ADDITIVE REGRESSION
AND OTHER NONPARAMETRIC MODELS

BY

CHARLES J. STONE

TECHNICAL REPORT NO. 33
OCTOBER 1984
(1/31/84)

RESEARCH PARTIALLY SUPPORTED BY
NATIONAL SCIENCE FOUNDATION GRANT MCS83-01257

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA
ADDITIVE REGRESSION AND OTHER NONPARAMETRIC MODELS

By Charles J. Stone
University of California, Berkeley

Summary

Let \((X,Y)\) be a pair of random variables such that \(X = (X_1, \ldots, X_J)\) and let \(f\) be a function that depends on the joint distribution of \((X,Y)\). A variety of parametric and nonparametric models for \(f\) are discussed in relation to flexibility, dimensionality, and interpretability. It is then supposed that each \(X_j \in [0,1]\), that \(Y\) is real valued with mean \(\mu\) and finite variance, and that \(f\) is the regression function of \(Y\) on \(X\). Let \(f^*\), of the form \(f^*(x_1, \ldots, x_J) = \mu + f_1^*(x_1) + \ldots + f_J^*(x_J)\), be chosen subject to the constraints \(E \ f_j^* = 0\) for \(1 \leq j \leq J\) to minimize \(E[(f(X) - f^*(X))^2]\). Then \(f^*\) is the closest additive approximation to \(f\), and \(f^* = f\) if \(f\) itself is additive. Spline estimates of \(f_j^*\) and its derivatives are considered based on a random sample from the distribution of \((X,Y)\). Under a common smoothness assumption on \(f_j^*\), \(1 \leq j \leq J\), and some mild auxiliary assumptions, these estimates achieve the same (optimal) rate of convergence for general \(J\) as they do for \(J = 1\).

1This research was supported in part by National Science Foundation Grant MCS 83-01257.

AMS 1980 subject classifications. Primary 62G20; secondary 62G05.

Key words and phrases. Nonparametric model, additivity, spline, rate of convergence.
1. Introduction. Let \((X,Y)\) be a pair of random variables, each ranging over some space, and let \(f\) be a function that depends on the joint distribution of \(X\) and \(Y\). As examples, \(f\) could be the regression function of \(Y\) on \(X\), the hazard rate of the conditional distribution of \(Y\) given \(X\), or the density of \(X\); alternatively, \(f\) could be the logarithm or logit of any of these functions, where \(\text{logit}(p) = \log(p/(1-p))\).

Suppose as usual that the joint distribution of \(X\) and \(Y\) is unknown and consider the problem of estimating \(f\) based on a random sample from this joint distribution. The parametric approach starts with the assumption of an a priori model for \(f\) that contains only finitely many unknown parameters, while the nonparametric approach eschews such an assumption.

Three fundamental aspects of statistical models are flexibility, dimensionality, and interpretability. Flexibility is the ability of the model to provide accurate fits in a wide variety of situations, inaccuracy here leading to bias in estimation. Dimensionality can be thought of in terms of the variance in estimation, the "curse of dimensionality" being that the amount of data required to avoid an unacceptably large variance increases rapidly with increasing dimensionality. In practice there is an inevitable trade-off between flexibility and dimensionality or, as usually put, between bias and variance. Interpretability lies in the potential for shedding light on the underlying structure.

In Section 2, parametric models for \(f\) are discussed in relation to these three aspects. Nonparametric and semiparametric models are similarly discussed in Section 3, where a heuristic dimensionality reduction principle
is presented. In Section 4, a precise special case of this principle is stated for additive regression functions and, more generally, for additive approximations to a not necessarily additive regression function. The proof is given in Sections 5 and 6.

2. Parametric models. Let \( X = (X_1, \ldots, X_J) \), where \( X_j \subseteq X_j \) for \( 1 \leq j \leq J \). Suppose each \( X_j \) is either a finite set or an interval (i.e., a nondegenerate subinterval of \( \mathbb{R} \)). If \( X_j \) is finite \( X_j \) is called a "factor", which can be either quantitative or qualitative; in either case the elements of \( X_j \) are referred to as the possible levels of \( X_j \). If there are any factors present, the remaining \( X_j \)'s are called "covariates".

Suppose now that the \( X_j \)'s are each intervals. Two common parametric models for \( f \) are

Model 1

\[
f(x_1, \ldots, x_J) = a + \sum_{j=1}^{J} b_j x_j
\]

and

Model 2

\[
f(x_j, \ldots, x_J) = a + \sum_{j=1}^{J} b_j x_j + \sum_{j=1}^{J} \sum_{k=j+1}^{J} c_{jk} x_j x_k.
\]

If \( f \) is the regression function, Model 1 corresponds to linear regression and Model 2 corresponds to quadratic regression; if \( f \) is the logit of the regression function, Model 1 corresponds to logistic regression; and if \( f \) is the logarithm of the density of \( X \), Model 2 corresponds to the multivariate normal density.

Suppose instead that the \( X_j \)'s are each finite sets. Here it is common to consider

Model 3

\[
f(x_1, \ldots, x_J) = \mu + \sum_{j=1}^{J} f_j(x_j),
\]
which is an additive function of the various factors, or

Model 4 \[ f(x_1, \ldots, x_j) = \mu + \sum_1^J f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k), \]

which also includes the two-factor interaction terms. If \( f \) is the regression function, these models are considered in analysis of variance (see Scheffé, 1959), while if \( f \) is the (discrete) density of \( X \), these models arise in discrete multivariate analysis (see Bishop et al., 1975 and Haberman, 1978). In such models the functions \( f_j \) and \( f_{jk} \) will be referred to as functional components.

More generally, suppose that the first \( L \) of the \( X_j \)'s are finite sets and the remaining \( X_j \)'s are intervals, where \( 1 \leq L < J \). (Such situations are especially common in Social Science.) In this context it is natural to consider models such as those that arise in the analysis of covariance; e.g.,

Model 5 \[ f(x_1, \ldots, x_j) = \mu + \sum_1^L f_j(x_j) + \sum_{L+1}^J b_j x_j. \]

These and similar models are the bread-and-butter of applied statistics, even though they are rarely justified on theoretical grounds. The orthodox approach to model selection is by means of hypothesis testing; e.g. Model 1 vs. Model 2, or Model 3 vs. Model 4. But this approach has its limitations. In particular one does not know in advance that Model 2 or Model 4 is valid; and there may not be enough data to perform such a test (see Scheffé, 1959, for a discussion of the one-degree-of-freedom test for additivity due to Tukey, 1949). Often there is not enough data to fit high-dimensional models such as Model 2 or 4, so lower-dimensional models such as Model 1, 3 or 5 are automatically assumed (as in models involving Latin and Greco-Latin squares and balanced incomplete blocks). In the chapter on Latin squares John (1971) discusses bias in estimation of treatment effects due to ignored
interaction effects. On the other hand, Fisher and McDonald (1978) assert that, when there is enough data to test for the presence of interaction effects, in a large fraction (but by no means all) of the time they are found not to be statistically significant.

