Linear and convex aggregation of density estimators

Philippe Rigollet, Alexandre Tsybakov

Linear and convex aggregation of density estimators

Philippe Rigollet Alexandre B. Tsybakov

Laboratoire de Probabilités et Modèles Aléatoires,
Université Paris 6,
4 pl. Jussieu, 75252 Paris Cedex 05, France,
{rigollet, tsybakov}@ccr.jussieu.fr

May 11, 2006

Abstract

We study the problem of linear and convex aggregation of M estimators of a density with respect to the mean squared risk. We provide procedures for linear and convex aggregation and we prove oracle inequalities for their risks. We also obtain lower bounds showing that these procedures are rate optimal in a minimax sense. As an example, we apply general results to aggregation of multivariate kernel density estimators with different bandwidths. We show that linear and convex aggregates mimic the kernel oracles in asymptotically exact sense for a large class of kernels including Gaussian, Silverman's and Pinsker's ones. We prove that, for Pinsker's kernel, the proposed aggregates are sharp asymptotically minimax simultaneously over a large scale of Sobolev classes of densities. Finally, we provide simulations demonstrating performance of the convex aggregation procedure.

1991 Mathematics Subject Classification. Primary 62G08, Secondary 62C20, 62G05, 62G20.
Key words and phrases: aggregation, oracle inequalities, statistical learning, nonparametric density estimation, sharp minimax adaptivity, kernel estimates of a density.
Short title: Aggregation of density estimators.

1 Introduction

Consider i.i.d. random vectors $X_1, \dots, X_n$ with values in $\mathbb{R}^d$ having an unknown common probability density $p \in L_2(\mathbb{R}^d)$ that we want to estimate. For an estimator $\hat{p}$ of p based on the sample $\mathbb{X}^n = (X_1, \dots, X_n)$, define the L_2-risk

$$R_n(\hat{p}, p) = E_p^n \|\hat{p} - p\|^2,$$

where E_p^n denotes the expectation w.r.t. the distribution P_p^n of $\mathbb{X}^n$ and, for a function $g \in L_2(\mathbb{R}^d)$,

$$\|g\| = \left(\int_{\mathbb{R}^d} g^2(x) \mathrm{d}x \right)^{1/2}.$$

Suppose that we have $M \geq 2$ estimators $\hat{p}_1, \dots, \hat{p}_M$ of the density p based on the sample $\mathbb{X}^n$. The problem that we study here is to construct a new estimator $\tilde{p}_n$ of p, called *aggregate*, which is approximately at least as good as the best linear or convex combination of $\hat{p}_1, \dots, \hat{p}_M$. The problems of linear and convex aggregation of density estimators under the L_2 loss can be stated as follows.

1. **Problem (L): linear aggregation.** Find a *linear aggregate*, i.e. an estimator $\tilde{p}_n^{\mathbf{L}}$ which satisfies

$$R_n(\tilde{p}_n^{\mathbf{L}}, p) \leq \inf_{\lambda \in \mathbb{R}^M} R_n(\mathsf{p}_\lambda, p) + \Delta_{n,M}^{\mathbf{L}} \tag{1.1}$$

for every p belonging to a large class of densities $\mathcal{P}$, where

$$\mathsf{p}_\lambda = \sum_{j=1}^{M} \lambda_j \hat{p}_j, \quad \lambda = (\lambda_1, \dots, \lambda_M),$$

and $\Delta_{n,M}^{\mathbf{L}}$ is a sufficiently small remainder term that does not depend on p.

2. **Problem (C): convex aggregation.** Find a *convex aggregate*, i.e. an estimator $\tilde{p}_n^{\mathbf{C}}$ which satisfies

$$R_n(\tilde{p}_n^{\mathbf{C}}, p) \le \inf_{\lambda \in H} R_n(\mathsf{p}_\lambda, p) + \Delta_{n,M}^{\mathbf{C}} \tag{1.2}$$

for every p belonging to a large class of densities $\mathcal{P}$, where $\Delta_{n,M}^{\mathbf{C}}$ is a sufficiently small remainder term that does not depend on p, and H is a convex compact subset of $\mathbb{R}^M$. We will discuss in more detail the case $H = \Lambda^M$ where Λ^M is a simplex,

$$\Lambda^M = \Big\{ \lambda \in \mathbb{R}^M : \ \lambda_j \ge 0, \sum_{j=1}^{M} \lambda_j \le 1 \Big\}.$$

Our aim is to find aggregates satisfying (1.1) or (1.2) with the smallest possible remainder terms $\Delta_{n,M}^{\mathbf{L}}$ and $\Delta_{n,M}^{\mathbf{C}}$. These remainder terms characterize the price to pay for aggregation.

Linear and convex aggregates mimic the best linear (respectively, convex) combinations of the initial estimators. Along with them, one may consider *model selection (MS) aggregates* that mimic the best among the initial estimators $\hat{p}_1, \dots, \hat{p}_M$. We do not analyze this type of aggregation here.

The study of convergence properties of aggregation methods has been initiated by Nemirovski (2000), Catoni (1999, 2004) and Yang (2000). Most of the results were obtained for the regression and Gaussian white noise models (see a recent overview in Bunea, Tsybakov and Wegkamp (2004)). Aggregation of density estimators has received less attention. The work on this subject is mainly devoted to the MS aggregation with the Kullback-Leibler divergence as a loss function [Catoni (1999, 2004), Yang (2000), Zhang (2003)], and is based on information-theoretical ideas close to the earlier papers of Barron (1987), Li and Barron (1999). Devroye and Lugosi (2001) developed a method of MS aggregation of density estimators satisfying certain complexity assumptions under the L_1 loss.

To our knowledge, linear aggregation of density estimators has not been previously studied. For convex aggregation, the only paper we are aware of is that of Birgé (2003) where this type of aggregation under the L_1 loss is considered, while we study here the L_2 loss. In his setup, Birgé (2003) proves an inequality which is weaker than (1.2), with the oracle risk on the right hand side multiplied by a constant which is much larger than 1.

We do not only suggest aggregates satisfying sharp oracle inequalities (1.1), (1.2), but also demonstrate their optimality. Namely, we introduce the notion of optimal rate of aggregation and show that our aggregates attain optimal rates. This extends to density estimation context some results of the paper of Tsybakov (2003) where optimal rates of aggregation for the regression model have been obtained.

The main purpose of aggregation is to improve upon the initial set of estimators $\hat{p}_1, \dots, \hat{p}_M$. This is a general tool that applies to various kinds of estimators satisfying very mild conditions (we only assume that they are square integrable). Consider, for example, the simplest case when we have only two estimators ($M = 2$), where $\hat{p}_1$ is a good parametric density estimator for some fixed regular parametric family and $\hat{p}_2$ is a nonparametric density estimator. If the underlying density p belongs to the parametric family, $\hat{p}_1$ is perfect: its risk converges with the parametric rate $O(1/n)$. But for densities which are not in this family it may not converge at all. As for $\hat{p}_2$, it converges with a slow nonparametric rate even if the underlying density is within the parametric family. Aggregation (cf. Section 2 below) allows one to construct procedures that combine the advantages of both $\hat{p}_1$ and $\hat{p}_2$: the convex or linear aggregates

converge with the parametric rate $O(1/n)$ if p is within the parametric family, and with a nonparametric rate otherwise. Similar use of aggregation can be done in the problem of adaptation to the unknown smoothness (cf. Sections 5 and 6). In this case the index j of $\hat{p}_j$ corresponds to a value of the smoothing parameter, and the adaptive estimators in the oracle or minimax sense can be obtained as linear or convex aggregates. Of course, there exists a large variety of other methods of adaptation to unknown smoothness. In the numerical examples that we consider, our aggregates are comparable to benchmarks, and show somewhat more stable behavior for densities with highly inhomogeneous smoothness (cf. Section 7). It is important to note that aggregation can be used for adaptation to other characteristics than smoothness, for example, to the dimension of the subspace where the data effectively lie, under dimension reduction models [cf. Samarov and Tsybakov (2005)].

In this paper, we consider only one example of application of our general results to the problem of adaptation to the unknown smoothness. Specifically, we deal with aggregation of multivariate kernel density estimators with different bandwidths. Here the number $M = M_n$ of the estimators depends on n and satisfies $M_n/n \to 0$, as $n \to \infty$. We show in Corollary 5.1 that linear and convex aggregates mimic the kernel oracles in sharp asymptotic sense. This corollary is in the spirit of Stone's (1984) theorem on asymptotic optimality of cross-validation, but it is more powerful in several aspects because it is obtained under weaker conditions on p and covers kernels with unbounded support including Gaussian, Silverman's and Pinsker's kernels. Another application of our results is that, for Pinsker's kernel, we construct aggregates that are sharp asymptotically minimax simultaneously over a large scale of Sobolev classes of densities in the multidimensional case.

To perform aggregation, we use a sample splitting scheme. The sample $\mathbf{X}^n$ is split into two independent subsamples $\mathbf{X}_1^m$ (training sample) and $\mathbf{X}_2^\ell$ (validation sample) of sizes m and ℓ respectively where $m+\ell = n$ and usually $m \gg \ell$. The first subsample $\mathbf{X}_1^m$ is used to construct estimators $\hat{p}_j = \hat{p}_{m,j},\ j = 1, \dots, M$, while the second subsample $\mathbf{X}_2^\ell$ is used to aggregate them, i.e., to construct $\tilde{p}_n$ (thus, $\tilde{p}_n$ is measurable w.r.t. the whole sample $\mathbf{X}^n$). In a first analysis we will not consider sample splitting schemes but rather deal with a "pure aggregation" framework (as in most of the papers on the subject, cf. ,e.g., Nemirovski (2000), Juditsky and Nemirovski (2000) and Tsybakov (2003) for the regression problem) where the first subsample is frozen. This means that instead of the estimators $\hat{p}_1, \dots, \hat{p}_M$ we have fixed functions $p_1, \dots, p_M$ and that the expectations in oracle inequalities are taken only w.r.t. the second subsample.

This paper is organized as follows. In Section 2 we introduce linear and convex aggregation procedures and prove that they satisfy oracle inequalities of the type (1.1) and (1.2). Section 3 provides lower bounds showing optimality of the rates obtained in Section 2. Consequences for averaged aggregates are stated in Section 4. In Sections 5 and 6 we apply the results of Sections 2 and 4 to aggregation of kernel density estimators. Section 7 contains a simulation study. Throughout the paper we denote by c_i finite positive constants.

2 Oracle inequalities for linear and convex aggregates

In this section, $p_1, \dots, p_M$ are fixed functions, not necessarily probability densities. ¿From now on the notation p_λ for a vector $\lambda = (\lambda_1, \dots, \lambda_M) \in \mathbb{R}^M$ is understood in the following sense:

$$\mathsf{p}_\lambda \triangleq \sum_{j=1}^{M} \lambda_j p_j,$$

and, since for any fixed $\lambda \in \mathbb{R}^M$, the function p_λ is non-random, we have

$$R_n(\mathsf{p}_\lambda, p) = \|\mathsf{p}_\lambda - p\|^2.$$

Denote by $\mathcal{P}_0$ the class of all densities on $\mathbb{R}^d$ bounded by a constant $L > 0$:

$$\mathcal{P}_0 \triangleq \left\{ p : \mathbb{R}^d \to \mathbb{R} \,\middle|\, p \geq 0, \int_{\mathbb{R}^d} p(x)\mathrm{d}x = 1, \|p\|_\infty \leq L \right\},$$

where $\|\cdot\|_\infty$ stands for the $L_\infty(\mathbb{R}^d)$ norm. The constant L need not be known to the statistician.

