

# Finite dimensional approximations to nonparametric statistical experiments

BY ANDREW V. CARTER

*University of California, Santa Barbara*

Nonparametric statistical experiments such as estimation of an unknown density can be approximated by sub-experiments with finite-dimensional parameter sets (such as a multinomial with unknown probabilities) which are easier to analyze and compare to other experiments. The dimension of the parameter class in the sub-experiment and the smoothness conditions on the parameter class in the nonparametric experiment determine the accuracy of the approximation. The loss of information is measured via Le Cam's deficiency distance between experiments. These bounds are at least implicitly part of recent results which bound the deficiency distance between nonparametric experiments and establish the asymptotic equivalence of density estimation and Gaussian experiments. Also, many nonparametric techniques rely upon a reduction of the observations to a finite set of coefficients, and these results bound the necessary number of coefficients to retain certain asymptotic properties. I will describe a general strategy for deriving the bounds and then calculate a specific bound in the case of densities on the unit interval.

October 9, 2001

**1. Introduction** A statistical experiment is a set of possible distributions indexed by a set of parameters in order to model some observed data. "Nonparametric" experiments have parameter sets  $\mathcal{F}$  that are subsets of some infinite dimensional function space ( $L^2$ , Hölder classes, Besov spaces, Sobolev spaces, etc.)

Specifically, a density estimation experiment observes  $n$  independent observations from some distribution  $P_f$  on the space  $(\mathcal{X}, \mathcal{A})$  with an unknown density  $f = dP_f/dP_0$  where  $P_0$  dominates all the  $P_f$ . The possible joint distributions for these observations are  $P_f^n$  for  $f \in \mathcal{F}$ , a set of possible density functions. This set of distribution functions forms a statistical experiment  $\mathcal{P}$ .

There is a related Gaussian process experiment  $\mathcal{Q}$  that consists of the distributions of the  $Y_g$ , linear Gaussian processes on the probability spaces  $(\mathcal{X}, \mathcal{A}, P_0)$ . The  $Y_g$  are such that

$$Y_g(A) \sim \mathcal{N}(P_0[gA], P_0[A]/n) \quad \text{for } A \in \mathcal{A},$$

and  $Y_g(A)$  and  $Y_g(B)$  are independent if  $A$  and  $B$  are disjoint. The parameter set is  $\mathcal{G} \subset L^2(P_0)$  which contains functions drift functions  $g$  that are not necessarily densities.

The  $\mathcal{P}$  and  $\mathcal{Q}$  experiments are indexed by classes of smooth functions  $\mathcal{F}$  or  $\mathcal{G}$ , but they can be approximated by observations from distributions indexed by finite-

---

*AMS 1991 subject classifications.*  
*Key words and phrases.*

dimensional parameters. These new experiments are generated by restricting the observations to a partition  $\{\chi_1, \chi_2, \dots, \chi_m\}$  where  $P_0\chi_i = 1/m$ .

Let  $\theta_i = P_f\chi_i$  for  $i = 1, \dots, m$ . The experiment  $\bar{\mathcal{P}}$  observes only how many observations fall into each  $\chi_i$ ,

$$\bar{X}_i = \sum_{j=1}^n \mathbf{1}_{\{X_j \in \chi_i\}}.$$

The resulting multinomial distributions

$$(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_m) \sim \bar{\mathbb{P}}_f = \mathcal{M}(n, \theta_1, \theta_2, \dots, \theta_m)$$

are indexed by a parameter set  $\Theta$  that is a subset of the  $m$  dimensional simplex.

Likewise, the increments of the Gaussian process,  $Y_g(\chi_i)$ , form a finite-dimensional experiment  $\bar{\mathcal{Q}}$  which observes  $m$  independent normals with means  $\psi_i = P_g\chi_i$  and variances  $(nm)^{-1}$ . The parameter set  $\Psi$  is a subset of  $\mathbb{R}^m$ .

My objective is to measure the information lost in approximating the nonparametric experiments by their finite-dimensional counter parts. Intuitively, if the partition is sufficiently fine then the observations  $\bar{\mathcal{P}}$  (or  $\bar{\mathcal{Q}}$ ) should be nearly sufficient statistics for the model in  $\mathcal{P}$  (or  $\mathcal{Q}$ ).

