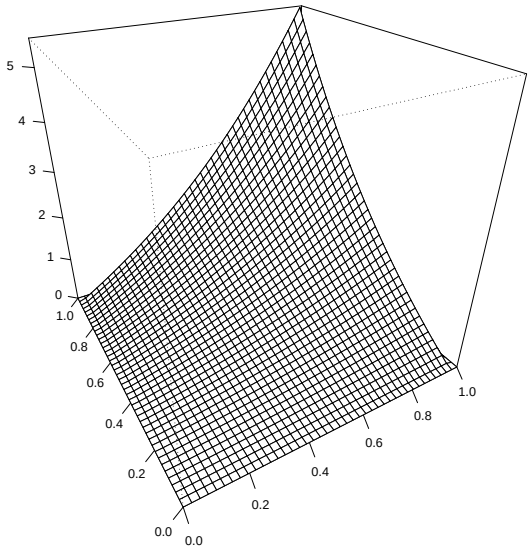


Figure 2: The true density of the first example.

# A Illustrations

**First example.** Figure 2 shows a product density with monotonic marginals. The density is $f(x_1, x_2) = g(x_1)g(x_2)$, where $g(t) = (e^t - 1)/(e - 2)I_{[0,1]}(t)$. We generated a sample of 300 from the distribution of this density. We took the the resolution parameter of the dyadic histogram to be $J = (4, 4)$. Figure 3 shows the histogram with $2^4 = 16$ equispaced bins in each direction. This histogram corresponds to the choice $\alpha = 0$ of the resolution parameter and thus the partition of this histogram is the finest allowed partition. Figure 4 shows the dyadic histogram with $\alpha = 0.025$.

One can see that the partition of the dyadic histogram has been adapted to the underlying distribution. The bins around the mode have the finest allowed resolution, but in the tails the small bins have been combined to reach optimal larger bins. The small bins in the tails have been joined mostly along the coordinate axis.

**Second example.** Figure 5 shows a density which is constant in the x-direction and monotonic in the y-direction. This is a prototypic example of a density which has anisotropic smoothness. The density is $f(x_1, x_2) = g(x_2)I_{[0,1]\times[0,1]}(x_1, x_2)$, where $g(t) = (e^t - 1)/(e - 2)$. We generated a sample

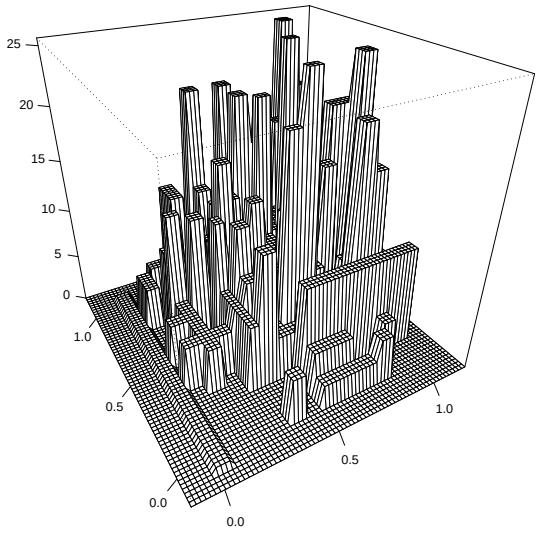


Figure 3: The dyadic histogram with $\alpha = 0$ for the first example.

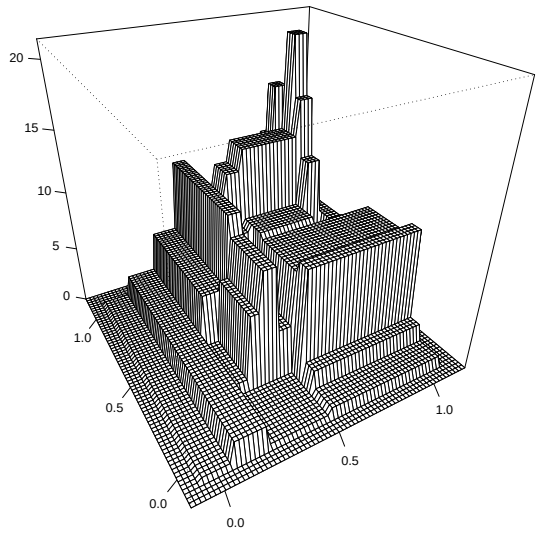


Figure 4: The dyadic histogram with $\alpha = 0.025$ for the first example.

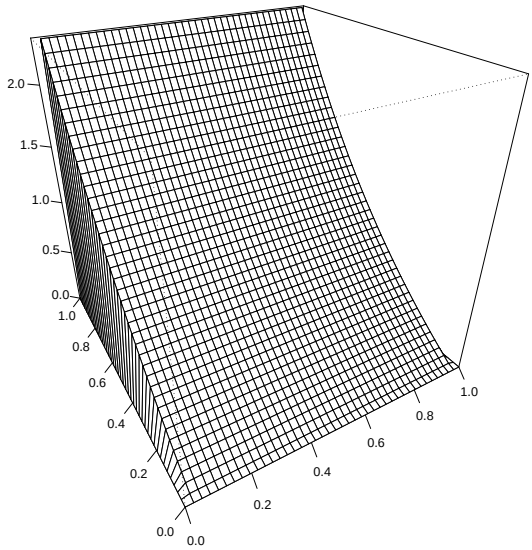


Figure 5: The true density of the second example.

of 300 from the distribution of this density. We took the the resolution parameter of the dyadic histogram to be $J = (4, 4)$. Figure 6 shows the histogram with $2^4 = 16$ equispaced bins in each direction. This histogram corresponds to the choice $\alpha = 0$ of the resolution parameter. Figure 7 shows the dyadic histogram with $\alpha = 0.05$.

