Simple and Honest Confidence Intervals in Nonparametric Regression*

Timothy B. Armstrong† Michal Kolesár‡

Yale University Princeton University

October 5, 2016

Abstract

We consider the problem of constructing honest confidence intervals (CIs) for a scalar parameter of interest, such as the regression discontinuity parameter, in nonparametric regression based on kernel or local polynomial estimators. To ensure that our CIs are honest, we derive and tabulate novel critical values that take into account the possible bias of the estimator upon which the CIs are based. We give sharp efficiency bounds of using different kernels, and derive the optimal bandwidth for constructing honest CIs. We show that using the bandwidth that minimizes the maximum mean-squared error results in CIs that are nearly efficient and that in this case, the critical value depends only on the rate of convergence. For the common case in which the rate of convergence is $n^{-4/5}$, the appropriate critical value for 95% CIs is 2.18, rather than the usual 1.96 critical value. We illustrate our results in an empirical application.

*We thank numerous seminar and conference participants for helpful comments and suggestions. All remaining errors are our own. The research of the first author was supported by National Science Foundation Grant SES-1628939. The research of the second author was supported by National Science Foundation Grant SES-1628878.

†email: timothy.armstrong@yale.edu

‡email: mkolesar@princeton.edu

1 Introduction

This paper considers the problem of constructing confidence intervals (CIs) for a scalar parameter $T(f)$ of a function f, which can be a conditional mean or a density. The scalar parameter may correspond, for example, to a conditional mean, or its derivatives at a point, the regression discontinuity parameter, or the value of a density or its derivatives at a point. The main requirement on the CIs we impose is that they be honest in the sense of Li (1989): they need to achieve asymptotically correct coverage for all possible model parameters, that is, be valid uniformly in f. This requires the researcher to be explicit about the parameter space $\mathcal{F}$ for f by spelling out the smoothness or shape restrictions imposed on f.

The CIs that we propose are simple to construct.[1] Given a desired confidence level $1-\alpha$, they take the form $\hat{T}(k;h) \pm \mathrm{cv}_{1-\alpha}(h;k)\widehat{\mathrm{se}}(\hat{T}(k;h))$, where $\hat{T}(k;h)$ is a kernel or local polynomial estimator based on a kernel k and bandwidth h, $\widehat{\mathrm{se}}(\hat{T}(k;h))$ is its standard error, and $\mathrm{cv}_{1-\alpha}(h;k)$ a critical value that which we derive and tabulate. To ensure that the CIs maintain coverage over the whole parameter space, the critical value takes into account the worst-case bias (over the parameter space $\mathcal{F}$) of the estimator. As a result, it is larger than $z_{1-\alpha/2}$, the usual critical value corresponding to the $(1-\alpha/2)$-quantile of a standard normal distribution. Asymptotically, these CIs correspond to a fixed-length CIs as defined by Donoho (1994). One-sided CIs can be constructed by subtracting the worst-case bias from $\hat{T}(k;h)$, in addition to subtracting the standard error times $z_{1-\alpha}$.

We derive three main results. First, we derive bandwidths that optimize the length of these CIs. We show that, asymptotically, picking the length-optimal bandwidth amounts to choosing the optimal bias-variance trade-off, which depends on the parameter $T(f)$ and the parameter space only through the rate of convergence r of the mean-squared error (MSE). Consequently, the amount of over- or undersmoothing relative to the MSE-optimal bandwidth (i.e. bandwidth that minimaxes the MSE) depends only on r and the desired confidence

[1] An R package implementing our CIs in regression discontinuity designs is available at `https://github.com/kolesarm/RDHonest`.

level $1-\alpha$. For 95% CIs, we find that the length-optimal bandwidth always *oversmooths* relative to the MSE-optimal bandwidth.

Second, we consider efficiency of CIs based on MSE-optimal bandwidth. We find that two-sided 95% CIs constructed around the MSE-optimal bandwidth are at least 99% efficient relative to using the optimal bandwidth, so long as the rate of convergence r is greater than $2/3$. This gives a particularly simple procedure for constructing honest CIs that are nearly asymptotically optimal: construct the CI around an estimator based on MSE-optimal bandwidth, adding and subtracting the standard error times a critical value that takes into account the possible bias of the estimator. Crucially, we show that if the bandwidth is chosen in this way, the critical value depends only on the rate of convergence r. When $r = 4/5$, for example, as is the case for estimation at a point or regression discontinuity when f is assumed to have two derivatives, the critical value for a 95% CI is 2.18, rather than the usual 1.96 critical value.

These results have implications for the common practice of constructing CIs based on estimators that undersmooth relative to the MSE-optimal bandwidth. Questions related to the optimality of this practice have been considered by Hall (1992) and Calonico et al. (2016). Importantly, these papers restrict attention to CIs that use the usual critical value $z_{1-\alpha/2}$. It then becomes necessary to choose a small enough bandwidth so that the bias is asymptotically negligible relative to the standard error, since this is the only way to achieve correct coverage. Our results imply that rather than choosing a smaller bandwidth, it is better to use a larger critical value that takes into account the potential bias, which ensures correct coverage regardless of the bandwidth. At the MSE- or length-optimal bandwidth, the resulting CIs shrink at the optimal rate $r/2$, in contrast to CIs based on undersmoothing, which shrink more slowly.

Third, we derive sharp efficiency bounds for one- and two-sided confidence intervals based on different kernels. We show that the kernel efficiency depends only on the parameter of interest and the parameter space, and not on the performance criterion. Consequently,

minimax MSE efficiencies of different kernels correspond directly to kernel efficiencies for constructing CIs. Furthermore, when the parameter space $\mathcal{F}$ is convex and symmetric (such as as when $\mathcal{F}$ only places restrictions on the derivatives of f), it follows from calculations in Donoho (1994) and Armstrong and Kolesár (2016) that our CIs, when constructed based on a highly efficient kernel and length-optimal or MSE-optimal bandwidth, are highly efficient among *all* CIs.

We specialize these results to the problem of inference about a nonparametric regression function at a point (i.e. inference about $f(x_0)$ for some x_0), and inference in sharp regression discontinuity (RD) designs. For inference at a point under a bound on the error of approximating f by a Taylor approximation around x_0, Fan (1993), Cheng et al. (1997), and Fan et al. (1997) calculate bounds on minimax MSE-efficiency of local polynomial estimators based on different kernels. In particular, Cheng et al. (1997) show that a local linear estimator with triangular kernel is 97% efficient for minimax MSE estimation at a boundary point under a bound on the error of the first order Taylor approximation. This result is often cited in recommending the use of this estimator in RD (see, e.g., Calonico et al., 2014). Our results show that, since the high efficiency of this estimator translates directly to the problem of constructing CIs, this recommendation can also be given when the goal is to construct CIs, as is often the case in practice.

Bounding the error from a Taylor approximation is one way to formalize the notion that the pth derivative of f at x_0 should be no larger than some constant M. In many applications, this restriction may too conservative, as it allows f to be non-smooth away from x_0. We therefore also consider the problem of inference under a Hölder class, which bounds the pth derivative globally. We derive an analytic expression for the maximum bias and kernel efficiencies of local polynomial estimators under this parameter space, and show that when the second derivative is bounded by a constant, a local linear estimator with triangular kernel is over 99.9% efficient at a boundary point. Furthermore, we show that, by bounding the second derivative globally, one can tighten the CIs by about 10–15%, with

the exact gain depending on the kernel.

We also consider coverage and efficiency of alternative CIs, in particular the usual CIs that use $z_{1-\alpha/2}$ as the critical value, and CIs based on the robust bias correction proposed recently by Calonico et al. (2014) and Calonico et al. (2016). We show that while at the MSE-optimal bandwidth, the usual CIs with nominal 95% coverage achieve honest coverage equal to 92.1%, the undercoverage problem can be quite severe if a larger bandwidth is used. We find that CIs based on robust bias correction have excellent coverage properties: a nominal 95% CI has asymptotic coverage equal to or just below 95%, depending on how one defines the parameter space. However, they are longer than the honest CIs at the length-optimal bandwidth that we propose by about 30% or shrink at a slower rate, again depending on how one defines the parameter space.

To illustrate the implementation of the honest CIs, we reanalyze the data from Ludwig and Miller (2007), who, using a regression discontinuity design, find a large and significant effect of receiving technical assistance to apply for Head Start funding on child mortality at a county level. However, this result is based on CIs that ignore the possible bias of the local linear estimator around which they are built, and an ad hoc bandwidth choice. We find that, if one bounds the second derivative globally by a constant M using a Hölder class the effect is not significant at the 5% level unless one is very optimistic about the constant M, allowing f to only be linear or nearly-linear.

Our results build on the literature on estimation of linear functionals in normal models with convex parameter spaces, as developed by Donoho (1994), Ibragimov and Khas'minskii (1985) and many others. As with the results in that literature, our setup gives asymptotic results for problems that are asymptotically equivalent to the Gaussian white noise model, including nonparametric regression (Brown and Low, 1996) and density estimation (Nussbaum, 1996). Our main results build on the "renormalization heuristics" of Donoho and Low (1992), who show that many nonparametric estimation problems have renormalization properties that allow easy computation of minimax mean squared error optimal kernels and

rates of convergence. Indeed, our results hold under essentially the same conditions (see Appendix B in the supplemental materials).

A drawback of our CIs is that they are non-adaptive: one needs to pick an a priori bound on the smoothness of f in order to implement them, and their length is determined by the conservativeness of this smoothness bound. When the parameter space is restricted by bounding the pth derivative by a constant M, for instance, then M need to be specified ex ante by the researcher. However, the results of Low (1997), Cai and Low (2004), and Armstrong and Kolesár (2016) imply that under smoothness restrictions that lead to convex, symmetric parameter spaces $\mathcal{F}$, this cannot be avoided, and therefore fully automatic nonparametric inference is not possible. In particular, their results show that honest CIs based on the worst possible smoothness constant M allowed are highly efficient at smooth functions relative to CIs that optimize their length at these smooth functions. Therefore, procedures that use data-driven rules to determine the smoothness of f in an attempt to "adapt" to f must either fail to improve upon our CIs, or else fail to maintain coverage over the whole parameter space.

On the other hand, it is often possible to form estimators that are adaptive in that they are close (up to a $\log n$ term) to the minimax MSE without knowing the smoothness constants of f, such as the number of its derivatives. It may therefore seem attractive to construct CIs that are centered at such adaptive estimators. Unfortunately, Cai and Low (2005) show that, in line with the non-adaptivity results cited above, not only is the rate of convergence of such CIs no faster than the rate corresponding to the worst possible smoothness class, in many cases it is strictly worse, or else such CIs have poor coverage properties. In contrast, our CIs are centered around estimators that are minimax for the worst possible smoothness constant allowed by the parameter space. We find that such CIs are not only optimal in rate, but are also close to optimal in the constant.

The rest of this paper is organized as follows. Section 2 gives the main results. Section 3 applies our results to inference at a point. Section 4 applies the results to RD, and presents

an empirical application based on Ludwig and Miller (2007). Proofs of the results in Section 2 are given in Appendix A. The supplemental materials contain further appendices and additional tables and figures. Appendix B verifies our regularity conditions for some examples, and includes proofs of the results in Section 3. Appendix C calculates the efficiency gain from using different bandwidths on either side of a cutoff in RD that is used in Section 4. Appendix D contains details on optimal kernel calculations discussed in Section 3.

2 General results

We are interested in a scalar parameter $T(f)$ of a function f, which is typically a conditional mean or density. The function f is assumed to lie in a function class $\mathcal{F}$, which places "smoothness" conditions on f. We have available a class of estimators $\hat{T}(h;k)$ based on a sample of size n, which depend on a bandwidth $h = h_n > 0$ and a kernel k. Let

$$\overline{\text{bias}}(\hat{T}) = \sup_{f\in\mathcal{F}} \left| E_f(\hat{T} - T(f)) \right|$$

denote the worst-case bias of an estimator $\hat{T}$, and let $\text{sd}_f(\hat{T}) = \text{var}_f(\hat{T})^{1/2}$ denote its standard deviation. We assume that the estimator $\hat{T}(h;k)$ is centered so that its maximum and minimum bias over $\mathcal{F}$ sum to zero, $\sup_{f\in\mathcal{F}} E_f(\hat{T}(h;k)-T(f)) = -\inf_{f\in\mathcal{F}} E_f(\hat{T}(h;k)-T(f))$.

