A new adaptive local polynomial density estimation procedure on complicated domains

Karine Bertin, Nicolas Klutchnikoff, Frédéric Ouimet

Submitted to Bernoulli

A new adaptive local polynomial density estimation procedure on complicated domains

KARINE BERTIN[1,a], NICOLAS KLUTCHNIKOFF[2,b] and FRÉDÉRIC OUIMET[3,4,c]

[1] *CIMFAV-INGEMAT, Universidad de Valparaíso, Valparaíso, Chile,* [a] *karine.bertin@uv.cl*
[2] *Univ Rennes, CNRS, IRMAR-UMR 6625, F-35000 Rennes, France,* [b] *nicolas.klutchnikoff@univ-rennes2.fr*
[3] *Centre de recherches mathématiques, Université de Montréal, Montreal, Canada*
[4] *Department of Mathematics and Statistics, McGill University, Montreal, Canada,* [c] *frederic.ouimet2@mcgill.ca*

This paper presents a novel approach for pointwise estimation of multivariate density functions on known domains of arbitrary dimensions using nonparametric local polynomial estimators. Our method is highly flexible, as it applies to both simple domains, such as open connected sets, and more complicated domains that are not star-shaped around the point of estimation. This enables us to handle domains with sharp concavities, holes, and local pinches, such as polynomial sectors. Additionally, we introduce a data-driven selection rule based on the general ideas of Goldenshluger and Lepski. Our results demonstrate that the local polynomial estimators are minimax under a L^2 risk across a wide range of Hölder-type functional classes. In the adaptive case, we provide oracle inequalities and explicitly determine the convergence rate of our statistical procedure. Simulations on polynomial sectors show that our oracle estimates outperform those of the most popular alternative method, found in the `sparr` package for the `R` software. Our statistical procedure is implemented in an online `R` package which is readily accessible.

Keywords: Adaptive estimation; complicated domain; concave domain; local polynomial; minimax; nonparametric density estimation; oracle inequality; pinched domain; pointwise risk; polynomial sector

1. Introduction

In practice, estimating a probability density function near or at the boundary of its support $\mathcal{D}$ is often challenging. This issue is particularly prominent in standard kernel density estimators, which suffer from a well-known bias that negatively impacts their performance. In situations where $\mathcal{D}$ represents a geometrically 'simple' domain in $\mathbb{R}^d$, this behavior is more indicative of limitations in these statistical procedures rather than an inherently difficult estimation problem. Consequently, dedicated estimators have been devised to reduce or eliminate the bias term, providing more accurate density estimates.

For instance, Bernstein estimators, introduced by Vitale (1975) and studied further among others by Babu and Chaubey (2006), Gawronski and Stadtmüller (1981), Ghosal (2001), Petrone (1999), Petrone and Wasserman (2002), utilize a specific choice of discrete kernel function whose shape adapts locally to the support of the density and leads to reduced bias near the boundaries. Asymmetric kernel estimators, put forward independently by Aitchison and Lauder (1985) and Chen (1999), extend the same idea to smooth kernels. The reflection method, proposed by Schuster (1985) and further investigated by Cline and Hart (1991), involves extending the support of the density function symmetrically beyond the observed data range and incorporating reflected data points into the estimation process. Boundary kernel estimators, initially proposed by Gasser and Müller (1979) and further refined by Gasser, Müller and Mammitzsch (1985), Jones (1993), Müller (1991), Zhang and Karunamuni (1998, 2000), assign higher weights to data points located near the boundaries, effectively giving greater emphasis to the boundary regions during estimation and improving accuracy. Local polynomial smoothers with variable bandwidths, introduced by Fan and Gijbels (1992) in the regression context, see also Cheng, Fan

and Marron (1997), Fan and Gijbels (1996), fit low-degree polynomials to local neighborhoods of data points, adaptively adjusting the polynomial degree and bandwidth to mitigate boundary bias. Other notable techniques include smoothing splines (Gu, 1993, Gu and Qiu, 1993), generalized jackknifing (Jones, 1993, Jones and Foster, 1996), and the transformation technique (Marron and Ruppert, 1994, Ruppert and Cline, 1994), which applies a transformation to the data before smoothing on a more suited domain. A combination that leverages the benefits of both reflection and transformation was proposed by Zhang, Karunamuni and Jones (1999). For a modern review of kernel density estimation methods and bandwidth selection for spatial data, refer to Davies, Marshall and Hazelton (2018).

In spite of the remarkable strides made throughout the years to alleviate the boundary bias of kernel density estimators, it often comes at the cost of producing other issues such as estimators that lack local adaptivity (i.e., are sensitive to outliers, generate spurious bumps, etc.), inadequate data-driven bandwidth selection rules (e.g., plug-in methods with reference/pilot densities that spill over boundaries), or density estimates that are not bona fide density functions (i.e., that are negative in some regions or do not integrate to one). Only recently, some methods have been able to address all these concerns at once.

For example, Botev, Grotowski and Kroese (2010) propose an adaptive kernel density estimator which is the solution, in a space-scale setting, to the Fokker-Planck partial differential equation (PDE) associated with an Itô diffusion process having the empirical density function as initial condition. The scale variable plays the role of the bandwidth and the shape specifics of the domain are taken into account directly using the boundary conditions of the PDE. Their proposed estimator is locally adaptive, has a bandwidth selection procedure which is free of the so-called normal reference rules (and thus is truly nonparametric), deals well with boundary bias and is always a bona fide density function. Further, the overall computational cost remains low. Yet, as pointed out by Ferraccioli et al. (2021), their estimator is unable to handle domains with complicated geometries.

In recent efforts to tackle the above concerns regarding kernel density estimators, specific focus has been directed towards estimating multivariate density functions supported on domains with complicated geometries. Notably, these efforts include Gaussian field approaches Bakka et al. (2019), Lindgren, Rue and Lindström (2011), Niu et al. (2019), PDE regularization schemes Arnone et al. (2022), Azzimonti et al. (2014, 2015), Ferraccioli et al. (2021), Sangalli (2021), shape constraints Feng et al. (2021), Xu and Samworth (2021), soap film smoothing Wood, Bravington and Hedley (2008), spline methods Guillas and Lai (2010), Lai and Schumaker (2007), Lai and Wang (2013), Miller and Wood (2014), Ramsay (2002), Sangalli, Ramsay and Ramsay (2013), Wang and Ranalli (2007), Yu et al. (2021), Zhou and Pan (2014), diffusion kernels Barry and McIntyre (2011, 2020), McIntyre and Barry (2018), McSwiggan, Baddeley and Nair (2017), and other hybrid methods such as the complex region spatial smoother (CReSS) of Scott-Hayward et al. (2015) based on improved geodesic distance estimation.

Unfortunately, all these methods are either intrinsically adapted for global estimation instead of pointwise estimation (and thus cannot always uphold the best choice of bandwidth near specific points), or they have only been developed, implemented and studied for spatial data in two or three dimensions. For instance, the state of the art PDE regularization method on Euclidean spaces is the one by Ferraccioli et al. (2021). In that paper, the authors introduce a log-likelihood approach with Laplacian operator regularization that penalizes high local curvatures and thus controls the global roughness of the estimates. Their method can handle complicated shapes, sharp concavities and holes. They also improve on the method of Botev, Grotowski and Kroese (2010) by discretizing the aforementioned Fokker-Planck equation and solving numerically using a forward Euler integration scheme, which allows for complicated domain shapes. The problem is that their global choice of regularization parameter limits refinements in the vicinity of any particular point in the domain. Furthermore, their general method relies crucially on a triangulation of the space, which is only numerically viable in low dimensions.

The other approaches have the same issue, as they also rely on a global regularization or a low domain dimension. For the few methods in arbitrary dimensions (e.g., Barry and McIntyre (2020), Niu et al. (2019)), some face practical challenges or lack theoretical development.

The present paper addresses these gaps in the literature by introducing a new adaptive nonparametric local polynomial estimator suitable for the pointwise estimation of multivariate density functions supported on known domains of arbitrary dimensions and investigating its theoretical properties and practical implementation. Our method is highly flexible, as it applies to both simple domains, such as open connected sets, and more complicated domains that are not star-shaped around the point of estimation. This enables us to deal with domains containing sharp concavities, holes, and local pinches, such as polynomial sectors. To elaborate, our approach involves estimating the density function locally using the minimizer of a continuous kernel-weighted least-squares problem on a space of polynomials. Our selection procedure follows the general ideas of Lepskii (1991) and Goldenshluger and Lepski (2008, 2009, 2011, 2014). It consists of jointly optimizing the polynomial degree and bandwidth over a discrete set and selecting the pair that minimizes the sum of upper bounds on the bias and a penalized version of the standard deviation. Importantly, our estimator does not require any tuning parameters, making it genuinely nonparametric. Moreover, it addresses the classical issues of kernel density estimators mentioned earlier, as it is locally adaptive, exhibits relatively small boundary bias regardless of the domain's shape, and is asymptotically a bona fide density function. For earlier advancements in density estimation on simple bounded domains using a similar Goldenshluger-Lepski selection procedure, refer to Bertin, El Kolei and Klutchnikoff (2019), Bertin and Klutchnikoff (2014), and also to Bertin and Klutchnikoff (2017), Bertin et al. (2020) under the assumption of weakly dependent observations.

Although our bandwidth selection procedure differs significantly, a similar concept was recently developed by Cattaneo, Jansson and Ma (2020) for estimating univariate density functions using local polynomials. Their approach involves taking the slope factor of the linear monomial in a kernel-weighted L^2 projection of the empirical cumulative distribution function (cdf) onto a polynomial space. In simpler terms, they project the cdf and identify the derivative of the projection (i.e., the density estimator), while our method directly projects the density function itself. Their estimator is boundary adaptive, does not require prior knowledge of the boundary points, and always yields a valid density function. Moreover, their implementation is simple, requiring no specific data modifications or additional tuning parameters. For the `R` software implementation, refer to Cattaneo, Jansson and Ma (2022), and for an extension of their method to conditional univariate density functions, see Cattaneo et al. (2023). However, one serious limitation of their method is its unclear generalizability to multivariate density functions, as explained in the following remark.

Remark 1. In private email communications, Cattaneo, Jansson, and Ma informed us that they had previously considered a multivariate version of their estimator but were unable to determine the asymptotic variance expression. Consequently, they could not implement the pointwise approximate MSE-optimal bandwidth selection procedure. We attempted to address this issue ourselves but were unsuccessful. Thus, the development of a multivariate version of their estimator remains an open problem, meriting further research.

Here is an outline of the paper. Section 2 provides prerequisite definitions and information on the pointwise L^2 risk measure, domains and Hölder classes considered for density functions, as well as the concept of adaptive rate of convergence. Section 3 presents a detailed description of our statistical procedure, which involves a kernel-weighted L^2 projection of the density function onto a polynomial space, and the application of the Goldenshluger-Lepski method for the joint selection of the polynomial degree and bandwidth. Section 4 states our main results. For the base local polynomial estimator, this corresponds to explicit bounds on its bias and variance, and its minimaxity on simple and complicated

domains. For our full statistical procedure (i.e., including the selection procedure), we derive an oracle-type inequality and the estimator is shown to achieve the optimal adaptive rate of convergence on simple and complicated domains. To aid understanding, the results for complicated domains are specifically stated for polynomial sectors. The case of unknown compact convex domains is also briefly discussed in Section 4.4. Section 5 complements the theoretical results with simulations on polynomial sectors. Our oracle estimates are shown to outperform those of the most popular alternative method, found in the `sparr` package Davies and Marshall (2023) for the `R` software. For convenience, our statistical procedure is implemented in an online `R` package called `densityLocPoly`, see Bertin, Klutchnikoff and Ouimet (2024). The proofs are all deferred to the appendices. Specifically, Appendix A contains preliminary technical lemmas and their proofs, whereas Appendix B delves into the main proofs.

2. Framework

Define the domain $\mathcal{D}$ as the closure of a nonempty open subset of $\mathbb{R}^d$ endowed with the sup norm, denoted by $\|\cdot\|_\infty$. Our focus will be on studying our estimator locally at a point $t \in \mathcal{D}$, inside its surrounding neighborhoods of radius $h \in (0,\infty)$,

$$\mathcal{V}(h) = \left\{u \in \mathbb{R}^d : t+u \in \mathcal{D} \text{ and } \|u\|_\infty \leq h\right\} = (\mathcal{D}-t) \cap (h\Delta),$$

where $\Delta = [-1,1]^d$. We assume that there exists a parameter $\rho \in (0, e^{-1}]$ such that the set $\mathcal{V}(\rho)$ is a neighborhood of the origin for the topology on $\mathcal{D}$ inherited from $\mathbb{R}^d$. Since $\mathcal{D}$, d, t and ρ are fixed throughout the paper, the dependence on these quantities is often omitted for readability.

For $n \in \mathbb{N} = \{1,2,\ldots\}$, let $\mathbb{X}_n = (X_1,\ldots,X_n)$ be a vector of n independent copies of a random variable X which is supported on $\mathcal{D}$ and has a probability density function f with respect to the Lebesgue measure on $\mathbb{R}^d$, hereafter denoted by $\mathrm{Leb}(\cdot)$. Our goal is to estimate the density function f at the fixed point $t \in \mathcal{D}$. By estimator, we mean any map $\tilde{f} = S \circ \mathbb{X}_n$ such that $S : \mathcal{D}^n \to \mathbb{R}^{\mathbb{R}^d}$ is Borel-measurable. The accuracy of such an estimator is measured by the pointwise risk

$$R_n(\tilde{f}, f) = \left(\mathsf{E}\left[\left\{\tilde{f}(t) - f(t)\right\}^2\right]\right)^{1/2},$$

where E denotes the expectation with respect to the probability measure P of the random sample $\mathbb{X}_n$.

Remark 2. Utilizing a pointwise L^2 risk measure is not inherently essential for deriving our findings. However, this choice facilitates the mathematical analysis across various proofs by decomposing the squared risk into two distinct components: the squared bias and the variance:

$$\left\{R_n(\tilde{f}, f)\right\}^2 = \left[\mathsf{E}\left\{\tilde{f}(t)\right\} - f(t)\right]^2 + \mathsf{Var}\left\{\tilde{f}(t)\right\}.$$

This breakdown simplifies the determination of a suitable bandwidth parameter when seeking the optimal convergence rate of our statistical procedure. Such arguments feature prominently in the proofs of our minimax and adaptivity results; see Propositions 4.4 and 4.6, as well as Theorems 4.5 and 4.7. More generally, one could explore a pointwise L^p risk measure for a fixed $p \geq 1$. However, this approach necessitates additional steps at various junctures to accommodate the imperfect decomposition:

$$\mathsf{E}\left[\left\{\tilde{f}(t) - f(t)\right\}^p\right] \leq 2^{p-1} \times \left\{\left|\mathsf{E}\left\{\tilde{f}(t)\right\} - f(t)\right| + \mathsf{E}\left(\left[\tilde{f}(t) - \mathsf{E}\left\{\tilde{f}(t)\right\}\right]^p\right)\right\}.$$

For instance, we would need to apply Khintchine-type inequalities to estimate the second expectation on the right-hand side in terms of the variance. Modifications to the proof of Theorem 4.3 would also

be required, shifting the focus from $\mathsf{E}(T^2)$ to $\mathsf{E}(T^p)$, etc. While it is possible to consider a more general risk measure, doing so would inevitably convolute the mathematical methods and ideas used in deriving our results with additional technical intricacies. Therefore, this avenue is not further explored here and is left open for future investigation.

Since there is no ambiguity, we use P and E instead of the more cumbersome notations $\mathsf{P}_{f,n}$ and $\mathsf{E}_{f,n}$. However, we will maintain the dependency on n for the pointwise risk R_n as we will investigate its asymptotic behavior in Section 4. Specifically, we will study the minimax and adaptive pointwise estimation of f over a large collection of Hölder-type functional classes. The remainder of this section covers the fundamental concepts that are used in this paper.

2.1. Hölder-type classes

First, consider a collection of polynomials, each having a degree which is less than or equal to a specified integer.

Definition 1. Let $m \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and define

$$\mathcal{P}_m = \text{Span}\left(\varphi_\alpha : \alpha \in \mathbb{N}_0^d \text{ and } |\alpha| \le m\right),$$

where, for any $u \in \mathbb{R}^d$, we have the monomials

$$\varphi_\alpha(u) = \prod_{j=1}^{d} u_j^{\alpha_j} \quad \text{with} \quad |\alpha| = \sum_{j=1}^{d} \alpha_j.$$

In particular, note that $\varphi_0 \equiv 1$.

Second, for any smoothness parameter $s \in (0, \infty)$ and any Lipschitz constant $L \in (0, \infty)$, define the Hölder-type functional class $\Sigma(s, L)$ inside which a density function f can be approximated by a polynomial of prescribed degree in a neighborhood of $t \in \mathcal{D}$.

Definition 2. Let $(s, L) \in (0, \infty)^2$ be given. A density function $f : \mathcal{D} \to \mathbb{R}$ belongs to the Hölder class $\Sigma(s, L)$ if there exists a polynomial $q \in \mathcal{P}_{\lfloor\!\lfloor s \rfloor\!\rfloor}$ such that, for any $u \in \mathcal{V}(\rho)$,

$$|f(t+u) - q(u)| \le L\|u\|_\infty^s. \tag{2.1}$$

Here, $\lfloor\!\lfloor s \rfloor\!\rfloor$ denotes the greatest integer less than s. In particular, $\lfloor\!\lfloor s \rfloor\!\rfloor = s - 1$ if $s \in \mathbb{N}$.

Remark 3. The functional class $\Sigma(s, L)$ is not commonly used in statistics. However, we believe that it offers more flexibility than classical Hölder spaces in the context of a domain $\mathcal{D}$ with a complicated boundary. A few comments are in order:

1. Knowing that a density function f belongs to the Hölder class $\Sigma(s, L)$ only provides local information (in a neighborhood of t) about the regularity of f. Note also that the polynomial q satisfies $q(0) = f(t)$.
2. If the partial derivatives $D^\alpha f = \partial^{|\alpha|} f / \partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}$ exist for all $|\alpha| \le \lfloor\!\lfloor s \rfloor\!\rfloor$ and if the interior of $\mathcal{V}(\rho)$, in the topology of $\mathbb{R}^d$, is star-shaped with respect to the origin, then the Taylor polynomial of f at t can be substituted for q in (2.1). In particular, this is true if the interior of $\mathcal{V}(\rho)$ is an open convex subset of $\mathbb{R}^d$ that contains the origin.

3. The existence of q does not imply the existence of directional derivatives for f. In particular, for some points on the boundary of the domain $\mathcal{D}$, the partial derivatives (in the directions given by the canonical basis) may not even make sense. Let us give two simple examples of densities defined on a disk sector to illustrate these assertions. Define

$$\mathcal{D} = \left\{(r\cos\theta, r\sin\theta) \in \mathbb{R}^2 : 0 \le r \le 1, \frac{\pi}{6} \le \theta \le \frac{\pi}{3}\right\}.$$

The density function $f_1(x,y) = 12/\pi \times \mathbb{1}_{\mathcal{D}}(x,y)$ is constant on $\mathcal{D}$, but the partial derivatives do not exist at $0 \in \mathcal{D}$. Similarly, the density function $f_2(x,y) = Ag(x+y)\mathbb{1}_{\mathcal{D}}(x,y)$, where $g(u) = 1+u+u^2+u^3\sin(1/u)$ and A is a normalization constant, can be approximated in a neighborhood of $0 \in \mathcal{D}$ by a polynomial of degree 2, even if no directional derivatives of order 2 exist at the origin.

Remark 4. Because the degrees of the polynomials in the basis $\mathcal{P}_{\lfloor\!\lfloor s \rfloor\!\rfloor}$ are restricted by the value $\lfloor\!\lfloor s \rfloor\!\rfloor$ across all dimensions, the functional class $\Sigma(s,L)$ can be categorized as isotropic. One might ponder the possibility of defining an anisotropic extension of $\Sigma(s,L)$ to provide greater flexibility. Here is one approach: For any integers $m_1,\dots,m_d \in \mathbb{N}_0$, define the new polynomial basis

$$\mathcal{P}_{m_1,\dots,m_d} = \text{Span}\left(\varphi_\alpha : \alpha \in \mathbb{N}_0^d \text{ and } \alpha_i \le m_i \ \ \forall i \in \{1,\dots,d\}\right).$$

Then, for any smoothness parameters $s_1,\dots,s_d \in (0,\infty)$ and any Lipschitz constant $L \in (0,\infty)$, let the new anisotropic Hölder class $\Sigma(s_1,\dots,s_d,L)$ be defined by the polynomials $q \in \mathcal{P}_{\lfloor\!\lfloor s_1 \rfloor\!\rfloor,\dots,\lfloor\!\lfloor s_d \rfloor\!\rfloor}$ such that, for any $u \in \mathcal{V}(\rho)$,

$$|f(t+u) - q(u)| \le L(|u_1|^{s_1} + \cdots + |u_d|^{s_d}).$$

This new definition would naturally lead in Section 3 to the adoption of a more general parameter $\gamma = (m_1,\dots,m_d,h_1,\dots,h_d)$ for our adaptive statistical procedure, where the m_i's and h_i's specify the estimator's smoothness and scale of estimation in each dimension, respectively. Consequently, numerous technical adjustments would be necessary in the statements of the results and their proofs, such as the additivity of the bias terms $h_i^{\beta_{m_i}(s_i)}$ in Proposition 4.1. While the conceptual framework would remain largely unchanged, these technical modifications would obscure the mathematical methodologies and ideas underpinning the derivation of the new results. This avenue is not further explored here, remaining open for future research. For recent advancements in multivariate density estimation within an anisotropic framework, refer to, e.g., Ammous, Dedecker and Duval (2023), Bertin, El Kolei and Klutchnikoff (2019), Liu and Wu (2019), Rebelles (2015), Varet et al. (2023).

2.2. Adaptive estimation

As mentioned at the beginning of Section 2, we are interested in the pointwise adaptive estimation, in a minimax framework, of the density function f over the collection of Hölder-type functional classes $\Sigma(s,L)$, where the nuisance parameter (s,L) is assumed to belong to $\mathcal{K} = (0,\infty)^2$. Since the results of Lepskii (1991), it is well-known that, in this context, it is impossible to construct a single estimation procedure $f_n^\star$ which attains the minimax rate of convergence

$$N_n(s,L) = \inf_{\tilde{f}_n} \sup_{f \in \Sigma(s,L)} R_n(\tilde{f}_n, f) \tag{2.2}$$

simultaneously for all functional classes, i.e., which satisfies

$$\sup_{f\in\Sigma(s,L)} R_n(f_n^\star, f) \lesssim N_n(s,L), \quad \text{for all } (s,L)\in\mathcal{K}. \tag{2.3}$$

In (2.2), the infimum is taken over all possible estimators $\tilde{f}_n$ of f, while the general notation $u_n \lesssim v_n$ used in (2.3) means that $u_n, v_n \geq 0$ and $\limsup_{n\to\infty} u_n/v_n < \infty$.

