

# Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss\*

**Cristina Butucea**

*LAMA (UPE-MLV), UPE, Marne La Vallée, France*  
*e-mail: [cristina.butucea@univ-mlv.fr](mailto:cristina.butucea@univ-mlv.fr)*

**Jean-François Delmas**

*CERMICS, École des Ponts, UPE, Champs-sur-Marne, France*  
*e-mail: [delmas@cermics.enpc.fr](mailto:delmas@cermics.enpc.fr); [fischerr@cermics.enpc.fr](mailto:fischerr@cermics.enpc.fr)*

**Anne Dufloy**

*EDF Research & Development, Industrial Risk Management Department, Palaiseau, France*  
*e-mail: [anne.dufloy@edf.fr](mailto:anne.dufloy@edf.fr)*

and

**Richard Fischer**

*LAMA (UPE-MLV), UPE, Marne La Vallée, France*

*CERMICS, École des Ponts, UPE, Champs-sur-Marne, France*  
*e-mail: [fischerr@cermics.enpc.fr](mailto:fischerr@cermics.enpc.fr)*

*EDF Research & Development, Industrial Risk Management Department, Palaiseau, France*

**Abstract:** We study the problem of aggregation of estimators with respect to the Kullback-Leibler divergence for various probabilistic models. Rather than considering a convex combination of the initial estimators  $f_1, \dots, f_N$ , our aggregation procedures rely on the convex combination of the logarithms of these functions. The first method is designed for probability density estimation as it gives an aggregate estimator that is also a proper density function, whereas the second method concerns spectral density estimation and has no such mass-conserving feature. We select the aggregation weights based on a penalized maximum likelihood criterion. We give sharp oracle inequalities that hold with high probability, with a remainder term that is decomposed into a bias and a variance part. We also show the optimality of the remainder terms by providing the corresponding lower bound results.

**MSC 2010 subject classifications:** Primary 62G07, 62M15; secondary 62G05.

---

\*This work is partially supported by the French “Agence Nationale de la Recherche”, CIFRE n° 1531/2012, and by EDF Research & Development, Industrial Risk Management Department

**Keywords and phrases:** Aggregation, Kullback-Leibler divergence, probability density estimation, sharp oracle inequality, spectral density estimation.

Received January 2016.

## Contents

1	Introduction . . . . .	2259
2	Aggregation procedures for the Kullback-Leibler divergence . . . . .	2264
2.1	Non-negative functions with given total mass . . . . .	2266
2.2	Non-negative functions . . . . .	2268
3	Applications and numerical implementation . . . . .	2270
3.1	Probability density model . . . . .	2270
3.2	Spectral density model . . . . .	2272
3.3	Numerical implementation . . . . .	2275
4	Lower bounds . . . . .	2277
4.1	Probability density estimation . . . . .	2277
4.2	Spectral density estimation . . . . .	2277
5	Proofs . . . . .	2278
5.1	Proofs of results in Section 2 . . . . .	2278
5.2	Proofs of results in Section 4 . . . . .	2284
	Appendix . . . . .	2289
	A.1 Results on Toeplitz matrices . . . . .	2289
	A.2 Proof of Lemma 3.4 . . . . .	2290
	References . . . . .	2293

## 1. Introduction

The pure aggregation framework focuses on the best way to combine estimators (seen as deterministic functions) in order to attain nearly the best risk of these estimators and in order to quantify the residual term. Given  $N$  estimators  $f_k, 1 \leq k \leq N$  and a sample  $X = (X_1, \dots, X_n)$  from the model  $f$ , the problem is to find an aggregated estimate  $\hat{f}$  which performs nearly as well as the best  $f_\lambda, \lambda \in \mathcal{U}$ , where:

$$f_\lambda = \sum_{k=1}^N \lambda_k f_k,$$

and  $\mathcal{U}$  is a certain subset of  $\mathbb{R}^N$  (we assume that linear combinations of the estimators are valid candidates). The performance of the estimator is measured by a loss function  $L$ . Common loss functions include  $L^p$  distance (with  $p = 2$  in most cases), Kullback-Leibler or other divergences, Hellinger distance, etc. The aggregation problem can be formulated as follows: find an aggregate estimator  $\hat{f}$  such that for some  $C \geq 1$  constant,  $\hat{f}$  satisfies an oracle inequality in expectation, i.e.:

$$\mathbb{E} \left[ L(f, \hat{f}) \right] \leq C \min_{\lambda \in \mathcal{U}} L(f, f_\lambda) + R_{n,N}, \quad (1)$$

or in deviation, i.e. for  $\varepsilon > 0$  we have with probability greater than  $1 - \varepsilon$ :

$$L(f, \hat{f}) \leq C \min_{\lambda \in \mathcal{U}} L(f, f_\lambda) + R_{n,N,\varepsilon}, \quad (2)$$

with remainder terms  $R_{n,N}$  and  $R_{n,N,\varepsilon}$  which do not depend on  $f$  or  $f_k, 1 \leq k \leq N$ . If  $C = 1$ , then the oracle inequality is sharp.

Historically, three types of problems were identified depending on the choice of  $\mathcal{U}$ . In the model selection problem, the estimator mimics the best estimator amongst  $f_1, \dots, f_N$ , that is  $\mathcal{U} = \{e_k, 1 \leq k \leq N\}$ , with  $e_k = (\lambda_j, 1 \leq j \leq N) \in \mathbb{R}^N$  the unit vector in direction  $k$  given by  $\lambda_j = \mathbf{1}_{\{j=k\}}$ . In the convex aggregation problem,  $f_\lambda$  are the convex combinations of  $f_k, 1 \leq k \leq N$ , i.e.  $\mathcal{U}$  is the simplex  $\Lambda^+ \subset \mathbb{R}^N$  with:

$$\Lambda^+ = \{\lambda = (\lambda_k, 1 \leq k \leq N) \in \mathbb{R}^N, \lambda_k \geq 0 \text{ and } \sum_{1 \leq k \leq N} \lambda_k = 1\}. \quad (3)$$

Finally in the linear aggregation problem,  $\mathcal{U} = \mathbb{R}^N$  is the entire linear span of the initial estimators. Recently, these problems were generalized to  $\ell_q$ -aggregation in [29], who define it as aggregation with  $\mathcal{U} = \ell_q(t_n)$ , i.e. a ball of radius  $t_n > 0$  in  $\ell_q$ -norm,  $0 \leq q \leq 1$ .

Early papers usually consider the  $L^2$  loss in expectation as in (1). For the regression model with random design, optimal bounds for the  $L^2$  loss in expectation for model selection aggregation was considered in [31] and [30], for convex aggregation in [20] with improved results for large  $N$  in [33], and for linear aggregation in [28], where knowledge of the density of the design is required. These results were extended to the case of regression with fixed design for the model selection aggregation in [15] and [16], and for affine estimators in the convex aggregation problem in [14]. A unified aggregation procedure which achieves near optimal loss for all three problems simultaneously was proposed in [8].

For density estimation, first results include [10] and [32] who independently considered the model selection aggregation under the Kullback-Leibler loss in expectation. For positive, integrable functions  $p, q$ , let  $D(p||q)$  denote the generalized Kullback-Leibler divergence given by:

$$D(p||q) = \int p \log(p/q) - \int p + \int q. \quad (4)$$

This is a Bregman divergence, introduced in [7], therefore  $D(p||q)$  is non-negative and  $D(p||q) = 0$  if and only if a.e.  $p = q$ . The Kullback-Leibler loss of an estimator  $\hat{f}$  is given by  $D(f||\hat{f})$ . In [10] and [32], the authors introduced the progressive mixture rule to give a series of estimators which verify oracle inequalities with optimal remainder terms. They introduced the sequential risk for estimating  $f$  with weights  $\delta$ :

$$R_{seq}(f, n, \delta) = \frac{1}{n+1} \sum_{i=0}^n \mathbb{E} \left[ D(f||\hat{f}_{\delta,i}) \right],$$

where the estimators  $\hat{f}_{\delta,i}$  depend only on the first  $i$  observations. They proved that there exists a progressive mixture procedure associated to the weights  $\delta_{seq}^*$  such that

$$R_{seq}(f, n, \delta_{seq}^*) \leq \inf_{j \geq 1} \left( \frac{1}{n+1} \log \frac{1}{\pi_j} + R_{seq}(f, n, \delta_j) \right).$$

This method was later generalized as the mirror averaging algorithm in [21] and applied to various problems. Corresponding lower bounds which ensure the optimality of this procedure were shown in [22]. The convex and linear aggregation problems for densities under the  $L^2$  loss in expectation were considered in [26].

While a lot of papers considered the expected value of the loss, relatively few papers address the question of optimality in deviation, that is with high probability as in (2). For the regression problem with random design, [1] shows that the progressive mixture method is deviation sub-optimal for the model selection aggregation problem, and proposes a new algorithm which is optimal for the  $L^2$  loss in deviation and expectation as well. Another deviation optimal method based on sample splitting and empirical risk minimization on a restricted domain was proposed in [23]. For the fixed design regression setting, [25] considers all three aggregation problems in the context of generalized linear models and gives constrained likelihood maximization methods which are optimal in both expectation and deviation with respect to the Kullback-Leibler loss. More recently, [13] extends the results of [25] for model selection by introducing the  $Q$ -aggregation method and giving a greedy algorithm which produces a sparse aggregate achieving the optimal rate in deviation for the  $L^2$  loss. More general properties of this method applied to other aggregation problems as well are discussed in [12].

For the density estimation, optimal bounds in deviation with respect to the  $L^2$  loss for model selection aggregation are given in [3]. The author gives a non-asymptotic sharp oracle inequality under the assumption that  $f$  and the estimators  $f_k$ ,  $1 \leq k \leq N$  are bounded, and shows the optimality of the remainder term by providing the corresponding lower bounds as well. The penalized empirical risk minimization procedure introduced in [3] inspired our current work. Here, we consider a more general framework which incorporates, as a special case, the density estimation problem. Moreover, we give results in deviation for the Kullback-Leibler loss instead of the  $L^2$  loss considered in [3].

The spectral density model is equivalent to estimating the covariance function  $cov(X, X_{+h})$  for all integers  $h$  from a stationary Gaussian sequence. It can be seen as the estimation of a Toeplitz covariance operator and has many applications.

Linear aggregation of lag window spectral density estimators with  $L^2$  loss was studied in [11]. The method we propose is more general as it can be applied to any set of estimators  $f_k$ ,  $1 \leq k \leq N$ , not only kernel estimators. Moreover, we establish oracle inequalities for the model selection type of risk, which has a less aggressive target than the linear aggregation risk (it aims the minimal risk over a finite choice instead the minimal risk over all

linear combinations), but leads to better performance than the linear aggregation due to smaller aggregation price. Also, this paper concerns optimal bounds in deviation for the Kullback-Leibler loss instead of the  $L^2$  loss in expectation.

We now present our main contributions. We propose aggregation schemes for the estimation of probability densities on  $\mathbb{R}^d$  and the estimation of spectral densities of stationary Gaussian sequences. We prove sharp oracle inequalities in deviation for the Kullback-Leibler loss. Indeed, for initial estimators  $f_k$ ,  $1 \leq k \leq N$ , we propose an aggregate estimator  $\hat{f}$  that verifies the following: for every  $f$  belonging to a large class of functions  $\mathcal{F}$ , with probability greater than  $1 - \exp(-x)$  for all  $x > 0$ ,

$$D(f \parallel \hat{f}) \leq \min_{1 \leq k \leq N} D(f \parallel f_k) + R_{n,N,x}.$$

We propose two methods of aggregation for non-negative estimators, see Propositions 2.4 and 2.6. Contrary to the usual approach of giving an aggregate estimator which is a linear or convex combination of the initial estimators, we consider an aggregation based on a convex combination of the logarithms of these estimators. The *aggregate estimators*  $\hat{f} = f_\lambda^D$  for the probability density model and  $\hat{f} = f_\lambda^S$  for the spectral density model with  $\hat{\lambda} = \hat{\lambda}(X_1, \dots, X_n) \in \Lambda^+$  maximize a penalized maximum likelihood criterion. The exact form of the convex aggregates  $f_\lambda^D$  and  $f_\lambda^S$  will be precised in later sections for each setup.

The first method concerns estimators with a given total mass and produces an aggregate  $f_\lambda^D$  which has also the same total mass. This method is particularly adapted for density estimation as it provides an aggregate which is also a proper density function. We use this method to propose an adaptive nonparametric density estimator for maximum entropy distributions of order statistics in [9]. The second method, giving the aggregate  $f_\lambda^S$ , does not have the mass conserving feature, but can be applied to a wider range of statistical estimation problems, in particular to spectral density estimation. We show that both procedures give an aggregate which verifies a sharp oracle inequality with a bias and a variance term. Indeed, the bias term is new in oracle inequalities. It appears from the estimation of linear functionals of  $f$  and not from estimation of  $f$  itself. In most cases, this bias is of order  $1/n$  or smaller under mild smoothness assumptions on  $f$  and on the estimators ( $f_k$ ,  $1 \leq k \leq N$ ). The smoothness parameter does not interfere at all with the rates, it only appears in the constants of the remaining term.

When applied to density estimation, our results provide sharp oracle inequalities with the optimal remainder term of order  $\log(N)/n$ . Theorem 3.1 proposes an aggregate estimator  $\hat{f}_*^D$  such that we have, for any  $x > 0$ , with probability higher than  $1 - \exp(-x)$ :

$$D(f \parallel \hat{f}_*^D) \leq \min_{1 \leq k \leq N} D(f \parallel f_k) + \frac{\beta(\log N + x)}{n}$$

where  $\beta$  is an explicit constant depending only on the infinity norm of the logarithms of  $f$  and  $f_k$ ,  $1 \leq k \leq N$ . In this case, the empirical measure provides an unbiased estimator of linear functionals of  $f$  and there is no bias term.

In the case of spectral density estimation, we need to assume a small amount of smoothness for the logarithm of the true spectral density and of the estimators. We require that there exists some  $r > 1/2$  such that the logarithms of the functions belong to the periodic Sobolev space  $W_r$ . We show that this also implies that the spectral density itself belongs to  $W_r$  and see that our assumption is slightly more restrictive than the usual assumption in the literature: that  $f$  belongs to  $W_{1/2}$ . Theorem 3.5 proposes an aggregate estimator  $\hat{f}_*^S$  such that, for any  $x > 0$ , with probability higher than  $1 - \exp(-x)$ :

$$D(f \| \hat{f}_*^S) \leq \min_{1 \leq k \leq N} D(f \| f_k) + \beta' \frac{\log(N) + x}{n} + \frac{\alpha}{n}$$

where  $\beta'$  and  $\alpha$  are constants which depend only on the regularity and the Sobolev norm of the logarithms of  $f$  and  $f_k$ ,  $1 \leq k \leq N$ .

To show the optimality in deviation of the aggregation procedures, we give the corresponding tight lower bounds as well, with the same remainder terms, see Propositions 4.2 and 4.3. This complements the results of [22] and [3] obtained for the density estimation problem. In [22] the lower bound for the expected value of the Kullback-Leibler loss was shown with the same order for the remainder term, while in [3] similar results were obtained in deviation for the  $L^2$  loss. The proof of Proposition 4.2 is quite close to the proof of lower bounds in expectation in [22]. However, this construction based on the Haar basis does not check the additional smoothness assumptions in Proposition 4.3. A new construction of test functions  $f_1, \dots, f_N$  and a new proof for the spectral density model can be found in Section 5.

