Simple and Honest Confidence Intervals in Nonparametric Regression*

Timothy B. Armstrong† Michal Kolesár‡

Yale University Princeton University

June 6, 2019

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 use critical values that take into account the possible bias of the estimator upon which the CIs are based. We show that this approach leads to CIs that are more efficient than conventional CIs that achieve coverage by undersmoothing or subtracting an estimate of the bias. 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 $n^{-2/5}$, the appropriate critical value for 95% CIs is 2.18, rather than the usual 1.96 critical value. We illustrate our results in a Monte Carlo analysis and an empirical application.

*We thank Don Andrews, Sebiastian Calonico, Matias Cattaneo, Max Farrell, Christoph Rothe and numerous seminar and conference participants for helpful comments and suggestions. We thank Kwok Hao Lee for research assistance. 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 or the regression kink parameter, or the value of a density or its derivatives at a point. A popular approach to estimation of $T(f)$ is to use kernel or local polynomial estimators. These estimators are both simple to implement, and highly efficient in terms of their mean squared error (MSE) properties (Fan, 1993; Cheng et al., 1997). CIs are typically formed by undersmoothing (choosing the bandwidth to shrink more quickly than the MSE optimal bandwidth) or bias-correction (subtracting an estimate of the estimator's bias).

In this paper, we propose a simple alternative approach to forming CIs based on these estimators that is more efficient than both undersmoothing and bias-correction in the sense that it leads to shorter CIs while maintaining coverage over the same parameter space $\mathcal{F}$ for f (which typically places bounds on derivatives of f). In particular, one simply adds and subtracts the estimator's standard error times a critical value that is larger than the usual normal quantile $z_{1-\alpha/2}$, and takes into account the possible bias of the estimator.[1] Asymptotically, these CIs correspond to fixed-length CIs as defined in Donoho (1994), and so we refer to them as fixed-length CIs. We show that the critical value depends only on (1) the order of the derivative that one bounds to define the parameter space $\mathcal{F}$; and (2) the criterion used to choose the bandwidth. In particular, if the MSE optimal bandwidth is used with a local linear estimator, computing our CI at the 95% coverage level amounts to replacing the usual critical value $z_{0.975} = 1.96$ with 2.18.

When the criterion for bandwidth choice is the length of the resulting CI, we show that the resulting bandwidth is in fact *larger* than the MSE optimal bandwidth. This contrasts with the work of Hall (1992) and Calonico et al. (2018) on optimality of undersmoothing. 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; this also ensures correct coverage regardless of the bandwidth sequence. While the fixed-length CIs shrink at the optimal rate, undersmoothed CIs shrink more slowly. We also show that fixed-length CIs are about 30% shorter than bias-corrected CIs, once the standard error is adjusted to take into account the variability of the bias estimate (Calonico et al. (2014) show that doing so is important for maintaining coverage).

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

The oversmoothing relative to the MSE optimal bandwidth is relatively modest: under a range of conditions most commonly used in practice, a fixed-length CI centered at the MSE optimal bandwidth is 99% efficient relative to using the CI optimal bandwidth. Therefore, a practically attractive implementation of our CIs is to simply center them around an estimator with MSE optimal bandwidth, rather than reoptimizing the bandwidth for length and coverage of the CI.

A key requirement that underlies our results is the notion of honesty: as in Li (1989), we require that the CIs cover the true parameter asymptotically at the nominal level uniformly over the parameter space $\mathcal{F}$. Furthermore, we allow this parameter space to grow with the sample size. The notion of honesty is closely related to the use of the minimax criterion used to derive the MSE efficiency results: in both cases, one requires good performance uniformly over the parameter space $\mathcal{F}$. The requirement that the CIs be honest is necessary for good finite-sample performance. In contrast, approaches to inference based on pointwise-in-f asymptotics, such as using bandwidths that optimize the pointwise-in-f asymptotic MSE can lead to arbitrarily poor finite-sample behavior, as we discuss further in Section 4.1. To illustrate the practical importance of this point, we conduct a Monte Carlo study in which we show that commonly used CIs based on plug-in bandwidths that attempt to estimate this pointwise-in-f optimal bandwidth exhibit severe undercoverage, even when combined with undersmoothing or bias-correction.

When the parameter space places a bound M on a derivative of f, our CIs require this bound to be specified explicitly. While this may appear to be a disadvantage of our particular approach, due to impossibility results of Low (1997), Cai and Low (2004), and Armstrong and Kolesár (2018a), this cannot be avoided, regardless of how one forms the CI, without making further restrictions on the function f. In particular, these papers show that, without additional assumptions on the parameter space, one cannot use a data-driven method to estimate M and maintain coverage over the whole parameter space—any other method that appears to avoid making this choice must do so *implicitly*. For example, an apparent advantage of undersmoothing is that it leads to correct coverage for any fixed smoothness constant M. However, as we discuss in detail in Section 4.2, a more accurate description of undersmoothing is that for each sample size n, it implicitly chooses a constant M_n under which coverage is controlled. Given a sequence of undersmoothed bandwidths, we show how M_n can be calculated explicitly. One can then obtain a shorter CI with the same coverage properties by computing a fixed-length CI for the corresponding M_n. Regardless of how one chooses M, the fixed-length CIs we propose are more efficient than undersmoothed or bias-corrected CIs that use the same (implicit or explicit) choice of M. In fact, it follows from the calculations in Donoho (1994) and Armstrong and Kolesár (2018a) that our CIs, when constructed using a length-optimal or MSE-optimal bandwidth, are highly efficient among *all* honest CIs: no other approach to inference can substantively improve on their length, while still maintaining coverage.

As an alternative to choosing M a priori, one can place additional conditions on the function f that allow for an upper bound on M to be obtained. To maintain efficiency of the resulting CI, however, care must be taken in doing so: if M is a bound on the pth derivative, and one imposes a bound $\tilde{M}$ on the $(p+1)$th derivative in order to estimate M, then the optimal CI will be based on a different estimator and will depend on the new bound $\tilde{M}$. To avoid such issues, we propose a regularity class that relates a global polynomial approximation to smoothness of the function f near the point of interest, and we show formally that, for this class, one can obtain a valid and highly efficient CI using a global polynomial rule of thumb suggested by Fan and Gijbels (1996). However, given the additional assumptions required by this (or any) data driven choice of M, we recommend that this approach be used as a starting point for sensitivity analysis allowing for other choices of M.

Another approach to data-driven choices of M is to use "self-similarity" conditions, as suggested by Giné and Nickl (2010), which relate the maximum and minimum bias at different bandwidths. Bull (2012) and Chernozhukov et al. (2014) have obtained rate optimal confidence bands under such conditions, which, like the CIs considered here, use a critical value based on an upper bound on the bias. While these results for confidence bands could be extended to cover the problem of constructing CIs for a scalar parameter, obtaining sharp critical values appears to be very difficult. Indeed, the results of Armstrong (2018) show that the sharp form of such CIs must depend to first order on auxiliary constants used to define self-similarity. Nonetheless, our approach of bounding local smoothness using a global polynomial approximation is inspired by the self-similarity approach taken by this literature, and we see it as being in the same spirit. Schennach (2015) also uses an upper bound on the bias based on an estimated smoothness constant. While the coverage of the resulting CIs is pointwise-in-f, it is plausible that the CIs are honest under additional auxiliary conditions, similar in spirit to self-similarity.

In addition to calculating the relative efficiency of CIs constructed using different bandwidths, our results allow us to calculate the relative efficiency of CIs constructed using different kernels. In particular, we show that the relative efficiency of kernels for the CIs we propose is *the same* as the relative efficiency of the estimates in terms of MSE. Thus, relative efficiency calculations for MSE, such as the ones in Fan (1993), Cheng et al. (1997), and Fan et al. (1997) for estimation of a nonparametric mean at a point (estimation of $f(x_0)$ for some x_0) that motivate much of empirical practice in the applied regression discontinuity literature, translate directly to CI construction. Despite their importance in motivating empirical practice, however, such results are subject to a technical critique about how the parameter space is specified: rather than placing a bound on a derivative of f (a Hölder condition), currently available relative efficiency results place assumptions directly on the error of a Taylor approximation at a particular point, so that some "nonsmooth" functions are in fact not ruled out.[2] To address this, we

[2]See Imbens and Wager (2019), as well as our discussion in Section 3.1 for an elaboration of this critique.

derive the minimax performance of local polynomial estimators under Hölder restrictions on f. These results confirm that the local polynomial estimators used in empirical practice are also highly efficient under Hölder restrictions on f. Furthermore, while we focus on asymptotic CIs and relative efficiency, these results include a derivation of the finite-sample worst-case bias of local polynomial estimators under Hölder restrictions, which was used by Kolesár and Rothe (2018) to form finite-sample valid CIs in a fixed-design regression setting. These findings may be of independent interest.

The requirement of honesty is also important to ensure that our concept of optimality is well-defined and consistent. As discussed above, it allows us to consider bandwidth or kernel efficiency for constructing CIs. In addition, it also allows us to formally show that using local polynomial regression of an order that's too high given the amount of smoothness imposed is suboptimal. In contrast, under pointwise-in-f asymptotics, high-order local polynomial estimates are superefficient at every point in the parameter space (see Chapter 1.2.4 in Tsybakov, 2009, and Brown et al., 1997).

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 uncertainty associated with the effect size is much larger than originally reported, 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. Our results hold under essentially the same conditions, which apply in many classical nonparametric settings. The CIs we consider in this paper are applications of the fixed-length CIs proposed by Donoho (1994), which have also been studied recently in Armstrong and Kolesár (2018a) and in contemporaneous and subsequent work by Kolesár and Rothe (2018) and Imbens and Wager (2019). In contrast to the finite-sample approach taken in these papers, we focus on asymptotic results. This allows for the additional simplifications and insights into relative efficiency that

are the subject of this paper.[3]

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, sharp and fuzzy RD, and it discusses practical implementation issues, including a rule of thumb for choosing M. Section 4 gives a theoretical comparison of our fixed-length CIs to other approaches, and Section 5 compares them in a Monte Carlo study. Finally, Section 6 presents an empirical application based on Ludwig and Miller (2007). Appendix A gives proofs of the results in Section 2. Additional results are collected in Supplemental Appendices.

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 $\mathcal{F} = \mathcal{F}(M)$, which places "smoothness" conditions on f, where M indexes the level of smoothness. We focus on classical nonparametric function classes, in which M corresponds to a bound on a derivative of f of a given order. We allow $M = M_n$ to grow with the sample size n.

We have available a class of estimators $\hat{T}(h;k)$, indexed by a bandwidth $h = h_n > 0$ and a kernel k. Let $\widehat{\text{se}}(h;k)$ denote the standard error of $\hat{T}(h;k)$, an estimate of its standard deviation $\text{sd}_f(\hat{T}(h;k))$. We assume that a central limit theorem applies to $\hat{T}(h;k)$, so that in large samples, the t-statistic $[\hat{T}(h;k) - T(f)]/\widehat{\text{se}}(h;k)$ will be approximately normal with variance 1 and mean given by the ratio of bias to standard deviation, $t_f = (E_f[\hat{T}(h;k) - T(f)])/\text{sd}_f(\hat{T}(h;k))$. Since t_f depends on the unknown function f, this ratio is unknown. Note, however, that we can bound $|t_f|$ by the worst-case ratio of bias to standard deviation (bias-sd ratio), $t_{\mathcal{F}} = \sup_{f\in\mathcal{F}}|E_f[\hat{T}(h;k) - T(f)]|/\text{sd}_f(\hat{T}(h;k))$. Therefore, if this bias-sd ratio can be computed up to asymptotically negligible terms, we can construct an honest CI as

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

where the approximate bias-sd ratio t satisfies $t = t_{\mathcal{F}}(1+o(1))$, and $\text{cv}_{1-\alpha}(t)$ is the $1-\alpha$ quantile of the folded normal distribution $|N(t,1)|$, or, equivalently, the square root of the $1-\alpha$ quantile of a χ^2 distribution with 1 degree of freedom, and non-centrality parameter t^2, which is readily available in statistical software. For easy reference, we list these critical values in Table 1 for selected values of t. Because the quantiles of a χ^2 distribution are increasing in its non-centrality

[3]While Donoho (1994) and Armstrong and Kolesár (2018a) give asymptotic forms of some of their efficiency bounds in our setting, these are different from and complementary to the ones given here: whereas we consider relative efficiency of estimators and fixed-length CIs based on different kernels and bandwidths, Donoho (1994) and Armstrong and Kolesár (2018a) bound the scope for efficiency gains from CIs that do not fall into this class. Donoho (1994) and Armstrong and Kolesár (2018a) find that the scope for further improvement is small, which motivates our focus on this class of estimators and CIs. See Remark 2.2 for further discussion.

parameter, replacing t_f with an upper bound that is valid for all $f \in \mathcal{F}$ yields a CI that is honest over $\mathcal{F}$. The CI in (1) is an approximate version of a fixed-length confidence interval (FLCI) studied in Donoho (1994), who replaces $\widehat{\text{se}}(h;k)$ with $\text{sd}_f(\hat{T}(h;k))$ in the definition of this CI, and assumes $\text{sd}_f(\hat{T}(h;k))$ is constant over f, in which case its length will be fixed. We thus refer to CIs of this form as "fixed-length", even though $\widehat{\text{se}}(h;k)$ is random.