It is therefore necessary to define a specific notion of adaptive rate of convergence (ARC) before trying to construct an optimal estimation procedure. Several definitions were proposed in the last decades, see, e.g., Lepskii (1991), Tsybakov (1998), Klutchnikoff (2014) and Rebelles (2015). In this paper we adopt the definition of Klutchnikoff (2014), which coincides in our framework with the refined version introduced by Rebelles (2015). In order to be self-contained, we recall the definition of admissibility of a collection of normalizations and the notion of ARC employed in these two papers. It is always assumed that collections such as $\phi = \{\phi_n(s,L) : (s,L)\in\mathcal{K}, n\in\mathbb{N}\}$ have nonnegative components, i.e., $\phi_n(s,L)\geq 0$ for all $(s,L)\in\mathcal{K}$ and all $n\in\mathbb{N}$.

Definition 3. A collection $\phi = \{\phi_n(s,L) : (s,L)\in\mathcal{K}, n\in\mathbb{N}\}$ is said to be admissible if there exists an estimator $\tilde{f}_n$ such that

$$\sup_{f\in\Sigma(s,L)} R_n(\tilde{f}_n, f) \lesssim \phi_n(s,L), \quad \text{for all } (s,L)\in\mathcal{K}.$$

Given two collections $\phi = \{\phi_n(s,L) : (s,L)\in\mathcal{K}, n\in\mathbb{N}\}$ and $\psi = \{\psi_n(s,L) : (s,L)\in\mathcal{K}, n\in\mathbb{N}\}$, we define two subsets of $\mathcal{K}$, namely

$$[\psi \ll \phi] = \left\{(s,L)\in\mathcal{K} : \lim_{n\to\infty} \frac{\psi_n(s,L)}{\phi_n(s,L)} = 0\right\},$$

and

$$[\psi \lll \phi] = \left\{(s,L)\in\mathcal{K} : \limsup_{n\to\infty} \frac{\psi_n(s,L)}{\phi_n(s,L)} \times \frac{\psi_n(s',L')}{\phi_n(s',L')} = \infty, \ \ \forall (s',L')\in[\psi\ll\phi]\right\}.$$

Equipped with these notations, the notion of ARC is defined as follows.

Definition 4. An admissible collection $\phi = \{\phi_n(s,L) : (s,L)\in\mathcal{K}, n\in\mathbb{N}\}$ is called an ARC if for any other admissible collection $\psi = \{\psi_n(s,L) : (s,L)\in\mathcal{K}, n\in\mathbb{N}\}$, we have

- $[\psi \ll \phi]$ is contained in a $(\dim(\mathcal{K})-1)$–dimensional manifold.
- $[\psi \lll \phi]$ contains an open subset of $\mathcal{K}$.

Remark 5. In our context, where $\mathcal{K} = (0,\infty)^2$, we can grasp the essence of Definition 4 heuristically as follows: Given an admissible collection ϕ, the set $[\psi \ll \phi]$ comprises collections ψ exhibiting a superior rate of convergence compared to ϕ. On the other hand, $[\psi \lll \phi]$ encapsulates collections that suffer from a worse rate of convergence than all the multiplicative inverse gaps observed in $[\psi \ll \phi]$. Consequently, Definition 4 designates ϕ as an ARC if $[\psi \ll \phi]$, akin to an optimal frontier, manifests as 1-dimensional. This implies that for a given s, the convergence rate of $\phi_n(s,L)$ cannot be enhanced. Conversely, the condition that $[\psi \lll \phi]$ contains an open subset of $(0,\infty)^2$ suggests ample space for collections exhibiting sub-optimal convergence rates in comparison to ϕ. To illustrate visually, envision s along the x-axis and L along the y-axis. It is shown in Theorem 4.5 that on simple domains, if ϕ is an ARC, then the set $[\psi \ll \phi]$ is equal to the vertical line $\{s_0\}\times(0,\infty)$ for an appropriate $s_0 > 0$, and $[\psi \lll \phi]$ corresponds to the half-space $(s_0,\infty)\times(0,\infty)$.

Let us recall that the ARC is unique up to asymptotic order, i.e., if ϕ and ψ are two ARCs then, for any $(s, L) \in \mathcal{K}$, we have

$$\phi_n(s, L) \lesssim \psi_n(s, L) \quad \text{and} \quad \psi_n(s, L) \lesssim \phi_n(s, L).$$

Henceforth, we refer to an ARC as *the* ARC using the above identification. Further, note that if the collection of minimax rates $N = \{N_n(s, L) : (s, L) \in \mathcal{K}, n \in \mathbb{N}\}$ is admissible, then it is the ARC.

3. Statistical procedure

In this section, we present a novel estimation procedure which is both simple to understand and completely free of any tuning parameters, thus eliminating the need for a sophisticated calibration step. The main idea behind our approach is to use polynomials of varying degrees to approximate the density function f in different neighborhoods of t and then to simultaneously select the optimal neighborhood and degree of approximation in a data-driven manner. Let us formalize this idea.

3.1. Local polynomial estimators

Our collection of local polynomial estimators will be indexed by a two-dimensional parameter

$$\gamma = (m, h) \in \Gamma = \mathbb{N}_0 \times (0, \rho],$$

where m represents the degree of the polynomial and h denotes the bandwidth, the latter of which determines the size of the neighborhoods. Let $\mathbb{1}_A$ be the indicator function for any set $A \subseteq \mathbb{R}^d$. Recall that $\Delta = [-1, 1]^d$, and consider the kernel function

$$K(u) = \mathbb{1}_\Delta(u), \quad u \in \mathbb{R}^d. \tag{3.1}$$

Note that the kernel K does not have to be a density function. While other positive kernels could be used, we limit our study to this particular kernel for the sake of simplicity. Let us define the polynomial p_γ as the solution to the following continuous kernel-weighted least-squares minimization problem:

$$p_\gamma = \underset{p \in \mathcal{P}_m}{\arg\min} \int_{\mathcal{V}(h)} \{f(t+u) - p(u)\}^2 w_h(u) \mathrm{d}u,$$

where

$$w_h(u) = h^{-d} K(h^{-1}u) \mathbb{1}_{\mathcal{D}}(t+u). \tag{3.2}$$

Notice that the support of w_h corresponds to the neighborhood $\mathcal{V}(h)$. If we consider the weighted space $L^2(w_h)$ endowed with its natural inner product

$$\langle g, \tilde{g} \rangle = \int_{\mathbb{R}^d} g(x) \tilde{g}(x) w_h(x) \mathrm{d}x,$$

then the polynomial p_γ can be interpreted as the orthogonal projection of the function $f(t + \cdot)$ onto the polynomial space $\mathcal{P}_m \subseteq L^2(w_h)$. A simple combinatorial argument shows that the dimension D_m of $\mathcal{P}_m$ is the number of d-combinations from a set of $m + d$ elements. For $\gamma = (m, h) \in \Gamma$, consider the collection of rescaled monomials

$$\Phi_\gamma(u) = \left\{ \varphi_\alpha \left(\frac{u}{h} \right) \right\}_{|\alpha| \leq m}$$

which we organize into a $D_m \times 1$ vector using the total order defined for $\alpha \neq \tilde{\alpha}$ by $\alpha \prec \tilde{\alpha}$ if and only if $|\alpha| < |\tilde{\alpha}|$, or $|\alpha| = |\tilde{\alpha}|$ and $\alpha_{i^\star} < \tilde{\alpha}_{i^\star}$ for $i^\star = \min(i \in \{1, \ldots, d\} : \alpha_i \neq \tilde{\alpha}_i)$. Now that $\Phi_\gamma(u)$ is a vector function, define the $D_m \times D_m$ Gram matrix

$$\mathcal{B}_\gamma = \int_{\mathbb{R}^d} \Phi_\gamma(u)\Phi_\gamma^\top(u) w_h(u) \mathrm{d}u. \tag{3.3}$$

By virtue of Lemma A.1 in Appendix A, the Gram matrix $\mathcal{B}_\gamma$ is symmetric positive definite, so its smallest eigenvalue

$$\lambda_\gamma = \min_{v^\top v = 1} v^\top \mathcal{B}_\gamma v \tag{3.4}$$

must be positive. Letting $\mathcal{B}_\gamma^{1/2}$ be the lower triangular matrix in the Cholesky decomposition of $\mathcal{B}_\gamma$ leads to the following orthonormal basis for the $L^2(w_h)$ space:

$$H_\gamma(u) = \mathcal{B}_\gamma^{-1/2}\Phi_\gamma(u), \quad u \in \mathbb{R}^d. \tag{3.5}$$

If a_γ denotes the $D_m \times 1$ coordinate vector of p_γ in this orthonormal basis, then

$$a_\gamma = \langle f(t+\cdot), H_\gamma \rangle = \mathsf{E}\left\{H_\gamma(X-t) w_h(X-t)\right\} \approx \frac{1}{n}\sum_{i=1}^n H_\gamma(X_i - t) w_h(X_i - t) =: \hat{a}_\gamma.$$

This leads to $\hat{p}_\gamma(u) = H_\gamma^\top(u)\hat{a}_\gamma$ being an estimator for p_γ. Therefore, for any $\gamma = (m, h) \in \Gamma$, we define our local polynomial estimator as follows:

$$\hat{f}_\gamma(t) = \hat{p}_\gamma(0) = \frac{1}{n}\sum_{i=1}^n H_\gamma^\top(0) H_\gamma(X_i - t) w_h(X_i - t). \tag{3.6}$$

3.2. Selection procedure

The data-driven procedure selects a local polynomial estimator from a collection $\{\hat{f}_\gamma(t) : \gamma \in \Gamma_n\}$, where Γ_n is a discrete one-parameter subset of the pairs of polynomial degrees and bandwidths in Γ, see (3.8) below. For any $\ell \in \mathbb{N}$, set

$$h_\ell = \rho \exp(-\ell), \tag{3.7}$$

and let $\{m_\ell : \ell \in \mathbb{N}\} \subseteq \{0, 1, \ldots, \lfloor \log n \rfloor\}$ be a nonincreasing sequence of integers. These choices put approximately $\log n$ bandwidth levels on a logarithmic scale down to a smallest neighborhood of radius at least ρn^{-1} asymptotically. Consider the set of indexed pairs

$$\Gamma_n = \{(m_\ell, h_\ell) : \ell \in \mathcal{L}_n\} \subseteq \Gamma, \tag{3.8}$$

where

$$\mathcal{L}_n = \left\{1 \leq \ell \leq \lfloor \log n \rfloor : n h_\ell^d W_{h_\ell} \geq (\log n)^3\right\} \quad \text{with} \quad W_h = \int_{\mathbb{R}^d} w_h(u) \mathrm{d}u. \tag{3.9}$$

To provide clarity for the selection process, we introduce certain observables. Specifically, for each $\gamma = (m, h) \in \Gamma_n$, the quantity

$$\hat{v}_\gamma = \frac{1}{n} \times \frac{1}{n}\sum_{i=1}^n \left\{H_\gamma^\top(0) H_\gamma(X_i - t) w_h(X_i - t)\right\}^2 \tag{3.10}$$

is an estimator of a natural upper bound on the variance of $\hat{f}_\gamma(t)$. Also, for some fixed constant $\delta > 1$, and the pairs $\gamma = (m,h) \in \Gamma_n$ and $(v,x) \in [0,\infty)^2$, define

$$c_\gamma = \frac{\sqrt{D_m}}{nh^d\lambda_\gamma}, \qquad \varepsilon_\gamma = \frac{(\delta-1)D_m W_h}{nh^d\lambda_\gamma^2}, \qquad \Lambda_\gamma = 2\,|\log(\lambda_\gamma)|, \tag{3.11}$$

and

$$r_\gamma(v,x) = \sqrt{2vx} + c_\gamma x, \qquad \mathrm{pen}(\gamma) = d\delta|\log h| + \Lambda_\gamma. \tag{3.12}$$

Equipped with these notations, consider an upper bound estimator of the bias of $\hat{f}_\gamma(t)$, i.e.,

$$\hat{A}_\gamma = \max_{\gamma' \in \Gamma_n} \left\{ \left| \hat{f}_{\gamma\vee\gamma'}(t) - \hat{f}_{\gamma'}(t) \right| - \hat{\mathbb{U}}_{\gamma\vee\gamma'} - \hat{\mathbb{U}}_{\gamma'} \right\}_+, \tag{3.13}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$ and $\gamma \vee \gamma'$ denotes the maximum of $\gamma = (m_\ell, h_\ell)$ and $\gamma' = (m_{\ell'}, h_{\ell'})$ with respect to the total order $\preceq$ defined on Γ_n by $\gamma \preceq \gamma'$ if $h_\ell \le h_{\ell'}$ (i.e., if $\ell \ge \ell'$). Moreover, consider a penalized version of the estimator of an upper bound on the standard deviation estimator of $\hat{f}_\gamma(t)$,

$$\hat{\mathbb{U}}_\gamma = r_\gamma\left\{\hat{v}_\gamma + \varepsilon_\gamma, \mathrm{pen}(\gamma)\right\}. \tag{3.14}$$

The final adaptive estimator is then defined by

$$\hat{f}(t) = \hat{f}_{\hat{\gamma}}(t), \quad \text{where} \quad \hat{\gamma} = \operatorname*{arg\,min}_{\gamma\in\Gamma_n} \left(\hat{A}_\gamma + \hat{\mathbb{U}}_\gamma\right). \tag{3.15}$$

Remark 6. Before we state our main results in Section 4, it is worthwhile to first discuss a few points that are particularly relevant:

1. The selection rule (3.15) is inspired by the so-called Goldenshluger-Lepski method introduced in a series of papers, see Lepskii (1991) and Goldenshluger and Lepski (2008, 2009, 2011, 2014). It consists of a trade-off between the bias upper bound estimators $\hat{A}_\gamma$ and the penalized standard deviation upper bound estimators $\hat{\mathbb{U}}_\gamma$. Finding tight upper variables $\hat{\mathbb{U}}_\gamma$ is the key point of this selection procedure. Ideally, we would like to choose

   $$\mathbb{U}_\gamma = r_\gamma\{v_\gamma, \mathrm{pen}(\gamma)\}, \tag{3.16}$$

   where the counterpart of $\hat{v}_\gamma$ in (3.10), namely

   $$v_\gamma = \frac{1}{n} \times \mathsf{E}\left[\left\{H_\gamma^\top(0) H_\gamma(X-t) w_h(X-t)\right\}^2\right], \tag{3.17}$$

   is a natural upper bound on the true variance of $\hat{f}_\gamma(t)$.
2. The form of $\mathbb{U}_\gamma$ can be easily explained. The function r_γ in (3.12) is used to apply a Bernstein inequality to control the bias term $\hat{A}_\gamma$, see Lemma A.4 and its proof for details in Appendix A. The penalty $\mathrm{pen}(\gamma)$ consists of two terms: the $|\log h|$-term is unavoidable in pointwise adaptive estimation (see Lepskii (1991)) while the quantity Λ_γ, which adapts to the geometry of the domain $\mathcal{D}$, is specific to our framework. To define $\hat{\mathbb{U}}_\gamma$, the variance bound v_γ is replaced by its estimator $\hat{v}_\gamma$ and the correction term ε_γ is added to ensure that, with large probability, $\hat{\mathbb{U}}_\gamma$ becomes larger than $\mathbb{U}_\gamma$ while staying close to it.
3. The Goldenshluger-Lepski method is generally defined using an order on Γ_n which is, roughly speaking, induced by the variance of the estimators $\hat{f}_\gamma(t)$. This explains the definition of the total order $\gamma \preceq \gamma'$ stated just below (3.13), which involves only the bandwidth h in the comparison of the parameter pairs (m,h) in Γ_n.

4. Main results

Our main results are presented in three steps. First, in Section 4.1, we obtain bounds on the bias and variance of the base local polynomial estimators $\hat{f}_\gamma(t)$ defined in (3.6), and we formulate an oracle-type inequality satisfied by the adaptive estimator $\hat{f}(t)$ defined in (3.15). Next, in Section 4.2, we focus on the most common case, where the geometry of $\mathcal{D}$ is 'simple' in a neighborhood of t, meaning that Assumption 1 is satisfied. In this situation, the minimax and adaptive rates of convergence are established over the whole collection of Hölder-type functional classes $\{\Sigma(s,L) : (s,L) \in (0,\infty)^2\}$. In Section 4.3, we consider a specific situation where the domain $\mathcal{D}$ has a more complicated geometry. To aid understanding, the results are specifically stated for polynomial sectors. We obtain upper bounds on the rate of convergence in each class $\Sigma(s,L)$, which depend on the geometry of $\mathcal{D}$, and we prove that, for small regularities s, these rates of convergence are minimax. Analogous adaptivity results are also derived. Finally, in Section 4.4, the case of unknown compact convex domains is briefly discussed.

4.1. General results

Our first two results pertain to the study of the bias and variance of each estimator in the collection $\{\hat{f}_\gamma(t) : \gamma \in \Gamma\}$. The proofs of these propositions are postponed to Appendix B.

Proposition 4.1 (Bound on the bias). *Let $\gamma = (m,h) \in \Gamma$ be given. Assume that the density function $f : \mathcal{D} \to \mathbb{R}$ belongs to the class $\Sigma(s,L)$ for some positive s and L. Then, there exists a positive real constant $\mathfrak{L}_{m,s,L}$ that depends on m, s and L such that*

$$\left|\mathsf{E}\left\{\hat{f}_\gamma(t)\right\} - f(t)\right| \leq \frac{W_h\sqrt{D_m}}{\lambda_\gamma} \times \mathfrak{L}_{m,s,L}\, h^{\beta_m(s)}, \tag{4.1}$$

where $\beta_m(s) = \min(m+1, s)$. Moreover, we have $\mathfrak{L}_{m,s,L} = L$ as soon as $m \geq \lfloor s \rfloor$.

Proposition 4.2 (Bound on the variance). *Let $\gamma = (m,h) \in \Gamma$ be given. Assume that the density function $f : \mathcal{D} \to \mathbb{R}$ satisfies*

$$\|f\|_{\mathcal{V}(\rho)} \coloneqq \sup_{u \in \mathcal{V}(\rho)} |f(t+u)| < \infty. \tag{4.2}$$

Also, recall the definition of v_γ from (3.17). Then, we have

$$\mathsf{Var}\left\{\hat{f}_\gamma(t)\right\} \leq v_\gamma \leq v_\gamma^\star, \quad \text{with} \quad v_\gamma^\star = \left(\frac{W_h\sqrt{D_m}}{\lambda_\gamma}\right)^2 \times \frac{\|f\|_{\mathcal{V}(\rho)}}{nh^d W_h}. \tag{4.3}$$

Let us highlight that both the bias bound (4.1) and the variance bound (4.3) can be broken down into two separate factors. The second factors, which are proportional to $h^{\beta_m(s)}$ and $1/(nh^dW_h)$, can be readily explained as they correspond (up to a constant) to the bias and variance bounds of multivariate histograms with adaptive bin shapes. These histograms are defined as $\hat{p}_\mathcal{B} = \sum_{i=1}^n \mathbb{1}_\mathcal{B}(X_i)/\{n\mathrm{Leb}(\mathcal{B})\}$ on each multivariate bin $\mathcal{B}$ belonging to a partition of $\mathcal{D}$, as documented for example by Klemelä (2009). Specifically, when considering the bin $\mathcal{B} = \mathcal{D} \cap (t + h\Delta)$, it is straightforward to show that the bias and the variance are proportional to $h^{\beta_m(s)}$ and $1/(nh^dW_h)$, respectively, just as we would with our estimator when the polynomial degree is set to $m = 0$. Notice however that in general the first factors introduce an additional quantity, the ratio W_h/λ_γ, which is not usually present and depends on

the geometry of the domain $\mathcal{D}$. In the case of simple domains, which are treated in Section 4.2, the ratio W_h/λ_γ remains asymptotically bounded, thereby exerting no influence on the bias/variance analysis.

In the case of complicated domains, which are treated in Section 4.3, we know that the ratio W_h/λ_γ encodes the complexity of the domain in some way. Unfortunately, we do not have a heuristic interpretation that would let us guess its asymptotic behavior in the variable n (or h) in general. This remains an open problem. However, in the specific case of the polynomial sectors $\mathcal{D}_k$ covered in Section 4.3, we do have a (potential) interpretation based on the calculations in Equations (B.25), (B.26) and (B.27) of Appendix B. Indeed, the quantity W_h is the area of the domain at the scale of the bandwidth parameter, so the specific result $W_h = h^{k-1}/(k+1)$ in (B.25) gives a 'unit' of the pointiness of the polynomial sectors. Equations (B.26) and (B.27) then show that each degree of a polynomial in our basis can capture one such unit. Since there are two dimensions and the polynomials in our space have degree at most m (recall that $\gamma = (m, h)$), we are left with W_h/λ_{γ_k} being asymptotic to $h^{-(k-1)\times 2m}$, where $\gamma_k \equiv \gamma$ is the notation tailored to the polynomial sector $\mathcal{D}_k$. One could therefore interpret $h^{(k-1)\times 2m}$ as the finest amount of pointiness that our polynomial basis can model. The finer the pointiness of the domain, the bigger the bias and variance become in Propositions 4.1 and 4.2.

For the forthcoming minimax and adaptive results in Sections 4.2 and 4.3, it is also worth noting that Lemma A.3 in Appendix A implies that $\|f\|_{\mathcal{V}(\rho)}$ remains uniformly bounded over $f \in \Sigma(s, L)$ in the variance bound (4.3).

The main result of this section is an oracle-type inequality which provides an explicit asymptotic bound for the pointwise L^2 risk of our adaptive estimator $\hat{f}(t)$.

