Institut für Angewandte Analysis
und Stochastik

im Forschungsverbund Berlin e.V.

Asymptotic equivalence of density estimation and white noise

M. Nussbaum

submitted: 23rd December 1992

Institut für Angewandte Analysis
und Stochastik
Hausvogteiplatz 5-7
D – O 1086 Berlin
Germany

Preprint No. 35
Berlin 1992

1991 Mathematics Subject Classification. Primary 62 G 07; Secondary 62 B 15, 62 G 20.
Key words and phrases. Nonparametric experiments, deficiency distance, curve estimation, likelihood ratio process, Hungarian construction, asymptotic minimax risk, exact constants, Hellinger loss, linear wavelet estimators.

Herausgegeben vom
Institut für Angewandte Analysis und Stochastik
Hausvogteiplatz 5-7
D – O 1086 Berlin

Fax: + 49 30 2004975
e-Mail (X.400): c=de;a=dbp;p=iaas-berlin;s=preprint
e-Mail (Internet): preprint@iaas-berlin.dbp.de

Asymptotic Equivalence of Density Estimation and White Noise

Michael Nussbaum
IAAS Berlin

December 1992

Abstract

Signal recovery in Gaussian white noise with variance tending to zero has served for some time as a representative model for nonparametric curve estimation, having all the essential traits in a purified form. The equivalence has mostly been stated informally, but an approximation in the sense of Le Cam's deficiency distance Δ would make it precise. Then two models are asymptotically equivalent for all purposes of statistical decision with bounded loss. In nonparametrics, a first result of this kind has recently been established for Gaussian regression (Brown and Low, 1992). We consider the analogous problem for the experiment given by n i. i. d. observations having density f on the unit interval. Our basic result concerns the parameter space of densities which are in a Sobolev class of order 4 and uniformly bounded away from zero. We show that an i. i. d. sample of size n with density f is globally asymptotically equivalent to a white noise experiment with trend $f^{1/2}$ and variance $\frac{1}{4}n^{-1}$. This represents a nonparametric analog of Le Cam's heteroskedastic Gaussian approximation in the finite dimensional case. The proof utilizes empirical process techniques, especially the Hungarian construction. White noise models on f and $\log f$ are also considered, allowing for various "automatic" asymptotic risk bounds in the i. i. d. model from white noise. As first applications we discuss linear wavelet estimators of a density and exact constants for Hellinger loss.

1 Introduction and main result

One of the basic principles of Le Cam's (1986) asymptotic decision theory is to approximate general experiments by simple ones. In particular, *weak convergence to Gaussian shift experiments* has now become a standard tool for establishing asymptotic risk bounds. The risk bounds implied by weak convergence are generally estimates from below, and in most of the literature the efficiency of procedures is more or less shown on an ad hoc basis. However, a systematic approach to the attainment problem is also made possible by Le Cam's theory, based on the notion of *strong convergence* which means proximity in the sense of the full

1990 *Mathematics Subject Classification.* Primary 62 G 07 ; Secondary 62 B 15, 62 G 20
Key words and phrases. Nonparametric experiments, deficiency distance, curve estimation, likelihood ratio process, Hungarian construction, asymptotic minimax risk, exact constants, Hellinger loss, linear wavelet estimators.

deficiency distance. But due to the inherent technical difficulties of handling the deficiency concept, this possibility is rarely made use of, even in root-n consistent parametric problems. In nonparametric curve estimation models of the 'inverse problem' class where there is no root-n consistency, a theory on *attainable exact risk bounds* is developing the origin of which has not been reduction to a limit experiment. Such an exact risk bound was first discovered by Pinsker (1980) in the problem of *signal recovery in Gaussian white noise*, which is by now recognized as the basic or "typical" nonparametric curve estimation problem. The cognitive value of this model had already been put forward by Ibragimov and Khasminski (1977). These risk bounds have been established since then in a variety of other problems, e. g. density, nonparametric regression, spectral density, see Efroimovich and Pinsker (1982), Golubev (1985), Nussbaum (1985)); and they have also been substantially extended conceptually (Donoho and Johnstone (1992)). The theory is now at a stage where the approximation of the various particular curve estimation problems by the white noise model could be made formal. An important step in this direction has been made by Brown and Low (1992) by relating Gaussian regression to the signal recovery problem. These models are essentially the continuous and discrete versions of each other. The aim of this paper is to attempt the *formal approximation by the white noise model* for the problem of density estimation from an i. i. d. sample.

To formulate our main result, define a basic parameter space Σ of densities as follows. Let

$$W_2^m(M) = \left\{ f \in L_2(0,1), \left\| f^{(m)} \right\|_2^2 \leq M \right\}$$

be an L_2-Sobolev class of order m. Let $\mathcal{F}_{\geq \epsilon}$ be the set of densities on $[0,1]$ bounded below by ϵ:

$$\mathcal{F}_{\geq \epsilon} = \left\{ f;\ \int_0^1 f = 1,\ f(x) \geq \epsilon,\ x \in [0,1] \right\} \tag{1}$$

Define an a priori set, for given $\epsilon > 0$, $M > 0$

$$\Sigma_{\epsilon,M} = W_2^4(M) \cap \mathcal{F}_{\geq \epsilon}.$$

For two sequences of experiments $\mathcal{P}_n$ and $\mathcal{Q}_n$ having the same parameter space we shall say that they are *asymptotically equivalent* if the respective deficiency Δ distance tends to zero, i. e. if $\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0$ as $n \to \infty$. Let dW denote the standard Gaussian white noise process on the unit interval.

1.1 Theorem. *For any $\epsilon > 0$, $M > 0$, the experiments given by observations*

$$y_i,\ i = 1, \ldots, n \qquad \textit{i. i. d. with density } f \tag{2}$$

$$dy(t) = f^{1/2}(t)dt + \frac{1}{2} n^{-1/2} dW(t),\ t \in [0,1] \tag{3}$$

with $f \in \Sigma_{\epsilon,M}$ are asymptotically equivalent .

This result is closely related to Le Cam's global asymptotic normality for parametric models. Let in the i. i. d. model f be in a parametric family $\{P_\vartheta,\ \vartheta \in \Theta\}$ where $\Theta \subset \mathbf{R}^k$, which is

sufficiently regular and has Fisher information matrix $I(\vartheta)$ at point ϑ. Then the i. i. d. model may be approximated by a heteroskedastic Gaussian experiment

$$y = \vartheta + n^{-1/2} I(t_n(\vartheta))^{-1/2} \eta \tag{4}$$

where η is a standard normal vector and $t_n(\vartheta)$ is a map which assigns to ϑ an element of a certain discrete net in Θ which becomes successively more dense with n (see Le Cam (1986), chap. 11). We see that (3) is a nonparametric analog of (4) when ϑ is identified with $f^{1/2}$. Indeed, the identity for the Fisher information matrix in the parametric case

$$\left\| f_{\vartheta'}^{1/2} - f_{\vartheta}^{1/2} \right\|^2 = \langle \vartheta' - \vartheta, 4I(\vartheta)(\vartheta' - \vartheta) \rangle + o\left(\|\vartheta' - \vartheta\|^2 \right)$$

formally describes $I(f^{1/2})$ as $\frac{1}{4}$ times the unit operator. But even for parametric families (3) seems to be an advantageous form of a global approximation, since the discretization map $t_n(\vartheta)$ is absent there. When deducing (4) from (3), which is possible for parametric families in $\Sigma_{\epsilon,M}$, one recognizes that this complication derives only from the "curved" nature of a general parametric family in the space of roots of densities.

We believe that the restriction to a parameter space $\Sigma_{\epsilon,M}$, in particular to densities of smoothness 4 is of preliminary nature. What matters in our view is that $\Sigma_{\epsilon,M}$ is genuinely global and nonparametric, and that approximation by the experiment (3) provides conceptual insight on asymptotic normality of i. i. d. models. White noise models with *fixed* variance do occur as local limits of experiments in $\sqrt{n}$ consistent nonparametric problems (Millar (1979)), and, via specific renormalizations, also in non root-n consistent curve estimation (Low (1992), Donoho and Low (1992)). Thus various central limit theorems for i. i. d. experiments can be imdedded in a relatively simple and closed form approximation by (3).

The paper is organized as follows. In section 2 we develop the basic approximation of likelihood ratios over local shrinking neighborhoods of a given density f_0. These neighborhoods $\Sigma_n(f_0)$ are already "nonparametric", in the sense of shrinking slower than $n^{-1/2}$. The technical part of the proof is in the appendix of the paper. Once in a Gaussian framework, in section 3 we manipulate likelihood ratios to obtain other approximations, in particular the one with trend $f^{1/2}$. For these experiments which are all Gaussian we use the methodology of Brown and Low (1992), who did compare the white noise model with its discrete version (the Gaussian regression model).

Piecing together local approximations to a global one by means of a preliminary estimator is the subject of section 4; the proof of theorem 1.1 is at the end of this section. Our method is somewhat different from Le Cam's which works in the parametric case; the concept of metric entropy or dimension and related theory is not utilized. But obviously these methods which already proved fruitful in nonparametrics (Birgé (1982), Van de Geer (1990)) have a potential application also here. The same holds true for the results of Mammen (1986) on the informational content of additional observations; this paper is also recommended for an accessible overwiew of some global asymptotic theory.

Some statistical consequences are discussed in section 5; here we focus on exact constants for L_2-loss. As an exercise we derive the result of Efroimovich and Pinsker (1982) result on density estimation from the white noise model; this then serves as a basis for extensions to linear wavelet density estimators and to exact constants for Hellinger loss.

The preliminary estimator required for the global approximation is treated in section 6, with an emphasis on existence. A more constructive theory of log-density estimation in exponential

families which was also instrumental for our approximation result has been developed recently by Barron and Sheu (1991).

2 The local approximation

Our model will be estimation of a density f on the unit interval [0,1]. Supppose we have i. i. d. observations X_i, $i = 1, \ldots, n$ distributed with Lebesgue density f, and it is known a priori that f belongs to a set Σ_0 of densities. Define Σ_0 to be the class of all densities on $[0,1]$ which are strictly positive and of bounded variation. Thus the logarithms of densities in Σ_0 exist, are bounded and of bounded variation.
Let $\|\cdot\|_p$ denote the norm in the space $L_p(0,1)$, $1 \leq p \leq \infty$, and $\|\cdot\|_{TV}$ be the total variation norm. Let a sequence $\tau_n \to \infty$ be given, and for any $f_0 \in \Sigma_0$ define a class $\Sigma_n(f_0)$ by

$$\Sigma_n(f_0) = \{f \in \Sigma_0, \quad \|\log f - \log f_0\|_\infty \leq \tau_n^{-1} n^{-1/3}, \quad \|\log f - \log f_0\|_{TV} \leq \tau_n n^{-1/3}\} \tag{5}$$

To define the approximating Gaussian shift experiment, assume a f_0 in Σ_0 fixed and let F_0 be the corresponding distribution function. Let B be the standard Brownian bridge on [0,1] and consider an observed process

$$y(t) = \int_0^t \log \frac{f}{f_0}(F_0^{-1}(u))du - tK(f_0\|f) + n^{-1/2}B(t), \ t \in [0,1]. \tag{6}$$

Let Q_{n,f,f_0} be the distribution of this process, and

$$\mathcal{Q}_{n,f_0} = \{Q_{n,f,f_0}, f \in \Sigma_n(f_0)\}$$

be the corresponding experiment when f varies in a neighborhood $\Sigma_n(f_0)$. Let $P_{n,f}$ be the joint distribution of the observations $X_i, i = 1, \ldots, n$, and let

$$\mathcal{P}_{n,f_0} = \{P_{n,f}, \ f \in \Sigma_n(f_0)\}$$

be the corresponding experiment around f_0. Let Δ denote Le Cam's deficiency distance.

