

# Sparse density estimation with $\ell_1$ penalties

Florentina Bunea<sup>1</sup>, Alexandre B. Tsybakov<sup>2</sup>, and Marten H. Wegkamp<sup>1</sup>

<sup>1</sup> Florida State University, Tallahassee FL 32306, USA,  
flori@stat.fsu.edu, wegwamp@stat.fsu.edu

<sup>2</sup> Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI, France  
tsybakov@ccr.jussieu.fr

**Abstract.** This paper studies oracle properties of  $\ell_1$ -penalized estimators of a probability density. We show that the penalized least squares estimator satisfies sparsity oracle inequalities, i.e., bounds in terms of the number of non-zero components of the oracle vector. The results are valid even when the dimension of the model is (much) larger than the sample size. They are applied to estimation in sparse high-dimensional mixture models, to nonparametric adaptive density estimation and to the problem of aggregation of density estimators.

## 1 Introduction

Let  $X_1, \dots, X_n$  be independent random variables with common unknown density  $f$  in  $\mathbb{R}^d$ . Let  $\{f_1, \dots, f_M\}$  be a finite set of functions with  $f_j \in L_2(\mathbb{R}^d)$ ,  $j = 1, \dots, M$ , called a dictionary. We consider estimators of  $f$  that belong to the linear span of  $\{f_1, \dots, f_M\}$ . We will be particularly interested in the case where  $M \gg n$ , where  $n$  is the sample size. Denote by  $f_\lambda$  the linear combinations

$$f_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x), \quad \lambda = (\lambda_1, \dots, \lambda_M) \in \mathbb{R}^M.$$

We provide below a number of examples where such estimates are of importance.

- *Estimation in sparse mixture models.* Assume that the density  $f$  can be represented as a finite mixture  $f = f_{\lambda^*}$  where  $f_j$  are known probability densities and  $\lambda^*$  is a vector of mixture probabilities. The number  $M$  can be very large, much larger than the sample size  $n$ , but we believe that the representation is sparse, i.e., that very few coordinates of  $\lambda^*$  are non-zero. Our goal is to estimate  $\lambda^*$  by a vector  $\tilde{\lambda}$  that adapts to this unknown sparsity.
- *Adaptive nonparametric density estimation.* Assume that the density  $f$  is a smooth function, and  $\{f_1, \dots, f_M\}$  are the first  $M$  functions from a basis in  $L_2(\mathbb{R}^d)$ . If the basis is orthonormal, a natural idea is to estimate  $f$  by an orthogonal series estimator which has the form  $f_{\tilde{\lambda}}$  with  $\tilde{\lambda}$  having the coordinates  $\tilde{\lambda}_j = n^{-1} \sum_{i=1}^n f_j(X_i)$ . However, it is well known that such estimators are very sensitive to the choice of  $M$ , and a data-driven selection of  $M$  or thresholding is needed to achieve adaptivity (cf., e.g., [25, 17, 3,

15]); moreover these methods have been applied with  $M \leq n$ . We would like to cover more general problems where the system  $\{f_j\}$  is not necessarily orthonormal, even not necessarily a basis,  $M$  is not necessarily smaller than  $n$ , but an estimate of the form  $f_{\hat{\lambda}}$  still achieves, adaptively, the optimal rates of convergence.

- *Aggregation of density estimators.* Assume now that  $f_1, \dots, f_M$  are some preliminary estimators of  $f$  constructed from a training sample independent of  $(X_1, \dots, X_n)$ , and we would like to aggregate  $f_1, \dots, f_M$ . This means that we would like to construct a new estimator, the aggregate, which is approximately as good as the best among  $f_1, \dots, f_M$  or approximately as good as the best linear or convex combination of  $f_1, \dots, f_M$ . Our aggregates will be of the form  $f_{\hat{\lambda}}$  with suitably chosen weights  $\hat{\lambda} = \hat{\lambda}(X_1, \dots, X_n) \in \mathbb{R}^M$ .

In this paper, we suggest a data-driven choice of  $\hat{\lambda}$  that can be used in all the examples mentioned above and also more generally. We define  $\hat{\lambda}$  as a minimizer of an  $\ell_1$ -penalized criterion, that we call SPADES (SPArse Density ESTimation). The idea of  $\ell_1$ -penalized estimation is widely used in the statistical literature, mainly in linear regression where it is usually referred to as the Lasso criterion [26, 7, 11, 10, 14, 21]. For Gaussian sequence models or for regression with orthogonal design matrix the Lasso is equivalent to soft thresholding [9, 20]. Recently, Lasso methods have been extended to nonparametric regression with general fixed or random design [4–6], as well as to some classification and other more general prediction type models [18, 19, 29].

We prove below oracle inequalities for the  $L_2$ -risk of the proposed SPADES estimator, and we obtain as corollaries some sparsity or optimality properties of this estimator for the three above mentioned examples.

