cemmap
centre for microdata methods and practice

A bias bound approach to nonparametric inference

Susanne M. Schennach

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP71/15

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

A bias bound approach to nonparametric inference

Susanne M. Schennach*
Brown University
smschenn@alum.mit.edu

First version: May 15, 2013.
This version: November 15, 2015

Abstract

The traditional approach to obtain valid confidence intervals for nonparametric quantities is to select a smoothing parameter such that the bias of the estimator is negligible relative to its standard deviation. While this approach is apparently simple, it has two drawbacks: First, the question of optimal bandwidth selection is no longer well-defined, as it is not clear what ratio of bias to standard deviation should be considered negligible. Second, since the bandwidth choice necessarily deviates from the optimal (mean squares-minimizing) bandwidth, such a confidence interval is very inefficient. To address these issues, we construct valid confidence intervals that account for the presence of a nonnegligible bias and thus make it possible to perform inference with optimal mean squared error minimizing bandwidths. The key difficulty in achieving this involves finding a strict, yet feasible, bound on the bias of a nonparametric estimator. It is well-known that it is not possible to consistently estimate the pointwise bias of an optimal nonparametric estimator (for otherwise, one could subtract it and obtain a faster convergence rate violating Stone's bounds on optimal convergence rate). Nevertheless, we find that, under minimal primitive assumptions, it is possible to consistently estimate an *upper bound* on the magnitude of the bias, which is sufficient to deliver a valid confidence interval whose length decreases at the optimal rate and which does not contradict Stone's results.

1 Introduction

While the classic topic of nonparametric inference is well-established and has been the subject of many reviews (for instance, Härdle and Linton (1994), Pagan and Ullah (1999),

*The author would like to thank Florian Gunsulius, Daniel Wilhelm, as well as seminar participants at the Harvard/MIT econometrics seminar, the Brown econometrics lunch seminar, the 2014 Cemmap Microdata Methods and Practice conference and the Conference in Honor of Jerry Hausman for helpful comments. Any errors are my own. Support from NSF grant SES-1357401 is gratefully acknowledged.

Ichimura and Todd (2007), Li and Racine (2007), Horowitz (2009)), the topic is still receiving considerable ongoing attention (Lewbel and Linton (2002), Giné and Nickl (2010), Low (1997), Cai, Low, and Xia (2013), Hoffmann and Nickl (2011), Cai, Low, and Ma (2014), Armstrong (2014), Chernozhukov, Chetverikov, and Kato (2014), Calonico, Cattaneo, and Titiunik (2014), Armstrong and Kolesár (2014), Calonico, Cattaneo, and Titiunik (2015), Pinkse (2001), Su and Ullah (2008), among many others), due to numerous open problems that have not found theoretical or practical solutions. One of these remaining problems is the following undesirable dilemma. When conducting statistical inference in traditional nonparametric settings, one can either use the optimal level of smoothing which minimizes the sum of the squared bias and the variance, but the resulting limiting distribution is shifted off center by an amount that is not asymptotically negligible and that is not consistently estimatable, thus affecting the validity of the resulting inference. Alternatively, one can select a bandwidth that undersmooths to obtain the asymptotically negligible bias necessary for valid inference, but at the expense of abandoning efficiency. The loss of efficiency is actually quite large in this case, because the mean squared error for an undersmoothed bandwidth sequence is, asymptotically, infinitely bigger than the one obtained at the optimal bandwidth.

This old dilemma has recently received renewed interest and various approaches towards a solution have been proposed (see, for instance, Hall and Horowitz (2013), Calonico, Cattaneo, and Farrell (2013), Chernozhukov, Chetverikov, and Kato (2014), Hansen (2014)), as discussed in Section 3. In this paper, we propose a different solution that is both simple to implement and relies on transparent primitive assumptions. We observe that there is no reason to expect that confidence intervals in a nonparametric context should take the standard form of a point estimate expanded by a multiple of some standard deviation. An arguably more appropriate approach is one that allows for an interval of possible values for the bias. That is, we obtain a feasible bound on the magnitude of the bias and take it into account during inference, rather than attempting to make it negligible. This is accomplished via a Fourier representation of the bias and a connection between the large frequency behavior of Fourier transforms with fundamental results in the theory of dynamical systems (Birkhoff (1931b)).

Our method relies on a nonparametric kernel point estimate $\hat{f}$, such as a density or a conditional expectation at a given point, estimated for the optimal (mean squared error minimizing) smoothing parameters. An interval including the true value of f can, theoretically, be obtained by taking the point $E\left[\hat{f}\right]$ and expanding it into an interval whose width is determined by an upper bound on the possible bias $\bar{b}$ of such an estimate:

$$\left[E\left[\hat{f}\right]-\bar{b}, E\left[\hat{f}\right]+\bar{b}\right].$$

As the estimate $\hat{f}$ obtained in practice is random, the boundaries of a confidence interval centered on $\hat{f}$ need to be further broadened by an appropriate multiple z of some estimated standard deviation $\hat{\sigma}$ to yield a confidence interval of the form:

$$\left[\hat{f}-\bar{b}-z\hat{\sigma}, \hat{f}+\bar{b}+z\hat{\sigma}\right].$$

By explicitly accounting for the unknown, but bounded, bias, one is not forced to use an undersmoothed suboptimal bandwidth. The key difficulty in proceeding in this fashion involves finding a strict bound on the bias of a nonparametric estimator that can be feasibly estimated. It is well-known that it is not possible to consistently estimate the pointwise bias of an optimal nonparametric estimator, for otherwise, one could subtract it and obtain a faster convergence rate violating Stone's bounds (Stone (1980), Stone (1982)) on the optimal convergence rate. Nevertheless, we show that it is possible to consistently estimate an *upper bound on the magnitude* of the bias under primitive regularity conditions, which is sufficient to deliver a valid confidence interval and that does not contradict Stone's results.

As a simple example to fix the ideas, if $\hat{f}$ is a kernel estimate of a density f whose second derivative f'' may not be continuous, but is known to be bounded by some constant $\bar{f}''$, then an upper bound on the bias is given by $\bar{b} = k_2 h^2 \bar{f}''/2$, where h is the bandwidth and k_2 is the second moment of the kernel. The lack of continuity in f'' precludes a more efficient estimation, but boundedness is sufficient to strictly bound the bias. Of course, having to specify an *a priori* bound on the second derivative would be unappealing in general, and much of our efforts below will be devoted to avoiding this via formal ways to estimate an upper bound on the bias that converges sufficiently fast so that it does not affect the asymptotic distribution. The key idea is to express the bias in terms of the Fourier transform of the unknown function of interest and find bounds on the latter that are implied by simple primitive smoothness conditions. These implied bounds take the form of power law bounds on the Fourier transform that can be shown to be reached (within an arbitrarily small tolerance) at almost periodic interval as frequency increases, thanks to a powerful result borrowed from the theory of dynamical systems. This enables both the consistent estimation of the bounds and their use in a simple estimator of the bias bound.

This paper is organized as follows: We first introduce the basic results needed to bound the bias, before considering the estimation of such bounds. We then use these results to generate confidence intervals after adding variance contributions. A comparison with other proposals to address the bias issue in nonparametrics inference can be found in Section 3. We discuss various extensions, for instance, estimation of derivatives, of quantiles, adaptive estimation, etc. All proofs can be found in the appendix.

2 Main results

2.1 Notation and definitions

To transparently cover the density and conditional mean cases jointly whenever possible, we focus on quantities of the form

$$f_{Y;X}(x) \equiv E[Y|X = x] f_X(x)$$

where X and Y are random variables and f_X denotes the density of X with respect to the Lebesgue measure. Note that this specializes to a density if $Y = 1$. Conditional expectation can be expressed as $E[Y|X = x] = f_{Y;X}(x)/f_{1;X}(x)$. For results that do not depend on

the specific variables Y and X involved, we will abbreviate $f_{Y;X}$ as f. Fourier transforms will generally be denoted by the corresponding greek letter, e. g., the Fourier transform of $f_{Y;X}(x)$ is denoted by $\phi_{Y;X}(\xi) \equiv \int f_{Y;X}(x) e^{i\xi x} dx$.

We consider kernel estimators of $f_{Y;X}(x)$:

$$\hat{f}_{Y;X}(x) = \hat{E}\left[Y\frac{1}{h}K\left(\frac{x-X}{h}\right)\right] = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{1}{h}K\left(\frac{x-X_i}{h}\right)$$

for some kernel $K(\cdot)$ and some bandwidth $h > 0$. We let $\hat{E}$ denote the sample average operator; we will also denote the usual variance estimator by $\widehat{\text{Var}}$. More generally, all estimators are denoted with hats and their dependence on n will be implicit. Similarly, the n-dependence of the bandwidth h will be omitted in the notation.

Kernel estimators are convenient in the present context because their bias admits a very simple expression, thus making the analysis more transparent. To study the bias, it is useful to define the expected value:

$$\bar{f}_{Y;X}(x) = E\left[\hat{f}_{Y;X}(x)\right] = \int \frac{1}{h}K\left(\frac{x-\tilde{x}}{h}\right) f_{Y;X}(\tilde{x})\, d\tilde{x},$$

where integrals are conventionally over the real line unless otherwise specified.

We then define a set of smooth functions that will play a central role in our approach. To do so, we recall the following well-known concept:

Definition 1 *The total variation $TV(f)$ of a function $f(x)$ is*

$$TV(f) = \sup_{m\in\mathbb{N}} \sup_{\substack{\{x_0,\ldots,x_m\}:\\ \text{partition of } \mathbb{R}}} \sum_{j=1}^{m} |f(x_j) - f(x_{j-1})|.$$

Intuitively, the total variation is the sum of the "up" and "down" absolute movements in the value of the function. For a function f whose differential exists and is integrable, $TV(f) = \int_{-\infty}^{\infty} \left|f^{(1)}(x)\right| dx$, but the above definition holds more generally (for instance, without assuming differentiability). The following definition specifies a hierarchy of increasingly smooth functions.

Definition 2 *For*[1] *$r \in \mathbb{N}\setminus\{0,1\}$ and $A, B \in \mathbb{R}^+$, let $\mathcal{F}^r_{A,B}$ be the set of all functions $f : \mathbb{R} \mapsto \mathbb{R}$ such that (i) $\int |f(x)|\, dx$ exists and does not exceed B, (ii) for $k = 0, \ldots, r-1$, the k-th derivative $f^{(k)}$ exists and satisfies $\lim_{x\to\pm\infty} f^{(k)}(x) = 0$, (iii) the total variation of $f^{(r-1)}$ is at most A and (iv) $f^{(r-1)}$ is absolutely continuous except over a finite nonempty set of points.*

[1] The requirement that $r \geq 2$ simply rules out discontinuous functions (for which the bias could never be uniformly bounded).

We present our theory within the class of functions admitting a finite (but arbitrarily large) number of derivatives, which is the setting that is traditionally used in the context of nonparametric estimation.[2] Definition 2 ensures that the $\mathcal{F}_{A,B}^{r}$ are disjoint (for two different values of r) by requiring that $f^{(r-1)}(x)$ fail to be smooth at least at one point (otherwise, a function that satisfies a certain level of smoothness would obviously satisfy the conditions for a set of functions having a lower level of smoothness, which would be notationally inconvenient). This is not the only or the most general way to accomplish this, but it is the most convenient for the present purpose.

The assumption $f \in \mathcal{F}_{A,B}^{r}$ places bounds on the tail behavior of the functions considered that are not very restrictive, because we work with $f_{Y;X}(x) \equiv E[Y|X=x] f_X(x)$, which is downweighted by a density, rather than with $E[Y|X=x]$ itself. This assumption ensures that various boundary terms arising from integration by parts of a kernel estimator vanish, which is a standard assumption of the nonparametric and semiparametric literature (for instance, Hardle and Stoker (1989)).

2.2 "Oracle" bias bounds

Our approach exploits the fact that the nonparametric bias takes a particularly simple form in Fourier representation. It is well-known that the bias $f_{Y;X}(x) - \bar{f}_{Y;X}(x)$ is given by the inverse Fourier transform

$$f_{Y;X}(x) - \bar{f}_{Y;X}(x) = \frac{1}{2\pi} \int (1 - \kappa(h\xi)) \phi_{Y;X}(\xi) e^{-i\xi x} d\xi, \tag{1}$$

where $\kappa(\cdot)$ denotes the Fourier transform of the kernel $K(\cdot)$ (Schennach (2004)). Hence, a simple upper bound on the bias can be obtained if one can find a bound on $\left|\phi_{Y;X}(\xi)\right|$:

$$\left|f_{Y;X}(x) - \bar{f}_{Y;X}(x)\right| \leq \frac{1}{2\pi} \int |1 - \kappa(h\xi)| \left|\phi_{Y;X}(\xi)\right| d\xi \tag{2}$$

We then rely on a direct relationship between the smoothness of a function and the rate of decay of its Fourier transform.

Lemma 1 $\sup_{f \in \mathcal{F}_{A,B}^{r}} \left|\int f(x) e^{\mathbf{i}\xi x} dx\right| = \min\left\{B, A|\xi|^{-r}\right\}$ *for all* $\xi \in \mathbb{R}$.

This Lemma, proven in the Appendix, is similar in spirit to the well-know Riemann-Lebesgue Lemma in that it relates smoothness of a function to the decay of its Fourier transform. It is more specific in that it provides a specific rate of decay that is related to

[2] As discussed in Section 4, the possibility of functions admitting an infinite number of derivatives (such as "supersmooth" functions) could be considered. However, since it is empirically very difficult to distinguish a function with large but finite number of derivatives from an infinitely differentiable function, the extra level of complexity may not appeal to most practitioners. Also, it is possible to construct finitely many time differentiable functions that do not belong to any of the $\mathcal{F}_{A}^{r}$, but such functions are somewhat peculiar. For instance, they are functions that are differentiable to some fractional order. Our focus on the classes $\mathcal{F}_{A,B}^{r}$ is driven by the ability to provide transparent primitive conditions.

the smoothness parameter r. By Lemma 1, if one knew which class $\mathcal{F}_{A,B}^{r}$ of functions f belongs to, one could obtain a bound on the bias. Of course, in practice, one would not have this information and we will turn shortly to the problem of devising data-driven rules to determine such information. For the moment, it is nevertheless useful to first consider "oracle" bias bounds. Note that the bound $A\left|\xi\right|^{-r}$ for $r>0$ becomes uninformative for very small ξ, so it is important to combine this bound with the trivial constant bound B from the assumption of absolute integrability to ensure a finite bound at all frequencies that yields a finite upper bound on the bias. The Lemma 2 below formalizes these ideas. But first, we need to more specific about the kernel used.