Theorem 4.3 (Oracle-type inequality). *Assume that the density function $f : \mathcal{D} \to \mathbb{R}$ satisfies (4.2). Assume further that the domain $\mathcal{D}$ satisfies the following property: there exists a positive real constant $b \in (0, \infty)$ such that*

$$\min_{\gamma=(m,h)\in\Gamma_n} (h^m)^{-b}\lambda_\gamma \geq 1. \tag{4.4}$$

Define, for any $\gamma \in \Gamma_n$,

$$\mathbb{B}_\gamma = \max_{\substack{\gamma'\in\Gamma_n \\ \gamma'\preceq\gamma}} \left|\mathsf{E}\left\{\hat{f}_{\gamma'}(t)\right\} - f(t)\right|, \quad \textit{and} \quad \mathbb{U}^\star_\gamma = r_\gamma\{v^\star_\gamma, \mathrm{pen}(\gamma)\}, \tag{4.5}$$

where recall $v^\star_\gamma$ is the bound on the variance in (4.3), and the quantities $r_\gamma(\cdot,\cdot)$ and $\mathrm{pen}(\cdot)$ are defined in (3.12). Recall also that $\delta > 1$ is a parameter that we fixed above (3.12). Then, we have

$$R_n(\hat{f}, f) \leq \min_{\gamma\in\Gamma_n}\left[5\mathbb{B}_\gamma + \left\{3 + 2\sqrt{(\delta-1)/\|f\|_{\mathcal{V}(\rho)}}\right\}\mathbb{U}^\star_\gamma\right] + O\left\{\sqrt{\frac{(\log n)^d}{n}}\right\}. \tag{4.6}$$

The notation $u_n = O(v_n)$ here means that $\limsup_{n\to\infty} |u_n/v_n| < \infty$.

Remark 7. On domains with simple geometries, which are characterized by Assumption 1 below, the technical condition (4.4) in Theorem 4.3 is automatically satisfied when m is constrained to a finite set of values, a situation often encountered in practical scenarios where the target density has a finite smoothness parameter s. This assertion is demonstrated in Lemma A.2 in Appendix A; see also the first paragraph in the proof of Theorem 4.5. However, on domains with complicated geometries, verifying the technical condition (4.4) becomes a case-by-case endeavor. For instance, in the polynomial sector example detailed in Section 4.3, explicit verification is provided within the proof of Theorem 4.7 to establish the adaptivity of our estimator.

4.2. Domains with simple geometries

Below, we introduce an assumption that qualifies what it means for a domain $\mathcal{D}$ to be geometrically 'simple' in a neighborhood of t.

Assumption 1. There exists an open subset $\Delta_0 \subseteq \Delta \equiv [-1,1]^d$ such that

$$h\Delta_0 \subseteq \mathcal{V}(h), \quad \text{for all } h \in (0,\rho].$$

This assumption is satisfied for example if $\mathcal{V}(h)$ is star-shaped with respect to the origin and has a non-empty interior. Assumption 1 is a relaxation of Assumption 5 in Section 2.2 of Bertin et al. (2020), which stipulates that there exists a finite set of linear transformations of determinant 1 such that for a hypercube $[0,r]^d$ of a small enough width r, you can, for any point $t \in \mathcal{D}$, choose a linear transformation in the finite set so that the linearly transformed hypercube (i.e., a rotated parallelotope) fits into $\mathcal{D} - t$. Assumption 1 above is weaker because we could simply take Δ_0 to be a small enough open cone contained in the rotated parallelotope that fits into $\mathcal{D} - t$.

Under Assumption 1, Proposition 4.4 below establishes that our local polynomial estimator $\hat{f}_\gamma(t)$ achieves the minimax rate of convergence over each individual class $\Sigma(s,L)$. Similarly, Theorem 4.5 proves that our adaptive statistical procedure $\hat{f}(t)$ attains the ARC over the whole collection $\{\Sigma(s,L) : (s,L) \in (0,\infty)^2\}$. The ARC here corresponds to the minimax rate of convergence up to a logarithmic factor.

Proposition 4.4 (Minimaxity of $\hat{f}_\gamma(t)$ on simple domains). *Suppose that Assumption 1 holds. Let $(s,L) \in (0,\infty)^2$ be given. Let $\gamma = (m,h) \in \Gamma$ be such that*

$$m = \lfloor\!\lfloor s \rfloor\!\rfloor \quad \text{and} \quad h = \left[\frac{1}{2s\mathfrak{L}^2_{\lfloor\!\lfloor s \rfloor\!\rfloor,s,L}} \times \frac{d\,\mathfrak{F}_{s,L}}{2^d\, n} \right]^{1/(2s+d)}, \tag{4.7}$$

where $\mathfrak{L}_{m,s,L}$ is a positive real constant that appears in the bias upper bound in Proposition 4.1, and $\mathfrak{F}_{s,L}$ is any upper bound on $\sup_{f\in\Sigma(s,L)} \|f\|_{\mathcal{V}(\rho)}$ (the existence of which is guaranteed by Lemma A.3 in Appendix A). If we set

$$N_n(s,L) = n^{-s/(2s+d)}, \tag{4.8}$$

then the following assertions hold:

1. *$\{N_n(s,L) : n \in \mathbb{N}\}$ is the minimax rate of convergence over $\Sigma(s,L)$.*
2. *The estimator $\hat{f}_\gamma(t)$ attains this rate of convergence. More precisely, for n large enough, we have*

$$N_n^{-1}(s,L) \sup_{f\in\Sigma(s,L)} R_n(\hat{f}_\gamma, f) \le C(s,L),$$

where the right-hand is defined by the positive real constant

$$C(s,L) = \frac{2^d\sqrt{D_m}}{\lambda_\star(m)} \sqrt{\left(\frac{d}{2s}+1\right)\left(\frac{2s}{4^s d}\right)^{d/(2s+d)}}\, \mathfrak{F}_{s,L}^{s/(2s+d)} (\mathfrak{L}_{\lfloor\!\lfloor s \rfloor\!\rfloor,s,L})^{d/(2s+d)} \tag{4.9}$$

and $\lambda_\star(m) > 0$ is a lower bound found in Lemma A.2 in Appendix A on the smallest eigenvalue λ_γ of the Gram matrix $\mathcal{B}_\gamma$.

Remark 8. Understanding the optimal parameters in (4.7) unfolds as follows. On simple domains, which are characterized by Assumption 1, it turns out that for a finite m, the ratios W_h/λ_γ in Proposition 4.1 and 4.2 remain uniformly bounded in the variable h. Therefore, the bound on the bias (4.1) depends on the smoothness parameter m asymptotically only through the factor $h^{\beta_m(s)}$, since D_m and $\mathfrak{L}_{m,s,L}$ are constants for a finite m. By the same argument, the bound on the variance (4.3) is not affected asymptotically by the choice of m. Therefore, given the decomposition of the squared risk measure as the sum between the squared bias and the variance of the estimator, recall Remark 2, it becomes clear that the optimal choice of m in this case will minimize $h^{\beta_m(s)} \equiv h^{\min(m+1,s)}$, which happens when $m \geq \lfloor\!\lfloor s \rfloor\!\rfloor$. Since the rate $h^{\beta_m(s)}$ does not improve for larger $m \geq \lfloor\!\lfloor s \rfloor\!\rfloor$ and D_m increases with m, the optimal choice of m for our pointwise L^2 risk measure is necessarily the smallest m that satisfies $m \geq \lfloor\!\lfloor s \rfloor\!\rfloor$, namely $m = \lfloor\!\lfloor s \rfloor\!\rfloor$. Once the optimal $m = \lfloor\!\lfloor s \rfloor\!\rfloor$ is selected, the optimization of the bandwidth h is analogous to the approach employed for classical kernel density estimators.

Heuristically, our local polynomial estimator wants to adapt its smoothness level m to the smoothness level s of the target density. The information of the target density is utilized optimally when m is the largest integer which is smaller than s and leaves room for a positive Hölder exponent in Definition 2. That value is necessarily $m = \lfloor\!\lfloor s \rfloor\!\rfloor$. In particular, if the practitioner decides to fix a suboptimal value, say $m = 0$ when $s > 1$, and then optimize over h, the bias term in Proposition 4.1 will simply not have the smallest rate of convergence possible.

On complicated domains, the optimization in h is more subtle since the ratio W_h/λ_γ has to be taken into account, which requires a case-by-case analysis; see, e.g., Proposition 4.6 and its proof.

In Section 3.2, we have constructed several statistical procedures depending on the choice of the discrete collection of polynomial degrees $\{m_\ell : \ell \in \mathbb{N}\}$. For the next result, we fix this collection by choosing

$$m_\ell = \left\lfloor \frac{\log n}{2\ell} \right\rfloor, \quad \ell \in \mathbb{N}. \tag{4.10}$$

Theorem 4.5 (Adaptivity of $\hat{f}(t)$ on simple domains). *Suppose that Assumption 1 holds. Let $\mathcal{K} = (0,\infty)^2$ and consider the set of Hölder-type functional classes $\{\Sigma(s,L) : (s,L) \in \mathcal{K}\}$. The following assertions hold:*

1. *The ARC is the admissible collection ϕ given by*

$$\phi_n(s,L) = \left(\frac{\log n}{n}\right)^{s/(2s+d)}, \quad (s,L) \in \mathcal{K}, \ \ n \in \mathbb{N}. \tag{4.11}$$

   More precisely, if ψ is another admissible collection of normalizations such that (s_0, L_0) belongs to $[\psi \ll \phi]$, then

$$[\psi \ll \phi] \subseteq \{s_0\} \times (0,\infty) \quad \text{and} \quad [\psi \ggg \phi] \supseteq (s_0,\infty) \times (0,\infty).$$

2. *The adaptive estimator $\hat{f}(t)$ is such that, for any $(s,L) \in \mathcal{K}$,*

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{f\in\Sigma(s,L)} \phi_n^{-1}(s,L) R_n(\hat{f},f) \\ &\leq C \left[4 + \left\{(\delta-1)/\|f\|_{\mathcal{V}(\rho)}\right\}^{1/2}\right] e^s \frac{2^d \sqrt{D^\star(s,d)}}{\lambda_\bullet(s,d)} \sqrt{d\,\delta}\, \mathfrak{F}_{s,L}^{s/(2s+d)} (\mathfrak{L}_{s,L}^\star)^{d/(2s+d)}, \end{aligned}$$

where $C > 0$ is a universal constant (independent of all parameters), $\delta > 1$ is a parameter that we fixed above (3.12), $\mathfrak{F}_{s,L}$ is any upper bound on $\sup_{f\in\Sigma(s,L)} \|f\|_{\mathcal{V}(\rho)}$ (the existence of which is guaranteed by Lemma A.3 in Appendix A), and

$$D^\star(s,d) = \max_{0\le m'\le\lfloor 2s+d\rfloor} D_{m'} \quad (\textit{recall that } D_{m'} = \binom{m'+d}{d})$$

$$\lambda_\bullet(s,d) = \min_{0\le m'\le\lfloor 2s+d\rfloor} \lambda_\star(m) > 0,$$

$$\mathfrak{L}^\star_{s,L} = \max_{0\le m'\le\lfloor 2s+d\rfloor} \mathfrak{L}_{m',s,L}.$$

Note that the positivity of $\lambda_\bullet(s,d)$ is guaranteed by Lemma A.2 in Appendix A.

4.3. Domains with complicated geometries

Set $\rho = 1$ and define the following polynomial sectors in $\mathbb{R}^2$:

$$\mathcal{D}_k = \left\{(x,y)\in\mathbb{R}^2 : 0\le x\le 1, 0\le y\le x^k\right\}, \quad k > 1. \tag{4.12}$$

This section is dedicated to the local estimation, at the origin $t = (0,0)$, of density functions supported on a given polynomial sector domain $\mathcal{D}_k$. In particular, we obtain minimaxity and adaptivity results similar to those in Section 4.2 but on the complicated domain $\mathcal{D}_k$, see Proposition 4.6 and Theorem 4.7, respectively.

Item 2 of Remark 9 below explains how our results could be extended to other complicated domains without much difficulty. We believe that treating these polynomial sector examples separately elucidates the overall arguments and techniques and thus provides greater clarity for the reader. Furthermore, our statements here complement the simulations presented in Section 5, which are also performed on polynomial sectors.

Proposition 4.6 (Minimaxity of $\hat{f}_\gamma(t)$ on $\mathcal{D}_k$). *Let $(s,L)\in(0,\infty)^2$ be given. Define*

$$s_k = \arg\max_{\beta\in(0,s]} \theta_k(\beta), \quad \textit{where} \quad \theta_k(\beta) = \frac{\beta - 2\lfloor\!\lfloor\beta\rfloor\!\rfloor(k-1)}{2\beta+k+1}. \tag{4.13}$$

The following assertions hold:

1. *There exists a positive real constant $C(s,L,k)$ such that, for any $f\in\Sigma(s,L)$, we have*

$$R_n(\hat{f}_{\gamma_k}, f) \le C(s,L,k) n^{-\theta_k(s_k)}, \tag{4.14}$$

   where $\gamma_k = (m,h)$ with $m = \lfloor\!\lfloor s_k\rfloor\!\rfloor$ and $h = n^{-1/(2s_k+k+1)}$.
2. *Assume further that $s\in(0,1]$, then there exists a positive real constant $c(s,L,k)$ such that*

$$\inf_{\tilde{f}} \sup_{f\in\Sigma(s,L)} R_n(\tilde{f}, f) \ge c(s,L,k) n^{-\theta_k(s)}. \tag{4.15}$$

   In particular, $\{n^{-s/(2s+k+1)} : n\in\mathbb{N}\}$ is the minimax rate of convergence on $\Sigma(s,L)$ as soon as $s\in(0,1]$; refer to Item 1 of Remark 9 below.

Remark 9. Let us make two comments:

1. Through the upper bound (4.14), Proposition 4.6 shows that the estimator is consistent as soon as $\theta_k(s_k) > 0$. This is always the case since, for any $\beta \in (0, \min(s,1)]$, we have $\theta_k(\beta) = \beta/(2\beta + k + 1) > 0$. Using similar arguments, it is easily seen that for any $s \in (0,1]$ and any $k > 1$, we have $s_k = s$. However, for $k > 2$ and $s > 1$, we always have $s - 2\lfloor\!\lfloor s \rfloor\!\rfloor(k-1) < 0$, which implies $s_k = 1 < s$. Moreover, note that, for any $s > 0$ and any $k > 1$, we have

$$\theta_k(s_k) < \frac{s}{2s+2}.$$

   Thus, the rate of convergence established in Proposition 4.6 is smaller compared with the one established for simple domains in Proposition 4.4. The estimation problem considered in the present section is more difficult because of the local pinch of $\mathcal{D}_k$ near the origin.
2. Results analogous to those of Proposition 4.6 can be obtained for other domains $\mathcal{D}$ with complicated geometries. In particular, for $\alpha, \beta \in \mathbb{N}_0^d$ and $h > 0$, define

$$I_{\mathcal{D}}(\alpha, \beta, h) = \int_{\mathbb{R}^d} \varphi_\alpha\left(\frac{u}{h}\right) \varphi_\beta\left(\frac{u}{h}\right) w_h(u) \mathrm{d}u.$$

   Assume that there exist a function $a : \mathbb{N}_0^d \to (0,\infty)$ and an increasing function $G : \mathbb{N}_0^d \to (0,\infty)$ such that

$$I_{\mathcal{D}}(\alpha, \beta, h) = a(\alpha+\beta) h^{G(\alpha+\beta)}, \quad \text{for any } \alpha, \beta \in \mathbb{N}_0^d.$$

   Now define

$$s_{\mathcal{D}} = \operatorname*{arg\,max}_{\beta \in (0,s]} \theta_{\mathcal{D}}(\beta), \quad \text{where} \quad \theta_{\mathcal{D}}(\beta) = \frac{\beta + G(0) - G(2\lfloor\!\lfloor \beta \rfloor\!\rfloor)}{2\beta + d + G(0)}.$$

   In this particular setting, it is possible to derive a result akin to (4.14), with the rate of convergence $n^{-\theta(s_{\mathcal{D}})}$ and the bandwidth $h = n^{-1/\{2s_{\mathcal{D}}+d+G(0)\}}$, by following the proof of Proposition 4.6.

Now consider the adaptive estimator $\hat{f}(t)$ constructed in Section 3.2 and choose $m_\ell = 0$ for all $\ell \in \mathbb{N}$. The following theorem ensures that it attains the ARC over the collection of Hölder-type functional classes $\{\Sigma(s,L) : (s,L) \in (0,1] \times (0,\infty)\}$. The ARC here corresponds to the minimax rate of convergence stated just below (4.15) up to a logarithmic factor.

Theorem 4.7 (Adaptivity of $\hat{f}(t)$ on $\mathcal{D}_k$). *The following assertions hold:*

1. *The ARC is the admissible collection ϕ given by*

$$\phi_n(s,L) = \left(\frac{\log n}{n}\right)^{s/(2s+k+1)}, \quad (s,L) \in (0,1] \times (0,\infty), \quad n \in \mathbb{N}.$$

2. *Let $(s,L) \in (0,1] \times (0,\infty)$ be given. There exists a positive real constant $\tilde{C}(s,L,k,\delta) > 0$ such that*

$$\limsup_{n\to\infty} \sup_{f \in \Sigma(s,L)} \left(\frac{\log n}{n}\right)^{-s/(2s+k+1)} R(\hat{f}, f) \le \tilde{C}(s,L,k,\delta). \tag{4.16}$$

Remark 10. Theorem 4.7 exclusively studies the case of a smoothness parameter s less than or equal to 1. We conjecture that a result similar to (4.16), with rate of convergence $(n^{-1}\log n)^{\theta_k(s_k)}$, can be obtained for any $s > 1$ by choosing

$$m_\ell = \left\lfloor \frac{\log n}{2\ell} - \frac{k+1}{2} \right\rfloor.$$

Nevertheless, this rate of convergence may be not the ARC for $s > 1$.

4.4. Extension to unknown domains

Following a question posed by two referees, we comment briefly in this section about the case in which the domain $\mathcal{D}$ is unknown.

Although the general problem of constructing optimal estimators in this setting is beyond the scope of this paper, we propose to consider a simple but common situation: assume that the support $\mathcal{D}$ is an unknown compact convex set with finite Lebesgue measure $\lambda(\mathcal{D})$, and assume further that the target density f is bounded, i.e., its supremum norm over $\mathbb{R}^d$, $\|f\|_\infty$, is finite. In order to control the error related to the estimation of the support, we have to consider an integrated risk (over $\mathbb{R}^d$) rather than a pointwise risk. That is, for any estimator $\tilde{f}$ of f, we are interested in bounding the L^1 integrated risk

$$IR_n(\tilde{f}, f) = \mathsf{E}\left\{\int_{\mathbb{R}^d} |\tilde{f}(x) - f(x)| \mathrm{d}x\right\}.$$

We propose to estimate f by splitting the data into two disjoint subsamples: we estimate the support of f using the first subsample and then use our adaptive estimation procedure on the second subsample for each point in the estimated domain. More precisely, for some proportion parameter $\alpha \in (0,1)$ satisfying $\alpha n \in \mathbb{N}$, define the two subsamples

$$\mathbb{X}_n^{(1)} = (X_1, \ldots, X_{(1-\alpha)n}), \quad \mathbb{X}_n^{(2)} = (X_{(1-\alpha)n+1}, \ldots, X_n),$$

so that $\mathbb{X}_n^{(1)} \cap \mathbb{X}_n^{(2)} = \emptyset$ and $\mathbb{X}_n^{(1)} \cup \mathbb{X}_n^{(2)} = \mathbb{X}_n$ in particular.

Define $\hat{\mathcal{D}}_n$ as the convex hull of the first subsample $\mathbb{X}_n^{(1)}$. This serves as an estimator for the convex support $\mathcal{D}$ of the target density f, a subject that has been extensively explored in the literature. For an in-depth review, we direct the reader to Brunel (2018). Now, we remark that, conditionally on the observations in $\mathbb{X}_n^{(1)}$, the observations in $\mathbb{X}_n^{(2)} \cap \hat{\mathcal{D}}_n$ are independent and identically distributed with a common density function given by

$$g(x) = \frac{f(x)\mathbb{1}_{\hat{\mathcal{D}}_n}(x)}{\int_{\hat{\mathcal{D}}_n} f(y)\mathrm{d}y}, \quad x \in \mathbb{R}^d,$$

where $f\mathbb{1}_{\hat{\mathcal{D}}_n} \leq f$ (since $\hat{\mathcal{D}}_n \subseteq \mathcal{D}$).

By definition, note that $\mathbb{X}_n^{(2)}$ contains αn observations, i.e., $\mathrm{card}(\mathbb{X}_n^{(2)}) = \alpha n$, and let

$$\hat{p}_n = \frac{\mathrm{card}(\mathbb{X}_n^{(2)} \cap \hat{\mathcal{D}}_n)}{\mathrm{card}(\mathbb{X}_n^{(2)})}$$

denote the proportion of those observations which fall into the estimated support $\hat{\mathcal{D}}_n$; $\hat{p}_n$ is the empirical counterpart of $p_n \equiv \int_{\hat{\mathcal{D}}_n} f(y)\mathrm{d}y$. Moreover, conditionally on $\mathbb{X}_n^{(1)}$, the support $\hat{\mathcal{D}}_n$ of g is known,

so our local polynomial estimation procedure can be used to obtain the adaptive estimator $\hat{g}(t)$ defined in (3.15) for each point $t \in \hat{\mathcal{D}}_n$. Given the above relationship between g and f, it is natural to define the estimator of f:

$$\hat{f}(x) = \hat{p}_n \hat{g}(x), \quad x \in \mathbb{R}^d.$$

Using the triangle inequality, Fubini's theorem, and Jensen's inequality, the integrated risk of $\hat{f}$ can be bounded as follows:

$$\begin{aligned} IR_n(\hat{f}, f) &= \mathsf{E}\left\{\int_{\mathcal{D}} \left|\hat{p}_n \hat{g}(x) - \hat{p}_n g(x) + \hat{p}_n g(x) - p_n g(x) + p_n g(x) - f(x)\right| \mathrm{d}x\right\} \\ &\leq \mathsf{E}\left[\int_{\hat{\mathcal{D}}_n} \mathsf{E}\left\{|\hat{g}(x) - g(x)| \,\middle|\, \mathbb{X}_n^{(1)}\right\} \mathrm{d}x\right] + \mathsf{E}\left(|\hat{p}_n - p_n|\right) + \mathsf{E}\left\{\int_{\mathcal{D}\setminus\hat{\mathcal{D}}_n} f(x)\mathrm{d}x\right\} \\ &\leq \mathsf{E}\left(\int_{\hat{\mathcal{D}}_n} \left[\mathsf{E}\left\{|\hat{g}(x) - g(x)|^2 \,\middle|\, \mathbb{X}_n^{(1)}\right\}\right]^{1/2} \mathrm{d}x\right) + \mathsf{E}\left(|\hat{p}_n - p_n|\right) + \mathsf{E}\left\{\int_{\mathcal{D}\setminus\hat{\mathcal{D}}_n} f(x)\mathrm{d}x\right\}. \end{aligned}$$

To bound $\mathsf{E}(|\hat{p}_n - p_n|)$, we use the fact that, conditionally to $\mathbb{X}_n^{(1)}$, the quantity $\alpha n \hat{p}_n$ has a Binomial distribution with parameters αn and p_n. To bound $\mathsf{E}\{\int_{\mathcal{D}\setminus\hat{\mathcal{D}}_n} f(x)\mathrm{d}x\}$, we use Theorem 7 of Brunel (2018) with the choice $x = \log\{(1-\alpha)n\} \geq 0$. As a result, for n large enough, we have

$$\begin{aligned} IR_n(\hat{f}, f) &\lesssim \lambda(\mathcal{D}) \times \mathsf{E}\left(\sup_{x\in\hat{\mathcal{D}}_n} \left[\mathsf{E}\left\{|\hat{g}(x) - g(x)|^2 \,\middle|\, \mathbb{X}_n^{(1)}\right\}\right]^{1/2} \mathbb{1}_{\{\hat{p}_n \geq 1/2\}}\right) \\ &\quad + \lambda(\mathcal{D}) \times \mathsf{P}(\hat{p}_n < 1/2) + \frac{1}{\sqrt{\alpha n}} + \left[\frac{\|f\|_\infty + 1}{\{(1-\alpha)n\}^{2/(d+1)}} + \frac{\log\{(1-\alpha)n\}}{(1-\alpha)n} + \frac{1}{(1-\alpha)n}\right]. \end{aligned}$$

By Theorem 7 of Brunel (2018), the probability $\mathsf{P}(\hat{p}_n < 1/2)$ is exponentially small in n. It remains to bound the conditional expectation in the last equation. Note that if n is large enough and the target density f belongs to $\Sigma(s, L)$, then g belongs to $\Sigma(s, L/p_n) \subseteq \Sigma(s, 2L)$ with probability larger than $1/\{(1-\alpha)n\}$, again by Theorem 7 of Brunel (2018). Hence, conditionally on $\mathbb{X}_n^{(1)}$, by applying item 2 of Theorem 4.5 for every $x = t \in \hat{\mathcal{D}}_n$ in $[\mathsf{E}\{|\hat{g}(x) - g(x)|^2 | \mathbb{X}_n^{(1)}\}]^{1/2} \mathbb{1}_{\{\hat{p}_n \geq 1/2\}}$, we obtain

$$IR_n(\hat{f}, f) \lesssim (n^{-1}\log n)^{s/(2s+d)} + n^{-2/(d+1)},$$

where $\lesssim$ depends on d, $\|f\|_\infty$, s L and α.

