Mach Learn (2007) 67:169–195
DOI 10.1007/s10994-006-5000-8

Density estimation with stagewise optimization of the empirical risk

Jussi Klemelä

Received: 9 May 2005 / Revised: 12 July 2006 / Accepted: 9 August 2006 /
Published online: 8 September 2006

Abstract We consider multivariate density estimation with identically distributed observations. We study a density estimator which is a convex combination of functions in a dictionary and the convex combination is chosen by minimizing the L_2 empirical risk in a stagewise manner. We derive the convergence rates of the estimator when the estimated density belongs to the L_2 closure of the convex hull of a class of functions which satisfies entropy conditions. The L_2 closure of a convex hull is a large non-parametric class but under suitable entropy conditions the convergence rates of the estimator do not depend on the dimension, and density estimation is feasible also in high dimensional cases. The variance of the estimator does not increase when the number of components of the estimator increases. Instead, we control the bias-variance trade-off by the choice of the dictionary from which the components are chosen.

Keywords Boosting · Empirical risk minimization · Greedy algorithms · Multivariate function estimation

1 Introduction

We study estimation of a multivariate density function $f : \mathbf{R}^d \to \mathbf{R}$ based on identically distributed random vectors $X^1, \ldots, X^n \in \mathbf{R}^d$. To analyze the estimator we assume that the observations are independent. We are interested in cases where dimension d is large.

We estimate the density applying a stagewise optimization of the L_2 empirical risk. The estimator is a convex combination of functions lying in a dictionary of simple functions. The algorithm starts by choosing a function from the dictionary which

Editor: Nicolo Cesa-Bianchi

J. Klemelä (✉)
University of Mannheim, Department of Economics, L7 3-5, 68131 Mannheim, Germany
e-mail: klemela@rumms.uni-mannheim.de

minimizes the empirical risk, and proceeds in a stagewise manner, adding new simple functions to the convex combination. The new terms of the convex combination will be chosen by minimizing the empirical risk in such a way that the new terms increase the accuracy of the estimate at those regions where the previous estimate was inaccurate.

We analyze the rate of convergence of the expected L_2 error of the estimator under the assumption that the true density belongs to the L_2 closure of the convex hull of a base class of functions. The main example will be the case where the base class is a finite dimensional manifold of functions. For these cases the estimator has the convergence rate $(n/\log n)^{-1/4}$ or $n^{-1/4}$. This rate does not depend on the dimension d. The curse of dimensionality can be overcome in the sense that we may estimate densities lying in the closure of the convex hull uniformly well, and this density class is a large non-parametric density class of practical relevance.

We show that the stagewise minimization estimator is essentially equally good as the convex combination which minimizes the empirical risk. By defining the estimator in a stagewise manner we enable efficient calculation of estimates. The definition of the estimator is semi-algorithmic in the sense that the algorithm involves an additional minimization problem at each stage. However, even the brute force method for solving the minimization problem at each stage is feasible in some cases.

Error bounds for minimization estimators have been usually given under entropy conditions for the density class. Here we do not consider entropy conditions for the density class but the entropy conditions are posed on the underlying class and the density class is defined as the closure of the convex hull of this underlying class.

The analysis of the estimator shows that we control the bias-variance balance by the choice of the dictionary of functions from which the convex combination is selected: rich dictionaries lead to a small bias and large variance. The dictionary is the smoothing parameter, or the regularizer, of the estimator. The number of terms in the estimate does not increase the variance, and we do not have to use any model selection procedure (like minimum description length) to choose the number of terms. This contrasts for example with the case of orthogonal series estimators where the number of terms in the expansion should be chosen to balance between the bias and variance.

Boosting applied to classification tasks is an algorithm defined in Freund and Schapire (1996), for example. Boosting can also be considered as a generic functional gradient descent algorithm, and then extended to regression and density function estimation. The interpretation of boosting as a gradient descent algorithm is discussed in Breiman (1998), Friedman, Hastie, and Tibshirani (2000), and Mason et al. (2000). The stagewise minimization algorithm which we consider does not involve the gradient of the empirical risk functional. We make simulation experiments with such a boosting algorithm and note that the algorithm gives slower convergence than the stagewise minimization estimator. In boosting one typically optimizes with respect to the size of the weights of the terms in the linear combination but the stagewise minimization estimator uses a fixed sequence of weights, avoiding the additional optimization step. In the literature concerning boosting for classification one has observed that the out-of-sample error decreases as the number of terms is increased, even after the training error of the linear combination has reached zero, see Breiman (1996). Our analysis is in conformity with this observation, but in boosting the number of terms is often considered to be the regularizer, see the discussion in Section 2.2.

The existence of approximation procedures with dimension-independent convergence rates was noticed by Jones (1992) and Barron (1993) for the particular case of the L_1 Fourier classes. The analysis of the L_1 Fourier classes is closely related to the analysis of mixture classes.

Stagewise minimization has been considered in density estimation previously only with the log-likelihood empirical risk. However, Rigollet and Tsybakov (2006) provide oracle inequalities for the aggregation in the context of density estimation with the L_2 empirical risk. Li and Barron (2000) derive error bounds for the Kullback-Leibler distance. They consider sieves which consists of convex combinations of M terms and both the bias and variance terms in the error bound depend on M. They consider complexity regularization with the complexity depending on the number of terms. In contrast, the bounds in this article are such that the variance term does not depend on the number of terms in the estimate, and thus we do not need a complexity regularization with respect to the number of terms. The use of the log-likelihood empirical risk requires the assumption of the boundedness of the logarithm of the density; the density is assumed to bounded and bounded away from 0, whereas the use of the L_2 empirical risk requires only the boundedness of the densities. Similarly to the current article, Rakhlin, Panchenko, and Mukherjee (2005) give error bounds which depend only on the entropy of the base class, but using the fact that Rademacher averages of a convex hull are equal to those of the base class, and they improve the bound in Li and Barron (2000) by removing the dependence of the variance bound on M. (We thank a referee for pointing this reference to us.) Ridgeway (2002) considers stagewise minimization with a dictionary of Gaussian functions and he finds the new members of the mixture by the EM algorithm with Newton-Raphson acceleration. Rosset and Segal (2002) apply a Taylor expansion of the log-likelihood which leads to a boosting algorithm where the weights of the empirical risk are adjusted at each step (observations are weighted by the reciprocal of the current estimate). They apply Bayesian networks as base learners. Projection pursuit density estimation as presented in Friedman et al. (1984) constructs an estimate of product form with a stagewise algorithm minimizing the negative log-likelihood criterion. Priebe (1994) considers an iterative algorithm with a stopping rule for estimating Gaussian mixtures.

The estimator is defined in Section 2. The main theorem is formulated in Section 3.1. Section 3.2 gives rates of convergence under specific assumptions on the underlying base class. Section 4 illustrates the properties of the estimator with simulation examples. The proofs are given in Section 5. A discussion is given in Section 6. Some of the proofs are given in the Appendix of the technical report Klemelä (2005).

We denote with $\|x\|$ the Euclidean norm of $x \in \mathbf{R}^d$, with $\|g\|_2$ the L_2 norm of $g : \mathbf{R}^d \to \mathbf{R}$, with respect to the Lebesgue measure, and $\|g\|_\infty = \sup_{x \in \mathbf{R}^d} |g(x)|$. We denote with $\#A$ the cardinality of a finite set A.

2 Definition of the estimator and related methods

2.1 Definition of the estimator

We assume to have a sequence $X^1, \ldots, X^n \in \mathbf{R}^d$ of identically distributed observations from the distribution of an unknown density $f : \mathbf{R}^d \to \mathbf{R}$. We define a model

free estimator which may be applied also for estimating the intensity function of a multivariate Poisson process. To derive the convergence rates we assume that the observations are i.i.d.

L_2 empirical risk. Define the empirical risk of a density estimator $\hat{f}:\mathbf{R}^d \to \mathbf{R}$ with

$$\gamma_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \gamma(\hat{f}, X^i) \tag{1}$$

where $\gamma(g, x)$ is the L_2 contrast function,

$$\gamma(g, x) = -2g(x) + \|g\|_2^2, \qquad g:\mathbf{R}^d \to \mathbf{R}, \quad x \in \mathbf{R}^d.$$

Minimization of $\|\hat{f} - f\|_2^2$ over estimators $\hat{f}$ is equivalent to the minimization of $\|\hat{f} - f\|_2^2 - \|f\|_2^2$, and minimization of $\gamma_n(\hat{f})$ amounts to the minimization of $\|\hat{f} - f\|_2^2 - \|f\|_2^2$, up to the approximation $\int_{\mathbf{R}^d} \hat{f} f \approx n^{-1}\sum_{i=1}^n \hat{f}(X^i)$. Indeed,

$$\begin{aligned}\|\hat{f} - f\|_2^2 - \|f\|_2^2 &= -2\int_{\mathbf{R}^d} f\hat{f} + \|\hat{f}\|_2^2 \\ &\approx -2n^{-1}\sum_{i=1}^{n} \hat{f}(X^i) + \|\hat{f}\|_2^2 \\ &= \gamma_n(\hat{f}).\end{aligned} \tag{2}$$

Stagewise minimization estimator. We define the estimator with a stagewise minimization algorithm. New functions are added to a convex combination by minimizing the empirical risk at each step. Functions are chosen from dictionary $\mathcal{D}$.

Definition 1. Stagewise minimization estimator $\hat{f}_n:\mathbf{R}^d \to \mathbf{R}$, with dictionary $\mathcal{D}$ of functions $\mathbf{R}^d \to \mathbf{R}$, with number of components $M \geq 1$, with mixing coefficients $0 < \pi_k < 1$, $k = 1, 2, \ldots, M-1$, and with the approximation bound $\epsilon > 0$, is defined recursively with the following rules.

