Supplemental Materials for "Simple and Honest Confidence Intervals in Nonparametric Regression"

Timothy B. Armstrong* Michal Kolesár†

Yale University Princeton University

September 4, 2019

These supplemental materials contain further appendices and additional tables and figures. Appendix B verifies our regularity conditions for some examples, and includes proofs of the results in Section 3.2. Appendix C discusses two additional applications: estimation of density at a point, and estimating a bidder valuation in first price auctions. Appendix D contains additional details for the applications in Section 3. Appendix E presents a formal analysis of the rule-of-thumb choice of M proposed in Section 3.3.

Appendix B Verification of regularity conditions

We verify the main condition (4) in some applications. Appendix B.1 gives sufficient conditions for (4) which do not require convergence of moments. Appendix B.2 shows that (4) holds in the Gaussian white noise model under a mild extension of conditions in Donoho and Low (1992). Thus, the results apply to estimating, among other things, a function or one of its derivatives evaluated at a given point, when the function is observed in the white noise model. By equivalence results in Brown and Low (1996) and Nussbaum (1996), our results also apply when the function of interest is a density or conditional mean. Appendix B.3 verifies (4) directly for local polynomial estimators in the nonparametric regression setting, and Appendix B.4 verifies it for in the fuzzy RD application.

*email: timothy.armstrong@yale.edu

†email: mkolesar@princeton.edu

B.1 Sufficient conditions for main regularity condition

This appendix gives sufficient conditions for the main condition (4). In particular, we show that a version of (2) stated in terms of convergence in distribution, rather than convergence of moments, suffices for (4) for the FLCI and OCI criteria, and for a truncated version of the RMSE criterion. Such conditions are appropriate for functionals that involve smooth nonlinear transformations, which preserve convergence in distribution but may not preserve convergence of moments: we show in Appendix B.1.1 that a version of the delta method can be used to verify our conditions in such cases.

As in the main text, we consider a general setup where, for each n (which typically denotes sample size), data are drawn from some distribution P_f, which also implicitly depends on n, for some f. Let $\mathcal{F}_n \subseteq \mathcal{F}$ be a sequence of function classes, and let $T : \mathcal{F} \to \mathbb{R}$. Let $\hat{T} = \hat{T}(h; k)$ be a sequence of estimators indexed implicitly by n, and by a kernel k and bandwidth $h = h_n$, which also depends on n. The function class $\mathcal{F}_n$ is indexed by a sequence of constants M_n.

To make concise statements about uniform-in-f convergence, we introduce some additional notation. For a random variable $W_{n,f}$ indexed by the sample size n and the distribution f, we use $W_{n,f_n} \overset{d}{\underset{f_n}{\to}} \mathcal{L}$ to denote that the distribution of W_{n,f_n} converges in distribution to $\mathcal{L}$ under the sequence f_n. When this holds for all sequences $f_n \in \mathcal{F}_n$ for some sequence of sets $\mathcal{F}_n$, we write $W_{n,f} \overset{d}{\underset{\mathcal{F}_n}{\to}} \mathcal{L}$, and we say that $W_{n,f}$ converges in distribution to $\mathcal{L}$ uniformly over $\mathcal{F}_n$. When the limiting law $\mathcal{L}$ is a point mass at some constant a, we write $W_{n,f_n} \overset{p}{\underset{f_n}{\to}} a$ and when the convergence holds for all $f_n \in \mathcal{F}_n$, we write $W_{n,f} \overset{p}{\underset{\mathcal{F}_n}{\to}} a$ and say that $W_{n,f}$ converges in probability to a uniformly over $\mathcal{F}_n$.

We make the following assumption on the estimators $\hat{T}(h; k)$. This assumption is similar to the condition (2) in the main text, but uses convergence in distribution rather than convergence of moments.

Assumption B.1. *For some sequences of random variables $Z_{n,h,f}$ and $b_{n,h,f}$, we have*

$$\hat{T}(h; k) = T(f) + h^{\gamma_b} M_n b_{n,h,f} + h^{\gamma_s} n^{-1/2} Z_{n,h,f}$$

where, for some sequence of constants $b^*_{n,h,f}$ and some $S(k)$ and $B(k)$, $|b_{n,h,f} - b^*_{n,h,f}| \overset{p}{\underset{\mathcal{F}_n}{\to}} 0$ and

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}_n} b^*_{n,h,f} = B(k), \quad \lim_{n\to\infty} \inf_{f\in\mathcal{F}_n} b^*_{n,h,f} = -B(k), \quad Z_{n,h,f} \overset{d}{\underset{\mathcal{F}_n}{\to}} N(0, S(k)^2).$$

We verify our main condition (4) for a class of performance criteria constructed as follows. Given a loss function $\ell \colon \mathbb{R} \to \mathbb{R}^+$, let $\tilde{r}_\ell(b_0, s) = E_{Z\sim N(0,1)}\ell(b_0 + sZ)$ denote the risk of an estimator that's normally distributed with standard deviation s and bias b_0. Let

$$\tilde{\rho}_\ell(b,s) = \sup_{b_0\in[-b,b]} \tilde{r}_\ell(b_0,s), \quad \text{and} \quad \tilde{R}_{\ell,\alpha}(b,s) = \inf\left\{\chi\colon \tilde{\rho}_\ell(b\chi^{-1}, s\chi^{-1}) \leq \alpha\right\}$$

denote its worst-case risk over the all biases bounded by b in absolute value, and the smallest scaling of the worst-case bias and the standard deviation such that its worst-case risk is bounded by α. Similarly, for an estimator $\hat{T}$ of $T(f)$, let

$$\rho_{\ell,\chi}\left(\hat{T};\mathcal{F}_n\right) = \sup_{f\in\mathcal{F}_n} E_f \ell\left(\chi^{-1}\left(\hat{T}-T(f)\right)\right), \quad \text{and} \quad R_{\ell,\alpha}(\hat{T};\mathcal{F}_n) = \inf\left\{\chi : \rho_{\ell,\chi}\left(\hat{T};\mathcal{F}_n\right) \leq \alpha\right\}.$$

Note that if we set $\ell_{\text{FLCI}}(x) = \mathrm{I}\{|x| > 1\}$, then $R_{\ell_{\text{FLCI}},\alpha}$ and $\tilde{R}_{\ell_{\text{FLCI}},\alpha}$ yield the performance criteria $R_{\text{FLCI},\alpha}$ and $\tilde{R}_{\text{FLCI},\alpha}$ as defined in the main text. Similarly, $R_{\ell_{\text{RMSE}},1}$ and $\tilde{R}_{\ell_{\text{RMSE}},1}$, where $\ell_{\text{RMSE}}(x) = x^2$, give the performance criteria R_{RMSE} and $\tilde{R}_{\text{RMSE}}$ given in the main text.

To cover performance criteria such as OCI which are constructed from requirements on multiple loss functions, we use the following construction. Let $\ell_1,\ldots,\ell_m$ be loss functions and let $\alpha_1,\ldots,\alpha_m$ be given. Let $\lambda\colon (0,\infty)^m \to (0,\infty)$ be continuous and homogeneous of degree one (i.e. it satisfies $\lambda(ax) = a\lambda(x)$ for any $a > 0$). If $m = 1$, one can take λ to be the identity function. Let

$$R(\hat{T}(h;k)) = \lambda(R_{\ell_1,\alpha_1}(\hat{T}(h;k)),\ldots,R_{\ell_m,\alpha_m}(\hat{T}(h;k))),$$
$$\tilde{R}(b,s) = \lambda(R_{\ell_1,\alpha_1}(b,s),\ldots,R_{\ell_m,\alpha_m}(b,s)).$$

Note that since $\tilde{R}_{\ell_j,\alpha_j}(tb,ts) = t\inf\{t^{-1}\chi\colon \tilde{\rho}_{\ell_j}(tb\chi^{-1}, ts\chi^{-1}) \leq \alpha_j\} = t\tilde{R}_{\ell_j,\alpha_j}(b,s)$, $\tilde{R}$ satisfies (5). To show how this generalization covers the OCI criterion $R_{\text{OCI},\alpha,\beta}$ defined in the main text, define $\ell_+(x) = \mathrm{I}\{x > 1\}$ and $\ell_-(x) = \mathrm{I}\{x < -1\}$. Then $R_{\ell_+,\alpha}(\hat{T};\mathcal{F}_n)$ is the smallest value of χ_+ such that $[\hat{T}-\chi_+,\infty)$ is a one-sided CI with coverage $1-\alpha$, since $\rho_{\ell_+,\chi_+}(\hat{T};\mathcal{F}_n) = \sup_{f\in\mathcal{F}_n} P_f(\chi_+^{-1}(\hat{T}-T(f)) > 1) = \sup_{f\in\mathcal{F}_n} P_f(\hat{T}-\chi_+ > T(f))$ gives the probability of not covering $T(f)$. The worst-case β quantile of excess length of this CI is the smallest value of χ_- such that $\inf_{f\in\mathcal{F}_n} P_f(T(f)-\hat{T}+\chi_+ \leq \chi_-) \geq \beta$, or equivalently, $\rho_{\ell_-,\chi_--\chi_+}(\hat{T};\mathcal{F}_n) = \sup_{f\in\mathcal{F}_n} P_f(T(f)-\hat{T} > \chi_- - \chi_+) = \sup_{f\in\mathcal{F}_n} P_f(T(f)-\hat{T}+\chi_+ > \chi_-) \leq 1-\beta$. Thus, the worst case β-quantile of excess length of a one-sided CI based on $\hat{T}$ is given by $R_{\ell_+,\alpha}(\hat{T};\mathcal{F}_n) + R_{\ell_-,1-\beta}(\hat{T};\mathcal{F}_n) = R_{\text{OCI},\alpha,\beta}(\hat{T})$. Similarly, $\tilde{R}_{\ell_+,\alpha}(b,s) + \tilde{R}_{\ell_-,1-\beta}(b,s)$ gives the criterion $\tilde{R}_{\text{OCI},\alpha,\beta}(b,s)$ as defined in the main text.

We make the following assumption on each of the loss functions ℓ.

Assumption B.2. *(i) $\ell : \mathbb{R} \to [0,\infty)$ is bounded, weakly decreasing on $(-\infty,0)$ and weakly increasing on $(0,\infty)$, and continuous almost everywhere, and there does not exist a constant function that is almost everywhere equal to ℓ. (ii) $\tilde{b} \mapsto \tilde{r}_\ell(\tilde{b},s)$ is quasiconvex.*

For symmetric loss functions, part (ii) follows from part (i) by Anderson's lemma.