2.1 Theorem. *Let M and $\tau_n \to \infty$ be given, and define $\Sigma_n(f_0)$ as in (5). Suppose $\tau_n = o(n^\epsilon)$ for any $\epsilon > 0$. Then*

$$\Delta(\mathcal{P}_{n,f_0}, \mathcal{Q}_{n,f_0}) \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

uniformly over $f_0 \in \Sigma_0$.

The proof is based upon the following principle, described in Le Cam and Yang (1991), p. 16. Consider two experiments $\mathcal{P} = \{P_\vartheta, \vartheta \in \Theta\}$ and $\mathcal{Q} = \{Q_\vartheta, \vartheta \in \Theta\}$. Assume there is some point $\vartheta_0 \in \Theta$ such that all the P_ϑ are dominated by P_{ϑ_0} and all the Q_ϑ are dominated by Q_{ϑ_0}. Construct the process $\Lambda^{(0)} = \{\Lambda^{(0)}(\vartheta), \vartheta \in \Theta\}$ where $\Lambda^{(0)}(\vartheta) = \frac{dP_\vartheta}{dP_{\vartheta_0}}$. Construct $\Lambda^{(1)}$ similarly with $\Lambda^{(1)}(\vartheta) = \frac{dQ_\vartheta}{dQ_{\vartheta_0}}$. Give $\Lambda^{(0)}$ the distribution induced by P_{ϑ_0} and $\Lambda^{(1)}$ the distribution induced by Q_{ϑ_0}

2.2 Proposition. *Suppose there is a pairing on some common probability space such that*

$$\sup_{\vartheta\in\Theta} \quad E|\Lambda^{(0)}(\vartheta) - \Lambda^{(1)}(\vartheta)| \leq \epsilon.$$

Then $\Delta(\mathcal{P}, \mathcal{Q}) \leq \epsilon/2$.

Let $(\Omega, \mathcal{B}, \mathbf{P})$ be the common probability space for the processes $\Lambda^{(i)}$. The proof of this proposition is easily argued by showing that the experiment $\mathcal{P}^* = \{P^*_\vartheta, \vartheta \in \Theta\}$ given by measures $dP^*_\vartheta = \Lambda^{(0)}(\vartheta)d\mathbf{P}$ is of the same type as $\mathcal{P}$ (likewise for $\mathcal{Q}$ and $\mathcal{Q}^*$ given by measures $dQ^*_\vartheta = \Lambda^{(1)}(\vartheta)d\mathbf{P}$), and that $E|\Lambda^{(0)}(\vartheta) - \Lambda^{(1)}(\vartheta)|$ is the total variation distance between P^*_ϑ and Q^*_ϑ. Thus the inequality for the deficiency follows from the total variation distance estimate for equivalent representations.
For our problem, we identify $\vartheta = f, \Theta = \Sigma_n(f_0), \mathcal{P} = \mathcal{P}_{n,f_0}, \mathcal{Q} = \mathcal{Q}_{n,f_0}$. Furthermore, we represent the observations X_i as $X_i = F^{-1}(Z_i)$, where Z_i are i. i. d. uniform (0,1) random variables and F is the distribution function for the density f (note that F are strictly monotone for $f \in \Sigma_0$). We will then make use of the *Hungarian construction* (see Shorack, Wellner (1986), chap. 12, section 1, theor. 2). Let $\mathbf{U}_n$ be the empirical process of $Z_1, \ldots, Z_n$, i. e.

$$\mathbf{U}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\chi_{[0,t]}(Z_i) - t), \; t \in [0,1].$$

2.3 Proposition (The Hungarian construction). *There is a double array of independent uniform (0,1) random variables* $Z_{in}, i = 1, \ldots, n, \; n = 1, 2, \ldots$ *and a sequence of Brownian Bridges* $\mathbf{B}_n, n = 1, 2, \ldots$, *all defined on a common probability space* $(\Omega, \mathcal{B}, \mathbf{P})$, *such that for some positive constants* c_1, c_2, c_3 *not depending on* n *we have*

$$\mathbf{P}(\| \mathbf{U}_n - \mathbf{B}_n \|_\infty > (c_1 \log n + x)/\sqrt{n}) \leq c_2 \exp(-c_3 x) \tag{7}$$

Using the elements of this construction, we define the pairing of proposition 1 as follows. Note that the experiment $\mathcal{P}_{n,f_0}$ is dominated by P_{f_0}; then the likelihood ratio process $\Lambda^{(0)}$ is

$$\Lambda^{(0)}(f) = \exp \sum_{i=1}^{n} \log \left\{ \frac{f}{f_0}(F_0^{-1}(Z_{in})) \right\}$$

Note that

$$E \log \left\{ \frac{f}{f_0}(F_0^{-1}(Z_i)) \right\} = \int \log \frac{f}{f_0} dF_0 = -K(f_0 \parallel f),$$

where $K(f_0 \parallel f)$ is the Kullback-Leibler relative entropy. Defining

$$\kappa_{f,f_0}(t) = \log \left\{ \frac{f}{f_0}(F_0^{-1}(t)) \right\},$$

we then have the following representation of $\Lambda^{(0)}$:

$$\Lambda^{(0)}(f) = \exp \left\{ n \int \kappa_{f,f_0}(t) \frac{1}{\sqrt{n}} \mathbf{U}_n(dt) - nK(f_0 \parallel f) \right\}.$$

The corresponding likelihood ratio process for $\mathcal{Q}_{n,f_0}$ has the form (with Z a uniform (0,1) random variable)

$$\Lambda^{(1)}(f) = \exp\left\{ n \int \kappa_{f,f_0}(t) \frac{1}{\sqrt{n}} \mathbf{B}_n(dt) - \frac{n}{2} \mathrm{Var}\left(\kappa_{f,f_0}(Z)\right) \right\}. \tag{8}$$

2.4 Proposition. *Under the conditions of theorem 2.1 we have on the probability space* $(\Omega, \mathcal{B}, \mathbf{P})$ *for* $n \longrightarrow \infty$

$$E|\Lambda^{(0)}(f) - \Lambda^{(1)}(f)| \longrightarrow 0$$

uniformly over $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma_0$.

Theorem 1 now hinges on this proposition; the proof is in the appendix.

3 Further local approximations

We are now able to identify several more asymptotically equivalent models. Let $W(t)$ be the standard Wiener process on the positive half line. Let $\mathcal{Q}^{(2)}_{n,f_0}$ be the experiment given by observations

$$dy(t) = (f(t) - f_0(t))dt + n^{-1/2} f_0^{1/2}(t) dW(t),\ t \in [0,1] \tag{9}$$

when f varies in $\Sigma_n(f_0)$, and let $\mathcal{Q}^{(3)}_{n,f_0}$ correspondingly given by

$$dy(t) = (f^{1/2}(t) - f_0^{1/2}(t))dt + \frac{1}{2} n^{-1/2} dW(t),\ t \in [0,1]. \tag{10}$$

Let $\Lambda^{(i)}(f)$ be the likelihood ratio process for $\mathcal{Q}^{(i)}_{n,f_0}$ built in analogy to $\Lambda^{(1)}$ in (2.6), i. e. when the dominating element is the one with $f = f_0$.

3.1 Theorem. *The experiments* $\mathcal{Q}^{(i)}_{n,f_0}$, $i = 1,2,3$ *are all asymptotically equivalent. More specifically, there are versions of* $\Lambda^{(i)}$, $i = 1,2,3$, *all defined on the probability space* $(\Omega, \mathcal{B}, \mathbf{P})$ *of proposition 2.4 such that as* $n \to \infty$

$$E|\Lambda^{(1)}(f) - \Lambda^{(i)}(f)| \longrightarrow 0$$

uniformly over $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma_0$, *for* $i = 2,3$.

3.2 Remark. For fixed f_0, the type of the experiment is not changed when the additive term $f_0(t)dt$ in (7) is omitted, since this amounts to a translation of the observed process y. The same is true for the other variants, so that locally asymptotically equivalent experiments for $f \in \Sigma_n(f_0)$ are also given by

$$y_i,\ i = 1, \ldots, n \qquad \text{i. i. d. with density } f \tag{11}$$

$$dy(t) = \log f(F_0^{-1}(t))dt + n^{-1/2} dW(t),\ t \in [0,1] \tag{12}$$

$$dy(t) = f(t)dt + n^{-1/2} f_0^{1/2}(t) dW(t),\ t \in [0,1] \tag{13}$$

$$dy(t) = f^{1/2}(t)dt + \frac{1}{2} n^{-1/2} dW(t),\ t \in [0,1] \tag{14}$$

For the proof of theorem 3.1 we need

3.3 Lemma. *Let $g_i \in L_2(0,1)$, $i = 1,2$ and P_i be the distribution of $\int_0^t g_i + \sigma W(t)$, $t \in [0,1]$, $i = 1,2$. Then for the total variation norm $\|\cdot\|$ for measures we have*

$$\|P_1 - P_2\|_{TV} \leq \left(1 - \exp\left\{-\frac{1}{4\sigma^2}\|g_1 - g_2\|_2^2\right\}\right)^{1/2}$$

Proof. Let $\rho(P_1, P_2)$ be the Hellinger affinity

$$\rho(P_1, P_2) = \int \sqrt{dP_1 dP_2}.$$

Then we have (Le Cam and Yang, p. 25)

$$\|P_1 - P_2\|_{TV} \leq \left(1 - \rho^2(P_1, P_2)\right)^{1/2}.$$

For the Gaussian measures P_i we have

$$\rho(P_1, P_2) = \exp\left\{-\frac{1}{8\sigma^2}\|g_1 - g_2\|_2^2\right\} \square$$

Proof of Theorem 3.1. We shall frequently suppress the index n. Let us first exhibit the versions of $\Lambda^{(i)}$ on $(\Omega, \mathcal{B}, \mathbf{P})$. Let η be a standard normal random variable, defined on $(\Omega, \mathcal{B}, \mathbf{P})$ and independent of $\mathbf{B}$ and the sequence $\{Z_i\}$. Define $\mathbf{W}(t) = \mathbf{B}(t) + t\eta$; then, since $E\mathbf{B}(t)\mathbf{B}(u) = t \wedge u - tu$, it follows that $E\mathbf{W}(t)\mathbf{W}(u) = t \wedge u$, so that $\mathbf{W}$ is a Wiener process. It then readily follows that $\mathbf{B}(t) = \mathbf{W}(t) - t\mathbf{W}(1)$. (Let the stochastic integral $\int g d\mathbf{B}(t)$ be defined via this representation of $\mathbf{B}$). Set $\kappa_{f,f_0}^{(1)} = \kappa_{f,f_0}$; for $\Lambda^{(1)}$ we then have

$$\Lambda^{(1)}(f) = \exp\left\{n\int(\kappa_{f,f_0}^{(1)} - K(f_0\|f))\frac{1}{\sqrt{n}}\, d\mathbf{W} - \frac{n}{2}\|\kappa_{f,f_0}^{(1)} - K(f_0\|f)\|^2\right\}. \tag{15}$$

Furthermore let $\mathbf{W} \circ F_0(t) = \mathbf{W}(F_0(t))$ and define a process

$$\tilde{b}(t) = \int_0^t f_0^{-1/2} d\mathbf{W} \circ F_0.$$

This is a Gaussian process with independent increments, and $\tilde{b}$ has variance $\int_0^t f_0^{-1} dF_0 = t$. Hence $\tilde{b}$ is a Wiener process, and we have for every continuous g on $[0,1]$

$$\int g f_0^{1/2} d\tilde{b} = \int g d\mathbf{W} \circ F_0.$$

In (9), the distribution of y is absolutely continuous with respect to the distribution of $n^{-1/2}\int_0^t f_0^{1/2} d\tilde{b}$, with density

$$\Lambda^{(2)}(f) = \exp\left\{n\int(f - f_0)f_0^{-1}n^{-1/2}f_0^{1/2}d\tilde{b} - \frac{n}{2}\int(f - f_0)^2 f_0^{-1}\right\},$$

