

ADAPTATION TO LOWEST DENSITY REGIONS WITH APPLICATION TO SUPPORT RECOVERY¹

BY TIM PATSCHKOWSKI AND ANGELIKA ROHDE

Ruhr-Universität Bochum

A scheme for locally adaptive bandwidth selection is proposed which sensitively shrinks the bandwidth of a kernel estimator at lowest density regions such as the support boundary which are unknown to the statistician. In case of a Hölder continuous density, this locally minimax-optimal bandwidth is shown to be smaller than the usual rate, even in case of homogeneous smoothness. Some new type of risk bound with respect to a density-dependent standardized loss of this estimator is established. This bound is fully nonasymptotic and allows to deduce convergence rates at lowest density regions that can be substantially faster than $n^{-1/2}$. It is complemented by a weighted minimax lower bound which splits into two regimes depending on the value of the density. The new estimator adapts into the second regime, and it is shown that simultaneous adaptation into the fastest regime is not possible in principle as long as the Hölder exponent is unknown. Consequences on plug-in rules for support recovery are worked out in detail. In contrast to those with classical density estimators, the plug-in rules based on the new construction are minimax-optimal, up to some logarithmic factor.

1. Introduction. Adaptation in the classical context of nonparametric function estimation in Gaussian white noise has been extensively studied in the statistical literature. Since Nussbaum (1996) has established asymptotic equivalence in Le Cam's sense for the nonparametric models of density estimation and Gaussian white noise, a rigorous framework is provided which allows to carry over specific statistical results established for the Gaussian white noise model to the model of density estimation, at least in dimension one. Density estimation is as one of the most fundamental problems in statistics subject to a variety of recent studies; see, for example, Efromovich (2008), Gach, Nickl and Spokoiny (2013), Lepski (2013), Birgé (2014) and Liu and Wong (2014). It has become clear that under the conditions for the asymptotic equivalence to hold, minimax rates of convergence in density estimation with respect to pointwise or mean integrated squared error loss coincide with the optimal convergence rates obtained in the context of nonparametric regression, and the procedures are typically identical on the level of ideas. A main requisite on the density for Nussbaum's (1996) asymptotic equivalence is

Received September 2014; revised May 2015.

¹Supported by the DFG Priority Program SPP 1324, RO 3766/2-1.

MSC2010 subject classifications. 62G07.

Key words and phrases. Anisotropic density estimation, bandwidth selection, adaptation to lowest density regions, density dependent minimax optimality, support estimation.

the assumption that it is compactly supported and uniformly bounded away from zero on its support. If this assumption is violated, the density estimation experiment may produce statistical features which do not have any analog in the regression context. For instance, minimax estimation of noncompactly supported densities under L_p -loss bears striking differences to the compact case; see [Juditsky and Lambert-Lacroix \(2004\)](#), [Reynaud-Bouret, Rivoirard and Tuleau-Malot \(2011\)](#) and [Goldenshluger and Lepski \(2011, 2014\)](#). The minimax rates reflect an interplay of the regularity parameters and the parameter of the loss function, an effect which is caused by the tail behavior of the densities under consideration. In this article, we recover such an exclusive effect even for compactly supported densities. It turns out that minimax estimation in regions where the density is small is possible with higher accuracy although fewer observations are available, leading to rates which can be substantially faster than $n^{-1/2}$. Even more, this accuracy can be achieved to a large extent without a priori knowledge of these regions by a kernel density estimator with an adaptively selected bandwidth. As discovered by [Butucea \(2001\)](#), the exact constant of normalization for pointwise adaptive univariate density estimation on Sobolev classes depends increasingly on the density at the point of estimation itself. The crucial observation is that the classical bias variance trade-off does not reflect the dependence of the kernel estimator's variance on the density, which brings the idea of an estimated variance in the bandwidth selection rule into play. Although Butucea's interesting result requires the point of estimation to be fixed, it suggests that a potential gain in the rate might be possible at lowest density regions. In this paper, we investigate the problem of adaptation to lowest density regions under anisotropic Hölder constraints. A bandwidth selection rule is introduced which provably attains fast pointwise rates of convergence at lowest density regions. On this way, new weighted lower risk bounds over anisotropic Hölder classes are established, which split into two regimes depending on the value of the density. We show that the new estimator uniformly improves the global minimax rate of convergence, adapts to the second regime and finally that adaptation into the fastest regime is not possible in principle if the density's regularity is unknown. We identify the best possible adaptive rate of convergence

$$n^{-\bar{\beta}/(\bar{\beta}+d)}$$

(up to a logarithmic factor), where $\bar{\beta}$ is the unnormalized harmonic mean of the d -dimensional Hölder exponent.

This breakpoint determines the attainable speed of convergence of plug-in estimators for functionals of the density where the quality of estimation at the boundary is crucial. We exemplarily demonstrate it for the problem of support recovery. In order to line up with the related results of [Cuevas and Fraiman \(1997\)](#) about plug-in rules for support estimation and [Rigollet and Vert \(2009\)](#) on minimax analysis of plug-in level-set estimators, we measure the performance of the plug-in support estimator with respect to the global measure of symmetric difference of

sets under the margin condition [Polonik (1995); see also Mammen and Tsybakov (1999) and Tsybakov (2004)]. In contrast to level set estimation, however, plug-in rules for the support functional possess sub-optimal convergence rates when the classical kernel density estimator with minimax-optimal global bandwidth choice is used. We determine the optimal minimax rate for support recovery

$$n^{-\gamma\beta/(\beta+d)}$$

(up to a logarithmic factor), where γ denotes the margin exponent, d the dimension and β the isotropic Hölder exponent. Our result demonstrates that support recovery is possible with higher accuracy than level set estimation as already conjectured by Tsybakov (1997). We finally show that the performance of the plug-in support estimator resulting from our new density estimator turns out to be minimax-optimal up to a logarithmic factor.

The article is organized as follows. Section 2 contains the basic notation. In Section 3, the adaptive density estimator is introduced, new weighted lower pointwise risk bounds are derived and the optimality performance of the estimator is proved. Section 4 addresses the important problem of density support estimation as an example of a functional which substantially benefits from the new density estimator. The proofs are deferred to Section 5 and the supplemental article [Patschkowski and Rohde (2015)].

2. Preliminaries and notation. All our estimation procedures are based on a sample of n real-valued d -dimensional random vectors $X_i = (X_{i,1}, \dots, X_{i,d})$, $i = 1, \dots, n$ ($d \geq 1$ and if not stated otherwise $n \geq 2$), that are independent and identically distributed according to some unknown probability measure \mathbb{P} on \mathbb{R}^d with continuous Lebesgue density p . $\mathbb{E}_p^{\otimes n}$ denotes the expectation with respect to the n -fold product measure $\mathbb{P}^{\otimes n}$. Let

$$\hat{p}_{n,h}(t) = \hat{p}_{n,h}(t, X_1, \dots, X_n) := \frac{1}{n} \sum_{i=1}^n K_h(t - X_i),$$

denote the kernel density estimator with d -dimensional bandwidth $h = (h_1, \dots, h_d)$ at point $t \in \mathbb{R}^d$, where

$$K_h(x) := \left(\prod_{i=1}^d h_i \right)^{-1} K\left(\frac{x_1}{h_1}, \dots, \frac{x_d}{h_d}\right)$$

describes a rescaled kernel supported on $\prod_{i=1}^d [-h_i, h_i]$. The kernel function K is assumed to be compactly supported on $[-1, 1]^d$ and to be of product structure, that is, $K(x_1, \dots, x_d) = \prod_{i=1}^d K_i(x_i)$. Additionally, $K_{i,h_i}(x) := h_i^{-1} K_i(x/h_i)$, $i = 1, \dots, d$. The components K_i are assumed to integrate to one and to be continuous on its support with $K_i(0) > 0$. If not stated otherwise, they are symmetric and nonnegative, implying that the kernel is of first order. Recall that K is said to be

of k th order, $k = (k_1, \dots, k_d) \in \mathbb{N}^d$, if the functions $x \mapsto x_i^{j_i} K_i(x_i)$, $j_i \in \mathbb{N}$ with $1 \leq j_i \leq k_i, i = 1, \dots, d$, satisfy

$$\int x_i^{j_i} K_i(x_i) d\lambda(x_i) = 0,$$

where λ^d denotes the Lebesgue measure on \mathbb{R}^d throughout the article. The Lebesgue measure on \mathbb{R} is denoted by λ . For any function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, we define the univariate functions

$$(2.1) \quad \begin{aligned} f_{i,x} : \mathbb{R} &\longrightarrow \mathbb{R} \\ y &\longmapsto f(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_d) \end{aligned}$$

and denote by $P_{y,l}^{(f_{i,x})}$ the Taylor polynomial

$$(2.2) \quad P_{y,l}^{(f_{i,x})}(\cdot) := \sum_{k=0}^l \frac{f_{i,x}^{(k)}(y)}{k!} (\cdot - y)^k$$

of $f_{i,x}$ at the point $y \in \mathbb{R}$ of degree l (whenever it exists). Let $\mathcal{H}_d(\beta, L)$ be the anisotropic Hölder class with regularity parameters (β, L) , that is, any function f belonging to this class fulfills for all $y, y' \in \mathbb{R}$ the inequality

$$\sup_{x \in \mathbb{R}^d} |f_{i,x}(y) - f_{i,x}(y')| \leq L|y - y'|^{\beta_i}$$

for those $i \in \{1, \dots, d\}$ with $\beta_i \leq 1$, and in case $\beta_i > 1$ admits derivatives with respect to its i th coordinate up to the order $\lfloor \beta_i \rfloor := \max\{n \in \mathbb{N} : n < \beta_i\}$, such that the approximation by the Taylor polynomial satisfies

$$\sup_{x \in \mathbb{R}^d} |f_{i,x}(y) - P_{y', \lfloor \beta_i \rfloor}^{(f_{i,x})}(y)| \leq L|y - y'|^{\beta_i} \quad \text{for all } y, y' \in \mathbb{R}.$$

For adaptation issues, it is assumed that $\beta = (\beta_1, \dots, \beta_d) \in \prod_{i=1}^d [\beta_{i,l}^*, \beta_{i,u}^*]$ and $L \in [L_l^*, L_u^*]$ for some positive constants $\beta_{i,l}^* < \beta_{i,u}^*, i = 1, \dots, d$, and $L_l^* < L_u^*$. For short, we simply write β^* and L^* for the couples (β_l^*, β_u^*) and (L_l^*, L_u^*) , and finally $\mathcal{R}(\beta^*, L^*)$ for the rectangle $\prod_{i=1}^d [\beta_{i,l}^*, \beta_{i,u}^*] \times [L_l^*, L_u^*]$. It turns out that all rates of convergence emerging in an anisotropic setting involve the unnormalized harmonic mean of the smoothness parameters

$$\bar{\beta} := \left(\sum_{i=1}^d \frac{1}{\beta_i} \right)^{-1}.$$

To focus on rates only and for ease of notation, we denote by c positive constants that may change from line to line. All relevant constants will be numbered consecutively. Dependencies of the constants on the functional classes' parameters are always indicated and it should be kept in mind that the constants can potentially

depend on the chosen kernel, the loss function and the dimension as well. Furthermore, $\mathcal{P}_d(\beta, L)$ denotes the set of all probability densities in $\mathcal{H}_d(\beta, L)$. It is well known that any function $f \in \mathcal{P}_d(\beta, L)$ is uniformly bounded by a constant

$$(2.3) \quad c_1(\beta, L) = \sup\{\|p\|_{\text{sup}} : p \in \mathcal{P}_d(\beta, L)\}$$

depending on the regularity parameters only.

3. New lower risk bounds, adaptation to lowest density regions. The fully nonparametric problem of estimating a density p at some given point $t = (t_1, \dots, t_d)$ has quite a long history in the statistical literature and has been extensively studied. Considering different estimators, a very natural question is whether there is an estimator that is optimal and how optimality can be exactly described. A common concept of optimality is stated in a minimax framework. An estimator $T_n(t) = T_n(t, X_1, \dots, X_n)$ is called minimax-optimal over the class $\mathcal{P}_d(\beta, L)$ if its risk matches the minimax risk

$$\inf_{T_n(t)} \sup_{p \in \mathcal{P}_d(\beta, L)} \mathbb{E}_p^{\otimes n} |T_n(t) - p(t)|^r$$

for some $r \geq 1$, where the infimum is taken over all estimators. However, the minimax approach is often rated as quite pessimistic as it aims at finding an estimator which performs best in the worst situation. Different in spirit is the oracle approach. Within a pre-specified class \mathcal{T} of estimators, it aims at finding for any individual density the estimator $\hat{T}_n \in \mathcal{T}$ which is optimal, leading to oracle inequalities of the form

$$\mathbb{E}_p^{\otimes n} |\hat{T}_n(t) - p(t)|^r \leq c \inf_{T_n \in \mathcal{T}} \mathbb{E}_p^{\otimes n} |T_n(t) - p(t)|^r + R_n(t)$$

with a remainder term $R_n(t)$ depending on the class \mathcal{T} , the underlying density p and the sample size only. Besides having the drawback that there is no notion of optimality judging about the adequateness of the estimator’s class, an equally severe problem may be caused by the fact that the remainder term is uniform in \mathcal{T} , and thus a worst case remainder. The latter is responsible for the fact that our fast convergence rates cannot be deduced from the oracle inequality in Goldenshluger and Lepski (2013), the order for their remainder being unimprovable, however. In this article, we introduce the notion of best possible p -dependent minimax speed of convergence $\psi_{p(t), \beta, L}^n$ within the function class $\mathcal{P}_d(\beta, L)$ and aim at constructing an estimator $T_n(t)$ bounding the risk

$$\sup_{p \in \mathcal{P}_d(\beta, L)} \sup_{\substack{t \in \mathbb{R}^d: \\ p(t) > 0}} \mathbb{E}_p^{\otimes n} \left(\frac{|T_n(t) - p(t)|}{\psi_{p(t), \beta, L}^n} \right)^r$$

uniformly over a range of parameters (β, L) . First, this requires a suitable definition of the quantity $\psi_{p(t), \beta, L}^n$.

