{\sum_{i=1}^{N} L\left(\frac{\mathbf{s}_0-\mathbf{S}_i}{\|\mathbf{S}_{(k_N)}-\mathbf{s}_0\|}\right) K\left(\frac{\boldsymbol{\theta}_0-\boldsymbol{\Theta}_i}{h_N}\right)}{h_N^p \sum_{i=1}^{N} L\left(\frac{\mathbf{s}_0-\mathbf{S}_i}{\|\mathbf{S}_{(k_N)}-\mathbf{s}_0\|}\right)}.$$

Thus, $\hat{g}_N$ is the uniform kernel case of $\tilde{g}_N$. The asymptotic properties of $\tilde{g}_N$, which are beyond the scope of the present article, will be explored elsewhere by the authors. A good starting point are the papers by Moore and Yackel (1977a,b) and Mack and Rosenblatt (1979), who study various properties of similar kernel-type nearest neighbor procedures for density estimation.

4 Rates of convergence

In this section, we go one step further in the analysis of the ABC-companion estimate $\hat{g}_N$ by studying its mean integrated square error rates of convergence. We follow the notation of Section 3 and try to keep the assumptions on unknown mathematical objects as mild as possible. Introduce the multi-index notation

$$|\beta| = \beta_1 + \ldots + \beta_n, \quad \beta! = \beta_1! \ldots \beta_n!, \quad \mathbf{x}^\beta = x_1^{\beta_1} \ldots x_n^{\beta_n}$$

for $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$ and $\mathbf{x} \in \mathbb{R}^n$. If all the k-order derivatives of some function $\varphi : \mathbb{R}^n \to \mathbb{R}$ are continuous at $\mathbf{x}_0 \in \mathbb{R}^n$ then, by Schwarz's theorem, one can change the order of mixed derivatives at $\mathbf{x}_0$, so the notation

$$D^\beta \varphi(\mathbf{x}_0) = \frac{\partial^{|\beta|} \varphi(\mathbf{x}_0)}{\partial x_1^{\beta_1} \ldots \partial x_n^{\beta_n}}, \quad |\beta| \leq k$$

for the higher-order partial derivatives is justified in this situation.

In the sequel, we shall need the following sets of assumptions. Recall that the collection of all $\mathbf{s}_0 \in \mathbb{R}^m$ with $\int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s}) \mathrm{d}\mathbf{s} > 0$ for all $\delta > 0$ is called the support of $\bar{f}$.

Assumption [A1] The marginal probability density $\bar{f}$ has compact support with diameter $L > 0$ and is three times continuously differentiable.

Assumption [A2] The joint probability density f is in $L^2(\mathbb{R}^p \times \mathbb{R}^m)$. Moreover, for fixed $\mathbf{s}_0$, the functions

$$\begin{aligned} \boldsymbol{\theta}_0 &\mapsto \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathbf{s}_0)}{\partial \theta_{i_1} \partial \theta_{i_2}}, \quad 1 \leq i_1, i_2 \leq p \\ \text{and } \boldsymbol{\theta}_0 &\mapsto \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathbf{s}_0)}{\partial s_j^2}, \quad 1 \leq j \leq m \end{aligned}$$

are defined and belong to $L^2(\mathbb{R}^p)$.

Assumption [A3] The joint probability density f is three times continuously differentiable on $\mathbb{R}^p \times \mathbb{R}^m$ and, for any multi-index β satisfying $|\beta| = 3$,

$$\sup_{\mathbf{s} \in \mathbb{R}^m} \int_{\mathbb{R}^p} \left[D^\beta f(\boldsymbol{\theta}, \mathbf{s}) \right]^2 \mathrm{d}\boldsymbol{\theta} < \infty.$$

It is also necessary to put some mild additional restrictions on the kernel.

Assumption [K2] The kernel K is symmetric and belongs to $L^2(\mathbb{R}^p)$. Moreover, for any multi-index β satisfying $|\beta| \in \{1, 2, 3\}$,

$$\int_{\mathbb{R}^p} \left| \boldsymbol{\theta}^\beta \right| K(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty.$$

We finally define

$$\xi_0 = \inf_{0 < \delta \leq L} \frac{1}{\delta^m} \int_{\mathcal{B}_m(\mathbf{s}_0, \delta)} \bar{f}(\mathbf{s}) \mathrm{d}\mathbf{s},$$

and introduce the following quantities, which are related to the average distance between $\mathbf{s}_0$ and its k_N-th nearest neighbor (see Proposition 6.1 and Proposition 6.2):

$$\begin{aligned} D_m(k_N) &= \frac{m}{\xi_0^{2/m}(m-2)} \left(\frac{k_N + 1}{N + 1} \right)^{2/m} - \frac{L^{2-m}}{\xi_0(m/2 - 1)} \frac{k_N + 1}{N + 1}, \\ \Delta_m(k_N) &= \frac{m}{\xi_0^{4/m}(m-4)} \left(\frac{k_N + 1}{N + 1} \right)^{4/m} - \frac{L^{4-m}}{\xi_0(m/4 - 1)} \frac{k_N + 1}{N + 1}, \\ D(k_N) &= \frac{1}{\xi_0} \left(1 + \log \left(\xi_0 L^2 \frac{N+1}{k_N + 1} \right) \right) \frac{k_N + 1}{N + 1}, \\ \Delta(k_N) &= \frac{1}{\xi_0} \left(1 + \log \left(\xi_0 L^4 \frac{N+1}{k_N + 1} \right) \right) \frac{k_N + 1}{N + 1}. \end{aligned}$$

The next theorem makes precise the mean integrated square error rates of convergence of $\hat{g}_N(.)$ towards $g(.|\mathrm{s}_0)$.

Theorem 4.1 *Let K be a kernel satisfying assumptions* [**K1**] *and* [**K2**]*. Let* s_0 *be a Lebesgue point of* $\bar{f}$ *such that* $\bar{f}(\mathrm{s}_0) > 0$*. Assume that Assumptions* [**A1**]-[**A3**] *are satisfied. Then, letting*

$$\phi_1(\boldsymbol{\theta}_0, \mathrm{s}_0) = \frac{1}{2} \sum_{i_1, i_2 = 1}^{p} \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathrm{s}_0)}{\partial \theta_{i_1} \partial \theta_{i_2}} \int_{\mathbb{R}^p} \theta_{i_1} \theta_{i_2} K(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

$$\phi_2(\boldsymbol{\theta}_0, \mathrm{s}_0) = \frac{1}{2m+4} \sum_{j=1}^{m} \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathrm{s}_0)}{\partial s_j^2},$$

$$\phi_3(\mathrm{s}_0) = \frac{1}{2m+4} \sum_{j=1}^{m} \frac{\partial^2 \bar{f}(\mathrm{s}_0)}{\partial s_j^2},$$

and

$$\Phi_1(\mathrm{s}_0) = \frac{1}{\bar{f}^2(\mathrm{s}_0)} \int_{\mathbb{R}^p} \phi_1^2(\boldsymbol{\theta}_0, \mathrm{s}_0) \mathrm{d}\boldsymbol{\theta}_0,$$

$$\Phi_2(\mathrm{s}_0) = \frac{1}{\bar{f}^4(\mathrm{s}_0)} \int_{\mathbb{R}^p} \left[\phi_2(\boldsymbol{\theta}_0, \mathrm{s}_0) \bar{f}(\mathrm{s}_0) - \phi_3(\mathrm{s}_0) f(\boldsymbol{\theta}_0, \mathrm{s}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0,$$

$$\Phi_3(\mathrm{s}_0) = \frac{2}{\bar{f}^3(\mathrm{s}_0)} \int_{\mathbb{R}^p} \phi_1(\boldsymbol{\theta}_0, \mathrm{s}_0) \left[\phi_2(\boldsymbol{\theta}_0, \mathrm{s}_0) \bar{f}(\mathrm{s}_0) - \phi_3(\mathrm{s}_0) f(\boldsymbol{\theta}_0, \mathrm{s}_0)\right] \mathrm{d}\boldsymbol{\theta}_0,$$

one has:

1. **For $\boldsymbol{m = 2}$**,

$$\mathbb{E}\left[\int_{\mathbb{R}^p} \left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0 | \mathrm{s}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right]$$
$$= \left(\Phi_1(\mathrm{s}_0) h_N^4 + \Phi_2(\mathrm{s}_0) \Delta_2(k_N) + \Phi_3(\mathrm{s}_0) h_N^2 D(k_N) + \frac{\int_{\mathbb{R}^p} K^2(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{k_N h_N^p}\right)$$
$$\times \left(1 + \mathrm{o}(1)\right).$$

2. **For $\boldsymbol{m = 4}$**,

$$\mathbb{E}\left[\int_{\mathbb{R}^p} \left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0 | \mathrm{s}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right]$$
$$= \left(\Phi_1(\mathrm{s}_0) h_N^4 + \Phi_2(\mathrm{s}_0) \Delta(k_N) + \Phi_3(\mathrm{s}_0) h_N^2 D_4(k_N) + \frac{\int_{\mathbb{R}^p} K^2(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{k_N h_N^p}\right)$$
$$\times \left(1 + \mathrm{o}(1)\right).$$

3. **For $\boldsymbol{m \notin \{2,4\}}$**,

$$\mathbb{E}\left[\int_{\mathbb{R}^p}\left[\hat{g}_N(\boldsymbol{\theta}_0)-g(\boldsymbol{\theta}_0|\mathbf{s}_0)\right]^2\mathrm{d}\boldsymbol{\theta}_0\right]$$
$$=\left(\Phi_1(\mathbf{s}_0)h_N^4+\Phi_2(\mathbf{s}_0)\Delta_m(k_N)+\Phi_3(\mathbf{s}_0)h_N^2D_m(k_N)+\frac{\int_{\mathbb{R}^p}K^2(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{k_Nh_N^p}\right)$$
$$\times\left(1+\mathrm{o}(1)\right).$$

By balancing the terms in Theorem 4.1, we are led to the following useful corollary:

Corollary 4.1 (Rates of convergence) *Under the conditions of Theorem 4.1, one has:*

1. **For $\boldsymbol{m\in\{1,2,3\}}$**, *there exists a sequence $\{k_N\}$ with $k_N\propto N^{\frac{p+4}{p+8}}$ and a sequence $\{h_N\}$ with $h_N\propto N^{-\frac{1}{p+8}}$ such that*

$$\mathbb{E}\left[\int_{\mathbb{R}^p}\left[\hat{g}_N(\boldsymbol{\theta}_0)-g(\boldsymbol{\theta}_0|\mathbf{s}_0)\right]^2\mathrm{d}\boldsymbol{\theta}_0\right]$$
$$=\left(\frac{L^{4-m}\Phi_1(\mathbf{s}_0)}{\xi_0(1-m/4)}+\Phi_2(\mathbf{s}_0)+\int_{\mathbb{R}^p}K^2(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right)N^{-\frac{4}{p+8}}+\mathrm{o}\left(N^{-\frac{4}{p+8}}\right).$$

2. **For $\boldsymbol{m=4}$**, *there exists a sequence $\{k_N\}$ with $k_N\propto N^{\frac{p+4}{p+8}}$ and a sequence $\{h_N\}$ with $h_N\propto N^{-\frac{1}{p+8}}$ such that*

$$\mathbb{E}\left[\int_{\mathbb{R}^p}\left[\hat{g}_N(\boldsymbol{\theta}_0)-g(\boldsymbol{\theta}_0|\mathbf{s}_0)\right]^2\mathrm{d}\boldsymbol{\theta}_0\right]$$
$$=\frac{4\Phi_1(\mathbf{s}_0)}{\xi_0(p+8)}N^{-\frac{4}{p+8}}\log N+\mathrm{o}\left(N^{-\frac{4}{p+8}}\log N\right).$$