It is immediate that the loss functions ℓ_+, ℓ_-, and ℓ_{FLCI} satisfy this assumption. The loss ℓ_{RMSE}, on the other hand, does not satisfy this assumption because it is unbounded. However, note that, for any $c > 0$, Assumption B.2 holds for the loss function $\ell_c(x) = \min\{x^2, c^2\}$. Since $\lim_{c\to\infty} R_{\ell_c,1}(\hat{T}, \mathcal{F}_n) = R_{\ell_{\text{RMSE}},1}(\hat{T}, \mathcal{F}_n)$, and $\lim_{c\to\infty} \tilde{R}_{\ell_c,1}(b, s) = \tilde{R}_{\ell_{\text{RMSE}},1}(b, s)$, we may interpret this criterion as a truncated version of RMSE.

Theorem B.1. *Let h_n be a sequence with*

$$0 < \liminf_n h_n(nM^2)^{1/[2(\gamma_b-\gamma_s)]} \leq \limsup_n h_n(nM^2)^{1/[2(\gamma_b-\gamma_s)]} < \infty. \tag{S1}$$

Suppose that $\hat{T}(h;k)$ satisfies Assumption B.1 for the sequence $h = h_n$. Let $R(\hat{T}(h;k))$ and $\tilde{R}(b,s)$ be given above, where $\ell_1, \dots, \ell_m$ are loss functions satisfying Assumption B.2, and suppose that $\tilde{R}_{\ell_j,\alpha_j}(b,s) > 0$ for all $b \geq 0$ and $s > 0$ for $j = 1, \dots, m$. Then (4) holds for R and $\tilde{R}$. Furthermore, if $b_{n,h,f} = b^*_{n,h,f}$, $E_f Z_{n,h,f} = 0$ and $E_f Z^2_{n,h,f} \to S(k)^2$ uniformly over $f \in \mathcal{F}_n$, then $\sup_{f\in\mathcal{F}} E_f(\hat{T}(h;k) - T(f)) = -\inf_{f\in\mathcal{F}} E_f(\hat{T}(h;k) - T(f))(1 + o(1)) = h^{\gamma_b} B(k)(1 + o(1))$, and $\text{sd}_f(\hat{T}(h;k)) = h^{\gamma_s} n^{-1/2} S(k)(1 + o(1))$ uniformly over $f \in \mathcal{F}_n$, and (4) holds with R and $\tilde{R}$ given by R_{RMSE} and $\tilde{R}_{RMSE}$.

The theorem implies that if Assumption B.1 holds for bandwidth sequences h_n satisfying Eq. (S1), minimizing the criterion $\lim_{c\to\infty} \lim_{n\to\infty} n^{r/2} M^{r-1} R_{\ell_c}(\hat{T}(h;k))$ discussed in footnote 4 in the main text, where ℓ_c is the truncated squared error loss defined above, is equivalent to minimizing the asymptotic RMSE:

$$\begin{aligned}\lim_{c\to\infty} \lim_{n\to\infty} n^{r/2} M^{r-1} R_{\ell_c}(\hat{T}(h;k)) &= S(k)^r B(k)^{1-r} \lim_{c\to\infty} t^{r-1} \tilde{R}_{\ell_c}(t, 1) \\ &= S(k)^r B(k)^{1-r} t^{r-1} \tilde{R}_{\ell_{\text{RMSE}},1}(t, 1).\end{aligned}$$

Thus, under this criterion, the optimal bandwidth is given by h^*_{RMSE}.

To prove Theorem B.1, we first note some properties of loss and risk functions in our setup. Note that, under Assumption B.2, $E\ell(W_n) \to EW$ for any sequence of random variables $W_n \xrightarrow{d} W$ such that W is continuously distributed (this follows from the continuous mapping theorem and the fact that ℓ is bounded). This also implies that $\tilde{r}_\ell(\tilde{b}, s)$ is continuous in $\tilde{b}$ and s (since $s_n Z + \tilde{b}_n \xrightarrow{d} sZ + b$ for $Z \sim N(0,1)$ and $\tilde{b}_n \to \tilde{b}$, $s_n \to s$). Also, by part (ii), $\tilde{\rho}_\ell(\chi^{-1}b, \chi^{-1}s) = \max_{\tilde{b}\in\{-b,b\}} E_{Z\sim N(0,1)} \ell\left(\chi^{-1}\left(Zs + b\right)\right)$, which is continuous in (b, s, χ), and is strictly decreasing in χ (since $\ell(\chi^{-1}t)$ is weakly decreasing in χ for each t, and, for any $0 < \chi < \tilde{\chi}$, there is a positive measure set of values of t such that $\ell(\chi^{-1}t) > \ell(\tilde{\chi}^{-1}t)$ for t on this set). This implies that $\tilde{R}_{\ell,\alpha}(b, s)$, taken as a function of α, is the inverse of the strictly increasing function $\chi \mapsto \tilde{\rho}_\ell(b\chi^{-1}, s\chi^{-1})$. Since convergence of a sequence of strictly increasing functions to a continuous, strictly increasing function implies convergence of their inverse, this implies that $\tilde{R}_{\ell,\alpha}(b,s)$ is continuous in (b,s).

We will use the following lemma.

Lemma B.1. *Let b,s be given. Suppose that ℓ satisfies Assumption B.2. Suppose that, for any sequence f_n, there exists $\tilde{b} \in [-b,b]$ and a subsequence along which $a_n(\hat{T} - T(f_n)) \underset{f_n}{\overset{d}{\to}} N(\tilde{b}, s^2)$. Furthermore, suppose that there exists a sequence f_n such that $a_n(\hat{T} - T(f_n)) \underset{f_n}{\overset{d}{\to}} N(b, s^2)$, and a sequence f_n such that $a_n(\hat{T} - T(f_n)) \underset{f_n}{\overset{d}{\to}} N(-b, s^2)$. Then $\lim_{n\to\infty} \rho_{\ell,\chi/a_n}(\hat{T}; \mathcal{F}_n) = \tilde{\rho}_\ell(\chi^{-1}b, \chi^{-1}s)$ and $\lim_{n\to\infty} a_n R_{\ell,\alpha}(\hat{T}; \mathcal{F}_n) = \tilde{R}_{\ell,\alpha}(b,s)$.*

Proof. To show $\limsup_n \rho_{\ell,\chi/a_n}(\hat{T}; \mathcal{F}_n) \leq \tilde{\rho}_\ell(\chi^{-1}b, \chi^{-1}s)$ it suffices to show that, for every sequence f_n, there is a subsequence along which $E_{f_n}\ell\left(a_n\chi^{-1}\left(\hat{T} - T(f_n)\right)\right)$ converges to a constant that is no greater than $\tilde{\rho}_\ell(\chi^{-1}b, \chi^{-1}s)$. By assumption, there exists a $\tilde{b} \in [-b,b]$ and a subsequence along which $a_n(\hat{T} - T(f_n)) \underset{f_n}{\overset{d}{\to}} N(\tilde{b}, s^2)$, which, under the assumptions on the loss function, implies $E_{f_n}\ell\left(a_n\chi^{-1}\left(\hat{T} - T(f_n)\right)\right) \to \tilde{r}_\ell(\chi^{-1}\tilde{b}, \chi^{-1}s) \leq \rho_\ell(\chi^{-1}b, \chi^{-1}s)$ along this subsequence. To show that this lim sup is a limit and the inequality is an equality, note that, letting f_n be a sequence such that $a_n(\hat{T} - T(f_n)) \underset{f_n}{\overset{d}{\to}} N(b, s^2)$, we have $\rho_{\ell,\chi/a_n}(\hat{T}; \mathcal{F}_n) \geq E_{f_n}\ell\left(\chi^{-1}\left(\hat{T} - T(f_n)\right)\right) \to \tilde{r}_\ell(\chi^{-1}b, \chi^{-1}s)$. Similarly, taking a sequence for which the limiting distribution is $N(-b, s^2)$, we have $\liminf_n \rho_{\ell,\chi/a_n}(\hat{T}; \mathcal{F}_n) \geq \tilde{r}_\ell(-\chi^{-1}b, \chi^{-1}s)$. Noting that, under Assumption B.2, $\rho_\ell(\chi^{-1}b, \chi^{-1}s)$ is equal to either $\tilde{r}_\ell(\chi^{-1}\tilde{b}, \chi^{-1}s)$ or $\tilde{r}_\ell(-b\chi^{-1}\tilde{b}, \chi^{-1}s)$ (or both), it now follows that $\liminf_n \rho_{\ell,\chi/a_n}(\hat{T} : \mathcal{F}_n) \geq \tilde{\rho}_\ell(\chi^{-1}b, \chi^{-1}s)$. Thus, $\lim_{n\to\infty} \rho_{\ell,\chi/a_n}(\hat{T} : \mathcal{F}_n) = \tilde{\rho}_\ell(\chi^{-1}b, \chi^{-1}s)$.

To derive the limit of $R_{\ell,\alpha}(\hat{T}; \mathcal{F}_n)$, first note that $\rho_{\ell,\chi}(\hat{T}; \mathcal{F}_n)$ is weakly decreasing in χ for any $\chi > 0$ for each n, since $\ell(\chi^{-1}t)$ is weakly decreasing in χ for all t under Assumption B.2. Also, $\tilde{\rho}_\ell(\chi^{-1}b, \chi^{-1}s)$ is strictly decreasing in χ. Thus, for $\chi > \tilde{R}_{\ell,\alpha}(b,s)$, we have $\tilde{\rho}_\ell(\chi^{-1}b, \chi^{-1}s) < \alpha$ so that, for large enough n, we have $\rho_{\ell,\chi/a_n}(\hat{T}; \mathcal{F}_n) < \alpha$ for all $\tilde{\chi} \geq \chi$, which implies $R_{\ell,\alpha}(\hat{T}; \mathcal{F}_n) \leq \chi/a_n$. Similarly, for $\chi < \tilde{R}_{\ell,\alpha}(b,s)$, we have $\tilde{\rho}_\ell(\chi^{-1}b, \chi^{-1}s) > \alpha$ so that, for large enough n, we have $\rho_{\ell,\chi/a_n}(\hat{T}; \mathcal{F}_n) > \alpha$ for all $\tilde{\chi} \leq \chi$, which implies $R_{\ell,\alpha}(\hat{T}; \mathcal{F}_n) \geq \chi/a_n$. Thus, for any $\eta > 0$, we have, for large enough n, $\tilde{R}_{\ell,\alpha}(b,s) - \eta \leq a_n R_{\ell,\alpha}(\hat{T}; \mathcal{F}_n) \leq \tilde{R}_{\ell,\alpha}(b,s) + \eta$. It follows that $a_n R_{\ell,\alpha}(\hat{T}; \mathcal{F}_n) \to \tilde{R}_{\ell,\alpha}(b,s)$. □

We are now ready to prove Theorem B.1.