Assumption 1 *The Fourier transform of the kernel, $\kappa(\xi)$, satisfies $\kappa(\xi)=1$ in a neighborhood of the origin and $\kappa(\xi)\leq\bar{\kappa}<\infty$.*

Employing an "infinite order" kernel makes the method adaptive in the sense that one does not need to be concerned whether the order of the kernel is sufficient to exploit the level of smoothness of the function to be estimated. Examples of infinite order kernels can be found in (Politis and Romano (1999) and Schennach (2004)). The use of an infinite order kernel is not essential for the method to work, however. Any kernel of a sufficiently high order (so that it does not limit the convergence rate given the smoothness of the data generating process) would work as well. In fact, as we will see, our method also indicates, as a by-product, the level of smoothness of the function, which could be used to select a kernel of the appropriate order.

Lemma 2 *For a given $h>0, A_{Y;X}>0, B_{Y;X}>0$ and $r_{Y;X}\in\mathbb{N}\backslash\{0,1\}$, If $f_{Y;X}\in\mathcal{F}_{A_{Y;X},B_{Y;X}}^{r_{Y;X}}$ then*

$$f_{Y;X}(x)\in\left[\bar{f}_{Y;X}(x)-b_{Y;X},\bar{f}_{Y;X}(x)+b_{Y;X}\right]\text{ for all }x\in\mathbb{R}$$

where

$$b_{Y;X}=\frac{1}{2\pi}\int\left|1-\kappa(h\xi)\right|\min\left\{A_{Y;X}\left|\xi\right|^{-r_{Y;X}},B_{Y;X}\right\}d\xi \tag{3}$$

and where $\kappa(\cdot)$ denotes the Fourier transform of $K(\cdot)$. Moreover, for $x\in\mathbb{R}$,

$$\begin{aligned}\inf_{f_{Y;X}\in\mathcal{F}_{A_{Y;X},B_{Y;X}}^{r_{Y;X}}} f_{Y;X}(x) &= \bar{f}_{Y;X}(x)-b_{Y;X}\\ \sup_{f_{Y;X}\in\mathcal{F}_{A_{Y;X},B_{Y;X}}^{r_{Y;X}}} f_{Y;X}(x) &= \bar{f}_{Y;X}(x)+b_{Y;X}.\end{aligned}$$

Also, if Assumption 1 holds,[3] *then $b_{Y;X}=O\left(h^{r_{Y;X}-1}\right)$.*

For $Y=1$, Lemma 2 provides bias bounds for kernel-smoothed densities. Of course, when $f_{1;X}$ is a density or otherwise known to be positive, quantities such $\bar{f}_{Y;X}(x)-b_{Y;X}$ can be replaced by $\max\left\{\bar{f}_{Y;X}(x)-b_{Y;X},0\right\}$.

We can give a similar result for kernel regressions.

[3]If Assumption 1 did not hold, the first part of the Lemma would still be true, but the resulting $b_{Y;X}$ may be too large to be useful.

Lemma 3 *For any* $f_{Y;X} \in \mathcal{F}^{r_{Y;X}}_{A_{Y;X},B_{Y;X}}$ *and any* $f_{1;X} \in \mathcal{F}^{r_{1;X}}_{A_{1;X},B_{1;X}}$

$$E[Y|X=x] = \frac{f_{Y;X}(x)}{f_{1;X}(x)} \in \left[\frac{\bar{f}_{Y;X}(x) - b_{Y;X}}{\max\left\{\bar{f}_{1;X}(x) + \operatorname{sgn}\left(\bar{f}_{Y;X}(x) - b_{Y;X}\right) b_{1;X}, 0\right\}}, \right.$$
$$\left. \frac{\bar{f}_{Y;X}(x) + b_{Y;X}}{\max\left\{\bar{f}_{1;X}(x) - \operatorname{sgn}\left(\bar{f}_{Y;X}(x) + b_{Y;X}\right) b_{1;X}, 0\right\}} \right]$$

for all $x \in \mathbb{R}$*, where* $b_{Y;X}$ *and* $b_{1;X}$ *are given in Lemma 2 and where the function* $\operatorname{sgn}(u)$ *returns the sign* ($\{-1, 0, +1\}$) *of* u *while the function* $v/\max\{u, 0\}$ *has the interpretation*

$$\frac{v}{\max\{u,0\}} \equiv \begin{cases} +\infty & \text{if } u \leq 0 \text{ and } v > 0 \\ -\infty & \text{if } u \leq 0 \text{ and } v < 0 \\ 0 & \text{if } u \leq 0 \text{ and } v = 0 \\ \frac{v}{u} & \text{otherwise} \end{cases}$$

and an interval of the form $[\infty, \infty]$ *or* $[-\infty, -\infty]$ *is considered empty.*

2.3 Feasible bias bounds

In order to obtain a practically useful method, we now describe a way to empirically determine which class $\mathcal{F}^r_{A,B}$ a given estimated function belongs to. The idea is simple: As illustrated in Figure 1 and proven in Theorem 1 below, one simply considers the estimated log characteristic function $\ln\left|\hat{\phi}(\xi)\right|$ as function of log frequency $\ln\xi$ over some range $\left[\ln\underline{\xi}, \ln\bar{\xi}_n\right]$ of log frequencies (whose selection will be discussed later). One then finds the tightest linear upper bound $\ln\left|\hat{\phi}(\xi)\right| \leq \ln A - r\ln\xi$ over that interval. Tightness is quantified by the area under the bounding linear function. The intercept $\ln\hat{A}$ and slope $-\hat{r}$ of the tightest linear bound will be shown to provide consistent estimates of the A and r parameters, respectively. The upper bound $\bar{\xi}_n$ must increase at a controlled rate as sample size grows, because the empirical counterpart $\hat{\phi}(\xi)$ converges to $\phi(\xi)$ uniformly only over an expanding interval (and not uniformly over the whole real line). The lower bound $\underline{\xi}$ is introduced because it is the large-ξ behavior of the Fourier transform that is relevant for evaluating the asymptotic bias.

Such a simple scheme works, because any function in the class $\mathcal{F}^r_{A,B}$ we consider will be shown to have a useful property: The magnitude of its Fourier transform $|\phi(\xi)|$ actually visits a neighborhood of its upper bound $A|\xi|^{-r}$ no less often than a certain periodic interval. The practical implication of this is that not only does the Fourier transform of a function in $\mathcal{F}^r_{A,B}$ exhibits a power law behavior, but it also does so rather uniformly, so that this feature can be detected even if one has access to a consistent estimate of the spectrum only over some finite frequency range. This result will not be merely assumed, but will rather be proven from our primitive conditions, by exploiting a connection with ergodicity in dynamical systems (as explained further below and as detailed in the proof of Theorem 1).

To state our formal result regarding the estimation of the constants $A_{1;X}, A_{Y;X}$, $B_{Y;X}$, $r_{1;X}$, $r_{Y;X}$, we need a few assumptions. We consider iid settings for simplicity, although this could be relaxed.

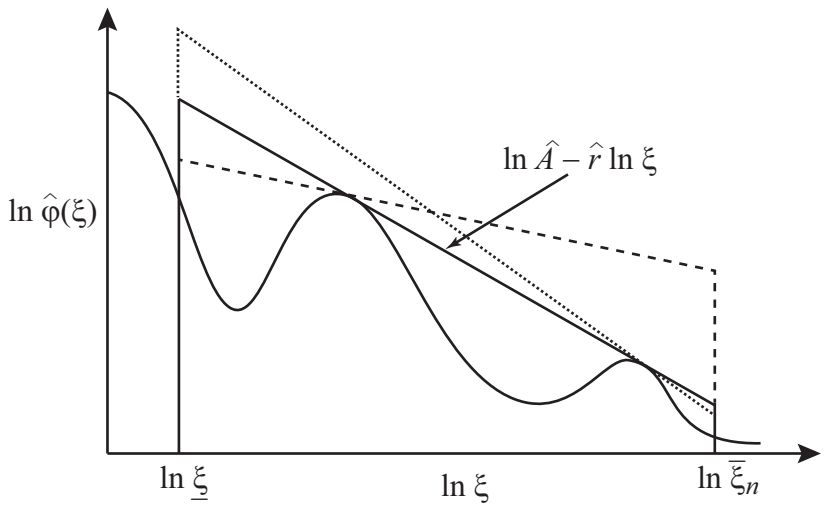

Figure 1: Estimation of the parameters A and r via determination of the tightest linear upper bound (solid line) on the log empirical Fourier transform $\ln \hat{\phi}(\xi)$ as a function of log frequency $\ln \xi$ over some interval $[\ln \underline{\xi}, \ln \overline{\xi}_n]$, based on the minization of the area under the line.

Assumption 2 *(Y_i, X_i) forms an iid sequence of random variables, each taking values in $\mathbb{R}^2$.*

Assumption 3 *$f_{Y;X} \in \mathcal{F}^{r_{Y;X}}_{A_{Y;X},B_{Y;X}}$ for some $r_{Y;X} \in \mathbb{N}\backslash\{0,1\}$ and $A_{Y;X}, B_{Y;X} \in \mathbb{R}^+$ (all of which do not need to be known a priori).*

The randomness in the dependent variable Y also needs to be minimally constrained.

Assumption 4 *$E[Y^2] < \infty$ and $\mathrm{Var}[Y|X = x] \leq C_V$ for some $C_V < \infty$ and $E[|YX|] < \infty$.*

Next, Theorem 1 below formally establishes the validity of the bounding scheme outlined earlier. For density estimation one invokes Theorem 1 for $Y = 1$, while for conditional expectations, one invokes Theorem 1 for both $Y = 1$ and for general Y.

Theorem 1 *Let Assumptions 2-4 hold. Let $\hat{B}_{Y;X} = \hat{E}[|Y|]$ and*

$$\left(\hat{A}_{Y;X}, \hat{r}_{Y;X}\right) = \arg \min_{(A,r)\in\mathcal{B}_{Y;X}} \int_{\ln \underline{\xi}}^{\ln \bar{\xi}_n} (\ln A - r\lambda)\, d\lambda \tag{4}$$

$$\mathcal{B}_{Y;X} = \left\{(A,r) : A \in \mathbb{R}^+ \text{ and } r \in \mathbb{N} \text{ and } \left|\hat{\phi}_{Y;X}(\xi)\right| \leq A|\xi|^{-r} \text{ for } \xi \in \left[\underline{\xi}, \bar{\xi}_n\right]\right\} \tag{5}$$

where $\hat{\phi}_{Y;X}(\xi) = \hat{E}\left[Ye^{\mathbf{i}\xi X}\right]$ and for some $\underline{\xi} > 0$ and $\bar{\xi}_n$ such that $\bar{\xi}_n \to \infty$, $\bar{\xi}_n = O\left(n^{1/4}\right)$ and $\Delta\phi_{Y;X,n}\bar{\xi}_n^{r_{Y;X}} \to 0$ where

$$\Delta\phi_{Y;X,n} = \begin{cases} \sqrt{7}\mu_Y n^{-1/2}(\ln n)^{1/2} & \text{for compactly supported } Y \\ \frac{7\sqrt{2}}{3}\mu_Y n^{-1/2}\ln n & \text{for general random } Y \end{cases}$$

and $\mu_Y = \left(E\left[Y^2\right]\right)^{1/2}$.
If $f_{Y;X} \in \mathcal{F}^{r_{Y;X}}_{A_{Y;X},B_{Y;X}}$ *for some* $A_{Y;X}, B_{A_{Y;X}} \in \mathbb{R}^+$ *and* $r_{Y;S} \in \mathbb{N}\backslash\{0,1\}$ *(but*[4] $f_{Y;X} \notin \mathcal{F}^{r_{Y;X}}_{A,B}$ *for either* $A < A_{Y;X}$ *or* $B < B_{Y;X}$*) then*

$$\begin{aligned} P\left[\hat{r}_{Y;X} = r_{Y,X}\right] &\rightarrow 1 \\ \hat{A}_{Y;X} - A_{Y;X} &= O_p(1) \\ \hat{B}_{Y;X} - B_{Y;X} &= O_p\left(n^{-1/2}\right) \end{aligned}$$

Proof. See Appendix. ■

This theorem can intuitively be understood as follows. First, one needs to establish that the Fourier transform $\phi(\xi)$ of a function in $\mathcal{F}^r_{A,B}$ has the property that its magnitude reaches the power law envelope of the form $A|\xi|^{-r}$ sufficiently often so that (i) observing $\phi(\xi)$ over a sufficiently long finite interval enables us to determine r and A and (ii) the power law behavior is indicative of actual behavior of the Fourier transform (i.e., replacing the true bias expression (1) by its bound (2) in terms of $A|\xi|^{-r}$ does not result in an overly pessimistic bias bound).

To show that the power law bound is indeed reached often, we observe that, by the definition of $\mathcal{F}^r_{A,B}$, the derivative $f^{(r-1)}(x)$ is absolutely continuous except over a finite nonempty set of points, and can therefore be decomposed as the sum of an absolutely continuous function and a finite sum of step functions. This conclusion, formalized in Lemma 5 in the Appendix, is a consequence of the Lebesgue decomposition theorem (see, for instance, Loève (1977)) applied to functions of bounded variations. After a Fourier transform of $f(x)$, these smooth and step-like components of $f^{(r-1)}(x)$ are mapped, respectively, into a rapidly decaying function ($o\left(|\xi|^{-r}\right)$, by Lemma 6) and a finite sum of the form $(i\xi)^{-r}\sum_{j=1}^{J} A_j e^{i\xi x_j}$ for some constants A_j, x_j. For large frequency ξ, the latter, oscillatory, terms dominate. As detailed in Lemma 7 in the Appendix, the oscillations $A_j e^{i\xi x_j}$ can be viewed as the solution to a simple system of first-order differential equations, and fundamental results from the theory of dynamical systems can thus be applied, namely Birkhoff's theorem on recurrence time (Birkhoff (1931b)). Birkhoff's theorem shows that the state of the system (described by the vector with entries $\psi_j \equiv A_j e^{i\xi x_j}$) visits any given open region of the set of all possible states at nearly periodic intervals. As a result, as $\bar{\xi}_n \rightarrow \infty$, one is assured that, eventually, over the interval $\left[\underline{\xi}, \bar{\xi}_n\right]$ one will see at least two local maxima in the $|\xi|^r \phi(\xi)$ that are within a small ε of the maximum possible value of the function.$|\xi|^r \phi(\xi)$. These two maxima then eventually pin down the value of r and A within a given tolerance.

Although Birkhoff's result does not predict how close these two maxima must be, such information is not needed to formally show consistency of our bias bound estimator, since estimation errors in $\hat{A}_{Y;X}$ and $\hat{r}_{Y;X}$ only have an effect on higher-order asymptotically negligible remainder terms, as discussed further below. The situation is entirely analogous to the widely used strategy of only showing consistency (but not the rate of convergence) of

[4]This qualification merely ensures that the values $A_{Y;X}$ and $B_{Y;X}$ are the smallest possible constant yielding a valid bound.

standard error estimators, because the associated estimation error has no effect on first-order asymptotics of t-statistics.[5]

To see how estimation of r and A is accomplished in practice, consider what happens when minimizing the area under the line, (given by Equation (4)), subject to the constraint (5) that it bounds the estimated function. As illustrated in Figure 1, for a value of r that is too small (shown by a dashed line), the area under the line will grow (as $\bar{\xi}_n \to \infty$) faster than for the correct value of r (shown by a solid line), because the dashed line remains constrained by the same low-frequency peak, regardless of the value of $\bar{\xi}_n$. Beyond that low-frequency peak, the dashed line will always lie above the solid line and thus, for sufficiently large $\bar{\xi}_n$, the corresponding area will be larger.