We first give an oracle inequality for linear aggregation. Denote by $\mathcal{L}$ the linear span of $p_1, \ldots, p_M$. Let $\phi_1, \ldots, \phi_{M'}$ with $M' \leq M$ be an orthonormal basis of $\mathcal{L}$ in $L_2(\mathbb{R}^d)$. Define a linear aggregate

$$\tilde{p}_n^{\mathbf{L}}(x) \triangleq \sum_{j=1}^{M'} \hat{\lambda}_j^{\mathbf{L}} \phi_j(x), \quad x \in \mathbb{R}^d, \tag{2.1}$$

where

$$\hat{\lambda}_j^{\mathbf{L}} = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i).$$

Theorem 2.1 *Assume that $p_1, \ldots, p_M \in L_2(\mathbb{R}^d)$ and $p \in \mathcal{P}_0$. Then*

$$R_n(\tilde{p}_n^{\mathbf{L}}, p) \leq \min_{\lambda \in \mathbb{R}^M} \|\mathsf{p}_\lambda - p\|^2 + \frac{LM}{n} \tag{2.2}$$

for any integers $M \geq 2$ and $n \geq 1$.

Proof. Consider the projection of p onto $\mathcal{L}$:

$$p_{\mathcal{L}}^* \triangleq \operatorname*{argmin}_{\mathsf{p}_\lambda \in \mathcal{L}} \|\mathsf{p}_\lambda - p\|^2 = \sum_{j=1}^{M'} \lambda_j^* \phi_j,$$

where $\lambda_j^* = (p, \phi_j)$, and $(\cdot, \cdot)$ is the scalar product in $L_2(\mathbb{R}^d)$. Using the Pythagorean theorem we get that, almost surely,

$$\|\tilde{p}_n^{\mathbf{L}} - p\|^2 = \sum_{j=1}^{M'} (\hat{\lambda}_j^{\mathbf{L}} - \lambda_j^*)^2 + \|p_{\mathcal{L}}^* - p\|^2.$$

To finish the proof it suffices to take expectations in the last equation and to note that $E_p^n(\hat{\lambda}_j^{\mathbf{L}}) = \lambda_j^*$ and

$$E_p^n\left[(\hat{\lambda}_j^{\mathbf{L}} - \lambda_j^*)^2\right] = \mathrm{Var}(\hat{\lambda}_j^{\mathbf{L}}) \leq \frac{1}{n} \int_{\mathbb{R}^d} \phi_j^2(x) p(x) \mathrm{d}x \leq \frac{L}{n} \,.$$

∎

Consider now convex aggregation. Its aim is to mimic the *convex oracle* defined as $\lambda^* = \operatorname{argmin}_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2$ where H is a given convex compact subset of $\mathbb{R}^M$. Clearly,

$$\|\mathsf{p}_\lambda - p\|^2 = \|\mathsf{p}_\lambda\|^2 - 2 \int_{\mathbb{R}^d} \mathsf{p}_\lambda p + \|p\|^2.$$

Removing here the term $\|p\|^2$ independent of λ and estimating $\int_{\mathbb{R}^d} p_j p$ by $n^{-1} \sum_{i=1}^n p_j(X_i)$ we get the following estimate of the oracle

$$\hat{\lambda}^{\mathbf{C}} = \operatorname*{argmin}_{\lambda \in H} \left\{ \|\mathsf{p}_\lambda\|^2 - \frac{2}{n} \sum_{i=1}^{n} \mathsf{p}_\lambda(X_i) \right\}. \tag{2.3}$$

Now, we define a *convex aggregate* $\tilde{p}_n^{\mathbf{C}}$ by

$$\tilde{p}_n^{\mathbf{C}} \triangleq \sum_{j=1}^{M} \hat{\lambda}_j^{\mathbf{C}} p_j = \mathsf{p}_{\hat{\lambda}^{\mathbf{C}}}.$$

Theorem 2.2 *Let H be a convex compact subset of $\mathbb{R}^M$. Assume that $p_1, \ldots, p_M \in L_2(\mathbb{R}^d)$ and $p \in \mathcal{P}_0$. Then the convex aggregate $\tilde{p}_n^{\mathbf{C}}$ satisfies*

$$R_n(\tilde{p}_n^{\mathbf{C}}, p) \le \min_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2 + \frac{4LM}{n} \tag{2.4}$$

for any integers $M \ge 2$ and $n \ge 1$.

Proof. We will write for brevity $\hat{\lambda} = \hat{\lambda}^{\mathbf{C}}$. First note that the mapping $\lambda \mapsto \|\mathsf{p}_\lambda\|^2 - \frac{2}{n}\sum_{i=1}^n \mathsf{p}_\lambda(X_i)$ is continuous, thus $\hat{\lambda}$ exists, and the oracle $\lambda^* = \operatorname{argmin}_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2$ also exists. The definition of $\hat{\lambda}$ implies that, for any $p \in \mathcal{P}_0$,

$$\|\mathsf{p}_{\hat{\lambda}} - p\|^2 \le \|\mathsf{p}_{\lambda^*} - p\|^2 + 2T_n \tag{2.5}$$

where

$$T_n \triangleq \frac{1}{n}\sum_{i=1}^n \mathsf{p}_{\hat{\lambda}-\lambda^*}(X_i) - \int_{\mathbb{R}^d} \mathsf{p}_{\hat{\lambda}-\lambda^*} p.$$

Introduce the notation

$$\mathcal{Z}_n \triangleq \sup_{\mu \in \mathbb{R}^M : \|\mathsf{p}_\mu\| \ne 0} \frac{\left|\frac{1}{n}\sum_{i=1}^n \mathsf{p}_\mu(X_i) - E_p^n[\mathsf{p}_\mu(X_1)]\right|}{\|\mathsf{p}_\mu\|}.$$

Using the Cauchy-Schwarz inequality, the identity $\mathsf{p}_{\hat{\lambda}-\lambda^*} = \mathsf{p}_{\hat{\lambda}} - \mathsf{p}_{\lambda^*}$ and the elementary inequality $2\sqrt{xy} \le ax + y/a,\ \forall\, x, y, a > 0$, we get

$$\begin{aligned} E_p^n|T_n| &\le E_p^n\left(\mathcal{Z}_n \|\mathsf{p}_{\hat{\lambda}-\lambda^*}\|\right) \\ &\le \sqrt{E_p^n(\mathcal{Z}_n^2)}\sqrt{E_p^n(\|\mathsf{p}_{\hat{\lambda}-\lambda^*}\|^2)} \\ &\le \frac{a}{2}E_p^n(\|\mathsf{p}_{\hat{\lambda}} - \mathsf{p}_{\lambda^*}\|^2) + \frac{1}{2a}E_p^n(\mathcal{Z}_n^2), \quad \forall\, a > 0. \end{aligned} \tag{2.6}$$

Representing p_μ in the form $\mathsf{p}_\mu = \sum_{l=1}^{M'} \nu_l \phi_l$ where $\nu_l \in \mathbb{R}$ and $\{\phi_l\}$ is an orthonormal basis in $\mathcal{L}$ (cf. proof of Theorem 2.1) we find

$$\mathcal{Z}_n \le \sup_{\nu \in \mathbb{R}^M \setminus \{0\}} \frac{|\sum_{l=1}^{M'} \nu_l \zeta_l|}{|\nu|} = \Big(\sum_{l=1}^{M'} \zeta_l^2\Big)^{1/2},$$

where $|\nu| = \Big(\sum_{l=1}^{M'} \nu_l^2\Big)^{1/2}$ and

$$\zeta_l = \frac{1}{n}\sum_{i=1}^n \phi_l(X_i) - E_p^n[\phi_l(X_1)].$$

Hence

$$E_p^n(\mathcal{Z}_n^2) \le \frac{M'}{n} \max_{l=1,\ldots,M'} E_p^n[\phi_l^2(X_1)] \le \frac{LM}{n}, \tag{2.7}$$

whenever $\|p\|_\infty \le L$. Since $\{\mathsf{p}_\lambda : \lambda \in H\}$ is a convex subset of $L_2(\mathbb{R}^d)$ and p_{λ^*} is the projection of p onto this set, we have

$$\|\mathsf{p}_\lambda - p\|^2 \ge \|\mathsf{p}_{\lambda^*} - p\|^2 + \|\mathsf{p}_\lambda - \mathsf{p}_{\lambda^*}\|^2, \quad \forall\ \lambda \in H,\ p \in L_2(\mathbb{R}^d). \tag{2.8}$$

Using (2.8) with $\lambda = \hat{\lambda}$, (2.6) and (2.7) we obtain

$$E_p^n |T_n| \le \frac{a}{2} \left\{ E_p^n (\|\mathsf{p}_{\hat{\lambda}} - p\|^2 - \|\mathsf{p}_{\lambda^*} - p\|^2) \right\} + \frac{LM}{2an}\ .$$

This and (2.5) yield that, for any $0 < a < 1$,

$$E_p^n (\|\mathsf{p}_{\hat{\lambda}} - p\|^2) \le \|\mathsf{p}_{\lambda^*} - p\|^2 + \frac{LM}{a(1-a)n}\ .$$

Now, (2.4) follows by taking the infimum of the right hand side of this inequality over $0 < a < 1$. ∎

3 Lower bounds and optimal aggregation

We first define the notion of *optimal rate of aggregation* for density estimation, similar to that for the regression problem given in Tsybakov (2003). It is related to the minimax behavior of the excess risk

$$\mathcal{E}(\tilde{p}_n, p, H) = R_n(\tilde{p}_n, p) - \inf_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2$$

for a given class H of weights λ.

Definition 3.1 *Let $\mathcal{P}$ be a given class of probability densities on $\mathbb{R}^d$, and let $H \subseteq \mathbb{R}^M$ be a given class of weights. A sequence of positive numbers $\psi_n(M)$ is called* **optimal rate of aggregation** *for H over $\mathcal{P}$ if*

- *for any functions $p_j \in L_2(\mathbb{R}^d), j = 1, \dots, M$, there exists an estimator $\tilde{p}_n$ of p (aggregate) such that*

$$\sup_{p \in \mathcal{P}} \left[R_n(\tilde{p}_n, p) - \inf_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2 \right] \le C\psi_n(M), \tag{3.1}$$

  for any integer $n \ge 1$ and for some constant $C < \infty$ independent of M and n,

and

- *there exist functions $p_j \in L_2(\mathbb{R}^d), j = 1, \dots, M$, such that for all estimators T_n of p, we have*

$$\sup_{p \in \mathcal{P}} \left[R_n(T_n, p) - \inf_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2 \right] \ge c\psi_n(M), \tag{3.2}$$

  for any integer $n \ge 1$ and for some constant $c > 0$ independent of M and n.