Proof of Theorem B.1. The last statement (regarding convergence of standard deviation and worst-case bias and RMSE) follows immediately from the assumptions. To show (4) for R and $\tilde{R}$ constructed from loss functions $\ell_1, \ldots, \ell_m$ satisfying Assumption B.2, it suffices to show that, for every subsequence, there exists a further subsequence along which $R(\hat{T}(h;k)) =$

$\tilde{R}(h^{\gamma_b}MB(k), h^{\gamma_s}n^{-1/2}S(k))(1+o(1))$. By the conditions on h_n, we can choose this subsequence so that $h_n(nM_n^2)^{1/[2(\gamma_b-\gamma_s)]} \to h_\infty$ for some $h_\infty > 0$.

Along this subsequence, we have

$$h_n^{\gamma_b}M_n = h_\infty^{\gamma_b}(nM_n^2)^{-\gamma_b/[2(\gamma_b-\gamma_s)]}M_n(1+o(1)) = h_\infty^{\gamma_b}M_n^{1-r}n^{-r/2}(1+o(1))$$

and

$$h_n^{\gamma_s}n^{-1/2} = h_\infty^{\gamma_s}(nM_n^2)^{-\gamma_s/[2(\gamma_b-\gamma_s)]}n^{-1/2}(1+o(1)) = h_\infty^{\gamma_s}n^{-r/2}M_n^{1-r}(1+o(1)).$$

Thus, on this subsequence, the conditions of Lemma B.1 hold with $a_n = M_n^{r-1}n^{r/2}$, $b = h_\infty^{\gamma_b}B(k)$ and $s = h_\infty^{\gamma_s}S(k)$, so that, for each $j = 1, \dots, m$,

$$M_n^{r-1}n^{r/2}R_{\ell_j,\alpha_j}(\hat{T}(h_n;k);\mathcal{F}_n) \to \tilde{R}_{\ell_j,\alpha_j}(h_\infty^{\gamma_b}B(k), h_\infty^{\gamma_s}S(k)).$$

Also, on this subsequence, using homogeneity and continuity of $\tilde{R}_{\ell,\alpha}$,

$$\begin{aligned}&M_n^{r-1}n^{r/2}\tilde{R}_{\ell_j,\alpha_j}(h_n^{\gamma_b}M_nB(k), h_n^{\gamma_s}n^{-1/2}S(k))\\ &\quad= \tilde{R}_{\ell_j,\alpha_j}(M_n^{r-1}n^{r/2}h_n^{\gamma_b}M_nB(k), M_n^{r-1}n^{r/2}h_n^{\gamma_s}n^{-1/2}S(k)) \to \tilde{R}_{\ell_j,\alpha_j}(h_\infty^{\gamma_b}B(k), h_\infty^{\gamma_s}S(k)).\end{aligned}$$

Combining this with the previous display and using homogeneity of the function λ, it follows that (4) holds along this subsequence, which gives the result. □

B.1.1 Delta method

Let $\mathcal{F}_n \subseteq \mathcal{F}$ be a sequence of function classes, and let $L\colon \mathcal{F} \to \mathbb{R}^m$. We are interested in a parameter $T(f) = \phi(L(f))$, where $\phi\colon \mathbb{R}^m \to \mathbb{R}$. To cover cases where ϕ may be nonlinear, we assume that $\mathcal{F}_n$ is localized around a particular value L^* in the range of L:

$$L(f_n) \to L^* \text{ for all sequences } f_n \in \mathcal{F}_n.$$

This localization of the parameter space plays a similar role to local asymptotic efficiency results in parametric and regular semiparametric settings (see, for example, Theorem 8.11 in van der Vaart, 1998).

We now show that, if $\hat{L}(h;k)$ satisfies a multivariate version of Assumption B.1 and ϕ is smooth, then Assumption B.1 holds for $\hat{T}(h;k) = \phi(\hat{L}(h;k))$, with $B(k)$ and $S(k)$ defined below. This is essentially a version of the delta method applied to our setup.

Assumption B.3. *The function ϕ is continuously differentiable at L^*, with Jacobian matrix*

$\phi'(L)$ and, for some sequences of random vectors $Z_{n,h,f}$ and $b_{n,h,f}$, we have

$$\hat{L}(h;k) = L(f) + h^{\gamma_b} M_n b_{n,h,f} + h^{\gamma_s} n^{-1/2} Z_{n,h,f},$$

where, for a uniformly bounded sequence of constant vectors $b^*_{n,h,f} \in \mathbb{R}^m$ and some $\Sigma(k)$ and $B(k)$, $|b_{n,h,f} - b^*_{n,h,f}| \xrightarrow[\mathcal{F}_n]{p} 0$ and

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}_n} \phi'(L^*) b^*_{n,h,f} = B(k), \quad \lim_{n\to\infty} \inf_{f\in\mathcal{F}_n} \phi'(L^*) b^*_{n,h,f} = -B(k), \quad Z_{n,h,f} \xrightarrow[\mathcal{F}_n]{d} N(0, \Sigma(k)).$$

Theorem B.2. *Suppose that Assumption B.3 holds, and put $S(k)^2 = \phi'(L^*)\Sigma(k)\phi'(L^*)'$. Then, if $h^{\gamma_b} M_n \to 0$ and $h^{\gamma_s} n^{-1/2} \to 0$, Assumption B.1 holds for $\hat{T}(h;k) = \phi(\hat{L}(h;k))$.*

Proof. First, note that the conditions on the bandwidth imply $\hat{L} \xrightarrow[\mathcal{F}_n]{p} L^*$. Then, by a Taylor expansion, for some $\tilde{L} = \tilde{L}(\hat{L}, L(f))$ on the line segment between $\hat{L}$ and $L(f)$, we have

$$\begin{aligned}\phi(\hat{L}) - \phi(L(f)) &= \phi'(\tilde{L})[\hat{L} - L(f)] \\ &= \phi'(\tilde{L})[h^{\gamma_b} M_n b_{n,h,f} + h^{\gamma_s} n^{-1/2} Z_{n,h,f}] = h^{\gamma_b} M_n \tilde{b}_{n,h,f} + h^{\gamma_s} n^{-1/2} \tilde{Z}_{n,h,f},\end{aligned}$$

where $\tilde{Z}_{n,h,f} = \phi'(\tilde{L}) Z_{n,h,f} \xrightarrow[\mathcal{F}_n]{d} N(0, S(k)^2)$ by the continuous mapping theorem and $\tilde{b}_{n,h,f} = \phi'(\tilde{L}) b_{n,h,f}$ satisfies $|\tilde{b}_{n,h,f} - \tilde{b}^*_{n,h,f}| = |\phi'(\tilde{L}) b_{n,h,f} - \phi'(L^*) b^*_{n,h,f}| \xrightarrow[\mathcal{F}_n]{p} 0$ where $\tilde{b}^*_{n,h,f} = \phi'(L^*) b^*_{n,h,f}$. Thus, Assumption B.1 holds with $\tilde{b}_{n,h,f}$ playing the role of $b_{n,h,f}$, and $\tilde{b}^*_{n,h,f}$ playing the role of $b^*_{n,h,f}$. □

If the function class $\mathcal{F}_n$ places separate restrictions on each mapping $x \mapsto f_j(x)$ for $j = 1, \ldots, m$, then the set of limits of the biases $b^*_{n,h,f}$ will take the form $[-\bar{B}_1(k), \bar{B}_1(k)] \times \cdots \times [-\bar{B}_m(k), \bar{B}_m(k)]$. In this case, the limiting worst-case bias takes the form

$$B(k) = \sum_{j=1}^{m} |\phi'_j(L^*) \bar{B}_j(k)|. \tag{S2}$$

Note that, while Theorem B.2 shows that Assumption B.1 is preserved under smooth nonlinear transformations, such a statement does not hold for a version of this assumption stated in terms of moments, rather than weak convergence. For such a result, one needs to either use truncation or place stronger conditions on the class of estimators. This is analogous to parametric and regular semiparametric settings such as instrumental variables, in which the asymptotic variance may only be finite if defined in terms of convergence in distribution.

B.2 Gaussian white noise model

The approximation (4) holds as an exact equality (i.e. with the $o(1)$ term equal to zero) for the RMSE, OCI, and FLCI criteria in the Gaussian white noise model whenever the problem renormalizes in the sense of Donoho and Low (1992). We show this below, using notation taken mostly from that paper. Consider a Gaussian white noise model

$$Y(dt) = (Kf)(t)\,dt + (\sigma/\sqrt{n})W(dt), \quad t \in \mathbb{R}^d.$$

We are interested in estimating the linear functional $T(f)$ where f is known to be in the class $\mathcal{F} = \{f\colon J_2(f) \leq C\}$ where $J_2(f) : \mathcal{F} \to \mathbb{R}$ and $C \in \mathbb{R}$ are given. Let $\mathcal{U}_{a,b}$ denote the renormalization operator $\mathcal{U}_{a,b}f(t) = af(bt)$. Suppose that T, J_2, and the inner product are homogeneous: $T(\mathcal{U}_{a,b}f) = ab^{s_0}T(f)$, $J_2(\mathcal{U}_{a,b}f) = ab^{s_2}J_2(f)$ and $\langle K\mathcal{U}_{a_1,b}f, K\mathcal{U}_{a_2,b}g\rangle = a_1a_2b^{2s_1}\langle Kf, Kg\rangle$. These are the same conditions as in Donoho and Low (1992) except for the last one, which is slightly stronger since it must hold for the inner product rather than just the norm.

Consider the class of linear estimators based on a given kernel k:

$$\hat{T}(h;k) = h^{s_h}\int (Kk(\cdot/h))(t)\,dY(t) = h^{s_h}\int [K\mathcal{U}_{1,h^{-1}}k](t)\,dY(t)$$

for some exponent s_h to be determined below. The worst-case bias of this estimator is

$$\overline{\text{bias}}(\hat{T}(h;k)) = \sup_{J_2(f)\leq C} |T(f) - h^{s_h}\langle Kk(\cdot/h), Kf\rangle|\,.$$

Note that $J_2(f) \leq C$ iff. $f = \mathcal{U}_{h^{s_2},h^{-1}}\tilde{f}$ for some $\tilde{f}$ with $J_2(\tilde{f}) = J_2(\mathcal{U}_{h^{-s_2},h}f) = J_2(f) \leq C$. This gives

$$\begin{aligned}\overline{\text{bias}}(\hat{T}(h;k)) &= \sup_{J_2(f)\leq C} |T(\mathcal{U}_{h^{s_2},h^{-1}}f) - h^{s_h}\langle Kk(\cdot/h), K\mathcal{U}_{h^{s_2},h^{-1}}f\rangle| \\ &= \sup_{J_2(f)\leq C} \left|h^{s_2-s_0}T(f) - h^{s_h+s_2-2s_1}\langle Kk(\cdot), Kf\rangle\right|.\end{aligned}$$

If we set $s_h = -s_0 + 2s_1$ so that $s_2 - s_0 = s_h + s_2 - 2s_1$, the problem will renormalize, giving

$$\overline{\text{bias}}(\hat{T}(h;k)) = h^{s_2-s_0}\,\overline{\text{bias}}(\hat{T}(1;k)).$$

The variance does not depend on f and is given by

$$\begin{aligned}\text{var}_f(\hat{T}(h;k)) &= h^{2s_h}(\sigma^2/n)\langle K\mathcal{U}_{1,h^{-1}}k, K\mathcal{U}_{1,h^{-1}}k\rangle = h^{2s_h-2s_1}(\sigma^2/n)\langle Kk, Kk\rangle \\ &= h^{-2s_0+2s_1}(\sigma^2/n)\langle Kk, Kk\rangle.\end{aligned}$$

Thus, Eq. (2) holds with $\gamma_b = s_2 - s_0$, $\gamma_s = s_1 - s_0$,

$$B(k) = \overline{\text{bias}}(\hat{T}(1; k)) = \sup_{J_2(f) \leq C} |T(f) - \langle Kk, Kf \rangle|,$$

and $S(k) = \sigma \|Kk\|$ and with both $o(1)$ terms equal to zero. This implies that (4) holds with the $o(1)$ term equal to zero, since the estimator is normally distributed.