## 2 Definition of SPADES

Consider the  $L_2(\mathbb{R}^d)$  norm

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

associated with the inner product

$$\langle g, h \rangle = \int_{\mathbb{R}^d} g(x)h(x) \, dx$$

for  $g, h \in L_2(\mathbb{R}^d)$ . Note that if the density  $f$  belongs to  $L_2(\mathbb{R}^d)$  and  $X$  has the same distribution as  $X_i$ , we have, for any  $g \in L_2$ ,

$$\langle g, f \rangle = \mathbb{E}g(X),$$

where the expectation is taken under  $f$ . Moreover

$$\|f - g\|^2 = \|f\|^2 + \|g\|^2 - 2\langle f, g \rangle = \|f\|^2 + \|g\|^2 - 2\mathbb{E}g(X). \quad (1)$$

In view of identity (1), minimizing  $\|f_\lambda - f\|^2$  in  $\lambda$  is the same as minimizing

$$\gamma(\lambda) = -2\mathbb{E}f_\lambda(X) + \|f_\lambda\|^2.$$

The function  $\gamma(\lambda)$  depends on  $f$  but can be approximated by its empirical counterpart

$$\hat{\gamma}(\lambda) = -\frac{2}{n} \sum_{i=1}^n f_\lambda(X_i) + \|f_\lambda\|^2.$$

This motivates the use of  $\hat{\gamma} = \hat{\gamma}(\lambda)$  as the empirical criterion, see, for instance, [3, 25, 30].

Let  $0 < \delta < 1/2$  be a small tuning parameter. We define the penalty

$$\text{pen}(\lambda) = 2 \sum_{j=1}^M \omega_j |\lambda_j| \quad \text{with} \quad \omega_j = 2L_j \sqrt{\frac{2 \log(M/\delta)}{n}} \quad (2)$$

and we propose the following data-driven choice of  $\lambda$ :

$$\begin{aligned} \hat{\lambda} &= \arg \min_{\lambda \in \mathbb{R}^M} \{\hat{\gamma}(\lambda) + \text{pen}(\lambda)\} \\ &= \arg \min_{\lambda \in \mathbb{R}^M} \left\{ -\frac{2}{n} \sum_{i=1}^n f_\lambda(X_i) + \|f_\lambda\|^2 + 2 \sum_{j=1}^M \omega_j |\lambda_j| \right\}. \end{aligned}$$

The estimator of the density  $f$ , henceforth called the *SPADES estimator*, is defined by

$$f^\spadesuit(x) = f_{\hat{\lambda}}(x), \quad \forall x \in \mathbb{R}^d.$$

Our estimate can be computed easily even if  $M \gg n$  and retains the desirable theoretical properties of other density estimators the computation of which may become problematic in such case. We refer to [30] for optimal bandwidth selection for kernel density estimators using the same empirical criterion as ours, to [8] for a thorough overview on combinatorial methods in density estimation, and to [2, 28] for density estimation using penalization by the dimension over a sequence of models and the same empirical criterion as ours.

### 3 Oracle inequalities for SPADES

For any  $\lambda \in \mathbb{R}^M$ , let

$$J(\lambda) = \{j \in \{1, \dots, M\} : \lambda_j \neq 0\}$$

be the set of non-zero indices of  $\lambda$  and

$$M(\lambda) = |J(\lambda)| = \sum_{j=1}^M I\{\lambda_j \neq 0\}$$

its cardinality. Here  $I\{\cdot\}$  denotes the indicator function. Set  $L_j = \|f_j\|_\infty$  for  $1 \leq j \leq M$  where  $\|\cdot\|_\infty$  is the  $L_\infty$  norm on  $\mathbb{R}^d$ . We begin with following preliminary lemma.

IV

**Lemma 1.** *Assume that  $L_j < \infty$  for  $j = 1, \dots, M$ . For all  $n \geq 1$  and  $\lambda \in \mathbb{R}^M$  we have*

$$\|f^\spadesuit - f\|^2 + \sum_{j=1}^M \omega_j |\widehat{\lambda}_j - \lambda_j| \leq \|f_\lambda - f\|^2 + 4 \sum_{j \in J(\lambda)} \omega_j |\widehat{\lambda}_j - \lambda_j| \quad (3)$$

with probability at least  $1 - 2\delta$ , for any  $0 < \delta < 1/2$ .

*Proof.* By the definition of  $\widehat{\lambda}$ ,

$$-\frac{2}{n} \sum_{i=1}^n f_{\widehat{\lambda}}(X_i) + \|f_{\widehat{\lambda}}\|^2 + 2 \sum_{j=1}^M \omega_j |\widehat{\lambda}_j| \leq -\frac{2}{n} \sum_{i=1}^n f_\lambda(X_i) + \|f_\lambda\|^2 + 2 \sum_{j=1}^M \omega_j |\lambda_j|$$

for all  $\lambda \in \mathbb{R}^M$ . We rewrite this inequality as

$$\begin{aligned} \|f^\spadesuit - f\|^2 &\leq \|f_\lambda - f\|^2 + 2 \sum_{j=1}^M \left( \frac{1}{n} \sum_{i=1}^n f_j(X_i) - \mathbb{E}f_j(X_i) \right) (\widehat{\lambda}_j - \lambda_j) \\ &\quad + 2 \sum_{j=1}^M \omega_j |\lambda_j| - 2 \sum_{j=1}^M \omega_j |\widehat{\lambda}_j|. \end{aligned}$$

Define the random variables

$$V_j = \frac{1}{n} \sum_{i=1}^n \{f_j(X_i) - \mathbb{E}f_j(X_i)\}.$$

By Hoeffding's inequality, it follows that the probability of the event

$$A = \bigcap_{j=1}^M \{2|V_j| \leq \omega_j\}$$

exceeds

$$1 - 2 \sum_{j=1}^M \exp\left(-\frac{n\omega_j^2}{8L_j^2}\right) = 1 - 2\delta.$$

Then, on the event  $A$ ,

$$\|f^\spadesuit - f\|^2 \leq \|f_\lambda - f\|^2 + \sum_{j=1}^M \omega_j |\widehat{\lambda}_j - \lambda_j| + 2 \sum_{j=1}^M \omega_j |\lambda_j| - 2 \sum_{j=1}^M \omega_j |\widehat{\lambda}_j|.$$

