

## EMPIRICAL RISK MINIMIZATION IN INVERSE PROBLEMS: EXTENDED TECHNICAL VERSION

BY JUSSI KLEMELÄ AND ENNO MAMMEN

*University of Oulu and University of Mannheim*

We study estimation of a multivariate function  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  when the observations are available from function  $Af$ , where  $A$  is a known linear operator. Both the Gaussian white noise model and density estimation are studied. We define an  $L_2$  empirical risk functional, which is used to define a  $\delta$ -net minimizer and a dense empirical risk minimizer. Upper bounds for the mean integrated squared error of the estimators are given. The upper bounds show how the difficulty of the estimation depends on the operator through the norm of the adjoint of the inverse of the operator, and on the underlying function class through the entropy of the class. Corresponding lower bounds are also derived. As examples we consider convolution operators and the Radon transform. In these examples the estimators achieve the optimal rates of convergence. Furthermore, a new type of oracle inequality is given for inverse problems in additive models.

**1. Introduction.** We consider estimation of a function  $f : \mathbf{R}^d \rightarrow \mathbf{R}$ , when a linear transform  $Af$  of the function is observed under stochastic noise. We consider both the Gaussian white noise model and density estimation with i.i.d. observations. We study two estimators: a  $\delta$ -net estimator which minimizes the  $L_2$  empirical risk over a minimal  $\delta$ -net of a function class, and a dense empirical risk minimizer which minimizes the empirical risk over the whole function class without restricting the minimization over a  $\delta$ -net. We call this estimator “dense minimizer” because it is defined as a minimizer over a possibly uncountable function class. The  $\delta$ -net estimator is more universal: it may be applied also for unsmooth functions and for severely ill-posed operators. On the other hand, the dense empirical minimizer is expected to work only for relatively smooth cases (the entropy integral has to converge). But because the minimization in the calculation of this estimator is not restricted to a  $\delta$ -net we have available a larger toolbox of algorithms for finding (an approximation of) the minimizer of the empirical risk.

Let  $(\mathbf{Y}, \mathcal{Y}, \nu)$  be a Borel space and let  $A : L_2(\mathbf{R}^d) \rightarrow L_2(\mathbf{Y})$  be a linear

---

*AMS 2000 subject classifications:* Primary 62G07

*Keywords and phrases:* deconvolution, empirical risk minimization, multivariate density estimation, nonparametric function estimation, Radon transform, tomography

operator, where  $L_2(\mathbf{R}^d)$  is the space of square integrable functions  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  (with respect to the Lebesgue measure), and  $L_2(\mathbf{Y})$  is the space of square integrable functions  $g : \mathbf{Y} \rightarrow \mathbf{R}$  (with respect to measure  $\nu$ ). In the density estimation model we have i.i.d. observations

$$(1) \quad Y_1, \dots, Y_n \in \mathbf{Y},$$

with common density function  $Af : \mathbf{Y} \rightarrow \mathbf{R}$ , where  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  is a density function which we want to estimate. In the Gaussian white noise model the observation is a realization of the process

$$(2) \quad dY_n(y) = (Af)(y) dy + n^{-1/2} dW(y), \quad y \in \mathbf{Y},$$

where  $W(y)$  is the Brownian process on  $\mathbf{Y}$ , that is, for  $h_1, h_2 \in L_2(\mathbf{Y})$ , the random vector  $(\int_{\mathbf{Y}} h_1 dW, \int_{\mathbf{Y}} h_2 dW)$  is a 2-dimensional Gaussian random vector with 0 mean, marginal variances  $\|h_1\|_{2,\nu}^2, \|h_2\|_{2,\nu}^2$ , and covariance  $\int_{\mathbf{Y}} h_1 h_2 d\nu$ . (In our examples  $\mathbf{Y}$  is either the Euclidean space or the product of the real line with the unit sphere, so that the existence of the Brownian process is guaranteed.) We want to estimate the signal function  $f : \mathbf{R}^d \rightarrow \mathbf{R}$ . The Gaussian white noise model is very useful in presenting the basic mathematical ideas in a transparent way. For the  $\delta$ -net estimator the treatment is almost identical for the Gaussian white noise model and for the density estimation, but when we consider the dense empirical risk minimization, then in the density estimation model we need to use bracketing numbers and empirical entropies with bracketing, instead of the usual  $L_2$  entropies. Our results for the Gaussian white noise model can also serve as first step for getting analogous results for inverse problems in regression or in other statistical models.

The  $L_2$  empirical risk is defined by

$$(3) \quad \gamma_n(g) = \begin{cases} -2 \int_{\mathbf{Y}} (Qg) dY_n + \|g\|_2^2, & \text{Gaussian white noise,} \\ -2n^{-1} \sum_{i=1}^n (Qg)(Y_i) + \|g\|_2^2, & \text{density estimation,} \end{cases}$$

where  $Q$  is the adjoint of the inverse of  $A$ :

$$(4) \quad \int_{\mathbf{R}^d} (A^{-1}h)g = \int_{\mathbf{Y}} h(Qg) d\nu,$$

for  $h \in L_2(\mathbf{Y})$ ,  $g \in L_2(\mathbf{R}^d)$ . The operator  $Q = (A^{-1})^*$  has the domain  $L_2(\mathbf{R}^d)$ , similarly as  $A$ . Minimizing  $\|\hat{f} - f\|_2^2$  with respect to estimators  $\hat{f}$  is equivalent to minimizing  $\|\hat{f} - f\|_2^2 - \|f\|_2^2$ , and we have, in the Gaussian

white noise model,

$$\begin{aligned}
\|\hat{f} - f\|_2^2 - \|f\|_2^2 &= -2 \int_{\mathbf{R}^d} f \hat{f} + \|\hat{f}\|_2^2 \\
&= -2 \int_{\mathbf{Y}} (Af)(Q\hat{f}) d\nu + \|\hat{f}\|_2^2 \\
&\approx -2 \int_{\mathbf{Y}} (Q\hat{f}) dY_n + \|\hat{f}\|_2^2 \\
(5) \qquad \qquad \qquad &= \gamma_n(\hat{f}).
\end{aligned}$$

The usual least squares estimator is defined as a minimizer of the the criterion

$$\begin{aligned}
\|A\hat{f} - Af\|_{\mathbf{Y}}^2 - \|Af\|_{\mathbf{Y}}^2 &\approx -2 \int_{\mathbf{Y}} (Ag) dY_n + \|Ag\|_{\mathbf{Y}}^2 \\
(6) \qquad \qquad \qquad &\stackrel{def}{=} \tilde{\gamma}_n(g).
\end{aligned}$$

See for example O'Sullivan (1986). In density estimation the log-likelihood empirical risk has been more common than the  $L_2$  empirical risk, and in the setting of inverse problems the log-likelihood is defined as  $\tilde{\gamma}_n(g) = -n^{-1} \sum_{i=1}^n \log(Ag)(Y_i)$ , analogously to (6). These alternative definitions of the empirical risk do not seem to lead to such elegant theory as the empirical risk in (3). The empirical risk in (3) has been used in deconvolution problems for projection estimators by Comte et al. (2005).

We give upper bounds for the mean integrated squared error (MISE) of the estimators. The upper bounds characterize how the rates of convergence depend on the entropy of the underlying function class  $\mathcal{F}$  and on smoothness properties of the operator  $A$ . Previously such characterizations have been given (up to our knowledge) in inverse problems only for the case of estimating real valued linear functionals  $L$ . In these cases the rates of convergence are determined by the modulus of continuity of the functional  $\omega(\epsilon) = \sup\{L(f) : f \in \mathcal{F}, \|Af\|_2 \leq \epsilon\}$ , see Donoho & Low (1992). For the case of estimating the whole function with a global loss function the rates of convergence depend on the largeness of the underlying function class in terms of the entropy and capacity, see Cencov (1972), Le Cam (1973), Ibragimov & Hasminskii (1980), Ibragimov & Hasminskii (1981), Birgé (1983), Hasminskii & Ibragimov (1990), Barron & Yang (1999), Ibragimov (2004).  $\delta$ -net estimators were considered e.g. by van der Laan et al. (2004). These papers consider direct statistical problems. We show that for inverse statistical problems the rate of convergence depends on the operator through the operator norm  $\varrho(Q, \mathcal{F}_\delta)$  of  $Q$ , over a minimal  $\delta$ -net  $\mathcal{F}_\delta$ , see (9) for the

definition of  $\varrho(Q, \mathcal{F}_\delta)$ . More precisely, the convergence rate  $\psi_n$  of the  $\delta$ -net estimator is the solution to the equation

$$n\psi_n^2 = \varrho^2(Q, \mathcal{F}_{\psi_n}) \log(\#\mathcal{F}_{\psi_n}),$$

where  $\#\mathcal{F}_{\psi_n}$  is the cardinality of a minimal  $\delta$ -net. For direct problems, when  $A$  is the identity operator,  $\varrho(Q, \mathcal{F}_\delta) \asymp 1$ . As examples of operators  $A$  we consider the convolution operator and the Radon transform. For these operators the estimators achieve the minimax rates of convergence over Sobolev classes.

The general framework for empirical risk minimization and the use of the empirical process machinery including entropy bounds for deriving optimal bounds seems to be new. Convolution and Radon transforms are discussed for illustrative purposes. These examples show that our results lead to optimal rates of convergence. As a new application we introduce the estimation of additive models in inverse problems. A new type of oracle inequality is presented, which gives the optimal rates of convergence also in “anisotropic” inverse problems.

*Contents.* Section 2 gives an upper bound for the MISE of the  $\delta$ -net estimator. Section 3 gives a lower bound for the MISE of any estimator. Section 4 gives an upper bound for the MISE of the dense empirical risk minimizer. Section 5 finds the adjoint of the inverse of  $A$ , when  $A$  is a convolution operator or the Radon transform. Section 6 proves that the  $\delta$ -net estimator achieves the optimal rate of convergence in the ellipsoidal framework and it contains an oracle inequality for additive models. Section 7 contains the proofs of the main results. The appendix contains calculations related to ellipsoids.

*Notation.* We use the notation  $\|\cdot\|$  to mean the Euclidean norm in  $\mathbf{R}^d$ . The  $L_2$  norm of a function  $g : \mathbf{R}^d \rightarrow \mathbf{R}$  will be denoted by  $\|g\|_2$ . The unit sphere in  $\mathbf{R}^d$  is denoted by  $\mathbf{S}_{d-1} = \{x \in \mathbf{R}^d : \|x\| = 1\}$ . The Lebesgue measure on  $\mathbf{S}_{d-1}$  is denoted by  $\mu$ . We will make use of the formula  $\mu(\mathbf{S}_{d-1}) = 2\pi^{d/2}/\Gamma(d/2)$ . By  $I_R$  we denote the indicator function, i.e.  $I_R(x) = 1$  when  $x \in R$  and  $I_R(x) = 0$  otherwise. We write  $a_n \asymp b_n$  to mean that  $0 < \liminf_{n \rightarrow \infty} a_n/b_n \leq \limsup_{n \rightarrow \infty} a_n/b_n < \infty$ , and  $a_n \gtrsim b_n$  means that  $\liminf_{n \rightarrow \infty} a_n/b_n > 0$ . The Fourier transform of a function  $g \in L_1(\mathbf{R}^d)$  is defined by

$$(Fg)(\omega) = \int_{\mathbf{R}^d} \exp\{ix^T\omega\}g(x) dx, \quad \omega \in \mathbf{R}^d,$$

where  $i$  is the imaginary unit. We use also the notation  $F_1g$  when  $g : \mathbf{R} \rightarrow \mathbf{R}$  is univariate. We have

$$g(x) = (2\pi)^{-d} \int_{\mathbf{R}^d} \exp\{-ix^T\omega\} (Fg)(\omega) d\omega, \quad x \in \mathbf{R}^d.$$

By Parseval's theorem, we have for  $f, g \in L_1(\mathbf{R}^d) \cap L_2(\mathbf{R}^d)$ ,

$$\int_{\mathbf{R}^d} fg = (2\pi)^{-d} \int_{\mathbf{R}^d} (Ff)(Fg).$$

Convolution of  $f$  and  $g$  is denoted by  $f * g(x) = \int_{\mathbf{R}^d} f(x-y)g(y) dy$ . We have that

$$(7) \quad F(f * g) = (Ff)(Fg).$$

The probability measures of the Gaussian white noise process  $Y_n$  and of the i.i.d. sequence  $(Y_1, \dots, Y_n)$  are denoted by  $P_{Af}^{(n)}$ .

## 2. $\delta$ -net minimizer.

*Definition of the estimator.* Let  $\mathcal{F}$  be a set of densities or signal functions  $f : \mathbf{R}^d \rightarrow \mathbf{R}$ . Let  $\mathcal{F}_\delta$  be a finite  $\delta$ -net of  $\mathcal{F}$  in the  $L_2$  metric, where  $\delta > 0$ . That is, for each  $f \in \mathcal{F}$  there is a  $\phi \in \mathcal{F}_\delta$  such that  $\|f - \phi\|_2 \leq \delta$ . Define the estimator  $\hat{f}$  by

$$\hat{f} = \operatorname{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi),$$

where  $\gamma_n(\phi)$  is defined in (3). Typically we would like to choose a  $\delta$ -net of minimal cardinality. We assume that  $\mathcal{F}$  is bounded in the  $L_2$  metric,

$$(8) \quad \sup_{g \in \mathcal{F}} \|g\|_2 \leq B_2,$$

where  $0 < B_2 < \infty$ .

*An upper bound to MISE.* Theorem 1 gives a bound for the mean integrated squared error of the estimate. We may identify the first term in the bound as a bias term and the second term as a variance term. The variance term depends on the operator norm of  $Q$  over the  $\delta$ -net  $\mathcal{F}_\delta$ . We define this operator norm as

$$(9) \quad \varrho(Q, \mathcal{F}_\delta) = \max_{\phi, \phi' \in \mathcal{F}_\delta, \phi \neq \phi'} \frac{\|Q(\phi - \phi')\|_2}{\|\phi - \phi'\|_2}, \quad \delta > 0,$$

where  $Q$  is defined by (4). In the case of density estimation we need the additional assumption that  $\varrho(Q, \mathcal{F}_\delta) \geq 1$  and that  $A\mathcal{F}$  and  $Q\mathcal{F}$  are bounded in the  $L_\infty$  metric:

$$(10) \quad \varrho(Q, \mathcal{F}_\delta) \geq 1, \quad \sup_{f \in \mathcal{F}} \|Af\|_\infty \leq B_\infty, \quad \sup_{f \in \mathcal{F}} \|Qf\|_\infty \leq B'_\infty,$$

where  $0 < B_\infty, B'_\infty < \infty$ .

**THEOREM 1.** *For the density estimation we assume that (10) is satisfied. We have that for  $f \in \mathcal{F}$ ,*

$$E \left\| \hat{f} - f \right\|_2^2 \leq C_1 \delta^2 + C_2 \frac{\varrho^2(Q, \mathcal{F}_\delta) \cdot (\log_e(\#\mathcal{F}_\delta) + 1)}{n},$$

where

$$(11) \quad C_1 = (1 - 2\xi)^{-1}(1 + 2\xi),$$

$$(12) \quad C_2 = (1 - 2\xi)^{-1}\xi C_\tau,$$

$$(13) \quad C_\tau > 0,$$

and  $\xi$  is such that

$$(14) \quad \begin{cases} C_\tau^{-1} \left( 4B'_\infty/3 + \sqrt{2[8(B'_\infty)^2/9 + C_\tau B_\infty]} \right) \leq \xi < 1/2, & \text{density estimation} \\ \sqrt{2/C_\tau} \leq \xi < 1/2, & \text{white noise.} \end{cases}$$

A proof of Theorem 1 is given in Section 7.2.

**REMARK 1.** Theorem 1 shows that the  $\delta$ -net estimator achieves the rate of convergence  $\psi_n$ , when  $\psi_n$  is the solution of the equation

$$(15) \quad \psi_n^2 \asymp n^{-1} \varrho^2(Q, \mathcal{F}_{\psi_n}) \log(\#\mathcal{F}_{\psi_n}).$$

We calculate the rate under the assumptions that  $\log(\#\mathcal{F}_\delta)$  and  $\varrho(Q, \mathcal{F}_\delta)$  increase polynomially as  $\delta$  decreases: we assume that one can find a  $\delta$ -net whose cardinality satisfies

$$\log(\#\mathcal{F}_\delta) = C\delta^{-b}$$

for some constants  $b, C > 0$  and we assume that

$$\varrho(Q, \mathcal{F}_\delta) = C'\delta^{-a}$$

for some  $a, C' > 0$  (in the direct case  $a = 0$  and  $C' = 1$ ). Then (15) can be written as  $\psi_n^2 \asymp n^{-1} \psi_n^{-2a-b}$  and the rate of the  $\delta$ -net estimator is

$$(16) \quad \psi_n \asymp n^{-1/[2(a+1)+b]}.$$

Let  $\mathcal{F}$  be a set of  $s$ -smooth  $d$ -dimensional functions, so that  $b = d/s$ . Then the rate is

$$\psi_n \asymp n^{-s/[2(a+1)s+d]},$$

which gives for the direct case  $a = 0$  the classical rate  $\psi_n \asymp n^{-s/(2s+d)}$ .

**3. A lower bound for MISE.** Theorem 2 gives a lower bound for the mean integrated squared error of any estimator, when estimating densities or signal functions  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  in the function class  $\mathcal{F}$ . Theorem 2 holds also for nonlinear operators.

**THEOREM 2.** *Let  $A$  be a possibly nonlinear operator. Assume that for each sufficiently small  $\delta > 0$  we find a finite set  $\mathcal{D}_\delta \subset \mathcal{F}$  for which*

$$(17) \quad \min\{\|f - g\|_2 : f, g \in \mathcal{D}_\delta, f \neq g\} \geq C_0\delta$$

and

$$(18) \quad \begin{cases} \max\{\|f - g\|_2 : f, g \in \mathcal{D}_\delta\} \leq C_1\delta, & \text{white noise,} \\ \max\{D_K(f, g) : f, g \in \mathcal{D}_\delta\} \leq C_1\delta, & \text{density estimation,} \end{cases}$$

where  $D_K^2(f, g) = \int \log_e(f/g) f$  is the Kullback-Leibler distance, and  $C_0, C_1$  are positive constants. Denote

$$\varrho_K(A, \mathcal{D}_\delta) = \begin{cases} \frac{1}{\sqrt{2}} \max_{f, g \in \mathcal{D}_\delta, f \neq g} \frac{\|A(f-g)\|_2}{\|f-g\|_2}, & \text{white noise,} \\ \max_{f, g \in \mathcal{D}_\delta, f \neq g} \frac{D_K(Af, Ag)}{\|f-g\|_2}, & \text{density estimation.} \end{cases}$$

Let  $\psi_n$  be such that

$$(19) \quad \log_e(\#\mathcal{D}_{\psi_n}) \asymp n\psi_n^2 \varrho_K^2(A, \mathcal{D}_{\psi_n}),$$

where  $a_n \asymp b_n$  means that  $\liminf_{n \rightarrow \infty} a_n/b_n > 0$ . Assume that

$$(20) \quad \lim_{n \rightarrow \infty} n\psi_n^2 \varrho_K^2(A, \mathcal{D}_{\psi_n}) = \infty.$$

Then,

$$\liminf_{n \rightarrow \infty} \psi_n^{-2} \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E\|f - \hat{f}\|_2^2 > 0,$$

where the infimum is taken over all estimators. That is,  $\psi_n$  is a lower bound for the minimax rate of convergence.

