Coverage Error Optimal Confidence Intervals for Local Polynomial Regression[*]

Sebastian Calonico[†] Matias D. Cattaneo[‡] Max H. Farrell[§]

May 28, 2020

Abstract

We characterize the minimax bound on coverage error of Wald-type confidence intervals for nonparametric local polynomial regression. This bound depends on the smoothness of the population regression function, the smoothness exploited by the inference procedure, and on whether the evaluation point of interest is in the interior or on the boundary of the support of the regression function. Our results also cover inference on derivatives of the regression function, in which case we find that the minimax coverage error bound does not depend on the order of the derivative being estimated. We show that robust bias corrected confidence intervals are able to attain the minimax rate when coupled with the principled, inference-optimal tuning parameter selections we propose. In addition, we show how the large-sample interval length can be further optimized through choice of the kernel function and other tuning parameters. Our main theoretical results rely on novel Edgeworth expansions that are proven to hold uniformly over relevant classes of data generating processes. These higher-order expansions allow for the uniform kernel and any derivative order, improving on previous technical results available in the literature.

Keywords: minimax bound, Edgeworth expansion, Cramér condition, nonparametric regression, robust bias correction, optimal inference.

[*] The second author gratefully acknowledges financial support from the National Science Foundation (SES 1357561, SES 1459931 and SES-1947805). The third author gratefully acknowledges financial support from the Richard N. Rosett and John E. Jeuck Fellowships. We thank Chris Hansen, Michael Jansson, Adam McCloskey, Rocio Titiunik, and participants at various seminars and conferences for comments.

[†] Department of Health Policy and Management, Columbia University.

[‡] Department of Operations Research and Financial Engineering, Princeton University.

[§] Booth School of Business, University of Chicago.

1 Introduction

In this paper we study the quality of local polynomial nonparametric inference in the general heteroskedastic nonparametric regression model:

$$Y = \mu(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon|X] = 0, \qquad \mathbb{E}[\varepsilon^2|X] = v(X), \tag{1}$$

where (Y, X) is a pair of random variables. The parameter of interest is the level or derivative of the regression function at a point $X = \mathsf{x}$:

$$\mu^{(\nu)} = \mu^{(\nu)}(\mathsf{x}) := \frac{\partial^\nu}{\partial x^\nu}\mathbb{E}\left[Y|X=x\right]\Big|_{x=\mathsf{x}}, \qquad \nu \in \mathbb{Z}_+, \tag{2}$$

where the evaluation point x may be in the interior or on the boundary of the support of X. We drop the evaluation point from the notation when possible, and employ the usual convention $\mu = \mu^{(0)}$. Derivatives at boundary points are defined as one-sided derivatives from the interior.

Our measure of inference quality will be how rapidly the coverage of a confidence interval collapses to its nominal level and, in particular, we are interested in a minimax result: characterizing the fastest (minimal) rate at which the worst-case (maximal) coverage error vanishes. The bound and its attainability will depend on the smoothness assumed on μ and that exploited by the local polynomial estimation procedure. To state the problem more formally, for a nominal level $(1 - \alpha)$, with fixed $\alpha \in (0, 1)$, we characterize the *minimax optimal coverage error decay rate bound*, which is the fastest vanishing sequence $r_\star = r_\star(n)$, $n \in \mathbb{N}$, such that

$$\liminf_{n\to\infty} r_\star^{-1} \inf_{I\in\mathscr{I}_p} \sup_{F\in\mathscr{F}_S} \left|\mathbb{P}_F\left[\mu^{(\nu)}(\mathsf{x}) \in I\right] - (1-\alpha)\right| > 0, \tag{3}$$

where $\mathbb{P}_F$ denotes the probability law of a random sample of size n from (Y, X) under a data generating process F, assumed to belong to a class $\mathscr{F}_S$, and where I denotes a confidence interval, assumed to belong to a class $\mathscr{I}_p$ of interval estimators. The latter two classes capture the key features of the problem: smoothness assumptions and bandwidth choices.

We restrict the classes $\mathscr{F}_S$ and $\mathscr{I}_p$ so that our results reflect empirical practice. We will assume that μ possesses at least S well-behaved derivatives and accordingly let $\mathscr{F}_S$ denote the researcher's set of plausible distributions for the data. Precise regularity conditions are given in Assumption 1 below. The class $\mathscr{I}_p$ collects interval estimators based on p-degree nonparametric local polynomial regression (Fan and Gijbels, 1996). More precisely, for centering $\hat{\theta}$ and scaling $\hat{\vartheta}$, based on local polynomial estimation, members of $\mathscr{I}_p$ are of the Wald-type form:

$$I = \left[\hat{\theta} - z_u\, \hat{\vartheta}\ ,\ \hat{\theta} - z_l\, \hat{\vartheta}\right], \tag{4}$$

where z_l and z_u denote an appropriate pair of quantiles. Each $I \in \mathscr{I}_p$ is determined by a choice of centering and scaling (as well as quantiles), and hence by a choice of bandwidth(s), kernel function(s) and any other tuning parameters, which are left implicit in the notation at this point. A detailed description is given in Section 2 below, including precise regularity conditions in Assumption 2.

We compare confidence interval estimators (in $\mathscr{I}_p$) by their worse case coverage error over the class of plausible data generating processes (in $\mathscr{F}_S$), building on an idea originally introduced by Hall and Jing (1995) for the specific case of one-sided confidence intervals in the i.i.d. parametric location model. A researcher's statistical model, accompanying assumptions, and the empirical regularities of the application of interest formalize a class of plausible distributions for the data encoded in $\mathscr{F}_S$. A researcher would like some assurances that the chosen confidence interval estimator is accurate in coverage level regardless of the specific data generating process. A "good" confidence interval is one for which this maximal error is minimized: a minimax result. At an intuitive level, this corresponds to the desire for similarity in testing: the confidence interval should have "similar" coverage over the set of plausible distributions. This notion of minimax optimality is specific to inference, and is conceptually distinct from the usual minimaxity considered for point estimation (Cheng et al., 1997; Fan et al., 1997).

In general terms, the interplay between the classes $\mathscr{F}_S$ and $\mathscr{I}_p$ is crucial for study of (3), and $\mathscr{F}_S$ and $\mathscr{I}_p$ should be neither too "large" nor too "small" in order to obtain useful and interesting results. The larger is $\mathscr{F}_S$, the more plausible a given data set is generated by some $F \in \mathscr{F}_S$, but well known results dating back at least to Bahadur and Savage (1956) show that if $\mathscr{F}_S$ is too large it is impossible to construct an "effective confidence interval" that controls the worst-case coverage. Similarly, the set $\mathscr{I}_p$ should be large enough to include both popular and useful interval estimators, but not so large as to include trivial estimators such as, for example, setting $I = (-\infty, \infty)$ with

$(1-\alpha)$ probability and empty otherwise. In the nonparametric regression setting we study, these concerns boil down to restrictions on the smoothness assumed on μ and the smoothness utilized by the estimation procedure, as those restrictions will ultimately affect the center and length of the confidence intervals.

Given $\mathscr{F}_S$ and $\mathscr{I}_p$, we derive $r_\star$ and show how the minimax rate depends on the relationship between p and S and on whether the point of evaluation x lies on the boundary of the support of X or its interior. This result, given in Section 3, is of both theoretical and practical interest because it provides a benchmark for the "best" possible confidence interval estimator in terms of uniformly fastest contraction rate of coverage error, as defined in Eqn. (3). This minimax bound applies to a large class of Wald-type confidence interval of the form (4) constructed using local polynomial estimation methods.

We then show that robust bias corrected confidence intervals (Calonico et al., 2014, 2018), reviewed in Section 2, can attain the minimax rate bound $r_\star$. See also Chen (2017) for a tutorial describing robust bias correction methods for Biostatistics and Epidemiology. In a nutshell, the idea underlying this inference approach is to estimate a bias correction term for the centering $\hat{\theta}$, but then also adjust the scale $\hat{\vartheta}$ to account for the additional variability introduced by the bias estimation. Implementation of the optimal interval requires bandwidth choices, and one byproduct of our theoretical results are accompanying inference-optimal bandwidth selectors. The need for inference optimal bandwidths has been appreciated theoretically and derived using Edgeworth expansions at least as far back as Hall (1992b), but they are not as often used in applications. In the context of kernel-based nonparametrics, interval estimators with good control of worst-case coverage are able to use larger bandwidths in general, and are thus shorter in large samples; an analogue to the adage that "similar tests have higher power".

Beyond the bandwidth choice, we further optimize the interval length through choice of kernel shape and other tuning parameters. We discuss these results in Section 4. Supoptimality of some common interval estimators, such as those constructed using classical undersmoothing, and other methodological findings are discussed in Section 5. We also highlight in that section the sometimes striking and underappreciated gap between point estimation and inference: the two may proceed at very different convergence rates. We illustrate by example that it is possible to perform optimal inference based on a point estimator that is not mean square consistent.

Our theoretical findings, as well as the practical recommendations that stem from them, are the results of new theoretical work on Edgeworth expansions, a long-standing tool for more detailed asymptotic analysis (Hall, 1992a). We prove novel Edgeworth expansions that (i) hold uniformly over $\mathscr{F}_S$, (ii) allow for the uniform kernel, and (iii) allow for any derivative ($\nu \geq 0$), all of which improve on prior literature as detailed and referenced in Section 5 below. An online supplement contains all proofs and complete derivations not reported in the main paper.

2 Model Assumptions and Interval Estimators

The minimax rate bound $r_\star$ in (3) depends on the class of data-generating processes, $\mathscr{F}_S$, the set interval estimators considered, $\mathscr{I}_p$, and whether x is on the boundary or in the interior of the support. We first formalize $\mathscr{F}_S$ through the following assumption, which imposes conditions that are not materially stronger than usual, beyond what is naturally required for uniform validity of Edgeworth expansions. Recall that derivatives at the boundary of the support of X correspond to one-sided derivatives from the interior of the support.

Assumption 1. *Let $\mathscr{F}_S$ be the set of distributions F for the pair (Y, X) which obey model (1) and the following. There exist constants $S \geq \nu$, $s \in (0, 1]$, $0 < c < C < \infty$, and a neighborhood of x on the support of X, none of which depend on F, such that for all x, x' in the neighborhood:*

- **(a)** *the Lebesgue density of (Y, X), $f_{yx}(\cdot)$, is continuous and $c \leq f_{yx}(\cdot) \leq C$; the Lebesgue density of X, $f(\cdot)$, is continuous and $c \leq f(x) \leq C$; $v(x) := \mathbb{V}[Y|X = x] \geq c$ and continuous; and $\mathbb{E}[|Y|^{8+c}|X = x] \leq C$, and*
- **(b)** *$\mu(\cdot)$ is S-times continuously differentiable and $|\mu^{(S)}(x) - \mu^{(S)}(x')| \leq C|x - x'|^s$.*

Throughout, $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ is a random sample from (Y, X).

Turning to interval estimation, we now define the various pieces of the class $\mathscr{I}_p$, which are of the form (4). We begin with the center of the interval, $\hat{\theta}$, which may account explicitly for the nonparametric bias (i.e., bias correction). The starting point is the standard local polynomial regression point estimate, which we index by $p \in \mathbb{Z}_+$, the order of the polynomial used, assumed to be at least $\nu \in \mathbb{Z}_+$. See Fan and Gijbels (1996) for an introduction to local polynomial regression.

Define $X_{h,i} = (X_i - \mathsf{x})/h$ and, similarly suppressing the dependence on x to simplify notation, set

$$\hat{\mu}_p^{(\nu)} = \nu! \boldsymbol{e}_\nu' \hat{\boldsymbol{\beta}}_p = \frac{1}{nh^\nu} \nu! \boldsymbol{e}_\nu' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Omega} \boldsymbol{Y}, \qquad \hat{\boldsymbol{\beta}}_p = \underset{\boldsymbol{b} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \boldsymbol{r}_p(hX_{h,i})' \boldsymbol{b})^2 K\left(X_{h,i}\right), \tag{5}$$

where $h = h(n) \to 0$ is a bandwidth sequence, $\boldsymbol{e}_\nu$ is the $(p+1)$-vector with a one in the $(\nu+1)^{\text{th}}$ position and zeros in the rest, $\boldsymbol{r}_p(u) = (1, u, u^2, \ldots, u^p)'$, K is a kernel or weighting function,

$$\boldsymbol{\Gamma} = \frac{1}{nh} \sum_{i=1}^{n} K\left(X_{h,i}\right) \boldsymbol{r}_p\left(X_{h,i}\right) \boldsymbol{r}_p\left(X_{h,i}\right)', \quad \boldsymbol{\Omega} = \frac{1}{h}\left[K\left(X_{h,1}\right) \boldsymbol{r}_p\left(X_{h,1}\right), \ldots, K\left(X_{h,n}\right) \boldsymbol{r}_p\left(X_{h,n}\right)\right],$$

and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$.

The inevitable nonparametric smoothing bias further determines the set of possible centerings and their properties. Our criterion for coverage is explicitly defined in terms of coverage of the true function, not the expectation of the estimator, as the true function is of direct scientific importance. One way or another smoothing bias must be removed. Two leading methods are undersmoothing and manual bias correction. To motivate these approaches, assume for the moment that $p - \nu$ is odd and $p \leq S - 1$, in which case the conditional bias of $\hat{\mu}_p^{(\nu)}$ is

$$\mathbb{E}\left[\hat{\mu}_p^{(\nu)} \middle| X_1, \ldots, X_n\right] - \mu^{(\nu)} = h^{p+1-\nu} \nu! \boldsymbol{e}_\nu' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Lambda} \frac{\mu^{(p+1)}}{(p+1)!} + o_{\mathbb{P}}(h^{p+1-\nu}), \tag{6}$$

with $\boldsymbol{\Lambda} = \boldsymbol{\Omega}[X_{h,1}^{p+1}, \cdots, X_{h,n}^{p+1}]'/n$. Asymptotic orders and their in-probability versions always hold uniformly in $\mathscr{F}_S$, as required by our framework, so that, for example $A_n = o_{\mathbb{P}}(a_n)$ means $\sup_{F \in \mathscr{F}_S} \mathbb{P}_F[|A_n/a_n| > \epsilon] \to 0$ for every $\epsilon > 0$. Limits are taken as $n \to \infty$ unless stated otherwise.

Robust bias correction involves subtracting an estimate of the leading term of (6), of which only $\mu^{(p+1)}$ is unknown, and accounting for the variability of this point estimate in standard errors (see (10) below). The bias corrected interval center is

$$\hat{\theta}_{\texttt{rbc}} := \hat{\mu}_p^{(\nu)} - h^{p+1-\nu} \nu! \boldsymbol{e}_\nu' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Lambda} \boldsymbol{e}_{p+1}' \hat{\boldsymbol{\beta}}_{p+1}, \tag{7}$$

where $\hat{\boldsymbol{\beta}}_{p+1}$ is defined via (5), but with a bandwidth $b := \rho^{-1} h$ instead of h. The parameter ρ will play a key role throughout: $\rho = 1$ means the same bandwidth is used in point estimation and bias correction ($h = b$), but we find in Section 4.2 that other values of ρ may be preferred from an inference perspective.