To motivate our main regularity condition (4) below that will facilitate studying the performance of these FLCIs and allow for an easy computation of the bias-sd ratio t, suppose that the standard deviation and the worst-case bias of the estimator $\hat{T}(h;k)$,

$$\overline{\text{bias}}(\hat{T}(h;k)) = \sup_{f\in\mathcal{F}} |E_f\hat{T}(h;k) - T(f)|,$$

scale as powers of h. In particular, suppose that, for some $\gamma_b > 0$, $\gamma_s < 0$, $B(k) > 0$ and $S(k) > 0$,

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

where the $o(1)$ term in the second equality is uniform over $f \in \mathcal{F}$. We show in Supplemental Appendix B that this condition will hold whenever the renormalization heuristics of Donoho and Low (1992) can be formalized. This includes most classical nonparametric problems, such as estimation of a density or a conditional mean, or its derivative, evaluated at a point (which may be a boundary point). In Section 3.1, we show that (2) holds with $\gamma_b = p$, and $\gamma_s = -1/2$ under mild regularity conditions when $\hat{T}(h;k)$ is a local polynomial estimator of a conditional mean at a point, and $\mathcal{F}(M)$ consists of functions with pth derivative bounded by M. The second condition in (2) implies that the standard deviation does not depend on the underlying function f asymptotically. In certain settings, such as density estimation (see Supplemental Appendix C.1), this may require choosing a localized sequence of parameter spaces $\mathcal{F}_n$, similar to local asymptotic minimax results in parametric settings (e.g., Section 8.7 in van der Vaart, 1998). While we allow for such dependence, we keep any dependence of $\mathcal{F}$ on n implicit in our notation in the main text.

Under (2), we can use the ratio $t = h^{\gamma_b-\gamma_s}MB(k)/(n^{-1/2}S(k))$ of the leading worst-case bias and standard deviation terms to compute the critical value $\text{cv}_{1-\alpha}(t)$ in (1). Analogously to the two-sided case, honest one-sided $1-\alpha$ CIs based on $\hat{T}(h;k)$ can be constructed by subtracting the standard error times a $1-\alpha$ quantile of the distribution $\mathcal{N}(t,1)$. This is asymptotically equivalent to the CI

$$[\hat{T}(h;k) - h^{\gamma_b}MB(k) - z_{1-\alpha}h^{\gamma_s}n^{-1/2}S(k)\ ,\ \infty), \tag{3}$$

which subtracts the maximum bias, in addition to subtracting $z_{1-\alpha}$ times the standard deviation,

from $\hat{T}(h;k)$.

Remark 2.1. One could also form honest two-sided 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, forming the CI as $\hat{T}(h;k) \pm (h^{\gamma_b} MB(k) + z_{1-\alpha/2} \cdot \widehat{\text{se}}(h;k))$. 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 (1).

To discuss the optimal choice of bandwidth h and compare efficiency of different kernels k in forming one- and two-sided CIs, and compare the results to the bandwidth and kernel efficiency results for estimation, it will be useful to introduce notation for a generic performance criterion. 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 $\hat{T} - T(f) \sim N(b, s^2)$. For FLCIs, we can take their half-length as the criterion, which leads to

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

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 quantiles of excess length. For CI of the form (3), its worst-case β quantile of excess length is given by $R_{\text{OCI},\alpha,\beta}(\hat{T}(h;k)) = \sup_{f\in\mathcal{F}} q_{f,\beta}(Tf - \hat{T}(h;k) + h^{\gamma_b} MB(k) + z_{1-\alpha} h^{\gamma_s} n^{-1/2} S(k))$, where $q_{f,\beta}(Z)$ is the β quantile of a random variable Z. 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 $\tilde{R}_{\text{OCI},\alpha,\beta}(b,s) = 2b + (z_{1-\alpha} + z_\beta)s$. Finally, to evaluate $\hat{T}(h;k)$ as an estimator we use root mean squared error (RMSE) as the performance criterion:

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

The key regularity condition that we impose on the class of estimators $\hat{T}(h;k)$ is that their performance can be approximated in large samples by the performance of a normally distributed estimator with bias and standard deviation that scale as powers of h,

$$R(\hat{T}(h;k)) = \tilde{R}(h^{\gamma_b} MB(k), h^{\gamma_s} n^{-1/2} S(k))(1 + o(1)). \tag{4}$$

For the performance criteria above, if the estimator $\hat{T}(h;k)$ satisfies an appropriate central limit theorem, and (2) holds, condition (4) will hold so long as the estimator is centered, so that, up to asymptotically negligible terms, its maximum and minimum bias over $\mathcal{F}$ sum to zero,

$\sup_{f\in\mathcal{F}} E_f(\hat{T}(h;k) - T(f)) = -\inf_{f\in\mathcal{F}} E_f(\hat{T}(h;k) - T(f))(1+o(1))$.[4] In Section 3.1, we verify it for the problem of estimation of a conditional mean at a point. For estimation of certain smooth non-linear functionals of the regression function or non-parametric density, including fuzzy regression discontinuity discussed in Section 3.3, and estimating a bidder valuation in first price auctions discussed in Supplemental Appendix C.2, moments of the estimator may not exist. In these cases, one can use Theorems B.1 and B.2 in Supplemental Appendix B to verify (4), which only require a weaker version of (2) stated in terms of convergence in distribution rather than moments, so long as one truncates unbounded loss functions.

We also assume that $\tilde{R}$ is homogeneous of degree one,

$$\tilde{R}(tb, ts) = t\tilde{R}(b, s) \quad \text{for all } t > 0. \tag{5}$$

This condition holds for all three criteria considered above. This allows us to simplify the right-hand side of (4). In particular, using the bias-sd ratio $t = h^{\gamma_b-\gamma_s}MB(k)/(n^{-1/2}S(k))$, write the bandwidth as $h = \left(tn^{-1/2}S(k)/(MB(k))\right)^{1/(\gamma_b-\gamma_s)}$. Substituting this expression in (4) and using (5) gives

$$\begin{aligned} R(\hat{T}(h;k)) &= \tilde{R}(t^r n^{-r/2} M^{1-r} S(k)^r B(k)^{1-r}, t^{r-1} n^{-r/2} M^{1-r} S(k)^r B(k)^{1-r})(1+o(1)) \\ &= n^{-r/2} M^{1-r} S(k)^r B(k)^{1-r} t^{r-1} \tilde{R}(t,1)(1+o(1)), \end{aligned} \tag{6}$$

where $r = \gamma_b/(\gamma_b - \gamma_s)$. Since the performance criterion converges at $n^{r/2}$ when M is fixed, we refer to r as the rate exponent (this matches the definition in, e.g., Donoho and Low 1992). Under (6), the asymptotically optimal bandwidth for a given performance criterion R is $h_R^* = (n^{-1/2}S(k)t_R^*/(MB(k)))^{1/(\gamma_b-\gamma_s)}$, with $t_R^* = \operatorname{argmin}_t t^{r-1}\tilde{R}(t,1)$.

Assuming t_R^* is finite and strictly greater than zero, the optimal bandwidth decreases at the rate $(nM^2)^{-1/[2(\gamma_b-\gamma_s)]}$ regardless of the performance criterion—the performance criterion only determines the optimal bandwidth constant. Since the approximation (4) 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 $(nM^2)^{-1/[2(\gamma_b-\gamma_s)]}$. For our main results, we assume directly that optimal bandwidth sequences decrease at this rate:

$$M^{r-1}n^{\frac{r}{2}}R(\hat{T}(h_n;k)) \to \infty \text{ for any } h_n \text{ with } h_n(nM^2)^{\frac{1}{2(\gamma_b-\gamma_s)}} \to \infty \text{ or } h_n(nM^2)^{\frac{1}{2(\gamma_b-\gamma_s)}} \to 0. \tag{7}$$

[4]This centering condition holds automatically by a symmetry argument for kernel or local polynomial estimators if f is a conditional mean or a density, $T(f)$ is its value or its derivative at a point, or a regression discontinuity parameter, and $\mathcal{F}$ bounds its derivatives. In other cases, (4) will hold when the estimator is recentered by subtracting $\mathfrak{B} = (\sup_{f\in\mathcal{F}} E_f(\hat{T}(h;k) - T(f)) + \inf_{f\in\mathcal{F}} E_f(\hat{T}(h;k) - T(f)))/2$, or an estimate $\hat{\mathfrak{B}}$ of $\mathfrak{B}$ that is consistent in the sense that $(\hat{\mathfrak{B}} - \mathfrak{B})/\widehat{\mathrm{se}}(h;k)$ converges in probability to zero, uniformly over $\mathcal{F}$. Recentering the estimator in this way improves the estimator's performance under the criteria that we consider.

Condition (7) 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.[5]

The next theorem collects implications of these derivations for the performance of different kernels. In particular, we consider minimax performance over bandwidth sequences, that is, bandwidth sequences h_n that achieve the asymptotically best possible worst-case performance in large samples in the sense that $M^{r-1}n^{r/2}(R(\hat{T}(h_n;k)) - \inf_{h>0} R(\hat{T}(h;k))) = o(1)$.

Theorem 2.1. *Let R be a performance criterion 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 (4) holds for any bandwidth sequence h_n with $\liminf_{n\to\infty} h_n(nM^2)^{1/[2(\gamma_b-\gamma_s)]} > 0$ and $\limsup_{n\to\infty} h_n(nM^2)^{1/[2(\gamma_b-\gamma_s)]} < \infty$, and suppose that Equations (5) and (7) hold. Define h_R^* and t_R^* as 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

$$\begin{aligned} M^{r-1}n^{r/2} \inf_{h>0} R(\hat{T}(h;k)) &= M^{r-1}n^{r/2} R(\hat{T}(h_R^*;k)) + o(1) \\ &= S(k)^r B(k)^{1-r} (t_R^*)^{r-1} \tilde{R}(t_R^*, 1) + o(1). \end{aligned}$$

(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 we consider two performance criteria R_1 and R_2 satisfying the conditions above, 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. If (2) holds, the asymptotically optimal bias-sd 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.

[5] In typical settings, we will need the optimal bandwidth h_R^* to shrink at a rate such that $(h_R^*)^{-2\gamma_s}n \to \infty$ and $h_R^* \to 0$. If M is fixed, this simply requires that $\gamma_b - \gamma_s > 1/2$, which basically amounts to a requirement that $\mathcal{F}(M)$ imposes enough smoothness so that the problem is not degenerate in large samples. If $M = M_n \to \infty$, then the condition also requires $n^{r/2}M^{r-1} \to \infty$, so that M does not increase too quickly.

Part (i) gives the optimal bandwidth formula for a given performance criterion. The performance criterion only determines the optimal bandwidth constant (the optimal bias-sd ratio) t^*_R.

Part (ii) shows that relative kernel efficiency results do not depend on the performance criterion. In particular, known kernel efficiency results under the RMSE criterion such as those in Fan (1993), Cheng et al. (1997), and Fan et al. (1997) apply unchanged to other performance criteria such as length of FLCIs, excess length of one-sided CIs, or expected absolute error.

Part (iii) shows that the optimal bias-sd ratio for a given performance criterion depends on $\mathcal{F}$ only through the rate exponent r, and does not depend on the kernel. The optimal bias-sd ratio for RMSE, FLCI and OCI, respectively, are

$$\begin{aligned}
t^*_{\mathrm{RMSE}} &= \operatorname*{argmin}_{t>0} t^{r-1}\tilde{R}_{\mathrm{RMSE}}(t,1) = \operatorname*{argmin}_{t>0} t^{r-1}\sqrt{t^2+1} = \sqrt{1/r-1},\\
t^*_{\mathrm{FLCI}} &= \operatorname*{argmin}_{t>0} t^{r-1}\tilde{R}_{\mathrm{FLCI},\alpha}(t,1) = \operatorname*{argmin}_{t>0} t^{r-1}\,\mathrm{cv}_{1-\alpha}(t), \qquad \text{and}\\
t^*_{\mathrm{OCI}} &= \operatorname*{argmin}_{t>0} t^{r-1}\tilde{R}_{\mathrm{OCI},\alpha,\beta}(t,1) = \operatorname*{argmin}_{t>0} t^{r-1}[2t+(z_{1-\alpha}+z_\beta)] = (1/r-1)\frac{z_{1-\alpha}+z_\beta}{2}.
\end{aligned}$$

Figures 1 and 2 plot these quantities as a function of r. Note that the optimal bias-sd ratio is larger for FLCIs (at levels $\alpha = .05$ and $\alpha = .01$) than for RMSE. Since h is increasing in t, it follows that, for FLCI, the optimal bandwidth *oversmooths* relative to the RMSE optimal bandwidth.