A proof of Theorem 2 is given in Section 7.3.

**REMARK 2.** Theorem 2 shows that one can get a lower bound  $\psi_n$  for the rate of converge by solving the equation

$$(21) \quad \psi_n^2 \varrho_K^2(A, \mathcal{D}_{\psi_n}) \asymp n^{-1} \log_e(\#\mathcal{D}_{\psi_n}).$$

The upper bound in Theorem 1 depends on the operator norm of  $Q$ , defined in (9), whereas the lower bound depends on the operator norm of  $A$ . Note also that the operator norm  $\varrho(Q, \mathcal{F}_{\psi_n})$  is on the different side of the equation in (15) than the operator norm  $\varrho_K(A, \mathcal{D}_{\psi_n})$  in the equation (21).

REMARK 3. In the density estimation case one can easily check assumptions (18) and (20) if one assumes that the functions in  $AD_\delta$  are bounded and bounded away from 0. Then,

$$(22) \quad C' \cdot \|A(f - g)\|_2 \leq D_K(Af, Ag) \leq C \cdot \|A(f - g)\|_2.$$

and (18) and (20) follow by the corresponding conditions with Hilbert norms instead of Kullback-Leibler distances.

**4. Dense minimizer.** The dense minimizer minimizes the empirical risk over the whole function class  $\mathcal{F}$ . In contrast to the  $\delta$ -net estimator the minimization is not restricted to a  $\delta$ -net. We call this estimator “dense minimizer” because it is defined as a minimizer over a possibly uncountable function class. The  $\delta$ -net estimator is more widely applicable: it may be applied also to estimate unsmooth functions and it may be applied when the operator is severely ill-posed. The dense minimizer may be applied only for relatively smooth cases (the entropy integral has to converge). Because it works without a restriction to a  $\delta$ -net we have available a larger toolbox of numerical algorithms that can be applied.

*Definition of the estimator.* Let  $\mathcal{F}$  be a collection of functions  $f : \mathbf{R}^d \rightarrow \mathbf{R}$ , which are bounded in the  $L_2$  metric as in (8), and let the estimator  $\hat{f}$  be a minimizer of the empirical risk over  $\mathcal{F}$ , up to  $\epsilon > 0$ :

$$\gamma_n(\hat{f}) \leq \inf_{g \in \mathcal{F}} \gamma_n(g) + \epsilon,$$

where  $\gamma_n(\phi)$  is defined in (3). For clarity, we present separate theorems for the Gaussian white noise model and for the density estimation model.

#### 4.1. Gaussian white noise.

*An upper bound to MISE.* Let  $\mathcal{F}_\delta$ ,  $\delta > 0$ , be a  $\delta$ -net of  $\mathcal{F}$ , with respect to the  $L_2$  norm. Define

$$(23) \quad \varrho(Q, \mathcal{F}_\delta) = \max \left\{ \frac{\|Q(f - g)\|_2}{\|f - g\|_2} : f \in \mathcal{F}_\delta, g \in \mathcal{F}_{2\delta}, f \neq g \right\}, \quad \delta > 0,$$

where  $Q$  is the adjoint of the inverse of  $A$ , defined by (4). Define the entropy integral

$$(24) \quad G(\delta) \stackrel{def}{=} \int_0^\delta \varrho(Q, \mathcal{F}_u) \sqrt{\log_e(\#\mathcal{F}_u)} du, \quad \delta \in (0, B_2],$$

where  $B_2$  is the  $L_2$  bound defined by (8).

THEOREM 3. *Assume that*

1. *the entropy integral in (24) converges,*
2.  *$G(\delta)/\delta^2$  is decreasing on the interval  $(0, B_2]$ ,*
3.  *$\varrho(Q, \mathcal{F}_\delta) = c\delta^{-a}$ , where  $0 \leq a < 1$  and  $c > 0$ ,*
4.  *$\lim_{\delta \rightarrow 0} G(\delta)\delta^{a-1} = \infty$ ,*
5.  *$\delta \mapsto \varrho(Q, \mathcal{F}_\delta)\sqrt{\log_e(\#\mathcal{F}_\delta)}$  is decreasing on  $(0, B_2]$ .*

Let  $\psi_n$  be such that

$$(25) \quad \psi_n^2 \geq C n^{-1/2} G(\psi_n),$$

where  $C$  is a positive constant, and assume that  $\lim_{n \rightarrow \infty} n\psi_n^{2(1+a)} = \infty$ . Then, for  $f \in \mathcal{F}$ ,

$$E \|\hat{f} - f\|_2^2 \leq C' (\psi_n^2 + \epsilon),$$

for a positive constant  $C'$ , for sufficiently large  $n$ .

A proof of Theorem 3 is given in Section 7.4

REMARK 4. Assumption 5 is a technical assumption which is used to replace a Riemann sum by an entropy integral. We prefer to write the assumptions in terms of the entropy integral in order to make them more readable.

REMARK 5. We may write  $\varrho(Q, \mathcal{F}_\delta)$  in a simpler way when there exists minimal  $\delta$ -nets  $\mathcal{F}_\delta$  which are nested:

$$\mathcal{F}_{2\delta} \subset \mathcal{F}_\delta.$$

Then we may define alternatively

$$\varrho(Q, \mathcal{F}_\delta) = \max_{f, g \in \mathcal{F}_\delta, f \neq g} \frac{\|Q(f - g)\|_2}{\|f - g\|_2}.$$

REMARK 6. Theorem 3 and Theorem 4 show that the rate of convergence of the dense minimizer is the solution of the equation

$$(26) \quad \psi_n^2 = n^{-1/2} G(\psi_n).$$

To get the optimal rate the net  $\mathcal{F}_\delta$  is chosen so that its cardinality is minimal. In the polynomial case one can find a  $\delta$ -net whose cardinality satisfies

$$\log(\#\mathcal{F}_\delta) = C\delta^{-b}$$

for some constants  $b, C' > 0$  and the operator norm satisfies

$$\varrho(Q, \mathcal{F}_\delta) = C' \delta^{-a}$$

for some  $a, C' > 0$ . (In the direct case  $a = 0$  and  $C' = 1$ .) Thus the entropy integral  $G(\delta)$  is finite when  $\int_0^\delta u^{-a-b/2} du < \infty$ , which holds when

$$(27) \quad a + b/2 < 1.$$

Then (26) leads to  $\psi_n^2 \asymp n^{-1/2} \psi_n^{-a-b/2+1}$  and the rate of the dense minimization estimator is

$$(28) \quad \psi_n \asymp n^{-1/[2(a+1)+b]}.$$

This is the same rate as the rate of the  $\delta$ -net estimator given in (16). We have the following example. Let  $\mathcal{F}$  be a set of  $s$ -smooth  $d$ -dimensional functions, so that  $b = d/s$ . Then condition (27) may be written as a condition for the smoothness index  $s$ :

$$s > \frac{d}{2(1-a)}.$$

When the problem is direct, then  $a = 0$ , and we have the classical condition  $s > d/2$ . The rate is  $\psi_n \asymp n^{-s/[2(a+1)s+d]}$ , which gives for the direct case  $a = 0$  the classical rate  $\psi_n \asymp n^{-s/(2s+d)}$ .

*4.2. Density estimation.* Let us call a  $\delta$ -bracketing net of  $\mathcal{F}$  with respect to the  $L_2$  norm a set of pairs of functions  $\mathcal{F}_\delta = \{(g_j^L, g_j^U) : j = 1, \dots, N_\delta\}$  such that

1.  $\|g_j^L - g_j^U\|_2 \leq \delta$ ,  $j = 1, \dots, N_\delta$ ,
2. for each  $g \in \mathcal{F}$  there is  $j = j(g) \in \{1, \dots, N_\delta\}$  such that  $g_j^L \leq g \leq g_j^U$ .

Let us denote  $\mathcal{F}_\delta^L = \{g_j^L : j = 1, \dots, N_\delta\}$  and  $\mathcal{F}_\delta^U = \{g_j^U : j = 1, \dots, N_\delta\}$ . Define

$$(29) \quad \varrho_{den}(Q, \mathcal{F}_\delta) = \max \left\{ \varrho(Q, \mathcal{F}_\delta^L, \mathcal{F}_\delta^U), \varrho(Q, \mathcal{F}_\delta^L, \mathcal{F}_{2\delta}^L) \right\},$$

where

$$\varrho(Q, \mathcal{F}_\delta^L, \mathcal{F}_\delta^U) = \max \left\{ \frac{\|Q(g^U - g^L)\|_2}{\|g^U - g^L\|_2} : g^L \in \mathcal{F}_\delta^L, g^U \in \mathcal{F}_\delta^U \right\}$$

and

$$\varrho(Q, \mathcal{F}_\delta^L, \mathcal{F}_{2\delta}^L) = \max \left\{ \frac{\|Q(f - g)\|_2}{\|f - g\|_2} : f \in \mathcal{F}_\delta^L, g \in \mathcal{F}_{2\delta}^L, f \neq g \right\},$$

for  $\delta > 0$ . Define the entropy integral

$$(30) \quad G(\delta) \stackrel{\text{def}}{=} \int_0^\delta \varrho_{den}(Q, \mathcal{F}_u) \sqrt{\log_e(\#\mathcal{F}_u)} du, \quad \delta \in (0, B_2],$$

where  $B_2 = \sup_{f \in \mathcal{F}} \|f\|_2$ .

**THEOREM 4.** *We make the Assumptions 1-5 of Theorem 3 (with operator norm  $\varrho_{den}(Q, \mathcal{F}_\delta)$  in place of  $\varrho(Q, \mathcal{F}_\delta)$ ), and in addition we assume that  $\sup_{f \in \mathcal{F}} \|Af\|_\infty < \infty$ ,  $\sup_{g \in \mathcal{F}_{B_2}^L \cup \mathcal{F}_{B_2}^U} \|Qg\|_\infty < \infty$ , and that the operator  $Q$  preserves positivity ( $g \geq 0$  implies that  $Qg \geq 0$ ). Let  $\psi_n$  be such that*

$$(31) \quad \psi_n^2 \geq C n^{-1/2} G(\psi_n),$$

for a positive constant  $C$ , and assume that  $\lim_{n \rightarrow \infty} n\psi_n^{2(1+a)} = \infty$ . Then, for  $f \in \mathcal{F}$ ,

$$E \left\| \hat{f} - f \right\|_2^2 \leq C' \left( \psi_n^2 + \epsilon \right),$$

for a positive constant  $C'$ , for sufficiently large  $n$ .

A proof of Theorem 4 is given in Section 7.5. An analogous discussion of optimal rates as in Remark 6 for the Gaussian white noise model also applies for dense density estimators.

**5. Examples of operators.** As examples for operators we consider convolution operators and the Radon transform. The definition of the empirical risk involves the adjoint of the inverse of the operator  $A$ , and we calculate the adjoint of the inverse of  $A$ , when  $A$  is a convolution operator or the Radon transform.

5.1. *Convolution.* The convolution operator  $A$  is defined by

$$Af = a * f, \quad f : \mathbf{R}^d \rightarrow \mathbf{R},$$

where  $a : \mathbf{R}^d \rightarrow \mathbf{R}$  is a known integrable function. The adjoint of the inverse of  $A$  is  $Q$ , defined for  $g : \mathbf{R}^d \rightarrow \mathbf{R}$ , by

$$(32) \quad Qg = F^{-1} \left( \frac{Fg}{Fa} \right),$$

where  $F$  denotes the Fourier transform. To derive this equation note that, for  $h : \mathbf{R}^d \rightarrow \mathbf{R}$ ,

$$FA^{-1}h = \frac{Fh}{Fa}.$$

Thus, for  $h : \mathbf{R}^d \rightarrow \mathbf{R}$ ,  $g : \mathbf{R}^d \rightarrow \mathbf{R}$ , applying two times Parseval's theorem give

$$\int_{\mathbf{R}^d} (A^{-1}h)g = (2\pi)^d \int_{\mathbf{R}^d} \frac{(Fh)(Fg)}{Fa} = \int_{\mathbf{R}^d} h(Qg).$$

Convolution operators appear in density estimation when the observations contain additional measurement errors. In the errors-in-variables model we observe  $Y_i = X_i + \epsilon_i$ ,  $i = 1, \dots, n$ , where  $X_i \sim f$ ,  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  is the unknown density which we want to estimate, and  $\epsilon_i \sim a$  are the measurement errors. The density of the observations  $Y_i$  is  $Af = a * f$ .

**5.2. Radon transform.** The Radon transform has been discussed in a series of papers and books including Deans (1983) and Natterer (2001). The Radon transform is defined as the integral of a  $d$ -dimensional function over  $d - 1$ -dimensional hyperplanes. We parameterize the  $d - 1$ -dimensional hyperplanes in the  $d$ -dimensional Euclidean space with the help of a direction vector  $\xi \in \mathbf{S}_{d-1}$  and a distance from the origin  $u \in [0, \infty)$ :

$$(33) \quad P_{\xi,u} = \{z \in \mathbf{R}^d : z^T \xi = u\}, \quad \xi \in \mathbf{S}_{d-1}, u \in [0, \infty).$$

Define the Radon transform for  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  as

$$(Af)(\xi, u) = \int_{P_{\xi,u}} f, \quad \xi \in \mathbf{S}_{d-1}, u \in [0, \infty),$$

where the integration is with respect to the  $d - 1$ -dimensional Lebesgue measure. We will take the Radon transform as a mapping from functions  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  to functions  $Af : \mathbf{Y} \rightarrow \mathbf{R}$ , where  $\mathbf{Y} = \mathbf{S}_{d-1} \times [0, \infty)$ , and the measure  $\nu$  of the Borel space  $(\mathbf{Y}, \mathcal{Y}, \nu)$  is taken to be  $d\nu(\xi, u) = u^{d-1} du d\mu(\xi)$ .

The adjoint of the inverse of  $A$  is  $Q$ , defined for  $g : \mathbf{R}^d \rightarrow \mathbf{R}$ , by

$$(34) \quad (Qg)(\xi, u) = (2\pi)^{d-1} \cdot (F_1^{-1} \mathcal{I}_\xi g)(u), \quad \xi \in \mathbf{S}_{d-1}, u \in [0, \infty),$$

where

$$(\mathcal{I}_\xi g)(t) = (Fg)(t\xi), \quad \xi \in \mathbf{S}_{d-1}, t \in [0, \infty).$$

To see this note first that, for  $h : \mathbf{S}_{d-1} \times [0, \infty) \rightarrow \mathbf{R}$ , we have that

$$(35) \quad (FA^{-1}h)(\omega) = (\mathcal{H}_{\omega/\|\omega\|} h)(\|\omega\|), \quad \omega \in \mathbf{R}^d,$$

where  $\mathcal{H}_\xi$  is the Fourier transform of  $h(\xi, \cdot)$  for fixed  $\xi \in \mathbf{S}_{d-1}$ :

$$\mathcal{H}_\xi h = F_1(h(\xi, \cdot)), \quad \xi \in \mathbf{S}_{d-1}.$$

Equation (35) follows directly from the projection theorem, see Natterer (2001).

Two applications of Parseval's theorem and (35) give for  $h : \mathbf{S}_{d-1} \times [0, \infty) \rightarrow \mathbf{R}$ ,  $g : \mathbf{R}^d \rightarrow \mathbf{R}$ , that

$$\begin{aligned} \int_{\mathbf{R}^d} (A^{-1}h)g &= (2\pi)^d \int_{\mathbf{R}^d} (\mathcal{H}_{\omega/\|\omega\|}h)(\|\omega\|)(Fg)(\omega) d\omega \\ &= (2\pi)^d \int_{\mathbf{S}_{d-1}} \int_0^\infty t^{d-1} (\mathcal{H}_\xi h)(t)(Fg)(t\xi) dt d\mu(\xi) \\ &= (2\pi)^{d-1} \int_{\mathbf{S}_{d-1}} \int_0^\infty u^{d-1} h(\xi, u)(F_1^{-1}I_\xi g)(u) du d\mu(\xi) \\ &= \int_{\mathbf{Y}} h(Qg). \end{aligned}$$

This shows (34).

*2D Radon transform.* In the 2D case we consider reconstructing a 2-dimensional function from observations of its integrals over lines. Let  $D = \{x \in \mathbf{R}^2 : \|x\| \leq 1\}$  be the unit disk in  $\mathbf{R}^2$ . The plane in (33) can be written as  $P_{\xi, u} = \{u\xi + t\xi^\perp : t \in \mathbf{R}\}$ , where  $\xi^\perp$  is a vector which is orthogonal to  $\xi$ . We can write  $\xi = (\cos \phi, \sin \phi)$  and  $\xi^\perp = (-\sin \phi, \cos \phi)$ . Thus we parameterize the lines by the length  $u \in [0, 1]$  of the perpendicular from the origin to the line and by the orientation  $\phi \in [0, 2\pi)$  of this perpendicular. A common way to define 2D Radon transform is

$$(36) \quad Af(u, \phi) = \frac{\pi}{2\sqrt{1-u^2}} \int_{\sqrt{1-u^2}}^{\sqrt{1-u^2}} f(u \cos \phi - t \sin \phi, u \sin \phi + t \cos \phi) dt,$$

where  $(u, \phi) \in \mathbf{Y} = [0, 1] \times [0, 2\pi]$ , and we suppose that  $f \in L_1(D) \cap L_2(D)$ . Now the Radon transform is  $\pi$  times the average of  $f$  over the line segment that intersects  $D$ . We consider  $Rf$  as the element of  $L_2(\mathbf{Y}, \nu)$ , where  $\nu$  is the measure defined by  $d\nu(u, \phi) = 2\pi^{-1}\sqrt{1-u^2} du d\phi$ .

*Tomography.* The positron emission tomography is a density estimation problem but the X-ray tomography is a regression type problem. In the setting of positron emission tomography events happen at points  $X_1, \dots, X_n \in \mathbf{R}^d$ , and these points are i.i.d. with density  $f$ . We do not observe the location of the points but only that an event has occurred on a hyperplane containing the point. We assume that the hyperplane is uniformly oriented, and that the distance of the hyperplane from the origin is given by the Radon transform:

$$(37) \quad S \sim \text{Unif}(\mathbf{S}_{d-1}), \quad U | S = \xi \sim (Af)(\xi, \cdot),$$

where hyperplanes are written as  $\{z \in \mathbf{R}^d : z^T S = U\}$ . We assume to observe i.i.d random variables  $Y_i = (S_i, U_i) \in \mathbf{S}_{d-1} \times [0, \infty)$ ,  $i = 1, \dots, n$ , which

are distributed as  $(S, U)$ , This is equivalent to observing the hyperplanes  $\{z \in \mathbf{R}^d : z^T S_i = U_i\}$ . We want to estimate the density  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  in (37). The density of the observations  $Y_i$  is equal to

$$(38) \quad (\tilde{A}f)(\xi, u) = \frac{1}{\mu(\mathbf{S}_{d-1})} (Af)(\xi, u), \quad \xi \in \mathbf{S}_{d-1}, \quad u \in [0, \infty).$$

## 6. Examples of function spaces.

6.1. *Ellipsoidal function spaces.* Since we are in the  $L_2$  setting it is natural to work in the sequence space; we define the function classes as ellipsoids. We shall apply singular value decompositions of the operators and wavelet-vaguelette systems in the calculation of the rates of convergence. In Section 6.1.1 we calculate the operator norms in the framework of singular value decomposition. In Section 6.1.2 we calculate the operator norms in the wavelet-vaguelette framework. Section 6.1.3 derives the rate of convergence of the  $\delta$ -net estimator for the case of a convolution operator and the Radon transform, and the lower bound for the rate of convergence of any estimator.