Our main assumption is that the variance and worst-case bias scale as powers of h. In particular, we assume that, for some $\gamma_b > 0$, $\gamma_s < 0$, $B(k) > 0$ and $S(k) > 0$,

$$\overline{\text{bias}}(\hat{T}(h;k)) = h^{\gamma_b} B(k)(1 + o(1)), \qquad \text{sd}_f(\hat{T}(h;k)) = h^{\gamma_s} n^{-1/2} S(k)(1 + o(1)), \quad (1)$$

where the $o(1)$ term in the second equality is uniform over $f \in \mathcal{F}$. Note that the second condition implies that the standard deviation does not depend on the underlying function f asymptotically.

In the remainder of this section, we derive our main results. Section 2.1 presents a

heuristic derivation of the results, while Section 2.2 gives formal statements with regularity conditions. Before continuing, we illustrate condition (1) with an example.

Example 2.1. For local linear estimation of a nonparametric regression function at an interior point under a second-order Taylor smoothness class, (1) essentially follows from calculations in Fan (1993). For expositional purposes, we give a full derivation of these results in a simplified setting. We normalize the point of interest to be 0, so that we are interested in $f(0)$. The second-order Taylor smoothness class comprises all functions for which the approximation error from a first-order Taylor approximation around 0 can be bounded by $Mx^2/2$, for some constant M,

$$\mathcal{F} = \{f\colon |r(x)| \leq Mx^2/2\},$$

where $r(x) = f(x) - f(0) - f'(0)x$. We assume that the regression error is homoskedastic, and that the design points are non-random, and equispaced on the interval $[-1/2, 1/2]$, so that the data follow the model

$$y_i = f(x_i) + u_i, \qquad \operatorname{var}(u_i) = \sigma^2, \qquad x_i = \begin{cases} -\frac{i-1}{2n} & i \text{ odd} \\ \frac{i}{2n} & i \text{ even,} \end{cases} \qquad i = 1, \ldots, n.$$

Assume that n is odd, so that the design points are symmetric around 0. Let k be a symmetric kernel. Because the design points are symmetric around zero and k is symmetric, the local linear and Nadaraya-Watson estimator are identical[2], and are both given by

$$\hat{T}(h; k) = \frac{\sum_{i=1}^n y_i k(x_i/h)}{\sum_{i=1}^n k(x_i/h)},$$

where $h = h_n$ is a bandwidth sequence with $h \to 0$ and $hn \to \infty$. The standard deviation is

[2]If the design points are not symmetric, the local linear and Nadaraya-Watson estimators are different, and the local linear estimator must be used to avoid infinite worst-case bias. See Section 3.

constant over f and is equal to

$$\mathrm{sd}_f(\hat{T}(h;k)) = \frac{\sigma n^{-1/2}h^{-1/2}\sqrt{\frac{1}{nh}\sum_{i=1}^n k(x_i/h)^2}}{\frac{1}{nh}\sum_{i=1}^n k(x_i/h)} = \frac{\sigma n^{-1/2}h^{-1/2}\sqrt{\int k(u)^2\,du}}{\int k(u)\,du}(1+o(1)),$$

where the last equality holds under mild regularity conditions on $k(\cdot)$. The bias at a function $f \in \mathcal{F}$ is $\frac{\sum_{i=1}^n [f(x_i)-f(0)]k(x_i/h)}{\sum_{i=1}^n k(x_i/h)} = \frac{\sum_{i=1}^n r(x_i)k(x_i/h)}{\sum_{i=1}^n k(x_i/h)}$, where the equality follows since $\sum_{i=1}^n x_i k(x_i/h) = 0$ by symmetry of k and the design points. The bias is maximized by taking $r(x_i) = (M/2)x^2 \cdot \mathrm{sign}(k(x_i/h))$, which gives

$$\overline{\mathrm{bias}}(\hat{T}(h;k)) = \frac{(M/2)h^2\frac{1}{nh}\sum_{i=1}^n (x_i/h)^2|k(x_i/h)|}{\frac{1}{nh}\sum_{i=1}^n k(x_i/h)} = \frac{(M/2)h^2\int u^2|k(u)|\,du}{\int k(u)\,du}(1+o(1)),$$

where the last equality holds under regularity conditions on $k(\cdot)$. Thus, under regularity conditions, Equation (1) holds with $\gamma_s = -1/2$, $\gamma_b = 2$, $S(k) = \frac{\sigma\sqrt{\int k(u)^2\,du}}{\int k(u)\,du}$ and $B(k) = \frac{(M/2)\int u^2|k(u)|\,du}{\int k(u)\,du}$. In Section 3, we show that this result generalizes to the case with heteroscedastic errors and general design points.

2.1 Overview of results

Let $t = h^{\gamma_b-\gamma_s}B(k)/(n^{-1/2}S(k))$ denote the ratio of the leading worst-case bias and standard deviation terms. Substituting $h = \left(tn^{-1/2}S(k)/B(k)\right)^{1/(\gamma_b-\gamma_s)}$ into (1), the approximate bias and standard deviation can be written as

$$h^{\gamma_b}B(k) = t^r n^{-r/2}S(k)^r B(k)^{1-r}, \qquad h^{\gamma_s}n^{-1/2}S(k) = t^{r-1}n^{-r/2}S(k)^r B(k)^{r-1} \tag{2}$$

where $r = \gamma_b/(\gamma_b - \gamma_s)$. Since bias and standard deviation converge at a $n^{r/2}$ rate, we refer to r as the rate exponent (note that this matches with the definition in, e.g., Donoho and Low 1992; see Appendix B in the supplemental materials). In Example 2.1, we have $r = 2/[2-(-1/2)] = 4/5$.

Computing the bias-standard deviation ratio t associated with a given bandwidth allows

easy computation of honest CIs. Let $\widehat{\text{se}}(h;k)$ denote the standard error, an estimate of $\text{sd}_f(\hat{T}(h;k))$. Assuming a central limit theorem applies to $\hat{T}(h;k)$, $[\hat{T}(h;k) - T(f)]/\widehat{\text{se}}(h;k)$ will be approximately distributed as a normal random variable with variance 1 and bias bounded by t. Thus, an approximate $1-\alpha$ CI is given by

$$\hat{T}(h;k) \pm \text{cv}_{1-\alpha}(t) \cdot \widehat{\text{se}}(h;k), \tag{3}$$

where

$$\text{cv}_{1-\alpha}(t) \text{ is the } 1-\alpha \text{ quantile of the } |N(t,1)| \text{ distribution.} \tag{4}$$

This is an approximate version of a fixed-length confidence interval (FLCI) as defined in Donoho (1994) (if $\text{sd}_f(\hat{T}(h;k))$ is constant over f instead of approximately constant, the CI with $\widehat{\text{se}}(h;k)$ replaced by $\text{sd}_f(\hat{T}(h;k))$ will have fixed length). Following this definition, we use the term fixed-length to refer to CIs of this form even though $\widehat{\text{se}}(h;k)$ is random. One could also form honest CIs by simply adding and subtracting the worst case bias, in addition to adding and subtracting the standard error times $z_{1-\alpha/2} = \text{cv}_{1-\alpha}(0)$, the $1-\alpha/2$ quantile of a standard normal distribution:

$$\hat{T}(h;k) \pm (\overline{\text{bias}}(\hat{T}(h;k)) + z_{1-\alpha/2} \cdot \widehat{\text{se}}(h;k)).$$

However, since the estimator $\hat{T}(h;k)$ cannot simultaneously have a large positive and a large negative bias, such CI will be conservative, and longer than the CI given in Equation (3).

The usual nonparametric CIs, $\hat{T}(h;k) \pm z_{1-\alpha/2} \cdot \widehat{\text{se}}(h;k)$, rely on "undersmoothing:" under the current setup, this means that the bandwidth needs to be chosen such that $t = 0$, so that the bias is asymptotically negligible relative to the standard deviation of the estimator (otherwise the CI will undercover). As a result, the CIs shrink at a slower rate than $r/2$. In contrast, the honest FLCIs in Equation (3) explicitly take into account the possible bias of the estimator by replacing the critical value with $\text{cv}_{1-\alpha}(t)$, thus allowing for larger

bandwidths to be used, which, for $0 < t < \infty$, leads to the CIs shrinking at the optimal rate $r/2$. Furthermore, one can choose the bandwidth in a way that optimizes the length of the CI, which is given by

$$2 \cdot \widehat{\text{se}}(h;k) \cdot \text{cv}_{1-\alpha}(t) \approx 2 \cdot t^{r-1} n^{-r/2} S(k)^r B(k)^{1-r} \cdot \text{cv}_{1-\alpha}(t). \tag{5}$$

The bias-standard deviation ratio minimizing this length is given by $t^*_{FLCI} = \text{argmin}_{t>0}\, t^{r-1} \cdot \text{cv}_{1-\alpha}(t)$, and the FLCI-optimal bandwidth is $h^*_{FLCI} = \left(t^*_{FLCI} n^{-1/2} S(k)/B(k)\right)^{1/(\gamma_b - \gamma_s)}$.

Let us compare h^*_{FLCI} to the optimal bandwidth for estimation under mean squared error loss. Since under (1), the leading variance term is independent of f, the maximum (over $\mathcal{F}$) MSE is approximately equal to the worst-case squared bias plus the variance. For comparison with CI length and other criteria, it will be convenient to consider the root mean squared error (RMSE)—the square root of the maximum MSE. Under (1), this is approximately equal to

$$\sqrt{\left[h^{\gamma_b} B(k)\right]^2 + \left[h^{\gamma_s} n^{-1/2} S(k)\right]^2} = \sqrt{(t^{2r} + t^{2r-2})} n^{-r/2} S(k)^r B(k)^{1-r}. \tag{6}$$

This is minimized by $t^*_{RMSE} = \text{argmin}_{t>0}(t^{2r} + t^{2r-2}) = \sqrt{1/r - 1}$, which gives the optimal bandwidth as

$$h^*_{RMSE} = \left(t^*_{RMSE} n^{-1/2} S(k)/B(k)\right)^{1/(\gamma_b - \gamma_s)} = \left(\sqrt{1/r - 1} \cdot n^{-1/2} S(k)/B(k)\right)^{1/(\gamma_b - \gamma_s)}.$$

These calculations have several useful consequences. First, note that both (5) and (6) depend on k only through multiplication by $S(k)^r B(k)^{1-r}$. Thus, the relative efficiency of two kernels k_1 and k_2 is given by $[S(k_1)^r B(k_1)^{1-r}]/[S(k_2)^r B(k_2)^{1-r}]$ regardless of whether we consider CI length or RMSE.

Second, the optimal bias-standard deviation ratios for RMSE and FLCI depend only on the rate exponent r: for nonparametric estimators that converge at rate $n^{-r/2}$, the optimal

bias-standard deviation ratio for RMSE is $t^*_{RMSE} = \sqrt{1/r - 1}$, and the optimal bias-standard deviation ratio for FLCI is $t^*_{FLCI} = \text{argmin}_{t>0} t^{r-1} \, \text{cv}_{1-\alpha}(t)$ (the latter quantity can be found numerically). Since h is increasing in t, it follows that the FLCI optimal bandwidth undersmooths relative to the RMSE optimal bandwidth (i.e. $h^*_{FLCI} < h^*_{RMSE}$) if $t^*_{FLCI} < t^*_{RMSE}$ and oversmooths if $t^*_{RMSE} < t^*_{FLCI}$. For 95% CIs and $r/2$ in the range of rates of convergence typically encountered in practice, it turns out that $t^*_{RMSE} < t^*_{FLCI}$: the FLCI optimal bandwidth oversmooths relative to the RMSE optimal bandwidth.

