A continuous Gaussian approximation to a nonparametric regression in two dimensions.

ANDREW V. CARTER

University of California Santa Barbara, Santa Barbara, CA 93106-3110.
Email: carter@pstat.ucsb.edu

Estimating the mean in a nonparametric regression on a two-dimensional regular grid of design points is asymptotically equivalent to estimating the drift of a continuous Gaussian process on the unit square. In particular, we provide a construction of a Brownian sheet process with a drift that is almost the mean function in the nonparametric regression. This can be used to apply estimation or testing procedures from the continuous process to the regression experiment as in Le Cam's theory of equivalent experiments. Our result is motivated by first looking at the amount of information lost in binning the data in a density estimation problem.

Keywords: Nonparametric regression, Asymptotic equivalence of experiments, density estimation

1 Introduction

The purpose of this paper is to establish a connection between a nonparametric regression on a two-dimensional set of design points and an appropriate continuous Gaussian approximation. This connection provides a bound on Le Cam's deficiency distance between the experiments and allows inference in the easier problem (the continuous case) to be applied to the practical problem (observations on a finite grid of points). The motivation for the form of this connection comes from a similar result: approximating the problem of estimating an unknown density from n independent observations by the experiment that observes only those observations aggregated into m bins.

Brown and Low (1996) showed that the nonparametric regression experiment

$$Y_i = g(x_i) + \sigma\xi_i \qquad i = 1, \ldots, n \tag{1}$$

with ξ_i independent standard normals, $x_i = i/(n+1)$, σ a known value, and g an unknown smooth function on $[0,1]$, is asymptotically equivalent to an observation of a continuous Gaussian process with an unknown drift function

$$dY_t = g(t)\, dt + \frac{\sigma}{\sqrt{n}}\, dW_t \tag{2}$$

where W_t is a standard Brownian Motion on $[0,1]$ whenever the class of possible mean functions $g \in \mathcal{G}$ is a subset of a Lipshitz(α) space with $\alpha > 1/2$.

This result does not immediately extend to a regression on a two-dimensional space. The higher dimensional result requires that the class of drift functions be smoother; in particular, they must be differentiable, $\alpha > 1$. We propose a construction that can take advantage of this added smoothness.

The idea of developing asymptotic results for nonparametric regression by appealing to a continuous Gaussian processes approximation is widely used, see for example Donoho et al. (1995) or Efromovich (1999, Chapt. 7). Donoho and Johnstone (1999) described a method for constructing wavelet coefficients from the nonparametric regression as if the original process was continuous. They showed that the squared-error loss was not significantly effected by the approximation. Our construction leads to the same statistical estimators. Brown et al. (2002) extended their original result to include regression with a random design, but our result assumes a fixed grid of equally-spaced design points. Grama and Nussbaum (1998) established equivalence for regression problems with non-normal errors.

Nussbaum (1996) showed that observing n independent observations from an unknown density is equivalent to a Brownian motion plus drift. Carter (2002) and Brown et al. (2004) constructed a connection between these experiments by comparing their behavior on a finite partition of the unit interval. It is also then necessary to bound the error in "discretizing" the continuous observations. The process in (2) is approximated by its increments $Y(j/m) - Y((j-1)/m)$, and the independent observations are approximated by the number falling in each of m equal subintervals. These two problems are related and the solution of the binning problem will motivate the approach used to go from the continuous to the finite-dimensional Gaussian process. The technical bounds in section 4 can essentially be used in both situations.

1.1 Le Cam's deficiency distance

The constructions described in sections 2 and 3 will bound the deficiency distance between the experiments. This statistical distance compares the relevant information about the parameter that is available in the two sets of distributions.

First, the total variation distance between two distributions is

$$\|P - Q\| = \sup_{A \in \mathcal{A}} P(A) - Q(A) \tag{3}$$

where P and Q are both measures on the σ-field $\mathcal{A}$. This distance also bounds the difference in the expectations of bounded functions $|g| < 1$,

$$\sup_{\{g:|g|\leq 1\}} |P(g) - Q(g)| \leq 2\|P - Q\|.$$

However, this distance is equal to 1 if P and Q do not have common support.

A statistical experiment $\mathcal{P}$ consists of a set of distributions $\{P_\theta : \theta \in \Theta\}$ indexed by a parameter set Θ for data $X \in \mathcal{X}$ with a σ-field $\mathcal{A}$. A second experiment $\mathcal{Q} = \{Q_\theta : \theta \in \Theta\}$ has the same parameter set but a different sample space $(\mathcal{Y}, \mathcal{B})$. To compare these two experiments, we need a way to connect the sample space $\mathcal{Y}$ to $\mathcal{X}$. A randomization of the data X can be described by the conditional distribution K_x on $(\mathcal{Y}, \mathcal{B})$ given $X = x$. Let $P_\theta K_x$ represent the marginal distribution on $(\mathcal{Y}, \mathcal{B})$. The randomization does not depend on θ so $P_\theta K_x$ cannot contain any additional information about θ. If $\|P_\theta K_x - Q_\theta\|$ is always small, then $\mathcal{Q}$ does not have much more information about θ than $\mathcal{P}$.