One can see that the partition of the dyadic histogram contains splits only in the y-direction: the anisotropicity of the underlying density has been detected.

# B Oracle inequality

## B.1 General setting

We will state an oracle inequality in a general setting, in order to simplify the notation and exposition. We denote

$$\hat{f}_{n,\alpha}(x) = \tilde{f}\left(x, \hat{\Lambda}_{n,\alpha}\right), \qquad x \in \mathbf{R}^d, \tag{51}$$

where

$$\tilde{f}(x, \Lambda) = \sum_{\phi \in \mathcal{D}} \lambda_\phi \phi(x), \qquad x \in \mathbf{R}^d, \tag{52}$$

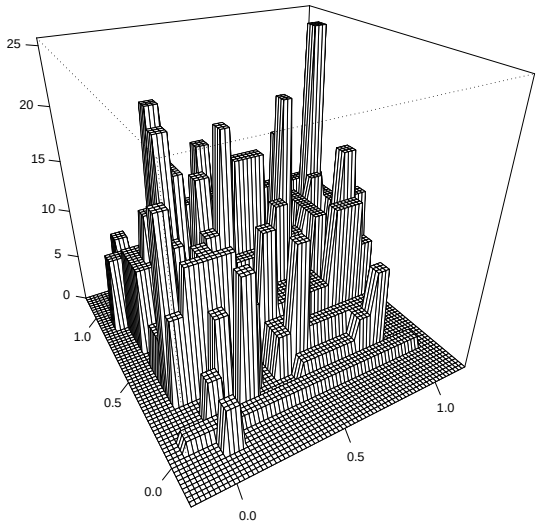


Figure 6: The dyadic histogram with $\alpha = 0$ for the second example.

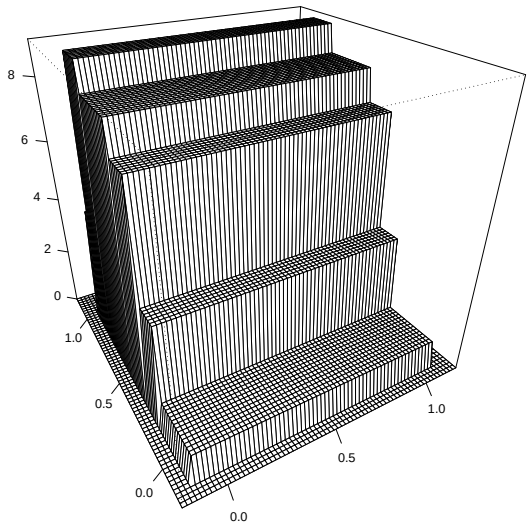


Figure 7: The dyadic histogram with $\alpha = 0.05$ for the second example.

where $\mathcal{D} \subset L_2([0,1]^d)$ is a collection of functions, we will assume that $\mathcal{D}$ has finite cardinality, and $\Lambda = (\lambda_\phi)_{\phi \in \mathcal{D}} \in \mathbf{R}^{\mathcal{D}}$ gives the coefficients of the expansion,

$$\hat{\Lambda}_{n,\alpha} = \text{argmin}_{\Lambda \in \mathbb{K}} \mathcal{E}_n(\Lambda, \alpha), \tag{53}$$
$$\mathcal{E}_n(\Lambda, \alpha) = \gamma_n \left( \tilde{f}(\cdot, \Lambda) \right) + \alpha \cdot D(\Lambda),$$

where $\gamma_n$ is defined in (4), $\alpha \geq 0$,

$$D(\Lambda) = \#\{\lambda_\phi : \lambda_\phi \neq 0\}, \tag{54}$$

and $\mathbb{K} \subset \mathbf{R}^{\mathcal{D}}$.

**Theorem 2** *We have for the estimator $\hat{f}_{n,\alpha}$ defined in (51), based on i.i.d. observations $X^1, \ldots, X^n$ from the distribution of a continuous density $f : [0,1]^d \to \mathbf{R}$, that*

$$E_f \left\| \hat{f}_{n,\alpha_n} - f \right\|_2^2 1_{\tilde{\Omega}} \leq C_1 \inf_{\Lambda \in \mathbb{K}_0} K\left(f, \Lambda, \alpha_n\right) + C_2 n^{-1}, \tag{55}$$

*where*

$$K(f, \Lambda, \alpha) = \left\| \tilde{f}(\,\cdot\,, \Lambda) - f \right\|_2^2 + \alpha \cdot D(\Lambda), \tag{56}$$

$$\mathbb{K}_0 = \left\{ \Lambda \in \mathbb{K} : \|\tilde{f}(\cdot, \Lambda)\|_\infty \leq 2B_\infty \right\}, \tag{57}$$