Third, we get formulas for CIs centered at the RMSE optimal estimate, and for their efficiency relative to the optimal FLCI. A fixed-length CI centered at $\hat{T}(h^*_{RMSE}; k)$ takes the form $\hat{T}(h^*_{RMSE}; k) \pm \widehat{\text{se}}(h^*_{RMSE}; k) \cdot \text{cv}_{1-\alpha}(\sqrt{1/r - 1})$. This modified critical value depends only on the rate r, and is given in Table 1 for some common values. By Equation (5), the length of this CI is approximately $2 \cdot (t^*_{RMSE})^{r-1} n^{-r/2} S(k)^r B(k)^{1-r} \cdot \text{cv}_{1-\alpha}(t^*_{RMSE})$. If the bandwidth were instead chosen to minimize the length of the CI, the length would be given by the minimum of (5) over t, which would decrease the length of the CI by a factor of

$$\frac{(t^*_{FLCI})^{r-1} \cdot \text{cv}_{1-\alpha}(t^*_{FLCI})}{(t^*_{RMSE})^{r-1} \cdot \text{cv}_{1-\alpha}(t^*_{RMSE})}. \tag{7}$$

Since t^*_{FLCI} and t^*_{RMSE} depend only on r, this depends only on r. Figure 1 plots this quantity as a function of r. It can be seen from the figure that if $r \geq 4/5$, CIs constructed around the RMSE optimal bandwidth are highly efficient.

In Example 2.1, $r = 4/5$ for estimation of the function at a point. The optimal bias-standard deviation ratio for RMSE is then $\sqrt{1/r - 1} = 1/2$, and a 95% CI centered at the RMSE optimal estimate adds and subtracts $\text{cv}_{.95}(1/2) \approx 2.18$ times the standard error, rather than $z_{.975} \approx 1.96$ times the standard error. Evaluating (7) for $r = 4/5$, we find that using the RMSE optimal bandwidth to construct a CI is over 99% efficient: the width of the CI centered at the FLCI optimal bandwidth is more than 0.99 times the width of this CI.

2.2 Formal results

We consider a slightly more general setup that encompasses other performance criteria, such as median absolute deviation and excess length of one-sided CIs. Let $R(\hat{T})$ denote the worst-case (over $\mathcal{F}$) performance of $\hat{T}$ according to a given criterion, and let $\tilde{R}(b,s)$ denote the value of this criterion when $\hat{T}-T(f)$ is $N(b,s^2)$. For RMSE, these are given by

$$R_{RMSE}(\hat{T}) = \sup_{f\in\mathcal{F}} \sqrt{E_f\left[\hat{T}-T(f)\right]^2}, \qquad \tilde{R}(b,s) = \sqrt{b^2+s^2}.$$

For FLCI,

$$\begin{aligned} R_{\text{FLCI},\alpha}(\hat{T}(h;k)) &= \inf\left\{\chi : P_f\left(|\hat{T}(h;k)-T(f)|\le\chi\right)\ge 1-\alpha \text{ all } f\in\mathcal{F}\right\}, \\ \tilde{R}_{\text{FLCI},\alpha}(b,s) &= \inf\left\{\chi : P_{Z\sim N(0,1)}\left(|sZ+b|\le\chi\right)\ge 1-\alpha\right\} = s\cdot \text{cv}_{1-\alpha}(b/s), \end{aligned}$$

where $\text{cv}_{1-\alpha}(t)$ is the $1-\alpha$ quantile of the absolute value of a $N(t,1)$ random variable, as defined in (4). Note that $\text{cv}_{1-\alpha}(t) = \tilde{R}_{FLCI}(t,1)$.

To evaluate one-sided CIs, one needs a criterion other than length, which is infinite. A natural criterion is expected excess length, or quantiles of excess length. We focus here on the worst-case quantiles of excess length. For CI of the form $[\hat{c},\infty)$, the worst-case β quantile of excess length is given by $\sup_{f\in\mathcal{F}} q_{f,\beta}(Tf-\hat{c})$, where $q_{f,\beta}(Z)$ is the β quantile of a random variable Z. Under (1) and a uniform-in-f central limit theorem for $\hat{T}(h;k)$, an honest one-sided $1-\alpha$ CI based on $\hat{T}(h;k)$ can be formed by subtracting the maximum bias, in addition to subtracting $z_{1-\alpha}$ times the standard deviation from $\hat{T}(h;k)$, leading to the interval

$$[\hat{T}(h;k) - h^{\gamma_b}B(k) - z_{1-\alpha}h^{\gamma_s}n^{-1/2}S(k)\ ,\ \infty).$$

We use $R_{\text{OCI},\alpha,\beta}(\hat{T}(h;k))$ to denote the worst-case β quantile of excess length of this CI. The worst-case β quantile of excess length based on an estimator $\hat{T}$ when $\hat{T}-T(f)$ is normal with variance s^2 and bias ranging between $-b$ and b is $\tilde{R}_{\text{OCI},\alpha,\beta}(b,s) \equiv 2b + (z_{1-\alpha}+z_\beta)s$.

When (1) holds and the estimator $\hat{T}(h;k)$ satisfies an appropriate central limit theorem, these performance criteria will satisfy

$$R(\hat{T}(h;k)) = \tilde{R}(h^{\gamma_b}B(k), h^{\gamma_s}n^{-1/2}S(k))(1+o(1)). \tag{8}$$

For our main results, we make this assumption directly. As we show in Section B, (8) holds with the $o(1)$ term equal to zero under the renormalization conditions of Donoho and Low (1992). Thus, verifying this condition in a given setting essentially amounts to verifying conditions for the renormalization heuristics of Donoho and Low (1992). We will also assume that $\tilde{R}$ scales linearly in its arguments (i.e. it is homogeneous of degree one): $\tilde{R}(tb, ts) = t\tilde{R}(b, s)$. This holds for all of the criteria considered above. Plugging in (2) and using scale invariance of $\tilde{R}$ gives

$$R(\hat{T}(h;k)) = n^{-r/2}S(k)^r B(k)^{1-r}t^{r-1}\tilde{R}(t,1)(1+o(1)) \tag{9}$$

where $t = h^{\gamma_b-\gamma_s}B(k)/(n^{-1/2}S(k))$ and $r = \gamma_b/(\gamma_b-\gamma_s)$, as defined in Section 2.1. Under (9), the asymptotically optimal bandwidth is given by $h_R^* = (n^{-1/2}S(k)t_R^*/B(k))^{1/(\gamma_b-\gamma_s)}$ where $t_R^* = \operatorname{argmin}_t t^{r-1}\tilde{R}(t,1)$. This generalizes the optimal bandwidth derivations based on (5) and (6) to other performance criteria: for $R = R_{FLCI}$, (9) essentially reduces to (5) (note that $\mathrm{cv}_{1-\alpha}(t) = \tilde{R}_{FLCI,\alpha}(t,1)$) and for $R = R_{RMSE}$, (9) reduces to (6). This gives the optimal bias-standard deviation ratios

$$t_{RMSE}^* = \operatorname*{argmin}_{t>0} t^{r-1}\tilde{R}_{RMSE}(t,1) = \operatorname*{argmin}_{t>0} t^{r-1}\sqrt{t^2+1} = \sqrt{1/r-1} \qquad \text{and}$$
$$t_{FLCI}^* = \operatorname*{argmin}_{t>0} t^{r-1}\tilde{R}_{FLCI,\alpha}(t,1) = \operatorname*{argmin}_{t>0} t^{r-1}\,\mathrm{cv}_{1-\alpha}(t),$$

and the corresponding optimal bandwidths, the same as in Section 2.1.

Assuming t_R^* is finite and strictly greater than zero, the optimal bandwidth decreases at the rate $n^{-1/[2(\gamma_b-\gamma_s)]}$ regardless of the performance criterion—the performance criterion

only determines the optimal bandwidth constant. Since the approximation (8) may not hold when h is too small or large relative to the sample size, we will only assume this condition for bandwidth sequences of order $n^{-1/[2(\gamma_b-\gamma_s)]}$. For our main results, we assume directly that optimal bandwidth sequences decrease at this rate:

$$n^{-r/2}R(\hat{T}(h_n;k)) \to \infty \text{ for any } h_n \text{ with } h_n/n^{1/[2(\gamma_b-\gamma_s)]} \to \infty \text{ or } h_n/n^{1/[2(\gamma_b-\gamma_s)]} \to 0. \quad (10)$$

Condition (10) will hold so long as it is suboptimal to choose a bandwidth such that the bias or the variance dominates asymptotically, which is the case in the settings considered here.

Using these conditions, we now give formal statements of the results obtained heuristically in Section 2.1.

Theorem 2.1. *Let R be a performance criterion that with $\tilde{R}(b,s) > 0$ for all $(b,s) \neq 0$ and $\tilde{R}(tb,ts) = t\tilde{R}(b,s)$ for all (b,s). Suppose that Equation (8) holds for any bandwidth sequence h_n with $\liminf_{n\to\infty} h_n/n^{1/[2(\gamma_b-\gamma_s)]} > 0$ and $\limsup_{n\to\infty} h_n/n^{1/[2(\gamma_b-\gamma_s)]} < \infty$, and suppose that Equation (10) holds. Let h_R^* and t_R^* be as defined above, and assume that $t_R^* > 0$ is unique and well defined. Then:*

(i) The asymptotic minimax performance of the kernel k is given by

$$n^{r/2} \inf_{h>0} R(\hat{T}(h;k)) = n^{r/2}R(\hat{T}(h_R^*;k)) + o(1) = S(k)^r B(k)^{1-r} \inf_t t^{r-1}\tilde{R}(t,1) + o(1),$$

where h_R^* is given above.

(ii) The asymptotic relative efficiency of two kernels k_1 and k_2 is given by

$$\lim_{n\to\infty} \frac{\inf_{h>0} R(\hat{T}(h;k_1))}{\inf_{h>0} R(\hat{T}(h;k_2))} = \frac{S(k_1)^r B(k_1)^{1-r}}{S(k_2)^r B(k_2)^{1-r}}.$$

It depends on the rate r but not on the performance criterion R.

(iii) If (1) holds, the asymptotically optimal bias-variance ratio is given by

$$\lim_{n\to\infty} \frac{\overline{\text{bias}}(\hat{T}(h_R^*;k))}{\text{sd}_f(\hat{T}(h_R^*;k))} = \operatorname*{argmin}_t t^{r-1}\tilde{R}(t,1) = t_R^*.$$

It depends only on the performance criterion R and rate exponent r. If we consider two performance criteria R_1 and R_2 such that these conditions hold, then the limit of the ratio of optimal bandwidths for these criteria is

$$\lim_{n\to\infty} \frac{h_{R_1}^*}{h_{R_2}^*} = \left(\frac{t_{R_1}^*}{t_{R_2}^*}\right)^{1/(\gamma_b-\gamma_s)}.$$

It depends only on γ_b and γ_s and the performance criteria.

Part (ii) shows that that relative efficiency results for RMSE apply unchanged to fixed-length CIs and minimax one-sided CIs. For example, Cheng et al. (1997) calculate bounds on the minimax MSE efficiency of local linear estimators for estimating a conditional mean and its derivatives at a boundary point. Theorem 2.1 shows that these calculations apply unchanged to give efficiency comparisons for CIs based on these estimators.

Part (iii) shows that the optimal bias-standard deviation ratio depends only on r and the performance criterion, and not on the kernel. For RMSE, we obtain $t_{\text{RMSE}}^* = \sqrt{1/r-1}$, using the same calculations as in Section 2.1. For one-sided CIs, $t_{\text{OCI},\alpha,\beta}^* = (1/r-1)(z_{1-\alpha}+z_\beta)$. For fixed-length CIs, t_{FLCI}^* can be evaluated numerically. Figures 2 and 3 plot these quantities as a function of r. As discussed in Section 2.1, the optimal bias-standard deviation ratio is larger for fixed-length CI construction (at levels $\alpha = .05$ and $\alpha = .01$) than for RMSE. Thus, for FLCI, the optimal bandwidth oversmooths relative to the RMSE optimal bandwidth.