These approximations will be evaluated using the deficiency distance of Le Cam (1964, 1986). This distance between statistical experiments can be bounded using a transformation  $T$  of the observations (which will also rely on some independent external randomization) such that the resulting image under  $\bar{\mathbb{P}}_f$  is approximately  $P_f^n$ . Thus,  $T$  is a randomized map from the sample space of  $\bar{\mathcal{P}}$  to the space  $\mathcal{X}^n$ , the sample space of  $P_f^n$ . This transformation is useful because it implies that any estimator  $\hat{f}(\mathbf{X})$  can be approximated by an estimator  $\hat{f}(T[\bar{\mathbf{X}}])$  such that the distributions of the two estimators are similar. The distance between the distributions will be measured using total-variation distance because the risks for any loss function bounded by 1 will differ by at most the total-variation distance between  $\bar{\mathbb{P}}_f T$  and  $P_f^n$ .

Le Cam's deficiency  $\delta(\mathcal{P}, \bar{\mathcal{P}})$  is bounded by

$$\delta(\mathcal{P}, \bar{\mathcal{P}}) \leq \inf_T \sup_f \|\bar{\mathbb{P}}_f T - P_f^n\|,$$

but it is not symmetric so let

$$\Delta(\mathcal{P}, \bar{\mathcal{P}}) = \max [\delta(\mathcal{P}, \bar{\mathcal{P}}), \delta(\bar{\mathcal{P}}, \mathcal{P})].$$

For sufficient statistics, the conditional distribution of the data given the statistic does not depend on the parameter. The map  $T$  plays a role similar to this conditional distribution. It is not a function of  $f$  and can be used to produce (nearly) observations from the other experiment. In this way the deficiency can be seen as measuring the distance the sub-experiments are from being sufficient statistics.

This approximation by finite-dimensional experiments plays a role in some asymptotic results in Carter (2001a,b) which establish a Gaussian approximation to  $\mathcal{P}$ . The experiments themselves are also of interest especially as  $\bar{\mathcal{Q}}$  is a sort of nonparametric regression experiment.

The more typical nonparametric regression experiment  $\mathcal{Q}^*$  observes  $n$  independent normal  $\mathcal{N}(g(x_i^*), 1/n)$  random variables where the  $x_i^*$  is a fixed point in  $\chi_i$  for each  $i$ . Brown and Low (1996) showed that if  $\mathcal{G}$  is a subset of a Hölder class on  $[0, 1]$  with exponent  $\alpha$ , then  $\mathcal{Q}$  is a good approximation for  $\mathcal{Q}^*$  as long as  $1/2 < \alpha \leq 1$ . This article essentially generalizes Brown and Low (1996) to get a better bound on the distance for  $1 < \alpha < 3/2$ , and extends the technique to the density estimation experiments.

A clear application of this result for the Gaussian experiment  $\mathcal{Q}$  is the work in Donoho and Johnstone (1999). They showed that a minimax rate derived for the continuous process experiment  $\mathcal{Q}$  is also valid for nonparametric regression experiments. They however were arguing over a larger class of functions including the Hölder classes with  $0 < \alpha \leq 1/2$ , but only for convergence under  $L^2$  loss. The asymptotic equivalence described here is only valid for  $\alpha > 1/2$  but is valid for all bounded loss functions. This difference points to a possible weakness of the deficiency distance in this problem. Regardless, their wavelet technique provides an interesting approach to our problem, and we will discuss this in section 4 in hopes of getting better convergence rates for certain function classes.

**2. Approximating density estimation by multinomials.** The deficiency  $\delta(\mathcal{P}, \bar{\mathcal{P}}) = 0$  because the multinomial observations can be produced by counting the number of observations in each set  $\chi_i$ . Therefore,  $\bar{\mathcal{P}}$  is a sub-experiment of  $\mathcal{P}$ .

The deficiency  $\delta(\bar{\mathcal{P}}, \mathcal{P})$  can be bounded using a transformation generated by the distributions  $K_1, \dots, K_m$ .

CONDITION A.  $K_i$  for  $i = 1, \dots, m$  are probability distributions such that  $K_i \ll P_0$ .