When (3.2) *holds, an aggregate $\tilde{p}_n$ satisfying* (3.1) *is called* **rate optimal aggregate** *for H over $\mathcal{P}$.*

Note that this definition applies to aggregation of any functions p_j in $L_2(\mathbb{R}^d)$, they are not necessarily supposed to be probability densities.

Theorems 2.1 and 2.2 provide upper bounds of the type (3.1) with the rate $\psi_n(M) = LM/n$ for linear and convex aggregates $\tilde{p}_n = \tilde{p}_n^{\mathbf{L}}$ and $\tilde{p}_n = \tilde{p}_n^{\mathbf{C}}$ when $\mathcal{P} = \mathcal{P}_0$ and $H = \mathbb{R}^M$ or H is a convex compact subset of $\mathbb{R}^M$. In this section we complement these results by lower bounds of the type (3.2) showing that $\psi_n(M) = LM/n$ is optimal rate of linear and convex aggregation. The proofs will be based on the following lemma which is adapted from Corollary 4.1 of Birgé (1986), p. 281.

Lemma 3.1 *Let $\mathcal{C}$ be a set of functions of the following type*

$$\mathcal{C} = \left\{ f + \sum_{i=1}^{r} \delta_i g_i,\ \delta_i \in \{0,1\},\ i = 1, \dots, r \right\},$$

where the g_i are functions on $\mathbb{R}^d$ with disjoint supports, such that $\int g_i(x)\mathrm{d}x = 0$, f is a probability density on $\mathbb{R}^d$ which is constant on the union of the supports of g_i's, and $f + g_i \geq 0$ for all i. Assume that

$$\min_{1 \leq i \leq r} \|g_i\|^2 \geq \alpha > 0 \quad \textit{and} \quad \max_{1 \leq i \leq r} h^2(f, f + g_i) \leq \beta < 1\,, \tag{3.3}$$

where $h^2(f,g) = (1/2)\int(\sqrt{f(x)} - \sqrt{g(x)})^2 dx$ is the squared Hellinger distance between two probability densities f and g. Then

$$\inf_{T_n} \sup_{p \in \mathcal{C}} R_n(T_n, p) \geq \frac{r\alpha}{4}(1 - \sqrt{2n\beta})$$

where $\inf_{T_n}$ denotes the infimum over all estimators.

Consider first a lower bound for linear aggregation of density estimators. We are going to prove (3.2) with $\psi_n(M) = LM/n$, $\mathcal{P} = \mathcal{P}_0$ and $H = \mathbb{R}^M$. Note first that for $\mathcal{P} = \mathcal{P}_0$ there is a natural limitation on the value $c\psi_n(M)$ on the right hand side of (3.2), whatever is H. In fact, $\inf_{T_n} \sup_{p \in \mathcal{P}_0} \left[R_n(T_n, p) - \inf_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2 \right] \leq \inf_{T_n} \sup_{p \in \mathcal{P}_0} R_n(T_n, p) \leq \sup_{p \in \mathcal{P}_0} R_n(0, p) = \sup_{p \in \mathcal{P}_0} \|p\|^2 \leq L$. Therefore, we must have $c\psi_n(M) \leq L$ where c is the constant in (3.2). For $\psi_n(M) = LM/n$ this means that only the values M such that $M \leq c_0 n$ are allowed, where $c_0 > 0$ is a constant. The upper bounds of Theorems 2.1 and 2.2 are too rough (non-optimal) when $M = M_n$ depends on n and the condition $M \leq c_0 n$ is not satisfied. In the sequel, we will apply those theorems with $M = M_n$ depending on n and satisfying $M_n/n \to 0$, as $n \to \infty$, so that the condition $M \leq c_0 n$ will obviously hold with any finite c_0 for n large enough.

Theorem 3.1 *Let the integers $M \geq 2$ and $n \geq 1$ be such that $M \leq c_0 n$ where c_0 is a positive constant. Then there exist probability densities $p_j \in L_2(\mathbb{R}^d)$, $j = 1, \dots, M$, such that for all estimators T_n of p we have*

$$\inf_{T_n} \sup_{p \in \mathcal{P}_0} \left[R_n(T_n, p) - \inf_{\lambda \in \mathbb{R}^M} \|\mathsf{p}_\lambda - p\|^2 \right] \geq cLM/n \tag{3.4}$$

where $c > 0$ is a constant depending only on c_0.

Proof. Set $r = M - 1 \geq 1$ and fix $0 < a < 1$. Consider the function $\tilde{g}$ defined for any $t \in \mathbb{R}$ by

$$\tilde{g}(t) \triangleq \frac{aL}{2} \mathbb{1}_{[0, \frac{1}{Lr}]}(t) - \frac{aL}{2} \mathbb{1}_{(\frac{1}{Lr}, \frac{2}{Lr})}(t),$$

where $\mathbb{1}_A(\cdot)$ denotes the indicator function of a set A. Let $\{\tilde{g}_j\}_{j=1}^r$ be the family of functions defined by $\tilde{g}_j(t) = \tilde{g}(t - 2(j-1)/Lr)$, $1 \leq j \leq r$. Define also the density $\tilde{f}(t) = (L/2)\mathbb{1}_{[0,2/L]}(t)$, $t \in \mathbb{R}$. For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ consider the functions

$$f(x) = \tilde{f}(x_1) \prod_{k=2}^{d} \mathbb{1}_{[0,1]}(x_k) \quad g_j(x) = \tilde{g}_j(x_1) \prod_{k=2}^{d} \mathbb{1}_{[0,1]}(x_k),\ j = 1, \dots, r\,.$$

Define the probability densities p_j by $p_1 = f$, $p_{j+1} = f + g_j$, $j = 1, \dots, M-1$.

Consider now the set of functions $\mathcal{Q} = \{q_\delta : q_\delta = f + \sum_{j=1}^r \delta_j g_j,\ \delta = (\delta_1, \dots, \delta_r) \in \{0,1\}^r\}$. Clearly, for any $\delta \in \{0,1\}^r$, q_δ satisfies $\int_{\mathbb{R}^d} q_\delta(x)\mathrm{d}x = 1$, $q_\delta \geq 0$ and $\|q_\delta\|_\infty \leq L$. Therefore $\mathcal{Q} \subset \mathcal{P}_0$. Also, $\mathcal{Q} \subset \{\mathsf{p}_\lambda, \lambda \in \mathbb{R}^M\}$. Thus,

$$\inf_{T_n} \sup_{p \in \mathcal{P}_0} \left[R_n(T_n, p) - \inf_{\lambda \in \mathbb{R}^M} \|\mathsf{p}_\lambda - p\|^2 \right] \geq \inf_{T_n} \sup_{p \in \mathcal{Q}} R_n(T_n, p)\,.$$

To prove that $\inf_{T_n} \sup_{p\in\mathcal{Q}} R_n(T_n,p) \geq cLM/n$ we check conditions (3.3) of Lemma 3.1. The first condition in (3.3) is obviously satisfied since

$$\|g_j\|^2 = \int_0^{\frac{2}{Lr}} \tilde{g}^2(t)\mathrm{d}t = \frac{a^2 L}{2r}\,, \quad j=1,\dots,r.$$

To check the second condition in (3.3), note that for $j=1,\dots,r$ we have

$$\begin{aligned} h^2(f,f+g_j) &= \frac{1}{2}\int_0^{\frac{2}{Lr}} \Big(\sqrt{L/2} - \sqrt{L/2+\tilde{g}(t)}\Big)^2 \mathrm{d}t \\ &= \frac{L}{4}\int_0^{\frac{2}{Lr}} \Big(1 - \sqrt{1+(2/L)\tilde{g}(t)}\Big)^2 \mathrm{d}t \\ &= \frac{L}{4}\left[\frac{4}{Lr} - 2\int_0^{\frac{2}{Lr}} \sqrt{1+(2/L)\tilde{g}(t)}\mathrm{d}t\right] \\ &= \frac{1}{r} - \frac{1}{2r}\Big(\sqrt{1+a}+\sqrt{1-a}\Big) \leq \frac{a^2}{2r} \end{aligned}$$

where we used the fact that $\sqrt{1+a}+\sqrt{1-a} \geq 2-a^2$ for $|a|\leq 1$. Define now $\tilde{c}_0 = \max(c_0,3)$ and choose $a^2 = M/(\tilde{c}_0 n) \leq 1$. Then $a^2/(2r) \leq (\tilde{c}_0 n)^{-1}$ for $M \geq 2$. Applying Lemma 3.1 with $\beta = (\tilde{c}_0 n)^{-1}$ and $\alpha = \dfrac{ML}{2\tilde{c}_0 nr}$ we get

$$\inf_{T_n}\sup_{p\in\mathcal{C}} R_n(T_n,p) \geq \frac{1}{8\tilde{c}_0}\left(1-\sqrt{\frac{2}{\tilde{c}_0}}\right)\frac{LM}{n}\,.$$

∎

Theorems 2.1 and 3.1 imply the following result.

Corollary 3.1 *Let the integers $M\geq 2$ and $n\geq 1$ be such that $M\leq c_0 n$ where c_0 is a positive constant. Then $\psi_n(M) = LM/n$ is optimal rate of linear aggregation over $\mathcal{P}_0$ (i.e. the optimal rate of aggregation for $H=\mathbb{R}^M$ over $\mathcal{P}_0$), and $\tilde{p}_n^{\mathbf{L}}$ defined in (2.1) is rate optimal aggregate for $\mathbb{R}^M$ over $\mathcal{P}_0$.*

Consider now a lower bound for convex aggregation. We analyze here only the case $H=\Lambda^M$. Other examples of convex sets H can be treated similarly.

Theorem 3.2 *Let the integers $M\geq 2$ and $n\geq 1$ be are such that $M\leq c_0 n$. Then there exist functions $p_j \in L_2(\mathbb{R}^d), j=1,\dots,M$, such that for all estimators T_n of p we have*

$$\inf_{T_n}\sup_{p\in\mathcal{P}_0}\Big[R_n(T_n,p) - \inf_{\lambda\in\Lambda^M}\|\mathsf{p}_\lambda - p\|^2\Big] \geq cLM/n \tag{3.5}$$

where $c>0$ is a constant depending only on c_0.

Proof. Consider the same family of densities $\mathcal{Q}$ as defined in the proof of Theorem 3.1. We may rewrite it in the form $\mathcal{Q} = \{q_\delta : q_\delta = \lambda_1 Mf + \sum_{j=1}^r \lambda_{j+1} M\delta_j g_j,\ \delta=(\delta_1,\dots,\delta_r)\in\{0,1\}^r\}$ where $\lambda_j = 1/M$, $j=1,\dots M$. Define now $p_1 = Mf$, $p_{j+1} = M(f+g_j)$, $j=1,\dots,M-1$. Since $\sum_{j=1}^M \lambda_j = 1$ we have $\mathcal{Q}\subset\big\{\mathsf{p}_\lambda, \lambda\in\Lambda^M\big\}$. The rest of the proof is identical to that of Theorem 3.1. ∎

Theorems 2.2 and 3.2 imply the following result.