In conclusion, we introduce two aggregated procedures based on penalized maximum likelihood criterion for exponential family of functions and show their sharp asymptotic optimality in deviation with respect to their Kullback-Leibler risk. The progressive mixture rule is a strong competitor of our procedure for the probability density model: it satisfies an oracle inequality in expectation but with far less restrictive assumptions and it behaves better numerically in our brief examples, for large enough sample sizes (larger than 200). No such competitor exists for the spectral density model that we also treat here.

The rest of the paper is organised as follows. In Section 2 we introduce the notation and give the basic definitions used in the rest of the paper. We present the two types of convex aggregation method for the logarithms in Sections 2.1 and 2.2. We give a general sharp oracle inequality in deviation for the Kullback-Leibler loss for each method and setup. In Section 3 we apply the methods for the probability density together with numerical implementation and the spectral density estimation problems. The results on the corresponding lower bounds can be found in Section 4 for both problems. Proofs were gathered in Section 5. We summarize the properties of Toeplitz matrices and periodic Sobolev spaces in the Appendix.

## 2. Aggregation procedures for the Kullback-Leibler divergence

In this section, we propose two convex aggregation methods, suited for models submitted to different type of constraints: non-negative functions with fixed given total mass and non-negative functions without mass restriction. First, we introduce the setups and the aggregation procedures suited for each type of constraint. Then, we state non-asymptotic oracle inequalities for the Kullback-Leibler divergence in a general form.

Let  $h : \mathbb{R}^d \rightarrow \mathbb{R}_+$  be a reference probability density with support  $\mathcal{H} := \{x \in \mathbb{R}^d : h(x) > 0\}$ . Note that  $\mathcal{H}$  can be compact or not. We consider the set  $\mathcal{G}$

$$\mathcal{G} = \{f : \mathbb{R}^d \rightarrow \mathbb{R}_+, \text{ measurable} : \|\log(f/h)\|_\infty < +\infty\},$$

with the convention that  $\log(0/0) = 0$ . Notice that any function in the set  $\mathcal{G}$  is integrable. Moreover, we have that  $\|\log(f/h)\|_\infty < \infty$  implies that necessarily  $f$  and  $h$  have the same support. Thus, the Kullback-Leibler divergence  $D(f' \| f)$ , defined in (4), is finite for any functions  $f', f$  belonging to  $\mathcal{G}$ .

We consider a probabilistic model  $\mathcal{P} = \{P_f; f \in \mathcal{G}\}$ , with  $P_f$  a probability distribution depending on  $f$ . We assume that we have a sample  $X = (X_1, \dots, X_n)$ ,  $n \in \mathbb{N}^*$  with distribution  $P_f$ . In the sequel, the examples we consider are models  $P_f$  that correspond either to a probability density model or to a spectral density model.

More precisely, the **probability density model** corresponds to a sample of i.i.d. random variables with common probability density functions (pdf)  $f$ . The random variables  $X = (X_1, \dots, X_n)$  has joint pdf  $f^{\otimes n}(x) = f(x_1) \cdot \dots \cdot f(x_n)$ , with  $x = (x_1, \dots, x_n)$  in  $(\mathbb{R}^d)^n$ .

The **spectral density model** corresponds to a sample issued from a stationary Gaussian sequence with spectral density function (sdf)  $f$ . Let  $1/(2\pi)\mathbf{1}_{[-\pi, \pi]}$  be the reference density and  $(X_k)_{k \in \mathbb{Z}}$  be a stationary, centered Gaussian sequence with covariance function  $\gamma_j = \text{Cov}(X_k, X_{k+j})$ , for  $j \in \mathbb{Z}$ . Under the standard assumption that  $\sum_{j=0}^{\infty} |\gamma_j| < +\infty$ , the spectral density  $f$  associated to the process is the even function defined on  $[-\pi, \pi]$  whose Fourier coefficients are  $\gamma_j$ :

$$f(x) = \sum_{j \in \mathbb{Z}} \frac{\gamma_j}{2\pi} e^{ijx} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{\infty} \gamma_j \cos(jx), \quad x \in [-\pi, \pi].$$

Note that,  $\gamma_0 = \int_{-\pi}^{\pi} f(x) dx$ .

Now, suppose we have  $(f_k, 1 \leq k \leq N)$ ,  $N$  distinct estimators belonging to  $\mathcal{G}$ . We propose two aggregation methods based on these estimators and the available sample, that behave, with high probability, as well as the best initial estimator  $f_{k^*}$  in terms of the Kullback-Leibler divergence, where  $k^*$  is defined as:

$$k^* = \operatorname{argmin}_{1 \leq k \leq N} D(f \| f_k). \quad (5)$$

The first aggregation method concerns setups where  $f$  as well as  $(f_k, 1 \leq k \leq N)$  have a given total mass supposed equal to 1 without loss of generality:

$\int f = \int f_1 = \dots = \int f_N = 1$ . This is the case in both the probability density model and in the spectral density model with given variance  $\gamma_0$ .

We set the classical notation for nonparametric exponential family of functions:

$$f(x) = e^{t(x)-\psi} \cdot h(x), \quad f_k(x) = e^{t_k(x)-\psi_k} \cdot h(x) \quad x \in \mathbb{R}^d, \quad (6)$$

where  $\psi = \int h \log(h/f)$  and  $t = \psi + \log(f/h)$ , giving that  $\int t h = 0$  and that  $\psi = \log(\int e^t h)$  is a normalization constant (when needed). Similarly for  $t_k$  and  $\psi_k$ .

We proceed in this case by aggregating  $(t_k, 1 \leq k \leq N)$  into a convex combination

$$t_\lambda = \sum_{k=1}^N \lambda_k t_k, \quad \text{with } \lambda \text{ in the simplex } \Lambda^+,$$

and by constructing

$$f_\lambda^D(x) = e^{t_\lambda(x)-\psi_\lambda} \cdot h(x), \quad \text{with } \psi_\lambda = \log\left(\int e^{t_\lambda} \cdot h\right),$$

where  $\psi_\lambda$  is the normalization constant.

The second aggregation method concerns setups where  $f$  and  $(f_k, 1 \leq k \leq N)$  are non-negative with arbitrary total mass, e.g. the more realistic spectral density model with unknown variance. In this case, we proceed by aggregating  $\log(f_k/h) =: g_k$  into a convex combination

$$g_\lambda = \sum_{k=1}^N \lambda_k g_k, \quad \text{with } \lambda \text{ in the simplex } \Lambda^+,$$

and leave the resulting function unnormalized:

$$f_\lambda^S(x) = e^{g_\lambda(x)} \cdot h(x).$$

It is easy to see that the normalized  $f_\lambda^S / \int f_\lambda^S$  is actually the same as  $f_\lambda^D$ . However, the criteria we estimate and penalize in order to define the optimal coefficients  $\hat{\lambda}_*^D$  and  $\hat{\lambda}_*^S$  are different and the two aggregation methods lead to two different estimators.

With previous notation, we check that, for all  $f$  in  $\mathcal{G}$ :

$$\int f \leq e^{\|\log(f/h)\|_\infty}, \quad |\psi| \leq \|\log(f/h)\|_\infty \quad \text{and} \quad \|t\|_\infty \leq 2 \|\log(f/h)\|_\infty. \quad (7)$$

*Remark 2.1.* At this point we notice that the assumption that  $f$  belongs to the class  $\mathcal{G}$  is quite restrictive. Indeed, assuming e.g. that  $h$  is the indicator function on  $[0,1]$  implies that there exist two constants  $0 < c \leq C$  such that  $c \leq f(x) \leq C$  for all  $x$  in  $[0,1]$ . However, if  $f$  tends to 0 in the probability density model, [32] suggested data augmentation to raise the function above 0. That means mixing the data with an auxiliary sample drawn from the distribution  $\varphi$  and use  $(f_k + \varphi)/2$  instead of  $f_k$ . We may choose  $\varphi$  uniformly distributed over a compact set that covers all data.



Moreover, the functions  $f$  and  $f_k$ ,  $1 \leq k \leq N$  may be unbounded, or tend to 0, through the reference probability density  $h$ . Consider e.g.  $h(x) = (1 - \alpha)/x^\alpha$  for  $x$  in  $[0,1]$  and some  $\alpha$  between 0 and 1, or an exponential distribution on non-negative real numbers, or a Gaussian distribution on the real line. These cases are not covered by the  $L^2$  aggregation procedure.

We fix some additional notation. We denote by  $I_n$  an integrable estimator of the function  $f$  measurable with respect to the sample  $X = (X_1, \dots, X_n)$ . The estimator  $I_n$  may be a biased estimator of  $f$  as is the periodogram in the spectral density model for example. We note  $\bar{f}_n$  the expected value of  $I_n$ :

$$\bar{f}_n = \mathbb{E}_f[I_n].$$

For a measurable function  $p$  on  $\mathbb{R}^d$  and a measure  $Q$  on  $\mathbb{R}^d$  (resp. a measurable function  $q$  on  $\mathbb{R}^d$ ), we write  $\langle p, Q \rangle = \int p(x)Q(dx)$  (resp.  $\langle p, q \rangle = \int pq$ ) when the integral is well defined. We shall consider the  $L^2(h)$  norm given by  $\|p\|_{L^2(h)} = (\int p^2 h)^{1/2}$ .

We describe now the implementation using data of our two aggregation methods of the estimators  $(f_k, 1 \leq k \leq N)$  and first results. For each of the two previously defined aggregation methods, we see how the Kullback-Leibler discrepancy between the estimated  $f$  and the convex combination  $f_\lambda$  (which is either  $f_\lambda^D$  or  $f_\lambda^S$ ) can be reduced to a linear functional of  $f$ . Such linear functionals can be estimated with fast rates in a biased or, sometimes, an unbiased way. We penalize this estimator to get the criterion  $H_n(\lambda)$ , a concave function of  $\lambda$  to be maximized over the simplex  $\Lambda^+$ . The resulting argument of this optimization,  $\hat{\lambda}_*$ , provides our aggregated procedure  $\hat{f}_* = f_{\hat{\lambda}_*}$  that is optimal in deviation.

Proofs are structured in two main steps, for each aggregation method. First,  $f$  as well as  $f_k$ ,  $1 \leq k \leq N$  belong to  $\mathcal{G}$  and Lemmas 2.3 and 2.5 show that the penalized criterion  $H_n(\lambda)$  has a unique maximizer provided that  $\{t_k, 1 \leq k \leq N\}$ , respectively  $\{g_k, 1 \leq k \leq N\}$ , are linearly independent. Moreover  $H_n(\lambda)$  is proved strongly concave around its maximum.

Next, we assume additionally that  $\|\log(f_k/h)\|_\infty$  are uniformly bounded by some constant  $K > 0$  for all  $1 \leq k \leq N$ , see Remark 2.1 on such constraint. We state in the following Propositions 2.4 and 2.6 oracle inequalities in deviation under quite general form. Indeed, the remainder term in these inequalities is decomposed for each method into a bias term for estimating the linear functional  $\langle \cdot, f \rangle$  of  $f$  and the maximum of  $N$  stochastic terms which are centered at second order. No additional smoothness assumptions are required here.

### 2.1. Non-negative functions with given total mass

In this Section, we shall consider non-negative functions with mass 1, but what follows can readily be adapted to functions with any given total mass, such as spectral density function with given variance.

We want to estimate  $f$  based on the estimators  $f_k \in \mathcal{G}$  for  $1 \leq k \leq N$  which we assume to be non-negative with mass 1. Recall the representation (6) of  $f$

and  $f_k$ . For  $\lambda \in \Lambda^+$  defined by (3), recall that the aggregated estimator  $f_\lambda^D$  is given by the convex combination of  $(t_k, 1 \leq k \leq N)$ :

$$f_\lambda^D = \exp(t_\lambda - \psi_\lambda) h \quad \text{with} \quad t_\lambda = \sum_{k=1}^N \lambda_k t_k \quad \text{and} \quad \psi_\lambda = \log \left( \int e^{t_\lambda} h \right).$$

Notice that  $f_\lambda^D$  is such that  $\|t_\lambda\|_\infty \leq \max_{1 \leq k \leq N} \|t_k\|_\infty < +\infty$ , that is  $f_\lambda^D \in \mathcal{G}$ . The Kullback-Leibler divergence for the estimator  $f_\lambda^D$  of  $f$  is given by:

$$D(f \| f_\lambda^D) = \int \log(f/f_\lambda^D) f = \langle t - t_\lambda, f \rangle + \psi_\lambda - \psi. \tag{8}$$

Minimizing the Kullback-Leibler distance is thus equivalent to maximizing  $\lambda \mapsto \langle t_\lambda, f \rangle - \psi_\lambda$ . Notice that  $\langle t_\lambda, f \rangle$  is linear in  $\lambda$  and the function  $\lambda \mapsto \psi_\lambda$  is convex since  $\nabla^2 \psi_\lambda$  is the covariance matrix of the random vector  $(t_k(Y_\lambda), 1 \leq k \leq N)$  with  $Y_\lambda$  having probability density function  $f_\lambda^D$ . Let  $I_n$  be a non-negative estimator of  $f$  based on the sample  $X = (X_1, \dots, X_n)$ , which may be the empirical measure or a biased estimator with  $\bar{f}_n = E_f[I_n]$ . We estimate the scalar product  $\langle t_\lambda, f \rangle$  by  $\langle t_\lambda, I_n \rangle$ . To select the aggregation weights  $\lambda$ , we consider on  $\Lambda^+$  the penalized empirical criterion  $H_n^D(\lambda)$  given by:

$$H_n^D(\lambda) = \langle t_\lambda, I_n \rangle - \psi_\lambda - \frac{1}{2} \text{pen}^D(\lambda),$$

with penalty term:

$$\text{pen}^D(\lambda) = \sum_{k=1}^N \lambda_k D(f_\lambda^D \| f_k) = \sum_{k=1}^N \lambda_k \psi_k - \psi_\lambda.$$

*Remark 2.2.* The penalty term in the definition of  $H_n^D$  can be multiplied by any constant  $\theta \in (0, 1)$  instead of  $1/2$ . The choice of  $1/2$  is optimal in the sense that it ensures that the constant  $\exp(-6K)/4$  in (14) of Proposition 2.4 is maximal, giving the sharpest result.

The penalty term is always non-negative and finite. Notice that  $H_n^D$  simplifies to:

$$H_n^D(\lambda) = \sum_{k=1}^N \lambda_k \cdot \left( \langle t_k, I_n \rangle - \frac{1}{2} \psi_k \right) - \frac{1}{2} \psi_\lambda, \tag{9}$$

which is obviously concave function of  $\lambda$ , as  $\psi_\lambda$  is convex in  $\lambda$ .

Lemma 2.3 below asserts that the function  $H_n^D$ , defined by (9), admits a unique maximizer on  $\Lambda^+$  and that it is strictly concave around this maximizer.

**Lemma 2.3.** *Let  $f$  and  $(f_k, 1 \leq k \leq N)$  be non-negative functions with mass 1, elements of  $\mathcal{G}$  such that  $(t_k, 1 \leq k \leq N)$  are linearly independent. Then there exists a unique  $\hat{\lambda}_*^D \in \Lambda^+$  such that:*

$$\hat{\lambda}_*^D = \operatorname{argmax}_{\lambda \in \Lambda^+} H_n^D(\lambda). \tag{10}$$

Furthermore, for all  $\lambda \in \Lambda^+$ , we have:

$$H_n^D(\hat{\lambda}_*^D) - H_n^D(\lambda) \geq \frac{1}{2} D\left(f_{\hat{\lambda}_*^D}^D \| f_\lambda^D\right). \quad (11)$$

Using  $\hat{\lambda}_*^D$  defined in (10), we set:

$$\hat{f}_*^D = f_{\hat{\lambda}_*^D}^D, \quad \hat{t}_*^D = t_{\hat{\lambda}_*^D} \quad \text{and} \quad \hat{\psi}_*^D = \psi_{\hat{\lambda}_*^D}. \quad (12)$$

We show that the convex aggregate estimator  $\hat{f}_*^D$  verifies almost surely the following non-asymptotic inequality with a bias and a variance term.