The next theorem gives conditions for the asymptotic validity and relative efficiency of a confidence interval centered at the MSE optimal bandwidth. Following the derivations in Section 2.1, this CI takes the form $\hat{T}(h_{RMSE}^*;k) \pm \widehat{\text{se}}(h_{RMSE}^*;k)\cdot \text{cv}_{1-\alpha}(\sqrt{1/r-1})$, and its relative efficiency is given by (7).

Theorem 2.2. *Suppose that the conditions of Theorem 2.1 hold for R_{RMSE} and for $R_{FLCI,\tilde{\alpha}}$ for all $\tilde{\alpha}$ in a neighborhood of α. Let $\widehat{\text{se}}(h^*_{RMSE};k)$ be such that $\widehat{\text{se}}(h^*_{RMSE};k)/\operatorname{sd}_f(h^*_{RMSE};k)$ converges in probability to 1 uniformly over $f \in \mathcal{F}$. Then*

$$\lim_{n\to\infty} \inf_{f\in\mathcal{F}} P_f\left(T(f) \in \left\{\hat{T}(h^*_{RMSE};k) \pm \widehat{\text{se}}(h^*_{RMSE};k)\cdot \text{cv}_{1-\alpha}(\sqrt{1/r-1})\right\}\right) = 1-\alpha.$$

The asymptotic efficiency of this CI relative to the one centered at the FLCI optimal bandwidth, defined as $\lim_{n\to\infty} \frac{\inf_{h>0} R_{FLCI,\alpha}(\hat{T}(h;k))}{R_{FLCI,\alpha}(\hat{T}(h^*_{RMSE};k))}$, is given by (7). It depends only on r.

Thus, for CIs centered at the RMSE optimal bandwidth, one forms a CI by simply adding and subtracting $\text{cv}_{1-\alpha}(\sqrt{1/r-1})$ times the standard error. Table 1 gives this quantity for some common values of r. The efficiency loss from using h^*_{RMSE} rather than h^*_{FLCI} is given by (7), and is plotted in Figure 1.

3 Inference at a point

In this section, we apply the general results from Section 2 to the problem of inference about a nonparametric regression function at a point, which we normalize to be zero, so that $T(f) = f(0)$. We allow the point of interest to be on the boundary on the parameter space. Because in sharp regression discontinuity (RD) designs, discussed in detail in Section 4, the parameter of interest can be written as the difference between two regression functions evaluated at boundary points, the efficiency results in this section generalize in a straightforward manner to sharp RD.

We write the nonparametric regression model as

$$y_i = f(x_i) + u_i, \quad i = 1,\ldots,n, \tag{11}$$

where the design points x_i are non-random, and the regression errors u_i are by definition mean-zero, with variance $\text{var}(u_i) = \sigma^2(x_i)$. We consider inference about $f(0)$ based on local

polynomial estimators of order q, which can be written as

$$\hat{T}_q(h;k) = \sum_{i=1}^{n} w_q^n(x_i;h,k)y_i,$$

where the weights $w_q^n(x_i;k,h)$ are given by

$$w_q^n(x;h,k) = e_1' Q_n^{-1} m_q(x) k(x/h).$$

Here $m_q(t) = (1, t, \ldots, t^q)'$, $k(\cdot)$ is a kernel with bounded support, e_1 is a vector of zeros with 1 in the first position, and

$$Q_n = \sum_{i=1}^{n} k(x_i/h) m_q(x_i) m_q(x_i)'.$$

In other words, $\hat{T}_q(h;k)$ corresponds to the intercept in a weighted least squares regression of y_i on $(1, x_i, \ldots, x_i^q)$ with weights $k(x_i/h)$. Local linear estimators correspond to $q = 1$, and Nadaraya-Watson (local constant) estimators to $q = 0$. It will be convenient to define the equivalent kernel

$$k_q^*(u) = e_1' \left(\int_{\mathcal{X}} m_q(t) m_q(t)' k(t) \, \mathrm{d}t \right)^{-1} m_q(u) k(u), \tag{12}$$

where the integral is over $\mathcal{X} = \mathbb{R}$ if 0 is an interior point, and over $\mathcal{X} = [0\,, \infty)$ if 0 is a (left) boundary point.

We assume the following conditions on the design points and regression errors u_i:

Assumption 3.1. *For some $d > 0$, the sequence $\{x_i\}_{i=1}^n$ satisfies $\frac{1}{nh_n} \sum_{i=1}^n g(x_i/h_n) \to d \int_{\mathcal{X}} g(u)\, du$ for any bounded function g with finite support and any sequence h_n with $0 < \liminf_n h_n n^{1/(2p+1)} < \limsup_n h_n n^{1/(2p+1)} < \infty$.*

Assumption 3.2. *The random variables $\{u_i\}_{i=1}^n$ are independent and normally distributed with $Eu_i = 0$ and $\mathrm{var}(u_i) = \sigma^2(x_i)$ where $\sigma^2(x)$ is continuous at $x = 0$.*

Assumption 3.1 requires that the empirical distribution of the design points is smooth around 0. When the support points are treated as random, the constant d typically corresponds to their density at 0. The assumption of normal errors in Assumption 3.2 is made for simplicity and could be replaced with the assumption that for some $\eta > 0$, $E[u_i^{2+\eta}] < \infty$.

Because the estimator is linear in y_i, its variance doesn't depend on f, and simply corresponds to the conditional variance of a weighted least squares estimator. Therefore, as we show in Appendix B.2 in the supplemental materials, under Assumptions 3.1 and 3.2,

$$\mathrm{sd}(\hat{T}_q(h;k))^2 = \sum_{i=1}^{n} w_q^n(x_i)^2 \sigma^2(x_i) = \left(\frac{\sigma^2(0)}{dnh} \int_{\mathcal{X}} k_q^*(u)^2 \, \mathrm{d}u \right) (1 + o(1)). \tag{13}$$

The condition on the standard deviation in Equation (1) thus holds with

$$\gamma_s = -1/2 \quad \text{and} \quad S(k) = d^{-1/2} \sigma(0) \sqrt{\int_{\mathcal{X}} k_q^*(u)^2 \, \mathrm{d}u}. \tag{14}$$

Tables S1 and S2 in the supplemental materials give the constant $\int_{\mathcal{X}} k_q^*(u)^2 \, \mathrm{d}u$ for some common kernels.

On the other hand, the worst-case bias will be driven primarily by the function class $\mathcal{F}$. We consider inference under two popular function classes. First, the p-order Taylor class, a generalization of the the second-order Taylor class from Example 2.1,

$$\mathcal{F}_{\mathrm{T},p}(M) = \left\{ f \colon \left| f(x) - \textstyle\sum_{j=0}^{p-1} f^{(j)}(0) x^j / j! \right| \leq M |x|^p / p! \;\; x \in \mathcal{X} \right\}.$$

This class consists of all functions for which the approximation error from a $(p-1)$-th order Taylor approximation around 0 can be bounded by $\frac{1}{p!} M |x|^p$. It formalizes the idea that the pth derivative of f at zero should be bounded by some constant M. Using this class of functions to derive optimal estimators goes back at least to Legostaeva and Shiryaev (1971), and it underlies much of existing minimax theory concerning local polynomial estimators (see Fan and Gijbels, 1996, Chapter 3.4–3.5).

While analytically convenient, the Taylor class may not be attractive in some empirical settings because it allows f to be non-smooth and discontinuous away from 0. We therefore also consider inference under Hölder class[3],

$$\mathcal{F}_{\text{Höl},p}(M) = \left\{ f\colon |f^{(p-1)}(x) - f^{(p-1)}(x')| \leq M|x - x'|,\ x, x' \in \mathcal{X} \right\},$$

This class is the closure of the family of p times differentiable functions with the pth derivative bounded by M, uniformly over $\mathcal{X}$, not just at 0. It thus formalizes the intuitive notion that f should be p-times differentiable with a bound on the pth derivative. The case $p = 1$ corresponds to the Lipschitz class of functions.

Theorem 3.1. *Suppose that Assumption 3.1 holds. Then, for a bandwidth sequence h_n with $0 < \liminf_n h_n n^{1/(2p+1)} < \limsup_n h_n n^{1/(2p+1)} < \infty$,*

$$\overline{\text{bias}}_{\mathcal{F}_{\text{T},p}(M)}(\hat{T}_q(h_n; k)) = \frac{M h_n^p}{p!} \mathcal{B}_{p,q}^T(k)(1 + o(1)), \qquad \mathcal{B}_{p,q}^T(k) = \int_{\mathcal{X}} |u^p k_q^*(u)|\, du$$

and

$$\overline{\text{bias}}_{\mathcal{F}_{\text{Höl},p}(M)}(\hat{T}_q(h_n; k)) = \frac{M h_n^p}{p!} \mathcal{B}_{p,q}^{\text{Höl}}(k)(1 + o(1)),$$
$$\mathcal{B}_{p,q}^{\text{Höl}}(k) = p \int_{t=0}^{\infty} \left| \int_{u \in \mathcal{X}, |u| \geq t} k_q^*(u)(|u| - t)^{p-1}\, du \right| dt.$$

Thus, the first part of (1) *holds with $\gamma_b = p$ and $B(k) = M\mathcal{B}_{p,q}(k)/p!$ where $\mathcal{B}_{p,q}(k) = \mathcal{B}_{p,q}^{\text{Höl}}(k)$ for $\mathcal{F}_{\text{Höl},p}(M)$, and $\mathcal{B}_{p,q}(k) = \mathcal{B}_{p,q}^{\text{T}}(k)$ for $\mathcal{F}_{\text{T},p}(M)$.*

If, in addition, Assumption 3.2 holds, then Equation (8) *holds for the RMSE, FLCI and OCI performance criteria, with γ_b and $B(k)$ given above and γ_s and $S(k)$ given in Equation (14).*

As we will see from the relative efficiency calculation below, the optimal order of the

[3]For simplicity, we focus on Hölder classes of integer order.

local polynomial regression is $q = p - 1$ for the kernels considered here. The theorem allows $q \geq p - 1$, so that we can examine the efficiency of local polynomial regressions that are of order that's too high relative to the smoothness class (when $q < p - 1$, the maximum bias is infinite).

Under the Taylor class $\mathcal{F}_{\mathrm{T},p}(M)$, the least favorable (bias-maximizing) function is given by $f(x) = M/p! \cdot \operatorname{sign}(w_q^n(x))|x|^p$. In particular, if the weights are not all positive, the least favorable function will be discontinuous away from the boundary. The first part of Theorem 3.1 then follows by taking the limit of the bias under this function. Assumption 3.1 ensures that this limit is well-defined.

Under the Hölder class $\mathcal{F}_{\mathrm{Höl},p}(M)$, it follows from an integration by parts identity that the bias under f can be written as a sample average of $f^{(p)}(x_i)$ times a weight function that depends on the kernel and the design points. The function that maximizes the bias is then obtained by setting the pth derivative to be M or $-M$ depending on whether this weight function is positive or negative. This leads to a pth order spline function maximizing the bias. See Appendix B.2 in the supplemental materials for details.

For kernels given by polynomial functions over their support, k_q^* also has the form of a polynomial, and therefore $\mathcal{B}_{p,q}^{\mathrm{T}}$ and $\mathcal{B}_{p,q}^{\mathrm{Höl}}$ can be computed analytically. Tables S1 and S2 in the supplemental materials give these constants for selected kernels.

