A SIMPLE BOOTSTRAP METHOD FOR CONSTRUCTING NONPARAMETRIC CONFIDENCE BANDS FOR FUNCTIONS

Peter Hall[1] and Joel Horowitz[2]

ABSTRACT. Standard approaches to constructing nonparametric confidence bands for functions are frustrated by the impact of bias, which generally is not estimated consistently when using the bootstrap and conventionally smoothed function estimators. To overcome this problem it is common practice to either undersmooth, so as to reduce the impact of bias, or oversmooth, and thereby introduce an explicit or implicit bias estimator. However, these approaches, and others based on nonstandard smoothing methods, complicate the process of inference, for example by requiring the choice of new, unconventional smoothing parameters and, in the case of undersmoothing, producing relatively wide bands. In this paper we suggest a new approach, which exploits to our advantage one of the difficulties that, in the past, has prevented an attractive solution to this problem—the fact that the standard bootstrap bias estimator suffers from relatively high-frequency stochastic error. The high frequency, together with a technique based on quantiles, can be exploited to dampen down the stochastic error term, leading to relatively narrow, simple-to-construct confidence bands.

KEYWORDS. Bandwidth, bias, confidence interval, conservative coverage, coverage error, kernel methods, statistical smoothing.

SHORT TITLE. Confidence bands.

1. INTRODUCTION

There is a particularly extensive literature, summarised at the end of this section, on constructing nonparametric confidence bands for functions. However, this work generally does not suggest practical solutions to the critical problem of choosing tuning parameters, for example smoothing parameters or the nominal coverage level for the confidence band, to ensure a high degree of coverage accuracy or to produce bands that err on the side of conservatism. In this paper we suggest new, simple

[1]Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia

[2]Department of Economics, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208, USA

bootstrap methods for constructing confidence bands using conventional smoothing parameter choices.

In particular, our approach does not need any undersmoothing or oversmoothing. The basic algorithm requires only a single application of the bootstrap, although a more refined, double bootstrap technique is also suggested. The greater part of our attention is directed to regression problems, but we also discuss the application of our methods to constructing confidence bands for density functions.

The key to our methodology is to exploit, to our advantage, a difficulty that in the past has hindered a simple solution to the confidence band problem. Specifically, if nonparametric function estimators are constructed in a conventional manner then their bias is of the same order as their error about the mean, and accommodating the bias has been a major obstacle to achieving good coverage accuracy. Conventional bootstrap methods can be used to estimate the bias and reduce its impact, but the bias estimators fail to be consistent, and in fact the stochastic noise from which they suffer is highly erratic; in the case of kernel methods it varies on the same scale as the bandwidth. However, as we show in this paper, the erratic behaviour is actually advantageous, since if we average over it then we can largely eliminate the negative impact that it has on the level accuracy of confidence bands. We do the averaging implicitly, not by computing means but by working with quantiles of the "distribution" of coverage.

To conclude we note some of the earlier literature on nonparametric confidence bands for functions. We summarise this literature largely in terms of whether it involves undersmoothing or oversmoothing; the technique suggested in the present paper is almost unique in that it requires neither of these approaches. Härdle and Bowman (1988), Härdle and Marron (1991), Hall (1992a), Eubank and Speckman (1993), Sun and Loader (1994), Härdle et al. (1995) and Xia (1998) suggested methods based on oversmoothing, using either implicit or explicit bias correction. Hall and Titterington (1988) also used explicit bias correction, in the sense that their bands required a known bound on an appropriate derivative of the target function. Bjerve et al. (1985), Hall (1992b), Hall and Owen (1993), Neumann (1995), Chen (1996), Neumann and Polzehl (1998), Picard and Tribouley (2000), Claeskens and Van Keilegom (2003) and McMurry and Politis (2008) employed methods that involve undersmoothing. There is also a theoretical literature which addresses the bias issue through consideration of the technical function class from which a regression mean or density came; see e.g. Low (1997) and Genovese and

Wasserman (2008). This work sometimes involves confidence balls, rather than bands, and in that respect is connected to research such as that of Li (1989), Eubank and Wang (1994) and Genovese and Wasserman (2005). Wang and Wahba (1995) considered spline and Bayesian methods.

2. METHODOLOGY

2.1. Model. Suppose we observe data pairs in a sample $\mathcal{Z} = \{(X_i, Y_i),\ 1 \le i \le n\}$, generated by the model

$$Y_i = g(X_i) + \epsilon_i\,, \tag{2.1}$$

where the experimental errors ϵ_i have zero mean. Our aim is to construct a pointwise confidence band for the true g in a closed region $\mathcal{R}$. A more elaborate, heteroscedastic model will be discussed in section 2.4; we omit it here only for the sake of simplicity.

2.2. Overview and intuition. Let $\hat{g}$ denote a conventional estimator of g. We assume that $\hat{g}$ incorporates smoothing parameters computed empirically from the data, using for example cross-validation or a plug-in rule, and that the variance of $\hat{g}$ can be estimated consistently by $s(\mathcal{X})^2\,\hat{\sigma}^2$, where $s(\mathcal{X})$ is a known function of the set of design points $\mathcal{X} = \{X_1, \ldots, X_n\}$ and the smoothing parameters, and $\hat{\sigma}^2$ is an estimator of the variance, σ^2, of the experimental errors ϵ_i, computed from the dataset $\mathcal{Z}$. The case of heteroscedasticity is readily accommodated too; see section 2.4. We write $\hat{g}^*$ for the version of $\hat{g}$ computed from a conventional bootstrap resample. For details of the construction of $\hat{g}^*$, see step 4 of the algorithm in section 2.3.

The smoothing parameters used for $\hat{g}$ would generally be chosen to optimise a measure of accuracy, for example in a weighted L_p metric where $1 \le p < \infty$, and we shall make this assumption implicitly in the discussion below. In particular it implies that the asymptotic effect of bias, for example as represented by the term $b(x)$ in (2.4) below, is finite and typically nonzero.

An asymptotic, symmetric confidence band for g, constructed naively without considering bias, and with nominal coverage $1 - \alpha$, has the form:

$$\mathcal{B}(\alpha) = \Big\{(x, y) : x \in \mathcal{R}\,,\ \hat{g}(x) - s(\mathcal{X})(x)\,\hat{\sigma}\,z_{1-(\alpha/2)} \le y \le \hat{g}(x) + s(\mathcal{X})(x)\,\hat{\sigma}\,z_{1-(\alpha/2)}\Big\}\,, \tag{2.2}$$

where $z_\beta = \Phi^{-1}(\beta)$ is the β-level critical point of the standard normal distribution, and Φ is the standard normal distribution function. Unfortunately, the coverage of

$\mathcal{B}(\alpha)$ at a point x, given by

$$\pi(x,\alpha) = P\{(x, g(x)) \in \mathcal{B}(\alpha)\}\,, \tag{2.3}$$

is incorrect even in an asymptotic sense, and in fact the band usually undercovers, often seriously, in the limit as $n \to \infty$. The reason is that the bias of $\hat{g}$, as an estimator of g, is of the same size as the estimator's stochastic error, and the confidence band allows only for the latter type of error. As a result the limit, as $n \to \infty$, of the coverage of the band is given by

$$\pi_{\lim}(x,\alpha) = \lim_{n\to\infty} \pi(x,\alpha) = \Phi\{z + b(x)\} - \Phi\{-z + b(x)\}\,, \tag{2.4}$$

where $z = z_{1-(\alpha/2)}$ and $b(x)$ describes the asymptotic effect that bias has on coverage. The right-hand side of (2.4) equals $\Phi(z) - \Phi(-z) = 1 - \alpha$ if and only if $b(x) = 0$. For all other values of $b(x)$, $\pi_{\lim}(x,\alpha) < 1 - \alpha$. This explains why the band at (2.2) almost always undercovers, unless some sort of bias correction is used.

The band potentially can be recalibrated, using the bootstrap, to correct for coverage errors caused by bias, but now another issue causes difficulty: the standard bootstrap estimator of bias, $E\{\hat{g}^*(x) \,|\, \mathcal{Z}\} - \hat{g}(x)$, is inconsistent, in the sense that the ratio of the estimated bias to its true value does not converge to 1 as $n \to \infty$. This problem can be addressed using an appropriately oversmoothed version of $\hat{g}$ when estimating bias, either explicitly or implicitly, but the degree of oversmoothing has to be determined from the data, and in practice this issue is awkward to resolve. Alternatively, the estimator $\hat{g}$ can be undersmoothed, so that the influence of bias is reduced, but now the amount of undersmoothing has to be determined, and that too is difficult. Moreover, confidence bands computed from an undersmoothed $\hat{g}$ are an order of magnitude wider than those at (2.2), and so the undersmoothing approach, although more popular than oversmoothing, is unattractive for at least two reasons.

However, a simpler bootstrap technique, described in detail in section 2.3, overcomes these problems. In summary it is implemented as follows. First, use standard bootstrap methods to construct an estimator $\hat{\pi}(x,\alpha)$ of the probability $\pi(x,\alpha)$, defined at (2.3), that the band $\mathcal{B}(\alpha)$ covers $(x, g(x))$. (For details, see step 5 in section 2.3.) Then, for a given desired level of coverage, $1 - \alpha_0$ say, define $\hat{\alpha}(x,\alpha_0)$ to be the solution of the equation $\hat{\pi}(x,\alpha) = 1 - \alpha_0$. Formally:

$$\hat{\pi}(x,\beta) = 1 - \alpha_0 \quad \text{when} \quad \beta = \hat{\alpha}(x,\alpha_0)\,. \tag{2.5}$$

In view of the undercoverage property discussed below (2.4), we expect $\hat\alpha(x,\alpha_0)$ to be less than α_0. Equivalently, we anticipate that the nominal coverage of the band has to be increased above $1-\alpha_0$ in order for the band to cover $(x,g(x))$ with probability at least $1-\alpha_0$. Conventionally we would employ $\hat\alpha(x,\alpha_0)$ as the nominal level, but, owing to the large amount of stochastic error in the bootstrap bias estimator that is used implicitly in this technique, it produces confidence bands with particularly poor coverage accuracy.

Instead we take the following approach. Given $\xi \in (0,\frac12)$, let $\hat\alpha_\xi(\alpha_0)$ be the $(1-\xi)$-level quantile of values of $\hat\alpha(x,\alpha_0)$, where $\hat\alpha(x,\alpha_0)$ is viewed as a function of $x \in \mathcal{R}$. (Details are given in step 6 in section 2.3.) The band $\mathcal{B}\{x,\hat\alpha_\xi(\alpha_0)\}$ is asymptotically conservative for all but at most a fraction ξ of pairs $(x,g(x))$, which can be identified from a plot of $\hat\alpha(x,\alpha_0)$. (They correspond to values x for which $|b(x)|$ is relatively large, or equivalently, to x for which $\hat\alpha(x,\alpha_0)$ tends to be relatively large.) In particular, taking $1-\xi \approx 0.7$ or 0.8 typically gives bands that err on the side of conservatism, but are not unduly conservative. See section 3.

Why does this work? It is clear that the coverage of the band $\mathcal{B}\{x,\hat\alpha_\xi(\alpha_0)\}$ increases with decreasing ξ. However, the more detailed claim made above, that the band is asymptotically conservative for all but at most a proportion ξ of pairs $(x,g(x))$, needs justification. This will be given in detail in section 4. In the remainder of section 2.2 we provide the reader with intuition, noting first that, although the bootstrap bias estimator is inconsistent, the estimator equals the true bias, plus negligible terms, together with a stochastic quantity which has zero mean and a symmetric distribution. This symmetry plays a major role.

For example, if the model at (2.1) obtains and $\hat g$ is a local linear estimator with bandwidth h and a compactly supported kernel, and if the design is univariate, then the bootstrap bias estimator is given by

$$E\{\hat g^*(x) \,|\, \mathcal{Z}\} - \hat g(x) = c_1\, g''(x)\, h^2 + (nh)^{-1/2}\, f_X(x)^{-1/2}\, W(x/h) + \text{negligible terms}\,, \tag{2.6}$$

where, here and below, c_j denotes a positive constant, c_1 and c_2 depend only on the kernel, f_X is the common density of the design points X_i, and W is a c_2-dependent, stationary Gaussian process with zero mean and unit variance. (Although the covariance structure, and hence the distribution, of W are fixed, a different version of W is used for each sample size, to ensure that (2.6) holds.) The value of c_2, in the claim of "c_2-dependence," depends on the length of the support of the kernel. The

first term, $c_1\, g''(x)\, h^2$, on the right-hand side of (2.6) is identical to the asymptotic bias of $\hat{g}(x)$, but the second term on the right makes the bias estimator, on the left-hand side of (2.6), inconsistent.

Still in the context of (2.6), the bandwidth h, which typically would have been chosen by a standard empirical method to minimise a version of L_p error where $1 \le p < \infty$, is asymptotic to $c_3\, n^{-1/5}$. The asymptotic variance of $\hat{g}(x)$ equals $c_4\, (nh)^{-1}\, f_X(x)^{-1}$, where f_X is the density of the design variables X_i. In this notation, $b(x)$ in (2.4) is given by

$$b(x) = -\lim_{h \to 0} \left[c_1\, g''(x)\, h^2 \big/ \left\{ c_4\, (nh)^{-1}\, f_X(x)^{-1} \right\}^{1/2} \right] = -c_5\, g''(x)\, f_X(x)^{1/2}\,. \quad (2.7)$$

However, the limiting form of the bootstrap estimator of bias is strongly influenced by the second term on the right-hand side of (2.6), as well as by the first term, with the result that

$$\hat{\pi}(x, \alpha) = \Phi\{z + b(x) + \Delta(x)\} - \Phi\{-z + b(x) + \Delta(x)\} + \text{negligible terms}\,, \quad (2.8)$$

where

$$\Delta(x) = -c_6\, W(x/h)\,. \quad (2.9)$$

Write $\beta = \alpha(x, \alpha_0) > 0$ for the solution of the equation

$$\Phi\{z_{1-(\beta/2)} + b(x)\} - \Phi\{-z_{1-(\beta/2)} + b(x)\} = 1 - \alpha_0\,. \quad (2.10)$$

Replacing $b(x)$ in (2.4) by $b(x) + \Delta(x)$ makes $\hat{\pi}(x, \alpha)$ greater, asymptotically, than $\pi_{\lim}(x, \alpha)$ in (2.4) if $|b(x) + \Delta(x)| > |b(x)|$; and makes $\hat{\pi}(x, \alpha)$ less than $\pi_{\lim}(x, \alpha)$ if $|b(x) + \Delta(x)| \le |b(x)|$. Since the stochastic process W is symmetric then, again asymptotically, these inequalities arise with equal probabilities, and moreover, $|b(x) + \Delta(x)| = |b(x)|$ if and only if $\Delta(x) = 0$. Observe too that, because Δ oscillates at a high frequency, methods based on quantiles of $|b(x) + \Delta(x)|$ for all $x \in \mathcal{R}$ are, in an asymptotic sense, based on quantiles of the "distribution" of $|b(x)|$ for $x \in \mathcal{R}$. Hence, if $\hat{\alpha}_\xi(\alpha_0)$ is taken to be the $(1 - \xi)$-level quantile of the "distribution" of $\hat{\alpha}(x, \alpha_0)$, as suggested four paragraphs above, then $\hat{\alpha}_\xi(\alpha_0)$ will converge, as $n \to \infty$, to a number $\beta = \alpha_\xi(\alpha_0)$ say, that solves the equation

$$\frac{1}{\int_{\mathcal{R}} dx} \int_{\mathcal{R}} I\{\alpha(x, \alpha_0) \le \beta\}\, dx = 1 - \xi\,, \quad (2.11)$$

or, in cases where the function b, in (4.8), vanishes on a set of nonzero measure, $\alpha_\xi(\alpha_0)$ equals the infimum of values β such that the left-hand side does not exceed

the right-hand side. Going back to (2.4) we see that this is exactly the value we need in order to ensure that the asymptotic coverage of the band equals $1 - \alpha_0$ for all x such that $\alpha(x, \alpha_0) \le \alpha_\xi(\alpha_0)$.