Corollary 3.2 *Let the integers $M \geq 2$ and $n \geq 1$ be such that $M \leq c_0 n$. Then $\psi_n(M) = LM/n$ is optimal rate of convex aggregation over $\mathcal{P}_0$ (i.e. the optimal rate of aggregation for $H = \Lambda^M$ over $\mathcal{P}_0$), and $\tilde{p}_n^{\mathbf{C}}$ is rate optimal aggregate for $H = \Lambda^M$ over $\mathcal{P}_0$.*

Inspection of the proofs of Theorems 3.2 and 3.1 reveals that the least favorable functions p_j used in the lower bound for linear aggregation are uniformly bounded by L, whereas this is not the case for least favorable functions in convex aggregation. It can be shown that, for convex aggregation of functions which are uniformly bounded by L, an elbow appears in the optimal rates of aggregation, with the bound (3.5) still remaining valid for $M \leq \sqrt{n}$. This issue will be treated in a forthcoming paper of the first author.

4 Sample splitting and averaged aggregates

We now come back to the original problem discussed in the introduction. Let $\mathbf{X}_1^m$ denote a subsample of $\mathbf{X}^n = (X_1, \ldots, X_n)$ of size $m \leq n$ (training sample). Take $m < n$ and construct estimators $\hat{p}_{m,1}, \ldots, \hat{p}_{m,M}$ of p based on $\mathbf{X}_1^m$. Then aggregate these estimators using the validation subsample $\mathbf{X}_2^\ell$ of $\mathbf{X}^n$ of size $\ell = n - m$,

$$\left(\mathbf{X}_1^m, \mathbf{X}_2^\ell\right) = \mathbf{X}^n = (X_1, \ldots, X_n).$$

For given $m < n$ the two subsamples can be obtained by different splits. The choice of split is arbitrary, and it may influence the result of estimation. In order to avoid the arbitrariness, we will use a jackknife type procedure averaging the aggregates over different splits. Define a *split* $\mathcal{S}$ of the initial sample $\mathbf{X}^n$ as a mapping

$$\mathcal{S} \ : \ \mathbf{X}^n \mapsto \left(\mathbf{X}_1^m, \mathbf{X}_2^\ell\right).$$

Denote by $\mathbf{X}_{1,\mathcal{S}}^m, \mathbf{X}_{2,\mathcal{S}}^\ell$ subsamples obtained for a fixed split $\mathcal{S}$ and consider an arbitrary set of splits $\mathbf{S}$. It can be, for example, the set of all splits. Define $\tilde{p}_n^{\mathcal{S}}$ as a linear or convex aggregate ($\tilde{p}_n^{\mathbf{L}}$ or $\tilde{p}_n^{\mathbf{C}}$ respectively) based on the validation sample $\mathbf{X}_{2,\mathcal{S}}^\ell$ and on the initial set of estimators $p_j = \hat{p}_{m,j}^{\mathcal{S}}, j = 1, \ldots, M$, where each of $\hat{p}_{m,j}^{\mathcal{S}}$'s is constructed from the training sample $\mathbf{X}_{1,\mathcal{S}}^m$. Introduce the following *averaged aggregate* estimator:

$$\tilde{p}_n^{\mathbf{S}} \triangleq \frac{1}{\mathrm{card}(\mathbf{S})} \sum_{\mathcal{S} \in \mathbf{S}} \tilde{p}_n^{\mathcal{S}}. \tag{4.1}$$

Let H be either $\mathbb{R}^M$ or a convex compact subset of $\mathbb{R}^M$. Define

$$\Delta_{\ell,M} = \begin{cases} LM/\ell & \text{if } H = \mathbb{R}^M, \\ 4LM/\ell & \text{if } H \text{ is a convex compact subset of } \mathbb{R}^M. \end{cases}$$

We get the following corollary of Theorems 2.1 and 2.2.

Corollary 4.1 *Let $m < n$, $\ell = n - m$, and let H be either $\mathbb{R}^M$ or a convex compact subset of $\mathbb{R}^M$. Let $\mathbf{S}$ be an arbitrary set of splits. Assume that $\hat{p}_{m,1}^{\mathcal{S}}, \ldots, \hat{p}_{m,M}^{\mathcal{S}} \in L_2(\mathbb{R}^d)$ for fixed $\mathbf{X}_{1,\mathcal{S}}^m$, $\forall \mathcal{S} \in \mathbf{S}$, and that $p \in \mathcal{P}_0$. Then the averaged aggregate* (4.1) *satisfies*

$$R_n(\tilde{p}_n^{\mathbf{S}}, p) \leq \inf_{\lambda \in H} R_m\Big(\sum_{j=1}^M \lambda_j \hat{p}_{m,j}, p\Big) + \Delta_{\ell,M} \tag{4.2}$$

for any integers $M \geq 2$ and $n \geq 1$.

Proof. For any fixed $\mathcal{S} \in \mathbf{S}$ and for a fixed training subsample $\mathbf{X}_{1,\mathcal{S}}^m$ inequalities (2.2) and (2.4) imply

$$E_p^{\ell,\mathcal{S}} \|\tilde{p}_n^{\mathcal{S}} - p\|^2 \leq \min_{\lambda \in H} \Big\| \sum_{j=1}^M \lambda_j \hat{p}_{m,j} - p \Big\|^2 + \Delta_{\ell,M}, \quad \forall\, p \in \mathcal{P}_0, \tag{4.3}$$

where $E_p^{\ell,\mathcal{S}}$ denotes the expectation w.r.t. the distribution of the validation sample $\mathbf{X}_{2,\mathcal{S}}^{\ell}$ when the true density is p. Taking expectations of both sides of (4.3) w.r.t. the training sample $\mathbf{X}_{1,\mathcal{S}}^{m}$ we get

$$R_n(\tilde{p}_n^{\mathcal{S}}, p) \le \inf_{\lambda \in H} R_m\Big(\sum_{j=1}^{M} \lambda_j \hat{p}_{m,j}, p\Big) + \Delta_{\ell,M}. \tag{4.4}$$

The right hand side here does not depend on $\mathcal{S}$. By Jensen's inequality,

$$R_n(\tilde{p}_n^{\mathbf{S}}, p) \le \frac{1}{\mathrm{card}(\mathbf{S})} \sum_{\mathcal{S} \in \mathbf{S}} R_n(\tilde{p}_n^{\mathcal{S}}, p).$$

This and (4.4) yield (4.2). ∎

5 Kernel aggregates for density estimation

Here we apply the results of the previous sections to aggregation of kernel density estimators. Let $\hat{p}_{m,h}$ denote a kernel density estimator based on $\mathbf{X}_1^m$ with $m \le n$,

$$\hat{p}_{m,h}(x) \triangleq \frac{1}{mh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \mathbb{1}_{\{\mathbf{X}_1^m\}}(X_i), \quad x \in \mathbb{R}^d, \tag{5.1}$$

where $h > 0$ is a bandwidth and $K \in L_2(\mathbb{R}^d)$ is a kernel. The notation $\hat{p}_{m,h}$ is slightly inconsistent with $\hat{p}_{m,j}$ used above but this will not cause ambiguity in what follows. In order to cover such examples as the sinc kernel we will not assume that K is integrable.

Define $h_0 = (n \log n)^{-1/d}$, $a_n = a_0 / \log n$, where $a_0 > 0$ is a constant, and M such that

$$M - 2 = \max\left\{j \in \mathbb{N} : h_0(1 + a_n)^j < 1\right\}.$$

It is easy to see that $M \le c_4 (\log n)^2$, where $c_4 > 0$ is a constant depending only on a_0 and d. Consider a grid $\mathcal{H}$ on $[0, 1]$ with a weakly geometrically increasing step:

$$\mathcal{H} \triangleq \{h_0, h_1, \dots, h_{M-1}\},$$

where $h_j = (1 + a_n)^j h_0$, $j = 1, \dots, M-2$, and $h_{M-1} = 1$. Fix now an arbitrary family of splits $\mathbf{S}$ such that, for $n \ge 3$,

$$m = \lfloor n\left(1 - (\log n)^{-1}\right) \rfloor \quad \text{and} \quad \ell = n - m \ge \frac{n}{\log n},$$

where $\lfloor x \rfloor$ denotes the integer part of x.

Define $\tilde{p}_n^{\mathbf{S},K}$ as the linear or convex (with $H = \Lambda^M$) averaged aggregate $\tilde{p}_n^{\mathbf{S}}$ where the initial estimators are taken in the form $p_j = \hat{p}_{m,h_{j-1}}, j = 1, \dots, M$, with $\hat{p}_{m,h}$ given by (5.1). Since $\Delta_{\ell,M} \le 4LM/\ell$ we get from (4.2) that, under the assumptions of Corollary 4.1,

$$R_n(\tilde{p}_n^{\mathbf{S},K}, p) \le \min_{h \in \mathcal{H}} R_m(\hat{p}_{m,h}, p) + \Delta_{\ell,M} \le \min_{h \in \mathcal{H}} R_m(\hat{p}_{m,h}, p) + \frac{4c_4 (\log n)^3}{n}. \tag{5.2}$$

We now give a theorem that extends (5.2) to the n-sample oracle risk $\inf_{h>0} R_n(\hat{p}_{n,h}, p)$ instead of $\min_{h \in \mathcal{H}} R_m(\hat{p}_{m,h}, p)$. Denote by $\mathcal{F}[f]$ the Fourier transform defined for $f \in L_2(\mathbb{R}^d)$ and normalized in such a way that its restriction to $f \in L_2(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$ has the form $\mathcal{F}[f](t) = \int_{\mathbb{R}^d} \mathrm{e}^{ix^T t} f(x) \mathrm{d}x$, $t \in \mathbb{R}^d$. In the sequel $\varphi = \mathcal{F}[p]$ denotes the characteristic function associated to p.

Theorem 5.1 *Assume that p satisfies $\|p\|_\infty \le L$ with $0 < L < \infty$ and let $K \in L_2(\mathbb{R}^d)$ be a kernel such that a version of its Fourier transform $\mathcal{F}[K]$ takes values in $[0,1]$ and satisfies the monotonicity condition $\mathcal{F}[K](h't) \ge \mathcal{F}[K](ht),\ \forall t \in \mathbb{R}^d,\ h > h' > 0$. Then there exists an integer $n_0 = n_0(L, \|K\|) \ge 4$ such that for $n \ge n_0$ the averaged aggregate $\tilde{p}_n^{\mathbf{S},K}$ satisfies the oracle inequality*

$$R_n(\tilde{p}_n^{\mathbf{S},K}, p) \le \left(1 + c_5(\log n)^{-1}\right) \inf_{h>0} R_n(\hat{p}_{n,h}, p) + c_6 \frac{(\log n)^3}{n}\,, \tag{5.3}$$

where c_5 is a positive constant depending only on d and a_0, and $c_6 > 0$ depends only on $L, \|K\|, d$ and a_0.

