Spatial aggregation of local likelihood estimates with applications to classification[*]

Belomestny, Denis

Weierstrass-Institute,
Mohrenstr. 39, 10117 Berlin, Germany
belomest@wias-berlin.de

Spokoiny, Vladimir

Weierstrass-Institute,
Mohrenstr. 39, 10117 Berlin, Germany
spokoiny@wias-berlin.de

Abstract

This paper presents a new method for spatially adaptive local (constant) likelihood estimation which applies to a broad class of nonparametric models, including the Gaussian, Poisson and binary response models. The main idea of the method is given a sequence of local likelihood estimates (”weak” estimates), to construct a new aggregated estimate whose pointwise risk is of order of the smallest risk among all “weak” estimates. We also propose a new approach towards selecting the parameters of the procedure by providing the prescribed behavior of the resulting estimate in the simple parametric situation. We establish a number of important theoretical results concerning the optimality of the aggregated estimate. In particular, our “oracle” results claims that its risk is up to some logarithmic multiplier equal to the smallest risk for the given family of estimates. The performance of the procedure is illustrated by application to the classification problem. A numerical study demonstrates its nice performance in simulated and real life examples.

Keywords: adaptive weights, local likelihood, exponential family, classification

AMS 2000 Subject Classification: 62G05, Secondary: 62G07, 62G08, 62H30

Running title: spatial aggregation for classification

[*]This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 Economic Risk.

1 Introduction

This paper presents a new method of spatially adaptive nonparametric estimation based on the aggregation of a family of local likelihood estimates. As a main application of the method we consider the problem of building a classifier on the base of the given family of k-NN or kernel classifiers.

The local likelihood approach has been intensively discussed in recent years, see e.g. Hastie and Tibshirani (1987), Staniswalis (1989), Loader (1996). We refer to Fan, Farmen and Gijbels (1998) for a nice and detailed overview of local maximum likelihood approach and related literature. Similarly to the nonparametric smoothing in regression or density framework, an important issue for the local likelihood modeling is the choice of localization (smoothing) parameters. Different types of model selection techniques based on the asymptotic expansion of the local likelihood are mentioned in Fan, Farmen and Gijbels (1998) which include global as well as variable bandwidth selection. However, the finite sample performance of estimators based on bandwidth or model selection is often rather unstable, see e.g. Breiman (1996). This point is particulary critical for the local or pointwise model selection procedures like Lepski's method (Lepski, 1990). In spite of the nice theoretical properties, see Lepski, Mammen and Spokoiny (1997), Lepski and Spokoiny (1997) or Spokoiny (1998), the resulting estimates suffer from a high variability due to a pointwise model choice, especially for a large noise level. This suggests that in some cases, the attempt to identify the true model is not necessarily the right thing to do. One approach to reduce a variability in adaptive estimation is model mixing or aggregation. Catoni (2001) and Yang (2004) among others have suggested global aggregating procedures that achieve the minimal estimation risks over the family of given "weak" estimates. In the regression setup Juditsky and Nemirovski (2000) have developed aggregation procedures which have a risk within a multiple of the smallest risk in the class of all convex combinations of "weak" estimates plus $\log(n)/n$. Tsybakov (2003) has discussed asymptotic minimax rates for the aggregation. The aggregation for density estimation has been studied by Li and Barron (1999) and more recently by Rigollet and Tsybakov (2005). To the best of our knowledge a pointwise aggregation has not yet been considered.

Our approach is based on the idea of the spatial (pointwise) aggregation of a family of local likelihood estimates ("weak" estimates) $\widetilde{\theta}^{(k)}$. The main idea is, given the sequence $\{\widetilde{\theta}^{(k)}\}$ to construct in a data driven way for every point x the "optimal" aggregated estimate $\widehat{\theta}(x)$. "Optimality" means that this estimate satisfies some kind of "oracle" inequality, that is, its pointwise risk does not exceed the smallest pointwise risk among all "weak" estimates up to a logarithmic multiple.

Our algorithm can be roughly described as follows. Let $\{\widetilde{\theta}^{(k)}(x)\}$, $k = 1, \ldots, K$, be a sequence of "weak" local likelihood estimates at a point x ordered according to their variability which decreases with k. Starting with $\widehat{\theta}^{(1)}(x) = \widetilde{\theta}^{(1)}(x)$, an aggregated estimate $\widehat{\theta}^{(k)}(x)$ at any step $1 < k \le K$ is constructed by mixing the previously constructed aggregated estimate $\widehat{\theta}^{(k-1)}(x)$ with the current "weak" estimate $\widetilde{\theta}^{(k)}(x)$:

$$\widehat{\theta}^{(k)}(x) = \gamma_k \widetilde{\theta}^{(k)}(x) + (1 - \gamma_k)\widehat{\theta}^{(k-1)}(x),$$

and $\widehat{\theta}^{(K)}(x)$ is taken as a final estimate. The mixing parameter γ_k (which may depend on the point x) is defined using a measure of statistical difference between $\widehat{\theta}^{(k-1)}(x)$ and $\widetilde{\theta}^{(k)}(x)$. In particular, γ_k is equal to zero if $\widehat{\theta}^{(k-1)}(x)$ lies outside the confidence set around $\widetilde{\theta}^{(k)}(x)$. In view of the sequential and pointwise nature of the algorithm, the suggested procedure is called *Spatial Stagewise Aggregation* (SSA). An important features of the procedure proposed are its simplicity and applicability to a variety of problems including Gaussian, binary, Poisson regression, density estimation, classification etc. The procedure does not require any splitting of the sample as many other aggregation procedures do, cf. Yang (2004). Besides that the theoretical properties of SSA can be rigorously studied. In particular, we establish precise nonasymptotic "oracle" results which are applicable under very mild conditions in a rather general set-up. We also show that the oracle property automatically implies spatial adaptivity of the proposed estimate.

Another important feature of the procedure is that it can be easily implemented and the problem of selecting the tuning parameters can be carefully addressed.

Our simulation study confirms a nice finite sample performance of the procedure for a broad class of different models and problems. We only show the results for the classification problem as the most interesting and difficult one. Some more examples for the univariate regression and density estimation can be found in our preprint Belomestny and Spokoiny (2005). Section 4 shows how the SSA procedure can be applied to aggregating kernel and k-NN classifiers in the classification problem. Although these two nonparametric classifiers are rather popular, the problem of selecting the smoothing parameter (the bandwidth for the kernel classifier or the number of neighbors for the k-NN method) has not been yet satisfactorily addressed. Again, the SSA-based classifier demonstrates the “oracle” quality in terms of the both pointwise and global misclassification errors. This application clearly shows one more important feature of the SSA method: it can be applied to an arbitrary design and arbitrary dimension of the design space. This is illustrated by simulated and real life classification examples in dimensions up to 10.

The procedure proposed in this paper is limited to aggregating the kernel type estimates which are based on the local constant approximation. The modern statistical literature usually considers the more general local linear (polynomial) approximation of the underlying function. However, for this paper we have decided by several reasons to restrict our attention to the local constant case. The most important one is that for the examples and applications we consider in this paper, the use of the local linear methods does not improve (and even degrade) the quality of estimation. Our experience strongly confirms that for the problems like classification, the local constant smoothing combined with the aggregation technique delivers a reasonable finite sample quality.

Our theoretical study is split into two big parts. Section 2 introduces the considered local parametric set-up and extends the parametric risk bounds to the local parametric and nonparametric situation under the so called “small modelling bias” condition. The main result (Corollary 2.6) claims that the parametric risk bounds continue to apply as long as this condition is fulfilled. One possible interpretation of our adaptive procedure is the search of the largest localizing scheme for which the ‘small modelling bias” condition still holds. Theoretical properties of the aggregation procedure are presented in Section 5. The main result states the “oracle” property of the SSA estimate: the risk of the aggregated estimate is within a log-multiple as small as the risk of the best “weak” estimate for the function at hand. The results are established in the precise nonasymptotic way for a rather general likelihood set-up under mild regularity conditions. Moreover, our approach allows to link the parametric and nonparametric theory. In particular, we show that the proposed method delivers the root-n accuracy in the parametric situation. In the nonparametric case, the quality corresponds to the best parametric approximation. Both the theoretical study and the motivation of the procedure employ some exponential bounds for the likelihood which are given in Section 2.2. An important feature of our theoretical study is that the problem of selecting the tuning parameters is also discussed in details. We offer a new approach in which the parameters of the procedure are selected to provide the desirable performance of the method in the simple parametric situation. This is similar to the hypothesis problem approach when the critical values are selected using the performance of the test statistic under the simple null hypothesis, see Section 3.3.1 for a detailed explanation.

2 Local constant modeling for exponential families

This section presents some results on local constant likelihood estimation. We begin by describing the model under consideration. Suppose we are given independent random data $Z_1, \ldots, Z_n$ of the form $Z_i = (X_i, Y_i)$. Here every X_i means a vector of "features" or explanatory variables which determines the distribution of the "observation" Y_i. For simplicity we assume that the X_i 's are valued in the finite dimensional Euclidean space $\mathcal{X} = I\!R^d$ and the Y_i 's belong to $I\!R$. The vector X_i can be viewed as a location and Y_i as the "observation at X_i". Our model assumes that the distribution of each Y_i is determined by a finite dimensional parameter θ which may depend on the location X_i.

More precisely, let $\mathcal{P} = (P_\theta, \theta \in \Theta)$ be a parametric family of distributions dominated by a measure P. In this paper we only consider the case when Θ is a subset of the real line. By $p(\cdot, \theta)$ we denote the corresponding density. We consider the regression-like model in which every "response" Y_i is, conditionally on $X_i = x$, distributed with the density $p(\cdot, f(x))$ for some unknown function $f(x)$ on $\mathcal{X}$ with values in Θ. The considered model can be written as

$$Y_i \sim P_{f(X_i)}.$$

The aim of the data-analysis is to infer on the "regression" function $f(x)$. For the related models see Fan and Zhang (1999) and Cai, Fan and Li (2000).