Remark 2.2. Theorem 2.1 does not address whether further efficiency improvements are possible by using estimators that do not fall into the class $\hat{T}(h;k)$, or by using variable length CIs. However, it follows from Donoho (1994) and Armstrong and Kolesár (2018a) that, in typical settings where our results hold, little further improvement is possible. In particular, these papers give efficiency bounds that, applied to our setting, yield asymptotic lower bounds for $R(\hat{T}^*)/R(\hat{T}(h^*;k^*))$, where $\hat{T}^*$ is the optimal estimator or CI among all procedures (for CIs, this includes variable length CIs, with performance measured in terms of expected length), and h^* and k^* are the optimal bandwidth and kernel. These asymptotic lower bounds depend only on the rate exponent r, and so can be used along with the bounds in Theorem 2.1 to obtain the efficiency of a particular kernel and bandwidth relative to the fully optimal procedure.

One can also form FLCIs centered at the estimator that is optimal for different performance criterion R as $\hat{T}(h^*_R;k) \pm \widehat{\mathrm{se}}(h^*_R;k)\cdot \mathrm{cv}_{1-\alpha}(t^*_R)$. The critical value $\mathrm{cv}_{1-\alpha}(t^*_R)$ depends only on the rate exponent r and the performance criterion R. In particular, the CI centered at the RMSE optimal estimator takes this form with $t^*_{\mathrm{RMSE}} = \sqrt{1/r-1}$, which yields the CI

$$\hat{T}(h^*_{\mathrm{RMSE}};k) \pm \mathrm{cv}_{1-\alpha}(\sqrt{1/r-1})\cdot \widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}};k), \tag{8}$$

Table 1 reports this critical value $\text{cv}_{1-\alpha}(\sqrt{1/r-1})$ for rate exponents r commonly encountered in practice. By (6), the resulting CI is wider than the one computed using the FLCI optimal bandwidth by a factor of

$$\frac{(t^*_{\text{FLCI}})^{r-1} \cdot \text{cv}_{1-\alpha}(t^*_{\text{FLCI}})}{(t^*_{\text{RMSE}})^{r-1} \cdot \text{cv}_{1-\alpha}(t^*_{\text{RMSE}})}. \tag{9}$$

Figure 3 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. For example, if $r = 4/5$, to construct an honest 95% FLCI based on an estimator with bandwidth chosen to optimize RMSE, one simply adds and subtracts the standard error multiplied by 2.18 (rather than the usual 1.96 critical value), and the corresponding CI is less than 1% longer than the one with bandwidth chosen to optimize CI length. The next theorem gives a formal statement.

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 $\widehat{\text{se}}(h^*_{\text{RMSE}}; k)$ be such that $\widehat{\text{se}}(h^*_{\text{RMSE}}; k)/[(h^*_{\text{RMSE}})^{\gamma_s} n^{-1/2} S(k)]$ converges in probability to 1 uniformly over $f \in \mathcal{F}$. Then*

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

The asymptotic efficiency of this CI relative to the one centered at the FLCI optimal bandwidth, defined as $\lim_{n\to\infty} \frac{\inf_{h>0} R_{FLCI,\alpha}(\hat{T}(h;k))}{R_{FLCI,\alpha}(\hat{T}(h^*_{\text{RMSE}};k))}$, is given by (9). It depends only on r.

3 Applications

In this section, we apply the general results from Section 2 to the problem of inference about a nonparametric regression function at a point, and to regression discontinuity. Supplemental Appendix C discusses two additional applications: estimation of a density at a point, and estimation of a bidder valuation in first-price auctions.

3.1 Inference at a point

Consider 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 to lie on the boundary of the support. We write the nonparametric regression model as

$$y_i = f(x_i) + u_i, \quad i = 1, \ldots, n, \tag{10}$$

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

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

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

$$w_q^n(x;h,k) = e_1' Q_n^{-1} m_q(x)k(x/h), \qquad Q_n = \sum_{i=1}^{n} k(x_i/h)m_q(x_i)m_q(x_i)'. \tag{11}$$

Here $m_q(t) = (1, t, \ldots, t^q)'$, $k(\cdot)$ is a kernel with bounded support, and e_1 is a vector of zeros with 1 in the first position. Thus, $\hat{T}_q(h;k)$ 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. *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 some $d > 0$, and for any bounded function g with finite support and any sequence h_n with $0 < \liminf_n h_n(nM^2)^{1/(2p+1)} < \limsup_n h_n(nM^2)^{1/(2p+1)} < \infty$.*

Assumption 3.2. *The random variables $\{u_i\}_{i=1}^n$ are independent with $Eu_i = 0$, $\text{var}(u_i) = \sigma^2(x_i)$ and $Eu_i^{2+\eta} \leq 1/\eta$ for some $\eta > 0$, and the variance function $\sigma^2(x)$ is continuous at $x = 0$ with $\sigma^2(0) > 0$.*

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.

Because the estimator is linear in y_i, its variance doesn't depend on f,

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

where the second equality holds under Assumptions 3.1 and 3.2, as we show in Supplemental Appendix B.3. The condition on the standard deviation in Equation (2) thus holds with $\gamma_s = -1/2$, and $S(k)$ given in the preceding display. Tables S1 and S2 in Supplemental Appendix give the constant $\int_{\mathcal{X}} k_q^*(u)^2\,du$ 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 Taylor class of order p,

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

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 (for simplicity, we focus on Hölder classes of integer order)

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

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 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 any bandwidth sequence h_n with $nh_n \to \infty$ and $0 < \liminf_n h_n(nM^2)^{1/(2p+1)} < \limsup_n h_n(nM^2)^{1/(2p+1)} < \infty$,*

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

and

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

Thus, the first part of (2) holds with $\gamma_b = p$ and $B(k) = \mathcal{B}_{p,q}(k)/p!$ where $\mathcal{B}_{p,q}(k) = \mathcal{B}_{p,q}^{\mathrm{Höl}}(k)$ for $\mathcal{F}_{\mathrm{Höl},p}(M)$, and $\mathcal{B}_{p,q}(k) = \mathcal{B}_{p,q}^{\mathrm{T}}(k)$ for $\mathcal{F}_{\mathrm{T},p}(M)$.

If, in addition, Assumption 3.2 holds, then Equation (4) 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 (13).

The theorem verifies the regularity conditions needed for the results in Section 2, and implies that $r = 2p/(2p+1)$ for $\mathcal{F}_{\mathrm{T},p}(M)$ and $\mathcal{F}_{\mathrm{Höl},p}(M)$. If $p = 2$, then we obtain $r = 4/5$. By Theorem 2.1(i), the optimal rate of convergence of a criterion R is $R(\hat{T}(h_R^*;k)) = O((n/M^{1/p})^{-p/(2p+1)})$. As we will see from the relative efficiency calculation below, the optimal order of the 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. Allowing for $q < p-1$ is not meaningful as in this case, the maximum bias is infinite.[6]

Under the Taylor class $\mathcal{F}_{\mathrm{T},p}(M)$, the least favorable (bias-maximizing) function is given by $f(x) = M/p! \cdot \mathrm{sign}(w_q^n(x))|x|^p$. In particular, if the weights are not all positive, it 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 $\mathcal{F}_{\mathrm{Höl},p}(M)$, the least favorable function takes the form of a pth order spline. See Supplemental Appendix B.3 for details.

These results 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

$$h^*_{\mathrm{RMSE}} = \left(\frac{1}{2pn}\frac{S(k)^2}{M^2 B(k)^2}\right)^{\frac{1}{2p+1}} = \left(\frac{\sigma^2(0)p!^2}{2pndM^2}\frac{\int_{\mathcal{X}} k_q^*(u)^2\,\mathrm{d}u}{\mathcal{B}_{p,q}(k)^2}\right)^{\frac{1}{2p+1}}, \tag{14}$$

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

3.1.1 Kernel efficiency

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

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

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 in Supplemental Appendix D.2. Table 2

[6] The smoothness classes $\mathcal{F}_{\mathrm{T},p}(M)$ and $\mathcal{F}_{\mathrm{Höl},p}(M)$ do not restrict derivatives of order $p-1$ and lower, so that, in order to achieve a finite worst-case bias, the estimator needs to be unbiased for polynomials of order $p-1$, which requires $q \geq p-1$.

compares 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 matches the kernel efficiency bounds in these papers. Table 2 shows 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. If the bandwidth picked is optimal for $\mathcal{F}_{\mathrm{T},q-1}(M)$, it 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 $\mathcal{F}_{\mathrm{Höl},p}(M)$ 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 (2018) for the boundary point problem. See Supplemental Appendix D.2 for details. 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 worst-case bias. Furthermore, the loss of efficiency from using a local polynomial estimator of order that's too high is smaller. Finally, local linear regression with a triangular kernel achieves high asymptotic efficiency under both $\mathcal{F}_{\mathrm{T},2}(M)$ and $\mathcal{F}_{\mathrm{Höl},2}(M)$, both at the interior and at a boundary, with efficiency at least 97%, giving a theoretical justification to this popular choice in empirical work.

3.1.2 Gains from imposing smoothness globally

The Taylor class $\mathcal{F}_{\mathrm{T},p}(M)$, only restricts the pth derivative locally to the point of interest, while the Hölder class $\mathcal{F}_{\mathrm{Höl},p}(M)$ restricts the pth derivative globally. How much can one tighten a confidence interval or reduce the RMSE due to this additional smoothness?

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

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

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 efficiency gain for common kernels. In general, the gains are greater for larger p, and greater at the boundary. For estimation at a boundary point with $p = 2$, for example, imposing global smoothness of f reduces CI length by about 13–15%, depending on the kernel, and about 10% if the optimal kernel is used.

3.2 Sharp regression discontinuity

We now consider the problem of inference in sharp regression discontinuity (RD) designs. Using data from the nonparametric regression model (10), the goal in sharp RD is to estimate the jump in the regression function f at a known cutoff, which we normalize to 0, so that $T(f) = \lim_{x\downarrow 0} f(x) - \lim_{x\uparrow 0} f(x)$. The cutoff 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:

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

Inference on $T(f)$ is then equivalent to inference on the difference between two regression functions evaluated at boundary points, and the results follow by a slight extension of the results for estimation at a boundary point in Section 3.1. We consider estimating $T(f)$ based on running a local linear regression on either side of the cutoff: the estimator $\hat{T}(h;k)$ is given by a difference between estimates from two local linear regressions with bandwidth h and kernel k 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. The estimator can be written as

$$\hat{T}(h;k) = \sum_{i=1}^{n} w^n(x_i;h,k)y_i, \qquad w^n(x;h,k) = w_+^n(x;h,k) - w_-^n(x;h,k), \tag{16}$$

with the weight w_+^n given by

$$w_+(x;h,k) = e_1' Q_{n,+}^{-1} m_1(x) k_+(x/h), \quad k_+(u) = k(u)\,\mathrm{I}\{u \geq 0\},$$

and $Q_{n,+} = \sum_{i=1}^n k_+(x_i/h) m_1(x_i) m_1(x_i)'$. The weights w_-^n, Gram matrix $Q_{n,-}$ and kernel k_- are defined similarly. Let $\sigma_+^2(x) = \sigma^2(x)\,\mathrm{I}\{x \geq 0\}$, and $\sigma_-^2(x) = \sigma^2(x)\,\mathrm{I}\{x < 0\}$.

In principle, one could allow the bandwidths for the two local linear regressions to be

different. We show in Supplemental Appendix D.1, however, that the loss in efficiency resulting from constraining the bandwidths to be the same is quite small unless the ratio of variances on either side of the cutoff, $\sigma_+^2(0)/\sigma_-^2(0)$, is quite large.