Undersmoothing, on the other hand, leaves the center of the interval at $\hat{\theta} = \hat{\mu}_p^{(\nu)}$ unchanged, and assumes instead that the bandwidth h vanishes rapidly enough to render the leading term of (6) negligible relative to the standard error of the point estimator. The term *under*smoothing refers to using less nonparametric smoothing than would be optimal from a mean squared error (MSE) point estimation point of view (Fan and Gijbels, 1996, Section 4). The MSE-optimal bandwidth choice is the most common by far, and indeed, the default in most software. With $p \leq S - 1$, the MSE-optimal bandwidth for $\hat{\mu}_p^{(\nu)}$ is well-defined whenever $\mu^{(p+1)}(\mathsf{x}) \neq 0$, and takes the form $h_{\texttt{mse}} = H_{\texttt{mse}} n^{-1/(2p+3)}$, where $H_{\texttt{mse}}$ is a constant that depends on the variance and bias at x of the local polynomial estimator. However, the MSE-optimal bandwidth is too "large" for standard Gaussian inference: the bias (6) remains first-order important when scaled by the standard deviation of the point estimator (Eqn. (9) below). Valid inference requires a bandwidth that vanishes faster than $n^{-1/(2p+3)}$ when the local polynomial estimator $\hat{\theta} = \hat{\mu}_p^{(\nu)}$ is used in (4) to construct the confidence interval.

With the center of $I \in \mathscr{I}_p$ defined, we turn to the scaling $\hat{\vartheta}$ in Eqn. (4). In contrast to first order distributional analysis, where only consistency is required, the choice of scaling is crucial for coverage error. Our detailed expansions show that, in general, there are two types of higher-order terms that arise due to Studentization. One is the unavoidable estimation error incurred when replacing any population quantity with a feasible counterpart. The second error arises from the difference between the population variability of the centering $\hat{\theta}$ and the population standardization chosen as the target. This second type of error can be removed entirely by employing "fixed-n" Studentization, also called "preasymptotic" by Fan and Yao (2005), which means choosing the Studentization to directly estimate $\mathbb{V}[\hat{\theta}|X_1, \ldots, X_n]$, a population quantity but not an asymptotic one, instead of employing an estimator of an asymptotic representation of $\mathbb{V}[\hat{\theta}|X_1, \ldots, X_n]$.

For centering $\hat{\theta}$ chosen to be either $\hat{\mu}_p^{(\nu)}$ or $\hat{\theta}_{\texttt{rbc}}$, the corresponding fixed-n variances are easy to compute. Using $\rho = h/b$, we can rewrite

$$\hat{\theta}_{\texttt{rbc}} = \frac{1}{nh^{\nu}} \nu! \boldsymbol{e}_{\nu}' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Omega}_{\texttt{rbc}} \boldsymbol{Y}, \qquad \boldsymbol{\Omega}_{\texttt{rbc}} = \boldsymbol{\Omega} - \rho^{p+1} \boldsymbol{\Lambda} \boldsymbol{e}_{p+1}' \bar{\boldsymbol{\Gamma}}^{-1} \bar{\boldsymbol{\Omega}}, \tag{8}$$

with $\bar{\boldsymbol{\Gamma}}$ and $\bar{\boldsymbol{\Omega}}$ defined akin to $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$, but with $p + 1$ and b in place of p and h, respectively.

Comparing to (5), the only change when $\hat{\theta} = \hat{\theta}_{\texttt{rbc}}$ is replacing $\mathbf{\Omega}$ with $\mathbf{\Omega}_{\texttt{rbc}}$, but both choices of centering have the same structure. Consequently, we first give details for $\hat{\theta} = \hat{\mu}_p^{(\nu)}$, as the same formula applies to $\hat{\theta}_{\texttt{rbc}}$ upon changing $\mathbf{\Omega}$ to $\mathbf{\Omega}_{\texttt{rbc}}$. The fixed-$n$ (conditional) variance is

$$nh^{1+2\nu}\, \mathbb{V}[\hat{\mu}_p^{(\nu)}|X_1,\ldots,X_n] = \nu!^2 \boldsymbol{e}_\nu' \mathbf{\Gamma}^{-1}(h\mathbf{\Omega}\mathbf{\Sigma}\mathbf{\Omega}'/n)\mathbf{\Gamma}^{-1}\boldsymbol{e}_\nu, \tag{9}$$

where $\mathbf{\Sigma}$ is the n-diagonal matrix of conditional variances $v(X_i)$ in (1). The fixed-n Studentization is obtained by replacing $\mathbf{\Sigma}$ with an appropriate plug-in estimator thereof.

More specifically, using Eqn. (9), we define the (square of the) fixed-n scalings

$$\begin{aligned} \hat{\vartheta}^2 &= \frac{\hat{\sigma}_p^2}{nh^{1+2\nu}}, & \hat{\sigma}_p^2 &:= \nu!^2 \boldsymbol{e}_\nu' \mathbf{\Gamma}^{-1}(h\mathbf{\Omega}\hat{\mathbf{\Sigma}}_p\mathbf{\Omega}'/n)\mathbf{\Gamma}^{-1}\boldsymbol{e}_\nu, \qquad \text{and} \\ \hat{\vartheta}^2 &= \hat{\vartheta}_{\texttt{rbc}}^2 := \frac{\hat{\sigma}_{\texttt{rbc}}^2}{nh^{1+2\nu}}, & \hat{\sigma}_{\texttt{rbc}}^2 &:= \nu!^2 \boldsymbol{e}_\nu' \mathbf{\Gamma}^{-1}(h\mathbf{\Omega}_{\texttt{rbc}}\hat{\mathbf{\Sigma}}_{\texttt{rbc}}\mathbf{\Omega}_{\texttt{rbc}}'/n)\mathbf{\Gamma}^{-1}\boldsymbol{e}_\nu, \end{aligned} \tag{10}$$

where $\hat{\mathbf{\Sigma}}_p$ and $\hat{\mathbf{\Sigma}}_{\texttt{rbc}}$ are the n-diagonal matrices of the squared residuals $\hat{v}(X_i) = (Y_i - \boldsymbol{r}_p(X_i)'\hat{\boldsymbol{\beta}}_p)^2$ and $\hat{v}(X_i) = (Y_i - \boldsymbol{r}_{p+1}(X_i)'\hat{\boldsymbol{\beta}}_{p+1})^2$, respectively. The above variance estimators separate explicitly the "constant" portions, denoted $\hat{\sigma}_p^2$ and $\hat{\sigma}_{\texttt{rbc}}^2$, which will be used in Section 4.2 for interval length optimization. More precisely, $\hat{\sigma}_p^2$ and $\hat{\sigma}_{\texttt{rbc}}^2$ will both be bounded and bounded away from zero in probability under our assumptions.

For comparison, consider approximating the variability of $\hat{\theta}$ by a limiting quantity instead of the fixed-n variance in (9). It is common practice to use the leading term of the asymptotic simplification $nh^{1+2\nu}\mathbb{V}[\hat{\theta}|X_1,\ldots,X_n] = v(\mathsf{x})f(\mathsf{x})^{-1}\nu!^2 \int \mathcal{K}(t)dt + o_{\mathbb{P}}(1)$, where $\mathcal{K}(t)$ is known as the equivalent kernel and is only a function of $K(\cdot)$, p, ν and, if required, ρ. This alternative, asymptotic variance representation is made feasible by using plug-in estimators of $f(\mathsf{x})$ and $v(\mathsf{x})$. When x is an interior point, standard kernel methods are used, while when x is on the boundary some changes are required. Using such asymptotic approximations to the variability of $\hat{\theta}$ can only increase coverage error, and thus we refer to Fan and Gijbels (1996) and Chen and Qin (2002) for theoretical and implementation details.

This completes the description of the elements of all $I \in \mathscr{I}_p$: each I is of the form (4), with any combination of the centerings and scalings given above, bandwidth choices h and b, if needed, and any chosen quantiles. Any such interval is *allowed*, but may not be *valid*. That is, for some

$I \in \mathscr{I}_p$, convergence of $\mathbb{P}_F\big[\mu^{(\nu)}(\mathsf{x}) \in I\big]$ to $(1-\alpha)$ fails either pointwise in F or uniformly in $F \in \mathscr{F}_S$. Examples include trivial cases such improper choices of quantiles or inconsistent variance estimators, but also choices such as using the MSE-optimal bandwidth sequence with centering $\hat{\theta} = \hat{\mu}_p^{(\nu)}$ and scaling $\hat{\vartheta}^2 = \hat{\sigma}_p^2/(nh^{1+2\nu})$. We make precise these claims in the upcoming sections.

Finally, we require the following regularity conditions on the kernel function and the polynomial degree, which formally restricts the class $\mathscr{I}_p$.

Assumption 2. *The kernel K is supported on $[-1,1]$, positive, bounded, and even. Further, $K(u)$ is either constant (the uniform kernel) or $(1, K(u)\boldsymbol{r}_{3(k+1)}(u))'$ is linearly independent on $[-1,0]$ and $[0,1]$, where $k = p$ if I is based on $\hat{\mu}_p^{(\nu)}$ and $\hat{\sigma}_p$ and $p+1$ if I uses $\hat{\theta}_{\mathtt{rbc}}$ and $\hat{\sigma}_{\mathtt{rbc}}$. The order p is at least ν.*

This assumption allows for standard choices such as the triangular and Epanechnikov kernels. A noteworthy technical innovation in this paper is that we allow for the uniform kernel: $K(u) = \mathbb{1}\{|u| < 1\}/2$. This technical innovation important because the uniform kernel not only is a popular choice in some domain-specific applied work, but also has length-optimality properties as discussed in Section 4.2 below. Prior work on coverage error expansions in kernel-based nonparametrics have explicitly ruled out the uniform kernel, essentially due to the failure of linear independence. See Section 5 for more details.

3 Minimax Coverage Error Decay Rates

With $\mathscr{F}_S$ and $\mathscr{I}_p$ defined in the previous section, we can now state the minimax bound.

Theorem 1. *Let Assumptions 1 and 2 hold.*

(i) Let x be an interior point in the support of X. If $p-\nu$ is odd, then Eqn. (3) holds with $r_\star = n^{-(p+3)/(p+4)}$ if $p \leq S-3$ and $r_\star = n^{-(S+s)/(S+s+1)}$ if $p \geq S-2$. If $p-\nu$ is even, then $r_\star = n^{-(p+2)/(p+3)}$ if $p \leq S-2$ and $r_\star = n^{-(S+s)/(S+s+1)}$ if $p \geq S-1$.

(ii) Let x be a boundary point of the support of X. Then, Eqn. (3) holds with $r_\star = n^{-(p+2)/(p+3)}$ if $p \leq S-2$ and $r_\star = n^{-(S+s)/(S+s+1)}$ if $p \geq S-1$.

This result gives the minimax rates for coverage error of the interval estimators in (4). Notice that because (3) is a pointwise in x asymptotic criterion, there is no need to consider a "region" of points within h of the boundary, as is sometimes done in local polynomial analyses, although this theorem could be extended via the usual localization-to-evaluation-point analysis (Fan and Gijbels, 1996, Section 3.2.5). As discussed above, p and S are key. It is the rate at which the bias vanishes which determines the separate cases, because the variance rate is always $(nh)^{1+2\nu}$, see Equation (10), provided ρ does not diverge. The fastest coverage error decay rate is limited by S and s. If these were known, then the researcher could choose p (and h and ρ) as a function of S and s to attain the minimax optimal rate $n^{-(S+s)/(S+s+1)}$ in all cases. This is the fastest possible rate: no choice of $I \in \mathscr{I}_p$, for any p, can attain a faster coverage error decay rate. However, this bound is of little practical importance because attaining it requires knowledge of S and s. In practice, S is not known and researchers first choose p and then conduct inference based on that choice (witness the ubiquity of local linear regression and cubic splines). Intervals which utilize more smoothness will yield better coverage error if such smoothness is available. If p, although fixed, is close enough to S, the same rate can be attained as in the known-S case. In the more empirically relevant case, S is unknown but taken to be larger than a fixed p, thereby making the coverage error rate depend only on p.

An important novel finding is that the coverage error rate bound does *not* depend on ν, the derivative being estimated. This highlights a perhaps under-appreciated gap between point estimation and inference: the two may proceed at different rates. In particular, the inference rate can be *faster*: the rate at which the distribution of $\hat{\theta}$ collapses to its asymptotic value (namely $\Phi(\cdot)$) can be *faster* than the rate at which $\hat{\theta}$ itself collapses to its asymptotic value (i.e. $\mu^{(\nu)}$). Indeed, Section 5.1 presents an extreme example of an interval that is minimax optimal but is based on a point estimator that is not mean-square consistent.

With the minimax rates known, we seek an interval estimator that can attain this rate. The next section shows that robust bias corrected confidence intervals attain the bounds in Theorem 1 in all cases, for any ν. Intuitively, this is because robust bias correction successfully exploits additional smoothness if it exists, but is not punished (in rates) if there is no such smoothness due to the change in Studentization. Beyond attaining optimality, it is also useful to show that certain other confidence interval estimators can never be optimal, as this gives empirical researchers tight guidance. We postpone the latter discussion until Section 5.1.

4 Attaining the Bound

We now show that robust bias correction is minimax coverage error rate optimal. To be specific, we consider the interval that uses centering $\hat{\theta} = \hat{\theta}_{\texttt{rbc}}$ in (7) and standard errors $\hat{\vartheta} = \hat{\vartheta}_{\texttt{rbc}}$ in (10). Further, we use the quantiles $z_l = -z_u = z_{\alpha/2} = \Phi^{-1}(\alpha/2)$, where $\Phi(\cdot)$ denotes the standard Gaussian distribution function, as our general results below (see Section 5) reaffirm the classical finding that symmetric intervals have superior coverage properties. We also discuss the choice of bandwidth h in practice, either to optimize the rate or to balance coverage error and interval length, and discuss how to select ρ and the kernel function to further optimize the length of the resulting confidence interval estimator. We let $I_{\texttt{rbc}}(h)$ denote the robust bias corrected confidence interval estimator with these choices, but now making explicit only the dependence on the bandwidth h.