Using the relation between $\tilde{b}$ and $\mathbf{W}$, we transform this to

$$\Lambda^{(2)}(f) = \exp\left\{n\left(\frac{f}{f_0} - 1\right)n^{-1/2}d\mathbf{W} \circ F_0 - -\frac{n}{2}\int\left(\frac{f}{f_0} - 1\right)^2 dF_0\right\},$$

and defining the function

$$\kappa^{(2)}_{f,f_0}(t) = \frac{f}{f_0}(F_0^{-1}(t)) - 1, \tag{16}$$

we obtain

$$\Lambda^{(2)}(f) = \exp\left\{n\int \kappa^{(2)}_{f,f_0} n^{-1/2} d\mathbf{W} - \frac{n}{2}\int (\kappa^{(2)}_{f,f_0})^2\right\}, \tag{17}$$

In view of the definition of $\mathbf{W}$, $\Lambda^{(2)}(f)$ is defined as a random variable on $(\Omega, \mathcal{B}, \mathbf{P})$. To obtain $\Lambda^{(3)}(f)$, we also use $\tilde{b}$ in (10) and obtain a likelihood ratio

$$\begin{aligned}\Lambda^{(3)}(f) &= \exp\left\{4n\int (f^{1/2} - f_0^{1/2})\frac{1}{2} n^{-1/2} d\tilde{\mathbf{W}} - \frac{4n}{2}\int (f^{1/2} - f_0^{1/2})^2\right\} \\ &= \exp\left\{2n\left(\frac{f^{1/2}}{f_0^{1/2}} - 1\right) n^{-1/2} d\mathbf{W} \circ F_0 - \frac{4n}{2}\int \left(\frac{f^{1/2}}{f_0^{1/2}} - 1\right)^2 dF_0\right\},\end{aligned}$$

and with

$$\kappa^{(3)}_{f,f_0}(t) = 2\frac{f^{1/2}}{f_0^{1/2}}(F_0^{-1}(t)) - 1$$

we obtain

$$\Lambda^{(3)}(f) = \exp\left\{n\int \kappa^{(3)}_{f,f_0} n^{-1/2} d\mathbf{W} - \frac{n}{2}\int (\kappa^{(3)}_{f,f_0})^2\right\}. \tag{18}$$

Now we apply lemma 3.3 for $\sigma = n^{-1/2}$ and observe that $E|\Lambda^1(f) - \Lambda^i(f)|$ is the total variation distance between the respective elements of $\mathcal{Q}^{(1)}_{n,f_0}$ and $\mathcal{Q}^{(i)}_{n,f_0}$ when these are construed as measures on $(\Omega, \mathcal{B}, \mathbf{P})$. It then remains to prove

$$\sup_{f\in\Sigma_n(f_0)} \|\kappa^{(1)}_{f,f_0} - K(f_0\|f) - \kappa^{(i)}_{f,f_0}\|_2^2 = o(n^{-1}) \tag{19}$$

uniformly over $f_0 \in \Sigma_0$. Using the expansion

$$\log x = \log(1 + x - 1) = x - 1 - \frac{1}{2}(x-1)^2 + o((x-1)^2) \tag{20}$$

and putting $x = \frac{f}{f_0}$, we note that for $f \in \Sigma_n(f_0)$

$$\left\|1 - \frac{f}{f_0}\right\| = O(n^{-1/3})$$

uniformly. (Actually we may write $o(n^{-1/3})$.) Consequently

$$\begin{aligned} K(f_0\|f) &= -\int \kappa^{(1)}_{f,f_0}(t)dt = \int \left(\frac{f}{f_0} - 1\right)(F_0^{-1}(t))dt + O(n^{-2/3}) \\ &= O(n^{-2/3}). \end{aligned} \tag{21}$$

Furthermore

$$\|\kappa^{(1)}_{f,f_0} - \kappa^{(2)}_{f,f_0}\|_2 = O\left(\left\|\left(1 - \frac{f}{f_0}\right)^2\right\|\right) = O(n^{-2/3}). \tag{22}$$

Now (21) and (22) imply

$$\|\kappa^{(1)}_{f,f_0} - K(f_0\|f) - \kappa^{(2)}_{f,f_0}\|_2 = O(n^{-2/3})$$

proving (19) for $i = 2$. To obtain it for $i = 3$, use (20) with $x = \left(\frac{f}{f_0}\right)^{1/2}$ to obtain

$$\left\| 1 - \frac{f^{1/2}}{f_0^{1/2}} \right\| = O(n^{-1/3}),$$

so that

$$\log \frac{f}{f_0}(t) = 2 \log \frac{f^{1/2}}{f_0^{1/2}}(t) = \kappa^{(3)}_{f,f_0}(t) + O(n^{-2/3}) \tag{23}$$

uniformly over $t \in [0,1]$. Now (23) and (21) imply (19) for $i = 3$. □

3.3 Remark. Note that to prove (19), it would have sufficed to have $o(n^{-1/2})$ in place of $O(n^{-2/3})$, so the asymptotic equivalence of the white noise models (12)-(14) is actually valid for f in neighborhoods of f_0

$$\left\{ f \in \Sigma_0, \quad \|\log f - \log f_0\| \le \tau_n^{-1} n^{-1/4} \right\}$$

for some $\tau_n \to \infty$. But we do not have then the equivalence to the density model over these.

4 From local to global results

The local result concerning a shrinking neighborhood of some f_0 is of limited value for statistical inference since in general such prior information cannot be assumed. It would now seem natural to construct an experiment where the prior information is furnished by a preliminary estimator, and subsequently the local Gaussian approximation is built around the value furnished by that estimator.
To formalize this approach, let $N(n)$ define a "fraction of the sample size", i. e. $N(n)$ is a sequence $N(n) \to \infty$, $N(n) < n$, and let the corresponding fraction of the sample be $S_1 = (X_1, \ldots, X_N)$. For the global result we need to restrict the densities to the set Σ defined in section 1. Let then $\hat{g}$ be an estimator of $\log f$ based on S_1 taking values in Σ and fulfilling

$$\inf_{f \in \Sigma} P_{n,f}(\hat{g} \in \Sigma_n(f)) \longrightarrow 1. \tag{24}$$

The set Σ must be chosen to guarantee its existence. If f is m times differentiable, we have for f an attainable rate in sup-norm $(n/\log n)^{-m/(2m+1)}$ (see Woodrofe (1967)), Bickel and Rosenblatt (1973)). Moreover, Barron and Sheu (1992) have shown that for estimating $\log f$, some of the attainable rate results for f carry over. Thus if we presuppose a smoothnes class with m larger than 1, we have reason to expect that there is an estimator $\hat{g}$ based on the whole sample ($N = n$) attaining the rate in sup-norm required in $\Sigma_n(f)$ (i.e. $\tau_n^{-1} n^{-1/3}$). However, the other norm occurrring in $\Sigma_n(f)$ is crucial. Note that the total variation norm is essentially an L_1-norm on the first derivative. For the k-th derivative we have an attainable rate $n^{-(m-k)/(2m+1)}$, which in our setting means that the rate $n^{-1/3}$ would be attainable from

$m = 4$ onwards. *Therefore, for the global result, we have to impose such a strong smoothness condition.* We believe that it is an artefact and due to the nonoptimal method of proof for the local approximation; i. e. the local result should actually hold for larger sets $\Sigma_n(f)$. As a further condition for the global result we need the uniform boundedness from below of our densities.
Let $\mathcal{F}_{\geq\epsilon}$ be the set of densities bounded below by ϵ (see (1)), and define an a priori set, for given $\epsilon > 0$, $M > 0$

$$\Sigma = W_2^4(M) \cap \mathcal{F}_{\geq\epsilon}.$$

In section 6 below we will prove the existence of the required preliminary estimator. In particular we will justify choices $N(n) \sim n/\log n$, $\tau_n = \log n$, and also $N \sim n/2$. In any case assume $n \asymp n - N$ henceforth.

The following construction of a global approximating experiment assumes such an estimator sequence fixed. Consider a process y which conditional upon S_1 is given by

$$y(t) = \int_0^t \log \frac{f}{\hat{g}}(\hat{G}^{-1}(u))du - tK(f_0\|f) + (n-N)^{-1/2}B(t),\ t \in [0,1]. \tag{25}$$

where $\hat{G}$ is the distribution function corresponding to the realized value $\hat{g}$. Call the conditional distribution of $y(t)$ given by (25) $Q^2_{n,f,\hat{g}}$. Denote the distribution of S_1 as $P^1_{n,f}$; define $\tilde{Q}_{n,f}$ to be the joint distribution of $\mathbf{S}_1$ and y . Define an experiment

$$\tilde{\mathcal{Q}}_n = \left\{\tilde{Q}_{n,f},\ f\ \in \Sigma.\right\}$$

4.2 Theorem. *Let an M and τ_n as in theorem 2.1 be given, and also an estimator sequence $\hat{g}$ fulfilling (24), depending only on the sample fraction $S_1 = (X_1, \ldots, X_N)$, $N = N(n)$, $n = 1, 2, \ldots$, and where $n - N \asymp n$. Then*

$$\Delta(\mathcal{P}_n, \tilde{\mathcal{Q}}_n) \longrightarrow 0 \ \text{ as } n \longrightarrow \infty.$$

Proof. Define $S_2 = (X_{N+1}, \ldots, X_n)$ as the second fraction of the sample, and let its distribution be $P^2_{n,f}$. We claim that there is a probability space $(\Omega_2, \mathcal{B}_2, \mathbf{P}_2)$ independent of n, and on it independent uniform random variables $Z^*_{in}, i = N+1, \ldots, n$, a sequence of Brownian Bridges $\mathbf{B}^*_n, n = 1, 2, \ldots$, such that if $\mathbf{U}^*_n$ denotes the empirical process of Z^*_{in}, $i = N+1, \ldots, n$ then $\mathbf{U}^*_n$ and $\mathbf{B}^*_n$ satisfy relation (7) with n replaced by $n-N$. Indeed this also follows from theorem 2 in Shorack, Wellner (1986), chap. 12, section 1; it suffices to consider X_i, $i = N+1, \ldots, n$ as the first fraction of a sample of size n. We may then identify as usual $X_i = F^{-1}(Z^*_{in})$, but also define $y(t)$ via the Brownian bridge $\mathbf{B}^*_n$, by taking $\mathbf{B}^*_n$ for B in (25).
Observe that given $g \in \Sigma$, the distribution $Q^2_{n,f,g}$ is absolutely continuous with respect to the distribution of pure noise, i. e. of $(n-N)^{-1/2}B$. This distribution may be written $Q^2_{n,g,g}$. Define then densities on $(\Omega_2, \mathcal{B}_2, \mathbf{P}_2)$

$$q^{2*}_{n,f,g}(\omega_2) = \frac{dQ^2_{n,f,g}}{dQ^2_{n,g,g}}(y(\omega_2))$$

where the random variable y has distribution $Q^2_{n,g,g}$. It was already argued in proposition 2.2 that $\left\{q^{2*}_{n,f,g}d\mathbf{P}_2, f \in \Sigma\right\}$ is equivalent to $\left\{Q^2_{n,f,g}\ ,\ f \in \Sigma\right\}$ (since the likelihood ratio distributions coincide). Analogously define for a random variable S_2 having distribution $P^2_{n,g}$

$$p^{2*}_{n,f,g}(\omega_2) = \frac{dP^2_{n,f}}{dP^2_{n,g}}(S_2(\omega_2)).$$

Then $\left\{p^{2*}_{n,f,g}d\mathbf{P}_2,\ f \in \Sigma\right\}$ is equivalent to $\left\{P^2_{n,f},\ f \in \Sigma\right\}$. Also in the local case (proposition 2.4) it was argued that if $f \in \Sigma_n(g)$ then

$$\int \mid q^{2*}_{n,f,g} - p^{2*}_{n,f,g} \mid d\mathbf{P}_2 \longrightarrow 0\ ,\ \text{as}\ n \longrightarrow \infty \tag{26}$$

uniformly over $f \in \Sigma_n(g)$ and $g \in \Sigma_0$. (It was shown for sample size n; but since $n - N \asymp n$, the argument remains valid for the now relevant diminished sample size). Hence (26) holds also uniformly over $g \in \Sigma_n(f)$ and $f \in \Sigma$. Now model the distribution of $(Z_1, \ldots, Z_N)$ on a probability space $(\Omega_1, \mathcal{B}_1, \mathbf{P}_1)$ such that S_1 has density $p^{1*}_{n,f}$, and take $(\Omega, \mathcal{B}, \mathbf{P})$ as the product of $(\Omega_1, \mathcal{B}_1, \mathbf{P}_1)$ and $(\Omega_2, \mathcal{B}_2, \mathbf{P}_2)$. On $(\Omega, \mathcal{B}, \mathbf{P})$ define a $\mathbf{P}$-density for $g = \hat{g}(\omega_1)$

$$q^*_{n,f} = q^{2*}_{n,f,g}\ p^{1*}_{n,f}$$

let $Q^*_{n,f}$ be the corresponding measure and $\mathcal{Q}^*_n = \left\{\ Q^*_{n,f},\ f \in \Sigma\ \right\}$.