Assuming further that the target density f is bounded away from zero on its support $\mathcal{D}$, i.e., $f_0 \equiv \inf_{x\in\mathcal{D}} f(x) > 0$, we also get a lower bound on the integrated risk of any estimator $\tilde{f}$ of f, including $\hat{f}$. Indeed, letting $\widetilde{\mathcal{D}}_n = \{x \in \mathbb{R}^d : \tilde{f}(x) \geq f_0/2\}$, we have

$$IR_n(\tilde{f}, f) = \mathsf{E}\left\{\int_{\mathbb{R}^d} |\tilde{f}(x) - f(x)| \mathrm{d}x\right\} \geq \frac{f_0}{2} \mathsf{E}\left\{\lambda(\mathcal{D} \triangle \widetilde{\mathcal{D}}_n)\right\} \gtrsim n^{-2/(d+1)},$$

where $\mathcal{D} \triangle \widetilde{\mathcal{D}}_n$ denotes the symmetric difference between $\mathcal{D}$ and $\widetilde{\mathcal{D}}_n$, and the last inequality follows from a direct application of Theorem 2 of Brunel (2018). If the domain $\mathcal{D}$ was known, we would also have $IR_n(\tilde{f}, f) \gtrsim n^{-s/(2s+d)}$.

Therefore, combining the findings from the last two paragraphs, it is reasonable to conjecture that in the case of unknown compact convex domains, both the minimax rate and the ARC for the L^1 integrated risk measure should be $n^{-s/(2s+d)} + n^{-2/(d+1)}$ when assuming some mild regularity conditions on the target density f. This problem is left open for future research.

5. Simulations

In this section, we conduct a concise investigation using simulated data, where we assess the performance of our estimator against the one provided in the `sparr` package Davies and Marshall (2023); see (Davies, Marshall and Hazelton, 2018) for a tutorial. The `sparr` package is extensively used within the `R` community for estimating bivariate density functions on complicated domains. It is worth noting that since 2020, based on the data gathered by David Robinson's Shiny app Robinson (2015), the download rate of the `sparr` package has consistently been approximately three times higher than that of the other commonly used `latticeDensity` package (Barry, 2021).

We consider two polynomial sector domains for our simulations, namely $\mathcal{D}_1$ and $\mathcal{D}_{2.1}$; recall (4.12). Notice that the linear sector $\mathcal{D}_1$ is a simple domain which satisfies Assumption 1, while $\mathcal{D}_{2.1}$ is a complicated domain which is not star-shaped around the point of estimation $t = (0,0)$. On each domain, we consider two different types of density functions, which we describe below.

The first two density functions, polynomial in nature, are defined by

$$f_k(x,y) = C_k\big\{(x-0.6)^2 + (y-0.2)^2\big\}\mathbb{1}_{\mathcal{D}_k}(x,y), \quad k \in \{1, 2.1\}, \tag{5.1}$$

where C_k is a positive normalizing constant which depends on k. The contour plots of these two densities are depicted in Figure 1.

The next two density functions are mixtures of truncated Gaussian distributions defined by

$$g_k(x,y) = A_k\left[\exp\left\{-\frac{(x-a_k)^2+(y-b_k)^2}{2(0.4)^2}\right\} + \exp\left\{-\frac{(x-c_k)^2+(y-d_k)^2}{2(0.15)^2}\right\}\right], \quad k \in \{1, 2.1\}, \tag{5.2}$$

where A_k is a positive normalizing constant which depends on k, and

$$a_k = 1/10, \quad b_k = (1/10)^k/2, \quad c_k = 3/4, \quad \text{and} \quad d_k = (3/4)^k/2.$$

The contour plots of these two densities are depicted in Figure 2.

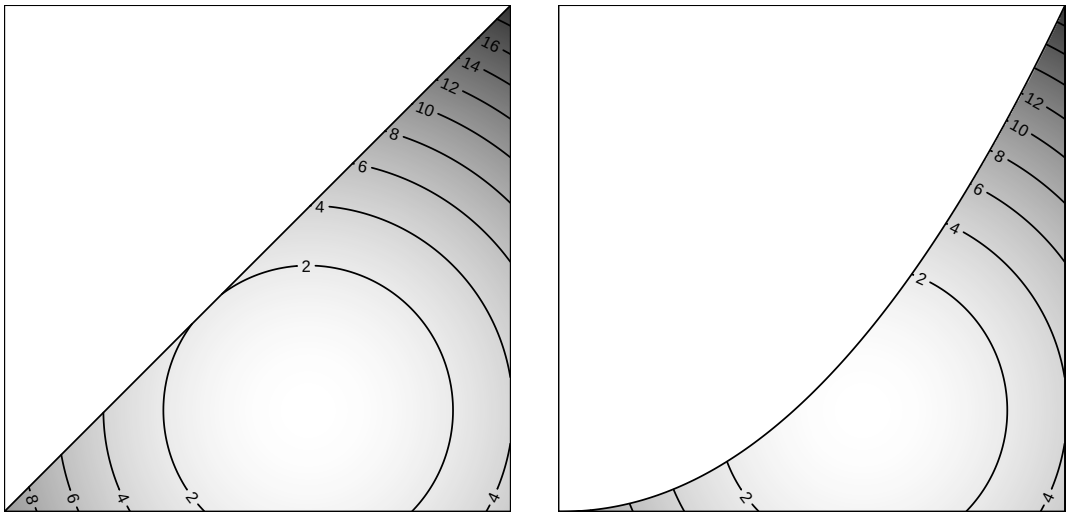

Figure 1. The polynomial density functions f_1 (left) and $f_{2.1}$ (right) defined in (5.1). Darker regions are associated with higher values of f_k.

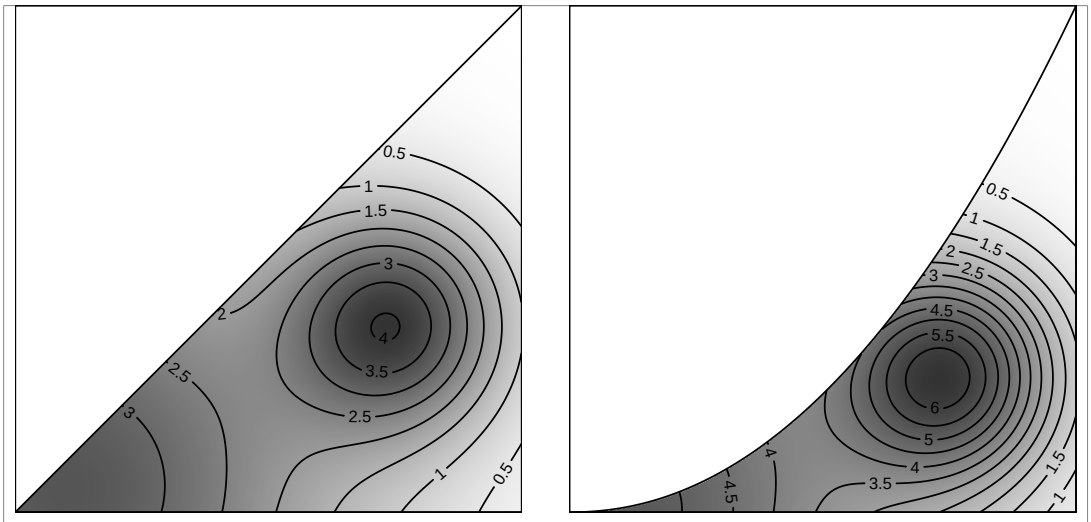

Figure 2. The mixtures of truncated Gaussian density functions g_1 (left) and $g_{2.1}$ (right) defined in (5.2). Darker regions are associated with higher values of g_k.

The `sparr` estimator $\hat{f}_h^{\text{SPARR}}$ depends on a tuning parameter h which plays the role of the bandwidth while our local polynomial estimator $\hat{f}_{m,h}^{\text{LP}}$ depends on both a bandwidth h and a polynomial degree m. Below, we consider the family of polynomial degrees $\mathcal{M} = \{0, 1, 2, 3, 4, 5\}$ and the family of bandwidths $\mathcal{H} = \{0.01 + 0.001 \times \ell : \ell = 0, \ldots, 599\}$. For each target density $f \in \{f_1, f_{2.1}, g_1, g_{2.1}\}$ and each sample size $n \in \{200, 500, 1000, 2000\}$, we aim to compare two oracle estimators defined by

$$\hat{f}_{m^*,h^*}^{\text{LP}} \quad \text{where} \quad (m^*, h^*) = \underset{(m,h)\in\mathcal{M}\times\mathcal{H}}{\arg\min}\, \mathsf{E}\left\{\left|\hat{f}_{m,h}^{\text{LP}}(t) - f(t)\right|^2\right\},$$

$$\hat{f}_{h^\circ}^{\text{SPARR}} \quad \text{where} \quad h^\circ = \underset{h\in\mathcal{H}}{\arg\min}\, \mathsf{E}\left\{\left|\hat{f}_h^{\text{SPARR}}(t) - f(t)\right|^2\right\}.$$

To accomplish this, we generate $R = 5000$ replications of the random samples of size n, and we compute for each replication $r \in \{1, \ldots, R\}$ the corresponding values of the estimators denoted by $[\hat{f}_{m,h}^{\text{LP}}(t)]_r$ and $[\hat{f}_h^{\text{SPARR}}(t)]_r$, respectively. Finally, we obtain the estimated oracles using

$$(\hat{m}^*, \hat{h}^*) = \underset{(m,h)\in\mathcal{M}\times\mathcal{H}}{\arg\min}\, \frac{1}{R}\sum_{r=1}^{R}\left|[\hat{f}_{m,h}^{\text{LP}}(t)]_r - f(t)\right|^2,$$

$$\hat{h}^\circ = \underset{h\in\mathcal{H}}{\arg\min}\, \frac{1}{R}\sum_{r=1}^{R}\left|[\hat{f}_h^{\text{SPARR}}(t)]_r - f(t)\right|^2.$$

The results are presented in Figures 3 and 4 below.

Given a random sample of size $n \in \{200, 500, 1000, 2000\}$ for each replication, the boxplots of the R estimated oracles using our local polynomial method, $[\hat{f}_{\hat{m}^*,\hat{h}^*}^{\text{LP}}(t)]_{r=1}^R$, are displayed in the left columns (LP) and the boxplots of the R estimated oracles using the `sparr` package, $[\hat{f}_{\hat{h}^\circ}^{\text{SPARR}}(t)]_{r=1}^R$, are displayed in the right columns (SPARR). Figure 3 corresponds with the polynomial-type densities f_k defined in (5.1) for $k = 1$ (left) and $k = 2.1$ (right), while Figure 4 corresponds with the mixtures of truncated Gaussian distributions g_k defined in (5.2) for $k = 1$ (left) and $k = 2.1$ (right).

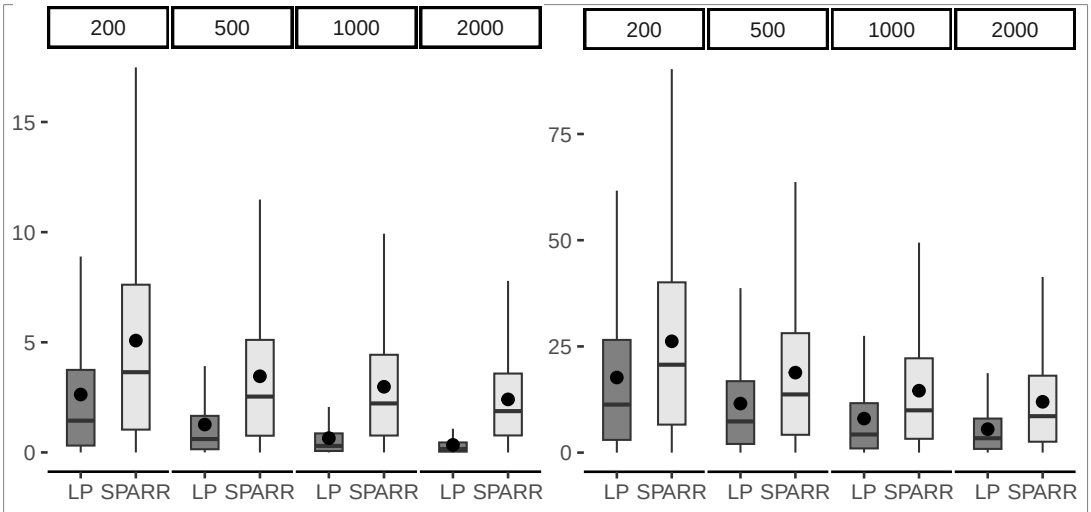

Figure 3. Boxplots of the R replications of the estimated oracles. The target densities are f_1 (left) and $f_{2.1}$ (right).

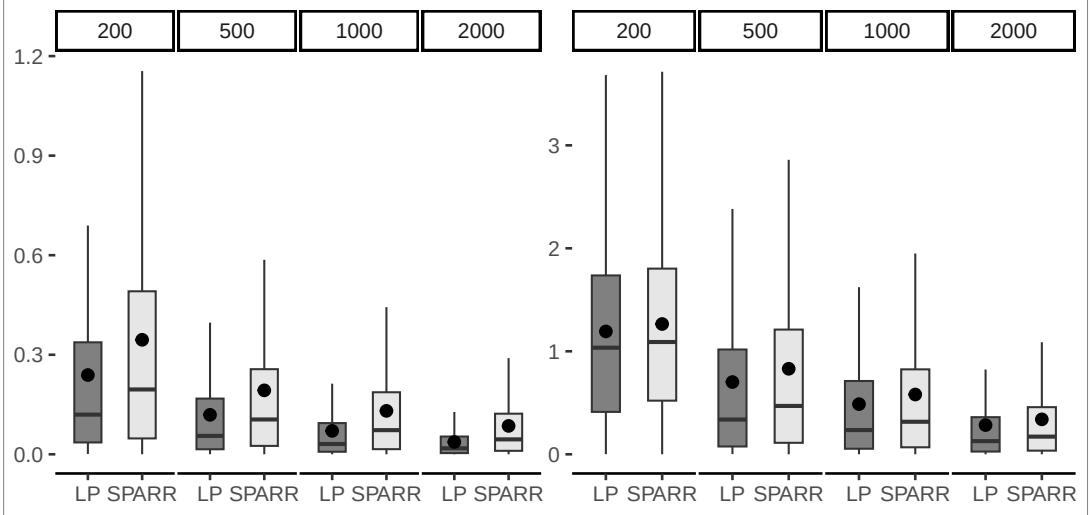

Figure 4. Boxplots of the R replications of the estimated oracles. The target densities are g_1 (left) and $g_{2.1}$ (right).

Together, Figures 3 and 4 reveal that, for both the simple domain $\mathcal{D}_1$ and the complicated domain $\mathcal{D}_{2.1}$, our oracle estimates almost always outperform those of the `sparr` package in mean (black dot inside each box), in median (horizontal line inside each box), and in interquartile range (height of each box), under both the polynomial densities f_k and the mixtures of truncated Gaussian densities g_k, and across all sample sizes n.

The only exception is observed under $g_{2.1}$ for $n = 200$, where the means, medians and interquartile ranges are close to being equal, still giving a slight edge to our method based on the means and quartiles.

Overall, out of the 16 cases considered, our oracle estimates outperform those of the `sparr` package in every case. The best polynomial degree selected was always either 1 or 2, which justifies our

restricted range of polynomial degrees $\mathcal{M} = \{0, 1, 2, 3, 4, 5\}$. It is typical of low degree polynomials to be optimal for local polynomial kernel estimators; see, e.g., p. 126 of Wand and Jones (1995).

Appendix A: Preliminary technical lemmas

Lemma A.1. *Let $\gamma = (m, h) \in \Gamma$ be given. The Gram matrix $\mathcal{B}_\gamma$ defined in (3.3) is symmetric positive definite. In particular, its smallest eigenvalue λ_γ, defined in (3.4), is positive.*

Proof of Lemma A.1. For all $v \in \mathbb{R}^{D_m}$, the fact that $w_h \geq 0$ implies

$$v^\top \mathcal{B}_\gamma v = \int_{\mathcal{V}(h)} v^\top \Phi_\gamma(u) \Phi_\gamma^\top(u) v \, w_h(u) \mathrm{d}u = \int_{\mathcal{V}(h)} \left\{ v^\top \Phi_\gamma(u) \right\}^2 w_h(u) \mathrm{d}u \geq 0. \tag{A.1}$$

Moreover, if some $v \in \mathbb{R}^{D_m}$ satisfies $v^\top \mathcal{B}_\gamma v = 0$, then $v^\top \Phi_\gamma(u) = 0$ for almost all $u \in \mathcal{V}(h)$. This means that the polynomial $v^\top \Phi_\gamma(u)$, with coefficients v_i, is zero for almost all $u \in \mathcal{V}(h) = \mathrm{Supp}(w_h)$. Since $\mathrm{Leb}\{\mathcal{V}(h)\} > 0$, we conclude that $v_i = 0$ for all $i \in \{1, \ldots, D_m\}$. □

Lemma A.2. *Suppose that Assumption 1 holds. First, we have*

$$\mathrm{Leb}(\Delta_0) \leq W_h \leq \mathrm{Leb}(\Delta) = 2^d, \quad \textit{for all } h \in (0, \rho].$$

Second, for all $m \in \mathbb{N}_0$, there exists $\lambda_\star(m) > 0$, which also depends on ρ, such that

$$\inf_{h \in (0,\rho]} \lambda_\gamma \geq \lambda_\star(m), \quad \textit{where } \gamma = (m, h).$$

Proof of Lemma A.2. Since $h\Delta_0 \subseteq \mathcal{V}(h) \subseteq h\Delta$ for all $h \in (0, \rho]$ by Assumption 1, we have

$$\int_{h\Delta_0} w_h(u) \mathrm{d}u \leq \int_{\mathcal{V}(h)} w_h(u) \mathrm{d}u \leq \int_{h\Delta} w_h(u) \mathrm{d}u.$$

Furthermore, $\mathrm{Supp}(w_h) = \mathcal{V}(h)$, so the middle integral is W_h as defined in (3.9). With a simple change of variable and the fact that K is identically equal to 1 on Δ according to (3.1), we get

$$\mathrm{Leb}(\Delta_0) \leq W_h \leq \mathrm{Leb}(\Delta) = 2^d, \quad \text{for all } h \in (0, \rho],$$

which proves the first claim of the lemma.

Next, using the fact that $\mathcal{V}(h) \supseteq h\Delta_0$ for all $h \in (0, \rho]$ by Assumption 1, and applying the linear change of variable $\tilde{u} = (\rho/h)u$, we have, for any $v \in \mathbb{R}^{D_m}$,

$$\begin{aligned} v^\top \mathcal{B}_\gamma v &= \int_{\mathcal{V}(h)} \left\{ v^\top \Phi_\gamma(u) \right\}^2 w_h(u) \mathrm{d}u \geq h^{-d} \int_{h\Delta_0} \left\{ v^\top \Phi_\gamma(u) \right\}^2 \mathrm{d}u \\ &= \rho^{-d} \int_{\rho\Delta_0} \left\{ v^\top \Phi_{(m,\rho)}(\tilde{u}) \right\}^2 \mathrm{d}\tilde{u} = \rho^{-d} v^\top \left\{ \int_{\rho\Delta_0} \Phi_{(m,\rho)}(\tilde{u}) \Phi_{(m,\rho)}^\top(\tilde{u}) \mathrm{d}\tilde{u} \right\} v. \end{aligned}$$

The same line of reasoning as in the proof of Lemma A.1 now shows that the integral on the right-hand side is a symmetric positive definite matrix. If $\mu(m, \rho) > 0$ denotes its smallest eigenvalue, then the last equation yields

$$\lambda_\gamma \geq \rho^{-d} \mu(m, \rho) =: \lambda_\star(m),$$

where the dependence on ρ is omitted. This proves the second claim of the lemma. □

Lemma A.3. *Let $(s, L) \in (0, \infty)^2$ be given. There exists a positive real constant $\mathfrak{F}_{s,L}$ (that also depends possibly on $\mathcal{D}$, t and ρ) such that*

$$\sup_{f \in \Sigma(s,L)} \|f\|_{\mathcal{V}(\rho)} \leq \mathfrak{F}_{s,L},$$

where recall from (4.2) that $\|f\|_{\mathcal{V}(\rho)} = \sup_{u \in \mathcal{V}(\rho)} |f(t+u)|$.