It follows from the results in Section 3.1 that if Assumptions 3.1 and 3.2 hold (with the requirement that $\sigma^2(x)$ is continuous 0 replaced by right- and left-continuity of $\sigma_+^2(x)$ and $\sigma_-^2(x)$), then the variance of the estimator doesn't depend on f and satisfies

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

with d defined in Assumption 3.1. Theorem 3.1 and arguments in Supplemental Appendix B.3 imply that the bias of $\hat{T}(h;k)$ is maximized at $f(x) = -Mx^2/2\cdot(\mathrm{I}\{x\geq 0\}-\mathrm{I}\{x<0\})$, so long as the kernel $k(\cdot)$ takes on nonnegative values. The worst-case bias therefore satisfies

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

It follows that for the RMSE, FLCI, and OCI criteria, Equation (4) holds with $\gamma_b = 2$, $\gamma_s = -1/2$, and $B(k)$ and $S(k)$ given in the displays above. Thus, the RMSE-optimal bandwidth is given by

$$h_{\mathrm{RMSE}}^* = \left(\frac{\int_0^\infty k_1^*(u)^2\,\mathrm{d}u}{(\int_0^\infty u^2k_1^*(u)\,\mathrm{d}u)^2}\cdot\frac{\sigma_+^2(0)+\sigma_-^2(0)}{4dnM^2}\right)^{1/5}. \tag{17}$$

The kernel efficiency results are analogous to those in Section 3.1.1.

3.3 Fuzzy regression discontinuity

In a fuzzy RD design, the treatment d_i is not entirely determined by whether the running variable x_i exceeds a cutoff. Instead, the cutoff induces a jump in the treatment probability. This fits into our framework if we let $f = (f_1, f_2)$ comprise two regression functions,

$$y_i = f_1(x_i) + u_{i1}, \qquad\qquad d_i = f_2(d_i) + u_{i2},$$

corresponding to the reduced-form regression of the outcome on the running variable, and the first-stage regression of the treatment on the running variable. Here $u_i = (u_{i1}, u_{i2})'$ is by definition mean zero, with covariance matrix $\operatorname{var}(u_i) = \Omega(x_i)$. The parameter of interest is given by the ratio of the reduced-form and first-stage sharp RD parameters,

$$T(f) = \frac{L_1(f)}{L_2(f)}, \qquad L(f) = \begin{pmatrix}\lim_{x\downarrow 0} f_1(x) - \lim_{x\uparrow 0} f_1(x)\\ \lim_{x\downarrow 0} f_2(x) - \lim_{x\uparrow 0} f_2(x)\end{pmatrix},$$

so that we can write $T(f) = \phi(L(f))$, with $\phi(L) = L_1/L_2$. If the regression functions of the potential outcomes and potential treatments are continuous at zero, and a monotonicity condition holds, then $T(f)$ measures the average treatment effect for individuals with $x_i = 0$ who are compliers (see Hahn et al., 2001). We assume that f lies in the class $\mathcal{F}_{\text{FRD}}(M_1, M_2) = \mathcal{F}_{\text{RD}}(M_1) \times \mathcal{F}_{\text{RD}}(M_2)$, so that both the reduced-form and the first-stage regression functions are assumed to have a bounded second derivative on either side of the cutoff,[7] and consider estimating $T(f)$ by its sample analog, replacing L_1 and L_2 with sharp RD local linear estimates, which are for simplicity assumed to be based on the same bandwidth, $\hat{T}(h;k) = \hat{L}_1(h;k)/\hat{L}_2(h;k)$, where $\hat{L}(h;k) = \sum_i w^n(x_i;h,k)\binom{y_i}{d_i}$, with the weight w^n given in (16).

Since the estimator is non-linear, to ensure that (4) holds, it will be necessary to consider a sequence of parameter spaces $\mathcal{F}_{\text{FRD},n}(M_1, M_2)$ localized around a particular value L^* with a non-zero jump in the first-stage regression $L_2^* \neq 0$. This allows us to apply a version of the delta method to $\hat{L}(h;k)$. We defer details to Supplemental Appendix B.4, where we show that under Assumption 3.1 and a version of Assumption 3.2, the distribution of $\hat{T}(h;k) - T(f)$ can in large samples be approximated by a normal distribution with variance

$$\operatorname{avar}(\hat{T}(h;k)) = \frac{S(k)^2}{nh} = \sum_{i=1}^{n} \frac{\varsigma^2(x_i;T(f))}{L_2(f)^2} \tilde{w}^n(x_i;h,k)^2(1+o(1)),$$

and mean bounded by

$$\overline{\text{abias}}(\hat{T}(h;k)) = M_1 h^2 B(k) = -\frac{M_1 + |T(f)|M_2}{2|L_2(f)|} \sum_{i=1}^{n} \tilde{w}^n(x_i;h,k)x_i^2(1+o(1)),$$

where $\tilde{w}^n(x_i;h,k) = w_+^n(x_i) + w_-^n(x_i)$, $\varsigma^2(x_i;T) = (1,-T)\Omega(x_i)(1,-T)'$,

$$B(k) = -\frac{\int_0^\infty u^2 k_1^*(u)\,\mathrm{d}u(1+|T(f)|M_2/M_1)}{|L_2(f)|},\ S(k)^2 = \frac{\int_0^\infty k_1^*(u)^2\,\mathrm{d}u}{d}\frac{\varsigma_+^2(0;T(f)) + \varsigma_-^2(0;T(f))}{L_2(f)^2},$$

$\varsigma_+^2(0;T) = \lim_{x\downarrow 0}\varsigma^2(x;T)$, and $\varsigma_-^2(0;T) = \lim_{x\uparrow 0}\varsigma^2(x;T)$.

It then follows that for the FLCI, OCI, and a truncated version of the RMSE criterion, Equation (4) holds with $M = M_1$, $\gamma_b = 2$, $\gamma_s = -1/2$, and $B(k)$ and $S(k)$ given in the preceding display. The RMSE-optimal bandwidth is therefore given by

$$h_{\text{RMSE}}^* = \left(\frac{\int_0^\infty k_1^*(u)^2\,\mathrm{d}u}{(\int_0^\infty u^2 k_1^*(u)\,\mathrm{d}u)^2} \cdot \frac{\varsigma^2(T(f))}{4dn(M_1 + |T(f)|M_2)}\right)^{1/5}. \tag{18}$$

[7] While we allow the bounds M_1 and M_2 to change with sample size, we assume that their ratio M_1/M_2 is fixed for simplicity.

Since $S(k)$ and $B(k)$ depend on the kernel k through the same quantities as for inference at a boundary point, the kernel efficiency results are analogous to those in Section 3.1.1.

Because the optimal bandwidth depends on $T(f)$, implementing a feasible version of it requires replacing it with an initial estimate. An alternative approach to the construction of two-sided CIs for $T(f)$ that doesn't require localization or the use of initial estimates is an Anderson and Rubin (1949) style construction studied by Noack and Rothe (2019). In particular, Noack and Rothe (2019) propose constructing, for each T_0, an auxiliary CI for the jump in the mean of $y_i - d_i T_0$ at the cutoff, using an approach similar to that in Section 3.2. The CI for $T(f)$ is then constructed by collecting all T_0's for which the auxiliary CI contains zero. This approach also has the additional advantage that it can allow for weak identification while it yields asymptotically equivalent CIs under strong identification.[8] See Noack and Rothe (2019) for a more detailed discussion.

3.4 Practical implementation

We now discuss some practical issues that arise when implementing our CIs for inference at a point, and in sharp and fuzzy RD studied in the previous subsections. To focus the discussion, we consider smoothness classes $\mathcal{F}_{\text{Höl},2}(M)$, $\mathcal{F}_{\text{RD}}(M)$, and $\mathcal{F}_{\text{FRD}}(M_1, M_2)$ that constrain the second derivative globally, so that, in the discussion below, $p = 2$. This choice implies optimality of estimators based on local linear regression, which is the most popular method in practice. In this case, both the Epanechnikov and the triangular kernel are nearly optimal.

3.4.1 Choice of M

Appropriate choice of the smoothness constant is key to implementing our method. Since the smoothness classes we consider are convex, the results of Low (1997), Cai and Low (2004) and Armstrong and Kolesár (2018a) imply that, to maintain honesty over the whole function class, a researcher must choose M a priori, rather than attempting to use a data-driven method.[9] We therefore recommend that, whenever possible, problem-specific knowledge be used to decide what choice of M is reasonable a priori, and that one consider a range of plausible values by way of sensitivity analysis.[10]

[8] Because we require that the sequence of parameter spaces $\mathcal{F}_{\text{FRD},n}(M_1, M_2)$ be localized around a value of L^* with $L_2^* \neq 0$, we rule out sequences in which the jump in the first-stage regression is arbitrarily close to zero (the term "weak identification" refers to such sequences). As a result, the CI we propose, unlike the CI proposed by Noack and Rothe (2019), is not honest over the original parameter space $\mathcal{F}_{\text{FRD}}(M_1, M_2)$.

[9] These negative results contrast with more positive results for estimation. See Lepski (1990), who proposes a data-driven method that automates the choice of both p and M.

[10] As is well-known, if the final bandwidth choice is influenced by such sensitivity analysis, the resulting CI may undercover, even if the estimator is unbiased. In this case, one can combine our method with the bandwidth snooping adjustment of Armstrong and Kolesár (2018b).

If one imposes additional restrictions on f that that make the parameter space for f non-convex, a data-driven method for choosing M may be feasible.[11] In Supplemental Appendix E, we consider a restriction which relates M to a global polynomial approximation to the regression function. In particular, the restriction formalizes the notion that the second derivative in a neighborhood of zero is bounded by the maximum second derivative of a global $\tilde{p}$th order global polynomial approximation. Heuristically, such restriction will hold if the local smoothness of f is no smaller than its smoothness at large scales.

This restriction allows us to calibrate M based on the following rule of thumb. For inference at a point, let $\breve{f}(x)$ be an estimate of f based on a global polynomial regression of order $\tilde{p}$, and let $[x_{\min}, x_{\max}]$ denote the support of x_i. Put $\hat{M}_{\text{ROT}} = \sup_{x\in[x_{\min},x_{\max}]}|\breve{f}^{(p)}(x)|$. This rule of thumb is similar to the suggestion of Fan and Gijbels (1996, Chapter 4.2), with the important distinction that their rule of thumb was designed to estimate the pointwise-in-f optimal bandwidth. We discuss the difference between this bandwidth and h^*_{RMSE} in Section 4. In sharp RD, the rule of thumb is analogous, except we define $\breve{f}^{(p)}(x)$ to be the global polynomial estimate of order $\tilde{p}$ in which the intercept and all coefficients are allowed to be different on either side of the discontinuity (that is, as regressors, we use $1, x_i, \ldots, x_i^{\tilde{p}}$, and their interactions with the indicator $\mathrm{I}\{x_i \geq 0\}$). For fuzzy RD, we use an analogous approach to separately calibrate M_1 and M_2 based on the reduced-form and first-stage regressions.

As a default choice, we set $\tilde{p} = p + 2 = 4$. In Supplemental Appendix E, we give a formal analysis of this rule, showing that the resulting CIs are honest and nearly optimal (over a regularity class that imposes the additional restriction f discussed above). In contrast, we expect that calibrating M based on local smoothness estimates may be difficult to justify, since estimating a local derivative of f is a harder problem than the initial problem of estimating its value at a point. We investigate the finite-sample performance of FLCIs based on $\hat{M}_{\text{ROT}}$ in a Monte Carlo exercise in Section 5.

3.4.2 Computation of h^*_{rmse}

Given a choice of M, one can compute a feasible version $\hat{h}^*_{\text{RMSE}}$ of h^*_{RMSE} by plugging this choice into the expressions (14), (17), and (18), along with consistent estimates of d, and of the variance at 0 (for fuzzy RD, one also needs a preliminary estimate of $T(f)$).

In the simulation exercise and empirical application below, we use an alternative approach based on directly minimizing the finite-sample RMSE over the bandwidth h. To describe it, let $\tilde{w}^n(x_i; h, k)$ denote the weights $w^n_1(x_i; h, k)$ given in (11) if the parameter of interest is the

[11] An alternative to restricting the parameter space is to change the notion of coverage. For example, in the context of constructing confidence bands for a regression function $f(x)$, Hall and Horowitz (2013) propose bands that have an average coverage property in that the bands achieve coverage of $f(x)$ for a random subset of values of x. This subset may vary with the unknown regression function and the realized sample.

conditional mean at a point, and let $\tilde{w}^n(x_i;h,k) = w^n_+(x_i) + w^n_-(x_i)$ if the parameter of interest is the sharp or fuzzy RD parameter. For inference at a point, or for sharp RD, the finite-sample RMSE takes the form

$$\mathrm{RMSE}(h)^2 = \frac{M^2}{4}\left(\sum_{i=1}^n \tilde{w}^n(x_i;h,k)x_i^2\right)^2 + \sum_{i=1}^n \tilde{w}^n(x_i;h,k)\sigma^2(x_i), \tag{19}$$

Since $\sigma^2(x_i)$ is typically unknown, one needs to replace it by an estimate. For inference at a point, the simplest choice is to use some estimate $\hat{\sigma}^2(x_i) = \hat{\sigma}^2$ that assumes homoskedasticity of the variance function. For sharp RD, one can use the estimate $\hat{\sigma}^2(x_i) = \hat{\sigma}^2_+(0)\,\mathrm{I}\{x \geq 0\} + \hat{\sigma}^2_-(0)\,\mathrm{I}\{x < 0\}$, where $\hat{\sigma}^2_+(0)$ and $\hat{\sigma}^2_-(0)$ are some preliminary variance estimates based on observations above and below the cutoff. We denote the resulting bandwidth by $\hat{h}^*_{\mathrm{RMSE},\tilde{M}}$, where $\tilde{M}$ denotes the chosen smoothness constant. This method was considered previously in Armstrong and Kolesár (2018a). It has the advantage that it avoids having to estimate d.

