Optimal kernel selection for density estimation

Matthieu Lerasle, Nelo Magalhães, Patricia Reynaud-Bouret

▶ **To cite this version:**

Matthieu Lerasle, Nelo Magalhães, Patricia Reynaud-Bouret. Optimal kernel selection for density estimation. High Dimensional Probability VII : The Cargese Volume, 71, Birkhauser, pp.425-460, 2016, Prog. Probab, 978-3-319-40519-3_19. 10.1007/978-3-319-40519-3_19 . hal-01224097

HAL Id: hal-01224097
https://hal.science/hal-01224097v1

Submitted on 6 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Optimal kernel selection for density estimation

M. Lerasle, N. Magalhães and P. Reynaud-Bouret

Abstract. We provide new general kernel selection rules thanks to penalized least-squares criteria. We derive optimal oracle inequalities using adequate concentration tools. We also investigate the problem of minimal penalty as described in [BM07].

Keywords. density estimation, kernel estimators, optimal penalty, minimal penalty, oracle inequalities.

1. Introduction

Concentration inequalities are central in the analysis of adaptive nonparametric statistics. They lead to sharp penalized criteria for model selection [Mas07], to select bandwidths and even approximation kernels for Parzen's estimators in high dimension [GL11], to aggregate estimators [RT07] and to properly calibrate thresholds [DJKP96].

In the present work, we are interested in the selection of a general kernel estimator based on a least-squares density estimation approach. The problem has been considered in L^1-loss by Devroye and Lugosi [DL01]. Other methods combining log-likelihood and roughness/smoothness penalties have also been proposed in [EL99b, EL99a, EL01]. However these estimators are usually quite difficult to compute in practice. We propose here to minimize penalized least-squares criteria and obtain from them more easily computable estimators. Sharp concentration inequalities for U-statistics [GLZ00, Ada06, HRB03] control the variance term of the kernel estimators, whose asymptotic behavior has been precisely described, for instance in [MS11, MS15, DO13]. We derive from these bounds (see Proposition 4.1) a penalization method to select a kernel which satisfies an asymptotically optimal oracle inequality, i.e. with leading constant asymptotically equal to 1.

This research was partly supported by the french Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration).

In the spirit of [GN09], we use an extended definition of kernels that allows to deal simultaneously with classical collections of estimators as projection estimators, weighted projection estimators, or Parzen's estimators. This method can be used for example to select an optimal model in model selection (in accordance with [Mas07]) or to select an optimal bandwidth together with an optimal approximation kernel among a finite collection of Parzen's estimators. In this sense, our method deals, in particular, with the same problem as that of Goldenshluger and Lepski [GL11] and we establish in this framework that a leading constant 1 in the oracle inequality is indeed possible.

Another main consequence of concentration inequalities is to prove the existence of a minimal level of penalty, under which no oracle inequalities can hold. Birgé and Massart shed light on this phenomenon in a Gaussian setting for model selection [BM07]. Moreover in this setting, they prove that the optimal penalty is twice the minimal one. In addition, there is a sharp phase transition in the dimension of the selected models leading to an estimate of the optimal penalty in their case (which is known up to a multiplicative constant). Indeed, starting from the idea that in many models the optimal penalty is twice the minimal one (this is the slope heuristic), Arlot and Massart [AM09] propose to detect the minimal penalty by the phase transition and to apply the "$\times 2$" rule (this is the slope algorithm). They prove that this algorithm works at least in some regression settings.

In the present work, we also show that minimal penalties exist in the density estimation setting. In particular, we exhibit a sharp "phase transition" of the behavior of the selected estimator around this minimal penalty. The analysis of this last result is not standard however. First, the "slope heuristic" of [BM07] only holds in particular cases as the selection of projection estimators, see also [Ler12]. As in the selection of a linear estimator in a regression setting [1], the heuristic can sometimes be corrected: for example for the selection of a bandwidth when the approximation kernel is fixed. In general since there is no simple relation between the minimal penalty and the optimal one, the slope algorithm of [AM09] shall only be used with care for kernel selection. Surprisingly our work reveals that the minimal penalty can be negative. In this case, minimizing an unpenalized criterion leads to oracle estimators. To our knowledge, such phenomenon has only been noticed previously in a very particular classification setting [FT06]. We illustrate all of these different behaviors by means of a simulation study.

In Section 2, after fixing the main notation, providing some examples and defining the framework, we explain our goal, describe what we mean by an *oracle inequality* and state the exponential inequalities that we shall need. Then we derive optimal penalties in Section 3 and study the problem of minimal penalties in Section 4. All of these results are illustrated for our three main examples : projection kernels, approximation kernels and weighted projection kernels. In Section 5, some simulations are performed in the approximation kernel case. The main proofs are detailed in Section 6 and technical results are discussed in the appendix.

2. Kernel selection for least-squares density estimation

2.1. Setting

Let $X, Y, X_1, \ldots, X_n$ denote i.i.d. random variables taking values in the measurable space $(\mathbb{X}, \mathcal{X}, \mu)$, with common distribution P. Assume P has density s with respect to μ and s is uniformly bounded. Hence, s belongs to L^2, where, for any $p \geq 1$,

$$L^p := \left\{ t : \mathbb{X} \to \mathbb{R}, \text{ s.t. } \|t\|_p^p := \int |t|^p \, d\mu < \infty \right\} .$$

Moreover, $\|\cdot\| = \|\cdot\|_2$ and $\langle \cdot, \cdot \rangle$ denote respectively the L^2-norm and the associated inner product and $\|\cdot\|_\infty$ is the supremum norm. We systematically use $x \vee y$ and $x \wedge y$ for $\max(x, y)$ and $\min(x, y)$ respectively, and denote $|A|$ the cardinality of the set A. Recall that $x_+ = x \vee 0$ and, for any $y \in \mathbb{R}^+$, $\lfloor y \rfloor = \sup\{n \in \mathbb{N} \text{ s.t. } n \leq y\}$.

Let $\{ k \}_{k \in \mathcal{K}}$ denote a collection of symmetric functions $k : \mathbb{X}^2 \to \mathbb{R}$ indexed by some given finite set $\mathcal{K}$ such that

$$\sup_{x \in \mathbb{X}} \int_{\mathbb{X}} k(x, y)^2 d\mu(y) \ \vee \sup_{(x,y) \in \mathbb{X}^2} |k(x, y)| < \infty \ .$$

A function k satisfying these assumptions is called a *kernel*, in the sequel. A kernel k is associated with an estimator $\widehat{s}_k$ of s defined for any $x \in \mathbb{X}$ by

$$\widehat{s}_k(x) := \frac{1}{n} \sum_{i=1}^{n} k(X_i, x) \ .$$

Our aim is to select a "good" $\widehat{s}_{\widehat{k}}$ in the family $\{\widehat{s}_k, k \in \mathcal{K}\}$. Our results are expressed in terms of a constant $\Gamma \geq 1$ such that for all $k \in \mathcal{K}$,

$$\sup_{x \in \mathbb{X}} \int_{\mathbb{X}} k(x, y)^2 d\mu(y) \ \vee \sup_{(x,y) \in \mathbb{X}^2} |k(x, y)| \leq \Gamma n \ . \tag{1}$$

This condition plays the same role as $\int |k(x, y)| s(y) d\mu(y) < \infty$, the milder condition used in [DL01] when working with L^1-losses. Before describing the method, let us give three examples of such estimators that are used for density estimation, and see how they can naturally be associated to some kernels. Section A of the appendix gives the computations leading to the corresponding Γ's.

Example 1: Projection estimators. Projection estimators are among the most classical density estimators. Given a linear subspace $S \subset L^2$, the projection estimator on S is defined by

$$\widehat{s}_S = \arg\min_{t \in S} \left\{ \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} t(X_i) \right\} \ .$$

Let $\mathcal{S}$ be a family of linear subspaces S of L^2. For any $S \in \mathcal{S}$, let $(\varphi_\ell)_{\ell \in \mathcal{I}_S}$ denote an orthonormal basis of S. The projection estimator $\widehat{s}_S$ can be computed and is equal to

$$\widehat{s}_S = \sum_{\ell \in \mathcal{I}_S} \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_\ell(X_i) \right) \varphi_\ell \ .$$

It is therefore easy to see that it is the estimator associated to the *projection kernel* k_S defined for any x and y in $\mathbb{X}$ by

$$k_S(x,y) := \sum_{\ell \in \mathcal{I}_S} \varphi_\ell(x)\varphi_\ell(y) \ .$$

Notice that k_S actually depends on the basis $(\varphi_\ell)_{\ell\in\mathcal{I}_S}$ even if $\widehat{s}_S$ does not. In the sequel, we always assume that some orthonormal basis $(\varphi_\ell)_{\ell\in\mathcal{I}_S}$ is given with S. Given a finite collection $\mathcal{S}$ of linear subspaces of L^2, one can choose the following constant Γ in (1) for the collection $(k_S)_{S\in\mathcal{S}}$

$$\Gamma = 1 \vee \frac{1}{n} \sup_{S\in\mathcal{S}} \sup_{f\in S, \|f\|=1} \|f\|_\infty^2 \ . \tag{2}$$

Example 2: Parzen's estimators. Given a bounded symmetric integrable function $K : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} K(u)du = 1$, $K(0) > 0$ and a bandwidth $h > 0$, the Parzen estimator is defined by

$$\forall x \in \mathbb{R}, \quad \widehat{s}_{K,h}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \ .$$

It can also naturally be seen as a kernel estimator, associated to the function $k_{K,h}$ defined for any x and y in $\mathbb{R}$ by

$$k_{K,h}(x,y) := \frac{1}{h} K\left(\frac{x-y}{h}\right) \ .$$

We shall call the function $k_{K,h}$ an approximation or Parzen kernel.
Given a finite collection of pairs $(K,h) \in \mathcal{H}$, one can choose $\Gamma = 1$ in (1) if,

$$h \geq \frac{\|K\|_\infty \|K\|_1}{n} \quad \text{for any } (K,h) \in \mathcal{H} \ . \tag{3}$$

Example 3: Weighted projection estimators. Let $(\varphi_i)_{i=1,\dots,p}$ denote an orthonormal system in L^2 and let $w = (w_i)_{i=1,\dots,p}$ denote real numbers in $[0,1]$. The associated weighted kernel projection estimator of s is defined by

$$\widehat{s}_w = \sum_{i=1}^p w_i \left(\frac{1}{n}\sum_{j=1}^n \varphi_i(X_j)\right)\varphi_i \ .$$

These estimators are used to derive very sharp adaptive results. In particular, Pinsker's estimators are weighted kernel projection estimators (see for example [Rig06]). When $w \in \{0,1\}^p$, we recover a classical projection estimator. A weighted projection estimator is associated to the *weighted projection kernel* defined for any x and y in $\mathbb{X}$ by

$$k_w(x,y) := \sum_{i=1}^p w_i \varphi_i(x)\varphi_i(y) \ .$$

Given any finite collection $\mathcal{W}$ of weights, one can choose in (1)

$$\Gamma = 1 \vee \left(\frac{1}{n} \sup_{x\in\mathbb{X}} \sum_{i=1}^p \varphi_i(x)^2\right) \ . \tag{4}$$

2.2. Oracle inequalities and penalized criterion

The goal is to estimate s in the best possible way using a finite collection of kernel estimators $(\widehat{s}_k)_{k\in\mathcal{K}}$. In other words, the purpose is to select among $(\widehat{s}_k)_{k\in\mathcal{K}}$ an estimator $\widehat{s}_{\widehat{k}}$ from the data such that $\left\|\widehat{s}_{\widehat{k}} - s\right\|^2$ is as close as possible to $\inf_{k\in\mathcal{K}} \|\widehat{s}_k - s\|^2$. More precisely our aim is to select $\widehat{k}$ such that, with high probability,

$$\left\|\widehat{s}_{\widehat{k}} - s\right\|^2 \leq C_n \inf_{k\in\mathcal{K}} \|\widehat{s}_k - s\|^2 + R_n \ , \tag{5}$$

where $C_n \geq 1$ is the leading constant and $R_n > 0$ is usually a remaining term. In this case, $\widehat{s}_{\widehat{k}}$ is said to satisfy an *oracle inequality*, as long as R_n is small compared to $\inf_{k\in\mathcal{K}} \|\widehat{s}_k - s\|^2$ and C_n is a bounded sequence. This means that the selected estimator does as well as the best estimator in the family up to some multiplicative constant. The best case one can expect is to get C_n close to 1. This is why, when $C_n \to_{n\to\infty} 1$, the corresponding oracle inequality is called *asymptotically optimal*. To do so, we study minimizers of *penalized least-squares criteria*. Note that in our three examples choosing $\widehat{k} \in \mathcal{K}$ amounts to choosing the smoothing parameter, that is respectively to choosing $\widehat{S} \in \mathcal{S}$, $(\widehat{K}, \widehat{h}) \in \mathcal{H}$ or $\widehat{w} \in \mathcal{W}$.

Let P_n denote the empirical measure, that is, for any real valued function t,

$$P_n(t) := \frac{1}{n}\sum_{i=1}^{n} t(X_i) \ .$$

For any $t \in L^2$, let also $P(t) := \int_{\mathbb{X}} t(x)s(x)d\mu(x)$.

