Minimal penalty for Goldenshluger-Lepski method

Claire Lacour, Pascal Massart

Minimal penalty for Goldenshluger-Lepski method

C. Lacour[(1)] & P. Massart[(1)]

Januar 2015

Abstract

This paper is concerned with adaptive nonparametric estimation using the Goldenshluger-Lepski methodology. This method is designed to select an estimator among a collection $(\hat{f}_h)_{h\in\mathcal{H}}$ by minimizing $B(h)+V(h)$ with $B(h)=\sup\{[\|\hat{f}_{h'}-\hat{f}_h\|-V(h')]_+,\ h'\in\mathcal{H}\}$ and $V(h)$ a variance term. In the case of density estimation with kernel estimators, it is shown that the procedure fails if the variance term is chosen too small: this gives for the first time a minimal penalty for the Goldenshluger-Lepski methodology. Some simulations illustrate the theoretical results.

1 Introduction

A challenging task in nonparametric estimation is the data-driven selection of an estimator among a collection. Selecting a bandwidth for kernel estimators, or a level resolution for wavelet estimators, is of crucial importance for theoretical and practical issues. The most (theoretically-justified) known methods for adaptive estimation are wavelet thresholding (Donoho et al., 1996), Lepski's method (Lepskiĭ, 1990), and model selection (Barron et al., 1999) (see also Birgé (2001) for the link between model selection and Lepski's method). A more recent procedure is the one introduced by Goldenshluger and Lepski (2008). This method proposes a data-driven choice of h to select an estimator among a collection $(\hat{f}_h)_{h\in\mathcal{H}}$. To sum up, the selected $\hat{h}$ is chosen as a minimizer of $B(h)+V(h)$ with

$$B(h)=\sup\{[\|\hat{f}_{h'}-\hat{f}_{h,h'}\|^2-V(h')]_+,\ h'\in\mathcal{H}\}$$

where x_+ denotes the positive part $\max(x,0)$ and where $\hat{f}_{h,h'}$ are oversmoothed auxiliary estimators and $V(h)$ is a penalty term (called "majorant") to be suitably chosen. They first develop their methodology in white noise model (Goldenshluger and Lepski, 2008, 2009), next for density estimation (Goldenshluger and Lepski, 2011) and then for various models (Goldenshluger and Lepski, 2013). Their initial objective was to provide an adaptive procedure for multivariate and anisotropic estimation. They use it to give minimax rates of convergence in very general framework (see Goldenshluger and Lepski, 2014). To this purpose, they have established oracle inequalities to ensure that, if $V(h)$ is large enough, the final estimator $\hat{f}_{\hat{h}}$ is almost as efficient as the best one in the collection. The Goldenshluger-Lepski methodology has already been fruitfully applied in various contexts: transport-fragmentation equations (Doumic et al., 2012), anisotropic deconvolution (Comte and Lacour, 2013), warped bases regression (Chagny, 2013) among others (see also

[(1)] Laboratoire de Mathématiques d'Orsay, Université Paris-Sud, France

Bertin et al. (2015) which contains some explanation on the methodology). We cannot close this paragraph without cite the nice work of Laurent et al. (2008), who have independently introduce a very similar method, in order to adapt model selection in a pointwise framework.

In this paper we focus on the calibration of the penaly term V. It is known that the method achieves good results for V large enough. But what is the minimal (and the optimal) value for V to keep this good behavior? We consider this issue from a theoretical point of view but actually it is decisive for a practical implementation of the method. The main contribution of this paper is to evidence an explosion phenomenon: if the penalty term V is chosen smaller than some critical V_0, the risk $\|f - \hat{f}_{\hat{h}}\|$ is proven to dramatically increase, though for $V > V_0$ this risk is quasi-optimal. Proofs are extensively based on concentration inequalities. In particular, left tail concentration inequalities are used to prove the explosion result. We also implement numerical simulations which corroborate this behavior.

We assume here that the function to estimate is univariate and we study the Goldenshluger-Lepski methodology without oversmoothing. That is to say that we do not use auxiliary estimators. Indeed, this is not the heart of the method, and only induces slight changes in the bias term in our context. Thus the precise procedure we study is the following one: the selected $\hat{h}$ is chosen as a minimizer of $B(h) + V(h)$ where

$$B(h) = \sup\{[\|\hat{f}_h - \hat{f}_{h'}\|^2 - V(h')]_+,\ h' \in \mathcal{H}\}.$$

The term $V(h)$ is chosen proportional to the variance of $\hat{f}_h$.

We first present some heuristics in Section 2 in order to well understand the working of the method. In Section 3 we recall the oracle inequality that can be obtained in the framework of density estimation. Then Section 4 contains our main theorem about minimal penalty. This result is illustrated by some simulations (Section 5). Finally, some proofs are gathered in Section 7 after some concluding remarks.

2 Heuristics in the Gaussian white noise model

In this section we consider the following Gaussian white noise model. Indeed, the Euclidean structure in this framework allows to better understand the phenomena at play. Let

$$dY_\varepsilon(x) = f(x)dx + \varepsilon dW(x)$$

where f is the signal to estimate, W the standard Brownian motion and ε the noise level. By projection on an orthonormal basis (φ_j), we derive the classical Gaussian sequence model

$$y_j = \int f(x)\varphi_j(x)dx + \varepsilon\xi_j$$

where the ξ_j are i.i.d. normal variables. We consider subspaces of L^2 for $1 \leq m \leq N$: $S_m := \mathrm{Span}(\varphi_j, \ldots, \varphi_{D_m})$ where the dimensions D_m are assumed to be ordered, such that $m \leq m' \Leftrightarrow S_m \subset S_{m'}$. Then we can introduce the estimators of f:

$$\hat{f}_m = \sum_{j=1}^{D_m} y_j\varphi_j,$$

and

$$f_m = \mathbb{E}(\hat{f}_m) = \sum_{j=1}^{D_m} \left(\int f(x)\varphi_j(x)dx \right) \varphi_j$$

the projection of f on the vectorial space $S_m := \text{Span}(\varphi_j, \ldots, \varphi_{D_m})$.

Denoting $\|.\|$ the L^2 norm, we are interested in the following procedure:

$$B(m) = \sup_{m' \geq m} \{\|\hat{f}_{m'} - \hat{f}_m\|^2 - a\varepsilon^2 D_{m'}\}_+,$$
$$\hat{m} = \arg\min_m \{B(m) + a\varepsilon^2 D_m\},$$

with a a tuning parameter of interest. This is simply the Goldenshluger-Lepski method in this context. Here $V(m) = a\varepsilon^2 D_m = a \int \text{Var}(\hat{f}_m)$. For the following computations, we shall use Pythagorean theorem which leads to

$$\forall m' \geq m \qquad \|\hat{f}_{m'} - \hat{f}_m\|^2 = \|\hat{f}_{m'}\|^2 - \|\hat{f}_m\|^2,$$

and

$$\forall m' \geq m \qquad \|f_{m'} - f_m\|^2 = \|f - f_m\|^2 - \|f - f_{m'}\|^2 = \|f_{m'}\|^2 - \|f_m\|^2.$$

In the sequel, we attempt to give some links between this Goldenshluger-Lepski method and others adaptive methods, and we provide some heuristic insight of the behavior of the method.