3.1 Kernel efficiency

It follows from Theorem 2.1 (ii) that the optimal equivalent kernel minimizes $S(k)^r B(k)^{1-r}$. Under the Taylor class $\mathcal{F}_{\mathrm{T},p}(M)$, this minimization problem is equivalent to minimizing

$$\left(\int_{\mathcal{X}} k^*(u)^2 \, \mathrm{d}u \right)^p \left(\int_{\mathcal{X}} |u^p k^*(u)| \, \mathrm{d}u \right), \tag{15}$$

The solution to this problem follows from Sacks and Ylvisaker (1978, Theorem 1) (see also Cheng et al. (1997)). We give details of the solution as well as plots of the optimal kernels in

Appendix D in the supplemental materials. In Table 2, we compare the asymptotic relative efficiency of local polynomial estimators based on the uniform, triangular, and Epanechnikov kernels to the optimal Sacks-Ylvisaker kernels.

Fan et al. (1997) and Cheng et al. (1997), conjecture that minimizing (15) yields a sharp bound on kernel efficiency. It follows from Theorem 2.1 (ii) that this conjecture is correct, and Table 2 match the kernel efficiency bounds in these papers. One can see from the tables that the choice of the kernel doesn't matter very much, so long as the local polynomial is of the right order. However, if the order is too high, $q > p-1$, the efficiency can be quite low, even if the bandwidth used was optimal for the function class or the right order, $\mathcal{F}_{\mathrm{T},p}(M)$, especially on the boundary. However, if the bandwidth picked is optimal for $\mathcal{F}_{\mathrm{T},q-1}(M)$, the bandwidth will shrink at a lower rate than optimal under $\mathcal{F}_{\mathrm{T},p}(M)$, and the resulting rate of convergence will be lower than r. Consequently, the relative asymptotic efficiency will be zero. A similar point in the context of pointwise asymptotics was made in Sun (2005, Remark 5, page 8).

The solution to minimizing $S(k)^r B(k)^{1-r}$ under $\mathcal{F}_{\mathrm{H\ddot{o}l},p}(M)$ is only known in special cases. When $p = 1$, the optimal estimator is a local constant estimator based on the triangular kernel. When $p = 2$, the solution is given in Fuller (1961) and Zhao (1997) for the interior point problem, and in Gao (2016) for the boundary point problem. See Appendix D in the supplemental materials for details, including plots of these kernels. When $p \geq 3$, the solution is unknown. Therefore, for $p = 3$, we compute efficiencies relative to a local quadratic estimator with a triangular kernel. Table 3 calculates the resulting efficiencies for local polynomial estimators based on the uniform, triangular, and Epanechnikov kernels. Relative to the class $\mathcal{F}_{\mathrm{T},p}(M)$, the bias constants are smaller: imposing smoothness away from the point of interest helps to reduce the maximum bias. Furthermore, the loss of efficiency from using a local polynomial estimator of order that's too high is smaller. Finally, one can see that local linear regression with a triangular kernel achieves high asymptotic efficiency under both $\mathcal{F}_{\mathrm{T},2}(M)$ and $\mathcal{F}_{\mathrm{H\ddot{o}l},2}(M)$, both at the interior and at a boundary, with efficiency

at least 97%, which shows that its popularity in empirical work can be justified on theoretical grounds. Under $\mathcal{F}_{\text{Höl},2}(M)$ on the boundary, the triangular kernel is nearly efficient.

3.2 Gains from imposing smoothness globally

The Taylor class $\mathcal{F}_{\text{T},p}(M)$, formalizes the notion that the pth derivative at 0, the point of interest, should be bounded by M, but doesn't impose smoothness away from 0. In contrast, the Hölder class $\mathcal{F}_{\text{Höl},p}(M)$ restricts the pth derivative to be at most M globally. How much can one tighten a confidence interval or reduce the maximum RMSE due to this additional smoothness?

It follows from Theorem 3.1 and from arguments underlying Theorem 2.1 that the risk of using a local polynomial estimator of order $p-1$ with kernel k_H and optimal bandwidth under $\mathcal{F}_{\text{Höl},p}(M)$ relative using an a local polynomial estimator of order $p-1$ with kernel k_T and optimal bandwidth under $\mathcal{F}_{\text{T},p}(M)$ is given by

$$\frac{\inf_{h>0} R_{\mathcal{F}_{\text{Höl},p}(M)}(\hat{T}(h;k_H))}{\inf_{h>0} R_{\mathcal{F}_{\text{T},p}(M)}(\hat{T}(h;k_T))} = \left(\frac{\int_{\mathcal{X}} k^*_{H,p-1}(u)^2\,\mathrm{d}u}{\int_{\mathcal{X}} k^*_{T,p-1}(u)^2\,\mathrm{d}u}\right)^{\frac{p}{2p+1}} \left(\frac{\mathcal{B}^{\text{Höl}}_{p,p-1}(k_H)}{\mathcal{B}^{T}_{p,p-1}(k_T)}\right)^{\frac{1}{2p+1}} (1+o(1)),$$

where $R_{\mathcal{F}}(\hat{T})$ denotes the worst-case performance of $\hat{T}$ over $\mathcal{F}$. If the same kernel is used, the first term equals 1, and the efficiency ratio is determined by the ratio of the bias constants $\mathcal{B}_{p,p-1}(k)$. Table 4 computes the resulting reduction in risk/CI length for common kernels. One can see that in general, the gains are greater for larger p, and greater at the boundary. In the case of estimation at a boundary point with $p=2$, for example, imposing global smoothness of f results in reduction in length of about 13–15%, depending on the kernel, and a reduction of about 10% if the optimal kernel is used.

3.3 RMSE and pointwise optimal bandwidth

We follow the literature on nonparametric efficiency bounds by using minimaxity within a smoothness class as our measure of efficiency: our relative efficiency comparisons are based

on the worst-case performance of $\hat{T}$ over a class $\mathcal{F}$, where $\mathcal{F}$ formalizes the notion that f should be "smooth." Since we take limits of bounds that hold for all $f \in \mathcal{F}$ for a given n, this approach can be called "uniform-in-f." Similarly, the honesty requirement on CIs requires that coverage converges to $1-\alpha$ uniformly over $f \in \mathcal{F}$. An alternative is to base relative efficiency comparisons and confidence statements on pointwise-in-f asymptotics. The pointwise approach has been criticized, since it can lead to "superefficient" estimators that perform poorly in finite samples (see Chapter 1.2.4 in Tsybakov, 2009). Thus, it is of interest to know for which questions these two approaches give substantively different answers. We now compare our optimal bandwidth calculations to optimal bandwidth calculations based on pointwise asymptotics.

The general results from Section 2 imply that given a kernel k and order of a local polynomial q, the RMSE-optimal bandwidth for $\mathcal{F}_{\mathrm{T},p}(M)$ and $\mathcal{F}_{\mathrm{Höl},p}(M)$ is given by

$$h^*_{\mathrm{RMSE}} = \left(\frac{1}{2pn}\frac{S(k)^2}{B(k)^2}\right)^{\frac{1}{2p+1}} = \left(\frac{\sigma^2(0)p!^2}{2pndM^2}\frac{\int_{\mathcal{X}} k_q^*(u)^2\,\mathrm{d}u}{\mathcal{B}_{p,q}(k)^2}\right)^{\frac{1}{2p+1}},$$

where $\mathcal{B}_{p,q}(k) = \mathcal{B}^{\mathrm{Höl}}_{p,q}(k)$ for $\mathcal{F}_{\mathrm{Höl},p}(M)$, and $\mathcal{B}_{p,q}(k) = \mathcal{B}^{\mathrm{T}}_{p,q}(k)$ for $\mathcal{F}_{\mathrm{T},p}(M)$.

In contrast, the optimal bandwidth based on pointwise asymptotics is obtained by minimizing the sum of the leading squared bias and variance terms under pointwise asymptotics for the case $q = p-1$. This bandwidth is given by (see, for example, Fan and Gijbels, 1996, Eq. (3.20))

$$h^*_{\mathrm{pointwise}} = \left(\frac{\sigma^2(0)p!^2}{2pdnf^{(p)}(0)^2}\frac{\int_{\mathcal{X}} k_q^*(u)^2\,\mathrm{d}u}{(\int_{\mathcal{X}} t^p k_q^*(t)\,\mathrm{d}t)^2}\right)^{\frac{1}{2p+1}}.$$

Thus, the pointwise optimal bandwidth replaces M with the pth derivative at zero, $f^{(p)}$, and $\mathcal{B}_{p,q}(k)$ with $\int_{\mathcal{X}} t^p k_q^*(t)\,\mathrm{d}t$. In general implementing this bandwidth is not feasible, because the pth derivative cannot be estimated without assuming the existence of more than p derivatives, and, if more than p derivatives are assumed to exist, setting the order of the local polynomial to $q = p-1$ is no longer optimal.

Suppose, therefore, that $f \in \mathcal{F}_p(M)$, where $\mathcal{F}_p(M)$ corresponds to either $\mathcal{F}_{\mathrm{T},p}(M)$ and

$\mathcal{F}_{\text{Höl},p}(M)$, and that it is known that the pth derivative at zero exists and equals M. Then both h^*_{RMSE} and $h^*_{\text{pointwise}}$ are feasible, and their ratio is given by

$$\frac{h^*_{\text{pointwise}}}{h^*_{\text{RMSE}}} = \left(\frac{\mathcal{B}_{p,q}(k)}{|\int_{\mathcal{X}} t^p k^*_q(t)\,\mathrm{d}t|}\right)^{\frac{2}{2p+1}} \geq 1. \tag{16}$$

The inequality obtains because the Taylor expansion used to derive the leading bias term under pointwise asymptotics effectively assumes that $f(x) = \pm Mx^p/p!$, which leads to the bias constant $|\int_{\mathcal{X}} t^p k^*_q(t)\,\mathrm{d}t|$. This choice of f is feasible under $\mathcal{F}_p(M)$, but may not maximize the bias in general.

Under $\mathcal{F}_{\text{T},p}(M)$, the inequality will be strict for $p \geq 2$, so that the pointwise optimal bandwidth will in general be too large. For example for $p = 2$ and local linear regression with the triangular kernel at the boundary, the ratio of bandwidths in Equation (16) evaluates to $\left(\frac{3/16}{1/10}\right)^{2/5} \approx 1.28588$, so that the pointwise optimal bandwidth is about 30% too large. Consequently, the minimax efficiency for root MSE is

$$(t^*_{MSE}/t_{\text{pointwise}})^{-1/5} \left(\frac{1+(t^*_{MSE})^2}{1+t^2_{\text{pointwise}}}\right)^{1/2} = (8/15)^{-1/5} \left(\frac{1+(1/2)^2}{1+(15/16)^2}\right)^{1/2} \approx 0.925.$$

On the other hand, under $\mathcal{F}_{\text{Höl},2}(M)$, Equation (16) holds with equality, so that the pointwise and minimax optimal bandwidths coincide, because, as we show in Appendix B.2 in the supplemental materials, the least favorable function is indeed given by $Mx^2/2$.

3.4 Confidence intervals based on pointwise asymptotics

Let us consider the performance of confidence intervals (CIs) justified by pointwise asymptotics. Suppose that the smoothness class is either $\mathcal{F}_{\text{T},p}(M)$ and $\mathcal{F}_{\text{Höl},p}(M)$ and denote it by $\mathcal{F}_p(M)$. Suppose, for concreteness that $p = 2$, and $q = 1$. A naïve, but popular way of constructing confidence intervals in practice is to center the CI around the estimator $\hat{T}_1(h;k)$, simply add and subtract $z_{1-\alpha/2}$ times its standard deviation, disregarding the possibility that

the estimator may be biased. If bandwidth used equals h^*_{RMSE}, then the resulting CIs are shorter than the 95% fixed-length CIs by a factor of $z_{0.975}/\operatorname{cv}_{0.95}(1/2) = 0.90$. Consequently, their coverage is 92.1% rather than the nominal 95% coverage. At the RMSE-optimal bandwidth, the worst-case bias-sd ratio equals 1/2, so disregarding the bias doesn't result in severe undercoverage. If one uses a larger bandwidth, however, the worst-case bias-sd ratio will be larger, and the undercoverage problem more severe: for example, if the bandwidth is 50% larger than h^*_{RMSE}, so that the worst-case bias-sd ratio equals $1/2 \cdot (1.5)^{(5/2)}$ the coverage is only 71.9%.