6.1.1. *Singular value decomposition.* We assume that the underlying function space  $\mathcal{F}$  consists of  $d$ -variate functions that are linear combinations of orthonormal basis functions  $\phi_j$  with multi-index  $j = (j_1, \dots, j_d) \in \{0, 1, \dots\}^d$ . Define the ellipsoid and the corresponding collection of functions by

$$(39) \quad \Theta = \left\{ \theta : \sum_{j_1=0, \dots, j_d=0}^{\infty} a_j^2 \theta_j^2 \leq L^2 \right\}, \quad \mathcal{F} = \left\{ \sum_{j_1=0, \dots, j_d=0}^{\infty} \theta_j \phi_j : \theta \in \Theta \right\}.$$

$\delta$ -net and  $\delta$ -packing set for polynomial ellipsoids. We assume that there exists positive constants  $C_1, C_2$  such that for all  $j \in \{0, 1, \dots\}^d$

$$(40) \quad C_1 \cdot |j|^s \leq a_j \leq C_2 \cdot |j|^s,$$

where  $|j| = j_1 + \dots + j_d$ . We construct a  $\delta$ -net  $\Theta_\delta$  and a  $\delta$ -packing set  $\Theta_\delta^*$  in Appendix A. Since the construction is in the sequence space we define the  $\delta$ -net and  $\delta$ -packing set of  $\mathcal{F}$  by

$$(41) \quad \mathcal{F}_\delta = \left\{ \sum_{j_1=0, \dots, j_d=0}^{\infty} \theta_j \phi_j : \theta \in \Theta_\delta \right\}, \quad \mathcal{D}_\delta = \left\{ \sum_{j_1=0, \dots, j_d=0}^{\infty} \theta_j \phi_j : \theta \in \Theta_\delta^* \right\}.$$

The set  $\Theta_\delta$  is such that for  $\theta \in \Theta_\delta$

$$\theta_j = 0, \quad \text{when } j \notin \{1, \dots, M\}^d,$$

where

$$(42) \quad M \asymp \delta^{-1/s}.$$

Set  $\Theta_\delta^*$  is such that for all  $\theta \in \Theta_\delta^*$

$$(43) \quad \theta_j = \theta_j^*, \quad \text{when } j \notin \{M^*, \dots, M\}^d,$$

where  $\theta^*$  is a fixed sequence with  $\sum_{|j| \geq 0} a_j^2 \theta_j^{*2} = L^* < L$ ,

$$M^* = \lfloor M/2 \rfloor.$$

Furthermore, it holds that

$$(44) \quad \log(\#\Theta_\delta) \leq C\delta^{-d/s}, \quad \log(\#\Theta_\delta^*) \geq C'\delta^{-d/s}.$$

*Operator norms.* We calculate the operator norms  $\varrho(Q, \mathcal{F}_\delta)$  and  $\varrho_K(A, \mathcal{D}_\delta)$  in the ellipsoidal framework, where  $\mathcal{F}_\delta$  and  $\mathcal{D}_\delta$  are defined in (41) and Appendix A. We apply the singular value decomposition of  $A$ . We assume that the domain of  $A$  is a separable Hilbert space  $H$  with inner product  $\langle \cdot, \cdot \rangle$ . The underlying function space  $\mathcal{F}$  satisfies  $\mathcal{F} \subset H$ . We denote with  $A^*$  the adjoint of  $A$ . We assume that  $A^*A$  is a compact operator on  $H$  with eigenvalues  $(b_j^2)$ ,  $b_j > 0$ ,  $j \in \{0, 1, \dots\}^d$ , with orthonormal system of eigenfunctions  $\phi_j$ . We assume that there exists positive constants  $q$  and  $C_1, C_2$  such that for all  $j \in \{0, 1, \dots\}^d$

$$(45) \quad C_1 \cdot |j|^{-q} \leq b_j \leq C_2 \cdot |j|^{-q}.$$

Let  $g, g'$  in  $\mathcal{F}_\delta$  or in  $\mathcal{D}_\delta$ , respectively. Write

$$g - g' = \sum_{j_1=1, \dots, j_d=1}^{\infty} (\theta_j - \theta'_j) \phi_j.$$

1. The functions  $Q\phi_j$  are orthogonal and  $\|Q\phi_j\|_2 = b_j^{-1}$ . Indeed,  $Q = (A^{-1})^*$ , and thus

$$\langle Q\phi_j, Q\phi_l \rangle = \langle \phi_j, A^{-1}(A^{-1})^*\phi_l \rangle = b_l^{-2} \langle \phi_j, \phi_l \rangle,$$

where we used the fact <sup>1</sup>

$$A^{-1}(A^{-1})^*\phi_l = A^{-1}(A^*)^{-1}\phi_l = (A^*A)^{-1}\phi_l = b_l^{-2}\phi_l.$$

---

<sup>1</sup>Note that when a bounded linear operator  $A$  between Banach spaces has a bounded inverse, then  $(A^{-1})^* = (A^*)^{-1}$ , see Dunford & Schwartz (1958), Section VI, Lemma 7, page 479.

Thus for  $g, g' \in \mathcal{F}_\delta$ ,

$$\begin{aligned}
\|Q(g - g')\|_2^2 &= \left\| \sum_{j_1=0, \dots, j_d=0}^M (\theta_j - \theta'_j)^2 Q\phi_j \right\|_2^2 \\
&= \sum_{j_1=0, \dots, j_d=0}^M (\theta_j - \theta'_j)^2 b_j^{-2} \\
(46) \quad &\leq CM^{2q} \sum_{j_1=0, \dots, j_d=0}^M (\theta_j - \theta'_j)^2,
\end{aligned}$$

where we used (45) to infer that when  $j \in \{0, \dots, M\}^d$ , then

$$b_j^{-2} \leq C_1^{-2} \cdot |j|^{2q} \leq C_1^{-2} \cdot (dM)^{2q}.$$

On the other hand,  $\|g - g'\|_2 = \sum_{j_1=0, \dots, j_d=0}^M (\theta_j - \theta'_j)^2$ . This gives the upper bound for the operator norm

$$(47) \quad \varrho(Q, \mathcal{F}_\delta) \leq CM^q \leq C'\delta^{-q/s},$$

by the definition of  $M$  in (42).

2. The functions  $A\phi_j$  are orthogonal and  $\|A\phi_j\|_2 = b_j$ . Indeed,

$$\langle A\phi_j, A\phi_l \rangle = \langle \phi_j, A^* A\phi_l \rangle = b_l^2 \langle \phi_j, \phi_l \rangle.$$

Thus for  $g, g' \in \mathcal{D}_\delta$ ,

$$\begin{aligned}
\|A(g - g')\|_2^2 &= \sum_{j_1=M^*, \dots, j_d=M^*}^M (\theta_j - \theta'_j)^2 \|A\phi_j\|_2^2 \\
&= \sum_{j_1=M^*, \dots, j_d=M^*}^M (\theta_j - \theta'_j)^2 b_j^2.
\end{aligned}$$

This and similar calculations as in (46) imply that

$$(48) \quad C'\delta^{q/s} \leq \varrho_K(A, \mathcal{D}_\delta) \leq C\delta^{q/s}.$$

6.1.2. *Wavelet-vaguelette decomposition.* We assume that the underlying function space  $\mathcal{F}$  consists of  $d$ -variate functions which are linear combinations of orthonormal wavelet functions  $(\phi_{jk})$ , where  $j \in \{0, 1, \dots\}$  and  $k \in \{0, \dots, 2^j - 1\}^d$ . The  $l_2$ -body and the corresponding class of functions can now be defined as

$$\Theta = \left\{ \theta : \sum_j 2^{2sj} \sum_k |\theta_{jk}|^2 \leq L^2 \right\}, \quad \mathcal{F} = \left\{ \sum_j \sum_k \theta_{jk} \phi_{jk} : \theta \in \Theta \right\},$$

where  $s > 0$ . We have already constructed a  $\delta$ -net and  $\delta$ -packing set for the  $l_2$ -bodies in (41), but in the current setting for  $\theta \in \Theta_\delta$

$$\theta_{jk} = 0, \quad \text{when } j \geq J + 1,$$

where

$$(49) \quad 2^J \asymp \delta^{-1/s}$$

and for  $\theta \in \Theta_\delta^*$

$$\theta_{jk} = \theta_{jk}^*, \quad \text{when } j \leq J^* \text{ or } j \geq J + 1,$$

where  $\theta^*$  is a fixed sequence with  $\sum_{j=0}^{\infty} \sum_k a_j^2 \theta_{jk}^{*2} = L^* < L$ , and  $J^* = J - 1$ .

*Operator norms.* We can apply the wavelet-vaguelette decomposition, as defined in Donoho (1995), to calculate the operator norms  $\varrho(Q, \mathcal{F}_\delta)$  and  $\varrho_K(A, \mathcal{D}_\delta)$ . We have available the following three sets of functions:  $(\phi_{jk})_{jk}$  is an orthogonal wavelet basis and  $(u_{jk})_{jk}$  and  $(v_{jk})_{jk}$  are near-orthogonal sets:

$$\left\| \sum_{jk} a_{jk} u_{jk} \right\|_2 \asymp \|(a_{jk})\|_{l_2}, \quad \left\| \sum_{jk} a_{jk} v_{jk} \right\|_2 \asymp \|(a_{jk})\|_{l_2},$$

where  $a \asymp b$  means that there exists positive constants  $C, C'$  such that  $Cb \leq a \leq C'a$ . The following quasi-singular relations hold:

$$A\phi_{jk} = \kappa_j v_{jk}, \quad A^* u_{jk} = \kappa_j \phi_{jk},$$

where  $\kappa_j$  are quasi-singular values. We assume that there exists positive constants  $q$  and  $C_1, C_2$  such that for all  $j \in \{0, 1, \dots\}$

$$(50) \quad C_1 \cdot 2^{-qj} \leq \kappa_j \leq C_2 \cdot 2^{-qj}.$$

1. Let  $g, g' \in \mathcal{F}_\delta$ . Write

$$g - g' = \sum_{j=0}^J \sum_k (\theta_{jk} - \theta'_{jk}) \phi_{jk}.$$

Since  $Q = (A^{-1})^*$ , then  $QA^* = (AA^{-1})^* = I$ . Thus,

$$\begin{aligned} \langle Q\phi_{jk}, Q\phi_{j'k'} \rangle &= \kappa_j^{-1} \kappa_{j'}^{-1} \langle QA^* u_{jk}, QA^* u_{j'k'} \rangle \\ &= \kappa_j^{-1} \kappa_{j'}^{-1} \langle u_{jk}, u_{j'k'} \rangle. \end{aligned}$$

Thus,

$$\begin{aligned}
\|Q(g - g')\|_2^2 &= \left\| \sum_{j=0}^J \sum_k (\theta_{jk} - \theta'_{jk}) Q\phi_{jk} \right\|_2^2 \\
&= \left\| \sum_{j=0}^J \kappa_j^{-1} \sum_k (\theta_{jk} - \theta'_{jk}) u_{jk} \right\|_2^2 \\
&\asymp \sum_{j=0}^J \kappa_j^{-2} \sum_k (\theta_{jk} - \theta'_{jk})^2 \\
(51) \quad &\leq C 2^{2qJ} \sum_{j=0}^J \sum_k (\theta_{jk} - \theta'_{jk})^2,
\end{aligned}$$

where we used (50) to infer that when  $j \in \{0, \dots, J\}$ , then

$$\kappa_j^{-2} \leq C_1^{-2} \cdot 2^{2qj} \leq C_1^{-2} \cdot 2^{2qJ}.$$

On the other hand,  $\|g - g'\|_2^2 = \sum_{j=0}^J \sum_k (\theta_{jk} - \theta'_{jk})^2$ . This gives the upper bound for the operator norm

$$\varrho(Q, \mathcal{F}_\delta) \leq C 2^{qJ} \leq C' \delta^{-q/s},$$

by the definition of  $J$  in (49).

2. We have  $\langle A\phi_{jk}, A\phi_{j'k'} \rangle = \kappa_j \kappa_{j'} \langle v_{jk}, v_{j'k'} \rangle$  and  $(v_{jk})$  is a near-orthogonal set. Thus, similarly as in (51), we get

$$C' \delta^{q/s} \leq \varrho_K(A, \mathcal{D}_\delta) \leq C \delta^{q/s}.$$

**6.1.3. Rates of convergence.** We derive the rates of convergence for the  $\delta$ -net estimator when the operator is a convolution operator and the Radon transform. It is also shown that the lower bounds have the same order as the upper bounds. We give examples in the setting of the Gaussian white noise model.

*Convolution.* Let  $A$  be a convolution operator:  $Af = a * f$  where  $a : \mathbf{R}^d \rightarrow \mathbf{R}$  is a known function. Denote

$$\phi_{jk}(x) = \prod_{i=1}^d \sqrt{2} [(1 - k_i) \cos(2\pi j_i x_i) + k_i \sin(2\pi j_i x_i)], \quad x \in [0, 1]^d,$$

where  $j \in \{0, 1, \dots\}^d$ ,  $k \in K_j$ , where

$$K_j = \left\{ k \in \{0, 1\}^d : k_i = 0, \text{ when } j_i = 0 \right\}.$$

The cardinality of  $K_j$  is  $2^{d-\alpha(j)}$ , where  $\alpha(j) = \#\{j_i : j_i = 0\}$ . The collection  $(\phi_{jk}), (j, k) \in \{0, 1, \dots\}^d \times K_j$ , is a basis for 1-periodic functions on  $L_2([0, 1]^d)$ . When the convolution kernel  $a$  is an 1-periodic function in  $L_2([0, 1]^d)$ , then we can write

$$a(x) = \sum_{j_1=0, \dots, j_d=0}^{\infty} \sum_{k \in K_j} b_{jk} \phi_{jk}(x).$$

The functions  $\phi_{jk}$  are the singular functions of the operator  $A$  and the values  $b_{jk}$  are the corresponding singular values. We assume that the underlying function space is equal to

$$(52) \quad \mathcal{F} = \left\{ \sum_{j_1=0, \dots, j_d=0}^{\infty} \sum_{k \in K_j} \theta_{jk} \phi_{jk}(x) : (\theta_{jk}) \in \Theta \right\},$$

where

$$(53) \quad \Theta = \left\{ \theta : \sum_{j_1=0, \dots, j_d=0}^{\infty} \sum_{k \in K_j} a_{jk}^2 \theta_{jk}^2 \leq L^2 \right\}.$$

We give the rate of convergence of the  $\delta$ -net estimator and show that the estimator achieves the optimal rate of convergence. Optimal rates of convergence has been previously obtained for the convolution problem in various settings in Ermakov (1989), Donoho & Low (1992), Koo (1993), Korostelev & Tsybakov (1993).

**COROLLARY 1.** *Let  $\mathcal{F}$  be the function class as defined in (52). We assume that the coefficients of the ellipsoid (53) satisfy*

$$C_0 |j|^s \leq a_{jk} \leq C_1 |j|^s.$$

*for some  $s > 0$  and  $C_0, C_1 > 0$ . We assume that the convolution filter  $a$  is 1-periodic function in  $L_2([0, 1]^d)$  and that the Fourier coefficients of filter  $a$  satisfy*

$$C_2 |j|^{-q} \leq b_{jk} \leq C_3 |j|^{-q}$$

*for some  $q \geq 0$ ,  $C_2, C_3 > 0$ . Then,*

$$\limsup_{n \rightarrow \infty} n^{2s/(2s+2q+d)} \sup_{f \in \mathcal{F}} E_f \left\| \hat{f} - f \right\|_2^2 < \infty,$$

*where  $\hat{f}$  is the  $\delta$ -net estimator. Also,*

$$\liminf_{n \rightarrow \infty} n^{2s/(2s+2q+d)} \inf_{\hat{g}} \sup_{f \in \mathcal{F}} E_f \left\| \hat{g} - f \right\|_2^2 > 0,$$

*where the infimum is taken over any estimators  $\hat{g}$ .*

*Proof.* For the upper bound we apply Theorem 1. Let  $\mathcal{F}_\delta$  be the  $\delta$ -net of  $\mathcal{F}$  as constructed in (41). We have shown in (47) that

$$\varrho(Q, \mathcal{F}_\delta) \leq C\delta^{-a},$$

where  $a = q/s$ . We have stated in (44) that the cardinality of the  $\delta$ -net satisfies

$$\log(\#\mathcal{F}_\delta) \leq C\delta^{-b},$$

where  $b = d/s$ . Thus we may apply (16) to get the rate

$$\psi_n = n^{-1/(2(a+1)+b)} = n^{-s/(2s+2q+d)}.$$

The upper bound is proved. For the lower bound we apply Theorem 2. Assumption (17) holds because  $\mathcal{D}_\delta$  in (41) is a  $\delta$ -packing set. Assumption (18) holds by the construction, see (94) in Appendix A. Assumptions (19) and (20) follow from (44) and (48). Thus the lower bound is proved.  $\square$

*Radon transform.* We consider the 2D Radon transform as defined in (36). The singular value decomposition of the Radon transform can be found in Deans (1983). Let

$$\tilde{\phi}_{jk}(r, \theta) = \pi^{-1/2}(j+k+1)^{1/2} Z_{j+k}^{|j-k|}(r) e^{i(j-k)\theta}, \quad (r, \theta) \in D = [0, 1] \times [0, 2\pi),$$

where  $Z_a^b$  denotes the Zernike polynomial of degree  $a$  and order  $b$ . Functions  $\tilde{\phi}_{jk}$ ,  $j, k = 0, 1, \dots$ ,  $(j, k) \neq (0, 0)$ , constitute an orthonormal complex-valued basis for  $L_2(D)$ . The corresponding orthonormal functions in  $L_2(\mathbf{Y}, \nu)$  are

$$\tilde{\psi}_{jk}(u, \phi) = \pi^{-1/2} U_{j+k}(u) e^{i(j-k)\phi}, \quad (u, \phi) \in \mathbf{Y} = [0, 1] \times [0, 2\pi),$$

where  $U_m(\cos \theta) = \sin((m+1)\theta)/\sin \theta$  are the Chebyshev polynomials of the second kind. We have

$$A\tilde{\phi}_{jk} = b_{jk}\tilde{\psi}_{jk},$$

where the singular values are

$$(54) \quad b_{jk} = \pi^{-1}(j+k+1)^{-1/2}.$$

We shall identify the complex bases with the equivalent real orthonormal bases by

$$\phi_{jk} = \begin{cases} \sqrt{2} \operatorname{Re}(\tilde{\phi}_{jk}) & \text{if } j > k \\ \tilde{\phi}_{jk} & \text{if } j = k \\ \sqrt{2} \operatorname{Im}(\tilde{\phi}_{jk}) & \text{if } j < k. \end{cases}$$

We assume that the underlying function space is equal to

$$(55) \quad \mathcal{F} = \left\{ \sum_{j_1=0, j_2=0, (j_1, j_2) \neq (0,0)}^{\infty} \theta_{j_1 j_2} \phi_{j_1 j_2}(x) : (\theta_{j_1 j_2}) \in \Theta \right\},$$

where

$$(56) \quad \Theta = \left\{ \theta : \sum_{j_1=0, j_2=0, (j_1, j_2) \neq (0,0)}^{\infty} a_{j_1 j_2}^2 \theta_{j_1 j_2}^2 \leq L^2 \right\}.$$

We give the rate of convergence of the  $\delta$ -net estimator and show that the estimator achieves the optimal rate of convergence. Optimal rates of convergence have been previously obtained in Johnstone & Silverman (1990), Korostelev & Tsybakov (1991), Donoho & Low (1992), Korostelev & Tsybakov (1993).

