Simple and Honest Confidence Intervals in Nonparametric Regression*

Timothy B. Armstrong† Michal Kolesár‡

Yale University Princeton University

March 18, 2018

Abstract

We consider the problem of constructing honest confidence intervals (CIs) for a scalar parameter of interest, such as the regression discontinuity parameter, in nonparametric regression based on kernel or local polynomial estimators. To ensure that our CIs are honest, we derive and tabulate novel critical values that take into account the possible bias of the estimator upon which the CIs are based. We 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 and numerous seminar and conference participants for helpful comments and suggestions. All remaining errors are our own. The research of the first author was supported by National Science Foundation Grant SES-1628939. The research of the second author was supported by National Science Foundation Grant SES-1628878.

†email: timothy.armstrong@yale.edu

‡email: mkolesar@princeton.edu

1 Introduction

This paper considers the problem of constructing confidence intervals (CIs) for a scalar parameter $T(f)$ of a function f, which can be a conditional mean or a density. The scalar parameter may correspond, for example, to a conditional mean, or its derivatives at a point, the regression discontinuity parameter, or the value of a density or its derivatives at a point. When the goal is to estimate $T(f)$, a popular approach 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).

In this paper, we show that one can also use these estimators to construct simple, and highly efficient confidence intervals (CIs): simply add and subtract its 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] We tabulate these critical values, and show that they depend only on (1) the optimal rate of convergence (equivalently, the order of the derivative that one bounds to obtain the asymptotic MSE); 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. Asymptotically, our CIs correspond to fixed-length CIs as defined in Donoho (1994), and so we refer to them as fixed-length CIs. We show that these CIs are near optimal in terms of their length in the class of honest CIs. As in Li (1989), we formalize the notion of honesty by requiring that the CI cover the true parameter asymptotically at the nominal level uniformly over a parameter space $\mathcal{F}$ for f (which typically places bounds on derivatives of 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 $\mathcal{F}$.

In deriving these results, we answer three main questions. First, how to optimally form a CI based on a given class of kernel-based estimators? Popular approaches include undersmoothing (choosing the bandwidth to shrink more quickly than the MSE optimal bandwidth) and bias-correction (subtracting an estimate of the bias). We show that widening the CI to take into account bias is more efficient (in the sense of leading to a smaller CI while maintaining coverage) than both of these approaches. In particular, we show that, in contrast to the practice of undersmoothing, the optimal bandwidth for CI construction is *larger* than the MSE optimal bandwidth. This contrasts with the work of Hall (1992) and Calonico et al. (2017) 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

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

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 the 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 in order to maintain coverage).

Second, since the MSE criterion is typically used for estimation, one may prefer to report a CI that is centered around the MSE optimal estimator, rather than reoptimizing the bandwidth for length and coverage of the CI. How much is lost by using the MSE optimal bandwidth to construct the CI? We show that, under a range of conditions most commonly used in practice, the loss in efficiency is very small: a fixed-length CI centered at the MSE optimal bandwidth is 99% efficient in these settings. Therefore, there is little efficiency loss from not re-optimizing the bandwidth for inference.

Third, how much is lost by using a kernel that is not fully optimal? We show that the relative kernel efficiency for the CIs that we propose, in terms of their length, 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. Moreover, it follows from calculations in Donoho (1994) and Armstrong and Kolesár (2017) 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.

The requirement of honesty, or uniformity the parameter space $\mathcal{F}$, that underlies our analysis, is important for two reasons. First, it leads to a well-defined and consistent concept of optimality. For example, it 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 (which do not require such uniformity), 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).

Second, it is necessary for good finite-sample performance. For example, as we show in Section 4.1, bandwidths that optimize the asymptotic MSE derived under pointwise-in-f asymptotics can lead to arbitrarily poor finite-sample behavior. This point is borne out in our Monte Carlo study, in which we show that commonly used plug-in bandwidths that attempt to estimate this pointwise-in-f optimal bandwidth can lead to severe undercoverage, even when combined

with undersmoothing or bias-correction. In contrast, fixed-length CIs perform as predicted by our theory.

Our approach requires an explicit definition of the parameter space $\mathcal{F}$. When the parameter space bounds derivatives of f, the parameter space will depend on this particular bound M. Unfortunately, the results of Low (1997), Cai and Low (2004), and Armstrong and Kolesár (2017) show that it is impossible to avoid choosing M a priori 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. We therefore recommend that, whenever possible, problem-specific knowledge be used to decide what choice of M is reasonable a priori. We also propose a data-driven rule of thumb for choosing M, although, by the above impossibility results, one needs to impose additional assumptions on f in order to guarantee honesty. 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.

While our results show that undersmoothing is inefficient, 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 it implicitly chooses a sequence $M_n \to \infty$ under which coverage is controlled. Given a sequence of undersmoothed bandwidths, we show how this sequence 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.

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

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

many classical nonparametric settings.

The rest of this paper is organized as follows. Section 2 gives the main results. Section 3 applies our results to inference at a point, Section 4 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 applies the results to RD, and presents an empirical application based on Ludwig and Miller (2007). Appendix A discusses implementation details and includes a proposal for a rule of thumb for choosing M. Appendix B gives proofs of the results in Section 2. The supplemental materials contain further appendices and additional tables and figures. Appendix C verifies our regularity conditions for some examples, and includes proofs of the results in Section 3. Appendix D calculates the efficiency gain from using different bandwidths on either side of a cutoff in RD that is used in Section 6. Appendix E contains details on optimal kernel calculations discussed in Section 3.

2 General results

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

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

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

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

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

where the $o(1)$ term in the second equality is uniform over $f \in \mathcal{F}$. Note that the second condition implies that the standard deviation does not depend on the underlying function f asymptotically. As we show in Appendix C in the supplemental materials, this condition (as well as the other conditions used in this section) holds 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 conditional mean, or its derivative, evaluated at a point (which may be a boundary point). In Section 3, we show that (1) 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.

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

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

where $r = \gamma_b/(\gamma_b - \gamma_s)$. Since the bias and the standard deviation both converge at rate $n^{r/2}$ when M is fixed, we refer to r as the rate exponent (this matches the definition in, e.g., Donoho and Low 1992; see Appendix C in the supplemental materials).