Proof of Lemma A.3. We adapt the proof found on p. 7 of Tsybakov (2004) to our framework. Let $f \in \Sigma(s, L)$ and let $\gamma_\rho = (\lfloor s \rfloor, \rho) \in \Gamma$. By definition of $\Sigma(s, L)$, there exists a polynomial $q \in \mathcal{P}_{\lfloor s \rfloor}$ such that, for any $u \in \mathcal{V}(\rho)$,

$$|f(t+u) - q(u)| \leq L\|u\|_\infty^s. \tag{A.2}$$

Let us write

$$q(u) = Q_{\lfloor s \rfloor}^\top \Phi_{\gamma_\rho}(u),$$

where $Q_{\lfloor s \rfloor}$ denotes the $D_{\lfloor s \rfloor} \times 1$ coordinate vector of q in the basis Φ_{γ_ρ}. Then, using the fact that $|\varphi_\alpha(u/\rho)| \leq 1$ for all $u \in \mathcal{V}(\rho)$ and all $\alpha \in \mathbb{N}_0^d$, we have

$$|q(u)| \leq \max_{|\alpha| \leq \lfloor s \rfloor} |\varphi_\alpha(u/\rho)| \times \|Q_{\lfloor s \rfloor}\|_1 \leq \sqrt{D_{\lfloor s \rfloor}}\, \|Q_{\lfloor s \rfloor}\|_2. \tag{A.3}$$

Since $\mathcal{P}_{\lfloor s \rfloor}$ is a vector space of finite dimension, the $L^1(w_\rho)$ and $L^2(w_\rho)$ norms are equivalent on that space. Thus, there exists some constant $c_{s,\rho} > 0$ such that

$$\begin{aligned} \int_{\mathcal{V}(\rho)} |q(u)| w_\rho(u) \mathrm{d}u &\geq c_{s,\rho} \sqrt{\int_{\mathcal{V}(\rho)} q^2(u) w_\rho(u) \mathrm{d}u} \\ &= c_{s,\rho} \sqrt{Q_{\lfloor s \rfloor}^\top \mathcal{B}_{\gamma_\rho} Q_{\lfloor s \rfloor}} \geq c_{s,\rho} \sqrt{\lambda_{\gamma_\rho}}\, \|Q_{\lfloor s \rfloor}\|_2, \end{aligned} \tag{A.4}$$

where the equality and the last inequality are a consequence of (A.1) and (3.4), respectively. Combining (A.3) and (A.4), followed by an application of the bound in (A.2), we obtain

$$\begin{aligned} |q(u)| &\leq \frac{\sqrt{D_{\lfloor s \rfloor}}}{c_{s,\rho}\sqrt{\lambda_{\gamma_\rho}}} \int_{\mathcal{V}(\rho)} |q(u)| w_\rho(u) \mathrm{d}u \\ &\leq \frac{\sqrt{D_{\lfloor s \rfloor}}}{c_{s,\rho}\sqrt{\lambda_{\gamma_\rho}}} \left\{ \int_{\mathcal{V}(\rho)} |f(t+u)| w_\rho(u) \mathrm{d}u + \int_{\mathcal{V}(\rho)} L\|u\|_\infty^s w_\rho(u) \mathrm{d}u \right\}. \end{aligned}$$

It remains to observe that $0 \leq w_\rho \leq \rho^{-d}$ and f is a density function to deduce

$$\|f\|_{\mathcal{V}(\rho)} \leq \sup_{u \in \mathcal{V}(\rho)} |q(u)| + L\rho^s \leq \frac{\sqrt{D_{\lfloor s \rfloor}}}{c_{s,\rho}\sqrt{\lambda_{\gamma_\rho}}} \left(1 + L\rho^s W_\rho\right) + L\rho^s. \tag{A.5}$$

The conclusion follows since $\lambda_{\gamma_\rho} > 0$ by Lemma A.1. □

Lemma A.4 (Concentration of $\hat{f}_\gamma(t)$ around its mean). *Let $\gamma = (m,h) \in \Gamma$ be given. Recall the definitions of c_γ and v_γ from (3.11) and (3.17). Assuming that $\|f\|_{\mathcal{V}(\rho)} < \infty$, we have, for all $x > 0$,*

$$\mathsf{P}\left[\left|\hat{f}_\gamma(t) - \mathsf{E}\left\{\hat{f}_\gamma(t)\right\}\right| > r_\gamma(v_\gamma, x)\right] \leq 2\exp(-x),$$

where $r_\gamma(v_\gamma, x) = \sqrt{2v_\gamma x} + c_\gamma x$.

Proof of Lemma A.4. By the definition of $\hat{f}_\gamma(t)$ in (3.6), observe that

$$\hat{f}_\gamma(t) - \mathsf{E}\left\{\hat{f}_\gamma(t)\right\} = \sum_{i=1}^{n} \left[Y_i(\gamma) - \mathsf{E}\{Y_i(\gamma)\}\right], \tag{A.6}$$

where

$$Y_i(\gamma) = \frac{1}{n}\left\{H_\gamma^\top(0) H_\gamma(X_i - t) w_h(X_i - t)\right\}.$$

This readily implies that $\sum_{i=1}^n \mathsf{E}\left\{Y_i^2(\gamma)\right\} = v_\gamma$. Furthermore, by Hölder's inequality, the submultiplicativity property of the spectral norm and the fact that $\|\Phi_\gamma(0)\Phi_\gamma^\top(0)\|_2 = 1$, note that, for all $u \in \mathcal{V}(h)$,

$$\begin{aligned}
\left\{H_\gamma^\top(0) H_\gamma(u)\right\}^2 &= \left\{\Phi_\gamma^\top(0)\mathcal{B}_\gamma^{-1}\Phi_\gamma(u)\right\}^2 \\
&= \Phi_\gamma^\top(u)\mathcal{B}_\gamma^{-1}\Phi_\gamma(0)\Phi_\gamma^\top(0)\mathcal{B}_\gamma^{-1}\Phi_\gamma(u) \\
&\leq \|\Phi_\gamma(u)\|_2 \|\mathcal{B}_\gamma^{-1}\Phi_\gamma(0)\Phi_\gamma^\top(0)\mathcal{B}_\gamma^{-1}\|_2 \|\Phi_\gamma(u)\|_2 \\
&\leq \|\Phi_\gamma(u)\|_2 \|\mathcal{B}_\gamma^{-1}\|_2 \|\Phi_\gamma(0)\Phi_\gamma^\top(0)\|_2 \|\mathcal{B}_\gamma^{-1}\|_2 \|\Phi_\gamma(u)\|_2 \\
&= \frac{1}{\lambda_\gamma^2}\|\Phi_\gamma(u)\|_2^2 \leq \frac{D_m}{\lambda_\gamma^2}.
\end{aligned} \tag{A.7}$$

By the bound (A.7), and the fact that $\mathrm{Supp}(w_h) = \mathcal{V}(h)$ and $0 \leq w_h \leq h^{-d}$ because of (3.2) and (3.1), we deduce

$$|Y_i(\gamma)| = \frac{w_h(X_i - t)}{n}\left|H_\gamma^\top(0) H_\gamma(X_i - t)\right| \leq \frac{1}{nh^d}\sqrt{\frac{D_m}{\lambda_\gamma^2}} = c_\gamma. \tag{A.8}$$

Using Bernstein's inequality (see Boucheron, Lugosi and Massart, 2013, Theorem 2.10), the conclusion follows. □

Lemma A.5 (Concentration of $\hat{v}_\gamma$ around v_γ). *Let $\gamma = (m,h) \in \Gamma$ be given. Recall the definitions of $\hat{v}_\gamma$ and v_γ from (3.10) and (3.17). Assuming that $nh^d W_h \geq (\log n)^3$ and $\|f\|_{\mathcal{V}(\rho)} < \infty$, there exists a constant $\kappa_1 > 0$ that depends on $\delta > 1$ and $\|f\|_{\mathcal{V}(\rho)}$ such that*

$$\mathsf{P}\left(|\hat{v}_\gamma - v_\gamma| > \varepsilon_\gamma\right) \leq 2\exp\left\{-\kappa_1(\log n)^3\right\},$$

where recall from (3.11) that $\varepsilon_\gamma = (\delta - 1)D_m W_h/(nh^d\lambda_\gamma^2)$.

Proof of Lemma A.5. Similarly to the proof of Lemma A.4, observe that

$$\hat{v}_\gamma - v_\gamma = \sum_{i=1}^{n}\left[Z_i(\gamma) - \mathsf{E}\{Z_i(\gamma)\}\right],$$

where

$$Z_i(\gamma) = \frac{1}{n^2}\left\{H_\gamma^\top(0) H_\gamma(X_i - t) w_h(X_i - t)\right\}^2.$$

By the bound (A.7), the fact that $w_h^4(X_i - t) = h^{-3d} w_h(X_i - t)$ and $0 \le w_h \le h^{-d}$ because of (3.2) and (3.1), and $W_h = \int_{\mathbb{R}^d} w_h(u)\mathrm{d}u$ as defined in (3.9), we have

$$\sum_{i=1}^n \mathsf{E}\{Z_i^2(\gamma)\} \le \frac{\|f\|_{\mathcal{V}(\rho)} D_m^2 W_h}{n^3 h^{3d} \lambda_\gamma^4} =: \tilde{v}_\gamma \quad \text{and} \quad |Z_i(\gamma)| \le \frac{D_m}{n^2 h^{2d} \lambda_\gamma^2} =: \tilde{c}_\gamma.$$

Using another version of Bernstein's inequality (see Boucheron, Lugosi and Massart, 2013, Corollary 2.11), we obtain

$$\mathsf{P}\left(|\hat{v}_\gamma - v_\gamma| > \varepsilon_\gamma\right) \le 2\exp\left\{-\frac{\varepsilon_\gamma^2}{2(\tilde{v}_\gamma + \tilde{c}_\gamma \varepsilon_\gamma)}\right\} = 2\exp\left\{-\frac{nh^d W_h(\delta-1)^2}{2(\|f\|_{\mathcal{V}(\rho)} + \delta - 1)}\right\}.$$

Since we assumed that $nh^d W_h \ge (\log n)^3$, the conclusion follows. □

Appendix B: Proofs of the propositions and theorems

Proof of Proposition 4.1. Set $t \in \mathcal{D}$ and $\gamma = (m, h) \in \Gamma$. Since $f \in \Sigma(s, L)$, there exists a polynomial $q \in \mathcal{P}_{\lfloor\!\lfloor s \rfloor\!\rfloor}$ such that, for any $u \in \mathcal{V}(h)$,

$$f(t+u) = q(u) + R(u) \quad \text{where} \quad |R(u)| \le L\|u\|_\infty^s. \tag{B.1}$$

Consider the decomposition

$$q(u) = \sum_{|\alpha| \le \lfloor\!\lfloor s \rfloor\!\rfloor} h^{|\alpha|} q_\alpha \varphi_\alpha\left(\frac{u}{h}\right) = Q_m^\top \Phi_\gamma(u) + \mathbb{1}_{[m+1,\infty)}(\lfloor\!\lfloor s \rfloor\!\rfloor) \sum_{m < |\alpha| \le \lfloor\!\lfloor s \rfloor\!\rfloor} q_\alpha \varphi_\alpha(u), \tag{B.2}$$

where Q_m denotes the vector of the first D_m components of the full $D_{\lfloor\!\lfloor s \rfloor\!\rfloor} \times 1$ coordinate vector of q in the basis $\Phi_{(\lfloor\!\lfloor s \rfloor\!\rfloor, h)}$. Given (B.1) and the decomposition of q in (B.2), we can write

$$f(t+u) = Q_m^\top \Phi_\gamma(u) + \widetilde{R}(u), \tag{B.3}$$

where

$$\begin{aligned}
|\widetilde{R}(u)| &= \left| \mathbb{1}_{[m+1,\infty)}(\lfloor\!\lfloor s \rfloor\!\rfloor) \sum_{m < |\alpha| \le \lfloor\!\lfloor s \rfloor\!\rfloor} q_\alpha \varphi_\alpha(u) + R(u) \right| \\
&\le \mathbb{1}_{[m+1,\infty)}(\lfloor\!\lfloor s \rfloor\!\rfloor) \left(\sum_{m < |\alpha| \le \lfloor\!\lfloor s \rfloor\!\rfloor} |q_\alpha| \right) \|u\|_\infty^{\beta_m(s)} + L\|u\|_\infty^s \le \mathfrak{L}_{m,s,L} \|u\|_\infty^{\beta_m(s)},
\end{aligned}$$

for some appropriate positive real constant $\mathfrak{L}_{m,s,L}$ that depends on m, s and L, and also denoting $\beta_m(s) = \min(m+1, s)$. Note that we can choose $\mathfrak{L}_{m,s,L} = L$ whenever $m \ge \lfloor\!\lfloor s \rfloor\!\rfloor$ because the indicator $\mathbb{1}_{[m+1,\infty)}(\lfloor\!\lfloor s \rfloor\!\rfloor)$ is simply equal to zero in that case.

By using the definition of $\hat{f}_\gamma(t)$ in (3.6) and the definition of H_γ in (3.5), followed by an application of (B.3) together with the fact that $\Phi_\gamma^\top(0)Q_m = f(t)$, we have

$$\begin{aligned}
\mathsf{E}\left\{\hat{f}_\gamma(t)\right\} &= \int_{\mathbb{R}^d} H_\gamma^\top(0) H_\gamma(u) w_h(u) f(t+u)\mathrm{d}u \\
&= \Phi_\gamma^\top(0)\mathcal{B}_\gamma^{-1} \int_{\mathbb{R}^d} \Phi_\gamma(u) w_h(u) f(t+u)\mathrm{d}u \\
&= f(t) + \Phi_\gamma^\top(0)\mathcal{B}_\gamma^{-1} \int_{\mathbb{R}^d} \Phi_\gamma(u) w_h(u) \widetilde{R}(u)\mathrm{d}u.
\end{aligned} \tag{B.4}$$

Since $\|\Phi_\gamma(0)\|_2 = 1$, $\mathrm{Supp}(w_h) = \mathcal{V}(h)$, $\max_{|\alpha|\le m} |\varphi_\alpha(u/h)| \le 1$ for all $u \in \mathcal{V}(h)$ and $\|u\|_\infty \le h$ for all $u \in \mathcal{V}(h)$, the error term on the right-hand side of (B.4) satisfies

$$\begin{aligned}
\left|\Phi_\gamma^\top(0)\mathcal{B}_\gamma^{-1} \int_{\mathbb{R}^d} \Phi_\gamma(u) w_h(u) \widetilde{R}(u)\mathrm{d}u\right| &\le \|\Phi_\gamma(0)\|_2 \left\|\mathcal{B}_\gamma^{-1} \int_{\mathbb{R}^d} \Phi_\gamma(u) w_h(u) \widetilde{R}(u)\mathrm{d}u\right\|_2 \\
&\le \left\|\mathcal{B}_\gamma^{-1}\right\|_2 \left\|\int_{\mathbb{R}^d} \Phi_\gamma(u) w_h(u) \widetilde{R}(u)\mathrm{d}u\right\|_2 \\
&\le \lambda_\gamma^{-1}\sqrt{D_m} \int_{\mathbb{R}^d} \max_{|\alpha|\le m} |\varphi_\alpha(u/h)| \times |\widetilde{R}(u)| w_h(u)\mathrm{d}u \\
&\le \lambda_\gamma^{-1}\sqrt{D_m}\, \mathfrak{L}_{m,s,L} \int_{\mathbb{R}^d} \|u\|_\infty^{\beta_m(s)} w_h(u)\mathrm{d}u \\
&\le \lambda_\gamma^{-1}\sqrt{D_m}\, \mathfrak{L}_{m,s,L}\, h^{\beta_m(s)} W_h.
\end{aligned}$$

This concludes the proof. □

Proof of Proposition 4.2. By using $w_h^2(u) = h^{-d} w_h(u)$ because of (3.2) and (3.1), the bound (A.7) from proof of Lemma A.4, and $W_h = \int_{\mathbb{R}^d} w_h(u)\mathrm{d}u$ as defined in (3.9), we have

$$\begin{aligned}
\mathsf{Var}\left\{\hat{f}_\gamma(t)\right\} \le v_\gamma &= \frac{1}{n}\int_{\mathbb{R}^d} \left\{H_\gamma^\top(0) H_\gamma(u) w_h(u)\right\}^2 f(t+u)\mathrm{d}u \\
&\le \frac{\|f\|_{\mathcal{V}(\rho)}}{nh^d} \int_{\mathbb{R}^d} \left\{H_\gamma^\top(0) H_\gamma(u)\right\}^2 w_h(u)\mathrm{d}u \le \frac{\|f\|_{\mathcal{V}(\rho)}}{nh^d} \times \frac{D_m W_h}{\lambda_\gamma^2}.
\end{aligned}$$

This concludes the proof. □

Proof of Theorem 4.3. Set $t \in \mathcal{D}$ and $\gamma = (m,h) \in \Gamma_n$. By following the line of proof of Theorem 1 (Steps 2 and 3) in Bertin and Klutchnikoff (2017), we obtain

$$|\hat{f}(t) - f(t)| \le 2\big(\hat{A}_\gamma + \hat{\mathbb{U}}_\gamma\big) + |\hat{f}_\gamma(t) - f(t)|,$$

where $\hat{A}_\gamma$ and $\hat{\mathbb{U}}_\gamma$ are defined in (3.13) and (3.14), respectively. Furthermore, the paper also establishes that

$$\hat{A}_\gamma \le 2\mathbb{B}_\gamma + 2T, \quad \text{where} \quad T = \max_{\gamma' \in \Gamma_n} \left[\left|\hat{f}_{\gamma'}(t) - \mathsf{E}\left\{\hat{f}_{\gamma'}(t)\right\}\right| - \hat{\mathbb{U}}_{\gamma'}\right]_+.$$

Recall that $(\cdot)_+ = \max\{\cdot, 0\}$. Together, the last two equations imply

$$R_n(\hat{f}, f) \leq 4\mathbb{B}_\gamma + 2\left\{\mathsf{E}(\hat{\mathbb{U}}_\gamma^2)\right\}^{1/2} + R_n(\hat{f}_\gamma, f) + 4\left\{\mathsf{E}(T^2)\right\}^{1/2}. \tag{B.5}$$

Using the triangle inequality for the $\mathsf{E}\{(\cdot)^2\}^{1/2}$ norm, $\mathsf{E}(\hat{v}_\gamma) = v_\gamma$, and the subadditivity of the function $x \mapsto x^{1/2}$, we obtain

$$\begin{aligned}
\left\{\mathsf{E}(\hat{\mathbb{U}}_\gamma^2)\right\}^{1/2} &\leq \sqrt{\mathsf{E}\left\{2(\hat{v}_\gamma + \varepsilon_\gamma)\mathrm{pen}(\gamma)\right\}} + c_\gamma \mathrm{pen}(\gamma) \\
&= \sqrt{2(v_\gamma + \varepsilon_\gamma)\mathrm{pen}(\gamma)} + c_\gamma \mathrm{pen}(\gamma) \\
&\leq \sqrt{2v_\gamma \mathrm{pen}(\gamma)} + c_\gamma \mathrm{pen}(\gamma) + \sqrt{2\varepsilon_\gamma \mathrm{pen}(\gamma)} \\
&= \mathbb{U}_\gamma + \sqrt{2\varepsilon_\gamma \mathrm{pen}(\gamma)} = \mathbb{U}_\gamma + \sqrt{(\delta-1)/\|f\|_{\mathcal{V}(\rho)}}\sqrt{2v_\gamma^\star \mathrm{pen}(\gamma)} \\
&\leq \left\{1 + \sqrt{(\delta-1)/\|f\|_{\mathcal{V}(\rho)}}\right\} \mathbb{U}_\gamma^\star,
\end{aligned} \tag{B.6}$$

where the quantities $\hat{v}_\gamma$, ε_γ, c_γ, $\mathrm{pen}(\gamma)$ and $\mathbb{U}_\gamma$ are all defined in Section 3.2, and $\mathbb{U}_\gamma^\star$ is defined in (4.5). Note also that, since δ and $|\log h|$ are both at least 1 by assumption (we assumed $\delta > 1$ in Section 3.2, and also $h \in (0, \rho] \subseteq (0, e^{-1}]$ at the beginning of Section 2), we have $\mathrm{pen}(\gamma) \geq 1$, and thus $\sqrt{v_\gamma} \leq \sqrt{2v_\gamma \mathrm{pen}(\gamma)} + c_\gamma \mathrm{pen}(\gamma) = \mathbb{U}_\gamma$. Therefore,

$$\begin{aligned}
R_n(\hat{f}_\gamma, f) &\leq \left|\mathsf{E}\left\{\hat{f}_\gamma(t)\right\} - f(t)\right| + \left(\mathsf{E}\left[\left|\hat{f}_\gamma(t) - \mathsf{E}\left\{\hat{f}_\gamma(t)\right\}\right|^2\right]\right)^{1/2} \\
&\leq \mathbb{B}_\gamma + \sqrt{v_\gamma} \leq \mathbb{B}_\gamma + \mathbb{U}_\gamma \leq \mathbb{B}_\gamma + \mathbb{U}_\gamma^\star,
\end{aligned} \tag{B.7}$$

where the last inequality follows from $v_\gamma \leq v_\gamma^\star$ in (4.3).