Conversely, for a value of r that is too large (shown by a dotted line), the area under the line will also grow (as $\bar{\xi}_n \to \infty$) faster than the for the correct value of r (shown by a solid line) because the dotted line is constrained by peaks lying at increasing frequencies (as $\bar{\xi}_n \to \infty$). As a result, the dotted will always lie above the solid one up to that constraining peak and the area under the line will grow faster than for the correct r. It follows that, for sufficiently large $\bar{\xi}_n$, only the correct r will minimize the area under the line. Consistency of the corresponding prefactor $\hat{A}$ then follows easily.

Since the exponent r takes one of a set of discrete values, the convergence properties of its estimator $\hat{r}$ are especially simple: It "snaps" to the correct value of r with probability approaching one. This implies that sampling fluctuations in $\hat{r}$ can be neglected for the purpose of asymptotic inference. Sampling fluctuations in the estimator $\hat{A}$ translate into higher-order effects on the asymptotic distribution and do not need to be accounted for during inference either. This follows form the fact that the asymptotic bias bound is proportional to Ah^{r-1}, so replacing A by a consistent estimator yields

$$\hat{A}h^{r-1} = \left(A + o_p\left(1\right)\right) h^{r-1} = Ah^{r-1} + o_p\left(h^{r-1}\right)$$

where the remainder term is asymptotically negligible. Intuitively, this phenomenon occurs because we obtain $\hat{A}$ from low frequencies $\left[\underline{\xi}, \bar{\xi}_n\right]$, but its noise is scaled down to a negligible level when we extrapolate the power law to higher frequencies to calculate the bias (since $\left|1 - \kappa\left(h\xi\right)\right|$ in Equation (3) takes nonnegligible values only for large ξ).

Since the method uses the information gathered for low frequencies to extrapolate the behavior at higher frequencies (that are not directly estimatable), this prompts the question of whether there are functions for which this extrapolation would be misleading. One of the key aspects of the proof of Theorem 1 is specifically to ensure that this does not happen, for sufficiently large samples, for functions in $\mathcal{F}^r_{A,B}$. It is possible, however, to construct

[5]It is also worth noting that, as the interval $\left[\underline{\xi}, \bar{\xi}_n\right]$ increases, even a fairly large value of ε (the allowed distance between a local maximum and the true maximum) becomes sufficient to pin down the correct value of r. Since the distance between two "near" maxima decreases as the tolerance ε is made larger, the recurrence of the event "$\left|\xi\right|^r \phi\left(\xi\right)$ comes within ε of its maximum" should then be a relatively common occurrence. Hence, we are using Birkoff's theorem in a regime where it has good predictive value, rather than in extreme cases (see for instance, Petersen (1983), Chap 2) where the recurrence only takes place over long intervals.

examples of functions, not belonging to $\mathcal{F}_{A,B}^{r}$, whose Fourier transforms go through an alternance of regimes with different rates of decay. The method would not apply in such cases, because extrapolation would fail and the resulting error is not necessarily in the direction of conservative inference. Fortunately, this situation presents itself only for rather contrived examples, for instance, a function that appears k times differentiable at a low resolution, but that reveals a different level of smoothness upon zooming. Such functions have been constructed in the field of fractal analysis Mandelbrot (1982), but are not typical in statistical and economics applications.

Note that the theorem accounts for the fact that one does not observe the true $\phi_{Y;X}(\xi)$ but rather an estimate $\hat{\phi}_{Y;X}(\xi)$. Indeed, it is possible to derive an almost sure uniform upper bound $\Delta\phi_{Y;X,n}$ on the error on an estimated Fourier transform $\hat{\phi}_{Y;X}(\xi)$ on an expanding interval (see Lemma 4 in the Appendix). As long as this error is asymptotically negligible relative to the true $\phi_{Y;X}(\xi)$, which is guaranteed by the requirement that $\Delta\phi_{Y;X,n}\bar{\xi}_n^{r_{Y;X}} \to 0$ in Theorem 1, the geometrical argument illustrated in Figure 1 still holds.

In practice, the parameter $\underline{\xi}$ and the sequence $\bar{\xi}_n$ needed to employ Theorem 1 can be determined by the following data-driven rule:

Theorem 2 *In addition to the conditions of Theorem 1, assume* $E\left[|X|^2\right] < \infty$. *Let* $\hat{\mu}_X = \left(\widehat{\mathrm{Var}}\left[X\right]\right)^{1/2}$ *and* $\underline{\xi} = \hat{\mu}_X^{-1}$. *Let* $\bar{\xi}_n$ *be the largest* $\xi \in \left[0, \hat{\mu}_X^{-1} n^{1/4}\right]$ *such that* $\Delta\hat{\phi}_{Y;X,n} / \left|\hat{\phi}_{Y;X}(\xi)\right| \leq (\ln n)^{-1}$ *where* $\Delta\hat{\phi}_{Y;X,n}$ *is as* $\Delta\phi_{Y;X,n}$ *given in Theorem 1 with* μ_Y *replaced by* $\hat{\mu}_Y \equiv \left(\hat{E}\left[Y^2\right]\right)^{1/2}$. *This choice satisfies*

$$\Delta\phi_{Y;X,n}\bar{\xi}_n^{r_{Y;X}} \leq C_n$$

with the inequality holding with probability approaching one for some deterministic sequence C_n *with the property* $C_n \to 0$. *Moreover, these choices yield inference results that are invariant to nondegenerate linear transformations of* X.

This rule (whose validity is proven in the appendix) proceeds by first ensuring that the range of ξ considered does not grow faster than $n^{1/4}$, which is important to secure a specific rate for the convergence of the estimated Fourier transform. Next, the test $\Delta\hat{\phi}_{Y;X,n} / \left|\hat{\phi}_{Y;X}(\xi)\right| \leq (\ln n)^{-1}$ ensures that the true error bound $\Delta\phi_{Y;X,n}$ decays sufficiently fast relative to the estimated Fourier transform $\hat{\phi}_{Y;X}(\xi)$, which, after simple manipulations, also implies that it is small relative to the true Fourier transform $\phi_{Y;X}(\xi)$, as desired. We introduce scalings by standard deviations in the various prefactors as a simple way to ensure invariance with respect to linear rescaling of the data.[6]

[6] In principe, replacing $\hat{\mu}_X$ by a multiple of it would also yield a valid method, but one has to realize that the upper bound $\hat{\mu}_X^{-1} n^{1/4}$ is rarely the binding constraint in selecting $\bar{\xi}_n$, so this does not really provide an avenue for researchers to influence the results. Similarly, low frequency behavior is rarely the determining factor in determining $\hat{r}$ and $\hat{A}$, so modifying the lower bound $\hat{\mu}_X^{-1}$ only has a limited effect on the final result.

2.4 Inference

The inference procedure (described further below) will be considerably simplified thanks to a few key facts.

First, although the bias is comparable in magnitude to the standard deviation at all sample sizes, the absolute length of the bias interval still converges to zero with increasing sample size. This implies that we can still use the "point estimate" $\hat{f}_{Y;X}(x)$ to build a consistent estimate of the asymptotic variance. This also implies that we can use a linearization of the ratio $\hat{f}_{Y;X}(x)/\hat{f}_{1;X}$ around $\hat{f}_{Y;X}(x) = f_{Y;X}(x)$ and $\hat{f}_{1;X}(x) = f_{1;X}(x)$ to calculate the variance of the estimator, just as in a conventional kernel regression estimator.

Another important point is that the estimated upper and lower bounds on the set of possible biases are perfectly correlated. This automatically avoids issues such having to consider the possibility that the bottom tail of the distribution of the upper bound may extend below the lower bound, in which case corrections to the critical values would have been needed.

A final observation is that the effect of estimation error on $A_{1;X}$, $A_{Y;X}$, $B_{Y;X}$, $r_{1;X}$, $r_{Y;X}$ on the bias interval is of a higher order relative to the standard deviation, as explained in the previous section. As a result, the corresponding estimation noise does not need to be accounted for in the asymptotic distribution. Securing this property requires that the selected bandwidth be "close" to the optimal mean squared error minimizing bandwidth h_{opt} in the sense of Assumption 5 below. Note that this condition allows for, but does not require, undersmoothing.

Assumption 5 $\operatorname{plim}\sup_{n\to\infty} \frac{h}{h_{opt}} < \infty$

We can now state our main inference results. For density estimation our regularity conditions on the density of X specialize to the following:

Assumption 6 $E[|X|] < \infty$.

Assumption 7 $f_{1;X} \in \mathcal{F}^{r_{1;X}}_{A_{1;X},1}$ *for some* $r_{1;X} \in \mathbb{N}\backslash\{0,1\}$ *and* $A_{1;X} \in \mathbb{R}^+$ *(all of which do not need to be known a priori).*

Theorem 3 *If Assumptions 2, 5, 6 and 7 hold, then, for a given* $x \in \mathbb{R}$ *and for a given* $f_X \in \mathcal{F}^r_{A_{1;X},1}$,

$$\lim_{n\to\infty} P\left[f_X(x) \in \hat{\mathcal{D}}(x)\right] \geq 1-\alpha \tag{6}$$

where

$$\hat{\mathcal{D}}(x) = \left[\max\left\{\hat{f}_X(x) - \hat{b}_{1;X} - z_{\alpha/2}\hat{\sigma}_X(x), 0\right\}, \max\left\{\hat{f}_X(x) + \hat{b}_{1;X} + z_{\alpha/2}\hat{\sigma}_X(x), 0\right\}\right]$$

where $z_{\alpha/2}$ *denotes the* $1-\alpha/2$ *quantile of a standard normal distribution.*

$$\begin{aligned}
\hat{b}_{1;X} &= \int_{-\infty}^{\infty} |1-\kappa(h\xi)| \min\left\{1, \hat{A}_{1;X}|\xi|^{-\hat{r}_{1;X}}\right\} d\xi \\
\hat{\sigma}_X(x) &= \left(\frac{\hat{f}_{1;X}(x)}{nh}\int_{-\infty}^{\infty} K^2(u)\,du\right)^{1/2}.
\end{aligned}$$

For conditional expectations, we have similar result:

Theorem 4 *If Assumptions 2-7 hold, then, for a given* $x \in \mathbb{R}$ *and for a given* $f_{Y;X} \in \mathcal{F}^{r_{Y;X}}_{A_{Y;X},B_{Y;X}}$ *and* $f_{1;X} \in \mathcal{F}^{r_{1;X}}_{A_{1;X},1}$,

$$\lim_{n\to\infty} P\left[E\left[Y_i|X_i = x\right] \in \widehat{\mathcal{C}}\left(x\right)\right] \geq 1-\alpha \tag{7}$$

where (under the conventions given in Lemma 3)

$$\widehat{\mathcal{C}}\left(x\right) = \left[\frac{\hat{f}_{Y;X}\left(x\right) - b_{Y;X}}{\max\left\{\hat{f}_{1;X}\left(x\right) + \operatorname{sgn}\left(\hat{f}_{Y;X}\left(x\right) - b_{Y;X}\right) b_{1;X}, 0\right\}} - z_{\alpha/2}\hat{\sigma}_{Y|X}\left(x\right)\ ,\right.$$
$$\left.\frac{\hat{f}_{Y;X}\left(x\right) + b_{Y;X}}{\max\left\{\hat{f}_{1;X}\left(x\right) - \operatorname{sgn}\left(\hat{f}_{Y;X}\left(x\right) + b_{Y;X}\right) b_{1;X}, 0\right\}} + z_{\alpha/2}\hat{\sigma}_{Y|X}\left(x\right)\right]$$

where

$$\begin{aligned}
\hat{b}_{Y;X} &= \int_{-\infty}^{\infty} \left|1-\kappa\left(h\xi\right)\right| \min\left\{\hat{B}_{Y;X}, \hat{A}_{Y;X}\left|\xi\right|^{-\hat{r}_{Y;X}}\right\} d\xi \\
\hat{b}_{1;X} &= \int_{-\infty}^{\infty} \left|1-\kappa\left(h\xi\right)\right| \min\left\{1, \hat{A}_{1;X}\left|\xi\right|^{-\hat{r}_{1;X}}\right\} d\xi \\
\hat{\sigma}_{Y|X} &= \left(\frac{\widehat{\operatorname{Var}}\left[Y|X=x\right]\hat{f}_{Y;X}\left(x\right)}{nh}\int_{-\infty}^{\infty} K^2\left(x\right)dx\right)^{1/2}
\end{aligned}$$

Note that the conventions given in Lemma 3 handle potential divisions by zero, by yielding uninformative confidence regions whenever appropriate.

Thanks to the fact that our bias bounds are uniform, we can readily provide uniform confidence bands (for instance, over $x \in [0,1]$, without loss of generality) by making use of the well-known uniform bands for kernel estimates (see, for instance, Härdle and Linton (1994)). This is accomplished as follows:

Corollary 1 *Under the assumption of Theorem 3, uniform confidence bands over an interval of length* L *can be obtained simply replacing* $z_{\alpha/2}$ *in (6) by*

$$\bar{z}_{\alpha/2} \equiv \left(\frac{c^*_\alpha}{\delta} + \delta + \frac{1}{2\delta}\ln\left(\frac{\int_{-\infty}^{\infty}\left(K^{(1)}\left(u\right)\right)^2 du}{4\pi^2\int_{-\infty}^{\infty}\left(K\left(u\right)\right)^2 du}\right)\right)$$

where $\delta = \sqrt{2\ln\left(L/h\right)}$ *and* $c^*_\alpha = -\ln\left(-\left(\ln\left(1-\alpha\right)\right)/2\right)$. *In the case of conditional expectations, a similar result holds with Equation (7), assuming that* $f_X\left(x\right) > 0$ *in the interval considered in addition to the assumptions of Theorem 4.*

3 Discussion

One may wonder if, somehow, the assumptions on the set $\mathcal{F}^r_{A,B}$ may be such that a more rapidly converging estimator could be devised, based on the idea that we could somehow locate the position and estimate the height of the discontinuities in the derivatives, subtract them and apply a kernel estimator to the, now smoother, remainder. However this scheme does not offer a way to speed up convergence, because the location of the discontinuities in the derivatives cannot be exactly determined in a finite sample (sample point spacings place a fundamental limit on the location accuracy). Any small error on their estimated location results in an imperfect cancellation of the discontinuities and induces a corresponding large error in estimating the value of the derivative over a small region near the discontinuity. Even though the size of the affected region decreases with sample size, the magnitude of the error does not decrease with sample size. This effect thus precludes the existence a more rapidly pointwise converging estimator.