Crucial to proving that such an interval is minimax optimal is that the bias vanishes at the best possible rate, given the smoothness assumed (S) and utililized (p). The bias depends on the location of x, the parity of $p-\nu$, the relationship between p and S, and, if p is large enough, also on s. In typical, first order nonparametric inference, the bias is studied conditional on $X_1, \ldots, X_n$, as in Equation (6). However, terms of coverage error (and Edgeworth) expansions must be nonrandom. Define $\boldsymbol{\beta}_{p+1}$ as the $p+2$ vector with $(j+1)$ element equal to $\mu^{(j)}(\mathsf{x})/j!$ for $j = 0, 1, \ldots, k$ as long as $j \leq S$, and zero otherwise, and $\boldsymbol{B}_{p+1}$ as the n-vector with i^{th} entry $[\mu(X_i) - \boldsymbol{r}_{p+1}(X_i - \mathsf{x})'\boldsymbol{\beta}_{p+1}]$. The bias term appearing in the expansions is then given by (recall (8))

$$\Psi_{\texttt{rbc},F} = \sqrt{nh}\ \nu! \boldsymbol{e}_\nu' \mathbb{E}[\boldsymbol{\Gamma}]^{-1} \left(\mathbb{E}[\boldsymbol{\Omega} \boldsymbol{B}_{p+1}] - \rho^{p+1} \mathbb{E}[\boldsymbol{\Lambda}] \boldsymbol{e}_{p+1}' \mathbb{E}[\bar{\boldsymbol{\Gamma}}]^{-1} \mathbb{E}[\bar{\boldsymbol{\Omega}} \boldsymbol{B}_{p+1}] \right). \tag{11}$$

This bias term is nonrandom but is otherwise nonasymptotic: all expectations are fixed-n and we have not done the typical Taylor expansion. This allows for more general results and will ease implementation.

Our starting point is the following result. See Theorem 3 in Section 5 for a more general uniformly valid Edgeworth expansions.

Theorem 2. *Let Assumptions 1 and 2 hold. Additionally suppose $nh/\log(nh)^{2+\gamma} \to \infty$ and $\Psi_{\texttt{rbc},F} \log(nh)^{1+\gamma} \to 0$, for any γ bounded away from zero uniformly in $\mathscr{F}_S$, and ρ is bounded and* *bounded away from zero uniformly in $\mathscr{F}_S$. Then*

$$\sup_{F\in\mathscr{F}_S} r_{\mathtt{rbc}}^{-1}\Big|\mathbb{P}_F\big[\mu^{(\nu)}(\mathsf{x})\in I_{\mathtt{rbc}}(h)\big]-(1-\alpha)-CE(I_{\mathtt{rbc}}(h),F)\Big|=o(1),$$

where $r_{\mathtt{rbc}}=\max\{(nh)^{-1},\Psi_{\mathtt{rbc},F}^2,(nh)^{-1/2}\Psi_{\mathtt{rbc},F}\}$*, and for quantities* $\omega_{4,\mathtt{rbc},F}$*,* $\omega_{5,\mathtt{rbc},F}$*, and* $\omega_{6,\mathtt{rbc},F}$*, given in the appendix,*

$$CE(I_{\mathtt{rbc}}(h),F)=\frac{1}{nh}2\omega_{4,\mathtt{rbc},F}+2\Psi_{\mathtt{rbc},F}^2\omega_{5,\mathtt{rbc},F}+(nh)^{-1/2}2\Psi_{\mathtt{rbc},F}\omega_{6,\mathtt{rbc},F}. \tag{12}$$

The quantities $\omega_{4,\mathtt{rbc},F}$, $\omega_{5,\mathtt{rbc},F}$, and $\omega_{6,\mathtt{rbc},F}$ are discussed briefly after Theorem 3 below, and more thoroughly in the supplement. They are non-zero uniformly in $\mathscr{F}_S$ and consistently estimable in practice. To fully utilize Theorem 2 we need to characterize the bias term, which depends on the location of x, the parity of $p-\nu$, and the relationship of p to S. We can then optimize $CE(I_{\mathtt{rbc}},F)$ with respect to h to obtain the minimax optimal coverage error decay rate of Theorem 1. This is done in the next subsection.

At this level of generality, we already have two interesting consequences for bandwidth selection. First, the coverage error rate of $I_{\mathtt{rbc}}(h)$ does not depend on ν, echoing the finding in Theorem 1. Because of this, the rate of the inference-optimal h will also not depend on ν, a key fact in constructing a feasible procedure which attains the coverage error minimax bound. Second, only implicit here, is that intervals with faster-decaying coverage error are able to employ larger bandwidths, and thus will have, in general, shorter length in large samples. The length of the interval (4) vanishes proportionally to $n^{-1/2}h^{-1/2-\nu}$, the rate of square root of the variance of centering, which depends on ν.

4.1 Coverage Error Rate Optimality: Choosing h

We now focus on choosing the bandwidth h optimally, leaving ρ and K to the next section. We continue to assume ρ is bounded and bounded away from zero. With pragmatism in mind, we restrict attention to bandwidth sequences that are polynomial in n, that is, of the form $h=Hn^{-\eta}$ for some constants $H>0$ and $\eta>0$. Most if not all practical implementations fall into this form, either in finite samples or asymptotically. Recall from above that an inference-optimal bandwidth (i.e., a choice of h that makes $I_{\mathtt{rbc}}(h)$ minimax optimal) has a decay rate now captured by the exponent η, which does not depend on ν.

However, η must depend on p or, for p large enough, on S and s in order to be optimal, since the key to attaining optimality is the smoothing bias. Correct pointwise coverage requires $\Psi_{\mathtt{rbc},F} = o(1)$, and Theorem 2 requires slightly more in order to obtain higher order results. The rate of convergence (to zero) of $\Psi_{\mathtt{rbc},F}$ can be deduced from Equation (11) by first expanding the $\mu(X_i)$ inside $\boldsymbol{B}_{p+1} = [\mu(X_i) - \boldsymbol{r}_{p+1}(X_i - \mathsf{x})'\boldsymbol{\beta}_{p+1} : i = 1, 2, \dots, n]'$ around x and then specializing to a given p and S. A complete derivation is given in the supplement. Briefly, for a point $\bar{x}$ we have

$$\mu(X_i) - \boldsymbol{r}_{p+1}(X_i - \mathsf{x})'\boldsymbol{\beta}_{p+1} = \sum_{k=S\wedge p+2}^{S} \frac{1}{k!}(X_i - \mathsf{x})^k \mu^{(k)}(\mathsf{x}) + \frac{1}{S!}(X_i - \mathsf{x})^S \left(\mu^{(S)}(\bar{x}) - \mu^{(S)}(\mathsf{x})\right),$$

where the summation is taken to be zero if $p + 1 \geq S$. Substituting this into $\Psi_{\mathtt{rbc},F}$ and extracting the leading terms yields the rate. The final rate will depend on the smoothness, location of x, parity of $p - \nu$, and the bandwidths h and $b = h/\rho$. However, the rate, and hence coverage itself, cannot be improved by letting ρ vanish or diverge: the first term of (11) is unchanged by ρ and if b vanishes faster than h, the effective sample size, and thus the precision, will be the smaller nb instead of the larger nh. We therefore assume ρ is bounded and bounded away from zero in this section, but Section 5 gives more general results.

With ρ bounded and bounded away from zero, the bias is always of the form $\Psi_{\mathtt{rbc},F} = O(\sqrt{nh}h^{\zeta})$ for an exponent ζ which depends on the location of x, the parity of $p - \nu$, and the smoothness. A complete list of ζ is shown in Table 1. From there, we see that if p is large enough relative to S (how large depends on the specific case), then $\zeta = S + s$, implying $\Psi_{\mathtt{rbc},F} = O(\sqrt{nh}h^{S+s})$. Therefore, setting $h = Hn^{-\eta}$ with $\eta = 1/(S + s + 1)$ renders $CE(I_{\mathtt{rbc}}(n^{-\eta}), F) = O(n^{-(S+s)/(S+s+1)}) = O(r_\star)$ uniformly in $\mathscr{F}_S$, implying minimax optimality. However, as mentioned above, this requires knowledge of S and s, to choose p large enough and then h accordingly, and hence is not valuable in applications.

The more empirically relevant case is to treat p as fixed and smaller than S, specifically $p \leq S-3$ for interior x with $p - \nu$ odd and $p \leq S - 2$ otherwise (i.e. for boundary points or if x is an interior point with $p - \nu$ even). In these cases, we can use the Taylor expansion above to characterize the leading term, and write $\Psi_{\mathtt{rbc},F} = \sqrt{nh}h^{\zeta}\psi_{\mathtt{rbc},F}[1 + o(1)]$ where $\zeta = p + 3$ for interior x with $p - \nu$ odd and $p + 2$ otherwise. The term $\psi_{\mathtt{rbc},F}$ will be referred to as the constant term for simplicity, though technically it is a nonrandom sequence with known form, uniformly bounded in $\mathscr{F}_S$, and

Location of x	Parity of $p-\nu$	Smoothness	ζ	$\psi_{\mathtt{rbc},F}$
Boundary	Odd or Even	$p+2 \leq S$	$p+2$	Equation (13a)
		$p+2 > S$	$S+s$	N/A
Interior	Even	$p+2 \leq S$	$p+2$	Equation (13b)
		$p+2 > S$	$S+s$	N/A
	Odd	$p+3 \leq S$	$p+3$	Equation (13c)
		$p+2 \geq S$	$S+s$	N/A

Table 1: Bias Terms in All Cases For Bias-Corrected Centering $\hat{\theta}_{\mathtt{rbc}}$. With ρ bounded and bounded away from zero, $\Psi_{\mathtt{rbc},F} = O(\sqrt{nh}h^{\zeta})$ and further, if p is small relative to S, $\Psi_{\mathtt{rbc},F} = \sqrt{nh}h^{\zeta}\psi_{\mathtt{rbc},F}[1+o(1)]$.

nonzero for some $F \in \mathscr{F}_S$. Referring to Table 1 for the different cases, $\psi_{\mathtt{rbc},F}$ can be

$$\begin{cases} \dfrac{\mu^{(p+2)}}{(p+2)!}\nu!\boldsymbol{e}_\nu'\mathbb{E}[\boldsymbol{\Gamma}]^{-1}\Big\{\mathbb{E}[\boldsymbol{\Lambda}_2] - \rho^{-1}\mathbb{E}[\boldsymbol{\Lambda}_1]\boldsymbol{e}_{p+1}'\mathbb{E}[\bar{\boldsymbol{\Gamma}}]^{-1}\mathbb{E}[\bar{\boldsymbol{\Lambda}}_1]\Big\}, & \text{(13a)} \\ \dfrac{\mu^{(p+2)}}{(p+2)!}\nu!\boldsymbol{e}_\nu'\mathbb{E}[\boldsymbol{\Gamma}]^{-1}\mathbb{E}[\boldsymbol{\Lambda}_2], \qquad \text{or} & \text{(13b)} \\ \nu!\boldsymbol{e}_\nu'\mathbb{E}[\boldsymbol{\Gamma}]^{-1}\Big\{\dfrac{\mu^{(p+2)}}{(p+2)!}\Big[h^{-1}\mathbb{E}[\boldsymbol{\Lambda}_2] - \rho^{-2}b^{-1}\mathbb{E}[\boldsymbol{\Lambda}_1]\boldsymbol{e}_{p+1}'\mathbb{E}[\bar{\boldsymbol{\Gamma}}]^{-1}\mathbb{E}[\bar{\boldsymbol{\Lambda}}_1]\Big] \\ \qquad\qquad + \dfrac{\mu^{(p+3)}}{(p+3)!}\Big[\mathbb{E}[\boldsymbol{\Lambda}_3] - \rho^{-2}\mathbb{E}[\boldsymbol{\Lambda}_1]\boldsymbol{e}_{p+1}'\mathbb{E}[\bar{\boldsymbol{\Gamma}}]^{-1}\mathbb{E}[\bar{\boldsymbol{\Lambda}}_2]\Big]\Big\}, & \text{(13c)} \end{cases}$$

where $\boldsymbol{\Lambda}_k = \boldsymbol{\Omega}[X_{h,1}^{p+k}, \cdots, X_{h,n}^{p+k}]'/n$ and $\bar{\boldsymbol{\Lambda}}_k = \bar{\boldsymbol{\Omega}}[X_{b,1}^{p+1+k}, \ldots, X_{b,n}^{p+1+k}]'/n$, and hence in particular $\boldsymbol{\Lambda}_1 \equiv \boldsymbol{\Lambda}$ as defined in Section 2.

For implementation purposes, we optimize $CE(I_{\mathtt{rbc}}, F)$ pointwise in F. The optimal bandwidths will be functions of F, and their implementations are functions of the data; neither depend explicitly upon $\mathscr{F}_S$. The resulting coverage error rates will still hold uniformly, because the bandwidths are of the form $h = Hn^{-\eta}$, where η does not depend on F and H is well-behaved in $\mathscr{F}_S$.

An obvious candidate for h in applications is the classical MSE-optimal choice, $h_{\mathtt{mse}} = H_{\mathtt{mse}}n^{-1/(2p+3)}$, where $H_{\mathtt{mse}}$ balances the variance against the squared bias. This choice is popular and readily available in most statistical software. It yields valid robust bias corrected inference, that is, $\sup_{F\in\mathscr{F}_S} |\mathbb{P}_F[\mu^{(\nu)}(\mathsf{x}) \in I_{\mathtt{rbc}}(h_{\mathtt{mse}})] - (1-\alpha)| \to 0$. An interesting consequence of Theorems 1 and 2 is that for interior points and local linear regression the mean-square optimal bandwidth is not only valid for robust bias corrected inference, but also rate optimal, in the sense that using $h_{\texttt{mse}}$ yields the best possible coverage error rate $n^{-(p+3)/(p+4)}$. However, this optimality does *not* hold for $p \neq 1$, or for x a boundary point, though $h_{\texttt{mse}}$ retains validity of the confidence intervals at a suboptimal coverge error rate decay. Even when it is not optimal for coverage, $I_{\texttt{rbc}}(h_{\texttt{mse}})$ has the advantage that it can be reported along with $\hat{\mu}_p^{(\nu)}(\mathsf{x}; h_{\texttt{mse}})$, pairing an optimal point estimator with a valid measure of uncertainty that uses the same samples.

When the goal is inference improved bandwidth choices can be developed. We discuss two natural criteria for choosing an inference-optimal h: first attaining coverage error optimality and second sacrificing optimality for a reduction in length.