**COROLLARY 2.** *Let  $\mathcal{F}$  be the function class as defined in (55). We assume that the coefficients of the ellipsoid (56) satisfy*

$$C_0 |j|^s \leq a_{jk} \leq C_1 |j|^s.$$

for some  $s > 0$  and  $C_0, C_1 > 0$ . Then, for  $d = 2$ ,

$$\limsup_{n \rightarrow \infty} n^{2s/(2s+2d-1)} \sup_{f \in \mathcal{F}} E_f \left\| \hat{f} - f \right\|_2^2 < \infty.$$

where  $\hat{f}$  is the  $\delta$ -net estimator. Also,

$$\liminf_{n \rightarrow \infty} n^{2s/(2s+2d-1)} \inf_{\hat{g}} \sup_{f \in \mathcal{F}} E_f \left\| \hat{g} - f \right\|_2^2 > 0,$$

where the infimum is taken over any estimators  $\hat{g}$ .

*Proof.* For the upper bound we apply Theorem 1. Let  $\mathcal{F}_\delta$  be the  $\delta$ -net of  $\mathcal{F}$  as constructed in (41). We have shown in (47) that

$$\varrho(Q, \mathcal{F}_\delta) \leq C \delta^{-a},$$

where  $a = q/s$  and  $q = 1/2$  (so that  $a = (d-1)/(2s)$ ), since the singular values are given in (54). We have stated in (44) that the cardinality of the  $\delta$ -net satisfies

$$\log(\#\mathcal{F}_\delta) \leq C \delta^{-b},$$

where  $b = d/s$ . Thus we may apply (16) to get the rate

$$\psi_n = n^{-s/(2s+2d-1)}.$$

The upper bound is proved. For the lower bound we apply Theorem 2 similarly as in the proof of Corollary 1.  $\square$

6.2. *Additive models.* In this section we will show that our approach can be used to prove oracle results for additive models. In additive models the unknown function  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  is assumed to have an additive decomposition  $f(x) = f_1(x_1) + \dots + f_d(x_d)$  with unknown additive components  $f_j : \mathbf{R} \rightarrow \mathbf{R}$ ,  $j = 1, \dots, d$ . We compare this model with theoretical oracle models where only one component function  $f_r$  is unknown, but the other functions  $f_j$  ( $j \neq r$ ) are known. We will show below that the function  $f$  can be estimated with the same rate of convergence as in the oracle model that has the slowest rate of convergence. In particular, if the rate of convergence is the same in all oracle models then the rate in the additive model remains the same. This is a well known fact for classical additive regression models, see e.g. Stone (1985). It efficiently avoids the curse of dimensionality in contrast to the full dimensional nonparametric model. Furthermore, it is practically important because it allows a flexible and nicely interpretable model for regression with high dimensional covariates, see e.g. Hastie & Tibshirani (1990) for a discussion of the additive and related models. Thus, our result will generalize the oracle result for additive models of Stone (1985) to inverse problems. For a theoretical discussion we will first use a slightly more general framework. We will come back to additive models afterwards.

6.2.1. *Abstract setting.* We assume that the function class  $\mathcal{F}$  is a subset of the direct sum of spaces  $\mathcal{F}_1, \dots, \mathcal{F}_p$ . All spaces contain functions from  $f : \mathbf{R}^d \rightarrow \mathbf{R}$ . At this stage, we do not assume that functions in  $\mathcal{F}_j$  ( $j = 1, \dots, p$ ) depend only on the argument  $x_j$ . An example of this more general set up are sums of smooth functions and indicator functions of convex sets or of sets with smooth boundary. We assume that a finite  $\delta$ -net  $\mathcal{F}_\delta$  of  $\mathcal{F}$  is a subset of the direct sum  $\mathcal{F}_{1,\delta} \oplus \dots \oplus \mathcal{F}_{p,\delta}$ , where  $\mathcal{F}_{j,\delta}$  are finite subsets of  $\mathcal{F}_j$ . We denote the number of elements of  $\mathcal{F}_{j,\delta}$  by  $\exp(\lambda_j)$ . Furthermore, we write  $\rho_j = \rho(Q, \mathcal{F}_{j,\delta})$ . We make the following essential geometrical assumption:

$$(57) \quad \|f_1 + \dots + f_p\|_2^2 \geq c \sum_{j=1}^p \|f_j\|_2^2$$

for a positive constant  $c > 0$ . For the  $\delta$ -net minimizer  $\hat{f}$  over the  $\delta$ -net  $\mathcal{F}_\delta$  we get the following result in the white noise model. (An additive model for density estimation would not make much sense.)

**THEOREM 5.** *We make assumption (57). In the white noise model the*

following bound holds for the  $\delta$ -net minimizer  $\hat{f}$ , for  $f \in \mathcal{F}$ ,

$$E \left( \|\hat{f} - f\|_2^2 \right) \leq 3\delta^2 + 32c^{-1}n^{-1} \left[ \sum_{j=1}^p \rho_j^2 \lambda_j + \left( \sum_{j=1}^p \rho_j \right)^2 \right].$$

A proof of Theorem 5 is given in Section 7.6.

**6.2.2. Application to additive models.** We now apply Theorem 5 for discussing additive models  $f(x) = f_1(x_1) + \dots + f_d(x_d)$ . In  $L_2(\mathbf{R}^d)$  we have  $\|f_1 + \dots + f_d\|_2^2 = \sum_{j=1}^d \|f_j\|_2^2$ , if the functions  $f_j$  are normed such that  $\int f_j(x_j) dx_j = 0$ . Thus (57) holds trivially. Assumption (57) also holds in other  $L_2$ -spaces with dominating measure differing from the Lebesgue measure. A discussion of condition (57) for these classes can be found e.g. Mammen et al. (1999). See also Bickel et al. (1993). Such  $L_2$ -spaces naturally arise in additive regression models. For a white noise model they come up if one assumes an additive model for transformed covariables. We assume that for the models  $\mathcal{F}_j$  one can find  $\delta_j$ -nets  $\mathcal{F}_{j,\delta_j}$  such that choosing  $\delta_j = \psi_{n,j}$  with

$$\psi_{n,j}^2 \asymp n^{-1} \rho^2(Q, \mathcal{F}_{j,\psi_{n,j}}) \log(\#\mathcal{F}_{j,\psi_{n,j}})$$

gives a rate optimal  $\delta$ -net minimizer in the model  $\mathcal{F}_j$ . Now,  $\mathcal{F}_\delta = \mathcal{F}_{1,\delta_1} \oplus \dots \oplus \mathcal{F}_{d,\delta_d}$  is a  $\delta$ -net of  $\mathcal{F}$  with  $\delta = \sum_{j=1}^d \delta_j$ . From Theorem 5 we get that the  $\delta$ -net minimizer  $\hat{f}$  over the net  $\mathcal{F}_\delta$  achieves the rate  $O(\psi_n)$  with  $\psi_n = \max_{1 \leq j \leq d} \psi_{n,j}$ . This is just the type of result we called oracle result at the beginning of this section.

In general, the oracle result does not follow from Theorem 1. The application of Theorem 1 leads to an assumption of the type

$$n^{-1} \max_{1 \leq j \leq d} \rho^2(Q, \mathcal{F}_{j,\psi_{n,j}}) \times \max_{1 \leq j \leq d} \log(\#\mathcal{F}_{j,\psi_{n,j}}) = O(\psi_n^2)$$

whereas Theorem 5 only requires that

$$n^{-1} \max_{1 \leq j \leq d} \left[ \rho^2(Q, \mathcal{F}_{j,\psi_{n,j}}) \log(\#\mathcal{F}_{j,\psi_{n,j}}) \right] = O(\psi_n^2).$$

This can make a big difference. First of all the entropy numbers of the additive classes  $\mathcal{F}_j$  may differ. Furthermore, the operator  $Q$  may act quite differently on the spaces  $\mathcal{F}_j$ .

6.2.3. *Ellipsoidal spaces and convolution.* As an example we now assume that the underlying function space is  $\mathcal{F} = \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_d$ , where

$$\mathcal{F}_k = \left\{ \sum_{j=0}^{\infty} \theta_{kj} \phi_{kj} : \theta_{k\cdot} \in \Theta_{s_k, L_k} \right\}$$

for basis functions  $\phi_{kj} : [0, 1] \rightarrow \mathbf{R}$  and the ellipsoids are defined by

$$(58) \quad \Theta_{s_k, L_k} = \left\{ \theta_{k\cdot} : \sum_{j=0}^{\infty} a_{kj}^2 \theta_{kj}^2 \leq L_k^2 \right\}, \quad k = 1, \dots, d,$$

where we assume that there exists positive constants  $C_1, C_2$  such that for all  $j \in \{0, 1, \dots\}$

$$(59) \quad C_1 \cdot j^{s_k} \leq a_{kj} \leq C_2 \cdot j^{s_k}.$$

Let  $A$  be a convolution operator:  $Af = a * f$  where  $a : \mathbf{R}^d \rightarrow \mathbf{R}$  is a known function. Then

$$Af = A_1 f_1 + \cdots + A_d f_d,$$

where  $f(x) = f_1(x_1) + \cdots + f_d(x_d)$  and

$$A_k f_k(x_k) = \int_{[0,1]^d} f_k(x_k - y_k) a_k(y_k) dy_k,$$

where

$$a_k(y_k) = \int_{[0,1]^d} a(y) \prod_{l=1, l \neq k}^d dy_l$$

is the  $k$ th marginal function of  $a$ . We can decompose  $Q$  accordingly:

$$Qg = Q_1 g_1 + \cdots + Q_d g_d.$$

Operators  $A_j$  and  $Q_j$  are restrictions of  $A$  and  $Q$  to  $\mathcal{F}_j$ . We apply the singular value decomposition for  $A_k$ . Denote

$$\phi_{kj}(t) = \sqrt{2} \cos(2\pi jt), \quad t \in [0, 1],$$

where  $j = 1, 2, \dots$  and  $\phi_0(t) = I_{[0,1]}(t)$ . The collection  $(\phi_{kj})$ ,  $j = 0, 1, \dots$ , is a basis for 1-periodic functions on  $L_2([0, 1])$ . When  $a_k$  are 1-periodic functions in  $L_2([0, 1])$ , then we can write

$$a_k(x_k) = \sum_{j=0}^{\infty} b_{kj} \phi_{kj}(x_k).$$

The functions  $\phi_{kj}$  are the singular functions of the operator  $A_k$  and the values  $b_{kj}$  are the corresponding singular values. We give the rate of convergence of the  $\delta$ -net estimator and show that the estimator achieves the optimal rate of convergence.

**COROLLARY 3.** *Let  $\mathcal{F} = \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_d$ . We assume that the coefficients of the ellipsoid satisfy (59). We assume that  $a_k$  are 1-periodic functions in  $L_2([0, 1])$  and that the Fourier coefficients of  $a_k$  satisfy*

$$C_2 j^{-q_k} \leq b_{kj} \leq C_3 j^{-q_k}$$

for some  $q_k \geq 0$ ,  $C_2, C_3 > 0$ . Then, in the white noise model,

$$\limsup_{n \rightarrow \infty} n^a \sup_{f \in \mathcal{F}} E_f \|\hat{f} - f\|_2^2 < \infty,$$

where  $\hat{f}$  is the  $\delta$ -net estimator and

$$a = \min_{k=1, \dots, d} \frac{2s_k}{2s_k + 2q_k + 1}.$$

Also,

$$\liminf_{n \rightarrow \infty} n^a \inf_{\hat{g}} \sup_{f \in \mathcal{F}} E_f \|\hat{g} - f\|_2^2 > 0,$$

where the infimum is taken over any estimators  $\hat{g}$  in the white noise model.

*Proof.* For the upper bound we apply Theorem 5. As in Section 6.1.1 we can find  $\delta$ -nets  $\mathcal{F}_{k,\delta}$  for  $\mathcal{F}_k$  whose cardinality is bounded by  $\log(\#\mathcal{F}_{k,\delta}) \leq C\delta^{-1/s_k}$  and  $\varrho(Q_k, \mathcal{F}_{k,\delta}) \leq C\delta^{-q_k/s_k}$ . The upper bound of Theorem 5 gives as the rate the maximum of the component rates  $n^{-2s_k/(2s_k+2q_k+1)}$ . For the lower bound we apply the lower bound of Corollary 1 in the case  $d = 1$  and the fact that one cannot do better in the additive model than in the model that has only one component.  $\square$

## 7. Proofs.

7.1. *A preliminary lemma.* We prove that the theoretical error of a minimization estimator may be bounded by the optimal theoretical error and an additional stochastic term.

**LEMMA 1.** *Let  $\mathcal{C} \subset L_2(\mathbf{R}^d)$ . Let  $\hat{f} \in \mathcal{C}$  be such that*

$$(60) \quad \gamma_n(\hat{f}) \leq \inf_{g \in \mathcal{C}} \gamma_n(g) + \varepsilon,$$

where  $\varepsilon \geq 0$ . Then for each  $f^0 \in \mathcal{C}$ ,

$$\|\hat{f} - f\|_2^2 \leq \|f^0 - f\|_2^2 + \varepsilon + 2\nu_n[Q(\hat{f} - f^0)]$$

where  $f$  is the true density or the true signal function, and  $\nu_n(g)$  is the centered empirical operator:

$$(61) \quad \nu_n(g) = \begin{cases} \int g dY_n - \int_{\mathbf{Y}} g(Af), & \text{white noise model,} \\ n^{-1} \sum_{i=1}^n g(Y_i) - \int_{\mathbf{Y}} g(Af), & \text{density estimation,} \end{cases}$$

where  $g : \mathbf{R}^d \rightarrow \mathbf{R}$ .

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

$$\begin{aligned} & \|g - f\|_2^2 - \gamma_n(g) \\ &= \begin{cases} \|f\|_2^2 - 2 \int_{\mathbf{R}^d} fg + 2 \int (Qg) dY_n, & \text{white noise model} \\ \|f\|_2^2 - 2 \int_{\mathbf{R}^d} fg + 2n^{-1} \sum_{i=1}^n (Qg)(Y_i), & \text{density estimation.} \end{cases} \end{aligned}$$

We have  $\int_{\mathbf{R}^d} fg = \int_{\mathbf{Y}} (Af)(Qg)$ . Thus,

$$(62) \quad \|\hat{f} - f\|_2^2 - \gamma_n(\hat{f}) + \gamma_n(f^0) - \|f^0 - f\|_2^2 = 2\nu_n[Q(\hat{f} - f^0)].$$

Thus,

$$\begin{aligned} & \|\hat{f} - f\|_2^2 - \|f^0 - f\|_2^2 \\ &= \|\hat{f} - f\|_2^2 - \gamma_n(\hat{f}) + \gamma_n(\hat{f}) - \|f^0 - f\|_2^2 \\ (63) \quad & \leq \|\hat{f} - f\|_2^2 - \gamma_n(\hat{f}) + \gamma_n(f^0) + \varepsilon - \|f^0 - f\|_2^2 \\ (64) \quad &= 2\nu_n[Q(\hat{f} - f^0)] + \varepsilon. \end{aligned}$$

In (63) we applied (60), and in (64) we applied (62).  $\square$

**7.2. Proof of Theorem 1.** Let  $f \in \mathcal{F}$  be the true density. Let  $\phi^0 \in \mathcal{F}_\delta$ . Denote

$$\zeta = C_1 \|\phi^0 - f\|_2^2 + C_2 n^{-1} \varrho^2(Q, \mathcal{F}_\delta) \log_e(\#\mathcal{F}_\delta),$$

where  $C_1$  is defined in (11) and  $C_2$  is defined in (12). We have that

$$\begin{aligned} & E\|\hat{f} - f\|_2^2 \\ &= \int_0^\infty P(\|\hat{f} - f\|_2^2 > t) dt \\ &\leq \zeta + \int_\zeta^\infty P(\|\hat{f} - f\|_2^2 > t) dt \\ (65) \quad &= \zeta + C_2 n^{-1} \varrho^2(Q, \mathcal{F}_\delta) \int_0^\infty P(\|\hat{f} - f\|_2^2 > C_2 n^{-1} \varrho^2(Q, \mathcal{F}_\delta)t + \zeta) dt. \end{aligned}$$

Denote

$$\tau_n = C_\tau n^{-1} \varrho^2(Q, \mathcal{F}_\delta) (\log_e(\#\mathcal{F}_\delta) + t),$$

where  $C_\tau$  is defined in (13). Then,

$$\begin{aligned} & P\left(\|\hat{f} - f\|_2^2 > C_2 n^{-1} \varrho^2(Q, \mathcal{F}_\delta) t + \zeta\right) \\ &= P\left(\|\hat{f} - f\|_2^2 > C_1 \|\phi^0 - f\|_2^2 + C_2 C_\tau^{-1} \tau_n\right) \\ &= P\left((1 - 2\xi)^{-1} \|\hat{f} - f\|_2^2 \right. \\ &\quad \left. > 2\xi(1 - 2\xi)^{-1} \|\hat{f} - f\|_2^2 + C_1 \|\phi^0 - f\|_2^2 + C_2 C_\tau^{-1} \tau_n\right) \\ (66) \quad &= P\left(\|\hat{f} - f\|_2^2 > 2\xi \|\hat{f} - f\|_2^2 + (1 + 2\xi) \|\phi^0 - f\|_2^2 + \xi \tau_n\right). \end{aligned}$$

We have by Lemma 1,

$$\|\hat{f} - f\|_2^2 \leq \|\phi^0 - f\|_2^2 + 2\nu_n[Q(\hat{f} - \phi^0)].$$

Denote

$$w(\phi) = \|\phi - f\|_2^2 + \|\phi^0 - f\|_2^2 + \tau_n/2.$$

Then we may continue (66) with

$$\begin{aligned} & P\left(\|\hat{f} - f\|_2^2 > C_2 n^{-1} \varrho^2(Q, \mathcal{F}_\delta) t + \zeta\right) \\ &= P\left(\nu_n[Q(\hat{f} - \phi^0)] > \xi \|\hat{f} - f\|_2^2 + \xi \|\phi^0 - f\|_2^2 + \xi \tau_n/2\right) \\ &= P\left(\nu_n[Q(\hat{f} - \phi^0)] > w(\hat{f}) \xi\right) \\ &\leq P\left(\max_{\phi \in \mathcal{F}_\delta, \phi \neq \phi^0} \frac{\nu_n[Q(\phi - \phi^0)]}{w(\phi)} > \xi\right) \\ (67) \quad &\stackrel{\text{def}}{=} P_{\max}. \end{aligned}$$

We prove that

$$(68) \quad P_{\max} \leq \exp(-t),$$

and this proves the theorem, when we combine (65) and (67).