It is instructive to compare the proposed method with recent proposals to address the bias issue in nonparametric inference. Bias correction methods have recently been advocated to address the bias issue in nonparametric inference (Calonico, Cattaneo, and Farrell (2013)). However, their main goal is to obtain statistics that are free of bias (to a sufficiently high order) and that minimize the coverage error of confidence intervals based on standard normal limiting distributions. Although this represents an extremely valuable and convenient tool for practitioners, the method does not exploit the unknown function's level of smoothness to the full extent to obtain the fastest possible convergence rate. In fact, the authors note (in Section 3.6) that implementability of the method becomes a problem when the order of the kernel matches the function's smoothness, because optimal bandwidth selection (for the bias correction method) requires the knowledge of constants that are no longer consistently estimatable. In contrast, by allowing for a nonnegligible bias, our approach allows for the use of the optimal mean-square minimizing bandwidth that yield optimal convergence rates (in the sense of reaching Stone's bounds (Stone (1980), Stone (1982)) for the function's level of smoothness), since we don't require the regularity conditions that would be necessary for consistent estimation of bias-related quantities.

Hall and Horowitz (2013) propose a bootstrap procedure for computing confidence bands around a nonparametric kernel regression that does not require the bias to be negligible. It proceeds by taking the bands obtained while ignoring bias and correcting them by a multiplicative factor determined so that the band's actual coverage is accurate over a given fraction $(1-\varepsilon)$ of the support. Bootstrap replications are used to estimate this coverage. While the method is appealing in its conceptual simplicity, it leads to a nonstandard notion of confidence bands, since the specified confidence level is not met everywhere, but instead over a fraction $(1-\varepsilon)$ of the support of the regressor the location of which the user has no a priori control over. In some sense, their method approaches the problem in a way that is complementary to ours. Our method seeks to bound the bias everywhere and is thus appropriately sensitive to portion of the functions that represent worst-case scenarios. In contrast, their method focuses on the portions of the function where the bias is the smallest (the authors observe that the excluded region is "typically close to the locations of peaks

and troughs", where the bias would be largest for the second-order smoother they use).[7]

Hansen (2014) also suggests a way to account for the bias in nonparametric inference. However, his approach is distinct from ours as it focuses on series estimation and, given the difficulties in obtaining bias expressions in this context, he relies on a high-level assumption specifying the exact form of the bias, while we provide a bias bound obtained from the primitive properties of the function to be estimated.

Our extrapolation property bears some superficial resemblance to the concept of self-similarity used in the adaptive inference literature (for instance, Giné and Nickl (2010), Chernozhukov, Chetverikov, and Kato (2014)). Self-similarity is traditionally expressed in terms of the behavior of wavelet coefficients, but the comparison with our method is facilitated by phrasing the self-similarity assumption in Fourier representation. Self-similarly essentially amounts to assuming the existence of both an upper and a *lower* bound on the Fourier transform that each take the form of a power law *with the same exponent*. In contrast, our condition does not assume a lower bound – it only implies that an upper bound can be proven to be reached often, i.e., quasi periodically. Our condition is also stated in terms of simple primitive smoothness conditions, instead of being directly assumed in terms of a specific power law behavior of wavelet or Fourier coefficients. Moreover, our approach is better adapted to calculate the bias, because the upper bound is reached often, whereas no such guarantee exists in "self-similarity". In fact, when assuming self-similarity, one may want to make the lower bound as small as possible and the upper bound as high as possible to allow for the largest possible class of self-similar functions. Yet, this situation is precisely one where the bias implied by the upper bound could be much larger than the actual bias, because the true function could happen to always remain close to the lower bound beyond a certain frequency.

4 Extensions

The above framework can be adapted to other similar problems. We did not include them in the above results to simplify and shorten the exposition. We briefly point out some possibilities below with their associated potential hurdles.

Derivatives of densities and conditional expectations can be handled in very similar ways. The only notable difference is that one loses one power of h in the order of the bias for every additional order of derivative.

Our treatment differs from the usual asymptotic of nonparametric estimators only in the way the bias is handled. Since the bias admits the same expression in time series settings, our results can be adapted to time series settings by using serial-dependence-robust expressions for the standard errors and by rederiving the convergence rate of the empirical characteristic function under more general conditions. While the first step is already available in the literature, the second may require more careful attention.

[7] An additional distinction with our method is that they do not exploit the smoothness of the unknown function to the full extent, because their Assumption 4.2 requires $2k$ Hölder-continuous derivatives, while requiring the use of a smoother with a local polynomial of order $2k-1$.

Quantile nonparametric regressions could be handled by simply redefining the dependent variable to be an indicator function $1(Y_i \leq q)$ for any given q. Of course, confidence bands on $E[1(Y_i \leq q)|X_i = x]$ have to be mapped onto confidence bands of quantiles in the usual way. Note that the fact that the indicator function is bounded enables the use of the tighter value of $\Delta\phi_{Y;X,n}$ in Theorem 1 for the error bound on the estimated Fourier transform.

The proposed method can be embedded into an adaptive estimation procedure (i.e. one that consistently detects the true smoothness of the function to be estimated). Adaptive bandwidth selection procedure are readily available in the literature (for instance, Politis (2003)) and are directly compatible with our approach. This follows from the fact that our bias bound estimator is adaptive as well, since it consistently estimates the decay exponent r reflecting the true smoothness of the unknown function to be estimated. This feature is shared by existing wavelet-based adaptive methods which have recently been proposed (for instance, Giné and Nickl (2010), Hoffmann and Nickl (2011), Cai, Low, and Ma (2014)), but the underlying assumptions regarding the generating process differ, as explained at the end of Section 3.

While biases proportional to an integral power of bandwidth are traditionally considered in nonparametric estimation (Härdle and Linton (1994)), one could also consider decay rates r of the Fourier transform having a nonintegral value. This would parallel the Holder differentiability conditions often made in the adaptive nonparametric estimation literature (Giné and Nickl (2010), Cai, Low, and Xia (2013), Hoffmann and Nickl (2011), Cai, Low, and Ma (2014), Armstrong (2014), Chernozhukov, Chetverikov, and Kato (2014)). Although allowing for general power laws for the rates of decay for the Fourier transform is straightforward, it may be difficult to formulate corresponding transparent sufficient conditions in real space. Considering fractional orders of differentiation may be necessary. Another difficulty lies in the fact that the estimated rate exponent, being nondiscrete, would no longer "snap" to the right value with probability one asymptotically. As a result, it may be necessary to include a safety margin around the estimated exponent for the upper bound on the bias to be asymptotically valid.

One could also allow for supersmoothness (i.e., certain classes of infinitely many time differentiable functions) by considering decay rate of the Fourier transform of the form $\exp\left(-\alpha|\xi|^{\beta}\right)$ with α and β to be estimated. This would parallel the settings sometimes considered in the nonparametric measurement error literature (Schennach (2004)). However, this possibility is not typically considered in the adaptive nonparametrics literature (and in the traditional finite-order kernel literature neither). For this extension, some of the same issues faced when allowing for fractional powers arise: Phrasing simple primitive conditions is difficult, as it would likely involve limits of sequences of differential operators of diverging order.

A multivariate extension is conceptually trivial, although there are many ways to approach the problem, depending on one's preference for complexity/flexibility trade-off. One could use a common smoothness parameter r for all elements of X or one could allow smoothness to differ along each dimension of the covariate. One could even allow for differing smoothness along different directions (not necessarily aligned with the coordinate axes).

Applying the proposed setup to series estimation would entail substantial reworking of the theory, however. Series estimators do not typically admit simple general expressions for the bias, let alone ones that can be uniformly bounded by estimatable quantities. An extension to local polynomial smoothers should be possible, thanks to the fact that such estimators can also be expressed in terms of convolutions.

Our procedure delivers asymptotically valid confidence bands for any bandwidth sequence satisfying Assumption 5, which allows for both undersmoothing and optimally smoothing bandwidths. Hence, a researcher who would have an a priori preference for bands whose width are determined to a larger extent by the standard deviation bands than by the bias bands could set up a loss function framework that would enable the optimization of his own criterion for the type of bands desired, while maintaining the asymptotic validity of the confidence bands.

5 Examples

We consider a simple example of density estimation of a triangular distribution. The data is generated by drawing X_i, $i = 1, \ldots, 1000$ from a triangular distribution that is the convolution of two uniform distributions on on $[0, 1]$. We use an infinite order kernel, so that the method automatically exploits as much smoothness as possible, regardless of the value of r that is estimated. The data-driven method of Theorem 2 is used to determine the interval $\left[\ln \underline{\xi}, \ln \bar{\xi}_n\right]$ of log frequencies used in the determination of $\hat{r}$ and $\hat{A}$ via the method of Theorem 1. As the left panels of Figure 2 show, the true density falls within the bands, as one would expect for a valid confidence band.

However, it is more instructive to specifically test the properties of the bias bound estimator, which can be done in a simulation context by repeating the estimation many times on randomly drawn samples and averaging the results (see right panel of Figure 2). In the resulting $E\left[\hat{f}_{1;X}(x)\right]$ (obtained with 200 replications), the statistical noise averages out and we are left with the bias only. The bias estimator was used using the data of only one replication picked at random, to illustrate the fact that bias bounds obtained even from a finite sample of 1000 observations can accurately reflect the population values. As expected, the true density then lies within the bias band around the expected value of the estimator. Of course, in a finite sample, it is possible that the bounds could be slightly exceeded, as our results are asymptotic, but we show that this error is asymptotically negligible. Nevertheless, since our bias bands are conservative, it is often the case that, even in a finite sample, the bounds are not exceeded. This is the case in this example.

It is interesting to see what happens when the bandwidth is decreased below the optimal bandwidth. The estimate (on the left) becomes more wiggly, but the truth still lies within the bands. If we now perform the same averaging experiment (on the right of Figure 2), we see that the averaged estimator has fewer wiggles, as expected, and that the truth lies exclusively within the bias bounds around the averaged estimator. It is clear from this experiment that a method that allows for the use of an optimal bandwidth is much preferable. Not only are the bands narrower than for the undersmoothed estimator, but they are also much less

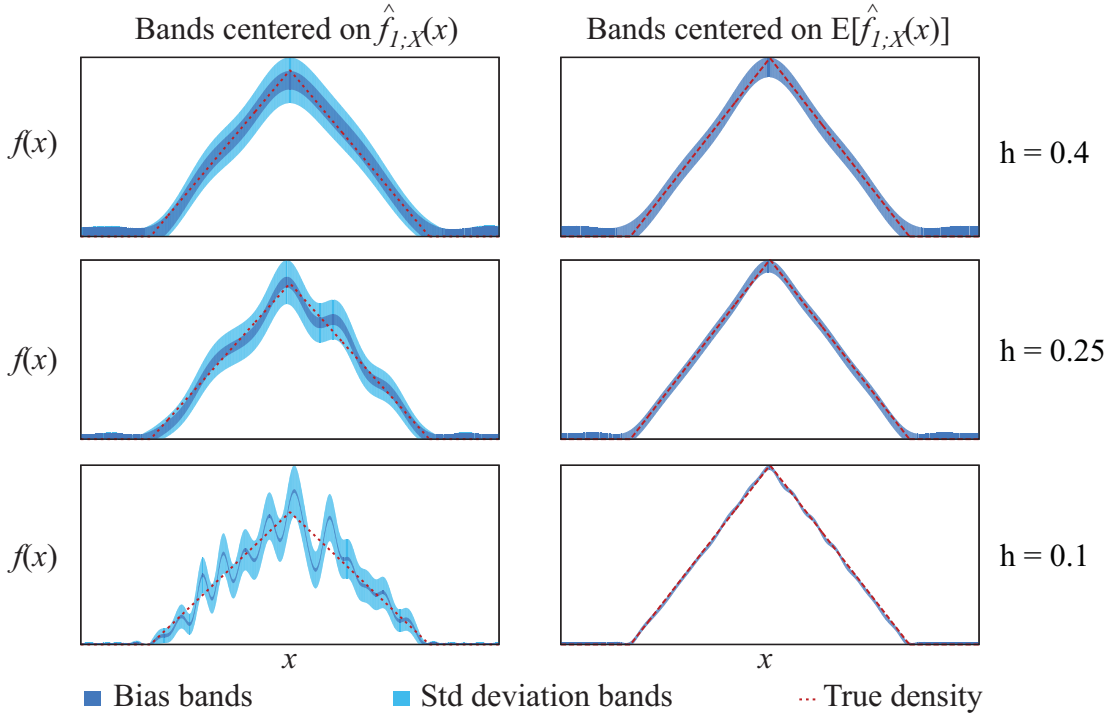

Figure 2: Example of nonparametric density estimation for a triangular density. Pointwise 95% confidence bands at various bandwidths (top: optimal, cross-validation minimizing, bandwidth; results two undersmoothed bandwidths shown below). The left column shows bands centered around a kernel estimate $\hat{f}_{1;X}(x)$ for one randomly generated sample while the right column show the bands centered on the expected value $E[\hat{f}_{1;X}(x)]$, calculated from 200 replications. The latter allows a more detailed study of the bias behavior, as the standard deviation bands are negligible.

contaminated by noise that can easily obscure clear trends in the data.[8] We do not explore larger-than-optimal bandwidths here because our theory does not guarantee that such bands guarantee a given coverage level (the estimation error in the bias estimator may then not be negligible relative to the standard deviation).

Figure 3 repeats a similar exercise for a smoother density: The convolution of three uniform distributions on $[0, 1]$. The results again confirm that the confidence bands are valid and that the bias-only bands correctly bound the bias. Note that in both cases, our data-driven rule for automatically determining the exponent of the power law decay of the Fourier transform was used and yields the correct value ($r = 2$ for the triangular case and $r = 3$ for the triple-convolution of a uniform density).

We can also investigate how the proposed method performs on real data. Here, we cannot

[8]It should be noted that we are not claiming any improvement in the accuracy of the level. Traditional undersmoothed bands already have asymptotically exact coverage. Bands that use bias bounds have an asymptotically conservative coverage. In spite of this, bias bound bands still asymptotically guarantee a given level of confidence and are asymptotically infinitely shorter than undersmoothed bands.