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

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

where $\text{cv}_{1-\alpha}(t)$ is the $1-\alpha$ quantile of the $|N(t,1)|$ distribution (see Table 1). This 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. One could also form honest CIs by simply adding and subtracting the worst case bias, in addition to adding and subtracting the standard error times $z_{1-\alpha/2} = \text{cv}_{1-\alpha}(0)$, the $1-\alpha/2$ quantile of a standard normal distribution, forming the CI as $\hat{T}(h;k) \pm (\overline{\text{bias}}(\hat{T}(h;k)) + z_{1-\alpha/2} \cdot \widehat{\text{se}}(h;k))$. However, since the estimator $\hat{T}(h;k)$ cannot simultaneously have a large positive and a large negative bias, such CI will be conservative, and longer than the CI given in Equation (3).

Honest one-sided $1-\alpha$ CIs based on $\hat{T}(h;k)$, can be constructed by simply subtracting the maximum bias, in addition to subtracting $z_{1-\alpha}$ times the standard deviation, from $\hat{T}(h;k)$:

$$[\hat{T}(h;k) - h^{\gamma_b}MB(k) - z_{1-\alpha}h^{\gamma_s}n^{-1/2}S(k)\ ,\ \infty). \tag{4}$$

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

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

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 (4), 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 $\tilde{R}_{\text{OCI},\alpha,\beta}(b,s) \equiv 2b + (z_{1-\alpha}+z_\beta)s$. Finally, to evaluate $\hat{T}(h;k)$ as an estimator we use root mean squared error (RMSE) as the performance criterion:

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

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

$$R(\hat{T}(h;k)) = \tilde{R}(h^{\gamma_b}MB(k), h^{\gamma_s}n^{-1/2}S(k))(1+o(1)). \tag{5}$$

To keep the statement of our main results simple, we make this assumption directly. As is the case for condition (1), we show in Appendix C in the supplemental materials that this condition will typically hold in most classical nonparametric problems. In Section 3, we verify it for the problem of estimation of a conditional mean at a point. We will also assume that $\tilde{R}$ scales linearly in its arguments (i.e. it is homogeneous of degree one): $\tilde{R}(tb,ts) = t\tilde{R}(b,s)$. This holds for all three criteria considered above. Plugging in (2) and using scale invariance of $\tilde{R}$ gives

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

where $t = h^{\gamma_b-\gamma_s}MB(k)/(n^{-1/2}S(k))$ is the bias-sd ratio and $r = \gamma_b/(\gamma_b-\gamma_s)$ is the rate exponent, as defined above. 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 (5) 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:

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

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.[2]

We collect some implications of these derivations in a theorem.

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 (5) 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 Equation (7) holds. Let h_R^* and t_R^* be as defined above, and assume that $t_R^* > 0$ is unique and well-defined. Then:*

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

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

where $h_R^* = (n^{-1/2}S(k)t_R^*/(MB(k)))^{1/(\gamma_b-\gamma_s)}$, and $t_R^* = \operatorname{argmin}_t t^{r-1}\tilde{R}(t,1)$.

(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.

[2]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.

(iii) If (1) 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. 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.

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

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

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.

One can also form FLCIs centered at the estimator that is optimal for different performance criterion R as $\hat{T}(h_R^*;k) \pm \widehat{\text{se}}(h_R^*;k)\cdot \text{cv}_{1-\alpha}(t_R^*)$. The critical value $\text{cv}_{1-\alpha}(t_R^*)$ 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 $t_{RMSE}^* = \sqrt{1/r-1}$. Table 1 reports this critical value $\text{cv}_{1-\alpha}(\sqrt{1/r-1})$ for some 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

$$\frac{(t^*_{FLCI})^{r-1} \cdot \mathrm{cv}_{1-\alpha}(t^*_{FLCI})}{(t^*_{RMSE})^{r-1} \cdot \mathrm{cv}_{1-\alpha}(t^*_{RMSE})}. \tag{8}$$

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 only about 3% 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 $\widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}}; k)$ be such that $\widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}}; k)/\operatorname{sd}_f(h^*_{\mathrm{RMSE}}; k)$ converges in probability to 1 uniformly over $f \in \mathcal{F}$. Then*

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

The asymptotic efficiency of this CI relative to the one centered at the FLCI optimal bandwidth, defined as $\lim_{n\to\infty} \frac{\inf_{h>0} R_{FLCI,\alpha}(\hat{T}(h;k))}{R_{FLCI,\alpha}(\hat{T}(h^*_{\mathrm{RMSE}};k))}$, is given by (8). It depends only on r.

3 Inference at a point

In this section, we apply the general results from Section 2 to the problem of inference about a nonparametric regression function at a point, which we normalize to be zero, so that $T(f) = f(0)$. We allow the point of interest to be on the boundary on the parameter space. Because in sharp regression discontinuity (RD) designs, discussed in detail in Section 6, the parameter of interest can be written as the difference between two regression functions evaluated at boundary points, the results in this section generalize naturally to sharp RD.

We write the nonparametric regression model as

$$y_i = f(x_i) + u_i, \quad i = 1, \ldots, n, \tag{9}$$

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

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

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

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

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. In particular, $\hat{T}_q(h; k)$ 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

$$k_q^*(u) = e_1' \left(\int_{\mathcal{X}} m_q(t) m_q(t)' k(t) \, \mathrm{d}t \right)^{-1} m_q(u) k(u), \tag{10}$$

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 and normally distributed with $Eu_i = 0$ and $\mathrm{var}(u_i) = \sigma^2(x_i)$, and the variance function $\sigma^2(x)$ is continuous at $x = 0$.*

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

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

$$\mathrm{sd}(\hat{T}_q(h; k))^2 = \sum_{i=1}^n w_q^n(x_i)^2 \sigma^2(x_i) = \left(\frac{\sigma^2(0)}{dnh} \int_{\mathcal{X}} k_q^*(u)^2 \, \mathrm{d}u \right) (1 + o(1)), \tag{11}$$

where the second equality holds under Assumptions 3.1 and 3.2, as we show in Appendix C.2 in the supplemental materials. The condition on the standard deviation in Equation (1) thus holds with

$$\gamma_s = -1/2 \quad \text{and} \quad S(k) = d^{-1/2} \sigma(0) \sqrt{\int_{\mathcal{X}} k_q^*(u)^2 \, \mathrm{d}u}. \tag{12}$$

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

On the other hand, the worst-case bias will be driven primarily by the function class $\mathcal{F}$. We consider inference under two popular function classes. First, the 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[3],

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

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 thus formalizes the intuitive notion that f should be p-times differentiable with a bound on the pth derivative. The case $p = 1$ corresponds to the Lipschitz class of functions.