The transformation produces  $\bar{X}_i$  independent observations from each  $K_i$  followed by a random permutation of the indices. This randomization produces observations  $X_j^*$  with conditional distributions

$$X_j^* | \bar{X}_i \sim \sum_{i=1}^m \frac{\bar{X}_i}{n} K_i$$

for  $j = 1, \dots, n$ .

More importantly, the marginal distributions of the  $X_j^*$  have a density

$$\prod_{j=1}^n \sum_{i=1}^m \theta_i \frac{dK_i}{dP_0}(x_j).$$

Let  $\hat{f} = \sum_{i=1}^m \theta_i \frac{dK_i}{dP_0}(x_j)$ . Thus the distribution  $\bar{\mathbb{P}}_f T = P_{\hat{f}}^n$ .

Therefore, a bound on the deficiency distance  $\delta(\bar{\mathcal{P}}, \mathcal{P})$  is the total variation distance between  $n$  independent observations from  $f$  and  $\hat{f}$ . In this setup, the problem bears resemblance to a problem of function estimation using kernels. However these kernels are necessarily positive so that they correspond to densities of distributions that fulfill Condition A. In Section 5, an interpolating kernel  $K_i$  will be used.

THEOREM 1.

$$\Delta(\mathcal{P}, \bar{\mathcal{P}}) \leq \sup_{f \in \mathcal{F}} C\sqrt{n} \left\| f^{1/2} - \hat{f}^{1/2} \right\|_2$$

where  $\hat{f}$  is defined in terms of the  $K_i$  probability measures above.

The theorem is a straightforward application of the randomization described above along with standard Hellinger distance bounds on total-variation distance.

If the densities are bounded away from zero, by  $f \geq \epsilon > 0$  for instance, then the Hellinger distance can be bounded by  $L^2$ ,

$$P_0 \left( f^{1/2} - \hat{f}^{1/2} \right)^2 = P_0 \frac{(f - \hat{f})^2}{(f^{1/2} + \hat{f}^{1/2})^2} \leq \epsilon^{-1} P_0 (f - \hat{f})^2. \quad (2.1)$$

**3. Gaussian drift experiments.** Let  $Y_0(x)$  be a standard Gaussian process which maps  $L^2(P_0)$  functions  $x$  and  $y$  to normal random variables with mean 0 and covariance  $P_0(xy)$ .

A Gaussian process with drift  $g \in L^2(P_0)$  and constant variance  $\sigma^2$  can be constructed via

$$Y_g(x) = \sigma Y_0(x) + P_0(gx).$$

A Gaussian drift experiment  $\mathcal{Q}$  is the set of Gaussian drift processes with drifts  $g \in \mathcal{G}$ .

The Gaussian drift experiment can be approximated by the values of

$$Y_g(\chi_1); Y_g(\chi_2); \cdots; Y_g(\chi_m). \quad (3.2)$$

A new experiment  $\bar{\mathcal{Q}}$  is generated by these observations with independent normal distributions having mean  $P_0(g\chi_i) = \psi_i$  and variance  $\sigma P_0(\chi_i^2) = \sigma/m$ .

The deficiency  $\delta(\mathcal{Q}, \bar{\mathcal{Q}}) = 0$  because the function described by (3.2) generates the distributions in  $\bar{\mathcal{Q}}$ .

The deficiency  $\delta(\bar{\mathcal{Q}}, \mathcal{Q})$  is bounded using a randomization on the finite collection of normals that is based on the same  $K_i$  as in Section 2 with one further condition imposed.

CONDITION B. *The set of distributions  $K_i$  must be such that*

$$P_0 = \frac{1}{m} \sum_{i=1}^m K_i.$$

This is typically fulfilled, especially when Condition A is true.

These  $K_i$  can define new Gaussian processes  $W_{K_i}$  with means 0 and covariance functions  $K_i(xy)$ . Likewise, a  $K_i$ -Brownian bridge is constructed by subtracting off the term corresponding to the identity  $\mathbf{1}$ .

$$B_{K_i}(x) = W_{K_i}(x) - K_i(x)W_{K_i}(\mathbf{1}) \quad \text{for } x \in L^2(P_0)$$

where  $W_{K_i}(\mathbf{1}) \sim \mathcal{N}(0, 1)$  because  $K_i$  is a probability distribution.