B.3 Local polynomial estimators in fixed design regression

This appendix proves Theorem 3.1 and Eq. (15) in Section 3.2.1.

We begin by deriving the worst-case bias of a general linear estimator

$$\hat{T} = \sum_{i=1}^{n} w(x_i) y_i$$

under Hölder and Taylor classes. For both $\mathcal{F}_{\text{T},p}(M)$ and $\mathcal{F}_{\text{Höl},p}(M)$ the worst-case bias is infinite unless $\sum_{i=1}^{n} w(x_i) = 1$ and $\sum_{i=1}^{n} w(x_i) x^j = 0$ for $j = 1, \dots, p-1$, so let us assume that $w(\cdot)$ satisfies these conditions. For $f \in \mathcal{F}_{\text{T},p}(M)$, we can write $f(x) = \sum_{j=0}^{p-1} x^j f^{(j)}(0)/j! + r(x)$ with $|r(x)| \leq M|x|^p/p!$. As noted by Sacks and Ylvisaker (1978), this gives the bias under f as $\sum_{i=1}^{n} w(x_i) r(x_i)$, which is maximized at $r(x) = M \operatorname{sign}(w(x)) |x|^p/p!$, giving $\overline{\text{bias}}_{\mathcal{F}_{\text{T},p}}(\hat{T}) = M \sum_{i=1}^{n} |w(x_i) x|^p/p!$.

For $f \in \mathcal{F}_{\text{Höl},p}(M)$, the $(p-1)$th derivative is Lipschitz and hence absolutely continuous. Furthermore, since $\sum_{i=1}^{n} w(x_i) = 1$ and $\sum_{i=1}^{n} w(x_i) x^j = 0$, the bias at f is the same as the bias at $x \mapsto f(x) - \sum_{j=0}^{p-1} x^j f^{(j)}(0)/j!$, so we can assume without loss of generality that $f(0) = f'(0) = \dots = f^{(p-1)}(0)$. This allows us to apply the following lemma.

Lemma B.2. *Let ν be a finite measure on $\mathbb{R}$ (with the Lebesgue σ-algebra) with finite support and let $w \colon \mathbb{R} \to \mathbb{R}$ be a bounded measurable function with finite support. Let f be $p-1$ times differentiable with bounded pth derivative on a set of Lebesgue measure 1 and with $f(0) = f'(0) = f''(0) = \dots = f^{(p-1)}(0) = 0$. Then*

$$\int_0^\infty w(x) f(x) \, d\nu(x) = \int_{s=0}^\infty \bar{w}_{p,\nu}(s) f^{(p)}(s) \, ds$$

and

$$\int_{-\infty}^0 w(x) f(x) \, d\nu(x) = \int_{s=-\infty}^0 \bar{w}_{p,\nu}(s) f^{(p)}(s) \, ds$$

where

$$\bar{w}_{p,\nu}(s) = \begin{cases} \int_{x=s}^\infty \frac{w(x)(x-s)^{p-1}}{(p-1)!} \, d\nu(x) & s \geq 0 \\ \int_{x=-\infty}^s \frac{w(x)(s-x)^{p-1}(-1)^p}{(p-1)!} \, d\nu(x) & s < 0. \end{cases}$$

Proof. By the Fundamental Theorem of Calculus and the fact that the first $p-1$ derivatives at 0 are 0, we have

$$f(x)=\int_{t_1=0}^{x}\int_{t_2=0}^{t_1}\cdots\int_{t_p=0}^{t_{p-1}} f^{(p)}(t_p)\,dt_p\cdots dt_2dt_1=\int_{s=0}^{x}\frac{f^{(p)}(s)(x-s)^{p-1}}{(p-1)!}\,ds.$$

Thus, by Fubini's Theorem,

$$\begin{aligned}\int_{x=0}^{\infty} w(x)f(x)\,d\nu(x)&=\int_{x=0}^{\infty} w(x)\int_{s=0}^{x}\frac{f^{(p)}(s)(x-s)^{p-1}}{(p-1)!}\,dsd\nu(x)\\&=\int_{s=0}^{\infty} f^{(p)}(s)\int_{x=s}^{\infty}\frac{w(x)(x-s)^{p-1}}{(p-1)!}\,d\nu(x)ds\end{aligned}$$

which gives the first display in the lemma. The second display in the lemma follows from applying the first display with $f(-x)$, $w(-x)$ and $\nu(-x)$ playing the roles of $f(x)$, $w(x)$ and $\nu(x)$. □

Applying Lemma B.2 with ν given by the counting measure that places mass 1 on each of the x_i's ($\nu(A)=\#\{i\colon x_i\in A\}$), it follows that the bias under f is given by $\int w(x)f(x)\,d\nu=\int\bar{w}_{p,\nu}(s)f^{(p)}(s)\,ds$. This is maximized over $f\in\mathcal{F}_{\text{Höl},p}(M)$ by taking $f^{(p)}(s)=M\operatorname{sign}(\bar{w}_{p,\nu}(s))$, which gives $\overline{\text{bias}}_{\mathcal{F}_{\text{Höl},p}(M)}(\hat{T})=M\int|\bar{w}_{p,\nu}(s)|\,ds$.

We collect these results in the following theorem.

Theorem B.3. *For a linear estimator* $\hat{T}=\sum_{i=1}^n w(x_i)y_i$ *such that* $\sum_{i=1}^n w(x_i)=1$ *and* $\sum_{i=1}^n w(x_i)x^j=0$ *for* $j=1,\ldots,p-1$,

$$\overline{\text{bias}}_{\mathcal{F}_{\text{T},p}(M)}(\hat{T})=M\sum_{i=1}^{n}|w(x_i)x|^p/p!\quad\textit{and}\quad\overline{\text{bias}}_{\mathcal{F}_{\text{Höl},p}(M)}(\hat{T})=M\int|\bar{w}_{p,\nu}(s)|\,ds$$

where $\bar{w}_{p,\nu}(s)$ *is as defined in Lemma B.2 with* ν *given by the counting measure that places mass 1 on each of the* x_i*'s.*

Note that, for $t>0$ and any q,

$$\begin{aligned}\int_{s=t}^{\infty}\overline{w}_{q,\nu}(s)\,ds&=\int_{s=t}^{\infty}\int_{x=s}^{\infty}\frac{w(x)(x-s)^{q-1}}{(q-1)!}\,d\nu(x)ds=\int_{x=t}^{\infty}\int_{s=t}^{x}\frac{w(x)(x-s)^{q-1}}{(q-1)!}\,dsd\nu(x)\\&=\int_{x=t}^{\infty}w(x)\left[\frac{-(x-s)^q}{q!}\right]_{s=t}^{x}d\nu(x)=\int_{x=t}^{\infty}\frac{w(x)(x-t)^q}{q!}\,d\nu(x)=\bar{w}_{q+1,\nu}(t).\end{aligned}\tag{S3}$$

Let us define $\bar{w}_{0,\nu}(x)=w(x)$, so that this holds for $q=0$ as well.

For the boundary case with $p=2$, the bias is given by (using the fact that the support of

ν is contained in $[0, \infty)$)

$$\int_0^\infty w(x)f(x)\,d\nu(x) = \int_0^\infty \bar{w}_{2,\nu}(x)f^{(2)}(x)\,dx \quad \text{where} \quad \bar{w}_{2,\nu}(s) = \int_{x=s}^\infty w(x)(x-s)\,d\nu(x).$$

For a local linear estimator based on a kernel with nonnegative weights and support $[-A, A]$, the equivalent kernel $w(x)$ is positive at $x = 0$ and negative at $x = A$ and changes signs once. From (S3), it follows that, for some $0 \le b \le A$, $\bar{w}_{1,\nu}(x)$ is negative for $x > b$ and nonnegative for $x < b$. Applying (S3) again, this also holds for $\bar{w}_{2,\nu}(x)$. Thus, if $\bar{w}_{2,\nu}(\tilde{s})$ were strictly positive for any $\tilde{s} > 0$, we would have to have $\bar{w}_{2,\nu}(s)$ nonnegative for $s \in [0, \tilde{s}]$. Since $\bar{w}_{2,\nu}(0) = \sum_{i=1}^n w(x_i)x_i = 0$, we have

$$0 < \bar{w}_{2,\nu}(0) - \bar{w}_{2,\nu}(\tilde{s}) = -\int_{x=0}^{\tilde{s}} w(x)(x-\tilde{s})\,d\nu(x)$$

which implies that $\int_{x=\underline{s}}^{\overline{s}} w(x)d\nu(x) < 0$ for some $0 \le \underline{s} < \overline{s} < \tilde{s}$. Since $w(x)$ is positive for small enough x and changes signs only once, this means that, for some $s^* \le \tilde{s}$, we have $w(x) \ge 0$ for $0 \le x \le s^*$ and $\int_{x=0}^{s^*} w(x)d\nu(x) > 0$. But this is a contradiction, since it means that $\bar{w}_{2,\nu}(s^*) = -\int_0^{s^*} w(x)(x-s^*)\,d\nu(x) < 0$. Thus, $\bar{w}_{2,\nu}(s)$ is weakly negative for all s, which implies that the bias is maximized at $f(x) = -(M/2)x^2$.