4.3 Lemma. *The experiment $\mathcal{Q}^*_n$ is of the same type as $\tilde{\mathcal{Q}}_n$.*

Proof. Let $P^1_{n,1}$ be the distribution of S_1 under the uniform density $\mathbf{1}$; and $Q^2_{n,1,1}$ be the distribution of pure noise in (25). Then $\tilde{Q}_{n,f}$ is absolutely continuous with respect to $P^1_{n,1} \otimes Q^2_{n,1,1}$, with density $\tilde{q}_{n,f}$, say. We already noted that $Q^2_{n,1,1} = Q^2_{n,g,g}$; hence

$$\tilde{q}_{n,f}(S_1, y) = \frac{dP^1_{n,f}}{dP^1_{n,1}}(S_1)\ \frac{dQ^2_{n,f,g(S_1)}}{dQ^2_{n,g(S_1),g(S_1)}}(y).$$

Now $q^*_{n,f}(\omega)$ is obtained from this density by plugging in $S_1(\omega_1) = (Z_1(\omega_1), \ldots, Z_N(\omega_1))$ and $y(\omega_2) = (n - N)^{-1/2}\mathbf{B}^*_n(\omega_2)$. The factorization criterion then says that the map $\omega \longmapsto (S_1(\omega_1), y(\omega_2))$ is a sufficient statistic in the experiment $\mathcal{Q}^*_n$, and $\tilde{\mathcal{Q}}_n$ is constituted by the distributions of this sufficient statistic. Thus $\tilde{\mathcal{Q}}_n$ and $\mathcal{Q}^*_n$ are equivalent. □

Having found a convenient representation of $\tilde{\mathcal{Q}}_n$, we will now complement it by one for $\mathcal{P}_n$ on the same probability space. Define a $\mathbf{P}$ - density

$$p^*_{n,f} = p^{2*}_{n,f,g}p^{1*}_{n,f}\ ,$$

a corresponding measure $P^*_{n,f}$ and an experiment $\mathcal{P}^*_n = \left\{P^*_{n,f},\ f \in \Sigma\right\}$.

4.4 Lemma. *The experiment $\mathcal{P}_n^*$ is of the same type as $\mathcal{P}_n$.*

Proof. The experiment $\mathcal{P}_n$ is dominated by the Lebesgue measure, which may be written $P_{n,1}^1 \otimes P_{n,1}^2$. But another dominating measure is $P_{n,1}^1 \otimes P_{n,g(S_1)}^2$, indeed this measure has Lebesgue density $\prod_{i=N+1}^n \hat{g}(x_1,\ldots,x_N)(x_i)$ which under our assumptions is positive and bounded. (Here the notation $\hat{g}(x_1,\ldots,x_N)(x_i)$ signifies dependence of the estimator $\hat{g}$ on $S_1 = (x_1,\ldots,x_N)$ and its dependence as a density on an argument x_i, $N+1 \le i \le n$.) Now the density of $P_{n,f} \in \mathcal{P}_n$ with respect to the measure $P_{n,1}^1 \otimes P_{n,g(S_1)}^2$ may be written

$$p_{n,f}(S_1,S_2) = \frac{dP_{n,f}^1}{dP_{n,1}^1}(S_1)\ \frac{dP_{n,f}^2}{P_{n,g(S_1)}^2}(S_2).$$

Now $p_{n,f}^*(\omega)$ is obtained from this density by plugging in $S_1(\omega_1) = (Z_1(\omega_1),\ldots,Z_N(\omega_1))$ and

$$S_2(\omega_1,\omega_2) = \left(\hat{G}_{S_1(\omega_1)}^{-1}(Z_{N+1}(\omega_2)),\ldots,\hat{G}_{S_1(\omega_1)}^{-1}(Z_n(\omega_2))\right).$$

The factorization criterion then says that the map $\omega \longmapsto (S_1(\omega_1), S_2(\omega_1,\omega_2))$ is a sufficient statistic in the experiment $\mathcal{P}_n^*$, and $\mathcal{P}_n$ is constituted by the distributions of this sufficient statistic. Thus $\mathcal{P}_n$ and $\mathcal{P}_n^*$ are equivalent. □

Theorem 4.2 then follows from

4.5 Lemma. *We have*

$$\int |q_{n,f}^* - p_{n,f}^*| d\mathbf{P} \longrightarrow 0 \qquad as\ n \longrightarrow \infty$$

uniformly over $f \in \Sigma$.

Proof. Define an event

$$A = \{\omega_1 \in \Omega_1 \,/\, \hat{g}(\omega_1) \in \Sigma_n(f)\}$$

Then

$$\int |q_{n,f}^* - p_{n,f}^*| d\mathbf{P} = \int |p_{n,f}^{1*} q_{n,f,g}^{2*} - p_{n,f}^{1*} p_{n,f,g}^{2*}| d\mathbf{P}$$
$$\le 2 \int_{A^c} p_{n,f}^{1*}(\omega_1) d\mathbf{P}_1(\omega_1) + \int_A \int |q_{n,f,g}^{2*} - p_{n,f,g}^{2*}| d\mathbf{P}_2\ p_{n,f}^{1*}\ d\mathbf{P}_1.$$

Here, for $\omega_1 \in A$, the inner integral in the second summand is $o(1)$ uniformly over $f \in \Sigma$, $\omega_1 \in A$ according to (26). The first summand is also $o(1)$ uniformly over $f \in \Sigma$, see (24). Hence the lemma. □

As in section 3, further white noise models may now be considered and used as global approximants. Remind that according to lemma 3.3, two white noise models with $\sigma = n^{-1}$ are asymptotically equivalent if the respective trend functions differ at $o(n^{-1/2})$ in L_2 (the Brown-Low approach). Furthermore, in the proof of the global approximation, (26) is crucial. Indeed the argument remains valid if the density $p_{n,f,g}^{2*}$, which is the likelihood ratio of (25), is replaced by another likelihood ratio also fulfilling (26). We may refer to (25) as "the second

part of the compound experiment"; this may now be replaced, on the basis of lemma 3.3, by another one.
In this sense, the Brownian bridge in (25) may at first be replaced by a Wiener process $W(t)$ (see proof of theorem 3.1; here the likelihood ratios coincide). As a next step, in (25) the term $tK(f_0\|f)$ may be omitted, since its derivative is $o(n^{-1/2})$ (see (21)). We arrive at the compound experiment with second part

$$dy(t) = \log \frac{f}{\hat{g}}(\hat{G}^{-1}(t))dt + (n-N)^{-1/2}dW(t),\ t \in [0,1]. \tag{27}$$

In the compound experiment $\hat{g}$ is observed. Hence addition of $\log \hat{g}(\hat{G}^{-1}(t))dt$ to (27) amounts to transforming the overall observations in a fixed given manner. This yields an experiment of the same type. Hence

4.6 Corollary. *The compound experiment with first part $y_1, \ldots, y_N$: i. i. d. with density f and second part*

$$dy(t) = \log f(\hat{G}^{-1}(t))dt + (n-N)^{-1/2}dW(t)\ t \in [0,1]$$

is asymptotically equivalent to $\mathcal{P}_n$.

In conjunction with remark 3.2 we obtain more generally

4.7 Corollary. *The compound experiments with first part $y_1, \ldots, y_N$: i. i. d. with density f and second parts*

$$y_i,\ i = N+1, \ldots, n \qquad \text{i. i. d. with density } f \tag{28}$$
$$dy(t) = \log f(\hat{G}^{-1}(t)) + (n-N)^{-1/2}dW(t),\ t \in [0,1] \tag{29}$$
$$dy(t) = f(t)dt + (n-N)^{-1/2}\hat{g}^{1/2}(t)dW(t),\ t \in [0,1] \tag{30}$$
$$dy(t) = f^{1/2}(t)dt + \frac{1}{2}(n-N)^{-1/2}dW(t),\ t \in [0,1] \tag{31}$$

are all asymptotically equivalent.

Proof of theorem 1.1. Note that the above corollary, in particular the result connected with (31) is not tied to the choice $N \asymp n/\log n$; it remains valid for larger N. Indeed theorem 4.2 spells out (24) as a condition; suppose we take $N = [n/2]$. Then (24) is still valid: the relevant neighborhoods $\Sigma_n(f)$ are of the same size in a rate sense (since $n - N \asymp n$), and the preliminary estimator only gets better. On the resulting compound experiment we may then operate again, reversing the roles of first and second part. Indeed first part $y_1, \ldots, y_N$ and second part (31) are independent, so we may in turn substitute $y_1, \ldots, y_N$ by a white noise model, using a preliminary estimator based on (31). The existence proof for such an estimator fulfilling (24) is entirely analogous to section 6; here we exploit the well known parallelism of density estimation and white noise on the rate of convergence level. Only we now have a white noise model on $f^{1/2}$ rather than on f, so a few things have to be taken

care of. Observe that for densities in $\mathcal{F}_{\geq\epsilon}$, $\epsilon > 0$, the seminorms $\left\|(f^{1/2})^{(m)}\right\|_2$ and $\left\|f^{(m)}\right\|_2$ are equivalent, and

$$\left\|\hat{f} - f\right\|_2 \leq \left\|\hat{f}^{1/2} - f^{1/2}\right\|_2 \left\|\hat{f}^{1/2} + f^{1/2}\right\|_\infty \leq C \left\|\hat{f}^{1/2} - f^{1/2}\right\|_2$$

(for estimators $\hat{f}^{1/2}$ of $f^{1/2}$ taking values in $W_2^m(M') \cap \mathcal{F}_{\geq\epsilon}$). Also

$$\left\|\hat{f}' - f'\right\|_2 = 2\left\|\left(\hat{f}^{1/2}\right)' - \left(f^{1/2}\right)'\right\|_2,$$

so that it is clear that from a white noise model with trend $f^{1/2}$, $f \in \Sigma$, the function f can be estimated with a rate as from trend f. Thus substituting $y_1, \ldots, y_N$ by white noise leads to an experiment with observations

$$dy_1(t) = f^{1/2}(t)dt + \frac{1}{2}N^{-1/2}dW(t),\ t \in [0,1] \tag{32}$$

$$dy_2(t) = f^{1/2}(t)dt + \frac{1}{2}(n-N)^{-1/2}dW(t),\ t \in [0,1]. \tag{33}$$

A sufficiency argument shows this equivalent to n i. i. d. processes with variance $1/4$, which in turn is equivalent to (3). □

5 An application: exact constants for L_2-risk

Let $\mathcal{F} \subset \Sigma$ be any a priori set for the density f, and L be a bounded loss function in an estimation problem:

$$L(g,f) \leq T \quad \text{for} \quad f \in \mathcal{F} \qquad \text{and for all possible estimator values g.}$$

Let $\mathcal{P}_n$ be the density experiment with full parameter space Σ, and $\rho_n(L,\mathcal{F})$ be the minimax risk there for restricted parameter space $\mathcal{F}$ and loss function L. Let $\mathcal{Q}_n$ be another experiment with parameter space Σ, and let $\rho_n^*(L,\mathcal{F})$ be the analogous minimax risk there.