- Let us first observe the link with the classical Lepski method (Lepskiĭ, 1990). Using the previous equations, we can rewrite

  $$B(m) = \sup_{m' \geq m} \{\|\hat{f}_{m'}\|^2 - \|\hat{f}_m\|^2 - a\varepsilon^2 D_{m'}\}_+.$$

  Thus B is a nonincreasing function with m. It vanishes from some point $\bar{m}$. This $\bar{m}$ is exactly the one of the Lepski method:

  $$\bar{m} = \min\{m, \quad \forall m' \geq m \quad \|\hat{f}_m - \hat{f}_{m'}\|^2 \leq a\varepsilon^2 D_{m'}\}.$$

  It is easy to see that $\hat{m} \leq \bar{m}$. Actually, when $a > 1$, $\hat{m} = \bar{m}$ with high probability.

- We can also find a link with the classical model selection method (Barron et al., 1999). If we remove the positive part, we can introduce:

  $$B_2(m) = \sup_{m' \geq m} \{\|\hat{f}_{m'} - \hat{f}_m\|^2 - a\varepsilon^2 D_{m'}\}$$

  which is equal to

  $$B_2(m) = \sup_{m' \geq m} \{\|\hat{f}_{m'}\|^2 - a\varepsilon^2 D_{m'}\} - \|\hat{f}_m\|^2.$$

  Then, denoting by $Cr(m) = -\|\hat{f}_m\|^2 + a\varepsilon^2 D_m$, we obtain

  $$\hat{m}_2 = \arg\min_m \sup_{m' \geq m} \{Cr(m) - Cr(m')\} = \arg\min_m Cr(m).$$

  This is the classical model selection method.

Now let us study the mean behavior of $B(m)$. For all $m' \geq m$,

$$\mathbb{E}\|\hat{f}_{m'} - \hat{f}_m\|^2 = \sum_{D_m+1}^{D_{m'}} \mathbb{E}y_j^2 = \|f_{m'} - f_m\|^2 + \varepsilon^2(D_{m'} - D_m)$$

and concentration results ensure that $\|\hat{f}_{m'} - \hat{f}_m\|^2$ is close to this expectation with high probability. Hence, with great probability $B(m)$ is close to

$$\sup_{m' \geq m} \{\|f_{m'} - f_m\|^2 + \varepsilon^2(D_{m'} - D_m) - a\varepsilon^2 D_{m'}\}_+$$

$$= \sup_{m' \geq m} \{\|f - f_m\|^2 - \|f - f_{m'}\|^2 + (1-a)\varepsilon^2 D_{m'} - \varepsilon^2 D_m\}_+$$

Now let us study the behavior of this quantity according to the value of $1 - a$.

• If $a \leq 1$: in this case the quantity $-\|f - f_{m'}\|^2 + (1-a)\varepsilon^2 D_{m'}$ is non-decreasing with m'. Therefore

$$\sup_{m' \geq m} \{\|f_{m'} - f_m\|^2 + (1-a)\varepsilon^2 D_{m'} - \varepsilon^2 D_m\}_+$$
$$= \{-\|f - f_N\|^2 + (1-a)\varepsilon^2 D_N + \|f - f_m\|^2 - \varepsilon^2 D_m\}_+$$

where N is the biggest model. Then, with great probability, the criterion to minimize is

$$\|f - f_m\|^2 - \varepsilon^2 D_m + a\varepsilon^2 D_m$$

(except when m is very close to N, where it is $a\varepsilon^2 D_m$). Then $\hat{m} \approx N$, which is the worst choice. We shall find again this result in Theorem 3.

• In the case where $a = 1$, we shall assume for the sake of simplicity that $f_N = f$ (that is f belongs to the largest model). Using previous computations, we can see that $B(m)$ is close to $\{\|f - f_m\|^2 - \varepsilon^2 D_m\}_+$. Introduce m^* the "oracle point" such that $\|f - f_{m^*}\|^2 \approx \varepsilon^2 D_{m^*}$ Then

$$B(m) \approx \{\|f - f_m\|^2 - \varepsilon^2 D_m\}_+ = \begin{cases} \|f - f_m\|^2 - \varepsilon^2 D_m & \text{for } m \leq m^*, \\ 0 & \text{for } m > m^*. \end{cases}$$

and the criterion to minimize is close to $\|f - f_m\|^2$ for $m \leq m^*$, and $\varepsilon^2 D_m$ for $m > m^*$. Thus $\hat{m} \approx m^*$.

• If $a > 1$: The penalty term $a\varepsilon^2 D_{m'}$ allows us to compensate the variance $\varepsilon^2(D_{m'} - D_m)$ so that $B(m)$ can be upper bounded with great probability by $\|f_{m'} - f_m\|^2 \leq \|f - f_m\|^2$. Then (1) appears as a bias variance tradeoff and the procedure is efficient, see the oracle inequality in Proposition 1.

This discussion shows that $a = 1$ seems to be a critical value for the penalty term: the method fails if $a < 1$, and succeeds as soon as $a \geq 1$. In the following we come back to the density model with kernel estimators. However we shall see that the behavior of the method is almost the same, replacing projection estimators by kernel ones, ε^2 by $1/n$ and $D_{m'} - D_m$ by $\|K_{h'} - K_h\|^2$.

3 Kernel density estimation framework and upper bound on the risk

We consider independent and identically distributed real variables $X_1, \dots, X_n$ with unknown density f. Let $\|.\|$ the L^2 norm for the Lebesgue measure. For h a bandwidth we can define the classical estimator

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

where K is a kernel and $K_h = K(./h)/h$. Now from $\{\hat{f}_h, h \in \mathcal{H}\}$ the collection of estimators, the procedure is the following. The bias is estimated by

$$B(h) = \sup_{h' \le h} \left[\|\hat{f}_{h'} - \hat{f}_h\|^2 - V(h') \right]_+ \qquad \text{with } V(h') = a\frac{\|K_{h'}\|^2}{n} \tag{1}$$

and the selected bandwidth is

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \{B(h) + V(h)\}. \tag{2}$$

This is the same procedure as for Gaussian white noise model with the classical equivalence $\varepsilon^2 = 1/n$ and also $m = 1/h$ and $\|K_h\|^2/n = \varepsilon D_m$ the variance term. We introduce the following notation:

$$f_h := \mathbb{E}(\hat{f}_h), \quad h_{\min} := \min \mathcal{H}, \quad h_{\max} := \max \mathcal{H}$$
$$D(h) := \max(\sup_{h' \le h} \|f_{h'} - f_h\|, \|f - f_h\|) \le 2 \sup_{h' \le h} \|f_{h'} - f\|$$

We assume that the kernel verifies assumption

(K0) $\int |K| = 1$, $\|K\| < \infty$ and

$$\forall\, 0 \le x \le 1 \qquad \frac{\langle K, K(x.)\rangle}{\|K\|^2} \ge 1.$$

This is verified for classical kernels (Gaussian kernel, rectangular kernel, Epanechnikov kernel, biweight kernel; see Lemma 4). This entails that for all $h' \le h$, $\|K_{h'} - K_h\|^2 \le \|K_{h'}\|^2 - \|K_h\|^2$ which is a key property for our results.

Let us now recall what can be obtained if a is well chosen.