The *least-squares contrast* is defined, for any $t \in L^2$, by

$$\gamma(t) := \|t\|^2 - 2t \ .$$

Then for any given function $\text{pen} : \mathcal{K} \to \mathbb{R}$, the *least-squares penalized criterion* is defined by

$$\mathcal{C}_{\text{pen}}(k) := P_n\gamma(\widehat{s}_k) + \text{pen}(k) \ . \tag{6}$$

Finally the selected $\widehat{k} \in \mathcal{K}$ is given by any minimizer of $\mathcal{C}_{\text{pen}}(k)$, that is,

$$\widehat{k} \in \arg\min_{k\in\mathcal{K}} \{\, \mathcal{C}_{\text{pen}}(k) \,\} \ . \tag{7}$$

As $P\gamma(t) = \|t - s\|^2 - \|s\|^2$, it is equivalent to minimize $\|\widehat{s}_k - s\|^2$ or $P\gamma(\widehat{s}_k)$. As our goal is to select $\widehat{s}_{\widehat{k}}$ satisfying an oracle inequality, an ideal penalty pen_{id} should satisfy $\mathcal{C}_{\text{pen}_{\text{id}}}(k) = P\gamma(\widehat{s}_k)$, i.e. criterion (6) with

$$\text{pen}_{\text{id}}(k) := (P - P_n)\gamma(\widehat{s}_k) = 2(P_n - P)(\widehat{s}_k) \ .$$

To identify the main quantities of interest, let us introduce some notation and develop $\text{pen}_{\text{id}}(k)$. For all $k \in \mathcal{K}$, let

$$s_k(x) := \int_{\mathbb{X}} k(y,x)s(y)d\mu(y) = \mathbb{E}\left[k(X,x)\right], \qquad \forall x \in \mathbb{X} \ ,$$

and

$$U_k := \sum_{i\neq j=1}^{n} \left(k(X_i,X_j) - s_k(X_i) - s_k(X_j) + \mathbb{E}\left[k(X,Y)\right]\right) \ .$$

Because those quantities are fundamental in the sequel, let us also define $\Theta_k(x) = A_k(x,x)$ where for $(x,y) \in \mathbb{X}^2$

$$A_k(x,y) := \int_{\mathbb{X}} k(x,z)k(z,y)d\mu(z) \ . \tag{8}$$

Denoting

$$\text{for all } x \in \mathbb{X}, \quad \chi_k(x) = k(x,x) \ ,$$

the ideal penalty is then equal to

$$\begin{aligned}&\text{pen}_{\text{id}}(k) = 2(P_n - P)(\widehat{s}_k - s_k) + 2(P_n - P)s_k \\ &= 2\left(\frac{P\chi_k - Ps_k}{n} + \frac{(P_n - P)\chi_k}{n} + \frac{U_k}{n^2} + \left(1 - \frac{2}{n}\right)(P_n - P)s_k \right) \ . \end{aligned} \tag{9}$$

The main point is that by using concentration inequalities, we obtain:

$$\text{pen}_{\text{id}}(k) \simeq 2\left(\frac{P\chi_k - Ps_k}{n} \right) \ .$$

The term Ps_k/n depends on s which is unknown. Fortunately, it can be easily controlled as detailed in the sequel. Therefore one can hope that the choice

$$\text{pen}(k) = 2\frac{P\chi_k}{n}$$

is convenient. In general, this choice still depends on the unknown density s but it can be easily estimated in a data-driven way by

$$\text{pen}(k) = 2\frac{P_n\chi_k}{n} \ .$$

The goal of Section 3 is to prove this heuristic and to show that $2P\chi_k/n$ and $2P_n\chi_k/n$ are optimal choices for the penalty, that is, they lead to an asymptotically optimal oracle inequality.

2.3. Concentration tools

To derive sharp oracle inequalities, we only need two fundamental concentration tools, namely a weak Bernstein's inequality and the concentration bounds for degenerate U-statistics of order two. We cite them here under their most suitable form for our purpose.

A weak Bernstein's inequality.

Proposition 2.1. *For any bounded real valued function f and any $X_1, \dots, X_n$ i.i.d. with distribution P, for any $u > 0$,*

$$\mathbb{P}\left((P_n - P)f \geq \sqrt{\frac{2P\left(f^2\right)u}{n}} + \frac{\|f\|_\infty u}{3n} \right) \leq \exp(-u) \ .$$

The proof is straightforward and can be derived from either Bennett's or Bernstein's inequality [BLM13].

Concentration of degenerate U-statistics of order 2.

Proposition 2.2. *Let $X, X_1, \dots X_n$ be i.i.d. random variables defined on a Polish space $\mathbb{X}$ equipped with its Borel σ-algebra and let $(f_{i,j})_{1\le i\neq j\le n}$ denote bounded real valued symmetric measurable functions defined on $\mathbb{X}^2$, such that for any $i \neq j$, $f_{i,j} = f_{j,i}$ and*

$$\forall\ i,j\ s.t.\ 1 \le i \neq j \le n, \qquad \mathbb{E}\left[f_{i,j}(x,X)\right] = 0 \qquad \text{for a.e. } x \text{ in } \mathbb{X}\ . \quad (10)$$

Let U be the following totally degenerate U-statistic of order 2,

$$U = \sum_{1\le i\neq j\le n} f_{i,j}(X_i, X_j)\ .$$

Let A be an upper bound of $|f_{i,j}(x,y)|$ for any i,j,x,y and

$$B^2 = \max\left(\sup_{i,x\in\mathbb{X}} \sum_{j=1}^{i-1} \mathbb{E}\left[f_{i,j}(x,X_j)^2\right], \sup_{j,t\in\mathbb{X}} \sum_{i=j+1}^{n} \mathbb{E}\left[f_{i,j}(X_i,t)^2\right] \right)$$

$$C^2 = \sum_{1\le i\neq j\le n} \mathbb{E}\left[f_{i,j}(X_i,X_j)^2\right]$$

$$D = \sup_{(a,b)\in\mathcal{A}} \mathbb{E}\left[\sum_{1\le i<j\le n} f_{i,j}(X_i,X_j)a_i(X_i)b_j(X_j) \right]\ ,$$

where $\mathcal{A} = \left\{ (a,b),\ \ s.t.\ \mathbb{E}\left[\sum_{i=1}^{n-1} a_i(X_i)^2\right] \le 1,\ \mathbb{E}\left[\sum_{j=2}^{n} b_j(X_j)^2\right] \le 1 \right\}$.
Then there exists some absolute constant $\kappa > 0$ such that for any $u > 0$, with probability larger than $1 - 2.7e^{-u}$,

$$U \le \kappa\left(C\sqrt{u} + Du + Bu^{3/2} + Au^2\right)\ .$$

The present result is a simplification of Theorem 3.4.8 in [GN15], which provides explicit constants for any variables defined on a Polish space. It is mainly inspired by [HRB03], where the result therein has been stated only for real variables. This inequality actually dates back to Giné, Latala and Zinn [GLZ00]. This result has been further generalized by Adamczak to U-statistics of any order [Ada06], though the constants are not explicit.

3. Optimal penalties for kernel selection

The main aim of this section is to show that $2P\chi_k/n$ is a theoretical optimal penalty for kernel selection, which means that if $\mathrm{pen}(k)$ is close to $2P\chi_k/n$, the selected kernel $\widehat{k}$ satisfies an asymptotically optimal oracle inequality.

3.1. Main assumptions

To express our results in a simple form, a positive constant Υ is assumed to control, for any k and k' in $\mathcal{K}$, all the following quantities.

$$(\Gamma(1+\|s\|_\infty)) \vee \sup_{k\in\mathcal{K}} \|s_k\|^2 \leq \Upsilon \ , \tag{11}$$

$$P\left(\chi_k^2\right) \leq \Upsilon n P\Theta_k \ , \tag{12}$$

$$\|s_k - s_{k'}\|_\infty \leq \Upsilon \vee \sqrt{\Upsilon n}\, \|s_k - s_{k'}\| \ , \tag{13}$$

$$\mathbb{E}\left[A_k(X,Y)^2\right] \leq \Upsilon P\Theta_k \ , \tag{14}$$

$$\sup_{x\in\mathbb{X}} \mathbb{E}\left[A_k(X,x)^2\right] \leq \Upsilon n \ , \tag{15}$$

$$v_k^2 := \sup_{t\in\mathbb{B}_k} Pt^2 \leq \Upsilon \vee \sqrt{\Upsilon P\Theta_k} \ , \tag{16}$$

where $\mathbb{B}_k$ is the set of functions t that can be written $t(x) = \int a(z)k(z,x)d\mu(z)$ for some $a \in L^2$ with $\|a\| \leq 1$.
These assumptions may seem very intricate. They are actually fulfilled by our three main examples under very mild conditions (see Section 3.3).

3.2. The optimal penalty theorem

In the sequel, $\square$ denotes a positive absolute constant whose value may change from line to line and if there are indices such as $\square_\theta$, it means that this is a positive function of θ and only θ whose value may change from line to line.

Theorem 3.1. *If Assumptions* (11), (12), (13), (14) (15), (16) *hold, then, for any $x \geq 1$, with probability larger than $1 - \square|\mathcal{K}|^2 e^{-x}$, for any $\theta \in (0,1)$, any minimizer $\widehat{k}$ of the penalized criterion* (6) *satisfies the following inequality*

$$\forall k \in \mathcal{K}, \qquad (1-4\theta)\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 \leq (1+4\theta)\|s - \widehat{s}_k\|^2 + \left(\mathrm{pen}(k) - 2\frac{P\chi_k}{n}\right) - \left(\mathrm{pen}\left(\widehat{k}\right) - 2\frac{P\chi_{\widehat{k}}}{n}\right) + \square\frac{\Upsilon x^2}{\theta n} \ . \tag{17}$$

Assume moreover that there exists $C > 0$, $\delta' \geq \delta > 0$ and $r \geq 0$ such that for any $x \geq 1$, with probability larger than $1 - Ce^{-x}$, for any $k \in \mathcal{K}$,

$$(\delta - 1)\frac{P\Theta_k}{n} - \square r\frac{\Upsilon x^2}{n} \leq \mathrm{pen}(k) - \frac{2P\chi_k}{n} \leq (\delta' - 1)\frac{P\Theta_k}{n} + \square r\frac{\Upsilon x^2}{n} \ . \tag{18}$$

Then for all $\theta \in (0,1)$ and all $x \geq 1$, the following holds with probability at least $1 - \square(C + |\mathcal{K}|^2)e^{-x}$,

$$\frac{(\delta \wedge 1) - 5\theta}{(\delta' \vee 1) + (4+\delta')\theta}\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 \leq \inf_{k\in\mathcal{K}} \|s - \widehat{s}_k\|^2 + \square\left(r + \frac{1}{\theta^3}\right)\frac{\Upsilon x^2}{n} \ .$$

Let us make some remarks.

- First, this is an oracle inequality (see (5)) with leading constant C_n and remaining term R_n given by

$$C_n = \frac{(\delta' \vee 1) + (4+\delta')\theta}{(\delta \wedge 1) - 5\theta} \quad \text{and} \quad R_n = \square C_n(r + \theta^{-3})\frac{\Upsilon x^2}{n} \ ,$$

as long as
- θ is small enough for C_n to be positive,
- x is large enough for the probability to be large and
- n is large enough for R_n to be negligible.

Typically, $r, \delta, \delta', \theta$ and Υ are bounded w.r.t. n and x has to be of the order of $\log(|\mathcal{K}| \vee n)$ for the remainder to be negligible. In particular, $\mathcal{K}$ may grow with n as long as (i) $\log(|\mathcal{K}| \vee n)^2$ remains negligible with respect to n and (ii) Υ does not depend on n.

- If $\text{pen}(k) = 2P\chi_k/n$, that is if $\delta = \delta' = 1$ and $r = C = 0$ in (18), the estimator $\widehat{s}_{\widehat{k}}$ satisfies an asymptotically optimal oracle inequality i.e. $C_n \to_{n\to\infty} 1$ since θ can be chosen as close to 0 as desired. Take for instance, $\theta = (\log n)^{-1}$.
- In general $P\chi_k$ depends on the unknown s and this last penalty cannot be used in practice. Fortunately, its empirical counterpart $\text{pen}(k) = 2P_n\chi_k/n$ satisfies (18) with $\delta = 1 - \theta$, $\delta' = 1 + \theta$, $r = 1/\theta$ and $C = 2|\mathcal{K}|$ for any $\theta \in (0, 1)$ and in particular $\theta = (\log n)^{-1}$ (see (34) in Proposition B.1). Hence, the estimator $\widehat{s}_{\widehat{k}}$ selected with this choice of penalty also satisfies an asymptotically optimal oracle inequality, by the same argument.
- Finally, we only get an oracle inequality when $\delta > 0$, that is when $\text{pen}(k)$ is larger than $(2P\chi_k - P\Theta_k)/n$ up to some residual term. We discuss the necessity of this condition in Section 4.

3.3. Main examples

This section shows that Theorem 3.1 can be applied in the examples. In addition, it provides the computation of $2P\chi_k/n$ in some specific cases of special interest.

Example 1 (continued).