3.1. *New weighted lower risk bound.* As we want to work out the explicit dependence on the value of the density, it seems suitable to fix an arbitrary constant $\varepsilon \in (0, 1)$, and to pick out maximal not necessarily disjoint subsets U_δ of $\mathcal{P}_d(\beta, L)$ with the following properties: $\cup U_\delta = \{p \in \mathcal{P}_d(\beta, L) : p(t) > 0\}$, and pairwise ratios $p(t)/q(t)$, $p, q \in U_\delta$, are bounded away from zero by ε and from infinity by $1/\varepsilon$. This motivates the construction of the subsequent theorem.

THEOREM 3.1 (New weighted lower risk bound). *For any $\beta = (\beta_1, \dots, \beta_d)$ with $0 < \beta_i \leq 2, i = 1, \dots, d, L > 0$ and $r \geq 1$, there exist constants $c_2(\beta, L, r) > 0$ and $n_0(\beta, L) \in \mathbb{N}$, such that for every $t \in \mathbb{R}$ the pointwise minimax risk over Hölder-smooth densities is bounded from below by*

$$\inf_{0 < \delta \leq c_1(\beta, L)} \inf_{T_n(t)} \sup_{\substack{p \in \mathcal{P}_d(\beta, L): \\ \delta/2 \leq p(t) \leq \delta}} \mathbb{E}_p^{\otimes n} \left(\frac{|T_n(t) - p(t)|}{\psi_{p(t), \beta}^n} \right)^r \geq c_2(\beta, L, r)$$

for all $n \geq n_0(\beta, L)$, where $\psi_{x, \beta}^n := x \wedge (x/n)^{\bar{\beta}/(2\bar{\beta}+1)}$ and $c_1(\beta, L)$ defined in (2.3).

REMARK 3.2. (i) The lower bound of the above theorem is attained by the oracle estimator

$$(3.1) \quad T_n(t) := \hat{p}_{n, h_{n, \delta}}(t) \cdot \mathbb{1}\{\delta \geq n^{-\bar{\beta}/(2\bar{\beta}+1)}\}$$

with $h_{n, \delta, i} = (\delta/n)^{(1/(2\bar{\beta}+1))(1/\beta_i)}$. Hence, $\psi_{p(t), \beta}^n$ cannot be improved in principle. We refer to it in the sequel as p -dependent speed of convergence within the functional class $\mathcal{P}_d(\beta, L)$.

(ii) Note that for the classical minimax rate $n^{-\bar{\beta}/(2\bar{\beta}+1)}$,

$$\lim_{n \rightarrow \infty} \inf_{0 < \delta \leq c_1(\beta, L)} \inf_{T_n(t)} \sup_{\substack{p \in \mathcal{P}_d(\beta, L): \\ \delta/2 \leq p(t) \leq \delta}} \mathbb{E}_p^{\otimes n} \left(\frac{|T_n(t) - p(t)|}{n^{-\bar{\beta}/(2\bar{\beta}+1)}} \right)^r = 0$$

as a direct consequence of the subsequently formulated Theorem 3.3. The p -dependent speed of convergence $\psi_{p(t), \beta}^n$ is of substantially smaller order than the classical one along a shrinking neighborhood of lowest density regions.

Note that the exponent $\bar{\beta}/(2\bar{\beta}+1)$ implicitly depends on the dimension d and coincides in case of isotropic smoothness with the well-known exponent $\beta/(2\beta+d)$. It splits into two regimes which are listed and specified in the following table.

	Regime	Rate $\psi_{x, \beta}^n$
(i)	$x \leq n^{-\bar{\beta}/(2\bar{\beta}+1)}$	x
(ii)	$n^{-\bar{\beta}/(2\bar{\beta}+1)} < x \leq c_1(\beta, L)$	$(\frac{x}{n})^{\bar{\beta}/(2\bar{\beta}+1)}$

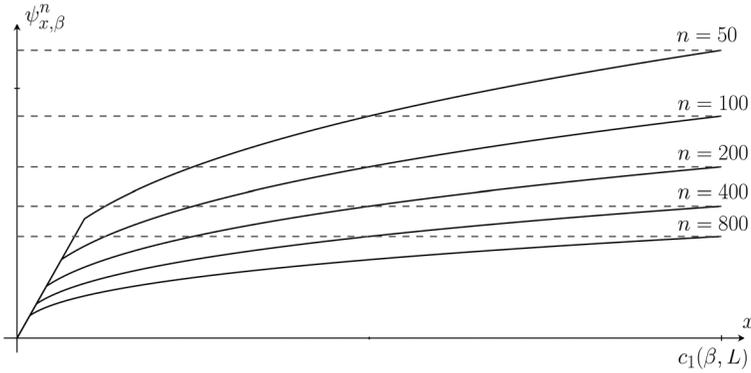


FIG. 1. *New lower bound (solid line), classical lower bound (dashed line).*

The worst p -dependent speed of convergence within $\mathcal{P}_d(\beta, L)$, namely

$$\sup_{0 < x \leq c_1(\beta, L)} \psi_{x, \beta}^n,$$

reveals the classical minimax rate $n^{-\bar{\beta}/(2\bar{\beta}+1)}$. The fastest rate in regime (ii) is of the order

$$n^{-\bar{\beta}/(\bar{\beta}+1)} \quad \text{for } x = n^{-\bar{\beta}/(\bar{\beta}+1)},$$

which is substantially smaller than the classical minimax risk bound. Figure 1 visualizes the split-up into the regimes and relates the new p -dependent rate of Theorem 3.1 to the classical minimax rate for different sample sizes from $n = 50$ to $n = 800$.

It becomes apparent from the proof that the lower bound actually even holds for the subset of (β, L) -regular densities with compact support. At first glance, however, the new lower bound is of theoretical value only, because the value of a density at some point to be estimated is unknown. The question is whether it is possible to improve the local rate of convergence of an estimator without prior knowledge in regions where fewer observations are available, that is, to which extent it is possible to adapt to lowest density regions.

3.2. Adaptation to lowest density regions. Adaptation is an important challenge in nonparametric estimation. Lepski (1990) introduced a sequential multiple testing procedure for bandwidth selection of kernel estimators in the Gaussian white noise model. It has been widely used and refined for a variety of adaptation issues over the last two decades. For recent references, see Giné and Nickl (2010), Chichignoud (2012), Goldenshluger and Lepski (2011, 2014), Chichignoud and Lederer (2014), Jirak, Meister and Reiß (2014), Dattner, Reiss and Trabs (2014) and Bertin, Lacour and Rivoirard (2014) and Lepski (2015) among many others. Our subsequently constructed estimator is based on the anisotropic bandwidth selection procedure of Kerkycharian, Lepski and Picard (2001), which has been de-

veloped in the Gaussian white noise model, but incorporates the new approach of adaptation to lowest density regions. Although Goldenshluger and Lepski (2013) pursue a similar goal via some kind of empirical risk minimization, their oracle inequality provides no faster rates than $n^{-1/2}$ times the average of the density over the unit cube around the point under consideration. They deduce from it adaptive minimax rates of convergence with respect to the L_p -risk over anisotropic Nikol'skii classes for density estimation on \mathbb{R}^d . As concerns adaptation to lowest density regions such as the unknown support boundary, this oracle inequality is not sufficient as no faster rates than $n^{-1/2}$ can be deduced from it, and it is not clear whether these faster rates are attainable for their estimator in principle. Besides having the drawback that there is no notion of optimality judging about the adequateness of the estimator's class, an equally severe problem of the oracle approach may be caused by the fact that the remainder term is uniform in the estimator's class, and thus a worst case remainder. The latter is responsible for the fact that our fast convergence rates cannot be deduced from the oracle inequality in Goldenshluger and Lepski (2013), the order for their remainder being unimprovable, however. It raises the question whether this imposes a fundamental limit on the possible range of adaptation (the corresponding inequality resulting from the bound on $\mathbb{P}^{\otimes n}(B_{1,m})$ has to be satisfied as well). We shall demonstrate in what follows that it is even possible to attain substantially faster rates, indeed that adaptation to the whole second regime of Theorem 3.1 is an achievable goal, and that this describes precisely the full range where adaptation to lowest density regions is possible as long as the density's regularity is unknown. Our procedure uses kernel density estimators $\hat{p}_{n,h}(t)$ with multivariate bandwidths $h = (h_1, \dots, h_d)$, which are able to deal with different degrees of smoothness in different coordinate directions. Note that optimal bandwidths for estimation of Hölder-continuous densities are typically derived by a bias-variance trade-off balancing the bias bound

$$(3.2) \quad |p(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,h}(t)| \leq c(\beta, L) \cdot \sum_{i=1}^d h_i^{\beta_i},$$

see (5.3) in Section 5 for details, against the rough variance bound

$$(3.3) \quad \text{Var}(\hat{p}_{n,h}(t)) \leq \frac{c_1(\beta, L) \|K\|_2^2}{n \prod_{i=1}^d h_i},$$

where $\|\cdot\|_2$ is the Euclidean norm [on $L_2(\lambda^d)$]. This bound leads to suboptimal rates of convergence whenever the density is small since it is not able to capture small values of p in a small neighborhood around t in contrast to the sharp convolution bound

$$(3.4) \quad \text{Var}(\hat{p}_{n,h}(t)) \leq \frac{1}{n} ((K_h)^2 * p)(t) =: \sigma_t^2(h).$$

Balancing (3.2) and (3.4) leads to smaller bandwidths at lowest density regions as compared to bandwidths resulting from the classical bias-variance trade-off between (3.2) and (3.3). The convolution bound (3.4) is unknown and it is natural to

replace it by its unbiased empirical version

$$\tilde{\sigma}_t^2(h) := \frac{1}{n^2 \prod_{i=1}^d h_i^2} \sum_{i=1}^n K^2\left(\frac{t - X_i}{h}\right).$$

However, $\tilde{\sigma}_t^2(h)$ concentrates extremely poorly around its mean if the bandwidth h is small, which is just the important situation at lowest density regions. Precisely, Bernstein's inequality provides the bound

$$(3.5) \quad \mathbb{P}^{\otimes n} \left(\left| \frac{\tilde{\sigma}_t^2(h)}{\sigma_t^2(h)} - 1 \right| \geq \eta \right) \leq 2 \exp \left(- \frac{3\eta^2}{2(3+2\eta)\|K\|_{\text{sup}}^2} \sigma_t^2(h) \cdot n^2 \prod_{i=1}^d h_i^2 \right),$$

which suggests to study the following truncated versions instead:

$$(3.6) \quad \begin{aligned} \sigma_{t,\text{trunc}}^2(h) &:= \max \left\{ \frac{\log^2 n}{n^2 \prod_{i=1}^d h_i^2}, \sigma_t^2(h) \right\}, \\ \tilde{\sigma}_{t,\text{trunc}}^2(h) &:= \max \left\{ \frac{\log^2 n}{n^2 \prod_{i=1}^d h_i^2}, \tilde{\sigma}_t^2(h) \right\}. \end{aligned}$$

Without the logarithmic term, the truncation level ensures tightness of the family of random variables $\tilde{\sigma}_{t,\text{trunc}}^2(h)/\sigma_{t,\text{trunc}}^2(h)$, because the exponent in (3.5) remains a nondegenerate function in η . The logarithmic term is introduced in order to guarantee sufficient concentration of $\sup_h |1 - \tilde{\sigma}_{t,\text{trunc}}^2(h)/\sigma_{t,\text{trunc}}^2(h)|$.

Construction of the adaptive estimator. Our estimation procedure is developed in the anisotropic setting, in which neither the variance bound nor the bias bound provides an immediate monotone behavior in the bandwidth. Unlike in the univariate or isotropic multivariate case, Lepski's (1990) idea of mimicking the bias-variance trade-off fails. Consequently, our estimation scheme imitates the anisotropic procedure of Kerkycharian, Lepski and Picard (2001) and Klutchnikoff (2005), developed in the Gaussian white noise model, with the following changes. First, their threshold given by the variance bound in the Gaussian white noise setting is replaced essentially with the truncated estimate in (3.6), which is sensitive to small values of the density. Moreover, it is crucial in the anisotropic setting that our procedure uses an ordering of bandwidths according to these estimated variances instead of an ordering according to the product of the bandwidth's components. The bandwidth selection scheme chooses a bandwidth in the set

$$\mathcal{H} := \left\{ h = (h_1, \dots, h_d) \in \prod_{i=1}^d (0, h_{\max,i}] : \prod_{i=1}^d h_i \geq \frac{\log^2 n}{n} \right\},$$

where for simplicity we set $(h_{\max,1}, \dots, h_{\max,d}) = (1, \dots, 1)$. Let furthermore

$$\mathcal{J} := \left\{ j = (j_1, \dots, j_d) \in \mathbb{N}_0^d : \sum_{i=1}^d j_i \leq \left\lfloor \log_2 \left(\frac{n}{\log^2 n} \right) \right\rfloor \right\}$$

be a set of indices and denote by

$$\mathcal{G} := \{(2^{-j_1}, \dots, 2^{-j_d}) : j \in \mathcal{J}\} \subset \mathcal{H}$$

the corresponding dyadic grid of bandwidths, that serves as a discretization for the multiple testing problem in Lepski’s selection rule. For ease of notation, we abbreviate dependences on the bandwidth $(2^{-j_1}, \dots, 2^{-j_d})$ by the multi-index j . Next, with $j \wedge m$ denoting the minimum by component, the set of admissible bandwidths is defined as

$$\begin{aligned} \mathcal{A} &= \mathcal{A}(t) \\ (3.7) \quad &:= \left\{ j \in \mathcal{J} : |\hat{p}_{n,j \wedge m}(t) - \hat{p}_{n,m}(t)| \leq c_3 \sqrt{\hat{\sigma}_t^2(m) \log n} \right. \\ &\quad \left. \text{for all } m \in \mathcal{J} \text{ with } \hat{\sigma}_t^2(m) \geq \hat{\sigma}_t^2(j) \right\}, \end{aligned}$$

with a properly chosen constant $c_3 = c_3(\beta^*, L^*)$ satisfying the constraint (5.17) appearing in the proof of Theorem 3.3. Here, both the threshold and the ordering of bandwidths are defined via the truncated variance estimator

$$\begin{aligned} (3.8) \quad \hat{\sigma}_t^2(h) &:= \min \left\{ \tilde{\sigma}_{t,\text{trunc}}^2(h), \frac{\|K\|_2^2 c_1}{n \prod_{i=1}^d h_i} \right\} \\ &= \min \left\{ \max \left[\frac{\log^2 n}{n^2 \prod_{i=1}^d h_i^2}, \frac{1}{n^2 \prod_{i=1}^d h_i^2} \sum_{i=1}^n K^2 \left(\frac{t - X_i}{h} \right) \right], \frac{\|K\|_2^2 c_1}{n \prod_{i=1}^d h_i} \right\}, \end{aligned}$$

where $c_1 = c_1(\beta^*, L^*)$ is an upper bound on $c_1(\beta, L)$ in the range of adaptation. The threshold in (3.7) could be modified by a further logarithmic factor to avoid the dependence of the constants on the range of adaptation. Recall again that this refined estimated threshold is crucial for our estimation scheme. The procedure selects the bandwidth among all admissible bandwidths with

$$(3.9) \quad \hat{j} = \hat{j}(t) \in \arg \min_{j \in \mathcal{A}} \hat{\sigma}_t^2(j).$$