In this paper we focus on the case when $\mathcal{P}$ is *an exponential family.* This means that the density functions $p(y,\theta) = \frac{dP_\theta}{dP}(y)$ are of the form $p(y,\theta) = p(y)e^{yC(\theta)-B(\theta)}$. Here $C(\theta)$ and $B(\theta)$ are some given nondecreasing functions on Θ and $p(y)$ is some nonnegative function on $I\!R$.

A natural parametrization for this family means the equality $\boldsymbol{E}_\theta Y = \int yp(y,\theta)P(dy) = \theta$ for all $\theta \in \Theta$. This condition is useful because the weighted average of observations is a natural unbiased estimate of θ. In what follows we assume that $\mathcal{P}$ also fulfills the following regularity conditions:

(A1) $\mathcal{P} = (P_\theta, \theta \in \Theta \subseteq I\!R)$ is an exponential family with a natural parametrization, and the functions $B(\cdot)$ and $C(\cdot)$ are continuously differentiable.

(A2) Θ is compact and convex and the Fisher information $I(\theta) := \boldsymbol{E}_\theta |\partial \log p(Y,\theta)/\partial\theta|^2$ fulfills for some $\varkappa \geq 1$

$$|I(\theta')/I(\theta'')|^{1/2} \leq \varkappa, \qquad \theta', \theta'' \in \Theta.$$

We illustrate this set-up with two examples relevant to the applications we consider below. Some more examples can be found in Fan, Farmen and Gijbels (1998) and Polzehl and Spokoiny (2005).

Example 2.1. (Inhomogeneous Bernoulli (Binary Response) model) Let $Z_i = (X_i, Y_i)$ with $X_i \in I\!R^d$ and Y_i being a Bernoulli r.v. with parameter $f(X_i)$, that is, $\boldsymbol{P}(Y_i = 1 \mid X_i = x) = f(x)$ and $\boldsymbol{P}(Y_i = 0 \mid X_i = x) = 1 - f(x)$. Such models arise in many econometric applications and are widely used in classification and digital imaging.

Example 2.2. (Inhomogeneous Poisson model) Suppose that every Y_i is valued in the set $\mathbb{N}$ of nonnegative integer numbers and $\boldsymbol{P}(Y_i = k \mid X_i = x) = f^k(x)e^{-f(x)}/k!$, that is, Y_i follows a Poisson distribution with parameter $\theta = f(x)$. This model is commonly used in the queueing theory, it occurs in positron emission tomography and also serves as an approximation for the density model obtained by a binning procedure.

In the parametric setup with $f(\cdot) \equiv \theta$ the distribution of every "observation" Y_i coincides with P_θ for some $\theta \in \Theta$ and the parameter θ can be well estimated using the parametric maximum likelihood method:

$$\widetilde{\theta} = \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \log p(Y_i, \theta).$$

In the nonparametric framework, one usually applies the localization idea. In the local constant set-up this means that the regression function $f(\cdot)$ can be well approximated by a constant within some neighborhood of every point x in the "feature" space $\mathcal{X}$. This leads to the local model concentrated in some neighborhood of the point x.

2.1 Localization

We use the *localization by weights* as a general method to describe a local model. Let, for a fixed x, a nonnegative weight $w_i = w_i(x) \le 1$ be assigned to the observation Y_i at X_i, $i = 1, \ldots, n$. The weights $w_i(x)$ determine a local model corresponding to the point x in the sense that, when estimating the local parameter $f(x)$, every observation Y_i is taken with the weight $w_i(x)$. This leads to the local (weighted) maximum likelihood estimate $\widetilde{\theta} = \widetilde{\theta}(x)$ of $f(x)$:

$$\widetilde{\theta}(x) = \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} w_i(x) \, \log p(Y_i, \theta) \; . \tag{2.1}$$

We mention now two possible ways of choosing the weights $w_i(x)$. *Localization by a bandwidth* is defined by weights of the form $w_i(x) = K_{\mathrm{loc}}(\boldsymbol{l}_i)$ with $\boldsymbol{l}_i = \rho(x, X_i)/h$ where h is a bandwidth, $\rho(x, X_i)$ is the Euclidean distance between x and the design point X_i and K_{loc} is a *location kernel. Localization by a window* simply restricts the model to a subset (window) $U = U(x)$ of the design space which depends on x, that is, $w_i(x) = \mathbf{1}(X_i \in U(x))$. Observations Y_i with X_i outside the region $U(x)$ are not used for estimating $f(x)$. This kind of localization arises e.g. in the classification with k-nearest neighbors method or in the regression tree approach. Sometime it is convenient to combine these two methods by defining $w_i(x) = K_{\mathrm{loc}}(\boldsymbol{l}_i)\mathbf{1}(X_i \in U(x))$. One example is given by the boundary corrected kernels.

We do not assume any special structure for the weights $w_i(x)$, that is, any configuration of weights is allowed. We also denote $W = W(x) = \{w_1(x), \ldots, w_n(x)\}$ and

$$L(W, \theta) = \sum_{i=1}^{n} w_i(x) \, \log p(Y_i, \theta).$$

To keep the notation short, we do not show the dependence of the weights on x explicitly in what follows.

2.2 Local likelihood estimation for an exponential family model

If $\mathcal{P} = (P_\theta)$ is an exponential family with the natural parametrization, the local log-likelihood and the local maximum likelihood estimates admit a simple closed form representation. For a given set of weights $W = \{w_1, \ldots, w_n\}$ with $w_i \in [0,1]$, denote

$$N = \sum_{i=1}^{n} w_i, \qquad S = \sum_{i=1}^{n} w_i Y_i \, .$$

Note that the both sums depend on the location x via the weights $\{w_i\}$.

Lemma 2.1 (Polzehl and Spokoiny, 2005). *It holds*

$$L(W,\theta) = \sum_{i=1}^{n} w_i \log p(Y_i,\theta) = SC(\theta) - NB(\theta) + R$$

where $R = \sum_{i=1}^{n} w_i \log p(Y_i)$. *Moreover,*

$$\widetilde{\theta} = S/N = \sum_{i=1}^{n} w_i Y_i \Big/ \sum_{i=1}^{n} w_i \tag{2.2}$$

and

$$L(W,\widetilde{\theta},\theta) := L(W,\widetilde{\theta}) - L(W,\theta) = N\mathcal{K}(\widetilde{\theta},\theta).$$

Now we present some exponential inequality for the "fitted log-likelihood" $L(W,\widetilde{\theta},\theta)$ which apply in the parametric situation $f(\cdot) \equiv \theta$ for arbitrary weighting scheme and arbitrary sample size.

Theorem 2.2 (Polzehl and Spokoiny, 2005). *Let* $W = \{w_i\}$ *be a localizing scheme such that* $\max_i w_i \le 1$. *If* $f(X_i) \equiv \theta^*$ *for all* X_i *with* $w_i > 0$ *then for any* $z > 0$

$$\boldsymbol{P}_{\theta^*}(L(W,\widetilde{\theta},\theta^*) > z) = \boldsymbol{P}_{\theta^*}\left(N\mathcal{K}(\widetilde{\theta},\theta^*) > z\right) \le 2e^{-z}.$$

Remark 2.1. Condition $A2$ ensures that the Kullback-Leibler divergence $\mathcal{K}$ fulfills $\mathcal{K}(\theta',\theta^*) \le I^*|\theta' - \theta^*|^2$ for any point θ' in a neighborhood of θ^*, where I^* is the maximum of the Fisher information over this neighborhood. Therefore, the result of Theorem 2.2 guarantees that $|\widetilde{\theta} - \theta^*| \le CN^{-1/2}$ with a high probability. Theorem 2.2 can be used for constructing the confidence intervals for the parameter θ^*.

Theorem 2.3. *If $\mathfrak{z}_\alpha$ satisfies $2e^{-\mathfrak{z}_\alpha} \le \alpha$, then*

$$\mathcal{E}_\alpha = \{\theta' : N\mathcal{K}\bigl(\widetilde{\theta}, \theta'\bigr) \le \mathfrak{z}_\alpha\}$$

is an α-confidence set for the parameter θ^*.

Theorem 2.2 claims that the estimation loss measured by $\mathcal{K}(\theta', \theta)$ is with high probability bounded by $\mathfrak{z}/N$ provided that $\mathfrak{z}$ is sufficiently large. Similarly, one can establish a risk bound for a power loss function.

Theorem 2.4. *Assume $A1$ and $A2$ and let Y_i be i.i.d. from P_{θ^*}. Then for any $r > 0$*

$$\boldsymbol{E}_{\theta^*} L^r(\widetilde{\theta}, \theta^*) \equiv N^r \boldsymbol{E}_{\theta^*} \mathcal{K}^r(\widetilde{\theta}, \theta^*) \le \tau_r .$$

where $\tau_r = 2r \int_{\mathfrak{z}\ge 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z} = 2r\Gamma(r)$. Moreover, for every $\lambda < 1$

$$\boldsymbol{E}_{\theta^*} \exp\bigl\{\lambda L(\widetilde{\theta}, \theta^*)\bigr\} \equiv \boldsymbol{E}_{\theta^*} \exp\bigl\{\lambda N\mathcal{K}(\widetilde{\theta}, \theta^*)\bigr\} \le 2(1-\lambda)^{-1}.$$

Proof. By Theorem 2.2

$$\begin{aligned}
\boldsymbol{E}_{\theta^*} L^r(\widetilde{\theta}, \theta^*) &\le -\int_{\mathfrak{z}\ge 0} \mathfrak{z}^r d\boldsymbol{P}_{\theta^*}(L(\widetilde{\theta}, \theta^*) > \mathfrak{z}) \\
&\le r\int_{\mathfrak{z}\ge 0} \mathfrak{z}^{r-1} \boldsymbol{P}_{\theta^*}(L(\widetilde{\theta}, \theta^*) > \mathfrak{z}) d\mathfrak{z} \le 2r \int_{\mathfrak{z}\ge 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z}
\end{aligned}$$

and the first assertion is fulfilled. The last assertion is proved similarly. □

2.3 Risk of estimation in nonparametric situation. "Small modeling bias" condition

This section extends the bound of Theorem 2.2 to the nonparametric situation when the function $f(\cdot)$ is not any longer constant even in a vicinity of the reference point x. We, however, suppose that the function $f(\cdot)$ can be well approximated by a constant θ at all points X_i with positive weights w_i. To measure the quality of the approximation, define for every θ

$$\Delta(W, \theta) = \sum_i \delta\bigl(\theta, f(X_i)\bigr) \mathbf{1}(w_i > 0), \tag{2.3}$$

where with $\ell(y, \theta, \theta') = \log \frac{p(y,\theta)}{p(y,\theta')}$

$$\delta(\theta, \theta') = \log E_\theta e^{-2\ell(Y,\theta,\theta')} = \log E_\theta \frac{p^2(Y,\theta')}{p^2(Y,\theta)}.$$

One can easily check that $\delta(\theta, \theta') \le I^* |\theta - \theta'|^2$, where $I^* = \max_{\theta'' \in [\theta, \theta']} I(\theta'')$.