We now provide a proof for Theorem 3.1 by proving the result for a more general sequence of estimators of the form

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

where $\tilde{k}_n$ satisfies $\frac{1}{nh}\sum_{i=1}^n \tilde{k}_n(x_i/h) = 1$ and $\frac{1}{nh}\sum_{i=1}^n \tilde{k}_n(x_i/h)x_i^j = 0$ for $j = 1, \dots, p-1$. We further assume

Assumption B.4. *The support and magnitude of $\tilde{k}_n$ are bounded uniformly over n, and, for some $\tilde{k}$, $\sup_{u\in\mathbb{R}} |\tilde{k}_n(u) - \tilde{k}(u)| \to 0$.*

Theorem B.4. *Suppose Assumption 3.1 and Assumption B.4 hold. Then for any bandwidth sequence h_n such that $nh_n \to \infty$, $\liminf_n h_n(nM^2)^{1/(2p+1)} > 0$, and $\limsup_n h_n(nM^2)^{1/(2p+1)} < \infty$,*

$$\overline{\text{bias}}_{\mathcal{F}_{\mathrm{T},p}(M)}(\hat{T}) = \frac{Mh_n^p}{p!}\tilde{\mathcal{B}}_p^T(\tilde{k})(1+o(1)), \qquad \tilde{\mathcal{B}}_p^T(\tilde{k}) = d\int_{\mathcal{X}} |u^p\tilde{k}(u)|\,du$$

and

$$\overline{\text{bias}}_{\mathcal{F}_{\text{Höl},p}(M)}(\hat{T}) = \frac{Mh_n^p}{p!}\tilde{\mathcal{B}}_p^{\text{Höl}}(\tilde{k})(1+o(1)),$$

$$\tilde{\mathcal{B}}_p^{\text{Höl}}(\tilde{k}) = dp\int_{t=0}^\infty \left|\int_{u\in\mathcal{X},|u|\ge t} \tilde{k}(u)(|u|-t)^{p-1}\,du\right|\,dt.$$

If Assumption 3.2 holds as well, then

$$\mathrm{sd}(\hat{T}) = h_n^{-1/2} n^{-1/2} S(\tilde{k})(1 + o(1)),$$

where $S(\tilde{k}) = d^{1/2}\sigma(0)\sqrt{\int_{\mathcal{X}} \tilde{k}(u)^2\,du}$, *and* (4) *holds for the RMSE, FLCI and OCI performance criteria with* $\gamma_b = p$ *and* $\gamma_s = -1/2$.

Proof. Let K_s denote the bound on the support of $\tilde{k}_n$, and K_m denote the bound on the magnitude of $\tilde{k}_n$.

The first result for Taylor classes follows immediately since

$$\overline{\mathrm{bias}}_{\mathcal{F}_{\mathrm{T},p}(M)}(\hat{T}) = \frac{M}{p!} h^p \frac{1}{nh} \sum_{i=1}^{n} |\tilde{k}_n(x_i/h)||x_i/h|^p = \left(\frac{M}{p!} h^p d \int_{\mathcal{X}} |\tilde{k}(u)||u|^p\,du \right)(1 + o(1)),$$

where the first equality follows from Theorem B.3 and the second equality follows from the fact that for any function $g(u)$ that is bounded over u in compact sets,

$$\begin{aligned}
&\left| \frac{1}{nh} \sum_{i=1}^{n} \tilde{k}_n(x_i/h) g(x_i/h) - d \int_{\mathcal{X}} k(u) g(u)\,du \right| \\
&\leq \left| \frac{1}{nh} \sum_{i=1}^{n} \tilde{k}(x_i/h) g(x_i/h) - d \int_{\mathcal{X}} k(u) g(u)\,du \right| + \frac{1}{nh} \sum_{i=1}^{n} \left| \tilde{k}_n(x_i/h) g(x_i/h) - \tilde{k}(x_i/h) g(x_i/h) \right| \\
&\qquad \leq o(1) + \frac{1}{nh} \sum_{i=1}^{n} \mathrm{I}\{|x_i/h| \leq K_s\} \sup_{u \in [-K_s, K_s]} |g(u)| \cdot \sup_{u \in [-K_s, K_s]} |\tilde{k}_n(u) - \tilde{k}(u)| = o(1), \qquad \text{(S4)}
\end{aligned}$$

where the second line follows by triangle inequality, the third line by Assumption 3.1 applied to the first summand (with $x \mapsto \tilde{k}(x) g(x)$ playing the role of $g(\cdot)$ in Assumption 3.1), and the last equality follows by Assumption 3.1 applied to the first term, and Assumption B.4 applied to the last term.

For Hölder classes,

$$\overline{\mathrm{bias}}_{\mathcal{F}_{\mathrm{H\ddot{o}l},p}(M)}(\hat{T}(h; \tilde{k}_n)) = M \int |\bar{w}_{p,\nu}(s)|\,ds$$

by Theorem B.3 where $\bar{w}_{p,\nu}$ is as defined in that theorem with $w(x) = \frac{1}{nh}\tilde{k}_n(x/h)$. We have, for $s > 0$,

$$\begin{aligned}
\bar{w}_{p,\nu}(s) &= \int_{x \geq s} \frac{\frac{1}{nh}\tilde{k}_n(x/h)(x - s)^{p-1}}{(p-1)!}\,d\nu(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{\tilde{k}_n(x_i/h)(x_i - s)^{p-1}}{(p-1)!} \mathrm{I}\{x_i \geq s\} \\
&= h^{p-1} \frac{1}{nh} \sum_{i=1}^{n} \frac{\tilde{k}_n(x_i/h)(x_i/h - s/h)^{p-1}}{(p-1)!} \mathrm{I}\{x_i/h \geq s/h\}.
\end{aligned}$$

Thus, by Eq. (S4), for $t \geq 0$, $h^{-(p-1)}\bar{w}_{p,\nu}(t \cdot h) \to d \cdot \bar{w}_p(t)$, where

$$\bar{w}_p(t) = \int_{u \geq t} \frac{\tilde{k}(u)(u-t)^{p-1}}{(p-1)!}\, du$$

(i.e. $\bar{w}_p(t)$ denotes $\bar{w}_{p,\nu}(t)$ when $w = \tilde{k}$ and ν is the Lebesgue measure). Furthermore,

$$|h^{-(p-1)}\bar{w}_{p,\nu}(t \cdot h)| \leq \left[\frac{K_m}{nh} \sum_{i=1}^{n} \frac{\mathrm{I}\{0 \leq x_i/h \leq K_s\}(x_i/h)^{p-1}}{(p-1)!} \right] \cdot \mathrm{I}\{t \leq K_s\} \leq K_1 \cdot \mathrm{I}\{t \leq K_s\},$$

where the last inequality holds for some K_1 by Assumption 3.1. Thus,

$$M \int_{s \geq 0} |\bar{w}_{p,\nu}(s)|\, ds = h^p M \int_{t \geq 0} |h^{-(p-1)}\bar{w}_{p,\nu}(t \cdot h)|\, dt = h^p M \left[d \int_{t \geq 0} |\bar{w}_p(t)|\, dt \right] (1 + o(1))$$

by the Dominated Convergence Theorem. Combining this with a symmetric argument for $t \leq 0$ gives the result.

For the second part of the theorem, the variance of $\hat{T}$ doesn't depend on f, and equals

$$\operatorname{var}(\hat{T}) = \frac{1}{n^2h^2} \sum_{i=1}^{n} \tilde{k}_n(x_i/h)^2 \sigma^2(x_i) = \frac{1}{nh}\tilde{S}_n^2, \qquad \text{where} \quad \tilde{S}_n^2 = \frac{1}{nh} \sum_{i=1}^{n} \tilde{k}_n(x_i/h)^2 \sigma^2(x_i).$$

By the triangle inequality,

$$\begin{aligned} &\left| \tilde{S}_n^2 - d\sigma^2(0) \int_{\mathcal{X}} \tilde{k}(u)^2\, du \right| \\ &\qquad \leq \sup_{|x| \leq hK_s} \left| \tilde{k}_n(x/h)^2\sigma^2(x) - \tilde{k}(x/h)^2\sigma^2(0) \right| \cdot \frac{1}{nh} \sum_{i=1}^{n} \mathrm{I}\{|x_i/h| \leq K_s\} \\ &\qquad\qquad + \sigma^2(0) \left| \frac{1}{nh} \sum_{i=1}^{n} \tilde{k}(x_i/h)^2 - d \int_{\mathcal{X}} \tilde{k}(u)^2\, du \right| = o(1), \end{aligned}$$

where the equality follows by Assumption 3.1 applied to the second summand and the second term of the first summand, and Assumption 3.2 and Assumption B.4 applied to the first term of the first summand. This gives the second display in the theorem.

To show the last statement (verification of Eq. (4)), we note that the above arguments show that Assumption B.1 holds with $b_{n,h,f} = b^*_{n,h,f}$ equal to the bias of the estimator and $E_f Z^2_{n,h,f} \to S(k)$ uniformly over $\mathcal{F}$, so long as we can verify the uniform central limit theorem for $Z_{n,h,f} = (nh)^{1/2}[\hat{T} - E_f\hat{T}] = (nh)^{-1/2}\sum_{i=1}^{n} \tilde{k}_n(x_i/h)u_i$. By the conditions on the errors u_i, this follows from the Lindeberg central limit theorem so long as $\max_i[(nh)^{-2}k_n(x_i/u)]^2/(nh)^{-1} = \max_i nhk_n(x_i/u)/(nh) \to 0$. By uniform boundedness of the kernel k_n, this holds so long as

$nh \to \infty$. $\square$

The local polynomial estimator takes the form given above with

$$\tilde{k}_n(u) = e_1' \left(\frac{1}{nh} \sum_{i=1}^{n} k(x_i/h) m_q(x_i/h) m_q(x_i/h)' \right)^{-1} m_q(u)k(u).$$

If k is bounded with bounded support, then, under Assumption 3.1 this sequence satisfies Assumption B.4 with

$$\tilde{k}(u) = e_1' \left(d \int_{\mathcal{X}} k(u) m_q(u) m_q(u)' \, du \right)^{-1} m_q(u)k(u) = d^{-1} k_q^*(u),$$

where k_q^* is the equivalent kernel defined in Eq. (14). Theorem 3.1 and Eq. (15) then follow immediately by applying Theorem B.4 with this choice of $\tilde{k}_n$ and $\tilde{k}$.