5.1 Proposition. *Let L_n be a uniformly bounded sequence of loss functions. Suppose that $\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0$. Then for any sequence of parameter spaces $\mathcal{F}_n \subset \Sigma$ the minimax risks fulfill*

$$\rho_n^*(L_n, \mathcal{F}_n) - \rho_n(L_n, \mathcal{F}_n) \to 0.$$

This is an immediate consequence of the characterization of the deficiency in terms of risk functions (see Le Cam and Yang (1990)). The interesting case is the one where L_n is derived from a normalized truncated loss function such as

$$L_{n,T}(\hat{f}, f) = \left(n^{1-r}\|\hat{f} - f\|_2^2\right) \wedge T, \tag{34}$$

where n^{r-1} is the optimal rate of convergence and T a truncation constant. By a suitable limit argument in which $T \to \infty$ after n, it will be possible to recover even the optimal constants in the known risk asymptotics for the nice nontruncated loss functions.

As a first exercise let us try to deduce the results of Efroimovich and Pinsker (1982) from the white noise approximation. Let $\varphi_j(x) = \sqrt{2}cos(2\pi jx)$, $j \geq 1$, $\varphi_j(x) = \sqrt{2}sin(2\pi jx)$, $j \leq 1$, $\varphi_0 \equiv 1$ be the classical Fourier basis, and $f_j = \langle f, \varphi_j \rangle$ be the Fourier coefficients of f. Consider a periodic Sobolev class

$$\tilde{W}_2^m(P) = \left\{ f,\ f_0 = 1,\ \sum_j (2\pi j)^{2m} f_j^2 \leq P \right\}.$$

and the set of all densities in it:

$$\mathcal{F}(m, P) = \tilde{W}_2^m(P) \cap \mathcal{F}_{\geq 0}.$$

We begin by stating Pinsker's (1980) minimax risk bound in the white noise model, for the full nontruncated L_2-loss function $L_{n,\infty}$. Let $\mathcal{Q}_n(\mathcal{F})$ be the experiment given by the distributions of y in a model (13) where $f \in F$ and $f_0 = \mathbf{1}$ (the uniform density). Let $L_{n,T}$ be given by (34). The ball in L_2 with center f and radius c will be denoted by $B(c, f)$.

5.2 Proposition. *For $r = \frac{1}{2m+1}$ and any sequence $\tau_n^* \to \infty$ the relation*

$$\rho_n^*(L_{n,\infty}, \tilde{W}_2^m(P) \cap B(\tau_n^* n^{(r-1)/2}, \mathbf{1})) \to \gamma(m) P^r,$$

holds, where $\gamma(m) = (2m+1)^r \left(\frac{m}{\pi(m+1)}\right)^{1-r}$ is the Pinsker constant.

The original result was stated as a nonlocal one, i. e. it referred to parameter space $\tilde{W}_2^m(P)$; but it is easily seen to be valid for the shrinking L_2-balls: see e. g. Golubev and Nussbaum (1990), section 4.
In this statement, $L_{n,\infty}$ may be substituted by L_{n,T_n} if T_n is an appropriate sequence tending to infinity. Indeed estimators may be assumed to take values in the ball $B(\tau_n^* n^{(r-1)/2}, f_0)$, whereupon $n^{1-r}\|\hat{f} - f\|_2^2 \leq 4\tau_n^{*2}$, so that any choice $T_n \geq 4\tau_n^{*2}$ is possible. As τ_n^* may grow arbitrarily slowly, we now have a sequence L_{n,T_n} in which $T_n \to \infty$ arbitrarily slowly. In conjunction with proposition 5.1 this already allows to state a risk convergence in the density model. We first use the local equivalence (remark 3.2); as it refers to local neighborhoods $\Sigma_n(f_0)$ the center of which is now known, we see that it is appropriate for *lower* asymptotic risk bounds. It remains to verify that $\Sigma_n(\mathbf{1})$ contains a parameter space as in proposition 5.2.

5.3 Lemma. *Let $m \geq 4$, and let the sequence τ_n in $\Sigma_n(f_0)$ be given. Then τ_n^* may be chosen such that for any n*

$$\tilde{W}_2^m(P) \cap B(\tau_n^* n^{(r-1)/2}, \mathbf{1}) \subset \Sigma_n(\mathbf{1})$$

Proof. First we demonstrate that for $f \in \tilde{W}_2^m(P) \cap B(\tau_n^* n^{(r-1)/2}, f_0)$ we have

$$\|f' - f_0'\|_2 = O\left(n^{-(m-1)/(2m+1)} \tau_n^{**}\right) \tag{35}$$

for $\tau_n^{**} = (\tau_n^*)^{1-1/m}$. We use the multiplicative imbedding inequality

$$\|f'\|_2 \le C_m \|f\|_2^{1-1/m} \left\|f^{(m)}\right\|_2^{1/m}$$

which implies

$$\|f' - f_0'\|_2 \le C_m \left(2\tau_n^* n^{(r-1)/2}\right)^{1-1/m} \left\|f^{(m)} - f_0^{(m)}\right\|_2^{1/m} \le$$

$$\le C_m (2\tau_n^*)^{1-1/m} n^{-(m-1)/(2m+1)} 2P^{1/2m}$$

so that (35) is proved. The remainder of the proof is now entirely analogous to the argument in section 6 (for the preliminary estimator $\hat{g}$ to be in $\Sigma_n(f)$ with high probability). Indeed (35) and the relation $\|f - f_0\|_2 = O(\tau_n^* n^{-m/(2m+1)})$ can be used to show that $f \in \Sigma_n(f_0)$ if τ_n^* is chosen to grow sufficiently slowly. To treat the logarithms, we may assume as in lemma 6.2 that $f, f_0 \in \mathcal{F}_{\ge\epsilon}$, since $f_0 = \mathbf{1}$ and (48) then implies $f \in \mathcal{F}_{\ge\epsilon}$ eventually.□

The lemma implies that functions in $\tilde{W}_2^m(P) \cap B(\tau_n^* n^{(r-1)/2}, \mathbf{1})$ are eventually positive and hence densities. This implies a lower risk bound in the density problem:

5.4 Proposition. *There is a sequence $T_n \to \infty$ such that in the density problem, for $m \ge 4$*

$$\liminf_n \ \rho_n(L_{n,T_n}, \mathcal{F}(m,P)) \ge \gamma(m) P^r$$

For the converse upper bound we shall invoke the global result of corollary 4.7. Take the model (30) and look what risk bounds are attainable there by linear estimators based on empirical Fourier coefficients. If $\hat{f} = \sum_j c_j \hat{f}_j \varphi_j$ where $\hat{f}_j = \int \varphi_j dy$ then

$$E\left\|\hat{f} - f\right\|_2^2 = \sum_j (1-c_j)^2 f_j^2 + (n-N)^{-1} \sum_j c_j^2 \int \varphi_j^2 \hat{g}.$$

(Note that $\hat{f}_j$ are not independent here if $\hat{g}$ is not the uniform density.) Observe that $\int(\varphi_j^2 + \varphi_{-j}^2)\hat{g} = 2\int \hat{g} = 2$. Hence if $c_j = c_{-j}$ we have

$$E\left\|\hat{f} - f\right\|_2^2 = \sum_j (1-c_j)^2 f_j^2 + (n-N)^{-1} \sum_j c_j^2 = R_n(c,f), \tag{36}$$

say. Thus we are essentially in the case of uniform variance function ($\hat{g} = \mathbf{1}$) provided we use estimators with $c_j = c_{-j}$. That is no real restriction if the parameter space fulfills a related kind of symmetry. Here is a formal argument.

Identify a set $\mathcal{F} \subset L_2(0,1)$ with its sequence of Fourier coefficients under the usual isometry, and for any i, let $\mathcal{G}_i$ be the group of transformation which act as follows: for any sequence f, the coefficients at level $|i|$ are permuted (i. e. f_i appears in place of f_{-i} and vice versa). Call $\mathcal{F}$ *level symmetric* if $\mathcal{F}$ is invariant under all $\mathcal{G}_i$, i natural. Call a sequence f level symmetric if the one element set $\{f\}$ is level symmetric, i. e. if $f_i = f_{-i}$, all i.

Following Donoho, Liu and MacGibbon (1990) (abbreviated DLM henceforth) we call $\mathcal{F}$ *orthosymmetric* if $f \in \mathcal{F}$ entails $\{\pm f_i\} \in \mathcal{F}$ for all possible combinations of $+$ and $-$.

5.5 Lemma. *Assume the set $\mathcal{F} \subset l_2$ is compact, orthosymmetric and level symmetric. Then for $R_n(c,f)$ given by (36) we have*

$$\inf_c \sup_{f\in\mathcal{F}} R_n(c,f) = \inf_{c \text{ level symmetric}} \sup_{f\in\mathcal{F}} R_n(c,f). \tag{37}$$

Proof. For any sequence $f \in l_2$ let f^* be the sequence $\{f_{-i}\}$ and $f^{(2)}$ be the sequence $\{f_i^2\}$. DLM in their theorem 11 show that if $\mathcal{F}$ is orthosymmetric and compact then the l. h. s. of (37) remains unchanged if $\mathcal{F}$ is replaced by its quadratically convex hull $Q\mathcal{F}$ (i. e. the set of f such that $f^{(2)}$ is in the convex hull of $g^{(2)}$, $g \in \mathcal{F}$). Hence

$$\inf_c \sup_{f\in Q\mathcal{F}} R_n(c,f) = \inf_c \sup_{f\in\mathcal{F}} R_n(c,f) \le \inf_{c \text{ l. s.}} \sup_{f\in\mathcal{F}} R_n(c,f). \tag{38}$$

For $f \in \mathcal{F}$ let $\bar{f}$ be a level symmetric sequence fulfilling $\bar{f} = \frac{1}{2}(f^{(2)} + f^{*(2)})$. For any level symmetric c we have $R_n(c,f) = R_n(c,\bar{f})$. Moreover, if $\mathcal{F}$ is level symmetric and compact then $f \in \mathcal{F}$ entails $f^* \in \mathcal{F}$, so that $\bar{f} \in Q\mathcal{F}$. Hence (38) may be continued by

$$= \inf_{c \text{ l. s.}} \sup_{f\in\mathcal{F}} R_n(c,\bar{f}) \le \inf_{c \text{ l. s.}} \sup_{f\in Q\mathcal{F},\, f \text{ l. s.}} R_n(c,f). \tag{39}$$

Clearly the abovementioned theorem 11 of DLM also applies to the r. h. s. of (39) and guarantees that $Q\mathcal{F}$ may be substituted there by $\mathcal{F}$. Furthermore, for any sequence c, let $\tilde{c}$ be the level symmetric sequence $\tilde{c} = \frac{1}{2}(c + c^*)$. Any level symmetric c may be represented as $\tilde{c}_0$ for some general c_0. Observe also that $R_n(c,f)$ is convex in c. Hence (39) may be continued by

$$= \inf_{c \text{ l. s.}} \sup_{f\in\mathcal{F},\, f \text{ l. s.}} R_n(c,f) = \tag{40}$$

$$= \inf_c \sup_{f\in\mathcal{F},\, f \text{ l. s.}} R_n(\tilde{c},f) \le \inf_c \sup_{f\in\mathcal{F},\, f \text{ l. s.}} \frac{1}{2}\left(R_n(c,f) + R_n(c^*,f)\right) = \tag{41}$$

$$= \inf_c \sup_{f\in\mathcal{F},\, f \text{ l. s.}} R_n(c,f) \le \inf_c \sup_{f\in Q\mathcal{F}} R_n(c,f) \tag{42}$$

The chain (38)-(42) shows that (38) is an equality. □

DLM also show that under the conditions of the lemma $R_n(c,f)$ coincides with the minimax linear risk, i. e. with the minimax risk for l_2-loss over *all linear* estimators.
For attainabiltiy in the white noise model we may disregard the restriction to f which are densities. Then evidently our periodic Sobolev class $\tilde{W}_2^m(P)$ fulfills all conditions of lemma 5.5. Hence the minimax linear risk, for nontruncated L_2-loss, in the model (30) with realized $\hat{g} \in \Sigma^*$ coincides with the one in the case $\hat{g} \equiv 1$. This latter one is the standard case covered by the original result of Pinsker (1980), where the asymptotics of the minimax linear risk is well known. This is thus an attainable risk in the full compound experiment; corollary 4.7 then implies

5.6 Proposition. *There is a sequence $T_n \to \infty$ such that in the density problem, for $m \geq 4$* and any $\epsilon > 0$

$$\rho_n(L_{n,T_n}, \mathcal{F}(m,P) \cap \mathcal{F}_{\geq \epsilon}) \to \gamma(m) P^r \quad as \quad n \to \infty.$$

Note that the lower bound of proposition 5.4 holds also over densities in $\mathcal{F}_{\geq \epsilon}$, in view of lemma 5.3.