Theorem 2.5. *Let $\mathcal{F}_W$ be a σ-field generated by the r.v. Y_i for which $w_i > 0$ and let $\Delta(W, \theta) \le \Delta$. Then it holds for any random variable ξ measurable w.r.t. $\mathcal{F}_W$*

$$\boldsymbol{E}_{f(\cdot)} \xi \le \bigl(\mathrm{e}^{\Delta} \boldsymbol{E}_\theta \xi^2\bigr)^{1/2}.$$

Proof. Define $Z_W(\theta) = \exp\Bigl\{-\sum_i \ell\bigl(Y_i, \theta, f(X_i)\bigr) \mathbf{1}(w_i > 0)\Bigr\}$. This value is nothing but the likelihood ratio of the measure $\boldsymbol{P}_{f(\cdot)}$ w.r.t. $\boldsymbol{P}_\theta$ upon restricting to the observations Y_i for which $w_i > 0$. Then for any $\xi \sim \mathcal{F}_W$, it holds $\boldsymbol{E}_{f(\cdot)} \xi = \boldsymbol{E}_\theta \xi Z_W(\theta)$. Independence of the Y_i's implies

$$\begin{aligned} \log \boldsymbol{E}_\theta Z_W^2(\theta) &= \sum_i \log \boldsymbol{E}_\theta e^{-2\ell(Y_i, \theta, f(X_i))} \mathbf{1}(w_i > 0) \\ &= \sum_i \delta\bigl(\theta, f(X_i)\bigr) \mathbf{1}(w_i > 0) \le \Delta. \end{aligned}$$

The result now follows from the Cauchy-Schwartz inequality $\boldsymbol{E}_\theta \xi Z_W(\theta) \le \bigl\{\boldsymbol{E}_\theta \xi^2 \boldsymbol{E}_\theta Z_W^2(\theta)\bigr\}^{1/2}$. □

This result implies that the bound for the risk of estimation $\boldsymbol{E}_{f(\cdot)} L^r(\widetilde{\theta}, \theta) \equiv N^r \boldsymbol{E}_{f(\cdot)} \mathcal{K}^r(\widetilde{\theta}, \theta)$ under the parametric hypothesis can be extended to the nonparametric situation provided that the value $\Delta(W, \theta)$ is sufficiently small.

Corollary 2.6. *For any $r > 0$ and any $\lambda < 1$, it holds*

$$\begin{aligned} N^r \boldsymbol{E}_{f(\cdot)} \bigl|\mathcal{K}(\widetilde{\theta}, \theta)\bigr|^r &\le \sqrt{e^{\Delta(W,\theta)} \tau_{2r}}, \\ N \Bigl\{\boldsymbol{E}_{f(\cdot)} \bigl|\mathcal{K}(\widetilde{\theta}, \theta)\bigr|^r\Bigr\}^{1/r} &\le \frac{1}{\lambda} \Bigl\{\log \frac{2}{1-\lambda} + \Delta(W, \theta) + 2(r-1)_+\Bigr\}. \end{aligned}$$

Proof. The first bound follows directly from Theorems 2.4 and 2.5. The proof of the second one utilizes the fact that for $r > 0$ the function $h(x) = \log^r\bigl(x + c_r\bigr)$ with $c_r = \mathrm{e}^{(r-1)_+}$ is concave on $(0, \infty)$ because

$$h''(x) = \frac{r \log^{r-2}\bigl(x + c_r\bigr)}{(x + c_r)^2} \Bigl\{r - 1 - \log\bigl(x + c_r\bigr)\Bigr\} \le 0$$

for $x \ge 0$. This implies with $\zeta = \lambda L(\widetilde{\theta}, \theta)/2$ by monotonicity of log and Jensen's inequality that $\boldsymbol{E}_{f(\cdot)} \zeta^r \le \boldsymbol{E}_{f(\cdot)} h(\mathrm{e}^{\zeta}) \le h(\boldsymbol{E}_{f(\cdot)} \mathrm{e}^{\zeta})$ and hence,

$$\boldsymbol{E}_{f(\cdot)}^{1/r} \zeta^r \le \log\bigl(\boldsymbol{E}_{f(\cdot)} \mathrm{e}^{\zeta} + c_r\bigr) \le \log \boldsymbol{E}_{f(\cdot)} \mathrm{e}^{\zeta} + (r-1)_+ \le \frac{1}{2} \log\bigl(\mathrm{e}^{\Delta(W,\theta)} \boldsymbol{E}_{\theta} \mathrm{e}^{2\zeta}\bigr) + (r-1)_+$$

and the assertion follows. □

Corollary 2.6 presents two bounds for the risk of estimation in the nonparametric situation which extend the similar parametric bounds by Theorem 2.5. The risk bound in the parametric situation can be interpreted as the bound for the variance of the estimate $\widetilde{\theta}$ while the term $\Delta(W, \theta)$ controls the bias of estimation, see the next section for more details. The both bounds formally apply whatever the "modeling bias" $\Delta(W, \theta)$ is. However, the results are meaningful only if this bias is not too large. The first bound could be preferable for small values of $\Delta(W, \theta)$, however, the multiplicative factor $e^{\Delta(W,\theta)/2}$ makes this bound useless for large $\Delta(W, \theta)$. The advantage of the second bound is that the "modeling bias" enters in the additive form.

In the rest of this section we briefly comment on relations between the results of Section 2.3 and the usual rate results under smoothness conditions on the function $f(\cdot)$ and the regularity conditions on the design $X_1, \ldots, X_n$. More precisely, we assume that the weights w_i are supported on the ball of a radius $h > 0$ with the center at x and the function $f(\cdot)$ is smooth within this ball in the sense that for $\theta^* = f(x)$

$$\delta^{1/2}\bigl(\theta^*, f(x+t)\bigr) \le Lh, \qquad \forall |t| \le h. \tag{2.4}$$

In view of the inequality $\delta(\theta, \theta') \le I^* |\theta - \theta'|^2$ this condition is equivalent to the usual Lipschitz property. Obviously, (2.4) implies with $\overline{N} = \sum_i \mathbf{1}(w_i > 0)$

$$\Delta(W, \theta^*) \le L^2 h^2 \overline{N}.$$

Combined with the result of Corollary 2.6 these bounds lead to the following rate results.

Theorem 2.7. *Assume (A1) and (A2) and let $\delta^{1/2}\bigl(\theta^*, f(x+t)\bigr) \le Lh$ for and all $|t| \le h$. Select $h = c(L^2 n)^{-1/(2+d)}$ for some $c > 0$ and let the localizing scheme W be such that $w_i = 0$ for all X_i with $|X_i - x| > h$, $N := \sum_i w_i \ge \mathfrak{d}_1 n h^d$ and*

$\overline{N} := \sum_i \mathbf{1}(w_i > 0) \le \mathfrak{d}_2 n h^d$ *with some constants* $\mathfrak{d}_1 < \mathfrak{d}_2$ *. Then*

$$\boldsymbol{E}_{f(\cdot)} \bigl| N \mathcal{K}\bigl(\widetilde{\theta}, \theta^*\bigr) \bigr|^{r/2} \le \bigl\{ \exp\bigl(c^{2+d} \mathfrak{d}_2\bigr) \tau_r \bigr\}^{1/2}.$$

Moreover, with $c_2 = c^{rd/2} \exp\bigl(c^{2+d} \mathfrak{d}_2 / 2\bigr) \mathfrak{d}_1^{-r/2}$ *, it holds*

$$\boldsymbol{E}_{f(\cdot)} \bigl| n^{1/(2+d)} \mathcal{K}\bigl(\widetilde{\theta}, \theta^*\bigr) \bigr|^{r/2} \le c_2 L^{rd/(2+d)} \tau_r^{1/2}.$$

This corresponds to the classical accuracy of nonparametric estimation for the Lipschitz functions, cf. Fan, Farmen and Gijbels (1998).

3 Description of the method

We start by describing the considered set-up. Let a point of interest x be fixed and the target of estimation is the value $f(x)$ of the regression function at x . The local parametric approach described in Section 2 and based on the local constant approximation of the regression function in a vicinity of the point x strongly relies on the choice of the local neighborhood, or more generally, of the set of weights (w_i) . The problem of selecting such weights and constructing an adaptive (data-driven) estimate is one of the main issues for practical applications and we focus on this problem in this section.

3.1 Local adaptive estimation. General setup

For a fixed x , we assume to be given an *ordered* set of localizing schemes $W^{(k)} = (w_i^{(k)})$ for $k = 1, ..., K$. The ordering condition means that $w_i^{(k)} \ge w_i^{(k')}$ for all i and all $k > k'$, that is, the degree of locality given by $W_i^{(k)}$ is weakened as k grows. See Section 3.3 for some examples. For the popular example of kernel weights $w_i^{(k)} = K\bigl((X_i - x)/h_k\bigr)$, this condition means that the bandwidth h_k grows with k . Let also $\{\widetilde{\theta}^{(k)}, \, k = 1, ..., K\}$ be the corresponding set of local likelihood estimates for $\theta = f(x)$:

$$\widetilde{\theta}^{(k)}(x) = \operatorname*{argmax}_{\theta \in \Theta} L(W, \theta) = \sum_{i=1}^{n} w_i^{(k)} Y_i \Big/ \sum_{i=1}^{n} w_i^{(k)}.$$

Due to Theorem 2.2 the value $1/N_k$ can be used to measure the variability of the estimate $\widetilde{\theta}^{(k)}$. The ordering condition particularly means that N_k grows and hence, the variability of $\widetilde{\theta}^{(k)}$ decreases with k .