Le Cam's deficiency (Le Cam, 1986, pp. 18–20) is

$$\delta(\mathcal{P}, \mathcal{Q}) = \inf_K \sup_{\theta \in \Theta} \|P_\theta K - Q_\theta\|$$

where K is a "transition" from the linear space including probability distributions P_θ to a space that includes the measures Q_θ. For experiments that depend on an increasing sample size n, if $\delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0$ then the sequence of experiments $\mathcal{P}_n$ are asymptotically as informative as $\mathcal{Q}_n$. Furthermore, if $\max\left[\delta(\mathcal{P}_n, \mathcal{Q}_n), \delta(\mathcal{Q}_n, \mathcal{P}_n)\right] \to 0$ then $\mathcal{P}_n$ and $\mathcal{Q}_n$ are termed asymptotically equivalent.

For our purposes, it is not necessary to think in terms of these general transitions. We will bound the deficiency using a transformation of the observations X from $\mathcal{P}$ that may also include an external randomization $T(X, W)$. The random variable W represents the external randomization that has the same distribution for all θ. Then K_x is the conditional distribution of $T(X, W)$ given $X = x$. Thus,

$$\begin{aligned}\delta(\mathcal{P}, \mathcal{Q}) &\leq \sup_{\theta \in \Theta} \|P_\theta K_x - Q_\theta\| \\ &= \sup_{\theta \in \Theta} \sup_{A \in \mathcal{A}} |P_\theta\{T(X, W) \in A\} - Q_\theta\{Y \in A\}|\end{aligned}$$

The usefulness of a bound of this type is in its flexibility. Suppose that $\sup_{\theta \in \Theta} \|P_\theta K_x - Q_\theta\| \leq \tau$. Then for any bounded loss function $|L(a, \theta)| \leq 1$, any decision procedure $d(Y)$ in the experiment $\mathcal{Q}$ which has risk $R(d(Y), \theta) = Q_\theta L(d(Y), \theta)$ generates a randomized decision procedure $d\left[T(X, W)\right]$ such that

$$R(d\left[T(X, W)\right], \theta) = P_\theta L(d\left[T(X, W)\right], \theta) \leq Q_\theta L(d(Y), \theta) + 2\tau.$$

In other words, the T transformation maps good decision procedures for $\mathcal{Q}$ to good decision procedures for the $\mathcal{P}$ experiment.

1.2 Main Results

The parameter sets in these nonparametric experiments are Lipshitz classes of differentiable functions $\mathcal{L}(\alpha, M)$ for $1 < \alpha \leq 2$ where $f \in \mathcal{L}(\alpha, M)$ implies that $|f(x)| \leq M$, $|f'(x)| \leq M$, and

$$|f'(x) - f'(y)| \leq M|x - y|^{\alpha - 1} \tag{4}$$

for every x and y in the sample space.

For $g\colon [0,1]^2 \mapsto \mathbb{R}$, the analogous conditions hold with the Euclidean norm in $\mathbb{R}^2$ replacing absolute values. The partial derivatives exist and the vector of partials, $g'(\mathbf{x})$, is bounded and smooth.

$$\mathcal{L}(\alpha, M) = \Big\{ g : \sup_{\mathbf{x}\in[0,1]^2} |g(\mathbf{x})| \le M, \sup_{\mathbf{x}\in[0,1]^2} |g'(\mathbf{x})| \le M, \sup_{\mathbf{x},\mathbf{y}\in[0,1]^2} |g'(\mathbf{x}) - g'(\mathbf{y})| \le M|\mathbf{x}-\mathbf{y}|^{\alpha-1} \Big\}$$

Theorem 1 *For n equally spaced design points $\mathbf{x}^*_{i,j}$ in $[0,1]^2$,*

$$\mathbf{x}^*_{i,j} = \left(\frac{2i-1}{2\sqrt{n}}, \frac{2j-1}{2\sqrt{n}} \right),$$

let $\bar{\mathbb{Q}}_g$ be the distribution of the

$$Y_{i,j} = g(\mathbf{x}^*_{i,j}) + \sigma\xi_{i,j}$$

where σ is known, the $\xi_{i,j}$ are independent standard normals, and g is an unknown function from $[0,1]^2$ to $\mathbb{R}$ that is in $\mathcal{L}(\alpha, M)$.

Let $\mathbb{Q}_g$ be the distribution of the Gaussian process

$$Y(\mathbf{t}) = \int_0^{t_1}\int_0^{t_2} g(\mathbf{x})\, dx_2\, dx_1 + \frac{\sigma}{\sqrt{n}} W(\mathbf{t}) \tag{5}$$

where $W(\mathbf{t})$ is a Brownian sheet on the unit square.

Then there exists a randomization K_y such that

$$\sup_{g\in\mathcal{L}(\alpha,M)} \|\bar{\mathbb{Q}}_g K_y - \mathbb{Q}_g\| \le \frac{6M}{\sqrt{2\pi\sigma^2}} \left(n^{1/2-\alpha/2} + n^{-1/4} \right).$$

For $\alpha > 1$, this implies that the error made by performing the inference in the continuous experiment is asymptotically negligible. This is a reasonably sharp result in that Brown and Zhang (1998) showed that the experiments are not equivalent for $\alpha = 1$. Theorem 1 is proven is section 3.

The experiments are asymptotically equivalent because there is a simple transformation in the other direction, taking the increments of the Gaussian process, that produces a smaller error than in Theorem 1. The details are in section 4.3.