B.4 Fuzzy RD

We consider the sequence of parameter spaces $\mathcal{F}_n \subseteq \mathcal{F}(M_1, M_2)$, such that $L(f_n) \to L^*$ for all sequences $f_n \in \mathcal{F}_n$. Here $L^* \in \mathbb{R}^2$ is a fixed vector such that $L_2^* \neq 0$. Let $M = M_1$, and suppose Assumption 3.1 holds (since the ratio M_1/M_2 is fixed, it suffices to verify the assumption for $M = M_1$). Assume also that the random variables $\{u_i\}_{i=1}^n$ are independent with $Eu_i = 0$, $\operatorname{var}(u_i) = \Omega(x_i)$ and $E(u_{1i}^2 + u_{2i}^2)^{1+\eta} \leq 1/\eta$ for some $\eta > 0$, and that the covariance function $\Omega(x)$ is left- and right- continuous at $x = 0$ with $\Omega_+(0) = \lim_{x \downarrow 0} \Omega(x) > 0$ and $\Omega_-(0) = \lim_{x \uparrow 0} \Omega(x) > 0$. It then follows by adapting arguments in the proof of Theorem 3.1 that for any bandwidth sequence h_n with $nh_n \to \infty$ and $0 < \liminf_n h_n(nM^2)^{1/(2p+1)} < \limsup_n h_n(nM^2)^{1/(2p+1)} < \infty$,

$$\hat{L}(h;k) = L(f) + h^2 \begin{pmatrix} M_1 b_{n,h,f,1}^* \\ M_2 b_{n,h,f,2}^* \end{pmatrix} + \frac{1}{\sqrt{nh}} Z_{n,h,f},$$

where $Z_{n,h,f}$ converges in distribution to $N(0, \Sigma(k))$ uniformly over $\mathcal{F}_n$ with

$$\Sigma(k) = \int_0^\infty k_1^*(u)^2 \, \mathrm{d}u \cdot (\Omega_+(0) + \Omega_-(0))/d,$$

and $b_{n,h,f,j}^* = \sum_{i=1}^n (w_+(x_i) + w_-(x_i)) f_j(x_i)/M_j$ for $j = 1, 2$, and the limits of these biases lie in the set $[\tilde{B}(k), -\tilde{B}(k)]^2$, where $\tilde{B}(k) = \int_0^\infty u^2 k_1^*(u) \, \mathrm{d}u$. From (S2), we obtain that Assumption B.3 holds with $\gamma_b = 2$, $\gamma_s = -1/2$, and

$$B(k) = -(|\phi_1'(L^*)| + M_2/M_1 |\phi_2'(L^*)|) \int_0^\infty u^2 k_1^*(u) \, \mathrm{d}u = -\frac{1 + M_2/M_1 |L_1^*/L_2^*|}{|L_2^*|} \int_0^\infty u^2 k_1^*(u) \, \mathrm{d}u.$$

Thus, by Theorem B.2, condition (4) holds for FLCI, OCI, and truncated RMSE with

$$S(k)^2 = \frac{\int_0^\infty k_1^*(u)^2 \,\mathrm{d}u}{d} \frac{\varsigma_+^2(0; L_1^*/L_2^*) + \varsigma_+^2(0; L_1^*/L_2^*)}{(L_2^*)^2},$$

where $\varsigma^2(x;T) = (1, -T)\Omega(x)(1, -T)'$, $\varsigma_+^2(0;T) = \lim_{x\downarrow 0} \varsigma^2(x;T)$, and $\varsigma_-^2(0;T) = \lim_{x\uparrow 0} \varsigma^2(x;T)$.

The expressions for $\operatorname{avar}(\hat{T}(h;k))$ and $\overline{\operatorname{abias}}(\hat{T}(h;k))$ in the main text then follow by observing that $\sum_{i=1}^n \tilde{w}^n(x_i;h,k)^2 \phi'(L(f))\Omega(x_i)\phi'(L(f))' = S(k)/nh(1+o(1))$, and $(|\phi_1'(L(f))|M_1 + |\phi_2'(L(f))|M_2) \sum_{i=1}^n \tilde{w}^n(x_i;h,k)/2 = M_1 h^2 B(k)(1+o(1))$.

Appendix C Additional applications

This appendix considers additional applications not considered in the main text, using the sufficient conditions from Appendix B.1. Appendix C.1 verifies our conditions in the density setting, and Appendix C.2 applies these results to a problem in the auctions literature.

C.1 Density estimation

Consider estimating a density at a point, which we normalize to 0. We observe $\{X_i\}_{i=1}^n$ iid with density f on the intersection of $\mathcal{X}$ and some neighborhood of 0, where either $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = [0, \infty)$. We are interested in $T(f) = f(0)$. Let $\hat{T} = \hat{T}(h;k) = \frac{1}{nh}\sum_{i=1}^n k(X_i/h)$ be a kernel estimate where k is a kernel with $\int_{\mathcal{X}} k(u)\,du = 1$ and finite support. Let $\mathcal{F} = \mathcal{F}(M)$ denote the Hölder class $\mathcal{F}_{\text{Höl},p}(M)$ or Taylor class $\mathcal{F}_{\text{T},p}(M)$ of order p, as defined in the paper. Assume that the kernel k satisfies $\int_{\mathcal{X}} u^j k(u)\,du = 0$ for $j = 1, \dots, p-1$. Let $f^* > 0$ be given, and let a_n be a sequence converging to zero more slowly than any polynomial. Let $\mathcal{F}(M, [-a, a])$ denote the class for which the Hölder or Taylor condition is imposed only for $x \in [-a, a] \cap \mathcal{X}$, and let $\mathcal{F}_n = \mathcal{F}(M_n; [-a_n, a_n]) \cap \{f : |f(x) - f^*| \le a_n \text{ all } x \in [-a_n, a_n] \cap \mathcal{X},\ f(x) \ge 0 \text{ all } x,\ \int f(x)\,dx = 1\}$.

We show that (4) holds for the performance criteria considered in the main text by verifying Assumption B.1. This gives a generalization of the results in Sacks and Ylvisaker (1981), who consider RMSE optimal kernels in Taylor classes, to performance criteria other than RMSE, and to cover Hölder classes in addition to Taylor classes. Note that $\mathcal{F}_n$ localizes the parameter space around a density with $T(f) = f^*$, similar to Appendix B.1.1. This differs slightly from Sacks and Ylvisaker (1981), who consider a fixed parameter space $\mathcal{F}$ which only places an upper bound f^* on $f(0)$. However, the result given below is essentially the same, since the worst-case risk over this class is taken in a shrinking neighborhood of f^* (i.e. the worst-case risk is the same as in our setup). Also, note that we only impose the Hölder or Taylor condition in the set $[-a_n, a_n]$, although we would obtain the same result if we did not impose this condition so long as M_n increases slowly enough so that the function can be extended to satisfy the smoothness condition outside of $[-a_n, a_n]$.

Theorem C.1. *For any bandwidth sequence with $h_n \to 0$, $h_n^p M_n \to 0$, $nh_n \to \infty$ and*

$$0 < \liminf_n h_n (nM^2)^{1/(2p-1)} \leq \limsup_n h_n (nM^2)^{1/(2p-1)} < \infty,$$

the kernel density estimator satisfies Assumption B.1 with $S(k) = \sqrt{f^* \int_{\mathcal{X}} k(u)^2\, du}$, $B(k)$ given in Theorem 3.1 and with $\gamma_b = p$ and $\gamma_s = -1/2$. In particular, (4) holds for the FLCI and OCI criteria. Furthermore, we can take $b_{n,h,f} = b^*_{n,h,f}$ to be nonrandom, and $E_f Z_{n,h,f} = 0$ and $E_f Z^2_{n,h,f} \to S(k)$ uniformly over $\mathcal{F}_n$, so that (4) holds for the RMSE criterion.

Proof. We have

$$\hat{T}(h;k) = T(f) + h^p M b_{n,h,f} + (nh)^{-1/2} Z_{n,h,f} \tag{S5}$$

where

$$b_{n,h,f} = h^{-p} M^{-1} [E_f \hat{T}(h;k) - T(f)] = h^{-p} M^{-1} \frac{1}{h} \int_{\mathcal{X}} k(x/h)[f(x) - f(0)]\, dx$$

is nonrandom and can be taken to be equal to $b^*_{n,h,f}$, and

$$Z_{n,h,f} = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} [k(X_i/h) - E_f k(X_i/h)].$$

Once h_n is small enough relative to a_n and f^*, the set of possible biases for the class $\mathcal{F}_n$ will be the same as for the Taylor or Hölder class $\mathcal{F}(M)$, without the additional local restriction of $f(x)$ for x near zero, or the restriction that f be a density (note, in particular, that, letting C be a bound on the support of the kernel k, the bias depends only on $f(x)$ for x in $[-Ch_n, Ch_n]$, and that the first $p-1$ derivatives of f at zero can be taken to be equal to zero without loss of generality, so that, for any function f satisfying the Hölder or Taylor condition, $f(x)$ is bounded from below by $f^* - a_n - \tilde{C} M_n h_n^p$ on this set for some constant $\tilde{C}$; this function can then be extrapolated so that it is positive on $[-a_n, a_n]$ while maintaining the Hölder or Taylor condition, and then defined outside of $[-a_n, a_n]$ so that it integrates to one), so that

$$\{b_{n,h,f} : f \in \mathcal{F}_n\} = \left\{ h^{-p} M^{-1} \frac{1}{h} \int_{\mathcal{X}} k(x/h)[f(x) - f(0)]\, dx : f \in \mathcal{F}(M) \right\}.$$

By the renormalization property of $\mathcal{F}$ ($f \in \mathcal{F}(1)$ iff. $x \mapsto h^p M f(x/h)$ is in $\mathcal{F}(M)$), the set in the above display remains the same if h and M are each replaced by 1. Thus, the expressions for asymptotic bias derived in Theorem 3.1 holds exactly with $\gamma_b = p$ and $B(k)$ given in Theorem 3.1

(with k playing the role of the equivalent kernel, k_q^*). For the variance, we have

$$\mathrm{var}_f(Z_{n,h,f}) = \frac{1}{h}\int_{\mathcal{X}} k(x/h)^2 f(x)\,dx - \frac{1}{h}\left[\int_{\mathcal{X}} k(x/h)f(x)\,dx\right]^2.$$

The second term converges to 0 uniformly over $\mathcal{F}_n$, and the first term converges to $f^*\int_{\mathcal{X}} k(u)^2\,du$ uniformly over $\mathcal{F}_n$. To verify the Lindeberg condition for asymptotic normality, note that $\frac{1}{nh}\sum_{i=1}^n E_f K(X_i/h)^2\,\mathrm{I}\{K(X_i/h)^2 \geq \varepsilon nh\} \to 0$ uniformly over $f \in \mathcal{F}_n$ since $nh \to \infty$. □

C.2 First price auctions

Our results for density estimation and nonparametric regression can be combined with the delta method (Theorem B.2) to verify our conditions for nonlinear functions of densities and nonparametric regression functions evaluated at finitely many points. To illustrate, we consider a setting from the auctions literature involving a nonlinear function of a density.