*where $B_\infty > \|f\|_\infty$ is a positive constant.*

$$\alpha_n = C_L \, B_\infty \, \frac{\log_e(\#\mathcal{D})}{n}, \tag{58}$$

*where $C_L$, $C_1$, $C_2$ are positive constants, and $1_{\tilde{\Omega}}$ is the indicator of the event*

$$\tilde{\Omega} = \left( \|\hat{f}_{n,\alpha_n}\|_\infty \leq 2B_\infty \right). \tag{59}$$

## B.2 Proof of Theorem 2

Denote $\hat{f} = \hat{f}_{n,\alpha_n}$ and $\hat{\Lambda} = \hat{\Lambda}_{n,\alpha_n}$. We condition on the set $\tilde{\Omega}$ so that $\hat{\Lambda} \in \mathbb{K}_0$. Let $f$ be the true density and let $\Lambda^0 \in \mathbb{K}_0$. Denote

$$\zeta = C_1 K(f, \Lambda^0, \alpha_n),$$

where $C_1$ is a positive constant to be chosen later. We have that

$$\begin{aligned} & E\|\hat{f} - f\|_2^2 \\ &= \int_0^\infty P\left( \|\hat{f} - f\|_2^2 > t \right) dt \\ &\leq \zeta + \int_\zeta^\infty P\left( \|\hat{f} - f\|_2^2 > t \right) dt \\ &= \zeta + C_2 n^{-1} \int_0^\infty P\left( \|\hat{f} - f\|_2^2 > C_2 n^{-1} t + \zeta \right) dt, \end{aligned} \tag{60}$$

where $C_2$ is a positive constant to be chosen later. Let $a > 0$. Now

$$\begin{aligned} A &\stackrel{def}{=} \left( \|\hat{f} - f\|_2^2 > C_2 n^{-1} t + \zeta \right) \\ &= \Big( (a+1)\|\hat{f} - f\|_2^2 \\ &\qquad > a\|\hat{f} - f\|_2^2 + C_1 K(f, \Lambda^0, \alpha_n) + C_2 n^{-1} t \Big) . \end{aligned} \tag{61}$$

Lemma 6 implies that the theoretical error-complexity of the minimization estimator may be bounded by the theoretical error-complexity of $f^0 = \tilde{f}(\cdot, \Lambda^0)$, with the additional empirical term:

$$\begin{aligned} & K(f, \hat{\Lambda}, \alpha_n) \leq K(f, \Lambda^0, \alpha_n) + 2\nu_n \left( \hat{f} - f^0 \right) \\ \Leftrightarrow \quad & \|\hat{f} - f\|_2^2 \leq K(f, \Lambda^0, \alpha_n) - \alpha_n D(\hat{\Lambda}) + 2\nu_n \left( \hat{f} - f^0 \right) . \end{aligned}$$

Thus we may continue (61) with

$$\begin{aligned} A &\subset \left( 2\nu_n \left( \hat{f} - f^0 \right) > \frac{a}{a+1} \|\hat{f} - f\|_2^2 + \alpha_n D(\hat{\Lambda}) \right. \\ &\qquad \left. + \left( \frac{C_1}{a+1} - 1 \right) K(f, \Lambda^0, \alpha_n) + \frac{C_2}{a+1} n^{-1} t \right) \\ &\subset \left( \nu_n \left( \hat{f} - f^0 \right) > w(\hat{\Lambda}) \xi \right) \\ &\subset \left( \sup_{\Lambda \in \mathbb{K}_0} \frac{\nu_n(\tilde{f}(\cdot, \Lambda)) - \nu_n(f^0)}{w(\Lambda)} > \xi \right) \\ &\stackrel{def}{=} B, \end{aligned}$$

where we used the fact that on $\tilde{\Omega}$, $\hat{\Lambda} \in \mathbb{K}_0$, and we denote

$$\xi = \frac{1}{2} \min \left\{ \frac{C_1}{a+1} - 1, \frac{a}{a+1} \right\},$$

$$\begin{aligned} w(\Lambda) &= \left\| \tilde{f}(\cdot, \Lambda) - f \right\|_2^2 + \left\| f - f^0 \right\|_2^2 + \frac{\tau(\Lambda)}{2n} , \\ \tau(\Lambda) &= C_\tau \left[ n\alpha_n \left( C_{\tau,1} D(\Lambda^0) + C_{\tau,2} D(\Lambda) \right) + t \right] , \\ C_\tau &= \frac{1}{\xi} \frac{C_2}{a+1} , \quad C_{\tau,1} = \frac{C_1 - a - 1}{C_2} , \quad C_{\tau,2} = \frac{a+1}{C_2} . \end{aligned} \tag{62}$$

We need to choose $C_1$, $C_2$, and $a$ so that $2C_\tau^{-1} \leq \xi^2$. This inequality will be needed in (71). This choice is possible: we take $2(a+1)/C_2 \leq \xi$. We need also $C_1/(a+1) - 1 > 0$ to guarantee $\xi > 0$. We have

$$P(A) \leq P(B). \tag{63}$$

We prove that

$$P(B) \leq C \exp\{-t(C_L B_\infty)^{-1}\}, \tag{64}$$

where $C$ and $C_L$ are positive constants. This proves the theorem, when we combine (60) and (63).