3. **For $\boldsymbol{m>4}$**, *there exists a sequence $\{k_N\}$ with $k_N\propto N^{\frac{p+4}{m+p+4}}$ and a sequence $\{h_N\}$ with $h_N\propto N^{-\frac{1}{m+p+4}}$ such that*

$$\mathbb{E}\left[\int_{\mathbb{R}^p}\left[\hat{g}_N(\boldsymbol{\theta}_0)-g(\boldsymbol{\theta}_0|\mathbf{s}_0)\right]^2\mathrm{d}\boldsymbol{\theta}_0\right]$$
$$=\left(\frac{m\Phi_1(\mathbf{s}_0)}{\xi_0^{4/m}(m-4)}+\Phi_2(\mathbf{s}_0)+\frac{m\Phi_3(\mathbf{s}_0)}{\xi_0^{2/m}(m-2)}+\int_{\mathbb{R}^p}K^2(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right)N^{-\frac{4}{m+p+4}}$$
$$+\,\mathrm{o}\left(N^{-\frac{4}{m+p+4}}\right).$$

Several remarks are in order:

1. From a practical perspective, the fundamental problem is that of the joint choice of k_N and h_N in the absence of *a priori* information regarding the posterior $g(.|\mathbf{s}_0)$. Various bandwidth selection rules for conditional density estimates have been proposed in the literature (see, e.g., Bashtannyk and Hyndman, 2001; Hall et al., 2004; Fan and Yim, 2004). However most if not all of these procedures pertain to kernel-type estimates and are difficult to adapt to our nearest-neighbor setting. Moreover, they are tailored to global statistical performance criteria, whereas the problem we are facing is local since $\mathbf{s}_0$ is held fixed. Devising a good methodology to automatically select both parameters k_N and h_N in function of $\mathbf{s}_0$ necessitates a specific analysis, which we believe is beyond the scope of the present paper.

2. Nevertheless, Corollary 4.1 provides a useful insight into the proportion of simulated values which should be accepted by the algorithm. For example, for $m > 4$, a rough rule of thumb is obtained by taking $k_N \approx N^{(p+4)/(m+p+4)}$, so that a fraction of about $k_N/N \approx N^{-m/(m+p+4)}$ ABC-simulations should not be rejected.

5 Proofs

5.1 Proof of Proposition 2.1

Denote by $(\tilde{\mathbf{\Theta}}_1, \tilde{\mathbf{S}}_1), \ldots, (\tilde{\mathbf{\Theta}}_k, \tilde{\mathbf{S}}_k)$ i.i.d. random couples with common probability density

$$\frac{1}{C_{d_{(k+1)}}}\mathbf{1}_{[\|\mathbf{s}-\mathbf{s}_0\| \leq d_{(k+1)}]} f(\boldsymbol{\theta}, \mathrm{s}),$$

where the normalizing constant $C_{d_{(k+1)}}$ is defined by

$$C_{d_{(k+1)}} = \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k+1)})} f(\boldsymbol{\theta}, \mathrm{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{ds}.$$

Note, since $\mathbf{s}_0$ belongs by assumption to the support of $\bar{f}$, that the constant $C_{d_{(k+1)}}$ is positive. To prove the first statement of the theorem, it is enough to establish that, for any test functions Φ and φ, with Φ symmetric in its arguments, one has

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left((\mathbf{\Theta}_{(1)}, \mathbf{S}_{(1)}), \ldots, (\mathbf{\Theta}_{(k)}, \mathbf{S}_{(k)})\right) \varphi(d_{(k+1)})\right] \\
&\quad = \mathbb{E}\left[\Phi\left((\tilde{\mathbf{\Theta}}_1, \tilde{\mathbf{S}}_1), \ldots, (\tilde{\mathbf{\Theta}}_k, \tilde{\mathbf{S}}_k)\right) \varphi(d_{(k+1)})\right].
\end{aligned}$$

This can be achieved by adapting the proof of Lemma A.1 in Cérou and Guyader (2006) to this context. Details are omitted.

To prove the second statement, it suffices to show that, for any test functions Φ and φ (with Φ not necessarily symmetric), one has

$$\begin{aligned}&\mathbb{E}[\Phi\left((\mathbf{\Theta}_{(1)}, \mathbf{S}_{(1)}), \ldots, (\mathbf{\Theta}_{(k)}, \mathbf{S}_{(k)})\right) \varphi(d_{(k+1)})] \\ &\quad = \mathbb{E}[\Phi\left((\tilde{\mathbf{\Theta}}_{(1)}, \tilde{\mathbf{S}}_{(1)}), \ldots, (\tilde{\mathbf{\Theta}}_{(k)}, \tilde{\mathbf{S}}_{(k)})\right) \varphi(d_{(k+1)})].\end{aligned}$$

The arguments of Cérou and Guyader (2006) may be repeated *mutatis mutandis* by replacing the k-combinations of $\{1, \ldots, N\}$ by the k-permutations.

5.2 Proof of Theorem 3.3

The proof strongly relies on Proposition 2.1. It is assumed throughout that $\mathbf{s}_0$ is a Lebesgue point of $\bar{f}$ (λ_m-almost all points satisfy this requirement) such that $\bar{f}(\mathbf{s}_0) > 0$. We note that this forces $\mathbf{s}_0$ to belong to the support of $\bar{f}$, so that the assumption of Proposition 2.1 is satisfied. The collection of valid $\mathbf{s}_0$ will vary during the proof, but only on subsets of Lebesgue measure 0. Similarly, we fix $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, up to subsets of Lebesgue measure 0 which will appear in the proof.

First observe that, according to Proposition 2.1,

$$\mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}] = \frac{1}{C_{d_{(k_N+1)}}} \int_{\mathbb{R}^p} K_{h_N}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) \left(\int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathrm{s}) \mathrm{ds} \right) \mathrm{d}\boldsymbol{\theta},$$

where, for any $\delta > 0$, $C_\delta = \int_{\mathcal{B}_m(\mathbf{s}_0, \delta)} \bar{f}(\mathbf{s}) \mathrm{d}\mathbf{s}$. Put differently, by Fubini's theorem,

$$\mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}] = \frac{1}{C_{d_{(k_N+1)}}} \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} K_{h_N}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) f(\boldsymbol{\theta}, \mathrm{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{ds}. \tag{5.1}$$

The proof starts with the variance-bias decomposition

$$\begin{aligned}\mathbb{E}\left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0)\right]^2 &= \mathbb{E}\left[\mathbb{E}\left[\left(\hat{g}_N(\boldsymbol{\theta}_0) - \mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}]\right)^2 \,\Big|\, d_{(k_N+1)}\right]\right] \\ &\quad + \mathbb{E}\left[\mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}] - g(\boldsymbol{\theta}_0)\right]^2. \end{aligned} \tag{5.2}$$

Our goal is to show that, under our assumptions, both terms on the right-hand side of (5.2) tend to 0 as $N \to \infty$. We start with the analysis of the second one, by noting that

$$\begin{aligned}&\left|\mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0)\,\middle|\,d_{(k_N+1)}]-g(\boldsymbol{\theta}_0)\right|\\&\quad=\left|\frac{1}{C_{d_{(k_N+1)}}}\int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathrm{s}_0,d_{(k_N+1)})}K_{h_N}(\boldsymbol{\theta}_0-\boldsymbol{\theta})f(\boldsymbol{\theta},\mathrm{s})\mathrm{d}\boldsymbol{\theta}\mathrm{ds}-\frac{f(\boldsymbol{\theta}_0,\mathrm{s}_0)}{\bar{f}(\mathrm{s}_0)}\right|,\end{aligned}$$

where we used (5.1) and the definition of $g(\boldsymbol{\theta}_0)$. Equivalently,

$$\begin{aligned}&\left|\mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0)\,\middle|\,d_{(k_N+1)}]-g(\boldsymbol{\theta}_0)\right|\\&\quad=\left|\frac{V_m\,d^m_{(k_N+1)}}{C_{d_{(k_N+1)}}}\frac{1}{V_m\,d^m_{(k_N+1)}}\int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathrm{s}_0,d_{(k_N+1)})}K_{h_N}(\boldsymbol{\theta}_0-\boldsymbol{\theta})f(\boldsymbol{\theta},\mathrm{s})\mathrm{d}\boldsymbol{\theta}\mathrm{ds}\right.\\&\qquad\left.-\frac{f(\boldsymbol{\theta}_0,\mathrm{s}_0)}{\bar{f}(\mathrm{s}_0)}\right|.\end{aligned}$$

For a fixed pair $(\boldsymbol{\theta}_0,\mathrm{s}_0)$ and all $h,\delta>0$, set

$$\psi_{\boldsymbol{\theta}_0,\mathrm{s}_0}(h,\delta)=\left|\frac{V_m\delta^m}{C_\delta}\frac{1}{V_m\delta^m}\int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathrm{s}_0,\delta)}K_h(\boldsymbol{\theta}_0-\boldsymbol{\theta})f(\boldsymbol{\theta},\mathrm{s})\mathrm{d}\boldsymbol{\theta}\mathrm{ds}-\frac{f(\boldsymbol{\theta}_0,\mathrm{s}_0)}{\bar{f}(\mathrm{s}_0)}\right|.$$

According to technical Lemma 6.1 (i), the quantity $V_m\delta^m/C_\delta$ tends to $1/\bar{f}(\mathrm{s}_0)$ as $\delta\to0$. Therefore, by the first statement of Theorem 3.2, we deduce that for $\lambda_p\otimes\lambda_m$-almost all pairs $(\boldsymbol{\theta}_0,\mathrm{s}_0)\in\mathbb{R}^p\times\mathbb{R}^m$, $\lim_{(h,\delta)\to(0,0)}\psi^2_{\boldsymbol{\theta}_0,\mathrm{s}_0}(h,\delta)=0$.