1. Choose $\tilde{f}_0 \in \mathcal{D}$ so that

$$\gamma_n(\tilde{f}_0) \leq \inf_{\phi\in\mathcal{D}} \gamma_n(\phi) + \epsilon. \tag{3}$$

2. For $k = 1, \ldots, M-1$, let

$$\tilde{f}_k = (1 - \pi_k)\tilde{f}_{k-1} + \pi_k\tilde{\phi}$$

where $\tilde{\phi} \in \mathcal{D}$ is chosen so that

$$\gamma_n(\tilde{f}_k) \leq \inf_{\phi\in\mathcal{D}} \gamma_n((1 - \pi_k)\tilde{f}_{k-1} + \pi_k\phi) + \pi_k\epsilon. \tag{4}$$

3. Let $\hat{f}_n = \tilde{f}_{M-1}$.

Remark 1 (Quantity to be minimized). We may write the quantity to be minimized in (4) as

$$\begin{aligned}&\gamma_n((1-\pi_k)\hat{f}_{k-1}+\pi_k\phi)\\&\quad=\gamma_n((1-\pi_k)\hat{f}_{k-1})+\gamma_n(\pi_k\phi)+2(1-\pi_k)\pi_k\int_{\mathbf{R}^d}\hat{f}_{k-1}\phi.\end{aligned}\tag{5}$$

Thus we try to choose the new additive component $\pi_k\phi$ in such a way that the usual empirical risk $\gamma_n(\pi_k\phi)$ with an additional "penalization term" which involves the inner product $\int_{\mathbf{R}^d}\hat{f}_{k-1}\phi$ is minimized. This penalization term has the effect that we minimize the empirical risk under the condition that the new component should be far from the current solution $\hat{f}_{k-1}$.

Remark 2 (Mixing coefficients). We may write the estimator as a mixture $\hat{f}_n=\sum_{k=0}^{M-1}p_k\tilde{f}_k$, where

$$p_k=\pi_k\cdot\prod_{i=k+1}^{M-1}(1-\pi_i),\qquad k=0,\ldots,M-1,\tag{6}$$

where we denote $\pi_0=1$. We have that $0<p_k<1$ and $\sum_{k=0}^{M-1}p_k=1$. The weights p_k of the mixture may be decreasing or increasing depending on the choice of the coefficients π_k. It would be possible to choose numbers π_k at each step so that the empirical risk is minimized. It will turn out that we get good error bounds with a fixed choice for the mixing coefficients, see Remark 6 below.

Remark 3 (Computational complexity). We say that the definition of the estimator in Definition 1 is semi-algorithmic because the definition contains the minimization problems (3) and (4). An advantage of the estimator is that the minimization problem over the convex hull is reduced to potentially simpler minimization problems (3) and (4). We may note that when $\mathcal{D}$ is a finite set, then the brute force algorithm for the calculation of the estimate would take $O(\#\mathcal{D}\cdot M\cdot C_{eva})$ steps, where C_{eva} is the cost of evaluating $F(\phi)=\gamma_n((1-\pi_k)\hat{f}_{k-1}+\pi_k\phi)$, $\phi\in\mathcal{D}$. Note that by (22) we take typically $M=n^{1/2}$.

Remark 4 (Minimization estimator over mixtures). To clarify the definition of the stagewise minimization estimator we compare this estimator with the estimator minimizing the empirical risk over the convex hull of the dictionary. This minimization estimator over mixtures has smaller mean integrated squared error than the stagewise minimization estimator, but the difference consists only of a term of order M^{-1}, as we state in Theorem 2. Indeed, Lemma 3 below shows that the empirical risk of the stagewise minimization estimator may be bounded with the minimal empirical risk and an additional term of order M^{-1}.

Definition 2. Minimization estimator over mixtures $\bar{f}_n : \mathbf{R}^d \to \mathbf{R}$, with dictionary $\mathcal{D}$ of functions $\mathbf{R}^d \to \mathbf{R}$, and with the approximation bound $\epsilon > 0$, satisfies

$$\gamma_n(\bar{f}_n) \leq \inf_{g \in CO(\mathcal{D})} \gamma_n(g) + \epsilon, \tag{7}$$

where $\mathrm{co}(\mathcal{D})$ is the convex hull of $\mathcal{D}$ (the collection of convex combinations $\sum_{i=1}^k \lambda_i \phi_i$, where $\phi_i \in \mathcal{D}$, $\sum_{i=1}^k \lambda_i = 1$, $\lambda_i \geq 0$, and $k = 1, 2, \ldots$).

2.2 Related methods

The stagewise minimization estimator is closely related to the family of boosting estimators, and we try to describe differences and similarities. We mention also convex minimization and the EM algorithm which provide alternative approaches.

Boosting. Boosting applied to classification tasks may be considered as a generic functional gradient descent algorithm for finding a linear combination of classifiers, see for example Mason et al. (2000). Let $(D\gamma_n)(f)$ be the gradient at f of the empirical risk γ_n. That is, $(D\gamma_n)(f)$ is a real valued functional and we have a first order approximation

$$\gamma_n(f + h) \approx \gamma_n(f) + (D\gamma_n)(f)(h).$$

Since we consider convex combinations we use the approximation

$$\begin{aligned} \gamma_n((1-\pi)f + \pi h) &= \gamma_n(f + \pi(h - f)) \\ &\approx \gamma_n(f) + \pi(D\gamma_n)(f)(h - f) \\ &= \gamma_n(f) + \pi[(D\gamma_n)(f)(h) - (D\gamma_n)(f)(f)]. \end{aligned}$$

Thus, when we want to minimize $\gamma_n((1-\pi)f + \pi h)$ with respect to h and π, we may first minimize $(D\gamma_n)(f)(h)$ with respect to h and then apply one dimensional minimization with respect to π.

Consider first the empirical risk γ_n as defined in (1), when $\gamma(g, x)$ is the L_2 contrast function, $\gamma(g, x) = -2g(x) + \|g\|_2^2$, $g : \mathbf{R}^d \to \mathbf{R}$, $x \in \mathbf{R}^d$. For the L_2 empirical risk we may define the gradient as

$$(D\gamma_n)(f)(h) = -\frac{2}{n}\sum_{i=1}^n h(X^i) + 2\int_{\mathbf{R}^d} fh. \tag{8}$$

Second, let $\gamma(g, x)$ be the log-likelihood contrast function, $\gamma(g, x) = -\log g(x)$, $g : \mathbf{R}^d \to \mathbf{R}$, $x \in \mathbf{R}^d$. We may define the gradient for the log-likelihood empirical risk as

$$(D\gamma_n)(f)(h) = -\frac{1}{n}\sum_{i=1}^n \frac{h(X^i)}{f(X^i)}. \tag{9}$$

Definition 3. Boosting estimator $\hat{f}_n : \mathbf{R}^d \to \mathbf{R}$, with dictionary $\mathcal{D}$ of functions $\phi : \mathbf{R}^d \to \mathbf{R}$, with number of components $M \geq 1$, and with empirical risk $\gamma_n : \mathrm{co}(\mathcal{D}) \to \mathbf{R}$, and its gradient $(D\gamma_n)(f)(\phi)$ as in (8) or (9), is defined recursively with the following rules.

1. Initialize $\tilde{f}_0 \in \mathcal{D}$.
2. For $k = 1, \ldots, M-1$,
   (a) apply a minimization algorithm to find an approximate solution ϕ_k to
   $$\phi_k = \mathrm{argmin}_{\phi \in \mathcal{D}} (D\gamma_n)(\tilde{f}_{k-1})(\phi),$$
   (b) apply one dimensional minimization to find an approximate solution π_k to
   $$\pi_k = \mathrm{argmin}_{\pi \in (0,1)} \gamma_n((1-\pi)\tilde{f}_{k-1} + \pi\phi_k),$$
   (c) set
   $$\tilde{f}_k = (1-\pi_k)\tilde{f}_{k-1} + \pi_k \phi_k.$$
3. Let $\hat{f}_n = \tilde{f}_{M-1}$.

Rosset and Segal (2002) considered the log-likelihood empirical risk and the gradient in (9).

When we use the L_2 empirical risk with the gradient (8), then in step 2(a) of the boosting algorithm we find $\phi \in \mathcal{D}$ minimizing

$$-\frac{2}{n}\sum_{i=1}^{n} \phi(X^i) + 2\int_{\mathbf{R}^d} \hat{f}_{k-1}\phi, \tag{10}$$

where $k = 1, \ldots, M-1$. In contrast, we noted in (5) that the stagewise minimization estimator finds $\phi \in \mathcal{D}$ minimizing

$$-\frac{2\pi_k}{n}\sum_{i=1}^{n} \phi(X^i) + \pi_k^2 \|\phi\|_2^2 + 2(1-\pi_k)\pi_k \int_{\mathbf{R}^d} \hat{f}_{k-1}\phi, \tag{11}$$

at steps $k = 1, \ldots, M-1$. The difference between (10) and (11) is that in (11) the coefficients π_k are involved, and that in (11) there appears term $\|\phi\|_2^2$. Term $\|\phi\|_2^2$ has a further regularization effect. For example, if we use a library of scaled functions $\phi(x) = \sigma^{-d}\psi((x-\mu)/\sigma)$, where $\psi : \mathbf{R}^d \to \mathbf{R}, \mu \in \mathbf{R}^d, \sigma > 0$, then $\|\phi\|_2^2 = \sigma^{-d}\|\psi\|_2^2$, which grows to infinity when $\sigma \to 0$. Thus adding term $\|\phi\|_2^2$ to the empirical risk excludes adding terms to the mixture with small σ.

In boosting one is not only optimizing the choice of a new term ϕ_k but also optimizing the choice of the weight π_k, whereas stagewise minimization uses fixed weights. Minimization with respect to the weights adds flexibility, but it might add also variance to the estimator. Indeed, one has often noticed that adding more terms to a boosting

estimator (or keeping the size of the weights of the terms large) increases variance of the estimator, see for example (Bühlmann, 2002). A high variance may have also come from the fact that one has used a version of boosting where the dictionary is not fixed but ϕ_k have been decision trees (so that the dictionary is empirical). To choose the parameters of a decision tree one needs additional estimation steps: decision trees are constructed by choosing splitting points empirically and applying sample averages to define the values of the function at the leaf nodes.

Convex minimization. We may try to solve the minimization problem in (7) by convex minimization. Indeed, the minimization of the L_2 empirical risk over weights of a convex combination amounts to the minimization of a convex functional over a convex domain. One may use for example simplex methods or interior point methods, see Gill, Murray, and Wright (1991); Nesterov and Nemirovskii (1994). Juditsky and Nemirovski (2000) considered the setting of regression estimation and they proposed a stochastic approximation algorithm for finding the minimizer over a convex hull.