Proof. Assume throughout that $n \ge 4$. First note that (5.3) deduces from (5.2) and from the following two inequalities that we are going to prove below:

$$\inf_{h\in[h_0,h_{M-1}]} R_n(\hat{p}_{n,h}, p) \le \inf_{h>0} R_n(\hat{p}_{n,h}, p) + \|K\|^2 \frac{\log n}{n}\,, \tag{5.4}$$

$$\min_{j=1,\dots,M} R_m(\hat{p}_{m,h_{j-1}}, p) \le \left(1 + c_5(\log n)^{-1}\right) \inf_{h\in[h_0,h_{M-1}]} R_n(\hat{p}_{n,h}, p) + \frac{c_5 L}{n \log n}\,. \tag{5.5}$$

In turn, (5.4) follows if we show that

$$\inf_{h\in[h_0,h_{M-1}]} R_n(\hat{p}_{n,h}, p) \le \inf_{0<h<h_0} R_n(\hat{p}_{n,h}, p), \tag{5.6}$$

$$\inf_{h\in[h_0,h_{M-1}]} R_n(\hat{p}_{n,h}, p) \le \inf_{h>h_{M-1}} R_n(\hat{p}_{n,h}, p) + \|K\|^2 \frac{\log n}{n}\,. \tag{5.7}$$

Thus, it remains to prove (5.5) – (5.7). We will use the following Fourier representation for MISE of kernel estimators that can be easily obtained from Plancherel's formula (it is a multivariate extension of the representation for $d = 1$ given, e.g., in Golubev (1992) and in Wand and Jones (1995), p.55):

$$\begin{aligned} R_n(\hat{p}_{n,h}, p) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \Big(& |1 - \mathcal{F}[K](ht)|^2 |\varphi(t)|^2 \\ & + \frac{1}{n}\left(1 - |\varphi(t)|^2\right) \left|\mathcal{F}[K](ht)\right|^2 \Big) \mathrm{d}t. \end{aligned} \tag{5.8}$$

Furthermore, using Plancherel's formula we get

$$\begin{aligned} &\int_{\mathbb{R}^d} |\varphi(t)|^2 \mathrm{d}t = (2\pi)^d \int_{\mathbb{R}^d} p^2(x) \mathrm{d}x \le (2\pi)^d L, \\ &\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\mathcal{F}[K](ht)|^2 \mathrm{d}t = h^{-d} \|K\|^2,\ \forall h > 0. \end{aligned} \tag{5.9}$$

Proof of (5.6). Using (5.8), (5.9) and the fact that $0 \le \mathcal{F}[K](t) \le 1, \forall t \in \mathbb{R}^d$, for any $h < h_0 = (n \log n)^{-1/d}$ we obtain

$$R_n(\hat{p}_{n,h}, p) \ge \frac{1}{n(2\pi)^d} \int_{\mathbb{R}^d} \left(1 - |\varphi(t)|^2\right) \left|\mathcal{F}[K](ht)\right|^2 \mathrm{d}t \ge \frac{\|K\|^2}{nh^d} - \frac{L}{n} \ge \|K\|^2 \log n - \frac{L}{n}. \tag{5.10}$$

On the other hand, since $h_{M-1} = 1$ we get

$$R_n(\hat{p}_{n,h_{M-1}}, p) \le \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(|1 - \mathcal{F}[K](t)|^2 |\varphi(t)|^2 + \frac{1}{n}\left|\mathcal{F}[K](t)\right|^2 \right) \mathrm{d}t \le L + \frac{\|K\|^2}{n}. \tag{5.11}$$

The right hand side of (5.10) is larger than that of (5.11) for $n \geq n_0$, where n_0 depends only on L and $\|K\|$. Thus, (5.6) is valid for $n \geq n_0$.

Proof of (5.7). Clearly, (5.7) follows if we show that

$$R_n(\hat{p}_{n,h'}, p) \leq \inf_{h > h_{M-1}} R_n(\hat{p}_{n,h}, p) + \|K\|^2 \frac{\log n}{n}$$

for $h' = (\log n)^{-1/d} \in [h_0, h_{M-1}]$. To prove this inequality, first note that, by the monotonicity of $h \mapsto \mathcal{F}[K](ht)$, we have

$$\int_{\mathbb{R}^d} |1 - \mathcal{F}[K](ht)|^2 |\varphi(t)|^2 \mathrm{d}t \geq \int_{\mathbb{R}^d} |1 - \mathcal{F}[K](h't)|^2 |\varphi(t)|^2 \mathrm{d}t, \quad \forall h > h_{M-1}.$$

This, together with (5.8) and the second equality in (5.9), yields that, for any $h > h_{M-1}$,

$$R_n(\hat{p}_{n,h}, p) \geq R_n(\hat{p}_{n,h'}, p) - \frac{1}{n(2\pi)^d} \int_{\mathbb{R}^d} \left(1 - |\varphi(t)|^2\right) \left|\mathcal{F}[K](h't)\right|^2 \mathrm{d}t \geq R_n(\hat{p}_{n,h'}, p) - \|K\|^2 \frac{\log n}{n}\,.$$

Proof of (5.5). We will show that for any $h \in [h_0, h_{M-1}]$ one has

$$R_m(\hat{p}_{m,\overline{h}}, p) \leq \left(1 + c_5 (\log n)^{-1}\right) R_n(\hat{p}_{n,h}, p) + \frac{c_5 L}{n \log n} \tag{5.12}$$

where $\overline{h} \triangleq \max\{h_j : h_j \leq h\}$. Clearly, this implies (5.5). To prove (5.12), note that if $h_j \leq h < h_{j+1}$ we have $\overline{h} = h_j$, $h/h_j \leq 1 + a_n = 1 + a_0/\log n$. Therefore, (5.8) and the monotonicity of $h \mapsto \mathcal{F}[K](ht)$ imply

$$\begin{aligned}
R_m(\hat{p}_{m,h_j}, p) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(\left[1 - \mathcal{F}[K](h_j t)\right]^2 |\varphi(t)|^2 + \frac{1}{m} \left[\mathcal{F}[K](h_j t)\right]^2 \right) \mathrm{d}t \\
&\qquad - \frac{1}{(2\pi)^d m} \int_{\mathbb{R}^d} |\varphi(t)|^2 \left[\mathcal{F}[K](h_j t)\right]^2 \mathrm{d}t \\
&\leq \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \Big(\left[1 - \mathcal{F}[K](ht)\right]^2 |\varphi(t)|^2 + \frac{1}{n} \left[\mathcal{F}[K](ht)\right]^2 \frac{nh^d}{mh_j^d} \Big) \mathrm{d}t \\
&\qquad - \frac{1}{(2\pi)^d n} \int_{\mathbb{R}^d} |\varphi(t)|^2 \left[\mathcal{F}[K](ht)\right]^2 \mathrm{d}t \\
&\leq \frac{nh^d}{mh_j^d} R_n(\hat{p}_{n,h}, p) + \left(\frac{nh^d}{mh_j^d} - 1 \right) \frac{1}{(2\pi)^d n} \int_{\mathbb{R}^d} |\varphi(t)|^2 \left[\mathcal{F}[K](ht)\right]^2 \mathrm{d}t.
\end{aligned}$$

Using here the fact that $(n/m)(h/h_j)^d \leq (1 - (\log n)^{-1} - n^{-1})(1 + a_0/\log n)^d \leq 1 + c_5 (\log n)^{-1}$ for $n \geq 4$ and for a constant $c_5 > 0$ depending only on d, a_0, and applying (5.9) we get (5.12). ∎

Corollary 5.1 *Let the assumptions of Theorem 5.1 be satisfied, and let* $\inf_{h>0} R_n(\hat{p}_{n,h}, p) \geq c n^{-1+\alpha}$, *for some* $c > 0, \alpha > 0$. *Then*

$$R_n(\tilde{p}_n^{\mathbf{S},K}, p) \leq \inf_{h>0} R_n(\hat{p}_{n,h}, p)(1 + o(1)), \qquad n \to \infty. \tag{5.13}$$

Using the argument as in Stone (1984) it is not hard to check that the assumption of Corollary 5.1 is valid for any non-negative kernel. In the one-dimensional case it also holds for any kernel satisfying the conditions of Lemma 4.1 in Rigollet (2006). On the difference to Rigollet (2006), Corollary 5.1 applies to multidimensional density estimation.

Theorem 5.1 and Corollary 5.1 show that linear or convex aggregate $\tilde{p}_n^{\mathbf{S},K}$ mimics the best kernel estimator, without being itself in the class of kernel estimators with data-driven bandwidth. Another method with such a property has been suggested recently by Rigollet (2006) in the one-dimensional case; it is based on a block Stein procedure in the Fourier domain.

The results of this section can be compared to the work on optimality of bandwidth selection in the L_2 sense for kernel density estimation. A key reference is the theorem of Stone (1984) establishing that, under some assumptions,

$$\lim_{n\to\infty} \frac{\|\hat{p}_{n,h_n} - p\|^2}{\inf_{h>0} \|\hat{p}_{n,h} - p\|^2} = 1, \quad \text{with probability 1,}$$

where h_n is a data-dependent bandwidth chosen by cross-validation. Our results are of a different type, because they treat convergence of expected risk rather than almost sure convergence. In addition, we provide oracle inequalities with precisely defined remainder terms that hold under mild assumptions on the density and on the kernel. Unlike Stone (1984), we do not require the one-dimensional marginals of the density p to be uniformly bounded. Wegkamp (1999) considers model selection approach to bandwidth choice for kernel density estimation. His main result is of the form of (5.13) with a model selection kernel estimator in place of $\tilde{p}_n^{\mathbf{S},K}$, but it is valid for bounded, nonnegative, Lipschitz kernels with compact support (similar assumptions on K are imposed by Stone (1984)). Our result covers kernels with unbounded support, for example, the Gaussian and Silverman's kernels that are often implemented, and Pinsker's kernel that gives sharp minimax adaptive estimators on Sobolev classes (cf. Section 6 below). In a recent work of Dalelane (2004) the choice of bandwidth and of the kernel by cross-validation is investigated for the one-dimensional case ($d = 1$). She provides an oracle inequality similar to (5.3) with a remainder term of the order $n^{\delta-1}$, $0 < \delta < 1$, instead of $(\log n)^3/n$ that we have here.

All these papers consider the model selection approach, i.e., they study estimators with a single data-driven bandwidth chosen from a set of candidate bandwidths. Our approach is different since we estimate the density by a linear or convex combination of kernel estimators with bandwidths in the candidate set. Simulations (see Section 7 below) show that in most cases one of these estimators gets highly dominant weight in the resulting mixture. However, inclusion of other estimators with some smaller weights allows one to treat more efficiently densities with inhomogeneous smoothness.

6 Sharp minimax adaptivity of kernel aggregates

In this section we show that the kernel aggregate defined in Section 5 is sharp minimax adaptive over a scale of Sobolev classes of densities.

For any $\beta > 0$, $Q > 0$ and any integer $d \geq 1$ define the Sobolev classes of densities on $\mathbb{R}^d$ by

$$\Theta(\beta, Q) \triangleq \left\{ p : \mathbb{R}^d \to \mathbb{R} \;\middle|\; p \geq 0, \int_{\mathbb{R}^d} p(x)\mathrm{d}x = 1, \int_{\mathbb{R}^d} \|t\|_d^{2\beta} |\varphi(t)|^2 \mathrm{d}t \leq Q \right\},$$

where $\|\cdot\|_d$ denotes the Euclidean norm in $\mathbb{R}^d$ and $\varphi = \mathcal{F}[p]$. Consider the Pinsker kernel K_β, i.e. the kernel having the Fourier transform

$$\mathcal{F}[K_\beta](t) \triangleq \left(1 - \|t\|_d^\beta\right)_+, \quad t \in \mathbb{R}^d,$$

where $x_+ = \max(x, 0)$. Set

$$C^* = \frac{[Q(2\beta + d)]^{\frac{d}{2\beta+d}}}{d(2\pi)^d} \left(\frac{\beta S_d}{\beta + d}\right)^{\frac{2\beta}{2\beta+d}} \tag{6.1}$$

where $S_d = 2\pi^{d/2}/\Gamma(d/2)$ is the surface of a sphere of radius 1 in $\mathbb{R}^d$. For $d = 1$ the value C^* equals to the Pinsker constant [Pinsker (1980), see also Tsybakov (2004), Chapter 3].