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

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

and

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

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

If, in addition, Assumption 3.2 holds, then Equation (5) *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 (12).*

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

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 (when $q < p-1$, the maximum bias is infinite).

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

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

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

3.1 Kernel efficiency

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

$$\left(\int_{\mathcal{X}} k^*(u)^2 \, \mathrm{d}u\right)^p \left(\int_{\mathcal{X}} |u^p k^*(u)| \, \mathrm{d}u\right), \tag{13}$$

The solution to this problem follows from Sacks and Ylvisaker (1978, Theorem 1) (see also Cheng et al. (1997)). We give details of the solution as well as plots of the optimal kernels in Appendix E in the supplemental materials. In Table 2, we compare the asymptotic relative efficiency of local polynomial estimators based on the uniform, triangular, and Epanechnikov kernels to the optimal Sacks-Ylvisaker kernels.

Fan et al. (1997) and Cheng et al. (1997), conjecture that minimizing (13) yields a sharp bound on kernel efficiency. It follows from Theorem 2.1 (ii) that this conjecture is correct, and

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

The solution to minimizing $S(k)^r B(k)^{1-r}$ under $\mathcal{F}_{\mathrm{Höl},p}(M)$ 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 Appendix E in the supplemental materials for details, including plots of these kernels. When $p \geq 3$, the solution is unknown. Therefore, for $p=3$, we compute efficiencies relative to a local quadratic estimator with a triangular kernel. Table 3 calculates the resulting efficiencies for local polynomial estimators based on the uniform, triangular, and Epanechnikov kernels. Relative to the class $\mathcal{F}_{\mathrm{T},p}(M)$, the bias constants are smaller: imposing smoothness away from the point of interest helps to reduce the worst-case bias. Furthermore, the loss of efficiency from using a local polynomial estimator of order that's too high is smaller. Finally, one can see that local linear regression with a triangular kernel achieves high asymptotic efficiency under both $\mathcal{F}_{\mathrm{T},2}(M)$ and $\mathcal{F}_{\mathrm{Höl},2}(M)$, both at the interior and at a boundary, with efficiency at least 97%, which shows that its popularity in empirical work can be justified on theoretical grounds. Under $\mathcal{F}_{\mathrm{Höl},2}(M)$ on the boundary, the triangular kernel is nearly efficient.

3.2 Gains from imposing smoothness globally

The Taylor class $\mathcal{F}_{\mathrm{T},p}(M)$, formalizes the notion that the pth derivative at 0, the point of interest, should be bounded by M, but doesn't impose smoothness away from 0. In contrast, the Hölder class $\mathcal{F}_{\mathrm{Höl},p}(M)$ restricts the pth derivative to be at most M 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 risk 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 reduction in risk/CI length for common kernels. One can see that in general, the gains are greater for larger p, and greater at the boundary. In the case of estimation at a boundary point with $p = 2$, for example, imposing global smoothness of f results in reduction in length of about 13–15%, depending on the kernel, and a reduction of about 10% if the optimal kernel is used.

3.3 Practical implementation

Given a smoothness class $\mathcal{F}_{\mathrm{T},p}(M)$ or $\mathcal{F}_{\mathrm{Höl},p}(M)$, Theorems 2.1, 2.2, and 3.1 imply that one can construct nearly efficient CIs for $f(0)$ as $\hat{T}_{p-1}(h^*_{\mathrm{RMSE}}; k) \pm \mathrm{cv}_{1-\alpha}(\sqrt{1/r-1}) \cdot \widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}}, k)$. Alternatively, one could use the critical value $\mathrm{cv}_{1-\alpha}(\overline{\mathrm{bias}}(\hat{T}_{p-1}(h^*_{\mathrm{RMSE}}; k))/\widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}}, k))$ based on the finite-sample bias-sd ratio (see Theorem C.1 in the supplemental materials for the finite-sample bias expression). To implement this CI, one needs to (i) choose p, M, and k; (ii) form an estimate $\widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}}, k)$ of the standard deviation of $\hat{T}_{p-1}(h^*_{\mathrm{RMSE}}; k)$; and (iii) form an estimate of h^*_{RMSE} (which depends on the unknown quantities $\sigma^2(0)$ and d). We now discuss these issues in turn, with reference to Appendix A for additional details.

The choice of p depends on the order of the derivative the researcher wishes to bound and it determines the order of local polynomial. Since local linear estimators are the most popular in practice, we recommend $p = 2$ as a default choice. In this case, both the Epanechnikov and the triangular kernel are nearly optimal. For M, the results of Low (1997), Cai and Low (2004) and Armstrong and Kolesár (2017) imply that to maintain honesty over the whole function class, a researcher must choose the constant a priori, rather than attempting to use a data-driven method. We therefore recommend that, whenever possible, problem-specific knowledge be used to decide what choice of M is reasonable a priori, and that one considers a range of plausible values by way of sensitivity analysis.[4] If additional restrictions on f are imposed, a data-driven method for choosing M may be feasible. In Appendix A.1, we describe a rule-of-thumb method based on the suggestion in Fan and Gijbels (1996, Chapter 4.2).

For the standard error $\widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}}, k)$, many choices are available in the literature. In our Monte Carlo and application, we use a nearest-neighbor estimator discussed in Appendix A.2. To compute h^*_{RMSE}, one can plug in the constant M (discussed above) along with estimates of d, and $\sigma^2(0)$. Alternatively, one can plug in M and an estimate of the function $\sigma^2(\cdot)$ to the formula for the finite-sample RMSE. See Appendix A.3 for details.

[4]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.

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 focus on the problem of inference about a nonparametric regression function at a point, as in Section 3. The first approach, that 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 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 $\hat{T}_q(h^*_{\text{RMSE}};k)$.