EM algorithm. When one applies the log-likelihood empirical risk and when the dictionary consists of Gaussian densities, or of densities belonging to an exponential family, one may solve the minimization steps (3) and (4) with the EM algorithm. This approach was considered by Ridgeway (2002).

3 Error bounds

3.1 A non-asymptotic error bound

The stagewise minimization estimator is a model free estimator. However, we want to analyze the performance of the estimator over certain test beds. We give first a general error bound when the density belongs to the L_2 closure of the convex hull of a set of functions $\mathbb{G}$. We give in Section 3.2 examples of set $\mathbb{G}$.

The collection of densities. Let $\mathbb{G}$ be a collection of functions $\mathbf{R}^d \to \mathbf{R}$ and denote with $\bar{\text{co}}(\mathbb{G})$ the L_2 closure of the convex hull of $\mathbb{G}$. We consider the density class

$$\mathcal{F} = \bar{\text{co}}(\mathbb{G}) \cap \mathcal{F}_{den}, \tag{12}$$

where $\mathcal{F}_{den} = \mathcal{F}_{den}(B_\infty)$ is the collection of bounded densities on $\mathbf{R}^d$:

$$\mathcal{F}_{den} = \left\{ f : \int_{\mathbf{R}^d} f = 1, \;\; 0 \le f \le B_\infty \right\}, \tag{13}$$

where $0 < B_\infty < \infty$.

The general error bound. We give a general bound to the mean integrated squared error of the stagewise minimization estimator. Let the mixing coefficients be

$$\pi_k = \frac{2}{k+2}, \qquad k = 1, 2, \ldots, M-1. \tag{14}$$

Theorem 1. *Let $X^1, \ldots, X^n$ be i.i.d. observations from the distribution of density $f : \mathbf{R}^d \to \mathbf{R}$. Let estimator $\hat{f}_n$ be defined in Definition* 1*, with mixing coefficients π_k given in* (14)*. Then, for $f \in \mathcal{F}$, when $\mathcal{F}$ is defined in* (12)*,*

$$E_f \| \hat{f}_n - f \|_2^2 \le \inf_{g \in \mathrm{co}(\mathcal{D})} \| g - f \|_2^2 + 4 E_f \sup_{\phi \in \mathcal{D}} |\nu_n(\phi)| + \frac{4B_2^2}{M+1} + \epsilon,$$

where $\mathrm{co}(\mathcal{D})$ *is the convex hull of $\mathcal{D}$, $\nu_n(\phi)$ is the centered empirical operator defined by*

$$\nu_n(\phi) = \frac{1}{n} \sum_{i=1}^{n} \phi(X^i) - \int_{\mathbf{R}^d} \phi f, \quad \phi : \mathbf{R}^d \to \mathbf{R}, \tag{15}$$

and $B_2 = \sup_{\phi \in \mathcal{D}} \|\phi\|_2$. We use the notation E_f to mean the expectation with respect to the distribution of $(X^1, \ldots, X^n)$, that is, with respect to the n-fold product measure with density $\prod_{i=1}^n f(x^i)$.

Theorem 1 is proved in Section 5.

Remark 5. The term $\inf_{g \in \mathrm{co}(\mathcal{D})} \| g - f \|_2^2$ may be identified as the bias term, the term $E_f \sup_{\phi \in \mathcal{D}} |\nu_n(\phi)|$ as the variance term, and the term $4B_2^2/(M+1) + \epsilon$ as the approximation term. The bias and variance terms are different depending on the choice of the dictionary $\mathcal{D}$. We consider two cases; (1) $\mathcal{D}$ is a δ-net of $\mathbb{G}$; (2) $\mathcal{D}$ is the base class: $\mathcal{D} = \mathbb{G}$. The first case is studied in Section 3.1.1 and the second case is studied in Section 3.1.2. In the first case the bias term is equal to δ and in the second case the bias term vanishes. The variance term has been studied extensively in the theory of empirical processes. Here it is important that the supremum in the variance term is over $\phi \in \mathcal{D}$ and not over $g \in \mathrm{co}(\mathcal{D})$. We give below examples of the bounds for the variance term.

Remark 6. With the choice of the mixing coefficients as in (14) we may write the estimator as a mixture $\hat{f}_n = \sum_{k=0}^{M-1} p_k \tilde{f}_k$, where $p_k = 2(k+1)/[(M(M+1)]$, $k = 0, \ldots, M-1$, where we used the formula given in (6). Thus the coefficients p_k of the mixture are linearly increasing with k and the new term gets always the largest weight.

Remark 7. The main interest of Theorem 1 lies in the case where $\mathbb{G}$ is a simple collection but $\bar{\mathrm{co}}(\mathbb{G})$ is nevertheless a large nonparametric class, although Theorem 1 holds also in the case $\mathbb{G} = \bar{\mathrm{co}}(\mathbb{G})$. The main example is the case where $\mathbb{G}$ is a finite dimensional manifold. This case is considered in Section 3.2. In this case we may

found reasonable algorithms for solving the minimization problems (3) and (4), and at the same time the convergence rate of the estimator is not unacceptably slow in high dimensional cases. When $\mathbb{G}$ is a Sobolev ball or a Hölder ball of multivariate functions, then $\mathbb{G} = \bar{\text{co}}(\mathbb{G})$, see Remark 10.

Minimization estimator over mixtures. We may clarify Theorem 1 by pointing out that when we consider the minimization estimator over the convex hull, defined in Definition 2, then we get a better bound, without term $4B_2^2/(M+1)$. However, in this case we do not have an algorithmic definition of the estimator.

Theorem 2. *Let $X^1, \ldots, X^n$ be i.i.d. observations from the distribution of density $f : \mathbf{R}^d \to \mathbf{R}$. Let $\bar{f}_n$ be defined Definition 2. Then, for $f \in \mathcal{F}$, when $\mathcal{F}$ is defined in* (12)*,*

$$E_f \|\bar{f}_n - f\|_2^2 \leq \inf_{g \in \mathbf{CO}(\mathcal{D})} \|g - f\|_2^2 + 4E_f \sup_{\phi \in \mathcal{D}} |\nu_n(\phi)| + \epsilon.$$

Theorem 2 is proved in Section 5; it follows directly from Lemma 4 given in Section 5.2.

3.1.1 δ-net dictionary

A natural choice for the dictionary $\mathcal{D}$ is to take it equal to a δ-net of the base class $\mathbb{G}$. When the base class $\mathbb{G}$ is L_2-bounded:

$$\sup_{g \in \mathbb{G}} \|g\|_2 < \infty, \tag{16}$$

then for each $\delta > 0$, there exists a finite δ-net $\mathcal{D}_\delta$ of $\mathbb{G}$; collection $\mathcal{D}_\delta$ is of finite cardinality and for each $\phi \in \mathbb{G}$ there is $\phi' \in \mathcal{D}_\delta$ such that $\|\phi - \phi'\|_2 \leq \delta$. Theorem 1 leads to the following corollary.

Corollary 1. *Let $X^1, \ldots, X^n$ be i.i.d. observations from the distribution of density $f : \mathbf{R}^d \to \mathbf{R}$. Let* (16) *hold. Let estimator $\hat{f}_n$ be defined in Definition* 1*, with mixing coefficients π_k given in* (14)*. Let $\delta > 0$ and let the dictionary be a δ-net $\mathcal{D} = \mathcal{D}_\delta$. Then, for $f \in \mathcal{F}$, when $\mathcal{F}$ is defined in* (12)*,*

$$E_f \|\hat{f}_n - f\|_2^2 \leq \delta^2 + 8 \cdot 2^{1/2} B_\infty \frac{\sqrt{\log_e(2\#\mathcal{D})}}{n^{1/2}} + \frac{4B_2^2}{M+1} + \epsilon,$$

where $B_2 = \sup_{\phi \in \mathcal{D}} \|\phi\|_2$, and $B_\infty = \sup_{\phi \in \mathcal{D}} \|\phi\|_\infty$.

Proof: Every convex combination of the functions in $\mathbb{G}$ may be approximated up to δ with some convex combination of the functions in dictionary $\mathcal{D}_\delta$:

$$\sup_{g \in \mathbf{CO}(\mathbb{G})} \inf_{h \in \mathbf{CO}(\mathcal{D}_\delta)} \|g - h\|_2 \leq \delta \tag{17}$$

where $\mathcal{D}_\delta$ is the dictionary defined in Assumption 16. The approximation (17) follows directly from the fact that $\mathcal{D}_\delta$ is an δ-net of $\mathbb{G}$ in the L_2 metric. For a proof of (17) see Appendix A. The fact that $\mathcal{F} \subset \bar{\text{co}}(\mathbb{G})$ implies

$$\sup_{f\in\mathcal{F}} \inf_{g\in \text{CO}(\mathbb{G})} \|f - g\|_2 = 0. \tag{18}$$

Equations (17) and (18) imply that

$$\sup_{f\in\mathcal{F}} \inf_{h\in \text{CO}(\mathcal{D}_\delta)} \|f - h\|_2 \leq \delta. \tag{19}$$

That is, when $\mathcal{D}_\delta$ is a δ-net for $\mathbb{G}$, then $\text{co}(\mathcal{D}_\delta)$ is a δ-net for $\mathcal{F}$. We have proved the bound for the bias term in Theorem 1 (the first term in the right hand side). It is left to prove a bound for the variance term. The cardinality of $\mathcal{D}_\delta$ is finite and we have

$$E_f \sup_{\phi\in\mathcal{D}} |\nu_n(\phi)| \leq 2B_\infty n^{-1/2} 2^{1/2} \sqrt{\log_e(2\,\#\,\mathcal{D}_\delta)},$$

see for example (Lugosi, 2002). We have proved the theorem. □

Remark 8 (Smoothing parameters of the estimate). We have identified term δ^2 as the bias term and term $(\log_e(\#\mathcal{D}_\delta)/n)^{1/2}$ as the variance term. We balance the bias and the variance of the estimator by the choice of dictionary $\mathcal{D}$. The choice of the number M of the terms does not affect the variance of the estimator. We may improve the estimator by choosing M large but this increases computational complexity.