LEMMA 1. *Under Conditions A and B, the process, constructed conditionally on the values of  $Y_g(\chi_i)$ ,*

$$\bar{Y}_g(x) = \sum_{i=1}^m \left[ Y_g(\chi_i) K_i(x) + \frac{\sigma}{\sqrt{m}} B_{K_i}(x) \right], \quad (3.3)$$

where  $B_{K_i}$  are independent  $K_i$ -Brownian bridges, is equal in distribution to the process

$$Y_{\hat{g}}(x) = P_0(x\hat{g}) + \sigma Y_0(x)$$

where  $\hat{g} = \sum \psi_i \frac{dK_i}{dP_0}$ .

PROOF OF LEMMA 1. The Brownian bridges can be written as

$$B_{K_i}(x) = W_{K_i}(x) - K_i(x)W_{K_i}(\mathbf{1})$$

for  $W_{K_i}(\mathbf{1}) \sim \mathcal{N}(0, 1)$  and independent of  $B_{K_i}(x)$ . Then the distribution of the observations  $Y_g(\chi_i) \sim \mathcal{N}(\psi_i, \sigma^2/m)$  have the same distributions as

$$\psi_i + \frac{\sigma}{\sqrt{m}} W_{K_i}(\mathbf{1}).$$

$\bar{Y}_g(x)$  therefore has the same distribution as

$$\sum_{i=1}^n \left[ \psi_i K_i(x) + \frac{\sigma}{\sqrt{m}} (K_i(x)W_{K_i}(\mathbf{1}) + B_{K_i}(x)) \right].$$

which has the same distribution as

$$\bar{Y}_g(x) \stackrel{D}{=} P_0 \left[ x \sum_{i=1}^m \psi_i \frac{dK_i}{dP_0} \right] + \frac{\sigma}{\sqrt{m}} \sum_{i=1}^m W_{K_i}(x).$$

The distribution of the sum of the  $W_{K_i}$  is still Gaussian with mean 0. The independence of the processes implies that the covariances are

$$\begin{aligned} \mathbb{E} \left( \frac{1}{\sqrt{m}} \sum_{i=1}^m W_{K_i}(x) \right) \left( \frac{1}{\sqrt{m}} \sum_{j=1}^m W_{K_j}(y) \right) &= \frac{1}{m} \sum_{i=1}^m \mathbb{E} [W_{K_i}(x)W_{K_i}(y)] \\ &= \frac{1}{m} \sum_{k=1}^m K_k(xy) \\ &= P_0(xy). \end{aligned}$$

Therefore  $\frac{1}{\sqrt{m}} \sum_{i=1}^m W_{K_i}(x) \stackrel{D}{=} Y_0(x)$  and the lemma follows.  $\square$

THEOREM 2.

$$\Delta(\bar{\mathcal{Q}}, \mathcal{Q}) \leq \sup_{g \in \mathcal{G}} \frac{C}{\sigma} \|g - \hat{g}\|_2.$$

PROOF OF THEOREM 2. Because  $\delta(\mathcal{Q}, \bar{\mathcal{Q}}) = 0$ , it is only necessary to bound  $\delta(\bar{\mathcal{Q}}, \mathcal{Q})$ . The transformation (3.3) produces the process

$$\bar{Y}_g(x) = \sum_{i=1}^m Y_g(\chi_i) K_i(x) + \frac{\sigma}{\sqrt{m}} B_{K_i}(x)$$

which is distributed as  $Y_{\bar{g}}$  as in Lemma 1.

The total-variation distance between a pair of Gaussian processes is bounded by the  $L^2$  distance between their means weighted by the variance. Therefore Theorem 2 follows from Lemma 1  $\square$

The variance of the Gaussian drift process will be typically  $\sigma^2 = 1/n$  to correspond with the result from Lemma 1.