Proposition 1. *Assume that f is uniformly bounded and K verifies* **(K0)**. *Assume that $a > 1$. Then, with probability larger than*

$$1 - 2\sum_{h \in \mathcal{H}} \sum_{h' \le h} \max(e^{-c\sqrt{n}}, e^{-c/h'}),$$

where c is a constant only depending on $\|K\|$, a and $\|f\|_\infty$, the following holds

$$\|\hat{f}_{\hat{h}} - f\| \le C_0 \inf_{h \in \mathcal{H}} \left\{ D(h) + \sqrt{a}\frac{\|K_h\|}{\sqrt{n}} \right\}$$

with $C_0 > 1 + \sqrt{\frac{2a}{a-1}}$. *Moreover, if* $h_{\max}^{-1} \leq \sqrt{n}$, *there exists a positive* $C = C(K, f)$ *such that*

$$\mathbb{E}\|\hat{f}_{\hat{h}} - f\|^2 \leq 2\left(\frac{3a-1}{a-1}\right)^2 \inf_{h\in\mathcal{H}}\left\{D^2(h) + a\frac{\|K_h\|^2}{n}\right\} + C\frac{|\mathcal{H}|^2}{h_{\min}} e^{-\frac{(1-a^{-1})^2}{Ch_{\max}}}$$

For $\mathcal{H} = \{e^{-k}, \lceil 2\log\log n\rceil \leq k \leq \lfloor \log n\rfloor\}$, *the remainder term is bounded by* $e^{-(1-a^{-1})^2(\log n)^2/C'}$.

We recognize in the right members the classical bias variance tradeoff. This oracle inequality shows that the Goldenshluger-Lepski methodology works when $a > 1$.

The proof of Proposition 1 is postponed in Section 7.1, and is based on the following concentration result (adapted from Klein and Rio (2005))

Lemme 2. *Let* $X_1, \ldots, X_n$ *be a sequence of i.i.d. variables and* $\nu(t) = n^{-1}\sum_{i=1}^n [g_t(X_i) - \mathbb{E}(g_t(X_i))]$ *for* t *belonging to a countable set of functions* $\mathcal{F}$. *Assume that for all* $t \in \mathcal{F}$ $\|g_t\|_\infty \leq b$ *and* $\mathrm{Var}(g_t(X_1)) \leq v$. *Denote* $H = \mathbb{E}(\sup_{t\in\mathcal{F}} \nu(t))$. *Then, for any* $\varepsilon > 0$, *for* $H' \geq H$,

$$\mathbb{P}(\sup_{t\in\mathcal{F}} \nu(t) \geq (1+\varepsilon)H') \leq \max\left(\exp\left(-\frac{\varepsilon^2}{6}\frac{nH'^2}{v}\right), \exp\left(-\frac{\min(\varepsilon,1)\varepsilon}{24}\frac{nH'}{b}\right)\right) \tag{3}$$

$$\mathbb{P}(\sup_{t\in\mathcal{F}} \nu(t) \leq H - \varepsilon H') \leq \max\left(\exp\left(-\frac{\varepsilon^2}{6}\frac{nH'^2}{v}\right), \exp\left(-\frac{\min(\varepsilon,1)\varepsilon}{24}\frac{nH'}{b}\right)\right) \tag{4}$$

Moreover

$$\mathrm{Var}(\sup_{t\in\mathcal{F}} \nu(t)) \leq \frac{v}{n} + 4\frac{bH}{n} \tag{5}$$

4 Minimal penalty

In this section, we are interested in finding a minimal penalty $V(h)$, beyond which the procedure fails. Indeed, if a and then $V(h)$ is too small, the minimization of the criterion amounts to minimize the bias, and then to choose the smallest possible bandwidth. This leads to the worst estimator and the risk explodes.

In the following result $h_{\min}$ denotes the smallest bandwidth in $\mathcal{H}$ and is of order $1/n$.

Theorem 3. *Assume that* f *is uniformly bounded. Choose* $\mathcal{H} = \{e^{-k}, \lceil 2\log\log n\rceil \leq k \leq \lfloor \log n\rfloor\}$ *as a set of bandwidths. Consider for* K *the Gaussian kernel, the rectangular kernel, the Epanechnikov kernel or the biweight kernel. If* $a < 1$ *where* a *is defined in* (1), *then, for* n *large enough (depending on* f *and* K), *the selected bandwidth* $\hat{h}$ *satisfies*

$$\exists C > 0 \qquad \mathbb{P}(\hat{h} \geq 3h_{\min}) \leq C(\log n)^2 \exp(-(\log n)^2/C)$$

i.e. $\hat{h} < 3h_{\min}$ *with high probability. Moreover*

$$\liminf_{n\to\infty} \mathbb{E}\|f - \hat{f}_{\hat{h}}\|^2 > 0$$

This theorem is proved in Section 7.2 for more general kernels and bandwith sets. It ensures that the critical value for the parameter a is 1. Beyond this value, the selected bandwidth $\hat{h}$ is of order $1/n$, which is very small (remember that for minimax study of a density with regularity α, the optimal bandwidth is $n^{-1/(2\alpha+1)}$), then the risk cannot tend to 0.

5 Simulations

In this Section, we illustrate the role of tuning parameter a, the constant in the penalty term V. The aim is to observe the evolution of the risk for various values of a. Is the critical value $a = 1$ observable in practice? To do this, we simulate data $X_1, \ldots, X_n$ for several densites f. Next, for a grid of values for a, we compute the selected bandwidth $\hat{h}$, the estimator $\hat{f}_{\hat{h}}$ and the integrated risk $\|\hat{f}_{\hat{h}} - f\|^2$.

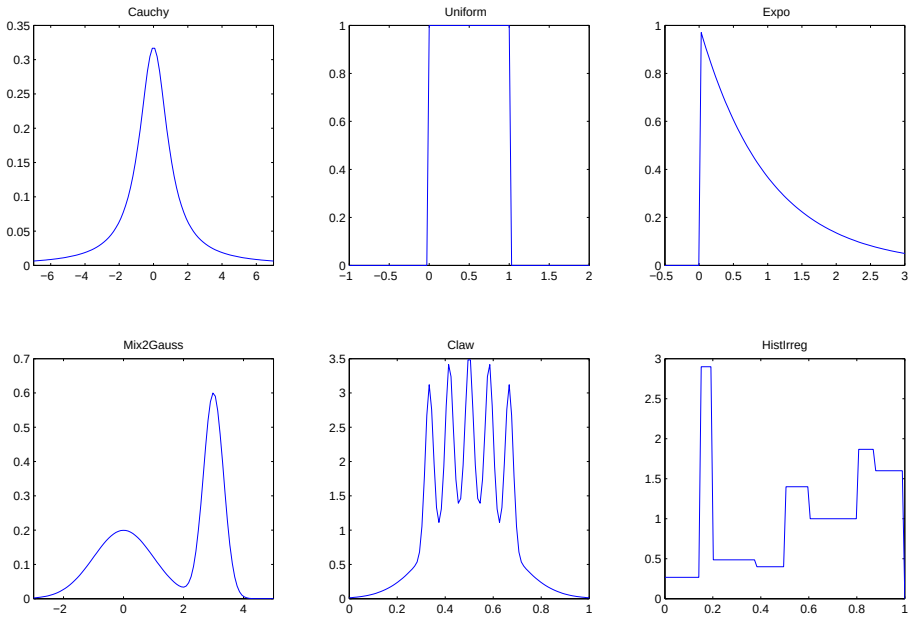

Figure 1: Plots of true density f for Examples 1–6

We consider the following examples, see Figure 1:

Example 1 f is the Cauchy density

Example 2 f is the uniform density $\mathcal{U}(0, 1)$

Example 3 f is the exponential density $\mathcal{E}(1)$

Example 4 f is a mixture of two normal densities $\frac{1}{2}\mathcal{N}(0, 1) + \frac{1}{2}\mathcal{N}(3, 9)$

Example 5 f is a mixture of normal densities sometimes called Claw

Example 6 f is a mixture of eight uniform densities

We implement the method for various kernels, but we only present results for Gaussian kernel, since the choice of kernel does not modify the results. On the other hand, the method is sensitive to the choice of bandwidths set $\mathcal{H}$: here we use

$$\mathcal{H} = \{e^{-k}, 3 \leq k \leq 10\} \cup \{0.002 + k \times 0.02, 0 \leq k \leq 24\}.$$

For $n = 5000$ and $n = 50000$, and several values of a, the Figure 2 plots

$$C_0 = \tilde{\mathbb{E}} \frac{\|\hat{f}_{\hat{h}} - f\|^2}{\min_{h \in \mathcal{H}} \|\hat{f}_h - f\|^2}$$

where $\tilde{\mathbb{E}}$ means the empirical mean on $N = 50$ experiments. Thus smaller C_0 better the estimation. Moreover, we also plot on Figure 3 the selected bandwidth compared to the optimal bandwidth in the selection (for $N = 1$ experiment), i.e.

$$\hat{h} - h_0 \qquad \text{where} \qquad \|\hat{f}_{h_0} - f\|^2 = \min_{h \in \mathcal{H}} \|\hat{f}_h - f\|^2.$$

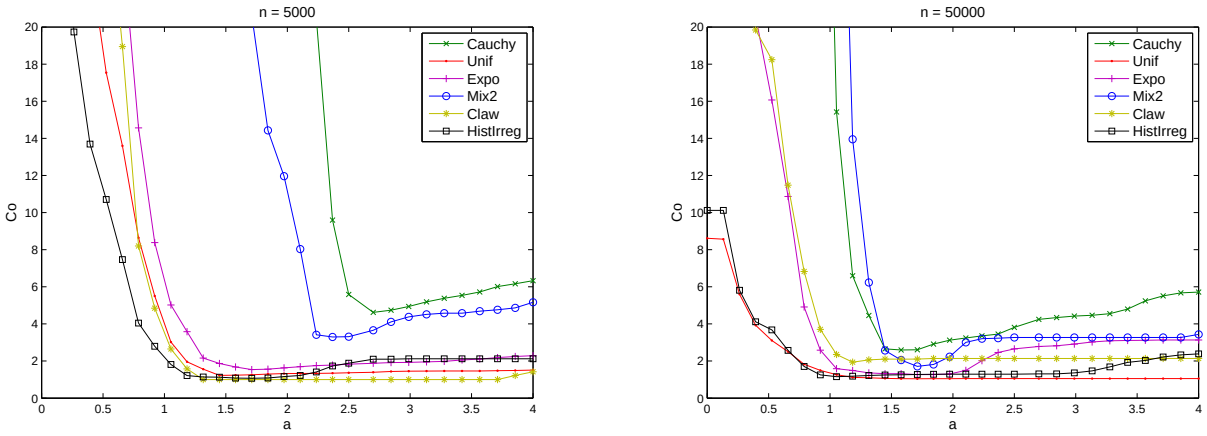

Figure 2: Oracle constant C_0 as a function of a, for Examples 1–6

We can observe that the risk (and then the oracle constant C_0) is very high for small values of a, as expected. Then it jumps to a small value, that indicates the method begins to work well. For too large values of a the risk finally goes back up. Thus we observe in practice what was announced by the theory. Notice that the theory is asymptotic. That is why in practice, the jump may be not exactly at $a = 1$, especially for small values of n. For irregular densities (examples 2, 5, 6), the optimal bandwidth is very low, then it is consistent to observe a smaller jump for the bandwidth choice. However the jump does exist and this is the interesting point.

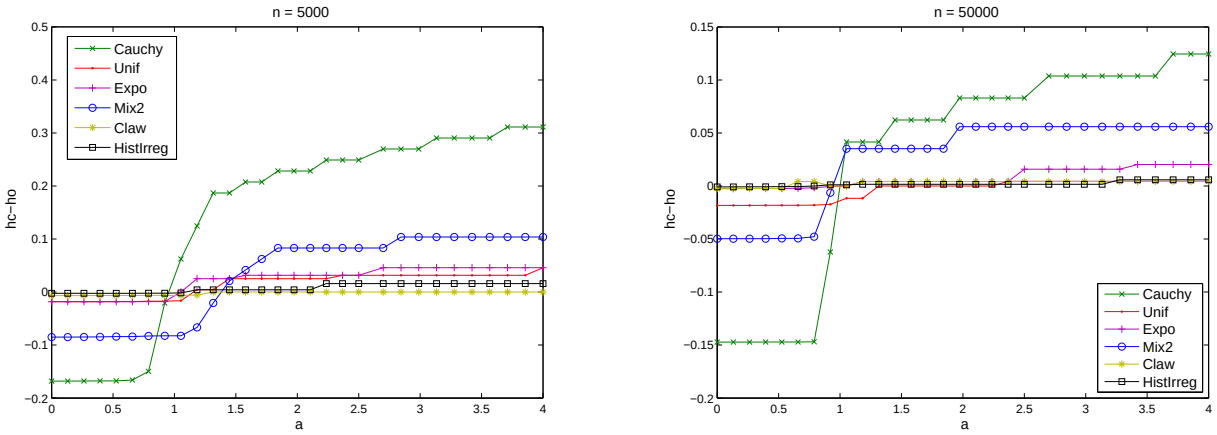

Figure 3: $\hat{h} - h_0$ as a function of a, for Examples 1–6

6 Discussion

To precisely calibrate the penalty V, we face a practical problem: just before $a = 1$, the risk explodes, and just after the result is optimal. Then we can consider another procedure:

$$B(h) = \sup_{h' \leq h} \left[\|\hat{f}_{h'} - \hat{f}_h\|^2 - a \frac{\|K_{h'}\|^2}{n} \right]_+ ,$$
$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ B(h) + b \frac{\|K_h\|^2}{n} \right\} .$$

with $b \neq a$ (here we just study the case $a = b$). The differentiation between a and b could enable a better calibration. Preliminar computations indicate that $a = 1$ and $b = 2$ may be optimal. But addressing the optimality issue requires further developpment that we do not want to give here. A good track for practical purpose seems to use the procedure of Section 3 to find a_0 where there is a jump in the risk: $a_0 = 1$ in the theory but could be slightly different in practice (simulations show that this jump is very perceptible), and then to choose $b = 2a_0$.