Add  $\sum_j \omega_j |\widehat{\lambda}_j - \lambda_j|$  to both sides of the inequality to obtain

$$\begin{aligned}
& \|f^\spadesuit - f\|^2 + \sum_{j=1}^M \omega_j |\widehat{\lambda}_j - \lambda_j| \\
& \leq \|f_\lambda - f\|^2 + 2 \sum_{j=1}^M \omega_j |\widehat{\lambda}_j - \lambda_j| + 2 \sum_{j=1}^M \omega_j |\lambda_j| - 2 \sum_{j=1}^M \omega_j |\widehat{\lambda}_j| \\
& \leq \|f_\lambda - f\|^2 + 2 \sum_{j \in J(\lambda)} \omega_j |\widehat{\lambda}_j - \lambda_j| + 2 \sum_{j=1}^M \omega_j |\lambda_j| - 2 \sum_{j \in J(\lambda)} \omega_j |\widehat{\lambda}_j| \\
& \leq \|f_\lambda - f\|^2 + 4 \sum_{j \in J(\lambda)} \omega_j |\widehat{\lambda}_j - \lambda_j|
\end{aligned}$$

where we used that  $\lambda_j = 0$  for  $j \notin J(\lambda)$  and the triangle inequality. ■

For any fixed integer  $M \geq 2$  we introduce the following notation. We denote by  $\Psi_M = (\langle f_i, f_j \rangle)_{1 \leq i, j \leq M}$  the Gram matrix associated with  $f_1, \dots, f_M$  and by  $I_M$  the  $M \times M$  identity matrix. The next theorem will be shown under the following assumption.

**ASSUMPTION (I).** *There exists  $\kappa_M > 0$  such that  $\Psi_M - \kappa_M I_M$  is positive semi-definite.*

**Theorem 1.** *Let Assumption (I) hold and let  $L_j < \infty$  for  $1 \leq j \leq M$ . Then, for all  $n \geq 1$ ,  $\alpha > 1$  and all  $\lambda \in \mathbb{R}^M$ , we have with probability at least  $1 - 2\delta$ ,*

$$\|f^\spadesuit - f\|^2 + \frac{\alpha}{\alpha - 1} \sum_{j=1}^M \omega_j |\widehat{\lambda}_j - \lambda_j| \leq \frac{\alpha + 1}{\alpha - 1} \|f_\lambda - f\|^2 + \frac{64\alpha^2}{\alpha - 1} \frac{G(\lambda) \log \frac{M}{\delta}}{n \kappa_M} \quad (4)$$

where  $G(\lambda) \triangleq \sum_{j \in J(\lambda)} L_j^2$ .

*Proof.* By Assumption (I) we have

$$\|f_\lambda\|^2 = \sum_{1 \leq i, j \leq M} \lambda_i \lambda_j \int_{\mathbb{R}^d} f_i(x) f_j(x) dx \geq \kappa_M \sum_{j \in J(\lambda)} \lambda_j^2.$$

By the definition of  $\omega_j$  and the Cauchy-Schwarz inequality, we find

$$\begin{aligned}
4 \sum_{j \in J(\lambda)} \omega_j |\widehat{\lambda}_j - \lambda_j| & \leq 4 \sqrt{\frac{8 \log(M/\delta)}{n}} \sum_{j \in J(\lambda)} L_j |\widehat{\lambda}_j - \lambda_j| \\
& \leq 8 \sqrt{\frac{2G(\lambda) \log(M/\delta)}{n \kappa_M}} \|f^\spadesuit - f_\lambda\|.
\end{aligned}$$

VI

Combination with Lemma 1 yields that with probability greater than  $1 - 2\delta$ ,

$$\begin{aligned} \|f^\spadesuit - f\|^2 + \sum_{j=1}^M \omega_j |\hat{\lambda}_j - \lambda_j| &\leq \|f_\lambda - f\|^2 + 8\sqrt{\frac{2G(\lambda)\log(M/\delta)}{n\kappa_M}} \|f^\spadesuit - f_\lambda\| \quad (5) \\ &\leq \|f_\lambda - f\|^2 + b(\|f^\spadesuit - f\| + \|f_\lambda - f\|) \end{aligned}$$

where  $b = 8\sqrt{2G(\lambda)\log(M/\delta)}/\sqrt{n\kappa_M}$ . This inequality is of the form  $v^2 + d \leq c^2 + vb + cb$  with

$$v = \|f^\spadesuit - f\|, \quad c = \|f_\lambda - f\|, \quad d = \sum_{j=1}^M \omega_j |\hat{\lambda}_j - \lambda_j|.$$

After applying the inequality  $2xy \leq x^2/\alpha + \alpha y^2$  ( $x, y \in \mathbb{R}$ ,  $\alpha > 1$ ) twice, we easily find

$$v^2 + d \leq v^2/(2\alpha) + \alpha b^2 + (2\alpha + 1)/(2\alpha) c^2,$$

whence

$$v^2 + d\{\alpha/(\alpha - 1)\} \leq \alpha/(\alpha - 1)\{b^2(\alpha/2) + c^2(\alpha + 1)/\alpha\}. \quad (6)$$

The claim of the theorem follows from (5) and (6).