Finally,

$$\hat{p}_n := \hat{p}_{n,\hat{j}} \wedge c_1$$

defines the adaptive estimator. In case of isotropic Hölder smoothness, it is sufficient to restrict the grid to bandwidths with equal components, and we even simplify the method by replacing the ordering by estimated variances in condition (3.8) “for all $m \in \mathcal{J}$ with $\hat{\sigma}_t^2(m) \geq \hat{\sigma}_t^2(j)$ ” by the classical order “for all $m \in \mathcal{J}$ with $m \geq j$ ” as the componentwise ordering is the same for all components.

Performance of the adaptive estimator: Clearly, the truncation in the threshold imposes serious limitations to which extent adaptation to lowest densities regions is possible. However, a careful analysis of the ratio

$$\sup_h \left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(h)}{\sigma_{t,\text{trunc}}^2(h)} - 1 \right|$$

rather than the difference $\sup_h |\tilde{\sigma}_{t,\text{trunc}}^2(h) - \sigma_{t,\text{trunc}}^2(h)|$ allows to prove indeed that adaptation is possible in the whole second regime.

THEOREM 3.3 (New upper bound). *For any rectangle $\mathcal{R}(\beta^*, L^*)$ with $[\beta_{i,l}^*, \beta_{i,u}^*] \subset (0, 2]$, $[L_l^*, L_u^*] \subset (0, \infty)$ and $r \geq 1$, there exists a constant $c_4(\beta^*, L^*, r) > 0$, such that the new density estimator \hat{p}_n with adaptively chosen bandwidth according to (3.9) satisfies*

$$\sup_{(\beta, L) \in \mathcal{R}(\beta^*, L^*)} \sup_{p \in \mathcal{P}_d(\beta, L)} \sup_{t \in \mathbb{R}^d} \mathbb{E}_p^{\otimes n} \left(\frac{|\hat{p}_n(t) - p(t)|}{\tilde{\psi}_{p(t), \beta}^n} \right)^r \leq c_4(\beta^*, L^*, r)$$

for all $n \geq 2$, where

$$\tilde{\psi}_{x, \beta}^n := [n^{-\bar{\beta}/(\bar{\beta}+1)} \vee (x/n)^{\bar{\beta}/(2\bar{\beta}+1)}] (\log n)^{3/2}.$$

The p -dependent speed of convergence $\tilde{\psi}_{p(t), \beta}^n$ (except the logarithmic factor) is plotted in Figure 2, which shows the superiority of the new estimator in low density regions. It also depicts that the new estimator is able to adapt to regime (ii) up to a logarithmic factor, and that it improves the rate of convergence significantly in both regimes as compared to the classical minimax rate. Besides, although not emphasized before, \hat{p}_n is fully adaptive to the smoothness in terms of Hölder regularity.

As ψ and $\tilde{\psi}$ coincide (up to a logarithmic factor) in regime (ii) but differ in regime (i), the question arises whether the breakpoint

$$n^{-\bar{\beta}/(\bar{\beta}+1)}$$

describes the fundamental bound on the range of adaptation to lowest density regions. The following result shows that this is indeed the case as long as the density’s regularity is unknown.

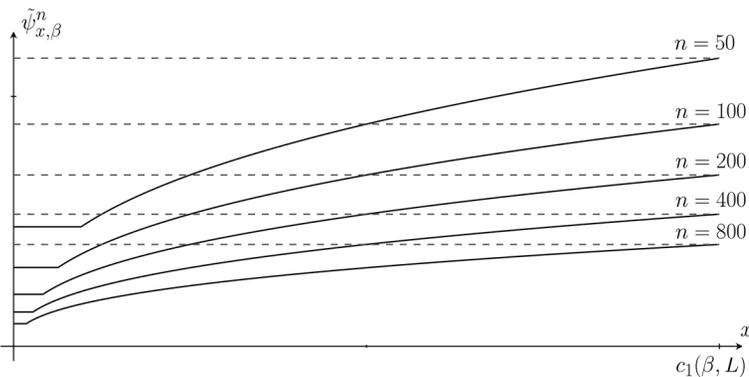


FIG. 2. *New upper bound without logarithmic factor (solid line), classical upper bound (dashed line).*

THEOREM 3.4. *For any $\beta_2 < \beta_1 \leq 2$ and any sequence $(\rho(n))$ converging to infinity with*

$$\rho(n) = O\left(n^{\frac{\beta_1 - \beta_2}{(2\beta_1 + 1)(\beta_2 + 1)}} (\log n)^{-3/2}\right),$$

there exist $L_1, L_2 > 0$ and densities $p_n \in \mathcal{P}_1(\beta_1, L_1)$ with

$$\frac{n^{-\beta_1/(\beta_1 + 1)}}{p_n(t)} = o(1)$$

as $n \rightarrow \infty$, such that for every estimator $T_n(t)$ satisfying

$$(3.10) \quad \mathbb{E}_{p_n}^{\otimes n} |T_n(t) - p_n(t)| \leq c_4(\beta_1^*, L_1^*, r) \left(\frac{p_n(t)}{n}\right)^{\beta_1/(2\beta_1 + 1)} (\log n)^{3/2},$$

there exist $n_0(\beta_1, \beta_2, L_1, L_2)$ and a constant $c > 0$ both independent of t , with

$$\sup_{\substack{q \in \mathcal{P}_d(\beta_2, L_2): \\ q(t) \leq c(n) \cdot n^{-\beta_2/(\beta_2 + 1)}}} \frac{\mathbb{E}_q^{\otimes n} |T_n(t) - q(t)|}{n^{-\beta_2/(\beta_2 + 1)}} \geq c$$

for all $n \geq n_0(\beta_1, \beta_2, L_1, L_2)$ and any sequence $(c(n))$ with $c(n) \geq \rho(n)^{-1}$.

The following consideration provides a heuristic reason why adaptation to regime (i) is not possible in principle. Consider the univariate and Lipschitz continuous triangular density $p : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto (1 - |x|)\mathbb{1}\{|x| \leq 1\}$. If $\delta_n < n^{-\beta/(\beta + 1)} = n^{-1/2}$, the expected number of observations in $\{p \leq \delta_n\}$ is less than one. Without the knowledge of the regularity, it is intuitively clear that it is impossible to predict whether local averaging is preferable to just estimating by zero.

3.2.1. Adaptation to lowest density regions when β is known. If the Hölder exponent $\beta \in (0, 2]$ is known to the statistician, the form of the oracle estimator (3.1) suggests that some further improvement in regime (i) might be possible by considering the truncated estimator

$$(3.11) \quad \hat{p}_n(\cdot) \cdot \mathbb{1}\{\hat{p}_n(\cdot) \geq n^{-\bar{\beta}/(\bar{\beta} + 1)} (\log n)^{\zeta_1}\}$$

for some suitable constant $\zeta_1 > 0$. In fact, elementary algebra shows that this threshold does not affect the performance in regime (ii) (up to a logarithmic term). For isotropic Hölder smoothness, we prove in the supplemental article [Patschkowski and Rohde (2015)] that the estimator (3.11) indeed attains the p -dependent speed of convergence

$$\vartheta_{p(t), \beta}^n = \psi_{p(t), \beta}^n \vee n^{-\zeta_2}$$

up to logarithmic terms, with $\psi_{x, \beta}^n$ as defined in Theorem 3.1. Here, the constant ζ_2 can be made arbitrarily large by enlarging c_3 and ζ_1 . That is, if the Hölder exponent is known, adaptation to regime (i) is possible to a large extent.

3.2.2. *Extension to $\beta > 2$.* As concerns an extension of Theorems 3.1 and 3.3 to arbitrary $\beta > 2$, Lemma 5.1(ii) demonstrates that the variance of the kernel density estimator never falls below the reference speed of convergence $\check{\psi}_{p(t),\beta}^n$. However, it can be substantially larger, resulting in a lower speed of convergence as compared to the reference speed of convergence. Therefore, it seems necessary to introduce a p -dependent speed of convergence which does not incorporate the value of the density $p(t)$ only but also information on the derivatives. An exception of outstanding importance are points t close to the support boundary, because not only $p(t)$ itself but also all derivatives are necessarily small. Theorem A.1, which is deferred to the supplemental article [Patschkowski and Rohde (2015)], reveals that our procedure then even reaches the fast adaptive speed of convergence at the support boundary for every $\beta > 0$. In fact, as $\beta \rightarrow \infty$, adaptive rates arbitrarily close to n^{-1} can be attained.

4. Application to support recovery. The phenomenon of faster rates of convergence in regions where the density is small may have strong consequences on plug-in rules for certain functionals of the density. As an application of the results of Section 3, we investigate the support plug-in functional. Support estimation has a long history in the statistical literature. Geffroy (1964) and Rényi and Sulanke (1963, 1964) are cited as pioneering reference most commonly, followed by further contributions of Chevalier (1976), Devroye and Wise (1980), Grenander (1981), Hall (1982), Groeneboom (1988), Tsybakov (1989, 1991, 1997), Cuevas (1990), Korostelev and Tsybakov (1993), Härdle, Park and Tsybakov (1995), Mammen and Tsybakov (1995), Cuevas and Fraiman (1997), Gayraud (1997), Hall, Nussbaum and Stern (1997), Baïllo, Cuevas and Justel (2000), Cuevas and Rodríguez-Casal (2004), Klemelä (2004), and Biau, Cadre and Pelletier (2008), Biau, Cadre, Mason and Pelletier (2009), Brunel (2013) and Cholaquidis, Cuevas and Fraiman (2014) as a by far nonexhaustive list of contributions. In order to demonstrate the substantial improvement in the rates of convergence for the plug-in support estimator based on the new density estimator, we first establish minimax lower bounds for support estimation under the margin condition which have not been provided in the literature so far. Theorems 4.4 and 4.5 then reveal that the minimax rates for the support estimation problem are substantially faster than for the level set estimation problem, as already conjectured in Tsybakov (1997). In fact, in the level set estimation framework, when β and L are given, the classical choice of a bandwidth of order $n^{-1/(2\beta+d)}$ in case of isotropic Hölder smoothness leads directly to a minimax-optimal plug-in level set estimator as long as the offset is suitably chosen [Rigollet and Vert (2009)]. In contrast, this bandwidth produces suboptimal rates in the support estimation problem, no matter how the offset is chosen. At first sight, this makes the plug-in rule as a by-product of density estimation inappropriate. We shall demonstrate subsequently, however, that our new density estimator avoids this problem. In order to line up with the results of Cuevas and Fraiman

(1997) and Rigollet and Vert (2009), we work essentially under the same type of conditions. The distance between two subsets A and B of \mathbb{R}^d is measured by

$$d_\Delta(A, B) := \lambda^d(A \Delta B),$$

where Δ denotes the symmetric difference of sets

$$A \Delta B := (A \setminus B) \cup (B \setminus A).$$

Subsequently, \bar{A} denotes the topological closure of a set $A \subset \mathbb{R}^d$. We impose the following condition, which characterizes the complexity of the problem. It was introduced by Polonik (1995) [see also Mammen and Tsybakov (1999), Tsybakov (2004) and Cuevas and Fraiman (1997)], where the latter authors referred to it as sharpness order.

DEFINITION 4.1 (Margin condition). A density $p : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to satisfy the κ -margin condition with exponent $\gamma > 0$, if

$$\lambda^d(\overline{\{x \in \mathbb{R}^d \mid 0 < p(x) \leq \varepsilon\}}) \leq \kappa_2 \cdot \varepsilon^\gamma$$

for all $0 < \varepsilon \leq \kappa_1$, where $\kappa = (\kappa_1, \kappa_2) \in (0, \infty)^2$.