Next, introduce $\pi^\star$ (respectively, $f^\star$), the maximal function defined in Theorem 3.1 (respectively, Theorem 3.2). Take any $\delta_0>0$. On the one hand, by the very definition of $f^\star$,

$$\sup_{h>0,\delta_0\geq\delta>0}[\psi_{\boldsymbol{\theta}_0,\mathrm{s}_0}(h,\delta)]\leq\sup_{0<\delta\leq\delta_0}\left[\frac{V_m\delta^m}{C_\delta}\right]f^\star(\boldsymbol{\theta}_0,\mathrm{s}_0)+\frac{f(\boldsymbol{\theta}_0,\mathrm{s}_0)}{\bar{f}(\mathrm{s}_0)}.$$

On the other hand, for $\delta>\delta_0$,

$$\psi_{\boldsymbol{\theta}_0,\mathrm{s}_0}(h,\delta)\leq\frac{1}{C_{\delta_0}}\int_{\mathbb{R}^p}K_h(\boldsymbol{\theta}_0-\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}+\frac{f(\boldsymbol{\theta}_0,\mathrm{s}_0)}{\bar{f}(\mathrm{s}_0)},$$

so that

$$\sup_{h>0,\delta>\delta_0}[\psi_{\boldsymbol{\theta}_0,\mathrm{s}_0}(h,\delta)]\leq\frac{\pi^\star(\boldsymbol{\theta}_0)}{C_{\delta_0}}+\frac{f(\boldsymbol{\theta}_0,\mathrm{s}_0)}{\bar{f}(\mathrm{s}_0)}.$$

Thus, putting all the pieces together, we infer that for $\lambda_p\otimes\lambda_m$-almost all pairs $(\boldsymbol{\theta}_0,\mathrm{s}_0)\in\mathbb{R}^p\times\mathbb{R}^m$,

$$\sup_{h>0,\delta>0}[\psi_{\boldsymbol{\theta}_0,\mathrm{s}_0}(h,\delta)]\leq\sup_{0<\delta\leq\delta_0}\left[\frac{V_m\delta^m}{C_\delta}\right]f^\star(\boldsymbol{\theta}_0,\mathrm{s}_0)+\frac{\pi^\star(\boldsymbol{\theta}_0)}{C_{\delta_0}}+\frac{2f(\boldsymbol{\theta}_0,\mathrm{s}_0)}{\bar{f}(\mathrm{s}_0)}.\quad(5.3)$$

In consequence, by Lemma 6.1 (*ii*), Theorem 3.1 (*ii*) and Theorem 3.2 (*ii*), for such pairs $(\boldsymbol{\theta}_0, \mathbf{s}_0)$,

$$\sup_{h>0,\delta>0} \left[\psi^2_{\boldsymbol{\theta}_0,\mathbf{s}_0}(h,\delta)\right] < \infty. \tag{5.4}$$

Now, since $d_{(k_N+1)} \to 0$ with probability 1 whenever $k_N/N \to 0$ (see, e.g., Lemma 5.1 in Devroye et al., 1996), we conclude by Lebesgue's dominated convergence theorem that the bias term in (5.2) tends to 0 as $N \to \infty$.

To finish the proof, it remains to show that the first term of (5.2) vanishes as $N \to \infty$. This is easier. Just note that, using again Proposition 2.1,

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{g}_N(\boldsymbol{\theta}_0) - \mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}]\right)^2 \,\Big|\, d_{(k_N+1)}\right] \\
&\quad = \frac{1}{k_N h_N^{2p}} \frac{1}{C_{d_{(k_N+1)}}} \int_{\mathbb{R}^p} K^2\left(\frac{\boldsymbol{\theta}_0 - \boldsymbol{\theta}}{h_N}\right) \left(\int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\mathbf{s}\right) \mathrm{d}\boldsymbol{\theta} \\
&\quad\quad - \frac{1}{k_N}\left(\mathbb{E}\left[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}\right]\right)^2.
\end{aligned} \tag{5.5}$$

Hence, if K is bounded by, say, $\|K\|_\infty$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{g}_N(\boldsymbol{\theta}_0) - \mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}]\right)^2 \,\Big|\, d_{(k_N+1)}\right] \\
&\quad \le \frac{1}{k_N h_N^{2p}} \frac{1}{C_{d_{(k_N+1)}}} \int_{\mathbb{R}^p} K^2\left(\frac{\boldsymbol{\theta}_0 - \boldsymbol{\theta}}{h_N}\right) \left(\int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\mathbf{s}\right) \mathrm{d}\boldsymbol{\theta} \\
&\quad \le \frac{1}{k_N h_N^{p}} \frac{\|K\|_\infty}{C_{d_{(k_N+1)}}} \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} K_{h_N}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{s}.
\end{aligned}$$

Thus, using (5.4), we obtain

$$\mathbb{E}\left[\left(\hat{g}_N(\boldsymbol{\theta}_0) - \mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}]\right)^2 \,\Big|\, d_{(k_N+1)}\right] \le \frac{C}{k_N h_N^p}$$

for some positive constant C depending on $\boldsymbol{\theta}_0$, $\mathbf{s}_0$ and K, but independent of h_N and k_N. This shows that the variance term goes to 0 as $k_N h_N^p \to \infty$ and concludes the proof of the theorem.

5.3 Proof of Theorem 3.4

We start as in the proof of Theorem 3.3 and write, using Fubini's theorem,

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathbb{R}^p} [\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0)]^2 \,\mathrm{d}\boldsymbol{\theta}_0\right] \\
&\quad = \mathbb{E}\left[\int_{\mathbb{R}^p} \mathbb{E}\left[\left(\hat{g}_N(\boldsymbol{\theta}_0) - \mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}]\right)^2 \,\Big|\, d_{(k_N+1)}\right] \mathrm{d}\boldsymbol{\theta}_0\right] \\
&\qquad + \mathbb{E}\left[\int_{\mathbb{R}^p} \left[\mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}] - g(\boldsymbol{\theta}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right]. \qquad (5.6)
\end{aligned}$$

It has already been seen that

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{g}_N(\boldsymbol{\theta}_0) - \mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}]\right)^2 \,\Big|\, d_{(k_N+1)}\right] \\
&\quad \leq \frac{1}{k_N h_N^{2p}} \frac{1}{C_{d_{(k_N+1)}}} \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} K^2\left(\frac{\boldsymbol{\theta}_0 - \boldsymbol{\theta}}{h_N}\right) f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{s}.
\end{aligned}$$

Consequently, by definition of $C_{d_{(k_N+1)}}$, we are led to

$$\int_{\mathbb{R}^p} \mathbb{E}\left[\left(\hat{g}_N(\boldsymbol{\theta}_0) - \mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}]\right)^2 \,\Big|\, d_{(k_N+1)}\right] \mathrm{d}\boldsymbol{\theta}_0 \leq \frac{\int_{\mathbb{R}^p} K^2(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{k_N h_N^p}.$$

This shows that the first term in (5.6) tends to 0 as $k_N h_N^p \to \infty$.

Let us now turn to the analysis of the bias term. With the notation of the proof of Theorem 3.3, we may write

$$\mathbb{E}\left[\int_{\mathbb{R}^p} \left[\mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}] - g(\boldsymbol{\theta}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right] = \mathbb{E}\left[\int_{\mathbb{R}^p} \psi^2_{\boldsymbol{\theta}_0, \mathbf{s}_0}(h_N, d_{(k_N+1)}) \mathrm{d}\boldsymbol{\theta}_0\right].$$

It is known from the proof of Theorem 3.3 that the limit of $\psi^2_{\boldsymbol{\theta}_0,\mathbf{s}_0}(h,\delta)$ is 0 for $\lambda_p \otimes \lambda_m$-almost all $(\boldsymbol{\theta}_0, \mathbf{s}_0) \in \mathbb{R}^p \times \mathbb{R}^m$, whenever $(h,\delta) \to (0,0)$. Take any $\delta_0 > 0$. Denoting by $f^\star$ (respectively, $\pi^\star$) the maximal function defined in Theorem 3.2 (respectively, Theorem 3.1), we also know (inequality (5.3)) that

$$\sup_{h>0,\delta>0} [\psi_{\boldsymbol{\theta}_0,\mathbf{s}_0}(h,\delta)] \leq \sup_{0<\delta\leq\delta_0} \left[\frac{V_m \delta^m}{C_\delta}\right] f^\star(\boldsymbol{\theta}_0, \mathbf{s}_0) + \frac{\pi^\star(\boldsymbol{\theta}_0)}{C_{\delta_0}} + \frac{2f(\boldsymbol{\theta}_0, \mathbf{s}_0)}{\bar{f}(\mathbf{s}_0)}.$$

Thus, because $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$,

$$\begin{aligned}
&\sup_{h>0,\delta>0} \left[\psi^2_{\boldsymbol{\theta}_0,\mathbf{s}_0}(h,\delta)\right] \\
&\quad \leq 3\left(\sup_{0<\delta\leq\delta_0} \left[\frac{V_m \delta^m}{C_\delta}\right] f^\star(\boldsymbol{\theta}_0, \mathbf{s}_0)\right)^2 + 3\left(\frac{\pi^\star(\boldsymbol{\theta}_0)}{C_{\delta_0}}\right)^2 + 12\left(\frac{f(\boldsymbol{\theta}_0, \mathbf{s}_0)}{\bar{f}(\mathbf{s}_0)}\right)^2.
\end{aligned}$$

By Lemma 6.1 (*ii*), the supremum on the right-hand side is bounded. Moreover, by assumption, f is in $L^2(\mathbb{R}^p \times \mathbb{R}^m)$. Therefore the function $\boldsymbol{\theta}_0 \mapsto f(\boldsymbol{\theta}_0, \mathbf{s}_0)$ is in $L^2(\mathbb{R}^p)$ as well for λ_m-almost all $\mathbf{s}_0 \in \mathbb{R}^m$. Similarly, for λ_m-almost all $\mathbf{s}_0$, by Theorem 3.2 (*iii*), the function $\boldsymbol{\theta}_0 \mapsto f^\star(\boldsymbol{\theta}_0, \mathbf{s}_0)$ is in $L^2(\mathbb{R}^p)$. Finally, $\pi^\star$ belongs to $L^2(\mathbb{R}^p)$ by Theorem 3.1 (*iii*). Since $d_{(k_N+1)} \to 0$ with probability 1 whenever $k_N/N \to 0$, the conclusion follows from Lebesgue's dominated convergence theorem.

5.4 Proof of Theorem 4.1

Throughout the proof, it is assumed that the Lebesgue point $\mathbf{s}_0$ is fixed and such that $\bar{f}(\mathbf{s}_0) > 0$. This forces $\mathbf{s}_0$ to belong to the support of $\bar{f}$.

As in the proofs of Theorem 3.3 and Theorem 3.4, we set, for any $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ and all $h, \delta > 0$,

$$\psi_{\boldsymbol{\theta}_0,\mathbf{s}_0}(h,\delta) = \left| \frac{V_m\delta^m}{C_\delta} \frac{1}{V_m\delta^m} \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} K_h(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) f(\boldsymbol{\theta}, \mathrm{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathrm{s} - \frac{f(\boldsymbol{\theta}_0, \mathrm{s}_0)}{\bar{f}(\mathrm{s}_0)} \right|,$$

where $C_\delta = \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s}$. With this notation, it is readily seen from identities (5.5) and (5.6) that

$$\begin{aligned} &\mathbb{E}\left[\int_{\mathbb{R}^p} \left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0|\mathbf{s}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right] \\ &\quad = \mathbb{E}\left[\int_{\mathbb{R}^p} \psi^2_{\boldsymbol{\theta}_0,\mathbf{s}_0}(h_N, d_{(k_N+1)})\mathrm{d}\boldsymbol{\theta}_0\right] + \frac{\int_{\mathbb{R}^p} K^2(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{k_N h_N^p} \\ &\quad\quad - \frac{1}{k_N}\mathbb{E}\left[\int_{\mathbb{R}^p} \left(\mathbb{E}\left[\hat{g}_N(\boldsymbol{\theta}_0) \,\middle|\, d_{(k_N+1)}\right]\right)^2 \mathrm{d}\boldsymbol{\theta}_0\right]. \end{aligned}$$

Recall that

$$\mathbb{E}[\hat{g}_N(\boldsymbol{\theta}_0) \,|\, d_{(k_N+1)}] = \frac{1}{C_{d_{(k_N+1)}}} \int_{\mathbb{R}^p} K_{h_N}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) \left(\int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathrm{s})\mathrm{d}\mathrm{s} \right) \mathrm{d}\boldsymbol{\theta},$$

and the same arguments as in the proof of Theorem 3.4 reveal that

$$\sup_{h_N>0, L \geq d_{(k_N+1)}>0} \left(\mathbb{E}\left[\hat{g}_N(\boldsymbol{\theta}_0) \,\middle|\, d_{(k_N+1)}\right]\right)^2 \leq \left(\sup_{0<\delta\leq L} \left[\frac{V_m\delta^m}{C_\delta}\right] f^\star(\boldsymbol{\theta}_0, \mathbf{s}_0)\right)^2.$$