3.1.2 $\mathbb{G}$ as dictionary

We may choose the base class $\mathbb{G}$ itself to be the dictionary. In order to apply Theorem 1 we need that the entropy integral of the base class converges. Let us call a δ-bracketing net of $\mathbb{G}$ with respect to the L_2 norm a set of pairs of functions $\mathbb{G}_\delta = \{(g_j^L, g_j^U) : j = 1, \ldots, N(\delta)\}$ such that

1. $\|g_j^L - g_j^U\|_2 \leq \delta$, $j = 1, \ldots, N(\delta)$,
2. for each $g \in \mathbb{G}$ there is $j = j(g) \in \{1, \ldots, N(\delta)\}$ such that $g_j^L \leq g \leq g_j^U$.

Define the entropy integral

$$G(B_2) = \int_0^{B_2} \sqrt{\log_e N(u)}\, du, \tag{20}$$

where $B_2 = \sup_{g\in\mathbb{G}} \|g\|_2$. Theorem 1 implies the following corollary.

Corollary 2. *Let $X^1, \ldots, X^n$ be i.i.d. observations from the distribution of density $f : \mathbf{R}^d \to \mathbf{R}$. Let estimator $\hat{f}_n$ be defined in Definition* 1*, with mixing coefficients π_k given in* (14)*. Let the dictionary of $\hat{f}_n$ be $\mathbb{G}$. Let* (16) *hold, let $B_\infty = \sup_{g\in\mathbb{G}} \|g\|_\infty < \infty$, and let the entropy integral $G(B_2)$ be finite. Then, for $f \in \mathcal{F}$, when $\mathcal{F}$ is defined*

in (12),

$$E_f \|\hat{f}_n - f\|_2^2 \leq \frac{C}{n^{1/2}} + \frac{4B_2^2}{M+1} + \epsilon,$$

where $B_2 = \sup_{g\in\mathbb{G}} \|g\|_2$, *and* C *is a positive constant depending on* B_2, B_∞, *and on the entropy integral* $G(B_2)$.

Proof: We noted already in (18) that the bias term is zero when $\mathcal{D} = \mathbb{G}$. For the variance term we have that

$$E_f \sup_{g\in\mathbb{G}} |\nu_n(g)| \leq \frac{C}{n^{1/2}}.$$

by applying a variant of exponential inequalities given by Ossiander (1987), Birgé and Massart (1993), Proposition 3, or van de Geer (2000), Theorem 8.13. □

Remark 9. One may also derive a bound for the variance term with the help of the empirical entropy:

$$E_f \sup_{g\in\mathbb{G}} |\nu_n(g)| \leq \frac{C}{n^{1/2}} E_f \int_0^{B_\infty} \sqrt{\log_e N(u, \mathbb{G}, \|\cdot\|_{2,n})}\, du,$$

where $N(\delta, \mathbb{G}, \|\cdot\|_{2,n})$ is the cardinality of the smallest δ-cover of $\mathbb{G}$ with respect to the empirical metric $\|g\|_{2,n}^2 = \sum_{i=1}^n g(X^i)^2$. See Pollard (1989) or van der Vaart and Wellner (1996). This type of bound was used in Rakhlin, Panchenko, and Mukherjee (2005). The expectation in the upper bound can further be bounded by using

$$N(\delta, \mathbb{G}, \|\cdot\|_{2,n}) \leq C'(1/\delta)^{2(V(\mathbb{G})-1)},$$

where $V(\mathbb{G})$ is the VC-dimension of $\mathbb{G}$, and C' is a positive constant depending on $V(\mathbb{G})$. See van der Vaart and Wellner (1996), Theorem 2.6.4.

3.2 Rates of convergence

We discuss cases where the estimator has the rate of convergence $(n/\log n)^{-1/4}$ or $n^{-1/4}$. This rate may be achieved in the case where the density belongs to the closure of the convex hull of a finite dimensional manifold. The case where the density is an infinite mixture of densities on a finite dimensional manifold may be reduced to this case.

3.2.1 Entropy condition

We consider the class of densities (12) and make restrictions to $\mathbb{G}$. When $\mathbb{G}$ is a k-dimensional class, then there exists a δ-net of $\mathbb{G}$ of cardinality $C\delta^{-k}$. Now we make concrete choices for the parameters of the stagewise minimization estimator. Let the

discretization parameter $\delta = \delta_n > 0$ satisfy

$$\delta_n \asymp \left(\frac{\log n}{n}\right)^{1/4}. \tag{21}$$

Let the component number $M = M_n \in \{1, 2, \ldots\}$ and the approximation bound $\epsilon = \epsilon_n > 0$ of the estimator satisfy

$$M_n^{-1} \asymp \epsilon_n = O(n^{-1/2}). \tag{22}$$

Corollary 3. *Assume that the collection $\mathbb{G}$ has a δ-net $\mathcal{D}_\delta$ of cardinality*

$$\log(\#\mathcal{D}_\delta) \leq C \log \delta^{-1}, \tag{23}$$

for a positive constant C. Let estimator $\hat{f}_n$ be defined in Definition 1 *with the component number M_n as in* (22)*, the mixing coefficients π_k as in* (14)*, and the approximation bound ϵ_n as in* (22)*.*

1. *Let $\delta = \delta_n > 0$ satisfy* (21) *and let the dictionary of the estimator be $\mathcal{D} = \mathcal{D}_{\delta_n}$, where $\mathcal{D}_\delta$ is as in Assumption* 16*. We have under the assumptions of Corollary* 1 *that*

$$\limsup_{n\to\infty} \left(\frac{n}{\log n}\right)^{1/2} \sup_{f\in\mathcal{F}} E_f \|\hat{f}_n - f\|_2^2 < \infty \tag{24}$$

   where $\mathcal{F}$ is defined in (12)*.*
2. *We have under the assumptions of Corollary* 2 *that*

$$\limsup_{n\to\infty} n^{1/2} \sup_{f\in\mathcal{F}} E_f \|\hat{f}_n - f\|_2^2 < \infty. \tag{25}$$

Proof: Corollary 3 follows directly by plugging the values (21), (22), (23) in the upper bound of Corollary 1, or in the upper bound of Corollary 2. □

Remark 10. When $\mathbb{G}$ is a Sobolev ball or a Hölder ball of multivariate functions, then there exists a δ-net of $\mathbb{G}$ with cardinality N_δ where

$$\log(N_\delta) \leq C\delta^{-d/s} \tag{26}$$

for a positive constant C, where $s > 0$ is the smoothness index, see Kolmogorov and Tikhomirov (1961). In these cases Corollary 1 gives the rate $n^{-s/(4s+d)}$. However, the optimal rate of convergence is known to be $n^{-s/(2s+d)}$ for the Sobolev or Hölder balls, and for example kernel estimators achieve this rate. In high dimensional cases a mixture class makes a stronger restriction to the density than the classical smoothness

conditions. We have that

$$(n/\log n)^{1/4} > n^{s/(2s+d)} \Leftrightarrow d > 2s. \tag{27}$$

Thus, in the high dimensional cases, when $d > 2s$, the rate $(n/\log n)^{-1/4}$ is better than the classical rate. The curse of dimensionality affects that accurate estimation is not possible in high dimensional cases if the true density is a worse case in a Sobolev ball, but if it happens that the true density lies in a mixture class accurate estimation is possible also in high dimensional cases.

Remark 11. van der Vaart and Wellner (1996) and Carl (1997) have shown that if there exists a δ-net of $\mathbb{G}$ of cardinality $C(1/\delta)^V$, then there exists a δ-net of $\mathcal{F}$ of cardinality $C'(1/\delta)^{2V/(V+2)}$, where C' depends only on the envelope of $\mathbb{G}$, on C and V. Here the δ-nets are with respect to the $L_2(Q)$ metric, where Q is a probability measure. Generalizations and better constants has been given by Carl, Kyrezi, and Pajor (1999) and Mendelson (2002). The results indicate that the rate $n^{-(V+2)/[4(V+1)]}$ can be achieved by a minimization estimator of Definition 2. For $V = 0$ the rate is $n^{-1/2}$ and for $V = \infty$ the rate is $n^{-1/4}$.

3.2.2 Convex closures of parametric families

We consider examples where condition (23) for $\mathbb{G}$ holds. Let $\mathcal{F}$ be defined by (12) where

$$\mathbb{G} = \{g(\cdot, \theta) : \theta \in \Theta\} \tag{28}$$

with $\Theta \subset \mathbf{R}^k$. Set $\mathbb{G}$ is a finite dimensional collection but set $\mathcal{F}$ is an infinite dimensional collection. With regularity conditions on $\theta \mapsto g(\cdot, \theta)$ we may guarantee a parametric bound for the entropy of $\mathbb{G}$.

Assumption 1. Assume that $\Theta \subset \mathbf{R}^k$ is bounded in the Euclidean metric, and for all $\theta, \theta' \in \Theta$,

$$\|g(\cdot, \theta) - g(\cdot, \theta')\|_2 \leq C\|\theta - \theta'\|$$

for a positive constant C.

Lemma 1. *Assumption* 1 *implies* (23)*, for* $\mathbb{G}$ *defined in* (28)*.*

Proof: It is enough to note that Θ may be covered with $C'\delta^{-k}$ balls of radius δ, see Kolmogorov and Tikhomirov (1961). □

Corollary 4. *Let Assumption* 1 *hold. Define estimator* $\hat{f}_n$ *similarly as in Corollary* 3*, except that the dictionary is defined by*

$$\mathcal{D}_\delta = \{g(\cdot, \theta) : \theta \in \Theta_\delta\}, \quad \delta > 0, \tag{29}$$

where Θ_δ, $\delta > 0$, is a δ-net of Θ in the Euclidean metric. Then (24) *holds, where $\mathcal{F}$ is defined in* (12) *with $\mathbb{G}$ as in* (28)*. If the dictionary is $\mathbb{G}$, then* (25) *holds.*

Proof: Corollary 4 follows from Corollary 3 and Lemma 1. □

3.2.3 Infinite mixture families

Estimating a density in a class of infinite mixtures may not be more difficult than estimating a density in the closure of a convex hull. Let $\mathbb{G}$ be defined in (28), that is, $\mathbb{G} = \{g(\cdot, \theta) : \theta \in \Theta\}$ with $\Theta \subset \mathbf{R}^k$. Assume that $\Theta = \Theta^{(1)} \times \Theta^{(2)}$, where $\Theta^{(i)} \subset \mathbf{R}^{k_i}$, $0 \le k_i \le k$, $i = 1, 2$, $k_1 + k_2 = k$. Let

$$\mathcal{G}(\mathbb{G}) = \left\{ \int_{\Theta^{(2)}} g\left(\cdot, \theta^{(1)}, \theta^{(2)}\right) dQ\left(\theta^{(2)}\right) : \theta^{(1)} \in \Theta^{(1)},\ Q \in \mathbb{Q}\left(\Theta^{(2)}\right) \right\} \tag{30}$$

where $g(\cdot, \theta) = g(\cdot, \theta^{(1)}, \theta^{(2)}) \in \mathbb{G}$ and $\mathbb{Q}(\Theta^{(2)})$ is the set of probability measures on $\Theta^{(2)}$. Let us state the additional regularity conditions.