This approach becomes steadily more conservative, for a proportion $1 - \xi$ of values of x, as we decrease ξ towards 0. (The proportion referred to here corresponds to the set of x such that $\alpha(x, \alpha_0) \le \alpha_\xi(\alpha_0)$; see (2.10).) Nevertheless, there may be some values of x for which the associated confidence interval for $g(x)$ is slightly anti-conservative. These are the $x \in \mathcal{R}$ for which $|b(x)|$ (where b is as in (2.4)) is relatively large. We use the word "slightly" since, if the function b is smooth, then, as $1 - \xi$ increases to 1, the effect of bias on the asymptotic coverage of confidence intervals for $g(x)$, when x is in the anti-conservative region of $\mathcal{R}$, converges uniformly to 0. To appreciate why, note that, for x in the anti-conservative region, the size of $|b(x)|$, which is causing the anti-conservatism, is close to its size in nearby places where conservatism is found, and so the coverage errors are close too.

If $\xi = 0$, so that $\hat\alpha_1(\alpha_0) = \sup_{x \in \mathcal{R}} \hat\alpha(x, \alpha_0)$, then we obtain a band which, in the one-dimensional example treated from (2.6) down, has width $O\{(nh)^{-1/2} (\log n)^{1/2}\}$ rather than $O\{(nh)^{-1/2}\}$. In comparison, for each fixed $\xi > 0$ the band is of width $O\{(nh)^{-1/2}\}$.

In summary, our algorithm produces confidence bands that, in asymptotic terms, cover $g(x)$ with probability at least $1 - \alpha_0$, for a given known α_0 such as $\alpha_0 = 0.05$, for at least a proportion $1 - \xi$ of values of $x \in \mathcal{R}$. If we take $\xi = \alpha_0$ then we have a method for constructing confidence bands "that cover $g(x)$ with probability at least $1 - \alpha_0$ for at least a proportion $1 - \alpha_0$ of values x."

2.3. The algorithm in detail

Step 1. Estimators of g and σ^2. Construct a conventional nonparametric estimator $\hat g$ of g. Use a standard empirical method (for example, cross-validation or a plug-in rule), designed to minimise mean L_p error for some p in the range $1 \le p < \infty$, to choose the smoothing parameters on which $\hat g$ depends. For example, if the design is univariate then a local linear estimator of $g(x)$ is given by

$$\hat g(x) = \frac{1}{n} \sum_{i=1}^{n} A_i(x)\, Y_i\,, \tag{2.12}$$

where

$$A_i(x) = \frac{S_2(x) - \{(x - X_i)/h\}\, S_1(x)}{S_0(x)\, S_2(x) - S_1(x)^2}\, K_i(x)\,, \tag{2.13}$$

$S_k(x) = n^{-1} \sum_i \{(x - X_i)/h\}^k K_i(x)$, $K_i(x) = h^{-1} K\{(x - X_i)/h\}$, K is a kernel function and h is a bandwidth.

There is an extensive literature on computing estimators $\hat{\sigma}^2$ of the error variance $\sigma^2 = \mathrm{var}(\epsilon)$; see, for example, Rice (1984), Buckley et al. (1988), Gasser et al. (1986), Müller and Stadtmüller (1987, 1992), Hall et al. (1990), Hall and Marron (1990), Seifert et al. (1993), Neumann (1994), Müller and Zhao (1995), Dette et al. (1998), Fan and Yao (1998), Müller et al. (2003), Munk et al. (2005), Tong and Wang (2005), Brown and Levine (2007), Cai et al. (2009) and Mendez and Lohr (2011). It includes residual-based estimators, which we introduce at (2.15) below, and methods based on differences and generalised differences. An example of the latter approach, in the case of univariate design, is the following estimator due to Rice (1984):

$$\hat{\sigma}^2 = \frac{1}{2\,(n-1)} \sum_{i=2}^{n} (Y_{[i]} - Y_{[i-1]})^2\,, \tag{2.14}$$

where $Y_{[i]}$ is the concomitant of $X_{(i)}$ and $X_{(1)} \le \ldots \le X_{(n)}$ is the sequence of order statistics derived from the design variables.

As in section 2.2, let $s(\mathcal{X})(x)^2\,\hat{\sigma}^2$ denote an estimator of the variance of $\hat{g}(x)$, where $s(\mathcal{X})(x)$ depends on the data only through the design points, and $\hat{\sigma}^2$ estimates error variance, for example being defined as at (2.14) or (2.15). In the local linear example, introduced in (2.12) and (2.13), we take $s(\mathcal{X})(x)^2 = \kappa/\{nh\,\hat{f}_X(x)\}$, where $\kappa = \int K^2$ and $\hat{f}_X(x) = (nh^1)^{-1} \sum_{1 \le i \le n} K_1\{(x - X_i)/h^1\}$ is a standard kernel density estimator, potentially constructed using a bandwidth h^1 and kernel K_1 different from those used for $\hat{g}$. There are many effective, empirical ways of choosing h^1, and any of those can be used.

Step 2. Computing residuals. Using the estimator $\hat{g}$ from step (1), calculate initial residuals $\tilde{\epsilon}_i = Y_i - \hat{g}(X_i)$, put $\bar{\epsilon} = n^{-1} \sum_i \tilde{\epsilon}_i$, and define the centred residuals by $\hat{\epsilon}_i = \tilde{\epsilon}_i - \bar{\epsilon}$.

A conventional, residual-based estimator of σ^2, alternative to the estimator at (2.14), is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i^2\,. \tag{2.15}$$

The estimator at (2.14) is root-n consistent for σ^2, whereas the estimator at (2.15) converges at a slower rate unless an undersmoothed estimator of $\hat{g}$ is used when computing the residuals. This issue is immaterial to the theory in section 4, although it tends to make the estimator at (2.14) a little more attractive.

Step 3. Computing bootstrap resample. Construct a resample $\mathcal{Z}^* = \{(X_i, Y_i^*), 1 \le i \le n\}$, where $Y_i^* = \hat{g}(X_i) + \epsilon_i^*$ and the ϵ_i^*s are obtained by sampling from $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ randomly, with replacement, conditional on $\mathcal{X}$. Note that, since regression is conventionally undertaken conditional on the design sequence, then the X_is are not resampled, only the Y_is.

Step 4. Bootstrap versions of $\hat{g}$, $\hat{\sigma}^2$ and $\mathcal{B}(\alpha)$. From the resample drawn in step 3, but using the same smoothing parameter employed to construct $\hat{g}$, compute the bootstrap version $\hat{g}^*$ of $\hat{g}$. (See section 2.4 for discussion of the smoothing parameter issue.) Let $\hat{\sigma}^{*2}$ denote the bootstrap version of $\hat{\sigma}^2$, obtained when the latter is computed from $\mathcal{Z}^*$ rather than $\mathcal{Z}$, and construct the bootstrap version of $\mathcal{B}(\alpha)$, at (2.2):

$$\mathcal{B}^*(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat{g}^*(x) - s(\mathcal{X})(x)\,\hat{\sigma}^*\,z_{1-(\alpha/2)} \le y \le \hat{g}^*(x) + s(\mathcal{X})(x)\,\hat{\sigma}^*\,z_{1-(\alpha/2)}\Big\}\,. \tag{2.16}$$

Note that $s(\mathcal{X})$ is exactly the same as in (2.2); again this is a consequence of the fact that we are conducting inference conditional on the design points.

If, as in the illustration in step 1, the design is univariate and local linear estimators are employed, then $\hat{g}^*(x) = n^{-1} \sum_{1 \le i \le n} A_i(x)\,Y_i^*$ where $A_i(x)$ is as at (2.13). The bootstrap analogue of the variance formula (2.14) is $\hat{\sigma}^{*2} = \{2\,(n-1)\}^{-1} \sum_{2 \le i \le n} (Y^*_{[i]} - Y^*_{[i-1]})^2$, where, if the ith largest order statistic $X_{(i)}$ equals X_j, then $Y^*_{[i]} = \hat{g}(X_j) + \epsilon_j^*$.

Step 5. Estimator of coverage error. The bootstrap estimator $\hat{\pi}(x,\alpha)$ of the probability $\pi(x,\alpha)$ that $\mathcal{B}(\alpha)$ covers $(x, g(x))$ is defined by:

$$\hat{\pi}(x,\alpha) = P\big\{(x,\hat{g}(x)) \in \mathcal{B}^*(\alpha) \,\big|\, \mathcal{X}\big\}\,, \tag{2.17}$$

and is computed, by Monte Carlo simulation, in the form

$$\frac{1}{B} \sum_{b=1}^{B} I\big\{(x,\hat{g}(x)) \in \mathcal{B}_b^*(x,\alpha)\big\}\,, \tag{2.18}$$

where $\mathcal{B}_b^*(x,\alpha)$ denotes the bth out of B bootstrap replicates of $\mathcal{B}^*(\alpha)$, where the latter is as at (2.16). The estimator at (2.17) is completely conventional, and in particular, no additional or nonstandard smoothing is needed.

Step 6. Constructing final confidence band. Define $\hat{\alpha}(x,\alpha_0)$ to be the solution, in α, of $\hat{\pi}(x,\alpha) = 1 - \alpha_0$, and let $\hat{\alpha}_\xi(\alpha_0)$ denote the $(1-\xi)$-level quantile of points in

the set $\{\hat\alpha(x,\alpha_0) : x \in \mathcal{R}\}$. Specifically:

> take $\mathcal{R}$ to be a subset of $\mathbb{R}^r$, superimpose on $\mathcal{R}$ a regular, r-dimensional, rectangular grid with edge width δ, let $x_1,\ldots,x_N \in \mathcal{R}$ be the grid centres, let $\hat\alpha_\xi(\alpha_0,\delta)$ denote the $(1-\xi)$-level empirical quantile of the points $\hat\alpha(x_1,\alpha_0),\ldots,\hat\alpha(x_N,\alpha_0)$, and, for $\xi \in (0,1)$, let $\hat\alpha_\xi(\alpha_0)$ denote the limit infimum, as $\delta \to 0$, of the sequence $\hat\alpha_\xi(\alpha_0,\delta)$. (2.19)

(We use the limit infimum to avoid ambiguity, although under mild conditions the limit exists.) For a value $\xi \in (0, \frac12]$, construct the band $\mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}$. In practice we have found that taking $1-\xi$ in the range 0.7–0.8 generally gives a slight to moderate degree of conservatism. It can result in noticeable anti-conservatism if the sample size is too small.

2.4. Two remarks on the algorithm.

Remark 1. Smoothing parameter for $\hat g^*$. An important aspect of step 4 is that we use the same empirical smoothing parameters for both $\hat g^*$ and $\hat g$, even though, in some respects, it might seem appropriate to use a bootstrap version of the smoothing parameters for $\hat g$ when estimating $\hat g^*$. However, since smoothing parameters should be chosen to effect an optimal tradeoff between bias and stochastic error, and the bias of $\hat g$ is not estimated accurately by the conventional bootstrap used in step 3 above, then the bootstrap versions of smoothing parameters, used to construct $\hat g^*$, are generally not asymptotically equivalent to their counterparts used for $\hat g$. This can cause difficulties. The innate conservatism of our methodology accommodates the slightly nonstandard smoothing parameter choice in step 4. Moreover, by not having to recompute the bandwidth at every bootstrap step we substantially reduce computational labour.

Remark 2. Heteroscedasticity. A heteroscedastic generalisation of the model at (2.1) has the form

$$Y_i = g(X_i) + \sigma(X_i)\,\epsilon_i\,, \tag{2.20}$$

where the ϵ_is have zero mean and unit variance, and $\sigma(x)$ is a nonnegative function that is estimated consistently by $\hat\sigma(x)$, say, computed from the dataset $\mathcal{Z}$ using either parametric or nonparametric methods. In this setting the variance of $\hat g(x)$ generally can be estimated by $s(\mathcal{X})^2\,\hat\sigma(x)^2$, where $s(\mathcal{X})$ is a known function of the design points, and the confidence band at (2.2) should be replaced by

$$\mathcal{B}(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat g(x) - s(\mathcal{X})(x)\,\hat\sigma(x)\,z_{1-(\alpha/2)} \le y \le \hat g(x) + s(\mathcal{X})(x)\,\hat\sigma(x)\,z_{1-(\alpha/2)}\Big\}\,.$$

The model for generating bootstrap data now has the form: $Y_i^* = \hat{g}(X_i) + \hat{\sigma}(X_i)\,\epsilon_i^*$, instead of: $Y_i^* = \hat{g}(X_i) + \epsilon_i^*$ in step 4; and the ϵ_i^*s are resampled conventionally from residual approximations to the ϵ_is.

With these modifications, the algorithm described in steps 1–6 can be implemented as before, and the resulting confidence bands have similar properties. In particular, if we redefine $\mathcal{B}^*(\alpha)$ by

$$\mathcal{B}^*(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat{g}^*(x) - s(\mathcal{X})(x)\,\hat{\sigma}^*(x)\,z_{1-(\alpha/2)} \le y \le \hat{g}^*(x) + s(\mathcal{X})(x)\,\hat{\sigma}^*(x)\,z_{1-(\alpha/2)}\Big\}$$

(compare (2.16)), and, using this new definition, continue to define $\hat{\pi}(x,\alpha)$ as at (2.17) (computed as at (2.18)); and if we continue to define $\alpha = \hat{\alpha}(x,\alpha_0)$ to be the solution of $\hat{\pi}(x,\alpha) = 1-\alpha_0$, and to define $\hat{\alpha}_\xi(\alpha_0)$ as in (2.19); then the confidence band $\mathcal{B}\{x, \hat{\alpha}_\xi(\alpha_0)\}$ is asymptotically conservative for at least a proportion $1-\xi$ of values $x \in \mathcal{R}$. (Moreover, as explained in the second-last paragraph of section 2.2, for the complementary proportion of values x the coverage of the confidence interval for $g(x)$ is close to 0.) This approach can be justified intuitively as in section 2.2, noting that, in the context of the model at (2.20), the expansion at (2.6) should be replaced by:

$$E\{\hat{g}^*(x) \,|\, \mathcal{Z}\} - \hat{g}(x) = c_1\,g''(x)\,h^2 + (nh)^{-1/2}\,\sigma(x)\,f_X(x)^{-1/2}\,W(x/h) + \text{negligible terms}\,.$$

2.5. Percentile bootstrap confidence bands. The methods discussed above are based on the symmetric, asymptotic confidence band $\mathcal{B}(\alpha)$, which in turn is founded on a normal approximation. The approach uses a single application of the bootstrap for calibration, so as to reduce coverage error. However, particularly if we would prefer the bands to be placed asymmetrically on either side of the estimator $\hat{g}$ so as to reflect skewness of the distribution of experimental errors, the initial confidence band $\mathcal{B}(\alpha)$, at (2.2), could be constructed using bootstrap methods, and a second iteration of the bootstrap, resulting in a double bootstrap method, could be used to refine coverage accuracy.

The first bootstrap implementation is undertaken using step 4 of the algorithm in section 2.3, and allows us to define the critical point $\hat{z}_\beta(x)$ by

$$P\big\{\hat{g}^*(x) - \hat{g}(x) \le s(\mathcal{X})\,\hat{z}_\beta \,|\, \mathcal{Z}\big\} = \beta\,, \tag{2.21}$$

for $\beta \in (0,1)$. The confidence band $\mathcal{B}(\alpha)$ is now re-defined as

$$\mathcal{B}(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat{g}(x) - s(\mathcal{X})(x)\,\hat{z}_{1-(\alpha/2)} \le y \le \hat{g}(x) + s(\mathcal{X})(x)\,\hat{z}_{1-(\alpha/2)}\Big\}\,. \quad (2.22)$$

The remainder of the methodology can be implemented in the following six-step algorithm.