By applying the bounds (B.6) and (B.7) back into (B.5), we get

$$R_n(\hat{f}, f) \leq \left[5\mathbb{B}_\gamma + \left\{3 + 2\sqrt{(\delta-1)/\|f\|_{\mathcal{V}(\rho)}}\right\} \mathbb{U}_\gamma^\star\right] + 4\left\{\mathsf{E}(T^2)\right\}^{1/2}. \tag{B.8}$$

To conclude, it remains to study the term $\mathsf{E}(T^2)$. To accomplish this, define the event

$$\mathcal{A} = \bigcap_{\gamma \in \Gamma_n} \left\{|\hat{v}_\gamma - v_\gamma| \leq \varepsilon_\gamma\right\}. \tag{B.9}$$

We decompose the expectation of interest as follows:

$$\mathsf{E}(T^2) = \mathsf{E}(T^2 \mathbb{1}_{\mathcal{A}}) + \mathsf{E}(T^2 \mathbb{1}_{\mathcal{A}^c}).$$

Let us bound the term $\mathsf{E}(T^2 \mathbb{1}_{\mathcal{A}^c})$ first. Notice that $D_m = \binom{m+d}{d} \leq \mathfrak{K}_d m^d$ for some positive real constant $\mathfrak{K}_d$, which implies

$$T \leq \max_{\gamma \in \Gamma_n} \left|\hat{f}_\gamma(t) - \mathsf{E}\left\{\hat{f}_\gamma(t)\right\}\right| \leq 2n \max_{\gamma \in \Gamma_n} c_\gamma \leq 2^{d+1}\sqrt{\mathfrak{K}_d}(\log n)^{d/2-3} n \max_{\gamma \in \Gamma_n} \frac{1}{\lambda_\gamma},$$

where the second inequality follows from (A.6) and (A.8), and the last inequality follows from the definition of c_γ in (3.11) and the aforementioned bound on D_m together with the fact that $m \leq \log n$

and $(nh^d)^{-1} \le W_h(\log n)^{-3} \le 2^d(\log n)^{-3}$ for $(m,h) \in \Gamma_n$; recall (3.9). Using our assumption on λ_γ in (4.4), together with that fact that $m \le \log n$ and $h \ge \rho n^{-1}$ (recall (3.7)), we obtain, for n large enough,

$$T \le 2^{d+1}\sqrt{\mathfrak{K}_d}(\log n)^{d/2-3} n \exp\left\{2b(\log n)^2\right\}.$$

By a union bound and the concentration bound on $\hat{v}_\gamma$ in Lemma A.5, this implies

$$\begin{aligned}\mathsf{E}(T^2 \mathbb{1}_{\mathcal{A}^c}) &\le 4^{d+1}\mathfrak{K}_d(\log n)^{d-6} n^2 \exp\left\{4b(\log n)^2\right\} \times \mathsf{P}(\mathcal{A}^c) \\ &\le 4^{d+1}\mathfrak{K}_d(\log n)^{d-6} n^2 \exp\left\{4b(\log n)^2\right\} \times \mathrm{card}(\Gamma_n) \times 2\exp\left\{-\kappa_1(\log n)^3\right\}.\end{aligned}$$

Since $\mathrm{card}(\Gamma_n) \le \log n$, we deduce that

$$\mathsf{E}(T^2 \mathbb{1}_{\mathcal{A}^c}) \lesssim \exp\left\{-(\kappa_1/2)(\log n)^3\right\}, \tag{B.10}$$

with much room to spare.

It remains to study $\mathsf{E}(T^2 \mathbb{1}_{\mathcal{A}})$. Note that under the event $\mathcal{A}$ defined in (B.9), we have $\hat{\mathbb{U}}_\gamma \ge \mathbb{U}_\gamma$ for all $\gamma \in \Gamma_n$ (to see this, compare (3.14) and (3.16)). This implies that

$$\mathsf{E}(T^2 \mathbb{1}_{\mathcal{A}}) \le \mathsf{E}(\tilde{T}^2)$$

where

$$\tilde{T} = \max_{\gamma' \in \Gamma_n} \left[\left|\hat{f}_{\gamma'}(t) - \mathsf{E}\left\{\hat{f}_{\gamma'}(t)\right\}\right| - \mathbb{U}_{\gamma'}\right]_+.$$

For simplicity of notations, define $r_{\gamma'}(x) = r_{\gamma'}(v_{\gamma'}, x)$; recall (3.12). Using integration by parts and the change of variable $u = r_{\gamma'}(x)$, we obtain

$$\begin{aligned}\mathsf{E}(\tilde{T}^2) &\le \sum_{\gamma' \in \Gamma_n} \mathsf{E}\left(\left[\left|\hat{f}_{\gamma'}(t) - \mathsf{E}\left\{\hat{f}_{\gamma'}(t)\right\}\right| - \mathbb{U}_{\gamma'}\right]_+^2\right) \\ &= \sum_{\gamma' \in \Gamma_n} \int_0^\infty 2u\mathsf{P}\left[\left|\hat{f}_{\gamma'}(t) - \mathsf{E}\left\{\hat{f}_{\gamma'}(t)\right\}\right| > \mathbb{U}_{\gamma'} + u\right] \mathrm{d}u \\ &= \sum_{\gamma' \in \Gamma_n} \int_0^\infty 2r_{\gamma'}(x) r'_{\gamma'}(x) \mathsf{P}\left[\left|\hat{f}_{\gamma'}(t) - \mathsf{E}\left\{\hat{f}_{\gamma'}(t)\right\}\right| > r_{\gamma'}\left\{\mathrm{pen}(\gamma')\right\} + r_{\gamma'}(x)\right] \mathrm{d}x.\end{aligned}$$

Now, using the fact that $r_{\gamma'}(\cdot)$ is a sub-additive function which also satisfies $x r'_{\gamma'}(x) \le r_{\gamma'}(x)$ for all $x > 0$, we get

$$\mathsf{E}(\tilde{T}^2) \le \sum_{\gamma' \in \Gamma_n} \int_0^\infty 2x^{-1} r^2_{\gamma'}(x) \mathsf{P}\left[\left|\hat{f}_{\gamma'}(t) - \mathsf{E}\left\{\hat{f}_{\gamma'}(t)\right\}\right| > r_{\gamma'}\left\{\mathrm{pen}(\gamma') + x\right\}\right] \mathrm{d}x.$$

As we noted earlier just above (B.7), we have $\mathrm{pen}(\gamma') \ge 1$ because of our assumptions $\delta > 1$ and $-|\log\rho| \ge 1$, which allows us to apply the concentration of $\hat{f}_{\gamma'}(t)$ around its mean in Lemma A.4. Together with the identity $r^2_{\gamma'}(x) \le 4v_{\gamma'}x + 2c^2_{\gamma'}x^2$ and $|\log h_{\ell'}| = |\log\rho - \ell'| \ge \ell' - |\log\rho| \ge \ell'$, this yields

$$\begin{aligned}\mathsf{E}(\tilde{T}^2) &\le \sum_{\gamma' \in \Gamma_n} \int_0^\infty 2x^{-1} r^2_{\gamma'}(x) \times 2\exp\left(-d\delta|\log h_{\ell'}| - \Lambda_{\gamma'} - x\right) \mathrm{d}x \\ &\le I_1 \sum_{\gamma' \in \Gamma_n} (2v_{\gamma'} + c^2_{\gamma'}) \exp\left(-d\delta\ell' - \Lambda_{\gamma'}\right),\end{aligned}$$

where $I_1 = \int_0^\infty 8\max(1,x)\exp(-x)\mathrm{d}x < \infty$. Since $nh_{\ell'}^d W_{h_{\ell'}} \geq (\log n)^3$, $\Lambda_{\gamma'} = 2|\log(\lambda_{\gamma'})|$ and $h_{\ell'} = \rho\exp(-\ell')$, we can write, using $v_\gamma \leq v_\gamma^\star$ in Proposition 4.2 and $c_{\gamma'}$ in (3.11),

$$\mathsf{E}(\tilde{T}^2) \leq \frac{I_1}{n\rho^d}\left\{2\,\|f\|_{\mathcal{V}(\rho)} + \frac{1}{(\log n)^3}\right\}\sum_{\ell'\in\mathcal{L}_n} D_{m_{\ell'}} W_{h_{\ell'}} \exp\{-d(\delta-1)\ell'\}\,.$$

Moreover, since $D_{m_{\ell'}} \leq \mathfrak{K}_d m_{\ell'}^d$, $m_{\ell'} \leq \log n$ for all $\ell' \in \mathcal{L}_n$, and $W_{h_{\ell'}} \leq 2^d$, we can write

$$\mathsf{E}(\tilde{T}^2) \leq \frac{I_1\mathfrak{K}_d(\log n)^d 2^d}{n\rho^d}\left\{2\,\|f\|_{\mathcal{V}(\rho)} + \frac{1}{(\log n)^3}\right\}\sum_{\ell'\in\mathcal{L}_n} \exp\{-d(\delta-1)\ell'\}\,.$$

Since $d(\delta-1)$ is a positive number (we assumed $\delta > 1$ in Section 3.2), the exponential terms above are summable, so we have

$$\mathsf{E}(T^2\mathbb{1}_{\mathcal{A}}) \leq \mathsf{E}(\tilde{T}^2) \lesssim \frac{(\log n)^d}{n}. \tag{B.11}$$

By combining (B.10) and (B.11), we conclude that $\mathsf{E}(T^2) \lesssim (\log n)^d/n$. Together with (B.8), the proof of (4.6) is complete. □

Proof of Proposition 4.4. In this proof, the parameters $\gamma = (m,h)$ are chosen to be the ones in (4.7). By combining the bounds on the bias and variance found in Propositions 4.1 and 4.2 together with Lemma A.2 ($W_h \leq 2^d$ and $\inf_{h\in(0,\rho]}\lambda_\gamma \geq \lambda_\star(m) > 0$ under Assumption 1) and Lemma A.3, we obtain

$$\begin{aligned} R_n(\hat{f}_\gamma, f) &\leq \frac{\sqrt{D_m}}{\lambda_\gamma} \times \left\{W_h^2 \mathfrak{L}_{m,s,L}^2\, h^{2s} + \frac{W_h\|f\|_{\mathcal{V}(\rho)}}{nh^d}\right\}^{1/2} \\ &\leq \frac{2^d\sqrt{D_m}}{\lambda_\star(m)} \times \left[\mathfrak{L}_{m,s,L}^2\, h^{2s} + \frac{\mathfrak{F}_{s,L}}{2^d nh^d}\right]^{1/2}. \end{aligned}$$

Our choice of h in (4.7) is simply the minimizer of the last bound. In particular, it should be noted that h decreases as n increases. Elementary computations show that, with this choice of h, the above yields

$$R_n(\hat{f}_\gamma, f) \leq C(s,L)N_n(s,L),$$

where $C(s,L)$ is the positive real constant defined in (4.9) and $N_n(s,L) = n^{-s/(2s+d)}$ is the rate of convergence defined in (4.8). This proves the upper bound.

To prove the lower bound, we use Lemma 3 of Lepski (2015). It is sufficient to construct two functions f_0 and f_1 that satisfy the following properties:

1. f_0 and f_1 are two density functions that belong to $\Sigma(s,L)$.
2. There exists a positive real constant $A = A(s,L) > 0$ such that $f_1(t) - f_0(t) = Ah^s$.
3. There exists a positive real constant $\mathfrak{a} = \mathfrak{a}(s,L) > 0$ such that

$$\mathsf{E}_{f_0,n}\left\{\prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)}\right\}^2 \leq \mathfrak{a},$$

where $\mathsf{E}_{f_0,n}$ denotes the expectation with respect to the law $\mathsf{P}_{f_0,n}$ of the random sample $\mathbb{X}_n$, provided f_0 is the true density of the observations X_i.

To construct these densities, we consider an auxiliary density function $\psi \in C^\infty(\mathbb{R}^d)$ whose support is Δ and which satisfies $\psi(0) = \mathrm{Leb}(\mathcal{D})$. Such a function can easily be constructed. We also define, for any $u \in \mathbb{R}^d$,

$$\psi_h(u) = \psi\left(\frac{u}{h}\right)\mathbb{1}_{\mathcal{V}(h)}(u) \quad \text{and} \quad c_h = \frac{1}{\mathrm{Leb}(\mathcal{D})}\int_{\mathbb{R}^d}\psi_h(u)\mathrm{d}u.$$

Furthermore, for all $x \in \mathbb{R}^d$, define

$$f_0(x) = \frac{\mathbb{1}_{\mathcal{D}}(x)}{\mathrm{Leb}(\mathcal{D})} \quad \text{and} \quad f_1(x) = f_0(x) + Ah^s\tilde{\psi}_h(x-t)\mathbb{1}_{\mathcal{D}}(x),$$

for some constant $A > 0$ to be chosen later, where

$$\tilde{\psi}_h(u) = \frac{\{\psi_h(u)/\mathrm{Leb}(\mathcal{D})\} - c_h}{1 - c_h}, \quad u \in \mathbb{R}^d.$$

Since $\tilde{\psi}(0) = 1$, the second point follows.

Now, since ψ is a density function, we have $c_h \le h^d/\mathrm{Leb}(\mathcal{D})$. Hence, using the fact that the function $x \mapsto -x/(1-x)$ is decreasing on $(0,1)$, and taking n large enough (and thus h small enough by (4.7)) so that $h^d/\mathrm{Leb}(\mathcal{D}) \le 1/2$, we have

$$\tilde{\psi}_h(u) \ge \frac{-c_h}{1-c_h} \ge \frac{-h^d/\mathrm{Leb}(\mathcal{D})}{1 - h^d/\mathrm{Leb}(\mathcal{D})} \ge -2\frac{h^d}{\mathrm{Leb}(\mathcal{D})},$$

which also implies that $f_1 \ge 0$ for large n. We easily deduce that f_1 is a density function using the fact that $\int_{\mathcal{D}-t}\tilde{\psi}_h(u)\mathrm{d}u = 0$. Since f_0 is also a density function, it remains to prove that both f_0 and f_1 belong to $\Sigma(s,L)$ to obtain the first point. This is obvious for f_0. Now, since $\psi \in C^\infty(\mathbb{R}^d)$ is supported on Δ, there exists a positive real constant $L_s(\psi) > 0$ such that $\psi \in \Sigma\{s, L_s(\psi)\}$. In turn, this implies that $f_1 \in \Sigma\{s, A \times 2\,L_s(\psi)/\mathrm{Leb}(\mathcal{D})\}$. By choosing the constant $A = \{2\,L_s(\psi)/\mathrm{Leb}(\mathcal{D})\}^{-1}L$, we deduce that $f_1 \in \Sigma(s,L)$.

Lastly, it remains to prove the third point. Note that

$$\begin{aligned}
\mathsf{E}_{f_0,n}\left\{\prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)}\right\}^2 &= \left\{\int_{\mathcal{D}}\frac{f_1^2(x)}{f_0(x)}\mathrm{d}x\right\}^n \\
&= \left[\int_{\mathcal{D}-t}\frac{\{1+\mathrm{Leb}(\mathcal{D})Ah^s\tilde{\psi}_h(u)\}^2}{\mathrm{Leb}(\mathcal{D})}\mathrm{d}u\right]^n \\
&= \left\{1 + \mathrm{Leb}(\mathcal{D})A^2h^{2s}\int_{\mathcal{D}-t}\tilde{\psi}_h^2(u)\mathrm{d}u\right\}^n.
\end{aligned} \tag{B.12}$$

To obtain the last equality, we used the fact that $\int_{\mathcal{D}-t}\tilde{\psi}_h(u)\mathrm{d}u = 0$. It remains to study the integral that appears on the last line. Using the elementary identity $(a+b)^2 \le 2a^2 + 2b^2$ together with the aforementioned bound $c_h \le h^d/\mathrm{Leb}(\mathcal{D})$, and taking n large enough (and thus h small enough by (4.7)) so that

$$1 - c_h \ge 1/2 \quad \text{and} \quad h^d \le 1/\mathrm{Leb}(\mathcal{D}),$$

we have

$$\begin{aligned}\int_{\mathcal{D}-t} \tilde{\psi}_h^2(u)\mathrm{d}x &\leq \frac{2}{\{\mathrm{Leb}(\mathcal{D})\}^2}\int_{\mathcal{V}(h)} \frac{\psi^2(u/h)}{(1-c_h)^2}\mathrm{d}u + 2\int_{\mathcal{D}} \frac{(-c_h)^2}{(1-c_h)^2}\mathrm{d}u \\ &\leq \frac{8h^d}{\{\mathrm{Leb}(\mathcal{D})\}^2}\int_{\Delta}\psi^2(u)\mathrm{d}u + \frac{8h^{2d}}{\mathrm{Leb}(\mathcal{D})} \\ &\leq \frac{8}{\{\mathrm{Leb}(\mathcal{D})\}^2}\left\{\int_{\Delta}\psi^2(u)\mathrm{d}u + 1\right\}h^d.\end{aligned} \tag{B.13}$$

Finally, since $1+x \leq e^x$, putting (B.13) into (B.12) yields the following bound:

$$\begin{aligned}\mathsf{E}_{f_0,n}\left\{\prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)}\right\}^2 &\leq \left[1 + \frac{8A^2}{\mathrm{Leb}(\mathcal{D})}\left\{\int_{\Delta}\psi^2(u)\mathrm{d}u + 1\right\}h^{2s+d}\right]^n \\ &\leq \exp\left[\frac{8A^2}{\mathrm{Leb}(\mathcal{D})}\left\{\int_{\Delta}\psi^2(u)\mathrm{d}u + 1\right\}nh^{2s+d}\right] \\ &= \exp\left[\frac{d\,\mathfrak{F}_{s,L}}{2^{d+1}s\mathfrak{L}_{m,s,L}^2} \times \frac{8A^2}{\mathrm{Leb}(\mathcal{D})}\left\{\int_{\Delta}\psi^2(u)\mathrm{d}u + 1\right\}\right] =: \mathfrak{a}.\end{aligned} \tag{B.14}$$

This concludes the proof. □

Proof of Theorem 4.5. We prove the upper bound first. Let $(s, L) \in (0,\infty)^2$ be given. Under Assumption 1, Lemma A.2 shows that

$$\lambda_\gamma \geq \min_{0\leq m' \leq \lfloor 2s+d \rfloor} \lambda_{(m',h)} \geq \lambda_\bullet(s,d) > 0,$$

where recall the quantity $\lambda_\bullet(s,d) = \min_{0\leq m'\leq \lfloor 2s+d\rfloor}\lambda_\star(m')$ depends only on s and d. In particular, the condition (4.4) of Theorem 4.3 is satisfied. Moreover, Lemma A.3 shows that there exists an appropriate constant $\mathfrak{F}_{s,L} > 0$ such that

$$\sup_{f\in\Sigma(s,L)} \|f\|_{\mathcal{V}(\rho)} \leq \mathfrak{F}_{s,L} < \infty.$$

Therefore, Theorem 4.3 can be applied to obtain

$$R_n(\hat{f}, f) \leq \min_{\gamma\in\Gamma_n}\left[5\mathbb{B}_\gamma + \left\{3 + 2\sqrt{(\delta-1)/\|f\|_{\mathcal{V}(\rho)}}\right\}\mathbb{U}_\gamma^\star\right] + O\left\{\sqrt{\frac{(\log n)^d}{n}}\right\}. \tag{B.15}$$

Now, define

$$\ell_0 = \left\lfloor \frac{1}{2s+d}\log\left\{\frac{(\mathfrak{L}_{s,L}^\star)^2}{\mathfrak{F}_{s,L}} \times \frac{n}{\log n}\right\} + \log\rho \right\rfloor,$$

where recall $\mathfrak{L}_{s,L}^\star = \max_{0\leq m'\leq\lfloor 2s+d\rfloor}\mathfrak{L}_{m',s,L}$ depends only on s, L and d. In the following, consider n large enough so that $\ell_0 \geq 1$. Using this definition, we have

$$1 \leq h_{\ell_0}\left\{\frac{\mathfrak{F}_{s,L}}{(\mathfrak{L}_{s,L}^\star)^2} \times \frac{\log n}{n}\right\}^{-1/(2s+d)} \leq e. \tag{B.16}$$

In the following, we consider $\gamma_0 = (m_{\ell_0}, h_{\ell_0})$ which belongs to Γ_n. From (4.10), recall that

$$m_{\ell_0} = \left\lfloor \frac{\log n}{2\ell_0} \right\rfloor.$$

Let us remark that if $\gamma = (m, h) \in \Gamma_n$ is such that $\gamma \preceq \gamma_0$, then $m \leq m_{\ell_0} \leq 2s + d$ for n large enough.