It will be instructive to first show the following result about density estimation experiments. Let $\mathbb{P}_f = P_f^n$ be the distribution of n independent observations from P_f a distribution on the unit interval with density f. Then let $\bar{\mathbb{P}}_f$

be the distribution that results from binning the n observations into m equal-length subintervals. The new observations have an m-dimensional multinomial distribution.

Theorem 2 *For a density $f \in \mathcal{F}(\alpha, M, \epsilon)$ such that $\mathcal{F}(\alpha, M, \epsilon) \subset \mathcal{L}(\alpha, M)$ and $f(x) > \epsilon$ for every x in the sample space, there exists a randomization K_x such that*

$$\sup_{f \in \mathcal{F}(\alpha, M, \epsilon)} \|\bar{\mathbb{P}}_f K_x - \mathbb{P}_f\| \leq 3M\epsilon^{-1/2} n^{1/2} \left(m^{-\alpha} + m^{-3/2} \right).$$

The implication is that the experiments are asymptotically equivalent as long as the number of bins is greater than $n^{-1/(2\alpha)}$ for $\alpha < 3/2$. The transformation in the other direction just bins the continuous observations to produce exactly $\bar{\mathbb{P}}_f$ from $\mathbb{P}_f$. Theorem 2 is proven in section 2.

Remark: The constants $6/\sqrt{2\pi}$ and 3 in these inequalities are not meant to be sharp, but they indicate that the statements are true for a reasonable size constant.

2 Binning in density estimation

Let $x_j^* = (2j-1)/2m$ denote the midpoint of the jth interval. For the n independent observations $X_1, \ldots, X_n$ with density f, let $X_i^* = x_j^*$ when X_i is in the sub-interval $[(j-1)/m, j/m]$. The $X_1^*, X_2^*, \ldots, X_n^*$ are independent observations from a discrete probability distribution P_θ with probabilities

$$\theta_j = \int_{\frac{j-1}{m}}^{\frac{j}{m}} f(x)\, dx \qquad j = 1, \ldots, m \tag{6}$$

on the points x_j^*.

The rounding off of observations onto a regular grid has computational advantages, see Silverman (1982) and Fan and Marron (1994). In Hall and Wand (1996), they demonstrated that kernel density estimators based on discretized versions of the data perform nearly as well as on the original data. They also discussed "common linear binning" which is related to our method, but differs in that their triangular kernel is part of the binning procedure while we use the kernel to smooth data that has already been binned.

A sufficient statistic for estimating the θ_j is the number of observations at each x_j^*. Let $\bar{X}_1, \ldots, \bar{X}_m$ be the counts at each x_j^*. The conditional distribution

of the X_i^* given the $\bar{X}_i$ is just the probability of each random ordering of n objects of m types. This conditional distribution does not depend on the θ_j, and the necessary randomization takes, for each j, $\bar{X}_j$ copies of x_j^* and then randomly assigns an index i to each of the n observations. Therefore, the first part of the randomization K_x chooses a random permutation to produce n independent observations from the discrete P_θ.

Theorem 2 can then be established by constructing a randomization from the X_i^* to the X_i. Working on each coordinate separately, the randomization chooses a new X_i to correspond to each X_i^*. By construction, if $X_i^* = x_j^*$ then X_i must have been in $[(j-1)/m, j/m]$, but we don't know anything else about X_i. Thus it seems appropriate to have K_x uniformly distributed over this subinterval. This results in a continuous marginal distribution that has a piecewise constant density (equal to $m\theta_j$ on the jth subinterval). A better approximation to the original distribution of f can be achieved using a K_x that is more spread resulting a smoother density (see section 4.)

Let K_j be the conditional distribution of the new X_i given that $X_i^* = x_j^*$. The marginal distribution $P_\theta K_x$ is a mixture of m distributions $K_1, \ldots, K_m$ with weights θ_j such that it has density

$$\frac{d(P_\theta K_x)}{d\lambda}(x) = \sum_{j=1}^{m} \theta_j \frac{dK_j}{d\lambda}(x) \equiv \hat{f}(x)$$

where λ is uniform on $[0,1]$. Repeating this randomization on the n independent observations defines K_x such that $\bar{\mathbb{P}}_f K_x = P_{\hat{f}}^n$. Therefore, the result of the randomization is n independent observations from a distribution with density $\hat{f}$.

The final step is to bound the total variation distance between n independent observations from f and $\hat{f}$ respectively. For product experiments it is worthwhile to use a Hellinger distance bound on the total variation distance

$$\|\mathbb{P} - \mathbb{Q}\| \leq \sqrt{2}\, H(\mathbb{P}, \mathbb{Q}) = \sqrt{2} \left(\frac{1}{2} \int \left(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}} \right)^2 \right)^{1/2},$$

because the squared Hellinger distance between product measures is bounded by the sum of the squared marginal distances,

$$H^2(P_{\hat{f}}^n, P_f^n) \leq n H^2(P_{\hat{f}}, P_f)$$

(see, for example, Strasser (1985, pp. 11–12).)