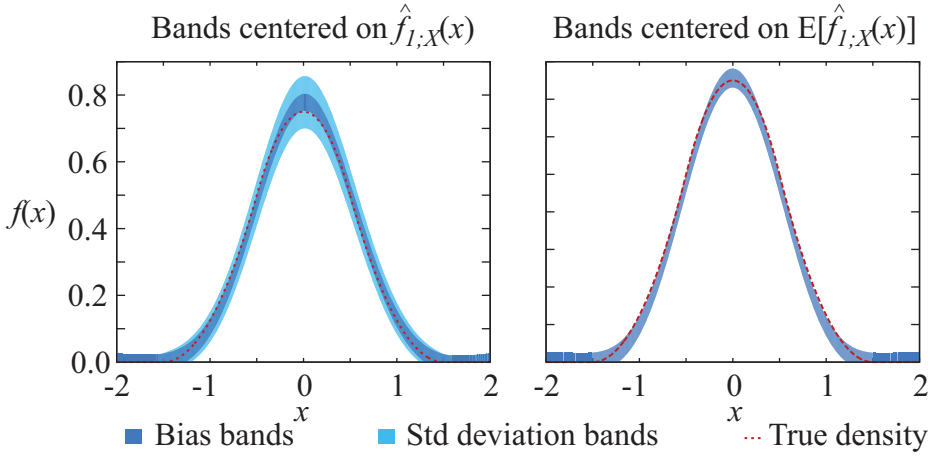

Figure 3: Example of nonparametric density estimation for a a twice differentiable density. Pointwise 95% confidence bands at the optimal, cross-validation minimizing, bandwidth are shown on the left while bias bands around the extpected value of the kernel estimator are shown on the right.

test the validity of the bands, but we can assess their width and nonoscillatory behavior. We employ data from the widely used Current Population Survey (CPS)/Social Security Earnings (SER) exact match file from March 1978. In Figure 4, we show a nonparametric estimate of the log income distribution of a subsample ($n = 551$) of women from the CPS. Avoiding undersmoothing clearly yields much narrower bands that are also qualitatively more plausible.[9]

6 Conclusion

We propose a simple and practical approach to perform nonparametric inference in kernel density and conditional mean estimation that avoids the traditional dilemma between efficient estimation at the optimal mean-square-error-minimizing bandwidth and valid inference at a suboptimal undersmoothed bandwidth. We achieve this by deriving an upper bound on the bias that can be consistently estimated under primitive assumptions and by accounting for this bias bound in the construction of confidence bands. The bias bound estimator is obtained via a combination of a Fourier representation of the bias, powerful results from the theory of dynamical systems and an asymptotically justified extrapolation procedure that infers the unknown high-frequency worst-case behavior of the function to be estimated from its observable low-frequency behavior.

[9]Note that the distribution is nonsmooth at the upper end because the high-income end of the sample is truncated for privacy reasons in this data set. However, as the truncation is done based on *exact* income, the distribution of *mismeasured* income plotted here drops to zero gradually rather than discontinuously.

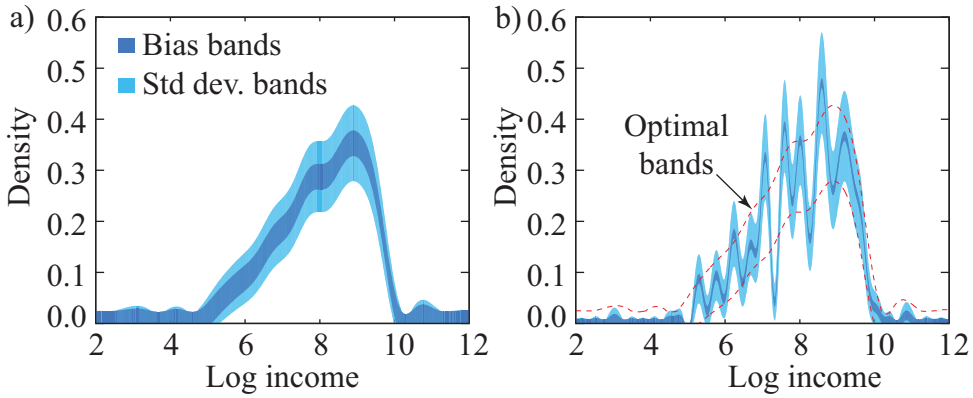

Figure 4: Example of application to the estimation of log income density. a) Bands at the optimal (cross-validation-score-minimizing) bandwidth. b) Bands for an undersmoothing bandwidth (optimal bands from a) are overlaid in dashed lines).

References

Armstrong, T. B. (2014): "Adaptive Testing on a Regression Function at a Point," Working Paper, Yale University.

Armstrong, T. B., and M. Kolesár (2014): "A Simple Adjustment for Bandwidth Snooping," Working Paper, Yale University.

Birkhoff, G. D. (1931a): "Proof of a recurrence theorem for strongly transitive systems," *Proc. Nat. Acad. Sci.*, 17, 650–655.

Birkhoff, G. D. (1931b): "Proof of the ergodic theorem," *Proc. Nat. Acad. Sci.*, 17, 656–660.

Cai, B. T. T., M. G. Low, and Y. Xia (2013): "Adaptive confidence intervals for regression Functions under shape constraints," *Annals of Statistics*, 41, 722–750.

Cai, T. T., M. Low, and Z. Ma (2014): "Adaptive Confidence Bands for Nonparametric Regression Functions," *Journal of the American Statistical Association*, 109, 1054–1070.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2013): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Estimation," Working Paper, University of Michigan.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2015): "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, forthcoming.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Anti-concentration and honest, adaptive confidence bands," *Annals of Statistics*, 42, 1787–1818.

Giné, E., and R. Nickl (2010): "Confidence bands in density estimation," *Annals of Statistics*, 38, 1122–1170.

Hall, P., and J. Horowitz (2013): "A simple bootstrap method for constructing nonparametric confidence bands for functions," *Annals of Statistics*, 41, 1892–1921.

Hansen, B. E. (2014): "Robust Inference," Working Paper, University of Wisconsin.

Härdle, W., and O. Linton (1994): "Applied Nonparametric Methods," in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden, vol. IV. Elsevier Science.

Hardle, W., and T. Stoker (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.

Hoffmann, M., and R. Nickl (2011): "On adaptive inference and confidence bands," *Annals of Statistics*, 39, 2383–2409.

Horowitz, J. (2009): *Semiparametric and Nonparametric Methods in Econometrics*. Springer.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B, chap. 74. Elsevier Science.

Lewbel, A., and O. Linton (2002): "Nonparametric Censored and Truncated Regression," *Econometrica*, 70, 765–779.

Li, Q., and J. Racine (2007): *Nonparametric Econometrics*. Princeton University Press.

Loève, M. (1977): *Probability Theory I*. New York: Springer.

Low, M. G. (1997): "On nonparametric confidence intervals," *Annals of Statistics*, 25, 2547–2554.

Mandelbrot, B. B. (1982): *The Fractal Geometry of Nature*. W. H. Freeman and Company, New York.

Pagan, A., and A. Ullah (1999): *Nonparametric Econometrics*. Cambridge University Press, Cambridge, UK.

Petersen, K. (1983): *Ergodic Theory*. Cambridge Studies in Advanced Mathematics, Cambridge University Press.

PINKSE, J. (2001): "Nonparametric Regression Estimation using Weak Separability," Working Paper, Penn State.

POLITIS, D. N. (2003): "Adaptive bandwidth choice," *J. Nonparam. Statist.*, 15, 517–533.

POLITIS, D. N., AND J. P. ROMANO (1999): "Multivariate Density Estimation with General Flat-Top Kernels of Infinite Order," *Journal of Multivariate Analysis*, 68, 1.

SCHENNACH, S. M. (2004): "Nonparametric Estimation in the Presence of Measurement Error," *Econometric Theory*, 20, 1046–1093.

STONE, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *Annals of Statistics*, 8, 1348–1360.

STONE, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals of Statistics*, 10, 1040–1053.

SU, L., AND A. ULLAH (2008): "Local polynomial estimation of nonparametric simultaneous equations models," *Journal of Econometrics*, 144, 193–218.

A Proofs

Proof of Lemma 1. First, we have

$$|\phi(\xi)| = \left| \int e^{i\xi x} f(x)\, dx \right| \leq \int \left| e^{i\xi x} \right| |f(x)|\, dx = \int |f(x)|\, dx = B$$

and this bound is reached whenever $f(x)$ is nonnegative.

To show that $|\phi(\xi)| \leq A|\xi|^{-r}$, note that, by integration by parts:

$$\phi(\xi) = \int e^{i\xi x} f(x)\, dx = \frac{1}{(-i\xi)} \int e^{i\xi x} df(x)$$

where the boundary terms vanish by requirement (ii) of the definition of $\mathcal{F}^r_{A,B}$ and where we have written the result as a Stieltjes integral. Repeating the process $(r-1)$ more times, we have

$$\phi(\xi) = \frac{1}{(-i\xi)^r} \int e^{i\xi x} df^{(r-1)}(x)$$

If $f^{(r-1)}(x)$ has bounded variation A then $f^{(r-1)}(x)$ can be written as $f^{(r-1)}_{\uparrow}(x) - f^{(r-1)}_{\downarrow}(x)$ where $f^{(r-1)}_{\uparrow}(x)$ and $f^{(r-1)}_{\downarrow}(x)$ are both increasing and such that $\lim_{x\to\infty} f^{(r-1)}_{\uparrow}(x) - \lim_{x\to-\infty} f^{(r-1)}_{\uparrow}(x) +$

$\lim_{x\to\infty} f_{\downarrow}^{(r-1)}(x) - \lim_{x\to-\infty} f_{\downarrow}^{(r-1)}(x) = A$. We can then write:

$$\begin{aligned}
|\xi|^r |\phi(\xi)| &= \left| \int e^{i\xi x} df^{(r-1)}(x) \right| \\
&= \left| \int e^{i\xi x} df_{\uparrow}^{(r-1)}(x) - \int e^{i\xi x} df_{\downarrow}^{(r-1)}(x) \right| \\
&\leq \int \left| e^{i\xi x} \right| df_{\uparrow}^{(r-1)}(x) + \int \left| e^{i\xi x} \right| df_{\downarrow}^{(r-1)}(x) \\
&\leq \int df_{\uparrow}^{(r-1)}(x) + \int df_{\downarrow}^{(r-1)}(x) \\
&= \lim_{x\to\infty} f_{\uparrow}^{(r-1)}(x) - \lim_{x\to-\infty} f_{\uparrow}^{(r-1)}(x) + \lim_{x\to\infty} f_{\downarrow}^{(r-1)}(x) - \lim_{x\to-\infty} f_{\downarrow}^{(r-1)}(x) = A
\end{aligned}$$

or $|\phi(\xi)| \leq A|\xi|^{-r}$.

Moreover, this upper bound is reached at all $\xi \in \mathbb{R}$ for some sequence $f_m(x)$ in $\mathcal{F}_{A,B}^r$ $(r > 0)$ with:

$$f_m(x) = h(x) g_m(x)$$

where

$$\begin{aligned}
h(x) &= \frac{A}{2\pi} \frac{x^{r-1}}{(r-1)!} \frac{1}{2} \operatorname{sgn}(x) \\
g_m(x) &= g(x/m).
\end{aligned}$$

where $g(x)$ is the inverse Fourier transform of some function $\gamma(\xi)$ satisfying the following: it is compactly supported, infinitely many times differentiable, absolutely integrable and such that $\int \gamma(\xi) d\xi = 1$. Note that the Fourier transform of $g_m(x)$ is $m\gamma(m\xi) \equiv \gamma_m(\xi)$.

Let $\eta(\xi)$ be the Fourier transform of $h(x)$, given by the moment theorem:

$$\eta(\xi) = \frac{i^{(r-1)}}{(r-1)!} \frac{d^{r-1}}{d\xi^{r-1}} \frac{A}{i\xi} = \frac{i^{(r-1)}}{(r-1)!} \frac{d^{r-1}}{d\xi^{r-1}} \frac{A}{i\xi} = A(i\xi)^{-r}.$$

Next, by the convolution theorem,

$$\phi_m(\xi) \equiv \int f_m(x) e^{i\xi x} dx = [\eta \otimes \gamma_m](\xi)$$

which is a standard convolution with a compactly supported kernel of shrinking width m^{-1} and thus, for any given $\xi \neq 0$, $\lim_{m\to\infty} [\eta \otimes \gamma_m](\xi) = \eta(\xi) = A(i\xi)^{-r}$ (note that for m sufficiently large, the support of $\gamma_m(\cdot - \xi)$ eventually excludes the origin, so the convolution integral is eventually finite).

■

Proof of Lemma 2. The exact (but unknown) bias can be bounded as

$$\begin{aligned}
\left|f_{Y;X}(x)-\bar{f}_{Y;X}(x)\right| &= \left|\frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_{Y;X}(\xi)\,e^{-i\xi x}d\xi-\frac{1}{2\pi}\int_{-\infty}^{\infty}\kappa(h\xi)\,\phi_{Y;X}(\xi)\,e^{-i\xi x}d\xi\right| \\
&= \frac{1}{2\pi}\left|\int_{-\infty}^{\infty}(1-\kappa(h\xi))\,\phi_{Y;X}(\xi)\,e^{-i\xi x}d\xi\right| \\
&\leq \frac{1}{2\pi}\int_{-\infty}^{\infty}|1-\kappa(h\xi)|\left|\phi_{Y;X}(\xi)\right|d\xi \\
&\leq \frac{1}{2\pi}\int_{-\infty}^{\infty}|1-\kappa(h\xi)|\min\left\{A_{Y;X}|\xi|^{-r_{Y;X}},B_{Y;X}\right\}d\xi.
\end{aligned}$$

by Lemma 1 and the first conclusion of the theorem follows.

Next, also by Lemma 1, the above bounds are actually least upper bounds and the second conclusion of the theorem follows.