Given the estimates $\widetilde{\theta}^{(k)}$, we consider a larger class of their convex combinations:

$$\widehat{\theta} = \alpha_1 \widetilde{\theta}^{(1)} + \ldots + \alpha_K \widetilde{\theta}^{(K)}, \quad \alpha_1 + ... + \alpha_K = 1, \quad \alpha_k \geq 0,$$

where the mixing coefficients α_k may depend on the point x. We aim at constructing a new estimate $\widehat{\theta}$ in this class which performs at least as good as the best one in the original family $\{\widetilde{\theta}^{(k)}\}$.

3.2 Stagewise aggregation procedure

The adaptive estimate $\widehat{\theta}$ of $\theta = f(x)$ is computed sequentially via the following algorithm.

1. Initialization: $\widehat{\theta}^{(1)} = \widetilde{\theta}^{(1)}$.

2. Stagewise aggregation: For $k = 2, ..., K$

$$\widehat{\theta}^{(k)} := \gamma_k \widetilde{\theta}^{(k)} + (1 - \gamma_k) \widehat{\theta}^{(k-1)},$$

with the mixing parameter γ_k being defined for some $\mathfrak{z}_k > 0$ and a kernel $K_{\mathrm{ag}}(\cdot)$ as

$$\gamma_k = K_{\mathrm{ag}}\bigl(\boldsymbol{m}^{(k)}/\mathfrak{z}_k\bigr), \qquad \boldsymbol{m}^{(k)} := N_k \mathcal{K}(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k-1)})$$

3. Loop: If $k < K$, then increase k by one and continue with step 2. Otherwise terminate and set $\widehat{\theta} = \widehat{\theta}^{(K)}$.

The idea behind the procedure is quite simple. We start with the first estimate $\widetilde{\theta}^{(1)}$ having the smallest degree of locality but the largest variability of order $1/N_1$. Next we consider estimates with larger values N_k. Every current estimate $\widetilde{\theta}^{(k)}$ is compared with the previously constructed estimate $\widehat{\theta}^{(k-1)}$. If the difference is not significant then the new estimate $\widehat{\theta}^{(k)}$ basically coincides with $\widetilde{\theta}^{(k)}$. Otherwise the procedure essentially keeps the previous value $\widehat{\theta}^{(k-1)}$. For measuring the difference between the estimates $\widetilde{\theta}^{(k)}$ and $\widehat{\theta}^{(k-1)}$, we use $\boldsymbol{m}^{(k)} := N_k \mathcal{K}(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k-1)})$ which is motivated by the results of Theorems 2.2 and 2.3. In particular, a large value of $\boldsymbol{m}^{(k)}$ means that $\widehat{\theta}^{(k-1)}$ does not belong to the confidence set corresponding to $\widetilde{\theta}^{(k)}$ and hence indicates a significant difference between these two estimates. To quantify this significance, the procedure utilizes the parameters (critical values) $\mathfrak{z}_k$. Their choice is discussed in Section 3.3.1.

Remark 3.1. If $K_{\mathrm{ag}}(\cdot)$ is the uniform kernel on $[0,1]$ then γ_k is either zero or one depending on the value of $\boldsymbol{m}^{(k)}$. This yields by induction arguments that the final estimate coincides with one of the "weak" estimates $\widetilde{\theta}^{(k)}$. In this case our method can be considered as a pointwise model selection method.

If the kernel K_{ag} is such that $K_{\mathrm{ag}}(t) = 1$ for $t \le b$ with some positive b, then the small values of the "test statistic" $\boldsymbol{m}^{(k)}$ lead to the aggregated estimate $\widehat{\theta}^{(k)} = \widetilde{\theta}^{(k)}$. This is an important feature of the procedure which will be used in our implementation and the theoretical study.

3.3 Parameter choice and implementation details

The implementation of the SSA procedure requires fixing a sequence of local likelihood estimates, the kernel K_{ag} and the parameters $\mathfrak{z}_k$. The next section gives some examples how the set of localizing schemes $W^{(k)}$ can be selected. The only important parameters of the method are "critical values" $\mathfrak{z}_k$ which normalize the "test statistics" $\boldsymbol{m}^{(k)}$. Section 3.3.1 describes in details how they can be selected in practice.

The kernel K_{ag} should satisfy $0 \le K_{\mathrm{ag}}(t) \le 1$, should be monotonously decreasing and have support on $[0,1]$. Besides that, there is a positive number b such that $K_{\mathrm{ag}}(t) = 1$ for $t \le b$. Our default choice is a piecewise linear kernel with $b = 1/6$ and $K_{\mathrm{ag}}(t) = \bigl(1 - (t-b)_+\bigr)_+$. Our numerical results (not shown here) indicate that the particular choice of the kernel K_{ag} has only a minor effect on the final results.

3.3.1 Choice of the parameters $\mathfrak{z}_k$

The "critical values" $\mathfrak{z}_k$ define the level of significance for the test statistics $\boldsymbol{m}^{(k)}$. A proper choice of these parameters is crucial for the performance of the procedure. We propose in this section one general approach for selecting them which is similar to the bootstrap idea in the hypothesis testing problem. Namely, we select these values to provide the prescribed performance of the procedure in the parametric situation (under the null hypothesis). For every step k, we require that the estimate $\widehat{\theta}^{(k)}$ is sufficiently close to the "oracle" estimate $\widetilde{\theta}^{(k)}$ in the parametric situation $f(\cdot) \equiv \theta$ in the sense that

$$\sup_{\theta^* \in \Theta} \boldsymbol{E}_{\theta^*} \bigl| N_k \mathcal{K}\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k)}\bigr) \bigr|^r \le \alpha \tau_r \tag{3.1}$$

for all $k = 2, \ldots, K$ with τ_r from Theorem 2.4. In some cases the risk $\boldsymbol{E}_{\theta^*} \bigl| N_k \mathcal{K}\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k)}\bigr) \bigr|^r$ does not depend on θ^*. This is, for example, the case when θ is a shift or scale parameter, as for Gaussian shift, exponential and volatility families. Then it sufficient to check (3.1) for any single point θ^*. In the general situation, the risk $\boldsymbol{E}_{\theta^*} \bigl| N_k \mathcal{K}\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k)}\bigr) \bigr|^r$ depends on the parameter value θ^*. However, our numerical results (not reported here) indicate that this dependence is minor and usually it suffices to check these conditions for one parameter θ^*. In particular, for the Bernoulli model considered in Section 4 we recommend to only check the condition (3.1) for the "least favorable" value $\theta^* = 1/2$ corresponding to the largest variance of the estimate $\widetilde{\theta}$.

The values α and r in (3.1) are two global parameters. The role of α is similar to the level of the test in the hypothesis testing problem while r describes the power of the loss function. A specific choice is subjective and depends on the particular application at hand. Taking a large r and small α would result in an increase of the critical values and therefore, improves the performance of the method in the parametric situation at cost of some loss of sensitivity to parameter changes. Theorem 5.1 presents some upper bounds for the critical values $\mathfrak{z}_k$ as functions of α and r in the form $a_0 + a_1 \log \alpha^{-1} + a_2 r (K - k)$ with some coefficients a_0, a_1 and a_2. We see that these bounds linearly depend on r and on $\log \alpha^{-1}$. For our applications to classification, we apply a relatively small value $r = 1/2$ because the misclassification error corresponds to the bounded loss function. We also apply $\alpha = 1$ although the other values in the range $[0.5, 1]$ lead to very similar results. Note that in general the such defined parameters $\mathfrak{z}_k$ depend on the model considered, design $X_1, \ldots, X_n$ and the localizing schemes $W^{(1)}, \ldots, W^{(K)}$ which in turn can differ from point to point. Therefore, an implementation of the suggested rule would require to compute the parameters separately for every point of estimation. However, in many situations, e.g. for the regular design, this variation from point to point is negligible, and a universal set of parameters can be used. Important is only that the conditions (3.1) are fulfilled for all the points.

3.3.2 Simplified parameter choice

The proposal (3.1) is not constructive: we have just $K-1$ conditions for choosing $K-1$ parameters. Here we present a simplified procedure which is rather simple for the implementation and based on the Monte Carlo simulations. It suggests to first identify the last value $\mathfrak{z}_K$ using the reduced aggregation procedure with only two estimates $\widetilde{\theta}^{(K-1)}$ and $\widetilde{\theta}^{(K)}$:

$$\sup_{\theta^* \in \Theta} \boldsymbol{E}_{\theta^*} \bigl| N_K \mathcal{K}\bigl(\widetilde{\theta}^{(K)}, \widehat{\theta}(\mathfrak{z}_K)\bigr) \bigr|^r \le \alpha \tau_r / (K-1)$$

where $\widehat{\theta}(\mathfrak{z}_K) = \gamma \widetilde{\theta}^{(K)} + (1-\gamma) \widetilde{\theta}^{(K-1)}$, $\gamma = K_{\mathrm{ag}}(\boldsymbol{m}/\mathfrak{z}_K)$ and $\boldsymbol{m} = N_K \mathcal{K}\bigl(\widetilde{\theta}^{(K)}, \widetilde{\theta}^{(K-1)}\bigr)$. The other values $\mathfrak{z}_k$ are found in the form $\mathfrak{z}_k = \mathfrak{z}_K + \iota(K-k)$ to provide (3.1). This suggestion is justified by the result of Theorem 5.1 from Section 5.1.

3.3.3 Examples of sequences of local likelihood estimates

This section presents some examples and recommendations for the choice of the localizing schemes $W^{(k)}$ which we also use in our simulation study. Note, however, that the choice of $W^{(k)}$'s is not a part of the SSA procedure. The procedure applies with any choice under some rather mild growth conditions.