Proposition 3.2. *Let $\{k_S, S \in \mathcal{S}\}$ be a collection of projection kernels. Assumptions* (11), (12), (14), (15) *and* (16) *hold for any $\Upsilon \geq \Gamma(1 + \|s\|_\infty)$, where Γ is given by* (2)*. In addition, Assumption* (13) *is satisfied under either of the following classical assumptions (see* [Mas07, Chapter 7]*):*

$$\forall S, S' \in \mathcal{S}, \qquad \text{either } S \subset S' \text{ or } S' \subset S \ , \tag{19}$$

or

$$\forall S \in \mathcal{S}, \qquad \|s_{k_S}\|_\infty \leq \frac{\Upsilon}{2} \ . \tag{20}$$

These particular projection kernels satisfy for all $(x, y) \in \mathbb{X}^2$

$$\begin{aligned} A_{k_S}(x, y) &= \int_{\mathbb{X}} k_S(x, z) k_S(y, z) d\mu(z) \\ &= \sum_{(i,j)\in\mathcal{I}_S^2} \varphi_i(x)\varphi_j(y) \int_{\mathbb{X}} \varphi_i(z)\varphi_j(z) d\mu(z) = k_S(x, y) \ . \end{aligned}$$

In particular, $\Theta_{k_S} = \chi_{k_S} = \sum_{i\in\mathcal{I}_S} \varphi_i^2$ and $2P\chi_{k_S} - P\Theta_{k_S} = P\chi_{k_S}$.
Moreover, it appears that the function Θ_{k_S} is constant in some linear spaces S of interest (see [Ler12] for more details). Let us mention one particular

case studied further on in the sequel. Suppose $\mathcal{S}$ is a collection of regular histogram spaces S on $\mathbb{X}$, that is, any $S \in \mathcal{S}$ is a space of piecewise constant functions on a partition $\mathcal{I}_S$ of $\mathbb{X}$ such that $\mu(i) = 1/D_S$ for any i in $\mathcal{I}_S$. Assumption (20) is satisfied for this collection as soon as $\Upsilon \geq 2\|s\|_\infty$. The family $(\varphi_i)_{i\in\mathcal{I}_S}$, where $\varphi_i = \sqrt{D_S}\mathbf{1}_i$ is an orthonormal basis of S and

$$\chi_{k_S} = \sum_{i\in\mathcal{I}_S} \varphi_i^2 = D_S \ .$$

Hence, $P\chi_{k_S} = D_S$ and $2D_S/n$ can actually be used as a penalty to ensure that the selected estimator satisfies an asymptotically optimal oracle inequality. Moreover, in this example it is actually necessary to choose a penalty larger than D_S/n to get an oracle inequality (see [Ler12] or Section 4 for more details).

Example 2 (continued).

Proposition 3.3. *Let $\{k_{K,h}, (K,h) \in \mathcal{H}\}$ be a collection of approximation kernels. Assumptions* (11), (12), (13), (14), (15) *and* (16) *hold with $\Gamma = 1$, for any*

$$\Upsilon \geq \max_K \left\{ \frac{K(0)}{\|K\|^2} \vee \left(1 + 2\|s\|_\infty \|K\|_1^2\right) \right\} \ ,$$

as soon as (3) *is satisfied.*

These approximation kernels satisfy, for all $x \in \mathbb{R}$,

$$\chi_{k_{K,h}}(x) = k_{K,h}(x,x) = \frac{K(0)}{h} \ ,$$
$$\Theta_{k_{K,h}}(x) = A_{k_{K,h}}(x,x) = \frac{1}{h^2}\int_{\mathbb{R}} K\left(\frac{x-y}{h}\right)^2 dy = \frac{\|K\|^2}{h} \ .$$

Therefore, the optimal penalty $2P\chi_{k_{K,h}}/n = 2K(0)/(nh)$ can be computed in practice and yields an asymptotically optimal selection criterion. Surprisingly, the lower bound $2P\chi_{k_{K,h}}/n - P\Theta_{k_{K,h}}/n = (2K(0) - \|K\|^2)/(nh)$ can be negative if $\|K\|^2 > 2K(0)$. In this case, a minimizer of (6) satisfies an oracle inequality, even if this criterion is not penalized. This remarkable fact is illustrated in the simulation study in Section 5.

Example 3 (continued).

Proposition 3.4. *Let $\{k_w, w \in \mathcal{W}\}$ be a collection of weighted projection kernels. Assumption* (11) *is valid for $\Upsilon \geq \Gamma(1 + \|s\|_\infty)$, where Γ is given by* (4). *Moreover* (11) *and* (1) *imply* (12), (13), (14), (15) *and* (16).

For these weighted projection kernels, for all $x \in \mathbb{X}$

$$\chi_{k_w}(x) = \sum_{i=1}^p w_i \varphi_i(x)^2, \qquad \text{hence} \qquad P\chi_{k_w} = \sum_{i=1}^p w_i P\varphi_i^2 \qquad \text{and}$$
$$\Theta_{k_w}(x) = \sum_{i,j=1}^p w_i w_j \varphi_i\varphi_j \int_{\mathbb{X}} \varphi_i(x)\varphi_j(x)d\mu(x) = \sum_{i=1}^p w_i^2 \varphi_i(x)^2 \leq \chi_{k_w}(x) \ .$$

In this case, the optimal penalty $2P\chi_{k_w}/n$ has to be estimated in general. However, in the following example it can still be directly computed.
Let $\mathbb{X} = [0,1]$, let μ be the Lebesgue measure. Let $\varphi_0 \equiv 1$ and, for any $j \geq 1$,

$$\varphi_{2j-1}(x) = \sqrt{2}\cos(2\pi jx), \qquad \varphi_{2j}(x) = \sqrt{2}\sin(2\pi jx) \ .$$

Consider some odd p and a family of weights $\mathcal{W} = \{w_i, i = 0, \ldots, p\}$ such that, for any $w \in \mathcal{W}$ and any $i = 1, \ldots, p/2$, $w_{2i-1} = w_{2i} = \tau_i$. In this case, the values of the functions of interest do not depend on x

$$\chi_{k_w} = w_0 + \sum_{j=1}^{p/2} \tau_j, \qquad \Theta_{k_w} = w_0^2 + \sum_{j=1}^{p/2} \tau_j^2 \ .$$

In particular, this family includes Pinsker's and Tikhonov's weights.

4. Minimal penalties for kernel selection

The purpose of this section is to see whether the lower bound $\mathrm{pen}_{\min}(k) := (2P\chi_k - P\Theta_k)/n$ is sharp in Theorem 3.1. To do so we first need the following result which links $\|s - \widehat{s}_k\|$ to deterministic quantities, thanks to concentration tools.

4.1. Bias-Variance decomposition with high probability

Proposition 4.1. *Assume $\{k\}_{k\in\mathcal{K}}$ is a finite collection of kernels satisfying Assumptions (11), (12), (13), (14) (15) and (16). For all $x > 1$, for all η in $(0,1]$, with probability larger than $1 - \square|\mathcal{K}|e^{-x}$*

$$\|s_k - \widehat{s}_k\|^2 \leq (1+\eta)\frac{P\Theta_k}{n} + \square\frac{\Upsilon x^2}{\eta n} \ ,$$

$$\frac{P\Theta_k}{n} \leq (1+\eta)\|s_k - \widehat{s}_k\|^2 + \square\frac{\Upsilon x^2}{\eta n} \ .$$

Moreover, for all $x > 1$ and for all η in $(0,1)$, with probability larger than $1 - \square|\mathcal{K}|e^{-x}$, for all $k \in \mathcal{K}$, each of the following inequalities hold

$$\|s - \widehat{s}_k\|^2 \leq (1+\eta)\left(\|s - s_k\|^2 + \frac{P\Theta_k}{n}\right) + \square\frac{\Upsilon x^2}{\eta^3 n} \ ,$$

$$\|s - s_k\|^2 + \frac{P\Theta_k}{n} \leq (1+\eta)\|s - \widehat{s}_k\|^2 + \square\frac{\Upsilon x^2}{\eta^3 n} \ .$$

This means that not only in expectation but also with high probability can the term $\|s - \widehat{s}_k\|^2$ be decomposed in a bias term $\|s - s_k\|^2$ and a "variance" term $P\Theta_k/n$. The bias term measures the capacity of the kernel k to approximate s whereas $P\Theta_k/n$ is the price to pay for replacing s_k by its empirical version $\widehat{s}_k$. In this sense, $P\Theta_k/n$ measures the complexity of the kernel k in a way which is completely adapted to our problem of density estimation. Even if it does not seem like a natural measure of complexity at first glance, note that in the previous examples, it is indeed always linked to a natural complexity. When dealing with regular histograms defined on $[0,1]$, $P\Theta_{k_S}$ is the dimension of the considered space S, whereas for approximation kernels $P\Theta_{k_{K,h}}$ is proportional to the inverse of the considered bandwidth h.

4.2. Some general results about the minimal penalty

In this section, we assume that we are in the asymptotic regime where the number of observations $n \to \infty$. In particular, the asymptotic notations refers to this regime.

From now on, the family $\mathcal{K} = \mathcal{K}_n$ may depend on n as long as both Γ and Υ remain absolute constants that do not depend on it. Indeed, on the previous examples, this seems a reasonable regime. Since $\mathcal{K}_n$ now depends on n, our selected $\widehat{k} = \hat{k}_n$ also depends on n.

To prove that the lower bound $\mathrm{pen}_{\min}(k)$ is sharp, we need to show that the estimator chosen by minimizing (6) with a penalty smaller than $\mathrm{pen}_{\min}$ does not satisfy an oracle inequality. This is only possible if the $\|s - \widehat{s}_k\|^2$'s are not of the same order and if they are larger than the remaining term $\square(r+\theta^{-3})\Upsilon x^2/n$. From an asymptotic point of view, we rewrite this thanks to Proposition 4.1 as for all $n \geq 1$, there exist $k_{0,n}$ and $k_{1,n}$ in $\mathcal{K}_n$ such that

$$\left\|s - s_{k_{1,n}}\right\|^2 + \frac{P\Theta_{k_{1,n}}}{n} \gg \left\|s - s_{k_{0,n}}\right\|^2 + \frac{P\Theta_{k_{0,n}}}{n} \gg \square\left(r + \frac{1}{\theta^3}\right)\frac{\Upsilon x^2}{n} \quad , \tag{21}$$

where $a_n \gg b_n$ means that $b_n/a_n \to_{n\to\infty} 0$. More explicitly, denoting by $\mathrm{o}(1)$ a sequence only depending on n and tending to 0 as n tends to infinity and whose value may change from line to line, one assumes that there exists c_s and c_R positive constants such that for all $n \geq 1$, there exist $k_{0,n}$ and $k_{1,n}$ in $\mathcal{K}_n$ such that

$$\left\|s - s_{k_{0,n}}\right\|^2 + \frac{P\Theta_{k_{0,n}}}{n} \leq c_s\ \mathrm{o}(1)\left(\left\|s - s_{k_{1,n}}\right\|^2 + \frac{P\Theta_{k_{1,n}}}{n}\right) \tag{22}$$

$$\frac{(\log(|\mathcal{K}_n| \vee n))^3}{n} \leq c_R\ \mathrm{o}(1)\left(\left\|s - s_{k_{0,n}}\right\|^2 + \frac{P\Theta_{k_{0,n}}}{n}\right) \quad . \tag{23}$$

We put a log-cube factor in the remaining term to allow some choices of $\theta = \theta_n \to_{n\to\infty} 0$ and $r = r_n \to_{n\to\infty} +\infty$.

But (22) and (23) (or (21)) are not sufficient. Indeed, the following result explains what happens when the bias terms are always the leading terms.

Corollary 4.2. *Let $(\mathcal{K}_n)_{n\geq 1}$ be a sequence of finite collections of kernels k satisfying Assumptions* (11), (12), (13), (14) (15), (16) *for a positive constant Υ independent of n and such that*

$$\frac{1}{n} = c_b\ \mathrm{o}(1) \inf_{k\in\mathcal{K}_n} \frac{\|s - s_k\|^2}{P\Theta_k} \quad , \tag{24}$$

for some positive constant c_b.

Assume that there exist real numbers of any sign $\delta' \geq \delta$ and a sequence $(r_n)_{n\geq 1}$ of nonnegative real numbers such that, for all $n \geq 1$, with probability larger than $1 - \square/n^2$, for all $k \in \mathcal{K}_n$,

$$\delta\frac{P\Theta_k}{n} - \square_{\delta,\delta',\Upsilon}\frac{r_n \log(n \vee |\mathcal{K}_n|)^2}{n}$$
$$\leq \mathrm{pen}(k) - \frac{2P\chi_k - P\Theta_k}{n} \leq \delta'\frac{P\Theta_k}{n} + \square_{\delta,\delta',\Upsilon}\frac{r_n \log(n \vee |\mathcal{K}_n|)^2}{n} \quad .$$

Then, with probability larger than $1 - \square/n^2$,

$$\left\| s - \widehat{s}_{\widehat{k}_n} \right\|^2 \leq$$

$$(1 + \square_{\delta,\delta',\Upsilon,c_b}\ \mathrm{o}(1)) \inf_{k \in \mathcal{K}_n} \|s - \widehat{s}_k\|^2 + \square_{\delta,\delta',\Upsilon} (r_n + \log n) \frac{\log(n \vee |\mathcal{K}_n|)^2}{n} \ .$$

The proof easily follows by taking $\theta = (\log n)^{-1}$ in (17), $\eta = 2$ for instance in Proposition 4.1 and by using Assumption (24) and the bounds on $\mathrm{pen}(k)$. This result shows that the estimator $\widehat{s}_{\widehat{k}_n}$ satisfies an asymptotically optimal oracle inequality when condition (24) holds, whatever the values of δ and δ' even when they are negative. This proves that the lower bound $\mathrm{pen}_{\min}$ is not sharp in this case.
Therefore, we have to assume that at least one bias $\|s - \widehat{s}_k\|^2$ is negligible with respect to $P\Theta_k/n$. Actually, to conclude, we assume that this happens for $k_{1,n}$ in (21).