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

$$\frac{1}{2}\int \left(f^{1/2} - \hat{f}^{1/2}\right)^2 = \int \frac{(f-\hat{f})^2}{2(f^{1/2}+\hat{f}^{1/2})^2} \leq \frac{1}{2\epsilon}\left\|f - \hat{f}\right\|_2^2.$$

In section 4.1, it is shown that for $f \in \mathcal{L}(\alpha, M)$,

$$\left\|\hat{f} - f\right\|_2^2 \leq M^2 m^{-3} + 9M^2 m^{-2\alpha}. \tag{7}$$

Therefore,

$$\sup_{f\in\mathcal{F}(\alpha,M,\epsilon)} \|P_{\hat{f}}^n - \mathbb{P}_f^n\| \leq 3M\epsilon^{-1/2} n^{1/2}(m^{-\alpha} + m^{-3/2}).$$

and Theorem 2 is established.

3 Nonparametric regression

The nonparametric regression experiment in (1) has n observations that can be thought of as approximating the increments of the Brownian Motion process in (2) $Y_i \approx n\left[Y(i/n) - Y([i-1]/n)\right]$. Brown and Low (1996) showed that for $g \in \mathcal{L}(\alpha, M)$, $\alpha > 1/2$, these experiments are equivalent. Using the Y_i as approximations to the increments, a continuous Gaussian process can be constructed by interpolating independent Brownian bridges $B_i(t)$ on $[(i-1)/n, i/n]$ between the points $Y(i/n)$,

$$Y_{\bar{g}}(t) = \int_0^t \sum_{i=1}^n Y_i \mathbf{1}_{((i-1)/n, i/n]}\, dt + \sigma\frac{1}{n}\sum_{i=1}^n B_i(t).$$

This process has mean $\bar{g}(t) = g(x_i)$ for $(i-1)/m \leq t < i/m$, and for $0 < s < t < 1$, $\mathrm{Cov}\left(Y_{\bar{g}}(t), Y_{\bar{g}}(s)\right) = \frac{\sigma^2 s}{n}$. This result can be extended to the unit square if differentiable drift functions can be properly exploited, much as in the density estimation case.

3.1 The construction

The constructed process is a function of the Y_i and some independent continuous Gaussian processes on $[0,1]^2$. These centered processes are determined by their covariance functions (Dudley, 2002, Theorem 12.1.3). The Brownian sheet $W(\mathbf{t})$ is a centered Gaussian process with covariance function $C(\mathbf{t}, \mathbf{s}) =$

$(s_1 \wedge t_1)(s_2 \wedge t_2)$. We also need K_i-Brownian sheets W_{Ki} where the K_i are probability measures on $[0,1]^2$. Let κ_i be the cumulative distribution function (CDF) for the measure K_i. The covariance function for W_{Ki} is

$$\mathrm{Cov}\left(W_{Ki}(\mathbf{s}), W_{Ki}(\mathbf{t})\right) = \kappa_i(\mathbf{s} \wedge \mathbf{t}) \qquad \mathbf{s}, \mathbf{t} \in [0,1]^2$$

where $\mathbf{s} \wedge \mathbf{t} = ((s_1 \wedge t_1), (s_2 \wedge t_2))$. Furthermore, a K_i-Brownian bridge B_{Ki} can be constructed via W_{Ki} by

$$B_{Ki}(\mathbf{t}) = W_{Ki}(\mathbf{t}) - \kappa_i(\mathbf{t}) W_{Ki}(1,1).$$

The K_i-Brownian bridge has mean zero and covariance

$$\mathrm{Cov}\left(B_{Ki}(\mathbf{s}), B_{Ki}(\mathbf{t})\right) = \kappa_i(\mathbf{s} \wedge \mathbf{t}) - \kappa_i(\mathbf{s})\kappa_i(\mathbf{t}). \tag{8}$$

Hence $\mathrm{Var}(B_{Ki}(\mathbf{t})) = \kappa_i(\mathbf{t})(1 - \kappa_i(\mathbf{t}))$.

The construction of the Gaussian process $Y^*(\mathbf{t})$ from the Y_i's first generates n independent processes $B_{K1}, \ldots, B_{Kn}$, and then

$$Y^*(\mathbf{t}) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i \kappa_i(\mathbf{t}) + \sigma B_{Ki}(\mathbf{t}) \right].$$

The mean of this process is $\frac{1}{n} \sum g(\mathbf{x}_i^*) \kappa_i(\mathbf{t})$ and the covariance is

$$\begin{aligned} \mathrm{Cov}(Y^*(\mathbf{t}),\, Y^*(\mathbf{s})) &= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 \kappa_i(\mathbf{t})\kappa_i(\mathbf{s}) + \sigma^2 \left(\kappa_i(\mathbf{s} \wedge \mathbf{t}) - \kappa_i(\mathbf{s})\kappa_i(\mathbf{t}) \right) \\ &= \frac{\sigma^2}{n} \left(\frac{1}{n} \sum_{i=1}^{n} \kappa_i(\mathbf{s} \wedge \mathbf{t}) \right). \end{aligned}$$

Assuming an additional condition on the K_i measures,

$$\frac{1}{n} \sum_{i=1}^{n} \kappa_i(\mathbf{t}) = t_1 t_2, \tag{9}$$

fixes the covariance of Y_g^* to be the same as Y_g. The condition in (9) seems reasonable as it implies that if all the Y_i are the same then the resulting drift function is constant over the whole square. The kernels described in section 4.2 meet this criterion as do uniform distributions over disjoint subsets.