Corollary 6.1 *For any integer $d \geq 1$ and any $\beta > d/2,\ Q > 0$, the averaged linear or convex kernel aggregate $\tilde{p}_n^{\mathbf{S},K_\beta}$ defined in Section 5 satisfies*

$$\sup_{p\in\Theta(\beta,Q)} R_n(\tilde{p}_n^{\mathbf{S},K_\beta}, p) \leq C^* n^{-\frac{2\beta}{2\beta+d}}(1+o(1)), \qquad n\to\infty,$$

where C^* is defined in (6.1).

Proof. Denote by $\hat{p}_{n,h}$ the kernel density estimator defined in (5.1) with $m = n$ and $K = K_\beta$. Using (5.8) and the fact that $0 \leq \mathcal{F}[K_\beta](t) \leq 1$, $\forall\, t \in \mathbb{R}^d$, we get

$$\begin{aligned} R_n(\hat{p}_{n,h}, p) &\leq \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left(|1-\mathcal{F}[K_\beta](ht)|^2|\varphi(t)|^2 + \frac{1}{n}\left|\mathcal{F}[K_\beta](ht)\right|^2\right)\mathrm{d}t \\ &\leq \frac{1}{(2\pi)^d}\left(Qh^{2\beta} + \frac{1}{n}\int_{\mathbb{R}^d}\left|\mathcal{F}[K_\beta](ht)\right|^2\mathrm{d}t\right), \forall\, h>0,\ p\in\Theta(\beta,Q). \end{aligned} \tag{6.2}$$

Now, choose h satisfying

$$\int_{\mathbb{R}^d}\|t\|_d^\beta \mathcal{F}[K_\beta](ht)\mathrm{d}t = Qnh^\beta. \tag{6.3}$$

The solution of (6.3) is

$$h = D^* n^{-\frac{1}{2\beta+d}} \quad \text{where} \quad D^* = \left(\frac{\beta S_d}{Q(\beta+d)(2\beta+d)}\right)^{\frac{1}{2\beta+d}}.$$

With h satisfying (6.3), inequality (6.2) becomes

$$\begin{aligned} R_n(\hat{p}_{n,h}, p) &\leq \frac{1}{(2\pi)^d n}\int_{\mathbb{R}^d}\mathcal{F}[K_\beta](ht)\left[\mathcal{F}[K_\beta](ht) + \|ht\|_d^\beta\right]\mathrm{d}t \\ &= \frac{1}{(2\pi)^d nh^d}\int_{\mathbb{R}^d}\mathcal{F}[K_\beta](t)\mathrm{d}t \\ &= \frac{1}{(2\pi)^d nh^d}\int_0^1 \left(1-r^\beta\right) r^{d-1} S_d \mathrm{d}r \\ &= C^* n^{-\frac{2\beta}{2\beta+d}}. \end{aligned}$$

Thus,

$$\inf_{h>0} R_n(\hat{p}_{n,h}, p) \leq C^* n^{-\frac{2\beta}{2\beta+d}}, \quad \forall\ p\in\Theta(\beta,Q). \tag{6.4}$$

Note that the kernel $K = K_\beta$ satisfies the conditions of Theorem 5.1, and it is easy to see that for $\beta > d/2$ there exists a constant $0 < L < \infty$ such that $\|p\|_\infty \leq L$ for all $p \in \Theta(\beta, Q)$. Thus, (5.3) holds, and to prove the corollary it suffices to take suprema of both sides of (5.3) over $p \in \Theta(\beta, Q)$ and to use (6.4). ∎

Along with Corollary 6.1, for any $\beta > d/2,\ Q > 0$ the following lower bound holds:

$$\inf_{T_n}\sup_{p\in\Theta(\beta,Q)} R_n(T_n, p) \geq C^* n^{-\frac{2\beta}{2\beta+d}}(1+o(1)), \qquad n\to\infty, \tag{6.5}$$

where C^* is defined in (6.1) and $\inf_{T_n}$ denotes the infimum over all estimators of p. For $d = 1$ the bound (6.5) can be deduced from the results of Golubev (1991, 1992); it is also proven explicitly in Schipper (1996) (for integer β) and in Rigollet (2006), Dalelane (2004) (for all $\beta > 1/2$). For $d > 1$ the bound (6.5) can be found for a slightly different but essentially analogous minimax setup in Efromovich (2000).

Corollary 6.1 and the lower bound (6.5) imply that the estimator $\tilde{p}_n^{\mathbf{S},K_\beta}$ is asymptotically minimax in the exact sense (with the constant) over the Sobolev class of densities $\Theta(\beta, Q)$ and is adaptive to Q for any given β. However, $\tilde{p}_n^{\mathbf{S},K_\beta}$ is not adaptive to the unknown smoothness β since the Pinsker kernel K_β depends on β.

To get adaptation to β, we need to push aggregation one step forward: we will aggregate kernel density estimators not only for different bandwidths but also for different kernels. To this end, we refine the notation $\hat{p}_{n,h}$ of (5.1) to $\hat{p}_{n,h,K}$, indicating the dependence of the density estimator both on kernel K and bandwidth h. For a family of $N \geq 2$ kernels, $\mathcal{K} = \{K_{(1)}, \dots, K_{(N)}\}$, define $\tilde{p}_n^{\mathbf{S},\mathcal{K}}$ as the linear or convex averaged aggregate where the initial estimators are taken in the collection of kernel density estimators $\{\hat{p}_{n,h,K}, K \in \mathcal{K}, h \in \mathcal{H}\}$. Thus, we aggregate now NM estimators instead of M. The following corollary is obtained by the same argument as Theorem 5.1, by merely inserting the minimum over $K \in \mathcal{K}$ in the oracle inequality and by replacing $\|K\|$ with its upper or lower bounds in the remainder terms.

Corollary 6.2 *Assume that p satisfies $\|p\|_\infty \leq L$ with $0 < L < \infty$ and let $\mathcal{K} = \{K_{(1)}, \dots, K_{(N)}\}$ be a family of kernels satisfying the assumptions of Theorem 5.1 and such that there exist constants $0 < \underline{c} < \bar{c} < \infty$ with $\underline{c} < \|K_{(j)}\| < \bar{c}$, $j = 1, \dots, N$. Then there exists an integer $n_1 = n_1(L, \underline{c}, \bar{c}) \geq 4$ such that for $n \geq n_1$ the averaged aggregate $\tilde{p}_n^{\mathbf{S},\mathcal{K}}$ satisfies the oracle inequality*

$$R_n(\tilde{p}_n^{\mathbf{S},\mathcal{K}}, p) \leq \left(1 + c_5(\log n)^{-1}\right) \min_{K \in \mathcal{K}} \inf_{h>0} R_n(\hat{p}_{n,h,K}, p) + c_7 \frac{N(\log n)^3}{n}, \tag{6.6}$$

where $c_5 > 0$ is the same constant as in Theorem 5.1, and $c_7 > 0$ depends only on $L, \underline{c}, \bar{c}, d$ and a_0.

Consider now a particular family of kernels $\mathcal{K}$. Define $\mathcal{B} = \{\beta_1, \dots, \beta_N\}$ where $\beta_1 = d/2$, $\beta_j = \beta_{j-1} + N^{-1/2}$, $j = 2, \dots, N$, and let $\mathcal{K}_\mathcal{B} = \{K_b,\ b \in \mathcal{B}\}$ be a family of Pinsker kernels indexed by $b \in \mathcal{B}$. We will later assume that $N = N_n \to \infty$, as $n \to \infty$, but for the moment assume that $N \geq 2$ is fixed. Note that $\mathcal{K} = \mathcal{K}_\mathcal{B}$ satisfies the assumptions of Corollary 6.2. In fact,

$$\|K_\beta\|^2 = S_d Q_d(\beta) \quad \text{where} \quad Q_d(\beta) = \frac{1}{d} - \frac{2}{\beta + d} + \frac{1}{2\beta + d},$$

and

$$\frac{1}{6d} \leq Q_d(\beta) \leq \frac{1}{d}, \quad \forall\ \beta \geq d/2. \tag{6.7}$$

Thus, the oracle inequality (6.6) holds with $\mathcal{K} = \mathcal{K}_\mathcal{B}$. We will now prove that, under the assumptions of Corollary 6.2 the linear or convex aggregate $\tilde{p}_n^{\mathbf{S},\mathcal{K}_\mathcal{B}}$ with the initial estimators in $\{\hat{p}_{n,h,K}, K \in \mathcal{K}_\mathcal{B}, h \in \mathcal{H}\}$ satisfies the following inequality where β in the oracle risk varies continuously:

$$R_n(\tilde{p}_n^{\mathbf{S},\mathcal{K}_\mathcal{B}}, p) \leq \left(1 + \frac{c_5}{\log n}\right)\left(1 + \frac{6}{\sqrt{N}}\right) \inf_{\substack{h>0 \\ d/2 < \beta < \beta_N}} R_n(\hat{p}_{n,h,K_\beta}, p) + c_8 \frac{N(\log n)^3}{n}. \tag{6.8}$$

Fix $\beta \in (d/2, \beta_N), Q > 0$ and $p \in \Theta(\beta, Q)$. Define $\bar{\beta} = \min\{\beta_j \in \mathcal{B} : \beta_j > \beta\}$. In view of (6.6) with $\mathcal{K} = \mathcal{K}_\mathcal{B}$, to prove (6.8) it is sufficient to show that for any $h > 0$ one has

$$R_n(\hat{p}_{n,h,K_{\bar{\beta}}}, p) \leq (1 + 6N^{-1/2}) \left(R_n(\hat{p}_{n,h,K_\beta}, p) + \frac{L}{n}\right). \tag{6.9}$$

Using (5.8) and the inequality $\bar{\beta} > \beta$ we get

$$R_n(\hat{p}_{n,h,K_{\bar{\beta}}}, p) \leq R_n(\hat{p}_{n,h,K_\beta}, p) + \mathcal{I}(\bar{\beta}) - \mathcal{I}(\beta) \tag{6.10}$$

where

$$\mathcal{I}(\beta) \triangleq \frac{1}{(2\pi)^d n} \int_{\mathbb{R}^d} (1 - \|ht\|_d^\beta)_+^2 \mathrm{d}t = \frac{\|K_\beta\|^2}{(2\pi)^d n h^d} = \frac{S_d}{(2\pi)^d n h^d} Q_d(\beta).$$

Now, $Q_d(\bar\beta) = Q_d(\beta) + (\bar\beta - \beta)Q_d'(b_0)$ for some $b_0 \in [\beta, \bar\beta]$. Using (6.7) and the inequality $|Q_d'(\beta)| \le 1/d^2$ valid for all $\beta > d/2$, we find that

$$Q_d(\bar\beta) \le Q_d(\beta) + 6(\bar\beta - \beta)Q_d(\beta) \le (1 + 6N^{-1/2})Q_d(\beta)\,.$$

Therefore,

$$\mathcal{I}(\bar\beta) \le (1 + 6N^{-1/2})\mathcal{I}(\beta). \tag{6.11}$$

Also, in view of (5.8) and (5.9) we have

$$\mathcal{I}(\beta) \le R_n(\hat p_{n,h,K_\beta}, p) + \frac{L}{n}. \tag{6.12}$$

Combining (6.10), (6.11) and (6.12) we obtain (6.9), thus proving (6.8).