Theorem 4.3. *Let $(\mathcal{K}_n)_{n \geq 1}$ be a sequence of finite collections of kernels satisfying Assumptions* (11), (12), (13), (14) (15), (16)*, with Υ not depending on n. Each $\mathcal{K}_n$ is also assumed to satisfy* (22) *and* (23) *with a kernel $k_{1,n} \in \mathcal{K}_n$ in* (22) *such that*

$$\left\| s - s_{k_{1,n}} \right\|^2 \leq c\ \mathrm{o}(1) \frac{P\Theta_{k_{1,n}}}{n} \ , \tag{25}$$

for some fixed positive constant c. Suppose that there exist $\delta \geq \delta' > 0$ and a sequence $(r_n)_{n \geq 1}$ of nonnegative real numbers such that $r_n \leq \square \log(|\mathcal{K}_n| \vee n)$ and such that for all $n \geq 1$, with probability larger than $1 - \square n^{-2}$, for all $k \in \mathcal{K}_n$,

$$\frac{2P\chi_k - P\Theta_k}{n} - \delta \frac{P\Theta_k}{n} - \square_{\delta,\delta',\Upsilon} \frac{r_n \log(|\mathcal{K}_n| \vee n)^2}{n} \leq \mathrm{pen}(k)$$
$$\leq \frac{2P\chi_k - P\Theta_k}{n} - \delta' \frac{P\Theta_k}{n} + \square_{\delta,\delta',\Upsilon} \frac{r_n \log(|\mathcal{K}_n| \vee n)^2}{n} \ . \tag{26}$$

Then, with probability larger than $1 - \square/n^2$, the following holds

$$P\Theta_{\widehat{k}_n} \geq \left(\frac{\delta'}{\delta} + \square_{\delta,\delta',\Upsilon,c,c_s,c_R}\ \mathrm{o}(1) \right) P\Theta_{k_{1,n}} \quad \textit{and} \tag{27}$$

$$\left\| s - \widehat{s}_{\widehat{k}_n} \right\|^2 \geq \left(\frac{\delta'}{\delta} + \square_{\delta,\delta',\Upsilon,c,c_s,c_R}\ \mathrm{o}(1) \right) \left\| s - \widehat{s}_{k_{1,n}} \right\|^2$$
$$\gg \left\| s - \widehat{s}_{k_{0,n}} \right\|^2 \geq \inf_{k \in \mathcal{K}_n} \|s - \widehat{s}_k\|^2 \ . \tag{28}$$

By (28), under the conditions of Theorem 4.3, the estimator $\widehat{s}_{\widehat{k}_n}$ cannot satisfy an oracle inequality, hence, the lower bound $(2P\chi_k - P\Theta_k)/n$ in Theorem 3.1 is sharp. This shows that $(2P\chi_k - P\Theta_k)/n$ is a minimal penalty in the sense of [BM07] for kernel selection. When

$$\mathrm{pen}(k) = \frac{2P\chi_k - P\Theta_k}{n} + \kappa \frac{P\Theta_k}{n} \ ,$$

the complexity $P\Theta_{\widehat{k}_n}$ presents a sharp phase transition when κ becomes positive. Indeed, when $\kappa < 0$ it follows from (27) that the complexity $P\Theta_{\widehat{k}_n}$ is asymptotically larger than $P\Theta_{k_{1,n}}$. But on the other hand, as a consequence of Theorem 3.1, when $\kappa > 0$, this complexity becomes smaller than

$$\square_\kappa n \inf_{k\in\mathcal{K}_n}\left(\|s-s_k\|^2+\frac{P\Theta_k}{n}\right)\le \square_\kappa\left(n\left\|s-s_{k_{0,n}}\right\|^2+P\Theta_{k_{0,n}}\right)$$
$$\ll \square_\kappa\left(n\left\|s-s_{k_{1,n}}\right\|^2+P\Theta_{k_{1,n}}\right)\le \square_\kappa P\Theta_{k_{1,n}}\ .$$

4.3. Examples

Example 1 (continued). Let $\mathcal{S}=\mathcal{S}_n$ be the collection of spaces of regular histograms on $[0,1]$ with dimensions $\{1,\dots,n\}$ and let $\hat{S}=\hat{S}_n$ be the selected space thanks to the penalized criterion. Recall that, for any $S\in\mathcal{S}_n$, the orthonormal basis is defined by $\varphi_i=\sqrt{D_S}\mathbf{1}_i$ and $P\Theta_{k_S}=D_S$. Assume that s is α-Hölderian, with $\alpha\in(0,1]$ with α-Hölderian norm L. It is well known (see for instance Section 1.3.3. of [Bir06]) that the bias is bounded above by

$$\|s-s_{k_S}\|^2\le \square_L D_S^{-2\alpha}\ .$$

In particular, if $D_{S_1}=n$,

$$\left\|s-s_{k_{S_1}}\right\|^2\le \square_L n^{-2\alpha}\ll 1=\frac{D_{S_1}}{n}=\frac{P\Theta_{k_{S_1}}}{n}\ .$$

Thus, (25) holds for kernel k_{S_1}. Moreover, if $D_{S_0}=\lfloor\sqrt{n}\rfloor$,

$$\frac{(\log(n\vee|\mathcal{S}_n|)^3}{n}\ll\left\|s-s_{k_{S_0}}\right\|^2+\frac{D_{S_0}}{n}\le\square_L\left(\frac{1}{n^\alpha}+\frac{1}{\sqrt{n}}\right)$$
$$\ll\left\|s-s_{k_{S_1}}\right\|^2+\frac{D_{S_1}}{n}\ .$$

Hence, (21) holds with $k_{0,n}=k_{S_0}$ and $k_{1,n}=k_{S_1}$. Therefore, Theorem 4.3 and Theorem 3.1 apply in this example. If $\mathrm{pen}(k_S)=(1-\delta)D_S/n$, the dimension $D_{k_{\widehat{S}_n}}\ge\square_\delta n$ and $\widehat{s}_{k_{\widehat{S}_n}}$ is not consistent and does not satisfy an oracle inequality. On the other hand, if $\mathrm{pen}(k_S)=(1+\delta)D_S/n$,

$$D_{\widehat{S}_n}\le\square_{L,\delta}\left(n^{1-\alpha}+\sqrt{n}\right)\ll D_{S_1}=n$$

and $\widehat{s}_{k_{\widehat{S}_n}}$ satisfies an oracle inequality which implies that, with probability larger than $1-\square/n^2$,

$$\left\|s-\widehat{s}_{k_{\widehat{S}_n}}\right\|^2\le\square_{\alpha,L,\delta}n^{-2\alpha/(2\alpha+1)}\ ,$$

by taking $D_S\simeq n^{1/(2\alpha+1)}$. It achieves the minimax rate of convergence over the class of α-Hölderian functions.

From Theorem 3.1, the penalty $\mathrm{pen}(k_S)=2D_S/n$ provides an estimator $\widehat{s}_{k_{\widehat{S}_n}}$ that achieves an asymptotically optimal oracle inequality. Therefore the optimal penalty is equal to 2 times the minimal one. In particular, the slope heuristics of [BM07] holds in this example, as already noticed in [Ler12].

Finally to illustrate Corollary 4.2, let us take $s(x)=2x$ and the collection of regular histograms with dimension in $\{1,\dots,\lfloor n^\beta\rfloor\}$, with $\beta<1/3$. Simple calculations show that

$$\frac{\|s-s_{k_S}\|^2}{D_S}\ge\square D_S^{-3}\ge\square n^{-3\beta}\gg n^{-1}.$$

Hence (24) applies and the penalized estimator with penalty $\mathrm{pen}(k_S) \simeq \delta \frac{D_S}{n}$ always satisfies an oracle inequality even if $\delta = 0$ or $\delta < 0$. This was actually expected since it is likely to choose the largest dimension which is also the oracle choice in this case.

Example 2 (continued). Let K be a fixed function, let $\mathcal{H} = \mathcal{H}_n$ denote the following grid of bandwidths

$$\mathcal{H} = \left\{ \frac{\|K\|_\infty \|K\|_1}{i} \quad / \quad i = 1, \dots, n \right\}$$

and let $\hat{h} = \hat{h}_n$ be the selected bandwidth. Assume as before that s is a density on $[0,1]$ that belongs to the Nikol'ski class $\mathcal{N}(\alpha, L)$ with $\alpha \in (0,1]$ and $L > 0$. By Proposition 1.5 in [Tsy09], if K satisfies $\int |u|^\alpha |K(u)|\, du < \infty$

$$\left\| s - s_{k_{K,h}} \right\|^2 \leq \square_{\alpha,K,L} h^{2\alpha} \ .$$

In particular, when $h_1 = \|K\|_\infty \|K\|_1 / n$,

$$\left\| s - s_{k_{K,h_1}} \right\|^2 \leq \square_{\alpha,K,L} n^{-2\alpha} \ll \frac{P\Theta_{k_{K,h_1}}}{n} = \frac{\|K\|^2}{\|K\|_\infty \|K\|_1} \ .$$

On the other hand, for $h_0 = \|K\|_\infty \|K\|_1 / \lfloor \sqrt{n} \rfloor$,

$$\begin{aligned} \frac{(\log n \vee |\mathcal{H}_n|)^2}{n} &\ll \left\| s - s_{k_{K,h_0}} \right\|^2 + \frac{P\Theta_{k_{K,h_0}}}{n} \\ &\leq \square_{K,\alpha,L} \left(\frac{1}{n^\alpha} + \frac{1}{\sqrt{n}} \right) \ll \left\| s - s_{k_{K,h_1}} \right\|^2 + \frac{P\Theta_{k_{K,h_1}}}{n} \ . \end{aligned}$$

Hence, (21) and (25) hold with kernels $k_{0,n} = k_{K,h_0}$ and $k_{1,n} = k_{K,h_1}$. Therefore, Theorem 4.3 and Theorem 3.1 apply in this example. If for some $\delta > 0$ we set $\mathrm{pen}(k_{K,h}) = (2K(0) - \|K\|^2 - \delta \|K\|^2)/(nh)$, then $\widehat{h}_n \leq \square_{\delta,K} n^{-1}$ and $\widehat{s}_{k_{K,\widehat{h}_n}}$ is not consistent and does not satisfy an oracle inequality. On the other hand, if $\mathrm{pen}(k_{K,h}) = (2K(0) - \|K\|^2 + \delta \|K\|^2)/(nh)$, then

$$\widehat{h}_n \geq \square_{\delta,K,L} \left(n^{1-\alpha} + \sqrt{n} \right)^{-1} \gg \square_{\delta,K,L} n^{-1} \ ,$$

and $\widehat{s}_{K,k_{\widehat{h}_n}}$ satisfies an oracle inequality which implies that, with probability larger than $1 - \square/n^2$,

$$\left\| s - \widehat{s}_{k_{K,\widehat{h}_n}} \right\|^2 \leq \square_{\alpha,K,L,\delta} n^{-2\alpha/(2\alpha+1)} \ ,$$

for $h = \|K\|_\infty \|K\|_1 / \left\lfloor n^{1/(2\alpha+1)} \right\rfloor \in \mathcal{H}$. In particular it achieves the minimax rate of convergence over the class $\mathcal{N}(\alpha, L)$. Finally, if $\mathrm{pen}(k_{K,h}) = 2K(0)/(nh)$, $\widehat{s}_{k_{K,\widehat{h}_n}}$ achieves an asymptotically optimal oracle inequality, thanks to Theorem 3.1.

The minimal penalty is therefore

$$\mathrm{pen}_{\min}(k_{K,h}) = \frac{2K(0) - \|K\|^2}{nh} \ .$$

In this case, the optimal penalty $\mathrm{pen}_{\mathrm{opt}}(k_{K,h}) = 2K(0)/(nh)$ derived from Theorem 3.1 is not twice the minimal one, but one still has, if $2K(0) \neq \|K\|^2$,

$$\mathrm{pen}_{\mathrm{opt}}(k_{K,h}) = \frac{2K(0)}{2K(0) - \|K\|^2}\,\mathrm{pen}_{\mathrm{min}}(k_{K,h}) \ ,$$

even if they can be of opposite sign depending on K. This type of nontrivial relationship between optimal and minimal penalty has already been underlined in [1] in regression framework for selecting linear estimators.
Note that if one allows two kernel functions K_1 and K_2 in the family of kernels such that $2K_1(0) \neq \|K_1\|^2$, $2K_2(0) \neq \|K_2\|^2$ and

$$\frac{2K_1(0)}{2K_1(0) - \|K_1\|^2} \neq \frac{2K_2(0)}{2K_2(0) - \|K_2\|^2} \ ,$$

then there is no absolute constant multiplicative factor linking the minimal penalty and the optimal one.