**4. Using a basis to describe the Gaussian process.** Another method for analyzing the approximations in the Gaussian case is to use wavelet bases. If  $\{\varphi_j\}$  is a complete orthonormal basis of  $L^2(P_0)$  such that

$$x = \sum_{j=0}^{\infty} \langle x, \varphi_j \rangle \varphi_j \quad \text{for any } x \in L^2(P_0),$$

then

$$Y_0(x) = \sum_{j=0}^{\infty} \langle x, \varphi_j \rangle \xi_j$$

is a standard Gaussian process if the  $\xi_j$  are all independent  $\mathcal{N}(0, 1)$ .

The Gaussian experiment  $\mathcal{Q}$  is therefore equivalent to observing a sequence of independent normals  $\xi_j \sim \mathcal{N}(P_g[\varphi_j], \sigma^2)$ . The two experiments are related by  $\xi_j = Y_g(\varphi_j)$  and  $Y_g(x) = \sum_j \langle x, \varphi_j \rangle \xi_j$  where the inner product is  $\langle x, \varphi_j \rangle = P_0(x\varphi_j)$ .

These experiments can be approximated by  $\tilde{\mathcal{Q}}$  which observes only the first  $m$  normals  $\xi_j$ . The rest of the coefficients can be produced randomly  $\tilde{\xi}_j \sim \mathcal{N}(0, \sigma^2)$ . Therefore, setting  $a_j = P_0(g\varphi_j)$ ,

$$\Delta(\mathcal{Q}, \tilde{\mathcal{Q}}) \leq \sup_{g \in \mathcal{G}} \frac{1}{\sigma} \left[ \sum_{j>m} a_j^2 \right]^{1/2}. \quad (4.4)$$

For  $P_0$  a measure on  $\mathbb{R}$ , let  $\mathcal{G}$  be the Lipschitz space  $\text{Lip}(\alpha, L^2(P_0))$  such that

$$\|g(x) - g(x+h)\|_2 \leq Mh^\alpha$$

for  $0 < \alpha \leq 1$ . Then a reasonable choice for  $\phi_j$  are the Haar basis functions. The first  $m$  basis functions span the partition  $\{\chi_i\}_{i=1}^m$  when  $m$  is a power of 2 so  $\Delta(\tilde{\mathcal{Q}}, \bar{\mathcal{Q}}) = 0$ . A version of Jackson's inequality implies that

$$\left[ \sum_{j>m} a_j^2 \right]^{1/2} \leq 2Mm^{-\alpha},$$

and therefore

$$\Delta(\mathcal{Q}, \mathcal{Q}^*) \leq CMm^{-\alpha}\sigma^{-1} \quad \text{for } \alpha \leq 1. \quad (4.5)$$

This is essentially the same construction and bound as Brown and Low (1996).

It is interesting to follow the example of Donoho and Johnstone (1999) and use smoother wavelet bases to get better bounds when  $\alpha > 1$ . However these experiments no longer are equivalent to  $\tilde{\mathcal{Q}}$  because the basis functions cannot be easily computed from the functions on the partition  $\{\chi_i\}$ . Thus,  $\Delta(\mathcal{Q}, \tilde{\mathcal{Q}}) \leq CMm^{-\alpha}/\sigma$  for wavelet functions with derivatives of order at least as large as  $\alpha$ , but  $\Delta(\tilde{\mathcal{Q}}, \tilde{\mathcal{Q}}) > 0$ .

The discrete experiment  $\mathcal{Q}^*$  that Donoho and Johnstone (1999) considers is a nonparametric function estimation experiment where the observations are  $Y_i \sim \mathcal{N}(f(x_i), \sigma^2)$  where  $x_i$  is the midpoint of each interval (or the end points of the intervals). Their proposal is to use linear interpolating functions  $K_i$  and then estimate the coefficients via

$$P_0 \left[ \sum_{i=1}^m Y_i K_i(x) \phi_j(x) \right].$$

These  $K_i$  still integrate to one and fulfill condition B, but they are allowed to be negative. The problems this causes with the variances seems to be negligible for large  $n$ . They bound the  $L^2$  distance between the normals constructed this way by  $O(m^{-\alpha})$  which is always smaller than the minimax risk

$$n^{-\alpha/(2\alpha+1)}$$

when  $m = n$ . For our purposes, however, the total variation distance is the  $L^2$  distance divided by the variance,  $n^{-1/2}$ , and therefore  $\alpha$  must be greater than  $1/2$  for asymptotic equivalence. This construction improves on Lemma 2 for  $\alpha > 3/2$ . It is still somewhat of a problem to go from the observations at a point  $Y_g(x_i)$  to observations over the interval  $Y_g(\chi_i)$ .