#### 4 Oracle inequalities for SPADES: the local mutual coherence assumption

When the dictionary  $\{f_1, \dots, f_M\}$  is over-complete (see, e.g., discussion in [10]) Assumption (I) may not be satisfied. Nevertheless, as discussed in [10], for many interesting dictionaries the Gram matrices satisfy the mutual coherence property, that is the correlations

$$\rho_M(i, j) = \frac{\langle f_i, f_j \rangle}{\|f_i\| \|f_j\|}, \quad i, j = 1, \dots, M,$$

admit a uniform (small) upper bound for all  $i \neq j$ . It can be shown that if this bound, called coherence, is relatively small, namely of the order  $O(1/M(\lambda))$  for some  $\lambda$ , then the oracle inequalities of the previous section remain valid for such  $\lambda$ . The assumption that the correlations are small for all  $i \neq j$  may still be too stringent a requirement in many situations. We relax this here by only imposing bounds on  $\rho_M(i, j)$  with  $j \in J(\lambda)$  and  $i \neq j$ . In our setting the correlations  $\rho_M(i, j)$  with  $i, j \notin J(\lambda)$  can be arbitrarily close to 1 or to  $-1$ . Note that such  $\rho_M(i, j)$  constitute the overwhelming majority of the elements of the correlation matrix if  $J(\lambda)$  is a set of small cardinality:  $M(\lambda) \ll M$ .

For  $\lambda \in \mathbb{R}^M$ , we define our first local coherence number (called *maximal local coherence*) by

$$\rho(\lambda) = \max_{i \in J(\lambda)} \max_{j \neq i} |\rho_M(i, j)|,$$

and we also define

$$F(\lambda) = \max_{j \in J(\lambda)} \frac{\|f_j\|_\infty}{\|f_j\|}.$$

**Theorem 2.** *Assume that  $L_j < \infty$  for  $1 \leq j \leq M$ . Then, with probability at least  $1 - 2\delta$ , for all  $n \geq 1$ ,  $\alpha > 1$  and  $\lambda \in \mathbb{R}^M$  that satisfy*

$$32F(\lambda)\rho(\lambda)M(\lambda) \leq 1, \quad (7)$$

we have the following oracle inequality:

$$\|f^\spadesuit - f\|^2 + \frac{1}{2} \frac{\alpha}{\alpha - 1} \sum_{j=1}^M \omega_j |\hat{\lambda}_j - \lambda_j| \leq \frac{\alpha + 1}{\alpha - 1} \|f_\lambda - f\|^2 + \frac{\alpha^2}{\alpha - 1} \{8F(\lambda)\}^2 M(\lambda) \frac{\log(M/\delta)}{n}.$$

*Proof.* In view of Lemma 1, we need to bound  $\sum_{j \in J(\lambda)} \omega_j |\hat{\lambda}_j - \lambda_j|$ . Set

$$u_j = \hat{\lambda}_j - \lambda_j, \quad U(\lambda) = \sum_{j \in J(\lambda)} |u_j| \|f_j\|, \quad U = \sum_{j=1}^M |u_j| \|f_j\|.$$

Then, by the definition of  $\omega_j$  and  $F(\lambda)$  we obtain

$$\sum_{j \in J(\lambda)} \omega_j |\hat{\lambda}_j - \lambda_j| \leq 2 \sqrt{\frac{2 \log(M/\delta)}{n}} F(\lambda) U(\lambda).$$

Clearly

$$\sum_{i,j \notin J(\lambda)} \langle f_i, f_j \rangle u_i u_j \geq 0$$

and so we obtain

$$\begin{aligned} \sum_{j \in J(\lambda)} u_j^2 \|f_j\|^2 &= \|f^\spadesuit - f_\lambda\|^2 - \sum_{i,j \notin J(\lambda)} u_i u_j \langle f_i, f_j \rangle - 2 \sum_{i \notin J(\lambda)} \sum_{j \in J(\lambda)} u_i u_j \langle f_i, f_j \rangle \\ &\quad - \sum_{i,j \in J(\lambda), i \neq j} u_i u_j \langle f_i, f_j \rangle \\ &\leq \|f^\spadesuit - f_\lambda\|^2 + 2\rho(\lambda) \sum_{i \notin J(\lambda)} |u_i| \|f_i\| \sum_{j \in J(\lambda)} |u_j| \|f_j\| \\ &\quad + \rho(\lambda) \sum_{i,j \in J(\lambda)} |u_i| |u_j| \|f_i\| \|f_j\| \\ &= \|f^\spadesuit - f_\lambda\|^2 + 2\rho(\lambda) U(\lambda) U - \rho(\lambda) U^2(\lambda). \end{aligned} \quad (8)$$

The left-hand side can be bounded by  $\sum_{j \in J(\lambda)} u_j^2 \|f_j\|^2 \geq U^2(\lambda)/M(\lambda)$  using the Cauchy-Schwarz inequality, and we obtain that

$$U^2(\lambda) \leq \|f^\spadesuit - f_\lambda\|^2 M(\lambda) + 2\rho(\lambda) M(\lambda) U(\lambda) U$$

VIII

and, using the properties of a function of degree two in  $U(\lambda)$ , we further obtain

$$U(\lambda) \leq 2\rho(\lambda)M(\lambda)U + \sqrt{M(\lambda)}\|f^\spadesuit - f_\lambda\|. \quad (9)$$