Since the estimate in fuzzy RD is non-linear, its moments, and hence the finite-sample RMSE do not exist. However, one can still employ an analogous approach by replacing $B(k)$ and $S(k)$ with finite-sample analogs in the expression for the asymptotic RMSE. As the asymptotic bias and the asymptotic standard deviation both scale with $L_2(f)$, this scaling doesn't affect the optimum, we can equivalently minimize

$$\mathrm{ARMSE}(h;M_1,M_2)^2 = \frac{(M_1 + |T(f)|M_2)^2}{4}\left(\sum_{i=1}^n \tilde{w}^n(x_i;h,k)\right)^2 + \sum_{i=1}^n w^n_q(x_i;h;k)^2\varsigma^2(x_i;T(f)),$$

with $\varsigma^2(x;T) = (1,-T)\Omega(x)(1,-T)'$. Since $\Omega(x_i)$ is unknown, one can again replace it with $\hat{\Omega}^2(x_i) = \hat{\Omega}^2_+(0)\,\mathrm{I}\{x \geq 0\} + \hat{\Omega}^2_-(0)\,\mathrm{I}\{x < 0\}$, where $\hat{\Omega}^2_+(0)$ and $\hat{\Omega}^2_-(0)$ are some preliminary variance estimates for observations above and below the cutoff. As a preliminary estimate of $T(f)$, one can take the estimate $\hat{T}(\hat{h}_0;k)$, where $\hat{h}_0$ minimizes the above expression at $T(f) = 0$. One can also use $\hat{h}_0$ directly as a simple bandwidth selector, which, while not RMSE optimal, has the advantage that it doesn't depend on the choice of M_2.

3.4.3 Construction of FLCIs

Given an estimate $\hat{h}^*_{\mathrm{RMSE}}$ of h^*_{RMSE}, such as the estimate $\hat{h}^*_{\mathrm{RMSE},\tilde{M}}$ discussed above, an honest FLCI can be constructed as

$$\hat{T}(\hat{h}^*_{\mathrm{RMSE}};k) \pm \mathrm{cv}_{1-\alpha}(t)\cdot\widehat{\mathrm{se}}(\hat{h}^*_{\mathrm{RMSE}};k), \tag{20}$$

where t is an estimate of the bias-sd ratio, and $\widehat{\mathrm{se}}(\hat{h}^*_{\mathrm{RMSE}};k)$ is an estimate of the standard error. For the standard error, many choices are available in the literature. For inference

at a point and sharp RD, the estimator $\hat{T}(\hat{h}^*_{\text{RMSE}}; k)$ is a weighted least squares estimator, and one can directly estimate its finite-sample conditional variance by the nearest neighbor variance estimator considered in Abadie and Imbens (2006) and Abadie et al. (2014). Given a bandwidth h, the estimator takes the form

$$\widehat{\text{se}}(h,k)^2 = \sum_{i=1}^{n} \tilde{w}^n(x_i; h, k)^2 \hat{\sigma}^2(x_i), \qquad \hat{\sigma}^2(x_i) = \frac{J}{J+1}\left(y_i - \frac{1}{J}\sum_{j=1}^{J} y_{j(i)}\right)^2, \tag{21}$$

for some fixed (small) $J \geq 1$, where $j(i)$ denotes the jth closest observation to i (for sharp RD $j(i)$ is only taken among units with the same sign of the running variable.). In contrast, the usual Eicker-Huber-White estimator sets $\hat{\sigma}^2(x_i) = \hat{u}_i^2$, where $\hat{u}_i$ is the regression residual, and it can be shown that this estimator will generally overestimate the conditional variance. For t, one can either use the asymptotic bias-sd ratio $t = 1/2$, or else an estimate of the finite-sample bias-sd ratio $t = -M \sum_{i=1}^{n} \tilde{w}^n(x_i; \hat{h}^*_{\text{RMSE}}, k) x_i^2 / 2\widehat{\text{se}}(\hat{h}^*_{\text{RMSE}}, k)$. We use the latter approach in the Monte Carlo and empirical application below. While both approaches are asymptotically equivalent when x_i is continuous, the latter approach has the advantage that it remains valid even when the covariates are discrete.[12]

For fuzzy RD, one can use an analogous approach to estimate the standard error as

$$\widehat{\text{se}}(h,k)^2 = \frac{1}{\hat{L}_2(h;k)^2} \sum_{i=1}^{n} \tilde{w}^n(x_i; h, k)^2 \hat{\varsigma}^2(x_i, \hat{T}(h;k)),$$

where $\hat{\varsigma}^2(x_i; T) = \frac{J}{J+1}(1, -T)(z_i - \frac{1}{J}\sum_{j=1}^{J} z_{j(i)})(z_i - \frac{1}{J}\sum_{j=1}^{J} z_{j(i)})'(1, -T)'$, $z_i = (y_i, d_i)'$, and $j(i)$ denotes that jth closest observation with the same sign of the running variable. For t, one can use $t = 1/2$, or else the finite-sample analog of the asymptotic bias-sd ratio,[13] $t = -(\tilde{M}_1 + |\hat{T}|\tilde{M}_2) \cdot \sum_{i=1}^{n} \tilde{w}^n(x_i; \hat{h}^*_{\text{RMSE}}, k) x_i^2 / 2\sqrt{\sum_{i=1}^{n} \hat{\varsigma}^2(x_i; \hat{T}) \tilde{w}^n(x_i; \hat{h}^*_{\text{RMSE}}, k)^2}$.

4 Comparison with other approaches

In this section, we compare our approach to inference about the parameter $T(f)$ to three other approaches to inference. To make the comparison concrete, we make the comparison in the context of inference about a nonparametric regression function at a point, as in Section 3.1. The first approach, which we term "conventional," ignores the potential bias of the estimator and constructs the CI as $\hat{T}_q(h,k) \pm z_{1-\alpha/2}\widehat{\text{se}}(h;k)$. The bandwidth h is typically chosen to

[12] See Armstrong and Kolesár (2018a), Kolesár and Rothe (2018) and Imbens and Wager (2019) for a more thorough discussion of the case with discrete covariates.

[13] For inference based on $\hat{T}(\hat{h}_0; k)$, it is necessary to use the finite-sample analog of the bias-sd ratio, since the bandwidth $\hat{h}_0$ is not RMSE optimal.

minimize the asymptotic mean squared error of $\hat{T}_q(h;k)$ under pointwise-in-f (or "pointwise", for short) asymptotics, as opposed to the uniform-in-f asymptotics that we consider. We refer to this bandwidth as h^*_{PT}. In undersmoothing, one chooses a sequence of smaller bandwidths, so that in large samples, the bias of the estimator is dominated by its standard error. Finally, in bias correction, one re-centers the conventional CI by subtracting an estimate of the leading bias term from $\hat{T}_q(h;k)$. In Section 4.1, we discuss the distinction between h^*_{PT} and h^*_{RMSE}. In Section 4.2, we compare the coverage and length properties of these CIs to the fixed-length CI (FLCI) based on $\hat{T}_q(h^*_{\text{RMSE}};k)$. Implementing any of these CIs in practice requires tuning parameter choices. For clarity of comparison, we keep implementation issues separate, and focus in this section on a theoretical comparison, assuming any tuning parameters are known. The Monte Carlo exercise in Section 5 below considers their finite-sample performance when the tuning parameters need to be chosen.

4.1 RMSE and pointwise optimal bandwidth

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))

$$h^*_{\text{PT}} = \left(\frac{\sigma^2(0) p!^2}{2pdnf^{(p)}(0)^2} \frac{\int_{\mathcal{X}} k^*_q(u)^2 \, \mathrm{d}u}{(\int_{\mathcal{X}} t^p k^*_q(t) \, \mathrm{d}t)^2} \right)^{\frac{1}{2p+1}}. \tag{22}$$

Comparing this expression with the RMSE optimal bandwidth under $\mathcal{F}_{\text{T},p}(M)$ and $\mathcal{F}_{\text{Höl},p}(M)$ given in (14), we see that the pointwise optimal bandwidth replaces M with the pth derivative at zero, $f^{(p)}(0)$, and it replaces $\mathcal{B}_{p,q}(k)$ with $\int_{\mathcal{X}} t^p k^*_q(t) \, \mathrm{d}t$.

Note that $\mathcal{B}_{p,q}(k) \geq |\int_{\mathcal{X}} t^p k^*_q(t) \, \mathrm{d}t|$ (this can be seen by noting that the right-hand side corresponds to the bias at the function $f(x) = \pm x^p/p!$, while the left-hand side is the supremum of the bias over functions with pth derivative bounded by 1). Thus, assuming that $f^{(p)}(0) \leq M$ (this holds by definition for any $f \in \mathcal{F}$ when $\mathcal{F} = \mathcal{F}_{\text{Höl},p}(M)$), we will have $h^*_{\text{PT}}/h^*_{\text{RMSE}} \geq 1$. The ratio $h^*_{\text{PT}}/h^*_{\text{RMSE}}$ can be arbitrarily large if M exceeds $f^{(p)}(0)$ by a large amount. It then follows from Theorem 2.1, that the RMSE efficiency of the estimator $\hat{T}_{p-1}(h^*_{\text{PT}};k)$ relative to $\hat{T}_{p-1}(h^*_{\text{RMSE}};k)$ may be arbitrarily low.

The bandwidth h^*_{PT} is intended to optimize RMSE at the function f itself, so one may argue that evaluating the resulting minimax RMSE is an unfair comparison. However, the mean squared error performance of $\hat{T}_{p-1}(h^*_{\text{PT}};k)$ at a given function f can be bad even if the same function f is used to calculate h^*_{PT}. For example, consider the function $f(x) = x^{p+1}$ if p is odd, or $f(x) = x^{p+2}$ if p is even. This is a smooth function with all derivatives bounded on the support of x_i. Since $f^{(p)}(0) = 0$, h^*_{PT} is infinite, and the resulting estimator is a global pth

order polynomial least squares estimator. Its RMSE will be poor, since the estimator is not even consistent.[14]

To address this problem, plug-in bandwidths that estimate h^*_{PT} include tuning parameters to prevent them from approaching infinity. The RMSE of the resulting estimator at such functions is then determined almost entirely by these tuning parameters. Furthermore, if one uses such a bandwidth as an input to an undersmoothed or bias-corrected CI, the coverage will be determined by these tuning parameters, and can be arbitrarily bad if the tuning parameters allow the bandwidth to be large. Indeed, we find in our Monte Carlo analysis in Section 5 that plug-in estimates of h^*_{PT} used in practice can lead to very poor coverage even when used as a starting point for a bias-corrected or undersmoothed estimator.

4.2 Efficiency and coverage comparison

Let us now consider the efficiency and coverage properties of conventional, undersmoothed, and bias-corrected CIs relative to the FLCI based on $\hat{T}_{p-1}(h^*_{\text{RMSE}}, k)$. To keep the comparison meaningful, and avoid the issues discussed in the previous subsection, we assume these CIs are also based on h^*_{RMSE}, rather than h^*_{PT} (in case of undersmoothing, we assume that the bandwidth is undersmoothed relative to h^*_{RMSE}). Suppose that the smoothness class is either $\mathcal{F}_{\text{T},p}(M)$ and $\mathcal{F}_{\text{Höl},p}(M)$ and denote it by $\mathcal{F}_p(M)$. For concreteness, let $p = 2$, and $q = 1$.

Consider first conventional CIs, given by $\hat{T}_1(h; k) \pm z_{1-\alpha/2}\widehat{\text{se}}(h; k)$. If the bandwidth h equals h^*_{RMSE}, then these CIs are shorter than the 95% FLCIs by a factor of $z_{0.975}/\,\text{cv}_{0.95}(1/2) = 0.90$. Consequently, their coverage is 92.1% rather than the nominal 95% coverage. At the RMSE-optimal bandwidth, the bias-sd ratio equals 1/2, so disregarding the bias doesn't result in severe undercoverage. If one uses a larger bandwidth, however, the 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 bias-sd ratio equals $1/2 \cdot (1.5)^{(5/2)}$, the coverage is only 71.9%.