7 Proofs

7.1 Proof of Proposition 1

The first step is to write, for some fixed $h \in \mathcal{H}$,

$$\|\hat{f}_{\hat{h}} - f\| \leq \|\hat{f}_{\hat{h}} - \hat{f}_h\| + \|\hat{f}_h - f\|.$$

The last term can be splitted in $\|\hat{f}_h - f_h\| + \|f_h - f\| \leq \|\hat{f}_h - f_h\| + D(h)$. Notice that for all $h' \leq h$, using (1), $\|\hat{f}_{h'} - \hat{f}_h\|^2 \leq B(h) + V(h')$, which can be written , for all h, h';

$$\|\hat{f}_{h'} - \hat{f}_h\|^2 \leq B(h \vee h') + V(h \wedge h')$$

where $h \vee h' = \max(h, h')$ and $h \wedge h' = \min(h, h')$. Then, using (2),

$$\|\hat{f}_{\hat{h}} - \hat{f}_h\|^2 \leq B(h \vee \hat{h}) + V(h \wedge \hat{h}) \leq B(h) + V(h) + \max(B(h), V(h)).$$

We obtain, for any $h \in \mathcal{H}$,

$$\|\hat{f}_{\hat{h}} - f\| \leq \sqrt{2B(h) + 2V(h)} + D(h) + \|\hat{f}_h - f_h\|.$$

Thus the heart of the proof is to control $B(h) = \sup_{h' \leq h} \left[\|\hat{f}_{h'} - \hat{f}_h\|^2 - V(h')\right]_+$ by a bias term. First we center the variables and write, for θ a real in $(0, a-1)$,

$$\|\hat{f}_{h'} - \hat{f}_h\|^2 \leq (1+\theta)\|\hat{f}_{h'} - f_{h'} - \hat{f}_h + f_h\|^2 + (1+\theta^{-1})\|f_{h'} - f_h\|^2$$

Moreover $\|\hat{f}_{h'} - f_{h'} - \hat{f}_h + f_h\| = \sup_{t \in B} \nu(t)$ where B is the unit ball in L^2 and

$$\nu(t) = \langle t, \hat{f}_{h'} - f_{h'} - \hat{f}_h + f_h \rangle = \frac{1}{n} \sum_{i=1}^{n} g_t(X_i) - \mathbb{E}(g_t(X_i))$$

with

$$g_t(X) = \int (K_{h'} - K_h)(x - X)t(x)dx.$$

We shall now use the concentration inequality stated in Lemma 2, with $\mathcal{F}$ a countable set in B such that $\sup_{t \in \mathcal{F}} \nu(t) = \sup_{t \in \mathcal{B}} \nu(t)$. To apply result (3), we need to compute b, H and v.

- For all $y \in \mathbb{R}$, since $t \in B$,

$$|g_t(y)| = |\int (K_{h'} - K_h)(x - y)t(x)dx| \leq \|K_{h'} - K_h\|\|t\| \leq \|K_{h'} - K_h\| \leq \|K_{h'}\|$$

  so that $b = \|K_{h'}\|$. We used assumption **(K0)** which implies, for $h' \leq h$, $\|K_{h'} - K_h\|^2 \leq \|K_{h'}\|^2 - \|K_h\|^2 \leq \|K_{h'}\|^2$.

- Jensen's inequality gives $H^2 \leq \mathbb{E}(\sup_{t \in \mathcal{F}} \nu^2(t))$. Now

$$\begin{aligned}
\sup_{t \in \mathcal{F}} \nu^2(t) &= \|\hat{f}_{h'} - f_{h'} - \hat{f}_h + f_h\|^2 \\
&= \|\frac{1}{n} \sum_{i=1}^{n} (K_{h'} - K_h)(x - X_i) - \mathbb{E}((K_{h'} - K_h)(x - X_i))\|^2 \\
\mathbb{E}(\sup_{t \in \mathcal{F}} \nu^2(t)) &= \int \mathrm{Var}(\frac{1}{n} \sum_{i=1}^{n} (K_{h'} - K_h)(x - X_i))dx \\
&= \frac{1}{n} \int \mathrm{Var}((K_{h'} - K_h)(x - X_1))dx \qquad (6) \\
&\leq \frac{1}{n} \int \mathbb{E}((K_{h'} - K_h)^2(x - X_1))dx \leq \frac{1}{n}\|K_{h'} - K_h\|^2 \leq \frac{1}{n}\|K_{h'}\|^2
\end{aligned}$$

  Then $H^2 \leq n^{-1}\|K_{h'}\|^2$.

- For the variance term, let us write

$$\begin{aligned}\mathrm{Var}(g_t(X_1)) &\leq \mathbb{E}\left[\left(\int (K_{h'}-K_h)(x-X)t(x)dx\right)^2\right]\\ &\leq \mathbb{E}\left[\int |K_{h'}-K_h|(x-X)dx\right]\mathbb{E}\left[\int |K_{h'}-K_h|(x-X)t^2(x)dx\right]\\ &\leq \|K_{h'}-K_h\|_1^2\|f\|_\infty\|t\|^2 \leq 4\|K\|_1^2\|f\|_\infty\|t\|^2\end{aligned}$$

since $\|K_{h'}-K_h\|_1 \leq 2\|K\|_1$. Then $v = 4\|K\|_1^2\|f\|_\infty = 4\|f\|_\infty$.

Finally, using (3), with probability larger than $1-\sum_{h'<h}\max(e^{-\frac{\varepsilon^2\wedge\varepsilon}{24}\sqrt{n}}, e^{-\frac{\varepsilon^2\|K\|^2}{24\|f\|_\infty}\frac{1}{h'}})$

$$\forall h' \leq h \in \mathcal{H} \qquad \|\hat{f}_{h'} - f_{h'} - \hat{f}_h + f_h\| \leq (1+\varepsilon)\frac{\|K_{h'}\|}{\sqrt{n}}$$

where ε is such that $a > (1+\varepsilon)(1+\theta)$. Then, with probability $1-\sum_{h\in\mathcal{H}}\sum_{h'\leq h}\max(e^{-\frac{\varepsilon^2\wedge\varepsilon}{24}\sqrt{n}}, e^{-\frac{\varepsilon^2\|K\|^2}{24\|f\|_\infty}\frac{1}{h'}})$ for any h,

$$B(h) \leq (1+\theta^{-1})D(h)^2$$

In the same way, choosing $\epsilon = \sqrt{a}-1$, we can prove with probability $1-\sum_{h\in\mathcal{H}}\max(e^{-\frac{\epsilon^2\wedge\epsilon}{24}\sqrt{n}}, e^{-\frac{\epsilon^2\|K\|^2}{6\|f\|_\infty}\frac{1}{h}})$ for any h,

$$\|\hat{f}_h - f_h\| \leq (1+\epsilon)\frac{\|K_h\|}{\sqrt{n}} = \sqrt{V(h)}$$

Finally, with high probability,

$$\begin{aligned}\|\hat{f}_{\hat{h}} - \hat{f}_h\| &\leq \sqrt{2(1+\theta^{-1})D(h)^2 + 2V(h)} + D(h) + \sqrt{V(h)}\\ &\leq C_0 D(h) + (\sqrt{2}+1)\sqrt{V(h)}\end{aligned}$$

with $C_0 > 1+\sqrt{2(1+(a-1)^{-1})} = 1+\sqrt{\frac{2a}{a-1}}$. Regarding the second result, note that the rough bound $\|\hat{f}_h\|^2 \leq \|K_h\|^2 \leq \|K\|^2/h_{\min}$ is valid for all h. Then, denoting A the set on which the previous oracle inequality is verified,

$$\mathbb{E}\|\hat{f}_{\hat{h}} - f\|^2 \leq \mathbb{E}\|\hat{f}_{\hat{h}} - f\|^2\mathbb{1}_A + 2(\|f\|^2 + \|K\|^2/h_{\min})\mathbb{P}(A^c)$$

with

$$\mathbb{P}(A^c) \leq 2\sum_{h,h'\in\mathcal{H}}\max(e^{-\frac{\varepsilon^2\wedge\varepsilon}{24}\sqrt{n}}, e^{-\frac{\varepsilon^2\|K\|^2}{24\|f\|_\infty}\frac{1}{h'}}) \leq 2|\mathcal{H}|^2 e^{-2C(K,f)\frac{\varepsilon^2\wedge\varepsilon}{h_{\max}}}$$

It is then sufficient to take $\varepsilon = \frac{a-1}{2a}$ and $C_0 = \frac{3a-1}{a-1}$.