If minimizing coverage error alone is the goal, we can construct an inference-optimal bandwidth using (12). To do so, we set $h_{\texttt{rbc}} = Hn^{-\eta_{\texttt{rbc}}}$ for $\eta_{\texttt{rbc}} = 1/(p+4)$ for interior x with $p - \nu$ odd and $\eta_{\texttt{rbc}} = 1/(p+3)$ otherwise. With these choices the rate $r_{\texttt{rbc}}$ of Theorem 2 equals $r_\star$ of Theorem 1. In terms of rates, $h_{\texttt{rbc}}$ balances the variance and bias of the point estimator, instead of the squared bias as in MSE optimality. Any H bounded and bounded away from zero uniformly in $\mathscr{F}_S$ will yield the minimax optimal rate of Theorem 1. A natural choice for practice is to minimize the constant portion of the coverage error of (12), where we apply $h_{\texttt{rbc}} = Hn^{-\eta_{\texttt{rbc}}}$ and substitute $\sqrt{nh}h^{\zeta}\psi_{\texttt{rbc},F}$ for $\Psi_{\texttt{rbc},F}$, then factor out $r_\star$:

$$H_{\texttt{rbc}} = \operatorname*{arg\,min}_{H>0} \left| H^{-1}\left\{2\omega_{4,\texttt{rbc},F}\right\} + H^{1+2\zeta}\left\{2\psi^2_{\texttt{rbc},F}\omega_{5,\texttt{rbc},F}\right\} + H^{\zeta}\left\{2\psi_{\texttt{rbc},F}\omega_{6,\texttt{rbc},F}\right\} \right|.$$

It is straightforward to give a data-driven version of $H_{\texttt{rbc}}$, and therefore of $h_{\texttt{rbc}}$, because all quantities involved can be estimated. We defer the details to the supplement to conserve space. In a nutshell, plug-in estimators can be constructed, denoted by $\hat{\omega}_{4,\texttt{rbc},F}$, $\hat{\omega}_{5,\texttt{rbc},F}$ and $\hat{\omega}_{6,\texttt{rbc},F}$, as well as an estimate of the bias constant, $\hat{\psi}_{\texttt{rbc},F}$. We then numerically solve

$$\hat{H}_{\texttt{rbc}} = \operatorname*{arg\,min}_{H>0} \left| H^{-1}\left\{2\hat{\omega}_{4,\texttt{rbc},F}\right\} + H^{1+2\zeta}\left\{2\hat{\psi}^2_{\texttt{rbc},F}\hat{\omega}_{5,\texttt{rbc},F}\right\} + H^{\zeta}\left\{2\hat{\psi}_{\texttt{rbc},F}\hat{\omega}_{6,\texttt{rbc},F}\right\} \right|.$$

Because this bandwidth depends on the specific data-generating process F, we view it as a rule-of-thumb implementation for minimax optimality.

As an alternative bandwidth selector, we look for a *trade-off* bandwidth $h_{\texttt{to}} = H_{\texttt{to}}n^{-\eta_{\texttt{to}}}$ such that not only does $I_{\texttt{rbc}}(h_{\texttt{to}})$ have uniformly correct coverage, but also its length $|I_{\texttt{rbc}}(h_{\texttt{to}})|$ contracts more quickly than $I_{\texttt{rbc}}(h_{\texttt{rbc}})$. From Theorem 2, the coverage of $I_{\texttt{rbc}}(n^{-\eta})$ is (uniformly) asymptotically correct for a wide range of η, but length is reduced for larger bandwidths, meaning smaller exponents η. For any $\eta > \eta_{\texttt{rbc}}$ (i.e. $h = o(h_{\texttt{rbc}})$), both the rate of coverage error decay and interval length contraction can be improved. Therefore, in considering trade-off bandwidths, we will only consider $\eta_{\texttt{to}} \in (1/(1+2\zeta), \eta_{\texttt{rbc}}]$, where recall that $\zeta = p + 3$ for interior points with $p - \nu$ odd and $\zeta = p + 2$ otherwise. There is no well-defined optimal choice in this range of asympotically valid options, as the choice must reflect each researcher's preference for length vs. coverage error. This range does not depend on ν, even though the resulting length will, see Eqn. (14) below. This may affect how the researcher wishes to trade off the two quantities.

To select the constant, $H_{\texttt{to}}$, note first that for $\eta < \eta_{\texttt{rbc}}$ the middle term of the coverage error (12) is dominant. This term, $n^{1-\eta_{\texttt{to}}(1+2\zeta)}\{2\psi^2_{T,F}\omega_{5,T,F}\}$, shares the rate of the scaled, squared bias. Therefore, it is natural to balance this against the square of interval length, to match the trade off that $h_{\texttt{rbc}}$ represents. The feasible choice of this constant, $\hat{H}_{\texttt{to}}$, will also be a direct plug-in rule that uses the estimators above and a pilot version of $\hat{\sigma}^2_{\texttt{rbc}}$, as well a researcher's choice of weight $\mathcal{H} \in (0, 1)$. Putting altogether, we can then set

$$\begin{aligned}
\hat{H}_{\texttt{to}} &= \operatorname*{arg\,min}_{H>0} \mathcal{H} \times H^{1+2\zeta}\{2\hat{\psi}^2_{\texttt{rbc},F}\hat{\omega}_{5,\texttt{rbc},F}\} + (1-\mathcal{H}) \times 4z^2_{\alpha/2}\frac{\hat{\sigma}^2_{\texttt{rbc}}}{H^{1+2\nu}} \\
&= \left(\frac{(1-\mathcal{H})(1+2\nu)4z^2_{\alpha/2}\hat{\sigma}^2_{\texttt{rbc}}}{\mathcal{H}(1+2\zeta)2\hat{\psi}^2_{\texttt{rbc},F}\hat{\omega}_{5,\texttt{rbc},F}}\right).
\end{aligned}$$

The resulting data-driven bandwidth choice is $\hat{h}_{\texttt{to}} = \hat{H}_{\texttt{to}} n^{-\eta_{\texttt{to}}}$, for a choice $\eta_{\texttt{to}} \in (1/(1+2\zeta), \eta_{\texttt{rbc}}]$, and trade-off weights $\mathcal{H} \in (0, 1)$. The supplement contains details and some additional results.

4.2 Interval Length Optimality: Choosing ρ and $K(\cdot)$

To further improve $I_{\texttt{rbc}}(h)$ we need to select the bias-correction bandwidth b, which we do in the form of $\rho = h/b$, and the kernel function $K(\cdot)$. Specifically, we optimize the length of the resulting confidence interval:

$$|I_{\texttt{rbc}}(h)| = 2z_{\alpha/2}\hat{\vartheta}_{\texttt{rbc}} = 2z_{\alpha/2}\frac{\hat{\sigma}_{\texttt{rbc}}}{\sqrt{nh^{1+2\nu}}}, \tag{14}$$

as a function of those two choices.

With ρ bounded and bounded away from zero, this choice affects only the constant portions of the coverage error expansion of $I_{\mathtt{rbc}}(h_{\mathtt{rbc}})$, in particular changing the shape of the *equivalent kernel* of $\hat{\theta}_{\mathtt{rbc}}$. To find this equivalent kernel, begin by writing $\hat{\theta}_{\mathtt{rbc}} = \nu!\boldsymbol{e}_\nu'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}_{\mathtt{rbc}}\boldsymbol{Y}/nh^\nu$ as a weighted average of the Y_i. Recall that $X_{h,i} = (X_i - \mathsf{x})/h$ and similarly for $X_{b,i}$. Then

$$\begin{aligned}\hat{\theta}_{\mathtt{rbc}} &= \frac{1}{nh^\nu}\nu!\boldsymbol{e}_\nu'\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\Omega} - \rho^{p+1}\boldsymbol{\Lambda}\boldsymbol{e}_{p+1}'\bar{\boldsymbol{\Gamma}}^{-1}\bar{\boldsymbol{\Omega}}\right)\boldsymbol{Y} \\ &= \frac{1}{nh^{1+\nu}}\sum_{i=1}^{n}\left\{\nu!\boldsymbol{e}_\nu'\boldsymbol{\Gamma}^{-1}\left(K(X_{h,i})\boldsymbol{r}_p(X_{h,i}) - \rho^{p+1}\frac{h}{b}\boldsymbol{\Lambda}\boldsymbol{e}_{p+1}'\bar{\boldsymbol{\Gamma}}^{-1}K(X_{b,i})\boldsymbol{r}_{p+1}(X_{b,i})\right)\right\}Y_i.\end{aligned}$$

The weights here depend on the sample, as $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, and $\bar{\boldsymbol{\Gamma}}$ are sample quantities. The equivalent kernel replaces these with their limiting versions (not, as elsewhere, their fixed-n expectations), which we shall denote $\boldsymbol{\mathsf{G}} = f(\mathsf{x})\int K(u)\boldsymbol{r}_p(u)\boldsymbol{r}_p(u)'du$, $\boldsymbol{\mathsf{L}} = f(\mathsf{x})\int K(u)\boldsymbol{r}_p(u)u^{p+1}du$, and $\bar{\boldsymbol{\mathsf{G}}} = f(\mathsf{x})\int K(u)\boldsymbol{r}_{p+1}(u)\boldsymbol{r}_{p+1}(u)'du$, respectively. The integrals are over $[-1,1]$ if x is an interior point and appropriately truncated when x is a boundary point. Under our assumptions, convergence to these limits is fast enough that, for the equivalent kernel $\mathcal{K}_{\mathtt{rbc}}(u;K,\rho,\nu)$ defined as

$$\mathcal{K}_{\mathtt{rbc}}(u;K,\rho,\nu) = \nu!\boldsymbol{e}_\nu'\boldsymbol{\mathsf{G}}^{-1}\left(K(u)\boldsymbol{r}_p(u) - \rho^{p+2}\boldsymbol{\mathsf{L}}\boldsymbol{e}_{p+1}'\bar{\boldsymbol{\mathsf{G}}}^{-1}K(u\rho)\boldsymbol{r}_{p+1}(u\rho)\right),$$

we have the representation

$$\hat{\theta}_{\mathtt{rbc}} = \frac{1}{nh^{1+\nu}}\sum_{i=1}^{n}\mathcal{K}_{\mathtt{rbc}}\big(X_{h,i};K,\rho,\nu\big)Y_i\left\{1 + o_{\mathbb{P}}(1)\right\}.$$

For more details on equivalent kernels, see Fan and Gijbels (1996, Sect. 3.2.2). It follows that the (constant portion of) the asymptotic length of $I_{\mathtt{rbc}}(h)$ depends on $K(\cdot)$ and ρ only through the specific functional $\int\big(\mathcal{K}_{\mathtt{rbc}}(u;K,\rho,\nu)\big)^2du$, which corresponds to the asymptotic variance of a local polynomial point estimator.

Cheng, Fan and Marron (1997) show that the asymptotic variance of a local polynomial point estimator at a boundary or interior point is minimized by employing the uniform kernel. Therefore, to minimize the constant term of interval length we choose ρ, depending on K, to make $\mathcal{K}_{\mathtt{rbc}}(u;K,\rho,\nu)$ as close as possible to the optimal equivalent kernel, i.e. the $\mathcal{K}_p^*(u)$ induced by the uniform kernel for a given p. If the uniform kernel is used initially, then $\rho^* = 1$ is optimal: that is,

$\mathcal{K}_{\texttt{rbc}}(\cdot; \mathbb{1}\{|u| < 1\}/2, 1, \nu) \equiv \mathcal{K}^*_{p+1}(\cdot)$. This highlights the importance of being able to accommodate the uniform kernel in our higher-order expansions. If a kernel other than uniform is used, we look for the optimal choice of ρ by minimizing the L_2 distance between the induced equivalent kernel and the optimal variance-minimizing equivalent kernel, solving

$$\rho^* = \operatorname*{arg\,min}_{\rho>0} \int \left|\mathcal{K}_{\texttt{rbc}}\big(u; K, \rho, \nu\big) - \mathcal{K}^*_{p+1}(u)\right|^2 du.$$

This is not a sample-dependent problem, only computational. For $p - \nu$ odd, the standard case in practice, Table 2 shows the optimal ρ^*, for boundary and interior points, respectively, the triangular kernel ($K(u) = (1 - |u|)\mathbb{1}(|u| \leq 1)$) and the Epanechnikov kernel ($K(u) = 0.75(1 - u^2)\mathbb{1}(|u| \leq 1)$). These two are popular choices and are MSE-optimal at boundary and interior points, respectively. The shapes of the resulting equivalent kernel, $\mathcal{K}_{\texttt{rbc}}(u; K, \rho^*, \nu)$, are shown in Figure 1 for $\nu = \{0, 1\}$. Note that although ρ^* itself does not vary with ν, the equivalent kernel shape does. Additional choices of p are illustrated in the supplement.

5 Uniformly Valid Edgeworth Expansions and Other Coverage Error Consequences

The primary technical ingredients behind Theorems 1 and 2, and the main technical contribution of this paper, are Edgeworth expansions that hold uniformly in $\mathscr{F}_S$. We now state and discuss these results, including the technical innovation of allowing the uniform kernel. We then use these general results to obtain the coverage error of any $I \in \mathscr{I}_p$, from which we can identify suboptimal interval estimators, further illustrate the gap between point and interval estimation, and reach other interesting conclusions.

Edgeworth expansions are a long-standing tool for more detailed analyses of asymptotic distributional approximations, named after the author of a series of papers on the idea, beginning with Edgeworth (1883) and treated more extensively in Edgeworth (1906). Relative to all previous technical work on Edgeworth expansions for nonparametric smoothing methods, we extend this higher-order distributional analysis in three directions: (i) to hold uniformly over $\mathscr{F}_S$, (ii) to allow for the uniform kernel, and (iii) to allow for any derivative $\nu \geq 0$.

Edgeworth expansions are almost always established pointwise in the underlying distribution, that is, for a single, fixed F, as $n \to \infty$. Indeed, standard references on the subject (Bhattacharya and Rao, 1976; Hall, 1992a) do not even mention uniformity. Uniformly valid expansions have some precedence in the literature when studying notions of optimality, e.g., Beran (1982), and most relevantly, Hall and Jing (1995), but these results are rare and confined to parametric models. In this section we establish uniformly valid Edgeworth expansions for kernel-based nonparametric t-statistics, which do not appear to have a direct antecedent in the literature.

The uniform kernel is explicitly ruled out in all prior work on Edgeworth expansions for kernel-based nonparametrics, both for density estimation (Hall, 1991, 1992b,a) and regression (Chen and Qin, 2000, 2002; Calonico, Cattaneo and Farrell, 2018). Other work on nonparametric regression (Hall, 1992c; Neumann, 1997) employed a fixed design, effectively assuming the issue away. In fact, Hall (1991, p. 218) conjectured that valid Edgeworth expansions would require techniques for lattice-valued random variables if the uniform kernel was used. On the contrary, we show here that this is not needed. Our new Edgeworth expansions allowing for the uniform kernel are important because it would not be possible to optimize the interval length as in Section 4.2 otherwise. Finally, allowing for the uniform kernel may have practical appeal because unweighted local least squares regression is a popular choice in some applications.