*Proof of (68).* Denote

$$\mathcal{G} = \left\{ \frac{Q(\phi - \phi^0)}{w(\phi)} : \phi \in \mathcal{F}_\delta, \phi \neq \phi^0 \right\}.$$

We have that

$$(69) \quad P_{max} \leq \sum_{g \in \mathcal{G}} P(\nu_n(g) > \xi).$$

Also,

$$w(\phi) \geq \frac{1}{2} \left( \|\phi - \phi^0\|_2^2 + \tau_n \right) \geq \|\phi - \phi^0\|_2 \tau_n^{1/2}$$

and thus

$$(70) \quad v_0 \stackrel{def}{=} \max_{g \in \mathcal{G}} \|g\|_2^2 \leq \frac{1}{\tau_n} \max_{\phi \in \mathcal{F}_\delta, \phi \neq \phi^0} \frac{\|Q(\phi - \phi_0)\|_2^2}{\|\phi - \phi_0\|_2^2} = \frac{\varrho^2(Q, \mathcal{F}_\delta)}{\tau_n}.$$

*Gaussian white noise.* When  $W \sim N(0, \sigma^2)$ , then we have  $P(W > \xi) \leq 2^{-1} \exp\{-\xi^2/(2\sigma^2)\}$  for  $\xi > 0$ , see for example Dudley (1999), Proposition 2.2.1. We have that  $\nu_n(g) \sim N(0, n^{-1}\|g\|_2^2)$ . Thus,

$$P(\nu_n(g) > \xi) \leq 2^{-1} \exp\left\{-\frac{n\xi^2}{2v_0}\right\} \leq 2^{-1} \exp\left\{-\frac{n\tau_n\xi^2}{2\varrho^2(Q, \mathcal{F}_\delta)}\right\}.$$

Thus, denoting  $C_\xi \stackrel{def}{=} \xi^2 C_\tau/2$ ,

$$\begin{aligned} P_{max} &\leq \#\mathcal{F}_\delta \cdot \exp\left\{-\frac{n\tau_n\xi^2}{2\varrho^2(Q, \mathcal{F}_\delta)}\right\} = \#\mathcal{F}_\delta \cdot \exp\{-C_\xi[\log_e(\#\mathcal{F}_\delta) + t]\} \\ &\leq \exp(-t), \end{aligned}$$

since  $C_\xi \geq 1$  by the choice of  $\xi$ .

*Density estimation.* Denote  $v = \sup_{g \in \mathcal{G}} \text{Var}_f(g(Y_1))$ , and  $b = \sup_{g \in \mathcal{G}} \|g\|_\infty$ . We have that

$$(71) \quad v \leq \|Af\|_\infty v_0 \leq B_\infty \frac{\varrho^2(Q, \mathcal{F}_\delta)}{\tau_n},$$

by (70). Also,

$$w(\phi) \geq \frac{\tau_n}{2}$$

and thus, because  $\varrho(Q, \mathcal{F}_\delta) \geq 1$ ,

$$(72) \quad b \leq 2B'_\infty \frac{2}{\tau_n} \leq 4B'_\infty \frac{\varrho^2(Q, \mathcal{F}_\delta)}{\tau_n}.$$

Applying Bernstein's inequality, applying (71) and (72),

$$\begin{aligned} P(\nu_n(g) > \xi) &\leq \exp\left\{\frac{-n\xi^2}{2(v + \xi b/3)}\right\} \\ &\leq \exp\left\{\frac{-n\xi^2\tau_n}{2\rho^2(Q, \mathcal{F}_\delta)(B_\infty + 4B'_\infty\xi/3)}\right\}. \end{aligned}$$

Continuing from (69),

$$\begin{aligned} P_{max} &\leq \#\mathcal{F}_\delta \cdot \exp\left\{\frac{-n\xi^2\tau_n}{2\rho^2(Q, \mathcal{F}_\delta)(B_\infty + 4B'_\infty\xi/3)}\right\} \\ &= \#\mathcal{F}_\delta \cdot \exp\{-C_\xi[\log_e(\#\mathcal{F}_\delta) + t]\} \\ &\leq \exp(-t) \end{aligned}$$

where

$$C_\xi \stackrel{def}{=} \frac{\xi^2 C_\tau}{2(B_\infty + 4B'_\infty\xi/3)},$$

and  $C_\xi \geq 1$  by the choice of  $\xi$ . We have proved (68) and thus the theorem.  $\square$

**7.3. Proof of Theorem 2.** To prove Theorem 2 we follow the approach of Hasminskii & Ibragimov (1990). We start with a useful lemma.

LEMMA 2. *Let  $\mathcal{D} \subset \mathcal{F}$  be a finite set for which*

$$(73) \quad \min\{\|f - g\|_2 : f, g \in \mathcal{D}, f \neq g\} \geq \delta$$

where  $\delta > 0$ . Assume that for some  $f_0 \in \mathcal{D}$ , and for all  $f \in \mathcal{D} \setminus \{f_0\}$ ,

$$(74) \quad P_{Af}^{(n)}\left(\frac{dP_{Af_0}^{(n)}}{dP_{Af}^{(n)}} \leq \tau\right) \leq \alpha,$$

where  $0 < \alpha < 1$ ,  $\tau > 0$ , and in the density estimation model  $P_{Af}^{(n)}$  is the product measure corresponding to density  $Af$ , and in the Gaussian white noise model  $P_{Af}^{(n)}$  is the measure of process  $Y_n$  in (2). Then,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} E_{Af} \|f - \hat{f}\|_2^2 \geq \frac{\delta^2}{4} (1 - \alpha) \frac{\tau(N_\delta - 1)}{1 + \tau(N_\delta - 1)},$$

where  $N_\delta = \#\mathcal{D} \geq 2$  and the infimum is taken over all estimators (either in the density estimation model or in the Gaussian white noise model).

*Proof.* Let  $f_n : \mathbf{R}^d \rightarrow \mathbf{R}$  be an estimator of  $f$ . Define a random variable  $\hat{\theta}$  taking values in  $\mathcal{D}$ :

$$\hat{\theta} = \operatorname{argmin}_{f \in \mathcal{D}} \|f_n - f\|_2.$$

Note that by (73),

$$\hat{\theta} \neq f \in \mathcal{D} \Rightarrow \|f_n - f\|_2 \geq \delta/2,$$

since  $\hat{\theta} \neq f$  for an  $f \in \mathcal{D}$  implies that  $f_n$  is closer to some other  $g \in \mathcal{D}$  than to  $f$ . Then, applying also Markov's inequality,

$$\begin{aligned} \sup_{f \in \mathcal{F}} E_{A_f} \|f_n - f\|_2^2 &\geq \max_{f \in \mathcal{D}} E_{A_f} \|f_n - f\|_2^2 \\ &\geq \frac{\delta^2}{4} \max_{f \in \mathcal{D}} P_{A_f}^{(n)} (\|f_n - f\|_2^2 \geq \delta^2/4) \\ &\geq \frac{\delta^2}{4} \max_{f \in \mathcal{D}} P_{A_f}^{(n)} (\hat{\theta} \neq f). \end{aligned}$$

The lemma follows by an application of Lemma 3 below.  $\square$

LEMMA 3. (*Tsybakov (1998), Theorem 6.*) Let  $\hat{\theta}$  be a random variable taking values on a finite set  $\mathbb{P}$  of probability measures. Denote  $\#\mathbb{P} = N$  and assume  $N \geq 2$ . Let  $\tau > 0$  and  $0 < \alpha < 1$ . Let for some  $P_0 \in \mathbb{P}$  and for all  $P \in \mathbb{P} \setminus \{P_0\}$ ,

$$(75) \quad P \left( \frac{dP_0}{dP} \leq \tau \right) \leq \alpha.$$

Then

$$\max_{P \in \mathbb{P}} P(\hat{\theta} \neq P) \geq (1 - \alpha) \frac{\tau(N - 1)}{1 + \tau(N - 1)}.$$

*Proof of Theorem 2.* For  $f, f_0 \in \mathcal{D}_{\psi_n}$ ,  $f \neq f_0$ ,

$$\begin{aligned} &P_{A_f}^{(n)} \left( \frac{dP_{A_{f_0}}^{(n)}}{dP_{A_f}^{(n)}} \leq \tau \right) \\ (76) \quad &\leq (\log \tau^{-1})^{-1} D_K^2(P_{A_f}^{(n)}, P_{A_{f_0}}^{(n)}) \\ (77) \quad &= \begin{cases} (\log \tau^{-1})^{-1} n D_K^2(Af, Af_0), & \text{density estimation} \\ (\log \tau^{-1})^{-1} \frac{n}{2} \|Af - Af_0\|_2^2, & \text{Gaussian white noise,} \end{cases} \end{aligned}$$

where in (76) we applied Markov's inequality, and in (77) we applied for the Gaussian white noise model the fact that under  $P_{A_f}^{(n)}$ ,

$$\frac{dP_{A_f}^{(n)}}{dP_{A_{f_0}}^{(n)}} = \exp \left\{ n^{1/2} \sigma Z + n\sigma^2/2 \right\},$$

where  $Z \sim N(0, 1)$  and  $\sigma = \|Af - Af_0\|_2$ . When we choose

$$\tau = \tau_n = \exp \left\{ -\alpha^{-1} n [C_1 \varrho_K(A, \mathcal{D}_{\psi_n}) \psi_n]^2 \right\},$$

for  $0 < \alpha < 1$ , then applying assumption (18),

$$\begin{aligned} P_{Af}^{(n)} \left( \frac{dP_{Af_0}^{(n)}}{dP_{Af}^{(n)}} \leq \tau \right) &\leq (\log \tau^{-1})^{-1} n \varrho_K^2(A, \mathcal{D}_{\psi_n}) \|f - f_0\|_2^2 \\ &\leq (\log \tau^{-1})^{-1} n [\varrho_K(A, \mathcal{D}_{\psi_n}) C_1 \psi_n]^2 \\ (78) \qquad \qquad \qquad &= \alpha. \end{aligned}$$

Applying Lemma 2, assumption (17), and (78) we get the lower bound

$$(79) \quad \inf_{\hat{f}} \sup_{f \in \mathcal{D}_{\psi_n}} \|f - \hat{f}\|_2^2 \geq \frac{(C_0 \psi_n)^2}{4} (1 - \alpha) \frac{\tau_n (N_{\psi_n} - 1)}{1 + \tau_n (N_{\psi_n} - 1)},$$

where  $N_{\psi_n} = \#\mathcal{D}_{\psi_n}$ . Let  $n$  be so large that  $\log_e N_{\psi_n} \geq C_2^2 n \varrho_K^2(A, \mathcal{D}_{\psi_n}) \psi_n^2$ , where  $C_2 > C_1$ . This is possible by (19). Then,

$$\begin{aligned} \tau_n N_{\psi_n} &= \exp \left\{ \log_e N_{\psi_n} - \alpha^{-1} n [C_1 \varrho_K(A, \mathcal{D}_{\psi_n}) \psi_n]^2 \right\} \\ &\geq \exp \left\{ n \varrho_K^2(A, \mathcal{D}_{\psi_n}) \psi_n^2 [C_2^2 - \alpha^{-1} C_1^2] \right\} \rightarrow \infty \end{aligned}$$

as  $n \rightarrow \infty$ , where we apply (20) and choose  $\alpha$  so that  $C_2^2 - \alpha^{-1} C_1^2 > 0$ , that is,  $(C_1/C_2)^2 < \alpha < 1$ . Then

$$\lim_{n \rightarrow \infty} \frac{\tau_n (N_{\psi_n} - 1)}{1 + \tau_n (N_{\psi_n} - 1)} = 1$$

and the theorem follows from (79).  $\square$

7.4. *Proof of Theorem 3.* Denote

$$\zeta = C_1 \epsilon + C_2 \psi_n^2,$$

where  $C_1 = (1 - 2\xi)^{-1}$ ,  $C_2 = 1 - 2\xi$ ,  $0 < \xi \leq (3 - \sqrt{5})/4$ . We have that

$$\begin{aligned} E \|\hat{f} - f\|_2^2 &= \int_0^\infty P(\|\hat{f} - f\|_2^2 > t) dt \\ &\leq \zeta + \int_\zeta^\infty P(\|\hat{f} - f\|_2^2 > t) dt \\ (80) \qquad \qquad &= \zeta + C_2 \psi_n^2 \int_0^\infty P(\|\hat{f} - f\|_2^2 > C_2 \psi_n^2 t + \zeta) dt. \end{aligned}$$

Denote

$$\tau_n = C_\tau \psi_n^2 (1+t), \quad C_\tau = \xi^{-1} (1-2\xi)^2.$$

Then,

$$\begin{aligned} & P\left(\|\hat{f} - f\|_2^2 > C_2 \psi_n^2 t + \zeta\right) \\ &= P\left(\|\hat{f} - f\|_2^2 > C_2 C_\tau^{-1} \tau_n + C_1 \epsilon\right) \\ &= P\left((1-2\xi)^{-1} \|\hat{f} - f\|_2^2 \right. \\ &\quad \left. > 2\xi(1-2\xi)^{-1} \|\hat{f} - f\|_2^2 + C_2 C_\tau^{-1} \tau_n + C_1 \epsilon\right) \\ (81) \quad &= P\left(\|\hat{f} - f\|_2^2 > 2\xi \|\hat{f} - f\|_2^2 + \xi \tau_n + \epsilon\right). \end{aligned}$$

We have by Lemma 1, choosing  $f^0 = f$ ,

$$\|\hat{f} - f\|_2^2 \leq 2\nu_n [Q(\hat{f} - f)] + \epsilon.$$

Denote

$$w(g) = \|g - f\|_2^2 + \tau_n/2.$$

Then we may continue (81) with

$$\begin{aligned} & P\left(\|\hat{f} - f\|_2^2 > C_2 \psi_n^2 t + \zeta\right) \\ &\leq P\left(\nu_n [Q(\hat{f} - f)] > \xi \|\hat{f} - f\|_2^2 + \xi \tau_n/2\right) \\ &= P\left(\nu_n [Q(\hat{f} - f)] > w(\hat{f}) \xi\right) \\ &\leq P\left(\sup_{g \in \mathcal{F}} \frac{\nu_n [Q(g - f)]}{w(g)} > \xi\right) \\ (82) \quad &\stackrel{def}{=} P_{sup}. \end{aligned}$$

We prove that

$$(83) \quad P_{sup} \leq \exp(-t \cdot \log_e 2),$$

and this proves the theorem, when we combine (80) and (82).

*Proof of (83).* We use the peeling device, see for example van de Geer (2000), page 69. Denote

$$a_0 = \tau_n/2, \quad a_j = 2^{2j} a_0, \quad b_j = 2^2 a_j, \quad j = 0, 1, \dots$$

Let  $\mathcal{G}_j$  be the set of functions

$$\mathcal{G}_j = \{g \in \mathcal{F} : a_j \leq w(g) < b_j\}, \quad j = 0, 1, \dots$$

and

$$\mathcal{F}_j = \{g \in \mathcal{F} : \|g - f\|_2^2 < b_j\}, \quad j = 0, 1, \dots$$

We have that

$$\mathcal{F} = \{g \in \mathcal{F} : w(g) \geq a_0\} = \bigcup_{j=0}^{\infty} \mathcal{G}_j.$$

Thus,

$$\begin{aligned} P_{sup} &\leq \sum_{j=0}^{\infty} P \left( \sup_{g \in \mathcal{G}_j} \frac{\nu_n[Q(g-f)]}{w(g)} > \xi \right) \\ &\leq \sum_{j=0}^{\infty} P \left( \sup_{g \in \mathcal{F}, w(g) < b_j} \nu_n[Q(g-f)] > \xi a_j \right) \\ (84) \quad &\leq \sum_{j=0}^{\infty} P \left( \sup_{g \in \mathcal{F}_j} \nu_n[Q(g-f)] > \xi a_j \right). \end{aligned}$$

By Assumption 4 of Theorem 3,  $\tilde{G}(\psi_n) = 24\sqrt{2}G(\psi_n)$ , where  $\tilde{G}$  is defined in (95), for sufficiently large  $n$ . Thus, by the choice of  $C = \xi^{-1}4 \cdot 24\sqrt{2}$  in (25),

$$\psi_n^2 \geq n^{-1/2}\xi^{-1}4\tilde{G}(\psi_n).$$

By the choice of  $\xi$  we have that  $C_\tau \geq 2$ , and thus  $a_0 = C_\tau\psi_n^2(1+t)/2 \geq \psi_n^2$ . Since  $G(\delta)/\delta^2$  is decreasing, by Assumption 2 of Theorem 3, then  $\tilde{G}(\delta)/\delta^2$  is decreasing, and

$$\xi n^{1/2}/4 \geq \tilde{G}(\psi_n)/\psi_n^2 \geq \tilde{G}(a_0^{1/2})/a_0 \geq \tilde{G}(b_j^{1/2})/b_j,$$

that is,

$$(85) \quad \xi a_j = \xi b_j/4 \geq n^{-1/2}\tilde{G}(b_j^{1/2}).$$

We may apply Lemma 4 given in Appendix B.1, with (85) to get

$$\begin{aligned} &P \left( \sup_{g \in \mathcal{F}_j} \nu_n[Q(g-f)] > \xi a_j \right) \\ (86) \quad &\leq \exp \left\{ -\frac{n(\xi a_j)^2 C'}{c^2 b_j^{1-a}} \right\} \\ &\leq \exp \left\{ -C'' 2^{2j(a+1)} n \psi_n^{2(1+a)} (1+t)^{1+a} \right\} \end{aligned}$$

$$(87) \quad \leq \exp \left\{ -C''(j+1)n\psi_n^{2(1+a)}(1+t)^{1+a} \right\},$$

where  $C'' = C'c^{-2}\xi^22^{2(a-1)}(C_\tau/2)^{1+a}$ , and we used the facts  $a_j^2/b_j^{1-a} = 2^{2(a-1)}a_j^{1+a} = 2^{2(a-1)}(2^{2j}a_0)^{1+a} = 2^{2(a-1)}[2^{2j}C_\tau\psi_n^2(1+t)/2]^{1+a}$  and  $2^{2j(a+1)} \geq j+1$ . When  $0 \leq b \leq 1/2$ , then  $\sum_{j=0}^{\infty} b^{j+1} = \sum_{j=1}^{\infty} b^j = b/(1-b) \leq 2b$ . When  $n\psi_n^{2(1+a)} \geq (\log_e 2)/C''$ , then  $\exp\{-C''n\psi_n^{2(1+a)}(1+t)^{1+a}\} \leq 1/2$ , and we combine (84) and (87) to get the upper bound

$$\begin{aligned} 2 \exp\left\{-C''n\psi_n^{2(1+a)}(1+t)^{1+a}\right\} &\leq 2 \exp\left\{-C''n\psi_n^{2(1+a)}(1+t)\right\} \\ &\leq \exp\{-t \log_e 2\}. \end{aligned}$$

We have proved (83) and thus we have proved Theorem 3 up to proving Lemma 4, which is done in Appendix B.1.