Assumption 2. Let $g(x, \theta^{(1)}, \cdot)$ be Riemann integrable for all $x \in \mathbf{R}^d$, $\theta^{(1)} \in \Theta^{(1)}$. Let

$$\sup_{\theta^{(2)} \in \Theta^{(2)}} \left\| g\left(\cdot, \theta^{(1)}, \theta^{(2)}\right) \right\|_2 < \infty, \quad \text{for all } \theta^{(1)} \in \Theta^{(1)}. \tag{31}$$

Lemma 2. *Let $\mathbb{G}$ be defined in* (28)*. Let Assumption* 2 *hold. Then,*

$$\mathcal{G}(\mathbb{G}) \subset \bar{\text{co}}(\mathbb{G}). \tag{32}$$

Proof: A proof of (32) is given in the technical report (Klemelä, 2005). The proof uses just the fact that the integral in the definition of $g \in \mathcal{G}(\mathbb{G})$ may be approximated with a Riemann sum. □

Corollary 5. *Let Assumption* 1 *and Assumption* 2 *hold. Define the estimator $\hat{f}_n$ similarly as in Corollary* 3*, except that the dictionary is defined by* (29)*. Then* (24) *holds, where $\mathcal{F}$ is defined by*

$$\mathcal{F} = \mathcal{G}(\mathbb{G}) \cap \mathcal{F}_{den}. \tag{33}$$

If the dictionary is $\mathbb{G}$, then (25) *holds.*

Proof: We apply Lemma 2 and Corollary 4. □

Remark 12. Genovese and Wasserman (2000) show that a maximum likelihood estimator can achieve the rate $(n/\log n)^{1/4}$ for Gaussian mixture models. Ghosal and van der Vaart (2001) show that for Gaussian mixture models a maximum likelihood estimator can achieve the almost parametric rate $n^{1/2}/(\log n)^\gamma$, $\gamma \ge 1$. Biau and

Devroye (2005) show that a complexity penalized minimum distance estimator achieves the parametric rate $n^{1/2}$.

4 Illustrations

We illustrate the behavior of the stagewise minimization estimator with one dimensional examples: a two-modal density, the standard log-normal density, and the claw density. The mean integrated squared error (MISE, $E\int|\hat{f}-f|^2$) and the mean integrated absolute error (MIAE, $E\int|\hat{f}-f|$) of the estimates were studied. The L_1 error may be more natural error criterion than the L_2 error since by Scheffé's lemma it is related to the total variation distance: $\int|\hat{f}-f| = \frac{1}{2}\sup_A|\int_A \hat{f} - \int_A f|$. We also compared the stagewise minimization estimator to a boosting estimator.

In the simulation experiments we did not choose the first term $\tilde{f}_0$ using the rule (3). Instead, the first term was chosen by performing the minimization only over the location μ and the standard deviation was fixed to unity: $\sigma = 1$. This improved the estimator. When the optimization in choosing the first term was performed over σ, then always the smallest value of σ was chosen. The boosting estimator was initialized by the same rule as the stagewise minimization estimator. The boosting estimator was otherwise defined by Definition 3, with the L_2 gradient (8), but the weights of the boosting estimator were not chosen by a minimization; the same fixed sequence of weights was used as for the stagewise minimization estimator.

A conclusion of the experiments is that the MISE of the stagewise minimization estimates decreases as the number of terms M is increased, and after achieving the minimum, the MISE does not increase significantly. The MIAE of the estimates behaves similarly. The stagewise minimization estimates detect the shape of the densities already when M is small, except for the claw density one needs large M. Small wiggles appear to estimates when M is large. The boosting estimates are much worse for small values of M, but for large values of M they have the same accuracy as the stagewise minimization estimates.

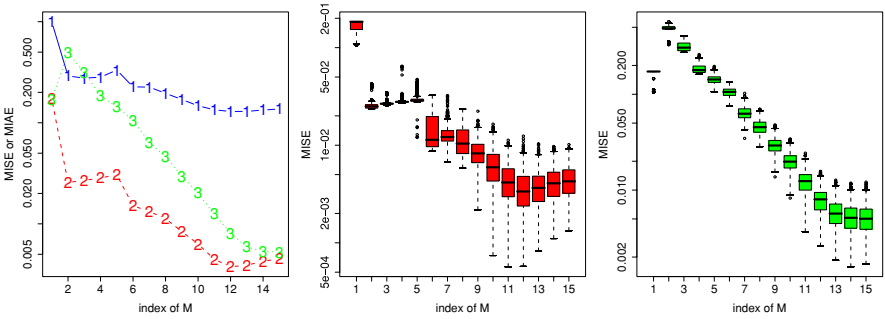

Fig. 1 2-modal density, sample size $n = 500$. Frame (a) shows the average of MIAE (blue "1") and the average of MISE (red "2") for the stagewise minimization estimator, for the 15 values of $M \in \mathcal{M}$, over 500 samples. The average of MISE for the boosting estimator is shown by green "3". Frame (b) shows the Box plots for each sample of MISE values for the stagewise minimization estimator and frame (c) shows the Box plots for the boosting estimator. A logarithmic scale is used for the y-axis

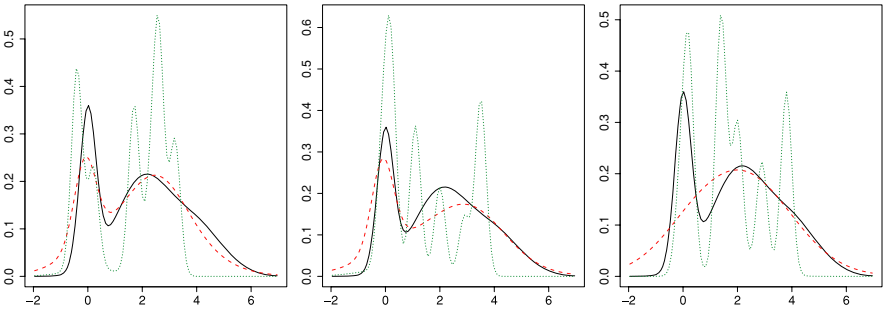

Fig. 2 2-modal density, sample size $n = 500$, number of terms $M = 7$. Frame (a) shows as dashed red graph the stagewise minimization estimate whose MISE value is equal to the 0.1 quantile among all MISE values for the 500 samples, as dotted green graph the boosting estimate whose MISE value was equal to the 0.1 quantile among all MISE values, and the true density as the black solid graph. Frame (b) shows the estimates corresponding to the median values of MISE. Frame (c) shows the estimates corresponding to the 0.9 quantiles of MISE

4.1 Two-modal density

We estimate a density which is a mixture of 3 univariate Gaussians, with means 0, 2, 4, with standard deviations 0.3, 1, 1, and with mixture weights 0.25, 0.5, 0.25. The solid black lines in Figs. 2–3 show the graph of the density.

We applied the dictionary of Gaussians $\phi((x-\mu)/\sigma)/\sigma$, where ϕ is the standard Gaussian density, $-1 \le \mu \le 5$ with stepsize 0.3, and $0.2 \le \sigma \le 2$ with stepsize 0.2. We generated 500 samples of size $n = 500$. We constructed estimates with the number of terms

$$M \in \mathcal{M} = \{1, 2, 3, 4, 5, 7, 9, 11, 15, 20, 30, 50, 100, 200, 300\}. \tag{34}$$

Figure 1(a) shows the average of MIAE (blue "1") and the average of MISE (red "2"), for the stagewise minimization estimator for the 15 values of $M \in \mathcal{M}$, over the 500 samples. The average of MISE for the boosting estimator is shown by green "3". Frame (b) shows the Box plots for each sample of MISE values of the stagewise

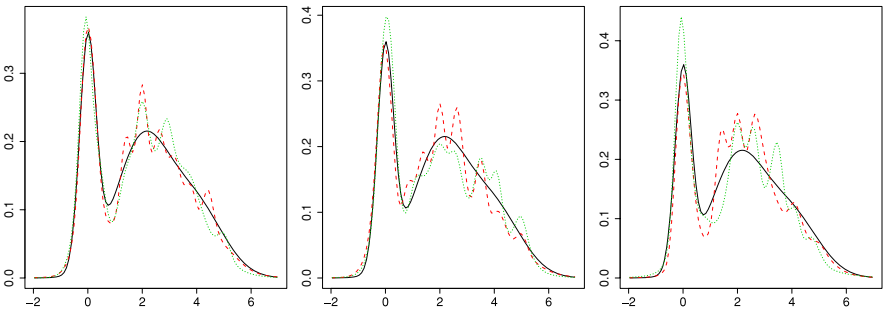

Fig. 3 2-modal density, sample size $n = 500$, number of terms $M = 300$. The setting is otherwise the same as in Fig. 2 but now $M = 300$; the dashed red graph shows a stagewise minimization estimate and the dotted green graph shows a boosting estimate

minimization estimator, and frame (c) shows the Box plots for the boosting estimator, over $M \in \mathcal{M}$.

Figure 2 shows estimates when the number of terms $M = 7$. The dashed red graphs in frames ((a)–(c)) show the stagewise minimization estimates corresponding to the 0.1-quantile, median, and the 0.9-quantile among the MISE values over 500 samples. The dotted green graphs in frames ((a)–(c)) show the boosting estimates whose MISE values were equal to the 0.1-quantile, median, and the 0.9-quantile. Thus frame (a) shows better than average estimates, frame (b) shows typical estimates, and frame (c) shows worse than average estimates.