**Proof of (64).** For $\Phi \subset \mathcal{D}$, let $\mathbb{K}_{0,\Phi}$ be the set of coefficients in $\mathbb{K}_0$ which are non-zero exactly at the positions given by set $\Phi$:

$$\mathbb{K}_{0,\Phi} = \{\Lambda \in \mathbb{K}_0 : \lambda_\phi \neq 0 \text{ if and only if } \phi \in \Phi\},$$

where we use again the notation $\Lambda = (\lambda_\phi)_{\phi \in \mathcal{D}}$. Let for $l = 1, 2, \ldots,$

$$\mathbb{D}_l = \{\Phi \subset \mathcal{D} : \#\Phi = l\}$$

be the set of subsets of $\mathcal{D}$ of cardinality $l$. We may write

$$\mathbb{K}_0 = \bigcup_{l=1}^{\infty} \bigcup_{\Phi \in \mathbb{D}_l} \mathbb{K}_{0,\Phi}.$$

That is, we make a countable partition of $\mathbb{K}_0$ and each member of the partition is the set of vectors $\Lambda$ which have exactly $l$ non-zero elements. We have that

$$B \subset \bigcup_{l=1}^{\infty} \bigcup_{\Phi \in \mathbb{D}_l} B_\Phi,$$

where

$$B_\Phi = \left( \sup_{\Lambda \in \mathbb{K}_{0,\Phi}} \frac{\nu_n(\tilde{f}(\cdot, \Lambda)) - \nu_n(f^0)}{w(\Lambda)} > \xi \right).$$

For $\Phi \in \mathbb{D}_l$,

$$P\left(B_\Phi\right) \leq 2 \exp\{-(C_L B_\infty)^{-1}(t + lL)\}, \tag{65}$$

where $C_L$ is a positive constant defined in (75) and

$$L = C_L B_\infty \log_e(\#\mathcal{D}).$$

We prove (65) below. We have that

$$\#\,\mathbb{D}_l = \left( \begin{array}{c} \#\mathcal{D} \\ l \end{array} \right) \leq \left( \frac{e\#\mathcal{D}}{l} \right)^l. \tag{66}$$

Thus,

$$\begin{aligned} P(B) &\leq 2\sum_{l=1}^{\infty}\sum_{\Phi\in\mathbb{D}_l}\exp\{-(C_L B_\infty)^{-1}(t+lL)\} \\ &\leq 2\sum_{l=1}^{\infty}\left(\frac{e\#\mathcal{D}}{l}\right)^{l}\exp\{-(C_L B_\infty)^{-1}(t+lL)\} \\ &\leq C\exp\{-(C_L B_\infty)^{-1}t\}, \end{aligned} \tag{67}$$

by the choice of $L$. We have proved (64) up to proving (65).

**Proof of (65).** Denote

$$Z = \sup_{g\in\mathcal{G}}\nu_n(g),$$

where

$$\mathcal{G} = \mathcal{G}_\Phi = \left\{\frac{\tilde{f}(\cdot,\Lambda)-f^0}{w(\Lambda)} : \Lambda\in\mathbb{K}_{0,\Phi}\right\}, \qquad \Phi\in\mathbb{D}_l.$$

Denote

$$v_0 = \sup_{g\in\mathcal{G}}\|g\|_2^2 = \sup_{\Lambda\in\mathbb{K}_{0,\Phi}}\frac{\|\tilde{f}(\cdot,\Lambda)-f^0\|_2^2}{w^2(\Lambda)}$$

and

$$\tau = C_\tau\left[n\alpha_n\left(C_{\tau,1}D(\Lambda^0)+C_{\tau,2}\,l\right)+t\right].$$

We have for $\Phi\in\mathbb{D}_l$, $\Lambda\in\mathbb{K}_{0,\Phi}$, that $\tau(\Lambda)=\tau$, where $\tau(\Lambda)$ is defined in (62). Thus, for $\Phi\in\mathbb{D}_l$ and $\Lambda\in\mathbb{K}_{0,\Phi}$,

$$w(\Lambda)\geq\frac{1}{2}\left(\left\|\tilde{f}(\cdot,\Lambda)-f^0\right\|_2^2+\frac{\tau}{n}\right)\geq\left\|\tilde{f}(\cdot,\Lambda)-f^0\right\|_2\left(\frac{\tau}{n}\right)^{1/2}. \tag{68}$$

Thus

$$v_0\leq\frac{n}{\tau}. \tag{69}$$

We have that

$$(EZ)^2\leq\|f\|_\infty\left(l+D(\Lambda^0)\right)\tau^{-1}. \tag{70}$$

We prove equation (70) below in page 30. Denote

$$\eta^2 = \tau^{-1}(t+C_{\tau,2}Ll).$$

Then we have

$$\begin{aligned} (EZ+\eta)^2 &\leq 2\left[(EZ)^2+\eta^2\right] \\ &\leq 2\tau^{-1}\left[B_\infty\left(l+D(\Lambda^0)\right)+t+C_{\tau,2}Ll\right] \\ &\leq 2C_\tau^{-1} \\ &\leq \xi^2 \end{aligned} \tag{71}$$

since

$$C_\tau^{-1}\tau \geq L\left(C_{\tau,1}D(\Lambda^0) + C_{\tau,2}\, l\right) + t \geq B_\infty D(\Lambda^0) + (B_\infty + C_{\tau,2}L)\, l + t,$$

where we used $n\alpha_n = L$ and $C_{\tau,1}C_L \log_e(\#\mathcal{D}) \geq 1$. Eq. (71) implies that

$$P(B_\Phi) = P\left(Z \geq \xi\right) \leq P\left(Z \geq EZ + \eta\right). \tag{72}$$