It is not all that clear exactly how to apply this strategy to the density estimation experiments. There is probably an analogous strategy which estimates the wavelet coefficients by

$$P_0 \left[ \phi_j \sum_{i=1}^m X_i K_i(x) \right]$$

where the  $K_i$  can be allowed to be negative. The distributions in this case are much more complicated and it is somewhat more difficult to randomize the coefficients from the wavelet observations.

## 5. Choosing suitable kernels $K_i$ .

5.1. *One dimension.* If the measure  $P_0$  is the uniform probability on  $[0, 1]$  and the sets are  $\chi_i = [(i-1)/m, i/m]$ . The parameter set is the set of densities  $f \in \mathcal{F}$  such that  $f > \epsilon$ , and the densities have a bounded derivative such that

$$|f'(t) - f'(s)| \leq M|t - s|^{\alpha-1}.$$

This can be used to bound the error in a Taylor expansion:

$$\begin{aligned} |f(t + \delta) - f(t) - \delta f'(t)| &= |f(t) + \delta f'(t^*) - f(t) - \delta f'(t)| \\ &\leq M\delta^\alpha \end{aligned} \tag{5.6}$$

because there exists a  $t^*$  such that  $|t^* - t| \leq \delta$  by the Mean Value Theorem.

Let  $x_i^*$  be the midpoint of each of these intervals,  $(i-1/2)/m$ .

The probability  $\theta_i$  can be written using a Taylor expansion

$$\begin{aligned} \theta_i &= \int_{\chi_i} f(x) dx \\ &= \int_{\chi_i} f(x_i^*) + (x - x_i^*)f'(x_i^*) + E dx \\ &= \frac{f(x_i^*)}{m} + \frac{E}{m} \end{aligned}$$

because  $\int_{\chi_i} (x - x_i^*) = 0$ . By the bound above,  $|E| \leq Mm^{-\alpha}$  because  $|x - x_i^*| \leq 1/m$  in  $\chi_i$ .

Following the argument in section 4, the conditional probabilities  $K_i$  will have densities equal to positive interpolating functions:

$$\frac{dK_i}{dP_0} : \quad \begin{array}{c} m \\ \diagup \quad \diagdown \\ x_{i-1}^* \quad x_i^* \quad x_{i+1}^* \end{array}$$

where  $x_{i-1}^* = x_i^* - 1/m$  and  $x_{i+1}^* = x_i^* + 1/m$ .

Other interpolating functions exist with greater smoothness, but they cannot be made into probability distributions because they are negative in some places (cf. Donoho and Johnstone (1999); Donoho (1992); Deslauriers and Dubuc (1989)).

For  $i = 1$ , I will need a different conditional density,

$$\frac{dK_1}{dP_0} : \quad \begin{array}{c} m \\ \square \quad \diagdown \\ 0 \quad \frac{1}{2m} \quad \frac{3}{2m} \end{array}$$

to avoid getting observations outside of  $[0, 1]$ . A similar conditional density will be used for  $i = m$  to avoid results greater than 1.

The resulting sums of these densities at a midpoint  $x_i^*$  is

$$\sum_{i=1}^m \theta_i \frac{dK_i}{dP_0}(x_i^*) = m\theta_i = f(x_i^*) + E.$$

Between successive  $x_i^*$ 's the density is linear. The original density is within  $Mm^{-\alpha}$  of a straight line between  $f(x_i^*)$  and  $f(x_{i+1}^*)$  by (5.6).

$$\begin{aligned} f(x) &= f(x_i^*) + (x - x_i^*)f'(x_i^*) + E \\ &= f(x_i^*) + (x - x_i^*)f'(\xi) + (x - x_i^*)(f'(x_i^*) - f'(\xi)) + E \end{aligned}$$

where  $\xi$  is the point between  $x_i^*$  and  $x$  such that  $f'(\xi)$  is the slope of the line between  $f(x_i^*)$  and  $f(x_{i+1}^*)$ .