(1) Calculate the uncentred bootstrap residuals, $\tilde{\epsilon}_i^* = Y_i^* - \hat{g}^*(X_i)$. (2) Centre them to obtain $\hat{\epsilon}_i^* = \tilde{\epsilon}_i^* - \bar{\epsilon}_i^*$, where $\bar{\epsilon}^* = n^{-1}\sum_i \tilde{\epsilon}_i^*$. (3) Draw a double-bootstrap resample, $\mathcal{Z}^{**} = \{(X_i, Y_i^{**}), 1 \le i \le n\}$, where $Y_i^{**} = \hat{g}^*(X_i) + \epsilon_i^{**}$ and the ϵ_i^{**}s are sampled randomly, with replacement, from the $\hat{\epsilon}_i^*$s. (4) Construct the bootstrap-world version $\mathcal{B}^*(\alpha)$ of the band $\mathcal{B}(\alpha)$ at (2.22), defined by

$$\mathcal{B}^*(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat{g}^*(x) - s(\mathcal{X})(x)\,\hat{z}^*_{1-(\alpha/2)} \le y \le \hat{g}^*(x) + s(\mathcal{X})(x)\,\hat{z}^*_{1-(\alpha/2)}\Big\}\,,$$

where, reflecting (2.21), $\hat{z}^*_\beta$ is defined by

$$P\big\{\hat{g}^{**}(x) - \hat{g}^*(x) \le s(\mathcal{X})\,\hat{z}^*_\beta \,|\, \mathcal{Z}^*\big\} = \beta\,,$$

and $\mathcal{Z}^*$ is defined as in step 3 of the algorithm in section 2.3. (5) For this new definition of $\mathcal{B}^*(\alpha)$, define $\hat{\pi}(x,\alpha)$ as at (2.17). (6) Define $\hat{\alpha}_\xi(\alpha_0)$ as in (2.19), and take the final confidence band to be $\mathcal{B}\{\hat{\alpha}_\xi(\alpha_0)\}$, where $\mathcal{B}(\alpha)$ is as at (2.22).

There is also a percentile-t version of this methodology, using our our quantile-based definition of $\hat{\alpha}_\xi(\alpha_0)$.

2.6. Confidence bands for probability densities. Analogous methods can be used effectively to construct confidence bands for probability densities. We consider here the version of the single-bootstrap technique introduced in section 2.3, when it is adapted so as to construct confidence bands for densities of r-variate probability distributions. Specifically, let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote a random sample drawn from a distribution with density f, let h be a bandwidth and K a kernel, and define the kernel estimator of f by

$$\hat{f}(x) = \frac{1}{nh^r}\sum_{i=1}^n K\Big(\frac{x - X_i}{h}\Big)\,.$$

This estimator is asymptotically normally distributed with variance $(nh^r)^{-1}\,\kappa\,f(x)$, where $\kappa = \int K^2$, and so a naive, pointwise confidence band for $f(x)$ is given by

$$\mathcal{B}(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat{f}(x) - \big[(nh^r)^{-1}\,\kappa\,\hat{f}(x)\big]^{1/2}\,z_{1-(\alpha/2)} \le y \le \hat{f}(x) + \big[(nh^r)^{-1}\,\kappa\,\hat{f}(x)\big]^{1/2}\,z_{1-(\alpha/2)}\Big\}\,;$$

compare (2.2).

To correct $\mathcal{B}(\alpha)$ for coverage error, draw a random sample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ from the distribution with density $\hat{f}_X$, and define $\hat{f}^*$ to be the corresponding kernel estimator of $\hat{f}$, based on $\mathcal{X}$ rather than $\mathcal{X}^*$:

$$\hat{f}^*(x) = \frac{1}{nh^r} \sum_{i=1}^{n} K\Big(\frac{x - X_i^*}{h}\Big) .$$

Importantly, we do not generate $\mathcal{X}^*$ simply by resampling from $\mathcal{X}$. Analogously to (2.16), the bootstrap version of $\mathcal{B}(\alpha)$ is

$$\mathcal{B}^*(\alpha) = \Big\{(x, y) : x \in \mathcal{R}\,,\ \hat{f}^*(x) - \big[(nh^r)^{-1}\,\kappa\,\hat{f}^*(x)\big]^{1/2}\, z_{1-(\alpha/2)} \le y \\ \le \hat{f}^*(x) + \big[(nh^r)^{-1}\,\kappa\,\hat{f}^*(x)\big]^{1/2}\, z_{1-(\alpha/2)}\Big\} .$$

For the reasons given in Remark 1 in section 2.4 we use the same bandwidth, h, for both $\mathcal{B}(\alpha)$ and $\mathcal{B}^*(\alpha)$.

Our bootstrap estimator $\hat{\pi}(x, \alpha)$ of the probability $\pi(x, \alpha) = P\{(x, f(x)) \in \mathcal{B}(\alpha)\}$ that $\mathcal{B}(\alpha)$ covers $(x, f(x))$, is given by $\hat{\pi}(x, \alpha) = P\{(x, \hat{g}(x)) \in \mathcal{B}^*(\alpha) \,|\, \mathcal{X}\}$. As in step 6 of the algorithm in section 2.3, for a given desired coverage level $1 - \alpha_0$, let $\alpha = \hat{\alpha}(x, \alpha_0)$ be the solution of $\hat{\pi}(x, \alpha) = 1 - \alpha_0$, and define $\hat{\alpha}_\xi(\alpha_0)$ as in (2.19). Our final confidence band is $\mathcal{B}\{\hat{\alpha}_\xi(\alpha_0)\}$. For a proportion of at least $1 - \xi$ of the values of $x \in \mathcal{R}$, the limit of the probability that this band covers $f(x)$ is not less than $1 - \alpha_0$, and for the remainder of values x the coverage error is close to 0.

In the cases $r = 1$ and 2, which are really the only cases where confidence bands can be depicted, theoretical results analogous to those in section 4, for regression, can be developed using Hungarian approximations to empirical distribution functions. See, for example, Theorem 3 of Kómlós, Major and Tusnády (1976) for the case $r = 1$, and Tusnády (1977) and Massart (1989) for $r \ge 2$. In the univariate case, the analogue of (2.6) is

$$E\{\hat{f}^*(x) \,|\, \mathcal{Z}\} - \hat{f}(x) = \tfrac{1}{2}\,\kappa_2\, f''(x)\, h^2 + (nh)^{-1/2}\, f(x)^{1/2}\, V(x/h) \\ + \text{negligible terms}\,,$$

and (2.8) holds as stated before; there, for constants c_7 and c_8, $\kappa_2 = \int u^2\, K(u)\, du$, we define $b(x) = -c_7\, f''(x)\, f(x)^{-1/2}$, $\Delta(x) = -c_8\, V(x)$, and V is a stationary Gaussian process with zero mean and covariance $K'' * K''$.

Alternative to the definition of $\mathcal{B}(\alpha)$ above, a confidence band based on the square-root transform, reflecting the fact that the asymptotic variance of $\hat{f}$ is proportional to f, could be used. Percentile and percentile-t methods, using our quantile-based method founded on $\hat{\alpha}_\xi(\alpha_0)$, can also be used.

4. NUMERICAL PROPERTIES

This section summarises the results of a simulation study addressing the finite-sample performance of the method described in section 2. In particular, we report empirical coverage probabilities of nominal 95% confidence intervals for $g(x)$, using different values of x, different quantiles (corresponding to different values of $1-\xi$), different choices of g and different design densities f_X, and different sample sizes n.

Data were generated randomly from the model at (2.1), where the experimental errors ϵ_i were distributed independently as $\mathrm{N}(0,1)$. We also simulated cases where the errors were distributed as $\mathrm{N}(0,0.04)$ and $\mathrm{N}(0,0.25)$, although, since the results were similar to those in the $\mathrm{N}(0,1)$ setting, they are not reported here.

We present results for $g(x) \equiv x^2$ and $g(x) \equiv \Phi(3x)$, where Φ denotes the standard normal distribution function. Graphs of these functions on the interval $[-1,1]$ are given in Figure 4. The X_is were sampled randomly from either the uniform distribution on $[-1,1]$ or from the distribution with probability density $f_X(x) = b\cos\{\frac{1}{2}(\pi - 0.01)\,x\}$ on $[-1,1]$, where b was chosen to make f_X a proper density on that interval. Below, we refer to that as the "cosine density."

The estimator $\hat{g}$ was constructed using local-linear kernel regression and the biweight kernel. The bandwidth was computed using the plug-in method of Ruppert et al. (1995). Almost identical results were obtained using least-squares cross validation bandwidths, and so for brevity they are not reported here. The variance of ϵ_i was estimated using (2.14), and f_X was estimated using a standard normal kernel with the bandwidth obtained by the "normal reference" method. The sample sizes were $n = 50$, 100, and 200. Each experiment involved 100 Monte Carlo replications, and 500 bootstrap resamples were used for each each replication.

The empirical coverage probabilities of nominal 95%, two-sided confidence intervals are shown in Figures 1 and 2. For the quantiles corresponding to $1-\xi = 0.7$ and 0.8 the coverage levels are slightly conservative, except for values of x near the boundaries of the support of f_X, where the coverage probabilities associated with these quantiles tend to be below 95% on account of data sparsity. (Near the boundaries the estimator is limited largely to data on just one side of the point being

estimated, but away from the boundaries it has access to data on both sides.) In the case of the cosine density the sparsity problems at boundaries are more serious, and so the anti-conservatism at boundaries is more noticeable.

Figure 3 shows average empirical critical values according to quantile level, $1-\xi$, with $n = 50$. These provide an indication of the widths of the confidence intervals for different quantile levels. The values in Figure 3 may be compared with the critical value of 1.96, which would be the correct asymptotic critical value if $\hat{g}$ had no asymptotic bias. (In captions to figures we refer to this as the "asymptotic" case.) The critical values at the quantiles corresponding to $1-\xi = 0.7$ and 0.8 are only slightly larger than 1.96, but the critical values are significantly larger for quantiles at levels above $1-\xi = 0.9$ or 0.95.

Figure 1 shows typical nominal pointwise 95% confidence intervals for $g(x) = x^2$, in the case of the cosine design density. The solid lines in the figure show the functions $g(x) = x^2$ and $g(x) = \Phi(3x)$. The dashed and dotted lines show confidence intervals using the asymptotic critical value and the critical values obtained by implementing the method of section 2, employing quantiles corresponding to $1-\xi = 0.7$ and 0.8. As can be seen, the three bands differ mainly in terms of width.

FIGURES ARE GIVEN AT END OF PAPER

4. THEORETICAL PROPERTIES

4.1. Theoretical background. In the present section we describe theoretical properties of bootstrap methods for estimating the distribution of $\hat{g}$. In section 4.2 we apply our results to underpin the arguments in section 2 that motivated our methodology. A proof of Theorem 4.1, below, is given in Appendix A of Hall and Horowitz (2012).

We take $\hat{g}(x)$ to be a local polynomial estimator of $g(x)$, defined by (A.16) and (A.17). The asymptotic variance, Avar, of the local polynomial estimator $\hat{g}$ at x is given by

$$\text{Avar}\{\hat{g}(x)\} = D_1\,\sigma^2\,f_X(x)^{-1}\,(nh_1^r)^{-1}\,, \tag{4.1}$$

where $D_1 > 0$ depends only on the kernel and $\sigma^2 = \text{var}(\epsilon)$. (If $r = k = 1$ then $D_1 = \kappa \equiv \int K^2$.) With this in mind we take the estimator $s(\mathcal{X})(x)^2\,\hat{\sigma}^2$, introduced in section 2.2, of the variance of $\hat{g}(x)$, to be $D_1\,\hat{\sigma}^2\,\hat{f}_X(x)^{-1}\,(nh^r)^{-1}$, where $\hat{f}_X$ is an estimator of the design density f_X and was introduced in step 1 of the algorithm in section 2.3.

We assume that:

(a) the data pairs (X_i, Y_i) are generated by the model at (2.1), where the design variables X_i are identically distributed, the experimental errors ϵ_i are identically distributed, and the design variables and errors are totally independent; (b) $\mathcal{R}$ is a closed, nondegenerate rectangular prism in $\mathbb{R}^r$; (c) the estimator $\hat{g}$ is constructed by fitting a local polynomial of degree $2k-1$, where $k \geq 1$; (d) $\hat{f}_X$ is weakly and uniformly consistent, on $\mathcal{R}$, for the common density f_X of the r-variate design variables X_i; (e) g has $2k$ Hölder-continuous derivatives on an open set containing $\mathcal{R}$; (f) f_X is bounded on $\mathbb{R}^r$, and Hölder continuous and bounded away from zero on an open subset of $\mathbb{R}^r$ containing $\mathcal{R}$; (g) the bandwidth, h, used to construct $\hat{g}$, is a function of the data in $\mathcal{Z}$ and, for constants $C_1, C_2 > 0$, satisfies $P\{|h - C_1\, n^{-1/(r+4k)}| > n^{-(1+C_2)/(r+4k)}\} \to 0$, and moreover, for constants $0 < C_3 < C_4 < 1$, $P(n^{-C_4} \leq h \leq n^{-C_3}) = 1 - O(n^{-C})$ for all $C > 0$; (h) the kernel used to construct $\hat{g}$, at (2.12), is a spherically symmetric, compactly supported probability density, and has C_5 uniformly bounded derivatives on $\mathbb{R}^r$, where the positive integer C_5 is sufficiently large and depends on C_2; and (j) the experimental errors satisfy $E(\epsilon) = 0$ and $E|\epsilon|^{C_6} < \infty$, where $C_6 > 2$ is chosen sufficiently large, depending on C_2. (4.2)

The model specified by (c) is standard in nonparametric regression. The assumptions imposed in (b), on the shape of $\mathcal{R}$, can be generalised substantially and are introduced here for notational simplicity. The restriction to polynomials of odd degree, in (c), is made so as to eliminate the somewhat anomalous behaviour in cases where the degree is even. See Ruppert and Wand (1994) for an account of this issue in multivariate problems. Condition (d) asks only that the design density be estimated uniformly consistently. The assumptions imposed on g and f_X in (e) and (f) are close to minimal when investigating properties of local polynomial estimators of degree $2k-1$. Condition (g) is satisfied by standard bandwidth choice methods, for example those based on cross-validation or plug-in rules. The assertion, in (g), that h be approximately equal to a constant multiple of $n^{-1/(r+2k)}$ reflects the fact that h would usually be chosen to minimise a measure of asymptotic mean L_p error, for $1 \leq p < \infty$. Condition (h) can be relaxed significantly if we have in mind a particular method for choosing h. Smooth, compactly supported kernels, such as those required by (h), are commonly used in practice. The moment condition imposed in (j) is less restrictive than, for example, the assumption of normality.

In addition to (4.2) we shall, on occasion, suppose that:

the variance estimators $\hat{\sigma}^2$ and $\hat{\sigma}^{*2}$ satisfy $P(|\hat{\sigma} - \sigma| > n^{-C_8}) \to 0$ and $P(|\hat{\sigma}^* - \hat{\sigma}| > n^{-C_8}) \to 0$ for some $C_8 > 0$. (4.3)

In the case of the estimators $\hat\sigma^2$ defined at (2.14) and (2.15), if (4.2) holds then so too does (4.3).

Let $h_1 = C_1\, n^{-1/(r+4k)}$ be the deterministic approximation to the empirical bandwidth h asserted in (4.2)(g). Under (4.2) the asymptotic bias of a local polynomial estimator $\hat g$ of g, evaluated at x, is equal to $h_1^{2k}\, \nabla\, g(x)$, where ∇ is a linear form in the differential operators $(\partial/\partial x^{(1)})^{j_1} \ldots (\partial/\partial x^{(r)})^{j_r}$, for all choices of $j_1, \ldots, j_r$ such that each j_s is an even, positive integer, $j_1 + \ldots + j_r = 2k$ (the latter being the number of derivatives assumed of g in (4.2)(e)), and $x = (x^{(1)}, \ldots, x^{(r)})$. For example, if $r = k = 1$ then $\nabla = \frac{1}{2}\, \kappa_2\, (d/dx)^2$, where $\kappa_2 = \int u^2\, K(u)\, du$.

Recall that σ^2 is the variance of the experimental error ϵ_i. Let $L = K * K$, denoting the convolution of K with itself, and put $M = L - K$. Let W_1 be a stationary Gaussian process with zero mean and the following covariance function:

$$\mathrm{cov}\{W_1(x_1), W_1(x_2)\} = \sigma^2\, (M * M)(x_1 - x_2)\,. \tag{4.4}$$

Note that, since h_1 depends on n, then so too does the distribution of W_1. Our first result shows that (4.2) is sufficient for the expansion (2.6) to hold in the case of local polynomial estimators.