Below we assume that the design $X_1, \ldots, X_n$ is supported on the unit cube $[-1,1]^d$. This condition can be easily provided by rescaling the design components. We mention two approaches for choosing the localizing scheme which are usually used in applications. One is based on a given sequence of bandwidths, one more is based on the nearest neighbor structure of the design. In both situations we assume that a *location* kernel K_{loc} is a nonnegative function on the unit cube in $[-1,1]^d$. In general we only assume that this kernel is decreasing alone any radial line, that is, $K_{\mathrm{loc}}(\rho x) \ge K_{\mathrm{loc}}(x)$ for any $x \in [-1,1]^d$ and $\rho \le 1$, and $K_{\mathrm{loc}}(x) = 0$ for $|x| \ge 1$. In the most of applications, one applies an isotropic kernel K_{loc} which only depends on the norm of x. The recommended choice is the Epanechnikov kernel $K_{\mathrm{loc}}(x) = (1-|x|^2)_+$.

Bandwidth-based localizing schemes: This way can be recommended for the univariate or bivariate equidistant design. Let $\{h_k\}_{k=1}^K$ be a finite set of *bandwidth-candidates*. We assume that this set is ordered, that is, $h_1 < h_2 < \ldots < h_K$. Every such bandwidth determines the collection of kernel weights $w_i^{(k)} = K_{\mathrm{loc}}\bigl((X_i - x)/h_k\bigr)$, $i = 1, \ldots, n$. This definition assumes that the same localizing bandwidth is applied for all the directions in the feature space. In all the examples below we apply a geometrically increasing sequence of "bandwidths" h_k, that is, $h_{k+1} = a h_k$ for some $a > 1$. This sequence is uniquely determined by the starting value h_1, the factor a and the total number K of local schemes. The recommended choice of a is $(1.25)^{1/d}$ although our numerical results (not reported here) indicate no significant change in the results when the other value of a in the range 1.1 to 1.3 is used. The value h_1 is to be selected in a way that the starting estimate $\widetilde{\theta}^{(1)}$ is well defined for all the points of estimation. In the case of a local constant approximation, this value can be taken very small because even one point can be sufficient for a preliminary estimation. In the case of a regular design, the value h_1 is of order $n^{-1/d}$. The number K of local schemes $W^{(k)}$ or, equivalently, of the "weak" estimates $\widetilde{\theta}^{(k)}$ is mostly determined by the values h_1 and a in such a way that $h_K = h_1 a^{K-1}$ is about one, that is, the last estimate behaves like a global parametric estimate from the whole sample. The formula $K = a \log(h_K/h_1)$ suggests that K is at most logarithmic in the sample size n.

k-NN based local schemes: If the design is irregular or the design space is high dimensional ($d > 2$) then it is useful to apply the local schemes based on the k-nearest neighbor structure of the design. For this approach, an increasing sequence $\{N_k\}$ of integers has to be fixed. For a fixed x and every $k \geq 1$, the bandwidth h_k is the minimal one for which the ball of radius h_k contains at least N_k design points. The weights are defined again by $w_i^{(k)} = K_{\mathrm{loc}}\bigl((X_i - x)/h_k\bigr)$. The sequence $\{N_k\}$ is selected similarly to the sequence $\{h_k\}$ in the bandwidth-based approach. One starts with a fixed N_1 and then multiplies it at every step with some factor $a > 1$: $N_{k+1} = a N_k$. The number of steps K is such that N_K is of order n.

One can easily check that the kernel and k-NN based local schemes coincide in the case of univariate regular design.

4 Application to classification

One observes a training sample (X_i, Y_i), $i = 1, \ldots, n$, with X_i valued in a Euclidean space $\mathcal{X} = I\!\!R^d$ with known class assignment $Y_i \in \{0, 1\}$. Our objective is to construct a discrimination rule assigning every point $x \in \mathcal{X}$ to one of the two classes. The classification problem can be naturally treated in the context of a binary response model. It is assumed that each observation Y_i at X_i is a Bernoulli r.v. with the parameter $\theta_i = f(X_i)$, that is, $\boldsymbol{P}(Y_i = 0|X_i) = 1 - f(X_i)$ and $\boldsymbol{P}(Y_i = 1|X_i) = f(X_i)$. The "ideal" Bayes discrimination rule is $\mathbf{1}\left(f(x) \geq 1/2\right)$. Since the function $f(x)$ is usually unknown it is replaced by its estimate $\widehat{\theta}$. If the distribution of X_i within the class k has density p_k then

$$\theta_i = \pi_1 p_1(X_i)/(\pi_0 p_0(X_i) + \pi_1 p_1(X_i)).$$

where π_k is the prior probability of k th population $k = 0, 1$.

Nonparametric methods of estimating the function θ are typically based on local averaging. Two typical examples are given by the k-nearest neighbor (k-NN) estimate and the kernel estimate. For a given k and every point x in $\mathcal{X}$, denote by $\mathcal{D}_k(x)$ the subset of the design $X_1, \ldots, X_n$ containing the k nearest neighbors of x. Then the k-NN estimate of $f(x)$ is defined by averaging the observations Y_i over $\mathcal{D}_k(x)$:

$$\widetilde{\theta}^{(k)}(x) = k^{-1} \sum_{X_i \in \mathcal{D}_k(x)} Y_i .$$

The definition of the kernel estimate of $f(x)$ involves a univariate kernel function $K(\cdot)$ and the bandwidth h:

$$\widetilde{\theta}^{(h)}(x) = \sum_{i=1}^{n} K\left(\frac{\rho(x, X_i)}{h}\right) Y_i \Big/ \sum_{i=1}^{n} K\left(\frac{\rho(x, X_i)}{h}\right).$$

Both methods require the choice of a smoothing parameter (the value k for k-NN and the bandwidth h for the kernel estimate).

Example 4.1. In this example we consider the binary classification problem with the corresponding class densities $p_0(x)$ and $p_1(x)$ given by two component normal mixtures

$$\begin{aligned} p_0(x) &= 0.2\phi(x; (-1, 0), 0.5\boldsymbol{I}_2) + 0.8\phi(x; (1, 0), 0.5\boldsymbol{I}_2) \\ p_1(x) &= 0.5\phi(x; (0, 1), 0.5\boldsymbol{I}_2) + 0.5\phi(x; (0, -1), 0.5\boldsymbol{I}_2) \end{aligned}$$

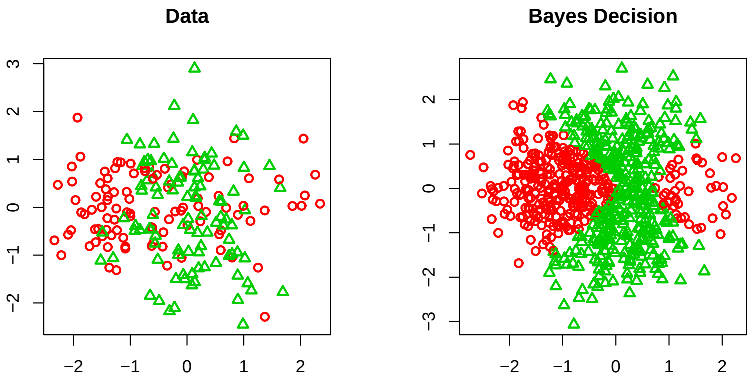

Figure 4.1: Sample from the binary response model with the normal mixture class densities (left) and results of applying the Bayes discrimination rule for this model (right).

where $\phi(\cdot;\mu,\Sigma)$ is the density of the multivariate normal distribution with the mean vector μ and the covariance matrix Σ and $\boldsymbol{I}_2$ is 2×2 unit matrix.

Figure 4.1 shows one typical realization of the training sample with 100 observations in each class (left) and the optimal Bayes classification for a testing sample with 1000 observations in each class (right). First, in order to illustrate the "oracle" property of the SSA we compute the pointwise misclassification errors for all week estimate and SSA estimate at four boundary points. They are obtained using training sample of size 400, k-NN weighting scheme with $N_1 = 5$, $N_K = 300$, $K = 30$ and $\alpha = 0.5$. Further, we have done 500 simulations runs generating each time 100 training points and 100 testing points. The rates of misclassification on testing sets have been averaged thereafter to give the *mean misclassification error* which is shown as a reference dotted line in Figure 4.3. We note here that the critical values

$$\mathfrak{z}_k = 0.0031 + 0.007 * (K-k), \quad k = 1,\ldots,K$$

have been computed only once for one design realization and least favorable parameter value $\theta^* = 0.5$ and then used in all runs. The same strategy is used in other examples as well. Next, two "weak" classification methods, k-NN and kernel classifiers, with

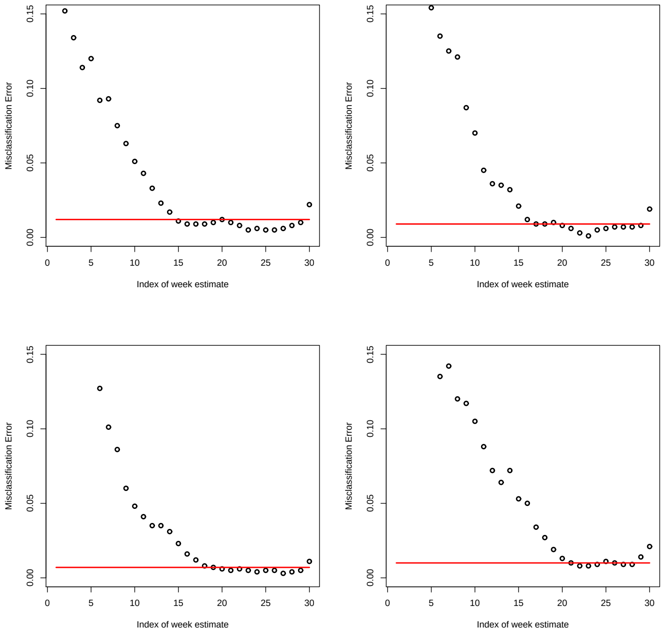

Figure 4.2: Pointwise misclassification errors (black dots) at four points for all weak estimates used in the example 4.1. The solid reference lines correspond to the SSA misclassification errors.

varying smoothing parameters are applied to the same data set Figure 4.3 (top) shows the dependence of the misclassification error on the bandwidth for kernel classifiers and on the number of nearest neighbors for the k-NN classifier.