Let κ_* denote the smallest positive ξ such that $\kappa(\xi)\neq 1$. We then have,

$$\begin{aligned}
b_{Y;X} &= 2\int_{0}^{\infty}|1-\kappa(h\xi)|\min\left\{A_{Y;X}|\xi|^{-r_{Y;X}},B_{Y;X}\right\}d\xi \\
&= 2\int_{\kappa_* h^{-1}}^{\infty}|1-\kappa(h\xi)|\min\left\{A_{Y;X}|\xi|^{-r_{Y;X}},B_{Y;X}\right\}d\xi \\
&\leq 2(\bar{\kappa}+1)\int_{\kappa_* h^{-1}}^{\infty}\min\left\{A_{Y;X}|\xi|^{-r_{Y;X}},B_{Y;X}\right\}d\xi \\
&= 2(\bar{\kappa}+1)\int_{\kappa_* h^{-1}}^{\infty}A_{Y;X}|\xi|^{-r_{Y;X}}\,d\xi \text{ for all } h \text{ sufficiently small} \\
&= 2(\bar{\kappa}+1)A_{Y;X}\left[\frac{|\xi|^{-r_{Y;X}+1}}{-r_{Y;X}+1}\right]_{\kappa_* h^{-1}}^{\infty} \\
&= 2(\bar{\kappa}+1)\frac{A_{Y;X}}{1-r_{Y;X}}h^{r_{Y;X}-1}
\end{aligned}$$

for $r_{Y;X}>1$. ■

Proof of Lemma 3. A ratio is always made larger by increasing the numerator. If the numerator is positive, decreasing a positive denominator (without going below zero) makes the ratio larger. If the numerator is negative, increasing a positive denominator also makes the ratio larger. Note that as soon as $\bar{f}_{1;X}(x)-b_{1;X}\leq 0$, there is a possibility that the true density in the denominator vanishes ($f_{1;X}(x)=0$), in which case the ratio $f_{Y;X}(x)/f_{Y;1}(x)$ could be arbitrarily large in magnitude, hence the infinite upper (lower) bounds if the numerator is positive (negative). These considerations lead to the expression for the bias. ■

Proof of Theorem 1. The maximum error on the empirical Fourier transform $\hat{\phi}_{Y;X}(\xi)$ satisfies $\sup_{\xi\in[0,\bar{\xi}_n]}\left|\hat{\phi}_{Y;X}(\xi)-\phi_{Y;X}(\xi)\right|\leq\Delta\phi_{Y;X,n}$ (for $\bar{\xi}_n=O\left(n^{1/4}\right)$) almost surely as

$n \to \infty$, by Lemma 4 below, a result we will apply repeatedly below. (We abbreviate $\hat{\phi}_{Y;X}(\xi)$ by $\hat{\phi}(\xi)$, etc.) Let

$$\begin{aligned}\hat{Q}_n(\hat{r}) &= \int_{\ln \underline{\xi}}^{\ln \bar{\xi}_n} \left(\ln \hat{A} - \hat{r}\lambda\right) d\lambda \\ Q_n &= \int_{\ln \underline{\xi}}^{\ln \bar{\xi}_n} (\ln A - r\lambda)\, d\lambda\end{aligned}$$

We have

$$\begin{aligned}\hat{Q}_n(\hat{r}) &= \left(\ln \bar{\xi}_n - \ln \underline{\xi}\right) \ln \hat{A} - \frac{\hat{r}}{2}\left(\left(\ln \bar{\xi}_n\right)^2 - \left(\ln \underline{\xi}\right)^2\right) \\ &= \left(\ln \bar{\xi}_n - \ln \underline{\xi}\right)\left(\ln \hat{A} - \frac{\hat{r}}{2}\left(\ln \bar{\xi}_n + \ln \underline{\xi}\right)\right)\end{aligned}$$

and similarly

$$Q_n = \left(\ln \bar{\xi}_n - \ln \underline{\xi}\right)\left(\ln A - \frac{r}{2}\left(\ln \bar{\xi}_n + \ln \underline{\xi}\right)\right).$$

Let $\hat{\xi}_n = \min \arg\max_{\xi \in \left[\underline{\xi}, \bar{\xi}_n\right]} |\xi|^{\hat{r}} \left|\hat{\phi}(\xi)\right|$ (that is, the point that is first just touching the line $\hat{A} - \hat{r} \ln \xi$). We then have:

$$\begin{aligned}\ln \hat{A} - \hat{r} \ln \hat{\xi}_n &= \ln \left|\hat{\phi}\left(\hat{\xi}_n\right)\right| \\ \ln \hat{A} &= \hat{r} \ln \hat{\xi}_n + \ln \left|\hat{\phi}\left(\hat{\xi}_n\right)\right|.\end{aligned}$$

Now, we consider

$$\begin{aligned}&\hat{Q}_n(\hat{r}) - Q_n \\ =\ & \left(\ln \bar{\xi}_n - \ln \underline{\xi}\right) \ln \hat{A} - \frac{\hat{r}}{2}\left(\left(\ln \bar{\xi}_n\right)^2 - \left(\ln \underline{\xi}\right)^2\right) - \left(\left(\ln \bar{\xi}_n - \ln \underline{\xi}\right) \ln A - \frac{r}{2}\left(\left(\ln \bar{\xi}_n\right)^2 - \left(\ln \underline{\xi}\right)^2\right)\right) \\ =\ & \left(\ln \bar{\xi}_n - \ln \underline{\xi}\right)\left(\ln \hat{A} - \ln A\right) + \left(\frac{r}{2} - \frac{\hat{r}}{2}\right)\left(\left(\ln \bar{\xi}_n\right)^2 - \left(\ln \underline{\xi}\right)^2\right) \\ =\ & \left(\ln \bar{\xi}_n - \ln \underline{\xi}\right)\left(\hat{r} \ln \hat{\xi}_n + \ln \left|\hat{\phi}\left(\hat{\xi}_n\right)\right| - \ln A\right) + \left(\frac{r}{2} - \frac{\hat{r}}{2}\right)\left(\left(\ln \bar{\xi}_n\right)^2 - \left(\ln \underline{\xi}\right)^2\right) \\ =\ & \left(\ln \bar{\xi}_n - \ln \underline{\xi}\right)\left(\hat{r} \ln \hat{\xi}_n + \ln \left|\hat{\phi}\left(\hat{\xi}_n\right)\right| - \ln A + \left(\frac{r}{2} - \frac{\hat{r}}{2}\right)\left(\ln \bar{\xi}_n + \ln \underline{\xi}\right)\right)\end{aligned}$$

or

$$\frac{\hat{Q}_n(\hat{r}) - Q_n}{\ln \bar{\xi}_n - \ln \underline{\xi}} = \hat{r} \ln \hat{\xi}_n + \ln \left|\hat{\phi}\left(\hat{\xi}_n\right)\right| - \ln A + \left(\frac{r}{2} - \frac{\hat{r}}{2}\right)\left(\ln \bar{\xi}_n + \ln \underline{\xi}\right)$$

If $\hat{r} < r$ then $\hat{\xi}_n = \hat{\xi}_{n^*}$ for some fixed n^* for all n sufficiently large. Then,

$$\frac{\hat{Q}_n(\hat{r}) - Q_n}{\ln \bar{\xi}_n - \ln \underline{\xi}} = \hat{r} \ln \hat{\xi}_{n^*} + \ln \left|\hat{\phi}\left(\hat{\xi}_{n^*}\right)\right| - \ln A + \left(\frac{r}{2} - \frac{\hat{r}}{2}\right)\left(\ln \bar{\xi}_n + \ln \underline{\xi}\right)$$

where then term $\hat{r}\ln\hat{\xi}_{n^*}+\ln\left|\hat{\phi}\left(\hat{\xi}_{n^*}\right)\right|-\ln A$ does not change with n (for n sufficiently large). Since $r-\hat{r}>0$ and $\bar{\xi}_n\to\infty$, we have that $\left(\hat{r}\ln\hat{\xi}_{n^*}+\ln\left|\hat{\phi}\left(\hat{\xi}_{n^*}\right)\right|-\ln A+\frac{1}{2}\left(r-\hat{r}\right)\left(\ln\bar{\xi}_n+\ln\underline{\xi}\right)\right)\to\infty$, and $\left(\hat{Q}_n\left(\hat{r}\right)-Q_n\right)/\left(\ln\bar{\xi}_n-\ln\underline{\xi}\right)\to\infty$ almost surely.

If $\hat{r}>r$ then $\hat{\xi}_n\to\infty$ and we have

$$\begin{aligned}
&\frac{\hat{Q}\left(\hat{r}\right)-Q_n\left(r\right)}{\ln\bar{\xi}_n-\ln\underline{\xi}}\\
=\;&\hat{r}\ln\hat{\xi}_n+\ln\left|\hat{\phi}\left(\hat{\xi}_n\right)\right|-\ln A+\left(\frac{r}{2}-\frac{\hat{r}}{2}\right)\left(\ln\bar{\xi}_n+\ln\underline{\xi}\right)\\
\geq\;&\hat{r}\ln\hat{\xi}_n+\ln\left(\left|\phi\left(\hat{\xi}_n\right)\right|-\Delta\phi_{Y;X,n}\right)-\ln A+\left(\frac{r}{2}-\frac{\hat{r}}{2}\right)\left(\ln\bar{\xi}_n+\ln\underline{\xi}\right)
\end{aligned}$$

Combining Lemmas 5, 6 and 7 (stated below) shows that $|\xi|^r\phi\left(\xi\right)$ must come within a small ε of its maximum value of A infinitely often, hence we eventually have that $\left|\phi\left(\hat{\xi}_n\right)\right|\geq\rho A\hat{\xi}_n^{-r}$ for some $\rho\in]0.5,1[$ and thus:

$$\begin{aligned}
\frac{\hat{Q}\left(\hat{r}\right)-Q_n\left(r\right)}{\ln\bar{\xi}_n-\ln\underline{\xi}}&\geq\hat{r}\ln\hat{\xi}_n+\ln\left(\rho A\hat{\xi}_n^{-r}-\Delta\phi_{Y;X,n}\right)-\ln A+\left(\frac{r}{2}-\frac{\hat{r}}{2}\right)\left(\ln\bar{\xi}_n+\ln\underline{\xi}\right)\\
&\geq\hat{r}\ln\hat{\xi}_n+\ln\left(\rho A\hat{\xi}_n^{-r}\right)+\ln\left(1-2\frac{\Delta\phi_{Y;X,n}}{A\hat{\xi}_n^{-r}}\right)-\ln A+\left(\frac{r}{2}-\frac{\hat{r}}{2}\right)\left(\ln\bar{\xi}_n+\ln\underline{\xi}\right)
\end{aligned}$$

Since $\bar{\xi}_n^r\Delta\phi_{Y;X,n}\to0$ by assumption, we have $\ln\left(1-2\frac{\Delta\phi_{Y;X,n}}{A\hat{\xi}_n^{-r}}\right)\leq\delta$ and for all n sufficiently large and it follows that

$$\begin{aligned}
\frac{\hat{Q}\left(\hat{r}\right)-Q_n\left(r\right)}{\ln\bar{\xi}_n-\ln\underline{\xi}}&\geq\hat{r}\ln\hat{\xi}_n+\ln\left(\rho A\hat{\xi}_n^{-r}\right)+\delta-\ln A+\left(\frac{r}{2}-\frac{\hat{r}}{2}\right)\left(\ln\bar{\xi}_n+\ln\underline{\xi}\right)\\
&=\left(\hat{r}-r\right)\ln\hat{\xi}_n+\ln\left(\rho A\right)-\ln A+\delta+\left(\frac{r}{2}-\frac{\hat{r}}{2}\right)\left(\ln\bar{\xi}_n+\ln\underline{\xi}\right)\\
&=\left(\hat{r}-r\right)\left(\ln\hat{\xi}_n-\frac{1}{2}\left(\ln\bar{\xi}_n+\ln\underline{\xi}\right)\right)+\ln\rho+\delta
\end{aligned}$$

where $\left(\hat{r}-r\right)>0$ and we show below that $\left(\ln\hat{\xi}_n-\frac{1}{2}\left(\ln\bar{\xi}_n+\ln\underline{\xi}\right)\right)\to\infty$.

When $\hat{r}>r$, the location of $\hat{\xi}_n$ remains constant over a range of values of n and then jumps up whenever the interval $\left[\underline{\xi},\bar{\xi}_n\right]$ includes a new value of $\hat{\xi}_n$. To capture this, let $\hat{\xi}_n\equiv\xi_{k_n}^*$ where the sequence ξ_k^* is strictly increasing in k and k_n satisfies $k_{n+1}-k_n\in\{0,1\}$. By combining Lemmas 5, 6 and 7 (below), we can find a bound on the distance between consecutive ξ_k^*. Lemma 5 shows that, for $f\in\mathcal{F}_{A,B}^r$, the function $f^{(r-1)}\left(x\right)$ can be decomposed as a sum of an absolutely continuous function and a finite sum of step functions. The Fourier

transform of a sum of J step functions has the form $(i\xi)^{-1}\sum_{j=1}^{J}A_j e^{i\xi x_j}$ for $A_j, x_j \in \mathbb{R}$. By Lemma 6, the Fourier transform of the absolutely continuous part is $o\left(|\xi|^{-1}\right)$, which is asymptotically negligible relative the Fourier transform of the step functions. These results imply that the Fourier transform of $f(x)$ has the form $\phi(\xi) = (i\xi)^{-r}\sum_{j=1}^{J}A_j e^{i\xi x_j} + o\left(|\xi|^{-r}\right)$. Lemma 7 then shows that $\sum_{j=1}^{J}A_j e^{i\xi x_j}$ reaches within a arbitrarily small ε of its maximum maximum value at quasi-periodic intervals. Specifically, it shows that:

$$\begin{aligned}
\left|\xi_{k+1}^{*} - \xi_{k}^{*}\right| &\leq \frac{1}{3}\left|\xi_{k}^{*}\right| \\
\xi_{k+1}^{*} &\leq \frac{4}{3}\xi_{k}^{*} \\
\ln \xi_{k+1}^{*} &\leq \ln\frac{4}{3} + \ln \xi_{k}^{*}.
\end{aligned}$$

Since $k_{n+1} - k_n \in \{0,1\}$, we have $k_{n+1} \leq k_n + 1$. Hence, $\hat{\xi}_{n+1} = \xi_{(k_{n+1})}^{*} \leq \xi_{(k_n)+1}^{*} \leq \ln\frac{4}{3} + \ln \xi_{k_n}^{*} = \ln\frac{4}{3} + \ln\hat{\xi}_n$ and we also have:

$$\begin{aligned}
\left(\ln\hat{\xi}_n - \frac{1}{2}\left(\ln\bar{\xi}_n + \ln\underline{\xi}\right)\right) &\geq \left(\ln\hat{\xi}_n - \frac{1}{2}\left(\ln\hat{\xi}_{n+1} + \ln\underline{\xi}\right)\right) \\
&\geq \left(\ln\hat{\xi}_n - \frac{1}{2}\left(\ln\frac{4}{3} + \ln\hat{\xi}_n + \ln\underline{\xi}\right)\right) \\
&= \left(\ln\hat{\xi}_n - \frac{1}{2}\ln\hat{\xi}_n - \frac{1}{2}\ln\underline{\xi} - \frac{1}{2}\ln\frac{4}{3}\right) \\
&= \left(\frac{1}{2}\ln\hat{\xi}_n - \frac{1}{2}\ln\underline{\xi} - \frac{1}{2}\ln\frac{4}{3}\right)
\end{aligned}$$

where $\hat{\xi}_n \to \infty$ and all other terms are constant. It follow that $\left(\hat{Q}_n(\hat{r}) - Q_n\right) / \left(\ln\bar{\xi}_n - \ln\underline{\xi}\right) \to \infty$ almost surely.