Even when there is enough data to fit models such as Model 2 or 4 and when the presence of interaction terms is detected, say, by hypothesis testing, additive models such as Models 1, 3 and 5 may be preferable because of greater interpretability. In particular, in the regression context additive models allow for the following interpretation: if $x_j$ is changed to $x'_j$, then all other variables remaining the same, the mean of $Y$ is increased by an amount which depends only on $x_j$ and $x'_j$; under the linear, additive Model 1 the amount of increase is a linear function of $(x'_j-x_j)$. In their discussion of additive regression models involving two factors, Box et al. (1978) state that

While blind faith in a particular model is foolhardy, refusal to associate data with any model is to eschew a powerful tool. As implied earlier, a middle course may be followed. On the one hand, inadequacies in proposed models should be looked for; on the other, if a model appears reasonably appropriate, advantage should be taken of the greater simplicity and clarity of interpretation that it provides.

Models 3 and 4 can be fitted by solving the normal equations for the associated dummy variable regression problem. Andrews et al. (1967) described a computer program, Multiple Classification Analysis (MCA), which uses instead a computationally efficient iterative method to fit the additive Model 3. Morgan and Sonquist (1963) proposed a tree structured alternative to Model 4 for detecting and interpreting interaction effects. Their automatic interaction detection technique was implemented in a program called AID (see Sonquist, Baker and Morgan, 1971; Fielding, 1977; and the discussion of a similar program, CART, in Breiman et al., 1984). Sonquist (1970) compared AID with MCA and suggested a strategy for using the two programs.
jointly. Both programs have been very popular among sociologists. They were developed by the Institute for Social Research, University of Michigan, and are currently available in the software package OSIRIS distributed by that organization. (SEARCH, in that package, is an outgrowth of AID.)

3. Nonparametric and semiparametric models. Suppose now that the $X_j$'s are each intervals. (The following discussion also applies, with obvious modifications, when some of these sets are intervals and the others are finite.) At the opposite end of the spectrum from Models 1 and 2, $f$ can be totally unrestricted (perhaps subject to some smoothness assumptions). Alternatively, one can consider, say, the additive Model 3 with unrestricted functional components. As a further step toward parametric modelling, one could impose parametric restrictions on some of these functional components as in Model 5. For example, if $f$ is the logarithm of the hazard rate of the conditional distribution of the failure time $Y$ given $X$, then the additive model given by

Model 6

$$f(x_1, \ldots, x_j, y) = g(y) + \sum_j b_j x_j$$


An interesting generalization of the additive model is

Model 7

$$f(x_1, \ldots, x_j) = \sum_j f(x_j(b_j x_j)),$$

which arises in projection pursuit regression (PPR) and (log)density estimation (see Friedman and Stuetzle, 1981; Friedman, Stuetzle and Schroeder, 1984; Friedman, Gross and Stuetzle, 1983; and Huber, 1983). Like Model 4, this model is more flexible than the additive model. Indeed, roughly
speaking, the projection pursuit model is so flexible that if \( V \) is permitted to be indefinitely large, it can yield an arbitrarily good approximation to any given function. The individual functional components \( f_\nu \) occurring in the projection pursuit model are readily graphed; but when \( V > 1 \), the model itself is hard to interpret. A less flexible, but more easily interpretable, generalization of the additive model is

\[
\text{Model 8} \quad f(x_1, \ldots, x_j) = g(\sum_j f_j(x_j)),
\]

which was considered by Winsberg and Ramsay (1980) (see also de Leeuw, et al., 1976; Young, et al., 1976; and Breiman and Friedman, 1985).

Friedman and Stuetzle (1981) also discussed "projection selection", which is PPR restricted to be additive, and they suggested strategies for using projection selection and unrestricted PPR jointly. Tibshirani (1983) discussed the extension of Cox's proportional hazards model (Model 6) to the unrestricted additive model. Hastie (1983) discussed the additive model for the logistic regression function when \( Y \) is an indicator variable, and Hastie (1984) pointed out the usefulness in graphical model diagnostics of additive estimates of the regression function or of its logit. Previously, Breiman and Stone (1978) proposed various additive regression estimates as modifications of (nonadditive) tree structured regression. They were motivated both by the successful meteorological application of an ad hoc additive regression technique (Zeldin and Thomas, 1975) and by the realization that unrestricted multivariate nonparametric regression is hard to interpret as well as being subject to the curse of dimensionality.

It is convenient to think of an arbitrary function of \( d \) real variables as being "d-dimensional". Consider a nonparametric model in which \( f \) is defined explicitly in terms of other functions, at least one of which is d-dimensional and none of which are more than d-dimensional.
Such a model will also be thought of as being d-dimensional. Accordingly, Models 3, 5, 6, 7 (for fixed $V$) and 8 are one-dimensional, while Model 4 is two-dimensional.

Model dimensionality, as just defined, is relevant to optimal rates of convergence for nonparametric function estimation (see Stone, 1982). In the absence of a restrictive model for $f$, the optimal rate of convergence in an $L^2$ norm is typically of the form $n^{-2r}$; here $r = (p-m)/(2p+J)$, $p$ being a measure of the assumed smoothness of $f$ and $m$ being the order of the derivative of $f$ that is being estimated ($m = 0$ if $f$ itself is being estimated). It is quite plausible that under the additional restriction of a $d$-dimensional model for $f$ ($d < J$), the optimal rate of convergence is of the same form with $J$ replaced by $d$ in the definition of $r$. But it is not at all obvious that this heuristic dimensionality reduction principle can be established rigorously as a general theorem applying, say, to projection pursuit regression. (Huber, 1985, points out that the sampling theory of PPR is practically nonexistent.) The discussion in Section 4 below suggests more general versions of this principle. The principle is implicit in Question 2 of Stone (1982).

4. Rates of convergence of additive regression. Suppose from now on that $X_j = [0,1]$ for $1 \leq j \leq J$ and that $EY^2 < \infty$. Let $f$ be the regression function of $Y$ on $X$; so that $f(x) = E(Y|X=x)$ for $x \in C = [0,1]^J$. Set $\mu = EY = Ef(X)$.

If $f$ is additive, it can be written in the form

$$f(x_1, \ldots, x_J) = \mu + \Sigma_1^J f_j(x_j),$$

where $Ef_j(X_j) = 0$ for $1 \leq j \leq J$. Under the mild Condition 1 below
the functional components \( f_j \) are uniquely determined up to sets of measure zero (see Lemma 1 in Section 5); and there is at most one continuous version of each such function.

Even if \( f \) is not genuinely additive, an additive approximation to \( f \) may be sufficiently accurate for a given application as well as being readily interpretable. Let \( f_j^* \), \( 1 \leq j \leq J \), be chosen subject to the constraints \( \text{Ef}_j^*(X_j) = 0 \) for \( 1 \leq j \leq J \) to minimize \( E[(f^*(X) - f(X))^2] \), where

\[
f^*(x_1, \ldots, x_J) = \mu + \sum_{j=1}^{J} f_j^*(x_j).
\]

(These functions exist under Condition 1 below--see Lemma 1 in Section 2.) Then \( f^* \) is the closest additive approximation to \( f \) in the sense of mean squared error. Again, under Condition 1 below, the functional components \( f_j^* \) are uniquely determined up to sets of measure zero and there is at most one continuous version of each such function. If \( f \) is genuinely additive, of course, then \( f^* = f \) and \( f_j^* = f_j \) for \( 1 \leq j \leq J \).