The total error is therefore

$$\left| f(x) - \sum_{i=1}^m \theta_i \frac{dK_i}{dP_0}(x) \right| \leq 4Mm^{-\alpha} \quad \text{for } \frac{1}{m} \leq x \leq \frac{m-1}{m}.$$

*Edge effects.* As before,

$$|\hat{f}(1/m) - f(1/m)| \leq Mm^{-\alpha},$$

and then

$$|f(x) - f(1/m)| \leq Mm^{-1} \quad \text{for } x < 1/(2m)$$

because the derivative is bounded by  $M$ . A similar bound is available for  $x$  near 1

$$|\hat{f}(x) - f(x)| \leq Mm^{-1} + Mm^{-\alpha} \quad \text{for } x > 1 - \frac{1}{2m}.$$

This error is larger than the error over the rest of the interval.

*Total Error.* The integral of the squared error is less than

$$\|\hat{f} - f\|_2^2 \leq m^{-1}Mm^{-2} + 4M^2m^{-2\alpha}$$

which is  $O(m^{-2\alpha})$  if  $\alpha \leq 3/2$ , or  $O(m^{-3})$  if  $\alpha > 3/2$ .

Therefore,  $\Delta(\mathcal{E}, \bar{\mathcal{E}}) \leq Cm^{-\alpha}n^{1/2}$  for  $\alpha < 3/2$  and  $\Delta(\mathcal{E}, \bar{\mathcal{E}}) \leq Cm^{-3/2}n^{1/2}$  otherwise.

5.2. *Experiments on the unit square.* If  $P_0$  is the measure uniform over a unit square, the partition  $\{\chi_i\}_{i=1}^m$  divides the space into squares of side length  $1/\sqrt{m}$ . One possible kernel would be to take independent copies of the  $K_i$  kernels in the  $x$  and  $y$  directions. The error in each interior is on the order of  $CMm^{-\alpha/2}$ .

The error at the edges is on the order of  $CMm^{-1/2}$ .

The total error is something like  $Cm^{-\alpha/2} + Cm^{-1}$  which is always on the order of  $Cm^{-\alpha/2}$  for  $\alpha \leq 2$ .

Therefore, for  $\alpha \leq 2$ , the distance  $\Delta(\mathcal{E}, \bar{\mathcal{E}}) \leq Cm^{-\alpha/2}n^{1/2}$ . So if these two experiments are to be asymptotically equivalent, then  $m$  must increase as a function of  $n$  so that  $n = o(m^\alpha)$ .

## REFERENCES

- BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

- CARTER, A. V. (2001a). Deficiency distance between multinomial and multivariate normal experiments. Available at [www.pstat.ucsb.edu/faculty/carter/research/](http://www.pstat.ucsb.edu/faculty/carter/research/).
- CARTER, A. V. (2001b). Deficiency distance between multinomial and multivariate normal experiments under smoothness constraints on the parameter set. Tech. rep., UCSB. Available at [www.pstat.ucsb.edu/faculty/carter/research/](http://www.pstat.ucsb.edu/faculty/carter/research/).
- DESLAURIERS, G. and DUBUC, S. (1989). Interpolation through an iterative scheme. *Constructive Approximation* **5** 49–68.
- DONOHO, D. L. (1992). Interpolating wavelet transforms. Tech. rep., Stanford University. Available at <http://www-stat.stanford.edu/~donoho/Reports>.
- DONOHO, D. L. and JOHNSTONE, I. M. (1999). Asymptotic minimaxity of wavelet estimators with sampled data. *Statistica Sinica* **9** 1–32.
- IBRAGIMOV, I. and KHASHINSKII, R. (1997). Some estimation problems in infinite-dimensional Gaussian white noise. In D. Pollard, E. Torgersen and G. L. Yang, eds., *Festschrift for Lucien Le Cam*. Springer, New York, 259–274.
- LE CAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35** 1419–1455.
- LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.
- LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics, Some Basic Concepts*. Springer-Verlag, New York, 2nd ed.
- NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.

ANDREW V. CARTER  
DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY  
UNIVERSITY OF CALIFORNIA, SANTA BARBARA  
SANTA BARBARA, CALIFORNIA 93106  
EMAIL: CARTER@PSTAT.UCSB.EDU