By Proposition 4.1, we have directly, for n large enough,

$$\begin{aligned} \mathbb{B}_{\gamma_0} &= \max_{\substack{\gamma \in \Gamma_n \\ \gamma \preceq \gamma_0}} \left| \mathsf{E}\left\{ \hat{f}_\gamma(t) \right\} - f(t) \right| \\ &\leq \max_{\substack{\gamma \in \Gamma_n \\ \gamma \preceq \gamma_0}} \frac{W_h \sqrt{D_m}}{\lambda_\gamma} \times \mathfrak{L}_{m,s,L}\, h^{\beta_m(s)} \leq C^\star(s,d) \max_{\substack{\gamma \in \Gamma_n \\ \gamma \preceq \gamma_0}} \mathfrak{L}_{m,s,L}\, h^{\beta_m(s)}, \end{aligned} \tag{B.17}$$

where

$$C^\star(s,d) = \frac{2^d \sqrt{D^\star(s,d)}}{\lambda_\bullet(s,d)} \quad \text{and} \quad D^\star(s,d) = \max_{0 \leq m' \leq 2s+d} D_{m'}. \tag{B.18}$$

It remains to bound the maximum that appears on the right-hand side of (B.17). To achieve this, fix $\gamma \in \Gamma_n$ such that $\gamma \preceq \gamma_0$. By definition, there exists $\ell \in \mathcal{L}_n$ such that $\gamma = (m_\ell, h_\ell)$. Assume first that $m_\ell + 1 < s$, so that $\beta_{m_\ell}(s) = m_\ell + 1$. Then, using $h_\ell = \rho e^{-\ell} \leq e^{-\ell}$, we have

$$\mathfrak{L}_{m_\ell,s,L} h_\ell^{\beta_{m_\ell}(s)} \leq \mathfrak{L}^\star_{s,L} \exp\left(-\frac{\ell \log n}{2\ell} \right) = \frac{\mathfrak{L}^\star_{s,L}}{\sqrt{n}}. \tag{B.19}$$

Similarly, if we assume $m_\ell + 1 \geq s$, then using the upper bound in (B.16), we have

$$\begin{aligned} \mathfrak{L}_{m_\ell,s,L} h_\ell^{\beta_{m_\ell}(s)} &\leq \mathfrak{L}^\star_{s,L} h_\ell^s \leq \mathfrak{L}^\star_{s,L} h_{\ell_0}^s \\ &\leq e^s\, \mathfrak{F}_{s,L}^{s/(2s+d)} (\mathfrak{L}^\star_{s,L})^{d/(2s+d)} \left(\frac{\log n}{n} \right)^{s/(2s+d)}. \end{aligned} \tag{B.20}$$

Therefore, by putting the estimates (B.19) and (B.20) back into (B.17), we have, for n large enough,

$$\mathbb{B}_{\gamma_0} \leq e^s\, C^\star(s,d)\, \mathfrak{F}_{s,L}^{s/(2s+d)} (\mathfrak{L}^\star_{s,L})^{d/(2s+d)} \left(\frac{\log n}{n} \right)^{s/(2s+d)}. \tag{B.21}$$

Now, let us bound $\mathbb{U}^\star_{\gamma_0}$. By combining (4.5) and (3.12), we can write

$$\mathbb{U}^\star_{\gamma_0} = \sqrt{2 v^\star_{\gamma_0} \mathrm{pen}(\gamma_0)} + c_{\gamma_0} \mathrm{pen}(\gamma_0), \quad \text{with} \quad \mathrm{pen}(\gamma_0) = d\delta |\log h_{\ell_0}| + 2 |\log(\lambda_{\gamma_0})|.$$

Let us first control the terms c_{γ_0} and $v^\star_{\gamma_0}$ defined (3.11) and (4.3), respectively. Using the notation in (B.18), we have

$$\begin{aligned} c_{\gamma_0} &\leq \frac{\sqrt{D^\star(s,d)}}{\lambda_{\gamma_0}} \times \frac{1}{n h_{\ell_0}^d} \leq \frac{C^\star(s,d)}{2^d} \times \frac{1}{n h_{\ell_0}^d}, \\ v^\star_{\gamma_0} &\leq \frac{2^d D^\star(s,d)}{\lambda^2_{\gamma_0}} \times \frac{\mathfrak{F}_{s,L}}{n h_{\ell_0}^d} \leq \frac{\{C^\star(s,d)\}^2 \mathfrak{F}_{s,L}}{2^d} \times \frac{1}{n h_{\ell_0}^d}. \end{aligned} \tag{B.22}$$

Now, using the subadditivity of the function $x \mapsto \sqrt{2v^\star_{\gamma_0}x} + c_{\gamma_0}x$, the four bounds in (B.22), and the fact that the functions $\lambda \mapsto |\log\lambda|/\lambda^k$ ($k \in \{1,2\}$) are decreasing for $\lambda \in (0,1)$ and uniformly bounded by 1 for $\lambda \geq 1$, we obtain

$$\begin{aligned}\mathbb{U}^\star_{\gamma_0} &\leq \sqrt{2v^\star_{\gamma_0} d\delta |\log h_{\ell_0}|} + c_{\gamma_0} d\delta |\log h_{\ell_0}| + \sqrt{4v^\star_{\gamma_0}|\log(\lambda_{\gamma_0})|} + 2c_{\gamma_0}|\log(\lambda_{\gamma_0})| \\ &\leq \sqrt{\frac{\{C^\star(s,d)\}^2 \mathfrak{F}_{s,L} d\delta}{2^{d-1}} \times \frac{|\log h_{\ell_0}|}{nh^d_{\ell_0}}} + \frac{C^\star(s,d)d\delta}{2^d} \times \frac{|\log h_{\ell_0}|}{nh^d_{\ell_0}} \\ &\quad + \sqrt{\frac{2^{d+2} D^\star(s,d)\mathfrak{F}_{s,L}}{\min\left\{1, \frac{\lambda^2_\bullet(s,d)}{|\log\lambda_\bullet(s,d)|}\right\}} \times \frac{1}{nh^d_{\ell_0}}} + \frac{2\sqrt{D^\star(s,d)}}{\min\left\{1, \frac{\lambda_\bullet(s,d)}{|\log\lambda_\bullet(s,d)|}\right\}} \times \frac{1}{nh^d_{\ell_0}}.\end{aligned}$$

As $n \to \infty$, note that $h_{\ell_0} \to 0$, $nh^d_{\ell_0} \to \infty$ and $|\log h_{\ell_0}|/(nh^d_{\ell_0}) \to 0$ by (B.16). Therefore, the first term on the right-hand side of the last equation leads the asymptotic behavior of the upper bound on $\mathbb{U}^\star_{\gamma_0}$. In particular, for n large enough, the second, third and fourth terms are all bounded from above by the first term, so that

$$\mathbb{U}^\star_{\gamma_0} \leq 4\sqrt{\frac{\{C^\star(s,d)\}^2 \mathfrak{F}_{s,L} d\delta}{2^{d-1}} \times \frac{|\log h_{\ell_0}|}{nh^d_{\ell_0}}} = \frac{C^\star(s,d)\sqrt{\mathfrak{F}_{s,L} d\delta}}{2^{(d-5)/2}} \times \left(\frac{|\log h_{\ell_0}|}{nh^d_{\ell_0}}\right)^{1/2}.$$

After straightforward calculations, this yields

$$\mathbb{U}^\star_{\gamma_0} \lesssim C^\star(s,d)\sqrt{d\delta}\ \mathfrak{F}^{s/(2s+d)}_{s,L} (\mathfrak{Q}^\star_{s,L})^{d/(2s+d)} \left(\frac{\log n}{n}\right)^{s/(2s+d)}, \tag{B.23}$$

where $\lesssim$ is independent of all parameters.

By taking the upper bounds on $\mathcal{B}_{\gamma_0}$ and $\mathbb{U}^\star_{\gamma_0}$ in (B.21) and (B.23) back into (B.15), we get

$$R_n(\hat{f}, f) \lesssim \mathfrak{c}(\delta, \|f\|_{\mathcal{V}(\rho)})\, e^s\, C^\star(s,d)\sqrt{d\delta}\ \mathfrak{F}^{s/(2s+d)}_{s,L} (\mathfrak{Q}^\star_{s,L})^{d/(2s+d)} \left(\frac{\log n}{n}\right)^{s/(2s+d)},$$

where $\mathfrak{c}(\delta, \|f\|_{\mathcal{V}(\rho)}) = 8 + 2\{(\delta-1)/\|f\|_{\mathcal{V}(\rho)}\}^{1/2}$ and $\lesssim$ is again independent of all parameters. This proves the upper bound.

Next, let us prove that the admissible collection ϕ defined in (4.11) is the ARC. To do so, we will use the result established in Section 5.6.1 of Rebelles (2015). In particular, it is sufficient to prove that for any $(s,L) \in \mathcal{K}$ and $(s',L') \in \mathcal{K}$ such that $s < s'$, there are two functions f_0 and f_1 that satisfy the following properties:

1. f_0 and f_1 are density functions that belong to $\Sigma(s',L')$ and $\Sigma(s,L)$, respectively.
2. $|f_1(t) - f_0(t)| \asymp \phi_n(s,L) = (n^{-1}\log n)^{s/(2s+d)}$. (Here, $\asymp$ means both $\lesssim$ and $\gtrsim$ hold.)
3. For any positive real τ such that $s/(2s+d) < \tau < s'/(2s'+d)$, we have

$$\mathsf{E}_{f_0,n}\left\{\prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)}\right\}^2 \lesssim n^\tau \phi_n(s,L).$$

Similarly to the proof of Proposition 4.4, let us define

$$f_0(x) = \frac{\mathbb{1}_{\mathcal{D}}(x)}{\mathrm{Leb}(\mathcal{D})} \quad \text{and} \quad f_1(x) = f_0(x) + Ah^s \tilde{\psi}_h(x-t)\mathbb{1}_{\mathcal{D}}(x),$$

where the quantities $\tilde{\psi}_h$ and A have already been introduced in the proof of Proposition 4.4, and h is defined in a slightly different way by

$$h = \left[\frac{1}{\frac{8A^2}{\mathrm{Leb}(\mathcal{D})} \left\{ \int_\Delta \psi^2(u) \mathrm{d}u + 1 \right\}} \times \frac{\log n^\varrho}{n} \right]^{1/(2s+d)}, \quad 0 < \varrho < \frac{s'}{2s'+d} - \frac{s}{2s+d}.$$

It is straightforward to verify the first and second points by employing arguments that are virtually identical to those presented in the proof of Proposition 4.4. For the third point, we have, by the same line of reasoning that led us to (B.14),

$$\begin{aligned} \mathsf{E}_{f_0,n} \left\{ \prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)} \right\}^2 &\leq \exp \left[\frac{8A^2}{\mathrm{Leb}(\mathcal{D})} \left\{ \int_\Delta \psi^2(u) \mathrm{d}u + 1 \right\} n h^{2s+d} \right] \\ &\leq n^\varrho \leq n^{\varrho + s/(2s+d)} \left(\frac{\log n}{n} \right)^{s/(2s+d)} \leq n^\tau \phi_n(s, L), \end{aligned} \tag{B.24}$$

where $\tau = \varrho + s/(2s+d)$. This shows that three conditions above, or equivalently the conditions of Lemma 4 of Rebelles (2015), are satisfied with $a_n^{-1} \asymp \phi_n(s, L)$ and $b_n = n^\tau$, which implies that the collection ϕ defined in (4.11) is the ARC. □

Proof of Proposition 4.6. Let $\rho = 1$ so that $h \in (0, 1]$. Also, throughout the proof, let $\gamma_k = (m, h)$ with $m = \lfloor\!\lfloor s_k \rfloor\!\rfloor$. Recall the definition of w_h from (3.2) and apply it to the domain $\mathcal{D}_k$ in (4.12) to obtain

$$W_h = \int_{\mathbb{R}^d} w_h(u) \mathrm{d}u = h^{-2} \int_0^h \int_0^{x^k} 1 \mathrm{d}y \mathrm{d}x = \frac{h^{k-1}}{k+1}. \tag{B.25}$$

Recall the definition of the Gram matrix $\mathcal{B}_{\gamma_k}$ from (3.3). By the simple change of variables $(\xi, \eta) = (x/h, y/h)$ compared with the previous integral, we have, for any $v \in \mathbb{R}^{D_m}$,

$$\begin{aligned} v^\top \mathcal{B}_{\gamma_k} v &= \int_0^1 \int_0^{h^{k-1}\xi^k} \left\{ v^\top \Phi_{m,1}(\xi, \eta) \right\}^2 \mathrm{d}\eta \mathrm{d}\xi \\ &= \sum_{|\alpha| \leq m} \sum_{|\beta| \leq m} v_\alpha v_\beta \int_0^1 \int_0^{h^{k-1}\xi^k} \xi^{\alpha_1+\beta_1} \eta^{\alpha_2+\beta_2} \mathrm{d}\eta \mathrm{d}\xi \\ &= \sum_{|\alpha| \leq m} \sum_{|\beta| \leq m} v_\alpha v_\beta \int_0^1 \xi^{\alpha_1+\beta_1} \frac{(h^{k-1}\xi^k)^{\alpha_2+\beta_2+1}}{\alpha_2+\beta_2+1} \mathrm{d}\xi \\ &= \sum_{|\alpha| \leq m} \sum_{|\beta| \leq m} c_k(\alpha, \beta) v_\alpha v_\beta h^{(k-1)(\alpha_2+\beta_2+1)}, \end{aligned} \tag{B.26}$$

where

$$c_k(\alpha, \beta) = [(\alpha_2+\beta_2+1)\{\alpha_1+\beta_1+k(\alpha_2+\beta_2+1)+1\}]^{-1} > 0.$$

Denote $A_{\alpha,\beta}(v) = c_k(\alpha,\beta)v_\alpha v_\beta$ and $\alpha^\star = \beta^\star = (0,m)$, where the dependence in k is omitted. By (B.26), we can write, for all $v \in \mathbb{R}^{D_m} \setminus \{(0,\dots,0)\}$,

$$v^\top \mathcal{B}_{\gamma_k} v = \sum_{|\alpha|\le m}\sum_{|\beta|\le m} A_{\alpha,\beta}(v) h^{(k-1)(\alpha_2+\beta_2+1)} = h^{(k-1)(2m+1)}\psi_v(1/h),$$

where ψ_v is a polynomial of degree $(k-1)(2m)$ which satisfies $\psi_v(0) = A_{\alpha^\star,\beta^\star}(v) > 0$. In particular, for any $v \in \mathbb{R}^{D_m}$ such that $v^\top v = 1$, we have $v^\top \mathcal{B}_{\gamma_k} v > 0$ and thus $\psi_v(x) > 0$ for any $x \in [0,\infty)$. Therefore, $\Psi_\star(v) = \min_{x\in[0,\infty)} \psi_v(x) > 0$ exists and is a continuous function of the coefficients of ψ_v, i.e., is a continuous function of v. Since $\mathbb{S}^1 = \{v \in \mathbb{R}^{D_m} : v^\top v = 1\}$ is a compact set, $\Psi_\star = \min_{v\in\mathbb{S}^1} \Psi_\star(v) > 0$ exists. Finally, we deduce

$$\lambda_{\gamma_k} = \min_{v\in\mathbb{S}^1} v^\top \mathcal{B}_{\gamma_k} v \ge \Psi_\star h^{(k-1)(2m+1)} > 0. \tag{B.27}$$

Since we assume $f \in \Sigma(s,L)$, a straightforward consequence of Definition 2 is that for all $\beta \in (0,s]$, there exists $\tilde{L}_{s,L,\beta} > 0$ such that $f \in \Sigma(\beta, \tilde{L}_{s,L,\beta})$. Indeed, if $f \in \Sigma(s,L)$, then there exists a polynomial $q = \sum_{|\alpha|\le \lfloor\!\lfloor s \rfloor\!\rfloor} q_\alpha \varphi_\alpha(\cdot) \in \mathcal{P}_{\lfloor\!\lfloor s \rfloor\!\rfloor}$ such that, for any $u \in \mathcal{V}(1)$,

$$\left| f(t+u) - \sum_{|\alpha|\le \lfloor\!\lfloor s \rfloor\!\rfloor} q_\alpha \varphi_\alpha(u) \right| \le L\|u\|_\infty^s.$$

This implies that $f \in \Sigma(\beta, \tilde{L}_{s,L,\beta})$ since

$$\begin{aligned}\left| f(t+u) - \sum_{|\alpha|\le \lfloor\!\lfloor \beta \rfloor\!\rfloor} q_\alpha \varphi_\alpha(u) \right| &\le L\|u\|_\infty^s + \sum_{\lfloor\!\lfloor \beta \rfloor\!\rfloor < |\alpha| \le \lfloor\!\lfloor s \rfloor\!\rfloor} |q_\alpha||\varphi_\alpha(u)| \\ &\le \left(L + \sum_{\lfloor\!\lfloor \beta \rfloor\!\rfloor < |\alpha| \le \lfloor\!\lfloor s \rfloor\!\rfloor} |q_\alpha| \right) \|u\|_\infty^\beta =: \tilde{L}_{s,L,\beta}\|u\|_\infty^\beta,\end{aligned}$$

where the last inequality follows from the fact that, for all $x \in [0,1]$, the map $\beta \mapsto x^\beta$ is decreasing on $(0,\infty)$ (notice that $u \in \mathcal{V}(1) \subseteq [-1,1]^d$ implies $x = \|u\|_\infty \in [0,1]$). In particular, for our choice of smoothness $\beta = s_k$ in (4.13), we have $f \in \Sigma(s_k, L_k)$, where we write $L_k = \tilde{L}_{s,L,s_k}$ for short. Therefore, by Proposition 4.1, Proposition 4.2 and Lemma A.3, together with the fact that $W_h = h^{k-1}/(k+1)$ from (B.25) and $\lambda_{\gamma_k} \ge \Psi_\star h^{(k-1)(2m+1)} > 0$ from (B.27), we have

$$\begin{aligned} R_n(\hat{f}_{\gamma_k}, f) &\le \frac{W_h\sqrt{D_{\lfloor\!\lfloor s_k \rfloor\!\rfloor}}}{\lambda_{\gamma_k}} \times \left\{ \mathfrak{L}^2_{\lfloor\!\lfloor s_k \rfloor\!\rfloor, s_k, L_k} h^{2s_k} + \frac{\mathfrak{F}_{s_k,L_k}}{W_h n h^2} \right\}^{1/2} \\ &\lesssim_{s,L,k} h^{-2\lfloor\!\lfloor s_k \rfloor\!\rfloor(k-1)} \times \left\{ \mathfrak{L}^2_{\lfloor\!\lfloor s_k \rfloor\!\rfloor, s_k, L_k} h^{2s_k} + \frac{(k+1)\mathfrak{F}_{s_k,L_k}}{n h^{k+1}} \right\}^{1/2}. \end{aligned}$$

By taking $h = n^{-1/(2s_k+k+1)}$, the upper bound (4.14) follows. To be more transparent, our choice of smoothness $s_k \in (0,s]$ in (4.13) is the one that minimizes the bound in the last equation if we were to replace s_k by a general smoothness parameter $\beta \in (0,s]$.

To prove the lower bound, we follow the scheme of proofs laid out in Sections 2.2 and 2.3 of Tsybakov (2009). Let $(s, L) \in (0, 1] \times (0, \infty)$ and $k > 1$ be given. Define the auxiliary function

$$\tilde{\psi}(x) = \exp\left(-\frac{1}{1-x^2}\right) \mathbb{1}_{(-1,1)}(x), \quad x \in \mathbb{R},$$

together with the positive real constant

$$b_k = \frac{\int_0^1 x^k(1-x)\mathrm{d}x}{\int_0^1 \tilde{\psi}(4x-3)x^k(1-x)\mathrm{d}x}.$$

Now, consider the function

$$\psi_k(x) = \{1 - b_k\tilde{\psi}(4x-3)\}(1-x)\mathbb{1}_{[0,1]}(x).$$

This is a bounded function which satisfies $\psi_k(0) = 1$, $\psi_k(1) = 0$ and $\int_0^1 x^k\psi_k(x)\mathrm{d}x = 0$. Moreover, ψ_k is s-Hölder-continuous on $[0, 1]$ for some positive Lipschitz constant $L_{k,s} > 0$. We now consider, for $u = (u_x, u_y) \in \mathcal{D}_k$,

$$f_0(u) = \frac{\mathbb{1}_{\mathcal{D}_k}(u)}{\mathrm{Leb}(\mathcal{D}_k)} \quad \text{and} \quad f_1(u) = f_0(u) + 2\rho_n\psi_k\left(\frac{u_x}{h_n}\right)\mathbb{1}_{\mathcal{V}(h_n)}(u), \tag{B.28}$$

where

$$h_n = n^{-1/(2s+k+1)} \quad \text{and} \quad \rho_n = \{L/(2L_{k,s})\}h_n^s. \tag{B.29}$$

Observe that both f_0 and f_1 are density functions on $\mathcal{D}_k$, for n large enough. This is obvious for the function f_0. Let us prove this assertion for f_1. First, note that for $u \in \mathcal{D}_k$,

$$f_1(u) \geq \frac{1}{\mathrm{Leb}(\mathcal{D}_k)} - 2\rho_n\|\psi_k\|_\infty,$$

which is positive for large n. Using the change of variable $\xi = x/h_n$, it can also be seen that

$$\begin{aligned}\int_{\mathcal{D}_k} f_1(u)\mathrm{d}u &= 1 + 2\rho_n\int_0^{h_n}\psi_k\left(\frac{x}{h_n}\right)\left(\int_0^{x^k} 1\mathrm{d}y\right)\mathrm{d}x \\ &= 1 + 2h_n^{k+1}\rho_n\int_0^1 \xi^k\psi_k(\xi)\mathrm{d}\xi \\ &= 1.\end{aligned}$$

Using the methodology developed in Section 2.5 of Tsybakov (2009), the proof of the lower bound (4.15) boils down to proving the following assertions:

1. $\{f_0, f_1\} \subseteq \Sigma(s, L)$.
2. $|f_1(0) - f_0(0)| \geq 2\rho_n$.
3. $nD_{\mathrm{KL}}(f_1, f_0) \leq A$ for some positive real constant $A = A(k, s, L) > 0$, where $D_{\mathrm{KL}}(\cdot, \cdot)$ denotes the Kullback-Leibler divergence between the associated probability measures.

To prove the first assertion, notice that the function

$$u = (u_x, u_y) \in \mathcal{D}_k \mapsto 2\rho_n\psi_k\left(\frac{u_x}{h_n}\right)\mathbb{1}_{\mathcal{V}(h_n)}(u)$$

belongs to $\Sigma(s, L)$, using the definitions of h_n and ρ_n in (B.29) and the aforementioned fact that ψ_k is s-Hölder-continuous on $[0,1]$ for some Lipschitz constant $L_{k,s} > 0$. The second assertion is satisfied because of (B.28) and the fact that $\psi_k(0) = 1$. It remains to prove the third assertion. Since $f_1(u)/f_0(u) = 1$ for all $u \in \mathcal{D}_k \setminus \mathcal{V}(h_n)$, observe that

$$nD_{\mathrm{KL}}(f_1, f_0) = n \int_{\mathcal{V}(h_n)} f_1(u) \log\left\{\frac{f_1(u)}{f_0(u)}\right\} \mathrm{d}u = B_1 + B_2, \tag{B.30}$$

where

$$B_1 = \frac{n}{\mathrm{Leb}(\mathcal{D}_k)} \int_{\mathcal{V}(h_n)} \log\left\{\frac{f_1(u)}{f_0(u)}\right\} \mathrm{d}u,$$

$$B_2 = 2n\rho_n \int_{\mathcal{V}(h_n)} \psi_k\left(\frac{u_x}{h_n}\right) \log\left\{\frac{f_1(u)}{f_0(u)}\right\} \mathrm{d}u.$$

Since ψ_k is bounded and ρ_n tends to zero as $n \longrightarrow \infty$, let n be large enough so that

$$\left|2\rho_n \mathrm{Leb}(\mathcal{D}_k)\psi_k\left(\frac{u_x}{h_n}\right)\right| \leq \frac{1}{2}.$$

Now, using

$$\int_{\mathcal{V}(h_n)} \psi_k\left(\frac{u_x}{h_n}\right) \mathrm{d}u = 0 \qquad \text{and} \qquad |\log(1+x) - x| \leq 2x^2 \quad \text{for } |x| \leq 1/2,$$

we have

$$\begin{aligned} |B_1| &\leq \frac{n}{\mathrm{Leb}(\mathcal{D}_k)} \int_{\mathcal{V}(h_n)} \left|\log\left\{\frac{f_1(u)}{f_0(u)}\right\} - 2\rho_n \mathrm{Leb}(\mathcal{D}_k)\psi_k\left(\frac{u_x}{h_n}\right)\right| \mathrm{d}u \\ &\leq 8n\rho_n^2 \mathrm{Leb}(\mathcal{D}_k) \int_{\mathcal{V}(h_n)} \psi_k^2\left(\frac{u_x}{h_n}\right) \mathrm{d}u \\ &\leq 8n\rho_n^2 h_n^{k+1} \mathrm{Leb}(\mathcal{D}_k) \int_0^1 \xi^k \psi_k^2(\xi) \mathrm{d}\xi. \end{aligned} \tag{B.31}$$

Similarly, but instead using

$$|\log(1+x)| \leq |x| \quad \text{for } x > 0,$$

we have

$$\begin{aligned} |B_2| &\leq 2n\rho_n \int_{\mathcal{V}(h_n)} \psi_k\left(\frac{u_x}{h_n}\right) \left|\log\left\{\frac{f_1(u)}{f_0(u)}\right\}\right| \mathrm{d}u \\ &\leq 4n\rho_n^2 \mathrm{Leb}(\mathcal{D}_k) \int_{\mathcal{V}(h_n)} \psi_k^2\left(\frac{u_x}{h_n}\right) \mathrm{d}u \\ &\leq 4n\rho_n^2 h_n^{k+1} \mathrm{Leb}(\mathcal{D}_k) \int_0^1 \xi^k \psi_k^2(\xi) \mathrm{d}\xi. \end{aligned} \tag{B.32}$$