We now show that for $\hat{r} = r$, $\left(\hat{Q}_n(\hat{r}) - Q_n\right) / \left(\ln\bar{\xi}_n - \ln\underline{\xi}\right)$ remains bounded, since it

reduces to:

$$\begin{aligned}
\frac{\hat{Q}(r) - Q_n(r)}{\ln \bar{\xi}_n - \ln \underline{\xi}} &= r \ln \hat{\xi}_n + \ln \left| \hat{\phi}\left(\hat{\xi}_n\right) \right| - \ln A \\
&\leq r \ln \hat{\xi}_n + \ln \left| \phi\left(\hat{\xi}_n\right) + \Delta\phi_{Y;X,n} \right| - \ln A \\
&\leq r \ln \hat{\xi}_n + \ln \left| A\hat{\xi}_n^{-r} + \Delta\phi_{Y;X,n} \right| - \ln A \\
&= r \ln \hat{\xi}_n + \ln A\hat{\xi}_n^{-r} + \ln \left| 1 + \frac{\Delta\phi_{Y;X,n}}{\rho A\hat{\xi}_n^{-r}} \right| - \ln A \\
&= r \ln \hat{\xi}_n - r \ln \hat{\xi}_n + \ln A + \ln \left| 1 + \frac{\Delta\phi_{Y;X,n}}{\rho A\hat{\xi}_n^{-r}} \right| - \ln A \\
&= \ln \left| 1 + \frac{\Delta\phi_{Y;X,n}}{\rho A\hat{\xi}_n^{-r}} \right| \to 0
\end{aligned}$$

almost surely.

Since for both $\hat{r} > r$ and $\hat{r} < r$, we have that $\hat{Q}_n(\hat{r}) - Q_n$ diverges faster than for $\hat{r} = r$, it follows that $\hat{r} = r$ with probability approaching one.

Next, for a given r (and in particular for the true value of r), we have

$$\hat{A}_{Y;X}(r) = \sup_{\xi \in \left[\underline{\xi}, \bar{\xi}_n\right]} \left| \hat{\phi}_{Y;X}(\xi) \right| \xi^r$$

and we also define

$$A^n_{Y;X}(r) = \sup_{\xi \in \left[\underline{\xi}, \bar{\xi}_n\right]} \left| \phi_{Y;X}(\xi) \right| \xi^r.$$

We then have

$$\left| \hat{A}_{Y;X}(\hat{r}) - A_{Y;X} \right| \leq \left| \hat{A}_{Y;X}(\hat{r}) - \hat{A}_{Y;X}(r) \right| + \left| \hat{A}_{Y;X}(r) - A^n_{Y;X}(r) \right| + \left| A^n_{Y;X}(r) - A_{Y;X} \right|$$

where (i) $\left| \hat{A}_{Y;X}(\hat{r}) - \hat{A}_{Y;X}(r) \right| \xrightarrow{p} 0$ because $\hat{r} = r$ with probability approaching one, (ii) almost surely $\left| \hat{A}_{Y;X}(r) - A^n_{Y;X}(r) \right| \leq \Delta\phi_{Y;X,n}\xi^r \to 0$ and (iii) $\left| A^n_{Y;X}(r) - A_{Y;X} \right| \to 0$ because, by construction, $A^n_{Y;X}(r)$ is increasing in n and is bounded above by $\sup_{\xi \in \mathbb{R}} \left| \phi_{Y;X}(\xi) \right| |\xi|^r = A_{Y;X}$ which is finite if $f \in \mathcal{F}^r_{A;B}$, thus implying that $\lim_{n\to\infty} A^n_{Y;X}(r)$ exists and is equal to $A_{Y;X}$. ■

Lemma 4 *If (X_i, Y_i) is iid, $E[|Y_i X_i|] < \infty$ and $\left(E\left[Y_i^2\right]\right)^{1/2} \equiv \mu_Y < \infty$, then, for any given $\bar{\xi} \in \mathbb{R}^+$ and $\kappa \in \mathbb{R}^+$*

$$\sup_{|\xi| \leq \bar{\xi} n^\kappa} \left| n^{-1} \sum_{i=1}^n Y_i e^{\mathbf{i}\xi X_i} - E\left[Y_i e^{\mathbf{i}\xi X_i}\right] \right| \leq \frac{2\sqrt{2}}{3} (3 + 2\kappa) \mu_Y n^{-1/2} \ln n$$

almost surely as $n \to \infty$. *If, in addition,* $|Y_i|$ *is bounded, then*

$$\sup_{|\xi| \leq \bar{\xi} n^{\kappa}} \left| n^{-1} \sum_{i=1}^{n} Y_i e^{\mathbf{i}\xi X_i} - E\left[Y_i e^{\mathbf{i}\xi X_i}\right] \right| \leq 2\left(3+2\kappa\right)^{1/2} \mu_Y n^{-1/2} \left(\ln n\right)^{1/2}$$

almost surely as $n \to \infty$.

Proof. Let

$$\Delta\phi_{Y;X}\left(\xi\right) = n^{-1} \sum_{i=1}^{n} Y_i e^{\mathbf{i}\xi X_i} - E\left[Y_i e^{\mathbf{i}\xi X_i}\right]$$

and let $C_\kappa = \frac{2\sqrt{2}}{3}\left(3+2\kappa\right)\sigma_Y$ (in the unbounded Y case). The proof proceeds by bounding $\Delta\phi_{Y;X}\left(\xi\right)$ on a finite mesh of equidistant points $\mathcal{M}_n = \left\{0, \frac{\bar{\xi} n^{\kappa}}{N_n}, \frac{2\bar{\xi} n^{\kappa}}{N_n}, \ldots, \bar{\xi} n^{\kappa}\right\}$ (with N_n to be determined) and bounding $d\left(\Delta\phi_{Y;X}\left(\xi\right)\right)/d\xi$ uniformly to ensure uniform convergence everywhere:

$$\begin{aligned}
& P\left[\Delta\phi_{Y;X}\left(\xi\right) \leq C_\kappa n^{-1/2} \ln n \text{ for all } |\xi| \leq \bar{\xi} n^{\kappa}\right] \\
\leq\ & P\left[\Delta\phi_{Y;X}\left(\xi\right) \leq C_\kappa n^{-1/2} \ln n \text{ for all } \xi \in \mathcal{M}_n \text{ and } \left|d\left(\Delta\phi_{Y;X}\left(\xi\right)\right)/d\xi\right| \leq D \text{ for all } |\xi| \leq \bar{\xi} n^{\kappa}\right] \\
\leq\ & P\left[\Delta\phi_{Y;X}\left(\xi\right) \leq C_\kappa n^{-1/2} \ln n \text{ for all } \xi \in \mathcal{M}_n \text{ and } \left|\hat{E}\left[Y_i X_i\right]\right| \leq D \text{ for all } |\xi| \leq \bar{\xi} n^{\kappa}\right] \\
\leq\ & \sum_{m=0}^{N_n} P\left[\Delta\phi_{Y;X}\left(\frac{m\bar{\xi} n^{\kappa}}{N_n}\right) \leq C_\kappa n^{-1/2} \ln n\right] + P\left[\left|\hat{E}\left[Y_i X_i\right]\right| \leq D\right] \qquad (8)
\end{aligned}$$

for $D = E\left[Y_i X_i\right] + \varepsilon_D$ with $\varepsilon_D > 0$. In the above display, we have used (i) a standard device to show uniform convergence by bounding the derivative $d\left(\Delta\phi_{Y;X}\left(\xi\right)\right)/d\xi$ between the mesh points at which we bound the function's value (ii) the fact that $\left|\frac{d}{d\xi} E\left[Y_i e^{\mathbf{i}\xi X_i}\right]\right| = \left|E\left[Y_i X_i e^{\mathbf{i}\xi X_i}\right]\right| \leq E\left[|Y_i X_i|\right]$, provided the latter exists and (iii) a Bonferroni-type bound.

To be able to use Bernstein's inequality to handle the first term of (8), we condition on the event that all $Y_i e^{\mathbf{i}\xi X}$ are bounded. We observe that

$$\begin{aligned}
\max_{i \in \{1,\ldots,n\}} \left|Y_i e^{\mathbf{i}\xi X}\right| &= \max_{i \in \{1,\ldots,n\}} |Y_i| = \left(\max_{i \in \{1,\ldots,n\}} |Y_i|^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^{n} |Y_i|^2\right)^{1/2} = n^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} |Y_i|^2\right)^{1/2} \\
&\leq n^{1/2}\left(E\left[|Y_i|^2\right] + o_{as}\left(1\right)\right)^{1/2} \\
&\leq n^{1/2}\left(E\left[|Y_i|^2\right] + \varepsilon_2\right)^{1/2} \equiv \bar{Y}_n \qquad (9)
\end{aligned}$$

almost surely as $n \to \infty$ for any $\varepsilon_2 > 0$. To allow for the fact that $\Delta\phi_{Y;X}(\xi)$ is complex, we write:

$$\begin{aligned} P[|\phi| > \varepsilon] &= P\left[(\mathrm{Re}\,\phi)^2 + (\mathrm{Im}\,\phi)^2 > \varepsilon^2\right] \\ &\leq P\left[(\mathrm{Re}\,\phi)^2 > \varepsilon^2/2 \text{ or } (\mathrm{Im}\,\phi)^2 > \varepsilon^2/2\right] \\ &\leq P\left[(\mathrm{Re}\,\phi)^2 > \varepsilon^2/2\right] + P\left[(\mathrm{Im}\,\phi)^2 > \varepsilon^2/2\right] \\ &= P\left[\mathrm{Re}\,\phi > \varepsilon/\sqrt{2}\right] + P\left[\mathrm{Im}\,\phi > \varepsilon/\sqrt{2}\right] \end{aligned}$$

By Bernstein's inequality:

$$P[\mathrm{Re}\,\phi > c] \leq 2\exp\left(-\frac{1}{2}\frac{c^2/2}{nE[Y_i^2] + \bar{Y}_n c/3}\right)$$

with $\bar{Y}_n$ being the almost sure bound on $|Y_i|$ defined in (9) and similarly for $P\left[\mathrm{Im}\,\phi > \varepsilon/\sqrt{2}\right]$. Next,

$$\begin{aligned} P\left[\mathrm{Re}\,\phi > Cn^{1/2}\ln n\right] &\leq 2\exp\left(-\frac{1}{2}\frac{\left(Cn^{1/2}\ln n\right)^2/2}{n\sigma_Y^2 + n^{1/2}(\sigma_Y + \varepsilon_2)\left(Cn^{1/2}\ln n\right)/3}\right) \\ &= 2\exp\left(-\frac{1}{2}\frac{C^2 n(\ln n)^2/2}{n\sigma_Y^2 + (\sigma_Y + \varepsilon_2)Cn(\ln n)/3}\right) \\ &= 2\exp\left(-\frac{1}{2}\frac{3C\ln n}{\frac{6n\sigma_Y^2}{Cn\ln n} + 2(\sigma_Y + \varepsilon_2)}\right) \\ &= 2\exp\left(-\frac{1}{2}\frac{3C\ln n}{2(\sigma_Y + \varepsilon_2)}\frac{1}{\frac{6n\sigma_Y^2}{2C(\sigma_Y+\varepsilon_2)n\ln n} + 1}\right) \\ &= 2\exp\left(-\frac{3}{4}\frac{C\ln n}{\sigma_Y + \varepsilon_2}\frac{1}{\frac{3\sigma_Y^2}{C(\sigma_Y+\varepsilon_2)\ln n} + 1}\right) \end{aligned}$$

Let $C = z\sigma_Y$ (where z is to be determined so that the overall probability of exceeding the bound decreases). We then have:

$$\begin{aligned} P\left[\operatorname{Re}\phi > Cn^{-1/2}\ln n\right] &\leq 2\exp\left(-\frac{3}{4}\frac{z\sigma_Y \ln n}{\sigma_Y+\varepsilon_2}\frac{1}{\frac{3\sigma_Y^2}{z\sigma_Y(\sigma_Y+\varepsilon_2)\ln n}+1}\right) \\ &= 2\exp\left(-\frac{3}{4}\frac{z\ln n}{1+\varepsilon_2/\sigma_Y}\frac{1}{\frac{3}{z(1+\varepsilon_2/\sigma_Y)\ln n}+1}\right) \\ &= 2\exp\left(-\frac{3}{4}z\ln n\frac{1}{1+\frac{3}{z\ln n}+\varepsilon_2/\sigma_Y}\right) \\ &\leq 2\exp\left(-\frac{3}{4}z\left(1-\varepsilon_3\right)\ln n\right) \\ &= 2n^{-\frac{3}{4}(1-\varepsilon_3)z} \equiv p_n \end{aligned}$$

for some arbitrarily small $\varepsilon_3 > 0$ and sufficiently large n.

To determine the necessary number of mesh points N_n, we balance the maximum possible deviation from mesh point values $\left|\hat{\phi}\left(\xi\right) - \hat{\phi}\left(\frac{m\bar{\xi}n^\kappa}{N_n}\right)\right|$ and the pointwise error in $\hat{\phi}\left(\xi\right)$:

$$\left(E\left[|YX|\right] + o_p\left(1\right)\right)\frac{\bar{\xi}n^\kappa}{N_n} = z\sigma_Y n^{-1/2}\ln n.$$

We thus need

$$N_n = \frac{E\left[|YX|\right]\bar{\xi}n^\kappa}{z\sigma_Y n^{-1/2}\ln n} = \frac{\bar{\xi}E\left[|YX|\right]}{z\sigma_Y}\frac{n^{\kappa+1/2}}{\ln n}$$

The total probability that $\operatorname{Re}\phi > Cn^{-1/2}\ln n$ for one of the mesh point is thus bounded by a Bonferroni-type bound:

$$\begin{aligned} N_n p_n &= \frac{\bar{\xi}E\left[|YX|\right]}{z\sigma_Y}\frac{n^{\kappa+1/2}}{\ln n}2n^{-\frac{3}{4}(1-\varepsilon_3)z} \\ &= O\left(\frac{1}{\ln n}n^{\kappa+\frac{1}{2}-\frac{3}{4}(1-\varepsilon_3)z}\right) \end{aligned}$$

We thus require $\kappa + \frac{1}{2} - \frac{3}{4}\left(1-\varepsilon_3\right)z < -1$, leading to

$$\left(1-\varepsilon_3\right)z > 2 + \frac{4}{3}\kappa.$$

For bounded Y, we similarly have

$$\begin{aligned} P\left[\operatorname{Re}\phi > Cn^{1/2}\sqrt{\ln n}\right] &\leq 2\exp\left(-\frac{1}{2}\frac{\left(Cn^{1/2}\sqrt{\ln n}\right)^2/2}{n\sigma_Y^2 + \left(Cn^{1/2}\ln n\right)/3}\right) \\ &\leq 2\exp\left(-\frac{1}{2}\left(1-\varepsilon_4\right)\frac{C^2 n\left(\ln n\right)/2}{n\sigma_Y^2}\right) \\ &= 2\exp\left(-\frac{1}{2}\left(1-\varepsilon_4\right)\frac{z^2\sigma_Y^2 n\left(\ln n\right)/2}{n\sigma_Y^2}\right) \\ &= 2\exp\left(-\frac{1}{4}\left(1-\varepsilon_4\right)z^2\left(\ln n\right)\right) \\ &= 2n^{-\frac{1}{4}(1-\varepsilon_4)z^2} \end{aligned}$$