7.5. *Proof of Theorem 4.* The proof goes similarly as the Proof of Theorem 3 until step (86). At this step we apply Lemma 5, given in Appendix B.2, to get

$$\begin{aligned} &P\left(\sup_{g \in \mathcal{F}_j} \nu_n[Q(g-f)] > \xi a_j\right) \\ &\leq \exp\left\{-\frac{n(\xi a_j)^2 C'}{c^2 b_j^{1-a}}\right\} + 2\#\mathcal{G}_{B_2} \exp\left\{-\frac{1}{12} \frac{n(\xi a_j)^2}{B_\infty c^2 b_j^{1-a} + 2\xi a_j B'_\infty/9}\right\}. \end{aligned}$$

The first term in the right hand side is handled similarly as in the Proof of Theorem 3. For the second term in the right hand side we have, for sufficiently large  $n$ ,

$$\begin{aligned} \exp\left\{-\frac{1}{12} \frac{n(\xi a_j)^2}{B_\infty c^2 b_j^{1-a} + 2\xi a_j B'_\infty/9}\right\} &= \exp\left\{-\frac{1}{12} \frac{n\xi^2 a_j}{B_\infty c^2 a_0^{-a} + 2\xi B'_\infty/9}\right\} \\ &\leq \exp\left\{-\frac{1}{12} \frac{n\xi^2 a_j a_0^a}{B_\infty c^2 + 2\xi B'_\infty/9}\right\} \\ &= \exp\left\{-n\psi_n^{2(1+a)}2^{2j}(1+t)^{1+a}C''\right\}, \end{aligned}$$

since  $a_j^{-a} = (2^{2j}a_0)^{-a} \leq a_0^{-a}$  and  $a_0^{-a} \geq 1$  for sufficiently large  $n$ , and we denote  $C' = \xi^2 C_\tau^{1+a}/[2^{1+a}12(B_\infty c^2 + 2\xi B'_\infty/9)]$ . The proof is finished similarly as the proof of Theorem 3.

7.6. *Proof of Theorem 5.* We proceed similarly as in the proof of Theorem 1. Choose  $f_\delta \in \mathcal{F}_\delta$  such that  $\|f - f_\delta\|_2 \leq \delta$ , where  $f$  is the underlying function in  $\mathcal{F}$ . Choose  $\xi < 1/2$  and put  $\zeta = \zeta_1 + \zeta_2$  with  $\zeta_1 =$

$(1-2\xi)^{-1}(1+2\xi)\|f-f_\delta\|_2^2$ ,  $\zeta_2 = \kappa n^{-1} \sum_{j=1}^p \rho_j^2 \lambda_j$  and  $\kappa = 4c^{-1}\xi^{-1}(1-2\xi)^{-1}$ . We have that

$$(88) \quad \begin{aligned} E\left(\|\hat{f}-f\|_2^2\right) &\leq \zeta + \int_\zeta^\infty P\left(\|\hat{f}-f\|_2^2 > t\right) dt \\ &\leq \zeta + \int_0^\infty P\left(\|\hat{f}-f\|_2^2 > t+\zeta\right) dt. \end{aligned}$$

For the integrand of the second term we have that

$$\begin{aligned} &P\left(\|\hat{f}-f\|_2^2 > t+\zeta\right) \\ &= P\left((1-2\xi)^{-1}\|\hat{f}-f\|_2^2 > 2\xi(1-2\xi)^{-1}\|\hat{f}-f\|_2^2 + t+\zeta\right) \\ &= P\left(\|\hat{f}-f\|_2^2 > 2\xi\|\hat{f}-f\|_2^2 + (1-2\xi)t + (1-2\xi)\zeta\right). \end{aligned}$$

We now use Lemma 1. This gives

$$\|\hat{f}-f\|_2^2 \leq \|f-f_\delta\|_2^2 + 2\nu_n\left(Q(\hat{f}-f_\delta)\right).$$

Together with the last equalities this gives

$$\begin{aligned} &P\left(\|\hat{f}-f\|_2^2 > t+\zeta\right) \\ &\leq P\left(\|f-f_\delta\|_2^2 + 2\nu_n\left(Q(\hat{f}-f_\delta)\right) > 2\xi\|\hat{f}-f\|_2^2 + (1-2\xi)(t+\zeta)\right) \\ &= P\left(\nu_n\left(Q(\hat{f}-f_\delta)\right) > \xi\|\hat{f}-f\|_2^2 + \xi\|f-f_\delta\|_2^2 + 2^{-1}(1-2\xi)(t+\zeta_2)\right) \\ &\leq P\left(\nu_n\left(Q(\hat{f}-f_\delta)\right) > 2^{-1}\xi\|\hat{f}-f_\delta\|_2^2 + 2^{-1}(1-2\xi)(t+\zeta_2)\right). \end{aligned}$$

Now, put  $w_j = \rho_j / \sum_{l=1}^p \rho_l$  and decompose  $f_\delta = f_{\delta,1} + \dots + f_{\delta,p}$  and  $\hat{f} = \hat{f}_1 + \dots + \hat{f}_p$  with  $f_{\delta,j}, \hat{f}_j \in \mathcal{F}_{j,\delta}$ . Using assumption (57) we get with  $\beta_j = 2^{-1}(1-2\xi)(w_j t + \kappa n^{-1} \rho_j^2 \lambda_j)$ ,

$$\begin{aligned} &P\left(\|\hat{f}-f\|_2^2 > t+\zeta\right) \\ &\leq P\left(\sum_{j=1}^p \nu_n\left(Q(\hat{f}_j-f_{\delta,j})\right) > 2^{-1}\xi c \sum_{j=1}^p \|\hat{f}_j-f_{\delta,j}\|_2^2 + \sum_{j=1}^p \beta_j\right) \\ &\leq \sum_{j=1}^p P\left(\nu_n\left(Q(\hat{f}_j-f_{\delta,j})\right) > 2^{-1}\xi c \|\hat{f}_j-f_{\delta,j}\|_2^2 + \beta_j\right) \\ &\leq \sum_{j=1}^p \sum_{g_j \in \mathcal{F}_{j,\delta}} P\left(\nu_n\left(Q(g_j-f_{\delta,j})\right) > 2^{-1}\xi c \|g_j-f_{\delta,j}\|_2^2 + \beta_j\right). \end{aligned}$$

We now use

$$P(\nu_n(h) > \xi) \leq 2^{-1} \exp\left(-\frac{n\xi^2}{2\|h\|_2^2}\right),$$

compare to the proof of Theorem 1. This gives

$$\begin{aligned} & P\left(\|\hat{f} - f\|_2^2 > t + \zeta\right) \\ & \leq \sum_{j=1}^p \sum_{g_j \in \mathcal{F}_{j,\delta}} 2^{-1} \exp\left[-\frac{n(2^{-1}\xi c\|g_j - f_{\delta,j}\|_2^2 + \beta_j)^2}{2\|Q(g_j - f_{\delta,j})\|^2}\right] \\ & \leq \sum_{j=1}^p \sum_{g_j \in \mathcal{F}_{j,\delta}} 2^{-1} \exp\left[-\frac{n\xi c\|g_j - f_{\delta,j}\|_2^2 \beta_j}{2\|Q(g_j - f_{\delta,j})\|^2}\right] \\ & \leq \sum_{j=1}^p \exp(\lambda_j) 2^{-1} \exp\left[-\frac{n\xi c \beta_j}{2\rho_j^2}\right] \\ & = \sum_{j=1}^p 2^{-1} \exp\left[-n\xi c 4^{-1}(1 - 2\xi)w_j \rho_j^{-2}t\right]. \end{aligned}$$

By plugging this into (88) we get

$$\begin{aligned} E\left(\|\hat{f} - f\|_2^2\right) & \leq \zeta + \sum_{j=1}^p \int_0^\infty \exp\left[-n\xi c 4^{-1}(1 - 2\xi)w_j \rho_j^{-2}t\right] dt \\ & \leq \zeta + \sum_{j=1}^p n^{-1} 4[\xi c(1 - 2\xi)w_j]^{-1} \rho_j^2 \\ & = \zeta + n^{-1} 4[\xi c(1 - 2\xi)]^{-1} \left(\sum_{j=1}^p \rho_j\right)^2. \end{aligned}$$

Choosing  $\xi = 4^{-1}$  gives the statement of Theorem 5.

**Acknowledgment.** We would like to thank referees for suggesting improvements and pointing out errors. Writing of this article was financed by Deutsche Forschungsgemeinschaft under project MA1026/8-1.

## References.

- Barron, A. & Yang, Y. (1999), ‘Information-theoretic determination of minimax rates of convergence’, *Ann. Statist.* **27**(5), 1564–1599.
- Bass, R. F. (1985), ‘Law of the iterated logarithm for set-indexed partial sum processes with finite variance’, *Z. Wahrsch. Verw. Gebiete* **65**, 181–237.
- Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, The Johns Hopkins Univ. Press, Baltimore.

- Birgé, L. (1983), ‘Approximation dans les espaces métriques et théorie de l’estimation’, *Z. Wahrsch. Verw. Gebiete* **65**, 181–237.
- Birgé, L. & Massart, P. (1993), ‘Rates of convergence for minimum contrast estimators’, *Probab. Theory Relat. Fields* **97**, 113–150.
- Cencov, N. N. (1972), *Statistical Decision Rules and Optimal Inference*, Nauka, Moscow.
- Comte, F., Taupin, M.-L. & Rozenholc, Y. (2005), ‘Penalized contrast estimator for density deconvolution’. Preprint 2003-2 MAP5, <http://www.math-info.univ-paris5.fr/~comte/publi.html>.
- Deans, S. R. (1983), *The Radon Transform and some of its Applications*, Wiley.
- Donoho, D. L. (1995), ‘Nonlinear solutions of linear inverse problems by wavelet-vaguelette decomposition’, *J. Applied and Comput. Harmonic Anal.* **2**, 101–126.
- Donoho, D. L. & Low, M. (1992), ‘Renormalization exponents and optimal pointwise rates of convergence’, *Ann. Statist.* **20**, 944–970.
- Dudley, R. M. (1999), *Uniform central limit theorems*, Cambridge University Press.
- Dunford, N. & Schwartz, J. T. (1958), *Linear Operators, Part I*, Interscience Publishers, New York.
- Ermakov, M. S. (1989), ‘Minimax estimation of the solution of an ill-posed convolution type problem’, *Problems Inform. Transmission* **25**, 191–200.
- Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd edn, Wiley.
- Hasminskii, R. Z. & Ibragimov, I. A. (1990), ‘On density estimation in the view of Kolmogorov’s ideas in approximation theory’, *Ann. Statist.* **18**, 999–1010.
- Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall, London.
- Ibragimov, I. A. (2004), ‘Estimation of multivariate regression’, *Theory Probab. Appl.* **48**(2), 256–272.
- Ibragimov, I. A. & Hasminskii, R. Z. (1980), ‘On estimate of the density function’, *Zap. Nauchn. Semin. LOMI* **98**, 61–85.
- Ibragimov, I. A. & Hasminskii, R. Z. (1981), ‘On the non-parametric density estimates’, *Zap. Nauchn. Semin. LOMI* **108**, 73–89.
- Johnstone, I. M. & Silverman, B. W. (1990), ‘Speed of estimation in positron emission tomography and related inverse problems’, *Ann. Statist.* **18**, 251–280.
- Kolmogorov, A. N. & Tikhomirov, V. M. (1961), ‘ $\epsilon$ -entropy and  $\epsilon$ -capacity of sets in function spaces’, *Translations of the American Math. Soc.* **17**, 277–364.
- Koo, J. Y. (1993), ‘Optimal rates of convergence for nonparametric statistical inverse problems’, *Ann. Statist.* **21**, 590–599.
- Korostelev, A. P. & Tsybakov, A. B. (1991), ‘Optimal rates of convergence of estimators in a probabilistic setup of tomography problem’, *Problems Inform. Transmission* **27**, 73–81.
- Korostelev, A. P. & Tsybakov, A. B. (1993), *Minimax Theory of Image Reconstruction, Lecture Notes in Statistics, 82*, Springer.
- Le Cam, L. (1973), ‘Convergence of estimates under dimensionality restrictions’, *Ann. Statist.* **1**, 38–53.
- Mammen, E., Linton, O. & Nielsen, J. (1999), ‘The existence and asymptotic properties of a backfitting projection algorithm under weak conditions’, *Ann. Statist.* **27**, 1443–1490.
- Natterer, F. (2001), *The Mathematics of Computerized Tomography*, Vol. 32 of *Classics in Applied Mathematics*, SIAM.
- Ossiander, M. (1987), ‘A central limit theorem under metric entropy with  $L_2$  bracketing’, *Ann. Probab.* **15**, 897–919.
- O’Sullivan, F. (1986), ‘A statistical perspective on ill-posed inverse problems’, *Statist.*

- Science* **1**(4), 502–527.
- Stone, C. J. (1985), ‘Additive regression and other nonparametric models’, *Ann. Statist.* **13**, 689–705.
- Tsybakov, A. B. (1998), ‘Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes’, *Ann. Statist.* **26**, 2420–2469.
- van de Geer, S. A. (2000), *Empirical Processes in M-Estimation*, Cambridge University Press.
- van der Laan, M. J., Dudoit, S. & van der Vaart, A. W. (2004), The cross-validated adaptive epsilon-net estimator, Working paper series, U.C. Berkeley Division of Biostatistics.

## APPENDIX A: ELLIPSOIDS

The ellipsoid has been defined in (39) and we assume that the  $a_j$  satisfy (40). We make the calculations now in the one dimensional case.

**A.1.  $\delta$ -net.** We shall construct a  $\delta$ -net  $\Theta_\delta$  for the ellipsoid in (39). The construction is similar to the construction of Kolmogorov & Tikhomirov (1961). Let

$$(89) \quad M = \lceil [(C_1^{-1} 2^{1/2} L \delta^{-1})^{1/s}] \rceil.$$

Let  $\Theta_\delta(M)$  be a  $\delta/2$ -net of

$$E_M = \left\{ (\theta_j)_{j \in \{1, \dots, M\}} : \sum_{j=1}^M a_j^2 \theta_j^2 \leq L^2 \right\}.$$

We can choose  $\Theta_\delta(M)$  in such a way that its cardinality satisfies

$$\#\Theta_\delta(M) \leq C \frac{\text{volume}(E_M)}{\text{volume}(B_\delta^{(M)})},$$

where  $B_\delta^{(M)}$  is a ball of radius  $\delta$  in the  $M$ -dimensional Euclidean space. Define the  $\delta$ -net by

$$\Theta_\delta = \left\{ (\theta_j)_{j \in \{1, \dots, \infty\}} : (\theta_j)_{j \in \{1, \dots, M\}} \in \Theta_\delta(M), \theta_j = 0, \text{ for } j \geq M + 1 \right\}.$$

( *$\delta$ -net property.*) We proof that  $\Theta_\delta$  is a  $\delta$ -net of the ellipsoid  $\Theta$ . For each  $\theta \in \Theta$  there is  $\theta_\delta \in \Theta_\delta$  such that  $\|\theta - \theta_\delta\|_{l_2} \leq \delta$ . Indeed, let  $\theta \in \Theta$ . Let  $\theta_\delta \in \Theta_\delta$  be such that

$$\sum_{j=1}^M (\theta_j - \theta_{\delta,j})^2 \leq \delta^2/2.$$

Then

$$\|\theta - \theta_\delta\|_{l_2}^2 = \sum_{j=1}^M (\theta_j - \theta_{\delta,j})^2 + \sum_{j=M+1}^{\infty} \theta_j^2 \leq \delta^2$$

where we used the fact

$$(90) \quad \sum_{j=M+1}^{\infty} \theta_j^2 \leq C_1^{-2} \cdot M^{-2s} \sum_{j=M+1}^{\infty} a_j^2 \theta_j^2 \leq C_1^{-2} M^{-2s} L^2 \leq \delta^2/2,$$

because, when  $j \notin \{1, \dots, M\}$ , then

$$a_j^{-2} \leq C_1^{-2} \cdot j^{-2s} \leq C_1^{-2} \cdot M^{-2s} \leq \delta^2/(2L^2).$$

(*Cardinality.*) We prove that

$$\log(\#\Theta_\delta) \leq C\delta^{-1/s}.$$

We have that

$$\text{volume}(E^{(M)}) = C_M \cdot L^M \prod_{j=1}^M a_j^{-1}$$

and

$$\text{volume}(B_\delta^{(M)}) = C_M \cdot \delta^M,$$

where  $C_M$  is the volume of the unit ball in the  $M$  dimensional Euclidean space. Thus the cardinality of  $\Theta_\delta$  satisfies

$$\#\Theta_\delta = \#\Theta_\delta(M) \leq C \frac{L^M \prod_{j=1}^M a_j^{-1}}{\delta^M}.$$

We have that

$$\prod_{j=1}^M a_j^{-1} \leq C \prod_{j=1}^M j^{-s} = C(M!)^{-s}.$$

Applying Feller (1968), pp. 50-53, we get

$$M! > M^{M+1/2} e^{-M}.$$

Thus

$$\begin{aligned} & \log(\#\Theta_\delta) \\ & \leq M \log(L) - s \log(M!) + M \log(\delta^{-1}) + C \\ & \leq M \log(L) - s(M + 1/2) \log M + sM + M \log(\delta^{-1}) + C \\ & \leq M(\log(L) + s) - sM \log M + M \log(\delta^{-1}) + C \\ & \leq M(\log(L) + s + C') + C \\ (91) \quad & \leq \delta^{-1/s} C'' + C, \end{aligned}$$

since  $M = C''' \delta^{-1/s}$ .