Theorem 4.1. *If* (4.2) *holds then, for each* n*, there exists a zero-mean Gaussian process* W*, having the distribution of* W_1 *and defined on the same probability space as the data* $\mathcal{Z}$*, such that for constants* $D_2, C_7 > 0$*,*

$$P\Bigg[\sup_{x\in\mathcal{R}} \Big|E\{\hat g^*(x)\,|\,\mathcal{Z}\} - \hat g(x) - \Big\{h_1^{2k}\, \nabla\, g(x) + D_2\, \big(nh_1^r\big)^{-1/2}\, f_X(x)^{-1/2}\, W(x/h_1)\Big\}\Big| > h_1^{2r}\, n^{-C_7}\Bigg] \to 0 \tag{4.5}$$

as $n \to \infty$*. If, in addition to* (4.2)*, we assume that* (4.3) *holds, then for some* $C_7 > 0$*,*

$$P\Bigg(\sup_{x\in\mathcal{R}} \sup_{z\in\mathbb{R}} \Big|P\Big[\hat g^*(x) - E\{\hat g^*(x)\,|\,\mathcal{Z}\} \le z\, \big\{D_1\, \hat\sigma^2\, \hat f_X(x)^{-1}\, (nh^r)^{-1}\big\}^{1/2}\ \Big|\ \mathcal{Z}\Big] - \Phi(z)\Big| > n^{-C_7}\Bigg) \to 0 \tag{4.6}$$

as $n \to \infty$*.*

Property (4.5) is a concise and more general version of the expansion at (2.6), which underpinned our heuristic motivation for our methodology. Result (4.6) asserts that the standard central limit theorem for $\hat g^*(x)$ applies uniformly in $x \in \mathcal{R}$.

In particular, the standard deviation estimator $\{D_1\,\hat\sigma^2\,\hat f_X(x)^{-1}\,(nh^r)^{-1}\}^{1/2}$, used to standardise $\hat g^* - E(\hat g^* \,|\, \mathcal{Z})$ on the left-hand side of (4.6), is none other than the conventional empirical form of the asymptotic variance of $\hat g$ at (4.1), and was used to construct the confidence bands described in sections 2.2 and 2.3. The only unconventional aspect of (4.6) is that the central limit theorem is asserted to hold uniformly in $x \in \mathcal{R}$, but this is unsurprising, given the moment assumption in (4.2)(j).

4.2. Theoretical properties of coverage error. Let $D_3 = D_1^{-1/2}\,\sigma^{-1}$ and $D_4 = D_2\,D_3$, and define

$$b(x) = -D_3\,f_X(x)^{1/2}\,\nabla\,g(x)\,, \quad \Delta(x) = -D_4\,W(x/h_1)\,, \tag{4.7}$$

where W is as in (4.5), these being the versions of $b(x)$ and $\Delta(x)$ at (2.7) and (2.9), respectively, in the present setting. Our first result in this section is a detailed version of (2.8):

Corollary 4.1. *If* (4.2) *and* (4.3) *hold then, with* $z = z_{1-(\alpha/2)}$ *and* $b(x)$ *and* $\Delta(x)$ *defined as above, we have for some* $C_9 > 0$,

$$P\biggl(\sup_{x\in\mathcal{R}}\Bigl|\hat\pi(x,\alpha) - \Bigl[\Phi\bigl\{z+b(x)+\Delta(x)\bigr\} - \Phi\bigl\{-z+b(x)+\Delta(x)\bigr\}\Bigr]\Bigr| > n^{-C_9}\biggr) \to 0 \tag{4.8}$$

as $n \to \infty$.

Next we give notation that enables us to assert, under specific assumptions, properties of coverage error of confidence bands stated in the paragraph immediately below (2.9). See particularly (4.11) in Corollary 4.2, below. Results (4.9) and (4.10) are used to derive (4.11), and are of interest in their own right because they describe large-sample properties of the quantities $\hat\alpha(x,\alpha_0)$ and $\hat\alpha_\xi(\alpha_0)$, respectively, in terms of which our confidence bands are defined; see section 2.3.

Given a desired coverage level $1-\alpha_0 \in (\frac12, 1)$, define $\hat\alpha(x,\alpha_0)$ and $\hat\alpha_\xi(\alpha_0)$ as at (2.5) and (2.19), respectively. Let $b(x)$ and $\Delta(x)$ be as at (4.7), put $d = b + \Delta$, and define $Z = Z(x,\alpha_0)$ to be the solution of

$$\Phi\{Z + d(x)\} - \Phi\{-Z + d(x)\} = 1 - \alpha_0\,.$$

Then $Z(x,\alpha_0) > 0$, and $A(x,\alpha_0) = 2\,[1-\Phi\{Z(x,\alpha_0)\}] \in (0,1)$. Define $\alpha = \alpha(x,\alpha_0)$ to be the solution of (2.11), and let $\alpha_\xi(\alpha_0)$ be the $(1-\xi)$-level quantile of the values of $\alpha(x,\alpha_0)$. Specifically, $\beta = \alpha_\xi(\alpha_0)$ solves equation (2.11). Define $\mathcal{R}_\xi(\alpha_0) = \{x \in \mathcal{R} : I[\alpha(x,\alpha_0) \le \alpha_\xi(\alpha_0)]\}$. Let the confidence band $\mathcal{B}(\alpha)$ be as at (2.2).

Corollary 4.2. *If* (4.2) *and* (4.3) *hold, then, for each* $C_{10}, C_{11} > 0$, *and as* $n \to \infty$,

$$P\Big\{ \sup_{x \in \mathcal{R} \,:\, |\Delta(x)| \le C_{10}} |\hat{\alpha}(x, \alpha_0) - A(x, \alpha_0)| > C_{11} \Big\} \to 0\,, \tag{4.9}$$

$$P\big\{ \hat{\alpha}_\xi(\alpha_0) \le \alpha_\xi(\alpha_0) + C_{11} \big\} \to 1\,, \tag{4.10}$$

for each $x \in \mathcal{R}_\xi(\alpha_0)$ *the limit infimum of the probability* $P[(x, g(x)) \in \mathcal{B}\{\hat{\alpha}_\xi(\alpha_0)\}]$, *as* $n \to \infty$, *is not less than* $1 - \alpha_0$. (4.11)

Property (4.10) implies that the confidence band $\mathcal{B}(\beta)$, computed using $\beta = \hat{\alpha}_\xi(\alpha_0)$, is no less conservative, in an asymptotic sense, than its counterpart when $\beta = \alpha_\xi(\alpha_0)$. This result, in company with (4.11), underpins our claims about the conservatism of our approach. Result (4.11) asserts that the asymptotic coverage of $(x, g(x))$ by $\mathcal{B}\{\hat{\alpha}_\xi(\alpha_0)\}$ is, for at most a proportion ξ of values of x, not less than $1 - \alpha_0$. Proofs of Corollaries 4.1 and 4.2 are given in Appendix B of Hall and Horowitz (2012).

REFERENCES

BJERVE, S., DOKSUM, K.A. AND YANDELL, B.S. (1985). Uniform confidence bounds for regression based on a simple moving average. *Scand. J. Statist.* **12**, 159–169.

BROWN, L.D. AND LEVINE, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *Ann. Statist.* **35**, 2219–2232.

BUCKLEY, M.J., EAGLESON, G.K. AND SILVERMAN, B.W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75**, 189–199.

CAI, T.T., LEVINE, M. AND WANG, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivar. Anal.* **100**, 126–136.

CHEN, S.X. (1996). Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika* **83**, 329–341.

CLAESKENS, G. AND VAN KEILEGOM, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.* **31**, 1852–1884.

DETTE, H., MUNK, A. AND WAGNER, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *J. Roy. Statist. Soc.* Ser. B **60**, 751–764.

EUBANK, R. L. AND SPECKMAN, P. L. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **88**, 1287–1301.

EUBANK, R.L. AND WANG, S. (1994). Confidence regions in non-parametric regression. *Scand. J. Statist.* **21**, 147–157.

FAN, J. AND YAO, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–660.

GASSER, T., SROKA, L. AND JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

GENOVESE, C. AND WASSERMAN, L. (2005). Nonparametric confidence sets for wavelet regression. *Ann. Statist.* **33**, 698–729.

GENOVESE, C. AND WASSERMAN, L. (2008). Adaptive confidence bands. *Ann. Statist.* **36**, 875–905.

HALL, P. (1992a). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20**, 675–694.

HALL, P. (1992b). On bootstrap confidence intervals in nonparametric regression. *Ann. Statist.* **20**, 695–711.

HALL, P. AND HOROWITZ, J. (2012). A simple bootstrap method for constructing confidence bands for functions—long version. Manuscript.

HALL, P., KAY, J.W. AND TITTERINGTON, D.M. (1990). Asymptotically optimal difference based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–528.

HALL, P. AND MARRON, J.S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77**, 415–419.

HALL, P. AND OWEN, A.B. (1993). Empirical likelihood confidence bands in density estimation. *J. Comput. Graph. Statist.* **2**, 273–289.

HALL, P. AND TITTERINGTON, D.M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multivariate Anal.* **27**, 228–254.

HÄRDLE, W. AND BOWMAN, A.W. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83**, 102–110.

HÄRDLE, W. HUET, S. AND MARRON, J.S. (1995). Better bootstrap confidence-intervals for regression curve estimation. *Statistics* **26**, 287–306.

HÄRDLE, W. AND MARRON, J.S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19**, 778–796.

KOMLÓS, J., MAJOR, P. AND TUSNÁDY, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **32**, 111–131.

KOMLÓS, J., MAJOR, P. AND TUSNÁDY, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **34**, 33–58.

LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17**, 1001–1008.

LOW, M.G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25**, 2547–2554.

MCMURRY, T.L. AND POLITIS, D.M. (2008). Bootstrap confidence intervals in nonparametric regression with built-in bias correction. *Statist. Probab. Lett.* **78**, 2463–2469.

MASSART, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT constructions. *Ann. Probab.* **17**, 266–291.

MENDEZ, G. AND LOHR, S. (2011). Estimating residual variance in random forest regression. *Comput. Statist. Data Anal.* **55**, 2937–2950.

MÜLLER, H.-G. AND STADTÜLLER, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610–635.

MÜLLER, H.-G. AND STADTÜLLER, U. (1992). On variance estimation with quadratic forms. *J. Statist. Plann. Inference* **35**, 213–231.

MÜLLER, H.-G. AND ZHAO, P.-L. (1995). On a semiparametric variance function model and a test for heteroscedasticity. *Ann. Statist.* **23**, 946–967.

MÜLLER, U.U., SCHICK, A. AND WEFELMEYER, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. *Statistics* **37**, 179–188.

MUNK, A., BISSANTZ, N., WAGNER, T. AND FRIETAG, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. Roy. Statist. Soc.* Ser. B **67**, 19–41.

NEUMANN, M.H. (1994). Fully data-driven nonparametric variance estimators. *Scand. J. Statist.* **25**, 189–212.

NEUMANN, M.H. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *Ann. Statist.* **23**, 1937–1959.

NEUMANN, M.H. AND POLZEHL, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *J. Nonparametric Statist.* **9**, 307–333.

PETROV, V.V. (1975). *Sums of Independent Random Variables*. Springer, Berlin.

PICARD, D. AND TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28**, 298–335.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215–1230.

RUPPERT, D. AND WAND, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.

SEIFERT, B., GASSER, T. AND WOLF, A. (1993). Nonparametric-estimation of residual variance revisited. *Biometrika* **80**, 373–383.

SUN, J. AND LOADER, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Ann. Statist.* **22**, 1328–1345.

TONG, T. AND WANG, Y. (2005). Estimating residual variance in nonparametric regression using least squares. *Biometrika* **92**, 821–830.

TUSNÁDY, G. (1977). A remark on the approximation of the sample DF in the multidimensional case. *Period. Math. Hungar.* **8**, 53–55.

WANG, Y.D. AND WAHBA, G. (1995). Bootstrap confidence-intervals and for smoothing splines and their comparison to Bayesian confidence-intervals. *J. Statist. Comput. Simul.* **51**, 263–279.

XIA, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. Roy. Statist. Soc.* Ser. B **60** 797–811.

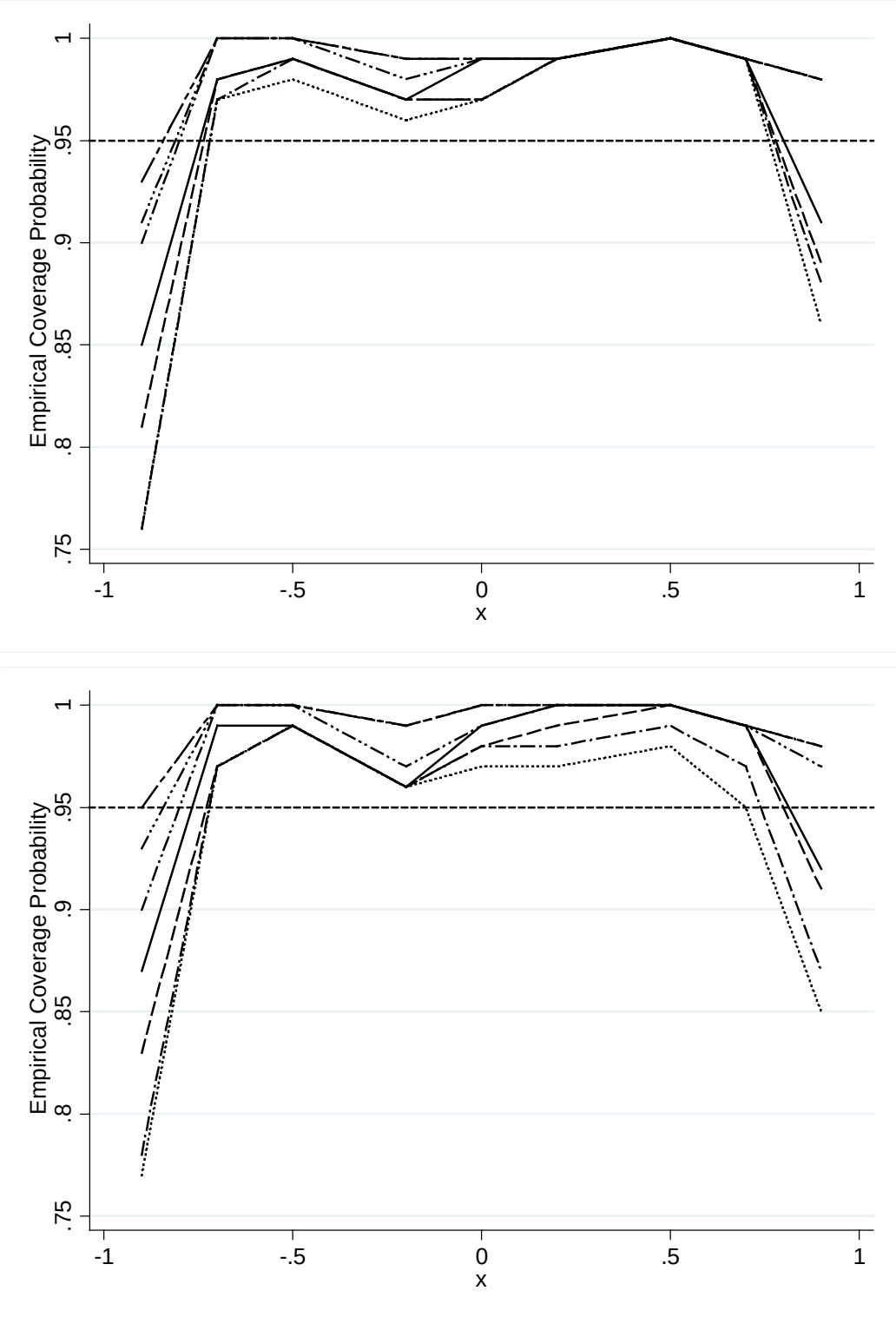

Figure 1: Empirical coverage probability as function of quantile and x for $n = 50$ and uniform density of X, with $g(x) = x^2$ in the top panel and $g(x) = \Phi(3x)$ in the bottom panel. The dashed horizontal line indicates a coverage probability of 0.95. Dots indicate 0.50 quantile, dash-dot indicates 0.60 quantile, dashes indicate 0.70 quantile, solid indicates 0.80 quantile, dash-dot-dot indicates 0.90 quantile, dash-dot-dot-dot indicates 0.95 quantile, and long dash-short dash indicates 0.98 quantile.