Therefore, Y^* actually has the same distribution as

$$Y_{\hat{g}}(\mathbf{t}) = \int_0^{t_1} \int_0^{t_2} \hat{g}(\mathbf{x})\, d\mathbf{x} + \frac{\sigma}{\sqrt{n}} W(\mathbf{t}).$$

where

$$\hat{g}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{x}_i^*)\, \frac{dK_i}{d\lambda}(\mathbf{x}).$$

This should be close in distribution to the Y_g if g and $\hat{g}$ are close.

3.2 A total-variation distance bound

Shifting two distributions by the same amount will not affect the total-variation distance between them. Thus, without loss of generality, we will find the total-variation distance between a process $Y_0(\mathbf{t})$ with mean 0 and variance σ^2/n, and a process $Y_\delta(\mathbf{t})$ with drift $\delta(\mathbf{t})$ and the same variance. Let $\mathbb{Q}_0$ and $\mathbb{Q}_\delta$ be the distributions of Y_0 and Y_δ respectively. The set of sample paths that achieves the supremum in (3) is the set where

$$\frac{d\mathbb{Q}_\delta}{d\mathbb{Q}_0} = \exp\left[\frac{n}{\sigma^2}\left(\int \delta(\mathbf{t})\, dY(\mathbf{t}) - \frac{1}{2}\|\delta\|_2^2\right)\right] > 1.$$

Let A be this set of continuous sample paths such that $\int\delta(\mathbf{t})\, dY(\mathbf{t}) > \frac{1}{2}\|\delta\|_2^2$. Under $\mathbb{Q}_0$, the integral $\int\delta(\mathbf{t})\, dY(\mathbf{t})$ has a normal distribution with mean 0 and variance $\sigma^2\|\delta\|_2^2/n$. Under $\mathbb{Q}_\delta$, the integral has mean $\|\delta\|_2^2$ and the same variance. Therefore the total variation distance is

$$\|\mathbb{Q}_0 - \mathbb{Q}_\delta\| = |\mathbb{Q}_0 A - \mathbb{Q}_\delta A| = 1 - 2\Phi(-\Delta/2) \tag{10}$$

where $\Delta = \sigma^{-1}n^{1/2}\|\delta\|_2$. The expression in (10) for the total variation distance is concave for positive Δ so a simple expansion gives

$$\|\mathbb{Q}_0 - \mathbb{Q}_\delta\| \leq \frac{1}{\sqrt{2\pi}}\Delta.$$

In the case of Y_g and $Y_{\hat{g}}$, the bound again depends essentially on the L_2 distance between the means,

$$\|\mathbb{Q}_g - \mathbb{Q}_{\hat{g}}\| \leq \frac{1}{\sqrt{2\pi}}\frac{\sqrt{n}}{\sigma}\|g - \hat{g}\|_2.$$

A bound on the L^2 distance is needed, in section 4.2 we will show that

$$\sup_{g\in\mathcal{L}(\alpha,M)} \|g - \hat{g}\|_2 \leq 6M\left(n^{-\alpha/2} + n^{-3/4}\right). \tag{11}$$

Therefore,

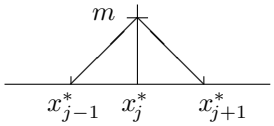
$$\sup_{g\in\mathcal{L}(\alpha,M)} \|\bar{\mathbb{Q}}_g K_y - \mathbb{Q}_g\| \leq \frac{6M}{\sqrt{2\pi\sigma^2}}\left(n^{1/2-\alpha/2} + n^{-1/4}\right)$$

which proves Theorem 1.

4 Bounding the L_2 distances

The K_j are probability measures with densities equal to linear interpolating functions:

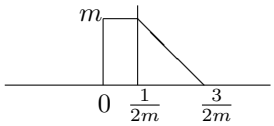
$\frac{dK_j}{d\lambda}$:



where $x^*_{j-1} = x^*_j - 1/m$ and $x^*_{j+1} = x^*_j + 1/m$.

For $j = 1$, we will need a different conditional density,

$\frac{dK_1}{d\lambda}$:



to avoid getting observations outside of $[0, 1]$. The analogous measure will be used for $j = m$ to avoid results greater than 1. Equivalently, we could use the triangular measures everywhere and then reflect any observations outside of the interval back in.

The average

$$\frac{1}{m}\sum_{j=1}^{m} a_j \, \frac{dK_j}{d\lambda}(x)$$

is a piecewise linear function that is equal to a_j at the midpoints x^*_j.

4.1 One dimension.

For the set of functions $f \in \mathcal{L}(\alpha, M)$ with continuous first derivatives ($\alpha > 1$), the error in a linear Taylor expansion is

$$|f(t+\delta) - f(t) - \delta f'(t)| = |f(t) + \delta f'(t^*) - f(t) - \delta f'(t)| \le M\delta^\alpha \qquad (12)$$

from (4). The existence of t^* such that $|t^* - t| \le \delta$ and $f(t+\delta) = f(t) + \delta f'(t^*)$ follows from the Mean Value Theorem.

Any x between $(2m)^{-1}$ and $1 - (2m)^{-1}$ lies between two of the grid points x^*_j and x^*_{j+1}. The difference between f and $\hat{f}$ at such an x is

$$\left|f(x) - \hat{f}(x)\right| \le \left|f(x^*_j) - \hat{f}(x^*_j)\right| + \left|x - x^*_j\right| \left|f'(x^*_j) - \hat{f}'(x^*_j)\right| + |E_1| + |E_2| \quad (13)$$

where E_1 and E_2 are the errors in expansions around x^*_j of f and $\hat{f}$ respectively. By (12), $|E_1| \le Mm^{-\alpha}$.