In an important recent paper, to improve the coverage properties of the naïve CI, Calonico et al. (2014) consider recentering $\hat{T}_1(h;k)$ by an estimate of the leading bias term, and adjusting the standard error estimate to account for the variability of the bias estimate. For simplicity, consider the case in which the main bandwidth and the pilot bandwidth (used to estimate the bias) are the same, and that the main bandwidth is chosen optimally in that it equals h^*_{RMSE}. In this case, their procedure amounts to using a local quadratic estimator, but with a bandwidth h^*_{RMSE}, optimal for a local linear estimator. The resulting CI obtains by adding and subtracting $z_{1-\alpha/2}$ times the standard deviation of the estimator. The maximum bias to standard deviation ratio of the estimator is given by

$$t_{\mathrm{CCT}} = (h^*_{\mathrm{RMSE}})^{5/2} \frac{M\mathcal{B}_{2,2}(k)/2}{\sigma(0)(\int k_2^*(u)^2\,\mathrm{d}u/dn)^{1/2}} = \frac{1}{2}\frac{\mathcal{B}_{2,2}(k)}{\mathcal{B}_{2,1}(k)} \left(\frac{\int_{\mathcal{X}} k_1^*(u)^2\,\mathrm{d}u}{\int_{\mathcal{X}} k_2^*(u)^2\,\mathrm{d}u}\right)^{1/2}. \tag{17}$$

The resulting coverage is given by $\Phi(t_{\mathrm{CCT}} + z_{1-\alpha/2}) - \Phi(t_{\mathrm{CCT}} - z_{1-\alpha/2})$. The CCT interval length relative to the fixed-length $1-\alpha$ CI around a local linear estimator with the same kernel and minimax MSE bandwidth is the same under both $\mathcal{F}_{\mathrm{T},p}(M)$, and $\mathcal{F}_{\mathrm{Höl},p}(M)$, and given by

$$\frac{z_{1-\alpha/2}\left(\int_{\mathcal{X}} k_2^*(u)^2\,\mathrm{d}u\right)^{1/2}}{\operatorname{cv}_{1-\alpha}(1/2)\left(\int_{\mathcal{X}} k_1^*(u)^2\,\mathrm{d}u\right)^{1/2}}(1+o(1)). \tag{18}$$

The resulting coverage and relative length is given in Table 5 for the class $\mathcal{F}_{\mathrm{T},2}(M)$, and in Table 6 for the class $\mathcal{F}_{\mathrm{Höl},2}(M)$ and $\alpha = 0.05$. One can see that although the coverage

properties are excellent (since t_{CCT} is quite low in all cases), the intervals are about 30% longer than the fixed-length CIs around the RMSE bandwidth.

Under the class $\mathcal{F}_{\text{Höl},2}(M)$, the CCT intervals are also reasonably robust to using a larger bandwidth: if the bandwidth used is 50% larger than h^*_{RMSE}, so that the bias-sd ratio in Equation (17) is larger by a factor of $(1.5)^{5/2}$, the resulting coverage is still at least 93.0% for the kernels considered in Table 6. Under $\mathcal{F}_{\text{T},2}(M)$, using a bandwidth 50% larger than h^*_{RMSE} yields coverage of about 80% on the boundary and 87% in the interior.

If one instead considers the classes $\mathcal{F}_{\text{T},3}(M)$ and $\mathcal{F}_{\text{Höl},3}(M)$ (but with h^*_{RMSE} still chosen to be MSE optimal for $\mathcal{F}_{\text{T},2}(M)$ or $\mathcal{F}_{\text{Höl},2}(M)$), then the CCT interval can be considered an undersmoothed CI based on a second order local polynomial estimator. In this case, the limiting bias-sd ratio is $t_{CCT} = 0$ and the limiting coverage is $1 - \alpha$ (this matches the pointwise-in-f coverage statements in CCT, which assume the existence of a continuous third derivative in the present context). Due to this undersmoothing, however, the CCT CI shrinks at a slower rate than the optimal CI. Thus, depending on the smoothness class, the 95% CCT CI has close to 95% coverage and efficiency loss of about 30%, or exactly 95% coverage at the cost of shrinking at a slower than optimal rate.

4 Application to sharp regression discontinuity

In this section, we apply the results for estimation at a boundary point from Section 3 to sharp regression discontinuity (RD), and illustrate them with an empirical application.

In a sharp RD, we are given data from a nonparametric regression model (11), and the goal is to estimate a jump in the regression function f at a known threshold, which we normalize to 0, so that the parameter of interest is

$$T(f) = \lim_{x \downarrow 0} f(x) - \lim_{x \uparrow 0} f(x).$$

The threshold determines participation in a binary treatment: units with $x_i \geq 0$ are treated;

units with $x_i < 0$ are controls. If the regression functions of potential outcomes are continuous at zero, then $T(f)$ measures the average effect of the treatment for units with $x_i = 0$ (Hahn et al., 2001).

For brevity, we focus on the most empirically relevant case in which the regression function f is assumed to lie in the class $\mathcal{F}_{\text{Höl},2}(M)$ on either side of the cutoff:

$$f \in \mathcal{F}_{\text{RD}}(M) = \{f_+(x)1(x \geq 0) - f_-(x)1(x < 0) \colon f_+, f_- \in \mathcal{F}_{\text{Höl},2}(M)\}.$$

We consider estimating $T(f)$ based on running a local linear regression on either side of the boundary. Given a bandwidth h and a second-order kernel k, the resulting estimator can be written as

$$\hat{T}(h;k) = \sum_{i=1}^{n} w_+^n(x_i;h,k)y_i - \sum_{i=1}^{n} w_-^n(x_i;h,k)y_i,$$

with the weight w_+^n given by

$$\begin{aligned} w_+(x;h,k) &= e_1' Q_{n,+}^{-1} m_1(x) k_+(x/h) \\ &= \frac{k_+(x/h)\sum_{i=1}^n k_+(x_i/h)(x_i^2 - x_i \cdot x)}{\sum_{i=1}^n k_+(x_i/h)\sum_{i=1}^n k_+(x_i/h)x_i^2 - (\sum_{i=1}^n k_+(x_i/h)x_i)^2}, \quad k_+(u) = k(u)1(u \geq 0), \end{aligned}$$

and $Q_{n,+} = \sum_{i=1}^n k_+(x_i/h)m_q(x_i)m_1(x_i)'$. The weights w_-^n, Gram matrix $\hat{Q}_{n,-}$ and kernel k_- are defined similarly. That is, $\hat{T}(h;k)$ is given by a difference between estimates from two local linear regressions at a boundary point, one for units with non-negative values running variable x_i, and one for units with negative values of the running variable. Let $\sigma_+^2(x) = \sigma^2(x)1(x \geq 0)$, and let $\sigma_-^2(x) = \sigma^2(x)1(x < 0)$.

In principle, one could allow the bandwidths for the two local linear regressions to be different. We show in Appendix C in the supplemental materials, however, that the loss in efficiency resulting from constraining the bandwidths to be the same is quite small unless the ratio of variances of Y_i on either side of the cutoff, $\sigma_+^2(0)/\sigma_-^2(0)$, is quite large.

It follows from the results in Section 3 that if Assumption 3.1 holds and the functions

$\sigma^2_+(x)$ and $\sigma^2_-(x)$ are right- and left-continuous, respectively, the variance of the estimator doesn't depend on f and satisfies

$$\mathrm{sd}(\hat{T}(h;k))^2 = \sum_{i=1}^{n} \left(w^n_+(x_i)^2 + w^n_-(x_i)^2\right)\sigma^2(x_i) = \frac{\int_0^\infty k_1^*(u)^2\,\mathrm{d}u}{dnh}\left(\sigma^2_+(0)+\sigma^2_-(0)\right)(1+o(1)),$$

with d defined in Assumption 3.1.

Because $\hat{T}(h;k)$ is given by the difference between two local linear regression estimators, it follows from Theorem 3.1 and arguments in Appendix B.2 in the supplemental materials that the bias of $\hat{T}(h;k)$ is maximized at the function $f(x) = -Mx^2/2\cdot(1(x\geq 0)-1(x<0))$. The worst-case bias therefore satisfies

$$\overline{\mathrm{bias}}(\hat{T}(h;k)) = -\frac{M}{2}\left(\sum_{i=1}^{n} w^n_+(x_i)x_i^2 + \sum_{i=1}^{n} w^n_-(x_i)x_i^2\right) = -Mh^2\cdot\int_0^\infty u^2 k_1^*(u)\,\mathrm{d}u\cdot(1+o(1)).$$

The RMSE-optimal bandwidth is given by

$$h^*_{RMSE} = \left(\frac{\int_0^\infty k_1^*(u)^2\,\mathrm{d}u}{(\int_0^\infty u^2 k_1^*(u)\,\mathrm{d}u)^2}\cdot\frac{\sigma^2_+(0)+\sigma^2_-(0)}{dn4M^2}\right)^{1/5}. \tag{19}$$

This definition is similar to the optimal bandwidth definition derived under pointwise asymptotics in Imbens and Kalyanaraman (2012), except that they replace $4M^2$ with $(f''_+(0) - f''_-(0))^2$, which gives infinite bandwidth if the second derivatives at zero are equal in magnitude and of opposite sign. Consequently, any feasible implementation of pointwise asymptotically optimal bandwidth will require an ad-hoc regularization term to avoid selecting an overly large bandwidth in practice[4].

The bias-standard deviation ratio at h^*_{RMSE} equals 1/2 in large samples; a two-sided CI around $\hat{T}(h^*_{RMSE};k)$ for a given kernel k can therefore be constructed as

$$\hat{T}(h^*_{RMSE};k) \pm \mathrm{cv}_{1-\alpha}(1/2)\cdot\mathrm{sd}(\hat{T}(h^*_{RMSE};k)). \tag{20}$$

[4]Furthermore, as pointed out in Section 3.3, it is not possible to estimate the second derivative without assuming the existence of more than 2 derivatives.

Alternatively, one could use the critical value $\text{cv}_{1-\alpha}(\overline{\text{bias}}(\hat{L}(h^*_{RMSE};k))/\,\text{sd}(\hat{L}(h^*_{RMSE};k)))$ based on the finite-sample bias-sd ratio.

In practice, this CI cannot be implemented directly because the variance function $\sigma^2(x)$ and the density d of x at 0 that are required to calculate h^*_{RMSE} and the standard error $\text{sd}(\hat{T}(h^*_{RMSE};k))$ are unknown. One therefore needs to replace h^*_{RMSE} and $\text{sd}(\hat{T}(h^*_{RMSE};k))$ in the previous display by their feasible versions.

Because $\text{sd}(\hat{T}(h^*_{RMSE};k))$ corresponds to the conditional variance of a weighted least squares estimator in a regression with potentially non-linear conditional expectation function f, it can be consistently estimated using the nearest neighbor variance estimator considered in Abadie and Imbens (2006) and Abadie et al. (2014); using the usual Eicker-Huber-White estimator will overestimate the conditional variance. To describe the estimator, given a bandwidth h, let $\hat{u}_i$ denote the estimated residuals, that is, for $x_i \geq 0$ $\hat{u}_i = y_i - m_1(x_i)Q^{-1}_{n,+}\sum_{j=1}^n m_1(x_j/h)k_+(x_i/h)y_i$, and $\hat{u}_i = y_i - m_1(x_i)Q^{-1}_{n,-}\sum_{j=1}^n m_1(x_j/h)k_-(x_i/h)y_i$ for $x_i < 0$. Then $\text{sd}(\hat{T}(h;k))$ can be estimated as $\widehat{\text{se}}(\hat{T}(h;k)) = \sum_{i=1}^n w^n_+(x_i)^2\hat{\sigma}^2(x_i) + \sum_{i=1}^n w^n_-(x_i)^2\hat{\sigma}^2(x_i)$, where

$$\hat{\sigma}^2(x_i) = \frac{J}{J+1}\left(Y_i - \frac{1}{J}\sum_{m=1}^{J} Y_{j(i)}\right)^2,$$

for some fixed (small) $J \geq 1$, where $j(i)$ denotes the j-th closes observation to i among units with the same sign of the running variable. In contrast, the usual Eicker-Huber-White estimator sets $\hat{\sigma}^2(x_i) = \hat{u}_i^2$.