Guerre et al. (2000) consider the problem of recovering valuations from bids in a first price auction setting. Here, we consider a simple version of their setting with no covariates, and the same number of bidders in each auction. We observe n total bids from symmetric independent private value sealed bid auctions with $I > 1$ bidders each, with independent valuations. The bids $\{X_i\}_{i=1}^n$ are then iid and, letting f denote their density, the valuation for a bidder with bid $X_i = x$ is given by

$$\xi(x; f, I) = x + \frac{1}{I-1}\frac{\int_{-\infty}^x f(t)\,dt}{f(x)}$$

(Equation (3) in Guerre et al., 2000). Consider the problem of estimating $T(f) = \xi(x_0; f, I)$ at a particular point x_0. Let $\mathcal{F}_{GPV,n}$ be defined in the same way as the class $\mathcal{F}_n$ defined in Appendix C.1 with $\mathcal{X} = \mathbb{R}$, but with an additional local restriction on the cumulative distribution function (CDF) $\int_{-\infty}^x f(t)\,dt$: $\mathcal{F}_{GPV,n} = \mathcal{F}_n \cap \{f : |\int_{-\infty}^x f(t)\,dt - F^*| \leq a_n\}$ where $F^* \in (0,1)$ is given.

Let $\hat{L}(h;k) = (\hat{L}_1(h;k), \hat{L}_2(h,k)) = \left(\frac{1}{n}\sum_{i=1}^n \mathrm{I}\{X_i \leq x_0\}, \frac{1}{nh}\sum_{i=1}^n k((X_i - x_0)/h)\right)$, where k is a kernel satisfying the conditions in Appendix C.1 and h satisfies the conditions of Theorem C.1 for some p. Let $\phi(L) = x_0 + \frac{1}{I-1}\frac{L_1}{L_2}$. Then a plug-in estimator of $T(f)$ is given by $\hat{T}(h;k) = \phi(\hat{L}(h;k))$. To verify (4), we verify Assumption B.3. First, note that, by a slight generalization of Theorem C.1, $\hat{L}_2(h;k)$ satisfies (S5), where $b_{n,h,f}$ is nonrandom and, for large enough n, ranges over the set $[-B_2(k), B_2(k)]$, with $B_2(k)$ given by $B(k)$ in Theorem 3.1, and with $Z_{n,h,f}$ converging to a $N(0, S_2(k))$ distribution uniformly over $\mathcal{F}_{GPV,n}$, where $S_2(k) = \sqrt{f^*\int k(u)\,du}$. (This follows from the arguments in Theorem C.1 along with the observation that the local restriction on $\int_{-\infty}^x f(t)\,dt$ does not restrict the set of possible biases $b_{n,h,f}$ for large enough n.) Also, $\hat{L}_1(h;k)$ satisfies $\hat{L}_1(h;k) = L_1(f) + h^{\gamma_b} M_n b_{n,h,f,1} + h^{\gamma_s} n^{-1/2} Z_{n,h,f,1}$

with $\gamma_s = -1/2$, where $b_{n,h,f,1} = 0$ and $Z_{n,h,f,1} = n^{1/2}h^{-\gamma_s}\left(\hat{L}_1(h;k) - L_1(h;k)\right)$ converges in probability to zero uniformly over $\mathcal{F}_{GPV,n}$. Thus, Assumption B.3 holds with $b_{n,h,f}$ ranging over the set $\{0\} \times [-B_2(k), B_2(k)]$ and with $\Sigma(k) = \left(\begin{smallmatrix} 0 & 0 \\ 0 & S_2(k) \end{smallmatrix}\right)$ and $\phi'(L^*) = \frac{1}{I-1}[\frac{1}{f^*}, -\frac{F^*}{f^*}]$. It follows that (4) holds for the FLCI and OCI criteria, with $\gamma_s = -1/2$ and $\gamma_b = p$, $B(k) = B_2(k)\frac{F^*}{(I-1)f^*}$, and $S(k) = S_2(k)\frac{F^{*2}}{(I-1)^2 f^{*2}}$. Note, however, that, since a density estimator appears in the denominator of the estimator of $T(f)$, the RMSE may not even be finite, and so truncation will be needed to apply our results to the RMSE criterion.

We note that the class $\mathcal{F}_{GPV,n}$ places assumptions conditions directly on the bid distribution, and does not incorporate additional restrictions that may arise from the assumption that f arises from an equilibrium in a first price auction model. We leave for future research whether such restrictions place sharper bounds on the bias, as well as the question of deriving primitive conditions on the value distribution for our smoothness assumptions on the bid distribution. Such questions are addressed by Guerre et al. (2000), although they focus on a slightly different setting, since they consider rate optimality in the supremum norm for estimation of the value distribution (rather than asymptotic constants for estimation of the function $\xi(x; f, I)$ at a given point x_0).

Appendix D Additional details for applications

This appendix gives additional details for applications in Section 3. Appendix D.1 calculates the efficiency gain from using different bandwidths on either side of the cutoff in sharp RD. Appendix D.2 gives details of optimal kernel calculations discussed in Section 3.2.1. Appendix D.3 gives the kernels constants $\int_{\mathcal{X}} k_q^*(u)^2 \, \mathrm{d}u$, and $\mathcal{B}_{p,q}(k)$ for selected kernels.

D.1 Regression discontinuity with different bandwidths on either side of the cutoff

We consider a slightly more general setup than that considered in Section 3.2.2. Consider estimating a parameter $T(f)$, $f \in \mathcal{F}$, using a class of estimators $\hat{T}(h_+, h_-; k)$ indexed by two bandwidths h_- and h_+. Suppose that the worst-case (over $\mathcal{F}$) performance of $\hat{T}(h_+, h_-; k)$ according to a given criterion satisfies

$$R(\hat{T}(h_+, h_-; k)) = \tilde{R}(MB(k)(h_-^{\gamma_b} + h_+^{\gamma_b}), n^{-1/2}(S_+(k)^2 h_+^{2\gamma_s} + S_-(k)^2 h_-^{2\gamma_s})^{1/2})(1 + o(1)), \quad \text{(S6)}$$

where $\tilde{R}(b, s)$ denotes the value of the criterion when $\hat{T}(h_+, h_-; k) - T(f) \sim N(b, s^2)$, and $S(k) > 0$ and $B(k) > 0$. Assume that $\tilde{R}$ satisfies (5).

In the RD application in Section 3.2.2, 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 Condition (S6) holds with $\gamma_s = -1/2$, $\gamma_b = 2$, $S_+(k) = \sigma_+^2(0)\int_0^\infty k_1^*(u)^2\,\mathrm{d}u/d$, $S_-(k) = \sigma_-^2(0)\int_0^\infty k_1^*(u)^2\,\mathrm{d}u/d$, and $B(k) = -\int_0^\infty u^2 k_1^*(u)\mathrm{d}u/2$.

Let $\rho = h_+/h_-$ denote the ratio of the bandwidths, and let t denote the ratio of the leading worst-case bias and standard deviation terms,

$$t = \frac{MB(k)(h_-^{\gamma_b} + h_+^{\gamma_b})}{n^{-1/2}(S_+(k)^2 h_+^{2\gamma_s} + S_-(k)^2 h_-^{2\gamma_s})^{1/2}} = h_-^{\gamma_b-\gamma_s}\frac{MB(k)(1+\rho^{\gamma_b})}{n^{-1/2}(S_+(k)^2\rho^{2\gamma_s} + S_-(k)^2)^{1/2}}.$$

Substituting $h_+ = \rho h_-$ and $h_- = (tn^{-1/2}(S_+(k)^2\rho^{2\gamma_s} + S_-(k)^2)^{1/2}M^{-1}B(k)^{-1}(1+\rho^{\gamma_b})^{-1})^{1/(\gamma_b-\gamma_s)}$ into (S6) and using linearity of $\tilde{R}$ gives

$$\begin{aligned} R(\hat{T}(h_+, h_-; k)) &= \tilde{R}(MB(k)h_-^{\gamma_b}(1+\rho^{\gamma_b}), h_-^{\gamma_s}n^{-1/2}(S_+(k)^2\rho^{2\gamma_s} + S_-(k)^2)^{1/2})(1+o(1)) \\ &= M^{1-r}n^{-r/2}(1+\varsigma(k)^2\rho^{2\gamma_s})^{r/2}\left(1+\rho^{\gamma_b}\right)^{1-r} S_-(k)^r B(k)^{1-r}\tilde{R}(t,1)(1+o(1)), \end{aligned}$$

where $r = \gamma_b/(\gamma_b - \gamma_s)$ is the rate exponent, and $\varsigma(k) = S_+(k)/S_-(k)$ is the ratio of the variance constants. Therefore, the optimal bias-sd ratio is given by $t_R^* = \operatorname{argmin}_{t>0}\tilde{R}(t,1)$, and depends only on the performance criterion. The optimal bandwidth ratio ρ is given by

$$\rho_* = \operatorname*{argmin}_{\rho}(1+\varsigma(k)^2\rho^{2\gamma_s})^{r/2}\left(1+\rho^{\gamma_b}\right)^{1-r} = \varsigma(k)^{\frac{2}{\gamma_b - 2\gamma_s}},$$

and doesn't depend on the performance criterion.

Consequently, inference that restricts the two bandwidths to be the same (i.e. restricting $\rho = 1$) has asymptotic efficiency given by

$$\begin{aligned} \lim_{n\to\infty}\frac{\min_{h_+,h_-} R(\hat{T}(h_+,h_-;k))}{\min_h R(\hat{T}(h;k))} &= \left(\frac{(1+\varsigma(k)^2\rho_*^{2\gamma_s})^{\gamma_b/2}\left(1+\rho_*^{\gamma_b}\right)^{-\gamma_s}}{(1+\varsigma(k)^2)^{\gamma_b/2}2^{-\gamma_s}}\right)^{\frac{1}{\gamma_b-\gamma_s}} \\ &= 2^{r-1}\frac{\left(1+\varsigma(k)^{\frac{2r}{2-r}}\right)^{1-r/2}}{(1+\varsigma(k)^2)^{r/2}}. \end{aligned}$$

In the RD application in Section 3.2.2, $\varsigma(k) = \sigma_+(0)/\sigma_-(0)$, and $r = 4/5$. The display above implies that the efficiency of restricting the bandwidths to be the same on either side of the cutoff is at least 99.0% if $2/3 \leq \sigma_+/\sigma_- \leq 3/2$, and the efficiency is still 94.5% when the ratio of standard deviations equals 3. There is therefore little gain from allowing the bandwidths to be different.

D.2 Optimal kernels for inference at a point

The optimal equivalent kernel under the Taylor class $\mathcal{F}_{\mathrm{T},p}(M)$ solves Eq. (17) in the main text. The solution is given by

$$k_{SY,p}(u) = \left(b + \textstyle\sum_{j=1}^{p-1} \alpha_j u^j - |u|^p\right)_+ - \left(b + \textstyle\sum_{j=1}^{p-1} \alpha_j u^j + |u|^p\right)_- ,$$

the coefficients b and α solving

$$\int_{\mathcal{X}} u^j k_{SY,p}(u)\,\mathrm{d}u = 0, \qquad j = 1, \ldots, p-1, \qquad \text{and} \qquad \int_{\mathcal{X}} k_{SY,p}(u)\,\mathrm{d}u = 1.$$

For $p = 1$, the triangular kernel $k_{\mathrm{Tri}}(u) = (1 - |u|)_+$ is optimal both in the interior and on the boundary. In the interior for $p = 2$, $\alpha_1 = 0$ solves the problem, yielding the Epanechnikov kernel $k_{\mathrm{Epa}}(u) = \frac{3}{4}(1 - u^2)_+$ after rescaling. For other cases, the solution can be easily found numerically. Figure S1 plots the optimal equivalent kernels for $p = 2$, 3, and 4, rescaled to be supported on $[0, 1]$ and $[-1, 1]$ in the boundary and interior case, respectively.