We require some additional notation to state results. The Edgeworth expansions are for the asymptotic distribution of the t statistics that are dual to the intervals $I \in \mathscr{I}_p$. For each interval of the form (4), we define

$$T = \frac{\hat{\theta} - \mu^{(\nu)}}{\hat{\vartheta}}.$$

The terms in the expansion are specific to a given T and F. All formulae are listed in the appendix and are derived and discussed in the supplement. First, we denote the bias of $\sqrt{nh}(\hat{\theta} - \mu^{(\nu)})$, the scaled numerator of T, by $\Psi_{T,F}$, which is computed for fixed-n, but is nonrandom. Equation (11) already showed this quantity for $\hat{\theta}_{\texttt{rbc}}$. The version for $\hat{\theta} = \hat{\mu}_p^{(\nu)}$ is $\Psi_{T_p,F} = \sqrt{nh}\ \nu!\boldsymbol{e}_\nu'\mathbb{E}[\boldsymbol{\Gamma}]^{-1}\mathbb{E}[\boldsymbol{\Omega}\boldsymbol{B}_p]$, where $\boldsymbol{B}_p$ is defined as in Equation (11) with p in place of $p+1$ ($\boldsymbol{\beta}_p$ is similarly defined). It holds that $\Psi_{T_p,F} = O(\sqrt{nh}h^{\zeta})$ uniformly in $\mathscr{F}_S$ where ζ varies depending on smoothness, parity of $p-\nu$, and location of x, similarly to the discussion in Section 4. The supplement gives complete details in all cases. The notation $\Psi_{T,F}$ is generic, encompassing all cases of smoothness, parity of $p-\nu$, and location of x. Second, there are six terms $\omega_{k,T,F}(z)$, $k = 1, 2, \ldots, 6$, which are functions of the chosen quantiles. All that is important at present is that they are nonrandom, known for all the t-statistics under consideration, bounded uniformly in $\mathscr{F}_S$, and bounded away from zero for at least some F. For coverage it is important that ω_1, ω_2, and ω_3 are even functions of z, while ω_4, ω_5, and ω_6 are odd.

We let $\lambda_{T,F}$ capture the mismatch between the variance of the numerator of the t-statistic and the population standardization chosen. There are too many possibilities in $\mathscr{I}_p$ to list all possible $\lambda_{T,F}$. Any of the asymptotic approximations to $\mathbb{V}[\hat{\theta}|X_1, \ldots, X_n]$ may yield nonzero $\lambda_{T,F}$. Specifically, at boundary points, Chen and Qin (2002) find $\lambda_{T,F}$ is of order exactly h at boundary points when using asymptotic approximations, whereas Calonico et al. (2018) prove that the fixed-n Studentizations (10) yield $\lambda_{T,F} \equiv 0$. For other Studentizations, the rates and constants may change, but control of worst-case coverage, our ultimate goal, cannot be improved beyond the fixed-n forms of $\hat{\sigma}_p$ and $\hat{\sigma}_{\texttt{rbc}}$. Let λ_T be such that $\sup_{F \in \mathscr{F}_S} \lambda_{T,F} = O(\lambda_T) = o(1)$.

With these terms all defined, let

$$\begin{aligned} E_{T,F}(z) = {} & \frac{1}{\sqrt{nh}} \omega_{1,T,F}(z) + \Psi_{T,F} \omega_{2,T,F}(z) + \lambda_{T,F} \omega_{3,T,F}(z) \\ & + \frac{1}{nh} \omega_{4,T,F}(z) + \Psi_{T,F}^2 \omega_{5,T,F}(z) + (nh)^{-1/2} \Psi_{T,F} \omega_{6,T,F}(z). \end{aligned} \tag{15}$$

The main technical result of this paper is the following Edgeworth expansion.

Theorem 3. *Let Assumptions 1 and 2 hold. Let $r_T = \max\{(nh)^{-1}, \Psi_{T,F}^2, (nh)^{-1/2}\Psi_{T,F}, \lambda_T\}$, i.e. the slowest vanishing of the rates. Assume, for any γ bounded away from zero uniformly in $\mathscr{F}_S$,*

$$nh / \log(nh)^{2+\gamma} \to \infty, \qquad \Psi_{T,F} \log(nh)^{1+\gamma} \to 0, \qquad \limsup_{n \to \infty} \sup_{F \in \mathscr{F}_S} \rho < \infty. \tag{16}$$

Then

$$\sup_{F \in \mathscr{F}_S} \sup_{z \in \mathbb{R}} r_T^{-1} \Big| \mathbb{P}_F \left[T < z \right] - \Phi(z) - E_{T,F}(z) \Big| = o\left(1\right).$$

This result is fully general, covering interior and boundary points, $p - \nu$ even and odd, any derivative $\nu \geq 0$, and all smoothness cases. Different settings yield different results in terms of the final rates, but they share a common structure as shown here. The assumptions are strengthened relative to typical pointwise first-order analyses only by $\log(nh)$ factors on the bandwidth and the other uniformity requirements of Assumption 1. To recover Theorem 2, specialize to the interval chosen there, $I_{\mathtt{rbc}}(h)$, define $CE(I_{\mathtt{rbc}}, F) = E_{T,F}(z_{\alpha/2}) - E_{T,F}(z_{\alpha/2})$, for the t statistic $(\hat{\theta}_{\mathtt{rbc}} - \mu^{(\nu)})/\hat{\vartheta}_{\mathtt{rbc}}$, and impose the additional conditions in that theorem.

A crucial piece in the proof of Theorem 3 is establishing that the appropriate Cramér's condition holds under Assumption 2. Hall (1991) appears to be the first to use this type of assumption to verify the appropriate Cramér's condition in the context of kernel smoothing, but because the uniform kernel is allowed herein, our result may be of independent interest. Prior work has often assumed Cramér's condition directly (Neumann, 1997; Calonico, Cattaneo and Farrell, 2018), or used linear independence of a vector of the form $(1, K(u), uK(u), \ldots)'$ either explicitly (Chen and Qin, 2000, 2002) or implicitly (Hall, 1991). This linear independence fails when K is uniform and u runs over the support of $K(u)$.

The key insight we exploit herein is that previous approaches ignored the region *outside* the support of $K(\cdot)$ but *inside* the neighborhood of Assumption 1. Loosely speaking, $(1, K(u), uK(u), \ldots)'$ may be linearly *dependent* on $u \in [-1, 1]$ (when K is uniform), but $(1, K(\frac{x-\mathsf{x}}{h}), (\frac{x-\mathsf{x}}{h})K(\frac{x-\mathsf{x}}{h}), \ldots)'$ is linearly *independent* on x in a fixed neighborhood of x. The following lemma gives a simplified case to illustrate illustrates our key observation and how this differs from approaches in the literature. The supplement gives the general result, which is used to prove Theorem 3.

Lemma 1. *Let the conditions of Theorem 3 hold and assume $v(x)$ is known. Consider T_p for $\nu = 0$, $p = 1$, and interior x. Let $\xi_Z(\boldsymbol{t})$ be the characteristic function of the random vector*

$$\boldsymbol{Z}_i = \Big([K(X_{h,i})](1, X_{h,i}, X_{h,i}^2)' \, , \; [K(X_{h,i})(Y_i - \mu - X_i\mu^{(1)})](1, X_{h,i})' \, , \; [K(X_{h,i})^2 v(X_i)](1, X_{h,i}, X_{h,i}^2)' \Big) \, .$$

For h sufficiently small, for all $C_1 > 0$ there is a $C_2 > 0$ such that

$$\sup_{|\boldsymbol{t}| > C_1} |\xi_Z(\boldsymbol{t})| < 1 - C_2 h.$$

Proof. The key first step is to bound the characteristic function separately depending on whether X_i is local to x. Note that h is fixed. The characteristic function of $\boldsymbol{Z}_i$ is

$$\xi_Z(\boldsymbol{t}) = \mathbb{E}[\exp\big(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i\big)] = \mathbb{E}\left[\exp\big(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i\big)\mathbb{1}\big(|X_{h,i}| > 1\big)\right] + \mathbb{E}\left[\exp\big(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i\big)\mathbb{1}\big(|X_{h,i}| \leq 1\big)\right], \qquad (17)$$

where $\mathrm{i} = \sqrt{-1}$. First, if $|X_{h,i}| > 1$, then $K(X_{h,i}) = 0$, and so $\boldsymbol{Z}_i$ is the zero vector and $\exp\big(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i\big) =$

1. Therefore, $\mathbb{E}[\exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i)\mathbb{1}(|X_{h,i}|>1)] = \mathbb{P}[|X_i - \mathsf{x}| > h]$. The neighborhood of Assumption 1 where the density of X is bounded and bounded away from zero contains $\{x : |x - \mathsf{x}| \leq h\}$, and therefore

$$\mathbb{P}[|X_i - \mathsf{x}| > h] = 1 - \int_{\mathsf{x}-h}^{\mathsf{x}+h} f(x)dx \leq 1 - h2\left(\min_{x:|x-\mathsf{x}|\leq h} f(x)\right) = 1 - C_3 h. \tag{18}$$

Next, consider the event that $|X_{h,i}| \leq 1$. Write $\boldsymbol{Z}_i = \boldsymbol{Z}_i(X_{h,i}, Y_i)$. By a change of variables $U = (X - \mathsf{x})/h$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i(X_{h,i},Y_i)\right)\mathbb{1}(|X_{h,i}|\leq 1)\right] &= \int\int_{\mathsf{x}-h}^{\mathsf{x}+h} \exp\left(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i(x,y)\right) f_{xy}(x,y)dxdy \\ &= h\int\int_{-1}^{1} \exp\left(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i(u,y)\right) f_{xy}(\mathsf{x}+uh,y)dudy.\end{aligned}$$

Suppose that K is not the uniform kernel. The assumption of linear independence implies that $\boldsymbol{Z}_i$ is a set of linearly independent and continuously differentiable functions $\{[-1,1]\}\cup\mathbb{R}$. Furthermore, the density of (U,Y), as random variables on $\{[-1,1]\}\cup\mathcal{Y}$, for some $\mathcal{Y}\subset\mathbb{R}$, is strictly positive. Therefore, by Bhattacharya (1977, Lemma 1.4), $\boldsymbol{Z}_i = \boldsymbol{Z}_i(U,Y)$ obeys Cramér's condition as a function of random variables on $\{[-1,1]\}\cup\mathcal{Y}$, and so there is some $C>0$ such that

$$\sup_{|\boldsymbol{t}|>C}\left|\int\int_{-1}^{1} \exp\left(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i(u,y)\right) f_{xy}(\mathsf{x}+uh,y)dudy\right| < 1, \tag{19}$$

according to Bhattacharya and Rao (1976, p. 207). Collecting Equations (17), (18), and (19) yields the result when the kernel is not uniform.

If K is the uniform kernel, Equation (19) still holds, as follows. The first element of $\boldsymbol{Z}_i(U,Y)$ is $K(U)$. Rewrite $\boldsymbol{Z}_i(U,Y)$ as $\boldsymbol{Z}_i(U,Y) := 2(K(U), \tilde{\boldsymbol{Z}}_i')'$ and partition $\boldsymbol{t}\in\mathbb{R}^{\dim(\boldsymbol{Z})}$ as $\boldsymbol{t} = (t_{(1)}, \tilde{\boldsymbol{t}}')'$. Then, because $K(U)\equiv 1/2$ for $U\in[-1,1]$,

$$\begin{aligned}\int\int_{-1}^{1} \exp\left(\mathrm{i}\boldsymbol{t}'\boldsymbol{Z}_i(u,y)\right) f_{xy}(\mathsf{x}+uh,y)dudy &= \int\int_{-1}^{1} \exp\left(\mathrm{i}\boldsymbol{t}'\left[(1,\tilde{\boldsymbol{Z}}_i')'\right]\right) f_{xy}(\mathsf{x}+uh,y)dudy \\ &= e^{\mathrm{i}t_1}\int\int_{-1}^{1} \exp\left(\mathrm{i}\tilde{\boldsymbol{t}}'\tilde{\boldsymbol{Z}}_i\right) f_{xy}(\mathsf{x}+uh,y)dudy.\end{aligned}$$

Bhattacharya (1977, Lemma 1.4) applies to $\tilde{\boldsymbol{Z}}_i$ and $|e^{it_1}|$ is bounded by one, thus yielding (19). □

5.1 Suboptimal Intervals and Other Findings

Theorem 3 is behind the characterization of the minimax bound in Theorem 1 and the fact that robust bias correction can always attain this bound, as established in Theorem 2. It is also useful to know that some other interval estimators are not optimal, helping practitioners avoid inferior inference. We now discuss such results.

Theorem 3 immediately implies that

$$\sup_{F\in\mathscr{F}_S} r_T^{-1}\Big|\mathbb{P}_F\big[\mu^{(\nu)}\in I\big]-(1-\alpha)-\big[E_{T,F}(z_u)-E_{T,F}(z_l)\big]\Big| = o\,(1)\,,$$

provided (i) z_u, z_l are chosen such that $\Phi(z_u)-\Phi(z_l)=(1-\alpha)$, (ii) the bandwidth(s) obey the requirements of (16). If these conditions fail, coverage error persists asymptotically:

$$\sup_{F\in\mathscr{F}_S}\Big|\mathbb{P}_F\big[\mu^{(\nu)}\in I\big]-(1-\alpha)\Big|\asymp 1,$$

where we employ the notation $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$.

We learn several generic lessons immediately. First, we see the well-known result that symmetric intervals, with $z_l = -z_u$, have superior coverage, due to the evenness of ω_1, ω_2, and ω_3 as functions of the quantile. Asymmetric choices that still have $\Phi(z_u)-\Phi(z_l) = 1-\alpha$ can still yield correct coverage, but the error will vanish more slowly. Bootstrap based quantiles will, in general, not improve coverage error rates in nonparametric contexts beyond the symmetric case (Hall, 1992b), and can in fact be detrimental for coverage error (Hall and Kang, 2001).

Second, interval estimators with $\lambda_{T,F}\equiv 0$ are superior. That is, there should not be a "mismatch" between the population variability of the centering and the population standardization. For any $\nu \geq 0$, and for both interior and boundary points, our expansions prove that the fixed-n standard errors of (10) achieve $\lambda_{T,F}\equiv 0$, and are therefore demonstrably superior choices from the point of view of controlling worst-case coverage. This result generalizes the pointwise finding of Calonico, Cattaneo and Farrell (2018) for $\mu^{(0)}$, which was the first theoretical proof that fixed-n (or "preasymptotic") Studentization is superior for inference.

This is particularly important at boundary points and has lead to confusion in the prior literature. A headline finding of Chen and Qin (2000) is that an empirical likelihood confidence interval estimator has coverage error of the same order at interior and boundary points, which is claimed (in the abstract) to be a "significant improvement over confidence intervals based directly on the asymptotic normal distribution". This claim is based on work by the same authors (Chen and Qin, 2002) who study, in our notation, the interval with centering $\theta = \hat{\mu}_1^{(0)}$ and scaling either $\hat{\vartheta} = (nh^{1+2\nu})^{-1/2}\hat{v}(\mathsf{x})\hat{f}(\mathsf{x})^{-1}\mathcal{V}$, for $\hat{v}(\mathsf{x})$ and $\hat{f}(\mathsf{x})$ given therein, or its population analogue $(nh^{1+2\nu})^{-1/2}v(\mathsf{x})f(\mathsf{x})^{-1}\mathcal{V}$, where $v(\mathsf{x})f(\mathsf{x})^{-1}\mathcal{V}$ is the probability limit of $\mathbb{V}[(nh^{1+2\nu})^{1/2}\hat{\theta} \mid X_1, \ldots, X_n]$. They find that, for both, $\lambda_T \asymp h$ at boundary points. This finding is entirely due to using an asymptotic standardization as opposed to a fixed-n one, and thus empirical likelihood, in particular, does *not* offer higher-order improvements over normality-based intervals.