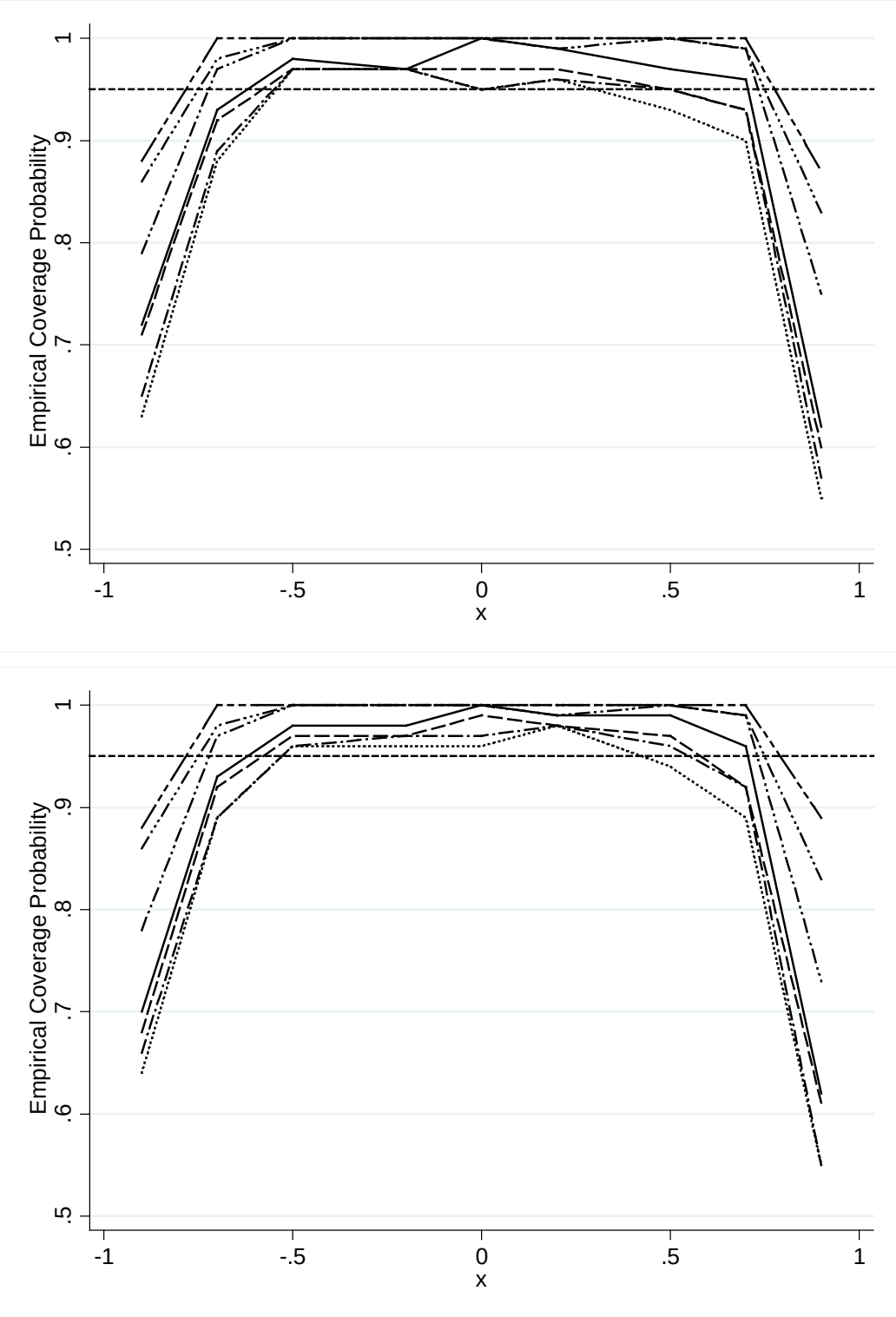

Figure 2: Empirical coverage probability as function of quantile and x for $n = 50$ and cosine density of X, with $g(x) = x^2$ in the top panel and $g(x) = \Phi(3x)$ in the bottom panel. The dashed horizontal line indicates a coverage probability of 0.95. Dots indicate 0.50 quantile, dash-dot indicates 0.60 quantile, dashes indicate 0.70 quantile, solid indicates 0.80 quantile, dash-dot-dot indicates 0.90 quantile, dash-dot-dot-dot indicates 0.95 quantile, and long dash-short dash indicates 0.98 quantile.

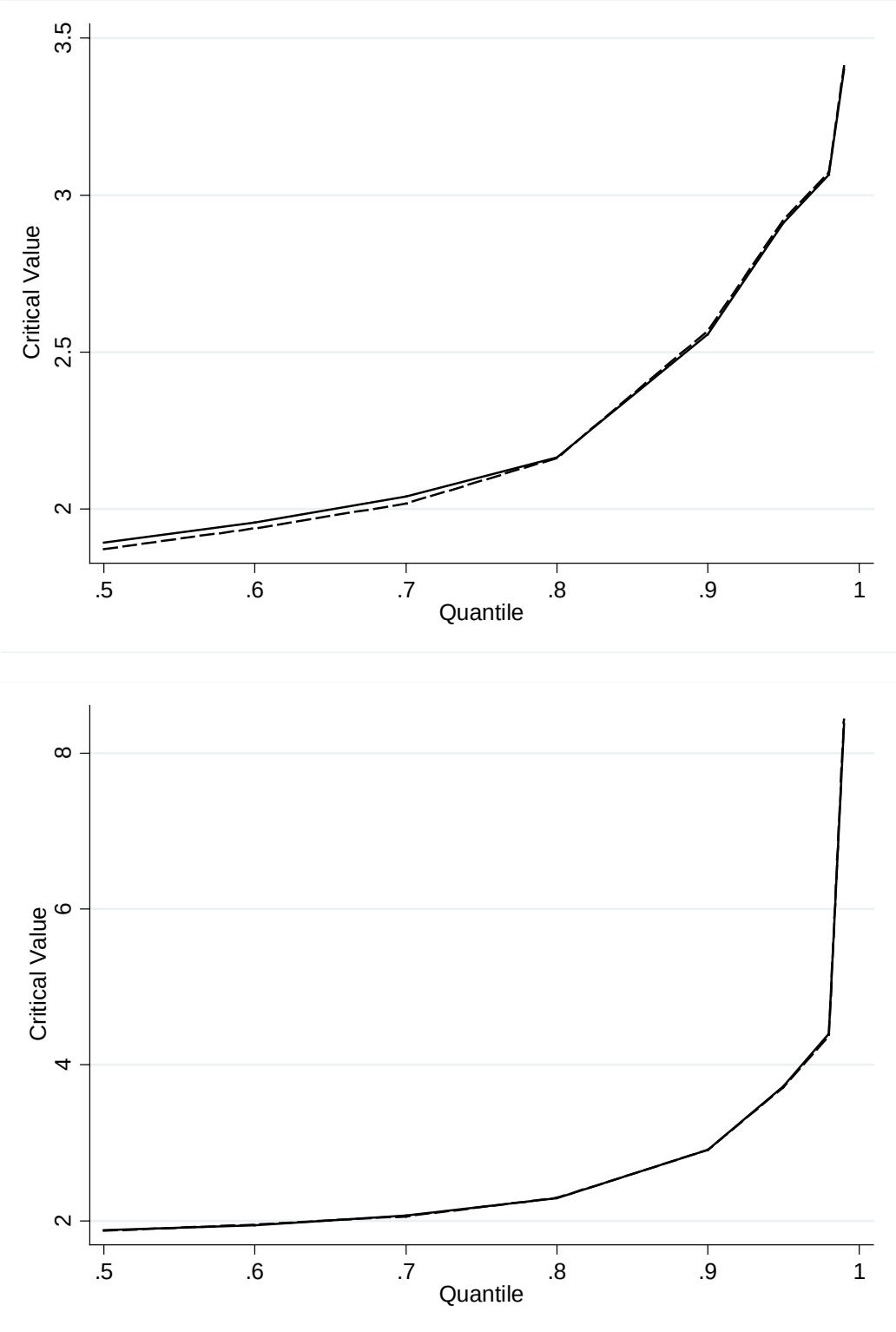

Figure 3: Average critical value as function of quantile for $n = 50$. Top panel is for uniform density of X; bottom panel is for cosine density. Solid line indicates $g(x) = x^2$. Dashed line indicates $g(x) = \Phi(3x)$.

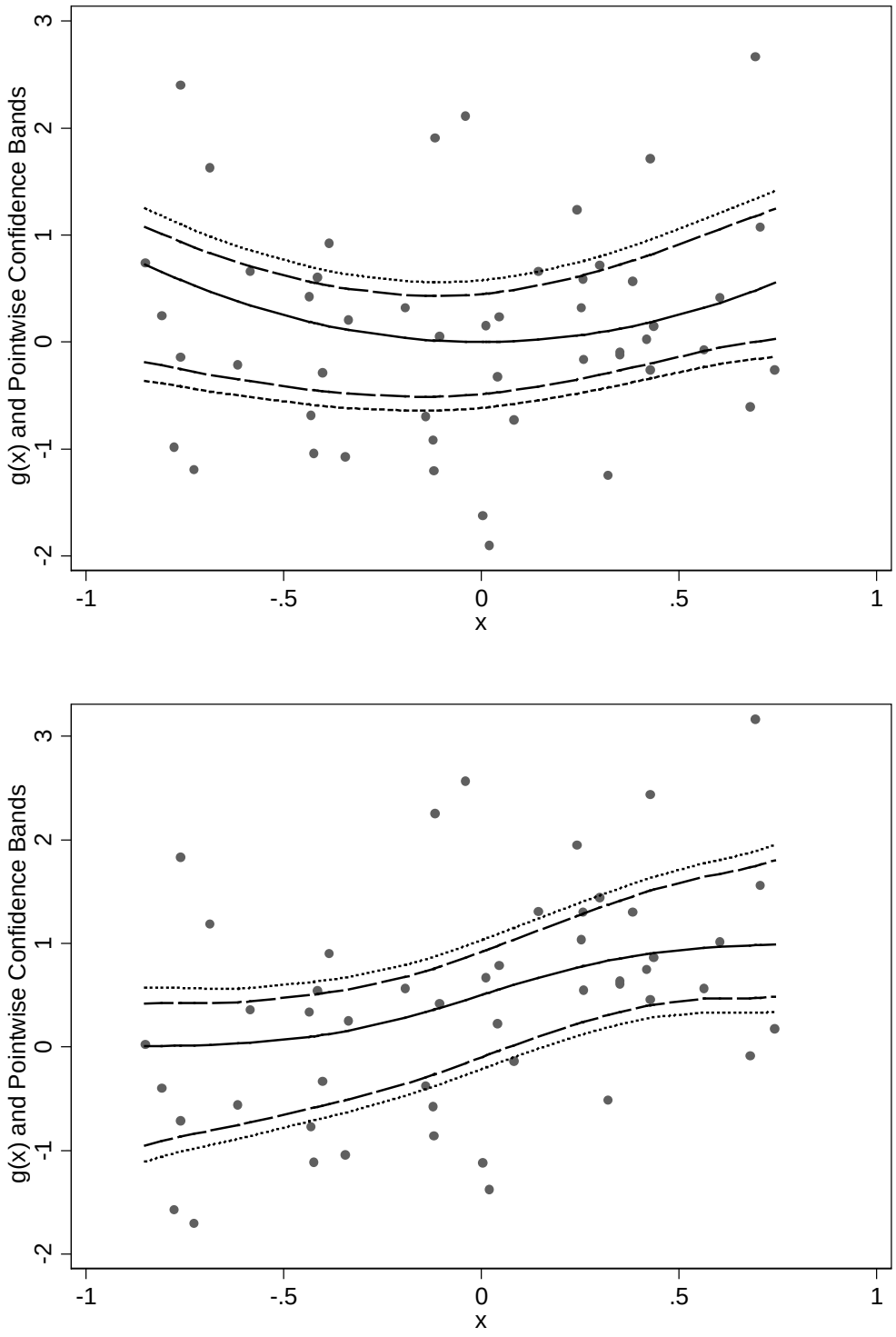

Figure 4: Nominal 95% pointwise confidence bands for $g(x) = x^2$ (top panel) and $g(x) = \Phi(3x)$ (bottom panel). Solid line indicates $g(x)$. Dashed line indicates confidence band at the 0.7 quantile. Dotted line indicates confidence band at the 0.8 quantile. Large dots are simulated data points.

APPENDICES A AND B
(NOT INTENDED FOR PUBLICATION)

APPENDIX A: PROOF OF THEOREM 4.1

We shall prove only (4.5), since (4.6) can be derived by adapting standard results on the rate of convergence in the central limit theorems for sums of independent random variables, for example Theorem 6, page 115 of Petrov (1975). In the present context the independent random variables are the quantities ϵ_i^* multiplied by weights depending only on $\mathcal{Z}$, which is conditioned on when computing the probability on the left-hand side of (4.6).

Step 1. Preliminaries. For the sake of clarity we give the proof only in the case $r = k = 1$, where $\hat{g}$ is defined by (2.12) and (2.13). However, in step 6 below we shall mention changes that have to be made for multivariate design and polynomials of higher degree. Define $\kappa_2 = \int u^2 K(u)\, du$ and $\kappa = \int K^2$.

Noting the model at (2.1), and defining

$$\begin{aligned} \tilde{g}(x) &= \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} A_{i_1}(x)\, A_{i_2}(X_{i_1})\, g(X_{i_2})\,, \\ e_1(x) &= \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} A_{i_1}(x)\, A_{i_2}(X_{i_1})\, \epsilon_{i_2}\,, \end{aligned} \tag{A.1}$$

where A_i as at (2.13), we have:

$$E\{\hat{g}^*(x) \,|\, \mathcal{Z}\} = \frac{1}{n} \sum_{i=1}^{n} A_i(x)\, \hat{g}(X_i) = \tilde{g}(x) + e_1(x)\,. \tag{A.2}$$

Writing $x_{i_1 i_2}$ for a quantity between 0 and $X_{i_2} - X_{i_1}$, and noting that $\sum_i A_i(x) \equiv n$ and $\sum_i A_i(x)\,(X_i - x) \equiv 0$, it can be shown that, for $x \in \mathcal{R}$,

$$\begin{aligned} \frac{1}{n} \sum_{i_2=1}^{n} A_{i_2}(X_{i_1})\, g(X_{i_2}) &= \frac{1}{n} \sum_{i_2=1}^{n} A_{i_2}(X_{i_1}) \Big\{ g(X_{i_1}) + (X_{i_2} - X_{i_1})\, g'(x) \\ &\qquad + \tfrac{1}{2}\,(X_{i_2} - X_{i_1})^2\, g''(X_{i_1} + x_{i_1 i_2}) \Big\} \\ &= g(X_{i_1}) + \tfrac{1}{2}\, h^2\, g''(X_{i_1}) + R(x, X_{i_1})\,, \end{aligned}$$

where

$$R(x, X_{i_1}) = \frac{1}{2n} \sum_{i_2=1}^{n} A_{i_2}(X_{i_1})\,(X_{i_2} - X_{i_1})^2 \left\{ g''(X_{i_1} + x_{i_1 i_2}) - g''(X_{i_1}) \right\}. \tag{A.3}$$

In this notation,

$$\begin{aligned}\tilde{g}(x) &= \frac{1}{n}\sum_{i=1}^{n} A_i(x)\left\{g(X_i) + \tfrac{1}{2}\,h^2\,g''(X_i) + R(x,X_i)\right\} \\ &= \hat{g}(x) + \tfrac{1}{2}\,h^2\,\kappa_2\,g''(x) - e_2(x) + \tfrac{1}{2}\,h^2\,R(x)\,, \end{aligned} \tag{A.4}$$

where

$$e_2(x) = \frac{1}{n}\sum_{i=1}^{n} A_i(x)\,\epsilon_i\,, \quad R(x) = \frac{1}{n}\sum_{i=1}^{n} A_i(x)\left\{R(x,X_i) + g''(X_i) - g(x)\right\}. \tag{A.5}$$