A smoothness assumption, Condition 3 below, will be imposed on the functions \( f_j^* \), \( 1 \leq j \leq J \). Then a special case of the dimensionality reduction principle will be established by determining the rates of convergence for spline estimates of \( f_j^* \) and its derivatives and for the corresponding estimates of \( f^* \) based on a random sample from the distribution of \( (X,Y) \).

Three conditions are required for the statements of the main results.

**CONDITION 1.** The distribution of \( X \) is absolutely continuous and its density \( g \) is bounded away from zero and infinity on \( \mathcal{C} \).

It follows from Condition 1 that the marginal density \( g_j \) of \( X_j \).
is bounded away from zero and infinity on \([0,1]\). Set 
\[ \text{Var}(Y|X=x) = E((Y-f(x))^2|X=x). \]

CONDITION 2. \( f \) and \( \text{Var}(Y|X=*) \) are bounded on \( C \).

Let \( k \) be a nonnegative integer, let \( \beta \in (0,1] \) be such that 
\( p = k + \beta > .5 \), and let \( M \in (0,\infty) \). Let \( H \) denote the collection of 
functions \( h \) on \([0,1]\) whose \( k \)th derivative, \( h^{(k)} \), exists and 
satisfies the Hölder condition with exponent \( \beta \):
\[ |h^{(k)}(t')-h^{(k)}(t)| \leq M|t'-t|^{\beta} \quad \text{for } 0 \leq t, t' \leq 1. \]

CONDITION 3. \( f^* \in H \) for \( 1 \leq j \leq J \).

Let \((X_1,Y_1),(X_2,Y_2),\ldots\) denote independent random pairs, each having 
the same distribution as \((X,Y)\) and write \( X_i \) as \((X_{i1},\ldots,X_{ij})\). We 
will consider additive spline estimates \( \hat{f}_n \) of \( f^* \) based on the random 
sample \((X_1,Y_1),\ldots,(X_n,Y_n)\) of size \( n \). Set \( \overline{Y}_n = (Y_1+\ldots+Y_n)/n \).

Let \( N_n \) denote a positive integer and let \( I_{n\nu} \), \( 1 \leq \nu \leq N_n \), 
denote the subintervals of \([0,1]\) defined by 
\( I_{n\nu} = [(\nu-1)/N_n,\nu/N_n) \) 
for \( 1 \leq \nu < N_n \) and \( I_{nN_n} = [1-N_n^{-1},1] \). Let \( \ell \) and \( \ell' \) be integers 
such that \( \ell \geq k \) and \( \ell > \ell' \). Let \( S_n \) denote the collection of functions 
\( s \) on \([0,1]\) such that
\[ \begin{align*}
(i) & \quad \text{the restriction of } s \text{ to } I_{n\nu} \text{ is a polynomial of degree} \\
& \quad \ell \text{ (or less) for } 1 \leq \nu \leq N_n; \\
\end{align*} \]
and, if \( \ell' > 0 \),
\[ \begin{align*}
(ii) & \quad s \text{ is } \ell' \text{-times continuously differentiable on } [0,1]. \\
\end{align*} \]

A function satisfying (i) is called a piecewise polynomial; if 
\( \ell = 0 \), it is piecewise constant. A function satisfying (i) and (ii)
is called a spline. Typically, splines are considered with \( l' = l - 1 \) and then called linear, quadratic or cubic splines accordingly as \( l = 1, 2 \) or \( 3 \). Wegman and Wright (1983) give a broad survey of works on splines in statistics.

Let \( \hat{f}_n \), of the form

\[
\hat{f}_n(x_1, \ldots, x_J) = \bar{\gamma}_n + \sum_{j=1}^{J} \hat{f}_{nj}(x_j),
\]

be chosen subject to the constraints that \( \hat{f}_{nj} \in \mathbb{S}_n \) and \( \sum_{j=1}^{n} \hat{f}_{nj}(x_{ij}) = 0 \) for \( 1 \leq j \leq J \) to minimize the residual sum of squares \( \sum_{i=1}^{n} (Y_i - \hat{f}_n(x_i))^2 \). We call the estimate \( \hat{f}_n \) of \( f^* \) an additive spline. If \( l = 0 \), we borrow from Tukey (1961) and call \( \hat{f}_n \) an additive regressogram. The numerical minimization is readily solved by using B-splines (see de Boor, 1978, or Powell, 1981); either the normal equations can be solved or iterative techniques can be employed as in MCA. It follows easily from Lemma 3 in Section 5 that under Condition 1 and the condition on \( N_n \) in Theorem 1 below, the solutions \( \hat{f}_{nj}, 1 \leq j \leq J \), to the constrained minimization problem are uniquely determined except on an event whose probability tends to zero with \( n \).

Given positive numbers \( a_n \) and \( b_n \), \( n \geq 1 \), let \( a_n \sim b_n \) mean that \( a_n/b_n \) is bounded away from zero and infinity. Given random variables \( Z_n, n \geq 1 \), let \( Z_n \in \mathbb{P}_{n}(b_n) \) mean that the random variables \( b_n^{-1}Z_n, n \geq 1 \), are bounded in probability or, equivalently, that

\[
\lim_{c \to \infty} \limsup_{n} \Pr(\{ |Z_n| > cb_n \}) = 0.
\]

Let \( \|\varphi\| \) denote the \( L^2 \) norm of a function \( \varphi \) on \( C \), defined by \( \|\varphi\|^2 = \mathbb{E} \varphi^2(x) = \int_{C} \varphi^2(x)g(x)dx \). For \( 1 \leq j \leq J \) let \( \|h\|_j^2 \)}
the $L^2$ norm of a function $h$ on $[0,1]$, defined by

$$\|h\|_2^2 = \int_0^1 h^2(x)g_j(x)dx.$$  

Set $\gamma = 1/(2p+1)$ and $r = p/(2p+1)$. 

Let $m$ be a nonnegative integer such that $m < k$ and set $r_m = (p-m)/(2p+1)$.

**THEOREM 1.** Suppose that Conditions 1-3 hold and that $N_n \sim n^\gamma$. Then

\begin{align*}
(1) & \quad E(\|\hat{f}_{m,n} - f_j^*\|_2^2 | x_1, \ldots, x_n) = O_pr(n^{-2r_m}) \quad \text{for } 1 \leq j \leq J.
\end{align*}

Since $Y$ has mean $\mu$ and finite second moment,

\begin{align*}
(2) & \quad E((\hat{Y}_n - \mu)^2 | x_1, \ldots, x_n) = O_pr(n^{-1}) = O_pr(n^{-2r}) .
\end{align*}

Thus Theorem 1 has the following consequence.

**COROLLARY 1.** Suppose that Conditions 1-3 hold and that $N_n \sim n^\gamma$. Then

\begin{align*}
(3) & \quad E(\|\hat{f}_{m,n} - f_j^*\|_2^2 | x_1, \ldots, x_n) = O_pr(n^{-2r}) \quad \text{for } 1 \leq j \leq J
\text{ and }
(4) & \quad E(\|\hat{f}_n - f_j^*\|_2^2 | x_1, \ldots, x_n) = O_pr(n^{-2r}) .
\end{align*}

The rates of convergence in (1), (3) and (4) do not depend on $J$. It is clear from the results in Stone (1982) for $J = 1$ that these rates are optimal.