Taking the above bounds (B.31) and (B.32) together in (B.30), we obtain

$$nD_{\mathrm{KL}}(f_1, f_0) \leq 12 \mathrm{Leb}(\mathcal{D}_k) \left\{\int_0^1 \xi^k \psi_k^2(\xi) \mathrm{d}\xi\right\} n\rho_n^2 h_n^{k+1}.$$

Using the definition of ρ_n and h_n in (B.29), the right-hand side of the last equation is easily seen to be bounded, which proves the third point. Moreover, we have $\rho_n \asymp n^{-s/(2s+k+1)} = n^{-\theta_k(s)}$ since $s_k = s$ for $s \in (0,1]$; see Item 1 of Remark 9. The lower bound (4.15) follows. □

Proof of Theorem 4.7. Let $(s, L) \in (0,1] \times (0,\infty)$ and $k > 1$ be given. Using (B.27) with the polynomial degree set to $m = 0$, there exists a positive real constant $\Psi_\star = \Psi_\star(k)$ which depends on k such that, for any $\gamma = (0,h) \in \Gamma_n$,

$$\lambda_\gamma \geq \Psi_\star h^{k-1} > 0. \tag{B.33}$$

Hence, condition (4.4) is satisfied and Theorem 4.3 can be applied to obtain

$$R_n(\hat{f}, f) \leq \min_{\gamma \in \Gamma_n} \left[5\mathbb{B}_\gamma + \left\{ 3 + 2\sqrt{(\delta-1)/\|f\|_{\mathcal{V}(\rho)}} \right\} \mathbb{U}_\gamma^\star \right] + \mathcal{O}\left\{ \sqrt{\frac{(\log n)^d}{n}} \right\}. \tag{B.34}$$

Now, similarly to the proof of Theorem 4.5, we define

$$\ell_0 = \left\lfloor \frac{1}{2s+k+1} \log\left(\frac{n}{\log n}\right) + \log \rho \right\rfloor \quad \text{and} \quad m_{\ell_0} = 0.$$

In the following, consider n large enough so that $\ell_0 \geq 1$. This choice of ℓ_0 implies

$$1 \leq h_{\ell_0} \left(\frac{\log n}{n}\right)^{-1/(2s+k+1)} \leq e.$$

By (B.25), we know that

$$\frac{W_h}{\lambda_\gamma} = \frac{h^{k-1}/(k+1)}{\lambda_\gamma} \leq \frac{1}{(k+1)\Psi_\star}. \tag{B.35}$$

Therefore, with the notation $\gamma_0 = (0, h_{\ell_0})$, Proposition 4.1 yields

$$\begin{aligned} \mathbb{B}_{\gamma_0} = \max_{\substack{\gamma \in \Gamma_n \\ \gamma \preceq \gamma_0}} \left| \mathsf{E}\left\{\hat{f}_\gamma(t)\right\} - f(t) \right| &\leq \max_{\substack{\gamma \in \Gamma_n \\ \gamma \preceq \gamma_0}} \frac{W_h \sqrt{D_0}}{\lambda_\gamma} \times \mathfrak{L}_{0,s,L}\, h^s \\ &\leq C_0(s,L,k) \max_{h \leq h_{\ell_0}} h^s \leq C_0(s,L,k)\, e^s \left(\frac{\log n}{n}\right)^{s/(2s+k+1)}, \end{aligned} \tag{B.36}$$

where $C_0(s,L,k) > 0$ is a positive real constant that depends on s, L and k. Following the same line of argument as in the proof of Theorem 4.5, we obtain, using (B.33), (B.35) and Lemma A.3,

$$\begin{aligned} \mathbb{U}_{\gamma_0}^\star &\leq C_1(s,L,k) \left(\sqrt{\frac{|\log h_{\ell_0}|}{n h_{\ell_0}^{k+1}}} + \frac{|\log h_{\ell_0}|}{n h_{\ell_0}^{k+1}} + \sqrt{\frac{1}{n h_{\ell_0}^{k+1}}} + \frac{1}{n h_{\ell_0}^{k+1}} \right) \\ &\leq C_2(s,L,k) \left(\frac{\log n}{n}\right)^{s/(2s+k+1)}, \end{aligned} \tag{B.37}$$

where $C_1(s,L,k)$ and $C_2(s,L,k)$ are positive real constants that depend on s, L and k. By plugging the bounds (B.36) and (B.37) back into (B.34), we get the upper bound (4.16).

To prove that the collection $\phi = \{\phi_n(s, L) : (s, L) \in (0, 1] \times (0, \infty), n \in \mathbb{N}\}$ is the ARC, we follow the same line of argument as in the proof of the adaptive lower bound in Theorem 4.5, but the density functions $f_0 \in \Sigma(s', L')$ and $f_1 \in \Sigma(s, L)$ are replaced by those defined in (B.28) together with the bandwidth

$$h = \left\{ \frac{1}{\{L/(2L_{k,s})\}^2 \mathrm{Leb}(\mathcal{D}_k) \int_0^1 \xi^k \psi_k^2(\xi) \mathrm{d}\xi} \times \frac{\log n^{\varrho}}{n} \right\}^{1/(2s+k+1)},$$

where $0 < \varrho < s'/(2s' + k + 1) - s/(2s + k + 1)$. To understand the choice of h here, simply look at (B.12) with $A = L/(2L_{k,s})$ and use the inequality $(1 + x) \leq e^x$, which is valid for $x > 0$, to obtain a bound analogous to (B.24). This concludes the proof. □

Acknowledgments

The authors express their sincere gratitude to the anonymous reviewers for their constructive comments. The quality and depth of the questions posed by the referees have significantly enhanced the caliber of this paper. It has been possible to conduct this research in part thanks to the support provided by Calcul Québec and the Digital Research Alliance of Canada.

Funding

K. Bertin is supported by FONDECYT regular grants 1221373 and 1230807 from ANID-Chile, and the Centro de Modelamiento Matemático (CMM) BASAL fund FB210005 for centers of excellence from ANID-Chile. N. Klutchnikoff acknowledges support from FONDECYT regular grant 1221373 from ANID-Chile. F. Ouimet was supported by a CRM-Simons postdoctoral fellowship from the Centre de recherches mathématiques (Montréal, Canada) and the Simons Foundation. F. Ouimet is currently funded through a contribution to Christian Genest's research program from the Trottier Institute for Science and Public Policy.

References

AITCHISON, J. and LAUDER, I. J. (1985). Kernel density estimation for compositional data. *J. Roy. Statist. Soc. Ser. C* **34** 129–137. https://doi.org/10.2307/2347365

AMMOUS, S., DEDECKER, J. and DUVAL, C. (2023). Adaptive directional estimator of the density in $\mathbb{R}^d$ for independent and mixing sequences. *Preprint* 1–28. arXiv:2305.13572v1.

ARNONE, E., FERRACCIOLI, F., PIGOLOTTI, C. and SANGALLI, L. M. (2022). A roughness penalty approach to estimate densities over two-dimensional manifolds. *Comput. Statist. Data Anal.* **174** Paper No. 107527, 14 pp. https://doi.org/10.1016/j.csda.2022.107527 MR4437662

AZZIMONTI, L., NOBILE, F., SANGALLI, L. M. and SECCHI, P. (2014). Mixed finite elements for spatial regression with PDE penalization. *SIAM/ASA J. Uncertain. Quantif.* **2** 305–335. https://doi.org/10.1137/130925426 MR3283911

AZZIMONTI, L., SANGALLI, L. M., SECCHI, P., DOMANIN, M. and NOBILE, F. (2015). Blood flow velocity field estimation via spatial regression with PDE penalization. *J. Amer. Statist. Assoc.* **110** 1057–1071. https://doi.org/10.1080/01621459.2014.946036 MR3420684

BABU, G. J. and CHAUBEY, Y. P. (2006). Smooth estimation of a distribution and density function on a hypercube using Bernstein polynomials for dependent random vectors. *Statist. Probab. Lett.* **76** 959–969. https://doi.org/10.1016/j.spl.2005.10.031 MR2270097

BAKKA, H., VANHATALO, J., ILLIAN, J. B., SIMPSON, D. and RUE, H. (2019). Non-stationary Gaussian models with physical barriers. *Spat. Stat.* **29** 268–288. https://doi.org/10.1016/j.spasta.2019.01.002 MR3903698

BARRY, R. (2021). latticeDensity: Density Estimation and Nonparametric Regression on Irregular Regions. R package version 1.2.6, available online at https://cran.r-project.org/src/contrib/Archive/latticeDensity [This link was accessed on 2024–05–15.].

BARRY, R. P. and MCINTYRE, J. (2011). Estimating animal densities and home range in regions with irregular boundaries and holes: A lattice-based alternative to the kernel density estimator. *Ecol. Model.* **222** 1666–1672. https://doi.org/10.1016/j.ecolmodel.2011.02.016

BARRY, R. P. and MCINTYRE, J. (2020). Lattice-based methods for regression and density estimation on complicated multidimensional regions. *Environ. Ecol. Stat.* **27** 571–589. https://doi.org/10.1007/s10651-020-00459-z

BERTIN, K., EL KOLEI, S. and KLUTCHNIKOFF, N. (2019). Adaptive density estimation on bounded domains. *Ann. Inst. Henri Poincaré Probab. Stat.* **55** 1916–1947. https://doi.org/10.1214/18-AIHP938 MR4029144

BERTIN, K. and KLUTCHNIKOFF, N. (2014). Adaptive estimation of a density function using beta kernels. *ESAIM Probab. Stat.* **18** 400–417. https://doi.org/10.1051/ps/2014010 MR3333996

BERTIN, K. and KLUTCHNIKOFF, N. (2017). Pointwise adaptive estimation of the marginal density of a weakly dependent process. *J. Statist. Plann. Inference* **187** 115–129. https://doi.org/10.1016/j.jspi.2017.03.003 MR3638047

BERTIN, K., KLUTCHNIKOFF, N. and OUIMET, F. (2024). densityLocPoly: Density Estimation on Complicated Domains; Implementation of the local polynomial density estimator in 2d. R package version 0.0.2, available online at https://github.com/klutchnikoff/densityLocPoly [This link was accessed on 2024–05–15.].

BERTIN, K., KLUTCHNIKOFF, N., LÉON, J. R. and PRIEUR, C. (2020). Adaptive density estimation on bounded domains under mixing conditions. *Electron. J. Stat.* **14** 2198–2237. https://doi.org/10.1214/20-EJS1682 MR4097810

BOTEV, Z. I., GROTOWSKI, J. F. and KROESE, D. P. (2010). Kernel density estimation via diffusion. *Ann. Statist.* **38** 2916–2957. https://doi.org/10.1214/10-AOS799 MR2722460

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199535255.001.0001 MR3185193

BRUNEL, V. E. (2018). Methods for estimation of convex sets. *Statist. Sci.* **33** 615–632. https://doi.org/10.1214/18-STS669 MR3881211

CATTANEO, M. D., JANSSON, M. and MA, X. (2020). Simple local polynomial density estimators. *J. Amer. Statist. Assoc.* **115** 1449–1455. https://doi.org/10.1080/01621459.2019.1635480 MR4143477

CATTANEO, M. D., JANSSON, M. and MA, X. (2022). lpdensity: Local Polynomial Density Estimation and Inference. *J. Stat. Softw.* **101** 1–25. https://doi.org/10.18637/jss.v101.i02

CATTANEO, M. D., CHANDAK, R., JANSSON, M. and MA, X. (2023). Boundary adaptive local polynomial conditional density estimators. *arXiv preprint* 69 pp. arXiv:2204.10359v4.

CHEN, S. X. (1999). Beta kernel estimators for density functions. *Comput. Statist. Data Anal.* **31** 131–145. https://doi.org/10.1016/S0167-9473(99)00010-9 MR1718494

CHENG, M. Y., FAN, J. and MARRON, J. S. (1997). On automatic boundary corrections. *Ann. Statist.* **25** 1691–1708. https://doi.org/10.1214/aos/1031594737 MR1463570

CLINE, D. B. H. and HART, J. D. (1991). Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics* **22** 69–84. https://doi.org/10.1080/02331889108802286 MR1097362

DAVIES, T. M., MARSHALL, J. C. and HAZELTON, M. L. (2018). Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk. *Stat. Med.* **37** 1191–1221. https://doi.org/10.1002/sim.7577 MR3777968

DAVIES, T. M. and MARSHALL, J. C. (2023). sparr: Spatial and Spatiotemporal Relative Risk. R package version 2.3-10, available online at https://cran.r-project.org/web/packages/sparr/index.html [This link was accessed on 2024–05–15.].

FAN, J. and GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20** 2008–2036. https://doi.org/10.1214/aos/1176348900 MR1193323

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. https://doi.org/10.1007/978-1-4899-3150-4 MR1383587

FENG, O. Y., GUNTUBOYINA, A., KIM, A. K. H. and SAMWORTH, R. J. (2021). Adaptation in multivariate log-concave density estimation. *Ann. Statist.* **49** 129–153. https://doi.org/10.1214/20-AOS1950 MR4206672

FERRACCIOLI, F., ARNONE, E., FINOS, L., RAMSAY, J. O. and SANGALLI, L. M. (2021). Nonparametric density estimation over complicated domains. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 346–368. https://doi.org/10.1111/rssb.12415 MR4250279

GASSER, T. and MÜLLER, H. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation (Proc. Workshop, Heidelberg, 1979). Lecture Notes in Math.* **757** 23–68. Springer, Berlin. https://doi.org/10.1007/BFb0098489 MR564251

GASSER, T., MÜLLER, H. G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252. https://doi.org/10.1111/j.2517-6161.1985.tb01350.x MR816088

GAWRONSKI, W. and STADTMÜLLER, U. (1981). Smoothing histograms by means of lattice and continuous distributions. *Metrika* **28** 155–164. https://doi.org/10.1007/BF01902889 MR638651

GHOSAL, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.* **29** 1264–1280. https://doi.org/10.1214/aos/1013203453 MR1873330

GOLDENSHLUGER, A. and LEPSKI, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150–1190. https://doi.org/10.3150/08-BEJ144 MR2543590

GOLDENSHLUGER, A. and LEPSKI, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields* **143** 41–71. https://doi.org/10.1007/s00440-007-0119-5 MR2449122

GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632. https://doi.org/10.1214/11-AOS883 MR2850214

GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on R^d. *Probab. Theory Related Fields* **159** 479–543. https://doi.org/10.1007/s00440-013-0512-1 MR3230001

GU, C. (1993). Smoothing spline density estimation: a dimensionless automatic algorithm. *J. Amer. Statist. Assoc.* **88** 495–504. https://doi.org/10.1080/01621459.1993.10476300 MR1224374

GU, C. and QIU, C. (1993). Smoothing spline density estimation: theory. *Ann. Statist.* **21** 217–234. https://doi.org/10.1214/aos/1176349023 MR1212174

GUILLAS, S. and LAI, M. J. (2010). Bivariate splines for spatial functional regression models. *J. Nonparametr. Stat.* **22** 477–497. https://doi.org/10.1080/10485250903323180 MR2662608

JONES, M. C. (1993). Simple boundary correction for kernel density estimation. *Stat Comput* . **3** 135–146. https://doi.org/10.1007/BF00147776

JONES, M. C. and FOSTER, P. J. (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statist. Sinica* **6** 1005–1013. https://www.jstor.org/stable/24306056. MR1422417

KLEMELÄ, J. (2009). Multivariate histograms with data-dependent partitions. *Statist. Sinica* **19** 159–176. https://www.jstor.org/stable/24308713. MR2487883

KLUTCHNIKOFF, N. (2014). Pointwise adaptive estimation of a multivariate function. *Math. Methods Statist.* **23** 132–150. https://doi.org/10.3103/S1066530714020045 MR3224636

LAI, M. J. and SCHUMAKER, L. L. (2007). *Spline Functions on Triangulations*. *Encyclopedia of Mathematics and its Applications* **110**. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511721588 MR2355272

LAI, M. J. and WANG, L. (2013). Bivariate penalized splines for regression. *Statist. Sinica* **23** 1399–1417. https://doi.org/10.5705/ss.2010.278 MR3114719

LEPSKI, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *Ann. Statist.* **43** 1178–1242. https://doi.org/10.1214/14-AOS1306 MR3346701

LEPSKII, O. V. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466. https://doi.org/10.1137/1135065

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x MR2853727

LIU, Y. and WU, C. (2019). Point-wise estimation for anisotropic densities. *J. Multivariate Anal.* **171** 112–125. MR3892029

MARRON, J. S. and RUPPERT, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **56** 653–671. https://doi.org/10.1111/j.2517-6161.1994.tb02006.x MR1293239

MCINTYRE, J. and BARRY, R. P. (2018). A lattice-based smoother for regions with irregular boundaries and holes. *J. Comput. Graph. Statist.* **27** 360–367. https://doi.org/10.1080/10618600.2017.1375935 MR3816271

McSwiggan, G., Baddeley, A. and Nair, G. (2017). Kernel density estimation on a linear network. *Scand. J. Stat.* **44** 324–345. https://doi.org/10.1111/sjos.12255 MR3658517

Miller, D. L. and Wood, S. N. (2014). Finite area smoothing with generalized distance splines. *Environ. Ecol. Stat.* **21** 715–731. https://doi.org/10.1007/s10651-014-0277-4 MR3279587

Müller, H. G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78** 521–530. https://doi.org/10.2307/2337021 MR1130920

Niu, M., Cheung, P., Lin, L., Dai, Z., Lawrence, N. and Dunson, D. (2019). Intrinsic Gaussian processes on complex constrained domains. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 603–627. https://doi.org/10.1111/rssb.12320 MR3961500

Petrone, S. (1999). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.* **27** 105–126. https://doi.org/10.2307/3315494 MR1703623

Petrone, S. and Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *J. Roy. Statist. Soc. Ser. B* **64** 79–100. https://doi.org/10.1111/1467-9868.00326 MR1881846

Ramsay, T. (2002). Spline smoothing over difficult regions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 307–319. https://doi.org/10.1111/1467-9868.00339 MR1904707

Rebelles, G. (2015). Pointwise adaptive estimation of a multivariate density under independence hypothesis. *Bernoulli* **21** 1984–2023. https://doi.org/10.3150/14-BEJ633 MR3378457

Robinson, D. (2015). View package downloads over time with Shiny. Blog post, available online at http://varianceexplained.org/r/cran-view [This link was accessed on 2024–05–15.].

Ruppert, D. and Cline, D. B. H. (1994). Bias reduction in kernel density estimation by smoothed empirical transformations. *Ann. Statist.* **22** 185–210. https://doi.org/10.1214/aos/1176325365 MR1272080

Sangalli, L. M. (2021). Spatial regression with partial differential equation regularisation. *Int. Stat. Rev.* **89** 505–531. https://doi.org/10.1111/insr.12444 MR4411916

Sangalli, L. M., Ramsay, J. O. and Ramsay, T. O. (2013). Spatial spline regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 681–703. https://doi.org/10.1111/rssb.12009 MR3091654

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Commun. Statist. Theor. Meth.* **14** 1123–1136. https://doi.org/10.1080/03610928508828965 MR797636

Scott-Hayward, L. A. S., Mackenzie, M. L., Ashe, E. and Williams, R. (2015). Modelling killer whale feeding behaviour using a spatially adaptive complex region spatial smoother (CReSS) and generalised estimating equations (GEEs). *J. Agric. Biol. Environ. Stat.* **20** 305–322. https://doi.org/10.1007/s13253-015-0209-2 MR3396553

Tsybakov, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.* **26** 2420–2469. https://doi.org/10.1214/aos/1024691478 MR1700239

Tsybakov, A. B. (2004). *Introduction à l'estimation non-paramétrique [french]*. *Mathématiques & Applications (Berlin) [Mathematics & Applications]* **41**. Springer-Verlag, Berlin. MR2013911

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. https://doi.org/10.1007/b13794 MR2724359

Varet, S., Lacour, C., Massart, P. and Rivoirard, V. (2023). Numerical performance of penalized comparison to overfitting for multivariate kernel density estimation. *ESAIM Probab. Stat.* **27** 621–667. https://doi.org/10.1051/ps/2022018 MR4602387

Vitale, R. A. (1975). Bernstein polynomial approach to density function estimation. In *Statistical inference and related topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 2; dedicated to Z. W. Birnbaum)* 87–99. Academic Press, New York. https://doi.org/10.1016/B978-0-12-568002-8.50011-2 MR0397977

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. *Monographs on Statistics and Applied Probability* **60**. Chapman and Hall, Ltd., London. https://doi.org/10.1007/978-1-4899-4493-1 MR1319818

Wang, H. and Ranalli, M. G. (2007). Low-rank smoothing splines on complicated domains. *Biometrics* **63** 209–217. https://doi.org/10.1111/j.1541-0420.2006.00674.x MR2345591

Wood, S. N., Bravington, M. V. and Hedley, S. L. (2008). Soap film smoothing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 931–955. https://doi.org/10.1111/j.1467-9868.2008.00665.x MR2530324

Xu, M. and Samworth, R. J. (2021). High-dimensional nonparametric density estimation via symmetry and shape constraints. *Ann. Statist.* **49** 650–672. https://doi.org/10.1214/20-aos1972 MR4255102

YU, S., WANG, G., WANG, L. and YANG, L. (2021). Multivariate spline estimation and inference for image-on-scalar regression. *Statist. Sinica* **31** 1463–1487. https://doi.org/10.5705/ss.202019.0188 MR4297712

ZHANG, S. and KARUNAMUNI, R. J. (1998). On kernel density estimation near endpoints. *J. Statist. Plann. Inference* **70** 301–316. https://doi.org/10.1016/S0378-3758(97)00187-0 MR1649872

ZHANG, S., KARUNAMUNI, R. J. and JONES, M. C. (1999). An improved estimator of the density function at the boundary. *J. Amer. Statist. Assoc.* **94** 1231–1241. https://doi.org/10.2307/2669937 MR1731485

ZHANG, S. and KARUNAMUNI, R. J. (2000). On nonparametric density estimation at the boundary. *J. Nonparametr. Statist.* **12** 197–221. https://doi.org/10.1080/10485250008832805 MR1752313

ZHOU, L. and PAN, H. (2014). Smoothing noisy data for irregular regions using penalized bivariate splines on triangulations. *Comput. Statist.* **29** 263–281. https://doi.org/10.1007/s00180-013-0448-z MR3260122