In particular, $\lambda^d(\partial\Gamma_p) = 0$ for every density which satisfies the margin condition, where $\partial\Gamma_p$ denotes the boundary of the support Γ_p . To highlight the line of ideas, we restrict the application to the important special case of isotropic smoothness. Let $\mathcal{H}_d^{\text{iso}}(\beta, L)$ denote the isotropic Hölder class with one-dimensional parameters β and L , which is for $0 < \beta \leq 1$ defined by

$$\mathcal{H}_d^{\text{iso}}(\beta, L) := \{f : \mathbb{R}^d \rightarrow \mathbb{R} \mid |f(x) - f(y)| \leq L \|x - y\|_2^\beta \text{ for all } x, y \in \mathbb{R}^d\}.$$

For $\beta > 1$, it is defined as the set of all functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ that are $\lfloor \beta \rfloor$ times continuously differentiable such that the following property is satisfied:

$$(4.1) \quad |f(x) - P_{y, \lfloor \beta \rfloor}^{(f)}(x)| \leq L \|x - y\|_2^\beta \quad \text{for all } x, y \in \mathbb{R}^d,$$

where

$$P_{y, l}^{(f)}(x) := \sum_{|k| \leq l} \frac{D^k f(y)}{k_1! \cdots k_d!} (x_1 - y_1)^{k_1} \cdots (x_d - y_d)^{k_d}$$

with $|k| := \sum_{i=1}^d k_i$ and the partial differential operator

$$D^k := \frac{\partial^{|k|}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}$$

denotes the multivariate Taylor polynomial of f at the point $y \in \mathbb{R}^d$ up to the l th order; see also (2.2) for the coinciding definition in one dimension. Correspondingly, $\mathcal{P}_d^{\text{iso}}(\beta, L)$ denotes the set of probability densities contained in $\mathcal{H}_d^{\text{iso}}(\beta, L)$. The following lemma demonstrates that not every combination of margin exponent and Hölder continuity is possible.

LEMMA 4.2. *There exists a compactly supported density in $\mathcal{P}_d^{\text{iso}}(\beta, L)$ satisfying a margin condition to the exponent γ if and only if $\gamma\beta \leq 1$.*

4.1. *Lower risk bounds for support recovery.* For any subset $A \subset \mathbb{R}^d$ and $\varepsilon > 0$, the closed outer parallel set of A at distance $\varepsilon > 0$ is given by

$$A^\varepsilon := \left\{ x \in \mathbb{R}^d : \inf_{y \in A} \|x - y\|_2 \leq \varepsilon \right\}$$

and the closed inner ε -parallel set by $A^{-\varepsilon} := \overline{((A^\varepsilon)^c)^c}$. Here, $\|\cdot\|_2$ denotes the Euclidean norm (on \mathbb{R}^d). A support satisfying

$$0 < \liminf_{\varepsilon \rightarrow 0} \frac{\lambda^d(\Gamma_p \setminus \Gamma_p^{-\varepsilon})}{\lambda^d(\Gamma_p^\varepsilon \setminus \Gamma_p)} \leq \limsup_{\varepsilon \rightarrow 0} \frac{\lambda^d(\Gamma_p \setminus \Gamma_p^{-\varepsilon})}{\lambda^d(\Gamma_p^\varepsilon \setminus \Gamma_p)} < \infty$$

is referred to as boundary regular support. Note that a support is always boundary regular if its Minkowski surface measure is well-defined (in the sense that outer and inner Minkowski content exist and coincide). The minimax lower bound is formulated under the assumption of Γ_p fulfilling the following complexity condition (to the exponent $\mu = \gamma\beta$), which even slightly weakens the assumption of boundary regularity under the margin condition.

DEFINITION 4.3 (Complexity condition). A set A is said to satisfy the ξ -complexity condition to the exponent $\mu > 0$ if for all $0 < \varepsilon \leq \xi_1$ there exists a disjoint decomposition $A = A_{1,\varepsilon} \cup A_{2,\varepsilon}$ such that

$$\frac{\lambda^d(A_{1,\varepsilon} \setminus A_{1,\varepsilon}) \vee \lambda^d(A_{2,\varepsilon})}{\varepsilon^\mu} \leq \xi_2,$$

where $\xi = (\xi_1, \xi_2) \in (0, \infty)^2$.

Note that a boundary regular support of a (β, L) -Hölder-smooth density satisfying the margin condition to the exponent γ fulfills the complexity condition to the exponent $\mu \geq \gamma\beta$ for the canonical decomposition $\Gamma_p = \Gamma_p \cup \emptyset$. Let us finally relate the margin condition (4.1) to the two-sided margin condition

$$\lambda^d \{x \in \mathbb{R}^d : 0 < |p(x) - \lambda| \leq \varepsilon\} \leq c\varepsilon^\gamma,$$

which is imposed in the context of density level set estimation for some level $\lambda > 0$; cf. Rigollet and Vert (2009). If $\Gamma_{p,\lambda} = \{x \in \mathbb{R}^d : p(x) > \lambda\}$ denotes the λ -level set at level $\lambda > 0$, the two-sided (κ, γ) -margin condition provides the bound

$$(4.2) \quad \lambda^d(\Gamma_{p,\lambda}^\varepsilon \setminus \Gamma_{p,\lambda}) \leq \kappa_2 (c\varepsilon^{\beta \wedge 1})^\gamma$$

for all $\varepsilon \leq \kappa_1$, where $c = L$ for $\beta \leq 1$ and $c = \sup_{x \in \mathbb{R}^d} \|\nabla p(x)\|_2$ for $\beta > 1$. In contrast, the margin condition at $\lambda = 0$ provides no bound on $\lambda^d(\Gamma_p^\varepsilon \setminus \Gamma_p)$. The complexity condition is a mild assumption which guarantees such type of bound.

For $\beta \leq 1$, the relation (4.2) for $\lambda = 0$ implies the complexity condition to the exponent $\mu = \gamma\beta$. Note that the typical situation is indeed

$$\lambda^d(\Gamma_p^\varepsilon \setminus \Gamma_p)/\varepsilon = \mathcal{O}(1) \quad \text{and} \quad \varepsilon/\lambda^d(\Gamma_p^\varepsilon \setminus \Gamma_p) = \mathcal{O}(1)$$

as $\varepsilon \rightarrow 0$. For instance, this holds true for any finite union of compact convex sets in \mathbb{R}^d as a consequence of the isoperimetric inequality [Theorem III.2.2, Chavel (2001)] and Theorem 3.1 [Bhattacharya and Rango Rao (1976)]. If it exists, the limit

$$\lim_{\varepsilon \searrow 0} \frac{\lambda^d(\Gamma_p^\varepsilon \setminus \Gamma_p)}{\varepsilon}$$

corresponds to the surface measure of the boundary if the latter is sufficiently regular. Due to the relation $\gamma\beta \leq 1$ by Lemma 4.2 and the decomposition into suitable subsets, the complexity condition relaxes this regularity condition on the surface area substantially. The subset of $\mathcal{P}_d^{\text{iso}}(\beta, L)$ consisting of densities satisfying the κ -margin condition to the exponent γ with support fulfilling the ξ -complexity condition to the exponent $\mu = \gamma\beta$ is denoted by $\mathcal{P}_d^{\text{iso}}(\beta, L, \gamma, \kappa, \xi)$.

THEOREM 4.4 (Minimax lower bound). *For any $\beta > 0$ and any margin exponent $\gamma > 0$ with $\gamma\beta \leq 1$, there exist $c_5(\beta, L) > 0$, $n_0(\beta, L, \gamma) \in \mathbb{N}$ and parameters $\kappa, \xi \in (0, \infty)$, such that the minimax risk with respect to the measure of symmetric difference of sets is bounded from below by*

$$\inf_{\hat{\Gamma}_n} \sup_{p \in \mathcal{P}_d^{\text{iso}}(\beta, L, \gamma, \kappa, \xi)} \mathbb{E}_p^{\otimes n} [d_\Delta(\hat{\Gamma}_n, \Gamma_p)] \geq c_5(\beta, L) \cdot n^{-\gamma\beta/(\beta+d)}$$

for all $n \geq n_0(\beta, L, \gamma)$.

4.2. Minimax-optimal plug-in rule. We use the plug-in support estimator with the kernel density estimator of Section 3. This density estimator improves the rate of convergence in particular at the support boundary. For the isotropic procedure, the index set \mathcal{J} is restricted to bandwidths coinciding in all components. We even simplify the ordering by estimated variances in condition (3.8) “for all $m \in \mathcal{J}$ with $\hat{\sigma}_t^2(m) \geq \hat{\sigma}_t^2(j)$ ” by the classical order “for all $m \in \mathcal{J}$ with $m \geq j$ ” as Lemma 5.2 shows that the relevant orderings are equivalent up to multiplicative constants for $0 < \beta \leq 2$. Furthermore, under isotropic smoothness it is natural to use a rotation invariant kernel, that is, $K(x) = \tilde{K}(\|x\|_2)$ with \tilde{K} supported on $[0, 1]$ and continuous on its support with $\tilde{K}(0) > 0$. The following theorem shows that the corresponding plug-in rule

$$\hat{\Gamma}_n = \overline{\{x \in \mathbb{R}^d : \hat{p}_n(x) > \alpha_n\}}$$

with offset level

$$(4.3) \quad \alpha_n := c_6(\beta, L) \left(\frac{(\log n)^{3/2}}{n} \right)^{\beta/(\beta+d)} \sqrt{\log n}$$

and constant $c_6(\beta, L)$ specified in the proof of the following theorem, is able to recover the support with minimax optimal rate, up to a logarithmic factor.

THEOREM 4.5 (Uniform upper bound). *For any $\beta \leq 2$, $\gamma > 0$ with $\gamma\beta \leq 1$ and $\kappa, \xi \in (0, \infty)$, there exist a constant $c_7 = c_7(\beta, L, \gamma, \kappa, \xi) > 0$ and $n_0 \in \mathbb{N}$, such that*

$$\sup_{p \in \mathcal{P}_d^{\text{iso}}(\beta, L, \gamma, \kappa, \xi)} \mathbb{E}_p^{\otimes n} [d_\Delta(\hat{\Gamma}_n, \Gamma_p)] \leq c_7 \cdot n^{-\gamma\beta/(\beta+d)} (\log n)^{2\gamma}$$

for all $n \geq n_0$.

As the rate already indicates, it is getting apparent from the proof that this result can be established only if the minimax optimal density estimator actually adapts up to the fastest rate in regime (ii).

REMARK 4.6. The results show the simultaneous optimality of the adaptive density estimator of Section 3 in the plug-in rule for support estimation. Correspondingly, they are restricted to $\beta \leq 2$. Whether the rate $n^{-\gamma\beta/(\beta+d)}$ is minimax optimal for $\beta > 2$ provided $\gamma\beta \leq 1$, and whether it can be attained by a plug-in rule in principle, remains open for the moment.

Let us finally point out two consequences. We have shown that the optimal minimax rates for support estimation are significantly faster than the corresponding rates for level set estimation

$$n^{-\gamma\beta/(2\beta+d)}$$

under the margin condition [Rigollet and Vert (2009)]. Although any level set of a fixed density satisfying the margin condition to the exponent γ fulfills the complexity condition to the exponent $\mu = \gamma\beta$ as long as $\beta \leq 1$, the hypotheses in the proof of the lower bounds of Rigollet and Vert (2009) do even satisfy this condition for some fixed ξ , uniformly in n , as well. Hence, their optimal minimax rates of convergence remain the same under our condition. On an intuitive level, this phenomenon can be nicely motivated by comparing the Hellinger distance $H(\mathbb{P}, \mathbb{Q})$ between the probability measure \mathbb{P} with Lebesgue density p and \mathbb{Q} whose Lebesgue density $q = p + \tilde{p}$ is a perturbation of p with a small function \tilde{p} around the level $\alpha \geq 0$; see Tsybakov (1997), Extension (E4). If $\alpha > 0$, then simple Taylor expansion of $\sqrt{p + \tilde{p}}$ yields $H^2(\mathbb{P}, \mathbb{Q}) \sim \int \tilde{p}^2 d\lambda^d$, whereas $H^2(\mathbb{P}, \mathbb{Q}) \sim \int \tilde{p} d\lambda^d$ in case $\alpha = 0$. Thus, perturbations at the boundary ($\alpha = 0$) can be detected with the higher accuracy resulting in faster attainable rates for support estimation than for level set estimation. Moreover, the rates for plug-in support estimators already established in the literature by Cuevas and Fraiman (1997) turn out to be always suboptimal in case of Hölder continuous densities of boundary regular support. To be precise, Cuevas and Fraiman (1997) establish in Theorem 1(c) a convergence

rate under the margin condition given in terms of $\rho_n = n^\rho$ and the offset level $\alpha_n = n^{-\alpha}$ (in their notation), which are assumed to satisfy $0 < \alpha < \rho$ and their condition (R2), namely

$$\rho_n \int |\hat{p}_n - p| d\lambda^d = o_{\mathbb{P}}(1) \quad \text{and} \quad \rho_n \alpha_n^{1+\gamma} = o(1) \quad \text{as } n \rightarrow \infty.$$

As a consequence, $\rho_n = o(n^{\beta/(2\beta+d)})$ for typical candidates $p \in \mathcal{P}_d^{\text{iso}}(\beta, L)$, that is, densities p which are locally not smoother than (β, L) -regular. Under the margin condition to the exponent $\gamma > 0$, this limits their rate of convergence $n^{-\rho+\alpha}$ to

$$d_{\Delta}(\Gamma_p, \hat{\Gamma}_n) = o_{\mathbb{P}}(n^{-(\beta/(2\beta+d))(\gamma/(1+\gamma))}),$$

which is substantially slower than the above established minimax rate. The crucial point is that even with the improved density estimator of Section 3, the above mentioned condition on ρ_n in (R2) cannot be improved, because any estimator can possess the improved performance at lowest density regions only. For this reason, the L_1 -speed of convergence of a density estimator is not an adequate quantity to characterize the performance of the corresponding plug-in support estimator.