Denote

$$b = \sup\{\|g\|_\infty : g \in \mathcal{G}\},$$

and

$$v = \sup\{\mathrm{Var}_f(g(X^1)) : g \in \mathcal{G}\}.$$

By Talagrand's theorem, as given in Bousquet (2002), Theorem 2.3, by applying the inequality $h(t) \geq t^2/(2+2t/3)$, $t > 0$, for $h(t) = (1+t)\log(1+t) - t$, we get

$$P\left(Z \geq EZ + \eta\right) \leq \exp\left\{\frac{-n\eta^2}{2[v + 2bEZ + \eta b/3]}\right\}. \tag{73}$$

We have

$$v \leq \|f\|_\infty v_0 \leq \|f\|_\infty \frac{n}{\tau}.$$

Also, for for $\Phi \in \mathbb{D}_l$ and $\Lambda \in \mathbb{K}_{0,\Phi}$,

$$w(\Lambda) \geq \frac{\tau}{2n} \tag{74}$$

and thus

$$b \leq \frac{2n}{\tau} \sup_{\Lambda \in \mathbb{K}_{0,\Phi}} \|\tilde{f}(\cdot, \Lambda) - f^0\|_\infty \leq \frac{4n}{\tau} B, \qquad B = 2\|f\|_\infty,$$

by the definition of $\mathbb{K}_0$. Thus, applying inequalities $EZ \leq \xi$, $\eta \leq \xi$,

$$\begin{aligned} v + 2bEZ + \eta b/3 &\leq \frac{n}{\tau}\|f\|_\infty\left(1 + 8 \cdot \xi(2 + 1/3)\right) \\ &\stackrel{def}{=} \frac{n}{\tau}\|f\|_\infty C_L/2. \end{aligned} \tag{75}$$

Thus

$$P\left(Z \geq EZ + \eta\right) \leq \exp\{-(t + lL)/(B_\infty C_L)\}. \tag{76}$$

Eq. (65) follows from (72) and (76).

**Proof of (70).** Let $\Lambda \in \mathbb{K}_{0,\Phi}$ where $\Phi \in \mathbb{D}_l$. Denote $\mathcal{D}(\Lambda, \Lambda^0) = \{\phi \in \mathcal{D} : \lambda_\phi \neq 0 \text{ or } \lambda^0_\phi \neq 0\}$. Let $\{\psi_1, \ldots, \psi_k\}$ be a basis of the span of $\mathcal{D}(\Lambda, \Lambda^0)$. We have $k \leq \#\mathcal{D}(\Lambda, \Lambda^0) \leq l + D(\Lambda^0)$ and, applying (68),

$$\sup_{\Lambda \in \mathbb{K}_{0,\Phi}} \frac{\|\tilde{f}(\cdot, \Lambda) - f^0\|_2^2}{w^2(\Lambda)} \leq \frac{n}{\tau}.$$

Thus we may apply Lemma 7 with $B_2^2 = n/\tau$ to get (70). □

## B.3 Proof of (32)

The series estimator $f^*_{n,\alpha}$ defined in (22) is equal to the estimator $\hat{f}_{n,\alpha}$ defined in (51), under suitable identifications. Indeed, we note that we can write

$$f^*_{n,\alpha}(x) = \tilde{f}\left(x, \hat{W}_{n,\alpha}, \hat{\Theta}_{n,\alpha}, \hat{\mathcal{B}}_{n,\alpha}\right), \qquad x \in \mathbf{R}^d, \tag{77}$$

where

$$\left(\hat{\mathcal{B}}_{n,\alpha}, \hat{\Theta}_{n,\alpha}, \hat{W}_{n,\alpha}\right) = \text{argmin}_{\mathcal{B}\in\mathcal{L}, \Theta\in\mathbf{R}^{\mathcal{B}}, W\in\mathcal{W}(\mathcal{B})} \mathcal{E}_n\left(W, \Theta, \mathcal{B}, \alpha\right). \tag{78}$$

We have that

$$f^*_{n,\alpha}(x) = \hat{f}_{n,\alpha}(x), \qquad x \in \mathbf{R}^d, \tag{79}$$

when we choose

$$\mathcal{D} = \left\{I_{[0,1]^d}\right\} \cup \bigcup_{\mathcal{B}\in\mathcal{L}} \mathcal{B},$$

and

$$\mathbb{K} \subset \left\{\Lambda \in \mathbf{R}^{\mathcal{D}} : \Lambda \in \mathcal{W}(\mathcal{B}) \text{ for some } \mathcal{B} \in \mathcal{L}\right\}. \tag{80}$$

Definition (80) is explained by the fact that the vectors $\Lambda = (\lambda_\phi)_{\phi\in\mathcal{D}}$ have to be such that components are non-zero only for a single basis $\mathcal{B}$: $\{\phi : \lambda_\phi \neq 0\} \subset \mathcal{B}$ for some $\mathcal{B} \in \mathcal{L}$. Eq. (79) holds since we may use the identification $\lambda_\phi = w_\phi \theta_\phi$.