5. Proof of Theorem 1. Throughout the remainder of the paper it will be assumed that Condition 1 is satisfied. Let $b$ and $B$ be positive numbers such that
\[ \delta = (1 - bB^{-1})^{1/2} < 1. \] Let \( \text{SD}(Z) \) denote the standard deviation of a random variable \( Z \).

**Lemma 1.** Let \( V_j = h_j(X_j) \) be random variables such that \( \sum_1^j V_j \) has finite second moment. Then each \( V_j \) has finite second moment. Also
\[
\text{SD}(V_1 + \ldots + V_j) \geq \left( \frac{1 - \delta}{2} \right)^{(j-1)/2} (\text{SD}(V_1) + \ldots + \text{SD}(V_j)) \quad \text{for } 1 \leq j \leq J.
\]

**Proof.** By Condition 1, the first conclusion reduces to that for independent \( V_j \)'s, which follows from Theorem 6.4.1 of Chung (1974). In proving the second conclusion it can be assumed that \( E V_j = 0 \) for \( 1 \leq j \leq J \). Set \( \sigma_j = \text{SD}(V_j) \) and \( \tau_j = \text{SD}(V_1 + \ldots + V_j) \). The desired result is trivially true for \( j = 1 \). Suppose it is true for \( j \), where \( 1 \leq j \leq J \). It will now be shown to hold for \( j + 1 \). If \( \tau_j = 0 \), then \( \sigma_1 = \ldots = \sigma_j = 0 \) and the desired conclusion reduces to the obvious inequality \( 0 \leq (1 - \delta)/2 \leq 1 \). If \( \sigma_{j+1} = 0 \), the desired conclusion follows from that for \( j \). Suppose instead that \( \tau_j > 0 \) and \( \sigma_{j+1} > 0 \), and let \( \rho \) denote the correlation between \( V_1 + \ldots + V_j \) and \( V_{j+1} \). Set \( W = (X_1, \ldots, X_j) \) and \( Z = X_{j+1} \). Let \( g_{W,Z} \) and \( g_Z \) denote respectively the joint density of \( W, Z \) and the marginal density of \( Z \). Then \( g_{W,Z} \geq b \) on \( [0,1]^{j+1} \) and \( g_Z \leq B \) on \( [0,1] \). Write \( V_1 + \ldots + V_j \) as \( \psi(W) \) and \( V_{j+1} \) as \( \psi(Z) \). Then
\[
(1 - \rho^2) \sigma_{j+1}^2 = \min_B E((\psi(Z) - B\psi(W))^2)
\]
\[
= \min_B \int \int (\psi(z) - B\psi(w))^2 g_{W,Z}(w,z) \, dw \, dz
\]
\[
> b B^{-1} \min_B \int I[(\psi(z) - B\psi(w))^2 g_Z(z)] \, dw
\]
\[
> b B^{-1} \sigma_{j+1}^2.
\]
so that $p^2 < 1 - bB^{-1}$ and hence $p > -\delta$. Consequently,

\[
\tau_{j+1}^2 = \tau_j^2 + 2p\tau_j\sigma_{j+1} + \sigma_{j+1}^2 \\
\geq \frac{1+p}{2} (\tau_j + \sigma_{j+1})^2 \\
\geq \frac{1-\delta}{2} ((1-\delta)(j-1)/2(\sigma_1 + \ldots + \sigma_j) + \sigma_{j+1})^2 \\
\geq (1-\delta)^2 (\sigma_1 + \ldots + \sigma_{j+1})^2.
\]

Therefore, the desired result holds for $j+1$. By induction it holds for $1 \leq j \leq J$ as desired.

Let $Pr_n$ correspond to the empirical distribution of $X_1, \ldots, X_n$; so that $Pr_n(X_j \in I_{nv})$ is $n^{-1}$ times the cardinality of $\{i: 1 \leq i \leq n$ and $X_{ij} \in I_{nv}\}$. Recall that $\gamma = 1/(2p+1) < .5$ since $p > .5$. Thus if $N_n \sim n^\gamma$, then

\[(6) \quad \lim_{n} n^{a-1} N_n^2 = 0 \quad \text{for some} \quad a > 0.\]

The next result follows easily from (5) and Bernstein's inequality (see Theorem 3 of Hoeffding, 1963) applied to the Binomial distribution.

**LEMMA 2.** Suppose (6) holds and let $\varepsilon > 0$. Then, except on an event whose probability tends to zero with $n$,

\[
|Pr_n(X_j \in I_{nv}) - Pr(X_j \in I_{nv})| \leq \varepsilon Pr(X_j \in I_{nv})
\]

for $1 \leq j \leq J$ and $1 \leq v \leq N_n$.

and
\[ |\Pr_n(X_j \in I_{n_1}, X_j \in I_{n_2}) - \Pr(X_j \in I_{n_1}, X_j \in I_{n_2})| \]
\[ \leq \epsilon \Pr(X_j \in I_{n_1}, X_j \in I_{n_2}) \]
for \( 1 \leq j_1, j_2 \leq J, \ j_1 \neq j_2 \), and \( 1 \leq \nu_1, \nu_2 \leq N_n \).

For \( 1 \leq j \leq J \) and \( s \) a function on \([0,1]\), let \( V_j(s) \) be the function on \( C \) defined by
\[ V_j(s)(x_1, \ldots, x_J) = s(x_j) \]

Let \( SD_n \) denote the standard deviation corresponding to the empirical distribution of \( X_1, \ldots, X_n \). The proof of the next result will be given in Section 6.

**Lemma 3.** Suppose (6) holds and let \( \delta_1 \in (\delta, 1) \). Then, except on an event whose probability tends to zero with \( n \), the following statement is valid for all choices of \( s_1, \ldots, s_J \in S_n \):
\[ SD_n(\sum_1^J V_j(s_j)) \geq \left( \frac{1-\delta_1}{2} \right)^{(J-1)/2} \sum_1^J SD_n(V_j(s_j)) \]

If the constrained minimization problem arising in the definition of \( \hat{f}_n \) has a unique solution, then
\[ \hat{f}_{nj}(x_j) = \sum_{i=1}^n W_{nji}(x_j) \gamma_i, \]
where the functions \( W_{nji} \) on \([0,1]\) are uniquely determined. Set
\[ |W_{nj}(x_j)|^2 = \sum_{i=1}^n W_{nji}^2(x_j). \]

**Lemma 4.** Suppose (6) holds. Then the constrained minimization problem has a unique solution, except on an event whose probability tends to zero.
with $n$. Moreover,

$$\sup_{0 \leq x_j \leq 1} |W_{nj}(x_j)|^2 = O_p(n^{-1} N_n) \quad \text{for } 1 \leq j \leq J.$$ 