5. Lemmas 5.1–5.7, proofs of Theorems 3.3 and 3.4. Due to space constraints, all remaining proofs are deferred to the supplemental article [Patschkowski and Rohde (2015)]. In the proof of Theorem 3.3, we frequently make use of the bandwidth

$$(5.1) \quad \bar{h}_i := c_8(\beta, L) \cdot \max \left\{ \left(\frac{\log n}{n} \right)^{(\bar{\beta}/(\bar{\beta}+1))(1/\beta_i)}, \left(\frac{p(t) \log n}{n} \right)^{(\bar{\beta}/(2\bar{\beta}+1))(1/\beta_i)} \right\}$$

for $i = 1, \dots, d$, with constant $c_8(\beta, L)$ of Lemma 5.1, which can be thought of as an optimal adaptive bandwidth. The truncation in the definition of \bar{h} results from the necessary truncation in $\sigma_{t, \text{trunc}}^2$. With the exponents

$$(5.2) \quad \bar{j}_i = \bar{j}_i(t) := \left\lfloor \log_2 \left(\frac{1}{\bar{h}_i} \right) \right\rfloor + 1, \quad i = 1, \dots, d$$

the bandwidth $2^{-\bar{j}_i}$ is an approximation of \bar{h}_i by the next smaller bandwidth on the grid \mathcal{G} such that $\bar{h}_i/2 \leq 2^{-\bar{j}_i} \leq \bar{h}_i$ for all $i = 1, \dots, d$.

Before turning to the proof of Theorem 3.3, we collect some technical ingredients. First, recall the classical upper bound on the bias of a kernel density estimator. With the notation provided in Section 2, and K of order $\max_i \beta_i$ at least, we obtain

$$\begin{aligned} b_t(h) &:= p(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,h}(t) \\ &= \int K(x)(p(t+hx) - p(t)) d\lambda^d(x) \\ &= \sum_{i=1}^d \int K(x)(p([t, t+hx]_{i-1}) - p([t, t+hx]_i)) d\lambda^d(x), \end{aligned}$$

using the notation $[x, y]_0 = y$, $[x, y]_d = x$, $[x, y]_i = (x_1, \dots, x_i, y_{i+1}, \dots, y_d)$, $i = 1, \dots, d - 1$ for two vectors $x, y \in \mathbb{R}^d$ and denoting by $hx = (h_1x_1, \dots, h_dx_d)$ the componentwise product. Taylor expansions for those components i with $\beta_i \geq 1$ lead to

$$\begin{aligned} & p([t, t + hx]_{i-1}) - p([t, t + hx]_i) \\ &= \sum_{k=1}^{\lfloor \beta_i \rfloor} p_{i, [t, t + hx]_i}^{(k)}(t_i) \frac{(h_i x_i)^k}{k!} \\ & \quad + (p([t, t + hx]_{i-1}) - P_{t_i, \lfloor \beta_i \rfloor}^{(p_{i, [t, t + hx]_i})}(t_i + h_i x_i)). \end{aligned}$$

Hence,

$$(5.3) \quad |b_t(h)| \leq L \sum_{i=1}^d c_{9,i}(\beta) h_i^{\beta_i} =: B_t(h)$$

with constants $c_{9,i}(\beta) := \int |x_i|^{\beta_i} |K(x)| d\lambda^d(x) < \infty$.

With a slight abuse of notation, dependencies on some bandwidth $h = 2^{-j}$ are subsequently expressed in terms of the corresponding grid exponent $j = (j_1, \dots, j_d)$, that is, $B_t(h)$ equals $B_t(j)$, etc. For any multi-index j , we use the abbreviation

$$|j| := \sum_{i=1}^d j_i.$$

The following lemmata are crucial ingredients for the proof of Theorem 3.3.

LEMMA 5.1. (i) For any (β, L) with $0 < \beta_i \leq 2$, $p \in \mathcal{P}_d(\beta, L)$, and for any bandwidth $h = (h_1, \dots, h_d)$ with $h_i \leq c_8(\beta, L) p(t)^{1/\beta_i}$, $i = 1, \dots, d$ with

$$c_8(\beta, L) := \min_{i=1, \dots, d} \left(\frac{2dL}{\|K\|_2^2} \int |x_i|^{\beta_i} K^2(x) d\lambda^d(x) \right)^{-1/\beta_i},$$

the following inequality chain holds true

$$\frac{1}{2} \frac{\|K\|_2^2}{n \prod_{i=1}^d h_i} p(t) \leq \frac{1}{n} ((K_h)^2 * p)(t) \leq \frac{3}{2} \frac{\|K\|_2^2}{n \prod_{i=1}^d h_i} p(t).$$

(ii) For any constant $c_{10} > 0$, there exists a constant $c_{11}(\beta, L) = c_{11}(\beta, L, c_{10}) > 0$, such that for any (β, L) , $0 < \beta_i < \infty$, $i = 1, \dots, d$, and $p \in \mathcal{P}_d(\beta, L)$,

$$\frac{c_{11}(\beta, L)}{n \prod_{i=1}^d h_i} p(t) \leq \frac{1}{n} ((K_h)^2 * p)(t)$$

for every bandwidth $h = (h_1, \dots, h_d)$ with $h_i \leq c_{10} p(t)^{1/\beta_i}$, $i = 1, \dots, d$.

(iii) For any density p with isotropic Hölder smoothness (β, L) , $0 < \beta < \infty$ and bandwidth h , we have

$$\frac{1}{n}((K_h)^2 * p)(t) \leq \frac{L \|K\|_2^2}{nh^d} \left(h + \inf_{y \in \Gamma_p^c} \|t - y\|_2 \right)^\beta,$$

where K is a rotation invariant kernel supported on the closed Euclidean unit ball.

Lemma 5.1(ii) provides an extension of the results of Rohde (2008, 2011).

LEMMA 5.2. There exists some constant $c_{12}(\beta, L) > 0$, such that for any $p \in \mathcal{P}_d(\beta, L)$, $0 < \beta_i \leq 2$, $i = 1, \dots, d$, and $t \in \mathbb{R}^d$ the inequality

$$\sigma_{t, \text{trunc}}^2(j \wedge m) \leq c_{12}(\beta, L) (\sigma_{t, \text{trunc}}^2(j) \vee \sigma_{t, \text{trunc}}^2(m))$$

holds true for all (nonrandom) indices $j = (j_1, \dots, j_d)$ and $m = (m_1, \dots, m_d)$ with $j \geq \bar{j}$ componentwise. If additionally $m \geq j$ componentwise, then

$$\sigma_{t, \text{trunc}}^2(j) \leq c_{12}(\beta, L) \sigma_{t, \text{trunc}}^2(m).$$

The next lemma carefully analyzes the ratio of the truncated quantities $\sigma_{t, \text{trunc}}^2$ and $\tilde{\sigma}_{t, \text{trunc}}^2$.

LEMMA 5.3. For the quantities $\sigma_{t, \text{trunc}}^2(h)$ and $\tilde{\sigma}_{t, \text{trunc}}^2(h)$ defined in (3.6) and any $\eta \geq 0$ holds

$$\mathbb{P}^{\otimes n} \left(\left| \frac{\tilde{\sigma}_{t, \text{trunc}}^2(h)}{\sigma_{t, \text{trunc}}^2(h)} - 1 \right| \geq \eta \right) \leq 2 \exp \left(- \frac{3\eta^2}{2(3 + 2\eta) \|K\|_{\text{sup}}^2} \log^2 n \right).$$

LEMMA 5.4. For any (β, L) with $0 < \beta_i \leq 2$, $i = 1, \dots, d$, there exist constants $c_{13}(\beta, L)$ and $c_{14}(\beta, L) > 0$ such that for the multi-index \bar{j} as defined in (5.2) and the bias upper bound B_t as given in (5.3),

$$(5.4) \quad B_t(\bar{j}) \leq c_{13}(\beta, L) \sqrt{\sigma_{t, \text{trunc}}^2(\bar{j}) \log n},$$

$$(5.5) \quad \sqrt{\sigma_{t, \text{trunc}}^2(\bar{j})} \leq c_{14}(\beta, L) \left\{ \left(\frac{\log n}{n} \right)^{\bar{\beta}/(\bar{\beta}+1)} \vee \left(\frac{p(t) \log n}{n} \right)^{\bar{\beta}/(2\bar{\beta}+1)} \right\}.$$

LEMMA 5.5. For any (nonrandom) index $j = (j_1, \dots, j_d)$, the tail probabilities of the random variable

$$Y := \frac{\hat{p}_{n,j}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,j}(t)}{\sqrt{\sigma_{t, \text{trunc}}^2(j) \log n}},$$

are bounded by

$$\mathbb{P}^{\otimes n} (|Y| \geq \eta) \leq 2 \exp \left(- \frac{\log n}{4} \cdot (\eta^2 \wedge \eta) \right)$$

for any $\eta \geq 0$, any $t \in \mathbb{R}^d$ and $n \geq n_0$ with n_0 depending on $\|K\|_{\text{sup}}$ only.

LEMMA 5.6. *Let Z be some nonnegative random variable satisfying*

$$\mathbb{P}(Z \geq \eta) \leq 2 \exp(-A\eta)$$

for some $A > 0$. Then

$$(\mathbb{E}Z^m)^{1/m} \leq c_{15} \frac{m}{A}$$

for any $m \in \mathbb{N}$, where the constant c_{15} does not depend on A and m .

LEMMA 5.7 [Klutchnikoff (2005)]. *For all $k, l \in \mathcal{J}$, the absolute value of the difference of bias terms is bounded by*

$$|b_t(k \wedge l) - b_t(l)| \leq 2B_t(k)$$

for all $t \in \mathbb{R}^d$.

PROOF OF THEOREM 3.3. Recall the notation of Section 3 and denote $\hat{p}_{n,\hat{j}} = \hat{p}_n$. In a first step, the risk

$$\mathbb{E}_p^{\otimes n} |\hat{p}_{n,\hat{j}}(t) - p(t)|^r$$

is decomposed as follows:

$$\begin{aligned} & \mathbb{E}_p^{\otimes n} |\hat{p}_{n,\hat{j}}(t) - p(t)|^r \\ (5.6) \quad &= \mathbb{E}_p^{\otimes n} [|\hat{p}_{n,\hat{j}}(t) - p(t)|^r \cdot \mathbb{1}\{\hat{\sigma}_t^2(\hat{j}) \leq \hat{\sigma}_t^2(\bar{j})\}] \\ &+ \mathbb{E}_p^{\otimes n} [|\hat{p}_{n,\hat{j}}(t) - p(t)|^r \cdot \mathbb{1}\{\hat{\sigma}_t^2(\hat{j}) > \hat{\sigma}_t^2(\bar{j})\}] \\ &=: R^+ + R^-. \end{aligned}$$

We start with R^+ , which is decomposed again as follows:

$$\begin{aligned} (5.7) \quad R^+ &\leq 3^{r-1} (\mathbb{E}_p^{\otimes n} [|\hat{p}_{n,\hat{j}}(t) - \hat{p}_{n,\hat{j} \wedge \bar{j}}(t)|^r \cdot \mathbb{1}\{\hat{\sigma}_t^2(\hat{j}) \leq \hat{\sigma}_t^2(\bar{j})\}]) \\ &+ \mathbb{E}_p^{\otimes n} [|\hat{p}_{n,\hat{j} \wedge \bar{j}}(t) - \hat{p}_{n,\bar{j}}(t)|^r \cdot \mathbb{1}\{\hat{\sigma}_t^2(\hat{j}) \leq \hat{\sigma}_t^2(\bar{j})\}] \\ &+ \mathbb{E}_p^{\otimes n} [|\hat{p}_{n,\bar{j}}(t) - p(t)|^r \cdot \mathbb{1}\{\hat{\sigma}_t^2(\hat{j}) \leq \hat{\sigma}_t^2(\bar{j})\}] \\ &=: 3^{r-1} (S_1 + S_2 + S_3), \end{aligned}$$

where we used the inequality $(x + y + z)^r \leq 3^{r-1}(x^r + y^r + z^r)$ for all $x, y, z \geq 0$. This decomposition bears the advantage that only kernel density estimators with well-ordered bandwidths are compared. We focus on the estimation of S_1, S_2 and S_3 and start with S_2 using the selection scheme's construction. Clearly, $\hat{j} \in \mathcal{A}$ as

defined in (3.7). As a consequence, the following inequality holds true:

$$\begin{aligned}
 S_2 &\leq c_3^r \mathbb{E}_p^{\otimes n} \left[(\hat{\sigma}_t^2(\bar{j}) \log n)^{r/2} \cdot \mathbb{1} \left\{ \left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(\bar{j})}{\sigma_{t,\text{trunc}}^2(\bar{j})} - 1 \right| < 1 \right\} \right] \\
 &\quad + c_3^r \mathbb{E}_p^{\otimes n} \left[(\hat{\sigma}_t^2(\bar{j}) \log n)^{r/2} \cdot \mathbb{1} \left\{ \left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(\bar{j})}{\sigma_{t,\text{trunc}}^2(\bar{j})} - 1 \right| \geq 1 \right\} \right] \\
 &\leq 2^{r/2} c_3^r \left(\min \left\{ \sigma_{t,\text{trunc}}^2(\bar{j}), \frac{\|K\|_2^2 c_1}{n 2^{-|\bar{j}|}} \right\} \log n \right)^{r/2} \\
 &\quad + c_3^r \left(\frac{\|K\|_2^2 c_1}{n 2^{-|\bar{j}|}} \log n \right)^{r/2} \mathbb{P}^{\otimes n} \left(\left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(\bar{j})}{\sigma_{t,\text{trunc}}^2(\bar{j})} - 1 \right| \geq 1 \right),
 \end{aligned}$$

where we used the condition in the indicator function in the first summand to bound the estimated truncated variance $\tilde{\sigma}_{t,\text{trunc}}^2$ from above by $2\sigma_{t,\text{trunc}}^2$, and additionally the upper truncation level in the second summand. By the deviation inequality of Lemma 5.3, we can further estimate S_2 by

$$\begin{aligned}
 S_2 &\leq 2^{r/2} c_3^r (\sigma_{t,\text{trunc}}^2(\bar{j}) \log n)^{r/2} \\
 &\quad + c_3^r \left(\frac{\|K\|_2^2 c_1}{n 2^{-|\bar{j}|}} \log n \right)^{r/2} \cdot 2 \exp \left(-\frac{3}{10 \|K\|_{\text{sup}}^2} \log^2 n \right).
 \end{aligned}$$