Corollary 6.3 *Assume that* $\mathrm{Card}(\mathcal{K}_\mathcal{B}) = N_n$ *where* $\lim_{n\to\infty} N_n = \infty$ *and* $\limsup_{n\to\infty} N_n/(\log n)^\nu < \infty$ *for some* $\nu > 0$. *Then for any integer* $d \ge 1$ *and any* $\beta > d/2$, $Q > 0$, *the averaged linear or convex kernel aggregate* $\tilde p_n^{\mathbf{S},\mathcal{K}_\mathcal{B}}$ *satisfies*

$$\sup_{p\in\Theta(\beta,Q)} R_n(\tilde p_n^{\mathbf{S},\mathcal{K}_\mathcal{B}}, p) \le C^* n^{-\frac{2\beta}{2\beta+d}}(1 + o(1)), \qquad n \to \infty,$$

where C^* *is defined in* (6.1).

Proof. Fix $\beta > d/2, Q > 0$. Let n be large enough to guarantee that $\beta < \beta_{N_n}$. Then the infimum on the right in (6.8) is smaller or equal to $C^* n^{-\frac{2\beta}{2\beta+d}}$ for all $p \in \Theta(\beta, Q)$ [cf. (6.4)]. To conclude the proof, it suffices to take suprema of both sides of (6.8) over $p \in \Theta(\beta, Q)$ and then pass to the limit as $n \to \infty$. ∎

Corollary 6.3 and the lower bound (6.5) imply that the aggregate $\tilde p_n^{\mathbf{S},\mathcal{K}_\mathcal{B}}$ is asymptotically minimax in the exact sense (with the constant) over all Sobolev classes of densities with $\beta > d/2$, $Q > 0$, and thus it is sharp adaptive (recall that its construction does not depend on the parameters Q and β of the class).

7 Simulations

Here we discuss the results of simulations for the averaged convex kernel aggregate with $H = \Lambda^M$ in the one-dimensional case. We focus on convex aggregation because simulations of linear aggregates show less numerical stability. The set of splits $\mathbf{S}$ is reduced to 10 random splits of the sample since we observed that the estimator is already stable for this number (cf. Figure 3). In the default simulations each sample is divided into two subsamples of equal sizes. The samples are drawn from 6 densities that can be classified in the following three groups.

- Common reference densities: the standard Gaussian density and the standard exponential density.
- Gaussian mixtures from Marron and Wand (1992) that are known to be difficult to estimate. We consider the Claw density and the Smooth Comb density.
- Densities with highly inhomogeneous smoothness. We consider two densities referenced to as dens1 and dens2 that are both mixtures of the standard Gaussian density $\varphi(\cdot)$ and of an oscillating density. They are defined as

  $$0.5\varphi(\cdot) + 0.5\sum_{i=1}^{T} \mathbb{I}_{\left(\frac{2(i-1)}{T}, \frac{2i-1}{T}\right]}(\cdot)\,,$$

  where $T = 14$ for dens1 and $T = 10$ for dens2.

We used the procedure defined in Section 5 to aggregate 6 kernel density estimators constructed with the Gaussian $\mathcal{N}(0,1)$ kernel K and with bandwidths h from the set $\mathcal{H} = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$. This procedure is further called *pure kernel aggregation* and quoted as `AggPure`. Another estimator that we analyze is `AggStein` procedure: it aggregates 7 estimators, namely the same 6 kernel estimators as for `AggPure` to which we add the block Stein density estimator described in Rigollet (2006). The optimization problem (2.3) that provides aggregates is solved numerically by a quadratic programming solver under linear constraints: here we used the package `quadprog` of R. Our simulation study shows that `AggPure` and `AggStein` have a good performance for moderate sample sizes and are reasonable competitors to kernel density estimators with common bandwidth selectors.

We start the simulation by a comparison of the Monte-Carlo mean integrated squared squared error (MISE) of `AggPure` and `AggStein` with benchmarks. The MISE has been computed by averaging integrated squared errors of 200 aggregate estimators calculated from different samples of size 50, 100, 200 and 500. We compared the performance of the convex aggregates and kernel estimators with common data-driven bandwidth selectors and Gaussian $\mathcal{N}(0,1)$ kernel. The following bandwidth selectors are taken from the default package `stats` of the R software.

- DPI that implements the direct plug-in method of Sheather and Jones (1991) to select the bandwidth using pilot estimation of derivatives.
- UCV and BCV that implement unbiased and biased cross-validation respectively (see, e.g., Wand and Jones (1995)).
- Nrd0 that implements Silverman's rule-of-thumb [cf. Silverman (1986), page 48]. It defaults the choice of bandwidth to 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power.

These descriptions correspond to the function `bandwidth` in R which also allows for another choice of rule-of-thumb called `Nrd`. It is a modification of `Nrd0` given by Scott (1992), using factor 1.06 instead of 0.9. In our case, on the tested densities and sample sizes, this always leads to a MISE greater than that of `Nrd0` except for the Gaussian density for which it is tailored. For this density, the performance of `Nrd` is presented instead of that of `Nrd0`.

The results are reported in Tables 1 to 3 where we included also the MISE of the block Stein density estimator described in Rigollet (2006) and the oracle risk which is defined as the minimum MISE of kernel density estimators over the grid $\mathcal{H}$. It is, in general, greater than the convex oracle risk, that is why it sometimes slightly exceeds the MISE of convex aggregates or of other estimators that mimic more powerful oracles for specific densities (such as DPI or `Nrd` for the Gaussian density).

	50	100	150	200	500
AggPure	0.020	0.011	0.008	0.006	0.002
AggStein	0.017	0.009	0.006	0.005	0.002
Stein	0.016	0.010	0.006	0.005	0.003
DPI	0.011	0.006	0.005	0.004	0.002
UCV	0.015	0.008	0.006	0.005	0.002
BCV	0.009	0.006	0.004	0.003	0.002
Nrd	0.010	0.006	0.004	0.003	0.002
Oracle	0.008	0.005	0.004	0.004	0.003

50	100	150	200	500
0.084	0.057	0.046	0.039	0.025
0.085	0.057	0.045	0.039	0.025
0.073	0.056	0.046	0.041	0.027
0.075	0.060	0.052	0.045	0.033
0.072	0.052	0.042	0.038	0.023
0.108	0.083	0.070	0.058	0.036
0.085	0.072	0.067	0.061	0.051
0.067	0.047	0.039	0.035	0.022

Table 1: *MISE for the Gaussian (left) and the exponential (right) densities*

It is well known (see, e.g., Wand and Jones (1995)) that bandwidth selection by cross-validation (UCV) is unstable and leads too often to undersmoothing. The DPI and BCV methods were proposed

	50	100	150	200	500
AggPure	0.058	0.041	0.034	0.029	0.014
AggStein	0.056	0.041	0.032	0.025	0.010
Stein	0.061	0.035	0.024	0.018	0.009
DPI	0.059	0.052	0.050	0.048	0.043
UCV	0.063	0.043	0.032	0.026	0.012
BCV	0.058	0.052	0.051	0.050	0.046
Nrd0	0.058	0.051	0.050	0.048	0.043
Oracle	0.058	0.037	0.029	0.025	0.012

50	100	150	200	500
0.064	0.042	0.034	0.029	0.017
0.061	0.042	0.033	0.028	0.017
0.057	0.041	0.033	0.028	0.017
0.070	0.054	0.046	0.042	0.029
0.057	0.038	0.031	0.026	0.016
0.101	0.083	0.066	0.055	0.027
0.088	0.078	0.072	0.069	0.057
0.064	0.038	0.030	0.025	0.016

Table 2: *MISE for the claw (left) and the smooth comb (right) densities*

	50	100	150	200	500
AggPure	0.145	0.125	0.111	0.100	0.067
AggStein	0.148	0.124	0.112	0.102	0.067
Stein	0.152	0.143	0.140	0.138	0.132
DPI	0.149	0.142	0.139	0.137	0.132
UCV	0.153	0.148	0.140	0.136	0.116
BCV	0.149	0.143	0.140	0.139	0.134
Nrd0	0.149	0.141	0.138	0.137	0.133
Oracle	0.148	0.144	0.142	0.133	0.067

50	100	150	200	500
0.142	0.119	0.102	0.093	0.061
0.148	0.141	0.103	0.092	0.060
0.154	0.143	0.140	0.137	0.132
0.147	0.140	0.138	0.136	0.132
0.154	0.142	0.133	0.126	0.074
0.146	0.141	0.139	0.138	0.134
0.146	0.140	0.137	0.136	0.132
0.145	0.128	0.109	0.101	0.062

Table 3: *MISE for dens1 (left) and dens2 (right)*

in order to bypass the problem of undersmoothing. However, sometimes they lead to oversmoothing as in the case of the Claw density while convex aggregation works well. For the normal density DPI, BCV and Nrd are better, which comes as no surprise since these estimators are designed to estimate this density well. For the other densities that are more difficult to estimate these data driven bandwidth selectors do not provide good estimators whereas the aggregation procedures remain stable. The block Stein estimator performs well in all the cases except for the highly inhomogeneous densities (cf. Table 3). In conclusion, the estimators `AggPure` and `AggStein` are very robust, as compared to other tested procedures: they are not far from the best performance for the densities that are easy to estimate and they are clear winners for densities with inhomogeneous smoothness for which other procedures fail.

`AggStein` is slightly better than `AggPure` for the Claw density and outperforms the other tested estimators in almost all the considered cases, so we studied this procedure in more detail. We focused on the Claw and Smooth Comb densities and a sample of size 500. Figure 1 gives a visual comparison of the `AggStein` procedure and the DPI procedure. It illustrates the oversmoothing effect of the DPI procedure and the fact that the `AggStein` procedure adapts to inhomogeneous smoothness. We finally comment on two other aspects of the `AggStein` procedure:

- the distribution of weights that are allocated to the aggregated estimators,
- the robustness to the number and size of the splits.