For h^*_{RMSE}, there are two feasible choices. One can either use a plug-in estimator that replaces the unknown quantities d, $\sigma^2_-(0)$, and $\sigma^2_+(0)$ by some consistent estimates $\hat{d}$, $\hat{\sigma}^2_-(0)$, and $\sigma^2_+(0)$. Alternatively, one can try to directly minimize the finite-sample MSE over the bandwidth h,

$$\text{MSE}(h) = \frac{M^2}{4}\left(\sum_{i=1}^n \left(w^n_+(x_i;h) + w^n_-(x_i;h)\right)x_i^2\right)^2 + \sum_{i=1}^n \left(w^n_+(x_i)^2 + w^n_-(x_i)^2\right)\sigma^2(x_i), \quad (21)$$

by replacing $\sigma^2(x)$ with the estimate $\hat{\sigma}^2(x_i) = \hat{\sigma}^2_+(0)1(x \geq 0) + \hat{\sigma}^2_-(0)1(x < 0)$. This method was considered previously in Armstrong and Kolesár (2016), who show that the resulting confidence intervals will be asymptotically valid and equivalent to the infeasible CI given in Equation (20). This method has the advantage that it avoids having to estimate d, and it can also be shown to work when the covariates are discrete.

4.1 Empirical illustration

To illustrate the implementation of feasible versions of the CIs (20), we use a subset of the dataset from Ludwig and Miller (2007).

In 1965, when the Head Start federal program launched, the Office of Economic Opportunity provided technical assistance to the 300 poorest counties in the United States to develop Head Start funding proposals. Ludwig and Miller (2007) use this cutoff in technical assistance to look at intent-to-treat effects of the Head Start program on a variety of outcomes using as a running variable the county's poverty rate relative to the poverty rate of the 300th poorest county (which had poverty rate equal to approximately 59.2%). We focus here on their main finding, the effect on child mortality due to causes addressed as part of Head Start's health services. The main health services provided by Head Start comprise vaccinations, screening, and medical referrals; this variable therefore measures deaths due to causes such as tuberculosis, meningitis, or respiratory causes, but excludes injuries and neoplasms. See the appendix in Ludwig and Miller (2007) for a detailed description of this variable.

Relative to the dataset used in Ludwig and Miller (2007), we remove two observations, one corresponding to a duplicate entry for Yellowstone County, MT, and an outlier that corresponds to Yellowstone National Park, MT. Mortality data is missing for counties in Alaska. We are therefore left with 3,103 observations that correspond to US counties, with 294 of them above the poverty cutoff.

Figure 4 plots the data. To estimate the discontinuity in mortality rates, Ludwig and

Miller (2007) use a uniform kernel[5] and consider bandwidths equal to 9, 18, and 36. This yields point estimates equal to -1.895, -1.198 and -1.114 respectively, which are large effects given that the average mortality rate for counties not receiving technical assistance was 2.15 per 100,000. The p-values reported in the paper, based on bootstrapping the t-statistic (which ignores any potential bias in the estimates), are 0.036, 0.081, and 0.027. The standard errors for these estimates, obtained using the nearest neighbor method described above (with $J = 3$) are 1.038, 0.696, and 0.522.

These bandwidth choices are optimal in the sense that they minimize the RMSE expression (21) if $M = 0.038$, 0.0076, and 0.0014, respectively. Thus, for bandwidths 18 or 36 to be optimal, one has to be very optimistic about the smoothness of the regression function. For these smoothness parameters, the finite-sample critical values based on $\text{cv}_{0.95}(\overline{\text{bias}}(\hat{L}(h^*_{RMSE};k))/\,\text{sd}(\hat{L}(h^*_{RMSE};k)))$ are given by 2.152, 2.201 and 2.115 respectively, which is very close to the asymptotic value $\text{cv}_{.95}(1/2) = 2.182$. The resulting 95% confidence intervals are given by

$$(-4.154, 0.297), \qquad (-2.729, 0.333), \qquad \text{and} \qquad (-2.219, -0.010),$$

respectively. The p-values based on these estimates are given by 0.091, 0.123, and 0.047. These values are higher than those reported in the paper, as they take into account the potential bias of the estimates. Thus, unless one is confident that the smoothness parameter M is very small, the results are not significant at 5% level.

Using a triangular kernel helps to tighten the confidence intervals by about 2% in length, as predicted by the relative asymptotic efficiency results from Table 3, yielding

$$(-4.196, 0.172), \qquad (-2.977, 0.055), \qquad \text{and} \qquad (-2.286, -0.091).$$

[5]The paper states that the estimates were obtained using a triangular kernel. However, due to a bug in the code, the results reported in the paper were actually obtained using a uniform kernel.

The underlying optimal bandwidths are given by 11.8, 22.8, and 45.7, respectively. The p-values associated with these estimates are 0.072, 0.059, and 0.033, tightening the p-values based on the uniform kernel. Thus, in contrast to the findings in the paper, these results indicate that, unless one is very optimistic about the smoothness of the regression function, the effect of Head Start assistance on child mortality is not significant at the 5% level.

Appendix A Proofs of theorems in Section 2

A.1 Proof of Theorem 2.1

Parts (ii) and (iii) follow from part (i) and simple calculations. To prove part (i), note that, if it did not hold, there would be a bandwidth sequence h_n such that

$$\liminf_{n\to\infty} n^{r/2} R(\hat{T}(h_n;k)) < S(k)^r B(k)^{1-r} \inf_t t^{r-1}\tilde{R}(t,1).$$

By Equation (10), the bandwidth sequence h_n must satisfy $\liminf_{n\to\infty} h_n/n^{1/[2(\gamma_b-\gamma_s)]} > 0$ and $\limsup_{n\to\infty} h_n/n^{1/[2(\gamma_b-\gamma_s)]} < \infty$. Thus, $n^{r/2}R(\hat{T}(h_n;k)) = S(k)^r B(k)^{1-r} t_n^{r-1}\tilde{R}(t_n,1) + o(1)$ where $t_n = h_n^{\gamma_b-\gamma_s}B(k)/(n^{-1/2}S(k))$. This contradicts the display above.

A.2 Proof of Theorem 2.2

The second statement (relative efficiency) is immediate from (9). For the first statement (coverage), fix $\varepsilon > 0$ and let $\mathrm{sd}_n = n^{-1/2}(h^*_{\mathrm{RMSE}})^{\gamma_s}S(k)$ so that, uniformly over $f \in \mathcal{F}$, $\mathrm{sd}_n \,/\, \mathrm{sd}_f(\hat{T}(h^*_{\mathrm{RMSE}};k)) \to 1$ and $\mathrm{sd}_n \,/\widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}};k) \xrightarrow{p} 1$. Note that, by Theorem 2.1 and the calculations above,

$$\tilde{R}_{\mathrm{FLCI},\alpha+\varepsilon}(\hat{T}(\hat{h}^*_{\mathrm{RMSE}};k)) = \mathrm{sd}_n \cdot \mathrm{cv}_{1-\alpha-\varepsilon}(\sqrt{1/r-1})(1+o(1))$$

and similarly for $\tilde{R}_{\text{FLCI},\alpha-\varepsilon}(\hat{T}(\hat{h}^*_{\text{RMSE}};k))$. Since $\text{cv}_{1-\alpha}(\sqrt{1/r-1})$ is strictly decreasing in α, it follows that there exists $\eta > 0$ such that, with probability approaching 1 uniformly over $f \in \mathcal{F}$,

$$R_{\text{FLCI},\alpha+\varepsilon}(\hat{T}(\hat{h}^*_{\text{RMSE}};k)) < \widehat{\text{se}}(\hat{T}(\hat{h}^*_{\text{RMSE}};k)) \cdot \text{cv}_{1-\alpha}(\sqrt{1/r-1}) \\ < (1-\eta) R_{\text{FLCI},\alpha-\varepsilon}(\hat{T}(\hat{h}^*_{\text{RMSE}};k)).$$

Thus,

$$\liminf_n \inf_{f\in\mathcal{F}} P\left(Tf \in \left\{\hat{T}(\hat{h}^*_{\text{RMSE}};k) \pm \widehat{\text{se}}(\hat{T}(\hat{h}^*_{\text{RMSE}};k)) \cdot \text{cv}_{1-\alpha}(\sqrt{1/r-1})\right\}\right) \\ \geq \liminf_n \inf_{f\in\mathcal{F}} P\left(Tf \in \left\{\hat{T}(\hat{h}^*_{\text{RMSE}};k) \pm R_{\text{FLCI},\alpha+\varepsilon}(\hat{T}(\hat{h}^*_{\text{RMSE}};k))\right\}\right) \geq 1-\alpha-\varepsilon$$

and

$$\limsup_n \inf_{f\in\mathcal{F}} P\left(Tf \in \left\{\hat{T}(\hat{h}^*_{\text{RMSE}};k) \pm \widehat{\text{se}}(\hat{T}(\hat{h}^*_{\text{RMSE}};k)) \cdot \text{cv}_{1-\alpha}(\sqrt{1/r-1})\right\}\right) \\ \leq \limsup_n \inf_{f\in\mathcal{F}} P\left(Tf \in \left\{\hat{T}(\hat{h}^*_{\text{RMSE}};k) \pm R_{\text{FLCI},\alpha-\varepsilon}(\hat{T}(\hat{h}^*_{\text{RMSE}};k))(1-\eta)\right\}\right) \leq 1-\alpha+\varepsilon,$$

where the last inequality follows by definition of $R_{\text{FLCI},\alpha-\varepsilon}(\hat{T}(\hat{h}^*_{\text{RMSE}};k))$. Taking $\varepsilon \to 0$ gives the result.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A., Imbens, G. W., and Zheng, F. (2014). Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association*, 109(508):1601–1614.

Armstrong, T. B. and Kolesár, M. (2016). Optimal inference in a class of regression models. arXiv:1511.06028.

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics*, 24(6):2384–2398.

Cai, T. T. and Low, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Annals of Statistics*, 32(5):1805–1840.

Cai, T. T. and Low, M. G. (2005). Adaptive estimation of linear functionals under different performance measures. *Bernoulli*, 11(2):341–358.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2016). On the effect of bias estimation on coverage accuracy in nonparametric inference. arXiv: 1508.02973.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cheng, M.-Y., Fan, J., and Marron, J. S. (1997). On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270.

Donoho, D. L. and Low, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *The Annals of Statistics*, 20(2):944–970.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21(1):196–216.

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(1):79–99.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC.

Fuller, A. T. (1961). Relay control systems optimized for various performance criteria. In Coales, J. F., Ragazzini, J. R., and Fuller, A. T., editors, *Automatic and Remote Control: Proceedings of the First International Congress of the International Federation of Automatic Control*, volume 1, pages 510–519. Butterworths, London.

Gao, W. (2016). Minimax linear estimation at a boundary point. Technical report. Unpublished manuscript, Yale University.

Hahn, J., Todd, P. E., and van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *The Annals of Statistics*, 20(2):675–694.

Ibragimov, I. A. and Khas'minskii, R. Z. (1985). On nonparametric estimation of the value of a linear functional in gaussian white noise. *Theory of Probability & Its Applications*, 29(1):18–32.

Imbens, G. W. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959.

Legostaeva, I. L. and Shiryaev, A. N. (1971). Minimax weights in a trend detection problem of a random process. *Theory of Probability & Its Applications*, 16(2):344–349.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008.

Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554.

Ludwig, J. and Miller, D. L. (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *Quarterly Journal of Economics*, 122(1):159–208.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics*, 24(6):2399–2430.

Sacks, J. and Ylvisaker, D. (1978). Linear estimation for approximately linear models. *The Annals of Statistics*, 6(5):1122–1137.