We thus require $\kappa + \frac{1}{2} - \frac{1}{4}\left(1-\varepsilon_4\right)z^2 < -1$, leading to

$$\left(1-\varepsilon_4\right)z^2 > 6 + 4\kappa.$$

■

Lemma 5 *For any function $f \in \mathcal{F}_{A,B}^r$, we have that $f^{(r-1)}(x)$ can be written as*

$$f^{(r-1)}(x) = s(x) + \sum_{j=1}^{J} A_j \mathbf{1}\left\{x \geq x_j\right\}$$

for some $J \in \mathbb{N}$, where $s(x)$ is absolutely continuous and $A_j \in \mathbb{R}$, and $1\{E\}$ denotes an indicator function of the event E. Also, the total variation of $f^{(r-1)}$ is given by

$$TV\left(f^{(r-1)}\right) = \int_{-\infty}^{\infty} \left|s(x)\right| dx + \sum_{j=1}^{J} \left|A_j\right|.$$

Proof. Let $x_1 < x_2 < \ldots < x_J$ be the points where $f^{(r)}(x)$ does not exist (and let $x_0 = -\infty$ and $x_{J+1} = \infty$, understood as appropriate limits). By the absolute continuity property, $f^{(r-1)}(x)$ is equal to the integral of its derivative for $x \notin \{x_1, \ldots, x_J\}$. Thus, for

$x \in]x_j, x_{j+1}[$, we have

$$\begin{aligned}
f^{(r-1)}(x) &= \lim_{u\to x_j^+} f^{(r-1)}(u) + \lim_{v\to x_j^+} \int_v^x f^{(r)}(u)\,du \\
&= \lim_{u\to x_j^-} f^{(r-1)}(u) - \lim_{u\to x_j^-} f^{(r-1)}(u) + \lim_{u\to x_j^+} f^{(r-1)}(u) + \lim_{v\to x_j^+} \int_v^x f^{(r)}(u)\,du \\
&= \left(\lim_{u\to x_{j-1}^+} f^{(r-1)}(u) + \lim_{\varepsilon_1,\varepsilon_2\to 0} \int_{x_{j-1}+\varepsilon_1}^{x_j-\varepsilon_2} f^{(r)}(u)\,du \right) \\
&\quad + \lim_{u\to x_j^+} f^{(r-1)}(u) - \lim_{u\to x_j^-} f^{(r-1)}(u) + \lim_{v\to x_j^+} \int_v^x f^{(r)}(u)\,du \\
&= \lim_{u\to x_{j-1}^+} f^{(r-1)}(u) + A_j + \int_{x_{j-1}^+}^{x_j^-} f^{(r)}(u)\,du + \int_{x_j^+}^x f^{(r)}(u)\,du \\
&= \lim_{u\to -\infty} f^{(r-1)}(u) + \sum_{k=1}^{j} A_k + \sum_{k=1}^{j} \int_{x_{k-1}^+}^{x_k^-} f^{(r)}(u)\,du + \int_{x_j^+}^x f^{(r)}(u)\,du \\
&= 0 + \sum_{k=1}^{J} A_k 1(x \geq x_k) + \int_{]-\infty,x]\setminus\{x_1,\ldots,x_J\}} f^{(r)}(u)\,du \\
&= \sum_{k=1}^{J} A_k 1(x \geq x_k) + s(x)
\end{aligned}$$

where $A_j = \lim_{u\to x_j^+} f^{(r-1)}(u) - \lim_{u\to x_j^-} f^{(r-1)}(u)$ and $s(x) = \int_{]-\infty,x]\setminus\{x_1,\ldots,x_J\}} f^{(r)}(u)\,du$. The bounded total variation assumption ensures that all these quantities are finite. ■

Remark 1 *This Lemma could also be shown using the Lebesgue decomposition theorem (see, for instance, Loève (1977)): Since the function $f^{(r-1)}(x)$ is of bounded variation, it can be written as the difference of two nondecreasing functions (for which the Lebesgue decomposition theorem applies). It follows that the measure $df^{(r-1)}(x)$ can be written as the sum of absolutely continuous, purely discrete and singular components. By assumption, $f^{(r)}(x)$ exists everywhere except at a finite number of points, so the decomposition reduces to the absolutely continuous component (corresponding to $s(x)$) and a purely discrete component (corresponding to $\sum_{k=1}^{J} A_k 1(x \geq x_k)$).*

Lemma 6 *The Fourier transform $\phi(\xi)$ of an absolutely continuous function $f(x)$ with bounded variation satisfies $\phi(\xi) = o\left(|\xi|^{-1}\right)$.*

Proof. Since f has bounded variations, the limits $f^+ = \lim_{x\to\infty} f(x)$ and $f^- = \lim_{x\to-\infty} f(x)$ exists and are finite. Let $s(x) = \frac{(f^+ + f_-)}{2} + \operatorname{erf}(x)\frac{(f^+ - f^-)}{2}$ and $f_0(x) = f(x) - s(x)$ and observe that the Fourier transform of $s(x)$ is $\sigma(\xi) = \frac{(f^+ + f_-)}{2}\delta(\xi) + \frac{(f^+ - f^-)}{2}\frac{1}{i\xi}2e^{-\frac{1}{4}\xi^2} = o(|\xi|)$.

Also, by construction $\lim_{|x|\to\infty} f_0(x) = 0$ and is absolutely continuous with bounded variations because f and s are.

We then have

$$\phi(\xi) = \int f(x) e^{i\xi x} dx = \int f_0(x) e^{i\xi x} dx + \int s(x) e^{i\xi x} dx = \int f_0(x) e^{i\xi x} dx + o(|\xi|)$$

If f_0 is absolutely continuous then it can be written as $f_0(a) + \int_a^x g(y)\, dy$ for some Lebesgue integrable function g. We then have, by integration by parts:

$$\begin{aligned}
\phi(\xi) &= \int \left(f_0(a) + \int_a^x g(y)\, dy \right) e^{i\xi x} dx + o\left(|\xi|^{-1}\right) \\
&= \left[f_0(x) \frac{e^{i\xi x}}{i\xi} \right]_{-\infty}^{\infty} - \int g(x) \frac{e^{i\xi x}}{i\xi} dx + o\left(|\xi|^{-1}\right) \\
&= 0 - \frac{1}{i\xi} \int g(x) e^{i\xi x} dx + o\left(|\xi|^{-1}\right)
\end{aligned}$$

Since f_0 has bounded variation, g is absolutely integrable. Then, by the Riemann-Lebesgue Lemma, $\lim_{|\xi|\to\infty} \int g(x) e^{i\xi x} dx = 0$, so $-\frac{1}{i\xi} \int g(x) e^{i\xi x} dx = o\left(|\xi|^{-1}\right)$ as well. ■

Theorem 5 *[Adapted from the Recurrence Theorem of Birkhoff (1931b)[10]]. Let $\psi \in \mathbb{R}^J$ be the state of a dynamical system evolving according to a measure-preserving transformation of the form*

$$\frac{d\psi_j}{dt} = \Psi_j(\psi_1, \ldots, \psi_J) \quad j = 1, \ldots, J, \tag{10}$$

where the functions Ψ_i are analytic. Let $\psi(t)$ denote the system's trajectory (solving (10)). Consider an analytic surface $\mathcal{S}$ in $\mathbb{R}^J$ and let $\xi_k(\mathcal{S})$ denote the time t where the trajectory $\psi(t)$ crosses $\mathcal{S}$ for the k-th time. Then, there exists $\lambda^*, \lambda_* \in \mathbb{R}^+$ with $|\lambda^* - \lambda_*|$ arbitrarily small such that, for all initial conditions $\psi(0)$ (except on a set of null Lebesgue measure), we have, for all $k \in \mathbb{N}$ greater than some k_0,

$$k\lambda_* \leq \xi_k(\mathcal{S}) \leq k\lambda^*. \tag{11}$$

Corollary 2 *Theorem 5 obviously holds for complex-valued ψ_j as well: One merely needs to consider the real and imaginary parts of each ψ_j as two distinct state variables and create an equivalent dynamical system consisting of $2J$ real-valued variables. Also, the results obviously holds if the surface $\mathcal{S}$ is the boundary of some set open $\mathcal{P}$ defined by analytic inequalities. Hence, a relation of the form (11) also holds (perhaps for different λ^*, λ_*) for the times $\xi_k(\mathcal{P})$ the trajectory enters $\mathcal{P}$ for the k-th time.*

[10] This article is an extension of Birkhoff (1931a), which is useful to consult first. We report here some of the paper's intermediate results rather than its final conclusion, because they are more useful for our purposes.

Lemma 7 *Let $\bar{\phi}(\xi) = \sum_{j=1}^{J} A_j e^{i\xi x_j}$ where $A_j \in \mathbb{C}\backslash\{0\}$ and $x_j \in \mathbb{R}$ for $j = 1, \ldots, J$. Then, for all values of $A_1, \ldots, A_J$ (except on a set of null Lebesgue measure), there exists, for any $c > 0$, a strictly increasing sequence ξ_k with $\xi_k \to \infty$ such that $\left|\bar{\phi}(\xi_k)\right| \geq \sum_{j=1}^{J} |A_j| - c$ and $\left|\xi_{k+1} - \xi_k\right| / |\xi_k| \leq 1/3$.*

Proof. We recast this question in a dynamical system framework to enable us to the invoke Birkhoff's Theorem (reproduced above as Theorem 5 and Corollary 2). Let the system's state be denoted $\psi \equiv (\psi_1, \ldots, \psi_J) \in \mathbb{C}^J$ which evolves (with increasing ξ) according to

$$\frac{d\psi_j}{d\xi} = ix_j\psi_j$$

from the initial conditions $\psi_j = A_j$ at $\xi = 0$. This dynamical system simply defines a rotation in space and thus satisfies Birkhoff Theorem's requirement of being measure-preserving and analytic. Observe that we then have $\psi_j(\xi) = A_j e^{i\xi x_j}$ so that $\bar{\phi}(\xi) = \sum_{j=1}^{J} \psi_j(\xi)$. Let $\mathcal{P} = \left\{\psi \in \mathbb{C}^J : \mathrm{Re}\left(\psi_j\right) \geq |A_j| - c/J \text{ for } j = 1, \ldots, J\right\}$ and note that $\psi \in \mathcal{P}$ is a sufficient condition for $\left|\bar{\phi}(\xi)\right| \geq \sum_{j=1}^{J} |A_j| - c$, since

$$\left|\bar{\phi}(\xi)\right| \geq \left|\mathrm{Re}\left(\bar{\phi}(\xi)\right)\right| \geq \mathrm{Re}\,\bar{\phi}(\xi) = \sum_{j=1}^{J} \mathrm{Re}\left(\psi_j(\xi)\right) \geq \sum_{j=1}^{J} \left(|A_j| - c/J\right) = \sum_{j=1}^{J} |A_j| - c.$$

Birkhoff ergodic theorem on recurrence time Birkhoff (1931b) then implies that, for all initial conditions A_j (except on a set of null Lebesgue measure), the "times" $\xi_k(\mathcal{P})$ when $\psi(\xi)$ enters $\mathcal{P}$ for the k-th time satisfy:

$$k\lambda_* \leq \xi_k(\mathcal{P}) \leq k\lambda^*$$

where $|\lambda^* - \lambda_*|$ can be chosen arbitrarily small (in particular, such that $\lambda^*/\lambda_* \leq 1/6$). Thus,

$$\frac{\xi_{k+1}(\mathcal{P}) - \xi_k(\mathcal{P})}{\xi_k(\mathcal{P})} \leq \frac{(k+1)\lambda^* - k\lambda_*}{k\lambda_*} = \frac{\lambda^*}{\lambda_*} + k^{-1}\frac{\lambda^*}{\lambda_*} - 1 \leq 2\frac{\lambda^*}{\lambda_*} \leq \frac{1}{3}.$$

■

Proof of Theorem 2. For $\xi \in \left[0, \hat{\mu}_X^{-1} n^{1/4}\right]$, Lemma 4 applies and the error on $\hat{\phi}_{Y;X}(\xi)$ is indeed bounded by $\Delta\phi_{Y;X,n}$ over that interval. Also, $\Delta\hat{\phi}_{Y;X,n}/\Delta\phi_{Y;X,n} \xrightarrow{p} 1$. Since $\frac{1}{\ln n} \geq \frac{\Delta\hat{\phi}_{Y;X,n}}{\left|\hat{\phi}_{Y;X}\left(\bar{\xi}_n\right)\right|}$ by construction, we then have

$$\frac{1}{\ln n} \geq \frac{\Delta\hat{\phi}_{Y;X,n}}{\left|\hat{\phi}_{Y;X}\left(\bar{\xi}_n\right)\right|} \geq \frac{\Delta\hat{\phi}_{Y;X,n}}{\left|\phi_{Y;X}\left(\bar{\xi}_n\right)\right| + \Delta\phi_{Y;X,n}} \geq \frac{\Delta\phi_{Y;X,n}/2}{\left|\phi_{Y;X}\left(\bar{\xi}_n\right)\right| + \Delta\phi_{Y;X,n}} = \frac{1/2}{\frac{\left|\phi_{Y;X}\left(\bar{\xi}_n\right)\right|}{\Delta\phi_{Y;X,n}} + 1}$$

with probability approaching one (wpa1). Next, since the function $(1/x + 1)^{-1}$ is increasing in x, we have, wpa1,

$$\frac{\Delta\phi_{Y;X,n}}{\left|\phi_{Y;X}\left(\bar{\xi}_n\right)\right|} \leq \frac{1}{\frac{1}{2}\ln n - 1} \to 0$$

Since $\left|\phi_{Y;X}(\xi)\right| \leq A_{Y;X}\xi^{-r_{y;X}}$,

$$\frac{\Delta\phi_{Y;X,n}}{\left|\phi_{Y;X}\left(\bar{\xi}_n\right)\right|} \geq \frac{\Delta\phi_{Y;X,n}}{A_{Y;X}\bar{\xi}_n^{-r_{Y;X}}}$$

and thus $\Delta\phi_{Y;X,n}\bar{\xi}_n^{r_{Y;X}} \to 0$. Since $\sup_{\xi\in\left[\underline{\xi},\bar{\xi}_n\right]}\Delta\phi_{Y;X,n}\xi^{r_{Y;X}} = \Delta\phi_{Y;X,n}\bar{\xi}_n^{r_{Y;X}} \to 0$, the conclusion follows. ■

Proof of Theorem 3. The result follows directly from our bias bounds and traditional standard deviation bands (for instance, Härdle and Linton (1994)). The only additional step is to observe that, since under assumption 5 the bias and the standard deviations are of the same order, so the fact that the estimation error on the bias bound is negligible relative to the bias itself implies that it is also negligible relative to the standard deviation. ■

Proof of Theorem 4. Similar to Theorem 3 and therefore omitted. ■