**PROOF.** Think of $X_1, \ldots, X_n$ as fixed. Consider an alternative experiment having $n^J$ cases $(\tilde{X}_i, \tilde{Y}_i)$, $i = (i_1, \ldots, i_J)$, where $1 \leq i_1, \ldots, i_J \leq n$ and $\tilde{X}_i \in \mathbb{R}^J$ is defined by $\tilde{X}_{i_j} = X_{i_j}$ for $1 \leq j \leq J$; so that the $j$th coordinate of $\tilde{X}_i$ is the same as the $j$th coordinate of $X_{i_j}$. Note that, for $1 \leq j \leq J$ and a function $h$ defined on $[0,1]$,

$$\Sigma_i \ h(\tilde{X}_{i_j}) = \Sigma_i \ h(X_{i_j}) = n^{J-1} \Sigma_{i=1}^n h(X_{i_j}).$$

Set $\tilde{Y}_n = n^{-J} \Sigma_i \tilde{Y}_i$ and

$$\tilde{Y}_{i_j} = n^{1-J} \Sigma_{i': i_j = i} \tilde{Y}_{i'} \quad \text{for } 1 \leq j \leq J \text{ and } 1 \leq i \leq n.$$ 

Then

$$\Sigma_{i=1}^n (\tilde{Y}_{i_j} - \tilde{Y}_n) = 0 \quad \text{for } 1 \leq j \leq J.$$ 

Let $\hat{f}_n$, of the form

$$\hat{f}_n(x_1, \ldots, x_J) = \overline{Y}_n + \Sigma_{i=1}^J \hat{f}_{nj}(x_{i_j}),$$

be chosen subject to the constraints that $\hat{f}_{nj} \in S_n$ and $\Sigma_i \hat{f}_{nj}(\tilde{X}_{i_j}) = 0$ for $1 \leq j \leq J$ to minimize the residual sum of squares

$$S^2 = \Sigma_i (\tilde{Y}_i - \hat{f}_n(\tilde{X}_i))^2.$$ 

Observe that the constraint $\Sigma_i \hat{f}_{nj}(\tilde{X}_{i_j}) = 0$ is equivalent to the constraint $\Sigma_{i=1}^n \hat{f}_{nj}(X_{i_j}) = 0$; observe also that

$$S^2 = n^{J-1} \Sigma_{i=1}^J \Sigma_{i'=1}^n (\tilde{Y}_{i_j} - \overline{Y}_n - \hat{f}_{nj}(X_{i_j}))^2 + \Sigma_i (\tilde{Y}_i - \overline{Y}_n - \Sigma_{i'=1}^J (\tilde{Y}_{i_j} - \overline{Y}_n))^2,$$

since the omitted cross-product terms equal zero. Thus $\hat{f}_{nj}$ minimizes

$$S^2 = \Sigma_{i=1}^n (\tilde{Y}_{i_j} - \overline{Y}_n - \hat{f}_{nj}(X_{i_j}))^2.$$
subject to the constraints that $\hat{f}_{nj} \in S_n$ and $\sum_1^n \hat{f}_{nj}(x_{ij}) = 0$; equivalently, $\hat{f}_{nj}$ minimizes $S_j^2$ subject to the constraint that $\hat{f}_{nj} \in S_n$.

Suppose now that $\varepsilon' = -1$ in the definition of $S_n$. Let $1 \leq j \leq J$ and $1 \leq \nu \leq N_{\nu}$ and set

$$I_{nj\nu} = \{i : 1 \leq i \leq n \text{ and } x_{ij} \in I_{\nu}\}.$$ 

Then $\hat{f}_{nj}$ reduces to a polynomial $\hat{p}_{nj\nu}$ of degree $\varepsilon$ (or less) on $I_{nj\nu}$; $\hat{p}_{nj\nu}$ minimizes

$$\sum_{I_{nj\nu}} (\bar{y}_{ij} - \bar{y}_n - p(x_{ij}))^2$$

among all such polynomials. Suppose this minimization problem has a unique solution. Then

$$\hat{f}_{nj}(x_j) = \hat{p}_{nj\nu}(x_j) = \sum_{I_{nj\nu}} \bar{v}_{nj}(x_j)(\bar{y}_{ij} - \bar{y}_n), \quad x_j \in I_{\nu},$$

where the functions $\bar{v}_{nj}$ on $I_{\nu}$ are uniquely determined. Set

$$|\bar{v}_{nj}(x_j)|^2 = \sum_{I_{nj\nu}} \bar{v}_{nj}^2(x_j), \quad x_j \in I_{\nu}.$$

It follows from (5), (6), Hoeffding's inequality (Theorem 1 of Hoeffding, 1963) and the argument on pages 1354-1355 of Stone (1980) that, except on an event whose probability tends to zero with $n$, these minimization problems all have a unique solution; and

$$(7) \quad \sup_{0 \leq x_j \leq 1} |\bar{v}_{nj}(x_j)|^2 = O_p(n^{-1}N_n) \quad \text{for } 1 \leq j \leq J.$$ 

Write
\[ \hat{\mathbf{f}}_{nj}(x_j) = \sum_i \mathbf{w}_{nj i}(x_j) \mathbf{y}_i \] and set
\[ |\mathbf{w}_{nj i}(x_j)|^2 = \sum_i |\mathbf{w}_{nj i}(x_j)|^2 \quad \text{for} \quad 0 \leq x_j \leq 1. \]

It follows from (7) that
\[ \sup_{0 \leq x_j \leq 1} |\mathbf{w}_{nj i}(x_j)|^2 = 0_{pr}(n^{-J} N_n) \quad \text{for} \quad 1 \leq j \leq J. \]

If \( \lambda' \) is changed from \(-1\) to a nonnegative number, then additional constraints are imposed on the overall minimization problem. Thus the probability of nonuniqueness and the quantity \( |\mathbf{w}_{nj i}(x_j)|^2 \) under uniqueness are both reduced or unaltered. Consequently, even without the supposition that \( \lambda' = -1 \) it is true that, except on an event whose probability tends to zero with \( n \), the overall constrained minimization problem has a unique solution; and it is true that (8) holds.

Consider a function \( \phi \) on \( C \) of the form
\[ \phi(x_1, \ldots, x_J) = u + \sum_{j=1}^{J} s_j(x_j), \]
where \( u \) is any real number and
\[ s_j \in S_n \quad \text{and} \quad \Sigma_{i=1}^{n} s_j(x_{ij}) = 0 \quad \text{for} \quad 1 \leq j \leq J. \]

Observe that
\[ \sum_i \phi(\tilde{x}_i) = n^J u^2 + n^{J-1} \sum_{j=1}^{J} \sum_{i=1}^{n} s_j^2(\tilde{x}_ij) \]

and
\[ \sum_{i=1}^{n} \phi^2(x_i) = nu^2 + \sum_{i=1}^{n} (\sum_{j=1}^{J} s_j(x_{ij}))^2. \]
Thus it follows from Lemma 3 that there is a $\gamma \in (0,1)$ such that, except on an event whose probability tends to zero with $n$, 

$$\sum_{i=1}^{n} \phi^2(x_i) \geq \gamma n^{1-j} \sum \phi^2(\tilde{x}_i)$$

for every real number $u$ and, for $1 \leq j \leq J$, for all choices of $s_j \in S_n$ such that $\sum s_j(x_{ij}) = 0$. The conclusion of the lemma now follows from (8) and Lemma 2.1 of Ehrenfeld (1956), a result involving the comparison of two experiments.