**A.2.  $\delta$ -packing set.** For a fixed sequence  $\theta^*$  with  $\sum_{j=0}^{\infty} a_j^2 \theta_j^{*2} = L^* < L$  let  $\Theta_\delta^*(M)$  be a  $\delta$ -packing set of

$$E_M^* = \left\{ (\theta_j)_{j \in \{M^*, \dots, M\}} : \sum_{j=M^*}^M a_j^2 \theta_j^2 \leq (L - L^*)^2 \right\}.$$

Here,  $M^* = \lceil M/2 \rceil$ . We can choose  $\Theta_\delta^*(M)$  in such a way that its cardinality satisfies

$$(92) \quad \log(\#\Theta_\delta^*(M)) \geq C^* \delta^{-1/s}.$$

Define

$$(93) \quad \Theta_\delta^* = \left\{ (\theta_j)_{j \in \{0, \dots, \infty\}} : \begin{aligned} & (\theta_j - \theta_j^*)_{j \in \{M^*, \dots, M\}} \in \Theta_\delta^*(M), \\ & \theta_j = \theta_j^*, \text{ for } j \notin \{M^*, \dots, M\} \end{aligned} \right\}.$$

The bound (92) follows similarly as the upper bound (91). In the white noise case one can use this construction with  $\theta^* = 0$  and  $L^* = 0$ . In the density case another choice of  $\theta^*$  may be appropriate to ensure that the functions in  $\mathcal{D}_\delta$  are bounded from above and from below. This would allow to use the bound (22) to carry over bounds on Hilbert norms to corresponding bounds on Kullback-Leibler distances. Note also that a similar calculation as in (90) shows that for  $\theta, \theta' \in \Theta_\delta^*$ ,

$$(94) \quad \|\theta - \theta'\|_{l_2}^2 = \sum_{i=M^*}^M (\theta_i - \theta'_i)^2 = \sum_{i=M^*}^{\infty} (\theta_i - \theta'_i)^2 \leq C\delta^2.$$

## APPENDIX B: LEMMAS RELATED TO EMPIRICAL PROCESS THEORY

**B.1. Gaussian white noise.** Lemma 4 gives an exponential tail bound for the Gaussian white noise model.

**LEMMA 4.** *Let  $\nu_n$  be the centered empirical operator of a Gaussian white noise process. Operator  $\nu_n$  is defined in (61). Let  $\mathcal{G} \subset L_2(\mathbf{R}^d)$  be such that  $\sup_{g \in \mathcal{G}} \|g\|_2 \leq R$  and denote with  $\mathcal{G}_\delta$  a  $\delta$ -net of  $\mathcal{G}$ ,  $\delta > 0$ . Assume that  $\delta \mapsto \varrho(Q, \mathcal{G}_\delta) \sqrt{\log_e(\#\mathcal{G}_\delta)}$  is decreasing on  $(0, R]$ , where  $\varrho(Q, \mathcal{G}_\delta)$  is defined in (23) and assume that the entropy integral  $G(R)$  defined in (24) is finite. Assume that  $\varrho(Q, \mathcal{G}_\delta) = c\delta^{-a}$ , where  $0 \leq a < 1$  and  $c > 0$ . Then for all*

$$(95) \quad \xi \geq n^{-1/2} \tilde{G}(R), \quad \tilde{G}(R) = \max \left\{ 24\sqrt{2}G(R), cR^{1-a} \sqrt{\log_e 2/C'} \right\}$$

where

$$(96) \quad C' = 12^{-2}(C'')^{-2}, \quad C'' = (1-a)^{-3/2}\Gamma(3/2)(\log_e 2)^{-3/2},$$

we have

$$P\left(\sup_{g \in \mathcal{G}} \nu_n(Qg) \geq \xi\right) \leq \exp\left\{-\frac{n\xi^2 C'}{c^2 R^{2-2a}}\right\}.$$

*Proof.* The proof uses the chaining technique. The chaining technique was developed by Kolmogorov. An analogous lemma in the direct case is for example Lemma 3.2 in van de Geer (2000). The basic difference to the direct case is visible in eq. (98). Let us denote  $R_k = 2^{-k}R$ ,  $N_k = \#\mathcal{G}_{R_k}$  and  $H_k = \log_e N_k$ , where  $k = 0, 1, \dots$ . For each  $g \in \mathcal{G}$ , let  $h_g^k$  be a member of  $R_k$  covering set of  $\mathcal{G}$  such that  $\|g - h_g^k\|_2 \leq R_k$ . We may write every  $g \in \mathcal{G}$  with telescoping as

$$g = \sum_{k=1}^{\infty} (h_g^k - h_g^{k-1})$$

where  $h_g^0 \equiv 0$  and the convergence is in  $L_2$ . Let  $\eta_k > 0$  be such that  $\sum_{k=1}^{\infty} \eta_k \leq 1$ . We will define  $\eta_k$  in (100). Then

$$(97) \quad P\left(\sup_{g \in \mathcal{G}} \nu_n(Qg) \geq \xi\right) \leq \sum_{k=1}^{\infty} P\left(\sup_{g \in \mathcal{G}} \nu_n(Q(h_g^k - h_g^{k-1})) \geq \xi \eta_k\right).$$

We have

$$\#\{h_g^k - h_g^{k-1} : g \in \mathcal{G}\} \leq N_k N_{k-1} \leq N_k^2.$$

We have

$$(98) \quad \begin{aligned} \max\left\{\|Q(h_g^k - h_g^{k-1})\|_2 : g \in \mathcal{G}\right\} &\leq T_k \max\left\{\|h_g^k - h_g^{k-1}\|_2 : g \in \mathcal{G}\right\} \\ &\leq 3T_k R_k, \end{aligned}$$

where we denote  $T_k = \varrho(Q, \mathcal{G}_{R_k})$ , when  $\varrho(Q, \mathcal{G}_\delta)$  is defined in (23), and we used the fact

$$\|h_g^k - h_g^{k-1}\|_2 \leq \|h_g^k - g\|_2 + \|h_g^{k-1} - g\|_2 \leq 2^{-k}R + 2^{-k+1}R = 3R_k.$$

When  $W \sim N(0, \sigma^2)$ ,  $\xi > 0$ , then  $P(W > \xi) \leq 2^{-1} \exp\{-\xi^2/(2\sigma^2)\}$ , see for example Dudley (1999), Proposition 2.2.1. We have that  $\nu_n(Q(h_g^k - h_g^{k-1})) \sim N(0, n^{-1}\|Q(h_g^k - h_g^{k-1})\|_2^2)$  and thus

$$(99) \quad P\left(\sup_{g \in \mathcal{G}} \nu_n(Q(h_g^k - h_g^{k-1})) \geq \xi \eta_k\right) \leq N_k^2 2^{-1} \exp\left\{-\frac{1}{2} \frac{n\xi^2 \eta_k^2}{3^2 T_k^2 R_k^2}\right\}.$$

Now we choose

$$(100) \quad \eta_k = 3T_k R_k \max \left\{ \frac{8^{1/2} H_k^{1/2}}{n^{1/2} \xi}, c^{-1} R^{a-1} (C' k)^{1/2} 2 \right\},$$

where  $C'$  is defined in (96). Then we may apply (99) to continue (97) with an upper bound

$$(101) \quad \frac{1}{2} \sum_{k=1}^{\infty} \exp \left\{ 2H_k - \frac{1}{2} \frac{n\xi^2 \eta_k^2}{3^2 T_k^2 R_k^2} \right\} \leq \frac{1}{2} \sum_{k=1}^{\infty} \exp \left\{ -\frac{1}{4} \frac{n\xi^2 \eta_k^2}{3^2 T_k^2 R_k^2} \right\}$$

$$(102) \quad \leq \frac{1}{2} \sum_{k=1}^{\infty} \exp \left\{ -\frac{n\xi^2 C' k}{c^2 R^{2-2a}} \right\}$$

$$(103) \quad \leq \exp \left\{ -\frac{n\xi^2 C'}{c^2 R^{2-2a}} \right\}.$$

In (101) we applied (100), which implies  $2H_k \leq n\xi^2 \eta_k^2 / (4 \cdot 3^2 T_k^2 R_k^2)$ , when we apply the first term in the maximum. In (102) we applied also (100), which implies  $\eta_k^2 / (4 \cdot 3^2 T_k^2 R_k^2) \geq C' k / [c^2 R^{2-2a}]$  where we applied the second term in the maximum. In (103) we applied that for  $0 \leq b \leq 1/2$ ,  $\sum_{k=1}^{\infty} b^k = b/(1-b) \leq 2b$ . Here we need that  $\exp \{-n\xi^2 C' / [c^2 R^{2-2a}]\} \leq 1/2$ , that is,  $\xi \geq cR^{1-a} \left( \frac{\log_e 2}{n C'} \right)^{1/2}$  which is implied by (95). We need to check that  $\sum_{k=1}^{\infty} \eta_k \leq 1$ . Since  $\delta \mapsto \varrho(Q, \mathcal{G}_\delta) \sqrt{\log_e (\#\mathcal{G}_\delta)}$  is decreasing,

$$(104) \quad \sum_{k=1}^{\infty} T_k R_k H_k^{1/2} = 2 \sum_{k=1}^{\infty} 2^{-k-1} R T_{2^{-k} R} \sqrt{\log_e (\#\mathcal{G}_{2^{-k} R})} \leq 2G(R).$$

We apply the assumption that  $T_k = T_{2^{-k} R} = cR^{-a} 2^{ak}$  to get

$$(105) \quad \begin{aligned} \sum_{k=1}^{\infty} k^{1/2} T_k R_k &= cR^{1-a} \sum_{k=1}^{\infty} k^{1/2} 2^{-(1-a)k} \\ &= cR^{1-a} \lim_{K \rightarrow \infty} K^{3/2} \int_0^1 t^{1/2} 2^{-(1-a)Kt} dt \\ &= cR^{1-a} (1-a)^{-3/2} \int_0^{\infty} u^{1/2} 2^{-u} du \end{aligned}$$

$$(106) \quad = cR^{1-a} C'',$$

where  $C''$  is defined in (96). We have from (104) and (106) that

$$\sum_{k=1}^{\infty} \eta_k \leq \frac{8^{1/2} 6G(R)}{n^{1/2} \xi} + 6\sqrt{C'} C'' \leq \frac{1}{2} + \frac{1}{2} = 1,$$

when  $\xi \geq 28^{1/2} 6G(R) n^{-1/2}$ , which is guaranteed by (95), and  $C'$  is chosen as in (96). The lemma follows from (97), (99), and (103).  $\square$

**B.2. Density estimation.** Lemma 5 gives an exponential bound for the tail probability in the case of density estimation.

LEMMA 5. *Let  $Y_1, \dots, Y_n \in \mathbf{R}^d$  be i.i.d. with density  $Af$ , and let the centered empirical process  $\nu_n$  be defined in (61). Assume that  $\|Af\|_\infty \leq B_\infty$ . Let  $\mathcal{G} \subset L_2(\mathbf{R}^d)$  be such that  $\sup_{g \in \mathcal{G}} \|g\|_2 \leq R$ . Denote with  $\mathcal{G}_\delta$  a  $\delta$ -bracketing net of  $\mathcal{G}$ ,  $\delta > 0$ . Denote with  $\mathcal{G}_\delta^L = \{g^L : (g^L, g^U) \in \mathcal{G}_\delta\}$  and  $\mathcal{G}_\delta^U = \{g^U : (g^L, g^U) \in \mathcal{G}_\delta\}$ . Assume that  $\sup_{g \in \mathcal{G}_R^L \cup \mathcal{G}_R^U} \|Qg\|_\infty \leq B'_\infty$ . Assume that  $\delta \mapsto \varrho_{den}(Q, \mathcal{G}_\delta) \sqrt{\log_e(\#\mathcal{G}_\delta)}$  is decreasing on  $(0, R]$ , where  $\varrho_{den}(Q, \mathcal{G}_\delta)$  is defined in (29) and assume that the entropy integral  $G(R)$  defined in (30) is finite. Assume that  $\varrho_{den}(Q, \mathcal{G}_\delta) = c\delta^{-a}$ , where  $0 \leq a < 1$  and  $c > 0$ . Then for all*

$$(107) \quad \xi \geq n^{-1/2} \tilde{G}(R),$$

where

$$(108) \quad \begin{aligned} \tilde{G}(R) &= B_\infty^{1/2} (9^2 + 96 \cdot 2^{-2a})^{1/2} \\ &\times \max \left\{ 24\sqrt{2}G(R), 4(\log_e(2))^{-1} (1-a)^{-3/2} \Gamma(3/2) c R^{1-a} \right\}, \end{aligned}$$

we have

$$\begin{aligned} &P \left( \sup_{g \in \mathcal{G}} \nu_n(Qg) \geq \xi \right) \\ &\leq 4 \exp \left\{ -\frac{n\xi^2 C'}{B_\infty c^2 R^{2-2a}} \right\} + 2\#\mathcal{G}_R \exp \left\{ -\frac{1}{12} \frac{n\xi^2}{B_\infty c^2 R^{2(1-a)} + 2\xi B'_\infty / 9} \right\}, \end{aligned}$$

where  $\nu_n$  is the centered empirical process defined in (61).

*Proof.* We use the chaining technique with truncation. The basic difference to the direct case is visible in (115) and (122). The technique was used in the direct case by Bass (1985), Ossiander (1987), Birgé & Massart (1993), Proposition 3, van de Geer (2000), Theorem 8.13. Let us denote  $R_k = 2^{-k}R$ ,  $N_k = \#\mathcal{G}_{R_k}$  and  $H_k = \log_e N_k$ , for  $k = 0, 1, \dots$ . Let us denote  $T_k = \varrho_{den}(Q, \mathcal{G}_{R_k})$ , where  $\varrho_{den}(Q, \mathcal{G}_\delta)$  is defined in (29). For each  $g \in \mathcal{G}$ , let  $(h_g^{k,L}, h_g^{k,U})$  be the member of the bracketing net  $\mathcal{G}_{R_k}$ , such that  $h_g^{k,L} \leq g \leq h_g^{k,U}$ . We may write every  $g \in \mathcal{G}$  with telescoping as

$$g = g - h_g^{\kappa_g, L} + \sum_{k=1}^{\kappa_g} (h_g^{k,L} - h_g^{k-1,L}) + h_g^{0,L},$$

where

$$\kappa_g = \begin{cases} \min \{0 \leq k \leq K-1 : Q\Delta_g^k \geq \beta_k\}, & \text{if } Q\Delta_g^k \geq \beta_k \text{ for some } 0 \leq k \leq K-1 \\ K, & \text{otherwise,} \end{cases}$$

where  $K \geq 1$  is defined in (123),

$$\Delta_g^k = h_g^{k,U} - h_g^{k,L},$$

and

$$(109) \quad \beta_k = \frac{12B_\infty T_k^2 R_k^2}{\xi}.$$

Then,

$$(110) \quad \begin{aligned} P \left( \sup_{g \in \mathcal{G}} \nu_n(Qg) \geq \xi \right) &\leq P \left( \sup_{g \in \mathcal{G}} \sum_{k=1}^{\kappa_g} \nu_n \left( Q(h_g^{k,L} - h_g^{k-1,L}) \right) \geq \xi/3 \right) \\ &\quad + P \left( \sup_{g \in \mathcal{G}} \nu_n \left( Q(g - h_g^{\kappa_g,L}) \right) \geq \xi/3 \right) \\ &\quad + P \left( \sup_{g \in \mathcal{G}} \nu_n \left( Qh_g^{0,L} \right) \geq \xi/3 \right) \\ &\stackrel{\text{def}}{=} P_I + P_{II} + P_{III}. \end{aligned}$$

*Term  $P_I$ .* We have

$$\begin{aligned} \sup_{g \in \mathcal{G}} \sum_{k=1}^{\kappa_g} \nu_n \left( Q(h_g^{k,L} - h_g^{k-1,L}) \right) &= \sup_{g \in \mathcal{G}} \sum_{k=1}^K I_{\{1, \dots, \kappa_g\}}(k) \nu_n \left( Q(h_g^{k,L} - h_g^{k-1,L}) \right) \\ &\leq \sum_{k=1}^K \sup_{g \in \mathcal{G}} I_{\{1, \dots, \kappa_g\}}(k) \nu_n \left( Q(h_g^{k,L} - h_g^{k-1,L}) \right). \end{aligned}$$

Let us denote

$$(111) \quad \eta_k = (9^2 + 96 \cdot 2^{-2a})^{1/2} T_k R_k \max \left\{ \frac{8^{1/2} B_\infty^{1/2} H_k^{1/2}}{n^{1/2} \xi}, c^{-1} R^{a-1} (C'k)^{1/2} 2 \right\},$$

where  $C'$  is defined by

$$(112) \quad C' = 4^{-2} (C'')^{-2} (9^2 + 96 \cdot 2^{-2a})^{-1}, \quad C'' = (1-a)^{-3/2} \Gamma(3/2) (\log_e 2)^{-3/2}.$$

We have defined  $\eta_k$  in (111) so that  $\eta_k > 0$  and  $\sum_{k=1}^{\infty} \eta_k \leq 1$ , which is proved in (134). Then,

$$(113) \quad P_I \leq \sum_{k=1}^K P \left( \sup_{g \in \mathcal{G}} I_{\{1, \dots, \kappa_g\}}(k) \nu_n \left( Q(h_g^{k,L} - h_g^{k-1,L}) \right) \geq \eta_k \xi / 3 \right).$$

We have

$$(114) \quad \# \{h_g^{k,L} - h_g^{k-1,L} : g \in \mathcal{G}\} \leq N_k N_{k-1} \leq N_k^2.$$

Also,

$$(115) \quad \begin{aligned} & \max \left\{ E \left| Q(h_g^{k,L} - h_g^{k-1,L}) \right|^2 : g \in \mathcal{G} \right\} \\ & \leq B_\infty \max \left\{ \left\| Q(h_g^{k,L} - h_g^{k-1,L}) \right\|_2^2 : g \in \mathcal{G} \right\} \\ & \leq B_\infty T_k^2 \max \left\{ \left\| h_g^{k,L} - h_g^{k-1,L} \right\|_2^2 : g \in \mathcal{G} \right\} \\ & \leq B_\infty 3^2 T_k^2 R_k^2, \end{aligned}$$

because

$$\left\| h_g^{k,L} - h_g^{k-1,L} \right\|_2 \leq \left\| h_g^{k,L} - g \right\|_2 + \left\| h_g^{k-1,L} - g \right\|_2 \leq 2^{-k} R + 2^{-k+1} R = 3R_k.$$

When  $k \leq \kappa_g$ , then

$$Q(h_g^{k,L} - h_g^{k-1,L}) \leq Q \Delta_g^{k-1} \leq \beta_{k-1},$$

which implies

$$(116) \quad \left| Q(h_g^{k,L} - h_g^{k-1,L}) - EQ(h_g^{k,L} - h_g^{k-1,L}) \right| \leq 2\beta_{k-1}.$$

Thus, applying (114), (115), (116), by Bernstein's inequality,

$$(117) \quad \begin{aligned} & P \left( \sup_{g \in \mathcal{G}} I_{\{1, \dots, \kappa_g\}}(k) \nu_n \left( Q(h_g^{k,L} - h_g^{k-1,L}) \right) \geq \xi \eta_k / 3 \right) \\ & \leq N_k^2 \exp \left\{ -\frac{1}{2} \frac{n(\xi \eta_k / 3)^2}{3^2 B_\infty T_k^2 R_k^2 + 2\beta_{k-1} \xi \eta_k / 9} \right\} \\ & \leq \exp \left\{ 2H_k - \frac{1}{2} \frac{n(\xi \eta_k)^2}{3^2 (3^2 + 24 \cdot 2^{2(1-a)} / 9) B_\infty T_k^2 R_k^2} \right\} \end{aligned}$$

$$(118) \quad \leq \exp \left\{ -\frac{1}{4} \frac{n(\xi \eta_k)^2}{(9^2 + 96 \cdot 2^{-2a}) B_\infty T_k^2 R_k^2} \right\}$$

$$(119) \quad \leq \exp \left\{ -\frac{n \xi^2 C' k}{c^2 B_\infty R^{2-2a}} \right\}.$$

In (117) we applied the fact  $\beta_{k-1}\xi\eta_k \leq 12B_\infty 2^{2(1-a)}T_k^2R_k^2$ , which follows since  $T_kR_k = cR_k^{1-a} = 2^{1-a}T_{k+1}R_{k+1}$ , which implies

$$(120) \quad \beta_k \leq \frac{12B_\infty 2^{2(1-a)}T_{k+1}^2R_{k+1}^2}{\eta_{k+1}\xi},$$

since  $0 < \eta_k \leq 1$ , where  $\eta_k$  is defined in (111). In (118) we applied the first term in the maximum in (111) which implies  $2H_k \leq 4^{-1}n(\xi\eta_k)^2/[(9^2 + 96 \cdot 2^{-2a})B_\infty T_k^2 R_k^2]$ . In (119) we applied the second term in the maximum in (111), which implies  $\eta_k^2/[4 \cdot (9^2 + 96 \cdot 2^{-2a})T_k^2 R_k^2] \geq C'k/[c^2 R^{2-2a}]$ . We may continue (113) with an upper bound

$$(121) \quad \sum_{k=1}^{\infty} \exp \left\{ -\frac{n\xi^2 C' k}{c^2 B_\infty R^{2-2a}} \right\} \leq 2 \exp \left\{ -\frac{n\xi^2 C'}{c^2 B_\infty R^{2-2a}} \right\}.$$

We applied the fact that for  $0 \leq a \leq 1/2$ ,  $\sum_{k=1}^{\infty} a^k = a/(1-a) \leq 2a$ . Here we need that  $\exp \{-n\xi^2 C' / [c^2 B_\infty R^{2-2a}]\} \leq 1/2$ , that is we need,  $\xi \geq \left(\frac{\log_e 2}{n C'}\right)^{1/2} c B_\infty^{1/2} R^{1-a}$  which is implied by (107).