Hence, by Lemma 1, we have with probability at least  $1 - 2\delta$ ,

$$\begin{aligned} & \|f^\spadesuit - f\|^2 + \sum_{j=1}^M \omega_j |\hat{\lambda}_j - \lambda_j| \\ & \leq \|f_\lambda - f\|^2 + 4 \sum_{j \in J(\lambda)} \omega_j |\hat{\lambda}_j - \lambda_j| \\ & \leq \|f_\lambda - f\|^2 + 8\sqrt{\frac{2\log(M/\delta)}{n}} F(\lambda)U(\lambda) \\ & \leq \|f_\lambda - f\|^2 + 8\sqrt{\frac{2\log(M/\delta)}{n}} F(\lambda) \left\{ 2\rho(\lambda)M(\lambda)U + \sqrt{M(\lambda)}\|f^\spadesuit - f_\lambda\| \right\} \\ & \leq \|f_\lambda - f\|^2 + 16F(\lambda)\rho(\lambda)M(\lambda) \sum_{j=1}^M \omega_j |\hat{\lambda}_j - \lambda_j| + 8F(\lambda)\sqrt{\frac{2\log(M/\delta)}{n}} \sqrt{M(\lambda)}\|f^\spadesuit - f_\lambda\|. \end{aligned}$$

For all  $\lambda \in \mathbb{R}^M$  that satisfy relation (7), we find that with probability exceeding  $1 - 2\delta$ ,

$$\begin{aligned} & \|f^\spadesuit - f\|^2 + \frac{1}{2} \sum_{j=1}^M \omega_j |\hat{\lambda}_j - \lambda_j| \\ & \leq \|f_\lambda - f\|^2 + 8F(\lambda)\sqrt{\frac{2\log(M/\delta)}{n}} \sqrt{M(\lambda)}\|f^\spadesuit - f_\lambda\|. \end{aligned}$$

This inequality is of the same form as (5), and we use (6) to conclude the proof.

Note that only a condition on the local coherence (7) is required to obtain the result of Theorem 2. However, even this weak condition can be too strong, because the bound on correlations is *uniform* over  $j \in J(\lambda), i \neq j$ , cf. definition of  $\rho(\lambda)$ . This excludes, for instance, the cases where the correlations can be relatively large for a small number of pairs  $(i, j)$  and almost zero otherwise. A possible solution is to require that the *cumulative local coherence*, rather than the maximal local coherence, be bounded, where the cumulative local coherence is defined as

$$\rho_*(\lambda) = \sum_{i \in J(\lambda)} \sum_{j > i} |\rho_M(i, j)|.$$

**Theorem 3.** *Assume that  $L_j < \infty$  for  $1 \leq j \leq M$ . Then, with probability at least  $1 - 2\delta$ , for all  $n \geq 1$ ,  $\alpha > 1$  and  $\lambda \in \mathbb{R}^M$  that satisfy*

$$32F(\lambda)\rho_*(\lambda)\sqrt{M(\lambda)} \leq 1, \quad (10)$$

*we have the following oracle inequality:*

$$\|f^\spadesuit - f\|^2 + \frac{1}{2} \frac{\alpha}{\alpha - 1} \sum_{j=1}^M \omega_j |\hat{\lambda}_j - \lambda_j| \leq \frac{\alpha + 1}{\alpha - 1} \|f_\lambda - f\|^2 + \frac{\alpha^2}{\alpha - 1} \{8F(\lambda)\}^2 M(\lambda) \frac{\log(M/\delta)}{n}.$$

*Proof.* The proof is similar to that of Theorem 2. With

$$U_*(\lambda) = \sqrt{\sum_{j \in J(\lambda)} u_j^2 \|f_j\|^2}$$

we obtain now the following analogue of (8):

$$\begin{aligned} U_*^2(\lambda) &\leq \|f^\spadesuit - f_\lambda\|^2 + 2\rho_*(\lambda) \max_{i \in J(\lambda), j > i} |u_i| \|f_i\| |u_j| \|f_j\| \\ &\leq \|f^\spadesuit - f_\lambda\|^2 + 2\rho_*(\lambda) U_*(\lambda) \sum_{j=1}^M |u_j| \|f_j\| \\ &= \|f^\spadesuit - f_\lambda\|^2 + 2\rho_*(\lambda) U_*(\lambda) U. \end{aligned}$$

Hence, as in the proof of Theorem 2, we have

$$U_*(\lambda) \leq 2\rho_*(\lambda) U + \|f^\spadesuit - f_\lambda\|,$$

and using the inequality  $U_*(\lambda) \geq U(\lambda)/\sqrt{M(\lambda)}$  we find

$$U(\lambda) \leq 2\rho_*(\lambda) \sqrt{M(\lambda)} U + \sqrt{M(\lambda)} \|f^\spadesuit - f_\lambda\|. \quad (11)$$

Note that (11) differs from (9) only in the fact that the factor  $2\rho(\lambda)M(\lambda)$  on the right hand side is now replaced by  $2\rho_*(\lambda)\sqrt{M(\lambda)}$ . The rest of the proof is identical to that of Theorem 2.

Theorem 3 is useful when we deal with sparse Gram matrices  $\Psi_M$ , i.e., matrices having only a small number  $N$  of non-zero off-diagonal entries. This number will be called a *sparsity index* of matrix  $\Psi_M$ , and is formally defined as

$$N = |\{(i, j) : i, j \in \{1, \dots, M\}, i > j \text{ and } \psi_M(i, j) \neq 0\}|,$$

where  $\psi_M(i, j)$  is the  $(i, j)$ th entry of  $\Psi_M$  and  $|A|$  denotes the cardinality of a set  $A$ . Clearly,  $N < M(M+1)/2$ . We get then the following immediate corollary of Theorem 3.