To bound the first term on the right in (13) The average over any interval $[(i-1)/m, i/m]$ can be approximated using an expansion around the midpoint x_j^*,

$$\begin{aligned}\hat{f}(x_j^*) &= m\int_{(i-1)/m}^{i/m} f(x)\,dx \\ &= m\int_{(i-1)/m}^{i/m} f(x_j^*) + (x-x_j^*)f'(x_j^*) + E_3\,dx \\ &= f(x_j^*) + E_4\end{aligned}$$

because $\int_{(i-1)/m}^{i/m}(x-x_j^*) = 0$. The bound in (12) implies that $|E_3| \leq Mm^{-\alpha}$ and thus $|E_4| \leq Mm^{-\alpha}$.

The second term on the right in (13) is problematic because the derivative $\hat{f}'(x_j^*)$ does not exist. However, if we are only interested in $x_j^* < x < x_{j+1}^*$ then defining $\hat{f}'(x_j^*) = m\left(\hat{f}(x_j^*) - \hat{f}(x_{j+1}^*)\right)$ makes $|E_2| \equiv 0$. Making two appeals to the Mean Value Theorem,

$$\begin{aligned}\hat{f}'(x_j^*) &= m\left(\hat{f}(x_j^*) - \hat{f}(x_{j+1}^*)\right) \\ &= m^2\int_{(i-1)/m}^{i/m} f(x) - f\left(x+m^{-1}\right)\,dx \\ &= m^2\int_{(i-1)/m}^{i/m} m^{-1}f'(\xi_x)\,dx \\ &= f'(\xi_j)\end{aligned}$$

where ξ_x and ξ_j are arbitrary points in the interval $[x_{j-1}^*, x_{j+1}^*]$ that make the statements true. Therefore, the second term is

$$\left|x-x_j^*\right|\left|f'(x_j^*) - \hat{f}'(x_j^*)\right| \leq \frac{1}{m}\left|f'(x_j^*) - f'(\xi_j)\right| \leq Mm^{-\alpha},$$

and each of the terms in (13) is bounded

$$\left|f(x) - \hat{f}(x)\right| \leq |E_4| + Mm^{-\alpha} + |E_1| + 0 \leq 3Mm^{-\alpha}. \tag{14}$$

This argument does not work at the edges. If $x < (2m)^{-1}$ then $\hat{f}(x) = \hat{f}([2m]^{-1})$ and

$$\begin{aligned}\left|f(x) - \hat{f}(x)\right| &\leq \left|\hat{f}([2m]^{-1}) - f([2m]^{-1})\right| + |x-x_j^*||f'(x_j^*)| + |E_1| \\ &\leq 2Mm^{-\alpha} + \frac{1}{2}Mm^{-1}\end{aligned} \tag{15}$$

because $|f'(x_j^*)| < M$ by assumption. The same argument works for $x > 1 - (2m)^{-1}$.

Squaring the point-wise bounds in (14) and (15) then integrating,

$$\left\|\hat{f} - f\right\|_2^2 = \int_0^{1/(2m)} (\hat{f} - f)^2 + \int_{1/(2m)}^{1-1/(2m)} (\hat{f} - f)^2 + \int_{1-1/(2m)}^{1} (\hat{f} - f)^2$$
$$\leq m^{-1} M^2 m^{-2} + 9M^2 m^{-2\alpha}. \tag{16}$$

Therefore, the bound needed in (7) is established.

4.2 Two dimensions.

For regression mean functions $g(\mathbf{x})$ on $[0,1]^2$, we will use kernels that are products of the one-dimensional kernels above. The CDF of the kernel $K_{i,j}$ is $\kappa_{i,j}(\mathbf{x}) = \kappa_i(x_1)\kappa_j(x_2)$ where κ_i and κ_j are the CDF's of K_i and K_j respectively.

In $\mathbb{R}^2$, the differentiability of g gives

$$g(\mathbf{x}) = g(\mathbf{x}_{i,j}^*) + (\mathbf{x} - \mathbf{x}_{i,j}^*)^T g'(\mathbf{x}_{i,j}^*) + E_1, \tag{17}$$

and the Lipshitz condition implies that the error is bounded by

$$|E_1| \leq |\mathbf{x} - \mathbf{x}_{i,j}^*||g'(\mathbf{x}_{i,j}^*) - g'(\boldsymbol{\xi})| \leq M|\mathbf{x} - \mathbf{x}_{i,j}^*|^\alpha \tag{18}$$

as in the one dimensional case.

The mean value of the constructed process $\hat{g}$ is no longer linear, but the quadratic part is small. Consider points $\mathbf{x}$ in the interior of the square formed by the four midpoints $\mathbf{x}_{i,j}^*, \mathbf{x}_{i+1,j}^*, \mathbf{x}_{i,j+1}^*$, and $\mathbf{x}_{i+1,j+1}^*$.

$$\hat{g}(\mathbf{x}) = g(\mathbf{x}_{i,j}^*) + \sqrt{n}\left(\mathbf{x} - \mathbf{x}_{i,j}^*\right)^T \begin{pmatrix} g(\mathbf{x}_{i+1,j}^*) - g(\mathbf{x}_{i,j}^*) \\ g(\mathbf{x}_{i,j+1}^*) - g(\mathbf{x}_{i,j}^*) \end{pmatrix} + E_2 \tag{19}$$

The differences can be written as

$$g(\mathbf{x}_{i+1,j}^*) - g(\mathbf{x}_{i,j}^*) = \begin{pmatrix} \frac{1}{\sqrt{n}} & 0 \end{pmatrix} g'(\mathbf{x}_{i,j}^*) + E_3,$$
$$g(\mathbf{x}_{i,j+1}^*) - g(\mathbf{x}_{i,j}^*) = \begin{pmatrix} 0 & \frac{1}{\sqrt{n}} \end{pmatrix} g'(\mathbf{x}_{i,j}^*) + E_4.$$

The bound in (18) means the errors $|E_3|$ and $|E_4|$ are less than $Mn^{-\alpha/2}$.