The second term is always of smaller order than the first term because $2^{-|\bar{j}|} \leq 1$ and, therefore, for $n \geq 2$,

$$\left(\frac{\|K\|_2^2 c_1}{n 2^{-|\bar{j}|}} \log n \right)^{r/2} \cdot 2 \exp \left(-\frac{3}{10 \|K\|_{\text{sup}}^2} \log^2 n \right) \leq c \left(\frac{\log^3 n}{n^2 (2^{-|\bar{j}|})^2} \right)^{r/2}$$

for some constant c depending on c_1, r and the kernel K only. Finally,

$$S_2 \leq c(\beta, L) (\sigma_{t,\text{trunc}}^2(\bar{j}) \log n)^{r/2}.$$

We will now turn to S_3 , the third term in (5.7). We split the risk into bias and stochastic error. It holds

$$(5.8) \quad S_3 \leq \mathbb{E}_p^{\otimes n} (|\hat{p}_{n,\bar{j}}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,\bar{j}}(t)| + B_t(\bar{j}))^r$$

and by Lemma 5.4

$$(5.9) \quad B_t(\bar{j}) \leq c_{13}(\beta, L) \sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n}.$$

Denoting by

$$(5.10) \quad Z_k := \frac{\hat{p}_{n,k}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,k}(t)}{\sqrt{\sigma_{t,\text{trunc}}^2(k) \log n}} \quad \text{for } k \in \mathcal{J},$$

the decomposition (5.8), the bias variance relation (5.9) and the inequality $(x + y)^r \leq 2^{r-1}(x^r + y^r)$, $x, y \geq 0$ together with Lemma 5.6 yields

$$\begin{aligned} S_3 &\leq (\sigma_{t,\text{trunc}}^2(\bar{j}) \log n)^{r/2} \cdot \mathbb{E}_p^{\otimes n} (|Z_{\bar{j}}| + c_{13}(\beta, L))^r \\ &\leq (\sigma_{t,\text{trunc}}^2(\bar{j}) \log n)^{r/2} \cdot 2^{r-1} \mathbb{E}_p^{\otimes n} (|Z_{\bar{j}}|^r + c_{13}(\beta, L)^r) \\ &\leq c(\beta, L) (\sigma_{t,\text{trunc}}^2(\bar{j}) \log n)^{r/2}. \end{aligned}$$

It remains to show an analogous result for S_1 , the first term in (5.7). Clearly,

$$\begin{aligned} S_1 &\leq \sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} [(|\hat{p}_{n,j}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,j}(t)| + |\hat{p}_{n,j \wedge \bar{j}}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,j \wedge \bar{j}}(t)| \\ (5.11) \quad &+ |b_t(j \wedge \bar{j}) - b_t(j)|)^r \cdot \mathbb{1}\{\hat{\sigma}_t^2(j) \leq \hat{\sigma}_t^2(\bar{j}), \hat{j} = j\}]. \end{aligned}$$

By Lemmas 5.7 and 5.4,

$$|b_t(j \wedge \bar{j}) - b_t(j)| \leq 2B_t(\bar{j}) \leq 2c_{13}(\beta, L) \sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n}.$$

On account of this inequality and in view of (5.11), it suffices to bound the expectations in the following expression:

$$\begin{aligned} S_1 &\leq 3^{r-1} (\sigma_{t,\text{trunc}}^2(\bar{j}) \log n)^{r/2} \\ &\times \left\{ \sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} \left[\left(\frac{|\hat{p}_{n,j}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,j}(t)|}{\sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n}} \right)^r \mathbb{1}\{\hat{\sigma}_t^2(j) \leq \hat{\sigma}_t^2(\bar{j}), \hat{j} = j\} \right] \right. \\ (5.12) \quad &+ \sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} \left[\left(\frac{|\hat{p}_{n,j \wedge \bar{j}}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,j \wedge \bar{j}}(t)|}{\sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n}} \right)^r \mathbb{1}\{\hat{\sigma}_t^2(j) \leq \hat{\sigma}_t^2(\bar{j}), \hat{j} = j\} \right] \\ &\left. + \sum_{j \in \mathcal{J}} 2^r c_{13}(\beta, L)^r \cdot \mathbb{P}^{\otimes n}(\hat{j} = j) \right\}. \end{aligned}$$

Denoting

$$(5.13) \quad A_{j,\bar{j}} := \left\{ \left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(j)}{\sigma_{t,\text{trunc}}^2(j)} - 1 \right| < \frac{1}{2} \text{ and } \left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(\bar{j})}{\sigma_{t,\text{trunc}}^2(\bar{j})} - 1 \right| < \frac{1}{2} \right\},$$

it follows

$$\begin{aligned} &\sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} \left[\left(\frac{|\hat{p}_{n,j}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,j}(t)|}{\sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n}} \right)^r \mathbb{1}\{\hat{\sigma}_t^2(j) \leq \hat{\sigma}_t^2(\bar{j}), \hat{j} = j\} \right] \\ &= \sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} \left[\left(\frac{|\hat{p}_{n,j}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,j}(t)|}{\sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n}} \right)^r \mathbb{1}\{\hat{\sigma}_t^2(j) \leq \hat{\sigma}_t^2(\bar{j}), \hat{j} = j\} \cdot \mathbb{1}_{A_{j,\bar{j}}} \right] \end{aligned}$$

$$+ \sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} \left[\left(\frac{|\hat{p}_{n,j}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,j}(t)|}{\sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n}} \right)^r \mathbb{1}\{\hat{\sigma}_t^2(j) \leq \hat{\sigma}_t^2(\bar{j}), \hat{j} = j\} \cdot \mathbb{1}_{A_{j,\bar{j}}^c} \right]$$

$$=: S_{1,1} + S_{1,2}.$$

Applying Lemma 5.6 and Hölder’s inequality for any $p > 1$,

$$\begin{aligned} S_{1,1} &\leq \left(\frac{3(1 \vee c_1 \|K\|_2^2)}{c_1 \|K\|_2^2} \right)^{r/2} \sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} [|Z_j|^r \cdot \mathbb{1}\{\hat{j} = j\}] \\ &\leq \left(\frac{3(1 \vee c_1 \|K\|_2^2)}{c_1 \|K\|_2^2} \right)^{r/2} \left(1 + \sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} [|Z_j|^r \mathbb{1}\{|Z_j| \geq 1\} \mathbb{1}\{\hat{j} = j\}] \right) \\ &\leq \left(\frac{3(1 \vee c_1 \|K\|_2^2)}{c_1 \|K\|_2^2} \right)^{r/2} \\ &\quad \times \left(1 + \sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} [|Z_j|^{rp} \mathbb{1}\{|Z_j| \geq 1\}]^{1/p} \cdot \mathbb{P}(\hat{j} = j)^{(p-1)/p} \right) \\ &\leq \left(\frac{3(1 \vee c_1 \|K\|_2^2)}{c_1 \|K\|_2^2} \right)^{r/2} \left(1 + c_{15}^r \left(\frac{8rp}{\log n} \right)^r \sum_{j \in \mathcal{J}} \mathbb{P}(\hat{j} = j)^{(p-1)/p} \right) \\ &\leq \left(\frac{3(1 \vee c_1 \|K\|_2^2)}{c_1 \|K\|_2^2} \right)^{r/2} \left(1 + c_{15}^r \left(\frac{8rp}{\log n} \right)^r \left(\sum_{j \in \mathcal{J}} \mathbb{P}(\hat{j} = j) \right)^{(p-1)/p} \cdot |\mathcal{J}|^{1/p} \right). \end{aligned}$$

By the constraint $2^{-|j|} \geq \log^2 n/n$ for any $j \in \mathcal{J}$, there exists some constant $c > 0$ such that $|\mathcal{J}| \leq c(\log n)^d$. Setting finally $p = d \log n$, yields $S_{1,1} \leq c(\beta^*, L^*)$. As concerns $S_{1,2}$, by the Cauchy–Schwarz inequality,

$$\begin{aligned} S_{1,2} &\leq \sum_{j \in \mathcal{J}} \left(\frac{\sigma_{t,\text{trunc}}^2(j)}{\sigma_{t,\text{trunc}}^2(\bar{j})} \right)^{r/2} \mathbb{E}_p^{\otimes n} [|Z_j|^r \mathbb{1}\{\hat{j} = j\} \mathbb{1}_{A_{j,\bar{j}}^c}] \\ &\leq \sum_{j \in \mathcal{J}} \left(\frac{\sigma_{t,\text{trunc}}^2(j)}{\sigma_{t,\text{trunc}}^2(\bar{j})} \right)^{r/2} \mathbb{E}_p^{\otimes n} [|Z_j|^{2r} \mathbb{1}\{\hat{j} = j\}]^{1/2} \\ &\quad \times \left\{ \mathbb{P}^{\otimes n} \left(\left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(j)}{\sigma_{t,\text{trunc}}^2(j)} - 1 \right| \geq \frac{1}{2} \right) + \mathbb{P}^{\otimes n} \left(\left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(\bar{j})}{\sigma_{t,\text{trunc}}^2(\bar{j})} - 1 \right| \geq \frac{1}{2} \right) \right\}^{1/2}. \end{aligned}$$

Via the lower and upper truncation levels in the definition of $\sigma_{t,\text{trunc}}^2$,

$$(5.14) \quad \frac{\sigma_{t,\text{trunc}}^2(k)}{\sigma_{t,\text{trunc}}^2(l)} \leq \frac{(1 \vee c_1 \|K\|_2^2)n^2}{\log^4 n} \quad \text{for any } k, l \in \mathcal{J},$$

and the remaining expectation $\sum_{j \in \mathcal{J}} \mathbb{E}_p^{\otimes n} [|Z_j|^{2r} \mathbb{1}\{\hat{j} = j\}]$ can be bounded by Lemma 5.6 as above. Finally, the probabilities compensate (5.14) by Lemma 5.3.

As concerns the expectation in (5.12), we proceed analogously using

$$\sigma_{t,\text{trunc}}^2(j \wedge \bar{j}) \leq c_{12}(\beta, L)(\sigma_{t,\text{trunc}}^2(\bar{j}) \vee \sigma_{t,\text{trunc}}^2(j))$$

by Lemma 5.2 and $\sigma_{t,\text{trunc}}^2(j) \leq c(\beta, L)\sigma_{t,\text{trunc}}^2(\bar{j})$ on $A_{j,\bar{j}} \cap \{\hat{\sigma}_t^2(j) \leq \hat{\sigma}_t^2(\bar{j})\}$. Combining the results for S_1, S_2 and S_3 proves that R^+ as defined in (5.6) is estimated by

$$R^+ \leq c(\beta, L)(\sigma_{t,\text{trunc}}^2(\bar{j}) \log n)^{r/2}.$$

To deduce a similar inequality for R^- , it remains to investigate the probability

$$\mathbb{P}^{\otimes n}(\hat{\sigma}_t^2(\hat{j}) > \hat{\sigma}_t^2(\bar{j})),$$

since \hat{p}_n and p are both upper bounded by c_1 . If $\hat{\sigma}_t^2(\hat{j}) > \hat{\sigma}_t^2(\bar{j})$, then \bar{j} cannot be an admissible exponent [see (3.7)], because \hat{j} had not been chosen in the minimization problem (3.9) otherwise. Hence, by definition there exists a multi-index $m \in \mathcal{J}$ with $\hat{\sigma}_t^2(m) \geq \hat{\sigma}_t^2(\bar{j})$ such that

$$|\hat{p}_{n,\bar{j} \wedge m}(t) - \hat{p}_{n,m}(t)| > c_3 \sqrt{\hat{\sigma}_t^2(m) \log n}.$$

Subsuming, we get

$$\begin{aligned} &\mathbb{P}^{\otimes n}(\hat{\sigma}_t^2(\hat{j}) > \hat{\sigma}_t^2(\bar{j})) \\ &\leq \sum_{m \in \mathcal{J}} \mathbb{P}^{\otimes n}(|\hat{p}_{n,\bar{j} \wedge m}(t) - \hat{p}_{n,m}(t)| > c_3 \sqrt{\hat{\sigma}_t^2(m) \log n}, \hat{\sigma}_t^2(m) \geq \hat{\sigma}_t^2(\bar{j})), \end{aligned}$$

and we divide the absolute value of the difference of the kernel density estimators as in (5.11) into the difference of biases $|b_t(\bar{j} \wedge m) - b_t(m)|$ and two stochastic terms $|\hat{p}_{n,\bar{j} \wedge m}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,\bar{j} \wedge m}(t)|$ and $|\hat{p}_{n,m}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,m}(t)|$. As before,

$$|b_t(\bar{j} \wedge m) - b_t(m)| \leq 2B_t(\bar{j}) \leq 2c_{13}(\beta, L) \sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n}$$

by Lemmas 5.7 and 5.4, leading to the inequality

$$\begin{aligned} &\mathbb{P}^{\otimes n}(\hat{\sigma}_t^2(\hat{j}) > \hat{\sigma}_t^2(\bar{j})) \\ &\leq \sum_{m \in \mathcal{J}} \mathbb{P}^{\otimes n}(|\hat{p}_{n,\bar{j} \wedge m}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,\bar{j} \wedge m}(t)| + |\hat{p}_{n,m}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,m}(t)| \\ &\quad > c_3 \sqrt{\hat{\sigma}_t^2(m) \log n} - 2c_{13}(\beta, L) \sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n}, \hat{\sigma}_t^2(m) \geq \hat{\sigma}_t^2(\bar{j})) \\ &\leq \sum_{m \in \mathcal{J}} (\mathbb{P}^{\otimes n}(B_{1,m}) + \mathbb{P}^{\otimes n}(B_{2,m})) \end{aligned}$$

with

$$\begin{aligned}
 B_{1,m} &:= \left\{ \left| \hat{p}_{n,\bar{j}\wedge m}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,\bar{j}\wedge m}(t) \right| \right. \\
 &\quad \left. > \frac{1}{2} \left(c_3 \sqrt{\hat{\sigma}_t^2(m) \log n} - 2c_{13}(\beta, L) \sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n} \right), \hat{\sigma}_t^2(m) \geq \hat{\sigma}_t^2(\bar{j}) \right\}, \\
 B_{2,m} &:= \left\{ \left| \hat{p}_{n,m}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,m}(t) \right| \right. \\
 &\quad \left. > \frac{1}{2} \left(c_3 \sqrt{\hat{\sigma}_t^2(m) \log n} - 2c_{13}(\beta, L) \sqrt{\sigma_{t,\text{trunc}}^2(\bar{j}) \log n} \right), \hat{\sigma}_t^2(m) \geq \hat{\sigma}_t^2(\bar{j}) \right\}.
 \end{aligned}$$