4.1 RMSE and pointwise optimal bandwidth

The general results from Section 2 imply that given a kernel k and order of a local polynomial q, the RMSE-optimal bandwidth for $\mathcal{F}_{\text{T},p}(M)$ and $\mathcal{F}_{\text{Höl},p}(M)$ in the conditional mean estimation problem in Section 3 is given by

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

where $\mathcal{B}_{p,q}(k) = \mathcal{B}^{\text{Höl}}_{p,q}(k)$ for $\mathcal{F}_{\text{Höl},p}(M)$, and $\mathcal{B}_{p,q}(k) = \mathcal{B}^{\text{T}}_{p,q}(k)$ for $\mathcal{F}_{\text{T},p}(M)$. In contrast, the optimal bandwidth based on pointwise asymptotics is obtained by minimizing the sum of the leading squared bias and variance terms under pointwise asymptotics for the case $q = p-1$. This bandwidth is given by (see, for example, Fan and Gijbels, 1996, Eq. (3.20))

$$h^*_{\text{PT}} = \left(\frac{\sigma^2(0)p!^2}{2pdnf^{(p)}(0)^2}\frac{\int_{\mathcal{X}} k^*_q(u)^2\,\mathrm{d}u}{(\int_{\mathcal{X}} t^p k^*_q(t)\,\mathrm{d}t)^2}\right)^{\frac{1}{2p+1}}. \tag{15}$$

Thus, 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, suppose that the support of x_i is finite and contains the point of interest $x = 0$. 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.

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 $\mathcal{F}(M)$. 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 this CIs are shorter than the 95% FLCIs by a factor of $z_{0.975}/\operatorname{cv}_{0.95}(1/2) = 0.90$. Consequently, their coverage is 92.1% rather than the nominal 95% coverage. At the RMSE-optimal bandwidth, the 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 $\hat{T}(h^n;k) \pm \mathrm{cv}_{1-\alpha}(0)\widehat{\mathrm{se}}(h_n;k) = \hat{T}(h^n;k) \pm z_{1-\alpha/2}\widehat{\mathrm{se}}(h_n;k)$ 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 constants $M_n \to \infty$ such that coverage is controlled under the sequence of parameter spaces $\mathcal{F}(M_n)$. 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 $\mathcal{F}(M_n)$.

To this end, given a sequence h_n, a better approximation to the finite-sample coverage of the CI $\hat{T}(h^n;k) \pm z_{1-\alpha/2}\widehat{\mathrm{se}}(h_n;k)$ over the parameter space $\mathcal{F}(M)$ 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 Appendix C. 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 $f \in \mathcal{F}(M_n)$: 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, their procedure amounts to using a local quadratic estimator, but with a bandwidth h^*_{RMSE}, optimal for a local linear estimator. The resulting CI obtains by adding and subtracting $z_{1-\alpha/2}$ times the standard deviation of the estimator. The bias-sd ratio of the estimator is given by

$$t_{\mathrm{RBC}} = (h^*_{\mathrm{RMSE}})^{5/2}\frac{M\mathcal{B}_{2,2}(k)/2}{\sigma(0)(\int k_2^*(u)^2\,\mathrm{d}u/dn)^{1/2}} = \frac{1}{2}\frac{\mathcal{B}_{2,2}(k)}{\mathcal{B}_{2,1}(k)}\left(\frac{\int_{\mathcal{X}} k_1^*(u)^2\,\mathrm{d}u}{\int_{\mathcal{X}} k_2^*(u)^2\,\mathrm{d}u}\right)^{1/2}. \tag{16}$$

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

$$\frac{z_{1-\alpha/2}\left(\int_{\mathcal{X}} k_2^*(u)^2\,\mathrm{d}u\right)^{1/2}}{\text{cv}_{1-\alpha}(1/2)\left(\int_{\mathcal{X}} k_1^*(u)^2\,\mathrm{d}u\right)^{1/2}}(1+o(1)). \tag{17}$$

The resulting coverage and relative length is given in Table 5. One can see that although the coverage properties are excellent (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 (16) 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.

If one instead considers the classes $\mathcal{F}_{\text{T},3}(M)$ and $\mathcal{F}_{\text{Höl},3}(M)$ (but with h^*_{RMSE} still chosen to be MSE optimal for $\mathcal{F}_{\text{T},2}(M)$ or $\mathcal{F}_{\text{Höl},2}(M)$), then the 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. 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.

5 Monte Carlo

To study the performance of the FLCI that we propose, and compare its performance to other approaches, we conduct a Monte Carlo analysis of the conditional mean estimation problem considered in Section 3. We consider Monte Carlo designs with conditional mean functions

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

where $M \in \{2, 6\}$, giving a total of 6 designs. In all cases, x_i is uniform on $[-1, 1]$, $u_i \sim N(0, 1/4)$, 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.

For each design, we implement the optimal FLCI centered at the MSE optimal estimate, as described in Section 3.3, for each choice of $M \in \{2, 6\}$, and with M calibrated using the rule-of-thumb (ROT) described Appendix A.1. 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 in order to have correct coverage (see Section 3.3).

In addition to these FLCIs, we consider five other methods of CI construction. The first four 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 CI uses a plug-in estimate of h^*_{PT} defined in (15), as implemented by Calonico et al. (2017), and an analogous estimate for the pilot bandwidth (this method is the default in their accompanying software package). The second CI, also implemented by Calonico et al. (2017), 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 third and fourth CIs, we set both the main and the pilot bandwidth to h^*_{RMSE} with $M = 2$, and $M = 6$, respectively. 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 3.3, it is impossible for the ROT choice of M (or any other data-driven choice) to do this uniformly over the whole function class, so one must take care in extrapolating these results to other designs. As predicted by the theory in Section 4, the RBC CI has good coverage when implemented using h^*_{RMSE}, although it is on average about 25% longer than the corresponding FLCI.

The other CIs all have very poor coverage for at least one of the designs. Our analysis in Sections 4 suggests that this is due to the use of plug-in bandwidths that estimate the pointwise MSE optimal bandwidth h^*_{PT}. Indeed, looking at the average of the bandwidth over the Monte Carlo draws (also reported in Table 6), it can be seen that the plug-in bandwidths used for these 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 construction (bias-correction or undersmoothing relative to plug-in bandwidths) can lead to severe undercoverage, and that plug-in bandwidths justified by pointwise-in-f asymptotics are the main culprit. 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 Application to sharp regression discontinuity

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

Using data from the nonparametric regression model (9), the goal in sharp RD is to estimate the jump in the regression function f at a known threshold, which we normalize to 0, so that $T(f) = \lim_{x\downarrow 0} f(x) - \lim_{x\uparrow 0} f(x)$. The threshold determines participation in a binary treatment: units with $x_i \geq 0$ are treated; units with $x_i < 0$ are controls. If the regression functions of potential outcomes are continuous at zero, then $T(f)$ measures the average effect of the treatment for units with $x_i = 0$ (Hahn et al., 2001).