Step 2. Bound for $|R(x)|$. The bound is given at (A.13) below. Let K be the kernel discussed in (4.2)(h). Since K is supported on a compact interval $[-B_1, B_1]$, for some $B_1 > 0$ (see (4.2)(h)), then $A_{i_2}(X_{i_1}) = 0$ unless $|X_{i_2} - X_{i_1}| \le 2B_1 h$, and therefore the contribution of the i_2th term to the right-hand side of (A.3) equals zero unless $|x_{i_1 i_2}| \le 2B_1 h$. However, g'' is Hölder continuous on an open set $\mathcal{O}$ containing $\mathcal{R}$ (see (4.2)(e)), and so $|g''(x_1) - g''(x_2)| \le B_2\,|x_1 - x_2|^{B_3}$ for all $x_1, x_2 \in \mathcal{O}$, where $B_2, B_3 > 0$. Hence, by (A.3),

$$|R(x, X_{i_1})| \le \frac{B_4\,h^{2+B_3}}{n}\sum_{i_2=1}^{n} |A_{i_2}(X_{i_1})|\,, \tag{A.6}$$

where $B_4 = \frac{1}{2}\,B_2\,(2B_1)^{2+B_3}$. Now,

$$\frac{1}{n}\sum_{i=1}^{n} |A_i(x)| \le \frac{S_0(x)\,S_2(x) + B_1\,S_0(x)\,|S_1(x)|}{S_0(x)\,S_2(x) - S_1(x)^2}\,. \tag{A.7}$$

We shall show in Lemma A.1, in step 9, that the open set $\mathcal{O}$ containing $\mathcal{R}$ can be chosen so that, for some $B_5 > 1$ and all $B_6 > 0$,

$$P\left\{\max_{j=0,1,2}\sup_{x\in\mathcal{O}} |S_j(x)| > B_5\right\} = O\big(n^{-B_6}\big)\,, \tag{A.8}$$

$$P\left[\min_{j=0,1,2}\inf_{x\in\mathcal{O}} \left\{S_0(x)\,S_2(x) - S_1(x)^2\right\} \le B_5^{-1}\right] = O\big(n^{-B_6}\big)\,. \tag{A.9}$$

Combining (A.7)–(A.9) we deduce that, for all $B_6 > 0$ and some $B_7 > 0$,

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} |A_i(x)| > B_5^2\right\} = O\big(n^{-B_7}\big)\,. \tag{A.10}$$

Hence, by (A.6), for all $B_6 > 0$,

$$P\left\{\sup_{x\in\mathcal{O}}\frac{1}{n}\sum_{i=1}^{n} |A_i(x)|\,|R(x,X_i)| > B_4\,B_5^2\,h^{B_3}\right\} = O\big(n^{-B_6}\big)\,. \tag{A.11}$$

More simply, since $A_i(x) = 0$ unless $|x - X_i| \le 2B_1 h$ then, for all $B_6 > 0$,

$$\sup_{x \in \mathcal{O}} \frac{1}{n} \sum_{i=1}^{n} |A_i(x)|\, |g''(X_i) - g''(x)| \le B_2\, (2B_1 h)^{B_3} \sup_{x \in \mathcal{O}} \frac{1}{n} \sum_{i=1}^{n} |A_i(x)|\,,$$

and so by (A.10), for all $B_6 > 0$,

$$P\biggl\{ \sup_{x \in \mathcal{O}} \frac{1}{n} \sum_{i=1}^{n} |A_i(x)|\, |g''(X_i) - g''(x)| > B_2\, (2B_1 h)^{B_3}\, B_5^2 \biggr\} = O\bigl(n^{-B_6}\bigr)\,. \quad \text{(A.12)}$$

Combining (A.11) and (A.12), and noting the definition of $R(x)$ at (A.5), we deduce that, for all $B_6 > 0$ and some $B_7 > 0$,

$$P\biggl\{ \sup_{x \in \mathcal{O}} |R(x)| > B_7\, h^{2+B_3} \biggr\} = O\bigl(n^{-B_6}\bigr)\,. \quad \text{(A.13)}$$

Step 3. Expansion of $e_1(x)$. Recall that $e_1(x)$ was defined at (A.1); our expansion of $e_1(x)$ is given at (A.23) below, and the terms R_1 and R_2 in (A.23) satisfy (A.22) and (A.25), respectively. The expansion is designed to replace h, the bandwidth in (4.2)(g), which potentially depends on the errors ϵ_i as well as on the design variables X_i, by a deterministic bandwidth h_1. If h were a function of the design sequence alone then this step would not be necessary.

Define $h_1 = C_1\, n^{-1/(r+4k)} = C_1\, n^{-1/5}$ where C_1 is as in (4.2)(g), put $\delta_1 = (h_1 - h)/h_1$, and note that, if $|\delta_1| \le \frac{1}{2}$,

$$\frac{h_1}{h} = (1 - \delta_1)^{-1} = 1 + \delta_1 + \tfrac{1}{2}\, \delta_1^2 + \tfrac{1}{3}\, \delta_1^3 + \ldots\,.$$

Therefore, if $\ell \ge 1$ is an integer, and if K has $\ell + 1$ uniformly bounded derivatives, then there exist constants $B_8, B_9 > 0$ such that, when $|\delta_1| \le \frac{1}{2}$,

$$\biggl| K\Bigl(\frac{u}{h}\Bigr) - \biggl\{ K\Bigl(\frac{u}{h_1}\Bigr) + \sum_{j_1} \sum_{j_2} c(j_1, j_2)\, \delta_1^{j_1 + j_2} \Bigl(\frac{u}{h_1}\Bigr)^{j_2} K^{(j_2)}\Bigl(\frac{u}{h_1}\Bigr) \biggr\} \biggr| \le B_8\, |\delta_1|^{\ell+1}\, I(|u/h_1| \le B_9)\,, \quad \text{(A.14)}$$

where $c(j_1, j_2)$ denotes a constant and the double summation is over j_1 and j_2 such that $j_1 \ge 0$, $j_2 \ge 1$ and $j_1 + j_2 \le \ell$. (This range of summation is assumed also in the double summations in (A.16) and (A.20) below.) The constant B_9 is chosen so that $K(u)$ and its derivatives vanish for $|u| > B_9$. More simply,

$$\frac{u}{h} = \frac{u}{h_1} \Bigl(1 + \delta_1 + \tfrac{1}{2}\, \delta_1^2 + \tfrac{1}{3}\, \delta_1^3 + \ldots\Bigr)\,. \quad \text{(A.15)}$$

Recall that $S_k(x) = n^{-1} \sum_i \{(x - X_i)/h\}^k K\{(x - X_i)/h\}$. Write this as $S_k(h, x)$, to indicate the dependence on h, and define

$$T_{kj}(x) = \frac{1}{nh_1} \sum_{i=1}^{n} \Big(\frac{x - X_i}{h_1}\Big)^{k+j} K^{(j)}\Big(\frac{x - X_i}{h_1}\Big) .$$

Results (A.14) and (A.15) imply that, for constants $c_k(j_1, j_2)$, and provided $|\delta_1| \le \frac{1}{2}$,

$$\Big| S_k(h, x) - \Big\{ S_k(h_1, x) + \sum_{j_1} \sum_{j_2} c_k(j_1, j_2)\, \delta_1^{j_1+j_2}\, T_{kj_2}(x) \Big\} \Big| \le B_{10}\, \frac{|\delta_1|^{\ell+1}}{nh_1} \sum_{i=1}^{n} I\Big(\Big| \frac{x - X_i}{h_1} \Big| \le B_9 \Big) . \quad \text{(A.16)}$$

The methods leading to (A.52), in the proof of Lemma A.1, can be used to show that there exists an open set $\mathcal{O}$, containing $\mathcal{R}$, such that for all $B_{11}, B_{12} > 0$, and each j and k,

$$P\Big\{ \sup_{x \in \mathcal{O}} \big| (1 - E)\, T_{kj}(x) \big| > (nh_1)^{-1/2}\, n^{B_{11}} \Big\} = O\big(n^{-B_{12}}\big) . \quad \text{(A.17)}$$

Additionally, the argument leading to (A.53) can be used to prove that

$$\sup_{x \in \mathcal{O}} |E\{T_{kj}(x)\} - \ell_{kj}(x)| \to 0 , \quad \text{(A.18)}$$

where $\ell_{kj}(x) = f_X(x) \int u^{k+j}\, K^{(j)}(u)\, du$.

The definition of $e_1(x)$, at (A.1), can be written equivalently as

$$\begin{aligned} e_1(x) = \frac{1}{(nh)^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} & \frac{S_2(h, x) - \{(x - X_{i_1})/h\}\, S_1(h, x)}{S_0(h, x)\, S_2(h, x) - S_1(h, x)^2} \\ & \times \frac{S_2(h, X_{i_1}) - \{(X_{i_1} - X_{i_2})/h\}\, S_1(h, X_{i_1})}{S_0(h, X_{i_1})\, S_2(h, X_{i_1}) - S_1(h, X_{i_1})^2} \\ & \times K\Big(\frac{x - X_{i_1}}{h}\Big)\, K\Big(\frac{X_{i_1} - X_{i_2}}{h}\Big)\, \epsilon_{i_2} . \end{aligned} \quad \text{(A.19)}$$

Write $A_i(h_1, x)$ for the version of $A_i(x)$, at (2.13), that would be obtained if h were replaced by h_1 in that formula, and in particular in the definitions of S_0, S_1, S_2 and K_i. Using (A.14) and (A.16) to substitute for $K(u/h)$ and $S_k(h, x)$, respectively, where $u = x - X_{i_1}$ or $X_{i_1} - X_{i_2}$ in the case of $K(u/h)$; and then Taylor expanding; it can be proved from (A.19), noting the properties at (A.16)–(A.18), that, provided $|\delta_1| \le \frac{1}{2}$,

$$\begin{aligned} e_1(x) = \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} & A_{i_1}(h_1, x)\, A_{i_2}(h_1, X_{i_1})\, \epsilon_{i_2} \\ & + \sum_{j_1} \sum_{j_2} c_1(j_1, j_2)\, \delta_1^{j_1+j_2}\, D_{j_1 j_2}(x) + |\delta_1|^{\ell+1}\, R_1(x) , \end{aligned} \quad \text{(A.20)}$$

where the constants $c_1(j_1, j_2)$ do not depend on h_1 or n and are uniformly bounded; each term $D_{j_1 j_2}(x)$ can be represented as

$$\frac{1}{nh} \sum_{i=1}^{n} J_i\Big(\frac{x - X_i}{h_1}\Big)\, \epsilon_i\,, \tag{A.21}$$

where the functions J_i depend on j_1, j_2 and x, on the design sequence $\mathcal{X}$ and the bandwidth h_1, but not on the errors or on h, and, for some B_{14} and B_{15}, and all B_{16}, satisfy

$$P\Big\{ \sup_{x\in\mathcal{O}} \sup_{|u|\le B_{13}} \max_{1\le i\le n} |J_i(u)| > B_{14} \Big\} = O\big(n^{-B_{16}}\big)\,,$$

$$P\Big\{ \sup_{x\in\mathcal{O}} \sup_{|u|> B_{14}} \max_{1\le i\le n} |J_i(u)| = 0 \Big\} = O\big(n^{-B_{16}}\big)\,,$$

$$P\Big\{ \sup_{x\in\mathcal{O}} \sup_{u_1,u_2\in\mathbb{R}} \max_{1\le i\le n} |J_i(u_1) - J_i(u_2)| \le B_{15}\,|u_1 - u_2| \Big\} = O\big(n^{-B_{16}}\big)\,;$$

and, also for some B_{14} and all B_{16},

$$P\Big\{ \sup_{x\in\mathcal{O}} |R_1(x)| > B_{14} \Big\} = O\big(n^{-B_{16}}\big)\,. \tag{A.22}$$

Combining the results from (A.20) down, and using Markov's inequality and lattice arguments to bound the quantity at (A.21), we deduce that, if $|\delta_1| \le \frac{1}{2}$,

$$e_1(x) = T(x) + \delta_1\, R_2(x) + |\delta_1|^{\ell+1}\, R_1(x)\,, \tag{A.23}$$

where

$$T(x) = \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} A_{i_1}(h_1, x)\, A_{i_2}(h_1, X_{i_1})\, \epsilon_{i_2}\,, \tag{A.24}$$

R_1 satisfies (A.22) and R_2 satisfies

$$P\Big\{ \sup_{x\in\mathcal{O}} |R_2(x)| > (nh_1)^{-1/2}\, n^{B_{17}} \Big\} = O\big(n^{-B_{18}}\big)\,. \tag{A.25}$$

In (A.25), for each fixed $B_{17} > 0$, B_{18} can be taken arbitrarily large, provided that $E|\epsilon|^{B_{19}} < \infty$ for sufficiently large B_{19}.

Step 4. Approximation to $T(x)$, defined at (A.24). The approximation is given by (A.33), with the remainder there controlled by (A.34).

Define

$$D_i(x) = \frac{1}{n} \sum_{i_1=1}^{n} A_{i_1}(h_1, x)\, A_i(h_1, X_{i_1})\,.$$

Result (A.53), derived during the proof of Lemma A.1, and assumption (4.2)(f) on f_X, imply that for a sequence of constants $\eta = \eta(n)$ decreasing to 0 at a polynomial rate as $n \to \infty$, and for all $B_{16} > 0$,

$$P\Big[\sup_{x\in\mathcal{O}} \max_{1\le i\le n} \Big\{\Big|h\,A_i(h_1,x)\,K\Big(\frac{x-X_i}{h_1}\Big)^{-1} - \frac{1}{f_X(x)}\Big|\Big\} > \eta\Big] = O\big(n^{-B_{16}}\big)\,. \quad \text{(A.26)}$$

Define

$$D(x_1,x_2) = \frac{1}{nh_1}\sum_{i=1}^{n} \frac{1}{f_X(X_i)}\,K\Big(\frac{x_1-X_i}{h_1}\Big)\,K\Big(\frac{x_2-X_i}{h_1}\Big)\,,$$

It follows from (A.26) and the compact support of K (see (4.2)(h)) that

$$D_i(x) = \frac{1+\Delta_i(x)}{h_1\,f_X(x)}\,D(x,X_i)\,, \quad \text{(A.27)}$$

where the random functions Δ_i are measurable in the sigma-field generated by $\mathcal{X}$ (we refer to this below as "$\mathcal{X}$ measurable") and satisfy, for some $B_{20} > 0$ and all $B_{16} > 0$,

$$P\Big\{\sup_{x\in\mathcal{O}} \max_{1\le i\le n} |\Delta_i(x)| > n^{-B_{20}}\Big\} = O\big(n^{-B_{16}}\big)\,. \quad \text{(A.28)}$$

Recall that $L = K * K$, and note that $E\{D(x_1,x_2)\} = L\{(x_1-x_2)h_1\}$, and that, since K is compactly supported, there exists $B_{21} > 0$ such that $D(x_1,x_2) = 0$ whenever $x_1 \in \mathcal{R}$ and $|x_1 - x_2| > B_{21}$. Furthermore, there exist B_{22} and $B_{23}(p)$, the latter for each choice of the integer $p \ge 1$, such that whenever $x_1 \in \mathcal{R}$,

$$\mathrm{var}\{D(x_1,x_2)\} \le \frac{B_{22}}{nh_1}\,I\Big(\Big|\frac{x_1-x_2}{h_1}\Big| \le B_{21}\Big)\,,$$

$$E\Big\{\frac{1}{f_X(X_i)}\,K\Big(\frac{x_1-X_i}{h_1}\Big)\,K\Big(\frac{x_2-X_i}{h_1}\Big)\Big\}^{2p} \le B_{23}(p)\,h_1\,I\Big(\Big|\frac{x_1-x_2}{h_1}\Big| \le B_{21}\Big)\,.$$

Hence, by Rosenthal's inequality, whenever $x_1 \in \mathcal{R}$ and $x_2 \in \mathbb{R}$,

$$E\big|(1-E)\,D(x_1,x_2)\big|^{2p} \le \frac{B_{24}(p)}{(nh_1)^p}\,I\Big(\Big|\frac{x_1-x_2}{h_1}\Big| \le B_{21}\Big)\,.$$

Therefore, by Markov's inequality, for each $B_{25}, B_{26} > 0$,

$$\sup_{x_1\in\mathcal{R},\,x_2\in\mathbb{R}} P\Big\{\big|(1-E)\,D(x_1,x_2)\big| > (nh_1)^{-1/2}\,n^{B_{25}}\Big\} = O\big(n^{-B_{26}}\big)\,.$$