Since f is in $L^2(\mathbb{R}^p \times \mathbb{R}^m)$ by Assumption [**A2**], this ensures that for λ_m-almost all $\mathbf{s}_0 \in \mathbb{R}^m$,

$$\mathbb{E}\left[\int_{\mathbb{R}^p} \left(\mathbb{E}\left[\hat{g}_N(\boldsymbol{\theta}_0) \,\middle|\, d_{(k_N+1)}\right]\right)^2 \mathrm{d}\boldsymbol{\theta}_0\right] < \infty$$

and

$$\frac{1}{k_N}\mathbb{E}\left[\int_{\mathbb{R}^p}\left(\mathbb{E}\left[\hat{g}_N(\boldsymbol{\theta}_0)\,\middle|\,d_{(k_N+1)}\right]\right)^2 \mathrm{d}\boldsymbol{\theta}_0\right] = \mathrm{O}\left(\frac{1}{k_N}\right).$$

In particular,

$$\frac{1}{k_N}\mathbb{E}\left[\int_{\mathbb{R}^p}\left(\mathbb{E}\left[\hat{g}_N(\boldsymbol{\theta}_0)\,\middle|\,d_{(k_N+1)}\right]\right)^2 \mathrm{d}\boldsymbol{\theta}_0\right] = \mathrm{o}\left(\frac{1}{k_N h_N^p}\right).$$

The rest of the proof is devoted to the study of the rate of convergence to 0 of the quantity

$$\mathbb{E}\left[\int_{\mathbb{R}^p}\psi^2_{\boldsymbol{\theta}_0,\mathbf{s}_0}(h_N, d_{(k_N+1)})\mathrm{d}\boldsymbol{\theta}_0\right].$$

By an elementary change of variables, using the symmetry of K,

$$\begin{aligned}\frac{1}{V_m\delta^m}&\int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} K_h(\boldsymbol{\theta}_0-\boldsymbol{\theta})f(\boldsymbol{\theta},\mathrm{s})\mathrm{d}\boldsymbol{\theta}\mathrm{ds}\\ &= \frac{1}{V_m}\int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathbf{0},1)} K(\boldsymbol{\theta})f(\boldsymbol{\theta}_0+h\boldsymbol{\theta},\mathrm{s}_0+\delta\mathrm{s})\mathrm{d}\boldsymbol{\theta}\mathrm{ds}.\end{aligned}$$

Next, by the multivariate Taylor's theorem applied to f around $(\boldsymbol{\theta}_0,\mathrm{s}_0)$ (which is valid here by Assumption [**A3**]),

$$\begin{aligned}f(\boldsymbol{\theta}_0+h\boldsymbol{\theta},\mathrm{s}_0+\delta\mathbf{s}) =& f(\boldsymbol{\theta}_0,\mathrm{s}_0)+\sum_{|\beta|=1}D^\beta f(\boldsymbol{\theta}_0,\mathrm{s}_0)(h\boldsymbol{\theta},\delta\mathbf{s})^\beta\\ &+\sum_{|\beta|=2}\frac{D^\beta f(\boldsymbol{\theta}_0,\mathrm{s}_0)}{\beta!}(h\boldsymbol{\theta},\delta\mathbf{s})^\beta\\ &+\sum_{|\beta|=3}R_\beta(\boldsymbol{\theta}_0+h\boldsymbol{\theta},\mathrm{s}_0+\delta\mathbf{s})(h\boldsymbol{\theta},\delta\mathbf{s})^\beta,\end{aligned}$$

where each component of the remainder term takes the form

$$R_\beta(\boldsymbol{\theta}_0+h\boldsymbol{\theta},\mathrm{s}_0+\delta\mathbf{s}) = \frac{3}{\beta!}\int_0^1(1-t)^2D^\beta f(\boldsymbol{\theta}_0+th\boldsymbol{\theta},\mathrm{s}_0+t\delta\mathbf{s})\mathrm{d}t.$$

In view of the symmetry of K and the ball $\mathcal{B}_m(\mathbf{0},1)$, it is clear that

$$\int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathbf{0},1)}K(\boldsymbol{\theta})\sum_{|\beta|=1}D^\beta f(\boldsymbol{\theta}_0,\mathrm{s}_0)(h\boldsymbol{\theta},\delta\mathbf{s})^\beta\mathrm{d}\boldsymbol{\theta}\mathrm{ds}=0.$$

Similarly, elementary calculations reveal that

$$\begin{aligned}\frac{1}{V_m}&\int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathbf{0},1)}K(\boldsymbol{\theta})\sum_{|\beta|=2}\frac{D^\beta f(\boldsymbol{\theta}_0,\mathrm{s}_0)}{\beta!}(h\boldsymbol{\theta},\delta\mathbf{s})^\beta\mathrm{d}\boldsymbol{\theta}\mathrm{ds}\\ &=\phi_1(\boldsymbol{\theta}_0,\mathrm{s}_0)h^2+\phi_2(\boldsymbol{\theta}_0,\mathrm{s}_0)\delta^2\end{aligned}$$

(where ϕ_1 is defined in the statement of Theorem 4.1), and

$$\phi_2(\boldsymbol{\theta}_0, \mathrm{s}_0) = \frac{1}{2V_m} \sum_{j=1}^{m} \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathrm{s}_0)}{\partial s_j^2} \int_{\mathcal{B}_m(\mathbf{0},1)} s_j^2 \mathrm{d}\mathrm{s}.$$

Using expression (3.2) of V_m, an elementary verification shows that

$$\frac{1}{V_m} \int_{\mathcal{B}_m(\mathbf{0},1)} s_j^2 \mathrm{d}\mathrm{s} = \frac{1}{m+2} \quad \text{and} \quad \phi_2(\boldsymbol{\theta}_0, \mathrm{s}_0) = \frac{1}{2m+4} \sum_{j=1}^{m} \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathrm{s}_0)}{\partial s_j^2}.$$

Let us now define $(\mathbf{h}, \boldsymbol{\delta}) = (h, \ldots, h, \delta, \ldots, \delta)$ (where h is replicated p times and δ is replicated m times) and care about the remainder term $R_\beta(\boldsymbol{\theta}_0 + h\boldsymbol{\theta}, \mathrm{s}_0 + \delta\mathbf{s})$. For any multi-index β with $|\beta| = 3$, it holds

$$\int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{0},1)} K(\boldsymbol{\theta}) R_\beta(\boldsymbol{\theta}_0 + h\boldsymbol{\theta}, \mathrm{s}_0 + \delta\mathbf{s})(h\boldsymbol{\theta}, \delta\mathbf{s})^\beta \mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathrm{s} = (\mathbf{h}, \boldsymbol{\delta})^\beta A_\beta(\boldsymbol{\theta}_0, h, \delta),$$

where, by definition,

$$A_\beta(\boldsymbol{\theta}_0, h, \delta) = \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{0},1)} K(\boldsymbol{\theta}) R_\beta(\boldsymbol{\theta}_0 + h\boldsymbol{\theta}, \mathrm{s}_0 + \delta\mathbf{s})(\boldsymbol{\theta}, \mathrm{s})^\beta \mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathrm{s}.$$

[Note that $A_\beta(\boldsymbol{\theta}_0, h, \delta)$ depends in fact upon $\mathbf{s}_0$ as well, but since this dependency is not crucial, we leave it out in the notation.] Finally,

$$\begin{aligned} &\frac{1}{V_m\delta^m} \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathrm{s}_0,\delta)} K_h(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) f(\boldsymbol{\theta}, \mathrm{s}) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathrm{s} \\ &\quad = f(\boldsymbol{\theta}_0, \mathrm{s}_0) + \phi_1(\boldsymbol{\theta}_0, \mathrm{s}_0)h^2 + \phi_2(\boldsymbol{\theta}_0, \mathrm{s}_0)\delta^2 + \sum_{|\beta|=3} (\mathbf{h}, \boldsymbol{\delta})^\beta A_\beta(\boldsymbol{\theta}_0, h, \delta). \end{aligned}$$

Considering now the function

$$\tau_{\mathbf{s}_0}(\delta) = \frac{C_\delta}{V_m\delta^m} = \frac{1}{V_m\delta^m} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} = \frac{1}{V_m} \int_{\mathcal{B}_m(\mathbf{0},1)} \bar{f}(\mathbf{s}_0 + \delta\mathbf{s})\mathrm{d}\mathbf{s},$$

and the asymptotic expansion of $1/\tau_{\mathbf{s}_0}$ around 0, a similar analysis shows that

$$\frac{V_m\delta^m}{C_\delta} = \frac{1}{\bar{f}(\mathbf{s}_0)} - \frac{\phi_3(\mathbf{s}_0)}{\bar{f}^2(\mathbf{s}_0)}\delta^2 + \delta^3\zeta_1(\delta)$$

(where ϕ_3 is defined in the statement of Theorem 4.1), and, with a slight abuse of notation, there exists $t \in (0, 1)$ such that $\zeta_1(\delta) = H(t\delta)/\tau_{\mathbf{s}_0}^4(t\delta)$. In this last expression, the function H depends only on the successive derivatives

$D^\beta \bar{f}(\mathbf{s}_0+t\delta\mathbf{s})$ for $0 \le |\beta| \le 3$ and is therefore bounded thanks to Assumption [**A1**]. Besides, by the very definition of ξ_0 and technical Lemma 6.3,

$$\tau_{\mathbf{s}_0}(t\delta) = \frac{1}{V_m (t\delta)^m} \int_{\mathcal{B}_m(\mathbf{s}_0, t\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} \ge \frac{\xi_0}{V_m} > 0.$$

Thus, the function $\zeta_1(\delta)$ is such that $\sup_{0<\delta\le L} \zeta_1(\delta) < \infty$. Putting all the pieces together, we conclude that

$$\psi_{\boldsymbol{\theta}_0,\mathbf{s}_0}(h,\delta) = \left|\phi_4(\boldsymbol{\theta}_0,\mathbf{s}_0)h^2 + \phi_5(\boldsymbol{\theta}_0,\mathbf{s}_0)\delta^2 + h^2\zeta_2(\boldsymbol{\theta}_0,h,\delta) + \delta^2\zeta_3(\boldsymbol{\theta}_0,h,\delta)\right|,$$

where

$$\phi_4(\boldsymbol{\theta}_0,\mathbf{s}_0) = \frac{\phi_1(\boldsymbol{\theta}_0,\mathbf{s}_0)}{\bar{f}(\mathbf{s}_0)} \quad \text{and} \quad \phi_5(\boldsymbol{\theta}_0,\mathbf{s}_0) = \frac{\phi_2(\boldsymbol{\theta}_0,\mathbf{s}_0)\bar{f}(\mathbf{s}_0) - \phi_3(\mathbf{s}_0) f(\boldsymbol{\theta}_0,\mathbf{s}_0)}{\bar{f}^2(\mathbf{s}_0)}.$$