Third, as long as p is small relative to S, the classic undersmoothing approach is dominated in a minimax sense. This can be seen in any given special case, using the appropriate bias calculation (see supplement). If $p - \nu$ is odd, the fastest the coverage error of an undersmoothed interval can vanish is $n^{-(p+1)/(p+2)}$, attained with $h \asymp n^{-1/(p+2)}$. If x is an interior point and $p - \nu$ is even, the best rate is $n^{-(p+1)/(p+2)}$. Comparing to Theorem 1, we find that no undersmoothed interval, for any choice of quantile, is minimax optimal in the sense of (3). This is a substantial strengthening and generalization of Calonico, Cattaneo and Farrell (2018), which proved, in a pointwise sense and only for the level of the regression function, that robust bias correction is as good or better than undersmoothing for coverage for a given bandwidth sequence. If both S and s are known, both p and h can be chosen in terms of these quantities to yield an undersmoothed interval that is minimax optimal.

Finally, we can see again the discrepancy between point estimation and inference. But beyond having different rates, as in Section 4, it is possible that coverage error may vanish even if mean square error does not, and vice versa. We have discussed one direction already: the coverage error of an interval using the MSE-optimal bandwidth $H_{\mathtt{mse}} n^{-1/(2p+3)}$ does not vanish,

$$\sup_{F \in \mathscr{F}_S} \left| \mathbb{P}_F \left[\mu^{(\nu)} \in \left\{ \hat{\mu}_p^{(\nu)} \pm z_{\alpha/2} \hat{\sigma}_p H_{\mathtt{mse}}^{-1/2} n^{-1/2 + (1+2\nu)/(4p+6)} \right\} \right] - (1-\alpha) \right| \asymp 1,$$

but this $\mu^{(\nu)}$ is the MSE-optimal point estimator.

The opposite direction may be more surprising and novel, and, intuitively, may occur if the variance of $\hat{\theta}$ is too large for mean-square consistency, but is captured well enough by $\hat{\vartheta}$ for inference.

For example, consider inference for the first derivative, $\mu^{(1)}(\mathsf{x})$, using local linear regression ($p = 1$) and conventional inference, with $\hat{\theta} = \hat{\mu}_p^{(\nu)}$ and $\hat{\vartheta}^2 = \hat{\sigma}_p^2/(nh^{1+2\nu})$. Choosing $h \asymp n^{-1/3}$ yields $r_T = n^{-2/3} \to 0$ (in fact, this is the fastest rate attainable by this I, i.e. $h \asymp n^{-1/3}$ is optimal undersmoothing, see below). However, $\mathbb{V}[\hat{\mu}_p^{(\nu)}|X_1, \ldots, X_n] = (nh^{1+2v})^{-1} \asymp 1$, thus $\hat{\mu}_p^{(\nu)}$ is not consistent in mean square. Therefore, we construct a confidence interval that is *optimal for coverage* of $\mu^{(1)}$, but implicitly relies on a point estimator that is *not even consistent* in mean square.

6 Simulation Study

This section presents results from a simulation study to examine the finite-sample performance of our methods. Additional results and implementation details can be found in the supplement. We focus on the performance of confidence intervals for $\mu(\mathsf{x})$ and $\mu^{(1)}(\mathsf{x})$ based on robust bias correction and traditional undersmoothing (i.e., centering $\hat{\mu}_p^{(\nu)}$ and scaling $\hat{\sigma}_p(nh^{1+2\nu})^{-1/2}$). Data is generated from model (1), with X_i uniformly distributed on $[-1, 1]$, ε standard normal, and

$$\mu(x) = \frac{\sin(3\pi x/2)}{1 + 18x^2(\text{sgn}(x) + 1)},$$

where $\text{sgn}(x) = -1$, 0, or -1 according to $x > 0$, $x = 0$ or $x < 0$, respectively. This function, which was also analyzed in Calonico et al. (2018), is displayed in Figure 2 together with $\mu^{(1)}(x)$. By looking at different evaluation points, we will be able to capture the performance of the methods under different levels of complexity.

We show results for sample sizes $n \in \{100, 250, 500, 750, 1000, 2000\}$, always with $5,000$ replications. We study inference at three evaluation points: $\mathsf{x} = -1$ (boundary point), $\mathsf{x} = -0.6$ (low curvature), and $\mathsf{x} = -0.2$ (high curvature). The supplement shows results for $\mathsf{x} \in \{0.2, 0.6, 1\}$. For implementation, we use $p = 1$ (for $\nu = 0$) and $p = 2$ (for $\nu = 1$) with the Epanechnikov kernel (the supplement gives results for the uniform kernel). Finally, we evaluate the performance of the confidence intervals using several bandwidth choices. First, following the results from Section 4, we use $\hat{h}_{\texttt{rbc}}$, a data-driven version of the inference-optimal bandwidth $h_{\texttt{rbc}}$. We also consider the analogous version for undersmoothed confidence intervals, denoted $\hat{h}_{\texttt{us}}$ (detailed in the supplement), and the standard choice in practice, $\hat{h}_{\texttt{mse}}$. Robust bias correction is implemented using $\rho = \rho^*$ according to

Table 2. All implementation details are available for R and Stata (Calonico et al., 2019).

Figures 3 and 4 present empirical coverage probabilities for $\nu = 0$ and $\nu = 1$, respectively, for each evaluation point and choice of bandwidth, as a function of the sample size. Overall, We can see that robust bias correction yields close to accurate coverage, improving over undersmoothing in almost every case. Performance is highly superior at points where the functions present high curvature and also at the boundary. Performance is never worse even when the function is quite linear and optimal bandwidths are (close to) ill-defined.

We also compare confidence interval performance in terms of length in Figure 5. We take coverage into account by looking at RBC and US confidence intervals implemented using their corresponding coverage error optimal bandwidth choices ($\hat{h}_{\texttt{rbc}}$ and $\hat{h}_{\texttt{us}}$, respectively), which is when they perform best in terms of coverage. We find that RBC confidence intervals are, on average, not larger than US, and sometimes even shorter. Lastly, Figure 6 shows the average estimated bandwidths at each point for each sample size, which behave as expected following our theory.

7 Conclusion

This paper derived the minimax bound for inference in nonparametric local polynomial regression, building on an idea in Hall and Jing (1995) to assess the quality of statistical inference through the rate at which the coverage of a confidence interval approaches its nominal value uniformly in over class of distributions for the data. After establishing the minimax rate bound, we demonstrated that robust bias corrected intervals, coupled with novel bandwidth choices, are minimax optimal. As a by-product, we also developed inference-optimal tuning parameter choices for local polynomial nonparametric regression.

Our main results measured coverage error symmetrically, but it is worth mentioning that the absolute loss function may be replaced by the "check" loss function, and thus studying the maximal coverage error $\sup_{F \in \mathscr{F}_S} \mathcal{L}(\mathbb{P}_F[\theta_F \in I] - (1-\alpha))$, with $\mathcal{L}(e) = \mathcal{L}_\tau(e) = e(\tau - \mathbb{1}\{e < 0\})$, and where $\tau \in (0,1)$ encodes the researcher's weight for over- and under-coverage. Setting $\tau = 1/2$ recovers the above, symmetric measure of coverage error. Guarding more against undercoverage (a preference for conservative intervals) requires choosing a $\tau < 1/2$. For example, setting $\tau = 1/3$ encodes the belief that undercoverage is twice as bad as the same amount of overcoverage. All our results can be established for this loss function; the supplement discusses adapting Theorem 3 to this case.

Finally, this paper studied the properties of confidence intervals at a fixed evaluation point x, but it would be of theoretical and practical interest to extent our results to the case of confidence band construction. Robust bias correction has recently been used to construct valid confidence bands for local polynomial estimation (Cheng et al., 2019) and linear sieve estimation (Cattaneo et al., 2020). Because the underlying distributional approximations for confidence band constructions are substantially more complex, obtaining minimax results similar to those presented herein will require substantial extension of the technical work here.

Table 2: L_2-Optimal Variance-Minimizing ρ

p	**Kernel**		
	Triangular	Epanechnikov	Uniform
0	0.778	0.846	1.000
1	0.850	0.898	1.000
2	0.887	0.924	1.000
3	0.909	0.940	1.000
4	0.924	0.950	1.000

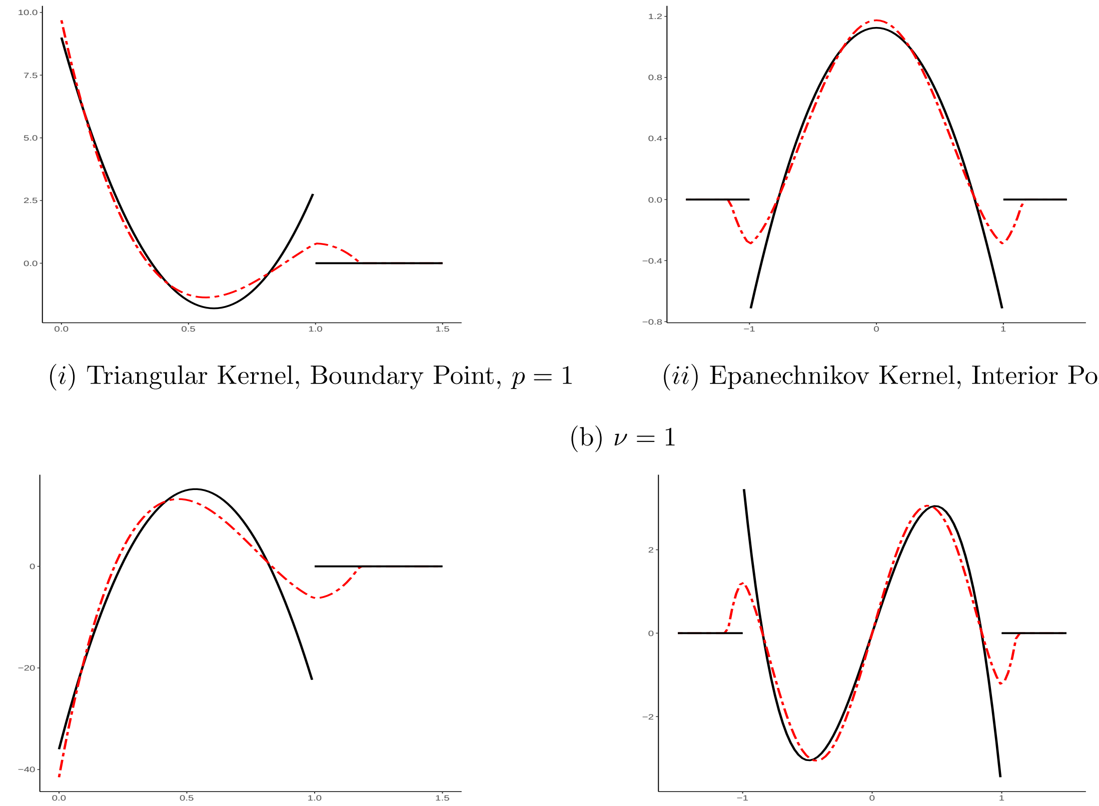
(a) Boundary point

p	**Kernel**		
	Triangular	Epanechnikov	Uniform
1	0.798	0.865	1.000
3	0.867	0.915	1.000
5	0.900	0.938	1.000
7	0.919	0.951	1.000

(b) Interior point

Note: Optimal ρ computed by minimizing the L_2 distance between the RBC induced equivalent kernel and the variance-minimizing equivalent kernel (Uniform Kernel).

Figure 1: $\mathcal{K}^*_{p+1}(u)$ vs. $\mathcal{K}_{\texttt{rbc}}(u; K, \rho^*, \nu)$



(a) $\nu = 0$

(i) Triangular Kernel, Boundary Point, $p = 1$

(ii) Epanechnikov Kernel, Interior Point, $p = 1$

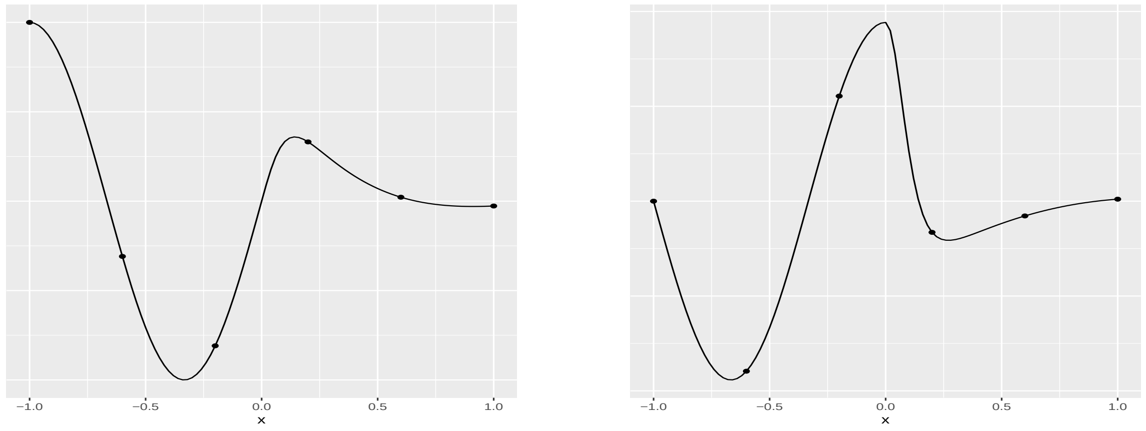
(b) $\nu = 1$

(iii) Triangular Kernel, Boundary Point, $p = 1$

(iv) Epanechnikov Kernel, Interior Point, $p = 2$

Notes: ━ $\mathcal{K}^*_{p+1}(u)$, -·- $\mathcal{K}_{\texttt{rbc}}(u; K, \rho^*, \nu)$