5. Small simulation study

In this section we illustrate on simulated data Theorem 3.1 and Theorem 4.3. We focus on approximation kernels only, since projection kernels have been already discussed in [Ler12].
We observe an $n = 100$ i.i.d. sample of standard gaussian distribution. For a fixed parameter $a \geq 0$ we consider the family of kernels

$$k_{K_a,h}(x,y) = \frac{1}{h}K_a\left(\frac{x-y}{h}\right) \qquad \text{with} \quad h \in \mathcal{H} = \left\{\frac{1}{2i},\ i = 1,\dots,50\right\} \ ,$$

where for $x \in \mathbb{R}$, $\quad K_a(x) = \dfrac{1}{2\sqrt{2\pi}}\left(e^{-\frac{(x-a)^2}{2}} + e^{-\frac{(x+a)^2}{2}}\right)$.
In particular the kernel estimator with $a = 0$ is the classical Gaussian kernel estimator. Moreover

$$K_a(0) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{a^2}{2}\right) \quad \text{and} \quad \|K_a\|^2 = \frac{1 + e^{-a^2}}{4\sqrt{\pi}} \ .$$

Thus, depending on the value of a, the minimal penalty $(2K_a(0) - \|K_a\|^2)/(nh)$ may be negative. We study the behavior of the penalized criterion

$$\mathcal{C}_{\mathrm{pen}}\left(k_{K_a,h}\right) = P_n\gamma(\widehat{s}_{k_{K_a,h}}) + \mathrm{pen}(k_{K_a,h})$$

with penalties of the form

$$\mathrm{pen}\left(k_{K_a,h}\right) = \frac{2K_a(0) - \|K_a\|^2}{nh} + \kappa\frac{\|K_a\|^2}{nh} \ , \tag{29}$$

for different values of κ ($\kappa = -1, 0, 1$) and a ($a = 0, 1.5, 2, 3$). On Figure 1 are represented the selected estimates by the optimal penalty $2K_a(0)/(nh)$ for the different values of a and on Figure 2 one sees the evolution of the different penalized criteria as a function of $1/h$. The contrast curves for $a = 0$ are classical on Figure 2. Without penalization, the criterion decreases and leads to the selection of the smallest bandwidth. At the minimal penalty, the curve is flat and at the optimal penalty one selects a meaningful bandwidth as shown on Figure 1.
When $a > 0$, despite the choice of those unusual kernels, the reconstructions on Figure 1 for the optimal penalty are also meaningful. However when $a = 2$ or $a = 3$, the criterion with minimal penalty is smaller than the unpenalized

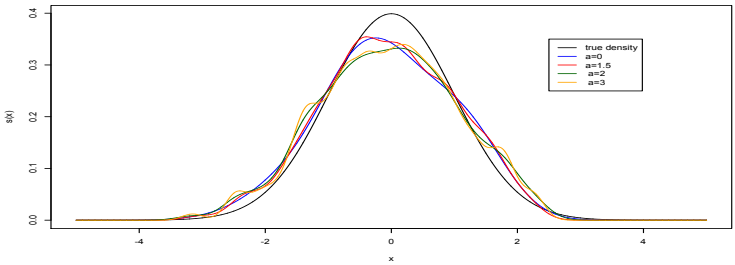

FIGURE 1. Selected approximation kernel estimators when the penalty is the optimal one i.e. $\frac{2K_a(0)}{nh}$.

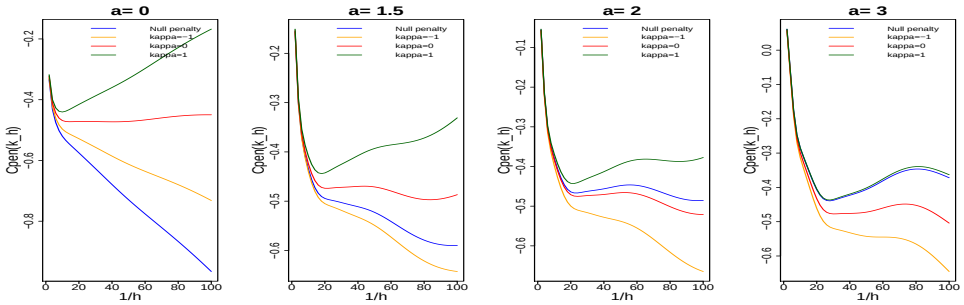

FIGURE 2. Behavior of $P_n\gamma(\widehat{s}_{k_{K_a,h}})$ (blue line) and $\mathcal{C}_{\text{pen}}(k_{K_a,h})$ as a function of $1/h$, which is proportional to the complexity $P\Theta_{k_{K_a,h}}$.

criterion, meaning that minimizing the latter criterion leads by Theorem 3.1 to an oracle inequality. In our simulation, when $a = 3$, the curves for the optimal criterion and the unpenalized one are so close that the same estimator is selected by both methods.

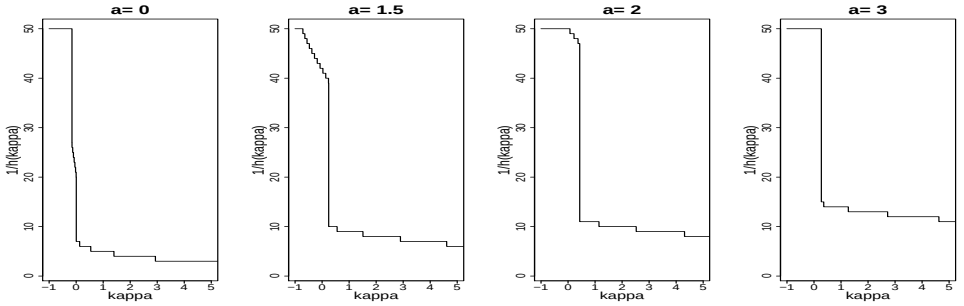

FIGURE 3. Behavior of $1/\hat{h}$, which is proportional to the complexity $P\Theta_{k_{K_a,h}}$, for the estimator selected by the criterion whose penalty is given by (29), as a function of κ.

Finally Figure 3 shows that there is indeed in all cases a sharp phase transition around $\kappa = 0$ i.e. at the minimal penalty for the complexity of the selected estimate.

6. Proofs

6.1. Proof of Theorem 3.1

The starting point to prove the oracle inequality is to notice that any minimizer $\widehat{k}$ of $\mathcal{C}_{\text{pen}}$ satisfies

$$\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 \leq \|s - \widehat{s}_k\|^2 + \left(\text{pen}(k) - \text{pen}_{\text{id}}(k)\right) - \left(\text{pen}\left(\widehat{k}\right) - \text{pen}_{\text{id}}\left(\widehat{k}\right)\right) \ .$$

Using the expression of the ideal penalty (9) we find

$$\begin{aligned}\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 \leq \|s - \widehat{s}_k\|^2 &+ \left(\text{pen}(k) - 2\frac{P\chi_k}{n}\right) - \left(\text{pen}\left(\widehat{k}\right) - 2\frac{P\chi_{\widehat{k}}}{n}\right) \\ &+ 2\frac{P(s_k - s_{\widehat{k}})}{n} + 2\left(1 - \frac{2}{n}\right)(P_n - P)(s_{\widehat{k}} - s_k) \\ &+ 2\frac{(P_n - P)(\chi_{\widehat{k}} - \chi_k)}{n} + 2\frac{U_{\widehat{k}} - U_k}{n^2} \ . \end{aligned} \tag{30}$$

By Proposition B.1 (see the appendix), for all $x > 1$, for all θ in $(0,1)$, with probability larger than $1 - (7.4|\mathcal{K}| + 2|\mathcal{K}|^2)e^{-x}$,

$$\begin{aligned}\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 \leq \|s - \widehat{s}_k\|^2 &+ \left(\text{pen}(k) - 2\frac{P\chi_k}{n}\right) - \left(\text{pen}\left(\widehat{k}\right) - 2\frac{P\chi_{\widehat{k}}}{n}\right) \\ &+ \theta\left\|s - s_{\widehat{k}}\right\|^2 + \theta\|s - s_k\|^2 + \square\frac{\Upsilon}{\theta n} \\ &+ \left(1 - \frac{2}{n}\right)\theta\left\|s - s_{\widehat{k}}\right\|^2 + \left(1 - \frac{2}{n}\right)\theta\|s - s_k\|^2 + \square\frac{\Upsilon x^2}{\theta n} \\ &+ \theta\frac{P\Theta_k}{n} + \theta\frac{P\Theta_{\widehat{k}}}{n} + \square\frac{\Upsilon x}{\theta n} + \theta\frac{P\Theta_k}{n} + \theta\frac{P\Theta_{\widehat{k}}}{n} + \square\frac{\Upsilon x^2}{\theta n}\end{aligned}$$

Hence

$$\begin{aligned}\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 \leq \|s - \widehat{s}_k\|^2 &+ \left(\text{pen}(k) - 2\frac{P\chi_k}{n}\right) - \left(\text{pen}\left(\widehat{k}\right) - 2\frac{P\chi_{\widehat{k}}}{n}\right) \\ &+ 2\theta\left[\left\|s - s_{\widehat{k}}\right\|^2 + \frac{P\Theta_{\widehat{k}}}{n}\right] + 2\theta\left[\|s - s_k\|^2 + \frac{P\Theta_k}{n}\right] + \square\frac{\Upsilon x^2}{\theta n} \ .\end{aligned}$$

This bound holds using (11), (12) and (13) only. Now by Proposition 4.1 applied with $\eta = 1$, we have for all $x > 1$, for all $\theta \in (0,1)$, with probability larger than $1 - (16.8|\mathcal{K}| + 2|\mathcal{K}|^2)e^{-x}$,

$$\begin{aligned}\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 \leq \|s - \widehat{s}_k\|^2 &+ \left(\text{pen}(k) - 2\frac{P\chi_k}{n}\right) - \left(\text{pen}\left(\widehat{k}\right) - 2\frac{P\chi_{\widehat{k}}}{n}\right) \\ &+ 4\theta\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 + 4\theta\|s - \widehat{s}_k\|^2 + \square\frac{\Upsilon x^2}{\theta n} \ .\end{aligned}$$

This gives the first part of the theorem.

For the second part, by the condition (18) on the penalty, we find for all $x > 1$, for all θ in $(0,1)$, with probability larger than $1 - (C + 16.8|\mathcal{K}| + 2|\mathcal{K}|^2)e^{-x}$,

$$(1-4\theta)\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 \leq$$
$$(1+4\theta)\left\|s - \widehat{s}_k\right\|^2 + (\delta'-1)_+\frac{P\Theta_k}{n} + (1-\delta)_+\frac{P\Theta_{\widehat{k}}}{n} + \square\left(r + \frac{1}{\theta}\right)\frac{\Upsilon x^2}{n} \ .$$

By Proposition 4.1 applied with $\eta = \theta$, we have with probability larger than $1 - (C + 26.2|\mathcal{K}| + 2|\mathcal{K}|^2)e^{-x}$,

$$(1-4\theta)\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 \leq (1+4\theta)\left\|s - \widehat{s}_k\right\|^2 + (\delta'-1)_+(1+\theta)\left\|s - \widehat{s}_k\right\|^2$$
$$+ (1-\delta)_+(1+\theta)\left\|s - \widehat{s}_{\widehat{k}}\right\|^2 + \square\left(r + \frac{1}{\theta^3}\right)\frac{\Upsilon x^2}{n} \ ,$$

that is

$$\left((\delta \wedge 1) - \theta(4 + (1-\delta)_+)\right)\left\|s - \widehat{s}_{\widehat{k}}\right\|^2$$
$$\leq \left((\delta' \vee 1) + \theta(4 + (\delta'-1)_+)\right)\left\|s - \widehat{s}_k\right\|^2 + \square\left(r + \frac{1}{\theta^3}\right)\frac{\Upsilon x^2}{n} \ .$$

Hence, because $1 \leq [(\delta' \vee 1) + (4 + (\delta'-1)_+)\theta] \leq (\delta' \vee 1) + (4 + \delta')\theta$, we obtain the desired result.