Moreover, one can check, using Assumption [**A2**] and the second statement of Assumption [**A3**] together with technical Lemma 6.2, that for $i = 2, 3$, $\zeta_i(\boldsymbol{\theta}_0, h, \delta) \to 0$ as $(h,\delta) \to (0,0)$, and

$$\sup_{0<h<M, 0<\delta\le L} \int_{\mathbb{R}^p} \zeta_i^2(\boldsymbol{\theta}_0,h,\delta)\mathrm{d}\boldsymbol{\theta}_0 < \infty$$

for all positive M. As a consequence,

$$\int_{\mathbb{R}^p} \psi^2_{\boldsymbol{\theta}_0,\mathbf{s}_0}(h,\delta)\mathrm{d}\boldsymbol{\theta}_0 = \Phi_1(\mathbf{s}_0)h^4 + \Phi_2(\mathbf{s}_0)\delta^4 + \Phi_3(\mathbf{s}_0)h^2\delta^2 + (h^2+\delta^2)^2\zeta_4(h,\delta)$$

(Φ_1, Φ_2 and Φ_3 are defined in the statement of Theorem 4.1). Besides, for all positive M,

$$\sup_{0<h<M, 0<\delta\le L} \zeta_4(h,\delta) < \infty \quad \text{and} \quad \lim_{(h,\delta)\to(0,0)} \zeta_4(h,\delta) = 0. \tag{5.7}$$

Finally,

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathbb{R}^p} \left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0|\mathbf{s}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right] \\
&\quad = \Phi_1(\mathbf{s}_0)h_N^4 + \Phi_2(\mathbf{s}_0)\mathbb{E}[d^4_{(k_N+1)}] + \Phi_3(\mathbf{s}_0)h_N^2\mathbb{E}[d^2_{(k_N+1)}] + \frac{\int_{\mathbb{R}^p} K^2(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{k_N h_N^p} \\
&\qquad + \mathbb{E}\left[(h_N^2 + d^2_{(k_N+1)})^2\zeta_4(h_N, d_{(k_N+1)})\right] + \mathrm{o}\left(\frac{1}{k_N h_N^p}\right).
\end{aligned}$$

The conclusion is then an immediate consequence of (5.7) and Assumption [**A1**], together with Proposition 6.1 and Proposition 6.2, which respectively provide upper bounds on $\mathbb{E}[d^2_{(k_N+1)}]$ and $\mathbb{E}[d^4_{(k_N+1)}]$ depending on the dimension m.

6 Some technical results

Lemma 6.1 *Let* $\mathrm{s}_0 \in \mathbb{R}^m$ *be a Lebesgue point of* $\bar{f}$ *such that* $\bar{f}(\mathrm{s}_0) > 0$. *For any* $\delta > 0$, *let* $C_\delta = \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s}$. *One has*

(*i*) $\lim_{\delta\to 0} V_m\delta^m/C_\delta = 1/\bar{f}(\mathrm{s}_0)$.

(*ii*) *For any* $\delta_0 > 0$, $\sup_{0<\delta\leq\delta_0} V_m\delta^m/C_\delta < \infty$.

Proof of Lemma 6.1 The first statement is an immediate consequence of Lebesgue's differentiation theorem (Wheeden and Zygmund, 1977, Theorem 7.2). Take now $\delta_0 > 0$. Since $\bar{f}(\mathrm{s}_0) > 0$, it is routine to verify that the mapping $\delta \mapsto \frac{V_m\delta^m}{C_\delta}$ is positive and continuous on $(0, \delta_0]$. Thus, by (*i*), we deduce that $\sup_{0<\delta\leq\delta_0} V_m\delta^m/C_\delta < \infty$. ■

Lemma 6.2 *Assume that the joint probability density* f *is three times continuously differentiable on* $\mathbb{R}^p \times \mathbb{R}^m$, *and let* β *be a multi-index satisfying* $|\beta| = 3$. *Assume that* $\sup_{\mathbf{s}\in\mathbb{R}^m} \int_{\mathbb{R}^p} \left[D^\beta f(\boldsymbol{\theta}, \mathbf{s})\right]^2 \mathrm{d}\boldsymbol{\theta} < \infty$, *and, for* $h, \delta > 0$, *consider the parameterized mapping* $\boldsymbol{\theta}_0 \mapsto A_\beta(\boldsymbol{\theta}_0, h, \delta)$, *where*

$$A_\beta(\boldsymbol{\theta}_0, h, \delta) = \int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathbf{0},1)} K(\boldsymbol{\theta})R_\beta(\boldsymbol{\theta}_0 + h\boldsymbol{\theta}, \mathrm{s}_0 + \delta\mathbf{s})(\boldsymbol{\theta}, \mathrm{s})^\beta \mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{s},$$

with

$$R_\beta(\boldsymbol{\theta}_0 + h\boldsymbol{\theta}, \mathrm{s}_0 + \delta\mathbf{s}) = \int_0^1 (1-t)D^\beta f(\boldsymbol{\theta}_0 + th\boldsymbol{\theta}, \mathrm{s}_0 + t\delta\mathbf{s})\mathrm{d}t.$$

Then

$$\sup_{h,\delta>0}\int_{\mathbb{R}^p} A_\beta^2(\boldsymbol{\theta}_0, h, \delta)\mathrm{d}\boldsymbol{\theta}_0 < \infty.$$

Proof of Lemma 6.2 The proof relies on an application of the generalized Minkowski's inequality (see, e.g., Theorem 202 in Hardy et al., 1988). Indeed,

$$\begin{aligned}&\left(\int_{\mathbb{R}^p} A_\beta^2(\boldsymbol{\theta}_0, h, \delta)\mathrm{d}\boldsymbol{\theta}_0\right)^{\frac{1}{2}}\\ &\quad\leq \int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathbf{0},1)}\int_0^1 \Sigma_\beta^{1/2}(\boldsymbol{\theta}, \mathrm{s}, t)(1-t)K(\boldsymbol{\theta})\left|(\boldsymbol{\theta}, \mathrm{s})^\beta\right| \mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{s}\mathrm{d}t,\end{aligned}$$

where

$$\Sigma_\beta(\boldsymbol{\theta}, \mathrm{s}, t) = \int_{\mathbb{R}^p} \left[D^\beta f(\boldsymbol{\theta}_0 + th\boldsymbol{\theta}, \mathrm{s}_0 + t\delta\mathbf{s})\right]^2 \mathrm{d}\boldsymbol{\theta}_0.$$

Letting $C^2 = \sup_{\mathbf{s}\in\mathbb{R}^m} \int_{\mathbb{R}^p} [D^\beta f(\boldsymbol{\theta}, \mathbf{s})]^2 \mathrm{d}\boldsymbol{\theta} < \infty$, we obtain

$$\left(\int_{\mathbb{R}^p} A_\beta^2(\boldsymbol{\theta}_0, h, \delta)\mathrm{d}\boldsymbol{\theta}_0\right)^{\frac{1}{2}} \leq C \int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{0},1)} \int_0^1 (1-t)K(\boldsymbol{\theta}) \left|(\boldsymbol{\theta}, \mathbf{s})^\beta\right| \mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{s}\mathrm{d}t.$$

This upper bound is finite thanks to Assumption [**K2**], and independent of h and δ. ■

Lemma 6.3 *Let $\mathbf{s}_0$ be a Lebesgue point of $\bar{f}$ such that $\bar{f}(\mathbf{s}_0) > 0$. Then, for all positive L,*

$$0 < \inf_{0<\delta\leq L} \frac{1}{\delta^m} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} < \infty.$$

Proof of Lemma 6.3 By exploiting the fact that $\mathbf{s}_0$ is a Lebesgue point of $\bar{f}$ satisfying $\bar{f}(\mathbf{s}_0) > 0$, we deduce that for some positive $\delta_0 < L$,

$$0 < \inf_{0<\delta\leq \delta_0} \frac{1}{\delta^m} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} < \infty.$$

Moreover,

$$\frac{1}{L^m} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta_0)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} \leq \inf_{\delta_0<\delta\leq L} \frac{1}{\delta^m} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} \leq \frac{1}{\delta_0^m}.$$

The quantity on the left-hand side is positive since $\mathbf{s}_0$ belongs to the support of $\bar{f}$. This concludes the proof. ■

Proposition 6.1 *Assume that the support of $\bar{f}$ is compact with diameter $L > 0$. Let $\mathbf{s}_0$ be a Lebesgue point of $\bar{f}$ such that $\bar{f}(\mathbf{s}_0) > 0$. Set*

$$\xi_0 = \inf_{0<\delta\leq L} \frac{1}{\delta^m} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s}.$$

Whenever $\frac{k_N+1}{N+1} \leq \xi_0 L^m$, one has:

1. *For $m = 2$,*

$$\mathbb{E}\left[d^2_{(k_N+1)}\right] \leq \frac{1}{\xi_0}\left(1 + \log\left(\xi_0 L^2 \frac{N+1}{k_N+1}\right)\right)\frac{k_N+1}{N+1}.$$

2. *For $m \neq 2$,*

$$\mathbb{E}\left[d^2_{(k_N+1)}\right] \leq \frac{m}{\xi_0^{2/m}(m-2)}\left(\frac{k_N+1}{N+1}\right)^{2/m} - \frac{L^{2-m}}{\xi_0(m/2-1)}\frac{k_N+1}{N+1}.$$

Proof of Proposition 6.1 First note, according to Lemma 6.3, that $0 < \xi_0 < \infty$. Next, observe that

$$\mathbb{E}\left[d^2_{(k_N+1)}\right] = \int_0^{L^2} \mathbb{P}\left\{d_{(k_N+1)} > \sqrt{\delta}\right\} \mathrm{d}\delta.$$

For some fixed $a \in (0, L^2)$, we use the decomposition

$$\begin{aligned}
&\int_0^{L^2} \mathbb{P}\left\{d_{(k_N+1)} > \sqrt{\delta}\right\} \mathrm{d}\delta \\
&\quad = \int_0^{a} \mathbb{P}\left\{d_{(k_N+1)} > \sqrt{\delta}\right\} \mathrm{d}\delta + \int_a^{L^2} \mathbb{P}\left\{d_{(k_N+1)} > \sqrt{\delta}\right\} \mathrm{d}\delta \\
&\quad \leq a + \int_a^{L^2} \mathbb{P}\left\{d_{(k_N+1)} > \sqrt{\delta}\right\} \mathrm{d}\delta.
\end{aligned}$$

Introduce $p_0(\sqrt{\delta}) = \int_{\mathcal{B}_m(\mathbf{s}_0,\sqrt{\delta})} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s}$, which is positive since $\mathbf{s}_0$ is in the support of $\bar{f}$. Using a binomial argument, we see that

$$\begin{aligned}
\mathbb{P}\left\{d_{(k_N+1)} > \sqrt{\delta}\right\} &= \sum_{j=0}^{k_N} \binom{N}{j} \left[p_0(\sqrt{\delta})\right]^j \left[1 - p_0(\sqrt{\delta})\right]^{N-j} \\
&= \frac{1}{p_0(\sqrt{\delta})} \sum_{j=0}^{k_N} \binom{N}{j} \left[p_0(\sqrt{\delta})\right]^{j+1} \left[1 - p_0(\sqrt{\delta})\right]^{N-j}.
\end{aligned}$$

By applying Lemma 3.1 in Biau et al. (2010), we obtain

$$\mathbb{P}\left\{d_{(k_N+1)} > \sqrt{\delta}\right\} \leq \frac{k_N+1}{N+1} \times \frac{1}{p_0(\sqrt{\delta})}.$$

Consequently,

$$\mathbb{E}\left[d^2_{(k_N+1)}\right] \leq a + \frac{1}{\xi_0}\frac{k_N+1}{N+1} \int_a^{L^2} \delta^{-m/2}\mathrm{d}\delta.$$