**Cardinality of the library.** To apply Theorem 2 we have to check that the smoothing parameter in (31) has the same order as the smoothing parameter in (58). This follows because the cardinality of the dictionary $\mathcal{D} = \{I_{[0,1]^d}\} \cup \cup_{\mathcal{B}\in\mathcal{L}(J)}\mathcal{B}$ satisfies

$$\#\,\mathcal{D} \leq C \cdot n^{a \log_2(2d)} \tag{81}$$

for a positive constant $C$. The calculation of the cardinality of the library is basically the same as the calculation in (23). Indeed, every function in $\mathcal{B}(\mathcal{T})$, defined in (15), may be obtained as a node of a large multitree which has 1 root node, where the number of children of each node is equal to $2d$, and the depth of the tree is equal to $|J_n|_{max} \leq \lceil a \log_2 n \rceil$. The number of the nodes of the tree in question is $\sum_{i=0}^{|J_n|_{max}-1} (2d)^i \leq (2d)^{|J_n|_{max}}$.

**Final detail.** Theorem 2 involves set $\tilde{\Omega}$ defined in (59). We need to show that the restriction to this set is not essential. We have the bound

$$E_f \left\| f^*_{n,\alpha_n} - f \right\|_2^2 1_{\tilde{\Omega}^c} \leq \left( \left\| f^*_{n,\alpha_n} \right\|_\infty + \|f\|_\infty \right)^2 P\left(\tilde{\Omega}^c\right), \tag{82}$$

where we applied the fact that $\|g\|_2^2 \leq \|g\|_\infty^2$, when the support of $g$ is contained in $[0,1]^d$. First, for all samples,

$$\|f^*_{n,\alpha_n}\|_\infty \leq n^\kappa \tag{83}$$

for some $\kappa > 0$. We may prove (83) by noting that by Lemma 1, $\|f^*_{n,\alpha}\|_\infty = \|\hat{f}_{n,\alpha}\|_\infty$ and $\|\hat{f}_{n,\alpha}\|_\infty \leq 2^{|J|}$, since $2^{|J|}$ is the minimal volume of the rectangles in the partition of histogram $\hat{f}_{n,\alpha}$. Second, we need that for sufficiently large $n$,

$$P\left(\tilde{\Omega}^c\right) \leq \delta_n^* \stackrel{def}{=} n^{\kappa'} \exp\left\{ -n^{1-a} \frac{3\|f\|_\infty}{8} \right\}, \tag{84}$$

for some $\kappa' > 0$, where $0 < a < 1$ is the fineness parameter in (28). Equation (84) follows from Bernstein's inequality. (Note that also in the proof of (84) we apply Lemma 1.) □

# C Proof of Lemma 4

As before, we denote $\mathcal{J}^* = \mathcal{J}_{h^*}$. We prove that

$$\sum_{m=0}^{M} \sum_{k \in K_{\mathcal{J}^*(m)}} \min\left\{ \tau_{mk}^2, \alpha \right\} \leq \left( 2 + \frac{2C_{p,L,d}}{2^{\beta_*} - 1} \right) \cdot \alpha^{2\sigma/(2\sigma+1)} \tag{85}$$

and

$$\sum_{m=M+1}^{\infty} \sum_{k \in K_{\mathcal{J}^*(m)}} \tau_{mk}^2 \leq \frac{2^d L^2}{2^{2\beta_*} - 1} \cdot \alpha^{2\sigma/(2\sigma+1)}, \tag{86}$$

when $0 < \alpha < 1$ is sufficiently small, where $\tau_{mk}$ is defined in (45), $C_{p,L,d} = \max_{l=1,\ldots,d}(2^{d/2}L)^{\tilde{p}_l}$, and $\beta_* = \min_{l=1,\ldots,d}(\sigma + 1/2 - 1/\tilde{p}_l) = \sigma + 1/2 - 1/p_*$, $p_* = \min_{l=1,\ldots,d} \tilde{p}_l$. This implies the lemma, when we apply Lemma 5 with $\mathcal{B} = \mathcal{B}^*_\alpha$ and with the basis $\mathcal{B}_\infty = \mathcal{B}^*$.

**Proof of (85).** Let $m^* \geq 1$ be defined by

$$m^* = \left\lceil \frac{1}{2\sigma + 1} \log_2 \alpha^{-1} \right\rceil. \tag{87}$$

Note that $m^* < M$ since $\tilde{a} > 1/(2\sigma+1)$ by the lower bound in (48). Write

$$\sum_{m=0}^{M} \sum_{k \in K_{\mathcal{J}^*(m)}} \min\left\{\tau_{mk}^2, \alpha\right\} \leq A + B, \tag{88}$$

where

$$\begin{aligned} A &\stackrel{def}{=} \sum_{m=0}^{m^*} \sum_{k \in K_{\mathcal{J}^*(m)}} \alpha = \alpha \sum_{m=0}^{m^*} 2^m = \alpha(2^{m^*+1} - 2) \\ &\leq 2\alpha\alpha^{-1/(2\sigma+1)} = 2\alpha^{2\sigma/(2\sigma+1)} \end{aligned} \tag{89}$$

by the definition of $m^*$ in (87), and

$$\begin{aligned} B &\stackrel{def}{=} \sum_{m=m^*+1}^{M} \sum_{k \in K_{\mathcal{J}^*(m)}} \min\left\{\tau_{mk}^2, \alpha\right\} \\ &\leq \sum_{m=m^*+1}^{M} \alpha^{1-\tilde{p}_{l_m^*}/2} \sum_{k \in K_{\mathcal{J}^*(m)}} |\tau_{mk}|^{\tilde{p}_{l_m^*}}, \end{aligned} \tag{90}$$