**Proposition 2.4.** *Let  $K > 0$ . Let  $f$  and  $(f_k, 1 \leq k \leq N)$  be non-negative functions with mass 1, elements of  $\mathcal{G}$  such that  $(t_k, 1 \leq k \leq N)$  are linearly independent and  $\max_{1 \leq k \leq N} \|t_k\|_\infty \leq K$ . Let  $X = (X_1, \dots, X_n)$  be a sample from the model  $P_f$ . Then the following inequality holds:*

$$D\left(f \| \hat{f}_*^D\right) - D\left(f \| f_{k^*}\right) \leq B_n\left(\hat{t}_*^D - t_{k^*}\right) + \max_{1 \leq k \leq N} V_n^D(e_k),$$

with the functional  $B_n$  given by, for  $\ell \in L^\infty(\mathbb{R})$ :

$$B_n(\ell) = \langle \ell, \bar{f}_n - f \rangle. \quad (13)$$

and the function  $V_n^D : \Lambda^+ \rightarrow \mathbb{R}$  given by:

$$V_n^D(\lambda) = \langle I_n - \bar{f}_n, t_\lambda - t_{k^*} \rangle - \frac{e^{-6K}}{4} \sum_{k=1}^N \lambda_k \|t_k - t_{k^*}\|_{L^2(h)}^2. \quad (14)$$

The bias term  $B_n$  is new in the aggregation results. We stress the fact that it is a bias for estimating the linear functional  $\langle \ell, f \rangle$  appearing in the definition of the risk and not a bias for estimating the function  $f$ . The bias for linear functionals can be made  $O(1/n)$ , and thus negligible in front of the variance, under very mild assumptions on the function  $f$ .

Obviously, when  $f$  is a pdf, we plug the empirical measure in  $\langle \ell, f \rangle$  and the resulting estimator  $n^{-1} \sum_{i=1}^n \ell(X_i)$  is unbiased.

## 2.2. Non-negative functions

In this Section, we shall consider non-negative functions. We want to estimate a function  $f \in \mathcal{G}$  based on the estimators  $f_k \in \mathcal{G}$  for  $1 \leq k \leq N$ . Since most of the proofs in this Section are similar to those in Section 2.1, we only give them when there is a substantial new element. Recall the representation (6) of  $f$  and  $f_k$ . For  $\lambda \in \Lambda^+$  defined by (3), we consider the aggregate estimator  $f_\lambda^S$  given by the convex aggregation of  $\log(f_k/h) =: g_k$ , for  $1 \leq k \leq N$ :

$$f_\lambda^S = e^{g_\lambda} \cdot h \quad \text{with} \quad g_\lambda = \sum_{k=1}^N \lambda_k g_k. \quad (15)$$

Notice that  $\|\log(f_\lambda^S/h)\|_\infty \leq \max_{1 \leq k \leq N} \|g_k\|_\infty < +\infty$ , that is  $f_\lambda^S \in \mathcal{G}$ . The Kullback-Leibler distance for the estimator  $f_\lambda^S$  of  $f$  is given by:

$$D(f \| f_\lambda^S) = \int f \log(f/f_\lambda^S) - \int f + \int f_\lambda^S = \langle g - g_\lambda, f \rangle - \int f + \int f_\lambda^S. \tag{16}$$

By our assumptions  $f$  and  $f_\lambda^S$  belong to  $\mathcal{G}$ , thus  $D(f \| f_\lambda^S) < \infty$  for all  $\lambda \in \Lambda^+$ .

Minimization of the Kullback-Leibler distance given in (16) is therefore equivalent to maximizing  $\lambda \mapsto \langle g_\lambda, f \rangle - \int f_\lambda^S$ . Notice that  $\langle g_\lambda, f \rangle$  is linear in  $\lambda$  and the function  $\lambda \mapsto \int f_\lambda^S$  is convex, since the Hessian matrix  $\nabla^2 \int f_\lambda^S$  is given by:  $[\nabla^2 \int f_\lambda^S]_{i,j} = \int g_i g_j f_\lambda^S$ , which is positive-semidefinite. As  $I_n$  is a non-negative estimator of  $f$  based on the sample  $X = (X_1, \dots, X_n)$ , we estimate the scalar product  $\langle g_\lambda, f \rangle$  by  $\langle g_\lambda, I_n \rangle$ . Here we select the aggregation weights  $\lambda$  based on the penalized empirical criterion  $H_n^S(\lambda)$  given by:

$$H_n^S(\lambda) = \langle g_\lambda, I_n \rangle - \int f_\lambda^S - \frac{1}{2} \text{pen}^S(\lambda),$$

with the penalty term:

$$\text{pen}^S(\lambda) = \sum_{k=1}^N \lambda_k D(f_\lambda^S \| f_k) = \sum_{k=1}^N \lambda_k \int f_k - \int f_\lambda^S.$$

The choice of the factor 1/2 for the penalty is justified by arguments similar to those given in Remark 2.2. The penalty term is always non-negative and finite. Notice that  $H_n^S$  simplifies to:

$$H_n^S(\lambda) = \sum_{k=1}^N \lambda_k \left( \langle g_k, I_n \rangle - \frac{1}{2} \int f_k \right) - \frac{1}{2} \int f_\lambda^S. \tag{17}$$

Thus, the new criterion  $H_n^S$  is also the sum of a linear term in  $\lambda$  and of a concave term  $-(1/2) \int f_\lambda^S$ . The proofs are very similar to those in the previous section, for this reason.

Lemma 2.5 below asserts that the function  $H_n^S$  admits a unique maximizer on  $\Lambda^+$  and that it is strictly concave around this maximizer.

**Lemma 2.5.** *Let  $f$  and  $(f_k, 1 \leq k \leq N)$  be elements of  $\mathcal{G}$  such that  $(g_k, 1 \leq k \leq N)$  are linearly independent. Let  $H_n^S$  be defined by (17). Then there exists a unique  $\hat{\lambda}_*^S \in \Lambda^+$  such that:*

$$\hat{\lambda}_*^S = \operatorname{argmax}_{\lambda \in \Lambda^+} H_n^S(\lambda). \tag{18}$$

Furthermore, for all  $\lambda \in \Lambda^+$ , we have:

$$H_n^S(\hat{\lambda}_*^S) - H_n^S(\lambda) \geq \frac{1}{2} D(f_{\hat{\lambda}_*^S}^S \| f_\lambda^S). \tag{19}$$

Using  $\hat{\lambda}_*^S$  defined in (18), we set:

$$\hat{f}_*^S = f_{\hat{\lambda}_*^S}^S \quad \text{and} \quad \hat{g}_*^S = g_{\hat{\lambda}_*^S}. \quad (20)$$

We show that the convex aggregate estimator  $\hat{f}_*^S$  verifies almost surely the following non-asymptotic inequality with a bias and a variance term.

**Proposition 2.6.** *Let  $K > 0$ . Let  $f$  and  $(f_k, 1 \leq k \leq N)$  be elements of  $\mathcal{G}$  such that  $(g_k, 1 \leq k \leq N)$  are linearly independent and  $\max_{1 \leq k \leq N} \|g_k\|_\infty \leq K$ . Let  $X = (X_1, \dots, X_n)$  be a sample from the model  $\mathbb{P}_f$ . Then the following inequality holds:*

$$D(f \parallel \hat{f}_*^S) - D(f \parallel f_{k^*}) \leq B_n(\hat{g}_*^S - g_{k^*}) + \max_{1 \leq k \leq N} V_n^S(e_k),$$

with the functional  $B_n$  given by (13), and the function  $V_n^S : \Lambda^+ \rightarrow \mathbb{R}$  given by:

$$V_n^S(\lambda) = \langle g_\lambda - g_{k^*}, I_n - \bar{f}_n \rangle - \frac{e^{-3K}}{4} \sum_{k=1}^N \lambda_k \|g_k - g_{k^*}\|_{L^2(h)}^2.$$

### 3. Applications and numerical implementation

In this section we apply the methods established in Section 2.1 and 2.2 to the problems of density estimation and spectral density estimation, respectively. By construction, the aggregate  $f_\lambda^D$  of Section 2.1 is more adapted for the density estimation problem as it produces a proper density function. In this model, we choose to use the unbiased empirical estimator of the linear functional of  $f$ . Next, Theorem 3.1 makes the remainder term explicit uniformly over pdf's  $f$  such that  $\|\log(f/h)\|_\infty \leq L$ , for some  $L > 0$ . A short numerical study might inspire the reader for future developments.

For the spectral density estimation problem, the aggregate  $f_\lambda^S$  will provide the optimal results. It is more convenient in this case to plug-in the periodogram  $I_n$  in order to get a biased estimator  $\langle \ell, I_n \rangle$  of  $\langle \ell, f \rangle$ . In order to control the bias and make it negligible with respect to the stochastic term, we assume that  $\log(f/h)$  belongs to a Sobolev ellipsoid of smoothness  $r$ , for some  $r > 1/2$  arbitrarily close to  $1/2$ . Note that  $r$  is not needed anywhere in the procedure, it only appears in the constants of the remainder term and does not affect the rate  $\log N/n$ . We show that our smoothness assumption implies that  $\sum_{j \geq 1} j^{2r} \text{Cov}(X_\cdot, X_{\cdot+j}) \leq L$  which is slightly more restrictive than the usual assumption in spectral density models that  $\sum_{j \geq 1} j^2 \text{Cov}(X_\cdot, X_{\cdot+j}) \leq L$ , when  $(X_j, j \in \mathbb{Z})$  is a stationary Gaussian sequence with spectral density function  $f$ .

#### 3.1. Probability density model

Recall that the model  $\{\mathbb{P}_f, f \in \mathcal{F}^D(L)\}$  corresponds to i.i.d. random sampling from a probability density  $f \in \mathcal{F}^D(L)$ , that is the random variable  $X =$

$(X_1, \dots, X_n)$  has density  $f^{\otimes n}(x) = \prod_{i=1}^n f(x_i)$ , with  $x = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ . We consider the following subset of probability density functions, for  $L > 0$ :

$$\mathcal{F}^D(L) = \{f \in \mathcal{G}; \|t_f\|_\infty \leq L \text{ and } \int f = 1\}.$$

Recall that for any fixed  $f$  in  $\mathcal{G}$ , the function  $t_f := \psi + \log(f/h)$  has the sup-norm bounded as follows  $\|t_f\|_\infty \leq 2\|\log(f/h)\|_\infty$  which is finite by assumption. We refer to Remark 2.1 for a discussion on this assumption. We want to establish here an oracle inequality uniformly over  $f$  which is the reason why we assume that  $\|t_f\|_\infty$  is uniformly bounded. This bound does not appear anywhere in the algorithm, but only in the remainder term that is a price for aggregation of estimators.

The aggregation procedure involves estimation of linear functionals  $\langle \ell, f \rangle$ , for functions  $\ell$  in  $L^1(h)$ . We estimate the probability measure  $f(x)dx$  by the empirical measure  $I_n$ , giving

$$\langle \ell, I_n \rangle = \frac{1}{n} \sum_{i=1}^n \ell(X_i) \quad (21)$$

which is an unbiased estimator of  $\langle \ell, f \rangle$ .

In the following Theorem, we give a sharp non-asymptotic oracle inequality in probability for the aggregation procedure  $\hat{f}_*^D$  with a remainder term of order  $\log(N)/n$ . We prove in Section 4.1 the lower bound giving that this remainder term is optimal.

**Theorem 3.1.** *Let  $L, K > 0$ . Let  $f \in \mathcal{F}^D(L)$  and  $(f_k, 1 \leq k \leq N)$  be elements of  $\mathcal{F}^D(K)$  such that  $(t_k, 1 \leq k \leq N)$  are linearly independent. Let  $X = (X_1, \dots, X_n)$  be an i.i.d. sample from  $f$  and  $I_n$  be defined in (21). Let  $\hat{f}_*^D$  be given by (12). Then for any  $x > 0$  we have with probability greater than  $1 - \exp(-x)$ :*

$$D(f \| \hat{f}_*^D) - D(f \| f_{k^*}) \leq \frac{\beta(\log(N) + x)}{n},$$

with  $\beta = 2 \exp(6K + 2L) + 4K/3$ .

*Remark 3.2.* We can also use the aggregation method of Section 2.2 and consider the normalized estimator  $\tilde{f}_*^S = \hat{f}_*^S / \int \hat{f}_*^S = f_{\hat{\lambda}_*^S}^D$ , which is a proper density function. Notice that the optimal weights  $\hat{\lambda}_*^D$  (which defines  $\hat{f}_*^D$ ) and  $\hat{\lambda}_*^S$  (which defines  $\tilde{f}_*^S$ ) maximize different criteria. Indeed, according to (18) the vector  $\hat{\lambda}_*^S$  maximizes:

$$H_n^S(\lambda) = \langle g_\lambda, I_n \rangle - \frac{1}{2} \int f_\lambda^S - \frac{1}{2} \sum_{k=1}^N \lambda_k \int f_k,$$

and according to (10) the vector  $\hat{\lambda}_*^D$  maximizes:

$$\begin{aligned}
H_n^D(\lambda) &= \langle t_\lambda, I_n \rangle - \frac{1}{2} \psi_\lambda - \frac{1}{2} \sum_{k=1}^N \lambda_k \psi_k \\
&= \langle g_\lambda, I_n \rangle - \frac{1}{2} \psi_\lambda + \frac{1}{2} \sum_{k=1}^N \lambda_k \psi_k \\
&= \langle g_\lambda, I_n \rangle - \frac{1}{2} \log \left( \int f_\lambda^S \right),
\end{aligned}$$

where we used the identity  $g_\lambda = t_\lambda - \sum_{k=1}^N \lambda_k \psi_k$  for the second equality and the equality  $\log(\int f_\lambda^S) = \log\left(\int e^{t_\lambda - \sum_{k=1}^N \lambda_k \psi_k} h\right) = \psi_\lambda - \sum_{k=1}^N \lambda_k \psi_k$  for the third. That shows that the resulting aggregated estimators are different.

### 3.2. Spectral density model

In this section we apply the convex aggregation scheme of Section 2.2 to spectral density estimation of stationary centered Gaussian sequences. Let  $h = 1/(2\pi)\mathbf{1}_{[-\pi, \pi]}$  be the reference density and  $(X_k)_{k \in \mathbb{Z}}$  be a stationary, centered Gaussian sequence with covariance function  $\gamma$  defined as, for  $j \in \mathbb{Z}$ :

$$\gamma_j = \text{Cov}(X_k, X_{k+j}).$$

Notice that  $\gamma_{-j} = \gamma_j$ . Then the joint distribution of  $X = (X_1, \dots, X_n)$  is a multivariate, centered Gaussian distribution with covariance matrix  $\Sigma_n \in \mathbb{R}^{n \times n}$  given by  $[\Sigma_n]_{i,j} = \gamma_{i-j}$  for  $1 \leq i, j \leq n$ . Notice the sequence  $(\gamma_j)_{j \in \mathbb{Z}}$  is semi-definite positive.