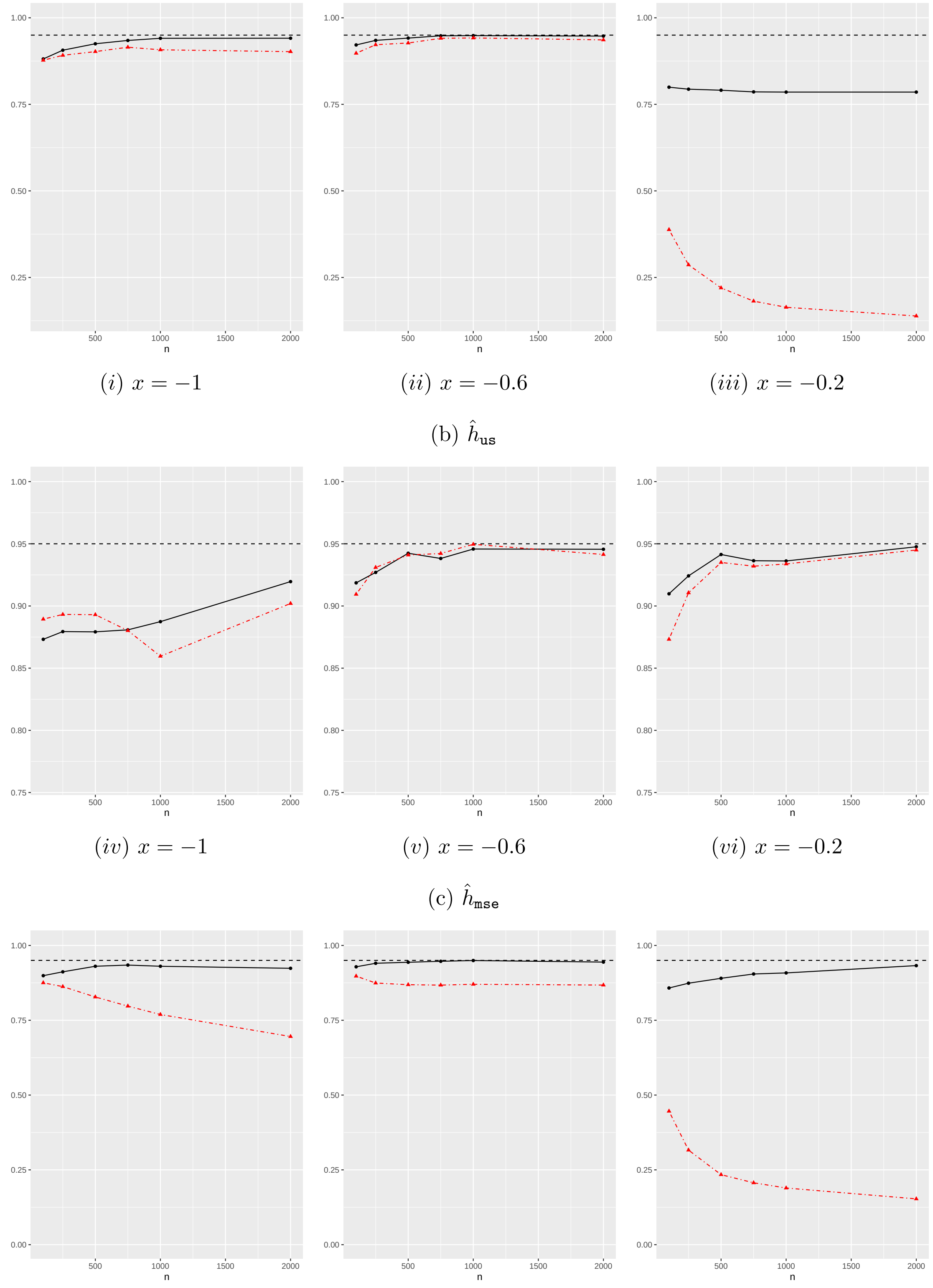
Figure 2: Conditional mean function and first derivative, $\mu^{(\nu)}(x)$



(a) $\nu = 0$

(b) $\nu = 1$

Figure 3: Empirical Coverage for 95% Confidence Intervals, $\nu = 0$



(a) $\hat{h}_{\texttt{rbc}}$

(*i*) $x = -1$ (*ii*) $x = -0.6$ (*iii*) $x = -0.2$

(b) $\hat{h}_{\texttt{us}}$

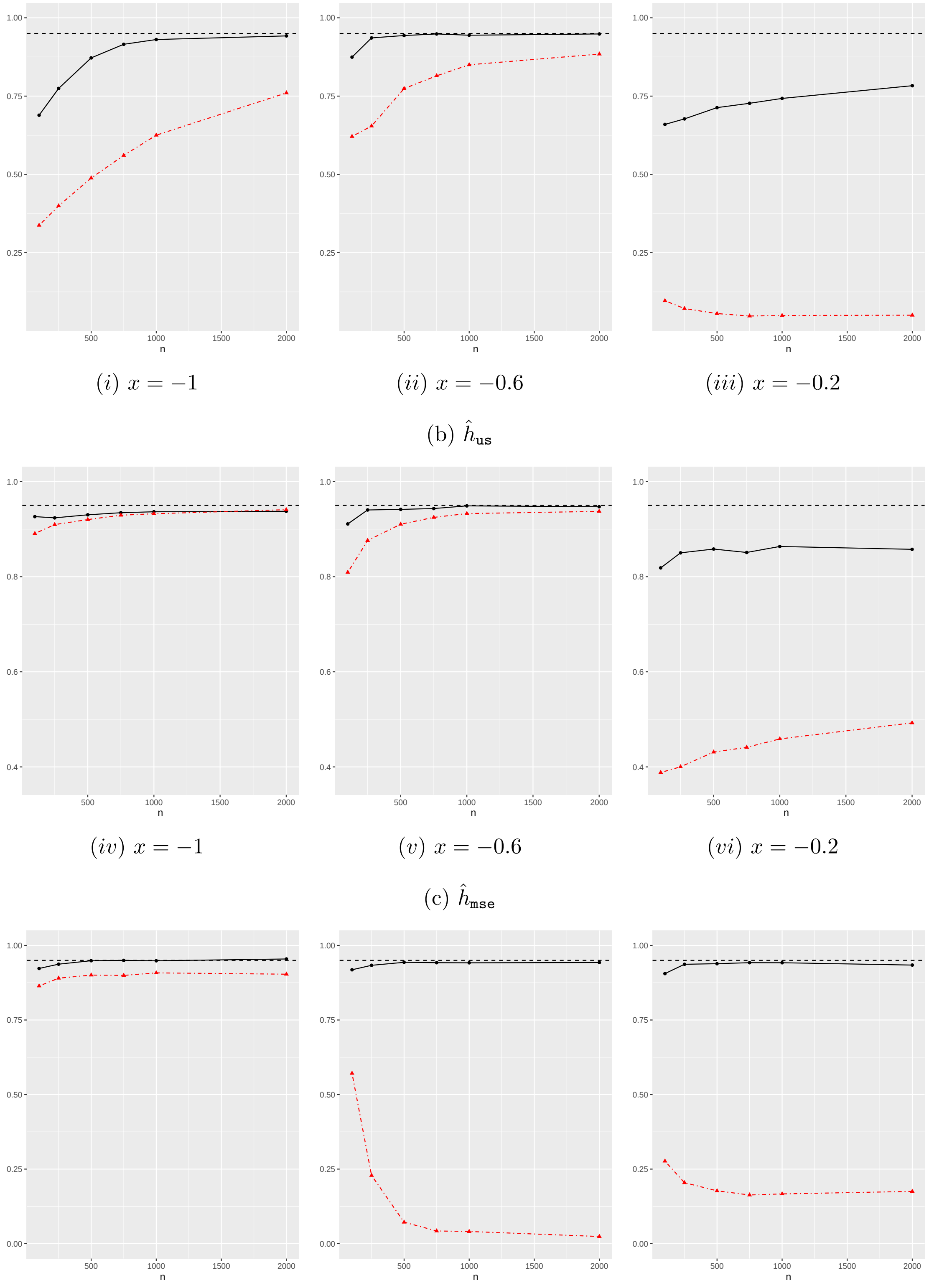
(*iv*) $x = -1$ (*v*) $x = -0.6$ (*vi*) $x = -0.2$

(c) $\hat{h}_{\texttt{mse}}$

(*vii*) $x = -1$ (*viii*) $x = -0.6$ (*ix*) $x = -0.2$

Notes: ━Robust Bias Correction, -·-·Undersmoothing; Epanechnikov Kernel

Figure 4: Empirical Coverage for 95% Confidence Intervals, $\nu = 1$



(a) $\hat{h}_{\texttt{rbc}}$

(i) $x = -1$ *(ii)* $x = -0.6$ *(iii)* $x = -0.2$

(b) $\hat{h}_{\texttt{us}}$

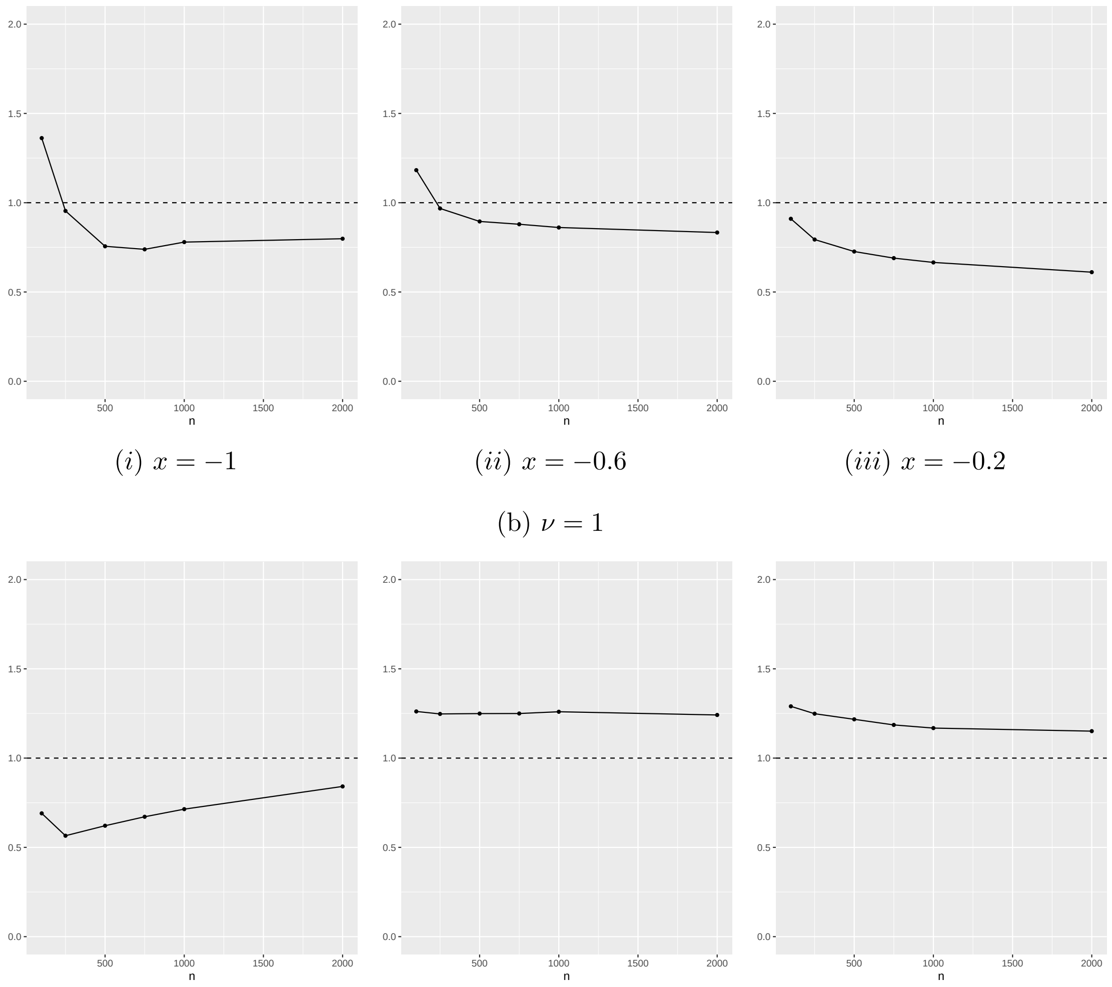
(iv) $x = -1$ *(v)* $x = -0.6$ *(vi)* $x = -0.2$

(c) $\hat{h}_{\texttt{mse}}$

(vii) $x = -1$ *(viii)* $x = -0.6$ *(ix)* $x = -0.2$

Notes: ━ Robust Bias Correction, -·- Undersmoothing; Epanechnikov Kernel

Figure 5: Average Interval Length for 95% Confidence Intervals



(a) $\nu = 0$

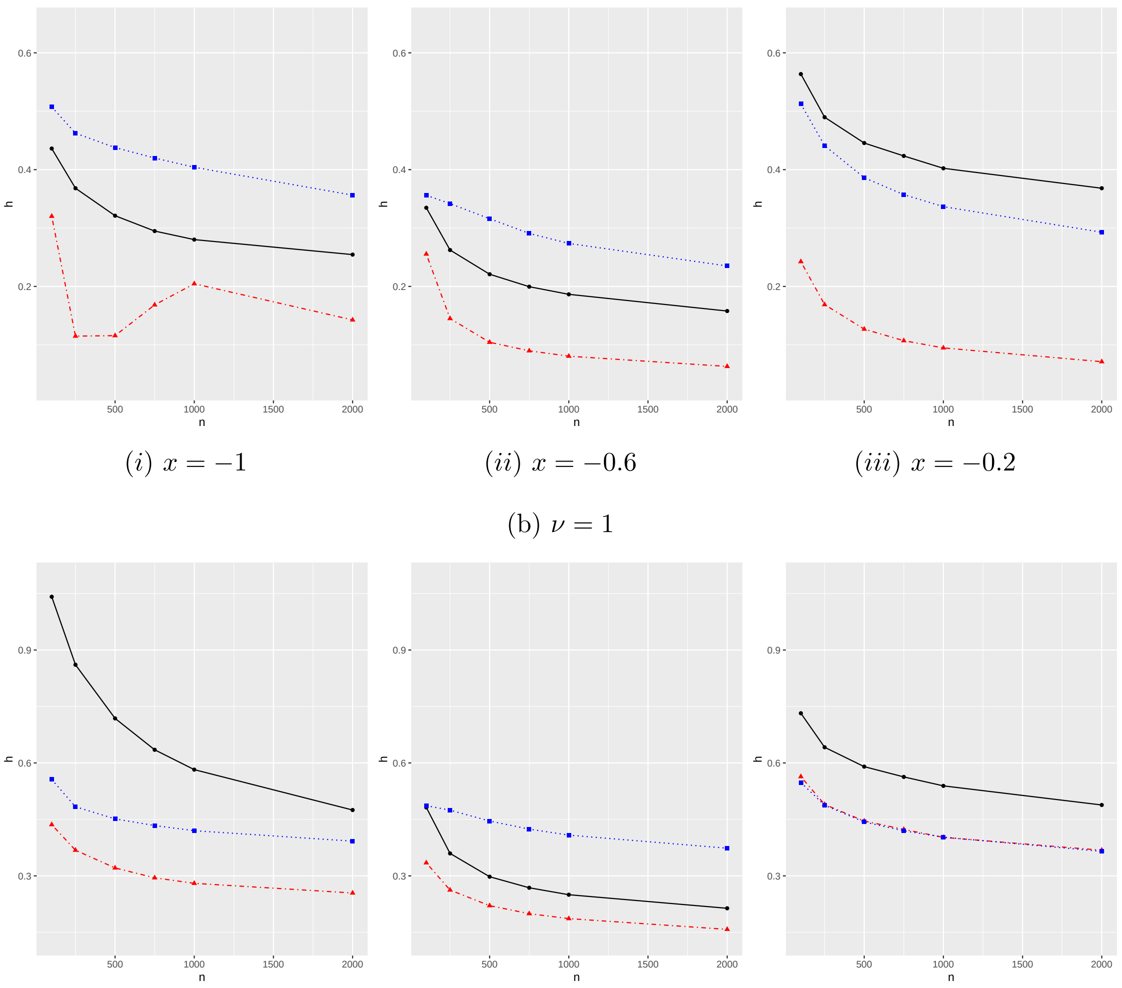
$(i)\ x = -1$ $(ii)\ x = -0.6$ $(iii)\ x = -0.2$