One can observe that a careful choice of the smoothing parameter is crucial for getting a reasonable quality of the classification. A wrong choice leads to a significant increase of the misclassification rate, especially for the kernel classifiers. At the same time, the optimal choice can lead to a reasonable quality of the classification which is only slightly worse than one of the Bayes decision rule.

Example 4.2. Now we consider the example 4.1 with additional 8 independent $\mathcal{N}(0,1)$ distributed nuisance components. So, now $X_i = (X_i^1, .., X_i^{10})$ where

$$(X_i^1, X_i^2) \sim p_{\mathrm{class}(i)}, \quad (X_i^3, .., X_i^{10}) \sim \mathcal{N}((\underbrace{0, ..., 0}_{8}), \boldsymbol{I}_8).$$

The SSA procedure is implemented now again using k-NN weights with the number of

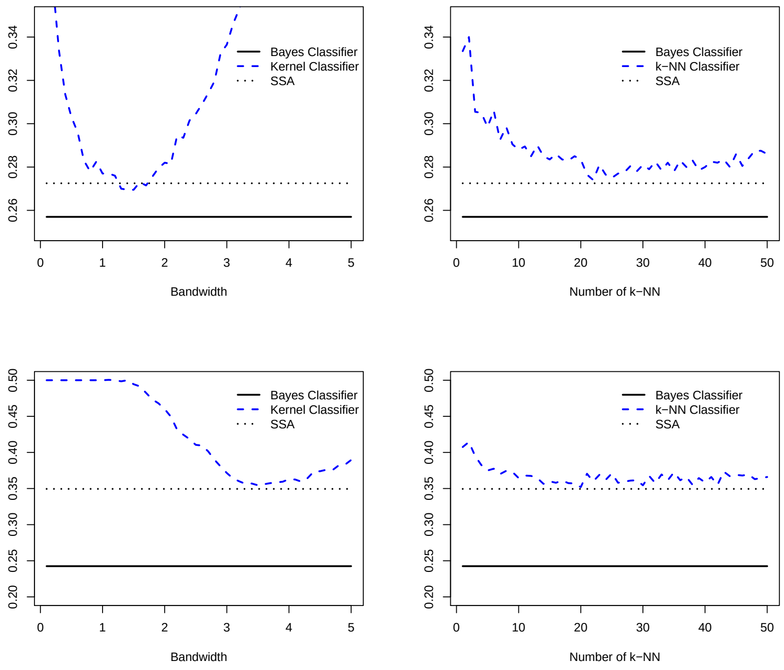

Figure 4.3: Misclassification errors as a functions of the main smoothing parameter for k-NN (right) and kernel (left) classifiers. SSA and Bayes misclassification errors are given as reference lines. Top: Example 4.1 (dimension 2). Bottom: Example 4.2 (dimension 10).

nearest neighbors exponentially increasing from 5 to 100. The results are shown in the bottom row of Figure 4.3. We observe again that the quality of the both standard classifiers depends significantly on the choice of the smoothing parameters. In the considered high dimensional situation, even under the optimal choice the quality of the dimension independent Bayes classifier is not attained. However, the SSA procedure performs again nearly as good as the best k-NN or kernel classifier.

Example 4.3. [BUPA liver disorders] We consider the dataset sampled by BUPA Medical Research Ltd. It consists of 7 variables and 345 observed vectors. The subjects are single male individuals. The first 5 variables are measurements taken by blood tests that are thought to be sensitive to liver disorders and might arise from excessive alcohol consumption. The sixth variable is a sort of selector variable. The seventh variable is the label indicating the class identity. Among all the observations, there are 145 people belonging to the liver-disorder group (corresponding to selector number 2) and 200 people belonging to the liver-normal group. The BUPA liver disorder data set is notoriously difficult for classifying with the usual error rates about 30% . We apply SSA, k -NN and kernel classifiers to tackle this problem. In SSA procedure the kNN weighting scheme was employed with number of k -NN ranging from 2 to 100 . Figure 4.4 shows the corresponding one-leave-out cross-validation errors for the above methods. One can see that the SSA method is uniformly better than kernel or k -NN classifiers.

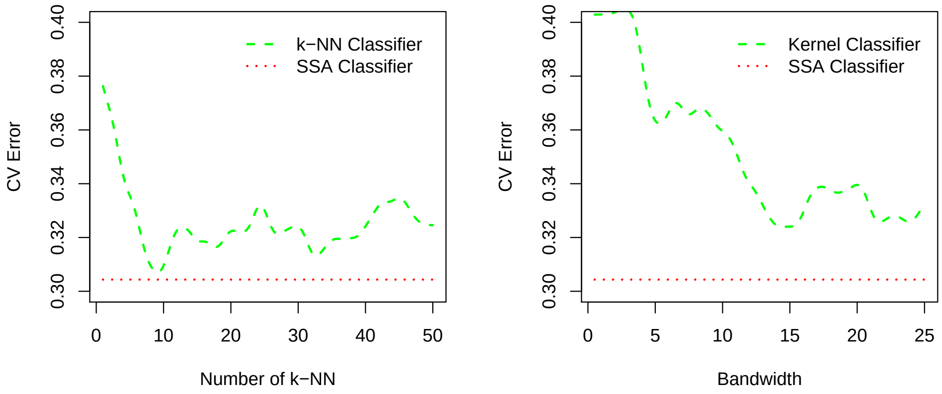

Figure 4.4: One-leave-out cross-validation errors as a functions of the main smoothing parameters for k -NN (right) and kernel (left) classifiers. The dotted line describes the error of SSA classifier.

Example 4.4. [Bankruptcy Data] The data set from the Compustat repository contains the statistics about bankruptcies (defaults) in private sector of USA economy during the period 2000-2005. There are 14 explanatory variables including different financial ratios, industry indicators and so on. First, the preliminary analysis is conducted and two most informative variables (equity/total assets ratio and net income/total assets ratio (profitability)) are selected. The projection of the default statistics on the corresponding plane is shown in Figure 4.5. Further, the performance of SSA procedure is compared to the performance of k-NN classifier with different numbers of nearest neighbors. Namely, the one-leave-out cross-validation errors are computed for both SSA and k-NN classification methods and the last one is presented in Figure 4.5 as a function of the number of nearest neighbors. Again as in previous examples, the quality of classification strongly depends on the choice of the parameter k. The adaptive SSA procedure provides the performance corresponding to the best possible choice of this parameter.

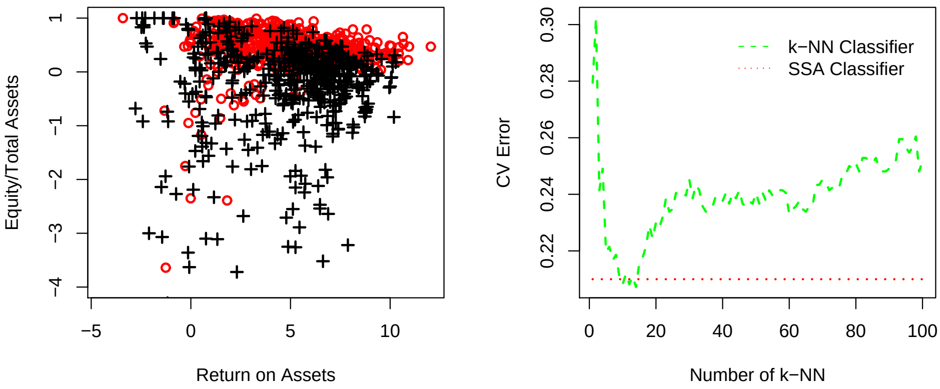

Figure 4.5: Left: Default events (crosses indicate defaulted firms and circles operating ones) are shown in dependence on the two characteristics of a firm. Right: One-leave-out cross-validation error for k-NN classifier as a function of the number of nearest neighbors. The CV error for SSA classifier is given as a red reference line.

5 Some theoretical properties of the SSA method

This section discusses some important theoretical properties of the proposed aggregating procedure. In particular we establish the "oracle" result which claims that the aggregated estimate is up to a log-factor as good as the best one among the considered family $\{\widetilde{\theta}^{(k)}\}$ of local constant estimates.

The majority of the results in the modern statistical literature are stated as asymptotic rate results. It is however well known that the rate optimality of an estimation procedure does not automatically imply its good finite sample properties and cannot be used for comparing different procedures. The rate results are also almost useless for selecting the parameters of the procedure. In our theoretical study we apply another approach which aims to link parametric and nonparametric inference with the focus on the adaptive behaviour of the proposed method. This means in particular that the SSA procedure attains the parametric accuracy if the parametric assumption is fulfilled. In the general situation the procedure attains (up to a unavoidable price for adaptation) the quality corresponding to the best possible local parametric approximation for the underlying model near the point of interest .

The "oracle" result is in its turn a consequence of two important properties of the aggregated estimate $\widehat{\theta}$: "propagation" and "stability". "Propagation" can be viewed as the oracle result in the parametric situation with $f(\cdot) \equiv \theta^*$. In this case the oracle choice would be the estimate with the largest value N_k, that is, the last estimate $\widetilde{\theta}^{(K)}$ in the family $\{\widetilde{\theta}^{(k)}\}$. The "propagation" property means that at every step k of the procedure the "aggregated" estimate $\widehat{\theta}^{(k)}$ is close to the "oracle" estimate $\widetilde{\theta}^{(k)}$. In other words, the "propagation" property ensures that at every step the degree of locality is relaxed and the local model applied for estimation is extended to a larger neighborhood described by the weights $W^{(k)}$. The "propagation" property can be naturally extended to a nearly parametric case when $\varDelta(W^{(k)}, \theta)$ is small for some fixed θ and all $k \le k^*$. The "propagation" feature of the procedure ensures that the quality of estimation improves and confidence bounds for $\widehat{\theta}^{(k)}$ become tighter as the number of iterations increases provided that the "small modeling bias" condition still holds. Finally, the "stability" property secures that the quality gained in the "propagation" stage will be kept for the final estimate.

Our theoretical study is done under assumptions A1 and A2 on the parametric family $\mathcal{P}$. Additionally we impose an assumption on the sequence of localizing schemes $W^{(k)}$ which was already mentioned in Section 3.