6.2. Proof of Proposition 4.1

First, let us denote for all $x \in \mathbb{X}$

$$F_{A,k}(x) := \mathbb{E}\left[A_k(X,x)\right], \qquad \zeta_k(x) := \int \left(k(y,x) - s_k(y)\right)^2 d\mu(y) \ ,$$

and

$$U_{A,k} := \sum_{i \neq j = 1}^{n} \left(A_k(X_i, X_j) - F_{A,k}(X_i) - F_{A,k}(X_j) + \mathbb{E}\left[A_k(X,Y)\right]\right) \ .$$

Some easy computations then provide the following useful equality

$$\left\|s_k - \widehat{s}_k\right\|^2 = \frac{1}{n}P_n\zeta_k + \frac{1}{n^2}U_{A,k} \ .$$

We need only treat the terms on the right-hand side, thanks to the probability tools of Section 2.3. Applying Proposition 2.1, we get, for any $x \geq 1$, with probability larger than $1 - 2\left|\mathcal{K}\right| e^{-x}$,

$$|(P_n - P)\zeta_k| \leq \sqrt{\frac{2x}{n}P\zeta_k^2} + \frac{\left\|\zeta_k\right\|_\infty x}{3n} \ .$$

One can then check the following link between ζ_k and Θ_k

$$P\zeta_k = \int \left(k(y,x) - s_k(x)\right)^2 s(y)d\mu(x)d\mu(y) = P\Theta_k - \left\|s_k\right\|^2 \ .$$

Next, by (1) and (11)

$$\left\|\zeta_k\right\|_\infty = \sup_{y \in \mathbb{X}} \int \left(k(y,x) - \mathbb{E}\left[k(X,x)\right]\right)^2 d\mu(x)$$
$$\leq 4 \sup_{y \in \mathbb{X}} \int k(y,x)^2 d\mu(x) \leq 4\Upsilon n \ .$$

In particular, since $\zeta_k \geq 0$,

$$P\zeta_k^2 \leq \|\zeta_k\|_\infty P\zeta_k \leq 4\Upsilon n P\Theta_k \ .$$

It follows from these computations and from (11) that there exists an absolute constant $\square$ such that, for any $x \geq 1$, with probability larger than $1-2\,|\mathcal{K}|\,e^{-x}$, for any $\theta \in (0,1)$,

$$|P_n\zeta_k - P\Theta_k| \leq \theta P\Theta_k + \square\frac{\Upsilon x}{\theta} \ .$$

We now need to control the term $U_{A,k}$. From Proposition 2.2, for any $x \geq 1$, with probability larger than $1 - 5.4\,|\mathcal{K}|\,e^{-x}$,

$$\frac{|U_{A,k}|}{n^2} \leq \frac{\square}{n^2}\left(C\sqrt{x} + Dx + Bx^{3/2} + Ax^2\right) \ .$$

By (1), (11) and Cauchy-Schwarz inequality,

$$A = 4\sup_{(x,y)\in\mathbb{X}^2}\int k(x,z)k(y,z)d\mu(z) \leq 4\sup_{x\in\mathbb{X}}\int k(x,z)^2 d\mu(z) \leq 4\Upsilon n \ .$$

In addition, by (15), $B^2 \leq 16\sup_{x\in\mathbb{X}}\mathbb{E}\left[A_k(X,x)^2\right] \leq 16\Upsilon n$.
Moreover, applying the Assumption (14),

$$C^2 \leq \sum_{i\neq j=1}^{n}\mathbb{E}\left[A_k(X_i,X_j)^2\right] = n^2\mathbb{E}\left[A_k(X,Y)^2\right] \leq n^2\Upsilon P\Theta_k \ .$$

Finally, applying the Cauchy-Schwarz inequality and proceeding as for C^2, the quantity used to define D can be bounded above as follows:

$$\mathbb{E}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i(X_i)b_j(X_j)A_k(X_i,X_j)\right] \leq n\sqrt{\mathbb{E}\left[A_k(X,Y)^2\right]} \leq n\sqrt{\Upsilon P\Theta_k} \ .$$

Hence for any $x \geq 1$, with probability larger than $1 - 5.4\,|\mathcal{K}|\,e^{-x}$,

$$\text{for any } \theta \in (0,1), \quad \frac{|U_{A,k}|}{n^2} \leq \theta\frac{P\Theta_k}{n} + \square\frac{\Upsilon x^2}{\theta n} \ .$$

Therefore, for all $\theta \in (0,1)$,

$$\left|\|\widehat{s}_k - s_k\|^2 - \frac{P\Theta_k}{n}\right| \leq 2\theta\frac{P\Theta_k}{n} + \square\frac{\Upsilon x^2}{\theta n} \ ,$$

and the first part of the result follows by choosing $\theta = \eta/2$. Concerning the two remaining inequalities appearing in the proposition, we begin by developing the loss. For all $k \in \mathcal{K}$

$$\|\widehat{s}_k - s\|^2 = \|\widehat{s}_k - s_k\|^2 + \|s_k - s\|^2 + 2\langle\widehat{s}_k - s_k, s_k - s\rangle \ .$$

Then, for all $x \in \mathbb{X}$

$$\begin{aligned} F_{A,k}(x) - s_k(x) &= \int s(y)\int k(x,z)k(z,y)d\mu(z)d\mu(y) - \int s(z)k(z,x)d\mu(z) \\ &= \int\left(\int s(y)k(z,y)d\mu(y) - s(z)\right)k(x,z)d\mu(z) \\ &= \int\left(s_k(z) - s(z)\right)k(z,x)d\mu(z) \ . \end{aligned}$$

Moreover, since $PF_{A,k} = \|s_k\|^2$, we find

$$\begin{aligned}\langle \widehat{s}_k - s_k, s_k - s\rangle &= \int (\widehat{s}_k(x)\,(s_k(x) - s(x)))\,d\mu(x) + \mathbb{E}\,[s_k(X)] - \|s_k\|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\int (k(x,X_i)\,(s_k(x) - s(x)))\,d\mu(x) + P(s_k - F_{A,k}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(F_{A,k}(X_i) - s_k(X_i)) + P(s_k - F_{A,k}) \\ &= (P_n - P)(F_{A,k} - s_k)\ .\end{aligned}$$

This expression motivates us to apply again Proposition 2.1 to this term. We find by (1), (11) and Cauchy-Schwarz inequality

$$\begin{aligned}\sup_{x\in\mathbb{X}}|F_{A,k}(x) - s_k(x)| &\le \|s - s_k\| \sup_{x\in\mathbb{X}}\int \frac{|s(z) - s_k(z)|}{\|s - s_k\|}k(x,z)d\mu(z) \\ &\le \|s - s_k\|\sqrt{\sup_{x\in\mathbb{X}}\int k(x,z)^2 d\mu(z)} \le \|s - s_k\|\sqrt{\Upsilon n}\ .\end{aligned}$$

Moreover,

$$\begin{aligned}P\,(F_{A,k} - s_k)^2 &\le \|s - s_k\|^2 P\left(\int \frac{|s(z) - s_k(z)|}{\|s - s_k\|}k(.,z)d\mu(z)\right)^2 \\ &\le \|s - s_k\|^2 v_k^2\ .\end{aligned}$$

Thus by (16), for any $\theta, u > 0$,

$$\begin{aligned}\sqrt{\frac{2P\,(F_{A,k} - s_k)^2\,x}{n}} &\le \theta\,\|s - s_k\|^2 + \frac{\left(\Upsilon \vee \sqrt{\Upsilon P\Theta_k}\right)x}{2\theta n} \\ &\le \theta\,\|s - s_k\|^2 + \frac{\Upsilon x}{\theta n} \vee \left(\frac{u}{\theta}\frac{P\Theta_k}{n} + \frac{\Upsilon x^2}{16\theta u n}\right)\ .\end{aligned}$$

Hence, for any $\theta \in (0,1)$ and $x \ge 1$, taking $u = \theta^2$

$$\sqrt{\frac{2P\,(F_{A,k} - s_k)^2\,x}{n}} \le \theta\left(\|s - s_k\|^2 + \frac{P\Theta_k}{n}\right) + \square\frac{\Upsilon x^2}{\theta^3 n}\ .$$

By Proposition 2.1, for all θ in $(0,1)$, for all $x > 0$ with probability larger than $1 - 2|\mathcal{K}|e^{-x}$,

$$\begin{aligned}2\,|\langle \widehat{s}_k - s_k, s_k - s\rangle| &\le 2\sqrt{\frac{2P\,(F_{A,k} - s_k)^2\,x}{n}} + 2\,\|s - s_k\|\sqrt{\Upsilon n}\frac{x}{3n} \\ &\le 3\theta\left(\|s - s_k\|^2 + \frac{P\Theta_k}{n}\right) + \square\frac{\Upsilon x^2}{\theta^3 n}\ .\end{aligned}$$

Putting together all of the above, one concludes that for all θ in $(0,1)$, for all $x > 1$, with probability larger than $1 - 9.4|\mathcal{K}|e^{-x}$

$$\|\widehat{s}_k - s\|^2 - \|s_k - s\|^2 \le 3\theta\,\|s - s_k\|^2 + (1 + 4\theta)\frac{P\Theta_k}{n} + \square\frac{\Upsilon x^2}{\theta^3 n}$$

and

$$\|\widehat{s}_k - s\|^2 - \|s_k - s\|^2 \ge -3\theta\left(\|s - s_k\|^2 + \frac{P\Theta_k}{n}\right) + (1 - \theta)\frac{P\Theta_k}{n} - \square\frac{\Upsilon x^2}{\theta^3 n}\ .$$

Choosing, $\theta = \eta/4$ leads to the second part of the result.

6.3. Proof of Theorem 4.3

It follows from (17) (applied with $\theta = \square(\log n)^{-1}$ and $x = \square \log(n \vee |\mathcal{K}_n|)$) and Assumption (26) that with probability larger than $1 - \square n^{-2}$ we have for any $k \in \mathcal{K}$ and any $n \geq 2$

$$\left\|\widehat{s}_{\widehat{k}_n} - s\right\|^2 \leq \left(1 + \frac{\square}{\log n}\right) \|\widehat{s}_k - s\|^2 - (1+\delta')\left(1 + \frac{\square}{\log n}\right)\frac{P\Theta_k}{n} + (1+\delta)\left(1 + \frac{\square}{\log n}\right)\frac{P\Theta_{\widehat{k}_n}}{n} + \square_{\delta,\delta',\Upsilon}\frac{\log(|\mathcal{K}_n| \vee n)^3}{n} \ . \quad (31)$$

Applying this inequality with $k = k_{1,n}$ and using Proposition 4.1 with $\eta = \square(\log n)^{-1/3}$ and $x = \square \log(|\mathcal{K}_n| \vee n)$ as a lower bound for $\left\|\widehat{s}_{\widehat{k}_n} - s\right\|^2$ and as an upper bound for $\left\|\widehat{s}_{k_{1,n}} - s\right\|^2$, we obtain asymptotically that with probability larger than $1 - \square n^{-2}$,

$$-\delta(1+\square_\delta\ \mathrm{o}(1))\frac{P\Theta_{\widehat{k}_n}}{n} \leq (1+\mathrm{o}(1))\left\|s_{k_{1,n}} - s\right\|^2 - \delta'(1+\square_{\delta'}\ \mathrm{o}(1))\frac{P\Theta_{k_{1,n}}}{n} + \square_{\delta,\delta',\Upsilon}\frac{\log(|\mathcal{K}_n| \vee n)^3}{n} \ .$$

By Assumption (25), $\left\|s_{k_{1,n}} - s\right\|^2 \leq c\ \mathrm{o}(1)\frac{P\Theta_{k_{1,n}}}{n}$ and by (22),

$$\frac{(\log(|\mathcal{K}_n| \vee n))^3}{n} \leq c_R c_s\ \mathrm{o}(1)\frac{P\Theta_{k_{1,n}}}{n} \ .$$

This gives (27). In addition, starting with the event where (31) holds and using Proposition 4.1, we also have with probability larger than $1 - \square n^{-2}$,

$$\left\|\widehat{s}_{\widehat{k}_n} - s\right\|^2 \leq \left(1 + \frac{\square}{\log n}\right)\left\|\widehat{s}_{k_{1,n}} - s\right\|^2 - (1+\delta')\frac{P\Theta_{k_{1,n}}}{n} + (1+\delta)\,(1+\mathrm{o}(1))\left\|\widehat{s}_{\widehat{k}_n} - s\right\|^2 + \square_{\delta,\delta',\Upsilon}\frac{\log(|\mathcal{K}_n| \vee n)^3}{n} \ .$$

Since $\left\|\widehat{s}_{k_{1,n}} - s\right\|^2 \simeq \frac{P\Theta_{k_{1,n}}}{n}$, this leads to

$$(-\delta + \square_\delta\ \mathrm{o}(1))\left\|\widehat{s}_{\widehat{k}} - s\right\|^2 \leq -(\delta' + \square_{\delta',c}\ \mathrm{o}(1))\left\|\widehat{s}_{k_{1,n}} - s\right\|^2 + \square_{\delta,\delta',\Upsilon}\ \frac{\log(|\mathcal{K}_n| \vee n)^3}{n} \ .$$

This leads to (28) by (21).