Let $\|h\|_\infty = \sup_{0 \leq t \leq 1} |h(t)|$ denote the supnorm of a function $h$ on $[0,1]$.

**Lemma 5.** For each $h \in H$ and $n \geq 1$ there is an $s \in S_n$ with $\|s-h\|_\infty \leq M_1n^{-p}$; here $M_1$ is some fixed positive constant.

This result is due to de Boor (1968); see also de Boor (1978) or Powell (1981).

**Lemma 6.** Consider a Hilbert space with norm $\| \|$. Let $P_j$ denote the orthogonal projection onto a subspace $V_j$ for $0 \leq j \leq J$. Suppose $v = \sum_{j=0}^{J} v_j$, where $v_j \in V_j$ for $0 \leq j \leq J$. Then

$$\|v\|^2 \leq \left( \max_{0 \leq j \leq J} \|v_j\| \right) \left( \sum_{j=0}^{J} \|P_jv\|^2 \right).$$

**Proof.** Let "\cdot" denote the Hilbert space inner product. Observe that $v_j \cdot (v - P_jv) = 0$ for $0 \leq j \leq J$. Thus

$$\|v\|^2 = \sum_{j} v_j \cdot v = \sum_{j} v_j \cdot P_jv \leq \sum_{j} \|v_j\| \|P_jv\|,$$

which yields the desired result.
Set \( \bar{\mu}_n = n^{-1} \sum_{i=1}^{n} f(X_i) \). Since \( f(X) \) has mean \( \mu \) and finite second moment,

\[
(\bar{\mu}_n - \mu)^2 = O_p(n^{-1}) = O_p(n^{-2r}) \,.
\]

Suppose (6) holds. Let \( f_n \), of the form

\[
\bar{f}_n(x_1, \ldots, x_J) = \bar{\mu}_n + \sum_{j=1}^{J} \bar{f}_n(x_j),
\]

be a solution to the constrained minimization problem with \( Y_i \) replaced by \( f(X_i) \); by the first conclusion to Lemma 4, this problem has a unique solution except on an event whose probability tends to zero with \( n \).

Let \( \| \phi \|_n \) denote the \( L^2 \) norm of a function \( \phi \) on \( C \) with respect to the empirical distribution of \( X_1, \ldots, X_n \); so that

\[
\| \phi \|_n^2 = n^{-1} \sum_{i=1}^{n} \phi^2(X_i) .
\]

Similarly, for \( 1 \leq j \leq J \), let \( \| h \|_{nj} \) denote the \( L^2 \) norm of a function \( h \) on \([0,1]\) with respect to the empirical distribution of \( X_{1j}, \ldots, X_{nj} \).

**LEMMA 7.** Suppose that Conditions 1 and 3 hold and that \( N_n \sim n^\gamma \).

Then

\[
\| \bar{f}_{nj} - f_n^* \|_{nj}^2 = 0_p(n^{-2r}) \quad \text{for} \quad 1 \leq j \leq J .
\]

**PROOF.** Since \( f = f^* + f - f^* \), \( (f^*)^* = f^* \) and \( (f-f^*)^* = 0 \), it suffices to verify the lemma when \( f = f^* \) and when \( f^* = 0 \).

Suppose first that \( f = f^* \) and set \( f_j = f^*_j \). Then

\[
f(x_1, \ldots, x_J) = \mu + \sum_{j=1}^{J} f_j(x_j) .
\]

For \( 1 \leq j \leq J \) choose \( s_{nj} \in S_n \) such that \( \| f_j - s_{nj} \|_\infty \leq M_1 n^{-r} \), which is possible by Lemma 5. Define \( f_n \) by

\[
f_n(x_1, \ldots, x_J) = \mu + \sum_{j=1}^{J} s_{nj}(x_j) .
\]

Then \( \| f_n - f \|_\infty = O(n^{-r}) \) and hence \( \| f_n - f \|_n^2 = O(n^{-2r}) \). Now \( \| \bar{f}_n - f_n \| \leq \| f_n - f \|_n \), so that
\[ \| \tilde{f}_n - f \|_n^2 = 0_{pr}(n^{-2r}) \] and hence

\[ (10) \quad \| \tilde{f}_n - f \|_n^2 = 0_{pr}(n^{-2r}). \]

For \(1 \leq j \leq J\) set \( \overline{\mu}_{nj} = n^{-1} \sum_1^n f_j(x_{ij}) \) and observe that \( f_j \) is bounded by Condition 3 and hence that \( \overline{\mu}_{nj}^2 = 0_{pr}(n^{-1}) = 0_{pr}(n^{-2r}) \); also set \( \overline{v}_{nj} = n^{-1} \sum_1^n s_{nj}(X_{ij}) \) and observe that \( \overline{v}_{nj}^2 = 0_{pr}(n^{-2r}) \).

It now follows from (10) and Lemma 3 that, for \(1 \leq j \leq J\),

\[ \| \tilde{f}_n - s_{nj} - \overline{v}_{nj} \|_{nj}^2 = 0_{pr}(n^{-2r}) \] and hence \( \| \tilde{f}_n - s_{nj} \|_{nj}^2 = 0_{pr}(n^{-2r}) \); consequently, \( \| \tilde{f}_n - f \|_{nj}^2 = \| \tilde{f}_n - f_{nj} \|_{nj}^2 = 0_{pr}(n^{-2r}) \) as desired.

Suppose next that \( f^* = 0 \) or, equivalently, that \( E(f(X)|X_j) = 0 \) for \(1 \leq j \leq J\) and hence that \( \mu = 0 \). Then \( \overline{\mu}_n = 0_{pr}(n^{-1}) = 0_{pr}(n^{-2r}) \) by (9). For \(1 \leq j \leq J\) let \( \tilde{f}_{nj} \), of the form \( \tilde{f}_{nj}(x_1, \ldots, x_J) = s_{nj}(x_j) \), be chosen subject to the constraint that \( s_{nj} \in S_n \) and \( n^{-1} \sum_1^n s_{nj}(X_{ij}) = 0 \) to minimize \( \| \tilde{f}_{nj} - f \|_n^2 \). It follows from the assumption that \( f^* = 0 \), the boundedness of \( f \) and Lemma 4 (with \( J = 1 \)) that

\[ (11) \quad \| \tilde{f}_{nj} \|_n^2 = 0_{pr}(n^{-2r}) \quad \text{for} \ 1 \leq j \leq J. \]

Temporarily, think of \( X_1, \ldots, X_n \) as fixed. Consider the Hilbert space of functions defined on the range of \( X_1, \ldots, X_n \) with the empirical norm \( \| \cdot \|_n \) defined by \( \| f \|_n^2 = n^{-1} \sum_1^n f^2(X_i) \). Let \( V_0 \) denote the space of constant functions and, for \(1 \leq j \leq J\), let \( V_j \) denote the space of functions of the form \( \phi(x_1, \ldots, x_J) = s(x_j) \), where \( s \in S_n \) and \( \sum_1^n s(X_{ij}) = 0 \). Let \( P_j \) denote the orthogonal projection onto \( V_j \) for \(0 \leq j \leq J\). Let \( v_0 \) denote the constant function \( \overline{\mu}_n \) and, for \(1 \leq j \leq J\), let \( v_j \) denote the function defined by \( v_j(x_1, \ldots, x_J) = \tilde{f}_{nj}(x_j) \). Set \( v = \tilde{f} = \sum_0^J v_j \). It follows from (9)
with \( \mu = 0 \) and (11) that

\[
(12) \quad \|P_j v_n\|^2 = o_p(n^{-2r}) \quad \text{for} \quad 0 \leq j \leq J.
\]