The boxplots represented in Figure 2 give the distributions of weights allocated to 7 estimators to be aggregated, the 6 kernel density estimators and the block Stein estimator. The boxplots are constructed from 2000 values of the vector of the weights (200 samples times 10 splits). We immediately notice that for the Claw density a median weight greater than 0.65 is allocated to the block Stein estimator. This can be explained by the fact that the block Stein estimator performs better than kernel density estimators on this density [cf. MISE of `AggPure` and Stein in Table 2 (left)], and the `AggStein` procedure takes advantage of it. On the other hand, for the Smooth Comb density, the block Stein estimator does not

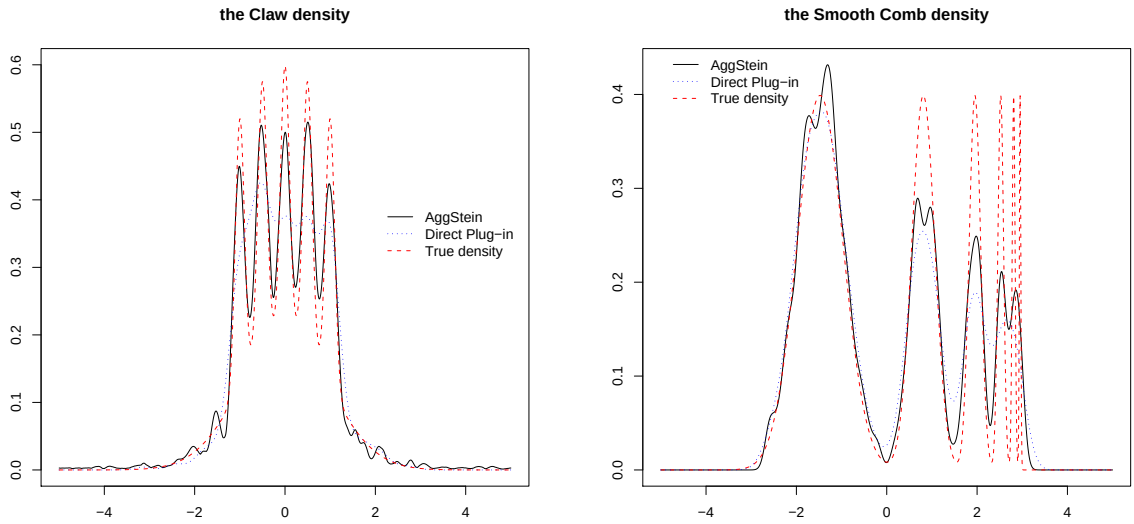

Figure 1: The Claw and Smooth Comb densities

perform significantly better than the kernel density estimators [see Table 2 (right)] and the `AggStein` procedure does not use it at all. For this sample size and this density, the procedures `AggStein` and `AggPure` are equivalent.

A free parameter of the aggregation procedures is the set of splits. In this study we choose random splits and we only have to specify their number and sizes. Obviously, we are interested to have less splits in order to make the procedure less time consuming. Figure 3 gives the sensibility of MISE both to the number of splits and to the size of the training sample in the case of dens1 and dens2 with the overall sample size 200. Two important conclusions are: (i) there exists a size of the training sample that achieves the minimum MISE, and (ii) there is essentially nothing to gain by producing more than 20 splits. Similar results are obtained for `AggPure`, and they are valid on the whole set of tested densities.

Acknowledgment: We would like to thank the referees for helpful remarks and Lucien Birgé for suggesting an improvement of the constants in Theorem 3.1 as well as a simplification of its proof. We refer to Birgé (2006) for comments on a previous version of this paper.

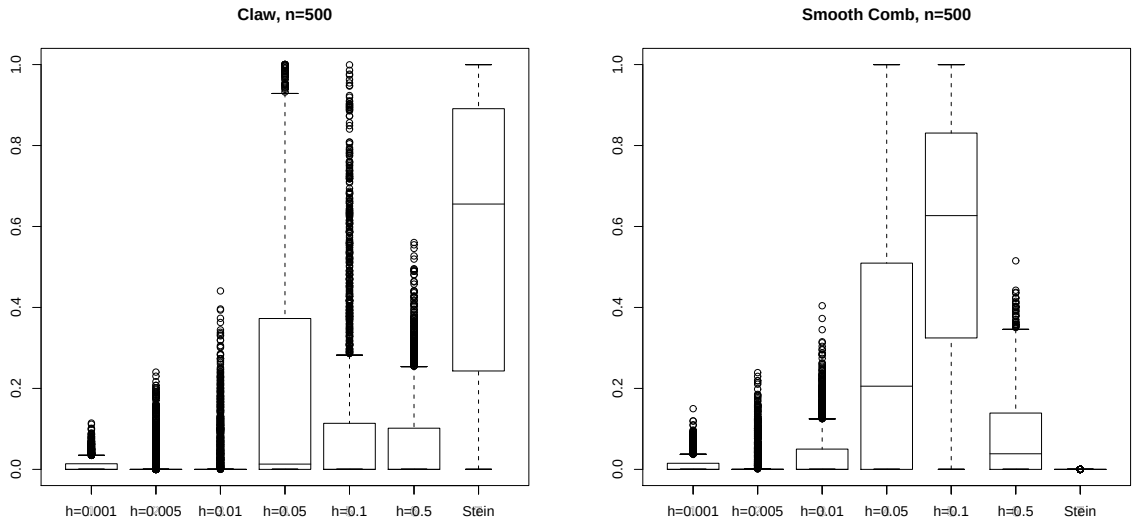

Figure 2: Boxplots for the Claw and Smooth Comb densities

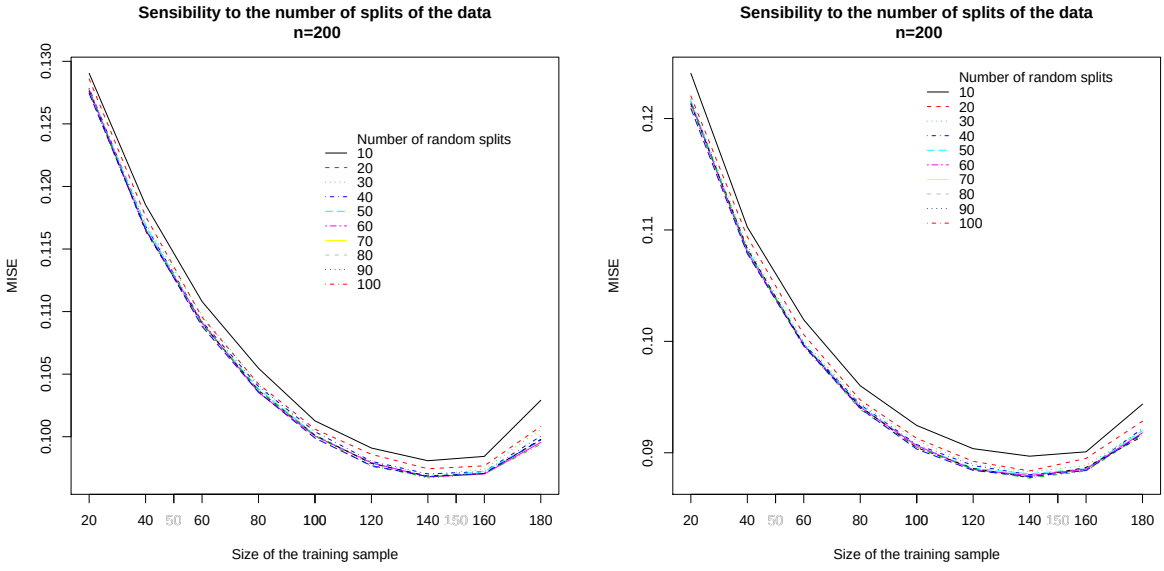

Figure 3: Sensibility to the number of splits for dens1 (left) and dens2 (right)

References

[1] Barron, A. (1987). Are Bayes rules consistent in information? In: *Open Problems in Communication and Computation,* T.M.Cover and B.Gopinath, eds. Springer, N.Y, 85-91.

[2] Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields*, **71**, 271-291.

[3] Birgé, L. (2003). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Preprint n.862, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. Available at http://www.proba.jussieu.fr/mathdoc/preprints.

[4] Birgé, L. (2006). The Brouwer Conference 2005: *Statistical estimation with model selection.* Available at arXiv:math.ST/0605187.

[5] Bunea, F., Tsybakov, A. and Wegkamp, M. (2004). Aggregation for regression learning. Preprint n.948, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. Available at http://www.proba.jussieu.fr/mathdoc/preprints and at arXiv:math.ST/0410214.

[6] Catoni O. (1999). "Universal" aggregation rules with exact bias bounds. Preprint n.510, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. Available at http://www.proba.jussieu.fr/mathdoc/preprints.

[7] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization. Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001.* Lecture Notes in Mathematics, vol.1851, Springer, New York.

[8] Dalelane C. (2004). *Data Driven Kernel Choice in Non-parametric Curve Estimation.* PhD Thesis, Technische Universität Braunschweig.

[9] Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation.* Springer, New-York.

[10] Efromovich, S. (2000). On sharp adaptive estimation of multivariate curves. *Math. Metods of Statist.*, **9**, 117-139.

[11] Golubev, G.K. (1991). LAN in nonparametric estimation of functions and lower bounds for quadratic risks. *Theory Probab. Appl.*, **36**, 152-157.

[12] Golubev, G.K. (1992). Nonparametric estimation of smooth probability densties in L_2. *Problems of Information Transmission*, **28**, 44-54.

[13] Juditsky, A., and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Annals of Statistics*, **28**, 681-712.

[14] Li, J.Q., and Barron, A. (1999). Mixture density estimation. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processings Systems*, 12, San Mateo, CA. Morgan Kaufmann Publishers.

[15] Marron, M.C. and Wand, M.P. (1992). Exact mean integrated square error. *Ann. Statist.*, **20**, 712-713.

[16] Nemirovski, A. (2000). Topics in Non-parametric Statistics. In: *Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998*, Lecture Notes in Mathematics, vol. 1738, Springer, New York.

[17] Pinsker, M.S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Information Transmission*, **16**, 120-133.

[18] Rigollet, P. (2006). Adaptive density estimation using the blockwise Stein method. *Bernoulli*, **12**, 351-370.

[19] Samarov, A. and Tsybakov, A. (2005). Aggregation of density estimators and dimension reduction. To appear in *Festschrift in Honor of Kjell Doksum*. Available at http://hal.ccsd.cnrs.fr/ccsd-00014122.

[20] Schipper, M. (1996). Optimal rates and constants in L_2-minimax estimation of probability density functions. *Math. Meth. Statist.*, **5** , 253-274.

[21] Scott (1992). *Multivariate Density Estimation.* John Wiley & Sons Inc., New York.

[22] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* (1991), **53**, 683-690.

[23] Silverman (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

[24] Stone, C. J.(1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, **12** , 1285-1297.

[25] Tsybakov, A. (2003). Optimal rates of aggregation. In: *Computational Learning Theory and Kernel Machines, Proc. 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines* (B.Schölkopf and M.Warmuth, eds.), Lecture Notes in Artificial Intelligence, v.2777. Springer, Heidelberg, 303-313.

[26] Tsybakov, A. (2004). *Introduction à l'estimation non paramétrique.* Springer-Verlag, Berlin.

[27] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

[28] Wegkamp, M.H. (1999). Quasi-universal bandwidth selection for kernel density estimators. *Canad. J. Statist.*, **27**, 409-420.

[29] Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.*, **28**, 75-87.

[30] Zhang, T. (2003). From epsilon-entropy to KL-complexity: analysis of minimum information complexity density estimation. Tech. Report RC22980, IBM T.J.Watson Research Center.