Figure 3 shows estimates when the number of terms is $M = 300$. We show again the estimates corresponding to the 0.1-quantile, median, and 0.9-quantile of the MISE values.

Figure 1 shows that the average MISE is minimized for the stagewise minimization estimator when $M = 100$, but the average MISE is not significantly larger for $M = 300$. The MIAE increases even less when M is increased from 100 to 300. The MISE values for the boosting estimates were larger than for the stagewise minimization estimates for small values of the number of terms M, but when M increases, then the MISE values approach each other. Figure 2 shows that when $M = 7$, then the stagewise minimization estimator gives good estimates, but the boosting estimate does not detect the shape of the density. Figure 3 shows that when $M = 300$, then both estimators produce good estimates, and the quality of the estimates does not have much variability. In general, estimates behave qualitatively similarly with respect to MISE and MIAE.

4.2 Log-normal density

To make a comparison with Priebe (1994) we considered the estimation of the standard log-normal density $(2\pi)^{-1/2}x^{-1}\exp\{-(\log_e x)^2/2\}$, $0 < x < \infty$. Priebe (1994) constructed a Gaussian mixture estimate with 27 terms when the sample size was 1000.

We applied the dictionary of Gaussians $\phi((x-\mu)/\sigma)/\sigma$, $-1 \le \mu \le 5$ with stepsize 0.3, and $0.2 \le \sigma \le 2$ with stepsize 0.2. We generated 500 samples of size $n = 50$. We constructed estimates with the number of terms $M \in \mathcal{M}$, when $\mathcal{M}$ is defined in (34).

Figure 4(a) shows the average of MIAE (blue "1") and the average of MISE (red "2"), for the stagewise minimization estimator for the 15 values of $M \in \mathcal{M}$, over the 500 samples. The average of MISE for the boosting estimator is shown by green "3". Box plots are shown only for the stagewise minimization estimator. Frame (b) shows the Box plots for each sample of MISE values, and frame (c) shows the Box plots for each sample of MIAE values, over $M \in \mathcal{M}$.

Figure 5 shows estimates when the number of terms $M = 7$. The dashed red graphs in frames ((a)–(c)) show the stagewise minimization estimates corresponding to the 0.1-quantile, median, and the 0.9-quantile among the MISE values over 500 samples. The dotted green graphs in frames ((a)–(c)) show the boosting estimates whose MISE values were equal to the 0.1-quantile, median, and the 0.9-quantile.

Figure 6 shows estimates when the number of terms is $M = 300$. We show again the estimates corresponding to the 0.1-quantile, median, and 0.9-quantile of the MISE values.

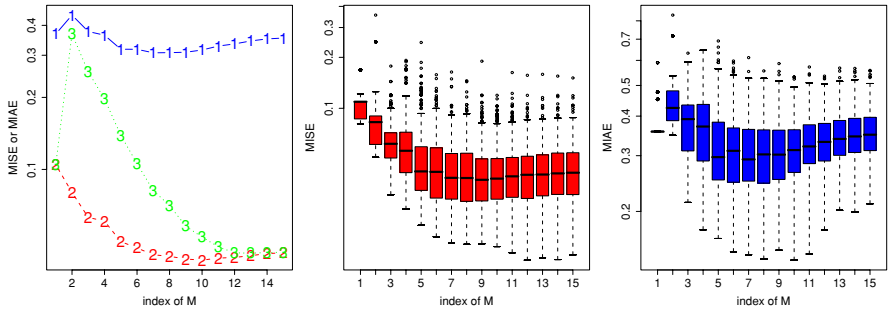

Fig. 4 Log-normal density, sample size $n = 50$. Frame (a) shows the average of MIAE (blue "1") and the average of MISE (red "2") for the stagewise minimization estimator, for the 15 values of $M \in \mathcal{M}$, over 500 samples. The average of MISE for the boosting estimator is shown by green "3". Frame (b) shows the Box plots for each sample of MISE values, for the stagewise minimization estimator, and frame (c) shows the Box plots for each sample of MIAE values, for the stagewise minimization estimator. A logarithmic scale is used for the y-axis

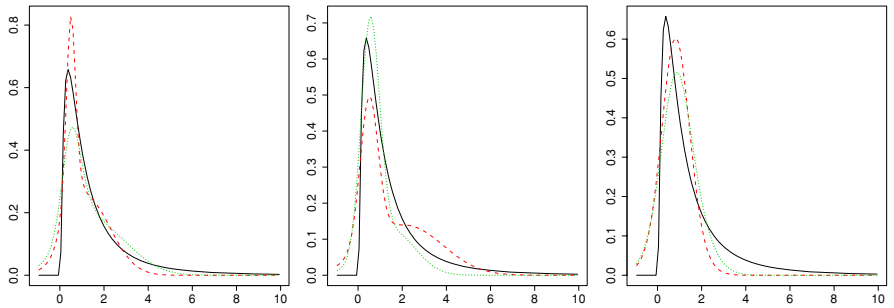

Fig. 5 Log-normal density, sample size $n = 50$, number of terms $M = 7$. Frame (a) shows as dashed red graph the stagewise minimization estimate whose MISE value is equal to the 0.1 quantile among all MISE values for the 500 samples, as dotted green graph the boosting estimate whose MISE value is equal to the 0.1 quantile among all MISE values, and the true density as the black solid graph. Frame (b) shows the estimates corresponding to the median values of MISE. Frame (c) shows the estimates corresponding to the 0.9 quantiles of MISE

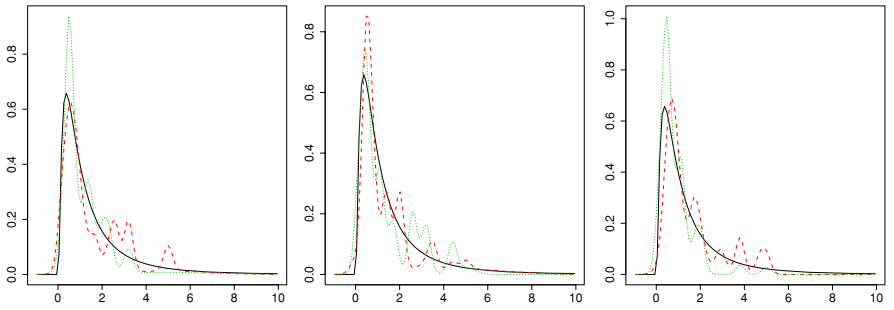

Fig. 6 Log-normal density, sample size $n = 50$, number of terms $M = 300$. The setting is otherwise the same as in Figure 5 but now $M = 300$

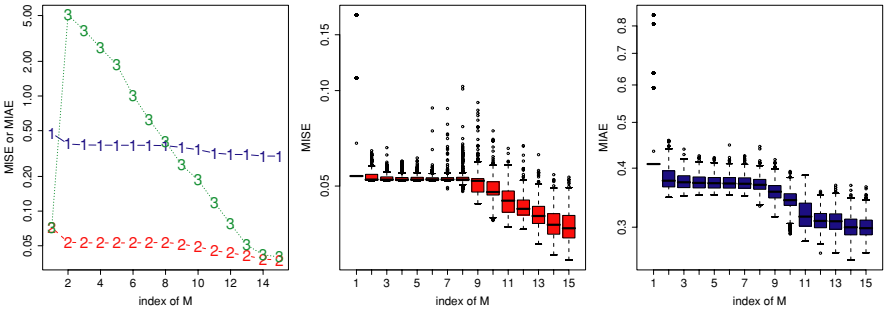

Fig. 7 Claw density, sample size $n = 500$. Frame (a) shows the average of MIAE (blue "1") and the average of MISE (red "2") for the stagewise minimization estimator, for the 15 values of $M \in \mathcal{M}$, over 500 samples. The average of MISE for the boosting estimator is shown by green "3". Frame (b) shows the Box plots for each sample of MISE values for the stagewise minimization estimator and frame (c) shows the Box plots for the boosting estimator. A logarithmic scale is used for the y-axis

Figure 4 shows that for the sample size $n = 50$ the behavior of MISE and MIAE is similar as for the two-modal density with sample size $n = 500$. The variability of the MIAE decreases slightly, when M increases, although the variability of the MISE stays the same. Figure 5 shows that qualitatively the estimates behave well when $M = 7$, although the heightness of the mode is not estimated accurately. Figure 6 shows that when $M = 300$, then small wiggles start to appear. In the best case the stagewise minimization estimator estimates the mode well, but the boosting estimate is less accurate.

4.3 The claw density

The claw density is the density number 10 in Marron and Wand (1992). It is a multi-modal density with 5 modes, shown in Fig. 8 as a black graph.

We applied the dictionary of Gaussians $\phi((x-\mu)/\sigma)/\sigma$, $-3 \le \mu \le 3$ with step-size 0.3, and $0.02 \le \sigma \le 1.1$ with stepsize 0.02. We generated 500 samples of size

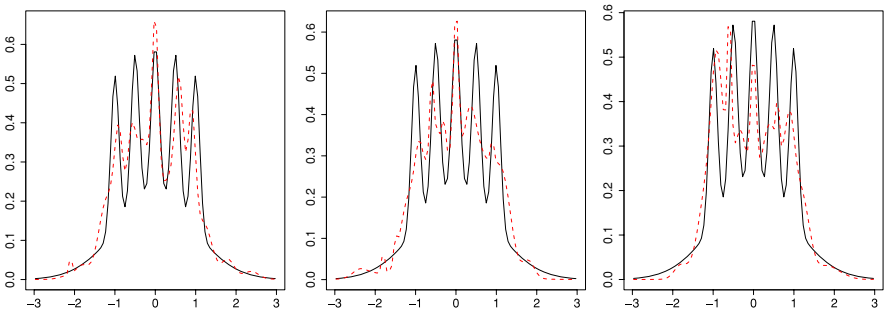

Fig. 8 Claw density, sample size $n = 500$, number of terms $M = 300$. Frame (a) shows as dashed red graph the stagewise minimization estimate whose MISE value is equal to the 0.1 quantile among all MISE values for the 500 samples, and the true density as the black solid graph. Frame (b) shows the estimates corresponding to the median value of MISE. Frame (c) shows the estimates corresponding to the 0.9 quantile of MISE

$n = 500$. We constructed estimates with the number of terms $M \in \mathcal{M}$, when $\mathcal{M}$ is defined in (34).