where $\tilde{p}_l$ is defined in (47), and we use the notation $l_m^* = h^*(m+1)$. Above we used the fact $\min\{\tau_{mk}^2, \alpha\} \leq \alpha^{1-\tilde{p}_l/2}|\tau_{mk}|^{\tilde{p}_l}$, for $l = 1, \ldots, d$. Indeed, we have that when $\tau_{mk}^2 \leq \alpha$, then $\alpha^{1-\tilde{p}_l/2}|\tau_{mk}|^{\tilde{p}_l} \geq \tau_{mk}^2$ and when $\tau_{mk}^2 > \alpha$, then $\alpha^{1-\tilde{p}_l/2}|\tau_{mk}|^{\tilde{p}_l} \geq \alpha$. We have from Lemma 3 that

$$\sum_{k \in K_{\mathcal{J}^*(m)}} |\tau_{mk}|^{\tilde{p}_{l_m^*}} \leq \left(2^{d/2} L 2^{-m\beta_m}\right)^{\tilde{p}_{l_m^*}}, \qquad \beta_m = \sigma + 1/2 - 1/\tilde{p}_{l_m^*}. \tag{91}$$

Continuing from (90),

$$\begin{aligned} B &\leq C_{p,L,d} \sum_{m=m^*+1}^{M} \alpha^{1-\tilde{p}_{l_m^*}/2} 2^{-\tilde{p}_{l_m^*} m\beta_m} && (92) \\ &= C_{p,L,d} \sum_{m=m^*+1}^{M} \alpha^{1-\tilde{p}_{l_m^*}/2} 2^{-\tilde{p}_{l_m^*} m^*\beta_m} 2^{\tilde{p}_{l_m^*}(m^*-m)\beta_m} \\ &\leq \frac{C_{p,L,d}}{2^{\beta_*} - 1} \cdot \alpha 2^{m^*} && (93) \\ &\leq \frac{2C_{p,L,d}}{2^{\beta_*} - 1} \cdot \alpha^{2\sigma/(2\sigma+1)}, && (94) \end{aligned}$$

where in (92) we applied (91). In (93) we applied

$$\alpha^{-\tilde{p}_{l_m^*}/2} 2^{-p_{l_m^*} m^*(\sigma+1/2)} \leq 1$$

which holds due to the choice of $m^*$ in (87), and we applied in (93) also the fact

$$\sum_{m=m^*+1}^{\infty} 2^{\tilde{p}_{l_m^*}(m^*-m)\beta_m} \leq \sum_{m=1}^{\infty} (2^{-\beta_*})^m = \frac{1}{2^{\beta_*}-1}$$

which holds because $\tilde{p}_{l_m^*} \geq 1$, because $\sum_{m=1}^{\infty} r^m = r/(1-r)$ for $0 < r < 1$, and because $\beta_* > 0$ which is assumed in (29). The claim (85) follows from (88), (89), and (94).

**Proof of (86).** We have that

$$\begin{aligned}
\sum_{m=M+1}^{\infty} \sum_{k \in K_{\mathcal{J}^*(m)}} \tau_{mk}^2 &\leq \sum_{m=M+1}^{\infty} \left( \sum_{k \in K_{\mathcal{J}^*(m)}} |\tau_{mk}|^{\tilde{p}_{l_m^*}} \right)^{2/\tilde{p}_{l_m^*}} && (95)\\
&\leq 2^d L^2 \sum_{m=M+1}^{\infty} 2^{-2m\beta_m} && (96)\\
&\leq \frac{2^d L^2}{2^{2\beta_*}-1} \cdot 2^{-2\beta_* M} && (97)\\
&\leq \frac{2^d L^2}{2^{2\beta_*}-1} \cdot \alpha^{2\beta_* \tilde{a}} && (98)\\
&\leq \frac{2^d L^2}{2^{2\beta_*}-1} \cdot \alpha^{2\sigma/(2\sigma+1)}, && (99)
\end{aligned}$$

where in (95) we applied the subadditivity of the function $x \mapsto x^{\tilde{p}_l/2}$, in (96) we applied (91), in (97) we applied that for $0 < r < 1$, $\sum_{m=M+1}^{\infty} r^m = r^{M+1}/(1-r)$ and the fact that $\beta_* > 0$, in (98) we applied the choice of $M$ in (49), and in (99) we applied the lower bound for $\tilde{a}$ in (48). We have proved Lemma 4. □

# D Auxiliary lemmas

## D.1 Complexity penalized approximation error

**Lemma 5** *Let $\mathcal{B}_\infty$ be a basis of $L_2([0,1]^d)$ such that $\mathcal{B} \subset \mathcal{B}_\infty$, where $\mathcal{B}$ is an orthonormal system. Then,*

$$\begin{aligned}
&\min_{W \in \{0,1\}^{\mathcal{B}}} K\left(f, W, \Theta_f(\mathcal{B}), \mathcal{B}, \alpha\right)\\
&= \alpha + \sum_{\phi \in \mathcal{B}} \min\left\{\theta_{f,\phi}^2, \alpha\right\} + \sum_{\phi \in \mathcal{B}_\infty \setminus \mathcal{B}} \theta_{f,\phi}^2,
\end{aligned}$$