(A3) the set $W^{(k)}$ is ordered in the sense that $w_i^{(k)} \ge w_i^{(k')}$ for all i and all $k > k'$. Moreover, for some constants u_0, u with $0 < u_0 \le u < 1$, values $N_k = \sum_{j=1}^{n} w_i^{(k)}$ satisfy for every $2 \le k \le K$

$$u_0 \le N_{k-1}/N_k \le u.$$

5.1 Behavior in the parametric situation

First we consider the homogeneous situation with the constant parameter value $f(x) = \theta^*$. Our first result claims that in this situation under condition $A3$ the parameters $\mathfrak{z}_k$ can be chosen in the form $\mathfrak{z}_k = \mathfrak{z}_K + \iota(K-k)$ to fulfill the "propagation" condition (3.1). The proof is given in the Appendix.

Theorem 5.1. *Assume $A1$, $A2$ and $A3$. Let $f(X_i) = \theta^*$ for all i. Then there are three constants a_0, a_1 and a_2 depending on r and u_0, u only such that the choice*

$$\mathfrak{z}_k = a_0 + a_1 \log \alpha^{-1} + a_2 r \log N_k$$

ensures (3.1) for all $k \le K$. Particularly, $\boldsymbol{E}_{\theta^*} \bigl| N_K \mathcal{K}\bigl(\widetilde{\theta}^{(K)}, \widehat{\theta}\bigr)\bigr|^r \le \alpha \tau_r$.

5.2 "Propagation" under "small modelling bias"

Now we extend the "propagation" result to the situation when the parametric assumption is not fulfilled any more but the deviation from the parametric structure within the considered local model is sufficiently small. This deviation can be measured for the localizing scheme $W^{(k)}$ by $\Delta(W^{(k)}, \theta)$ from (2.3).

We suppose that there is a number k^* such that the modeling bias $\Delta(W^{(k)}, \theta)$ is small for some θ and all $k \le k^*$. Consider the corresponding estimate $\widehat{\theta}^{(k^*)}$ obtained after the first k^* steps of the algorithm. Theorem 2.5 implies in this situation the following result.

Theorem 5.2. *Assume $A1$, $A2$ and $A3$. Let θ and k^* be such that $\Delta(W^{(k)}, \theta) \le \Delta$ for some $\Delta \ge 0$ and all $k \le k^*$. Then*

$$\begin{aligned} \boldsymbol{E}_{f(\cdot)} \bigl| N_{k^*} \mathcal{K}\bigl(\widetilde{\theta}^{(k^*)}, \widehat{\theta}^{(k^*)}\bigr)\bigr|^{r/2} &\le \sqrt{\alpha \tau_r e^{\Delta}}, \\ \boldsymbol{E}_{f(\cdot)} \bigl| N_{k^*} \mathcal{K}\bigl(\widetilde{\theta}^{(k^*)}, \theta\bigr)\bigr|^{r/2} &\le \sqrt{\tau_r e^{\Delta}}. \end{aligned}$$

5.3 "Stability after propagation" and "oracle" results

Due to the "propagation" result, the procedure performs well as long as the "small modeling bias" condition $\Delta(W^{(k)}, \theta) \le \Delta$ is fulfilled. To establish the accuracy result for the final estimate $\widehat{\theta}$, we have to check that the aggregated estimate $\widehat{\theta}^{(k)}$ does not vary much at the steps "after propagation" when the divergence $\varDelta(W^{(k)}, \theta)$ from the parametric model becomes large.

Theorem 5.3. *Under* $A1$*,* $A2$ *and* $A3$*, it holds for every* $k \le K$

$$N_k \mathcal{K}\bigl(\widehat{\theta}^{(k)}, \widehat{\theta}^{(k-1)}\bigr) \le \mathfrak{z}_k. \tag{5.1}$$

Moreover, under $A3$*, it holds for every* k' *with* $k < k' \le K$

$$N_k \mathcal{K}\bigl(\widehat{\theta}^{(k')}, \widehat{\theta}^{(k)}\bigr) \le \varkappa^2 c_u^2 \,\bar{\mathfrak{z}}_k \tag{5.2}$$

with $c_u = (u^{-1/2} - 1)^{-1}$ *and* $\bar{\mathfrak{z}}_k = \max_{l \ge k} \mathfrak{z}_l$.

Remark 5.1. An interesting feature of this result is that it is fulfilled with probability one, that is, the control of stability "works" not only with a high probability, it always applies. This property follows directly from the construction of the procedure.

Proof. By convexity of the Kullback-Leibler divergence $\mathcal{K}(u, v)$ w.r.t. the first argument

$$\mathcal{K}\bigl(\widehat{\theta}^{(k)}, \widehat{\theta}^{(k-1)}\bigr) \le \gamma_k \mathcal{K}\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k-1)}\bigr).$$

If $\mathcal{K}\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k-1)}\bigr) \ge \mathfrak{z}_k/N_k$, then $\gamma_k = 0$ and (5.1) follows. Now, Assumption A2 and Lemma 6.1 yield

$$\mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k')}, \widehat{\theta}^{(k)}\bigr) \le \varkappa \sum_{l=k+1}^{k'} \mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(l)}, \widehat{\theta}^{(l-1)}\bigr) \le \varkappa \sum_{l=k+1}^{k'} \bigl(\mathfrak{z}_l/N_l\bigr)^{1/2}.$$

The use of Assumption A3 leads to the bound

$$\mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k')}, \widehat{\theta}^{(k)}\bigr) \le \varkappa\bigl(\bar{\mathfrak{z}}_k/N_k\bigr)^{1/2} \sum_{l=k+1}^{k'} u^{(l-k)/2} \le \varkappa\sqrt{u}(1-\sqrt{u})^{-1}\bigl(\bar{\mathfrak{z}}_k/N_k\bigr)^{1/2}$$

which proves (5.2). □

Combination of the "propagation" and "stability" statements implies the main result concerning the properties of the adaptive estimate $\widehat{\theta}$.

Theorem 5.4. *Assume* $A1$*,* $A2$ *and* $A3$*. Let* k^* *be a "good" choice in the sense that*

$$\max_{k \le k^*} \varDelta(W^{(k)}, \theta) \le \varDelta$$

for some θ *and some value* Δ. *Then*

$$\boldsymbol{E}_{f(\cdot)}\bigl|N_{k^*}\mathcal{K}\bigl(\widetilde{\theta}^{(k^*)},\widehat{\theta}\bigr)\bigr|^{r/2} \le 2^{(r-1)_+}\varkappa^r\bigl\{\sqrt{\alpha\tau_r e^{\Delta}} + \bigl(c_u^2\,\bar{\mathfrak{z}}_{k^*}\bigr)^{r/2}\bigr\}$$

where c_u *is the constant from Theorem 5.3.*

Proof. Just observe that by Lemma 6.1

$$\mathcal{K}^{1/2}\bigl(\widehat{\theta},\widetilde{\theta}^{(k^*)}\bigr) \le \varkappa\Bigl\{\mathcal{K}^{1/2}\bigl(\widetilde{\theta}^{(k^*)},\widehat{\theta}^{(k^*)}\bigr) + \sum_{l=k^*+1}^{\widehat{k}}\mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(l)},\widehat{\theta}^{(l-1)}\bigr)\Bigr\}$$

and follow the proof of Theorem 5.3. □

We also present a corollary of the "oracle" result concerning the risk of the adaptive estimate $\widehat{\theta}$ for the special case with $r = 1$. The other values of r can be considered as well, one only has to update the constants depending on r. We also assume that $\alpha \le 1$.

Corollary 5.5. *Let* $\max_{k\le k^*}\Delta(W^{(k)},\theta) \le \Delta$ *for some* θ *and some* Δ. *Then*

$$N_{k^*}^{1/2}\boldsymbol{E}_{f(\cdot)}\mathcal{K}^{1/2}\bigl(\widehat{\theta},\theta\bigr) \le \varkappa\Bigl(2\sqrt{\tau_1 e^{\Delta}} + \sqrt{c_u^2\bar{\mathfrak{z}}_{k^*}}\Bigr).$$

Remark 5.2. Recall that in the parametric situation, the risk $\boldsymbol{E}_{\theta^*}\bigl|N_{k^*}\mathcal{K}\bigl(\widetilde{\theta}^{(k^*)},\theta^*\bigr)\bigr|^{1/2}$ of $\widetilde{\theta}^{(k^*)}$ is bounded by $\tau_{1/2}$, cf. Theorem 2.2. In the nonparametric situation, the result is only slightly worse: the value $\tau_{1/2}$ is replaced by $\sqrt{\tau_1 e^{\Delta}}$ which takes into account the modeling bias. There is also an additional term proportional to $\sqrt{\bar{\mathfrak{z}}_{k^*}}$ which can be considered as the payment for adaptation. Due to Theorem 5.1, $\bar{\mathfrak{z}}_{k^*}$ is bounded from above by $\mathfrak{z}_K + \iota(K-k^*)$. By Theorem 5.1 K is only logarithmic in the sample size n.

Therefore, the risk of the aggregated estimate corresponds to the best possible risk among the family $\{\widetilde{\theta}^{(k)}\}$ for the choice $k = k^*$ up to a logarithmic factor. Lepski, Mammen and Spokoiny (1997) established a similar result in the regression setup for the pointwise adaptive Lepski procedure. Combining the result of Corollary 5.5 with Theorem 2.7 yields the rate of adaptive estimation $\bigl(n^{-1}\log n\bigr)^{1/(2+d)}$ under Lipschitz smoothness of the function f and the usual design regularity, see Polzehl and Spokoiny (2005) for more details. It was shown by Lepski (1990) that in the problem of pointwise adaptive estimation this rate is optimal and cannot be improved by any estimation method. This gives an indirect proof of the optimality of our procedure: the factor $\bar{\mathfrak{z}}_{k^*}$

in the accuracy of estimation cannot be removed or reduced in the rate because otherwise the similar improvement would appear in the rate of estimation.

6 Appendix: Proof of Theorem 5.1

The proof utilizes the following simple "metric like" property of $\mathcal{K}^{1/2}(\cdot,\cdot)$.