We make the following standard assumption on the covariance function  $\gamma$ :

$$\sum_{j=0}^{\infty} |\gamma_j| < +\infty. \quad (22)$$

The spectral density function  $f$  associated to the process is the even function defined on  $[-\pi, \pi]$  whose Fourier coefficients are  $\gamma_j$ :

$$f(x) = \sum_{j \in \mathbb{Z}} \frac{\gamma_j}{2\pi} e^{ijx} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{\infty} \gamma_j \cos(jx).$$

Condition (22) ensures that the spectral density is well-defined, continuous and bounded. It is also even and non-negative as  $(\gamma_j)_{j \in \mathbb{Z}}$  is semi-definite positive. The function  $f$  completely characterizes the model as:

$$\gamma_j = \int_{-\pi}^{\pi} f(x) e^{ijx} dx = \int_{-\pi}^{\pi} f(x) \cos(jx) dx \quad \text{for } j \in \mathbb{Z}. \quad (23)$$

For  $\ell \in L^1(h)$ , we define the corresponding Toeplitz matrix  $T_n(\ell)$  of size  $n \times n$  by:

$$[T_n(\ell)]_{j,k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ell(x) e^{i(j-k)x} dx.$$

Notice that  $T_n(2\pi f) = \Sigma_n$ . Some properties of the Toeplitz matrix  $T_n(\ell)$  are collected in Section A.1.

We choose the following estimator of  $f$ , for  $x \in [-\pi, \pi]$ :

$$I_n(x) = \sum_{|j| < n} \frac{\hat{\gamma}_j}{2\pi} e^{jx} = \frac{\hat{\gamma}_0}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{n-1} \hat{\gamma}_j \cos(jx), \tag{24}$$

with  $(\hat{\gamma}_j, 0 \leq j \leq n - 1)$  the empirical estimates of the covariances  $(\gamma_j, 0 \leq j \leq n - 1)$ :

$$\hat{\gamma}_j = \frac{1}{n} \sum_{i=1}^{n-j} X_i X_{i+j}. \tag{25}$$

The function  $I_n$  is a biased estimator, where the bias is due to two different sources: truncation of the infinite sum up to  $n$ , and renormalization in (25) by  $n$  instead of  $n - j$  (but it is asymptotically unbiased when  $n$  goes to infinity as condition (22) is satisfied). The expected value  $\bar{f}_n$  of  $I_n$  is given by:

$$\bar{f}_n(x) = \sum_{|j| < n} \left(1 - \frac{|j|}{n}\right) \frac{\gamma_j}{2\pi} e^{jx} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{n-1} \frac{(n-j)}{n} \gamma_j \cos(jx).$$

In order to be able to apply Proposition 2.6, we assume that  $f$  and the estimators  $f_1, \dots, f_N$  of  $f$  belongs to  $\mathcal{G}$  (they are in particular positive and bounded) and are even functions. In particular the estimators  $f_1, \dots, f_N$  and the convex aggregate estimator  $\hat{f}_*^S$  defined in (20) are proper spectral densities of stationary Gaussian sequences.

*Remark 3.3.* By choosing  $h = 1/(2\pi)\mathbf{1}_{[-\pi, \pi]}$ , we restrict our attention to spectral densities that are bounded away from  $+\infty$  and 0, see [24] and [6] for the characterization of such spectral densities. Note that we can apply the aggregation procedure to non even functions  $f_k, 1 \leq k \leq N$ , but the resulting estimator would not be a proper spectral density in that case.

To prove a sharp oracle inequality for the spectral density estimation, since  $I_n$  is a biased estimator of  $f$ , we shall assume some regularity on the functions  $f$  and  $f_1, \dots, f_N$  in order to be able to control the bias term. More precisely those conditions will be Sobolev conditions on their logarithm, that is on the functions  $\log(f/h)$  and  $\log(f_1/h), \dots, \log(f_N/h)$ .

For  $\ell \in L^2(h)$ , the corresponding Fourier coefficients are defined for  $k \in \mathbb{Z}$  by  $a_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikx} \ell(x) dx$ . From the Fourier series theory, we deduce that  $\sum_{k \in \mathbb{Z}} |a_k|^2 = \|\ell\|_{L^2(h)}^2$  and a.e.  $\ell(x) = \sum_{k \in \mathbb{Z}} a_k e^{ikx}$ . If furthermore  $\sum_{k \in \mathbb{Z}} |a_k|$  is finite, then  $\ell$  is continuous,  $\ell(x) = \sum_{k \in \mathbb{Z}} a_k e^{ikx}$  for  $x \in [-\pi, \pi]$  and  $\|\ell\|_{\infty} \leq \sum_{k \in \mathbb{Z}} |a_k|$ .

For  $r > 0$ , we define the Sobolev norm  $\|\ell\|_{2,r}$  of  $\ell$  as:

$$\|\ell\|_{2,r}^2 = \|\ell\|_{L^2(h)}^2 + \{\ell\}_{2,r}^2 \quad \text{with} \quad \{\ell\}_{2,r}^2 = \sum_{k \in \mathbb{Z}} |k|^{2r} |a_k|^2.$$



The corresponding Sobolev space is defined by:

$$W_r = \{\ell \in L^2(h); \|\ell\|_{2,r} < +\infty\}.$$

For  $r > 1/2$ , we can bound the supremum norm of  $\ell$  by its Sobolev norm:

$$\|\ell\|_\infty \leq \sum_{k \in \mathbb{Z}} |a_k| \leq C_r \|\ell\|_{2,r} \leq C_r \|\ell\|_{2,r}, \quad (26)$$

where we used Cauchy-Schwarz inequality for the second inequality with

$$C_r^2 = \sum_{k \in \mathbb{Z}^*} |k|^{-2r} < +\infty. \quad (27)$$

The proof of the following Lemma seems to be part of the folklore, but since we didn't find a proper reference, we give it in Section A.2.

**Lemma 3.4.** *Let  $r > 1/2$ ,  $K > 0$ . There exists a finite constant  $C(r, K)$  such that for any  $g \in W_r$  with  $\|g\|_{2,r} \leq K$ , then we have  $\|\exp(g)\|_{2,r} \leq C(r, K)$ .*

For  $r > 1/2$ , we consider the following subset of functions:

$$\mathcal{F}_r^S(L) = \{f \in \mathcal{G} : \|g_f\|_{2,r} \leq L/C_r \text{ and } g_f \text{ even}\}, \quad \text{with } g_f = \log(f/h). \quad (28)$$

For  $f \in \mathcal{F}_r^S(L)$ , we deduce from (26) that  $g_f$  is continuous (and bounded by  $L$ ). This implies that  $f$  is a positive, continuous, even function and thus a proper spectral density. Notice that  $2\pi \|f\|_\infty \leq \exp(L)$ . We deduce from (23) that  $\gamma_k = \int_{-\pi}^{\pi} e^{-ikx} f(x) dx$  and thus:

$$\|f\|_{2,r}^2 = \frac{\gamma_0^2}{4\pi^2} + \frac{1}{2\pi^2} \sum_{k=1}^{\infty} (1 + k^{2r}) \gamma_k^2.$$

Thus Lemma 3.4 and (26) imply also that the covariance function associated to  $f \in \mathcal{F}_r^S(L)$  satisfies (22). We also get that  $\sum_{j=1}^{\infty} j \gamma_j^2 < +\infty$ , which is a standard assumption for spectral density estimation.

The following Theorem is the main result of this section.

**Theorem 3.5.** *Let  $r > 1/2$ ,  $K, L > 0$ . Let  $f \in \mathcal{F}_r^S(L)$  and  $(f_k, 1 \leq k \leq N)$  be elements of  $\mathcal{F}_r^S(K)$  such that  $(g_k, 1 \leq k \leq N)$  are linearly independent. Let  $X = (X_1, \dots, X_n)$  be a sample of a stationary centered Gaussian sequence with spectral density  $f$  and  $I_n$  be defined in (24). Let  $\hat{f}_*^S$  be given by (15). Then for any  $x > 0$ , we have with probability higher than  $1 - \exp(-x)$ :*

$$D(f \| \hat{f}_*^S) - D(f \| f_{k^*}) \leq \frac{\beta(\log(N) + x)}{n} + \frac{\alpha}{n},$$

with  $\beta = 4(K e^L + e^{2L+3K})$  and  $\alpha = 4KC(r, L)/C_r$ .

*Remark 3.6.* When the value of  $\gamma_0$  is given, we shall use the aggregation method of Section 2.1 after normalizing the estimators  $f_k$ ,  $1 \leq k \leq N$  by dividing  $f_k$  by  $\int f_k$ . The final estimator of  $f$  would take the form  $\hat{f}_{\hat{\lambda}_*^D}^D = \gamma_0 f_{\hat{\lambda}_*^D}^D$  and verifies a similar sharp oracle inequality as  $\hat{f}_*^S$  (that is without the term  $\alpha/n$  of Theorem 3.5). When the value of  $\gamma_0$  is unknown, it could be estimated empirically by  $\hat{\gamma}_0 = \frac{1}{n} \sum_{i=1}^n X_i^2$ . Then we could use  $\hat{\gamma}_0 f_{\hat{\lambda}_*^D}^D$  to estimate  $f$ . However the empirical estimation of  $\gamma_0$  introduces an error term of order  $1/\sqrt{n}$ , which leads to a suboptimal remainder term for this aggregation method.

### 3.3. Numerical implementation

In this section we apply the aggregation method  $f_\lambda^D$  established in Section 2.1 to the problem of density estimation. We compare the performance of our aggregation scheme to the progressive mixture method introduced by [10] and [32], which proves to be a strong competitor.

To the best of our knowledge, there is no progressive mixture rule for spectral density estimation.

We consider two examples of probability densities to estimate: a shifted and truncated version of the Claw distribution and of the Smooth Comb distribution used in [26]. A random variable is distributed according to the Claw distribution if its density  $f_{Claw}$  is given by:

$$f_{Claw}(x) = \alpha \left( \frac{1}{2} \varphi(2x - 1) + \sum_{k=0}^4 0.1 \cdot \varphi \left( \frac{2x - 0.5k}{0.1} \right) \right) \mathbf{1}_{[0,1]}(x),$$

where  $\varphi$  is the density function of a standard normal random variable, and  $\alpha$  a normalizing constant. The Smooth Comb distribution has density function  $f_{SmComb}$  defined as, for  $x \in \mathbb{R}$ :

$$f_{SmComb}(x) = a \sum_{k=0}^5 \varphi \left( \frac{6x - 3 - (65 + 96 \cdot 2^{-k})/21}{32 \cdot 2^{-k}/63} \right) \mathbf{1}_{[0,1]}(x),$$

with normalizing constant  $a$ .

Given a sample of the underlying distribution, we use the first part of it to create kernel density estimators with different bandwidths, then use the second part to perform the aggregation. We consider 5 different estimators with the normal kernel and bandwidths  $h = 0.005, 0.0075, 0.01, 0.02, 0.05$ . We create 100 samples of size 1000, and then use  $n = 50, 100, 200, 500, 1000$  of the available size to estimate the true density, allocating 80% of the sample to creating the kernel estimators and 20% for the aggregation part.

The results of the estimations can be found in Figure 1. We can observe that for small sample sizes the convex aggregation method performs better in terms of risk and dispersion, but as the sample size increases, the progressive mixture method yields better results.

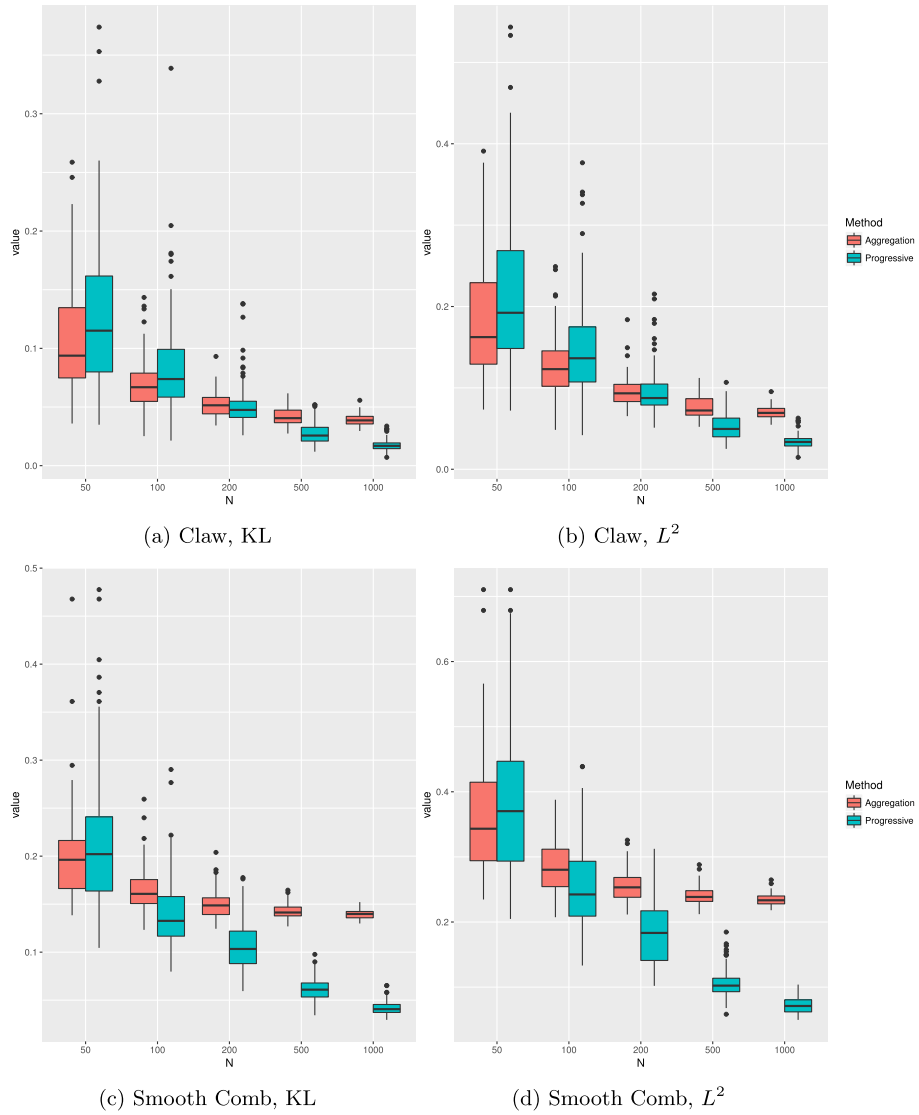


FIG 1. Boxplots of the Kullback-Leibler (KL) and  $L^2$  distance of the estimators for the Claw and the Smooth Comb density with different sample sizes.

Further numerical studies may include a hybrid procedure that would benefit from our aggregated procedure for small samples and from the progressive mixture procedure for large enough samples. The latter might also be introduced for the spectral density model, which characterizes the covariance of stationary series and is of great interest for many applications.

### 4. Lower bounds

In this section we show that the aggregation procedure given in Section 2 is optimal by giving a lower bound corresponding to the upper bound of Theorem 3.1 and 3.5 for the estimation of the probability density function as well as for the spectral density.

#### 4.1. Probability density estimation

In this section we suppose that the reference density is the uniform distribution on  $[0, 1]^d$ :  $h = \mathbf{1}_{[0,1]^d}$ .