The conclusion is easily obtained by optimizing the right-hand side with respect to the parameter a. ■

Proposition 6.2 *Assume that the support of $\bar{f}$ is compact with diameter $L > 0$. Let $\mathbf{s}_0$ be a Lebesgue point of $\bar{f}$ such that $\bar{f}(\mathbf{s}_0) > 0$. Set*

$$\xi_0 = \inf_{0<\delta\leq L} \frac{1}{\delta^m} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s}.$$

Whenever $\frac{k_N+1}{N+1} \leq \xi_0 L^m$, one has:

1. For $m = 4$,

$$\mathbb{E}\left[d^4_{(k_N+1)}\right] \leq \frac{1}{\xi_0}\left(1 + \log\left(\xi_0 L^4 \frac{N+1}{k_N+1}\right)\right)\frac{k_N+1}{N+1}.$$

2. For $m \neq 4$,

$$\mathbb{E}\left[d^4_{(k_N+1)}\right] \leq \frac{m}{\xi_0^{4/m}(m-4)}\left(\frac{k_N+1}{N+1}\right)^{4/m} - \frac{L^{4-m}}{\xi_0(m/4-1)}\frac{k_N+1}{N+1}.$$

Proof of Proposition 6.2 Proof is similar to the one of Proposition 6.1, and is therefore omitted. ■

A Complements on singular integrals

Recall that the convolution (Wheeden and Zygmund, 1977, Chapter 6) of two measurable functions f and g in $\mathbb{R}^n$ is defined by

$$(f \star g)(\mathbf{x}) = \int_{\mathbb{R}^n} f(\mathbf{y})g(\mathbf{x}-\mathbf{y})\mathrm{d}\mathbf{y}, \quad \mathbf{x} \in \mathbb{R}^n,$$

provided the integral exists. This appendix is devoted to the study of some properties of convolution when $\mathbb{R}^n = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and g is of the form

$$\varphi_{\varepsilon_1,\varepsilon_2}(\mathbf{x}) = \frac{1}{\varepsilon_1^{n_1}\varepsilon_2^{n_2}}\varphi_1\left(\frac{\mathbf{x}_1}{\varepsilon_1}\right)\varphi_2\left(\frac{\mathbf{x}_2}{\varepsilon_2}\right), \quad \mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}.$$

More precisely, the question of interest is to analyze the effect of letting ε_1 and ε_2 go independently to 0 in the expression $(f \star \varphi_{\varepsilon_1,\varepsilon_2})(\mathbf{x})$. We prove in particular (Theorem A.1) that $(f \star \varphi_{\varepsilon_1,\varepsilon_2})(\mathbf{x}) \to f(\mathbf{x})$ for λ_n-almost all $\mathbf{x}$ if f and φ are suitably restricted.

The issues discussed in the present appendix fall within the field of maximal functions and approximation of the identity (Stein, 1970; Wheeden and Zygmund, 1977). The novelty is that we allow the family $\{\varphi_{\varepsilon_1,\varepsilon_2} : \varepsilon_1 > 0, \varepsilon_2 > 0\}$ (the so-called approximation of the identity) to depend upon **two independent** parameters ε_1 and ε_2. Interestingly, the real analysis literature offers little help with respect to this important question, which is however fundamental in the study of multivariate nonparametric estimates. Valuable ideas and comments in this respect are included in Devroye and Krzyżak (2002).

Let φ be an integrable function on $\mathbb{R}^n = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, termed "the kernel" hereafter. It is assumed throughout that φ is a product kernel, of the form

$$\varphi(\mathbf{x}) = \varphi_1(\mathbf{x}_1)\varphi_2(\mathbf{x}_2), \quad \mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}. \tag{A.1}$$

For $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, we set

$$\varphi_{\varepsilon_1,\varepsilon_2}(\mathbf{x}) = \frac{1}{\varepsilon_1^{n_1}\varepsilon_2^{n_2}}\varphi_1\left(\frac{\mathbf{x}_1}{\varepsilon_1}\right)\varphi_2\left(\frac{\mathbf{x}_2}{\varepsilon_2}\right).$$

We will need the following assumption:

Assumption [K] For $i = 1, 2$, the functions

$$\psi_i(\mathbf{x}_i) = \sup_{\|\mathbf{y}_i\| \geq \|\mathbf{x}_i\|} |\varphi_i(\mathbf{y}_i)|, \quad \mathbf{x}_i \in \mathbb{R}^{n_i},$$

are in $L^1(\mathbb{R}^{n_i})$, with

$$\int_{\mathbb{R}^{n_i}} \psi_i(\mathbf{x}_i)\mathrm{d}\mathbf{x}_i \leq \sqrt{A} < \infty.$$

If f is a locally integrable function in $\mathbb{R}^n$, we also denote by $M_{12}f$ the associated Hardy-Littlewood maximal function with two degrees of freedom. It is defined for $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ by

$$(M_{12}f)(\mathbf{x}) = \sup_{\varepsilon_1,\varepsilon_2>0}\left[\frac{1}{V_{n_1}\varepsilon_1^{n_1}V_{n_2}\varepsilon_2^{n_2}}\int_{\mathcal{B}_{n_1}(\mathbf{x}_1,\varepsilon_1)}\int_{\mathcal{B}_{n_2}(\mathbf{x}_2,\varepsilon_2)}|f(\mathbf{y}_1,\mathbf{y}_2)|\,\mathrm{d}\mathbf{y}_1\mathrm{d}\mathbf{y}_2\right],$$

where $\mathcal{B}_{n_1}(\mathbf{x}_1, \varepsilon_1)$ (respectively, $\mathcal{B}_{n_2}(\mathbf{x}_2, \varepsilon_2)$) is the closed ball in $\mathbb{R}^{n_1}$ (respectively, $\mathbb{R}^{n_2}$), with center at $\mathbf{x}_1$ (respectively, $\mathbf{x}_2$) and radius ε_1 (respectively, ε_2), and V_{n_1} (respectively, V_{n_2}) is the volume of the unit ball in $\mathbb{R}^{n_1}$ (respectively, $\mathbb{R}^{n_2}$).

Our objective is to prove the following theorem, which is a more general version of Theorem 3.2.

Theorem A.1 *Let f be a measurable function in $\mathbb{R}^n$ satisfying*

$$\int_{\mathbb{R}^n} |f(\mathbf{x})|\left(1 + \log^+|f(\mathbf{x})|\right)\mathrm{d}\mathbf{x} < \infty, \tag{A.2}$$

and let φ be a product kernel of the form (A.1) satisfying Assumption [**K**]. *Assume, in addition, that $\int_{\mathbb{R}^n}\varphi(\mathbf{x})\mathrm{d}\mathbf{x} = 1$.*

(i) For λ_n-almost all $\mathbf{x} \in \mathbb{R}^n$, $\lim_{\varepsilon_1,\varepsilon_2\to 0}(f \star \varphi_{\varepsilon_1,\varepsilon_2})(\mathbf{x}) = f(\mathbf{x})$.

(ii) For λ_n-almost all $\mathbf{x} \in \mathbb{R}^n$,

$$\sup_{\varepsilon_1,\varepsilon_2>0}|(f \star \varphi_{\varepsilon_1,\varepsilon_2})(\mathbf{x})| \leq A(M_{12}f)(\mathbf{x}) < \infty,$$

where A is the constant of Assumption [**K**].

(iii) Moreover, if f is in $L^q(\mathbb{R}^n)$, $1 < q \leq \infty$, then $M_{12}f$ is in $L^q(\mathbb{R}^n)$ and

$$\|M_{12}f\|_q \leq c_q\|f\|_q,$$

where the constant c_q depends only on q and the dimension n.

Proof of Theorem A.1 To prove the theorem, we will need some general results on singular integrals and Hardy-Littlewood maximal functions. As shown in page 50 of de Guzmán (1975), for all $\alpha > 0$ and a locally integrable f,

$$\lambda_n\left(\{\mathbf{x} \in \mathbb{R}^n : (M_{12}f)(\mathbf{x}) > \alpha\}\right) \leq c \int_{\mathbb{R}^n} \frac{|f(\mathbf{x})|}{\alpha}\left(1 + \log^+ \frac{|f(\mathbf{x})|}{\alpha}\right) \mathrm{d}\mathbf{x}, \quad \text{(A.3)}$$

where c is a constant independent of f and α. This result will be crucial in our proof. It easily follows that whenever

$$\int_{\mathbb{R}^n} |f(\mathbf{x})| \left(1 + \log^+ |f(\mathbf{x})|\right) \mathrm{d}\mathbf{x} < \infty,$$

then $(M_{12}f)(\mathbf{x}) < \infty$ at λ_n-almost all $\mathbf{x}$.

Proof of *(ii)* The proof follows arguments of pages 63-64 of Stein (1970). For $i = 1, 2$, with a slight abuse of notation, we write $\psi_i(r_i) = \psi_i(\mathbf{x}_i)$ if $r_i = \|\mathbf{x}_i\|$. This should cause no confusion since each ψ_i is anyway radial. Observe that, for $i = 1, 2$,

$$\int_{r_i/2 \leq \|\mathbf{x}_i\| \leq r_i} \psi_i(\mathbf{x}_i)\mathrm{d}\mathbf{x}_i \geq \psi_i(r_i) \int_{r_i/2 \leq \|\mathbf{x}_i\| \leq r_i} \mathrm{d}\mathbf{x}_i \propto \psi_i(r_i) r_i^{n_i}.$$

Therefore, the assumption $\psi_i \in L^1(\mathbb{R}^{n_i})$ proves that $r_i^{n_i}\psi_i(r_i) \to 0$, as $r_i \to 0$ or $r_i \to \infty$. To prove *(ii)*, it is enough to show that for all nonnegative f satisfying (A.2), all $\varepsilon_1 > 0, \varepsilon_2 > 0$,

$$(f \star \psi_{\varepsilon_1,\varepsilon_2})(\mathbf{x}) \leq A(M_{12}f)(\mathbf{x}), \quad \text{(A.4)}$$

where

$$\psi_{\varepsilon_1,\varepsilon_2}(\mathbf{x}) = \frac{1}{\varepsilon_1^{n_1}\varepsilon_2^{n_2}}\psi_1\left(\frac{\mathbf{x}_1}{\varepsilon_1}\right)\psi_2\left(\frac{\mathbf{x}_2}{\varepsilon_2}\right), \quad \mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^n.$$

Set $\psi = \psi_1\psi_2$. Since assertion (A.4) is clearly translation invariant (with respect to f) and also dilatation invariant (with respect to ψ), it suffices to show that

$$(f \star \psi)(\mathbf{0}) \leq A(M_{12}f)(\mathbf{0}).$$

Moreover, recalling (A.3), we may clearly assume that $(M_{12}f)(\mathbf{0}) < \infty$. For $i = 1, 2$, denote by $S^{n_i - 1}$ the unit $(n_i - 1)$-sphere in $\mathbb{R}^{n_i}$ and let σ_i be the corresponding spherical measure. We set as well

$$\begin{aligned}
\ell(r_1, r_2) &= \int_{S^{n_1-1}} \int_{S^{n_2-1}} f(r_1\mathbf{x}_1, r_2\mathbf{x}_2) \mathrm{d}\sigma_1(\mathbf{x}_1)\mathrm{d}\sigma_2(\mathbf{x}_2), \\
\Lambda_1(r_1, r_2) &= \int_0^{r_1} \ell(u_1, r_2) u_1^{n_1-1} \mathrm{d}u_1 = \int_0^{r_2} \Lambda_1(r_1, u_2) u_2^{n_2-1} \mathrm{d}u_2 \\
&= \int_0^{r_1} \int_0^{r_2} \ell(u_1, u_2) u_1^{n_1-1} u_2^{n_2-1} \mathrm{d}u_1 \mathrm{d}u_2,
\end{aligned}$$

and will repeatedly use the inequality

$$\Lambda(r_1, r_2) = \int_{\mathcal{B}_{n_1}(\mathbf{0}, r_1)} \int_{\mathcal{B}_{n_2}(\mathbf{0}, r_2)} f(\mathbf{x}) \mathrm{d}\mathbf{x} \le V_{n_1}.V_{n_2} r_1^{n_1} r_2^{n_2} (M_{12}f)(\mathbf{0}). \qquad \text{(A.5)}$$