Lemma 6.1 (Polzehl and Spokoiny, 2005, Lemma 5.2). *Under condition A2 it holds for every sequence* $\theta_0, \theta_1, \ldots, \theta_m$ *that*

$$\begin{aligned}
\mathcal{K}^{1/2}(\theta_1,\theta_2) &\le \varkappa\bigl\{\mathcal{K}^{1/2}(\theta_1,\theta_0)+\mathcal{K}^{1/2}(\theta_2,\theta_0)\bigr\},\\
\mathcal{K}^{1/2}(\theta_0,\theta_m) &\le \varkappa\bigl\{\mathcal{K}^{1/2}(\theta_0,\theta_1)+\ldots+\mathcal{K}^{1/2}(\theta_{m-1},\theta_m)\bigr\}.
\end{aligned}$$

With the given constants $\mathfrak{z}_k$, define for $k>1$ the random sets

$$\mathcal{A}_k = \{N_k\,\mathcal{K}(\widetilde{\theta}^{(k)},\widetilde{\theta}^{(k-1)}) \le b\mathfrak{z}_k\}, \qquad \mathcal{A}^{(k)} = \mathcal{A}_2 \cap \ldots \cap \mathcal{A}_k\,,$$

where b enters in the construction of K_{ag}: $K_{\mathrm{ag}}(t)=1$ for $t\le b$.

Note first that $\widehat{\theta}^{(k)}=\widetilde{\theta}^{(k)}$ on $\mathcal{A}^{(k)}$ for all $k\le K$. This fact can be proved by induction in k. For $k=1$, the assertion is trivial because $\widehat{\theta}^{(1)}=\widetilde{\theta}^{(1)}$. Now suppose that $\widehat{\theta}^{(k-1)}=\widetilde{\theta}^{(k-1)}$. Then it holds on $\mathcal{A}_k$ that $\boldsymbol{m}^{(k)}=N_k\mathcal{K}(\widetilde{\theta}^{(k)},\widehat{\theta}^{(k-1)})=N_k\mathcal{K}(\widetilde{\theta}^{(k)},\widetilde{\theta}^{(k-1)})\le b\mathfrak{z}_k$ and thus, $\gamma_k=K_{\mathrm{ag}}(\boldsymbol{m}^{(k)}/\mathfrak{z}_k)\ge K_{\mathrm{ag}}(b)=1$ yielding $\widehat{\theta}^{(k)}=\widetilde{\theta}^{(k)}$.

Therefore, it remains to bound the risk of $\widehat{\theta}^{(k)}$ on the complement $\overline{\mathcal{A}}^{(k)}$ of $\mathcal{A}^{(k)}$. Define $\mathcal{B}_k=\mathcal{A}^{(k-1)}\setminus\mathcal{A}^{(k)}$. On the event $\mathcal{B}_k$, the index k is the first one for which the condition $N_k\,\mathcal{K}(\widetilde{\theta}^{(k)},\widetilde{\theta}^{(k-1)})\le b\mathfrak{z}_k$ is violated. It is obvious that $\overline{\mathcal{A}}^{(k)}=\bigcup_{l<k}\mathcal{B}_l$. First we bound the probability $\boldsymbol{P}_{\theta^*}\bigl(\mathcal{B}_l\bigr)$. Applying assumption A3 and Lemma 6.1 yields for every l

$$\begin{aligned}
N_l\,\mathcal{K}(\widetilde{\theta}^{(l)},\widetilde{\theta}^{(l-1)}) &\le 2\varkappa^2 N_l\bigl\{\mathcal{K}(\widetilde{\theta}^{(l)},\theta^*)+\mathcal{K}(\widetilde{\theta}^{(l-1)},\theta^*)\bigr\}\\
&\le 2\varkappa^2\bigl\{N_l\,\mathcal{K}(\widetilde{\theta}^{(l)},\theta^*)+u_0^{-1}N_{l-1}\mathcal{K}(\widetilde{\theta}^{(l-1)},\theta^*)\bigr\}.
\end{aligned}$$

Therefore, by Theorem 2.2,

$$\boldsymbol{P}_{\theta^*}\bigl(\mathcal{B}_l\bigr)\le\boldsymbol{P}_{\theta^*}\bigl(N_l\,\mathcal{K}(\widetilde{\theta}^{(l)},\widetilde{\theta}^{(l-1)})>b\mathfrak{z}_l\bigr)\le 2\exp\Bigl(-\frac{u_0 b}{4\varkappa^2}\mathfrak{z}_l\Bigr).$$

On the set $\mathcal{B}_l$, it holds $\widehat{\theta}^{(l-1)} = \widetilde{\theta}^{(l-1)}$ and thus, for every $k > l$ the aggregated estimate $\widehat{\theta}^{(k)}$ by construction is a convex combination of $\widetilde{\theta}^{(l-1)}, \ldots, \widetilde{\theta}^{(k)}$. Convexity of the Kullback-Leibler divergence w.r.t. the second argument, the definition of $\widehat{\theta}^{(k)}$ and Lemma 6.1 ensure that

$$\begin{aligned}
\mathcal{K}^{1/2}\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k)}\bigr)\mathbf{1}\bigl(\mathcal{B}_l\bigr) &\le \max_{l'=l-1,\ldots,k-1} \mathcal{K}^{1/2}\bigl(\widetilde{\theta}^{(k)}, \widetilde{\theta}^{(l')}\bigr) \\
&\le \varkappa \max_{l'=l-1,\ldots,k-1} \bigl\{\mathcal{K}^{1/2}(\widetilde{\theta}^{(k)}, \theta^*) + \mathcal{K}^{1/2}(\widetilde{\theta}^{(l')}, \theta^*)\bigr\} \\
&\le 2\varkappa \max_{l'=l-1,\ldots,k} \mathcal{K}^{1/2}(\widetilde{\theta}^{(l')}, \theta^*).
\end{aligned}$$

This and Theorem 2.4 imply for every r

$$\begin{aligned}
\boldsymbol{E}_{\theta^*}\mathcal{K}^r\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k)}\bigr)\mathbf{1}\bigl(\mathcal{B}_l\bigr) &\le (2\varkappa)^{2r}\boldsymbol{E}_{\theta^*} \sum_{l'=l-1}^{k} \mathcal{K}^r(\widetilde{\theta}^{(l')}, \theta^*)\mathbf{1}\bigl(\mathcal{B}_l\bigr) \\
&\le (2\varkappa)^{2r} \sum_{l'=l-1}^{k} \boldsymbol{E}_{\theta^*}^{1/2}\mathcal{K}^{2r}(\widetilde{\theta}^{(l')}, \theta^*)\boldsymbol{P}_{\theta^*}^{1/2}\bigl(\mathcal{B}_l\bigr) \\
&\le (2\varkappa)^{2r}\tau_{2r}^{1/2} \sum_{l'=l-1}^{k} N_{l'}^{-r} 2\exp\Bigl(-\frac{u_0 b}{8\varkappa^2}\mathfrak{z}_l\Bigr) \\
&\le C_1 N_l^{-r}\tau_{2r}^{1/2}\exp\bigl(-c_2\mathfrak{z}_l\bigr)
\end{aligned}$$

for some fixed constants C_1 and c_2. Therefore,

$$\boldsymbol{E}_{\theta^*}\mathcal{K}^r\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k)}\bigr) \le \sum_{l=2}^{k} \boldsymbol{E}_{\theta^*}\mathcal{K}^r\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k)}\bigr)\mathbf{1}\bigl(\mathcal{B}_l\bigr) \le \sum_{l=2}^{k} C_1 N_l^{-r}\tau_{2r}^{1/2}\exp\bigl(-c_2\mathfrak{z}_l\bigr).$$

It remains to check that the choice $\mathfrak{z}_k = a_0 + a_1\log\alpha^{-1} + a_2 r\log(N_K/N_k)$ with properly selected a_0, a_1 and a_2 provide the required bound $\boldsymbol{E}_{\theta^*}\bigl|N_k\mathcal{K}\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k)}\bigr)\bigr|^r \le \alpha\tau_r$.

References

[1] Bickel, P.J., C.A.J. Klaassen, Y. Ritov and J.A. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models*, 1998, Springer.

[2] Breiman, L. (1996). Stacked regression. *Machine Learning*, **24** 49–64.

[3] Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inference for varying coefficients models. *J. Amer. Statist. Ass.*, **95** 888–902.

[4] Catoni, O. (2004). *Statistical learning theory and stochastic optimization.* Lecture Notes in Mathematics, **1851**. Springer-Verlag, Berlin.

[5] Fan, J., and Gijbels, I. (1995). Data driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statist. Soc.* Ser. B, **57** 371–394.

[6] Fan, J., Farmen, M. and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *J. Royal Statist. Soc.* Ser. B, **60** 591–608.

[7] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman & Hall, London.

[8] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518.

[9] Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient models (with discussion). *J. Royal Statist. Soc. Ser. B*, **55** 757–796.

[10] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.*, **28** 682–712.

[11] Lepski, O., Mammen, E. and Spokoiny, V. (1997). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Annals of Statistics*, **25** 929–947.

[12] Li, J. and Barron, A. (1999). Mixture density estimation. In. S.A. Sola, T.K. Leen, and K.R. Mueller, editors, *Advances in Neural Inforamtion proceedings systems.*

[13] Loader, C. R. (1996). *Local likelihood density estimation.* Academic Press.

[14] Polzehl, J. and Spokoiny, V. (2005). Propagation-Separation Approach for Local Likelihood Estimation. *Probab. Theory Related Fields*, DOI: 10.1007/s00440-005-0464-1.

[15] Rigollet, Ph. and Tsybakov, A. (2005). Linear and convex aggregation of density estimators. Manuscript.

[16] Spokoiny, V. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Ann. Statist.*, **26** 1356–1378.

[17] Staniswalis, J.C. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, **84** 276–283.

[18] Tibshirani, J.R., and Hastie, T.J. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82** 559–567.

[19] Tsybakov, A. (2003). Optimal rates of aggregation. Computational Learning Theory and Kernel Machines. B.Scholkopf and M.Warmuth, eds. *Lecture Notes in Artificial Intelligence*, **2777** Springer, Heidelberg, 303-313.

[20] Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** 25–47