(b) $\nu = 1$

$(iv)\ x = -1$ $(v)\ x = -0.6$ $(vi)\ x = -0.2$

Notes: ratio of average interval length for $I_{\texttt{rbc}}(\hat{h}_{\texttt{rbc}})$ over $I_{\texttt{us}}(\hat{h}_{\texttt{us}})$; Epanechnikov Kernel

Figure 6: Average Estimated Bandwidth



(a) $\nu = 0$

(*i*) $x = -1$

(*ii*) $x = -0.6$

(*iii*) $x = -0.2$

(b) $\nu = 1$

(*iv*) $x = -1$

(*v*) $x = -0.6$

(*vi*) $x = -0.2$

Notes: ━ $\hat{h}_{\texttt{rbc}}$, -·- $\hat{h}_{\texttt{us}}$, ···· $\hat{h}_{\texttt{mse}}$; Epanechnikov Kernel

A Appendix: Terms of the Edgeworth Expansion

We now give the precise forms of the terms in the Edgeworth expansion, $E_{T,F}(z)$. This amounts to defining the terms ω_k, $k = 1, 2, \ldots, 6$, and $\lambda_{T,F}$. The bias terms are discussed in the text. Derivation, details, and discussion of all objects is given in the supplement. Several examples of $\lambda_{T,F}$, and further discussion, is also in the supplement, but there are too many to give a comprehensive list. In general, if the chosen standard errors are consistent, $\lambda_{T,F}$ has the form $\lambda_{T,F} = l_n L$, for a rate $l_n \to 0$ and a sequence L that is bounded and bounded away from zero, a "constant", capturing the difference between the variance of the numerator of the t-statistic and the population standardization chosen. Fixed-n standard errors yield $\lambda_{T,F} \equiv 0$. Traditional explicit bias correction, where the point estimate (or numerator of T) is bias-corrected but it is assumed that σ_p provides valid standardization (this requires $\rho \to 0$), we find that $\lambda_{T,F} = \rho^{p+2}(L_1 + \rho^{p+2} L_2)$, where L_1 captures the (scaled) covariance between $\hat{\mu}^{(\nu)}$ and $\hat{\mu}^{(p+1)}$ and L_2 the variance of $\hat{\mu}^{(p+1)}$; see Calonico et al. (2018) and its companion supplement for the exact expressions. For another example, for inference at the boundary when using the asymptotic variance for standardization (i.e. the probability limit of the conditional variance of the numerator), one finds $l_n = h$ and L capturing the difference between the conditional variance and its limit, based on the localization of the kernel; see Chen and Qin (2002) for the exact expression.

It remains to define ω_k, $k = 1, 2, \ldots, 6$. First, define the following objections, all calculated in a fixed-n sense, bounded uniformly in $\mathscr{F}_S$, and nonzero for some $F \in \mathscr{F}_S$. As shorthand, let a tilde accent denote a fixed-n expectation, so that $\tilde{\boldsymbol{\Gamma}} = \mathbb{E}[\boldsymbol{\Gamma}]$, $\tilde{\boldsymbol{\Lambda}}_1 = \mathbb{E}[\boldsymbol{\Lambda}_1]$, and so forth. Let

$$\begin{aligned}
\ell^0_{T_p}(X_i) &= \nu! \boldsymbol{e}_\nu' \tilde{\boldsymbol{\Gamma}}^{-1} (K\boldsymbol{r}_p)(X_{h,i}); \\
\ell^0_{T_{\texttt{rbc}}}(X_i) &= \ell^0_{T_p}(X_i) - \rho^{p+1} \nu! \boldsymbol{e}_\nu' \tilde{\boldsymbol{\Gamma}}^{-1} \tilde{\boldsymbol{\Lambda}}_1 \boldsymbol{e}_{p+1}' \tilde{\tilde{\boldsymbol{\Gamma}}}^{-1} (K\boldsymbol{r}_{p+1})(X_{b,i}); \\
\ell^1_{T_p}(X_i, X_j) &= \nu! \boldsymbol{e}_\nu' \tilde{\boldsymbol{\Gamma}}^{-1} \left(\mathbb{E}[(K\boldsymbol{r}_p \boldsymbol{r}_p')(X_{h,j})] - (K\boldsymbol{r}_p \boldsymbol{r}_p')(X_{h,j}) \right) \tilde{\boldsymbol{\Gamma}}^{-1} (K\boldsymbol{r}_p)(X_{h,i}); \\
\ell^1_{T_{\texttt{rbc}}}(X_i, X_j) &= \ell^1_{T_p}(X_i, X_j) - \rho^{p+1} \nu! \boldsymbol{e}_\nu' \tilde{\boldsymbol{\Gamma}}^{-1} \Big\{ \left(\mathbb{E}[(K\boldsymbol{r}_p \boldsymbol{r}_p')(X_{h,j})] - (K\boldsymbol{r}_p \boldsymbol{r}_p')(X_{h,j}) \right) \tilde{\boldsymbol{\Gamma}}^{-1} \tilde{\boldsymbol{\Lambda}}_1 \boldsymbol{e}_{p+1}' \\
&\qquad + \Big((K\boldsymbol{r}_p)(X_{h,j}) X_{h,i}^{p+1} - \mathbb{E}[(K\boldsymbol{r}_p)(X_{h,j}) X_{h,i}^{p+1}] \Big) \boldsymbol{e}_{p+1}' \\
&\qquad + \tilde{\boldsymbol{\Lambda}}_1 \boldsymbol{e}_{p+1}' \tilde{\tilde{\boldsymbol{\Gamma}}}^{-1} \left(\mathbb{E}[(K\boldsymbol{r}_{p+1} \boldsymbol{r}_{p+1}')(X_{b,j})] - (K\boldsymbol{r}_{p+1} \boldsymbol{r}_{p+1}')(X_{b,j}) \right) \Big\} \tilde{\tilde{\boldsymbol{\Gamma}}}^{-1} (K\boldsymbol{r}_{p+1})(X_{b,i}).
\end{aligned}$$

Then define $\tilde{\sigma}_T^2 = \mathbb{E}[h^{-1} \ell^0_T(X)^2 v(X)]$ and denote the standard Normal density as $\phi(z)$. Then we define

$$\begin{aligned}
\omega_{1,T,F}(z) &= \phi(z)\tilde{\sigma}_T^{-3}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^3\varepsilon_i^3\right]\left\{(2z^2-1)/6\right\}, \\
\omega_{2,T,F}(z) &= -\phi(z)\tilde{\sigma}_T^{-1}, \\
\omega_{3,T,F}(z) &= -\phi(z)\left\{z/2\right\}, \\
\omega_{5,T,F}(z) &= -\phi(z)\tilde{\sigma}_T^{-2}\left\{z/2\right\}, \\
\omega_{6,T,F}(z) &= \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}[h^{-1}\ell_T^0(X_i)^3\varepsilon_i^3]\left\{z^3/3\right\}.
\end{aligned}$$

For ω_3, it is not quite as simple to state a generic version. Let $\tilde{\boldsymbol{G}}$ stand in for $\tilde{\boldsymbol{\Gamma}}$ or $\tilde{\tilde{\boldsymbol{\Gamma}}}$, $\tilde{p}$ stand in for p or $p+1$, and d_n stand in for h or b, all depending on if $T = T_p$ or $T_{\texttt{rbc}}$. Note however, that h is still used in many places, in particular for stabilizing fixed-n expectations, for $T_{\texttt{rbc}}$. Indexes i, j, and k are always distinct (i.e. $X_{h,i} \neq X_{h,j} \neq X_{h,k}$).

$$\begin{aligned}
\omega_{4,T,F}(z) = {} & \phi(z)\tilde{\sigma}_T^{-6}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^3\varepsilon_i^3\right]^2\left\{z^3/3 + 7z/4 + \tilde{\sigma}_T^2 z(z^2-3)/4\right\} \\
& + \phi(z)\tilde{\sigma}_T^{-2}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)\ell_T^1(X_i,X_i)\varepsilon_i^2\right]\left\{-z(z^2-3)/2\right\} \\
& + \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^4(\varepsilon_i^4 - v(X_i)^2)\right]\left\{z(z^2-3)/8\right\} \\
& - \phi(z)\tilde{\sigma}_T^{-2}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^2\boldsymbol{r}_{\tilde{p}}(X_{d_n,i})'\tilde{\boldsymbol{G}}^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,i})\varepsilon_i^2\right]\left\{z(z^2-1)/2\right\} \\
& - \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^3\boldsymbol{r}_{\tilde{p}}(X_{d_n,i})'\tilde{\boldsymbol{G}}^{-1}\varepsilon_i^2\right]\mathbb{E}\left[h^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,i})\ell_T^0(X_i)\varepsilon_i^2\right]\left\{z(z^2-1)\right\} \\
& + \phi(z)\tilde{\sigma}_T^{-2}\mathbb{E}\left[h^{-2}\ell_T^0(X_i)^2(\boldsymbol{r}_{\tilde{p}}(X_{d_n,i})'\tilde{\boldsymbol{G}}^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,j}))^2\varepsilon_j^2\right]\left\{z(z^2-1)/4\right\} \\
& + \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-3}\ell_T^0(X_j)^2\boldsymbol{r}_{\tilde{p}}(X_{d_n,j})'\tilde{\boldsymbol{G}}^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,i})\ell_T^0(X_i)\boldsymbol{r}_{\tilde{p}}(X_{d_n,j})'\tilde{\boldsymbol{G}}^{-1}(K\boldsymbol{r}_{\tilde{p}})(X_{d_n,k})\ell_T^0(X_k)\varepsilon_i^2\varepsilon_k^2\right] \\
& \qquad\qquad \times \left\{z(z^2-1)/2\right\} \\
& + \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\ell_T^0(X_i)^4\varepsilon_i^4\right]\left\{-z(z^2-3)/24\right\} \\
& + \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\left(\ell_T^0(X_i)^2v(X_i) - \mathbb{E}[\ell_T^0(X_i)^2v(X_i)]\right)\ell_T^0(X_i)^2\varepsilon_i^2\right]\left\{z(z^2-1)/4\right\} \\
& + \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-2}\ell_T^1(X_i,X_j)\ell_T^0(X_i)\ell_T^0(X_j)^2\varepsilon_j^2v(X_i)\right]\left\{z(z^2-3)\right\} \\
& + \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-2}\ell_T^1(X_i,X_j)\ell_T^0(X_i)\left(\ell_T^0(X_j)^2v(X_j) - \mathbb{E}[\ell_T^0(X_j)^2v(X_j)]\right)\varepsilon_i^2\right]\left\{-z\right\} \\
& + \phi(z)\tilde{\sigma}_T^{-4}\mathbb{E}\left[h^{-1}\left(\ell_T^0(X_i)^2v(X_i) - \mathbb{E}[\ell_T^0(X_i)^2v(X_i)]\right)^2\right]\left\{-z(z^2+1)/8\right\}.
\end{aligned}$$

8 References

Bahadur, R. R., and Savage, L. J. (1956), “The Nonexistence of Certain Statistical Procedures in Nonparametric Problems,” *Annals of Mathematical Statistics*, 27, 1115–1122.

Beran, R. (1982), “Estimated Sampling Distributions: The Bootstrap and Competitors,” *Annals of Statistics*, 10, 212–225.

Bhattacharya, R. N. (1977), “Refinements of the Multidimensional Central Limit Theorem and Applications,” *Annals of Probability*, 5, 1–27.

Bhattacharya, R. N., and Rao, R. R. (1976), *Normal Approximation and Asymptotic Expansions*, John Wiley and Sons.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018), “On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference,” *Journal of the American Statistical Association*, 113, 767–779.

——— (2019), “`nprobust`: Nonparametric Kernel-Based Estimation and Robust Bias-Corrected Inference,” *Journal of Statistical Software*, 91, 1–33.

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

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020), “Large Sample Properties of Partitioning-Based Estimators,” *Annals of Statistics*, forthcoming.

Chen, S. X., and Qin, Y. S. (2000), “Empirical likelihood confidence intervals for local linear smoothers,” *Biometrika*, 87, 946–953.

Chen, S. X., and Qin, Y. S. (2002), “Confidence Intervals Based on Local Linear Smoother,” *Scandinavian Journal of Statistics*, 29, 89–99.

Chen, Y.-C. (2017), “A Tutorial on Kernel Density Estimation and Recent Advances,” *Biostatistics & Epidemiology*, 1, 161–187.

Cheng, G., Chen, Y.-C. et al. (2019), “Nonparametric Inference via Bootstrapping the Debiased Estimator,” *Electronic Journal of Statistics*, 13, 2194–2256.

Cheng, M.-Y., Fan, J., and Marron, J. S. (1997), “On Automatic Boundary Corrections,” *Annals of Statistics*, 25, 1691–1708.

Edgeworth, F. Y. (1883), “The law of error,” *The London, Edinburgh and Dublin Philosophical Magazine*, 5, 300–309.

——— (1906), “The generalised law of error, or law of great numbers,” *Journal of the Royal Statistical Society*, 69, 497–539.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1997), “Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency,” *Annals of the Institute of Statistical Mathematics*, 49, 79–99.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, London: Chapman and Hall.

Fan, J., and Yao, Q. (2005), *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer.

Hall, P. (1991), "Edgeworth Expansions for Nonparametric Density Estimators, with Applications," *Statistics*, 22, 215–232.

——— (1992a), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

——— (1992b), "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density," *Annals of Statistics*, 20, 675–694.

——— (1992c), "On Bootstrap Confidence Intervals in Nonparametric Regression," *Annals of Statistics*, 20, 695–711.

Hall, P., and Jing, B.-Y. (1995), "Uniform Coverage Error Bounds for Confidence Intervals and Berry-Esseen Theorems for Edgeworth Expansion," *Annals of Statistics*, 23, 363–375.

Hall, P., and Kang, K.-H. (2001), "Bootstrapping Nonparametric Density Estimators with Empirically Chosen Bandwidths," *Annals of Statistics*, 29, 1443–1468.

Neumann, M. H. (1997), "Pointwise confidence intervals in nonparametric regression with heteroscedastic error structure," *Statistics*, 29, 1–36.