According to Lemma 3 there is a constant \( \gamma \in (0,1) \) such that, except on an event whose probability tends to zero with \( n \),

\[
\|\sum_1^J v_j\|^2 \geq \gamma \sum_1^J v_j n\|^2 \quad \text{and hence}
\]

\[
\|v\|^2 = \|\sum_1^J v_j\|^2 = \|v_0\|^2 + \|\sum_1^J v_j\|^2 \geq \gamma \max_{0 \leq j \leq J} v_j n\|^2.
\]

Thus, by Lemma 6, \( \|v\|^2 \leq \gamma^{-1}(\sum_0^J P_j v\|^2) \); so, by (12), \( \|v\|^2 = o_p(n^{-2r}) \).

It now follows from Lemma 3 applied to the norm \( \|v\|_n \) that

\[\|v_j\|^2 = o_p(n^{-2r}) \quad \text{for} \quad 1 \leq j \leq J, \quad \text{which is equivalent to the conclusion of the lemma when} \quad f^* = 0.\]

The proofs of the next two results will be given in Section 6.

**LEMMA 8.** There is an \( M_2 \in (0,\infty) \) such that

\[
\|s(m) - h(m)\|^2_j \leq M_2(N_n^{2(m-p)} + N_n^{2m}\|s-h\|^2_j)
\]

for \( 1 \leq j \leq J \), \( h \in H \), \( n \geq 1 \) and \( s \in S_n \).

**LEMMA 9.** Suppose that

\[
(13) \quad \lim_{n} n^{a-1} N_n = 0 \quad \text{for some} \quad a > 0.
\]

Then there is an \( M_3 \in (0,\infty) \) such that, except on an event whose probability tends to zero with \( n \),

\[
\|s-h\|^2_j \leq M_3(N_n^{2p + 2m}\|s-h\|^2_j)
\]

for \( 1 \leq j \leq J \), \( h \in H \) and \( s \in S_n \).
The next result follows from Lemmas 7-9.

**LEMMA 10.** Suppose that Conditions 1 and 3 hold and that $N_n \sim n^\gamma$. Then

$$
\|\hat{f}_{nj} - (f^*_j)(m)\|_j^2 = 0_{pr}(n^{-2r_m}) \text{ for } 1 \leq j \leq J .
$$

**PROOF OF THEOREM 1.** Suppose that Conditions 1-3 hold and that $N_n \sim n^\gamma$. Now

$$(14) \quad \hat{f}_{nj} - (f^*_j)(m) = \hat{f}_{nj} - \hat{f}_{nj} + \hat{f}_{nj} - (f^*_j)(m) .$$

It follows from Lemma 4 that

$$E(\|\hat{f}_{nj} - \hat{f}_{nj} \|_j^2|X_1, \ldots, X_n) = 0_{pr}(n^{-2r}) .$$

Thus, by Lemma 8,

$$(15) \quad E(\|\hat{f}_{nj} - \hat{f}_{nj} \|_j^2|X_1, \ldots, X_n) = 0_{pr}(n^{-2r_m}) .$$

The desired conclusion follows from (14), (15) and Lemma 10.

6. Proofs of Lemmas 3, 8 and 9. The next result follows easily from a compactness argument and change of variables.

**LEMMA 11.** If $Q$ is a polynomial of degree $k$ (or less), $I$ is an interval of finite positive length $\tau$, and $t_0 \in I$, then

$$
(\sum_0^m \frac{m}{m!} |Q^{(m)}(t_0)|)^2 \leq c_k^{2\tau-1} \int_I Q^2(t) dt .
$$

Here $c_k$ is some positive constant depending only on $k$. 

PROOF OF LEMMA 8. Let $t_0 \in I = I_{n_0} = [0,1]$, where $I$ has length $\tau = N^{-1}$. Let $Q$ be the Taylor polynomial of degree $k$ about $t_0$ corresponding to $h$. Since $h \in H$, $|h - Q| \leq M_k^P/k!$ on $I$ and hence $\int_I (h - Q)^2 \leq M^2 \tau^{2p+1}/(k!)^2$. Thus by Lemma 11

\[(s(m)(t_0) - h(m)(t_0))^2 = (s(m)(t_0) - Q(m)(t_0))^2 \leq M_2^2 \tau^{-(2m+1)}(\tau^{2p+1} + \int_I (s-h)^2)\]

for some $M_2 \in (0,\infty)$. Consequently

\[\int_I (s(m) - h(m))^2 \leq M_2 \tau^{-2m}(\tau^{2p+1} + \int_I (s-h)^2)\]

and hence

\[\int_0^1 (s(m) - h(m))^2 \leq M_2 (\tau^{2(p-m)} + \tau^{-2m}) \int_0^1 (s-h)^2\]

Since each $g_j$ is bounded away from 0 and $\infty$ by Condition 1, the desired result holds.

Let $I$ denote an interval of finite positive length $\tau$ and let $(T,U)$ denote a pair of $I$-valued random variables each having an absolutely continuous distribution. Let $E'$ be the expectation operator corresponding to the empirical distribution based on a random sample of size $n_0$ from the distribution of $(T,U)$. Similarly, let Cov' denote the covariance operator corresponding to this empirical distribution.

LEMMA 12. Suppose that the marginal densities of $T$ and $U$ are each bounded below by $\beta/\tau$ on $I$, where $\beta > 0$. Let $t > 0$. Then, except on an event having probability at most $d_k \exp(-2n_0 t^2)$, the following inequalities hold simultaneously for all polynomials $Q, R$.
of degree $k$:

$$|E' Q(T) - E Q(T)| \leq t \beta^{-\frac{1}{4}k} SD(Q(T));$$

$$|E' Q^2(T) - E Q^2(T)| \leq t \beta^{-\frac{1}{2}k} Var(Q(T));$$

$$|E' R(U) - E R(U)| \leq t \beta^{-\frac{1}{4}k} SD(R(U));$$

and

$$|\text{Cov}'(Q(T), R(U)) - \text{Cov}(Q(T), R(U))| \leq t(t+3) \beta^{-\frac{1}{2}k} SD(Q(T))SD(R(U)).$$

Here $d_k$ is some positive constant depending only on $k$.