Figure 7(a) shows the average of MIAE (blue "1") and the average of MISE (red "2"), for the stagewise minimization estimator for the 15 values of $M \in \mathcal{M}$, over the 500 samples. The average of MISE for the boosting estimator is shown by green "3". Frame (b) shows the Box plots for each sample of MISE values of the stagewise minimization estimator, and frame (c) shows the Box plots for the MIAE values of the stagewise minimization estimator, over $M \in \mathcal{M}$.

Figure 8 shows estimates when the number of terms $M = 300$. The dashed red graphs in frames ((a)–(c)) show the stagewise minimization estimates corresponding to the 0.1-quantile, median, and the 0.9-quantile among the MISE values over 500 samples. The true density is shown as the black solid graph.

Figure 7 shows that the MISE and MIAE start decreasing when $M = 20$; for smaller values of M the modes are not detected at all. Figure 8 shows that qualitatively the modes are detected by the stagewise minimization estimator, although the height of the modes is not estimated accurately. The variability of the estimates is small.

5 Proofs

First we show in Section 5.1 that the stagewise minimization estimator has not much larger empirical risk than the minimization estimator defined in Definition 2. Second, we derive in Section 5.2 an upper bound for the integrated squared error of the stagewise minimization estimator in terms of the optimal approximation error, that is, we give an oracle inequality.

Proof of Theorem 1: Theorem 1 follows combining Lemmas 3 and 4; we choose $\varepsilon = 4B_2^2/(M+1) + \epsilon$ in Lemma 4.

Proof of Theorem 2: Theorem 2 follows directly from Lemma 4 given in Section 5.2.

Notation. Denote

$$f(x, \Lambda) = f(x, \Lambda, \mathcal{D}) = \sum_{\phi \in \mathcal{D}} \lambda_\phi \phi(x), \qquad x \in \mathbf{R}^d, \tag{35}$$

where $\Lambda = (\lambda_\phi)_{\phi \in \mathcal{D}} \in \mathcal{W}$, and $\mathcal{W}$ is the set of vectors of coefficients of finite convex combinations:

$$\mathcal{W} = \mathcal{W}(\mathcal{D}) = \left\{ (\lambda_\phi)_{\phi \in \mathcal{D}} \in \mathbf{R}^{\mathcal{D}} : \lambda_\phi \geq 0, \ \sum_{\phi \in \mathcal{D}} \lambda_\phi = 1, \ \ \#\{\lambda_\phi > 0\} < \infty \right\},$$

where $\mathbf{R}^{\mathcal{D}}$ denotes the set of vectors indexed by the possibly infinite set $\mathcal{D}$. The dictionaries $\mathcal{D}$ which we consider in Section 3.2 have finite cardinality and condition $\#\{\lambda_\phi > 0\} < \infty$ is superfluous in these cases. Note that we have $\{f(\cdot, \Lambda) : \Lambda \in \mathcal{W}\} = \mathrm{co}(\mathcal{D})$, where we denote with $\mathrm{co}(\mathcal{D})$ the convex hull of $\mathcal{D}$.

5.1 Empirical risk of the stagewise minimization estimator

We prove that the stagewise minimization estimator $\hat{f}_n$ does not have much larger empirical risk than the global minimum of the empirical risk over $f(\cdot, \Lambda)$, $\Lambda \in \mathcal{W}$.

The fact that the minimization over a convex hull may be solved in a stagewise manner was noted by Jones (1992) and Barron (1993), Lemma 2, in the Hilbert space context. Lemma 3 may be considered as an empirical version of these results. A result of Maurey given in Pisier (1981) is referred as an origin of this algorithm. Breiman (1993), Lee, Bartlett, and Williamson (1996), Theorem 2, considered the L_2 regression estimation. In the context of density estimation with the log-likelihood empirical risk the stagewise procedure was analyzed in Li and Barron (2000). A general version of the algorithm is analyzed in Zhang (2003).

Lemma 3. *We have for the estimator $\hat{f}_n$ defined in Definition* 1 *that*

$$\gamma_n(\hat{f}_n) \le \inf_{\Lambda \in \mathcal{W}} \gamma_n(f(\cdot, \Lambda)) + \frac{4B_2^2}{M+1} + \epsilon,$$

where γ_n is defined in (1)*, $B_2 = \sup_{\phi \in \mathcal{D}} \|\phi\|_2$, $M \ge 1$ is the number of terms in $\hat{f}_n$, and the mixing coefficients are $\pi_k = 2/(k+2)$, $k = 1, \ldots, M-1$.*

Proof: Let $0 < \delta < B_2^2$ and let $f^* \in \{f(\cdot, \Lambda) : \Lambda \in \mathcal{W}\}$ be such that

$$\gamma_n(f^*) \le \inf_{\Lambda \in \mathcal{W}} \gamma_n(f(\cdot, \Lambda)) + \delta. \tag{36}$$

We prove that for $k = 0, 1, \ldots, M-1$,

$$\gamma_n(\tilde{f}_k) \le \gamma_n(f^*) + \frac{4B_2^2}{k+2} + \epsilon. \tag{37}$$

The lemma follows from (37) by letting $\delta \to 0$, because $\tilde{f}_{M-1} = \hat{f}_n$. Write

$$f^* = \sum_{i=1}^{N} p_i \phi_i,$$

where f^* is defined in (36), $p_i \ge 0$, $\sum_{i=1}^N p_i = 1$, and $\phi_i \in \mathcal{D}$. We have that

$$\begin{aligned} &\sum_{i=1}^{N} p_i \gamma_n((1-\pi_k)\tilde{f}_{k-1} + \pi_k \phi_i) - \gamma_n(f^*) \\ &\quad \le (1-\pi_k)[\gamma_n(\tilde{f}_{k-1}) - \gamma_n(f^*)] + \pi_k^2 B_2^2. \end{aligned} \tag{38}$$

(For a proof of (38) see Klemelä (2005).) From (38) it follows that there is such $\phi^* \in \{\phi_1, \ldots, \phi_N\} \subset \mathcal{D}$ that

$$\begin{aligned}&\gamma_n((1-\pi_k)\tilde{f}_{k-1} + \pi_k\phi^*) - \gamma_n(f^*) \\ &\quad\leq (1-\pi_k)\left[\gamma_n(\tilde{f}_{k-1}) - \gamma_n(f^*)\right] + \pi_k^2 B_2^2. \end{aligned} \tag{39}$$

We prove (37) with induction. From the definition of $\tilde{f}_0$ and (39) it follows that

$$\gamma_n(\tilde{f}_0) \leq \gamma_n(\phi^*) + \epsilon \leq \gamma_n(f^*) + B_2^2 + \epsilon.$$

Thus the case $k = 0$ in (37) is proved. We make the inductive hypothesis that for $k \geq 1$,

$$\gamma_n(\tilde{f}_{k-1}) - \gamma_n(f^*) \leq \frac{4B_2^2}{k+2} + \epsilon \tag{40}$$

and prove the inductive step. We have

$$\begin{aligned}\gamma_n(\tilde{f}_k) - \gamma_n(f^*) &\leq \gamma_n((1-\pi_k)\tilde{f}_{k-1} + \pi_k\phi^*) - \gamma_n(f^*) + \pi_k\epsilon && (41)\\ &\leq (1-\pi_k)\left[\frac{4B_2^2}{k+2} + \epsilon\right] + \pi_k^2 B_2^2 + \pi_k\epsilon && (42)\\ &= \frac{4B_2^2(k+1)}{(k+2)^2} + \epsilon && (43)\\ &\leq \frac{4B_2^2}{k+2} + \epsilon.\end{aligned}$$

In (41) we applied the definition of the $\tilde{f}_k$. In (42) we applied (39) and the inductive hypothesis (40). In (43) we applied the choice $\pi_k = 2/(k+2)$. We have proved (37) and thus the lemma. □

5.2 Oracle inequality

We prove that the theoretical error of a minimization estimator may be bounded by the optimal theoretical error and an additional stochastic term.

Lemma 4. *Let* $\hat{f} \in \{f(\cdot, \Lambda) : \Lambda \in \mathcal{W}\}$ *be such that*

$$\gamma_n(\hat{f}) \leq \inf_{\Lambda\in\mathcal{W}} \gamma_n(f(\cdot, \Lambda)) + \varepsilon, \tag{44}$$

where $\varepsilon > 0$. *Then*

$$\|\hat{f} - f\|_2^2 \leq \inf_{\Lambda\in\mathcal{W}} \|f(\cdot, \Lambda) - f\|_2^2 + \varepsilon + 4\sup_{\phi\in\mathcal{D}} |\nu_n(\phi)|$$

where f is the true density, $f(\cdot, \Lambda)$ is defined in (35)*, and $\nu_n(\phi)$ is the centered empirical operator defined in* (15)*.*

Proof: Let $\varepsilon' > 0$ and let $f^0 \in \{f(\cdot, \Lambda) : \Lambda \in \mathcal{W}\}$ be such that

$$\|f^0 - f\|_2^2 \leq \inf_{\Lambda \in \mathcal{W}} \|f(\cdot, \Lambda) - f\|_2^2 + \varepsilon'.$$

We have for $g = \hat{f}$, $g = f^0$,

$$\|g - f\|_2^2 - \gamma_n(g) = \|f\|_2^2 - 2\int_{\mathbf{R}^d} fg + \frac{2}{n}\sum_{i=1}^n g(X^i).$$

Thus,

$$\|\hat{f} - f\|_2^2 - \gamma_n(\hat{f}) + \gamma_n(f^0) - \|f^0 - f\|_2^2 = 2\nu_n(\hat{f} - f^0). \tag{45}$$