With this notation, we have

$$\begin{aligned}
(f \star \psi)(\mathbf{0}) &= \int_0^\infty \int_0^\infty \ell(r_1, r_2) \psi_1(r_1) r_1^{n_1-1} \psi_2(r_2) r_2^{n_2-1} \mathrm{d}r_1 \mathrm{d}r_2 \\
&= \lim_{\substack{\varepsilon_1 \to 0 \\ N_1 \to \infty \\ \varepsilon_2 \to 0 \\ N_2 \to \infty}} \int_{\varepsilon_2}^{N_2} \left[\int_{\varepsilon_1}^{N_1} \ell(r_1, r_2) \psi_1(r_1) r_1^{n_1-1} \mathrm{d}r_1 \right] \psi_2(r_2) r_2^{n_2-1} \mathrm{d}r_2.
\end{aligned}$$

Denote by $I_1(\varepsilon_1, N_1)$ the integral inside the brackets. We may write, using an integration by parts (in the sense of Stieltjès-Lebesgue),

$$I_1(\varepsilon_1, N_1) = \int_{\varepsilon_1}^{N_1} \Lambda_1(r_1, r_2) \mathrm{d}\left(-\psi_1(r_1)\right) + \Lambda_1(N_1, r_2)\psi_1(N_1) - \Lambda_1(\varepsilon_1, r_2)\psi_1(\varepsilon_1).$$

Consequently,

$$\begin{aligned}
&\int_{\varepsilon_2}^{N_2} I_1(\varepsilon_1, N_1) \psi_2(r_2) r_2^{n_2-1} \mathrm{d}r_2 = I_A + I_B - I_C \\
&\quad = \int_{\varepsilon_2}^{N_2} \int_{\varepsilon_1}^{N_1} \Lambda_1(r_1, r_2) \mathrm{d}\left(-\psi_1(r_1)\right) \psi_2(r_2) r_2^{n_2-1} \mathrm{d}r_2 \\
&\qquad + \int_{\varepsilon_2}^{N_2} \Lambda_1(N_1, r_2) \psi_1(N_1) \psi_2(r_2) r_2^{n_2-1} \mathrm{d}r_2 \\
&\qquad - \int_{\varepsilon_2}^{N_2} \Lambda_1(\varepsilon_1, r_2) \psi_1(\varepsilon_1) \psi_2(r_2) r_2^{n_2-1} \mathrm{d}r_2.
\end{aligned}$$

Each term of the sum is analyzed separately. Using again an integration by parts, we are led to

$$\begin{aligned} I_A &= \int_{\varepsilon_1}^{N_1} \Big[\int_{\varepsilon_2}^{N_2} \Lambda(r_1, r_2) \mathrm{d}\left(-\psi_2(r_2)\right) + \Lambda(r_1, N_2)\psi_2(N_2) \\ &\qquad\qquad - \Lambda(r_1, \varepsilon_2)\psi_2(\varepsilon_2) \Big] \mathrm{d}\left(-\psi_1(r_1)\right) \\ &= \int_{\varepsilon_1}^{N_1} \int_{\varepsilon_2}^{N_2} \Lambda(r_1, r_2) \mathrm{d}\left(-\psi_1(r_1)\right) \mathrm{d}\left(-\psi_2(r_2)\right) \\ &\quad + \int_{\varepsilon_1}^{N_1} \Lambda(r_1, N_2)\psi_2(N_2) \mathrm{d}\left(-\psi_1(r_1)\right) \\ &\quad - \int_{\varepsilon_1}^{N_1} \Lambda(r_1, \varepsilon_2)\psi_2(\varepsilon_2) \mathrm{d}\left(-\psi_1(r_1)\right) \\ &= A_1 + A_2 - A_3. \end{aligned}$$

The main term, A_1, is handled as follows via inequality (A.5):

$$\begin{aligned} A_1 &\leq V_{n_1}.V_{n_2}(M_{12}f)(\mathbf{0}) \int_0^\infty \int_0^\infty r_1^{n_1} r_2^{n_2} \mathrm{d}\left(-\psi_1(r_1)\right) \mathrm{d}\left(-\psi_2(r_2)\right) \\ &\leq A(M_{12}f)(\mathbf{0}) \end{aligned}$$

since for $i = 1, 2$, we have

$$V_{n_i} \int_0^\infty r_i^{n_i} \mathrm{d}\left(-\psi_i(r_i)\right) = \int_{\mathbb{R}^{n_i}} \psi_i(\mathbf{x}_i) \mathrm{d}\mathbf{x}_i \leq \sqrt{A},$$

by Assumption [**K**]. The remaining terms, A_2 and A_3, converge to 0. To see this, just note that

$$A_2 \leq V_{n_1}.V_{n_2}(M_{12}f)(\mathbf{0}) \times N_2^{n_2}\psi_2(N_2) \int_0^\infty r_1^{n_1} \mathrm{d}\left(-\psi_1(r_1)\right),$$

which goes to 0 since the integral is convergent and $N_2^{n_2}\psi_2(N_2) \to 0$ as $N_2 \to \infty$. Similarly,

$$A_3 \leq V_{n_1}.V_{n_2}(M_{12}f)(\mathbf{0}) \times \varepsilon_2^{n_2}\psi_2(\varepsilon_2) \int_0^\infty r_1^{n_1} \mathrm{d}\left(-\psi_1(r_1)\right).$$

The term on the right-hand side tends to 0 since $\varepsilon_2^{n_2}\psi_2(\varepsilon_2) \to 0$ as $\varepsilon_2 \to 0$. Using similar arguments, it is easy to prove that I_B and I_C go to 0 as $\varepsilon_1, \varepsilon_2 \to 0$ and $N_1, N_2 \to \infty$. Proof of (ii) is therefore complete.

Proof of (i) For the sake of clarity, the proof is divided into three steps.

Step 1 If f is continuous and has compact support, then the result is easy to verify. Indeed, we have in this case

$$(f \star \varphi_{\varepsilon_1,\varepsilon_2})(\mathbf{x}) = \int_{\mathbb{R}^{n_1}} \int_{\mathbb{R}^{n_2}} f(\mathbf{x}_1 - \varepsilon_1 \mathbf{y}_1, \mathbf{x}_2 - \varepsilon_2 \mathbf{y}_2) \varphi(\mathbf{y}_1, \mathbf{y}_2) \mathrm{d}\mathbf{y}_1 \mathrm{d}\mathbf{y}_2,$$

whence, using the fact that $\int_{\mathbb{R}^n} \varphi(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$,

$$\begin{aligned} &|(f \star \varphi_{\varepsilon_1,\varepsilon_2})(\mathbf{x}) - f(\mathbf{x})| \\ &\quad \leq \int_{\mathbb{R}^{n_1}} \int_{\mathbb{R}^{n_2}} |f(\mathbf{x}_1 - \varepsilon_1 \mathbf{y}_1, \mathbf{x}_2 - \varepsilon_2 \mathbf{y}_2) - f(\mathbf{x})| \,.\, |\varphi(\mathbf{y}_1, \mathbf{y}_2)| \,\mathrm{d}\mathbf{y}_1 \mathrm{d}\mathbf{y}_2 \\ &\quad \leq \sup_{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2} |f(\mathbf{x}_1 - \varepsilon_1 \mathbf{y}_1, \mathbf{x}_2 - \varepsilon_2 \mathbf{y}_2) - f(\mathbf{x})| \int_{\mathbb{R}^{n_1}} \int_{\mathbb{R}^{n_2}} |\varphi(\mathbf{y}_1, \mathbf{y}_2)| \,\mathrm{d}\mathbf{y}_1 \mathrm{d}\mathbf{y}_2. \end{aligned}$$

Since f is uniformly continuous, this term tends to 0.

Step 2 We establish that $\lim_{\varepsilon_1,\varepsilon_2 \to 0} (f \star \varphi_{\varepsilon_1,\varepsilon_2})(\mathbf{x})$ exists for λ_n-almost all $\mathbf{x} \in \mathbb{R}^n$. As for now, to ease the notation, we set $g^\star_{\varepsilon_1,\varepsilon_2}(\mathbf{x}) = (g \star \varphi_{\varepsilon_1,\varepsilon_2})(\mathbf{x})$, and let

$$(\Omega g)(\mathbf{x}) = \left| \limsup_{\varepsilon_1,\varepsilon_2 \to 0} g^\star_{\varepsilon_1,\varepsilon_2}(\mathbf{x}) - \liminf_{\varepsilon_1,\varepsilon_2 \to 0} g^\star_{\varepsilon_1,\varepsilon_2}(\mathbf{x}) \right|.$$

Let $\alpha > 0$ and $\delta > 0$ be arbitrary. Thanks to Proposition A.1 at the end of the section, we may write $f = h + g$, where h is continuous with compact support and g is such that

$$\int_{\mathbb{R}^n} \frac{|g(\mathbf{x})|}{\alpha} \left(1 + \log^+ \frac{|g(\mathbf{x})|}{\alpha}\right) \mathrm{d}\mathbf{x} \leq \delta.$$

By (ii), we have at λ_n-almost all $\mathbf{x}$, $(\Omega g)(\mathbf{x}) \leq 2A(M_{12}g)(\mathbf{x})$. Thus, by (A.3),

$$\lambda\left(\{\mathbf{x} \in \mathbb{R}^n : (\Omega g)(\mathbf{x}) > 2A\alpha\}\right) \leq c \int_{\mathbb{R}^n} \frac{|g(\mathbf{x})|}{\alpha} \left(1 + \log^+ \frac{|g(\mathbf{x})|}{\alpha}\right) \mathrm{d}\mathbf{x} \leq c\delta.$$

Clearly, $\Omega f \leq \Omega g + \Omega h$ and, by Step 1, $\Omega h \equiv 0$. Therefore

$$\lambda\left(\{\mathbf{x} \in \mathbb{R}^n : (\Omega f)(\mathbf{x}) > 2A\alpha\}\right) \leq c\delta.$$

Since α and δ are arbitrary, we conclude that $\lambda\left(\{\mathbf{x} \in \mathbb{R}^n : (\Omega f)(\mathbf{x}) > 0\}\right) = 0$.