**Corollary 1.** *Let  $\Psi_M$  be a sparse matrix with sparsity index  $N$ . Then Theorem 3 continues to hold with condition (10) replaced by*

$$32F(\lambda)N\sqrt{M(\lambda)} \leq 1. \quad (12)$$

## 5 Sparse estimation in mixture models

In this section we assume that the true density  $f$  can be represented as a finite mixture

$$f(x) = \sum_{j=1}^M \lambda_j^* f_j(x), \quad (13)$$

X

for some  $\lambda^* \in \Lambda^M$  where  $\Lambda^M$  is a simplex in  $\mathbb{R}^M$ :

$$\Lambda^M = \{\lambda \in \mathbb{R}^M : \lambda_j \geq 0, \sum_{j=1}^M \lambda_j = 1\}$$

and  $f_j$  are known probability densities. The number  $M$  can be very large, much larger than the sample size  $n$ , but we believe that the representation (13) is sparse, i.e., there are very few non-zero coefficients  $\lambda_j^*$ , in other words  $M(\lambda^*) \ll M$ . If the representation (13) is not unique, we consider  $\lambda^*$  corresponding to the most parsimonious representation, i.e., such that  $M(\lambda^*) = \min \left\{ \sum_{j=1}^M I_{\{\lambda_j \neq 0\}} : f = \sum_{j=1}^M \lambda_j f_j \right\}$ .

From Theorems 1 and 2, using that  $\min_{\alpha > 1} \alpha^2 / (\alpha - 1) = 4$ , we easily get the following result.

**Theorem 4.** (i) Let (13) and Assumption (I) hold and let  $L_j < \infty$  for  $1 \leq j \leq M$ . Then, for all  $n \geq 1$ , we have with probability at least  $1 - 2\delta$ ,

$$\|f^\spadesuit - f\|^2 \leq \frac{256 G(\lambda^*) \log(M/\delta)}{n \kappa_M}. \quad (14)$$

(ii) Let (13) hold,  $L_j < \infty$  for  $1 \leq j \leq M$ , and let  $\lambda^* \in \mathbb{R}^M$  satisfy

$$32F(\lambda^*)\rho(\lambda^*)M(\lambda^*) \leq 1. \quad (15)$$

Then, for all  $n \geq 1$ , we have with probability at least  $1 - 2\delta$ ,

$$\|f^\spadesuit - f\|^2 \leq 256 F^2(\lambda^*)M(\lambda^*) \frac{\log(M/\delta)}{n}. \quad (16)$$

**Example.** Let  $f_j$ 's be Gaussian densities in  $\mathbb{R}^d$  with means  $\mu_j$  and unit covariance matrices, such that  $|\mu_j - \mu_k| \geq \tau > 0$ ,  $k \neq j$ , where  $|\cdot|$  stands for the Euclidean distance. Then, for all  $\lambda$ , the mutual coherence satisfies  $\rho(\lambda) \leq \exp(-\tau^2/4)$ , and also  $F(\lambda) \equiv 2^{-d/2} \pi^{-d/4}$ . So, for  $\tau$  large enough (15) is satisfied, and we can apply Theorem 4. It is interesting that the dimension ‘‘helps’’ here: the larger is  $d$ , the smaller is  $F(\lambda)$ . The large constant 256 in (16) is compensated by the small value of  $F(\lambda)$  when the dimension is sufficiently high, say  $d \geq 8$ .

## 6 SPADES for adaptive nonparametric density estimation

We assume in this section that the density  $f$  is defined on  $[0, 1]$ . Let  $f_1, \dots, f_M$  be the first  $M$  functions of the Fourier basis  $\{f_j\}_{j=0}^\infty$  in  $L_2[0, 1]$  defined by  $f_1(x) \equiv 1$ ,  $f_{2k}(x) = \sqrt{2} \cos(2\pi kx)$ ,  $f_{2k+1}(x) = \sqrt{2} \sin(2\pi kx)$  for  $k = 1, 2, \dots$ ,  $x \in [0, 1]$ . Then  $f^\spadesuit$  is a nonparametric estimator of density  $f$ . The following oracle inequality is a direct consequence of Theorem 1.

**Theorem 5.** Let  $f_1, \dots, f_M$  be as defined above, and set  $\omega_j \equiv 4\sqrt{\frac{\log(M/\delta)}{n}}$  for some  $0 < \delta < 1/2$ . Then for all  $n \geq 1$ ,  $\varepsilon > 0$  and all  $\lambda \in \mathbb{R}^M$ , we have with probability at least  $1 - 2\delta$ ,

$$\|f^\spadesuit - f\|^2 \leq (1 + \varepsilon)\|f_\lambda - f\|^2 + C(\varepsilon)\frac{M(\lambda)\log(M/\delta)}{n} \quad (17)$$

where  $C(\varepsilon) > 0$  is a constant depending only on  $\varepsilon$ .

This is a very general inequality that allows one to show that the estimator  $f^\spadesuit$  attains minimax rates of convergence, up to a logarithmic factor simultaneously on various functional classes. In fact, since (17) holds with arbitrary  $\lambda$ , we may use (17) with  $\lambda$  such that  $\lambda_j = 0$  if  $j \geq n^{1/(2\beta+1)}$ , for some  $\beta > 0$ , and thus show in a standard way that  $f^\spadesuit$  attains the minimax rate, up to logarithms, on usual smoothness classes of densities, such as Sobolev or Hölder classes with smoothness index  $\beta$ . Since the rates are attained on one and the same estimator  $f^\spadesuit$  which does not depend on  $\beta$ , this means adaptivity of  $f^\spadesuit$  on the corresponding scales of classes. Results of such type, and even more pointed (without extra logarithmic factors in the rate and sometimes with exact asymptotic minimax constants) are known for various other adaptive density estimators, see, e.g., [3, 12, 15–17, 23, 24] and the references therein.