*Remark 4.1.* If the reference density is not the uniform distribution on  $[0, 1]^d$ , then we can apply the Rosenblatt transformation, see [27], to reduce the problem to this latter case. More precisely, according to [27], if the random variable  $Z$  has probability density  $h$ , then there exists two maps  $T$  and  $T^{-1}$  such that  $U = T(Z)$  is uniform on  $[0, 1]^d$  and a.s.  $Z = T^{-1}(U)$ . Then if the random variable  $X$  has density  $f = \exp(g)h$ , we deduce that  $T(X)$  has density  $f^T = \exp(g \circ T^{-1})\mathbf{1}_{[0,1]^d}$ . Furthermore, if  $f_1$  and  $f_2$  are two densities (with respect to the reference density  $h$ ), then we have  $D(f_1 \| f_2) = D(f_1^T \| f_2^T)$ .

We give the main result of this Section. Let  $\mathbb{P}_f$  denote the probability measure when  $X_1, \dots, X_n$  are i.i.d. random variable with density  $f$ .

**Proposition 4.2.** *Let  $N \geq 2$ ,  $L > 0$ . Then there exist  $N$  probability densities  $(f_k, 1 \leq k \leq N)$ , with  $f_k \in \mathcal{F}^D(L)$  such that for all  $n \geq 1$ ,  $x \in \mathbb{R}^+$  satisfying:*

$$\frac{\log(N) + x}{n} < 3(1 - e^{-L})^2, \tag{29}$$

we have:

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}^D(L)} \mathbb{P}_f \left( D(f \| \hat{f}_n) - \min_{1 \leq k \leq N} D(f \| f_k) \geq \frac{\beta' (\log(N) + x)}{n} \right) \geq \frac{1}{24} e^{-x},$$

with the infimum taken over all estimators  $\hat{f}_n$  based on the sample  $X_1, \dots, X_n$ , and  $\beta' = 2^{-17/2}/3$ .

#### 4.2. Spectral density estimation

In this section we give a lower bound for aggregation of spectral density estimators. Let  $\mathbb{P}_f$  denote the probability measure when  $(X_n)_{n \in \mathbb{Z}}$  is a centered Gaussian sequence with spectral density  $f$ . Recall the set of positive even function  $\mathcal{F}_r^S(L) \subset \mathcal{G}$  defined by (28) for  $r \in \mathbb{R}$ .

**Proposition 4.3.** *Let  $N \geq 2$ ,  $r > 1/2$ ,  $L > 0$ . There exist a constant  $C(r, L)$  free of  $N$  and  $N$  spectral densities  $(f_k, 1 \leq k \leq N)$  belonging to  $\mathcal{F}_r^S(L)$  such that for all  $n \geq 1$ ,  $x \in \mathbb{R}^+$  satisfying:*

$$\frac{\log(N) + x}{n} < \frac{C(r, L)}{\log(N)^{2r}} \tag{30}$$

we have:

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}_r^S(L)} \mathbb{P}_f \left( D(f \parallel \hat{f}_n) - \min_{1 \leq k \leq N} D(f \parallel f_k) \geq \frac{\beta' (\log(N) + x)}{n} \right) \geq \frac{1}{24} e^{-x}, \tag{31}$$

with the infimum taken over all estimators  $\hat{f}_n$  based on the sample sequence  $X = (X_1, \dots, X_n)$ , and  $\beta' = 8^{-5/2}/3$ .

## 5. Proofs

### 5.1. Proofs of results in Section 2

*Proof of Lemma 2.3.* Consider the form (9) of  $H_n^D(\lambda)$ . Obviously, the function

$$\lambda \mapsto \sum_{k=1}^N \lambda_k (\langle t_k, I_n \rangle - \frac{1}{2} \psi_k)$$

is linear in  $\lambda$  and that  $\lambda \mapsto \psi_\lambda$  is convex. Notice that  $\nabla \psi_\lambda = (\langle t_k, f_\lambda^D \rangle, 1 \leq k \leq N)$ . This implies that for all  $\lambda, \lambda' \in \Lambda^+$ :

$$\begin{aligned} & (\lambda - \lambda') \cdot \nabla \psi_{\lambda'} + D(f_{\lambda'}^D \parallel f_\lambda^D) \\ &= \sum_{k=1}^N (\lambda_k - \lambda'_k) \langle t_k, f_{\lambda'}^D \rangle + \langle t_{\lambda'} - t_\lambda, f_{\lambda'}^D \rangle + \psi_\lambda - \psi_{\lambda'} \\ &= \psi_\lambda - \psi_{\lambda'}. \end{aligned} \tag{32}$$

Since  $\psi_\lambda$  is convex and differentiable, we deduce from (9) that  $H_n^D$  is concave and differentiable. We also have by the linearity of  $L_n^D$  and (32) that for all  $\lambda, \lambda' \in \Lambda^+$ :

$$H_n^D(\lambda) - H_n^D(\lambda') = (\lambda - \lambda') \cdot \nabla H_n^D(\lambda') - \frac{1}{2} D(f_{\lambda'}^D \parallel f_\lambda^D). \tag{33}$$

The concave function  $H_n^D$  on a compact set attains its maximum at some points  $\Lambda^* \subset \Lambda^+$ . For  $\hat{\lambda}_* \in \Lambda^*$ , we have for all  $\lambda \in \Lambda^+$ :

$$(\lambda - \hat{\lambda}_*) \cdot \nabla H_n^D(\hat{\lambda}_*) \leq 0, \tag{34}$$

see for example Equation 4.21 of [5]. Using (33) with  $\lambda' = \hat{\lambda}_*$  and (34), we get (11). Let  $\hat{\lambda}_*^1$  and  $\hat{\lambda}_*^2$  be elements of  $\Lambda^*$ . Then by (11), we have:

$$0 = H_n^D(\hat{\lambda}_*^1) - H_n^D(\hat{\lambda}_*^2) \geq \frac{1}{2} D(f_{\hat{\lambda}_*^1}^D \parallel f_{\hat{\lambda}_*^2}^D),$$

which implies that a.e.  $f_{\hat{\lambda}_*^1}^D = f_{\hat{\lambda}_*^2}^D$ . By the linear independence of  $(t_k, 1 \leq k \leq N)$ , this gives  $\hat{\lambda}_*^1 = \hat{\lambda}_*^2$ , giving the uniqueness of the maximizer.  $\square$

*Proof of Proposition 2.4.* Using (8), we get:

$$D\left(f\|\hat{f}_*^D\right) - D\left(f\|f_{k^*}\right) = \langle t_{k^*} - \hat{t}_*^D, f \rangle + \hat{\psi}_*^D - \psi_{k^*}.$$

By the definition of  $k^*$ , together with  $\text{pen}^D(e_k) = 0$  for all  $1 \leq k \leq N$  and the strict concavity (11) of  $H_n^D$  at  $\hat{\lambda}_*^D$  with  $\lambda = e_{k^*}$ , we get:

$$\begin{aligned} D\left(f\|\hat{f}_*^D\right) - D\left(f\|f_{k^*}\right) &\leq \langle t_{k^*} - \hat{t}_*^D, f \rangle + \hat{\psi}_*^D - \psi_{k^*} \\ &\quad + H_n^D(\hat{\lambda}_*^D) - H_n^D(e_{k^*}) - \frac{1}{2} D\left(\hat{f}_*^D\|f_{k^*}\right) \\ &= \langle \hat{t}_*^D - t_{k^*}, I_n - f \rangle - \frac{1}{2} D\left(\hat{f}_*^D\|f_{k^*}\right) - \frac{1}{2} \text{pen}^D(\hat{\lambda}_*^D) \\ &= B_n\left(\hat{t}_*^D - t_{k^*}\right) + A_n^D, \end{aligned}$$

with:

$$A_n^D = \langle \hat{t}_*^D - t_{k^*}, I_n - \bar{f}_n \rangle - \frac{1}{2} D\left(\hat{f}_*^D\|f_{k^*}\right) - \frac{1}{2} \sum_{k=1}^N \hat{\lambda}_{*,k}^D D\left(\hat{f}_*^D\|f_k\right). \quad (35)$$

We recall, see Lemma 1 of [2], that for any non-negative integrable functions  $p$  and  $q$  on  $\mathbb{R}^d$  satisfying  $\|\log(p/q)\|_\infty < +\infty$ , we have:

$$D(p\|q) \geq \frac{1}{2} e^{-\|\log(p/q)\|_\infty} \int p(\log(p/q))^2. \quad (36)$$

We have:

$$\begin{aligned} D\left(\hat{f}_*^D\|f_k\right) &\geq \frac{1}{2} e^{-\|\log(\hat{f}_*^D/f_k)\|_\infty} \int \hat{f}_*^D \left(\log(\hat{f}_*^D/f_k)\right)^2 \\ &\geq \frac{1}{2} e^{-4K - \|\hat{f}_*^D - \hat{\psi}_*^D\|_\infty} \int h \left(\log(\hat{f}_*^D/f_k)\right)^2 \\ &\geq \frac{1}{2} e^{-6K} \left(\|\hat{t}_*^D - t_k\|_{L^2(h)}^2 + (\hat{\psi}_*^D - \psi_k)^2\right) \\ &\geq \frac{1}{2} e^{-6K} \|\hat{t}_*^D - t_k\|_{L^2(h)}^2, \end{aligned}$$

where we used (36) for the first inequality, the fact that  $\psi \leq \|t_f\|_\infty$ , as well as  $\int t_f h = 0$ . By using this lower bound on  $D\left(\hat{f}_*^D\|f_k\right)$  to both terms on the right hand side of (35), we get:

$$\begin{aligned} A_n^D &\leq \langle \hat{t}_*^D - t_{k^*}, I_n - \bar{f}_n \rangle - \frac{e^{-6K}}{4} \|\hat{t}_*^D - t_{k^*}\|_{L^2(h)}^2 \\ &\quad - \frac{e^{-6K}}{4} \sum_{k=1}^N \hat{\lambda}_{*,k}^D \|\hat{t}_*^D - t_k\|_{L^2(h)}^2 \\ &= \langle \hat{t}_*^D - t_{k^*}, I_n - \bar{f}_n \rangle - \frac{e^{-6K}}{4} \sum_{k=1}^N \hat{\lambda}_{*,k}^D \|t_k - t_{k^*}\|_{L^2(h)}^2 \\ &= V_n^D(\hat{\lambda}_*^D), \end{aligned}$$

where the first equality is due to the following bias-variance decomposition equality which holds for all  $\ell \in L^2(h)$  and  $\lambda \in \Lambda^+$ :

$$\sum_{k=1}^N \lambda_k \|t_k - \ell\|_{L^2(h)}^2 = \|t_\lambda - \ell\|_{L^2(h)}^2 + \sum_{k=1}^N \lambda_k \|t_\lambda - t_k\|_{L^2(h)}^2. \tag{37}$$

The function  $V_n^D$  is affine in  $\lambda$ , therefore it takes its maximum on  $\Lambda^+$  at some  $e_k$ ,  $1 \leq k \leq N$ , giving:

$$D(f \| \hat{f}_*^D) - D(f \| f_{k^*}) \leq B_n (\hat{t}_*^D - t_{k^*}) + \max_{1 \leq k \leq N} V_n^D(e_k).$$

This concludes the proof. □

*Proof of Lemma 2.5.* Notice that for all  $\lambda, \lambda' \in \Lambda^+$ :

$$\int f_\lambda - \int f_{\lambda'} = (\lambda - \lambda') \cdot \nabla \int f_{\lambda'} + D(f_{\lambda'} \| f_\lambda). \tag{38}$$

The proof is then similar to the proof of Lemma 2.3 using (38) instead of (32). □

*Proof of Proposition 2.6.* Similarly to the proof of Proposition 2.4 we obtain that:

$$D(f \| \hat{f}_*^S) - D(f \| f_{k^*}) \leq B_n (\hat{g}_*^S - g_{k^*}) + A_n^S,$$

with:

$$A_n^S = \langle \hat{g}_*^S - g_{k^*}, I_n - \bar{f}_n \rangle - \frac{1}{2} D(\hat{f}_*^S \| f_{k^*}) - \frac{1}{2} \sum_{k=1}^N \hat{\lambda}_{*,k}^S D(\hat{f}_*^S \| f_k). \tag{39}$$

Since  $\|\log(\hat{f}_*^S / f_k)\|_\infty = \|\log \hat{\lambda}_*^S - g_k\| \leq 2K$  for  $1 \leq k \leq N$ , we can apply (36) with  $\hat{f}_*^S$  and  $f_k$ :

$$\begin{aligned} D(\hat{f}_*^S \| f_k) &\geq \frac{1}{2} e^{-\|\log(\hat{f}_*^S / f_k)\|_\infty} \int \hat{f}_*^S (\log(\hat{f}_*^S / f_k))^2 \\ &\geq \frac{1}{2} e^{-2K - \|\hat{g}_*^S\|_\infty} \int h (\hat{g}_*^S - g_k)^2 \\ &\geq \frac{1}{2} e^{-3K} \|\hat{g}_*^S - g_k\|_{L^2(h)}^2, \end{aligned} \tag{40}$$

where in the second and third inequalities we use that  $\|\hat{g}_*^S\|_\infty \leq \max_{1 \leq k \leq N} \|g_k\|_\infty \leq K$ . Applying (40) to both terms on the right hand side of (39) gives:

$$\begin{aligned} A_n(\hat{\lambda}_*^S) &\leq \langle \hat{g}_*^S - g_{k^*}, I_n - \bar{f}_n \rangle - \frac{e^{-3K}}{4} \|\hat{g}_*^S - g_{k^*}\|_{L^2(h)}^2 \\ &\quad - \frac{e^{-3K}}{4} \sum_{k=1}^N \hat{\lambda}_{*,k}^S \|\hat{g}_*^S - g_k\|_{L^2(h)}^2 \end{aligned}$$

$$\begin{aligned} &= \langle \hat{g}_*^S - g_{k^*}, I_n - \bar{f}_n \rangle - \frac{e^{-3K}}{4} \sum_{k=1}^N \hat{\lambda}_{*,k}^S \|g_k - g_{k^*}\|_{L^2(h)}^2 \\ &= V_n^S(\hat{\lambda}_*^S), \end{aligned}$$

where we used (37) for the second equality. The function  $V_n^S$  is affine in  $\lambda$ , therefore it takes its maximum on  $\Lambda^+$  at some  $e_k$ ,  $1 \leq k \leq N$ , giving:

$$D(f \|\hat{f}_*^S) - D(f \|\hat{f}_{k^*}^S) \leq B_n(\hat{g}_*^S - g_{k^*}) + \max_{1 \leq k \leq N} V_n^S(e_k).$$

This concludes the proof. □

*Proof of Theorem 3.1.* By Proposition 2.4, we have that:

$$D(f \|\hat{f}_*^D) - D(f \|\hat{f}_{k^*}^D) \leq B_n(\hat{t}_*^D - t_{k^*}) + \max_{1 \leq k \leq N} V_n^D(e_k). \tag{41}$$

Since  $I_n(dy)$  is an unbiased estimator of  $f(y)dy$ , we get  $B_n(\hat{t}_*^D - t_{k^*}) = 0$ . Notice that

$$\mathbb{P}\left(V_n^D(e_k) \geq \frac{\beta(\log(N) + x)}{n}\right) \leq \frac{e^{-x}}{N} \quad \text{for all } 1 \leq k \leq N, \tag{42}$$

implies

$$\mathbb{P}\left(\max_{1 \leq k \leq N} V_n^D(e_k) \geq \frac{\beta(\log(N) + x)}{n}\right) \leq e^{-x},$$

which will provide a control of the second term on the right hand side of (41). Thus, the proof of the theorem will be complete as soon as (42) is proved.