*Term  $P_{II}$ .* We have

$$g - h_g^{k,L} \leq \Delta_g^k$$

and thus

$$\nu_n \left( Q(g - h_g^{\kappa_g, L}) \right) \leq \nu_n \left( Q\Delta_g^{\kappa_g} \right) + 2E \left| Q\Delta_g^{\kappa_g} \right|.$$

Here we used the assumption that operator  $Q$  preserves positivity ( $g \geq 0$  implies  $Qg \geq 0$ ). We have for  $k = 0, \dots, K$ ,

$$(122) \quad \begin{aligned} \max \left\{ E \left| Q\Delta_g^k \right|^2 : g \in \mathcal{G} \right\} &\leq B_\infty \max \left\{ \left\| Q\Delta_g^k \right\|_2^2 : g \in \mathcal{G} \right\} \\ &\leq B_\infty T_k^2 \max \left\{ \left\| \Delta_g^k \right\|_2^2 : g \in \mathcal{G} \right\} \\ &\leq B_\infty T_k^2 R_k^2. \end{aligned}$$

When  $\kappa_g = k$ , then  $Q\Delta_g^{\kappa_g} \geq \beta_k$ , for  $k = 0, \dots, K-1$ . Thus, for  $\kappa_g = k$ ,  $k = 0, \dots, K-1$ , using (109),

$$E|Q\Delta_g^{\kappa_g}| \leq \beta_k^{-1} E|Q\Delta_g^k|^2 \leq \beta_k^{-1} B_\infty T_k^2 R_k^2 \leq \xi/12,$$

and for  $\kappa_g = K$ ,

$$E|Q\Delta_g^{\kappa_g}| \leq \left( E|Q\Delta_g^K|^2 \right)^{1/2} \leq B_\infty^{1/2} T_K R_K \leq \xi/12,$$

when we choose

$$(123) \quad K = \min \left\{ k \geq 1 : 12B_\infty^{1/2} T_k R_k < \xi \right\}.$$

Thus

$$P \left( \sup_{g \in \mathcal{G}} 2E|Q\Delta_g^{\kappa_g}| > \xi/6 \right) = 0.$$

Define

$$\mathcal{G}^{(k)} = \{g \in \mathcal{G} : \kappa_g = k\}, \quad k = 0, \dots, K,$$

so that  $\mathcal{G} = \bigcup_{k=0}^K \mathcal{G}^{(k)}$ . Then,

$$(124) \quad \begin{aligned} P_{II} &\leq P \left( \sup_{g \in \mathcal{G}} \nu_n(Q\Delta_g^{\kappa_g}) \geq \xi/6 \right) \leq \sum_{k=0}^K P \left( \sup_{g \in \mathcal{G}^{(k)}} \nu_n(Q\Delta_g^k) \geq \xi/6 \right) \\ &= P_{II}^{(0)} + P_{II}^{(1)}, \end{aligned}$$

where

$$P_{II}^{(0)} = P \left( \sup_{g \in \mathcal{G}^{(0)}} \nu_n(Q\Delta_g^0) \geq \xi/6 \right), \quad P_{II}^{(1)} = \sum_{k=1}^K P \left( \sup_{g \in \mathcal{G}^{(k)}} \nu_n(Q\Delta_g^k) \geq \xi/6 \right).$$

We have

$$(125) \quad \# \left\{ \Delta_g^k : g \in \mathcal{G}^{(k)} \right\} \leq \# \left\{ \Delta_g^k : g \in \mathcal{G} \right\} \leq N_k.$$

It holds that

$$(126) \quad \left| Q\Delta_g^0 - EQ\Delta_g^0 \right| \leq 4B'_\infty.$$

We have, using (122), (125), (126), by Bernstein's inequality,

$$(127) \quad P_{II}^{(0)} \leq N_0 \exp \left\{ -\frac{1}{2} \frac{n(\xi/6)^2}{B_\infty T_0^2 R_0^2 + 2B'_\infty \xi/9} \right\}.$$

Let us turn to  $P_{II}^{(1)}$ . For  $\kappa_g = k$  (that is, when  $g \in \mathcal{G}^{(k)}$ ), for  $k = 1, \dots, K$ ,

$$Q\Delta_g^k \leq Q\Delta_g^{k-1} \leq \beta_{k-1},$$

which implies

$$(128) \quad \left| Q\Delta_g^k - EQ\Delta_g^k \right| \leq 2\beta_{k-1}.$$

Thus, using (122), (125), (128), the fact that  $0 < \eta_k \leq 1$ , where  $\eta_k$  is defined in (111), by Bernstein's inequality, for  $k = 1, \dots, K$ ,

$$\begin{aligned} P\left(\sup_{g \in \mathcal{G}^{(k)}} \nu_n(Q\Delta_g^k) \geq \xi/6\right) &\leq P\left(\sup_{g \in \mathcal{G}^{(k)}} \nu_n(Q\Delta_g^k) \geq \xi\eta_k/6\right) \\ &\leq N_k \exp\left\{-\frac{1}{2} \frac{n(\xi\eta_k/6)^2}{B_\infty T_k^2 R_k^2 + \beta_{k-1} \xi\eta_k/9}\right\} \\ (129) \quad &\leq \exp\left\{H_k - \frac{1}{2} \frac{n(\xi\eta_k)^2}{6^2(1 + 2^{2(2-a)}/3)B_\infty T_k^2 R_k^2}\right\} \end{aligned}$$

$$(130) \quad \leq \exp\left\{-\frac{1}{4} \frac{n(\xi\eta_k)^2}{(6^2 + 48 \cdot 2^{-2a})B_\infty T_k^2 R_k^2}\right\}$$

$$(131) \quad \leq \sum_{k=1}^{\infty} \exp\left\{-\frac{n\xi^2 C' k}{B_\infty c^2 R^{2-2a}}\right\}.$$

In (129) we applied the fact  $\beta_{k-1} \xi\eta_k \leq 12B_\infty 2^{2(1-a)} T_k^2 R_k^2$ , which follows by using (120). In (130) we applied the first term in the maximum in (111) which implies  $H_k \leq 4^{-1} n(\xi\eta_k)^2 / [(6^2 + 48 \cdot 2^{-2a}) B_\infty T_k^2 R_k^2]$ , since  $2^{-1}(6^2 + 48 \cdot 2^{-2a}) \leq 9^2 + 96 \cdot 2^{-2a}$ . In (131) we applied the second term in the maximum in (111), which implies  $\eta_k^2 / [4 \cdot (6^2 + 48 \cdot 2^{-2a}) T_k^2 R_k^2] \geq C' k / [c^2 R^{2-2a}]$ . We get

$$(132) \quad P_{II}^{(1)} \leq \sum_{k=1}^{\infty} \exp\left\{-\frac{n\xi^2 C' k}{B_\infty c^2 R^{2-2a}}\right\} \leq 2 \exp\left\{-\frac{n\xi^2 C'}{B_\infty c^2 R^{2-2a}}\right\}.$$

In (132) we applied that for  $0 \leq a \leq 1/2$ ,  $\sum_{k=1}^{\infty} a^k = a/(1-a) \leq 2a$ . Here we need that  $\exp\{-n\xi^2 C' / [B_\infty c^2 R^{2-2a}]\} \leq 1/2$ , that is we need,  $\xi \geq \left(\frac{\log_e 2}{n C'}\right)^{1/2} B_\infty c R^{1-a}$  which is implied by (107).

*Term P<sub>III</sub>.* We have first,

$$\#\{h_g^{0,L} : g \in \mathcal{G}\} \leq N_0,$$

second

$$\sup_{g \in \mathcal{G}} E \left| Qh_g^{0,L} \right|^2 \leq B_\infty \sup_{g \in \mathcal{G}} \left\| Qh_g^{0,L} \right\|_2^2 \leq B_\infty T_0^2 R_0^2 = B_\infty c^2 R^{2-2a},$$

and third

$$\sup_{g \in \mathcal{G}} \left\| Qh_g^{0,L} \right\|_\infty \leq B'_\infty.$$

Thus, by Bernstein's inequality

$$(133) \quad P_{III} \leq N_0 \exp \left\{ -\frac{1}{2} \frac{n(\xi/3)^2}{B_\infty T_0^2 R_0^2 + \xi 2B'_\infty/9} \right\}.$$

*Finishing the proof.* The lemma follows from (110), (121), (127), (132), and (133), after checking some final facts. We need to check that  $\sum_{k=1}^{\infty} \eta_k \leq 1$ . Applying the calculations in (104) and (106) we get

$$(134) \quad \sum_{k=1}^{\infty} \eta_k \leq (9^2 + 96 \cdot 2^{-2a})^{1/2} \left( \frac{8^{1/2} B_\infty^{1/2} 2G(R)}{n^{1/2} \xi} + 2\sqrt{C' C''} \right) \leq \frac{1}{2} + \frac{1}{2} = 1,$$

when  $\xi \geq 2 \cdot 8^{1/2} 6G(R)n^{-1/2}(9^2 + 96 \cdot 2^{-2a})^{1/2} B_\infty^{1/2}$ , which is guaranteed by (107), and  $C'$  is chosen as in (112).  $\square$

REMARK 7. When in addition  $\xi$  satisfies

$$(135) \quad 2\sqrt{44 \log_e(\#\mathcal{G}_R) B_\infty^{1/2} c R^{1-a} n^{-1/2}} \leq \xi \leq B_\infty c^2 R^{2(1-a)} / B'_\infty,$$

then

$$\#\mathcal{G}_R \exp \left\{ -\frac{1}{12} \frac{n\xi^2}{B_\infty c^2 R^{2(1-a)} + 2\xi B'_\infty/9} \right\} \leq \exp \left\{ -\frac{n\xi^2 C'}{B_\infty c^2 R^{2(1-a)}} \right\}.$$

Indeed, we may continue (127) by

$$(136) \quad \begin{aligned} P_{II}^{(1)} &\leq N_0 \exp \left\{ -\frac{1}{2} \frac{n(\xi/6)^2}{B_\infty T_0^2 R_0^2 + 2B'_\infty \xi/9} \right\} \\ &\leq \exp \left\{ H_0 - \frac{1}{2} \frac{n\xi^2}{6^2(1+2/9)B_\infty c^2 R^{2(1-a)}} \right\} \end{aligned}$$

$$(137) \quad \leq \exp \left\{ -\frac{1}{4} \frac{n\xi^2}{44B_\infty c^2 R^{2(1-a)}} \right\}.$$

In (136) we applied the upper bound in (135) and the fact  $T_0 R_0 = cR^{1-a}$ . In (137) we applied the lower bound in (135) which implies the fact  $H_0 \leq 4^{-1} n\xi^2 / [44B_\infty c^2 R^{2(1-a)}]$ . Also, we may continue (133) by

$$(138) \quad \begin{aligned} P_{III} &\leq N_0 \exp \left\{ -\frac{1}{2} \frac{n(\xi/3)^2}{B_\infty T_0^2 R_0^2 + \xi 2B'_\infty/9} \right\} \\ &\leq \exp \left\{ H_0 - \frac{1}{2} \frac{n\xi^2}{3^2(1+2/9)B_\infty c^2 R^{2-2a}} \right\} \end{aligned}$$

$$(139) \quad \leq \exp \left\{ -\frac{1}{4} \frac{n\xi^2}{11B_\infty c^2 R^{2-2a}} \right\}.$$

In (138) where we applied the upper bound in (135). In (139) we applied the lower bound in (135) which implies  $H_0 \leq 4^{-1}n\xi^2/(11B_\infty c^2 R^{2-2a})$ .

### APPENDIX C: INTRODUCTORY REMARKS

We add a short introduction to the setting of the article, in order to make the article more accessible to PhD students.

A quite general inverse problem could be described as a problem where we want to recover a function  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  when we have only available some transform  $Af$  of the function. An important example is the sampling operator  $Af = (f(x_1), \dots, f(x_n)) \in \mathbf{R}^n$ , where  $x_1, \dots, x_n \in \mathbf{R}^d$  are fixed points. Classical methods for recovering  $f$  in this case include piecewise constant interpolation and various ways to linearly interpolate the observed function values. In statistics some kind of sampling operator is always involved and thus recovering  $f$  from noisy data  $f(x_i) + \epsilon_i$ ,  $i = 1, \dots, n$ , where  $\epsilon_i$  are error terms, would not be called an inverse problem in statistics. We mention three classical statistical inverse problems, where function  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  has to be estimated and  $A$  is a fixed operator mapping functions  $\mathbf{R}^d \rightarrow \mathbf{R}$  to functions  $\mathbf{Y} \rightarrow \mathbf{R}$ , where  $\mathbf{Y}$  is some general space.

1. (*Regression function estimation.*) We observe data

$$Y_i = (Af)(X_i) + \epsilon_i \in \mathbf{R}, \quad i = 1, \dots, n,$$

where  $\epsilon_i \in \mathbf{R}$  are random errors and  $X_i \in \mathbf{Y}$  are random design points.

2. (*Density estimation.*) We observe identically distributed observations

$$Y_1, \dots, Y_n \in \mathbf{Y},$$

whose common density is  $Af$ .

3. (*Gaussian white noise model.*) We observe a realization of the process

$$dY_n(y) = (Af)(y) dt + n^{-1/2} dW(y), \quad y \in \mathbf{Y},$$

where  $W(y)$  is a Wiener process on  $\mathbf{Y}$ . When  $\mathbf{Y} = \mathbf{R}$ , then we can define the process by

$$Y_n(y) = \int_{-\infty}^y (Af)(t) dt + n^{-1/2} W(y),$$

where  $W$  is the Brownian motion, or Wiener process, on the real line. The Gaussian white noise model is rather close to the regression function estimation when the error terms  $\epsilon_i$  are Gaussian and the design points  $X_i$  are uniformly distributed in the unit square. However, in the

Gaussian white noise model we have eliminated the problems related to interpolation since the function  $Af$  is observed continuously and not in a finite number of design points. Since the assumption of continuous observation is quite far from reality, we can use inference in the Gaussian white noise model only as a first approximation. In addition, the assumption of the exact Gaussian distribution is very restrictive. Due to the central limit theorem the Gaussian white noise model is a relevant approximation also for the model of density estimation and for the model of regression function estimation under non-Gaussian noise.

Let us now consider the estimation of a regression function (item 1 of the above list). A common approach for regression function estimation is to find the estimator  $\tilde{f}$  as a solution of the minimization problem

$$(140) \quad \tilde{f} = \operatorname{argmin}_{g \in \mathcal{F}} \sum_{i=1}^n (Y_i - (Ag)(X_i))^2,$$

where  $\mathcal{F}$  is some class of functions  $\mathbf{R}^d \rightarrow \mathbf{R}$ . Note that estimator  $\hat{f}$  is a special case of the linear regression estimator

$$\tilde{f}(x) = \hat{\beta}_0 + \hat{\beta}_1^T x, \quad (\hat{\beta}_0, \hat{\beta}_1) = \operatorname{argmin}_{\beta_0 \in \mathbf{R}, \beta_1 \in \mathbf{R}^d} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1^T X_i)^2,$$

when  $A$  is the identity operator and  $\mathcal{F} = \{\beta_0 + \beta_1^T x : \beta_0 \in \mathbf{R}, \beta_1 \in \mathbf{R}^d\}$ . Estimator  $\tilde{f}$ , defined in (140), can be defined also as

$$(141) \quad \tilde{f} = \operatorname{argmin}_{g \in \mathcal{F}} \left( -\frac{2}{n} \sum_{i=1}^n Y_i \cdot (Ag)(X_i) + \frac{1}{n} \sum_{i=1}^n (Ag)^2(X_i) \right).$$

The estimator which we have considered can be defined, assuming now for simplicity that the design points  $X_i$  have a known distribution  $\nu$  on  $\mathbf{Y}$ ,

$$(142) \quad \hat{f} = \operatorname{argmin}_{g \in \mathcal{F}} \left( -\frac{2}{n} \sum_{i=1}^n Y_i \cdot (Qg)(X_i) + \int_{[0,1]^d} g^2 \right),$$

where  $Q = (A^{-1})^*$  is the adjoint of the inverse, for the space  $L_2(\nu)$ . Note that when  $g : \mathbf{R}^d \rightarrow \mathbf{R}$ , then  $Qg : \mathbf{Y} \rightarrow \mathbf{R}$ .

When operator  $B : H_1 \rightarrow H_2$  is defined as a mapping from a Hilbert space  $H_1$  to an another Hilbert space  $H_2$ , then the adjoint  $B^*$  is defined as the operator satisfying the equality

$$\langle Bx, y \rangle_2 = \langle x, B^*y \rangle_1,$$

where  $\langle \cdot, \cdot \rangle_i$  are the inner products of the Hilbert spaces. In the case when the Hilbert spaces are the Euclidean space:  $H_1 = H_2 = \mathbf{R}^d$ , then the operators are  $d \times d$  real matrices, and we have  $\langle Bx, y \rangle = \langle x, B^T y \rangle$ , where  $B^T$  is the transpose of matrix  $B$ , and thus the adjoint is equal to the transpose. We have given further examples of adjoints in (32), where the adjoint of the inverse of a convolution operator is given, and in (34), where the adjoint of the inverse of the Radon transform is given.

The estimator defined by (140) and (141) seems quite natural but we can justify the estimator defined in (142) by the following calculation. We have, similarly as in (5),

$$\begin{aligned} \|\hat{f} - f\|_2^2 - \|f\|_2^2 &= -2 \int_{\mathbf{R}^d} f \hat{f} + \|\hat{f}\|_2^2 \\ &= -2 \int_{\mathbf{Y}} (Af)(Q\hat{f}) d\nu + \|\hat{f}\|_2^2 \\ &\approx -\frac{2}{n} \sum_{i=1}^n Y_i \cdot (Q\hat{f})(X_i) + \|\hat{f}\|_2^2. \end{aligned}$$

The last approximation in the above calculation uses the fact that the distribution of the design points is  $\nu$ .

JUSSI KLEMELÄ  
UNIVERSITY OF OULU  
DEPARTMENT OF MATHEMATICAL SCIENCES  
P. O. BOX 3000  
90014 UNIVERSITY OF OULU  
FINLAND  
FAX: +358-8-5531730  
E-MAIL: klemela@oulu.fi

ENNO MAMMEN  
UNIVERSITY OF MANNHEIM, DEPARTMENT OF ECONOMICS  
L7 3-5,  
68131 MANNHEIM, GERMANY  
FAX +49-621-1811931  
E-MAIL: emammen@rumms.uni-mannheim.de