Step 3 We finally prove that, for λ_n-almost all $\mathbf{x} \in \mathbb{R}^n$,

$$\lim_{\varepsilon_1,\varepsilon_2 \to 0} f^\star_{\varepsilon_1,\varepsilon_2}(\mathbf{x}) = f(\mathbf{x}).$$

Set $f_1(\mathbf{x}) = \lim_{\varepsilon_1,\varepsilon_2\to 0} f^\star_{\varepsilon_1,\varepsilon_2}(\mathbf{x})$ (this limit exists λ_n-almost everywhere by Step 2). Fix $\alpha > 0$, $\delta > 0$, and choose h continuous with compact support as in Step 2 such that

$$\int_{\mathbb{R}^n} \frac{|(f-h)(\mathbf{x})|}{\alpha}\left(1+\log^+ \frac{|(f-h)(\mathbf{x})|}{\alpha}\right)\mathrm{d}\mathbf{x} \le \delta.$$

For λ_n-almost all $\mathbf{x} \in \mathbb{R}^n$,

$$|f(\mathbf{x}) - f_1(\mathbf{x})| \le |f(\mathbf{x}) - h(\mathbf{x})| + |\lim_{\varepsilon_1,\varepsilon_2\to 0} h^\star_{\varepsilon_1,\varepsilon_2}(\mathbf{x}) - \lim_{\varepsilon_1,\varepsilon_2\to 0} f^\star_{\varepsilon_1,\varepsilon_2}(\mathbf{x})| = A_1 + A_2.$$

By *(ii)*,

$$A_2 \le \sup_{\varepsilon_1,\varepsilon_2>0} \left|(f-h)^\star_{\varepsilon_1,\varepsilon_2}(\mathbf{x})\right| \le A\left(M_{12}|f-h|\right)(\mathbf{x}).$$

Thus,

$$\begin{aligned}
&\lambda\left(\{\mathbf{x}\in\mathbb{R}^n : |f(\mathbf{x}) - f_1(\mathbf{x})| > 2A\alpha\}\right)\\
&\quad \le \lambda\left(\{\mathbf{x}\in\mathbb{R}^n : |f(\mathbf{x}) - h(\mathbf{x})| > A\alpha\}\right)\\
&\qquad + \lambda\left(\{\mathbf{x}\in\mathbb{R}^n : \left(M_{12}|f-h|\right)(\mathbf{x}) > \alpha\}\right)\\
&\quad \le \frac{\|f-h\|_1}{A\alpha} + c\int_{\mathbb{R}^n} \frac{|(f-h)(\mathbf{x})|}{\alpha}\left(1+\log^+ \frac{|(f-h)(\mathbf{x})|}{\alpha}\right)\mathrm{d}\mathbf{x}\\
&\quad \le \left(\frac{1}{A} + c\right)\delta.
\end{aligned}$$

In the second inequality, we used Markov's inequality together with inequality (A.3). Since both α and δ can be chosen arbitrarily, we conclude that

$$\lambda\left(\{\mathbf{x}\in\mathbb{R}^n : |f(\mathbf{x}) - f_1(\mathbf{x})| > 0\}\right) = 0.$$

Proof of *(iii)* The proof is adapted from page 307 of Zygmund (1959). Let the partial maximal functions be defined for $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ by

$$(M_1 f)(\mathbf{x}) = \sup_{\varepsilon_1>0}\left[\frac{1}{V_{n_1}\varepsilon_1^{n_1}}\int_{\mathcal{B}_{n_1}(\mathbf{x}_1,\varepsilon_1)} |f(\mathbf{y}_1,\mathbf{x}_2)|\mathrm{d}\mathbf{y}_1\right]$$

and

$$(M_2 f)(\mathbf{x}) = \sup_{\varepsilon_2>0}\left[\frac{1}{V_{n_2}\varepsilon_2^{n_2}}\int_{\mathcal{B}_{n_2}(\mathbf{x}_2,\varepsilon_2)} |f(\mathbf{x}_1,\mathbf{y}_2)|\mathrm{d}\mathbf{y}_2\right].$$

From these definitions, it is clear that $(M_{12}f)(\mathbf{x}) \le \left(M_1(M_2 f)\right)(\mathbf{x})$. But, for $1 < q \le \infty$, $f_1 \in L^q(\mathbb{R}^{n_1})$, $f_2 \in L^q(\mathbb{R}^{n_2})$, it is known (see, e.g., Stein, 1970, Theorem 1, page 5), that

$$\|M_1 f\|_q \le c_{1,q}\,\|f_1\|_q \;\text{ and }\; \|M_2 f\|_q \le c_{2,q}\|f_2\|_q,$$

where the constants $c_{1,q}$ and $c_{2,q}$ depend only on n_1, n_2 and q. It immediately follows that $\|M_{12}f\|_q^q \leq c_{1,q}^q c_{2,q}^q \|f\|_q^q$. This concludes the proof of the theorem. ■

Proposition A.1 *Let $\Phi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a continuous and nondecreasing function satisfying $\Phi(0) = 0$, and let f be a measurable function from $\mathbb{R}^n$ to $\mathbb{R}$ such that $\int_{\mathbb{R}^n} \Phi\left(|f(\mathbf{x})|\right) \mathrm{d}\mathbf{x} < \infty$. Then, for all $\delta > 0$, there exists a function h continuous with compact support such that*

$$\int_{\mathbb{R}^n} \Phi\left(|f(\mathbf{x}) - h(\mathbf{x})|\right) \mathrm{d}\mathbf{x} \leq \delta.$$

Proof of Proposition A.1 First, assume that $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$. Take $\{f_t\}$ a sequence of nonnegative continuous functions, each with compact support and such that $0 \leq f_t(\mathbf{x}) \uparrow f(\mathbf{x})$ at λ_n-almost all $\mathbf{x} \in \mathbb{R}^n$. For such an $\mathbf{x}$, by the continuity of Φ at 0, one has $\Phi(f(\mathbf{x}) - f_t(\mathbf{x})) \rightarrow \Phi(0) = 0$. Since $\Phi(f(\mathbf{x}) - f_t(\mathbf{x})) \leq \Phi(f(\mathbf{x}))$ and $\Phi(f)$ is in $L^1(\mathbb{R}^n)$ by assumption, we may apply Lebesgue's dominated convergence theorem and conclude that

$$\int_{\mathbb{R}^n} \Phi\left(f(\mathbf{x}) - f_t(\mathbf{x})\right) \mathrm{d}\mathbf{x} \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

If we drop the assumption that $f(\mathbf{x}) \geq 0$, we may split f into positive and negative part and apply the above result. ■

Acknowledgments We thank two anonymous referees for valuable comments and insightful suggestions.

References

I.S. Abramson. On bandwidth variation in kernel estimates—a square root law. *The Annals of Statistics*, 10:1217–1223, 1982.

D.M. Bashtannyk and R.J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics and Data Analysis*, 36:279–298, 2001.

M. Beaumont, J.-M. Cornuet, J.-M. Marin, and C.P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96:983–990, 2009.

M.A. Beaumont, W. Zhang, and D.J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.

G. Biau, F. Cérou, and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11:687–712, 2010.

M. Blum. Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association*, 105:1178–1187, 2010.

L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19:135–144, 1977.

F. Cérou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355, 2006.

T.M. Cover. Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:50–55, 1968.

M. de Guzmán. *Differentiation of Integrals in* $\mathbb{R}^n$, volume 481 of *Lecture Notes in Mathematics*. Springer, Berlin, 1975.

L. Devroye. Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61:467–481, 1982.

L. Devroye and A. Krzyżak. New multivariate product density estimates. *Journal of Multivariate Analysis*, 82:88–110, 2002.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

J. Fan and T.H. Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 94:819–834, 2004.

O.P. Faugeras. A quantile-copula approach to conditional density estimation. *Journal of Multivariate Analysis*, 100:2083–2099, 2009.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B*, 74:419–474, 2012.

E. Fix and J.L. Hodges. *Discriminatory analysis—Nonparametric discrimination: Consistency properties*. Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, pages 261-279, Randolph Field, 1951.

Y.X. Fu and W.H. Li. Estimating the age of the common ancestor of a sample of DNA sequences. *Journal of Molecular Biology and Evolution*, 14:195–199, 1997.

L. Györfi and M. Kohler. Nonparametric estimation of conditional distributions. *IEEE Transactions on Information Theory*, 53:1872–1879, 2007.

P. Hall and J.S. Marron. Variable window width kernel estimates of probability densities. *Probability Theory and Related Fields*, 80:37–49, 1988.

P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99:1015–1026, 2004.

B.H. Hansen. *Nonparametric conditional density estimation.* Technical Report, University of Wisconsin, 2004. `http://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf`.

G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities.* Cambridge University Press, Cambridge, 1988.

W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

R.J. Hyndman, D.M. Bashtannyk, and G.K. Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5:315–336, 1996.

B. Jessen, J. Marcinkiewicz, and A. Zygmund. Note on the differentiability of multiple integrals. *Fundamenta Mathematicae*, 25:217–234, 1935.

M.C. Jones. Variable kernel density estimates and variable kernel density estimates. *Australian Journal of Statistics*, 32:361–371, 1990.

P. Joyce and P. Marjoran. Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(26), 2008.

E. Kaufmann and R.-D. Reiss. On conditional distributions of nearest neighbors. *Journal of Multivariate Analysis*, 42:67–76, 1992.

D.O. Loftsgaarden and C.P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36: 1049–1051, 1965.

Y.P. Mack and M. Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9:1–15, 1979.

J.M. Marin and C.P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer, New York, 2007.

J.M. Marin, N. Pillai, C.P. Robert, and J. Rousseau. *Relevant statistics for Bayesian model choice*. arXiv:1110.4700, 2011.

J.M. Marin, P. Pudlo, C.P. Robert, and R. Ryder. Approximate Bayesian Computational methods. *Statistics and Computing*, 22:1167–1180, 2012.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

D.S. Moore and J.W. Yackel. Consistency properties of nearest neighbor density function estimators. *The Annals of Statistics*, 5:143–154, 1977a.

D.S. Moore and J.W. Yackel. Large sample properties of nearest neighbor density function estimators. In S.S. Gupta and D.S. Moore, editors, *Statistical Decision Theory and Related Topics II: Proceedings of a Symposium Held at Purdue University, May 17-19, 1976*, pages 269–279, New York, 1977b. Academic Press.

E.A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.

E.A. Nadaraya. On nonparametric estimates of density functions and regression curves. *Theory of Probability and its Applications*, 10:186–190, 1965.

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

J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun, and M.W. Feldman. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16:1791–1798, 1999.

B.D. Ripley. *Stochastic Simulation*. John Wiley & Sons, New York, 1982.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods (2nd ed.)*. Springer, New York, 2004.

C.P. Robert, J.-M. Cornuet, J.-M. Marin, and N.S. Pillai. Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108:15112–15117, 2011.

M. Rosenblatt. Conditional probability density and regression estimates. In P.R. Krishnaiah, editor, *Multivariate Analysis II*, pages 25–31, New York, 1969. Academic Press.

R.M. Royall. *A class of non-parametric estimates of a smooth regression function*. Technical Report 14, Stanford University, 1966.

D. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12:1151–1172, 1984.

S.A. Sisson, Y. Fan, and M.M. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104:1760–1765, 2007.

E.M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, Princeton, 1970.

C.J. Stone. Consistent nonparametric regression (with discussion). *The Annals of Statistics*, 5:595–645, 1977.

S. Tavaré, D. Balding, R. Griffith, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518, 1997.

G.S. Watson. Smooth regression analysis. *Sankhyā Series A*, 26:359–372, 1964.

R.L. Wheeden and A. Zygmund. *Measure and Integral. An Introduction to Real Analysis*. Marcel Dekker, New York, 1977.

R.D. Wilkinson. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12:129–141, 2008.

A. Zygmund. *Trigonometric Series. Vol. II.* Cambridge University Press, Cambridge, 1959.