To prove (42), we use the concentration inequality of Proposition 5.3 in [3] which states that for  $Y_1, \dots, Y_n$  independent random variables with finite variances such that  $|Y_i - \mathbb{E}Y_i| \leq b$  for all  $1 \leq i \leq n$ , we have for all  $u > 0$  and  $a > 0$ :

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (Y_i - \mathbb{E}Y_i - a \text{Var } Y_i) > \left(\frac{1}{2a} + \frac{b}{3}\right) \frac{u}{n}\right) \leq e^{-u}. \tag{43}$$

Let us choose  $Y_i = t_k(X_i) - t_{k^*}(X_i)$  for  $1 \leq i \leq n$ . Then, since  $f_k$  and  $f_{k^*}$  belong to  $\mathcal{F}^D(K)$ , we have  $|Y_i - \mathbb{E}Y_i| \leq 4K$ , and:

$$\text{Var } Y_i \leq \int (t_k - t_{k^*})^2 f \leq e^{2L} \|t_k - t_{k^*}\|_{L^2(h)}^2. \tag{44}$$

Applying (43) with  $a = \exp(-6K - 2L)/4$ ,  $b = 4K$  and  $u = \log(N) + x$ , we obtain:

$$\begin{aligned} \frac{e^{-x}}{N} &\geq \mathbb{P}\left(\langle t_k - t_{k^*}, I_n - \bar{f}_n \rangle - \frac{e^{-6K-2L}}{4} \text{Var } Y_1 > \frac{\beta(\log(N) + x)}{n}\right) \\ &\geq \mathbb{P}\left(\langle t_k - t_{k^*}, I_n - \bar{f}_n \rangle - \frac{e^{-6K}}{4} \|t_k - t_{k^*}\|_{L^2(h)}^2 > \frac{\beta(\log(N) + x)}{n}\right) \\ &= \mathbb{P}\left(V_n^D(e_k) > \frac{\beta(\log(N) + x)}{n}\right), \end{aligned}$$



where the second inequality is due to (44). This proves (42) and completes the proof.  $\square$

*Proof of Theorem 3.5.* Using Proposition 2.6 and the notations defined there, we have that:

$$D(f\|\hat{f}_*^S) - D(f\|f_{k^*}) \leq B_n(\hat{g}_*^S - g_{k^*}) + \max_{1 \leq k \leq N} V_n^S(e_k). \tag{45}$$

*First step: Concentration inequality for  $\max_{1 \leq k \leq N} V_n^S(e_k)$ .*

We shall prove that

$$\mathbb{P}\left(\max_{1 \leq k \leq N} V_n^S(e_k) \geq \frac{\beta(\log(N) + x)}{n}\right) \leq e^{-x}. \tag{46}$$

It is enough to prove that for each  $1 \leq k \leq N$ :

$$\mathbb{P}\left(V_n^S(e_k) \geq \frac{\beta u}{n}\right) \leq e^{-u}. \tag{47}$$

Indeed take  $u = \log(N) + x$  and the union bound over  $1 \leq k \leq N$  to deduce (46) from (47).

The end of this first step is devoted to the proof of (47). Recall definition (63) of Toeplitz matrices associated to Fourier coefficients. We express the scalar product  $\langle \ell, I_n \rangle$  for  $\ell \in \mathbb{L}^\infty([-\pi, \pi])$  in a matrix form:

$$\langle \ell, I_n \rangle = \frac{1}{2\pi n} \sum_{i=1}^n \sum_{j=1}^n X_i X_j \int_{-\pi}^{\pi} \ell(x) \cos((i-j)x) dx = \frac{1}{n} X^T T_n(\ell) X. \tag{48}$$

We have the following expression of the covariance matrix of  $X$ :  $\Sigma_n = 2\pi T_n(f)$ . Since  $f$  is positive, we get that  $\Sigma_n$  is positive-definite. Set  $\xi = \Sigma_n^{-1/2} X$  so that  $\xi$  is a centered  $n$ -dimensional Gaussian vector whose covariance matrix is the  $n$ -dimensional identity matrix. By taking the expected value in (48), we obtain:

$$\mathbb{E} \langle \ell, I_n \rangle = \langle \ell, \bar{f}_n \rangle = \frac{1}{n} \text{tr}(R_n(\ell)),$$

where  $\text{tr}(A)$  denotes the trace of the matrix  $A$ , and  $R_n(\ell) = \Sigma_n^{\frac{1}{2}} T_n(\ell) \Sigma_n^{\frac{1}{2}}$ . Therefore the difference  $\langle \ell, I_n - \bar{f}_n \rangle$  takes the form:

$$\langle \ell, I_n - \bar{f}_n \rangle = \frac{1}{n} (\xi^T R_n(\ell) \xi - \text{tr}(R_n(\ell))).$$

We shall take  $\ell = g_k - g_{k^*}$ . For this reason, we assume that  $\ell$  is even and  $\|\ell\|_\infty \leq 2K$ . Let  $\eta = (\eta_i, 1 \leq i \leq n)$  denote the eigenvalues of the symmetric matrix  $R_n(\ell)$ , with  $\eta_1$  having the largest absolute value. Similarly to Lemma 4.2. of [4], we have that for all  $a > 0$ :

$$e^{-u} \geq \mathbb{P}\left(\langle \ell, I_n - \bar{f}_n \rangle \geq \frac{2|\eta_1|u}{n} + \frac{2\|\eta\|\sqrt{u}}{n}\right)$$

$$\geq \mathbb{P} \left( \langle \ell, I_n - \bar{f}_n \rangle \geq \frac{2|\eta_1|u}{n} + \frac{\|\eta\|^2}{an} + \frac{au}{n} \right), \tag{49}$$

where we used for the second inequality that  $2\sqrt{vw} \leq v/a + aw$  for all  $v, w, a > 0$ . Let us give upper bounds for  $|\eta_1|$  and  $\|\eta\|^2$ . We note  $\rho(A)$  for  $A \in \mathbb{R}^{n \times n}$  the spectral radius of the matrix  $A$ . Then by the well-known properties of the spectral radius, we have that:

$$|\eta_1| = \rho(R_n(\ell)) \leq \rho(\Sigma_n)\rho(T_n(\ell))$$

We deduce from (64) that  $\rho(\Sigma_n) = \rho(2\pi T_n(f)) \leq 2\pi\|f\|_\infty \leq \exp(L)$  and  $\rho(T_n(\ell)) \leq \|\ell\|_\infty \leq 2K$ . Therefore we obtain:

$$|\eta_1| \leq 2K e^L. \tag{50}$$

As for  $\|\eta\|^2$ , we have:

$$\|\eta\|^2 = \text{tr}(R_n^2(\ell)) = \text{tr}((\Sigma_n T_n(\ell))^2) \leq \rho(\Sigma_n)^2 \text{tr}(T_n^2(\ell)) \leq e^{2L} n \|\ell\|_{L^2(h)}^2, \tag{51}$$

where we used (65) for the last inequality. Using (50) and (51) in (49) gives:

$$\begin{aligned} e^{-u} &\geq \mathbb{P} \left( \langle \ell, I_n - \bar{f}_n \rangle \geq \frac{4K e^L u}{n} + \frac{e^{2L} \|\ell\|_{L^2(h)}^2}{a} + \frac{au}{n} \right) \\ &\geq \mathbb{P} \left( \langle \ell, I_n - \bar{f}_n \rangle - \frac{e^{-3K}}{4} \|\ell\|_{L^2(h)}^2 \geq \frac{\beta u}{n} \right), \end{aligned}$$

where for the second inequality we set  $a = 4 \exp(2L + 3K)$ . This proves (47), thus (46).

*Second step: Upper bound for the bias term  $B_n(\hat{g}_*^S - g_{k^*})$*

We set  $\ell_* = \hat{g}_*^S - g_{k^*}$  and we have  $\|\ell_*\|_{2,r} \leq 2K/C_r$ . Let  $(a_k)_{k \in \mathbb{Z}}$  be the corresponding Fourier coefficients, which are real as  $\ell_*$  is even. We decompose the the bias term as follows:

$$B_n(\ell_*) = \langle \bar{f}_n - f, \ell_* \rangle = \langle \bar{f}_{n,1} - f, \ell_* \rangle - \langle \bar{f}_{n,2}, \ell_* \rangle, \tag{52}$$

with  $\bar{f}_{n,1}, \bar{f}_{n,2}$  given by, for  $x \in [-\pi, \pi]$ :

$$\bar{f}_{n,1}(x) = \sum_{|j| < n} \frac{\gamma_j}{2\pi} e^{ijx} \quad \text{and} \quad \bar{f}_{n,2}(x) = \frac{1}{n} \sum_{|j| < n} \frac{|j|\gamma_j}{2\pi} e^{ijx}.$$

For the first term of the right hand side of (52) notice that:

$$\bar{f}_{n,1}(x) - f(x) = - \sum_{|j| \geq n} \frac{\gamma_j}{2\pi} e^{ijx}.$$

We deduce that  $\langle \bar{f}_{n,1} - f, \ell_* \rangle = \langle \bar{f}_{n,1} - f, \bar{\ell}_* \rangle$ , with  $\bar{\ell}_* = \sum_{|j| \geq n} a_j e^{jjx}$ . Then, by the Cauchy-Schwarz inequality, we get:

$$|\langle \bar{f}_{n,1} - f, \bar{\ell}_* \rangle| \leq \|\bar{f}_{n,1} - f\|_{L^2(h)} \|\bar{\ell}_*\|_{L^2(h)}.$$

Thanks to Lemma 3.4, we get:

$$\begin{aligned} \|\bar{f}_{n,1} - f\|_{L^2(h)}^2 &= \sum_{|j| \geq n} \frac{\gamma_j^2}{4\pi^2} \leq \sum_{|j| \geq n} \frac{|j|^{2r}}{n^{2r}} \frac{\gamma_j^2}{4\pi^2} \\ &\leq \frac{1}{n^{2r}} \{f\}_{2,r}^2 \leq \frac{1}{n^{2r}} \|f\|_{2,r}^2 \leq \frac{C(r,L)^2}{n^{2r}}. \end{aligned}$$

This gives  $\|\bar{f}_{n,1} - f\|_{L^2(h)} \leq C(r,L)n^{-r}$ . Similarly, we have  $\|\bar{\ell}_*\|_{L^2(h)} \leq n^{-r} \{\ell_*\}_{2,r} \leq n^{-r} \|\ell_*\|_{2,r} \leq 2Kn^{-r}/C_r$ . We deduce that:

$$|\langle \bar{f}_{n,1} - f, \bar{\ell}_* \rangle| \leq \frac{2KC(r,L)}{C_r} n^{-2r}. \quad (53)$$

For the second term on the right hand side of (52), we have:

$$\langle \bar{f}_{n,2}, \ell_* \rangle = \frac{1}{n} \sum_{|j| < n} \frac{|j|\gamma_j}{2\pi} a_j.$$

Using the Cauchy-Schwarz inequality and then Lemma 3.4, we get as  $r > 1/2$ :

$$|\langle \bar{f}_{n,2}, \ell_* \rangle| \leq \frac{1}{n} \{\ell_*\}_{2,1/2} \{f\}_{2,1/2} \leq \frac{1}{n} \|\ell_*\|_{2,r} \|f\|_{2,r} \leq \frac{2KC(r,L)}{C_r} n^{-1}. \quad (54)$$

Therefore combining (53) and (54), we obtain the following upper bound for the bias:

$$|B_n(\ell_*)| \leq \frac{4KC(r,L)}{C_r} n^{-1}. \quad (55)$$

*Third step: Conclusion*

Use (46) and (55) in (45) to get the result.  $\square$

## 5.2. Proofs of results in Section 4

In the following proof, we shall use the Hellinger distance which is defined as follows. For two non-negative integrable functions  $p$  and  $q$ , the Hellinger distance  $H(p, q)$  is defined as:

$$H(p, q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}.$$

A well known property of this distance is that its square is smaller than the Kullback-Leibler divergence defined by 4, that is for all non-negative integrable functions  $p$  and  $q$ , we have:

$$H^2(p, q) \leq D(p||q).$$

*Proof of Proposition 4.2.* Since the probability densities  $(f_k, 1 \leq k \leq N)$  belongs to  $\mathcal{F}^D(L)$ , we have:

$$\begin{aligned} \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}^D(L)} \mathbb{P}_f \left( D(f \|\hat{f}_n) - \min_{1 \leq k \leq N} D(f \| f_k) \geq \frac{\beta'(\log(N) + x)}{n} \right) \\ \geq \inf_{\hat{f}_n} \max_{1 \leq k \leq N} \mathbb{P}_{f_k} \left( D(f_k \|\hat{f}_n) \geq \frac{\beta'(\log(N) + x)}{n} \right) \\ \geq \inf_{\hat{f}_n} \max_{1 \leq k \leq N} \mathbb{P}_{f_k} \left( H^2(f_k, \hat{f}_n) \geq \frac{\beta'(\log(N) + x)}{n} \right). \end{aligned}$$

For the choice of  $(f_k, 1 \leq k \leq N)$ , we follow the choice given in the proof of Theorem 2 of [22]. Let  $D$  be the smallest positive integer such that  $2^{D/8} \geq N$  and  $\Delta = \{0, 1\}^D$ . For  $0 \leq j \leq D - 1, s \in \mathbb{R}$ , we set:

$$\alpha_j(s) = \frac{T}{D} \mathbf{1}_{(0, \frac{1}{2}]}(Ds - j) - \frac{T}{D} \mathbf{1}_{(\frac{1}{2}, 1]}(Ds - j),$$

where  $T$  verifies  $0 < T \leq D(1 - e^{-L})$ . Notice the support of the function  $\alpha_j$  is  $(j/D, (j + 1)/D]$ . Then for any  $\delta = (\delta_1, \dots, \delta_D) \in \Delta$ , the function  $f^\delta$  defined by:

$$f^\delta(y) = 1 + \sum_{j=0}^{D-1} \delta_j \alpha_j(y_1), \quad y = (y_1, \dots, y_d) \in [0, 1]^d,$$

is a probability density function with  $e^L \geq 1 + T/D \geq f \geq 1 - T/D \geq e^{-L}$ . This implies that  $f^\delta \in \mathcal{F}^D(L)$ . As shown in the proof of Theorem 2 in [22], there exists  $N$  probability densities  $(f_k, 1 \leq k \leq N)$  amongst  $\{f^\delta, \delta \in \Delta\}$  such that for any  $i \neq j$ , we have:

$$H^2(f_i, f_j) \geq \frac{8^{-3/2} T^2}{4D^2},$$

and  $f_1$  can be chosen to be the density of the uniform distribution on  $[0, 1]^d$ . Recall the notation  $p^{\otimes n}$  of the  $n$ -product probability density corresponding to the probability density  $p$ . Then we also have (see the proof of Theorem 2 of [22]) for all  $1 \leq i \leq N$ :

$$D(f_i^{\otimes n} \| f_1^{\otimes n}) \leq \frac{nT^2}{D^2}.$$

Let us take  $T = D\sqrt{(\log(N) + x)/3n}$ , so that with condition (29) we indeed have  $T \leq D(1 - e^{-L})$ . With this choice, and the definition of  $\beta'$ , we have for  $1 \leq i \neq j \leq N$

$$H^2(f_i, f_j) \geq 4 \frac{\beta'(\log(N) + x)}{n} \quad \text{and} \quad D(f_i^{\otimes n} \| f_1^{\otimes n}) \leq \frac{\log(N) + x}{3}.$$

Now we apply Corollary 5.1 of [3] with  $m = N - 1$  and with the squared Hellinger distance instead of the  $L^2$  distance to get that for any estimator  $\hat{f}_n$ :

$$\begin{aligned} \max_{1 \leq k \leq N} \mathbb{P}_{f_k} \left( H^2(f_k, \hat{f}_n) \geq \frac{\beta'(\log(N) + x)}{n} \right) &\geq \frac{1}{12} \min \left( 1, (N - 1) e^{-(\log(N) + x)} \right) \\ &\geq \frac{1}{24} e^{-x}. \end{aligned}$$