*where $\theta_{f,\phi} = \int_{\mathbf{R}^d} f\phi$, $\Theta_f(\mathcal{B}) = (\theta_{f,\phi})_{\phi \in \mathcal{B}}$.*

## D.2 Pre-oracle inequality

Let

$$\mathcal{C} = \{g_\kappa : \kappa \in \mathbb{K}\}, \tag{100}$$

where $g_\kappa : \mathbf{R}^d \to \mathbf{R}$ and $\mathbb{K}$ is a set of parameters. Let $D : \mathbb{K} \to [0, \infty)$ be a penalization term and define the complexity penalized empirical risk as

$$\mathcal{E}_n(\kappa, \alpha) = \gamma_n\left(g_\kappa\right) + \alpha\, D(\kappa), \tag{101}$$

where $\gamma_n(g)$ is defined in (4) and $\alpha \geq 0$ is the smoothing parameter controlling the amount of penalization. We assume that $D(\kappa)$ takes larger values for more complex $g_\kappa$. Let $\hat{\kappa}$ be such that

$$\mathcal{E}_n\left(\hat{\kappa}, \alpha\right) \leq \inf_{\kappa \in \mathbb{K}} \mathcal{E}_n(\kappa, \alpha) + \epsilon, \tag{102}$$

where $\epsilon > 0$ and define the minimization estimator by

$$\hat{f} = g_{\hat{\kappa}}. \tag{103}$$

**Lemma 6** *Let $\mathcal{C} \subset L_2(\mathbf{R}^d)$ be parameterized by (100) and let $\hat{f} = g_{\hat{\kappa}} \in \mathcal{C}$ where $\hat{\kappa}$ satisfies (102). Then for each $f^0 = g_{\kappa^0} \in \mathcal{C}$,*

$$K(f, \hat{\kappa}, \alpha) \leq K(f, \kappa^0, \alpha) + \varepsilon + 2\nu_n\left(\hat{f} - f^0\right),$$

*where $f$ is the true density,*

$$K(f, \kappa, \alpha) = \|g_\kappa - f\|_2^2 + \alpha \cdot D(\kappa),$$

*and $\nu_n(g)$ is the centered empirical operator defined by*

$$\nu_n(g) = n^{-1} \sum_{i=1}^{n} g(X^i) - \int_{\mathbf{R}^d} gf,$$

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

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

Thus,

$$\left\|\hat{f} - f\right\|_2^2 - \gamma_n\left(\hat{f}\right) + \gamma_n\left(f^0\right) - \left\|f^0 - f\right\|_2^2 = 2\nu_n\left(\hat{f} - f^0\right). \tag{104}$$

We have

$$\begin{aligned}
&K(f,\hat{\kappa},\alpha) - K(f,\kappa^0,\alpha) \\
&= K(f,\hat{\kappa},\alpha) - \mathcal{E}_n(\hat{\kappa},\alpha) + \mathcal{E}_n(\hat{\kappa},\alpha) - K(f,\kappa^0,\alpha) \\
&\leq K(f,\hat{\kappa},\alpha) - \mathcal{E}_n(\hat{\kappa},\alpha) + \mathcal{E}_n(\kappa^0,\alpha) + \varepsilon - K(f,\kappa^0,\alpha) \qquad (105) \\
&= \left\|\hat{f} - f\right\|_2^2 - \gamma_n(\hat{f}) + \gamma_n(f^0) + \varepsilon - \left\|f^0 - f\right\|_2^2 \\
&= 2\nu_n\left(\hat{f} - f^0\right) + \varepsilon. \qquad (106)
\end{aligned}$$

In (105) we applied (102) and in (106) we applied (104). □

## D.3 Expectation of the supremum

Let $\mathcal{G}$ be a set of linear combinations of an orthonormal system:

$$\mathcal{G} = \left\{ \sum_{j=1}^{k} \theta_j \phi_j : \sum_{j=1}^{k} \theta_j^2 \leq B_2^2 \right\}, \qquad (107)$$

where $\{\phi_1, \ldots, \phi_k\}$ is an orthonormal system and $0 < B_2 < \infty$. We have a bound for $E \sup_{g \in \mathcal{G}} \nu_n(g)$ which depends essentially from $\sqrt{k/n}$.

**Lemma 7** *Let $\mathcal{G}$ be defined in (107). We have that*

$$E \sup_{g \in \mathcal{G}} \nu_n(g) \leq B_2 \|f\|_\infty^{1/2} (k/n)^{1/2}.$$

*Proof.* By the Cauchy-Schwartz inequality, for $g = \sum_{j=1}^{k} \theta_j \phi_j \in \mathcal{G}$,

$$\nu_n(g) = \sum_{j=1}^{k} \theta_j \nu_n(\phi_j) \leq \left( \sum_{j=1}^{k} \theta_j^2 \sum_{j=1}^{k} \nu_n(\phi_j)^2 \right)^{1/2}.$$

We have $E|X|^{1/2} \leq (E|X|)^{1/2}$. Thus,

$$E \sup_{g \in \mathcal{G}} \nu_n(g) \leq B_2 \left( \sum_{j=1}^{k} E\nu_n(\phi_j)^2 \right)^{1/2}.$$

We have

$$E\nu_n(\phi_j)^2 \leq \|f\|_\infty n^{-1},$$

which implies the lemma. □