■

7.2 Proof of Theorem 3

We shall prove a more general version of the theorem, where several bandwidths sets $\mathcal{H}$ and kernels K are possible. We denote $\mathrm{Crit}(h) := B(h) + V(h)$ and $E_{\mathcal{H}} = \min\{h/h'; h \in \mathcal{H}, h' \in \mathcal{H}, h > h'\}$. We assume that $E_{\mathcal{H}}$ does not depend on n and is larger than 1 ($\mathcal{H} = \{e^{-k}, a_n \leq k \leq b_n\}$ suits with $E_{\mathcal{H}} = e$). Let us define

$$\phi(x) = \|K\|^{-2}\|K - K_x\|^2 = 1 + \frac{1}{x} - 2\frac{\langle K, K(x.)\rangle}{\|K\|^2}.$$

We assume that the kernel K satisfies :

(K1) the function ϕ is bounded from below over $[E_{\mathcal{H}}, +\infty)$,

(K2) for $0 < \mu < 1$, the function $\phi(x) - \frac{\mu}{x}$ tends to $+\infty$ when $x \to 0$ and is decreasing in some neighborhood of 0,

(K3) for $0 < \mu < 1$, the function $\phi(x) + \frac{\mu}{x}$ is increasing for $x \geq 2$.

These assumptions are mild, as shown in the following Lemma, proved in Section 7.3.

Lemme 4. *The following kernels satisfy assumptions* **(K0–K3)***:*

a - *Gaussian kernel:* $K(x) = e^{-x^2/2}/\sqrt{2\pi}$

b - *Rectangular kernel:* $K(x) = \mathbb{1}_{[-1,1]}(x)/2$

c - *Epanechnikov kernel:* $K(x) = (3/4)(1-x^2)\mathbb{1}_{[-1,1]}(x)$

d - *Biweight kernel:* $K(x) = (15/16)(1-x^2)^2\mathbb{1}_{[-1,1]}(x)$

The general result is:

Theorem 5. *Assume* **(K0–K3)** *and that f is uniformly bounded. Assume that $E_{\mathcal{H}}$ does not depend on n and $h_{\max} \to 0$ when $n \to \infty$. We also assume that there exist $\theta_1 < \theta_2$ reals such that $\theta_2 \geq 2$, $\theta_1.h_{\min} \in \mathcal{H}$ and $\phi(\theta_2) - \phi(\theta_1) \geq 1/\theta_1 - 1/\theta_2$.*

Then, if $a < 1$, for n large enough (depending on $f, \mathcal{H}, K$),

$$\exists C > 0 \qquad \mathbb{P}(\hat{h} \geq \theta_2 h_{\min}) \leq \sum_{h\in\mathcal{H}} \sum_{h'<h} \max(e^{-C\varepsilon^2/\sqrt{n}}, e^{-C\varepsilon^2\|K_{h'} - K_h\|^2})$$

where $\varepsilon < 1 - a^{1/3}$. If $\mathcal{H} = \{e^{-k}, a_n \leq k \leq b_n\}$ and the kernel is Gaussian, rectangular, Epanechnikov or biweight, $\theta_1 = e$ and $\theta_2 = 3$ work.

This results implies Theorem 3, since under **(K1)**, $\|K_{h'} - K_h\|^2 = \frac{\|K\|^2}{h'}\phi(h/h') \geq (\min_{E_{\mathcal{H}}} \phi)\frac{\|K\|^2}{h'}$ as soon as $h > h'$, so that

$$\sum_{h\in\mathcal{H}} \sum_{h'<h} e^{-C\|K_{h'} - K_h\|^2} \leq |\mathcal{H}|^2 e^{-C'/h_{\max}}.$$

Let $\varepsilon \in (0,1)$ such that $a < (1-\varepsilon)^3$ and

$$\varepsilon^3 + 3\varepsilon < \frac{\phi(\theta_2) - \phi(\theta_1) - a/\theta_1 + a/\theta_2}{\phi(\theta_2) + \phi(\theta_1)} \tag{7}$$

(possible since $a < 1 \leq (\phi(\theta_2) - \phi(\theta_1))/(1/\theta_1 - 1/\theta_2)$). Let us decompose

$$\hat{f}_{h'} - \hat{f}_h = (\hat{f}_{h'} - f_{h'} - \hat{f}_h + f_h) + (f_{h'} - f_h) = S(h, h') + (f_{h'} - f_h)$$

with

$$S(h, h') = \frac{1}{n} \sum_{i=1}^{n} (K_{h'} - K_h)(x - X_i) - \mathbb{E}((K_{h'} - K_h)(x - X_i))$$

and the bias term $\|f_{h'} - f_h\| \leq \sup_{h' \leq h} \|K_{h'} * f - K_h * f\| = D(h)$. First write

$$(1 - \varepsilon)\|S(h, h')\|^2 - \left(\frac{1}{\varepsilon} - 1\right) D(h)^2 \leq \|\hat{f}_{h'} - \hat{f}_h\|^2 \leq (1 + \varepsilon)\|S(h, h')\|^2 + \left(1 + \frac{1}{\varepsilon}\right) D(h)^2$$

Now we shall prove that with high probability

$$(1 - \varepsilon)\frac{\|K_{h'} - K_h\|}{\sqrt{n}} \leq \|S(h, h')\| \leq (1 + \varepsilon)\frac{\|K_{h'} - K_h\|}{\sqrt{n}}.$$

First, we can prove as in Section 3 that for all $h' < h$

$$\begin{aligned} &\mathbb{P}\left(\|S(h, h')\| \geq (1 + \varepsilon)\frac{\|K_{h'} - K_h\|}{\sqrt{n}}\right) \\ &\leq \max\left(\exp\left(-\frac{\varepsilon^2 \wedge \varepsilon}{24}\sqrt{n}\right), \exp\left(-\frac{\varepsilon^2}{24\|f\|_\infty}\|K_{h'} - K_h\|^2\right)\right). \end{aligned}$$