This concludes the proof. □

*Proof of Proposition 4.3.* Similarly to the proof of Proposition 4.2, the left hand side of (31) is greater than:

$$\inf_{\hat{f}_n} \max_{1 \leq k \leq N} \mathbb{P}_{f_k} \left( H^2(f_k, \hat{f}_n) \geq \frac{\beta'(\log(N) + x)}{n} \right).$$

We shall choose a set of spectral densities  $(f_k, 1 \leq k \leq N)$  similarly as in the proof of Proposition 4.2 such that  $f_k \in \mathcal{F}_r^S(L)$ . Let us define  $\varphi : [0, \pi] \rightarrow \mathbb{R}$  as, for  $x \in [0, \pi]$ :

$$\varphi(x) = \zeta(x) \mathbf{1}_{[0, \pi/2]}(x) - \zeta(x) \mathbf{1}_{[\pi/2, \pi]}(x) \quad \text{with} \quad \zeta(x) = e^{-1/(x(\frac{\pi}{2} - x))}.$$

We have that  $\varphi \in C^\infty(\mathbb{R})$  and:

$$\|\varphi\|_\infty = e^{-16/\pi^2}, \quad \int_0^\pi \varphi = 0. \tag{56}$$

Let  $D$  be the smallest integer such that  $2^{D/8} \geq N$  and  $\Delta = \{0, 1\}^D$ . For  $1 \leq j \leq D$ ,  $x \in [0, \pi]$ , let  $\bar{\alpha}_j(x)$  be defined as:

$$\bar{\alpha}_j(x) = \varphi(Dx - (j - 1)\pi),$$

and for any  $\delta = (\delta_1, \dots, \delta_D) \in \Delta$  and  $s \geq 0$ , let the function  $f_s^\delta$  be defined by:

$$2\pi f_s^\delta(y) = 1 + s \sum_{j=1}^D \delta_j \bar{\alpha}_j(|y|), \quad y \in [-\pi, \pi]. \tag{57}$$

Since  $\int_0^\pi \varphi = 0$ , we get:

$$\frac{1}{2\pi} \int_{-\pi}^\pi f_s^\delta(x) dx = 1 \quad \text{and} \quad 1 - s \|\varphi\|_\infty \leq 2\pi f_s^\delta \leq 1 + s \|\varphi\|_\infty. \tag{58}$$

We assume that  $s \in [0, 1/2]$ , so that  $2\pi f_s^\delta \geq 1/2$ . Let us denote  $g_s^\delta = g_{f_s^\delta} = \log(2\pi f_s^\delta)$ . We first give upper bounds for  $\|(g_s^\delta)^{(p)}\|_{L^2(h)}$  with  $p \in \mathbb{N}$ .

For  $p = 0$ , we have by (58):

$$\|g_s^\delta\|_{L^2(h)} \leq \log \left( \frac{1}{1 - s \|\varphi\|_\infty} \right) \leq \frac{s \|\varphi\|_\infty}{1 - s \|\varphi\|_\infty} \leq 2s. \tag{59}$$

For  $p \geq 1$ , we get by Faà di Bruno's formula that:

$$\|(g_s^\delta)^{(p)}\|_{L^2(h)} = \left\| \sum_{k \in \mathcal{K}_p} \frac{p!}{k_1! k_2! \dots k_p!} \frac{(-1)^{\bar{k}+1} \bar{k}!}{(2\pi f_s^\delta)^{\bar{k}}} \prod_{\ell=1}^p \left( \frac{(2\pi f_s^\delta)^{(\ell)}}{\ell!} \right)^{k_\ell} \right\|_{L^2(h)}, \tag{60}$$

with  $\mathcal{K}_p = \{k = (k_1, \dots, k_p) \in \mathbb{N}^p; \sum_{\ell=1}^p \ell k_\ell = p\}$  and  $\bar{k} = \sum_{\ell=1}^p k_\ell$ . The  $\ell$ -th derivative of  $2\pi f_s^\delta$  is given by, for  $y \in [0, \pi]$ :

$$(2\pi f_s^\delta(y))^{(\ell)} = sD^\ell \sum_{j=1}^D \delta_j \varphi^{(\ell)}(Dy - (j-1)\pi).$$

Therefore we have the following bound for this derivative:

$$\|(2\pi f_s^\delta(y))^{(\ell)}\|_\infty \leq sD^\ell \|\varphi^{(\ell)}\|_\infty.$$

From  $\varphi \in C^\infty(\mathbb{R})$ , we deduce that  $\|\varphi^{(\ell)}\|_\infty$  is finite for all  $\ell \in \mathbb{N}^*$ . Since  $s \in [0, 1/2]$  and  $2\pi f_s^\delta \geq 1 - s \|\varphi\|_\infty \geq 1/2$ , there exists a constant  $\bar{C}_p$  depending on  $p$  (and not depending on  $N$ ), such that:

$$\|(g_s^\delta)^{(p)}\|_{L^2(h)} \leq s\bar{C}_p D^p \leq s\bar{C}_p \frac{16^p}{\log(2)^p} \log(N)^p. \tag{61}$$

In order to have  $f_s^\delta \in \mathcal{F}_r^S(L)$ , we need to ensure that  $\|g_s^\delta\|_{2,r} \leq L/\mathcal{C}_r$ . For  $r \in \mathbb{N}^*$ , we have:

$$\|g_s^\delta\|_{2,r} = \sqrt{\|g_s^\delta\|_{L^2(h)}^2 + \|(g_s^\delta)^{(r)}\|_{L^2(h)}^2}.$$

Therefore if  $s \in [0, s_{r,L}]$  with  $s_{r,L} \in [0, 1/2]$  given by:

$$s_{r,L} = \log(N)^{-r} \bar{C}_{r,L}, \quad \text{with} \quad \bar{C}_{r,L} = \min\left(\frac{\log(2)^r}{2}, \frac{\log(2)^r L}{\sqrt{8}\mathcal{C}_r}, \frac{\log(2)^r L}{\sqrt{2}\mathcal{C}_r 16^r \bar{C}_r}\right),$$

then by (59) and (61) we get:

$$\|g_s^\delta\|_{2,r} \leq \sqrt{\frac{L^2}{2\mathcal{C}_r^2} + \frac{L^2}{2\mathcal{C}_r^2}} = \frac{L}{\mathcal{C}_r}.$$

Let  $\lceil r \rceil$  and  $\lfloor r \rfloor$  denote the unique integers such that  $\lceil r \rceil - 1 < r \leq \lceil r \rceil$  and  $\lfloor r \rfloor \leq r < \lfloor r \rfloor + 1$ . For  $r \notin \mathbb{N}^*$ , Hölder's inequality yields:

$$\begin{aligned} \|g_s^\delta\|_{2,r} &= \sqrt{\|g_s^\delta\|_{L^2(h)}^2 + \{g_s^\delta\}_{2,r}^2} \\ &\leq \sqrt{\|g_s^\delta\|_{L^2(h)}^2 + \{g_s^\delta\}_{2,\lceil r \rceil}^{2(r-\lfloor r \rfloor)} \{g_s^\delta\}_{2,\lfloor r \rfloor}^{2(\lceil r \rceil-r)}} \\ &= \sqrt{\|g_s^\delta\|_{L^2(h)}^2 + \|(g_s^\delta)^{(\lceil r \rceil)}\|_{L^2(h)}^{2(r-\lfloor r \rfloor)} \|(g_s^\delta)^{(\lfloor r \rfloor)}\|_{L^2(h)}^{2(\lceil r \rceil-r)}}. \end{aligned}$$

Using (61) and (61) with  $p = \lceil r \rceil$  and  $p = \lfloor r \rfloor$ , we obtain:

$$\|(g_s^\delta)^{(\lceil r \rceil)}\|_{L^2(h)}^{2(r-\lfloor r \rfloor)} \|(g_s^\delta)^{(\lfloor r \rfloor)}\|_{L^2(h)}^{2(\lceil r \rceil-r)} \leq s^2 \bar{C}_{\lceil r \rceil}^{2(r-\lfloor r \rfloor)} \bar{C}_{\lfloor r \rfloor}^{2(\lceil r \rceil-r)} \frac{16^{2r}}{\log(2)^{2r}} \log N^{2r}.$$

Hence if  $s \in [0, s_{r,L}]$  with  $s_{r,L} \in [0, 1/2]$  given by:

$$s_{r,L} = \log(N)^{-r} \bar{C}_{r,L},$$

with

$$\bar{C}_{r,L} = \min \left( \frac{\log(2)^r}{2}, \frac{\log(2)^r L}{\sqrt{8}C_r}, \frac{\log(2)^r L}{\sqrt{2}C_r 16^r \bar{C}_{[r]}^{r-[r]} \bar{C}_{[r]}^{[r]-r}} \right),$$

we also have  $\|g_s^\delta\|_{2,r} \leq L/C_r$ , providing  $f_s^\delta \in \mathcal{F}_r^S(L)$ .

Mimicking the proof of Theorem 2 in [22] and omitting the details, we first obtain (see last inequality of p.975 in [22]) that for  $\delta, \delta' \in \Delta$ :

$$H^2(f_s^\delta, f_s^{\delta'}) \geq 8^{-3/2} \frac{\sigma(\delta, \delta')}{D} \frac{2}{\pi} s^2 \int_0^\pi \varphi^2,$$

with  $\sigma(\delta, \delta')$  the Hamming distance between  $\delta$  and  $\delta'$ , and then deduce that there exist  $(\delta^k, 1 \leq k \leq N)$  in  $\Delta$  with  $\delta^1 = 0$  such that for any  $1 \leq i \neq j \leq N$  and  $s \in [0, s_{r,L}]$ , we have (see first inequality of p.976 in [22]):

$$H^2(f_s^{\delta^i}, f_s^{\delta^j}) \geq \frac{2 \cdot 8^{-5/2}}{\pi} s^2 \int_0^\pi \varphi^2.$$

Notice  $f_s^{\delta^1} = f_s^0 = h$  is the density of the uniform distribution on  $[-\pi, \pi]$ .

With a slight abuse of notation, let us denote by  $P_f$  the joint probability density of the centered Gaussian sequence  $X = (X_1, \dots, X_n)$  corresponding to the spectral density  $f$ . Assume  $X$  is standardized (that is  $\text{Var}(X_1) = 1$ ), which implies  $\int f = 1$ . Let  $\Sigma_{n,f}$  denote the corresponding covariance matrix. Since  $h = (1/2\pi)\mathbf{1}_{[-\pi,\pi]}$ , we have  $\Sigma_{n,h} = \mathcal{I}_n$  the  $n \times n$ -dimensional identity matrix. We compute:

$$\begin{aligned} D(P_f \| P_h) &= \int_{\mathbb{R}^n} P_f(x) \log \left( \frac{P_f(x)}{P_h(x)} \right) dx \\ &= \int_{\mathbb{R}^n} P_f(x) \log \left( \frac{1}{\sqrt{\det(\Sigma_{n,f})}} \exp \left( -\frac{1}{2} x^T (\Sigma_{n,f}^{-1} - \mathcal{I}_n) x \right) \right) dx \\ &= -\frac{1}{2} \log(\det(\Sigma_{n,f})) - \frac{1}{2} E_f [X^T (\Sigma_{n,f}^{-1} - \mathcal{I}_n) X]. \end{aligned}$$

The expected value in the previous equality can be written as:

$$E_f [X^T (\Sigma_{n,f}^{-1} - \mathcal{I}_n) X] = \text{tr} \left( (\Sigma_{n,f}^{-1} - \mathcal{I}_n) E_f [X^T X] \right) = \text{tr} (\mathcal{I}_n - \Sigma_{n,f}) = 0,$$

where for the last equality, we used that the Gaussian random variables are standardized. This yields  $D(P_f \| P_h) = -\frac{1}{2} \log(\det(\Sigma_{n,f}))$ . We can use this last equality for  $f = f_s^\delta$  since  $\int f_s^\delta = 1$  thanks to (56), and obtain:

$$D(P_{f_s^\delta} \| P_{f_s^0}) = -\frac{1}{2} \log(\det(\Sigma_{n,f_s^\delta})).$$

Notice that for  $s \in [0, s_{r,L}]$ , we have  $3/2 \geq 1 + s \|\varphi\|_\infty \geq 2\pi f_s^\delta \geq 1 - s \|\varphi\|_\infty \geq 1/2$  thanks to (58) and (56). Therefore we have:

$$D(P_{f_s^\delta} \| P_{f_s^0}) \leq \frac{n}{2} \|2\pi f_s^\delta - 1\|_{L^2(h)}^2 \leq \frac{n}{2} \frac{s^2}{\pi} \int_0^\pi \varphi^2, \tag{62}$$

where we used  $\Sigma_{n, f_s^\delta} = T_n(2\pi f_s^\delta)$  and Lemma A.2 with  $\ell = 2\pi f_s^\delta$  for the first inequality, and (57) for the second inequality. We set:

$$C(r, L) = \frac{3\bar{C}_{r,L}^2 \int_0^\pi \varphi^2}{2\pi} \quad \text{and} \quad s = \sqrt{\frac{2\pi}{3 \int_0^\pi \varphi^2}} \sqrt{\frac{\log(N) + x}{n}},$$

so that (30) holds for  $s \in [0, s_{r,L}]$ . We obtain for all  $\delta^1, \delta^2 \in \bar{\Delta}$ ,  $\delta \in \Delta$ :

$$H^2(f_s^{\delta_1}, f_s^{\delta_2}) \geq 4 \frac{\beta'(\log(N) + x)}{n} \quad \text{and} \quad D(P_{f_s^\delta} \| P_{f_s^0}) \leq \frac{\log(N) + x}{3}.$$

We conclude the proof as in the end of the proof of Proposition 4.2. □

## Appendix

### A.1. Results on Toeplitz matrices

Let  $\ell \in L^1(h)$  be a real function with  $h = 1/(2\pi)\mathbf{1}_{[-\pi, \pi]}$ . We define the corresponding Toeplitz matrix  $T_n(\ell)$  of size  $n \times n$  of its Fourier coefficients by:

$$[T_n(\ell)]_{j,k} = \frac{1}{2\pi} \int_{-\pi}^\pi \ell(x) e^{i(j-k)x} dx \quad \text{for } 1 \leq j, k \leq n. \tag{63}$$

Notice that  $T_n(\ell)$  is Hermitian. It is also real if  $\ell$  is even. Recall that  $\rho(A)$  denotes the spectral density of the matrix  $A$ .

**Lemma A.1.** *Let  $\ell \in L^2(h)$  be a real function.*

1. *All the eigenvalues of  $T_n(\ell)$  belong to  $[\min \ell, \max \ell]$ . In particular, we have the following upper bound on the spectral radius  $\rho(T_n(\ell))$  of  $T_n(\ell)$ :*

$$\rho(T_n(\ell)) \leq \|\ell\|_\infty. \tag{64}$$

2. *For the trace of  $T_n(\ell)$  and  $T_n^2(\ell)$ , we have:*

$$\text{tr}(T_n(\ell)) = \frac{n}{2\pi} \int_{-\pi}^\pi \ell(x) dx \quad \text{and} \quad \text{tr}(T_n^2(\ell)) \leq n \|\ell\|_{L^2(h)}^2. \tag{65}$$

*Proof.* For Property (1), see Equation (6) of Section 5.2 in [19]. For Property (2), the first part is clear and for the second part, see Lemma 3.1 of [17]. □

We use the following elementary result.