PROOF. Let $t_0$ denote the left endpoint of $I$. It follows from Hoeffding's inequality that, for $m \geq 1$,

$$\Pr\left(\left|E' \left[\left(\frac{T-t_0}{\tau}\right)^m\right] - E \left[\left(\frac{T-t_0}{\tau}\right)^m\right]\right| > t\right) \leq \exp(-2n_0 t^2).$$

Thus by Lemma 11, except on an event having probability at most $k \exp(-2n_0 t^2)$.

$$|E' Q(T) - E Q(T)| = \left| \sum_{m=0}^{k} \frac{Q^{(m)}(t_0)}{m!} \tau^m \left(\left(\frac{T-t_0}{\tau}\right)^m\right) - E \left[\left(\frac{T-t_0}{\tau}\right)^m\right]\right|$$

$$\leq t \sum_{m=0}^{k} \frac{\tau^m}{m!} |Q^{(m)}(t_0)|$$

$$\leq tc_k \int_{I} Q^2(t)dt \frac{1}{\tau^k}$$

$$\leq t\beta^{-\frac{1}{2}} (E Q^2(T))^\frac{1}{2}$$

$$= t\beta^{-\frac{1}{2}} SD(Q(T))$$

for all polynomials $Q$ of degree $k$ such that $E Q(T) = 0$; and hence
the first inequality of the lemma holds for all polynomials $Q$ of degree $k$. The proofs of the remaining inequalities are similar.

PROOF of LEMMA 9. Let $T_j$ denote the $j^{th}$ coordinate of $X_i$. Observe that $I_{j} = I_{n, j}$ has length $\tau = N^{-1}_n$. Set

$I_{j} = \{i: 1 \leq i \leq n \text{ and } T_i \in I_{j}\}$ and let $|I_{j}|$ denote the number of elements in $I_{j}$. It follows from (5), (13) and Lemma 2 that, except on an event whose probability tends to zero with $n$, for $n$ sufficiently large

$$|I_{j}| \geq b\tau /2 \geq n^a \quad \text{for all } j.$$ 

here $a$ is some positive constant. Choose $t_j \in I_{j}$ and let $Q_j$ be the Taylor polynomial of degree $k$ about $t_j$ corresponding to $h$. Then $|h - Q_j| \leq \frac{M_{\tau^j}}{k!}$ on $I_{j}$. It follows from Lemma 12 (the second inequality) that, except on an event whose probability tends to zero with $n$, for $n$ sufficiently large and for all $j$

$$\frac{1}{2\delta} \int_{I_{j}} (s - Q_j)^2 g_j \leq \frac{1}{|I_{j}|} \sum_{I_{j}} (s(T_i) - Q_j(T_i))^2 \leq 2 \left( \left( \frac{M_{\tau^j}}{k!} \right)^2 + \frac{1}{|I_{j}|} \sum_{I_{j}} (s(T_i) - h(T_i))^2 \right)$$

and hence

$$\frac{1}{2\delta} \int_{I_{j}} (s - h)^2 g_j \leq 6 \left( \left( \frac{M_{\tau^j}}{k!} \right)^2 + \frac{1}{|I_{j}|} \sum_{I_{j}} (s(T_i) - h(T_i))^2 \right).$$

Consequently by (5),

$$\int_{I_{j}} (s-h)^2 g_j \leq 12B\tau \left( \frac{M_{\tau^j}}{k!} \right)^2 + \frac{1}{|I_{j}|} \sum_{I_{j}} (s(T_i) - h(T_i))^2.$$
The conclusion to Lemma 9 now follows easily by summing on $v$.

Let $(\Omega, P)$ be a probability space. Let $\Lambda$ be a positive integer and let $A_1, \ldots, A_\Lambda$ be a finite partition of $\Omega$ which is nondegenerate in the sense that $p_\lambda = P(A_\lambda) > 0$ for $1 \leq \lambda \leq \Lambda$. Let $V_1, V_2$ be a pair of random variables having finite second moment. Set $\mu_j = EV_j$, $\sigma_{j\varepsilon} = \text{Cov}(V_j, V_\varepsilon)$ and $\sigma_j^2 = \text{Var}(V_j)$. Also set $\mu_{j|\lambda} = E(V_j|A_\lambda)$, $\sigma_{j\varepsilon|\lambda} = \text{Cov}(V_j, V_\varepsilon|A_\lambda) = E((V_j - \mu_{j|\lambda})(V_\varepsilon - \mu_\varepsilon|\lambda)|A_\lambda)$ and $\sigma_{j|\lambda} = \sqrt{\sigma_{j\varepsilon|\lambda}}$. Let $(\Omega', P')$ be a second probability space having a nondegenerate partition $A'_1, \ldots, A'_\Lambda$ and a pair $V'_1, V'_2$ of random variables each having finite second moment. Define $p'_{\lambda}$, $\mu'_j$, etc. as before.

**Lemma 13.** Given $\varepsilon > 0$, there is a $\delta > 0$ such that the following statement is valid: If $|p'_{\lambda} - p_{\lambda}| \leq \delta p_{\lambda}$, $|\mu'_j - \mu_{j|\lambda}| \leq \delta \sigma_{j|\lambda}$ and $|\sigma'_{j\varepsilon|\lambda} - \sigma_{j\varepsilon|\lambda}| \leq \delta \sigma_{j|\lambda} \sigma_{\varepsilon|\lambda}$ for $1 \leq j, \varepsilon \leq 2$ and $1 \leq \lambda \leq \Lambda$, then

$$|\sigma'_{j\varepsilon} - \sigma_{j\varepsilon}| \leq \varepsilon \sigma_j \sigma_{\varepsilon}$$

for $1 \leq j, \varepsilon \leq 2$.

**Proof.** Observe that

$$\mu_j = \sum p_\lambda \mu_{j|\lambda},$$

$$\sigma_j^2 = \sum p_\lambda (\sigma_{j|\lambda}^2 + (\mu_{j|\lambda} - \mu_j)^2)$$

and

$$\sigma_{j\varepsilon} = \sum p_\lambda (\sigma_{j\varepsilon|\lambda}^2 + (\mu_{j|\lambda} - \mu_j)(\mu_{\varepsilon|\lambda} - \mu_\varepsilon))$$

and that similar formulas hold for the second probability space. The desired conclusion now follows in a straightforward manner (Schwarz's inequality is used several times).
PROOF OF LEMMA 3. Choose \( \varepsilon \) with \( 0 < \varepsilon < ((1-\delta)/2)^{-1} \). It follows easily from Lemmas 2, 12 and 13 that, except on an event whose probability tends to zero with \( n \), the following inequality holds for all choices of \( s_1, \ldots, s_J \in S_n \) and the corresponding choices of \( V_j = V_j(s_j) \):

\[
|\text{Cov}_n(V_j, V_\ell) - \text{Cov}(V_j, V_\ell)| < \varepsilon \text{SD}(V_j) \text{SD}(V_\ell) \quad \text{for } 1 \leq j, \ell \leq J;
\]

in particular

\[
\text{SD}_n(V_j) \leq (1+\varepsilon)^{1/2} \text{SD}(V_j) \quad \text{for } 1 \leq j \leq J
\]

and hence by Lemma 1

\[
\text{Var}_n(V_1 + \ldots + V_J) \geq \text{Var}(V_1 + \ldots + V_J) - \varepsilon (\sum_{1}^{J} \text{SD}(V_j))^2
\]

\[
\geq \left( \frac{1-\delta}{2} - \varepsilon \right) (\sum_{1}^{J} \text{SD}(V_j))^2
\]

\[
\geq \left( \frac{1-\delta}{1+\varepsilon} - \varepsilon \right) (\sum_{1}^{J} \text{SD}_n(V_j))^2.
\]

Since \( \varepsilon \) can be made arbitrarily small, the desired result holds.
REFERENCES