Approximating to $(1-E)\,D(x_1,x_2)$ on a polynomially fine lattice of pairs (x_1,x_2), with $x_1 \in \mathcal{R}$ and $|x_1-x_2| \le B_{21}$, we deduce that the supremum here can be placed inside the probability statement: for each $B_{25}, B_{26} > 0$,

$$P\Big\{\sup_{x_1\in\mathcal{R},\,x_2\in\mathbb{R}} \big|(1-E)\,D(x_1,x_2)\big| > (nh_1)^{-1/2}\,n^{B_{25}}\Big\} = O\big(n^{-B_{26}}\big)\,. \quad \text{(A.29)}$$

Combining (A.27)–(A.29) we deduce that

$$D_i(x) = \frac{1+\Delta_i(x)}{h_1\, f_X(x)}\, L\Big(\frac{x-X_i}{h_1}\Big) + \frac{\Theta_i(x)}{h_1\,(nh_1)^{1/2}}\, I\Big(\Big|\frac{x-X_i}{h_1}\Big| \le B_{21}\Big)\,, \tag{A.30}$$

where the function Θ_i is $\mathcal{X}$-measurable and satisfies, for each B_{25}, $B_{26} > 0$,

$$P\Big\{ \sup_{x\in\mathcal{R}} \max_{i\,:\,|x-X_i|\le B_{21}h_1} |\Theta_i(x)| > n^{B_{25}} \Big\} = O\big(n^{-B_{16}}\big)\,. \tag{A.31}$$

By (A.24) and (A.30),

$$T(x) = \frac{1}{n}\sum_{i=1}^n D_i(x)\,\epsilon_i = \frac{1}{nh_1\, f_X(x)}\sum_{i=1}^n L\Big(\frac{x-X_i}{h_1}\Big)\,\epsilon_i + T_1(x) + T_2(x)\,, \tag{A.32}$$

where

$$T_1(x) = \frac{1}{nh_1\, f_X(x)}\sum_{i=1}^n \Delta_i(x)\, L\Big(\frac{x-X_i}{h_1}\Big)\,\epsilon_i\,,$$
$$T_2(x) = \frac{1}{(nh_1)^{3/2}\, f_X(x)}\sum_{i=1}^n \Theta_i(x)\, I\Big(\Big|\frac{x-X_i}{h_1}\Big| \le B_{21}\Big)\,\epsilon_i\,.$$

Using he fact that the functions Δ_i and Θ_i are $\mathcal{X}$-measurable, as well as (A.28), (A.31), and approximations on polynomially fine lattices, it can be proved that if B_{27} (large) and B_{28} (small) are given then, provided $E|\epsilon|^{B_{29}} < \infty$ where B_{29} depends on B_{27} and B_{28}, we have for some $B_{30} > 0$,

$$P\Big\{ \sup_{x\in\mathcal{R}} |T_1(x)| > (nh_1)^{-1/2}\, n^{-B_{30}} \Big\} = O\big(n^{-B_{27}}\big)\,,$$
$$P\Big\{ \sup_{x\in\mathcal{R}} |T_2(x)| > (nh_1)^{-1}\, n^{B_{28}} \Big\} = O\big(n^{-B_{27}}\big)\,.$$

Therefore, by (A.32),

$$T(x) = \frac{1}{n}\sum_{i=1}^n D_i(x)\,\epsilon_i = \frac{1}{nh_1\, f_X(x)}\sum_{i=1}^n L\Big(\frac{x-X_i}{h_1}\Big)\,\epsilon_i + R_3(x)\,, \tag{A.33}$$

where the following property holds for $j = 3$: for some $B_{31} > 0$, if $B_{32} > 0$ is given then, provided $E|\epsilon|^{B_{33}} < \infty$, with B_{33} depending on B_{32},

$$P\Big\{ \sup_{x\in\mathcal{R}} |R_j(x)| > (nh_1)^{-1/2}\, n^{-B_{31}} \Big\} = O\big(n^{-B_{32}}\big)\,. \tag{A.34}$$

Step 5. Approximation to $e_2(x)$, defined at (A.5). Property (A.53), and arguments similar to those in step 5, permit us to show that

$$e_2(x) = \frac{1}{nh_1\, f_X(x)}\sum_{i=1}^n L\Big(\frac{x-X_i}{h_1}\Big)\,\epsilon_i + R_4(x)\,, \tag{A.35}$$

where (A.34) holds for $j = 4$.

Step 6. Gaussian approximation to $e_1(x) - e_2(x)$. The approximation is given at (A.42). Recall that $M = L - K = K * K - K$. By (A.23), (A.25) and (A.33)–(A.35),

$$e_1(x) - e_2(x) = \frac{1}{nh_1\, f_X(x)} \sum_{i=1}^{n} M\Big(\frac{x - X_i}{h_1}\Big)\, \epsilon_i + |\delta_1|^{\ell+1}\, R_1(x) + R_5(x)\,, \quad \text{(A.36)}$$

where R_1 and R_5 satisfy (A.22) and (A.34), respectively.

Next we use an approximation due to Kómlós, Major and Tusnády (1976). Theorem 4 there implies that if B_{34} (small) and B_{35} (large) are given then there exists $B_{36} > 0$, depending on B_{34} and B_{35}, such that, if $E|\epsilon|^{B_{36}} < \infty$, then it is possible to construct a sequence of Normal random variables $Z_1, Z_2, \ldots$ with $E(Z_i) = E(\epsilon_i) = 0$ and $E(Z_i^2) = E(\epsilon_i)^2 = \sigma^2$, and for which

$$P\Big\{ \max_{1 \le i \le n} \Big| \sum_{i_1=1}^{i} (\epsilon_{i_1} - Z_{i_1}) \Big| > n^{B_{34}} \Big\} = O\big(n^{-B_{35}}\big)\,. \quad \text{(A.37)}$$

Define $M_i(x) = M\{(x - X_i)/h_1\}$ for $1 \le i \le n$, $M_{n+1} = 0$, $V_i = \sum_{1 \le i_1 \le i} \epsilon_{i_1}$ and $N_i = \sum_{1 \le i_1 \le i} Z_{i_1}$, and note that, using Euler's method of summation,

$$\sum_{i=1}^{n} M\Big(\frac{x - X_i}{h_1}\Big)\, \epsilon_i = \sum_{i=1}^{n} M_i(x)\, \epsilon_i = \sum_{i=1}^{n} \{M_i(x) - M_{i+1}(x)\}\, V_i\,.$$

Therefore,

$$\sum_{i=1}^{n} M\Big(\frac{x - X_i}{h_1}\Big)\, (\epsilon_i - Z_i) = \sum_{i=1}^{n} \{M_i(x) - M_{i+1}(x)\}\, (V_i - N_i)\,. \quad \text{(A.38)}$$

Let $\mathcal{T} = \mathcal{T}(h_1)$ denote the set of all points $x_1 \in \mathbb{R}$ such that $(x - x_1)/h_1$ lies within the support of K for some $x \in \mathcal{R}$. Then $\mathcal{T}$ depends on n, and, for $n \ge B_{37}$ say, is a subset of the open set $\mathcal{O}$ introduced in step 9. Hence, if $x \in \mathcal{T}$ and $n \ge B_{37}$ then $f_X(x) > B_{38}$, where $B_{38} > 0$ is a lower bound for f_X on the open set referred to in (4.2)(f). Let ν denote the number of X_is, for $1 \le i \le n$, that lie in $\mathcal{T}$. Order the X_is so that these X_is are listed first in the sequence $X_1, \ldots, X_n$, and moreover, such that $X_1 \le \ldots \le X_\nu$. Let $X_{\nu+1}$ be the X_i that is nearest to X_ν and is not one of $X_1, \ldots, X_\nu$. Using properties of spacings of order statistics from a distribution the density of which is bounded away from zero, we deduce that if $B_{39} < 1$ then, for all $B_{40} > 0$,

$$P\Big(\max_{1 \le i \le \nu} |X_i - X_{i+1}| > n^{-B_{39}} \Big) = O\big(n^{-B_{40}}\big)\,. \quad \text{(A.39)}$$

If $1 \le i \le \nu$ and $x \in \mathcal{R}$ then

$$|M_i(x) - M_{i+1}(x)| = \Big|M\Big(\frac{x - X_i}{h_1}\Big) - M\Big(\frac{x - X_i}{h_1} + \frac{X_i - X_{i+1}}{h_1}\Big)\Big| \\ \le h_1^{-1} \, (\sup |M'|) \, |X_i - X_{i+1}| \le h_1^{-1} \, (\sup |M'|) \, n^{-B_{39}} \,,$$

where the identity and the first inequality hold with probability 1, and, by (A.39), the second inequality holds with probability $1 - O(n^{-B_{40}})$. If $n \ge B_{37}$ and $x \in \mathcal{R}$ then all the indices i for which $M_i(x) - M_{i+1}(x) \ne 0$ are in the range from 1 to ν, and therefore the second series on the right-hand side of (A.38) can be restricted to a sum from $i = 1, \ldots, \nu$. Combining the results in this paragraph we deduce that for all $B_{40} > 0$,

$$P\Big\{ \max_{1 \le i \le n} |M_i(x) - M_{i+1}(x)| \le h_1^{-1} \, (\sup |M'|) \, n^{-B_{39}} \Big\} = 1 - O\big(n^{-B_{40}}\big) \,. \quad \text{(A.40)}$$

In multivariate cases, where the number of dimensions, r, satisfies $r \ge 2$, the spacings argument above should be modified by producing an ordering $X_1, \ldots, X_n$ of the X_is which is such that $\|X_i - X_{i+1}\|$ is small, for $1 \le i \le n-1$, where $\|\cdot\|$ is the Euclidean metric. We do this by taking $B < 1/r$ and, first of all, constructing a regular, rectangular lattice within $\mathcal{R}$ where the total number of cells, or lattice blocks, is bounded above and below by constant multiples of $n^{B+\delta}$ for a given $\delta \in (0, \frac{1}{3}\,(r^{-1} - B))$. (The sizes of the faces of the cells are in proportion to the sizes of the faces of $\mathcal{R}$.) We order the points X_i within each given cell so that $\|X_i - X_{i+1}\| \le n^B$ is small. (With probability converging to 1 at a polynomial rate, this can be done simultaneously for each of the cells.) Then we choose one representative point X_i in each cell (it could be the point nearest to the cell's centre), and draw a path linking that point in one cell to its counterpart in an adjacent cell, such that those linked points are no further than $n^{B+2\delta}$ apart, and each cell is included in the chain after just $n-1$ links have been drawn. Again this can be achieved with probability converging to 1 at a polynomial rate. Once the linkage has been put in place, the n design points can be reordered so that $\|X_i - X_{i+1}\| \le n^{B+3\delta}$ for $1 \le i \le n-1$. By taking $B > r$, but very close to r, and then choosing $\delta > 0$ but very close to 0, we see that, for any given $B' > r$, we can, with probability converging to 1 at a polynomial rate, construct an ordering $X_1, \ldots, X_n$ so that $\|X_i - X_{i+1}\| \le n^{B'}$ for $1 \le i \le n-1$.

Result (A.38) implies that, if $x \in \mathcal{R}$,

$$\Big| \frac{1}{n h_1 \, f_X(x)} \sum_{i=1}^n M\Big(\frac{x - X_i}{h_1}\Big) \, (\epsilon_i - Z_i) \Big|$$

$$\leq \frac{1}{nh_1 B_{38}} \left\{ \max_{1\leq i\leq n} |M_i(x) - M_{i+1}(x)| \right\} \left\{ \max_{1\leq i\leq n} \left| \sum_{i_1=1}^{i} (\epsilon_{i_1} - Z_{i_1}) \right| \right\}. \quad \text{(A.41)}$$

Combining (A.37), (A.40) and (A.41), recalling from step 3 that $h_1 = C_1\, n^{-1/5}$, and taking $B_{35} > 1$ in (A.37), we conclude that for all $B_{40} > 0$,

$$P\Bigg\{ \sup_{x\in\mathcal{R}} \left| \frac{1}{nh_1\, f_X(x)} \sum_{i=1}^{n} M\Big(\frac{x - X_i}{h_1}\Big) (\epsilon_i - Z_i) \right| \leq \frac{\sup |M'|}{B_{38}\, C_1^2\, n^{(3/5)+B_{39}-B_{34}}} \Bigg\} = 1 - O\big(n^{-B_{40}}\big)\,.$$

Hence, by (A.36),

$$e_1(x) - e_2(x) = \zeta(x) + |\delta_1|^{\ell+1}\, R_1(x) + R_5(x) + R_6(x)\,, \quad \text{(A.42)}$$

where

$$\zeta(x) = \frac{1}{nh_1\, f_X(x)} \sum_{i=1}^{n} M\Big(\frac{x - X_i}{h_1}\Big)\, Z_i\,, \quad \text{(A.43)}$$

R_1 and R_5 satisfy (A.22) and (A.34), respectively, and, for some $B_{41} > 0$ and all $B_{40} > 0$,

$$P\Big\{ \sup_{x\in\mathcal{R}} |R_6(x)| \leq B_{41}\, n^{-(3/5)-B_{39}+B_{34}} \Big\} = 1 - O\big(n^{-B_{40}}\big)\,. \quad \text{(A.44)}$$

Step 7. Approximation to $\zeta(x)$ in terms of a Gaussian process. The approximation is given at (A.45). Conditional on the design sequence $\mathcal{X}$ the process ζ, at (A.43), is itself Gaussian, with zero mean and covariance

$$\operatorname{cov}\{\zeta(x_1), \zeta(x_2) \,|\, \mathcal{X}\} = \frac{\sigma^2}{\{nh_1\, f_X(x)\}^2} \sum_{i=1}^{n} M\Big(\frac{x_1 - X_i}{h_1}\Big)\, M\Big(\frac{x_2 - X_i}{h_1}\Big)\,,$$

and standard arguments show that for some $B_{42} > 0$ and all $B_{43} > 0$,

$$P\Bigg\{ \sup_{x_1,\, x_2\in\mathcal{R}} \Big| nh_1\, f_X(x) \operatorname{cov}\{\zeta(x_1), \zeta(x_2) \,|\, \mathcal{X}\} - (M * M)\Big(\frac{x_1 - x_2}{h_1}\Big) \Big| > n^{-B_{42}} \Bigg\} = O\big(n^{-B_{43}}\big)\,.$$

Hence, for each n there exists a Gaussian stationary process W, with zero mean and covariance given by (4.4), such that for some $B_{44} > 0$ and all $B_{43} > 0$,

$$P\Big\{ (nh_1)^{1/2} \sup_{x\in\mathcal{R}} \big|\zeta(x) - f_X(x)^{-1/2}\, W(x)\big| > n^{-B_{44}} \Big\} = O\big(n^{-B_{43}}\big)\,. \quad \text{(A.45)}$$

Step 8. Completion of proof of Theorem 4.1, except for Lemma A.1. Combining (A.2) and (A.4) we deduce that

$$E\{\hat{g}^*(x) \,|\, \mathcal{Z}\} = \hat{g}(x) + \tfrac{1}{2}\, h^2\, \kappa_2\, g''(x) + e_1(x) - e_2(x) + \tfrac{1}{2}\, h^2\, R(x)\,. \tag{A.46}$$

Combining (A.42) and (A.45), using the bounds at (A.22), (A.34) and (A.44) on the remainder terms R_1, R_5 and R_6 on the right-hand side of (A.42), and noting that, in view of (4.2)(g) and the definition $\delta_1 = (h_1 - h)/h_1$, $P(|\delta_1| > n^{-C_2/(r+4k)}) \to 0$, we see that if the exponent $\ell+1$ in (A.42) can be taken sufficiently large (depending on $C_2 > 0$, and enabled by taking C_5 sufficiently large in (4.2)(h)), then for some $B_{45} > 0$,

$$P\biggl\{(nh_1)^{1/2} \sup_{x\in\mathcal{R}} \bigl|e_1(x) - e_2(x) - f_X(x)^{-1/2}\, W(x)\bigr| > n^{-B_{45}}\biggr\} \to 0\,. \tag{A.47}$$

In view of the approximation to h by $h_1 = C_1\, n^{-1/5}$ asserted in (4.2)(g), (A.13) implies that

$$P\biggl\{\sup_{x\in\mathcal{R}} |R(x)| > B_7\, h_1^2\, n^{-B_{46}}\biggr\} \to 0\,. \tag{A.48}$$

Result (4.5) follows on combining (A.46)–(A.48).

Step 9. Derivation of (A.8) *and* (A.9).