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

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

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

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

with the weight w^n_+ given by

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

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 $\hat{Q}_{n,-}$ and kernel k_- are defined similarly. That is, $\hat{T}(h;k)$ is given by a difference between estimates from two local linear regressions at a boundary point, one for units with non-negative values running variable x_i, and one for units with negative values of the running variable. Let $\sigma^2_+(x) = \sigma^2(x)1(x \geq 0)$,

and let $\sigma^2_-(x) = \sigma^2(x)1(x < 0)$.

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

It follows from the results in Section 3 that if Assumption 3.1 holds and the functions $\sigma^2_+(x)$ and $\sigma^2_-(x)$ are right- and left-continuous, respectively, the variance of the estimator doesn't depend on f and satisfies

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

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

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

The RMSE-optimal bandwidth is given by

$$h^*_{\mathrm{RMSE}} = \left(\frac{\int_0^\infty k_1^*(u)^2\,\mathrm{d}u}{(\int_0^\infty u^2 k_1^*(u)\,\mathrm{d}u)^2} \cdot \frac{\sigma^2_+(0) + \sigma^2_-(0)}{dn4M^2}\right)^{1/5}. \tag{18}$$

Similar to the discussion in Section 4.1, this expression is similar to the optimal bandwidth definition derived under pointwise asymptotics (Imbens and Kalyanaraman, 2012), except that $4M^2$ is replaced with $(f''_+(0) - f''_-(0))^2$, which gives infinite bandwidth if the second derivatives at zero are equal in magnitude and of opposite sign. Consequently, the critique in Section 4.1 applies to this bandwidth as well.

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

$$\hat{T}(h^*_{\mathrm{RMSE}};k) \pm \mathrm{cv}_{1-\alpha}(1/2) \cdot \mathrm{sd}(\hat{T}(h^*_{\mathrm{RMSE}};k)). \tag{19}$$

One can use the critical value $\mathrm{cv}_{1-\alpha}(\overline{\mathrm{bias}}(\hat{T}(h^*_{\mathrm{RMSE}};k))/\,\mathrm{sd}(\hat{T}(h^*_{\mathrm{RMSE}};k)))$ based on the finite-sample bias-sd ratio. The choice of M, and computation of the standard error and h^*_{RMSE} are similar to the conditional mean case, and are discussed in Appendix A.

6.1 Empirical illustration

To illustrate the implementation of feasible versions of the CIs (19), 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 estimate the discontinuity in mortality rates, Ludwig and Miller (2007) use a uniform kernel[5] and consider bandwidths equal to 9, 18, and 36. This yields point estimates equal to -1.90, -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 (22) 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 Appendix A.1 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 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. Thus, unless one is confident that the smoothness parameter M is very

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

small, the results are not significant at 5% level.

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.

Appendix A Implementation details

This section discusses implementation details. We focus on the nonparametric regression setting of Section 3, with additional details for the RD setting of Section 6 where relevant.

A.1 Rule of thumb for M

Fan and Gijbels (1996) suggest using a global polynomial estimate of order $p+2$ to estimate the pointwise-in-f optimal bandwidth. We apply this approach to estimate M, thereby giving an analogous rule-of-thumb estimate of the minimax optimal bandwidth. To calibrate M, let $\breve{f}(x)$ be the global polynomial estimate of order $p+2$, and let $[x_{\min}, x_{\max}]$ denote the support of x_i. We define the rule-of-thumb choice of M to be the supremum of $|\breve{f}^{(p)}(x)|$ over $x \in [x_{\min}, x_{\max}]$. The resulting minimax RMSE optimal bandwidth is given by (14) with the rule-of-thumb M plugged in. In contrast, the rule-of-thumb bandwidth proposed by Fan and Gijbels (1996, Chapter 4.2) plugs in $\breve{f}^{(p)}(0)$ to the pointwise-in-f optimal bandwidth formula (15).

We conjecture that, for any M, the resulting CI will be asymptotically honest over the intersection of $\mathcal{F}(M)$ and an appropriately defined set of regression functions that formalizes the notion that the pth derivative in a neighborhood of zero is bounded by the maximum pth derivative of the global $p+2$ polynomial approximation to the regression function. We leave this question, as well as optimality of the resulting CI for this class, for future research.

In the RD setting in Section 6, the regression function has a discontinuity at a point on the support of x_i, which is normalized to zero. In this case, we define $\breve{f}^{(p)}(x)$ to be the global polynomial estimate of order $p+2$ in which the intercept and all coefficients are allowed to be different on either side of the discontinuity (that is, we add the indicator $I(x_i > 0)$ for observation i being above the discontinuity, as well as interactions of this indicator with each order of the polynomial). We then take the supremum of $|\breve{f}^{(p)}(x)|$ over $x \in [x_{\min}, x_{\max}]$ as our rule-of-thumb choice of M, as before.

A.2 Standard errors

Because the local linear estimator $\hat{T}_1(h^*_{\text{RMSE}}; k)$ is a weighted least squares estimator, one can consistently 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

$$\widehat{\text{se}}(h,k)^2 = \sum_{i=1}^{n} w_1^n(x_i; h, k)^2 \hat{\sigma}^2(x_i), \qquad \hat{\sigma}^2(x_i) = \frac{J}{J+1}\left(y_i - \frac{1}{J}\sum_{j=1}^{J} y_{j(i)}\right)^2, \tag{20}$$

for some fixed (small) $J \geq 1$, where $j(i)$ denotes the jth closest observation to i. 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. In the RD setting, the standard error can be estimated using the same formula with the corresponding weight function $w^{(n)}(x_i; h, k)$ for the local linear RD estimator, except that the jth closest observation to i, $j(i)$, is only taken among units with the same sign of the running variable.