Although Theorem 5 is somewhat less precise than the benchmarks for these standard classes of densities, it can be used to show adaptivity of  $f^\spadesuit$  on a wider scale of classes than those traditionally considered. In particular, Theorem 5 holds for unbounded densities  $f$ , and even for densities  $f \notin L_2[0, 1]$ .

For example, let  $f$  belong to a subset of  $L_2[0, 1]$  containing possibly unbounded densities and such that  $\|f_{\lambda^*(k)} - f\| \leq a_k, \forall k \leq M$ , for some sequence  $a_k$  tending to 0 very slowly, where  $\lambda^*(k)$  is the vector with components  $\lambda_1 = \langle f, f_1 \rangle, \dots, \lambda_k = \langle f, f_k \rangle, \lambda_j = 0, j > k$ . Then choosing  $k^*$  as a solution of  $a_k \sim (k \log M)/n$  and using (17) with  $\lambda = \lambda^*(k^*)$  we get that our estimator  $f^\spadesuit$  achieves some (slow) convergence rates even for such “bad” classes of unbounded densities.

Another example is given by the  $\mathcal{L}_0$ -classes. Assume that  $f$  belongs to one of the classes

$$\mathcal{L}_0(k) = \left\{ f : [0, 1] \rightarrow \mathbb{R} : f \text{ is a probability density and } |\{j : \langle f, f_j \rangle \neq 0\}| \leq k \right\}$$

where  $k \leq M$  is an unknown integer and  $|A|$  denotes the cardinality of a set  $A$ . We have the following minimax adaptive result.

**Corollary 2.** Let the assumptions of Theorem 5 hold with  $\delta = n^{-2}$  and  $M \leq n^s$  for some  $s > 0$ . Then

$$\sup_{f \in \mathcal{L}_0(k)} \mathbb{P} \left\{ \|f^\spadesuit - f\|^2 \geq b(s) \left( \frac{k \log n}{n} \right) \right\} \leq 2n^{-2}, \quad \forall k \leq M, \quad (18)$$

where  $b(s) > 0$  is a constant depending on  $s$  only.

This can be viewed as an extension to density estimation problem of the adaptive minimax results for  $\mathcal{L}_0$ -classes obtained in the Gaussian sequence space model [1, 13] and in random design regression model [6].

## 7 SPADES for aggregation of density estimators

In this section we assume that  $f_1, \dots, f_M$  are density estimators constructed from a preliminary sample that will be considered as frozen in further discussion.

The aim of aggregation is to construct a new estimator, called aggregate, which is approximately as good as the best among  $f_1, \dots, f_M$  (model selection, or MS-aggregation) or approximately as good as the best linear or convex combination of  $f_1, \dots, f_M$  (L-aggregation and C-aggregation respectively) or approximately as good as the best linear combination of  $D \leq M$  estimators among  $f_1, \dots, f_M$  (subset selection, or S-aggregation). We refer to [4, 22–24, 27] for discussion of aggregation methods. Each type of aggregation corresponds to a particular set  $H^M$  where the weights  $\lambda$  are allowed to lie. The set  $H^M$  is either the whole  $\mathbb{R}^M$  (for L-aggregation), or the simplex  $A^M$  (for C-aggregation), or the set of all vertices of  $A^M$ , except the vertex  $(0, \dots, 0) \in \mathbb{R}^M$  (for MS-aggregation). For subset selection, or S-aggregation we put  $H^M = A^{M,D}$ , where  $A^{M,D}$  denotes the set of all  $\lambda \in \mathbb{R}^M$  having at most  $D$  non-zero coordinates. The corresponding oracles are the values of  $\lambda$  minimizing the risk on these sets.

Using Theorem 1 we obtain the following oracle inequalities for these four types of aggregation.

**Theorem 6.** *Let Assumption (I) be satisfied and  $L_j \leq L < \infty$  for  $1 \leq j \leq M$ . Let  $f^\spadesuit$  be the SPADES estimator with  $\delta = (Mn)^{-1}$ . Then for all  $\varepsilon > 0$  there exists a constant  $C_\varepsilon = C(\varepsilon, L, \kappa_M) > 0$  such that for all integers  $n \geq 1$ ,  $M \geq 2$  and  $1 \leq D \leq M$  we have, with probability greater than  $1 - 2\delta$ ,*

$$\|f^\spadesuit - f\|^2 \leq (1 + \varepsilon) \inf_{1 \leq j \leq M} \|f_j - f\|^2 + C_\varepsilon \frac{\log(Mn)}{n}. \quad (19)$$

$$\|f^\spadesuit - f\|^2 \leq (1 + \varepsilon) \inf_{\lambda \in A^{M,D}} \|f_\lambda - f\|^2 + C_\varepsilon \frac{D \log(Mn)}{n}. \quad (20)$$

$$\|f^\spadesuit - f\|^2 \leq (1 + \varepsilon) \inf_{\lambda \in \mathbb{R}^M} \|f_\lambda - f\|^2 + C_\varepsilon \frac{M \log(Mn)}{n}. \quad (21)$$

$$\|f^\spadesuit - f\|^2 \leq (1 + \varepsilon) \inf_{\lambda \in A^M} \|f_\lambda - f\|^2 + C_\varepsilon \bar{\psi}_n^C(M), \quad (22)$$

where

$$\bar{\psi}_n^C(M) = \begin{cases} (M \log n)/n & \text{if } M \leq \sqrt{n}, \\ \sqrt{(\log M)/n} & \text{if } M > \sqrt{n}. \end{cases}$$

This theorem follows from Theorem 1 via arguments analogous to those used in the regression estimation context in [4], proof of Corollary 3.2. For brevity we do not repeat the proof here.