Sun, Y. (2005). Adaptive estimation of the regression discontinuity model. working paper.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.

Zhao, L. H. (1997). Minimax linear estimation in a white noise problem. *Annals of Statistics*, 25(2):745–755.

		$1-\alpha$		
r	b	0.01	0.05	0.1
	0.0	2.576	1.960	1.645
	0.1	2.589	1.970	1.653
	0.2	2.626	1.999	1.677
	0.3	2.683	2.045	1.717
	0.4	2.757	2.107	1.772
6/7	0.408	2.764	2.113	1.777
4/5	0.5	2.842	2.181	1.839
	0.6	2.934	2.265	1.916
	0.7	3.030	2.356	2.001
2/3	0.707	3.037	2.362	2.008
	0.8	3.128	2.450	2.093
	0.9	3.227	2.548	2.187
1/2	1.0	3.327	2.646	2.284
	1.5	3.826	3.145	2.782
	2.0	4.326	3.645	3.282

Table 1: Critical values $\mathrm{cv}_{1-\alpha}(b)$ and $\mathrm{cv}_{1-\alpha}(\sqrt{1/r-1})$ for selected confidence levels, values of maximum absolute bias b, and values of r. For $b \geq 2$, $\mathrm{cv}_{1-\alpha}(b) \approx b + z_{1-\alpha/2}$ up to 3 decimal places for these values of $1-\alpha$.

		Boundary Point			Interior point		
Kernel	Order	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
Uniform $1(\lvert u\rvert \leq 1)$	0	0.9615			0.9615		
	1	0.5724	0.9163		0.9615	0.9712	
	2	0.4121	0.6387	0.8671	0.7400	0.7277	0.9267
Triangular $(1-\lvert u\rvert)_+$	0	1			1		
	1	0.6274	0.9728		1	0.9943	
	2	0.4652	0.6981	0.9254	0.8126	0.7814	0.9741
Epanechnikov $\frac{3}{4}(1-u^2)_+$	0	0.9959			0.9959		
	1	0.6087	0.9593		0.9959	1	
	2	0.4467	0.6813	0.9124	0.7902	0.7686	0.9672

Table 2: Relative efficiency of local polynomial estimators of different orders for the function class $\mathcal{F}_{\mathrm{T},p}(M)$, relative to the optimal equivalent kernel k^*_{SY}. Functional of interest is value of f at a point.

		Boundary Point			Interior point		
Kernel	Order	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
Uniform $1(\lvert u\rvert \leq 1)$	0	0.9615			0.9615		
	1	0.7211	0.9711		0.9615	0.9662	
	2	0.5944	0.8372	0.9775	0.8800	0.9162	0.9790
Triangular $(1-\lvert u\rvert)_+$	0	1			1		
	1	0.7600	0.9999		1	0.9892	
	2	0.6336	0.8691	1	0.9263	0.9487	1
Epanechnikov $\frac{3}{4}(1-u^2)_+$	0	0.9959			0.9959		
	1	0.7471	0.9966		0.9959	0.9949	
	2	0.6186	0.8602	0.9974	0.9116	0.9425	1

Table 3: Relative efficiency of local polynomial estimators of different orders for the function class $\mathcal{F}_{\text{Höl},p}(M)$. Functional of interest is value of f at a point. For $p = 1, 2$, efficiency is relative to optimal kernel, for $p = 3$, efficiency is relative to local quadratic estimator with triangular kernel.

	Boundary Point			Interior point		
Kernel	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
Uniform	1	0.855	0.764	1	1	0.848
Triangular	1	0.882	0.797	1	1	0.873
Epanechnikov	1	0.872	0.788	1	1	0.866
Optimal	1	0.906		1	0.995	

Table 4: Gains from imposing global smoothness: asymptotic risk of local polynomial estimators of order $p-1$ and a given kernel under the class $\mathcal{F}_{\text{Höl},p}(M)$ relative to risk under $\mathcal{F}_{\text{T},p}(M)$. "Optimal" refers to using optimal kernel under given smoothness class.

Kernel	Length	Coverage	t_{CCT}
Boundary			
Uniform	1.35	0.931	0.400
Triangular	1.32	0.932	0.391
Epanechnikov	1.33	0.932	0.393
Interior			
Uniform	1.35	0.941	0.279
Triangular	1.27	0.940	0.297
Epanechnikov	1.30	0.940	0.298

Table 5: Performance of CCT CIs that use minimax MSE bandwidth for local linear regression under $\mathcal{F}_{\text{T},2}$. Coverage (Coverage), bias-sd ratio (t_{CCT}), and length (Length) relative to 95% fixed-length CIs around local linear estimator that uses the same kernel and minimax MSE bandwidth.

Kernel	Length	Coverage	t_{CCT}
Boundary			
Uniform	1.35	0.948	0.138
Triangular	1.32	0.947	0.150
Epanechnikov	1.33	0.947	0.148
Interior			
Uniform	1.35	0.949	0.086
Triangular	1.27	0.949	0.110
Epanechnikov	1.30	0.949	0.105

Table 6: Performance of CCT CIs that use minimax MSE bandwidth for local linear regression under $\mathcal{F}_{\text{Höl},2}$. Coverage (Coverage), bias-sd ratio (t_{CCT}), and length (Length) relative to 95% fixed-length CIs around local linear estimator that uses the same kernel and minimax MSE bandwidth.

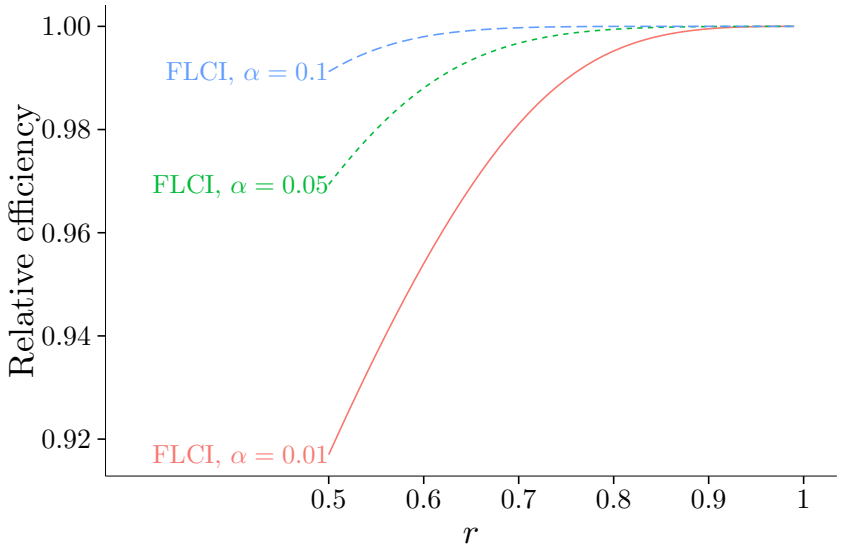

Figure 1: Efficiency of fixed-length CIs based on minimax MSE bandwidth relative to fixed-length CIs based on optimal bandwidth.

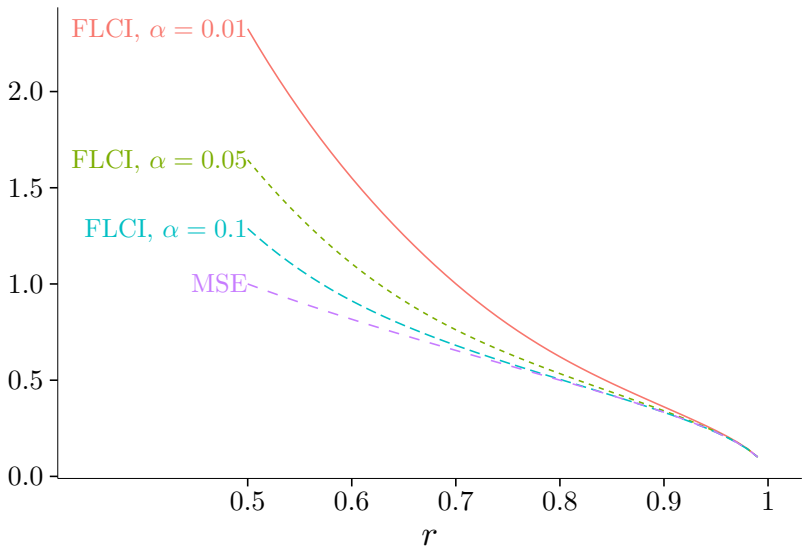

Figure 2: Optimal ratio of maximum bias to standard deviation for fixed length CIs (FLCI), and maximum MSE (MSE) performance criteria.

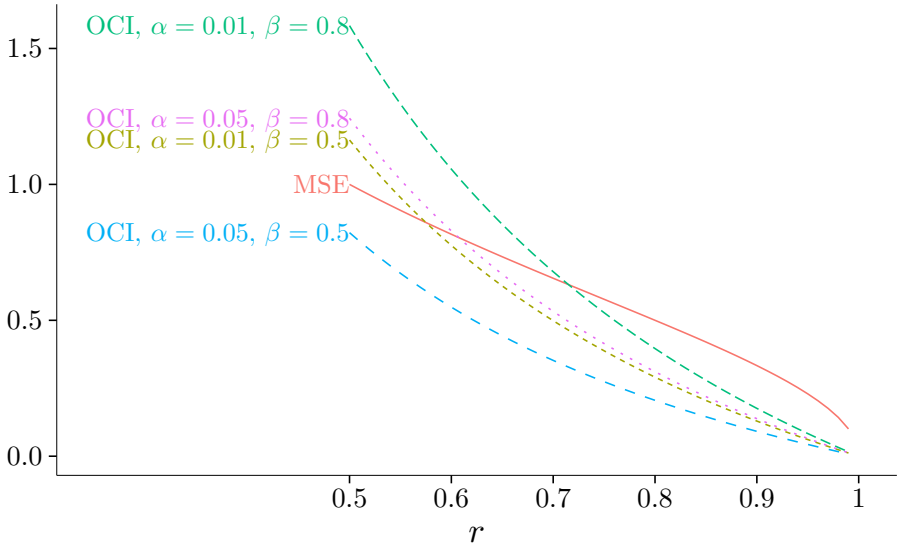

Figure 3: Optimal ratio of maximum bias to standard deviation for one-sided CIs (OCI), and maximum MSE (MSE) performance criteria.

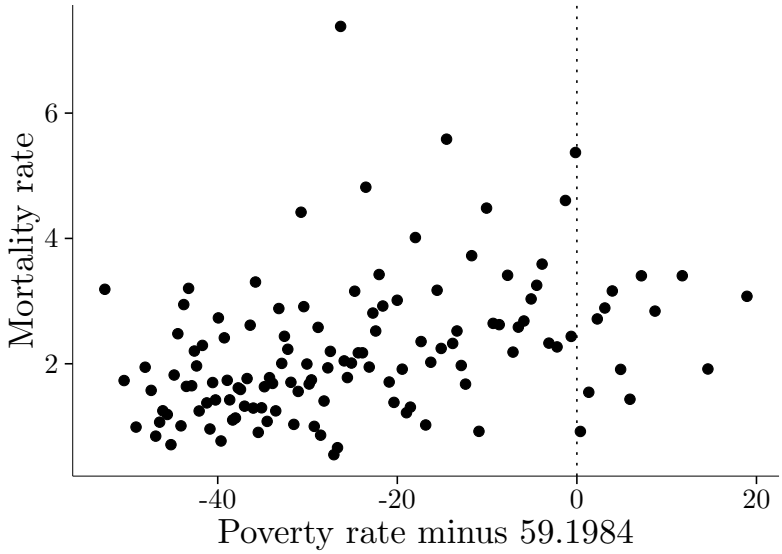

Figure 4: Average county mortality rate per 100,000 for children aged 5–9 over 1973–83 due to causes addressed as part of Head Start's health services (labeled "Mortality rate") plotted against poverty rate in 1960 relative to 300th poorest county. Each point corresponds to an average for 25 counties. Data are from Ludwig and Miller (2007).