Second, consider undersmoothing. This amounts to choosing a bandwidth sequence h_n such that $h_n/h^*_{\text{RMSE}} \to 0$, so that for any fixed M, the bias-sd ratio $t_n = h_n^{\gamma_b - \gamma_s} MB(k)/(n^{-1/2}S(k))$ approaches zero, and the CI $\hat{T}(h^n; k) \pm \text{cv}_{1-\alpha}(0)\widehat{\text{se}}(h_n; k) = \hat{T}(h^n; k) \pm z_{1-\alpha/2}\widehat{\text{se}}(h_n; k)$ will consequently have proper coverage in large samples. However, the CIs shrink at a slower rate than $n^{r/2} = n^{4/5}$, and thus the asymptotic efficiency of the undersmoothed CI relative to the optimal FLCI is zero.

On the other hand, an apparent advantage of the undersmoothed CI is that it appears to avoid specifying the smoothness constant M. However, a more accurate description of undersmoothing is that the bandwidth sequence h_n implicitly chooses a sequence of smoothness

[14] To ensure consistency and finiteness of h^*_{PT}, it is standard to assume that $f^{(p)} \neq 0$. However, the RMSE can still be arbitrarily poor whenever the pth derivative is locally small, but non-zero, and large globally, such as when $f(x) = x^{p+1} + \eta x^p$ for p odd and $f(x) = x^{p+2} + \eta x^p$ if p is even, provided η is sufficiently small.

constants $M_n \to \infty$ such that coverage is controlled under the sequence of parameter spaces $\mathcal{F}_p(M_n)$. We can improve on the coverage and length of the resulting CI by making this sequence explicit and computing an optimal (or near-optimal) FLCI for $\mathcal{F}_p(M_n)$.

To this end, given a sequence h_n, a better approximation to the finite-sample coverage of the CI $\hat{T}(h_n;k) \pm z_{1-\alpha/2}\widehat{\text{se}}(h_n;k)$ over the parameter space $\mathcal{F}_p(M)$ is $P_{Z\sim N(0,1)}(|Z+t_n(M)| \geq z_{1-\alpha/2})$ where $t_n(M) = h_n^{\gamma_b-\gamma_s}MB(k)/(n^{-1/2}S(k))$ is the bias-sd ratio for the given choice of M. This approximation is exact in idealized settings, such as the white noise model in Supplemental Appendix B. For a given level of undercoverage $\eta = \eta_n$, one can then compute M_n as the greatest value of M such that this approximation to the coverage is at least $1-\alpha-\eta$. In order to trust the undersmoothed CI, one must be convinced of the plausibility of the assumption $f \in \mathcal{F}_p(M_n)$: otherwise the coverage will be worse than $1-\alpha-\eta$. This suggests that, in the interest of transparency, one should make this smoothness constant explicit by reporting M_n along with the undersmoothed CI. However, once the sequence M_n is made explicit, a more efficient approach is to simply report an optimal or near-optimal CI for this sequence, either at the coverage level $1-\alpha-\eta$ (in which case the CI will be strictly smaller than the undersmoothed CI while maintaining the same coverage) or at level $1-\alpha$ (in which case the CI will have better finite-sample coverage and may also be shorter than the undersmoothed CI).

Finally, let us consider bias correction. It is known that re-centering conventional CIs by an estimate of the leading bias term often leads to poor coverage (Hall, 1992). In an important paper, Calonico et al. (2014, CCT hereafter) show that the coverage properties of this bias-corrected CI are much better if one adjusts the standard error estimate to account for the variability of the bias estimate, which they call robust bias correction (RBC). 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, the bias-corrected local linear estimator coincides with a local quadratic estimator. As a result, the RBC procedure in this case 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.

To ensure that the bias be estimable, the theory of bias correction requires that the conditional mean function is sufficiently smooth, which requires $q < p-1$ (thus, assuming that f is sufficiently smooth to ensure that the bias of $\hat{T}_1(h;k)$ can be estimated implies that the polynomial order $q=1$ of the original estimator is not optimal). Suppose, therefore, that the smoothness class is given by $\mathcal{F}_3(M)$ (with $q=1$, and $h = h^*_{\text{RMSE}}$ still chosen to be MSE optimal for $\mathcal{F}_2(M)$). In this case the RBC interval can be considered an undersmoothed CI based on a second order local polynomial estimator. Following the discussion of undersmoothed CIs above, the limiting coverage is $1-\alpha$ when M is fixed (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 RBC CI shrinks at a slower rate than the optimal CI.

It is also interesting to consider the case when the order $q = 1$ of the local polynomial of the estimator $\hat{T}_1(h^*_{\text{RMSE}}; k)$ is optimal under the maintained smoothness assumption, so that the smoothness class is given by $\mathcal{F}_2(M)$. In this case, the smoothness of the conditional mean function is too low for the bias to be estimable: the bias of the bias-corrected estimator will be of the same order as the bias of the original estimator. Consequently, the estimator will remain asymptotically biased, even after the bias correction. In particular, bias-sd ratio of the estimator is given by

$$t_{\text{RBC}} = (h^*_{\text{RMSE}})^{5/2} \frac{M\mathcal{B}_{2,2}(k)/2}{\sigma(0)(\int k_2^*(u)^2 \, \mathrm{d}u/dn)^{1/2}} = \frac{1}{2}\frac{\mathcal{B}_{2,2}(k)}{\mathcal{B}_{2,1}(k)} \left(\frac{\int_{\mathcal{X}} k_1^*(u)^2 \, \mathrm{d}u}{\int_{\mathcal{X}} k_2^*(u)^2 \, \mathrm{d}u} \right)^{1/2}. \tag{23}$$

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

$$\frac{z_{1-\alpha/2} \left(\int_{\mathcal{X}} k_2^*(u)^2 \, \mathrm{d}u \right)^{1/2}}{\text{cv}_{1-\alpha}(1/2) \left(\int_{\mathcal{X}} k_1^*(u)^2 \, \mathrm{d}u \right)^{1/2}} (1 + o(1)). \tag{24}$$

The resulting coverage and relative length is given in Table 5. One can see that although the undercoverage is very mild, (since t_{RBC} is quite low in all cases), the intervals are about 30% longer than the FLCIs around the RMSE bandwidth.

Under the class $\mathcal{F}_{\text{Höl},2}(M)$, the RBC 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 (23) 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 5. 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. Thus, depending on the smoothness class, the 95% RBC 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.

Our asymptotic efficiency comparisons focus on minimizing length among CIs with coverage at least $1-\alpha$ for all $f \in \mathcal{F}$, which follows the usual definition of coverage. One may also consider a criterion that also penalizes CIs that cover "too much", by placing an upper bound $1 - \underline{\alpha}$ on coverage. For the CIs considered in this paper, the maximum coverage occurs when the bias is zero, and is given by $P_{Z\sim N(0,1)}(|Z| \leq \text{cv}_{1-\alpha}(t)) = 1 - 2\Phi(-\text{cv}_{1-\alpha}(t))$ where t is the asymptotic bias-sd ratio. In particular, when $\mathcal{F} = \mathcal{F}_{\text{T},2}(M)$ or $\mathcal{F} = \mathcal{F}_{\text{Höl},2}(M)$ and the RMSE optimal bandwidth is used, the maximum coverage of a FLCI with 95% (minimum) coverage is $1 - 2\Phi(-2.18) = .971$. If one wants the maximum coverage to be smaller, then undersmoothing

(or subtracting an estimate of the bias) will be necessary, and Edgeworth expansions may be needed to deal with higher order approximation terms if one wants $\alpha - \underline{\alpha} \to 0$ quickly enough with the sample size (see Calonico et al., 2019). The resulting CIs will be much larger than the CIs proposed in the present paper, which do not penalize "overcoverage."

5 Monte Carlo

To study the finite-sample performance of the FLCI that we propose, and compare its performance to other approaches, this section conducts a Monte Carlo analysis of the conditional mean estimation problem considered in Section 3.1.

We consider Monte Carlo designs with conditional mean functions

$$
\begin{aligned}
f_1(x) &= \frac{M}{2}(x^2 - 2\mathsf{s}(|x| - 0.25)),\\
f_2(x) &= \frac{M}{2}(x^2 - 2\mathsf{s}(|x| - 0.2)^2 + 2\mathsf{s}(|x| - 0.5) - 2\mathsf{s}(|x| - 0.65)),\\
f_3(x) &= \frac{M}{2}((x+1)^2 - 2\mathsf{s}(x + 0.2) + 2\mathsf{s}(x - 0.2) - 2\mathsf{s}(x - 0.4) + 2\mathsf{s}(x - 0.7) - 0.92),
\end{aligned}
$$

where $\mathsf{s}(x) = (x)_+^2 = \max\{x, 0\}^2$ is the square of the plus function, and $M \in \{2, 6\}$, giving a total of 6 designs. In all cases, x_i is drawn from a uniform distribution with support $[-1, 1]$ (so that the design is random), $u_i \sim N(0, 1/4)$, and the sample size is $n = 500$. Figure 5 plots these designs. The regression function for each design lies in $\mathcal{F}_{\text{Höl},2}(M)$ for the corresponding M. To ensure that our results, discussed below, are not sensitive to the choice of the error distribution or the distribution for the running variable, in Supplemental Appendix F, we also consider designs with x_i drawn from a beta distribution, designs with log-normal and heteroskedastic errors, and designs with different error variance. Finally, we also show in the appendix that the results remain effectively the same when the function $\mathbf{s}(\cdot)$ is replaced by a smooth approximating function.[15]

For each design, we implement the optimal FLCI centered at a local linear estimate with a triangular kernel and MSE optimal bandwidth, as described in Section 3.4, for each choice of $M \in \{2, 6\}$, and with M calibrated using the rule-of-thumb (ROT) described in Section 3.4. The implementations with $M \in \{2, 6\}$ allow us to gauge the effect of using an appropriately calibrated M, compared to a choice of M that is either too conservative or too liberal by a factor of 3. The ROT calibration chooses M automatically, but requires additional conditions

[15]The RBC method considered below assumes that the conditional mean function be at least three times continuously differentiable in the neighborhood of 0. Since the functions f_1, f_2 and f_3 are not globally three times continuously differentiable, depending on the neighborhood definition, this assumption is arguably violated. The results in the appendix are nearly identical to those reported here, implying that the performance of the RBC method is not driven by this lack of smoothness.

in order to have correct coverage (see Section 3.4).

In addition to these FLCIs, we consider six other CIs (Supplemental Appendix F considers two further methods). The first five are different implementations of the robust bias-corrected (RBC) CIs proposed by CCT (discussed in Section 4). Implementing these CIs requires two bandwidth choices: a bandwidth for the local linear estimator, and a pilot bandwidth that is used to construct an estimate of its bias. The first two CIs use bandwidth choices justified by pointwise-in-f asymptotics. The first CI uses a plug-in estimate of h^*_{PT} defined in (22), as implemented by Calonico et al. (2018), and an analogous estimate for the pilot bandwidth[16]. The second CI, also implemented by Calonico et al. (2018), uses bandwidth estimates for both bandwidths that optimize the pointwise asymptotic coverage error (CE) among CIs that use usual $z_{1-\alpha/2}$ critical value. This CI can be considered a particular form of undersmoothing. For the next three CIs, we consider bandwidths justified by uniform-in-f asymptotics. For the third and fourth CIs, we set both the main and the pilot bandwidth to h^*_{RMSE} with $M = 2$, and $M = 6$, respectively. For the fifth CI, we set both bandwidths to $\hat{h}^*_{\text{RMSE},\hat{M}_{\text{ROT}}}$. Finally, we consider a conventional CI centered at a plug-in bandwidth estimate of h^*_{PT}, using the rule-of-thumb estimator of Fan and Gijbels (1996, Chapter 4.2). All CIs are computed at the nominal 95% coverage level.

Table 6 reports the results. The FLCIs perform well when the correct M is used. As expected, they suffer from undercoverage if M is chosen too small, or suboptimal length when M is chosen too large. The ROT choice of M appears to do a reasonable job of having good coverage and length in these designs without requiring knowledge of the true smoothness constant. However, as discussed in Section 3.4, this ROT choice imposes additional restrictions on the parameter space, so one must take care in extrapolating these results to other designs.

As predicted by the theory in Section 4, the RBC CIs also have good coverage when implemented using the h^*_{RMSE} bandwidth, and they are less sensitive to the choice of M than the corresponding FLCIs, at the expense of being on average about 25% longer. RBC CIs with bandwidth given by $\hat{h}^*_{\text{RMSE},\hat{M}_{\text{ROT}}}$ also achieve good coverage, but they are again about 25% longer than the corresponding FLCIs.