The remainder terms on the right hand side of inequalities (19), (21) and (22) in Theorem 6 are optimal up to logarithmic factors. This follows from the corresponding lower bounds and the expressions optimal rates of aggregation [24, 23]. We conjecture that the remainder term in (21) is optimal as well, and that this can be shown by a technique similar to that of [4].

In conclusion, SPADES is obtained via one procedure and achieves near optimal aggregation of all four types: model selection, subset selection, convex and linear aggregation.

## References

1. ABRAMOVICH, F, BENJAMINI, Y., DONOHO, D.L. and JOHNSTONE, I.M.. (2006). Adapting to unknown sparsity by controlling the False Discovery Rate. *Annals of Statistics* **34** 584 - 653.
2. BARRON, A., BIRGÉ, L. and MASSART, P. (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113** 301-413.
3. BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. *Festschrift for Lucien LeCam: Research Papers in Probability and Statistics*, D. Pollard, E. Torgersen, and G. Yang, Eds., 55-87. Springer, New York.
4. BUNEA, F., TSYBAKOV, A.B. and WEGKAMP, M.H. (2005). Aggregation for Gaussian regression. Preprint Department of Statistics, Florida State University. *Annals of Statistics*, to appear.
5. BUNEA, F., TSYBAKOV, A.B. and WEGKAMP, M.H. (2006a). Aggregation and sparsity via  $\ell_1$ -penalized least squares. *Proceedings of 19th Annual Conference on Learning Theory, COLT 2006. Lecture Notes in Artificial Intelligence* **4005** 379–391. Springer-Verlag, Heidelberg.
6. BUNEA, F., TSYBAKOV, A.B. and WEGKAMP, M.H. (2006b). Sparsity oracle inequalities for the Lasso. Submitted.
7. CHEN, S., DONOHO, D. and SAUNDERS, M. (2001) Atomic decomposition by basis pursuit. *SIAM Review* **43** 129 - 159.
8. DEVROYE, L. and LUGOSI, G. (2000) *Combinatorial Methods in density estimation*, Springer.
9. DONOHO, D.L. (1995) Denoising via soft-thresholding. *IEEE Trans. Info. Theory* **41** 613-627.
10. DONOHO, D.L., ELAD, M. and TEMLYAKOV, V. (2004). Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. *Manuscript*.
11. DONOHO, D.L. and HUO, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Transactions Inform. Theory* **47** 2845–2862.
12. GOLUBEV, G.K. (1992). Nonparametric estimation of smooth probability densities in  $L_2$ . *Problems of Information Transmission* **28** 44-54.
13. GOLUBEV, G.K. (2002). Reconstruction of sparse vectors in white Gaussian noise. *Problems of Information Transmission* **38** 65–79.
14. GREENSHTEIN, E. and RITOV, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988.
15. HALL, P., KERKYACHARIAN, G., and PICARD, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Annals of Statistics* **26** 922–942.

16. HÄRDLE, W., KERKYACHARIAN, G., PICARD, D., and TSYBAKOV, A. (1998). *Wavelets, Approximation and Statistical Applications*. Lecture Notes in Statistics, vol. 129. Springer, New York.
17. KERKYACHARIAN, G., PICARD, D. and TRIBOULEY, K. (1996).  $L^p$  adaptive density estimation. *Bernoulli* **2** 229–247.
18. KOLTCHINSKII, V. (2005). Model selection and aggregation in sparse classification problems. *Oberwolfach Reports* **2** 2663–2667, Mathematisches Forschungsinstitut Oberwolfach.
19. KOLTCHINSKII, V. (2006). Sparsity in penalized empirical risk minimization. *Submitted*.
20. LOUBES, J. – M. and VAN DE GEER, S. A. (2002). Adaptive estimation in regression, using soft thresholding type penalties. *Statistica Neerlandica* **56** 453 – 478.
21. MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462.
22. NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In P. Bernard, editor, *Ecole d'Été de Probabilités de Saint-Flour 1998*, volume XXVIII of *Lecture Notes in Mathematics*. Springer, New York.
23. RIGOLLET, PH. (2006). Inégalités d'oracle, agrégation et adaptation. PhD thesis, University of Paris 6.
24. RIGOLLET, PH., and TSYBAKOV, A. B. (2004). Linear and convex aggregation of density estimators. <https://hal.ccsd.cnrs.fr/ccsd-00068216>.
25. RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9** 65 –78.
26. TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B.* **58** 267–288.
27. TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *Lecture Notes in Artificial Intelligence*, volume 2777 of *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines*. Springer-Verlag, Heidelberg.
28. VAPNIK, V. N. (1998). *Statistical Learning Theory*. Wiley, New York.
29. VAN DE GEER, S.A. (2006). High dimensional generalized linear models and the Lasso. Research report No.133. Seminar für Statistik, ETH, Zürich.
30. WEGKAMP, M. H. (1999). Quasi-Universal Bandwidth Selection for Kernel Density Estimators. *Canadian Journal of Statistics* **27** 409 – 420.