Appendix A. Proofs for the examples

A.1. Computation of the constant Γ for the three examples

We have to show for each family $\{k\}_{k\in\mathcal{K}}$ (see (8) and (1)) that there exists a constant $\Gamma \geq 1$ such that for all $k \in \mathcal{K}$

$$\sup_{x\in\mathbb{X}} |\Theta_k(x)| \leq \Gamma n, \quad \text{and} \quad \sup_{(x,y)\in\mathbb{X}^2} |k(x,y)| \leq \Gamma n \ .$$

Example 1: Projection kernels. First, notice that from Cauchy-Schwarz inequality we have for all $(x,y) \in \mathbb{X}^2$ $|k_S(x,y)| \leq \sqrt{\chi_{k_S}(x)\chi_{k_S}(y)}$ and by orthonormality, for any $(x,x') \in \mathbb{X}^2$,

$$A_{k_S}(x,x') = \sum_{(i,j)\in\mathcal{I}_S^2} \varphi_i(x)\varphi_j(x') \int_{\mathbb{X}} \varphi_i(y)\varphi_j(y)d\mu(y) = k_S(x,x') \ .$$

In particular, for any $x \in \mathbb{X}$, $\Theta_{k_S}(x) = \chi_{k_S}(x)$. Hence, projection kernels satisfy (1) for $\Gamma = 1 \vee n^{-1} \sup_{S\in\mathcal{S}} \|\chi_{k_S}\|_\infty$. We conclude by writing

$$\|\chi_{k_S}\|_\infty = \sup_{x\in\mathbb{X}} \sum_{i\in\mathcal{I}_S} \varphi_i(x)^2 = \sup_{\substack{(a_i)_{i\in\mathcal{I}} \text{ s.t.} \\ \sum_{i\in\mathcal{I}_S} a_i^2 = 1}} \sup_{x\in\mathbb{X}} \left(\sum_{i\in\mathcal{I}_S} a_i\varphi_i(x) \right)^2 \ .$$

For $f \in S$ we have $\|f\|^2 = \sum_{i\in\mathcal{I}} \langle f, \varphi_i\rangle^2$. Hence with $a_i = \langle f, \varphi_i\rangle$,

$$\|\chi_{k_S}\|_\infty = \sup_{f\in S, \|f\|=1} \|f\|_\infty^2 \ .$$

Example 2: Approximation kernels. First, $\sup_{(x,y)\in\mathbb{X}^2} |k_{K,h}(x,y)| \leq \|K\|_\infty / h$. Second, since $K \in L^1$

$$\Theta_{k_{K,h}}(x) = \frac{1}{h^2} \int_{\mathbb{X}} K\left(\frac{x-y}{h}\right)^2 dy = \frac{\|K\|^2}{h} \leq \frac{\|K\|_\infty \|K\|_1}{h} \ .$$

Now $K \in L^1$ and $\int K(u)du = 1$ implies $\|K\|_1 \geq 1$, hence (1) holds with $\Gamma = 1$ if one assumes that $h \geq \|K\|_\infty \|K\|_1 / n$.

Example 3: Weighted projection kernels. For all $x \in \mathbb{X}$

$$\Theta_{k_w}(x) = \sum_{i,j=1}^p w_i\varphi_i(x)w_j\varphi_j(x) \int_{\mathbb{X}} \varphi_i(y)\varphi_j(y)d\mu(y) = \sum_{i=1}^p w_i^2\varphi_i(x)^2 \ .$$

From Cauchy-Schwarz inequality, for any $(x,y) \in \mathbb{X}^2$,

$$|k_w(x,y)| \leq \sqrt{\Theta_{k_w}(x)\Theta_{k_w}(y)} \ .$$

We thus find that k_w verifies (1) with $\Gamma \geq 1 \vee n^{-1} \sup_{w\in\mathcal{W}} \|\Theta_{k_w}\|_\infty$. Since $w_i \leq 1$ we find the announced result which is independent of $\mathcal{W}$.

A.2. Proof of Proposition 3.2

Since $\|s_{k_S}\|^2 \leq \|s\|^2 \leq \|s\|_\infty$, we find that (11) only requires $\Upsilon \geq \Gamma(1+\|s\|_\infty)$. Assumption (12) holds: this follows from $\Upsilon \geq \Gamma$ and

$$\mathbb{E}\left[\chi_{k_S}(X)^2\right] \leq \|\chi_{k_S}\|_\infty P\chi_{k_S} \leq \Gamma n P\Theta_{k_S} \ .$$

Now for proving Assumption (14), we write

$$\begin{aligned}
\mathbb{E}\left[A_{k_S}(X,Y)^2\right] = \mathbb{E}\left[k_S(X,Y)^2\right] &= \int_{\mathbb{X}} \mathbb{E}\left[k_S(X,x)^2\right] s(x)d\mu(x) \\
&\leq \|s\|_\infty \sum_{(i,j)\in\mathcal{I}_S^2} \mathbb{E}\left[\varphi_i(X)\varphi_j(X)\right] \int_{\mathbb{X}} \varphi_i(x)\varphi_j(x)d\mu(x) \\
&= \|s\|_\infty P\Theta_{k_S} \leq \Upsilon P\Theta_{k_S} \ .
\end{aligned}$$

In the same way, Assumption (15) follows from $\|s\|_\infty \Gamma \leq \Upsilon$. Suppose (19) holds with $S = S + S'$ so that the basis $(\varphi_i)_{i\in\mathcal{I}}$ of S' is included in the one $(\varphi_i)_{i\in\mathcal{J}}$ of S. Since $\|\chi_{k_S}\|_\infty \leq \Gamma n$ we have

$$s_{k_S}(x) - s_{k_{S'}}(x) = \sum_{j\in\mathcal{J}\setminus\mathcal{I}} (P\varphi_j)\,\varphi_j(x) \leq \sqrt{\sum_{j\in\mathcal{J}\setminus\mathcal{I}} (P\varphi_j)^2 \sum_{j\in\mathcal{J}\setminus\mathcal{I}} \varphi_j(x)^2}$$
$$\leq \left\|s_{k_S} - s_{k_{S'}}\right\| \|\chi_{k_S}\|_\infty^{1/2} \leq \left\|s_{k_S} - s_{k_{S'}}\right\| \sqrt{\Gamma n} \ .$$

Hence, (13) holds in this case. Assuming (20) implies that (13) holds since

$$\left\|s_{k_S} - s_{k_{S'}}\right\|_\infty \leq \|s_{k_S}\|_\infty + \left\|s_{k_{S'}}\right\|_\infty \leq \Upsilon \ .$$

Finally for (16), for any $a \in L^2$,

$$\int_{\mathbb{X}} a(x) k_S(x,y) d\mu(x) = \sum_{i\in\mathcal{I}} \langle a, \varphi_i\rangle \varphi_i(y) = \Pi_S(a) \ .$$

is the orthogonal projection of a onto S. Therefore, $\mathbb{B}_{k_S}$ is the unit ball in S for the L^2-norm and, for any $t \in \mathbb{B}_{k_S}$, $\mathbb{E}\left[t(X)^2\right] \leq \|s\|_\infty \|t\|^2 \leq \|s\|_\infty$.

A.3. Proof of Proposition 3.3

First, since $\|K\|_1 \geq 1$

$$\left\|s_{k_{K,h}}\right\|^2 = \int_{\mathbb{X}} \left(\int_{\mathbb{X}} s(y) \frac{1}{h} K\left(\frac{x-y}{h}\right) dy \right)^2 dx$$
$$= \int_{\mathbb{X}} \left(\int_{\mathbb{X}} s(x+hz) K(z)\, dz \right)^2 dx$$
$$\leq \|K\|_1^2 \int_{\mathbb{X}} \left(\int_{\mathbb{X}} s(x+hz) \frac{|K(z)|}{\|K\|_1} dz \right)^2 dx$$
$$\leq \|K\|_1^2 \int_{\mathbb{X}^2} s(x+hz)^2 \frac{|K(z)|}{\|K\|_1} dx dz \leq \|s\|_\infty \|K\|_1^2 \ .$$

Hence, Assumption (11) holds if $\Upsilon \geq 1 + \|s\|_\infty \|K\|_1^2$. Now, we have

$$P\left(\chi^2_{k_{K,h}}\right) = \frac{K(0)^2}{h^2} = P\Theta_{k_{K,h}} \frac{K(0)^2}{\|K\|^2 h} \leq nP\Theta_{k_{K,h}} \frac{K(0)^2}{\|K\|^2 \|K\|_\infty \|K\|_1} \ ,$$

so it is sufficient to have $\Upsilon \geq K(0)/\|K\|^2$ (since $K(0) \leq \|K\|_\infty$) to ensure (12). Moreover, for any $h \in \mathcal{H}$ and any $x \in \mathbb{X}$,

$$s_{k_{K,h}}(x) = \int_{\mathbb{X}} s(y) \frac{1}{h} K\left(\frac{x-y}{h}\right) dy = \int_{\mathbb{X}} s(x+zh) K(z) dz \leq \|s\|_\infty \|K\|_1 \ .$$

Therefore, Assumption (13) holds for $\Upsilon \geq 2 \|s\|_\infty \|K\|_1$. Then on one hand

$$\begin{aligned}\left|A_{k_{K,h}}(x,y)\right| &\leq \frac{1}{h^2}\int_{\mathbb{X}} \left|K\left(\frac{x-z}{h}\right) K\left(\frac{y-z}{h}\right)\right| dz \\ &\leq \frac{1}{h}\int_{\mathbb{X}} \left|K\left(\frac{x-y}{h}-u\right) K(u)\right| du \\ &\leq \frac{\|K\|^2}{h} \wedge \frac{\|K\|_\infty \|K\|_1}{h} \leq P\Theta_{k_{K,h}} \wedge n \ .\end{aligned}$$

And on the other hand

$$\begin{aligned}\mathbb{E}\left[\left|A_{k_{K,h}}(X,x)\right|\right] &\leq \frac{1}{h}\int_{\mathbb{X}^2} \left|K\left(\frac{x-y}{h}-u\right) K(u)\right| du\ s(y)dy \\ &= \int_{\mathbb{X}^2} |K(v)\,K(u)|\, s(x+h(v-u))dudv \leq \|s\|_\infty \|K\|_1^2 \ .\end{aligned}$$

Therefore,

$$\begin{aligned}\sup_{x\in\mathbb{X}} \mathbb{E}\left[A_{k_{K,h}}(X,x)^2\right] &\leq \sup_{(x,y)\in\mathbb{X}^2} \left|A_{k_{K,h}}(x,y)\right| \sup_{x\in\mathbb{X}} \mathbb{E}\left[\left|A_{k_{K,h}}(X,x)\right|\right] \\ &\leq \left(P\Theta_{k_{K,h}} \wedge n\right) \|s\|_\infty \|K\|_1^2 \ ,\end{aligned}$$

and $\mathbb{E}\left[A_{k_{K,h}}(X,Y)^2\right] \leq \sup_{x\in\mathbb{X}} \mathbb{E}\left[A_{k_{K,h}}(X,x)^2\right] \leq \|s\|_\infty \|K\|_1^2 P\Theta_{k_{K,h}}$. Hence Assumption (14) and (15) hold when $\Upsilon \geq \|s\|_\infty \|K\|_1^2$. Finally let us prove that Assumption (16) is satisfied. Let $t \in \mathbb{B}_{k_{K,h}}$ and $a \in L^2$ be such that $\|a\| = 1$ and $t(y) = \int_{\mathbb{X}} a(x)\frac{1}{h}K\left(\frac{x-y}{h}\right)dx$ for all $y \in \mathbb{X}$. Then the following follows from Cauchy-Schwarz inequality

$$t(y) \leq \frac{1}{h}\sqrt{\int_{\mathbb{X}} a(x)^2 dx}\sqrt{\int_{\mathbb{X}} K\left(\frac{x-y}{h}\right)^2 dx} \leq \frac{\|K\|}{\sqrt{h}} \ .$$

Thus for any $t \in \mathbb{B}_{k_{K,h}}$

$$Pt^2 \leq \|t\|_\infty \langle |t|, s\rangle \leq \frac{\|K\|}{\sqrt{h}} \|s\| = \|s\| \sqrt{P\Theta_{k_{K,h}}} \leq \sqrt{\Upsilon P\Theta_{k_{K,h}}} \ .$$

We conclude that all the assumptions hold if

$$\Upsilon \geq \left(K(0)/\|K\|^2\right) \vee \left(1 + 2\|s\|_\infty \|K\|_1^2\right) \ .$$

A.4. Proof of Proposition 3.4

Let us define for convenience $\Phi(x) := \sum_{i=1}^p \varphi_i(x)^2$, so $\Gamma \geq 1 \vee n^{-1}\|\Phi\|_\infty$. Then we have for these kernels: $\Phi(x) \geq \chi_{k_w}(x) \geq \Theta_{k_w}(x)$ for all $x \in \mathbb{X}$. Moreover, denoting by Πs the orthogonal projection of s onto the linear span of $(\varphi_i)_{i=1,\ldots,p}$,

$$\|s_{k_w}\|^2 = \sum_{i=1}^p w_i^2 (P\varphi_i)^2 \leq \|\Pi s\|^2 \leq \|s\|^2 \leq \|s\|_\infty \ .$$

Assumption (11) holds for this family if $\Upsilon \geq \Gamma(1 + \|s\|_\infty)$. We prove in what follows that all the remaining assumptions are valid using only (1) and (11).