A.3 Computation of $h^*_{\mathbf{rmse}}$

For h^*_{RMSE}, there are two feasible choices. The first is to use a plug-in estimator that replaces the unknown quantities d, and $\sigma^2(0)$, by some consistent estimates. Alternatively, one can directly minimize the finite-sample RMSE over the bandwidth h, which for $\mathcal{F}_{\text{Höl},2}(M)$ takes the form

$$\text{RMSE}(h)^2 = \frac{M^2}{4}\left(\sum_{i=1}^{n} w_1^n(x_i; h, k) x_i^2\right)^2 + \sum_{i=1}^{n} w_1^n(x_i; h, k)^2 \sigma^2(x_i). \tag{21}$$

For $\mathcal{F}_{\text{T},2}(M)$, the sum $\sum_{i=1}^n w_1^n(x_i; h, k) x_i^2$ is replaced by $\sum_{i=1}^n |w_1^n(x_i; h) x_i^2|$. Since $\sigma^2(x_i)$ is typically unknown, one can replace it by an estimate $\hat{\sigma}^2(x_i) = \hat{\sigma}^2$ that assumes homoscedasticity of the variance function. For the RD setting with the class $\mathcal{F}_{\text{RD}}(M)$, the finite-sample RMSE takes the form

$$\text{RMSE}(h)^2 = \frac{M^2}{4}\left(\sum_{i=1}^{n} \left(w_+^n(x_i; h) + w_-^n(x_i; h)\right) x_i^2\right)^2 + \sum_{i=1}^{n} \left(w_+^n(x_i)^2 + w_-^n(x_i)^2\right) \sigma^2(x_i), \tag{22}$$

and h^*_{RMSE} can be chosen to minimize this expression with $\sigma^2(x)$ replaced with the estimate $\hat{\sigma}^2(x_i) = \hat{\sigma}^2_+(0) 1(x \geq 0) + \hat{\sigma}^2_-(0) 1(x < 0)$, where $\hat{\sigma}^2_+(0)$ and $\hat{\sigma}^2_-(0)$ are some preliminary variance estimates for observations above and below the cutoff.

This method was considered previously in Armstrong and Kolesár (2017), who show that the resulting confidence intervals will be asymptotically valid and equivalent to the infeasible CI based on minimizing the infeasible RMSE (21). This method has the advantage that it avoids having to estimate d, and it can also be shown to work when the covariates are discrete.

Appendix B Proofs of theorems in Section 2

B.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,

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

where $t_n = h_n^{\gamma_b-\gamma_s}B(k)/(n^{-1/2}S(k))$. This contradicts the display above.

B.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, uniformly over $f \in \mathcal{F}$, $\mathrm{sd}_n \,/\, \mathrm{sd}_f(\hat{T}(h^*_{\mathrm{RMSE}};k)) \to 1$ and $\mathrm{sd}_n \,/\widehat{\mathrm{se}}(h^*_{\mathrm{RMSE}};k) \xrightarrow{p} 1$. Note that, by Theorem 2.1 and the calculations above,

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

and similarly for $\tilde{R}_{\mathrm{FLCI},\alpha-\varepsilon}(\hat{T}(h^*_{\mathrm{RMSE}};k))$. 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 $f \in \mathcal{F}$,

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

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

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

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

References

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

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

Armstrong, T. B. and Kolesár, M. (2017). Optimal inference in a class of regression models. *Econometrica*, forthcoming.

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *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.

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

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2017). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, forthcoming.

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

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

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

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

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

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1997). Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics*, 49(1):79–99.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. 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.

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

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

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

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

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

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.

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

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

Sun, Y. (2005). Adaptive estimation of the regression discontinuity model. Technical report. Uinivesity of California, San Diego.

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

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

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

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

Notes: Critical values $\text{cv}_{1-\alpha}(t)$ and $\text{cv}_{1-\alpha}(\sqrt{1/r-1})$, correspond to the $1-\alpha$ quantiles of the $|N(t,1)|$ and $|N(\sqrt{1/r-1},1)|$ distribution, where b is the worst-case bias-standard deviation ratio, and r is the exponent r. For $b \geq 2$, $\text{cv}_{1-\alpha}(b) \approx b + z_{1-\alpha/2}$ up to 3 decimal places for these values of $1-\alpha$.

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

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

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

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}—worst-case bias-standard deviation ratio;