LINEAR WAVELET ESTIMATORS

The principle that for level symmetric sets and L_2-loss the case of heteroskedastic white noise may be reduced to the homoskedastic one may be extended. Above it was developed for linear Fourier series estimators; an extension to linear wavelet estimators is now straightforward. Although recent developments in Wavelet estimation mostly concern *nonlinear* estimators (Donoho, Johnstone (1992)), still linear estimators are efficient in a number of cases.
Suppose a doubly indexed orthonormal wavelet basis of $L_2(0,1)$ be given: ϕ_0, ψ_{ij}, $i = 0, 1, \ldots$, $j = 1, \ldots 2^i$, with corresponding wavelet expansion of a function f:

$$f = f_0 \phi_0 + \sum_{i=0}^{\infty} \sum_{j=1}^{2^i} f_{ij} \psi_{ij}$$

Here the *i-th level* is naturally defined as the set of coefficients for given i. Suppose an a priori class for f is formulated in terms of wavelet coefficients. Such a class may be called level symmetric if all coefficients of the same level are treated the same way, or more formally, if the set is invariant under all groups $\mathcal{G}_i$ which permute the coefficients of level i. The Besov smoothness classes are level symmetric:

$$B^s_{p,q}(P) = \left\{ f, \left(\sum_{i=0}^{\infty} \left(2^{js} \left(\sum_{j=1}^{2^i} f_{ij}^p \right)^{1/p} \right)^q \right)^{1/q} \leq P \right\}$$

The background is that a restriction which defines smoothness should force the coefficients of high frequencies to be small; but in wavelet analysis all coefficient of a given level i correspond to the same frequency. Now identify the double arrays $\{f_{ij}\}$ with the sequence space l_2; *then lemma 5.5 above also holds in the wavelet framework.* Since the Besov class is also orthosymmetric, and compact if the restriction $f_0 = 1$ is added, the minimax linear risk is attained by level symmetric smoothing coefficients c. Thus if $\hat{f} = \phi_0 + \sum_{i,j} c_{ij} f_{ij} \psi_{ij}$ where c_{ij} does not depend on j then in the model (30) we have

$$E \left\| \hat{f} - f \right\|_2^2 = o(n^{-1}) + \sum_{i,\,j} (1 - c_{ij})^2 f_{ij}^2 + (n-N)^{-1} \sum_{i,\,j} c_{ij}^2 \int \psi_{ij}^2 \hat{g}.$$

$$= o(n^{-1}) + \sum_{i,\,j} (1 - c_{i\cdot})^2 f_{ij}^2 + (n-N)^{-1} \sum_{i} c_{i\cdot}^2 \sum_{j=1}^{2^i} \int \psi_{ij}^2 \hat{g}.$$

Here $\psi_{ij}^2(x) = 2^i \psi^2(2^i x - j)$ where ψ is the mother wavelet; hence

(43) $$\sum_{j=1}^{2^i} \int \psi_{ij}^2 \hat{g} \approx \sum_{j=1}^{2^i} \hat{g}(j2^{-i}) \approx 2^i \int \hat{g} = 2^i$$

for large i. (Here we did simplify slightly about the wavelet basis, but (43) is still true if we use the correct version for $[0,1]$ given by Meyer (1992)). Consequently, in nonparametric settings where the relevant i are large

$$E \left\| \hat{f} - f \right\|_2^2 = o(n^{-1}) + \sum_{i,\, j} (1 - c_{i\cdot})^2 f_{ij}^2 + (n-N)^{-1} \sum_{i,\, j} c_{i\cdot}^2$$

which means that the same risk as for $\hat{g} \equiv 1$ is attainable. So indeed we have the same phenomenon as for linear Fourier series estimators; the heteroskedasticity of the noise does not influence the L_2-risk of linear estimators.
For linear wavelet estimators of a density see also Kerkyacharian and Picard (1992).

EXACT CONSTANTS FOR HELLINGER LOSS

The basic white noise approximation of theorem 1.1 in conjunction with the result on the L_2-risk for the density over Sobolev classes discussed above immediately suggests a result on the exact asymptotics of the Hellinger risk. Consider an a priori class

$$\mathcal{F}^H(m,P) = \left\{ f,\ f \text{ a density},\ f^{1/2} \in W_2^m(P) \right\}.$$

Define a truncated (squared) Hellinger loss as

$$L_{n,T}^H(\hat{f}, f) = \left(n^{1-r} \| \hat{f}^{1/2} - f^{1/2} \|_2^2 \right) \wedge T,$$

where $r = 1/(2m+1)$, and let ρ_n as before be the minimax risk in the i. i. d. model.

5.6 Proposition. *There is a sequence $T_n \to \infty$ such that in the density problem, for $m \geq 4$ and any $\epsilon > 0$*

$$\rho_n(L_{n,T_n}^H, \mathcal{F}^H(m,P) \cap \mathcal{F}_{\geq \epsilon}) \to 2^{2(r-1)} \gamma(m) P^r \quad as \quad n \to \infty.$$

Proof. The attainability of the bound is obvious, in view of theorem 1.1, and our previous reasoning on how to carry over the Pinsker bound to the density model. Also, it is to be noted that the variance σ^2 of the white noise (which in our case is $1/4$) appears with an exponent $1-r$ in the risk asymptotics. For the lower bound we have to take into account that $f^{1/2}$ is now restricted to the unit sphere in L_2. For this we use proposition 5.2 and restrict $f^{1/2}$ to a shrinking ball $B(\tau_n^* n^{(r-1)/2}, \mathbf{1})$ around the uniform density $\mathbf{1}$. Let then $\Pi(f^{1/2})$ be the projection of $f^{1/2}$ to the tangent space of the unit sphere in $\mathbf{1}$. Then obviously

(44) $$\left\| f^{1/2} - \Pi(f^{1/2}) \right\| \leq (\tau_n^* n^{(r-1)/2})^2$$

uniformly over $f^{1/2} \in B(\tau_n^* n^{(r-1)/2}, \mathbf{1})$. Now $n^{r-1} = n^{-2m/(2m+1)}$; thus the r. h. s. of (44) is $o(n^{-1/2})$ for $m > 1/2$ and τ_n^* growing not too fast. We may then apply lemma 3.3 for $\sigma^2 = n^{-1}$, to show that in the white noise model the trend $f^{1/2}$ may be substituted by

$\Pi(f^{1/2})$, with asympotic equivalence of the experiments. Then $\Pi(f^{1/2})$ varies in an affine subspace of L_2, and since it can be represented $\Pi(f^{1/2}) = f^{1/2} + c_f \mathbf{1}$ for some number c_f, its m-th derivative coincides with that of $f^{1/2}$. Hence $h = \Pi(f^{1/2})$ varies fully within $W_2^m(P) \cap B(\tau_n^* n^{(r-1)/2}, \mathbf{1})$, subject only to the affine restriction $(h, \mathbf{1}) = 1$. But that is essentially the case covered by proposition 5.2, so we may infer the lower bound in the white noise model. It is then carried over to the density model as before. □

6 The preliminary estimator

Recall that the a priori set Σ was defined as

$$\Sigma = \left\{ f;\ \int f = 1,\ f(x) \geq \epsilon,\ x \in [0,1],\ f \in W_2^m(M) \right\}$$

where we did impose the smoothness condition $m = 4$. For the sake of clarity we will retain here the general m in the notation. First let us recap a standard type result on attainable rates for the density f itself, based on the whole sample $X_1, \ldots, X_n$.

6.1 Lemma. *Let the density f be in a Sobolev class $W_2^m(M)$, $m \geq 1$. Then there exists an estimator $\tilde{f}_n$, which almost surely is a function from $W_2^m(M)$, such that*

$$\limsup_n \sup_{f \in W_2^m(M)} P\left(n^{m/(2m+1)} \left\| \tilde{f}_n - f \right\|_2 + n^{(m-1)/(2m+1)} \left\| \tilde{f}_n' - f' \right\|_2 \geq t \right) \to 0,\ t \to \infty$$

Proof. Suppose first that f fulfills periodicity conditions, and use a truncated Fourier series estimator. Form empirical Fourier coefficients $\hat{f}_j = n^{-1} \sum_{i=1}^n \varphi_j(x_i)$. Then

$$E\hat{f}_j = f_j,\ Var\ \hat{f}_j = n^{-1} \left(\int \varphi_j^2 f - f_j^2 \right) \leq n^{-1} \|f\|_\infty .$$

Define $k = \left[n^{m/(2m+1)} \right]$ and the estimator $\sum_{|j| \leq k} \hat{f}_j \varphi_j$. Then, with the usual bias/variance decomposition,

$$E \left\| \tilde{f}_n - f \right\|_2^2 = E \sum_{-\infty}^{\infty} \left(\hat{f}_j - f_j \right)^2 \leq \sum_{|j| \geq k} f_j^2 + n^{-1}(2k+1) \|f\|_\infty ,$$

$$E \left\| \tilde{f}_n' - f' \right\|_2^2 = E \sum_{-\infty}^{\infty} (2\pi j)^2 \left(\hat{f}_j - f_j \right)^2 \leq \sum_{|j| \geq k} (2\pi j)^2 f_j^2 + n^{-1} \sum_{|j| \leq k} (2\pi j)^2 \|f\|_\infty$$

Obviously $\sum_{|j| \leq k} (2\pi j)^2 \leq ck^3$. Furthermore, by imbedding inequalities, we have $\|f\|_\infty \leq c$, say, on $W_2^m(M)$. The equality

$$\left\| f^{(m)} \right\|_2^2 = \sum_j (2\pi j)^{2m} f_j^2 \tag{45}$$

implies that for $f \in W_2^m(M)$

$$\sum_{|j| \geq k} f_j^2 \leq (2\pi k)^{-2m} M,\ \sum_{|j| \geq k} (2\pi j)^2 f_j^2 \leq (2\pi k)^{-2(m-1)} M.$$

Consequently

$$E\left\|\tilde{f}_n - f\right\|_2^2 \leq C\left(k^{-2m} + n^{-1}k\right),$$

$$E\left\|\tilde{f}_n' - f'\right\|_2^2 \leq C\left(k^{-2(m-1)} + n^{-1}k^3\right) = Ck^2\left(k^{-2m} + n^{-1}k\right).$$

Now the usual argument to choose $k \asymp n^{-1}k$ is seen to yield $k \asymp n^{-1/(2m+1)}$ for both risks, with respective optimal rates $n^{-2m/(2m+1)}$, $n^{-2(m-1)/(2m+1)}$. Chebyshev's inequality then implies the assertion for the periodic case.
For the class $W_2^m(M)$ without periodic boundary conditions, it suffices to replace the Fourier basis with another suitable basis. One possibility is to use Wavelets, e. g. the orthonormal basis in $L_2(0,1)$ of Meyer (1991) (see Donoho, Johnstone (1992) for a statistical context). The equality (45) which expresses $\left\|f^{(m)}\right\|_2$ in terms of Fourier coefficients is then to be replaced by the equivalence of this seminorm with the corresponding seminorm in a Besov space $B_{2,2}^m$, which in turn can be expressed (up to equivalence) in terms of Wavelet coefficients.□