The CIs based on bandwidths justified by pointwise-in-f asymptotics (rows 1, 2, and 6 for each design in the table) all have very poor coverage for at least one of the designs. Our analysis in Sections 4 suggests that this is due to the tuning parameter choices required by these bandwidths. Indeed, looking at the average of the bandwidth over the Monte Carlo draws (also reported in Table 6), it can be seen that the bandwidths tend to be much larger than those that estimate h^*_{RMSE}. This is even the case for the CE bandwidth, which is intended to minimize coverage errors.

Overall, the Monte Carlo analysis suggests that default approaches to nonparametric CI

[16]This is the default in version 0.1.1 of their accompanying software package `nprobust`

construction (bias-correction or undersmoothing relative to plug-in bandwidths) can lead to severe undercoverage when implemented using bandwidths justified by pointwise-in-f asymptotics. Bias-corrected CIs such as the one proposed by CCT can have good coverage if one starts from the minimax RMSE bandwidth, although they will be wider than FLCIs proposed in this paper.

6 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. See Ludwig and Miller (2007) for a detailed description of this variable. Relative to the dataset used in Ludwig and Miller (2007), we remove one duplicate entry and one outlier, which after discarding counties with partially missing data leaves us with 3,103 observations, with 294 of them above the poverty cutoff.

Figure 4 plots the data (to reduce the noise in the outcome variable, we plot bin averages of size 25). To estimate the discontinuity in mortality rates, Ludwig and Miller (2007) use a uniform kernel[17] and consider bandwidths equal to 9, 18, and 36. This yields point estimates equal to -1.90, -1.20 and -1.11 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 (with $J = 3$) are 1.04, 0.70, and 0.52.

These bandwidth choices are optimal in the sense that they minimize the RMSE expression (19) if $M = 0.040$, 0.0074, and 0.0014, respectively. Thus, for these bandwidths to be optimal, one has to be very optimistic about the smoothness of the regression function. In comparison, the rule of thumb method for estimating M discussed in Section 3.4 yields $\hat{M}_{\text{ROT}} = 0.299$, implying h^*_{RMSE} estimate 4.0, and the point estimate -3.17. For these smoothness parameters, the critical values based on the finite-sample bias-sd ratio are given by 2.165, 2.187, 2.107 and

[17] Ludwig and Miller (2007) state 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.

2.202 respectively, which is very close to the asymptotic value $\text{cv}_{.95}(1/2) = 2.181$. The resulting 95% confidence intervals are given by

$$(-4.143, 0.353), \qquad (-2.720, 0.323), \qquad (-2.215, -0.013), \qquad \text{and} \qquad (-6.352, 0.010),$$

respectively. The p-values based on these estimates are given by 0.100, 0.125, 0.047, and 0.051. These values are higher than those reported in the paper, as they take into account the potential bias of the estimates.

Using a triangular kernel helps to tighten the confidence intervals by about 2–4% in length, as predicted by the relative asymptotic efficiency results from Table 3, yielding

$$(-4.138, 0.187), \qquad (-2.927, 0.052), \qquad (-2.268, -0.095), \qquad \text{and} \qquad (-5.980, -0.322)$$

The underlying optimal bandwidths are given by 11.6, 23.1, 45.8, and 4.9 respectively. The p-values associated with these estimates are 0.074, 0.059, 0.033, and 0.028, tightening the p-values based on the uniform kernel.

These results indicate that unless one is very optimistic about the smoothness of the regression function, the uncertainty associated with the magnitude of the effect of Head Start assistance on child mortality is much higher than originally reported. This is due mainly to the relatively large bandwidths used by Ludwig and Miller (2007), which imply an optimistic bound on the smoothness of the regression function if we assume that such bandwidths are close to optimal for MSE. Interestingly, while the more conservative smoothness bound in our benchmark specification leads to much wider CIs, the point estimate is larger in magnitude, so that one still finds a statistically significant effect at a 5 or 10% level, depending on the kernel.

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

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

By Equation (7), the bandwidth sequence h_n must satisfy $\liminf_{n\to\infty} h_n(nM^2)^{1/[2(\gamma_b-\gamma_s)]} > 0$ and $\limsup_{n\to\infty} h_n(nM^2)^{1/[2(\gamma_b-\gamma_s)]} < \infty$. Thus, by Equation (6),

$$M^{r-1}n^{r/2}R(\hat{T}(h_n;k)) = S(k)^r B(k)^{1-r} t_n^{r-1}\tilde{R}(t_n,1) + o(1)$$

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 (6). 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 $\mathrm{sd}_n/\widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}};k) \xrightarrow{p} 1$ uniformly over $f \in \mathcal{F}$. Note that, by Theorem 2.1 and the fact that $t^*_{\mathrm{RMSE}} = \sqrt{1/r-1}$,

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

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

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

Thus,

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

and

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

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

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63.

Armstrong, T. B. (2018). Adaptation bounds for confidence bands under self-similarity. ArXiv: 1810.09762.

Armstrong, T. B. and Kolesár, M. (2018a). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.

Armstrong, T. B. and Kolesár, M. (2018b). A simple adjustment for bandwidth snooping. *Review of Economic Studies*, 85(2):732–765.

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

Brown, L. D., Low, M. G., and Zhao, L. H. (1997). Superefficiency in nonparametric function estimation. *The Annals of Statistics*, 25(6):2607–2625.

Bull, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics*, 6:1490–1516.

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

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(522):767–779.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2019). Coverage Error Optimal Confidence Intervals for Local Polynomial Regression. *arXiv:1808.01398 [econ, math, stat]*.

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.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818.

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*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, New York, NY.

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. Y. (2018). Minimax linear estimation at a boundary point. *Journal of Multivariate Analysis*, 165:262–269.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *The Annals of Statistics*, 38(2):1122–1170.

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.

Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics*, 41(4):1892–1921.

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. and Wager, S. (2019). Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278.

Kolesár, M. and Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, 108(8):2277–2304.

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.

Lepski, O. V. (1990). On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466.

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.

Noack, C. and Rothe, C. (2019). Bias-aware inference in fuzzy regression discontinuity designs. Unpublished manuscript, University of Mannheim.

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.

Schennach, S. M. (2015). A bias bound approach to nonparametric inference. Working Paper CWP71/15, Cemmap.

Sun, Y. (2005). Adaptive estimation of the regression discontinuity model. Unpublished manuscript, Univesity of California, San Diego.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York, NY.

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

Table 1: Critical values $\mathrm{cv}_{1-\alpha}(\cdot)$

		α		
r	t	0.01	0.05	0.1
	0.0	2.576	1.960	1.645
6/7	0.408	2.764	2.113	1.777
4/5	0.5	2.842	2.181	1.839
2/3	0.707	3.037	2.362	2.008
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

Notes: Critical values $\mathrm{cv}_{1-\alpha}(t)$ and $\mathrm{cv}_{1-\alpha}(\sqrt{1/r-1})$, for the FLCIs in (1) and (8), corresponding to the $1-\alpha$ quantiles of the $|N(t,1)|$ and $|N(\sqrt{1/r-1},1)|$ distributions, where t is the bias-sd ratio, and r is the rate exponent. For $t \geq 2$, $\mathrm{cv}_{1-\alpha}(t) \approx t + z_{1-\alpha/2}$ up to 3 decimal places for these values of $1-\alpha$.

Table 2: Relative efficiency of local polynomial estimators for the function class $\mathcal{F}_{\mathrm{T},p}(M)$.

		Boundary Point			Interior point		
Kernel	Order	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
Uniform $\mathrm{I}\{\lvert u\rvert \leq 1\}$	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

Notes: Efficiency is relative to the optimal equivalent kernel k^*_{SY}. The functional Tf corresponds to the value of f at a point.

Table 3: Relative efficiency of local polynomial estimators for the function class $\mathcal{F}_{\text{Höl},p}(M)$.

		Boundary Point			Interior point		
Kernel	Order	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
Uniform $\mathrm{I}\{\lvert u\rvert \leq 1\}$	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

Notes: For $p = 1, 2$, efficiency is relative to the optimal kernel, for $p = 3$, efficiency is relative to the local quadratic estimator with triangular kernel. The functional Tf corresponds to the value of f at a point.

Table 4: Gains from imposing global smoothness

	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	

Notes: Table gives the relative 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 the risk under $\mathcal{F}_{\text{T},p}(M)$. "Optimal" refers to using the optimal kernel under a given smoothness class.

Table 5: Performance of RBC CIs based on h^*_{RMSE} bandwidth for local linear regression under $\mathcal{F}_{\text{T},2}$ and $\mathcal{F}_{\text{Höl},2}$.

	$\mathcal{F}_{\text{T},2}$			$\mathcal{F}_{\text{Höl},2}$		
Kernel	Length	Coverage	t_{RBC}	Length	Coverage	t_{RBC}
Boundary						
Uniform	1.35	0.931	0.400	1.35	0.948	0.138
Triangular	1.32	0.932	0.391	1.32	0.947	0.150
Epanechnikov	1.33	0.932	0.393	1.33	0.947	0.148
Interior						
Uniform	1.35	0.941	0.279	1.35	0.949	0.086
Triangular	1.27	0.940	0.297	1.27	0.949	0.110
Epanechnikov	1.30	0.940	0.298	1.30	0.949	0.105

Legend: Length—CI length relative to 95% FLCI based on a local linear estimator and the same kernel and bandwidth h^*_{RMSE}; t_{RBC}—ratio of the worst-case bias to standard deviation;

Table 6: Monte Carlo simulation: Inference at a point.

		$M=2$					$M=6$				
Method	Bandwidth	Bias	SE	$E[h]$	Cov	RL	Bias	SE	$E_m[h]$	Cov	RL
Design 1											
RBC	$h=\hat{h}^*_{\text{PT}}$, $b=\hat{b}^*_{\text{PT}}$	0.063	0.035	0.75	55.6	0.73	0.157	0.036	0.62	0.1	0.61
RBC	$h=\hat{h}_{\text{CE}}$, $b=\hat{b}_{\text{CE}}$	0.030	0.041	0.45	85.8	0.85	0.059	0.045	0.34	72.4	0.76
RBC	$h=b=\hat{h}^*_{\text{RMSE},2}$	0.001	0.061	0.36	94.5	1.27	0.002	0.061	0.36	94.5	1.01
RBC	$h=b=\hat{h}^*_{\text{RMSE},6}$	0.000	0.076	0.23	94.2	1.58	0.000	0.075	0.23	94.2	1.26
RBC	$h=b=\hat{h}^*_{\text{RMSE},\hat{M}_{\text{ROT}}}$	0.000	0.078	0.22	93.9	1.64	0.000	0.097	0.14	93.4	1.63
Conventional	$\hat{h}^*_{\text{PT,ROT}}$	0.032	0.036	0.56	76.6	0.76	0.049	0.046	0.31	77.4	0.77
FLCI, $M=2$	$\hat{h}^*_{\text{RMSE},2}$	0.021	0.043	0.36	94.9	1.00	0.065	0.043	0.36	75.2	0.80
FLCI, $M=6$	$\hat{h}^*_{\text{RMSE},6}$	0.009	0.054	0.23	96.6	1.25	0.028	0.053	0.23	94.7	1.00
FLCI, $M=\hat{M}_{\text{ROT}}$	$\hat{h}^*_{\text{RMSE},\hat{M}_{\text{ROT}}}$	0.008	0.056	0.22	95.6	1.29	0.010	0.069	0.14	96.3	1.30
Design 2											
RBC	$h=\hat{h}^*_{\text{PT}}$, $b=\hat{b}^*_{\text{PT}}$	0.043	0.035	0.77	75.9	0.72	0.129	0.035	0.77	4.6	0.58
RBC	$h=\hat{h}_{\text{CE}}$, $b=\hat{b}_{\text{CE}}$	0.028	0.040	0.49	87.4	0.83	0.074	0.041	0.44	54.1	0.69
RBC	$h=b=\hat{h}^*_{\text{RMSE},2}$	0.002	0.061	0.36	94.5	1.27	0.006	0.061	0.36	94.4	1.01
RBC	$h=b=\hat{h}^*_{\text{RMSE},6}$	0.000	0.076	0.23	94.2	1.58	0.000	0.075	0.23	94.2	1.26
RBC	$h=b=\hat{h}^*_{\text{RMSE},\hat{M}_{\text{ROT}}}$	0.001	0.068	0.30	94.0	1.43	0.000	0.083	0.20	93.8	1.38
Conventional	$\hat{h}^*_{\text{PT,ROT}}$	0.032	0.032	0.78	74.4	0.67	0.073	0.040	0.44	53.0	0.66
FLCI, $M=2$	$\hat{h}^*_{\text{RMSE},2}$	0.020	0.043	0.36	95.1	1.00	0.061	0.043	0.36	78.1	0.80
FLCI, $M=6$	$\hat{h}^*_{\text{RMSE},6}$	0.009	0.054	0.23	96.6	1.25	0.028	0.053	0.23	94.7	1.00
FLCI, $M=\hat{M}_{\text{ROT}}$	$\hat{h}^*_{\text{RMSE},\hat{M}_{\text{ROT}}}$	0.013	0.048	0.30	94.3	1.13	0.020	0.059	0.20	94.3	1.10
Design 3											
RBC	$h=\hat{h}^*_{\text{PT}}$, $b=\hat{b}^*_{\text{PT}}$	-0.043	0.035	0.77	75.7	0.72	-0.123	0.035	0.74	9.9	0.59
RBC	$h=\hat{h}_{\text{CE}}$, $b=\hat{b}_{\text{CE}}$	-0.026	0.040	0.49	88.1	0.83	-0.063	0.043	0.43	64.2	0.71
RBC	$h=b=\hat{h}^*_{\text{RMSE},2}$	-0.002	0.061	0.36	94.5	1.27	-0.007	0.061	0.36	94.4	1.01
RBC	$h=b=\hat{h}^*_{\text{RMSE},6}$	0.000	0.076	0.23	94.2	1.58	0.000	0.075	0.23	94.2	1.26
RBC	$h=b=\hat{h}^*_{\text{RMSE},\hat{M}_{\text{ROT}}}$	0.000	0.074	0.25	94.2	1.54	0.000	0.092	0.16	93.6	1.54
Conventional	$\hat{h}^*_{\text{PT,ROT}}$	-0.032	0.033	0.72	74.7	0.69	-0.065	0.042	0.39	62.0	0.70
FLCI, $M=2$	$\hat{h}^*_{\text{RMSE},2}$	-0.020	0.043	0.36	95.0	1.00	-0.060	0.043	0.36	78.1	0.80
FLCI, $M=6$	$\hat{h}^*_{\text{RMSE},6}$	-0.009	0.054	0.23	96.5	1.25	-0.027	0.053	0.23	94.7	1.00
FLCI, $M=\hat{M}_{\text{ROT}}$	$\hat{h}^*_{\text{RMSE},\hat{M}_{\text{ROT}}}$	-0.010	0.052	0.25	95.6	1.22	-0.013	0.065	0.16	96.1	1.22