Thus,

$$\begin{aligned}
&\|\hat{f} - f\|_2^2 - \|f^0 - f\|_2^2 \\
&\quad = \|\hat{f} - f\|_2^2 - \gamma_n(\hat{f}) + \gamma_n(\hat{f}) - \|f^0 - f\|_2^2 \\
&\quad \leq \|\hat{f} - f\|_2^2 - \gamma_n(\hat{f}) + \gamma_n(f^0) + \varepsilon - \|f^0 - f\|_2^2 \qquad (46)\\
&\quad = 2\nu_n(\hat{f} - f^0) + \varepsilon. \qquad (47)
\end{aligned}$$

In (46) we applied (44), and in (47) we applied (45). Denote the vectors of the empirical and theoretical coefficients as

$$\Theta_n = \left(\frac{1}{n}\sum_{i=1}^n \phi(X^i)\right)_{\phi \in \mathcal{D}}, \qquad \Theta_f = \left(\int_{\mathbf{R}^d} f\phi\right)_{\phi \in \mathcal{D}}.$$

Let $\hat{\Lambda}_n = (\hat{\lambda}_\phi)_{\phi \in \mathcal{D}} \in \mathcal{W}$, $\Lambda^0 = (\lambda_\phi^0)_{\phi \in \mathcal{D}} \in \mathcal{W}$ be such that

$$\hat{f} = f(\cdot, \hat{\Lambda}_n), \qquad f^0 = f(\cdot, \Lambda^0).$$

Then

$$\begin{aligned}
\nu_n(\hat{f} - f^0) &= (\Theta_n - \Theta_f)^T(\hat{\Lambda}_n - \Lambda^0) \\
&\leq \left[\|\hat{\Lambda}_n\|_{l_1} + \|\Lambda^0\|_{l_1}\right] \|\Theta_n - \Theta_f\|_{l_\infty} \\
&\leq 2\|\Theta_n - \Theta_f\|_{l_\infty} \\
&= 2\sup_{\phi \in \mathcal{D}} |\nu_n(\phi)|. \qquad (48)
\end{aligned}$$

Here we applied the notation $\|\Lambda\|_{l_1} = \sum_{\phi\in\mathcal{D}} |\lambda_\phi|$ and $\|\Lambda\|_{l_\infty} = \sup_{\phi\in\mathcal{D}} |\lambda_\phi|$, and the fact that $\|\hat{\Lambda}_n\|_{l_1}, \|\Lambda^0\|_{l_1} \leq 1$. The lemma follows from (47) and (48) by letting $\varepsilon' \to 0$. □

6 Discussion

Estimators based on local averaging, like kernel estimators, suffer from the curse of dimensionality. Stagewise minimization can however work in high dimensional cases, but even this estimator cannot work uniformly well over such large classes as Sobolev balls. The mixture classes provide however an example of a density class where the stagewise minimization estimator behaves uniformly well, and this class is a large non-parametric density class of practical interest.

The stagewise minimization estimator is related to the family of boosting estimators. Unlike usual boosting estimators, the stagewise minimization estimator uses a fixed sequence of weights. When using the stagewise minimization estimator one does not need to choose the number of terms (or the size of the weights) using a model selection procedure (regularization).

Appendix A: Proof of (17)

Using the notation in (35) Eq. (17) may be written as

$$\sup_{\Lambda\in\mathcal{W}(\mathbb{G})} \inf_{\Lambda'\in\mathcal{W}(\mathcal{D}_\delta)} \left\| f(\cdot, \Lambda, \mathbb{G}) - f(\cdot, \Lambda', \mathcal{D}_\delta)\right\|_2 \leq \delta.$$

Let Φ_ψ, $\psi \in \mathcal{D}_\delta$, be the collection of those $\phi \in \mathbb{G}$ for which ψ is the closest member of $\mathcal{D}_\delta$:

$$\{\phi \in \mathbb{G} : \psi = \text{argmin}_{\psi'\in\mathcal{D}_\delta} \|\psi' - \phi\|_2\}.$$

We may solve ties arbitrarily to make a partition of $\mathbb{G}$:

$$\mathbb{G} = \cup_{\psi\in\mathcal{D}_\delta} \Phi_\psi, \qquad \Phi_\psi \cap \Phi_{\psi'},\ \psi \neq \psi'.$$

Let $\Lambda \in \mathcal{W}(\mathbb{G})$, $\Lambda = (\lambda_\phi)_{\phi\in\mathbb{G}}$. Let $\Lambda' \in \mathcal{W}(\mathcal{D}_\delta)$, $\Lambda' = (\lambda'_\psi)_{\psi\in\mathcal{D}_\delta}$ be such that

$$\lambda'_\psi = \sum_{\phi\in\Phi_\psi} \lambda_\phi.$$

Now

$$f(\cdot, \Lambda, \mathbb{G}) = \sum_{\phi\in\mathbb{G}} \lambda_\phi \phi = \sum_{\psi\in\mathcal{D}_\delta} \sum_{\phi\in\Phi_\psi} \lambda_\phi \phi$$

and

$$f(\cdot, \Lambda', \mathcal{D}_\delta) = \sum_{\psi \in \mathcal{D}_\delta} \lambda'_\psi \psi = \sum_{\psi \in \mathcal{D}_\delta} \left(\sum_{\phi \in \Phi_\psi} \lambda_\phi \right) \psi.$$

Thus

$$\begin{aligned} \left\| f(\cdot, \Lambda, \mathbb{G}) - f(\cdot, \Lambda', \mathcal{D}_\delta) \right\|_2 &\leq \sum_{\psi \in \mathcal{D}_\delta} \sum_{\phi \in \Phi_\psi} \lambda_\phi \|\phi - \psi\|_2 \\ &\leq \max_{\psi \in \mathcal{D}_\delta} \sup_{\phi \in \Phi_\psi} \|\phi - \psi\|_2 \leq \delta. \end{aligned}$$

We have proved (17). □

Acknowledgments Writing of this article was financed by Deutsche Forschungsgemeinschaft under projects MA1026/8-1 and MA1026/8-2. I wish to thank the referees for helpful comments and pointing out references.

References

Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Information Theory*, *39*, 930–945.

Biau, G., & Devroye, L. (2005). Density estimation by the penalized combinatorial method. *J. Multivariate Anal.*, *94*, 196–208.

Birgé, L., & Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields*, *97*, 113–150.

Breiman, L. (1993). Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans. Inform. Theory*, *39*(3), 999–1013.

Breiman, L. (1996). Bias, variance, and arcing classifiers, Technical report 460, Department of Statistics, Univ. Calif. at Berkeley.

Breiman, L. (1998). Arcing classifiers. *Ann. Statist.*, *26*(3), 801–824.

Bühlmann, P. (2002). Consistency for L_2Boosting and matching pursuit with trees and tree-type basis functions, Research report 109, ETH Zürich.

Carl, B. (1997). Metric entropy of convex hulls in Hilbert spaces. *Bull. London Math. Soc.*, *29*, 452–458.

Carl, B., Kyrezi, I., & Pajor, A. (1999). Metric entropy of convex hulls in Banach spaces. *J. London Math. Soc.*, *60*(2), 871–896.

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Machine learning*: *Proceedings of the thirteenth international conference* (pp. 148–156) San Fransisco: Morgan Kaufman.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, *28*(2), 337–407. With discussion.

Friedman, J. H., Stuetzle, W., & Schroeder, A. (1984). Projection pursuit density estimation. *Amer. Statist. Assoc.*, *79*, 599–608.

Genovese, C. R., & Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.*, *28*(4), 1105–1127.

Ghosal, S., & van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, *29*(5), 1233–1263.

Gill, P. E., Murray, W., & Wright, M. H. (1991). *Numerical linear algebra and optimization*. Addison Wesley.

Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rate for projection pursuit regression. *Ann. Statist.*, *20*(1), 608–613.

Juditsky, A., & Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.*, *28*(3), 681–712.

Klemelä, J. (2005). Density estimation with stagewise optimization of the empirical risk. Technical report, http://www.denstruct.net.

Kolmogorov, A. N., & Tikhomirov, V. M. (1961). ϵ-entropy and ϵ-capacity of sets in function spaces. *Translations of the American Math. Soc.*, *17*, 277–364.

Lee, W. S., Bartlett, R. C., & Williamson, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Information Theory*, *42*(6), 2118–2132.

Li, J., & Barron, A. (2000). Mixture density estimation. In: S. Solla, T. Leen and K. Müller (Eds), *Advances in Neural Information Processing systems* (vol. 12). MIT Press.

Lugosi, G. (2002). Pattern classification and learning theory. In: L. Györfi (Eds.), *Principles of nonparametric learning* (pp. 1–56) Springer.

Marron, J. S., & Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.*, *20*, 712–736.

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (2000). Functional gradient techniques for combining hypothesis. In A.J. Smola, P.J. Bartlett, B. Schölkopf & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 221–246). Cambridge, MA, MIT Press.

Mendelson, S. (2002). On the size of convex hulls of small sets. *J. Machine Learning Res.*, *2*, 1–18.

Nesterov, Y., & Nemirovskii, A. (1994). Interior Point Polynomial Algorithms in Convex Programming, SIAM studies in applied mathematics.

Ossiander, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.*, *15*, 897–919.

Pisier, G. (1981). Remarque sur un resultat non publie de B. Maurey. In Seminaire d'analyse fonctionelle 1980-1981, Ecole Polytechnique, Palaiseau, pp. 1–12.

Pollard, D. (1989). Asymptotics via empirical processes. *Statist. Science*, *4*(4), 341–366.

Priebe, C. E. (1994). Adaptive mixtures. *J. Amer. Statist. Assoc.*, *89*, 796–806.

Rakhlin, A., Panchenko, D., & Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: Probab. and Statist.*, *9*, 220–229.

Ridgeway, G. (2002). Looking for lumps: Boosting and bagging for density estimation. *Comput. Statist. Data Anal.*, *38*, 379–392.

Rigollet, P., & Tsybakov, A. B. (2006). Linear and convex aggregation of density estimators, Technical report, Université Paris 6.

Rosset, S., & Segal, E. (2002). Boosting density estimation. In *Proceedings of the 16th international conference on neural information processing systems* (NIPS).

van de Geer, S. A. (2000). *Empirical processes in M-estimation*. Cambridge University Press.

van der Vaart, A. D., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Zhang, T. (2003). Sequential greedy approximation for certain convex optimization problems. *IEEE Trans. Information Theory*, *49*, 682–691.