Introduce a family of norms $\nu(\cdot, a, b)$

$$\nu^2(f, a, b) = a^2 \|f\|_2^2 + b^2 \|f'\|_2^2.$$

These are all Hilbertian norms for the Sobolev space W_2^1. Consider sequences
$\gamma_{1n} = n^{m/(2m+1)}$, $\gamma_{2n} = n^{(m-1)/(2m+1)}$. The statement of lemma 6.1 is then equivalent to

$$\nu(\tilde{f}_n - f, \gamma_{1n}, \gamma_{2n}) = O_P(1), \text{ uniformly in } f \in W_2^m(M). \tag{46}$$

Define now $\hat{f}$ to be the projection with respect to $\nu(\cdot, \gamma_{1n}, \gamma_{2n})$ of $\tilde{f}$ onto the convex set $\Sigma \subset W_2^1$. Then

$$\nu(\hat{f} - f, \gamma_{1n}, \gamma_{2n}) = \nu(\tilde{f} - f, \gamma_{1n}, \gamma_{2n}) \tag{47}$$

which implies that (46) is valid for $\hat{f}$ in place of $\tilde{f}$, and $\hat{f}$ is now a bona fide probability density taking values in Σ.
The neighborhood $\Sigma_n(f_0)$ involves involves the sup-norm. To avoid the technical argument connected with the correct rate of convergence involving a log term, we shall simply invoke a multiplicative Sobolev imbedding inequality:

$$\|f\|_\infty \leq C \|f\|_2^{1/2} \|f'\|_2^{1/2} \tag{48}$$

(a special case of general multiplicative imbedding inequalities, see Donoho and Liu (1991) for references). Define another family of norms

$$\nu^*(f, a, b) = a \|f\|_\infty + b \|f'\|_1$$

and a sequence $\gamma_{1n}^* = \gamma_{1n}^{1/2} \gamma_{2n}^{1/2} = n^{(m-1/2)/(2m+1)}$. It immediately follows from (46)-(48) and from $\|f'\|_1 \leq \|f'\|_2$ that

$$\nu^*(\hat{f} - f, \gamma_{1n}^*, \gamma_{2n}) = O_P(1), \text{ uniformly in } f \in W_2^m(M). \tag{49}$$

Now form $\log \hat{f}$; the next step is to prove (49) for the log-densities.

6.2 Lemma. *We have*

$$\nu^*(\log \hat{f} - \log f, \gamma_{1n}^*, \gamma_{2n}) = O_P(1), \quad \text{uniformly in } f \in \Sigma.$$

Proof. Since $f \in \Sigma$, $\hat{f} \in \Sigma$, we have

$$\left|\log \hat{f}(t) - \log f(t)\right| \leq \left|\hat{f}(t) - f(t)\right| \max\left(\hat{f}^{-1}(t), f^{-1}(t)\right) \leq \left|\hat{f}(t) - f(t)\right| M.$$

Furthermore

$$\left|(\log \hat{f})'(t) - (\log f)'(t)\right| = \left|\frac{\hat{f}'}{\hat{f}}(t) - \frac{f'}{f}(t)\right| \leq$$

$$\left|\frac{\hat{f}' - f'}{\hat{f}}(t)\right| + \left|\frac{(\hat{f} - f)f'}{\hat{f}\, f}(t)\right| \leq \left|(\hat{f}' - f')(t)\right| M + \left|(\hat{f} - f)(t)\right| M^2 \left\|f'\right\|_\infty .$$

By an imbedding inequality, we have $\|f'\|_\infty \leq CM$ on $W_2^m(M)$. Consequently

$$\nu^*(\log \hat{f} - \log f, \gamma_{1n}^*, \gamma_{2n}) \leq \gamma_{1n}^* M \left\|\hat{f} - f\right\|_\infty + \gamma_{2n} M \left\|\hat{f}' - f'\right\|_1 + \gamma_{2n} M^3 C \left\|\hat{f} - f\right\|_\infty .$$

Since $\gamma_{2n} < \gamma_{1n}^*$, we see that (49) implies the lemma. □

Observe that for differentiable f, we have $\|f\|_{TV} = \|f'\|_1$. The rate implied by lemma 6.2 for $\log f$ in terms of the seminorm $\|\cdot\|_{TV}$ is $n^{-(m-1)/(2m+1)}$; *to achieve the* $n^{-1/3}$ *rate connected with* $\Sigma_n(f_0)$ *we have to assume* $m \geq 4$. Hence the restrictive assumption $f \in \Sigma$.
Lemma 6.2 implies that an achievable rate in $\|\cdot\|_\infty$ is $n^{-3.5/9} \ll n^{-1/3}$. Thus if τ_n in $\Sigma_n(f_0)$ does not grow too fast we already have

$$\sup_{f \in \Sigma} P_{n,f}\left(\hat{f}_n \in \Sigma_n(f)\right) \to 1, \; n \to \infty.$$

However, $\hat{f}_n$ is based on the whole sample. It now remains to deal with achievability in terms of sample size $N_n \ll n$.

6.3 Lemma. *Suppose* $\Sigma_n(f)$ *is defined in terms of* $\tau_n = \log n$, *while* $N_n = n/\log n$. *Then for the estimator* $\hat{g}_n = \hat{f}_N$ *based on a sample fraction* $(X_1, \ldots, X_N)$ *we have*

$$\sup_{f \in \Sigma} P_{n,f}\left(\hat{g}_n \in \Sigma_n(f)\right) \to 1, \; n \to \infty.$$

Proof. Consider lemma 6.2 for a sample size N_n. It then suffices to show $\gamma_{1N}^*\, \tau_n^{-1} n^{-1/3} \to \infty$, $\gamma_{2N}\, \tau_n n^{-1/3} \to \infty$. Since $\gamma_{1N}^* = n^{-\alpha}$ for $\alpha = 3.5/9 > 1/3$, we have

$$\gamma_{1N}^*\, \tau_n^{-1} n^{-1/3} = n^\alpha / (\log n)^{1+\alpha} n^{1/3} \to \infty.$$

Furthermore, $\gamma_{2N} = N^{1/3}$, so

$$\gamma_{2N}\, \tau_n n^{-1/3} = (\log n)^{2/3} \to \infty. \square$$

Other choices of N_n and τ_n are also possible, in particular $N_n = [n/2]$, $\tau_n = \log n$.

7 Appendix

Proof of proposition 2.4. Define

$$\begin{aligned} T_{11} &= nK(f_0\|f), \qquad T_{12} = \frac{n}{2}Var(\kappa_{f,f_0}(Z)), \\ T_{21} &= n\int \kappa_{f,f_0}(t)\frac{1}{\sqrt{n}}\,\mathbf{U}_n(dt), \qquad T_{22} = n\int \kappa_{f,f_0}(t)\frac{1}{\sqrt{n}}\,\mathbf{B}_n(dt). \end{aligned}$$

Then

$$\begin{aligned} &E|\Lambda^{(0)}(f) - \Lambda^{(1)}(f)| = E|\exp(T_{21} - T_{11}) - \exp(T_{22} - T_{12})| \\ &\leq E|\exp(T_{21}) - \exp(T_{22})|\exp(-T_{11}) \\ &+E\exp(T_{22})|\exp(-T_{11}) - \exp(-T_{12})|. \end{aligned}$$

Furthermore, T_{1i} are nonrandom, and T_{22} is a normal random variable with expectation 0 and variance $nVar(\kappa_{f,f_0}(Z))$. Hence

$$E\exp(T_{22}) = \exp\left\{\frac{n}{2}Var(\kappa_{f,f_0}(Z))\right\} = \exp T_{12}, \tag{50}$$

and we obtain

$$\begin{aligned} E|\Lambda^{(0)}(f) - \Lambda^{(1)}(f)| &\leq E|\exp(T_{21}) - \exp(T_{22})|\exp(-T_{11}) \\ &+ |\exp(T_{12} - T_{11}) - 1| \end{aligned}$$

It now suffices to prove that uniformly over $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma_0$

$$T_{12} - T_{11} \to 0, \tag{51}$$

$$E\,|\exp(T_{21}) - \exp(T_{22})|\exp(-T_{11}) \to 0,\ n \to \infty. \tag{52}$$

We begin with (51). Consider the expansion

$$\log x = \log(1 + x - 1) = x - 1 - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 + o((x-1)^3) \tag{53}$$

and put $x = \frac{f}{f_0}$ .Then it follows that for $f \in \Sigma_n(f_0)$

$$\left\|1 - \frac{f}{f_0}\right\|_\infty \leq 2\left\|\log\frac{f}{f_0}\right\|_\infty \leq 2\tau_n^{-1}n^{-1/3} \tag{54}$$

for n large enough, uniformly. Consequently (as in section 3, (21))

$$K(f_0\|f) = O(n^{-2/3}). \tag{55}$$

Now expand T_{12} and T_{11} by means of (53), with $x = \frac{f}{f_0}\left(F_0^{-1}(t)\right)$. Then

$$
\begin{aligned}
T_{12} &= \frac{n}{2}\int\left(\log\frac{f}{f_0}\left(F_0^{-1}(t)\right)\right)^2 dt - \frac{n}{2}\left(K(f_0\|f)\right)^2, \\
&= \frac{n}{2}\int(x-1)^2 - \frac{n}{2}(x-1)^3 + nO\left(n^{-4/3}\right), \\
T_{11} &= nK(f_0\|f) = -n\int\log\frac{f}{f_0}\left(F_0^{-1}(t)\right)dt \\
&= -n\int(x-1) + \frac{n}{2}\int(x-1)^2 - \frac{n}{3}\int(x-1)^3 + nO\left(n^{-4/3}\right)
\end{aligned}
$$

Now, since

$$
\int(x-1) \doteq \int\left(\frac{f}{f_0}-1\right)(F_0^{-1}(t))dt = 0,
$$

we have

$$
T_{12} - T_{11} = \frac{n}{6}\int(x-1)^3 + O\left(n^{-1/3}\right)
$$

Then (54) with $\tau_n \to \infty$ implies (51).
To prove the analog for the random parts (52), define $T_0 = T_{21} - T_{22}$. We then have

$$
T_0 = \sqrt{n}\int\kappa_{f,f_0}d(\mathbf{U}_n - \mathbf{B}_n) = \sqrt{n}\int\log\frac{f}{f_0}d(\mathbf{U}_n\circ F_0 - \mathbf{B}_n\circ F_0).
$$

With a partial integration we obtain

$$
|T_0| = \left|\sqrt{n}\int(\mathbf{U_n}\circ F_0 - \mathbf{B_n}\circ F_0)\,d(\log f - \log f_0)\right|
$$

$$
\leq \sqrt{n}\,\|\mathbf{U_n}\circ F_0 - \mathbf{B_n}\circ F_0\|_\infty\,\|\log f - \log f_0\|_{TV}\,.
$$

Now taking into account $\|\mathbf{U_n}\circ F_0 - \mathbf{B_n}\circ F_0\|_\infty = \|\mathbf{U_n} - \mathbf{B_n}\|_\infty$ and the definition of the class $\Sigma_n(f_0)$, we get for $\delta_n = \tau_n n^{-1/3}$

$$
|T_0| \leq \sqrt{n}\,\|\mathbf{U_n} - \mathbf{B_n}\|_\infty\,\delta_n \tag{56}
$$