To start with the second probability, we intersect event $B_{2,m}$ with $A_{m,\bar{j}}$ as defined in (5.13). Obviously,

$$\mathbb{P}^{\otimes n}(B_{2,m}) \leq \mathbb{P}^{\otimes n}(B_{2,m} \cap A_{m,\bar{j}}) + \mathbb{P}^{\otimes n}(A_{m,\bar{j}}^c).$$

The definition of c_3 and Lemma 5.5 allow to bound the probability

$$\begin{aligned}
 \mathbb{P}^{\otimes n}(B_{2,m} \cap A_{m,\bar{j}}) &\leq \mathbb{P}^{\otimes n} \left(\frac{|\hat{p}_{n,m}(t) - \mathbb{E}_p^{\otimes n} \hat{p}_{n,m}(t)|}{\sqrt{\sigma_{t,\text{trunc}}^2(m) \log n}} > c_{16}(\beta, L) \right) \\
 (5.15) \qquad \qquad \qquad &\leq 2 \exp \left(- \frac{c_{16}(\beta, L)^2 \wedge c_{16}(\beta, L)}{4} \log n \right)
 \end{aligned}$$

with

$$(5.16) \quad c_{16}(\beta, L) := \left(\frac{c_3}{2} - c_{13}(\beta, L) \sqrt{2 \frac{1 \vee c_1 \|K\|_2^2}{c_1 \|K\|_2^2}} \right) \cdot \sqrt{\frac{1}{2} \frac{c_1 \|K\|_2^2}{1 \vee c_1 \|K\|_2^2}}.$$

At this point, we specify a lower bound on c_3 . Precisely, c_3 has to be chosen large enough to guarantee that

$$(5.17) \quad \frac{c_{16}(\beta, L)^2 \wedge c_{16}(\beta, L)}{4} \geq \frac{r \bar{\beta}}{\bar{\beta} + 1} + 1$$

for any β in the range of adaptation. Finally, by means of Lemma 5.3,

$$\begin{aligned}
 &\mathbb{P}^{\otimes n}(A_{m,\bar{j}}^c) \\
 (5.18) \quad &\leq \mathbb{P}^{\otimes n} \left(\left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(\bar{j})}{\sigma_{t,\text{trunc}}^2(\bar{j})} - 1 \right| \geq \frac{1}{2} \right) + \mathbb{P}^{\otimes n} \left(\left| \frac{\tilde{\sigma}_{t,\text{trunc}}^2(m)}{\sigma_{t,\text{trunc}}^2(m)} - 1 \right| \geq \frac{1}{2} \right) \\
 &\leq 4 \exp \left(- \frac{3}{32 \|K\|_{\text{sup}}^2} \log^2 n \right),
 \end{aligned}$$

which is of smaller order than the bound in (5.15). Altogether, with this restriction on c_3 ,

$$\mathbb{P}^{\otimes n}(B_{2,m}) \leq c(\beta, L)(\sigma_{t,\text{trunc}}^2(\bar{j}) \log n)^{r/2}.$$

By Lemma 5.2, the probability $\mathbb{P}^{\otimes n}(B_{1,m})$ can be bounded in the same way using additionally

$$\sigma_{t,\text{trunc}}^2(\bar{j} \wedge m) \leq c_{12}(\beta, L)(\sigma_{t,\text{trunc}}^2(\bar{j}) \vee \sigma_{t,\text{trunc}}^2(m)) = c(\beta, L)\sigma_{t,\text{trunc}}^2(m),$$

because $\sigma_{t,\text{trunc}}^2(\bar{j}) \leq c(\beta, L)\sigma_{t,\text{trunc}}^2(m)$ on the event $A_{m,\bar{j}} \cap \{\hat{\sigma}_t^2(m) \geq \hat{\sigma}_t^2(\bar{j})\}$. Summarizing,

$$(5.19) \quad \mathbb{P}^{\otimes n}(\hat{\sigma}_t^2(\hat{j}) > \hat{\sigma}_t^2(\bar{j})) \leq c(\beta, L)(\sigma_{t,\text{trunc}}^2(\bar{j}) \log n)^{r/2}.$$

Finally, by Lemma 5.4,

$$\begin{aligned} & (\mathbb{E}_p^{\otimes n} |\hat{p}_{n,\hat{j}}(t) - p(t)|^r)^{1/r} \\ & \leq c(\beta, L) \left\{ \left(\frac{\log n}{n} \right)^{\bar{\beta}/(\bar{\beta}+1)} \vee \left(\frac{p(t) \log n}{n} \right)^{\bar{\beta}/(2\bar{\beta}+1)} \right\} \sqrt{\log n}. \end{aligned}$$

This completes the proof of Theorem 3.3. \square

PROOF OF THEOREM 3.4. Before we construct the densities p_n and q_n , we first specify their amplitudes Δ_n and δ_n in t , respectively. Let

$$\begin{aligned} (5.20) \quad \Delta_n & := n^{-\beta_1/(\beta_1+1)} \cdot \varrho(n), \\ \delta_n & := 4c_4(\beta_1^*, L_1^*, r) \left(\frac{\Delta_n}{n} \right)^{\beta_1/(2\beta_1+1)} (\log n)^{3/2} \\ & = 4c_4(\beta_1^*, L_1^*, r) \Delta_n \cdot \varrho(n)^{-(\beta_1+1)/(2\beta_1+1)} (\log n)^{3/2}, \end{aligned}$$

for

$$\varrho(n) := n^{\frac{\beta_1 - \beta_2}{(\beta_1+1)(\beta_2+1)}}$$

converging to infinity. Note first that with this choice of $\varrho(n)$ it holds that $\Delta_n = n^{-\beta_2/(\beta_2+1)}$, and hence tends to zero as n goes to infinity. The amplitude δ_n is smaller than Δ_n for sufficiently large n , and hence also tends to zero. Furthermore, it holds

$$\delta_n = 4c_4(\beta_1^*, L_1^*, r) \cdot n^{-\frac{\beta_2}{\beta_2+1}} \cdot n^{\frac{\beta_2 - \beta_1}{(2\beta_1+1)(\beta_2+1)}} \cdot (\log n)^{3/2} = o\left(n^{-\frac{\beta_2}{\beta_2+1}}\right).$$

Denote by $K(\cdot; \beta_i)$, $i = 1, 2$ the univariate, symmetric and nonnegative functions to the Hölder exponent β_i , respectively, as defined in the supplemental article [Patschkowski and Rohde (2015)], Section A.4, normalized by appropriate choices of $c_{17}(\beta_i)$ such that both functions integrate to one. Let $\tilde{L}_i = \tilde{L}_i(\beta_i)$, $i = 1, 2$ be

such that $K(\cdot; \beta_i) \in \mathcal{P}_1(\beta_i, \tilde{L}_i)$. Note that $K(\cdot; h, \beta_i) := h^{\beta_i} K(\cdot/h; \beta_i)$ has the same Hölder regularity as K [as opposed to $K_h(\cdot; \beta_i) := h^{-1} K(\cdot/h; \beta_i)$, which has the same Hölder parameter β_i but not necessarily the same \tilde{L}_i].

To ensure that $p_n(t) = \Delta_n$ we use the scaled version $K(\cdot - t; g_{1,n}, \beta_1)$ for some bandwidth $g_{1,n}$ defined below, preserving the Hölder regularity. In order to re-establish integrability to one, a second part is added alongside. The density q_n is then defined as p_n with a perturbation added and subtracted around t , that is,

$$p_n(x) = K(x - t; g_{1,n}, \beta_1) + K(x - t - g_{1,n} - g_{2,n}; g_{2,n}, \beta_1) \in \mathcal{P}_1(\beta_1, L_1),$$

$$q_n(x) = p_n(x) - K(x - t; h_n, \beta_2) + K(x - t - 2h_n; h_n, \beta_2) \in \mathcal{P}_1(\beta_2, L_2),$$

with

$$g_{1,n} := \left(\frac{\Delta_n}{K(0; \beta_1)} \right)^{1/\beta_1},$$

$$g_{2,n} := (1 - g_{1,n}^{\beta_1+1})^{1/(\beta_1+1)},$$

$$h_n := \left(\frac{\Delta_n - \delta_n}{K(0; \beta_2)} \right)^{1/\beta_2}$$

and suitable constants L_1 and L_2 independent of n . The construction of the hypotheses is depicted in Figure 3. Recall that the particular construction of $K(\cdot; h, \beta)$ does not change the Hölder parameters and note that the classes $\bigcup_{L>0} \mathcal{C}_c \cap \mathcal{P}_1(\beta, L)$, $0 < \beta \leq 2$, are nested (\mathcal{C}_c denotes the set of continuous functions from \mathbb{R} to \mathbb{R} of compact support). The bandwidth $g_{1,n}$ tends to zero, and hence $g_{2,n}$ converges to one. In particular, $g_{2,n}$ is positive for sufficiently large n . In turn, h_n ensures that $q_n(t) = \delta_n$. Note furthermore that $\Delta_n > \Delta_n - \delta_n$ and $K(0; \beta_1) < K(0; \beta_2)$ since the constant $c_{17}(\beta)$ is monotonously increasing in β and $\beta_2 < \beta_1$. Thus, h_n is smaller than $g_{1,n}$ and consequently q_n is nonnegative for sufficiently large n .

Let $T_n(t)$ be an arbitrary estimator with property (3.10). Note first that we can pass on to the consideration of the estimator

$$\tilde{T}_n(t) := T_n(t) \cdot \mathbb{1}\{T_n(t) \leq 2\Delta_n\},$$

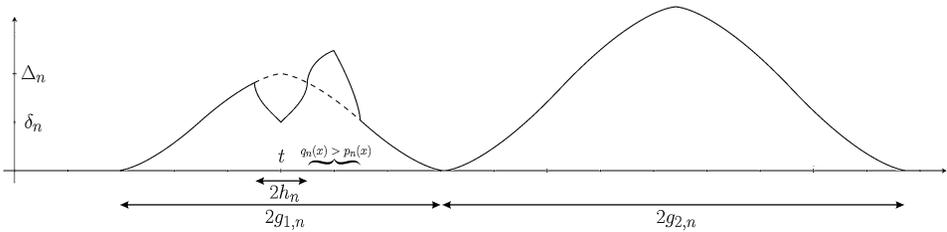


FIG. 3. Construction of p_n (dashed line) and q_n (solid line).

since it both improves the quality of estimation of $p_n(t)$ and $q_n(t)$: Obviously,

$$\begin{aligned} \mathbb{E}_{p_n}^{\otimes n} |\tilde{T}_n(t) - p_n(t)| &= \mathbb{E}_{p_n}^{\otimes n} [p_n(t) \cdot \mathbb{1}\{T_n(t) - p_n(t) > p_n(t)\}] \\ &\quad + \mathbb{E}_{p_n}^{\otimes n} [|T_n(t) - p_n(t)| \cdot \mathbb{1}\{T_n(t) - p_n(t) \leq p_n(t)\}] \\ &\leq \mathbb{E}_{p_n}^{\otimes n} |T_n(t) - p_n(t)| \end{aligned}$$

and because of $q_n(t) \leq p_n(t)$ also

$$\mathbb{E}_{q_n}^{\otimes n} |\tilde{T}_n(t) - q_n(t)| \leq \mathbb{E}_{q_n}^{\otimes n} |T_n(t) - q_n(t)|.$$

As in the proof of the constrained risk inequality in Cai, Low and Zhao (2007), by reverse triangle inequality holds

$$\mathbb{E}_{q_n}^{\otimes n} |\tilde{T}_n(t) - q_n(t)| \geq (\Delta_n - \delta_n) - \mathbb{E}_{q_n}^{\otimes n} |\tilde{T}_n(t) - p_n(t)|.$$

In contrast to their proof, we need the decomposition:

$$\begin{aligned} &\mathbb{E}_{q_n}^{\otimes n} |\tilde{T}_n(t) - q_n(t)| \\ (5.21) \quad &\geq (\Delta_n - \delta_n) - \mathbb{E}_{q_n}^{\otimes n} [|T_n(t) - p_n(t)| \mathbb{1}_{B_n}] \\ &\quad - \mathbb{E}_{q_n}^{\otimes n} [|\tilde{T}_n(t) - p_n(t)| \mathbb{1}_{B_n^c}] \\ &=: (\Delta_n - \delta_n) - S_1 - S_2, \end{aligned}$$

where

$$B_n := \left\{ x = (x_1, \dots, x_n) \in \mathbb{R}^n : \prod_{i=1}^n \frac{q_n(x_i)}{p_n(x_i)} \leq \frac{\Delta_n}{\delta_n} \right\}.$$