The error term E_2 in (19) is the quadratic component of the mean function. Let $(\zeta_1, \zeta_2) = (\mathbf{x} - \mathbf{x}_{i,j}^*)^T$, then

$$E_2 = n\zeta_1\zeta_2\left[g(\mathbf{x}_{i,j}^*) - g(\mathbf{x}_{i+1,j}^*) - g(\mathbf{x}_{i,j+1}^*) + g(\mathbf{x}_{i,j+1}^*)\right]$$

The size of $|E_2|$ is bounded using $n|\zeta_1\zeta_2| \leq 1$ and

$$\left|g(\mathbf{x}^*_{i,j}) - g(\mathbf{x}^*_{i+1,j}) - g(\mathbf{x}^*_{i,j+1}) + g(\mathbf{x}^*_{i+1,j+1})\right| \leq n^{-1/2}|g'(\mathbf{x}^*_{i,j+1}) - g'(\mathbf{x}^*_{i,j})| + |E_3| + |E_5|$$

where E_5 is the error in approximating the difference $g(\mathbf{x}^*_{i+1,j+1}) - g(\mathbf{x}^*_{i,j+1})$ by the derivative at $\mathbf{x}^*_{i,j+1}$. Thus, $|E_5| \leq Mn^{-\alpha/2}$ and $|E_2| \leq 3Mn^{-\alpha/2}$.

Putting together (17) and (19)

$$\begin{aligned}|g(\mathbf{x}) - \hat{g}(\mathbf{x})| \leq \Big| & g(\mathbf{x}^*_{i,j}) + (\mathbf{x} - \mathbf{x}^*_{i,j})^T g'(\mathbf{x}^*_{i,j}) + E_1 + \\ - & \left[g(\mathbf{x}^*_{i,j}) + (\mathbf{x} - \mathbf{x}^*_{i,j})^T g'(\mathbf{x}^*_{i,j}) + n^{1/2}\zeta_1 E_3 + n^{1/2}\zeta_2 E_4 + E_2\right] \Big| \leq 6Mn^{-\alpha/2}\end{aligned}$$

for $\mathbf{x}$ not within $\frac{1}{2}n^{-1/2}$ of the edge of the square. The contribution from the center of the square to $\|g - \hat{g}\|_2^2$ is therefore less than $36M^2n^{-\alpha}$.

To see what happens near the edges of the square, consider a point $\mathbf{x}$ where the first coordinate is less than $\frac{1}{2}n^{-1/2}$, then the $\hat{g}$ function is exactly

$$\hat{g}(\mathbf{x}) = g(\mathbf{x}^*_{i,j}) + \sqrt{n}\left(\mathbf{x} - \mathbf{x}^*_{i,j}\right)^T \begin{pmatrix} g(\mathbf{x}^*_{i+1,j}) - g(\mathbf{x}^*_{i,j}) \\ 0 \end{pmatrix}$$

Thus,

$$|g(\mathbf{x}) - \hat{g}(\mathbf{x})| \leq |E_1| + |E_3| + \frac{M}{2}n^{-1/2} \leq \frac{1}{2}Mn^{-1/2} + \frac{3}{2}Mn^{-\alpha/2}$$

where $\frac{1}{2}Mn^{-1/2}$ bounds the contribution from the second coordinate of $g'(\mathbf{x})$ because $|g'(\mathbf{x})| \leq M$. An analogous bound can be put on the errors along the other sides.

Finally, if $x_1 < \frac{1}{2}n^{-1/2}$ and $x_2 < \frac{1}{2}n^{-1/2}$, then $\hat{g} = g(\mathbf{x}^*_{i,j})$ and thus $|g(\mathbf{x}) - g(\mathbf{x}^*_{i,j})| \leq M(2n)^{-1/2}$. The same bound applies in the other three corners.

The total contribution to $\|g - \hat{g}\|_2^2$ from all the area near the edges is less than $M^2n^{-3/2}/2$ because the total area is less than $2n^{-1/2}$.

Therefore,

$$\|g - \hat{g}\|_2^2 \leq \frac{1}{2}M^2n^{-3/2} + 36M^2n^{-\alpha}$$

which establishes (11).