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	0.60
RBC	$h=\hat{h}_{\text{CE}}$, $b=\hat{b}_{\text{CE}}$	0.030	0.041	0.45	85.9	0.85	0.059	0.045	0.34	72.5	0.75
RBC	$h=b=\hat{h}^*_{\text{RMSE},M=2}$	0.001	0.061	0.36	94.5	1.27	0.002	0.061	0.36	94.5	1.00
RBC	$h=b=\hat{h}^*_{\text{RMSE},M=6}$	0.000	0.076	0.23	94.2	1.58	0.000	0.075	0.23	94.2	1.24
Conventional	$\hat{h}^*_{\text{PT,ROT}}$	0.032	0.036	0.56	76.6	0.76	0.049	0.046	0.31	77.4	0.76
FLCI, $M=2$	$\hat{h}^*_{\text{RMSE},M=2}$	0.021	0.043	0.36	94.9	1.00	0.065	0.043	0.36	75.2	0.79
FLCI, $M=6$	$\hat{h}^*_{\text{RMSE},M=6}$	0.009	0.054	0.23	96.6	1.25	0.028	0.053	0.23	94.7	0.99
FLCI, $M=\hat{M}_{\text{ROT}}$	$\hat{h}^*_{\text{RMSE},M=\hat{M}_{\text{ROT}}}$	0.008	0.056	0.22	95.6	1.29	0.010	0.069	0.14	96.3	1.28
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	0.57
RBC	$h=\hat{h}_{\text{CE}}$, $b=\hat{b}_{\text{CE}}$	0.028	0.040	0.49	87.5	0.83	0.074	0.041	0.44	54.3	0.68
RBC	$h=b=\hat{h}^*_{\text{RMSE},M=2}$	0.002	0.061	0.36	94.5	1.27	0.006	0.061	0.36	94.4	1.00
RBC	$h=b=\hat{h}^*_{\text{RMSE},M=6}$	0.000	0.076	0.23	94.2	1.58	0.000	0.075	0.23	94.2	1.24
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, $M=2$	$\hat{h}^*_{\text{RMSE},M=2}$	0.020	0.043	0.36	95.1	1.00	0.061	0.043	0.36	78.1	0.79
FLCI, $M=6$	$\hat{h}^*_{\text{RMSE},M=6}$	0.009	0.054	0.23	96.6	1.25	0.028	0.053	0.23	94.7	0.99
FLCI, $M=\hat{M}_{\text{ROT}}$	$\hat{h}^*_{\text{RMSE},M=\hat{M}_{\text{ROT}}}$	0.013	0.048	0.30	94.3	1.13	0.020	0.059	0.20	94.3	1.08
Design 3											
RBC	$h=\hat{h}^*_{\text{PT}}$, $b=\hat{b}^*_{\text{PT}}$	-0.043	0.035	0.77	75.7	0.72	-0.122	0.035	0.74	10.2	0.58
RBC	$h=\hat{h}_{\text{CE}}$, $b=\hat{b}_{\text{CE}}$	-0.026	0.040	0.49	88.2	0.83	-0.063	0.043	0.43	64.6	0.71
RBC	$h=b=\hat{h}^*_{\text{RMSE},M=2}$	-0.002	0.061	0.36	94.5	1.27	-0.007	0.061	0.36	94.4	1.00
RBC	$h=b=\hat{h}^*_{\text{RMSE},M=6}$	0.000	0.076	0.23	94.2	1.58	0.000	0.075	0.23	94.2	1.24
Conventional	$\hat{h}^*_{\text{PT,ROT}}$	-0.032	0.033	0.72	74.7	0.69	-0.065	0.042	0.39	62.0	0.69
FLCI, $M=2$	$\hat{h}^*_{\text{RMSE},M=2}$	-0.020	0.043	0.36	95.0	1.00	-0.060	0.043	0.36	78.1	0.79
FLCI, $M=6$	$\hat{h}^*_{\text{RMSE},M=6}$	-0.009	0.054	0.23	96.5	1.25	-0.027	0.053	0.23	94.7	0.99
FLCI, $M=\hat{M}_{\text{ROT}}$	$\hat{h}^*_{\text{RMSE},M=\hat{M}_{\text{ROT}}}$	-0.010	0.052	0.25	95.6	1.22	-0.013	0.065	0.16	96.1	1.21

Legend: $E[h]$—average (over Monte Carlo draws) bandwidth; SE—average standard error, Cov—coverage of CIs (in %); RL—relative (to optimal FLCI) length.
Bandwidth (bw) descriptions: $\hat{h}^*_{\text{PT}}$—plugin estimate of pointwise MSE optimal 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. (2017) is used for all four bws. $\hat{h}^*_{\text{RMSE},M=2}$, $\hat{h}^*_{\text{RMSE},M=6}$—RMSE optimal bw, assuming $M=2$, and $M=6$, respectively. $\hat{h}^*_{\text{PT,ROT}}$—Fan and Gijbels (1996) rule of thumb; $\hat{h}^*_{\text{RMSE},M=\hat{M}_{\text{ROT}}}$—RMSE optimal bw, using ROT for M. See Appendix A for detailed description of $\hat{h}^*_{\text{RMSE},M=2}$, $\hat{h}^*_{\text{RMSE},M=6}$, $\hat{h}^*_{\text{RMSE},M=\hat{M}_{\text{ROT}}}$, and $\hat{h}^*_{\text{PT,ROT}}$. 50,000 Monte Carlo draws.

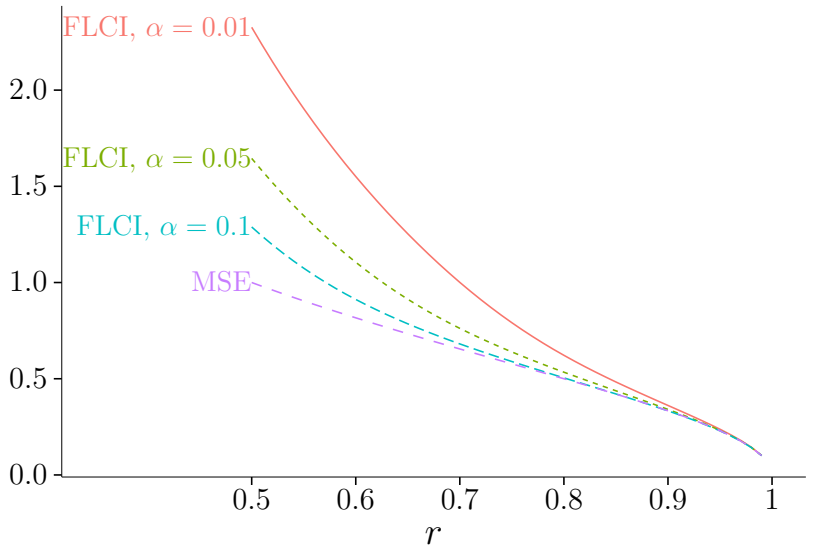

Figure 1: Optimal worst-case bias-standard deviation ratio for fixed length CIs (FLCI), and maximum MSE (MSE) performance criteria.

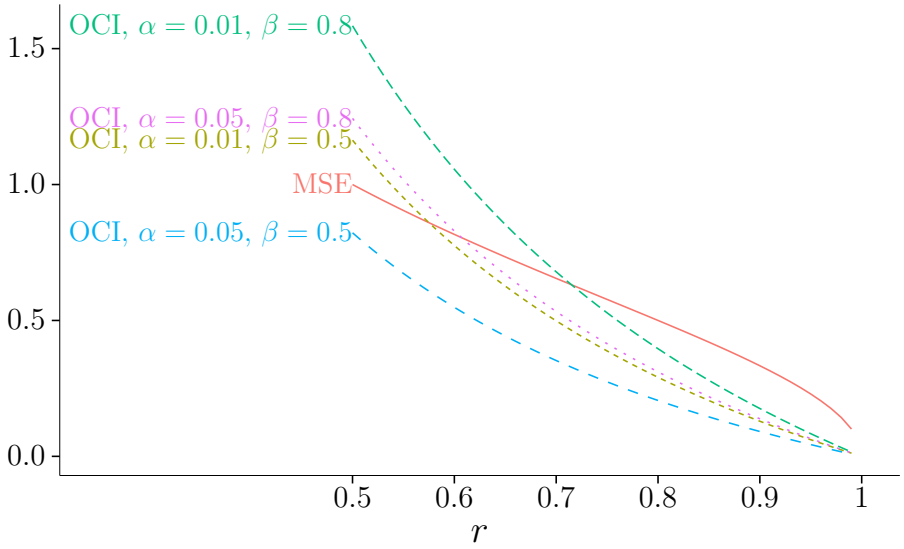

Figure 2: Optimal worst-case bias-standard deviation ratio for one-sided CIs (OCI), and maximum MSE (MSE) performance criteria.