Next, we shall use (4) in Lemma 2 in order to lowerbound $\|S(h, h')\|$. Recall that $\|S(h, h')\| = \sup_{t \in B} \nu(t)$ where B is the unit ball in L^2 and $\nu(t) = \frac{1}{n}\sum_{i=1}^{n} g_t(X_i) - \mathbb{E}(g_t(X_i))$ with $g_t(X) = \int (K_{h'} - K_h)(x - X)t(x)dx$. With notations of Lemma 2, we have $b = \|K_{h'} - K_h\|$, $H'^2 = n^{-1}\|K_{h'} - K_h\|^2$ and $v = 4\|K\|_1^2\|f\|_\infty$. It remains to lowerbound H. First, remark, that (6) provides $n\mathbb{E}(\sup_{t \in B} \nu^2(t)) = \|K_{h'} - K_h\|^2 - \|(K_{h'} - K_h) * f\|^2$. Next, using (5)

$$\mathbb{E}(\sup_{t \in B} \nu^2(t)) \leq \frac{v}{n} + 4\frac{bH}{n} + H^2 \leq \frac{v}{n} + \left(H + \frac{2b}{n}\right)^2.$$

Then

$$n\left(H + \frac{2b}{n}\right)^2 \geq n\mathbb{E}(\sup_{t \in B} \nu^2(t)) - v = \|K_{h'} - K_h\|^2 - \|(K_{h'} - K_h) * f\|^2 - 4\|K\|_1^2\|f\|_\infty$$

which implies

$$\sqrt{n}\left(H + \frac{2b}{n}\right) \geq \sqrt{\|K_{h'} - K_h\|^2 - 4\|K\|_1^2(\|f\|_\infty + \|f\|^2)}.$$

Since $b = \|K_{h'} - K_h\|$,

$$H \geq \sqrt{\frac{\|K_{h'} - K_h\|^2 - 4\|K\|_1^2(\|f\|_\infty + \|f\|^2)}{n}} - \frac{2\|K_{h'} - K_h\|}{n}$$

Now, for $h' < h$

$$H \geq \frac{\|K_{h'} - K_h\|}{\sqrt{n}}\left(\sqrt{1 - \frac{4\|K\|_1^2(\|f\|_\infty + \|f\|^2)}{\|K_{h'} - K_h\|^2}} - \frac{2}{\sqrt{n}}\right)$$

so

$$H - \frac{\varepsilon}{3} H' \geq H' \left(\sqrt{1 - \frac{4\|K\|_1^2 (\|f\|_\infty + \|f\|^2)}{\|K_{h'} - K_h\|^2}} - \frac{2}{\sqrt{n}} - \frac{\varepsilon}{3} \right).$$

From **(K1)**, $\|K_{h'} - K_h\|^2 = \frac{\|K\|^2}{h'} \phi(h/h') \geq (\min_{E_{\mathcal{H}}} \phi) \frac{\|K\|^2}{h'} \geq \frac{C}{h_{\max}} \to \infty$ and, in consequence, for n large enough

$$H - \frac{\varepsilon}{3} H' \geq H' (1 - \varepsilon).$$

Thus for n large enough

$$\begin{aligned} &\mathbb{P}\left(\|S(h,h')\| \leq (1-\varepsilon) \frac{\|K_{h'} - K_h\|}{\sqrt{n}} \right) \\ &\leq \max\left(\exp\left(-\frac{\varepsilon^2 \wedge (3\varepsilon)}{24 \times 9} \sqrt{n} \right), \exp\left(-\frac{\varepsilon^2}{24 \times 9 \|f\|_\infty} \|K_{h'} - K_h\|^2 \right) \right) \end{aligned} \tag{8}$$

Let $\delta(h,h) = 0$ and, if $h \neq h'$,

$$\delta(h,h') = 2 \max\left(\exp\left(-\frac{\varepsilon^2 \wedge (3\varepsilon)}{24 \times 9} \sqrt{n} \right), \exp\left(-\frac{\varepsilon^2}{24 \times 9 \|f\|_\infty} \|K_{h'} - K_h\|^2 \right) \right).$$

We just proved that for n large enough, with probability larger than $1 - \delta(h,h')$

$$(1-\varepsilon)^2 \frac{\|K_{h'} - K_h\|^2}{n} \leq \|S(h,h')\|^2 \leq (1+\varepsilon)^2 \frac{\|K_{h'} - K_h\|^2}{n}.$$

Next, with probability larger than $1 - \sum_{h' \leq h} \delta(h,h')$

$$\begin{cases} B(h) \geq \sup_{h' \leq h} \left[(1-\varepsilon)^3 \frac{\|K_{h'} - K_h\|^2}{n} - a \frac{\|K_{h'}\|^2}{n} \right]_+ - \left(\frac{1}{\varepsilon} - 1 \right) D(h)^2 \\ B(h) \leq \sup_{h' \leq h} \left[(1+\varepsilon)^3 \frac{\|K_{h'} - K_h\|^2}{n} - a \frac{\|K_{h'}\|^2}{n} \right]_+ + \left(1 + \frac{1}{\varepsilon} \right) D(h)^2 \end{cases}$$

But, if $h_{\min}$ small enough, for $\lambda > a$

$$\sup_{h' \leq h} \left[\lambda \frac{\|K_{h'} - K_h\|^2}{n} - a \frac{\|K_{h'}\|^2}{n} \right]_+ = \lambda \frac{\|K_{h_{\min}} - K_h\|^2}{n} - a \frac{\|K_{h_{\min}}\|^2}{n}$$

Indeed, for $x = h'/h \leq 1$

$$\begin{aligned} \lambda \frac{\|K_{h'} - K_h\|^2}{n} - a \frac{\|K_{h'}\|^2}{n} &= \lambda \frac{\|K\|^2}{nh} \left(1 + \frac{1 - a/\lambda}{x} - 2 \frac{\langle K, K(x.) \rangle}{\|K\|^2} \right) \\ &= \lambda \frac{\|K\|^2}{nh} \left(\phi(x) - \frac{a/\lambda}{x} \right) \end{aligned}$$

and the function $\phi(x) - \frac{a/\lambda}{x}$ tends to $+\infty$ when $x \to 0$ and is decreasing in some neighborhood of 0 (assumption **(K2)**). Then with probability larger than $1 - \sum_h \sum_{h' \leq h} \delta(h,h')$, for all h

$$\begin{cases} \mathrm{Crit}(h) \geq \frac{\|K\|^2}{nh_{\min}} \left[-a + (1-\varepsilon)^3 \phi(h/h_{\min}) + \frac{a}{h/h_{\min}} \right] - \left(\frac{1}{\varepsilon} - 1 \right) D(h)^2 \\ \mathrm{Crit}(h) \leq \frac{\|K\|^2}{nh_{\min}} \left[-a + (1+\varepsilon)^3 \phi(h/h_{\min}) + \frac{a}{h/h_{\min}} \right] + \left(1 + \frac{1}{\varepsilon} \right) D(h)^2 \end{cases}$$

In particular, for $h = \theta_1 h_{\min}$,

$$\mathrm{Crit}(\theta_1 h_{\min}) \leq \frac{\|K\|^2}{n h_{\min}} \left[-a + (1+\varepsilon)^3 \phi(\theta_1) + \frac{a}{\theta_1} \right] + \left(1 + \frac{1}{\varepsilon}\right) \sup_h D(h)^2. \tag{9}$$

Moreover, since $a < (1-\varepsilon)^3$, $(1-\varepsilon)^3\phi(x) + \frac{a}{x}$ is increasing for $x \geq 2$ (assumption **(K3)**). This implies that

$$\forall h \geq \theta_2 h_{\min}, \quad \mathrm{Crit}(h) \geq \frac{\|K\|^2}{n h_{\min}} \left[-a + (1-\varepsilon)^3 \phi(\theta_2) + \frac{a}{\theta_2} \right] - \left(\frac{1}{\varepsilon} - 1\right) \sup_h D(h)^2. \tag{10}$$