4.3 The other direction

Asymptotic equivalence also requires a transformation in the other direction: a way to generate the n regression observations from the continuous Gaussian

process. The transformation uses the increments of the process over each square,

$$Y^*_{i,j} = n\left[Y\left(\tfrac{i}{\sqrt{n}}, \tfrac{j}{\sqrt{n}}\right) - Y\left(\tfrac{i-1}{\sqrt{n}}, \tfrac{j}{\sqrt{n}}\right) - Y\left(\tfrac{i}{\sqrt{n}}, \tfrac{j-1}{\sqrt{n}}\right) + Y\left(\tfrac{i-1}{\sqrt{n}}, \tfrac{j-1}{\sqrt{n}}\right)\right].$$

Let $I_{i,j} = \left\{(x_1, x_2) : \frac{i-1}{\sqrt{n}} < x_1 \leq \frac{i}{\sqrt{n}}, \frac{j-1}{\sqrt{n}} < x_2 \leq \frac{j}{\sqrt{n}}\right\}$. Then the increments $Y^*_{i,j}$ are independent with variance σ and have mean $n \iint_{Iij} g \, d\mathbf{x}$.

The only error then is the difference between $g(\mathbf{x}^*_{i,j})$ and this average.

$$\begin{aligned} g(\mathbf{x}^*_{i,j}) - n \iint_{I_{i,j}} g(\mathbf{x}) \, d\mathbf{x} &= n \iint_{I_{i,j}} g(\mathbf{x}^*_{i,j}) - g(\mathbf{x}) \, d\mathbf{x} \\ &= n \iint_{I_{i,j}} (\mathbf{x} - \mathbf{x}^*_{i,j})^T g'(\mathbf{x}^*_{i,j}) + E_1 \, d\mathbf{x} \\ &= n \iint_{I_{i,j}} E_1 \, d\mathbf{x} \end{aligned}$$

Therefore, by (18), $(g(\mathbf{x}^*_{i,j}) - n \iint_{I_{i,j}} g(\mathbf{x}) \, d\mathbf{x})^2 \leq Mn^{-\alpha}$. The total variation distance between the joint distributions of all n observations is therefore less than $(\sqrt{2\pi}\sigma)^{-1} M n^{-\alpha/2+1/2}$.

This error is less than that made in the other direction and therefore whenever the bound in Theorem 1 goes to 0, the experiments are asymptotically equivalent.

4.4 Higher dimensions

The same technique could provide a solution in the three-dimensional case if there is an added restriction to the class of parameter functions in order to lessen the edge effects.

The kernel would still be a product of the one-dimensional kernels in each of the dimensions $K(\mathbf{x}) = K_i(x_1)K_j(x_2)K_k(x_3)$. The errors in the inner part of the cube are of order $O(n^{-\alpha/3})$.

The points near the edge of the cube however would make an error of order $n^{-1/3}$ in at least one of the three coordinates. This contributes a term of order $O(n^{-1})$ to the squared difference between g and $\hat{g}$. The result is that the total-variation distance between the Gaussian processes will not converge to 0. It is possible to impose conditions on the drift functions to minimize these edge effects. For example, imposing a periodic boundary condition and adjusting the edge kernels to be periodic makes the error $O(n^{-\alpha/3})$ everywhere. This is sufficient for asymptotic equivalence if $\alpha > 3/2$.

Higher dimensions would require $\alpha > 2$ (see Brown and Zhang (1998)), and our methods cannot take advantage of this added smoothness. Higher-order interpolating kernels are negative in places, and the requirement that the kernels have a positive density function is critical in the construction.

Using a Fourier series expansion, Rohde (2004) was able to take advantage of arbitrary amounts of smoothness in the case of periodic functions in the nonparametric regression problem, and it is likely those techniques could be extended to higher dimensional problems.

References

Brown, L., Cai, T., Low, M. and Zhang, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30** 688–707.

Brown, L. D., Carter, A. V., Low, M. G. and Zhang, C.-H. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.* **32** 2074–2097.

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

Brown, L. D. and Zhang, C.-H. (1998). Asymptotic nonequivalence of nonparametric experiments when the smoothness index is 1/2. *Ann. Statist.* **26** 279–287.

Carter, A. V. (2002). Deficiency distance between multinomial and multivariate normal experiments. *Ann. Statist.* **30** 708–730.

Donoho, D. L. and Johnstone, I. M. (1999). Asymptotic minimaxity of wavelet estimators with sampled data. *Statistica Sinica* **9** 1–32.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymoptopia? *Journal of the Royal Statistical Society Series B* **57** 301–369.

Dudley, R. M. (2002). *Real analysis and probability.* Cambridge University Press.

EFROMOVICH, S. (1999). *Nonparametric Curve Estimation, Methods, Theory, and Application.* Springer-Verlag.

FAN, J. and MARRON, J. S. (1994). Fast implementation of nonparametric curve estimators. *J. Comput. Graph. Statist.* **3** 35–56.

GRAMA, I. and NUSSBAUM, M. (1998). Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Related Fields* **111** 167–214.

HALL, P. and WAND, M. P. (1996). On the accuracy of binned kernel density estimators. *Journal of Multivariate Analysis* **56** 165–184.

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer–Verlag, New York.

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.

ROHDE, A. (2004). On the asymptotic equivalence and rate of convergence of nonparametric regression and Gaussian white noise. *Statistics & Decisions* **22** 235–243.

SILVERMAN, B. W. (1982). Algorithm AS 176: Kernel density estimation using the fast Fourier transform. *Applied Statistics* **31** 93–99.

STRASSER, H. (1985). *Mathematical Theory of Statistics, Statistical Experiments and Asymptotic Decision Theory*, vol. 7 of *de Gruyter Studies in Mathematics.* Walter de Gruyter, New York.