By definition of Δ_n and δ_n in (5.20) and the risk bound (3.10) the first two summands in (5.21) can be further estimated by

$$\begin{aligned} &(\Delta_n - \delta_n) - S_1 \\ &\geq (\Delta_n - \delta_n) - \mathbb{E}_{p_n}^{\otimes n} |T_n(t) - p_n(t)| \cdot \frac{\Delta_n}{\delta_n} \\ &\geq (\Delta_n - \delta_n) \left(1 - \frac{c_4(\beta_1^*, L_1^*, r)(\Delta_n/n)^{\beta_1/(2\beta_1+1)}(\log n)^{3/2}(\Delta_n/\delta_n)}{\Delta_n - \delta_n} \right) \\ &= \delta_n \left(\frac{\varrho(n)^{(\beta_1+1)/(2\beta_1+1)}(\log n)^{-3/2}}{4c_4(\beta_1^*, L_1^*, r)} - 1 \right) \\ &\quad \times \left(1 - \frac{c_4(\beta_1^*, L_1^*, r)(\Delta_n/n)^{\beta_1/(2\beta_1+1)}(\log n)^{3/2}(\Delta_n/\delta_n)}{\Delta_n(1 - 4c_4(\beta_1^*, L_1^*, r) \cdot \varrho(n)^{-(\beta_1+1)/(2\beta_1+1)}(\log n)^{3/2})} \right), \end{aligned}$$

which is lower bounded by

$$\begin{aligned} (\Delta_n - \delta_n) - S_1 &\geq \delta_n \frac{\varrho(n)^{(\beta_1+1)/(2\beta_1+1)} (\log n)^{-3/2}}{8c_4(\beta_1^*, L_1^*, r)} \\ &\quad \times \left(1 - \frac{2c_4(\beta_1^*, L_1^*, r) (\Delta_n/n)^{\beta_1/(2\beta_1+1)} (\log n)^{3/2}}{\delta_n} \right) \\ &= \delta_n \frac{\varrho(n)^{(\beta_1+1)/(2\beta_1+1)} (\log n)^{-3/2}}{16c_4(\beta_1^*, L_1^*, r)} \end{aligned}$$

for sufficiently large n . Furthermore,

$$S_2 \leq 2\Delta_n \cdot \mathbb{Q}_n^{\otimes n}(B_n^c) = \delta_n \frac{\varrho(n)^{(\beta_1+1)/(2\beta_1+1)} (\log n)^{-3/2}}{2c_4(\beta_1^*, L_1^*, r)} \cdot \mathbb{Q}_n^{\otimes n}(B_n^c),$$

and it remains to show that $\mathbb{Q}_n^{\otimes n}(B_n^c)$ tends to zero. By Markov's inequality,

$$\begin{aligned} \mathbb{Q}_n^{\otimes n}(B_n^c) &= \mathbb{Q}_n^{\otimes n} \left(\prod_{i=1}^n \frac{q_n(X_i)}{p_n(X_i)} > \frac{\Delta_n}{\delta_n} \right) \\ &\leq \frac{\delta_n}{\Delta_n} \left(\mathbb{E}_{q_n} \frac{q_n(X_1)}{p_n(X_1)} \right)^n \\ &\leq \frac{\delta_n}{\Delta_n} \left(1 + \int \frac{q_n(x)}{p_n(x)} q_n(x) \mathbb{1}_{\{q_n(x) > p_n(x)\}} dx \right)^n \\ &\leq \frac{\delta_n}{\Delta_n} \left(1 + \frac{(2\Delta_n - \delta_n)^2}{K(3h_n; g_{1,n}, \beta_1)} \cdot 2h_n \right)^n \\ &\leq \frac{\delta_n}{\Delta_n} \left(1 + \frac{4\Delta_n^2}{g_{1,n}^{\beta_1} K(3h_n/g_{1,n}; \beta_1)} \cdot 2h_n \right)^n \\ &\leq \frac{\delta_n}{\Delta_n} (1 + c(\beta_1, \beta_2) \Delta_n^{(\beta_2+1)/\beta_2})^n \end{aligned}$$

for sufficiently large n , where the last inequality is due to

$$h_n/g_{1,n} = c(\beta_1, \beta_2) \Delta_n^{(\beta_1-\beta_2)/(\beta_1\beta_2)} \longrightarrow 0,$$

that is, $K(3h_n/g_{1,n}; \beta_1)$ stays uniformly bounded away from zero. Finally,

$$\begin{aligned} \mathbb{Q}_n^{\otimes n}(B_n^c) &\leq \frac{\delta_n}{\Delta_n} \exp(n \log(1 + c(\beta_1, \beta_2) \Delta_n^{(\beta_2+1)/\beta_2})) \\ &\leq \frac{\delta_n}{\Delta_n} \exp(n \cdot c(\beta_1, \beta_2) \Delta_n^{(\beta_2+1)/\beta_2}) \end{aligned}$$

and

$$n \Delta_n^{(\beta_2+1)/\beta_2} = 1,$$

such that $\mathbb{Q}_n^{\otimes n}(B_n^c) \leq c(\beta_1, \beta_2) \cdot \delta_n/\Delta_n \longrightarrow 0$. \square

Acknowledgements. We are very grateful to two anonymous referees and an Associate Editor for three constructive and detailed reports which led to a substantial improvement of our presentation and stimulated further interesting research.

SUPPLEMENTARY MATERIAL

Supplement to “Adaptation to lowest density regions with application to support recovery” (DOI: [10.1214/15-AOS1366SUPP](https://doi.org/10.1214/15-AOS1366SUPP); .pdf). Supplement A is organized as follows. Section A.1 contains the proofs of Lemmas 5.1–5.6, which are central ingredients for the proof of Theorem 3.3. Section A.2 is concerned with the remaining proofs of Section 3. Section A.3 contains the proofs of Section 4. Section A.4 introduces a specific construction of a kernel function with prescribed Hölder regularity, which is frequently used throughout the article.

REFERENCES

- BAÍLLO, A., CUEVAS, A. and JUSTEL, A. (2000). Set estimation and nonparametric detection. *Canad. J. Statist.* **28** 765–782. [MR1821433](#)
- BERTIN, K., LACOUR, C. and RIVOIRARD, V. (2014). Adaptive pointwise estimation of conditional density function. Available at [arXiv:1312.7402](https://arxiv.org/abs/1312.7402).
- BHATTACHARYA, R. N. and RANGA RAO, R. (1976). *Normal Approximation and Asymptotic Expansions*. Wiley, New York. [MR0436272](#)
- BIAU, G., CADRE, B. and PELLETIER, B. (2008). Exact rates in density support estimation. *J. Multivariate Anal.* **99** 2185–2207. [MR2463383](#)
- BIAU, G., CADRE, B., MASON, D. M. and PELLETIER, B. (2009). Asymptotic normality in density support estimation. *Electron. J. Probab.* **14** 2617–2635. [MR2570013](#)
- BIRGÉ, L. (2014). Model selection for density estimation with \mathbb{L}_2 -loss. *Probab. Theory Related Fields* **158** 533–574. [MR3176358](#)
- BRUNEL, V.-E. (2013). Adaptive estimation of convex and polytopal density support. *Probab. Theory Related Fields*. To appear. Available at [arXiv:1309.6602](https://arxiv.org/abs/1309.6602).
- BUTUCEA, C. (2001). Exact adaptive pointwise estimation on Sobolev classes of densities. *ESAIM Probab. Stat.* **5** 1–31 (electronic). [MR1845320](#)
- CAI, T. T., LOW, M. G. and ZHAO, L. H. (2007). Trade-offs between global and local risks in nonparametric function estimation. *Bernoulli* **13** 1–19. [MR2307391](#)
- CHAVEL, I. (2001). *Isoperimetric Inequalities: Differential Geometric and Analytic Perspectives*. *Cambridge Tracts in Mathematics* **145**. Cambridge Univ. Press, Cambridge. [MR1849187](#)
- CHEVALIER, J. (1976). Estimation du support et du contour du support d’une loi de probabilité. *Ann. Inst. H. Poincaré Sect. B (N.S.)* **12** 339–364. [MR0451491](#)
- CHICHIGNOUD, M. (2012). Minimax and minimax adaptive estimation in multiplicative regression: Locally Bayesian approach. *Probab. Theory Related Fields* **153** 543–586. [MR2948686](#)
- CHICHIGNOUD, M. and LEDERER, J. (2014). A robust, adaptive M-estimator for pointwise estimation in heteroscedastic regression. *Bernoulli* **20** 1560–1599. [MR3217454](#)
- CHOLAQUIDIS, A., CUEVAS, A. and FRAIMAN, R. (2014). On Poincaré cone property. *Ann. Statist.* **42** 255–284. [MR3189486](#)
- CUEVAS, A. (1990). On pattern analysis in the nonconvex case. *Kybernetes* **19** 26–33. [MR1084947](#)
- CUEVAS, A. and FRAIMAN, R. (1997). A plug-in approach to support estimation. *Ann. Statist.* **25** 2300–2312. [MR1604449](#)
- CUEVAS, A. and RODRÍGUEZ-CASAL, A. (2004). On boundary estimation. *Adv. in Appl. Probab.* **36** 340–354. [MR2058139](#)

- DATTNER, I., REISS, M. and TRABS, M. (2014). Adaptive quantile estimation in deconvolution with unknown error distribution. *Bernoulli*. To appear.
- DEVROYE, L. and WISE, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.* **38** 480–488. [MR0579432](#)
- EFROMOVICH, S. (2008). Adaptive estimation of and oracle inequalities for probability densities and characteristic functions. *Ann. Statist.* **36** 1127–1155. [MR2418652](#)
- GACH, F., NICKL, R. and SPOKOINY, V. (2013). Spatially adaptive density estimation by localised Haar projections. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 900–914. [MR3112439](#)
- GAYRAUD, G. (1997). Estimation of functionals of density support. *Math. Methods Statist.* **6** 26–46. [MR1456645](#)
- GEFFROY, J. (1964). Sur un problème d'estimation géométrique. *Publ. Inst. Statist. Univ. Paris* **13** 191–210. [MR0202237](#)
- GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. [MR2604707](#)
- GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632. [MR2850214](#)
- GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on R^d . *Probab. Theory Related Fields* **159** 479–543. [MR3230001](#)
- GRENDER, U. (1981). *Abstract Inference*. Wiley, New York. [MR0599175](#)
- GROENEBOOM, P. (1988). Limit theorems for convex hulls. *Probab. Theory Related Fields* **79** 327–368. [MR0959514](#)
- HALL, P. (1982). On estimating the endpoint of a distribution. *Ann. Statist.* **10** 556–568. [MR0653530](#)
- HALL, P., NUSSBAUM, M. and STERN, S. E. (1997). On the estimation of a support curve of indeterminate sharpness. *J. Multivariate Anal.* **62** 204–232. [MR1473874](#)
- HÄRDLE, W., PARK, B. U. and TSYBAKOV, A. B. (1995). Estimation of non-sharp support boundaries. *J. Multivariate Anal.* **55** 205–218. [MR1370400](#)
- JIRAK, M., MEISTER, A. and REISS, M. (2014). Adaptive function estimation in nonparametric regression with one-sided errors. *Ann. Statist.* **42** 1970–2002. [MR3262474](#)
- JUDITSKY, A. and LAMBERT-LACROIX, S. (2004). On minimax density estimation on \mathbb{R} . *Bernoulli* **10** 187–220. [MR2046772](#)
- KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121** 137–170. [MR1863916](#)
- KLEMELÄ, J. (2004). Complexity penalized support estimation. *J. Multivariate Anal.* **88** 274–297. [MR2025614](#)
- KLUTCHNIKOFF, N. (2005). Sur l'estimation adaptative de fonctions anisotropes. Ph.D. Thesis, Univ. Aix-Marseille I.
- KOROSTELÉV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction. Lecture Notes in Statistics* **82**. Springer, New York. [MR1226450](#)
- LEPSKI, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 459–470. [MR1091202](#)
- LEPSKI, O. (2013). Multivariate density estimation under sup-norm loss: Oracle approach, adaptation and independence structure. *Ann. Statist.* **41** 1005–1034. [MR3099129](#)
- LEPSKI, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *Ann. Statist.* **43** 1178–1242. [MR3346701](#)
- LIU, L. and WONG, W. H. (2014). Multivariate density estimation based on adaptive partitioning: Convergence rate, variable selection and spatial adaptation. Available at [arXiv:1401.2597](#).
- MAMMEN, E. and TSYBAKOV, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502–524. [MR1332579](#)
- MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. [MR1765618](#)

- NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430. [MR1425959](#)
- PATSCHKOWSKI, T. and ROHDE, A. (2015). Supplement to “Adaptation to lowest density regions with application to support recovery.” DOI:10.1214/15-AOS1366SUPP.
- POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—An excess mass approach. *Ann. Statist.* **23** 855–881. [MR1345204](#)
- RÉNYI, A. and SULANKE, R. (1963). Über die konvexe Hülle von n zufällig gewählten Punkten. *Z. Wahrsch. Verw. Gebiete* **2** 75–84. [MR0156262](#)
- RÉNYI, A. and SULANKE, R. (1964). Über die konvexe Hülle von n zufällig gewählten Punkten. II. *Z. Wahrsch. Verw. Gebiete* **3** 138–147. [MR0169139](#)
- REYNAUD-BOURET, P., RIVOIRARD, V. and TULEAU-MALOT, C. (2011). Adaptive density estimation: A curse of support? *J. Statist. Plann. Inference* **141** 115–139. [MR2719482](#)
- RIGOLLET, P. and VERT, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli* **15** 1154–1178. [MR2597587](#)
- ROHDE, A. (2008). Adaptive goodness-of-fit tests based on signed ranks. *Ann. Statist.* **36** 1346–1374. [MR2418660](#)
- ROHDE, A. (2011). Optimal calibration for multiple testing against local inhomogeneity in higher dimension. *Probab. Theory Related Fields* **149** 515–559. [MR2776625](#)
- TSYBAKOV, A. B. (1989). Optimal estimation accuracy of nonsmooth images. *Problems of Information Transmission* **25** 180–191. [MR1021196](#)
- TSYBAKOV, A. B. (1991). Nonparametric techniques in image estimation. In *Nonparametric Functional Estimation and Related Topics (Spetses, 1990)* (G. Roussas, ed.). *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* **335** 669–677. Kluwer Academic, Dordrecht. [MR1154358](#)
- TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969. [MR1447735](#)
- TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. [MR2051002](#)

FAKULTÄT FÜR MATHEMATIK
RUHR-UNIVERSITÄT BOCHUM
44780 BOCHUM
GERMANY
E-MAIL: tim.patschkowski@ruhr-uni-bochum.de
angelika.rohde@ruhr-uni-bochum.de