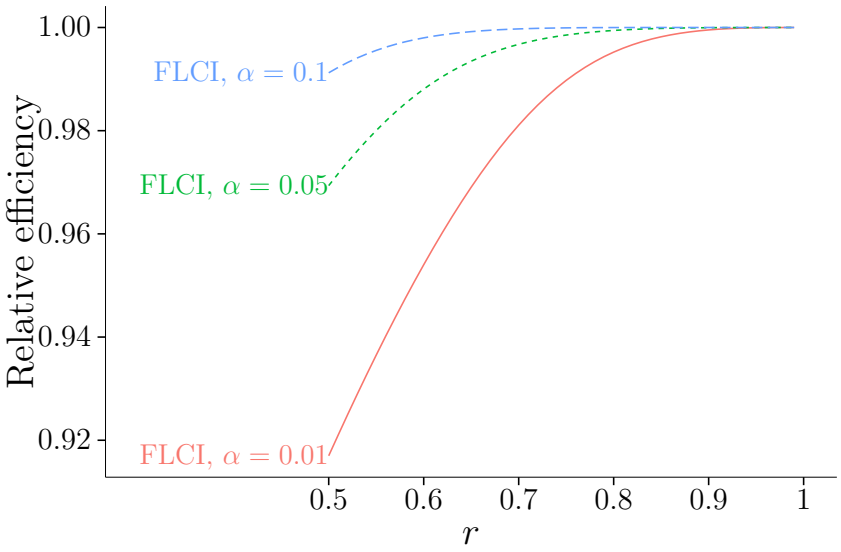

Figure 3: Efficiency of fixed-length CIs based on minimax MSE bandwidth relative to fixed-length CIs based on optimal bandwidth.

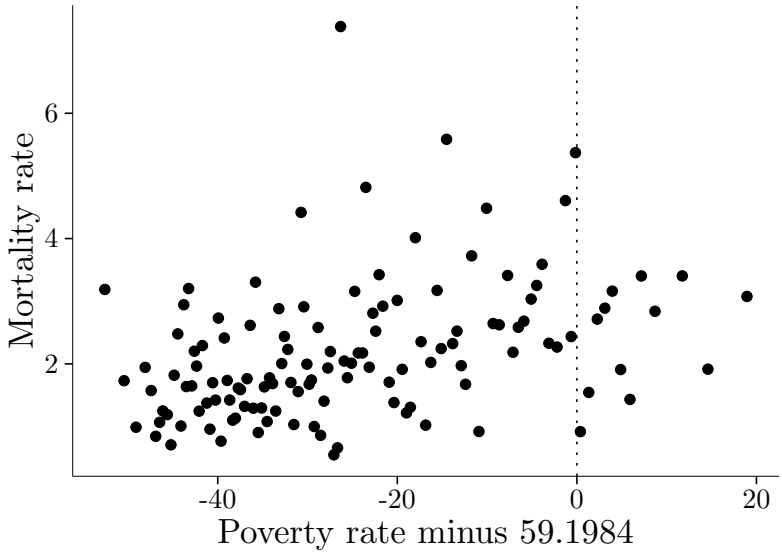

Figure 4: Average county mortality rate per 100,000 for children aged 5–9 over 1973–83 due to causes addressed as part of Head Start's health services (labeled "Mortality rate") plotted against poverty rate in 1960 relative to 300th poorest county. Each point corresponds to an average for 25 counties.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

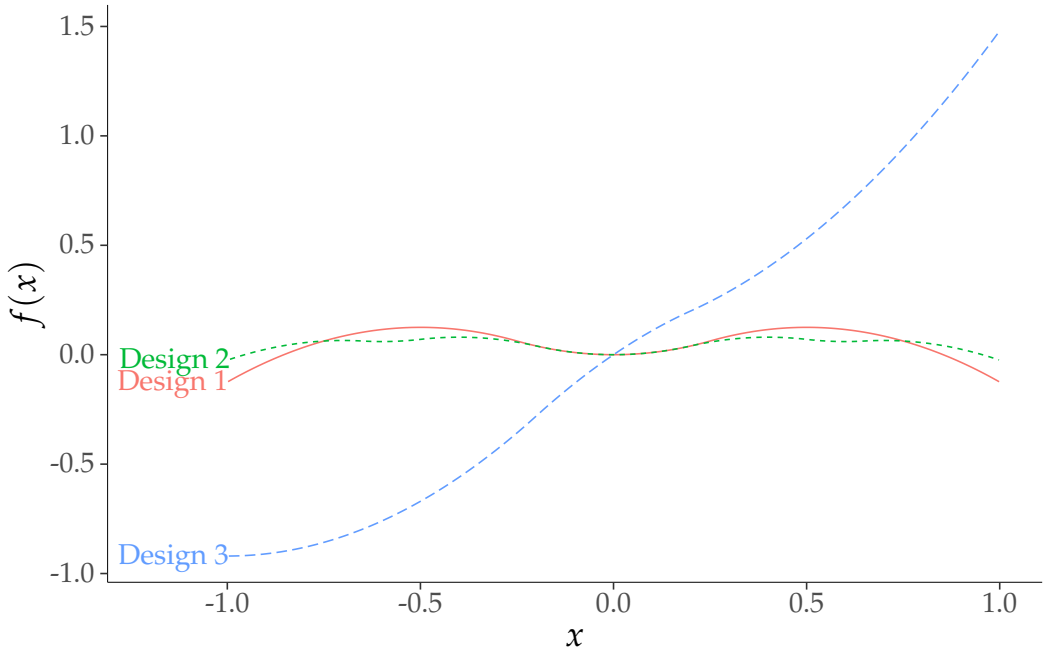

Figure 5: Monte Carlo simulation Designs 1–3, and $M = 2$.