First, it follows from Cauchy-Schwarz inequality that, for any $x \in \mathbb{X}$, $\chi_{k_w}(x)^2 \leq \Phi(x)\Theta_{k_w}(x)$. Assumption (12) is then automatically satisfied from the definition of Γ

$$\mathbb{E}\left[\chi_{k_w}(X)^2\right] \leq \|\Phi\|_\infty P\Theta_{k_w} \leq \Gamma n P\Theta_{k_w} \ .$$

Now let w and w' be any two vectors in $[0,1]^p$, we have

$$s_{k_w} = \sum_{i=1}^p w_i (P\varphi_i)\varphi_i, \qquad s_{k_w} - s_{k_{w'}} = \sum_{i=1}^p (w_i - w_i')\left(P\varphi_i\right)\varphi_i \ .$$

Hence $\left\|s_{k_w} - s_{k_{w'}}\right\|^2 = \sum_{i=1}^p (w_i - w_i')^2 \left(P\varphi_i\right)^2$ and, by Cauchy-Schwarz inequality, for any $x \in \mathbb{X}$,

$$\left|s_{k_w}(x) - s_{k_{w'}}(x)\right| \leq \left\|s_{k_w} - s_{k_{w'}}\right\| \sqrt{\Phi(x)} \leq \left\|s_{k_w} - s_{k_{w'}}\right\| \sqrt{\Gamma n} \ .$$

Assumption (13) follows using (11). Concerning Assumptions (14) and (15), let us first notice that by orthonormality, for any $(x,x') \in \mathbb{X}^2$,

$$A_{k_w}(x,x') = \sum_{i=1}^p w_i^2 \varphi_i(x)\varphi_i(x') \ .$$

Therefore, Assumption (15) holds since

$$\begin{aligned}
\mathbb{E}\left[A_{k_w}(X,x)^2\right] &= \int_{\mathbb{X}} \left(\sum_{i=1}^p w_i^2 \varphi_i(y)\varphi_i(x)\right)^2 s(y)d\mu(y) \\
&\leq \|s\|_\infty \sum_{1\leq i,j\leq p} w_i^2 w_j^2 \varphi_i(x)\varphi_j(x) \int_{\mathbb{X}} \varphi_i(y)\varphi_j(y)d\mu(y) \\
&= \|s\|_\infty \sum_{i=1}^p w_i^4 \varphi_i(x)^2 \leq \|s\|_\infty \Phi(x) \leq \|s\|_\infty \Gamma n \ .
\end{aligned}$$

Assumption (14) also holds from similar computations:

$$\begin{aligned}
\mathbb{E}\left[A_{k_w}(X,Y)^2\right] &= \int_{\mathbb{X}} \mathbb{E}\left[\left(\sum_{i=1}^p w_i^2 \varphi_i(X)\varphi_i(x)\right)^2\right] s(x)d\mu(x) \\
&\leq \|s\|_\infty \sum_{1\leq i,j\leq p} w_i^2 w_j^2 \mathbb{E}\left[\varphi_i(X)\varphi_j(X)\right] \int_{\mathbb{X}} \varphi_i(x)\varphi_j(x)d\mu(x) \\
&\leq \|s\|_\infty P\Theta_{k_w} \ .
\end{aligned}$$

We finish with the proof of (16). Let us prove that $\mathbb{B}_{k_w} = \mathcal{E}_{k_w}$, where

$$\mathcal{E}_{k_w} = \left\{ t = \sum_{i=1}^p w_i t_i \varphi_i, \ \text{ s.t. } \ \sum_{i=1}^p t_i^2 \leq 1 \right\} \ .$$

First, notice that any $t \in \mathbb{B}_{k_w}$ can be written

$$\int_{\mathbb{X}} a(x) k_w(x,y)d\mu(x) = \sum_{i=1}^p w_i \langle a, \varphi_i\rangle \varphi_i(y) \ .$$

Then, consider some $t \in \mathcal{E}_{k_w}$. By definition, there exists a collection $(t_i)_{i=1,\dots,p}$ such that $t = \sum_{i=1}^p w_i t_i \varphi_i$, and $\sum_{i=1}^p t_i^2 \leq 1$. If $a = \sum_{i=1}^p t_i \varphi_i$, $\|a\|^2 = \sum_{i=1}^p t_i^2 \leq 1$ and $\langle a, \varphi_i\rangle = t_i$, hence $t \in \mathbb{B}_{k_w}$. Conversely, for $t \in \mathbb{B}_{k_w}$, there exists some function $a \in L^2$ such that $\|a\|^2 \leq 1$, and $t = \sum_{i=1}^p w_i \langle a, \varphi_i\rangle \varphi_i$. Since $(\varphi_i)_{i=1,\dots,p}$ is an orthonormal system, one can take $a = \sum_{i=1}^p \langle a, \varphi_i\rangle \varphi_i$.

With $t_i = \langle a, \varphi_i \rangle$, we find $\|a\|^2 = \sum_{i=1}^p t_i^2$ and $t \in \mathcal{E}_{k_w}$. For any $t \in \mathbb{B}_{k_w} = \mathcal{E}_{k_w}$, $\|t\|^2 = \sum_{i=1}^p w_i^2 t_i^2 \leq \sum_{i=1}^p t_i^2 \leq 1$. Hence $Pt^2 \leq \|s\|_\infty \|t\|^2 \leq \|s\|_\infty$.

Appendix B. Concentration of the residual terms

The following proposition gathers the concentration bounds of the remaining terms appearing in (30).

Proposition B.1. *Let $\{k\}_{k\in\mathcal{K}}$ denote a finite collection of kernels satisfying* (1) *and suppose that Assumptions* (11), (12) *and* (13) *hold. Then*

$$\forall \theta \in (0,1), \qquad 2\frac{P(s_{\widehat{k}} - s_k)}{n} \leq \theta \left\|s - s_{\widehat{k}}\right\|^2 + \theta \|s - s_k\|^2 + \frac{2\Upsilon}{\theta n} \ . \tag{32}$$

For any $x \geq 1$, with probability larger than $1 - 2|\mathcal{K}|^2 e^{-x}$, for any $(k,k') \in \mathcal{K}^2$, for any $\theta \in (0,1)$,

$$|2(P_n - P)(s_k - s_{k'})| \leq \theta \left(\|s - s_{k'}\|^2 + \|s - s_k\|^2 \right) + \square \frac{\Upsilon x^2}{\theta n} \ . \tag{33}$$

For any $x \geq 1$, with probability larger than $1 - 2|\mathcal{K}| e^{-x}$, for any $k \in \mathcal{K}$,

$$\forall \theta \in (0,1), \qquad |2(P_n - P)\chi_k| \leq \theta P \Theta_k + \square \frac{\Upsilon x}{\theta} \ . \tag{34}$$

For any $x \geq 1$, with probability larger than $1 - 5.4|\mathcal{K}| e^{-x}$, for any $k \in \mathcal{K}$,

$$\forall \theta \in (0,1), \qquad \frac{2|U_k|}{n^2} \leq \theta \frac{P\Theta_k}{n} + \square \frac{\Upsilon x^2}{\theta n} \ . \tag{35}$$

Proof First for (32), notice that, by (13), for any $\theta \in (0,1)$

$$\begin{aligned} 2\frac{P(s_{\widehat{k}} - s_k)}{n} &\leq 2\frac{\left\|s_{\widehat{k}} - s_k\right\|_\infty}{n} \leq \frac{2}{n}\left(\Upsilon \vee \left(\frac{\theta}{4} n \left\|s_k - s_{\widehat{k}}\right\|^2 + \frac{\Upsilon}{\theta} \right) \right) \\ &\leq \frac{\theta}{2} \left\|s_k - s_{\widehat{k}}\right\|^2 + \frac{2\Upsilon}{\theta n} \leq \theta \left\|s - s_{\widehat{k}}\right\|^2 + \theta \|s - s_k\|^2 + \frac{2\Upsilon}{\theta n} \ . \end{aligned}$$

Then, by Proposition 2.1, with probability larger than $1 - |\mathcal{K}|^2 e^{-x}$,

$$\text{for any } (k,k') \in \mathcal{K}^2, \quad (P_n - P)(s_k - s_{k'}) \leq \sqrt{\frac{2P(s_k - s_{k'})^2 x}{n}} + \frac{\|s_k - s_{k'}\|_\infty x}{3n} \ .$$

Since by (11) $P(s_k - s_{k'})^2 \leq \|s\|_\infty \|s_k - s_{k'}\|^2 \leq \Upsilon \|s_k - s_{k'}\|^2$,

$$\sqrt{\frac{2P(s_k - s_{k'})^2 x}{n}} \leq \frac{\theta}{4} \|s_k - s_{k'}\|^2 + \frac{2\Upsilon x}{\theta n} \ .$$

Moreover, by (13) $\frac{\|s_k - s_{k'}\|_\infty x}{3n} \leq \frac{\theta}{4} \|s_k - s_{k'}\|^2 + \square \frac{\Upsilon x^2}{\theta n}$. Hence, for $x \geq 1$, with probability larger than $1 - |\mathcal{K}|^2 e^{-x}$

$$\begin{aligned} (P_n - P)(s_k - s_{k'}) &\leq \frac{\theta}{2} \|s_k - s_{k'}\|^2 + \square \frac{\Upsilon x^2}{\theta n} \\ &\leq \theta \left(\|s - s_{k'}\|^2 + \|s - s_k\|^2 \right) + \square \frac{\Upsilon x^2}{\theta n} \ , \end{aligned}$$

which gives (33). Now, using again Proposition 2.1, with probability larger than $1-|\mathcal{K}|\,e^{-x}$, for any $k\in\mathcal{K}$,

$$(P_n-P)\chi_k \le \sqrt{\frac{2P\left(\chi_k\right)^2 x}{n}}+\frac{\|\chi_k\|_\infty x}{3n}\ .$$

By (1) and (11), for any $k\in\mathcal{K}$, $\|\chi_k\|_\infty \le \sup_{(x,y)\in\mathbb{X}^2}|k(x,y)|\le \Gamma n\le \Upsilon n$.
Concerning (34), we get by (12), $P\chi_k^2\le \Upsilon n P\Theta_k$, hence, for any $x\ge 1$ we have with probability larger than $1-|\mathcal{K}|\,e^{-x}$

$$(P_n-P)\chi_k \le \theta P\Theta_k+\left(\frac{1}{3}+\frac{1}{2\theta}\right)\Upsilon x\ .$$

For (35), we apply Proposition 2.2 to obtain with probability larger than $1-2.7\,|\mathcal{K}|\,e^{-x}$, for any $k\in\mathcal{K}$,

$$\frac{U_k}{n^2}\le \frac{\square}{n^2}\left(C\sqrt{x}+Dx+Bx^{3/2}+Ax^2\right)\ ,$$

where A,B,C,D are defined accordingly to Proposition 2.2. Let us evaluate all these terms. First, $A\le 4\sup_{(x,y)\in\mathbb{X}^2}|k(x,y)|\le 4\Upsilon n$ by (1) and (11).
Next, $C^2\le \square n^2\mathbb{E}\left[k(X,Y)^2\right]\le \square n^2\|s\|_\infty P\Theta_k\le \square n^2\Upsilon P\Theta_k$.
Using (1), we find $B^2\le 4n\sup_{x\in\mathbb{X}}\int k(x,y)^2 s(y)d\mu(y)\le 4n\|s\|_\infty\Gamma$.
By (11), we consequently have $B^2\le 4\Upsilon n$. Finally, using Cauchy-Schwarz inequality and proceeding as for C^2,

$$\mathbb{E}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}a_i(X_i)b_j(X_j)k(X_i,X_j)\right]\le n\sqrt{\mathbb{E}\left[k(X,Y)^2\right]}\le n\sqrt{\Upsilon P\Theta_k}\ .$$

Hence, $D\le n\sqrt{\Upsilon P\Theta_k}$ which gives (35).

References

[1] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

[Ada06] R. Adamczak. Moment inequalities for U-statistics. *Ann. Probab.*, 34(6):2288–2314, 2006.

[AM09] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279, 2009.

[Bir06] L. Birgé. Statistical estimation with model selection. *Indag. Math. (N.S.)*, 17(4):497–537, 2006.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013.

[BM07] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[DJKP96] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, 2001.

[DO13] P. Deheuvels and S. Ouadah. Uniform-in-bandwidth functional limit laws. *J. Theoret. Probab.*, 26(3):697–721, 2013.

[EL99a] P. P. B. Eggermont and V. N. LaRiccia. Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Trans. Inform. Theory*, 45(4):1321–1326, 1999.

[EL99b] P. P. B. Eggermont and V. N. LaRiccia. Optimal convergence rates for good's nonparametric maximum likelihood density estimator. *Ann. Statist.*, 27(5):1600–1615, 1999.

[EL01] P. P. B. Eggermont and V. N. LaRiccia. *Maximum penalized likelihood estimation*, volume I of *Springer Series in Statistics*. Springer-Verlag, New York, 2001.

[FT06] M. Fromont and C. Tuleau. Functional classification with margin conditions. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 94–108. Springer, Berlin, 2006.

[GL11] A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

[GLZ00] E. Giné, R. Latała, and J. Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, 2000.

[GN09] E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. *Ann. Probab.*, 37(4):1605–1646, 2009.

[GN15] E Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University press, 2015.

[HRB03] C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 55–69. Birkhäuser, Basel, 2003.

[Ler12] M. Lerasle. Optimal model selection in density estimation. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(3):884–908, 2012.

[Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School in Saint-Flour.

[MS11] D. M. Mason and J. W. H. Swanepoel. A general result on the uniform in bandwidth consistency of kernel-type function estimators. *TEST*, 20(1):72–94, 2011.

[MS15] D. M. Mason and J. W. H. Swanepoel. Erratum to: A general result on the uniform in bandwidth consistency of kernel-type function estimators . *TEST*, 24(1):205–206, 2015.

[Rig06] P. Rigollet. Adaptive density estimation using the blockwise Stein method. *Bernoulli*, 12(2):351–370, 2006.

[RT07] P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.

[Tsy09] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

M. Lerasle
Univ. Nice Sophia Antipolis LJAD CNRS UMR 7351
06100 Nice France
e-mail: mlerasle@unice.fr

N. Magalhães
INRIA, Select Project
Univ. Paris-Sud 11
Departement de Mathematiques d'Orsay
91405 Orsay Cedex - France
e-mail: nelo.moltermagalhaes@gmail.com

P. Reynaud-Bouret
Univ. Nice Sophia Antipolis LJAD CNRS UMR 7351
06100 Nice France
e-mail: Patricia.REYNAUD-BOURET@unice.fr