Since (K_h) is an approximation to the identity, $\|f - K_h * f\|$ tends to 0 when h tends to 0. This implies that $D(h) \leq 2 \sup_{h' \leq h} \|f - K_{h'} * f\|$ tends to 0 and $\sup_{h \in \mathcal{H}} D(h)$ tends to 0, as soon as $h_{\max}$ tends to 0. Now (7) leads to $\Delta := (1-\varepsilon)^3\phi(\theta_2) + \frac{a}{\theta_2} - (1+\varepsilon)^3\phi(\theta_1) - \frac{a}{\theta_1} > 0$. Then, for n large enough, $(2/\varepsilon)\sup_h D(h)^2 < \frac{\|K\|^2}{n h_{\min}} \Delta$ so that

$$\begin{aligned} &\frac{\|K\|^2}{n h_{\min}} \left[-a + (1+\varepsilon)^3 \phi(\theta_1) + \frac{a}{\theta_1} \right] + \left(1 + \frac{1}{\varepsilon}\right) \sup_h D(h)^2 \\ &< \frac{\|K\|^2}{n h_{\min}} \left[-a + (1-\varepsilon)^3 \phi(\theta_2) + \frac{a}{\theta_2} \right] - \left(\frac{1}{\varepsilon} - 1\right) \sup_h D(h)^2 \end{aligned} \tag{11}$$

Finally, combining (9) and (10) and (11) gives $\hat{h} < \theta_2 h_{\min}$ with probability larger than $1 - \sum_h \sum_{h' \leq h} \delta(h, h')$.

Let us now prove the second part of Theorem 3, that is the lower bound on the risk. Let $A_n = \{\hat{h} \leq 3h_{\min}\}$ and $B_n = \cap_{h \in \mathcal{H}} \{\|f_h - \hat{f}_h\| \geq \frac{1}{2} \frac{\|K_h\|}{\sqrt{n}}\}$. We have just proved that $\mathbb{P}(A_n^c) \leq C(\log n)^2 \exp(-(\log n)^2/C)$. In the same way that (8), we can write for n large enough

$$\begin{aligned} &\mathbb{P}\left(\|f_h - \hat{f}_h\| \leq (1-\varepsilon) \frac{\|K_h\|}{\sqrt{n}} \right) \\ &\leq \max\left(\exp\left(-\frac{\varepsilon^2 \wedge (3\varepsilon)}{24 \times 9} \sqrt{n} \right), \exp\left(-\frac{\varepsilon^2}{6 \times 9 \|f\|_\infty} \|K_h\|^2 \right) \right) \end{aligned}$$

which implies $\mathbb{P}(B_n^c) \leq C'(\log n) \exp(-(\log n)^2/C')$ and then

$$\mathbb{P}(A_n \cap B_n) \geq 1 - o(1).$$

Then we can write

$$\begin{aligned} \|f - \hat{f}_{\hat{h}}\| &\geq \|f_{\hat{h}} - \hat{f}_{\hat{h}}\| \mathbb{1}_{A_n \cap B_n} - \|f - f_{\hat{h}}\| \\ &\geq \min_{h \leq 3h_{\min}} \|f_h - \hat{f}_h\| \mathbb{1}_{A_n \cap B_n} - \max_h \|f - f_h\| \\ &\geq \min_{h \leq 3h_{\min}} \frac{1}{2} \frac{\|K_h\|}{\sqrt{n}} \mathbb{1}_{A_n \cap B_n} - \max_h \|f - f_h\| \\ &\geq \frac{\|K\|}{2\sqrt{3}} \frac{1}{\sqrt{n h_{\min}}} \mathbb{1}_{A_n \cap B_n} - \max_h \|f - f_h\| \end{aligned}$$

But $\max_h \|f - f_h\| \to 0$ (since $h_{\max} \to 0$), and $nh_{\min} \to 1$ when $n \to \infty$. Hence

$$\mathbb{E}\|f - \hat{f}_{\hat{h}}\| \quad \geq \quad \frac{\|K\|}{2\sqrt{3}} \frac{\mathbb{P}(A_n \cap B_n)}{\sqrt{1+o(1)}} - o(1)$$

which proves that $\mathbb{E}\|f - \hat{f}_{\hat{h}}\| \geq \frac{\|K\|}{4\sqrt{3}}$ for n large enough.

■

7.3 Proof of Lemma 4

To prove Lemma 4, it is sufficient to do computations on integrals. We obtain:

a - if K is the Gaussian kernel,

$$\frac{\langle K, K(x.)\rangle}{\|K\|^2} = \sqrt{\frac{2}{1+x^2}}.$$

b - if K is the rectangular kernel,

$$\frac{\langle K, K(x.)\rangle}{\|K\|^2} = \frac{1}{x} \wedge 1.$$

c - if K is the Epanechnikov kernel,

$$\frac{\langle K, K(x.)\rangle}{\|K\|^2} = \frac{5}{4}\left[\left(\frac{1}{x} \wedge 1\right) - \frac{x^2}{5}\left(\frac{1}{x} \wedge 1\right)^5\right].$$

d - if K is the biweight kernel:

$$\frac{\langle K, K(x.)\rangle}{\|K\|^2} = \frac{1}{16}\left[21\left(\frac{1}{x} \wedge 1\right) - 6x^2\left(\frac{1}{x} \wedge 1\right)^5 + x^4\left(\frac{1}{x} \wedge 1\right)^9\right].$$

These formulas permit to verify all the assumptions.

■

References

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

Bertin, K., Lacour, C., and Rivoirard, V. (2015). Adaptive pointwise estimation of conditional density function. *Ann. Inst. H. Poincaré Probab. Statist.* To appear.

Birgé, L. (2001). *An alternative point of view on Lepski's method*, volume Volume 36 of *Lecture Notes–Monograph Series*, pages 113–133. Institute of Mathematical Statistics, Beachwood, OH.

Chagny, G. (2013). Penalization versus Goldenshluger-Lepski strategies in warped bases regression. *ESAIM: Probability and Statistics*, 17:328–358.

Comte, F. and Lacour, C. (2013). Anisotropic adaptive kernel deconvolution. *Ann. Inst. H. Poincar Probab. Statist.*, 49(2):569–609.

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

Doumic, M., Hoffmann, M., Reynaud-Bouret, P., and Rivoirard, V. (2012). Nonparametric estimation of the division rate of a size-structured population. *SIAM Journal on Numerical Analysis*, 50(2):925–950.

Goldenshluger, A. and Lepski, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190.

Goldenshluger, A. and Lepski, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields*, 143(1-2):41–71.

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

Goldenshluger, A. and Lepski, O. (2014). On adaptive minimax density estimation on R^d. *Probab. Theory Related Fields*, 159(3-4):479–543.

Goldenshluger, A. V. and Lepski, O. V. (2013). General selection rule from a family of linear estimators. *Theory Probab. Appl.*, 57(2):209–226.

Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077.

Laurent, B., Ludena, C., and Prieur, C. (2008). Adaptive estimation of linear functionals by model selection. *Electronic Journal of Statistics*, 2:993–1020.

Lepskiĭ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.*, 35(3):454–466.