Lemma A.1. *If* (4.2) *holds then there exists an open set* $\mathcal{O}$*, containing* $\mathcal{R}$*, such that* (A.8) *and* (A.9) *obtain.*

To derive the lemma, recall that

$$S_k(x) = \frac{1}{nh} \sum_{i=1}^n \Bigl(\frac{x - X_i}{h}\Bigr)^k K\Bigl(\frac{x - X_i}{h}\Bigr)\,.$$

Let $\mathcal{H} = [n^{-B_{49}}, n^{-B_{48}}]$, where $0 < B_{48} < B_{49} < 1$. As noted in (4.2), the bandwidth h is a function of the data in $\mathcal{Z}$, but initially we take h to be deterministic, denoting it by h_2 and denoting the corresponding value of $S_k(x)$ by $S_k(h_2, x)$. Then by standard calculations, for each integer $j \ge 1$, and for $k = 0, 1, 2$,

$$\sup_{h_2\in\mathcal{H}} \sup_{x\in\mathbb{R}} (nh_2)^j\, E\bigl[\{(1-E)\, S_k(h_2, x)\}^{2j}\bigr] \le B(j)\,, \tag{A.49}$$

where $B(j)$ does not depend on n. Here we have used the uniform boundedness of f_X, asserted in (4.2)(f). Result (A.49), and Markov's inequality, imply that for all $B_{50}, B_{51} > 0$, and for $k = 0, 1, 2$,

$$\sup_{h_2\in\mathcal{H}} \sup_{x\in\mathbb{R}} P\Bigl[\bigl|(1-E)\, S_k(h_2, x)\bigr| > (nh_2)^{-1/2}\, n^{B_{50}}\Bigr] = O\bigl(n^{-B_{51}}\bigr)\,. \tag{A.50}$$

Let $B_{52} > 0$. It follows from (A.50) that if $\mathcal{S}$, contained in the open set referred to in (4.2)(f), is a compact subset of $\mathbb{R}$, if $\mathcal{S}(n)$ is any subset of $\mathcal{S}$ such that $\#\mathcal{S}(n) = O(n^{B_{52}})$, and if $\mathcal{H}(n)$ is any subset of $\mathcal{H}$ such that $\#\mathcal{H}(n) = O(n^{B_{52}})$, then for all $B_{50}, B_{51} > 0$, and for $k = 0, 1, 2$,

$$P\Bigg[\sup_{h_2 \in \mathcal{H}(n)} \sup_{x \in \mathcal{S}(n)} \big|(1-E)\, S_k(h_2, x)\big| > (nh_2)^{-1/2}\, n^{B_{50}}\Bigg] = O\big(n^{-B_{51}}\big)\,. \tag{A.51}$$

Approximating to $S_k(h_2, x)$ on a polynomially fine lattice of values of h_2 and x, we deduce from (A.51) that, for all $B_{50}, B_{51} > 0$, and for $k = 0, 1, 2$,

$$P\Bigg[\sup_{h_2 \in \mathcal{H}} \sup_{x \in \mathcal{S}} \big|(1-E)\, S_k(h_2, x)\big| > (nh_2)^{-1/2}\, n^{B_{50}}\Bigg] = O\big(n^{-B_{51}}\big)\,. \tag{A.52}$$

Choose $\mathcal{S}$ sufficiently large to contain an open set, $\mathcal{O}$, which contains $\mathcal{R}$ and has the property that, for some $\delta > 0$, the set of all closed balls of radius δ and centred at a point in $\mathcal{O}$ is contained in $\mathcal{S}$. Since K is compactly supported and $h_2 \in \mathcal{H}$ satisfies $h_2 \le n^{-B_{48}}$, it can be proved from (4.2)(f) that from some $B_A > 0$,

$$\sup_{h_2 \in \mathcal{H}} \sup_{x \in \mathcal{O}} \Big[|E\{S_0(x)\} - f_X(x)| + |E\{S_1(x)\}| + |E\{S_2(x)\} - \kappa_2\, f_X(x)|\Big] = O\big(n^{-B_A}\big)\,.$$

Therefore, defining $\ell_k(x) = f_X(x), 0, \kappa_2\, f_X(x)$ according as $k = 0, 1, 2$, respectively, we deduce from (A.52) that for a sequence $\eta = \eta(n)$ decreasing to 0 at a polynomial rate in n as $n \to \infty$, for $k = 0, 1, 2$, and for all $B_{51} > 0$,

$$P\Bigg[\sup_{h_2 \in \mathcal{H}} \sup_{x \in \mathcal{O}} \big|S_k(h_2, x) - \ell_k(x)\big| > \eta\Bigg] = O\big(n^{-B_{51}}\big)\,. \tag{A.53}$$

Results (A.8) and (A.9) follow from (A.53) on noting the properties on h in (4.2)(g); we can take C_3 and C_4 there to be B_{48} and B_{49} above.

APPENDIX B: OUTLINE PROOFS OF COROLLARIES 4.1 AND 4.2

B.1. Proof of Corollary 4.1. Define

$$\hat{d}^*(x) = \frac{\hat{g}(x) - E\{\hat{g}^*(x) \,|\, \mathcal{Z}\}}{\{D_1\, \hat{\sigma}^{*2}\, \hat{f}_X(x)^{-1}\, (nh^r)^{-1}\}^{1/2}}\,, \quad \hat{d}(x) = \frac{\hat{g}(x) - E\{\hat{g}^*(x) \,|\, \mathcal{Z}\}}{\{D_1\, \sigma^2\, f_X(x)^{-1}\, (nh_1^r)^{-1}\}^{1/2}}\,.$$

Recall that, motivated by the variance formula (4.1), we take $s(\mathcal{X})(x)^2\, \hat{\sigma}^2$, in the definition of the confidence band $\mathcal{B}(\alpha)$ at (2.2), to be $D_1\, \hat{\sigma}^2\, \hat{f}_X(x)^{-1}\, (nh^r)^{-1}$. The

bootstrap estimator $\hat\pi(x,\alpha)$, defined at (2.17), of the probability $\pi(x,\alpha)$, at (2.3), that the band $\mathcal{B}(\alpha)$ covers the the point $(x, g(x))$, is given by

$$\begin{aligned}\hat\pi(x,\alpha) &= P\Big\{\hat g^*(x) - s(\mathcal{X})(x)\,\hat\sigma^*\,z_{1-(\alpha/2)} \le \hat g(x) \le \hat g^*(x) + s(\mathcal{X})(x)\,\hat\sigma^*\,z_{1-(\alpha/2)} \;\Big|\; \mathcal{Z}\Big\} \\ &= P\left[- z_{1-(\alpha/2)} \le \frac{\hat g^*(x) - \hat g(x)}{\big\{D_1\,\hat\sigma^{*2}\,\hat f_X(x)^{-1}\,(nh^r)^{-1}\big\}^{1/2}} \le z_{1-(\alpha/2)} \;\middle|\; \mathcal{Z}\right] \\ &= P\left[- z_{1-(\alpha/2)} + \hat d^*(x) \le \frac{\hat g^*(x) - E\{\hat g^*(x)\,|\,\mathcal{Z}\}}{\big\{D_1\,\hat\sigma^{*2}\,\hat f_X(x)^{-1}\,(nh^r)^{-1}\big\}^{1/2}}\right. \\ &\qquad\qquad \left. \le z_{1-(\alpha/2)} + \hat d^*(x) \;\middle|\; \mathcal{Z}\right]. \qquad \text{(B.1)}\end{aligned}$$

If both (4.2) and (4.3) hold then, by (4.5), (4.6), (B.1) and minor additional calculations,

$$\begin{aligned}P\bigg(\sup_{x\in\mathcal{R}} \Big|\hat\pi(x,\alpha) - \Big[&\Phi\big\{z_{1-(\alpha/2)} + \hat d(x)\big\} \\ &- \Phi\big\{ - z_{1-(\alpha/2)} + \hat d(x)\big\}\Big]\Big| > n^{-C_9}\bigg) \to 0\,. \qquad \text{(B.2)}\end{aligned}$$

Now, $-\hat d(x) = D_3\,f_X(x)^{1/2}\,\nabla\,g(x) + D_4\,W(x/h_1)$ where $D_3 = D_1^{-1/2}\,\sigma^{-1}$ and $D_4 = D_2\,D_3$, and so (4.8) follows from (B.2).

B.2. Proof of Corollary 4.2. Result (4.9) follows from (4.8). Shortly we shall outline a proof of (4.10); at present we use (4.10) to derive (4.11). To this end, recall that $\beta = \alpha_\xi(\alpha_0)$ solves equation (2.11) when $z = z_{1-(\beta/2)}$, and $\beta = \alpha(x,\alpha_0) > 0$ denotes the solution of equation (2.10). If (4.10) holds then (4.11) will follow if we establish that result when $\hat\alpha_\xi(\alpha_0)$, in the quantity $P[(x,g(x)) \in \mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}]$ appearing in (4.11), is replaced by $\alpha_\xi(\alpha_0)$. Call this property (P). Now, the definition of $\alpha_\xi(\alpha_0)$, and the following monotonicity property,

$$\begin{aligned}&\Phi(z+b) - \Phi(-z+b) \text{ is a decreasing (respectively, increasing) function of } b \\ &\text{for } b > 0 \text{ (respectively, } b < 0) \text{ and for each } z > 0,\end{aligned} \qquad \text{(B.3)}$$

ensure that

$$\liminf_{n\to\infty}\; P[(x,g(x)) \in \mathcal{B}\{\alpha_\xi(\alpha_0)\}] \ge 1 - \alpha_0$$

whenever $\alpha(x,\alpha_0) \le \alpha_\xi(\alpha_0)$, or equivalently, whenever $x \in \mathcal{R}_\xi(\alpha_0)$. This establishes (P).

Finally we derive (4.10), for which purpose we construct a grid of edge width δ, where δ is sufficiently small (see (B.4) below), and show that if this grid is used

to define $\hat{\alpha}_\xi(\alpha_0)$ (see (2.19)) then (4.10) holds. Let $x'_1, \ldots, x'_{N_1}$ be the centres of the cells, in a regular rectangular grid in $\mathbb{R}^r$ with edge width δ_1, that are contained within $\mathcal{R}$. (For simplicity we neglect here cells that overlap the boundaries of $\mathcal{R}$; these have negligible impact.) Within each cell that intersects $\mathcal{R}$, construct the smaller cells (referred to below as subcells) of a subgrid with edge width $\delta = m^{-1}\delta_1$, where $m = m(\delta_1) \geq 1$ is an integer and $m \sim \delta_1^{-c}$ for some $c > 0$. Put $N = m^r\, N_1$; let $x_{j\ell}$, for $j = 1, \ldots, N_1$ and $\ell = 1, \ldots, m^r$, denote the centres of the subcells that are within the cell that has centre x'_j; and let $x_1, \ldots, x_N$ be an enumeration of the values of $x_{j\ell}$, with $x_{11}, \ldots, x_{1m}$ listed first, followed by $x_{21}, \ldots, x_{2m}$, and so on. Recalling the definition of $\hat{\alpha}_\xi(\alpha_0)$ at (2.19), let $\hat{\alpha}_\xi(\alpha_0, \delta)$ denote the $(1-\xi)$-level quantile of the sequence $\hat{\alpha}(x_1, \alpha_0), \ldots, \hat{\alpha}(x_N, \alpha_0)$.

Let $h_1 = C_1\, n^{-1/(r+4k)}$ represent the asymptotic size of the bandwidth asserted in (4.2)(g), and assume that

$$\delta = O\bigl(n^{-B_1}\bigr)\,, \quad 1/(r+4k) < B_1 < \infty\,. \tag{B.4}$$

Then

$$\delta = O\bigl(h_1\, n^{-B_2}\bigr) \tag{B.5}$$

for some $B_2 > 0$. In particular, δ is an order of magnitude smaller than h_1.

Recall that $A(x, \alpha_0) = 2\,[1 - \Phi\{Z(x, \alpha_0)\}] \in (0, 1)$, where $Z = Z(x, \alpha_0) > 0$ is the solution of

$$\Phi\{Z + b(x) + \Delta(x)\} - \Phi\{-Z + b(x) + \Delta(x)\} = 1 - \alpha_0\,,$$

and $\Delta(x) = -D_4\, W(x/h_1)$; and that $\beta = \alpha(x, \alpha_0) > 0$ solves $\Phi\{\beta + b(x)\} - \Phi\{-\beta + b(x)\} = 1 - \alpha_0$. Define $e(x, \alpha_0) = 2\,[1 - \Phi\{\alpha(x, \alpha_0)\}]$. Given a finite set $\mathcal{S}$ of real numbers, let $\mathrm{quant}_{1-\xi}(\mathcal{S})$ and $\mathrm{med}(\mathcal{S}) = \mathrm{quant}_{1/2}(\mathcal{S})$ denote, respectively, the $(1-\xi)$th empirical quantile and the empirical median of the elements of $\mathcal{S}$. Noting (B.3), and the fact that the stationary process W is symmetric (W is a zero-mean Gaussian process the distribution of which does not depend on n), it can be shown that $P\{Z(x, \alpha_0) > \alpha(x, \alpha_0)\} = P\{Z(x, \alpha_0) \leq \alpha(x, \alpha_0)\} = \frac{1}{2}$. Therefore the median value of the random variable $A(x, \alpha_0)$ equals $e(x, \alpha_0)$. Hence, since the lattice subcell centres $x_{j1}, \ldots, x_{jm^r}$ are clustered regularly around x_j, it is unsurprising, and can be proved using (B.5), that the median of $A(x_{j1}, \alpha_0), \ldots, A(x_{jm^r}, \alpha_0)$ is closely approximated by $e(x, \alpha_0)$, and in particular that for some $B_3 > 0$ and all $B_4 > 0$,

$$P\biggl\{ \max_{j=1,\ldots,N_1} \Bigl| \mathrm{med}\{A(x_{j1}, \alpha_0), \ldots, A(x_{jm^r}, \alpha_0)\}$$

$$- e(x_j, \alpha_0)\Big| > n^{-B_3} \Big\} = O\big(n^{-B_4}\big)\,.$$

Therefore, since the $(1-\xi)$-level quantile of the points in the set

$$\bigcup_{j=1}^{N_1} \{A(x_{j1}, \alpha_0), \ldots, A(x_{jm^r}, \alpha_0)\}$$

is bounded below by $\{1 + o_p(1)\}$ multiplied by the $(1-\xi)$-level quantile of the N_1 medians

$$\mathrm{med}\{A(x_{j1}, \alpha_0), \ldots, A(x_{jm^r}, \alpha_0)\}\,, \quad 1 \le j \le N_1\,,$$

then for all $\eta > 0$,

$$P\Big[\mathrm{quant}_{1-\xi}\,\{e(x, \alpha_0) : x \in \mathcal{R}\} \le \mathrm{quant}_{1-\xi}\,\{A(x, \alpha_0) : x \in \mathcal{R}\} + \eta\Big] \to 1\,.$$

Since $\mathrm{quant}_{1-\xi}\,\{e(x, \alpha_0) : x \in \mathcal{R}\} = \alpha_\xi(\alpha_0)$ then

$$P\Big[\mathrm{quant}_{1-\xi}\,\{A(x, \alpha_0) : x \in \mathcal{R}\} \le \alpha_\xi(\alpha_0) + \eta\Big] \to 1\,. \tag{B.6}$$

In view of (4.9),

$$P\Big[\Big|\mathrm{quant}_{1-\xi}\,\{A(x, \alpha_0) : x \in \mathcal{R}\} - \mathrm{quant}_{1-\xi}\,\{\hat\alpha(x, \alpha_0) : x \in \mathcal{R}\}\Big| > \eta\Big] \to 0 \tag{B.7}$$

for all $\eta > 0$, and moreover, if δ satisfying (B.4) is chosen sufficiently small,

$$\mathrm{quant}_{1-\xi}\,\{\hat\alpha(x, \alpha_0) : x \in \mathcal{R}\} - \hat\alpha_\xi(\alpha_0) \to 0 \tag{B.8}$$

in probability. (This can be deduced from the definition of $\hat\alpha_\xi(\alpha_0)$ at (2.19).) Combining (B.6)–(B.8) we deduce that $P\big\{\hat\alpha_\xi(\alpha_0) \le \alpha_\xi(\alpha_0) + \eta\big\} \to 1$ for all $\eta > 0$, which is equivalent to (4.10).