Now consider the inequality (7) of the Hungarian construction. In (7) set $x = \frac{u_n}{\delta_n}$, where $u_n = \tau_n^{-1/2}$. Then we obtain

$$
P\left(\sqrt{n}\delta_n\,\|\mathbf{U_n} - \mathbf{B_n}\|_\infty > c_1\delta_n\log n + u_n\right) \leq c_2\exp\left(-c_3\frac{u_n}{\delta_n}\right).
$$

Now $\delta_n\log n = \tau_n\log n\; n^{-1/3} \ll u_n$ (since $\tau_n = o(n^\epsilon)$ for any $\epsilon > 0$ by assumption), while $u_n/\delta_n = n^{1/3}/\tau_n^{3/2}$. Hence for n large enough

$$
P\left(\sqrt{n}\delta_n\,\|\mathbf{U_n} - \mathbf{B_n}\|_\infty > 2u_n\right) \leq c_2\exp\left(-c_3\frac{n^{1/3}}{\tau_n^{3/2}}\right). \tag{57}
$$

Now we claim

$$
E\exp 2\left(T_{2i} - T_{1i}\right) \leq \exp\left(2\tau_n^{-2}n^{1/3}\right),\ i = 1,2. \tag{58}
$$

For $i = 2$ we have analogously to (50)

$$E \exp 2T_{22} = \exp(4T_{12}),$$

Hence

$$E \exp 2(T_{22} - T_{12}) \leq \exp(2T_{12}) \leq \exp\left(n \int \kappa_{f,f_0}^2\right).$$

Now $f \in \Sigma_n(f_0)$ implies $\left\|\kappa_{f,f_0}\right\|_\infty \leq \tau_n^{-1} n^{-1/3}$, so that (58) is proved for $i = 2$. For the case $i = 1$, observe

$$E \exp 2(T_{21} - T_{11}) = E \prod_{i=1}^{n} \left(\frac{f}{f_0}\left(F_0^{-1}(Z_i)\right)\right)^2 = \left(\int \left(\frac{f}{f_0}\right)^2 dF_0\right)^n.$$

Now with $\kappa_{f,f_0}^{(2)}$ from section 3, (16)

$$\int \left(\frac{f}{f_0}\right)^2 dF_0 = \int \left(1 - \kappa_{f,f_0}^{(2)}\right)^2 = 1 + \int \left(\kappa_{f,f_0}^{(2)}\right)^2 \leq 1 + 2\tau_n^{-2} n^{-2/3}$$

according to (54). Consequently

$$E \exp 2(T_{21} - T_{11}) \leq \left(1 + 2\tau_n^{-2} n^{-2/3}\right)^n \leq \exp\left(n 2\tau_n^{-2} n^{-2/3}\right) = \exp\left(2\tau_n^{-2} n^{1/3}\right)$$

so that (58) is established for $i = 1$.
Define an event

$$A = \{\omega : |T_{21} - T_{22}| \leq 2u_n\}.$$

Then (56) and (57) imply the estimate

$$P(A^c) \leq c_2 \exp\left(-c_3 n^{1/3}/\tau_n^{3/2}\right) \tag{59}$$

To prove (52), split the expectation there into $E\chi_A \cdot$ and $E\chi_{A^c} \cdot$, and consider

$$E\chi_A \left|\exp(T_{21}) - \exp(T_{22})\right| \exp(-T_{11}) = E\chi_A \left|1 - \exp(T_{22} - T_{21})\right| \exp(T_{21} - T_{11}).$$

Observe that on $\omega \in A$

$$|1 - \exp(T_{22} - T_{21})| \leq |T_{22} - T_{21}| \exp(T_{22} - T_{21}) \leq 2u_n \exp 2u_n = o(1),$$

so that, since $E \exp(T_{21} - T_{11}) = E\Lambda^{(0)}(f) = 1$,

$$E\chi_A \left|\exp(T_{21}) - \exp(T_{22})\right| \exp(-T_{11}) \leq 2u_n \exp 2u_n E \exp(T_{21} - T_{11}) = o(1). \tag{60}$$

For the other part, use Cauchy-Schwartz to obtain

$$E\chi_{A^c} \left|\exp(T_{21}) - \exp(T_{22})\right| \exp(-T_{11}) \leq$$

$$\left(P(A^c) E \exp 2(T_{21} - T_{11})\right)^{1/2} + \left(P(A^c) E \exp 2(T_{22} - T_{12})\right)^{1/2} \exp(T_{12} - T_{11}). \tag{61}$$

Here the first term on the r. h. s. can be bounded from above by (see (58), (59))

$$\left(c_2 \exp\left(-c_3 n^{1/3}\tau_n^{-3/2}\right) \exp\left(2\tau_n^{-2}n^{1/3}\right)\right)^{1/2} = \left(c_2 \exp -n^{1/3}\tau_n^{-3/2}\left(c_3 - 2\tau_n^{-1/2}\right)\right)^{1/2}.$$

Since $\tau_n^{-1} \to 0$, $n^{1/3}\tau_n^{-3/2} \to \infty$, we see that this term is $o(1)$. The second term on the r. h. s. of (61) is estimated analogously, using (58) for $i = 2$, and in addition (51) for the term $\exp(T_{12} - T_{11})$. Now (60) and (61) prove (52).□

Acknowledgement. The author wishes to thank David Donoho for encouraging discussions, and Enno Mammen for his knowledgeable and constructive advice on global asymptotic normality. He suggested to eliminate the initial sample splitting and obtain a closed form global result (3).

References

[1] Barron, A. R. and Sheu, C. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347-1369

[2] Bickel, P. J., Rosenblatt, M., (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071-1095.

[3] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. Z. Wahrsch. Verw. Gebiete **65** 181-237

[4] Brown, L. D. and Low, M. (1992). Asymptotic equivalence of nonparametric regression and white noise. Mss.

[5] Donoho, D. L, Johnstone, I., (1992). Minimax estimation via wavelet shrinkage. Technical Report, Department of Statistics, Stanford University.

[6] Donoho, D. L. and Liu, R. (1991). Geometrizing rates of convergence, III. *Ann. Statist.* **19** 668-701.

[7] Donoho, D. L., Liu, R. and MacGibbon, B. (1990). Minimax risk for hyperrectangles. *Ann. Statist.* **18** 1416-1437.

[8] Donoho, D. L. and Low, M. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944-970

[9] Efroimovich, S. Yu. and Pinsker, M. S. (1982). Estimating a square integrable probability density of a random variable (in Russian). *Problems Inform. Transmission* **18**, No. 3, 19-38

[10] Golubev, G. K. (1984). On minimax estimation of regression (in Russian). *Problems Inform. Transmission* **20**, No. 1, 56-64

[11] Golubev, G. K. and Nussbaum, M. (1990). A risk bound in Sobolev class regression. *Ann. Statist.* **18** 758-778

[12] Ibragimov, I. A. and Khasminski, R. Z. (1977). On the estimation of an infinite dimensional parameter in Gaussian white noise. *Soviet Math. Dokl.* **236**, No. 5, 1053-1055.

[13] Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. *Statistics and Probability Letters* **13** 15-24

[14] Le Cam , L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

[15] Le Cam , L. and Yang, G. (1990). *Asymptotics in Statistics*. Springer-Verlag, New York.

[16] Low, M. (1992). Renormalization and white noise approximation for nonparametric functional estimation problems. *Ann. Statist.* **20** 545- 554

[17] Mammen, E. (1986). The statistical information contained in additional observations. *Ann. Statist.* **14** 665-678

[18] Meyer, Y. (1992). Ondelettes sur l'intervalle. *Revista Matématica Ibero-Americana* **7**, 115-133.

[19] Millar, P. W. (1979). Asymptotic minimax theorems for the sample distribution function. *Z. Wahrsch. verw. Gebiete*, **48**, 233-252

[20] Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13** 984-997

[21] Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission* (1980) 120-133

[22] Shorack, G., Wellner, J. (1986). *Empirical Processes with Applications to Statistics*. Wiley. New York.

[23] Van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907-924

[24] Woodrofe, M. (1967). On the maximum deviation of the sample density. *Ann. Math. Statist.* **2** 475-481

INSTITUTE OF APPLIED ANALYSIS AND STOCHASTICS
HAUSVOGTEIPLATZ 5-7
D-O 1086 BERLIN
GERMANY
E-MAIL NUSSBAUM@IAAS-BERLIN.DBP.DE

Veröffentlichungen des Instituts für Angewandte Analysis und Stochastik

Preprints 1992

1. D.A. Dawson and J. Gärtner: Multilevel large deviations.
2. H. Gajewski: On uniqueness of solutions to the drift–diffusion–model of semiconductor devices.
3. J. Fuhrmann: On the convergence of algebraically defined multigrid methods.
4. A. Bovier and J.-M. Ghez: Spectral properties of one–dimensional Schrödinger operators with potentials generated by substitutions.
5. D.A. Dawson and K. Fleischmann: A super–Brownian motion with a single point catalyst.
6. A. Bovier, V. Gayrard: The thermodynamics of the Curie–Weiss model with random couplings.
7. W. Dahmen, S. Prößdorf, R. Schneider: Wavelet approximation methods for pseudodifferential equations I: stability and convergence.
8. A. Rathsfeld: Piecewise polynomial collocation for the double layer potential equation over polyhedral boundaries. Part I: The wedge, Part II: The cube.
9. G. Schmidt: Boundary element discretization of Poincaré–Steklov operators.
10. K. Fleischmann, F. I. Kaj: Large deviation probability for some rescaled superprocesses.
11. P. Mathé: Random approximation of finite sums.
12. C.J. van Duijn, P. Knabner: Flow and reactive transport in porous media induced by well injection: similarity solution.
13. G.B. Di Masi, E. Platen, W.J. Runggaldier: Hedging of options under discrete observation on assets with stochastic volatility.
14. J. Schmeling, R. Siegmund–Schultze: The singularity spectrum of self-affine fractals with a Bernoulli measure.
15. A. Koshelev: About some coercive inequalities for elementary elliptic and parabolic operators.
16. P.E. Kloeden, E. Platen, H. Schurz: Higher order approximate Markov chain filters.

17. H.M. Dietz, Y. Kutoyants: A minimum–distance estimator for diffusion processes with ergodic properties.

18. I. Schmelzer: Quantization and measurability in gauge theory and gravity.

19. A. Bovier, V. Gayrard: Rigorous results on the thermodynamics of the dilute Hopfield model.

20. K. Gröger: Free energy estimates and asymptotic behaviour of reaction–diffusion processes.

21. E. Platen (ed.): Proceedings of the 1^{st} workshop on stochastic numerics.

22. S. Prößdorf (ed.): International Symposium "Operator Equations and Numerical Analysis" September 28 – October 2, 1992 Gosen (nearby Berlin).

23. K. Fleischmann, A. Greven: Diffusive clustering in an infinite system of hierarchically interacting diffusions.

24. P. Knabner, I. Kögel–Knabner, K.U. Totsche: The modeling of reactive solute transport with sorption to mobile and immobile sorbents.

25. S. Seifarth: The discrete spectrum of the Dirac operators on certain symmetric spaces.

26. J. Schmeling: Hölder continuity of the holonomy maps for hyperbolic basic sets II.

27. P. Mathé: On optimal random nets.

28. W. Wagner: Stochastic systems of particles with weights and approximation of the Boltzmann equation. The Markov process in the spatially homogeneous case.

29. A. Glitzky, K. Gröger, R. Hünlich: Existence and uniqueness results for equations modelling transport of dopants in semiconductors.

30. J. Elschner: The *h-p*–version of spline approximation methods for Mellin convolution equations.

31. R. Schlundt: Iterative Verfahren für lineare Gleichungssysteme mit schwach besetzten Koeffizientenmatrizen.

32. G. Hebermehl: Zur direkten Lösung linearer Gleichungssysteme auf Shared und Distributed Memory Systemen.

33. G.N. Milstein, E. Platen, H. Schurz: Balanced implicit methods for stiff stochastic systems: An introduction and numerical experiments.

34. M.H. Neumann: Pointwise confidence intervals in nonparametric regression with heteroscedastic error structure.