**Lemma A.2.** *Let  $\ell \in L^2(h)$  such that  $\int \ell h = 1$  and  $\ell(x) \in [1/2, 3/2]$ , then we have:*

$$\log(\det(T_n(\ell))) \geq -n \|\ell - 1\|_{L^2(h)}^2. \tag{66}$$



*Proof.* Notice that by Property (1), the eigenvalues  $(\nu_i, 1 \leq i \leq n)$  of  $T_n(\ell)$  verify  $\nu_i \in [1/2, 3/2]$ . For  $t \in [-1/2, 1/2]$ , we have  $\log(1+t) \geq t - t^2$ , giving that:

$$\begin{aligned} \log(\det(T_n(\ell))) &= \sum_{i=1}^n \log(\nu_i) \geq \sum_{i=1}^n (\nu_i - 1) - (\nu_i - 1)^2 \\ &= -\operatorname{tr}(T_n^2(\ell - 1)) \geq -n \|\ell - 1\|_{L^2(h)}^2, \end{aligned}$$

where we used that  $T_n(\ell - 1) = T_n(\ell) - \mathcal{I}_n$  for the second equality and Property (2) for the second inequality.  $\square$

### A.2. Proof of Lemma 3.4

The next Lemma is inspired by the work of [18] on fractional Sobolev spaces. For  $r \in (0, 1)$  and  $\ell \in L^2(h)$ , we define:

$$I_r(\ell) = \frac{1}{2\pi} \int_{[-\pi, \pi]^2} \frac{|\ell(x+y) - \ell(x)|^2}{|y|^{1+2r}} dx dy,$$

where we set  $\ell(z) = \ell(z - 2\pi)$  for  $z \in (\pi, 2\pi]$  and  $\ell(z) = \ell(z + 2\pi)$  for  $z \in [-2\pi, -\pi)$ .

**Lemma A.3.** *Let  $r \in (0, 1)$ . Then there exist constants  $0 < c_r < C_r < \infty$  depending on  $r$ , such that for all  $\ell$  in  $L^2(h)$  we have:*

$$c_r \{\ell\}_{2,r}^2 \leq I_r(\ell) \leq C_r \{\ell\}_{2,r}^2. \quad (67)$$

*Proof.* Using the Fourier representation of  $\ell$ , we get:

$$I_r(\ell) = \sum_{k \in \mathbb{Z}} |a_k|^2 \int_{-\pi}^{\pi} \frac{|1 - e^{iky}|^2}{|y|^{1+2r}} dy = \sum_{k \in \mathbb{Z}} |k|^{2r} |a_k|^2 \int_{-|k|\pi}^{|k|\pi} \frac{|1 - e^{iz}|^2}{|z|^{1+2r}} dz.$$

For  $r \in (0, 1)$  and  $k \in \mathbb{Z}^*$ , we have

$$0 < c_r := \int_{-\pi}^{\pi} \frac{|1 - e^{iz}|^2}{|z|^{1+2r}} dz \leq \int_{-|k|\pi}^{|k|\pi} \frac{|1 - e^{iz}|^2}{|z|^{1+2r}} dz \leq \int_{\mathbb{R}} \frac{|1 - e^{iz}|^2}{|z|^{1+2r}} dz =: C_r < +\infty.$$

This yields (67).  $\square$

*First step:  $r \in (1/2, 1)$*

Let  $r \in (1/2, 1)$  and set  $L = C_r K$ . Let  $f = e^g$  with  $g \in W_r$  such that  $\|g\|_{2,r} \leq K$ . Thanks to (26), we have  $\|g\|_{\infty} \leq C_r K = L$ . Using that  $|e^x - e^y| \leq e^L |x - y|$  for  $x, y \in [-L, L]$ , we deduce that:

$$I_r(f) = I_r(e^g) \leq e^{2L} I_r(g) \quad \text{and} \quad \|f\|_{L^2(h)}^2 \leq e^{2L}. \quad (68)$$

Using (67) twice, we get:

$$\|f\|_{2,r}^2 \leq e^{2L} \left(1 + \frac{C_r}{c_r} \{g\}_{2,r}^2\right) \leq e^{2C_r K} \left(1 + \frac{C_r}{c_r} K^2\right).$$

Which proves the Lemma for  $r \in (1/2, 1)$ .

*Second step:  $r \in \mathbb{N}^*$*

Let  $r \in \mathbb{N}^*$ . For  $\ell \in W_r$ , the  $r$ -th derivative of  $\ell$ , say  $\ell^{(r)}$ , exists in  $L^2(h)$  and:

$$\{\ell\}_{2,r}^2 = \|\ell^{(r)}\|_{L^2(h)}^2 \quad \text{as well as} \quad \|\ell\|_{2,r}^2 = \|\ell\|_{L^2(h)}^2 + \|\ell^{(r)}\|_{L^2(h)}^2.$$

According to (26), we also get that for all  $p \in \mathbb{N}$  with  $p < r$  we have  $\|\ell^{(p)}\|_\infty \leq C_{r-p} \{\ell^{(r)}\}_{2,r} \leq C_1 \{\ell^{(r)}\}_{2,r}$ .

Set  $L = C_r K$ . Let  $f = e^g$  with  $\|g\|_{2,r} \leq K$ . We have  $\|g^{(p)}\|_\infty \leq C_1 K$  for all integer  $p < r$ . According to Leibniz's rule, we get that  $f^{(r)} = g^{(r)} f + P_r(g^{(1)}, \dots, g^{(r-1)})f$ , where  $P_r$  is a polynomial function of maximal degree  $r$  such that:

$$\max_{x_1, \dots, x_{r-1} \in [-C_1 K, C_1 K]} |P_r(x_1, \dots, x_{r-1})| \leq C_{r,1} K^r. \tag{69}$$

for some finite constant  $C_{r,1}$ . We deduce that:

$$\|f^{(r)}\|_{L^2(h)} \leq e^L \|g^{(r)}\|_{L^2(h)} + e^L C_{r,1} K^r.$$

Then use that  $\|f\|_{L^2(h)} \leq e^L$  to get the Lemma for  $r \in \mathbb{N}^*$ .

*Third step:  $r > 1, r \notin \mathbb{N}^*$*

Let  $r > 1$  such that  $r \notin \mathbb{N}^*$ . Set  $p = [r] \in \mathbb{N}^*$  the integer part of  $r$  and  $s = r - p \in (0, 1)$ . For  $\ell \in W_r$ , the  $p$ -th derivative of  $\ell$ , say  $\ell^{(p)}$ , exists in  $L^2(h)$  and:

$$\{\ell\}_{2,r}^2 = \{\ell^{(p)}\}_{2,s}^2 \quad \text{as well as} \quad \|\ell\|_{2,r}^2 = \|\ell\|_{L^2(h)}^2 + \{\ell^{(p)}\}_{2,s}^2. \tag{70}$$

Thanks to (67) (twice) and the triangle inequality, we have for all measurable function  $t$ :

$$c_s \{\ell t\}_{2,s}^2 \leq I_s(\ell t) \leq \|t\|_\infty^2 I_s(\ell) + J_s(\ell, t) \leq \|t\|_\infty^2 C_s \{\ell\}_{2,s}^2 + J_s(\ell, t), \tag{71}$$

with

$$J_s(\ell, t) = \frac{1}{2\pi} \int_{[-\pi, \pi]^2} \ell(x)^2 \frac{|t(x+y) - t(x)|^2}{|y|^{1+2s}} dx dy.$$

Let  $K > 0$  and set  $L = C_r K$ . Let  $f = e^g$  with  $g \in W_r$  such that  $\|g\|_{2,r} \leq K$ . Following the proof of Lemma A.3, we first give an upper bound of  $J_s(\ell, f)$  in this

context under the only condition that  $\ell \in L^2(h)$ . Using that  $|e^x - e^y| \leq e^L |x - y|$  for  $x, y \in [-L, L]$ , we deduce that:

$$\int_{-\pi}^{\pi} \frac{|f(x+y) - f(x)|^2}{|y|^{1+2s}} dy \leq e^{2L} \int_{-\pi}^{\pi} \frac{|g(x+y) - g(x)|^2}{|y|^{1+2s}} dy.$$

Since a.e.  $g(x) = \sum_{k \in \mathbb{Z}} a_k e^{ikx}$ , we deduce that:

$$J_s(\ell, f) \leq \frac{e^{2L}}{2\pi} \int_{-\pi}^{\pi} dx \ell(x)^2 \sum_{k, j \in \mathbb{Z}} |a_k| |a_j| \int_{-\pi}^{\pi} \frac{|(1 - e^{iky})(1 - e^{-ijy})|}{|y|^{1+2s}} dy.$$

Let  $\varepsilon \in (0, 1/2)$  such that  $s + \varepsilon \leq 1$ . Since  $|1 - e^{ix}| \leq 2|x|^{s+\varepsilon}$  for all  $x \in \mathbb{R}$ , we deduce that:

$$\int_{-\pi}^{\pi} \frac{|(1 - e^{iky})(1 - e^{-ijy})|}{|y|^{1+2s}} dy \leq C_{2,\varepsilon} |k|^{s+\varepsilon} |j|^{s+\varepsilon},$$

for some constant  $C_{2,\varepsilon}$  depending only on  $\varepsilon$ . Using Cauchy-Schwarz inequality and the fact that  $r - s - \varepsilon > 1/2$ , we get:

$$\sum_{k \in \mathbb{Z}} |k|^{s+\varepsilon} |a_k| \leq C_{r-s-\varepsilon} \{g\}_{2,r}.$$

We deduce that:

$$J_s(\ell, f) \leq e^{2L} \|\ell\|_{L^2(h)}^2 C_{2,\varepsilon} C_{r-s-\varepsilon}^2 \{g\}_{2,r}^2. \tag{72}$$

According to Leibniz's rule, we get that  $f^{(p)} = \ell f + g^{(p)} f$  with  $\ell = P_p(g^{(1)}, \dots, g^{(p-1)})$ . We get:

$$\begin{aligned} c_s \{\ell f\}_{2,s}^2 &\leq \|f\|_{\infty}^2 C_s \{\ell\}_{2,s}^2 + J_s(\ell, f) \\ &\leq e^{2L} C_s \{f\}_{2,s}^2 + e^{2L} \|\ell\|_{L^2(h)}^2 C_{2,\varepsilon} C_{r-s-\varepsilon}^2 \{g\}_{2,r}^2, \end{aligned} \tag{73}$$

where we used (71) for the first inequality and (72) for the latter. Then use (69) with  $r$  replaced by  $p$  to get that  $\|\ell\|_{L^2(h)} \leq \|\ell\|_{\infty} \leq C_{p,1} K^p$ . Notice also that:

$$\{f\}_{2,s}^2 \leq e^{2L} \frac{C_s}{c_s} \{g\}_{2,s}^2,$$

using (67) twice and (68) (with  $s$  instead of  $r$ ). We deduce that  $\{\ell f\}_{2,s}$  is bounded by a constant depending only on  $K, r$  and  $\varepsilon$ .

The upper bound of  $\{g^{(p)} f\}_{2,s}^2$  is similar. Using (71) and (72), we get:

$$\begin{aligned} c_s \{g^{(p)} f\}_{2,s}^2 &\leq \|f\|_{\infty}^2 I_s(g^{(p)}) + J_s(g^{(p)}, f) \\ &\leq e^{2L} C_s \{g^{(p)}\}_{2,s}^2 + e^{2L} \|g^{(p)}\|_{L^2(h)}^2 C_{2,\varepsilon} C_{r-s-\varepsilon}^2 \{g\}_{2,r}^2. \end{aligned}$$

We deduce that  $\{g^{(p)} f\}_{2,s}$ , and thus  $f^{(p)}$ , is bounded by a constant depending only on  $K, r$  and  $\varepsilon$ . Then use (70) and that  $\|f\|_{L^2(h)} \leq \|f\|_{\infty} \leq e^L$  to get the Lemma for  $r > 1$  and  $r \notin \mathbb{N}$ . This concludes the proof.

## References

- [1] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 41–48. Curran Associates, Inc., 2008.
- [2] A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991. [MR1126328](#)
- [3] P. Bellec. Optimal exponential bounds for aggregation of density estimators. *arXiv:1405.3907*, 2014. [MR3556772](#)
- [4] J. Bigot, R. B. Lirio, J.-M. Loubes, and L. M. Alvarez. Adaptive estimation of spectral densities via wavelet thresholding and information projection. *arXiv preprint arXiv:0912.2026*, 2009.
- [5] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004. [MR2061575](#)
- [6] R. C. Bradley. On positive spectral density functions. *Bernoulli*, 8(2):175–193, 2002. [MR1895889](#)
- [7] L. M. Brègman. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vychisl. Mat. i Mat. Fiz.*, 7:620–631, 1967. [MR0215617](#)
- [8] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 08 2007. [MR2351101](#)
- [9] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Fast adaptive estimation of log-additive exponential models in Kullback-Leibler divergence. *arXiv:1604.06304*, 2016.
- [10] O. Catoni. Universal aggregation rules with exact bias bounds. Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris. *Preprint*, 510, 1999.
- [11] C. Chang and D. Politis. Aggregation of spectral density estimators. *Statistics & Probability Letters*, 94:204–213, 2014. [MR3257381](#)
- [12] D. Dai, P. Rigollet, L. Xia, and T. Zhang. Aggregation of affine estimators. *Electron. J. Statist.*, 8(1):302–327, 2014. [MR3192554](#)
- [13] D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy  $Q$ -aggregation. *Ann. Statist.*, 40(3):1878–1905, 06 2012. [MR3015047](#)
- [14] A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 08 2012. [MR3059085](#)
- [15] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007. [MR2397581](#)
- [16] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1–2):39–61, 2008.
- [17] R. B. Davies. Asymptotic inference in stationary Gaussian time-series. *Advances in Appl. Probability*, 5:469–497, 1973. [MR0341699](#)
- [18] E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker’s guide to the fractional Sobolev spaces. *Bull. Sci. Math.*, 136(5):521–573, 2012. [MR2944369](#)

- [19] U. Grenander and G. Szegő. *Toeplitz forms and their applications*, volume 321. Univ of California Press, 1958. [MR0094840](#)
- [20] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 05 2000. [MR1792783](#)
- [21] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 10 2008. [MR2458184](#)
- [22] G. Lecué. Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.*, 7:971–981, 2006. [MR2274393](#)
- [23] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145(3–4):591–613, 2009. [MR2529440](#)
- [24] C. C. Moore. The degree of randomness in a stationary time series. *Ann. Math. Statist.*, 34:1253–1258, 1963. [MR0159405](#)
- [25] P. Rigollet. Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665, 04 2012. [MR2933661](#)
- [26] P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007. [MR2356821](#)
- [27] M. Rosenblatt. Remarks on a multivariate transformation. *Ann. Math. Statist.*, 23(3):470–472, 09 1952. [MR0049525](#)
- [28] A. B. Tsybakov. Optimal rates of aggregation. In B. Schölkopf and M. K. Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 303–313. Springer Berlin Heidelberg, 2003.
- [29] Z. Wang, S. Paterlini, F. Gao, and Y. Yang. Adaptive minimax regression estimation over sparse  $\ell_q$ -hulls. *J. Mach. Learn. Res.*, 15:1675–1711, 2014. [MR3225246](#)
- [30] M. Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 02 2003. [MR1962506](#)
- [31] Y. Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74(1):135–161, 2000. [MR1790617](#)
- [32] Y. Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 02 2000. [MR1762904](#)
- [33] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 02 2004. [MR2044592](#)