Legend: SE—average standard error; $E[h]$—average (over Monte Carlo draws) bandwidth; Cov—coverage of CIs (in %); RL—relative (to optimal FLCI) length.
Bandwidth descriptions: $\hat{h}^*_{\text{PT}}$—plugin estimate of pointwise MSE optimal bandwidth (bw); $\hat{b}^*_{\text{PT}}$—analog for estimate of the bias; $\hat{h}_{\text{CE}}$—plugin estimate of coverage error optimal bw; $\hat{b}_{\text{CE}}$—analog for estimate of the bias; The implementation of Calonico et al. (2018) is used for all four bws. $\hat{h}^*_{\text{RMSE},2}$, $\hat{h}^*_{\text{RMSE},6}$—RMSE optimal bw, assuming $M=2$, and $M=6$, respectively. $\hat{h}^*_{\text{PT,ROT}}$—Fan and Gijbels (1996) rule of thumb; $\hat{h}^*_{\text{RMSE},\hat{M}_{\text{ROT}}}$—RMSE optimal bw, using rule-of-thumb for M. 50,000 Monte Carlo draws.

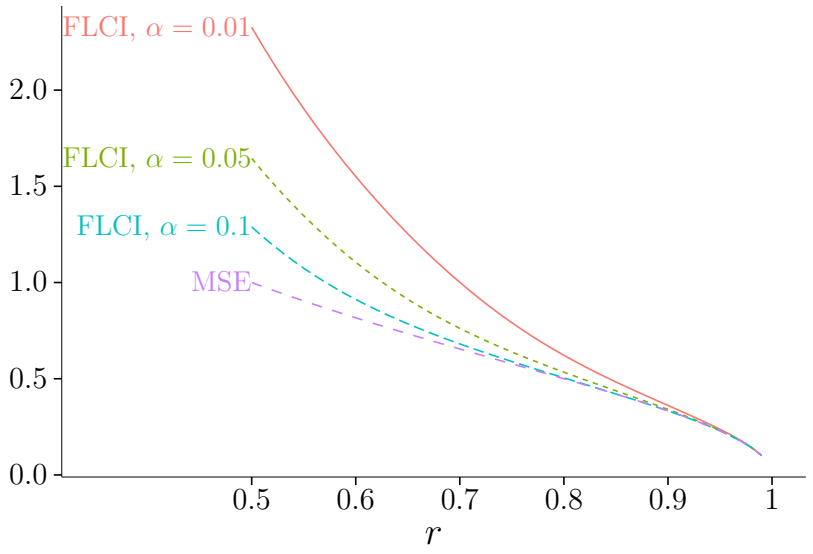

Figure 1: Optimal ratio of the worst-case bias to standard deviation for fixed length CIs (FLCI), and maximum MSE (MSE) performance criteria.

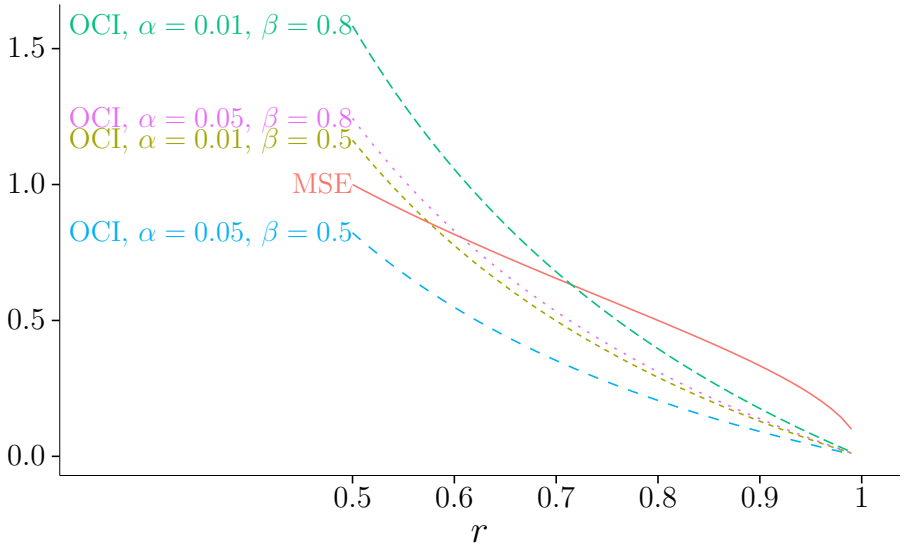

Figure 2: Optimal ratio of the worst-case bias to standard deviation for one-sided CIs (OCI), and maximum MSE (MSE) performance criteria.

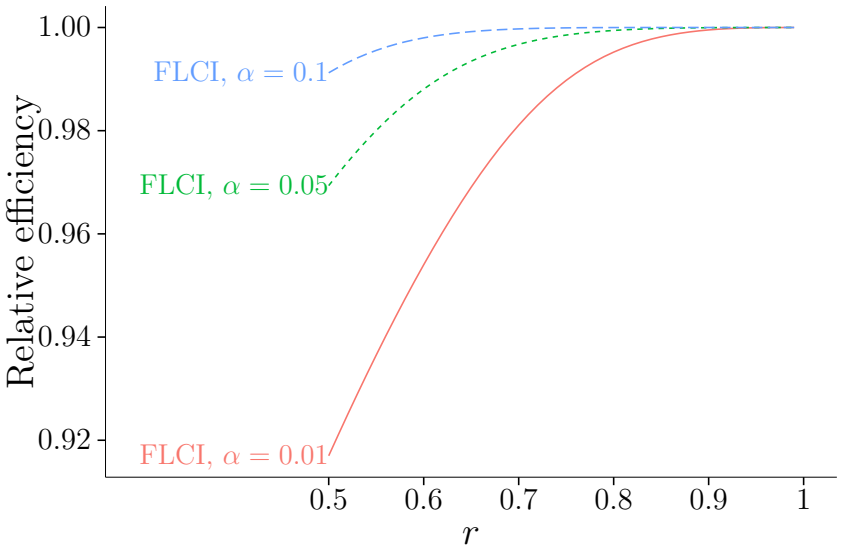

Figure 3: Efficiency of fixed-length CIs based on minimax MSE bandwidth relative to fixed-length CIs based on optimal bandwidth.

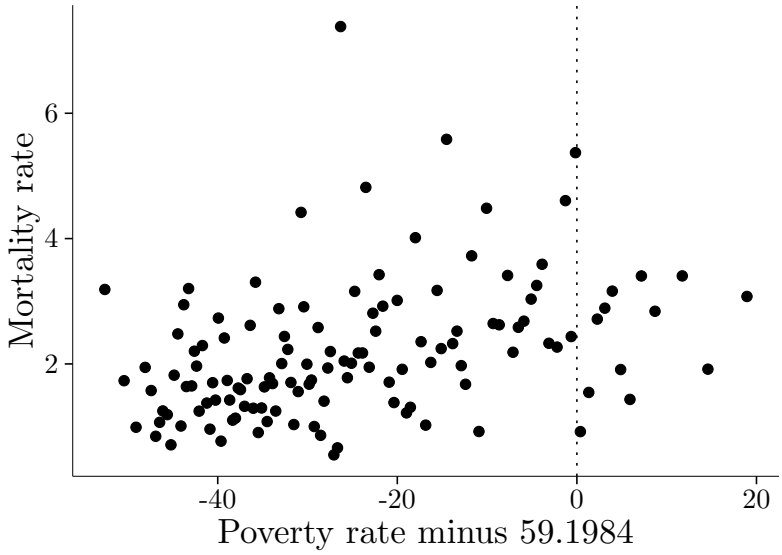

Figure 4: Average county mortality rate per 100,000 for children aged 5–9 over 1973–83 due to causes addressed as part of Head Start's health services (labeled "Mortality rate") plotted against poverty rate in 1960 relative to the 300th poorest county. Each point corresponds to an average for 25 counties.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

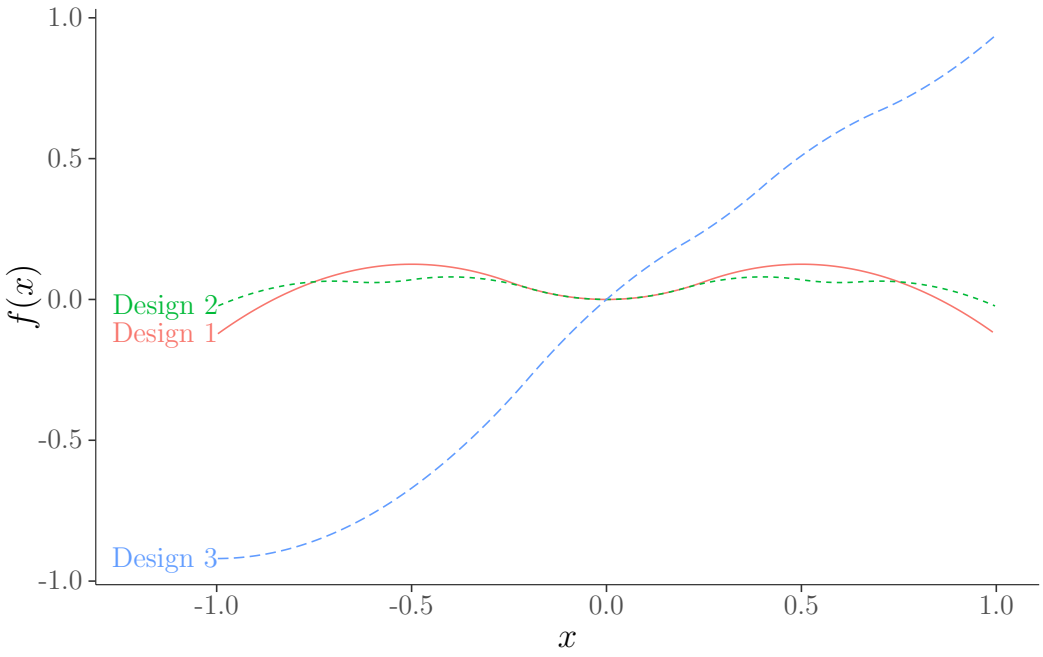

Figure 5: Monte Carlo simulation Designs 1–3, and $M = 2$.