The optimal equivalent kernel under the Hölder class $\mathcal{F}_{\text{Höl},2}(M)$ has the form of a quadratic spline with infinite number of knots on a compact interval. In particular, in the interior, the optimal kernel is given by $f^{\mathrm{Int}}_{\text{Höl},2}(u) / \int_{-\infty}^{\infty} f^{\mathrm{Int}}_{\text{Höl},2}(u)\,\mathrm{d}u$, where

$$f^{\mathrm{Int}}_{\text{Höl},2}(u) = 1 - \frac{1}{2}x^2 + \sum_{j=0}^{\infty} (-1)^j (|x| - k_j)^2_+,$$

and the knots k_j are given by $k_j = \frac{(1+q)^{1/2}}{1-q^{1/2}}(2 - q^{j/2} - q^{(j+1)/2})$, where q is a constant $q = (3 + \sqrt{33} - \sqrt{26 + 6\sqrt{33}})^2/16$.

At the boundary, the optimal kernel is given by $f^{\mathrm{Bd}}_{\text{Höl},2}(u) / \int_{-\infty}^{\infty} f^{\mathrm{Bd}}_{\text{Höl},2}(u)\,\mathrm{d}u$, where

$$f^{\mathrm{Bd}}_{\text{Höl},2}(u) = (1 - x_0 x + x^2/2)\,\mathrm{I}\{0 \le x \le x_0\} + (1 - x_0^2) f^{\mathrm{Int}}_{\text{Höl},2}((x - x_0)/(x_0^2 - 1))\,\mathrm{I}\{x > x_0\},$$

with $x_0 \approx 1.49969$, so that for $x > x_0$, the optimal boundary kernel is given by a rescaled version of the optimal interior kernel. The optimal kernels are plotted in Figure S2.

D.3 Kernel constants

For the uniform, triangular, and Epanechnikov kernels, the kernel constants $\int_{\mathcal{X}} k^*_q(u)^2\,\mathrm{d}u$, $\mathcal{B}^{\mathrm{T}}_{p,q}(k)$, and $\mathcal{B}^{\text{Höl}}_{p,q}(k)$ discussed in Section 3.2.1 involve integrals that can be computed in closed form. Table S1 gives these constants for the case in which the point of interest is an interior point, and Table S2 gives them for the boundary case.

Appendix E Data-driven Bandwidths

This appendix considers CIs with the bandwidth chosen based on the data, with the smoothness constant M treated as unknown. In particular, we formalize the statements in Section 3.3 regarding honesty and near-optimality of CIs based on the rule-of-thumb bandwidth suggested in that section, over a regularity class that imposes further restrictions.

Consider the regression setting in Section 3.1. Let $\mathcal{F}(M)$ denote the Taylor or Hölder class defined in Section 3.2.1, which places the bound M on the pth derivative of the regression function. Let $\mathcal{F}(M;\eta)$ denote the class that imposes this bound only over $x \in [-\eta, \eta]$. We note that all of our asymptotic results for $\mathcal{F}(M)$ hold for $\mathcal{F}(M;\eta)$ as well. Let $\hat{T}_q(h;k)$ denote the qth order local polynomial estimator, with $q \geq p-1$. Let $h_n = h(M) = (n^{-1/2}S(k)t/(MB(k)))^{1/(\gamma_b-\gamma_s)}$ denote a sequence of bandwidths corresponding to bias-sd ratio t. Here, $B(k)$ and $S(k)$ are given in Theorem 3.1 and $\gamma_b = p$ and $\gamma_s = -1/2$. Let $r = 2p/(2p-1)$ denote the rate exponent. It follows from the results in the main text that the CI $\{\hat{T}_q(h_n;k) \pm \widehat{\mathrm{se}}(h_n;k) \cdot \mathrm{cv}_{1-\alpha}(t)\}$ has correct asymptotic coverage, and it is near-optimal if highly efficient choices for t and k are used.

We consider the CI $\{\hat{T}_q(\hat{h};k) \pm \widehat{\mathrm{se}}(\hat{h};k) \cdot \mathrm{cv}_{1-\alpha}(t)\}$, which uses a data-driven bandwidth $\hat{h}$ to estimate the optimal bandwidth $h_n = h(M)$, thereby avoiding the requirement of prior knowledge of M. As discussed in the main text, results from Low (1997), Cai and Low (2004) and Armstrong and Kolesár (2018) imply that it is impossible for such a CI to achieve coverage and near-optimality over $\mathcal{F}(M;\eta)$ when M is unknown. We therefore consider a class $\mathcal{G}(M) \subsetneq \mathcal{F}(M;\eta)$ that imposes additional conditions that allow M to be estimated consistently. We allow $\mathcal{G}(M)$ to depend directly on the sample size as well, but we leave this implicit in the notation. Appendix E.1 presents results under high level consistency conditions on $\hat{h}$ over the class $\mathcal{G}(M)$. Appendix E.2 defines a particular class $\mathcal{G}(M)$ that formalizes the notion that local smoothness of f is no smaller than its smoothness at large scales, and verifies that the rule-of-thumb bandwidth suggested in Section 3.3 leads to honest CIs over this class. Appendix E.3 derives asymptotic efficiency bounds that show formally that the CI with rule-of-thumb bandwidth considered in Appendix E.2 is highly efficient over the class $\mathcal{G}(M)$. In particular, it is impossible to substantively improve upon this CI using the additional restrictions in the class $\mathcal{G}(M)$. Appendix E.4 presents auxiliary results and intuition for the efficiency bounds presented in Appendix E.3.

E.1 General results for estimated h

We maintain Assumptions 3.1 and 3.2. We make the following additional assumptions on the kernel.

Assumption E.1. *The kernel k is bounded and Lipschitz continuous with finite support.*

Theorem E.1. *Let $h(M) = (n^{-1/2}S(k)t/(MB(k)))^{2/(2p+1)}$ where $t > 0$. Let $\hat{h}$ be a bandwidth sequence, which may depend on the data, such that $\hat{h}/h(M) \xrightarrow{p} 1$ and $nh(M) \to \infty$ uniformly over $\cup_{M \in [\underline{M}_n, \overline{M}_n]} \mathcal{G}(M)$, where $\mathcal{G}(M) \subset \mathcal{F}(M;\eta)$. Let $\widehat{\text{se}}(h;k)$ be a standard error such that $\widehat{\text{se}}(\hat{h};k)/\operatorname{sd}_f(\hat{h};k)$ converges in probability to one uniformly over $\cup_{M \in [\underline{M}_n, \overline{M}_n]} \mathcal{G}(M)$. Let Assumption 3.2 and Assumption E.1 hold, and let Assumption 3.1 hold for any sequence $M_n \in [\underline{M}_n, \overline{M}_n]$. Then*

$$\liminf_{n\to\infty} \inf_{f \in \cup_{M \in [\underline{M}_n, \overline{M}_n]} \mathcal{G}(M)} P_f \left(T(f) \in \left(\hat{T}_q(\hat{h};k) \pm \widehat{\text{se}}(\hat{h};k) \operatorname{cv}_{1-\alpha}(t) \right) \right) \geq 1 - \alpha.$$

The length of the CI satisfies

$$\lim_{n\to\infty} \sup_{M \in [\underline{M}_n, \overline{M}_n]} \sup_{f \in \mathcal{G}(M)} P_f \left(\left| \frac{2\widehat{\text{se}}(\hat{h};k) \operatorname{cv}_{1-\alpha}(t)}{2n^{-r/2} M^{1-r} S(k)^r B(k)^{1-r} t^{r-1} \operatorname{cv}_{1-\alpha}(t)} - 1 \right| > \delta \right) \to 0$$

for any $\delta > 0$.

To prove this theorem, let $M_n \in [\underline{M}_n, \overline{M}_n]$ be given, and let f_n be a sequence of functions in $\mathcal{G}(M_n)$. Let $h_n = h(M_n)$. For any sequence $c_n \to 0$, the coverage probability under f_n is bounded from below by

$$P_{f_n}\left(\left| \frac{\hat{T}_q(h_n;k) - T(f_n)}{\widehat{\text{se}}(\hat{h};k)} \right| \leq \operatorname{cv}_{1-\alpha}(t)(1-c_n) \right) - P_{f_n}\left(\left| \frac{\hat{T}_q(\hat{h};k) - \hat{T}_q(h_n;k)}{\widehat{\text{se}}(\hat{h};k)} \right| > \operatorname{cv}_{1-\alpha}(t) c_n \right).$$

For the first term, we first note that Theorem 2.2 continues to hold with $\sqrt{1/r-1}$ replaced by t and h^*_{RMSE} replaced by h_n, with obvious modifications to the proof. The first term is asymptotically bounded from below by $1-\alpha$ by Theorem 3.1 and this generalization of Theorem 2.2, applied with $\widehat{\text{se}}(\hat{h};k)(1-c_n)$ playing the role of the standard error in Theorem 2.2 (note that, by Theorem 3.1 and the assumptions on $\hat{h}$, $\widehat{\text{se}}(\hat{h};k)/[n^{-1/2}h_n^{-1/2}S(k)]$ converges in probability to one under f_n). The second term will converge to zero for c_n decreasing slowly enough so long as $\sqrt{nh_n}\left(\hat{T}_q(\hat{h};k) - \hat{T}_q(h_n;k)\right)$ converges in probability to zero (again using the fact that $\widehat{\text{se}}(\hat{h};k)/[n^{-1/2}h_n^{-1/2}S(k)]$ converges in probability to one).

Let

$$a_n(h) = \left(\frac{1}{nh} \sum_{i=1}^n k(x_i/h) m_q(x_i/h) m_q(x_i/h)' \right)^{-1} e_1, \quad b_n(x_i;h) = \frac{1}{nh} m_q(x_i/h) k(x_i/h)$$

and let $w_q^n(x;h,k) = a_n(h)'b_n(x_i;h)$. We have

$$\sqrt{nh_n}\left[\hat{T}_q(h_n;k) - \hat{T}_q(\hat{h};k)\right] = \sqrt{nh_n}\sum_{i=1}^n [w_q^n(x_i;h_n,k) - w_q^n(x_i;\hat{h},k)]y_i$$
$$= \sqrt{nh_n}\sum_{i=1}^n [w_q^n(x_i;h_n,k) - w_q^n(x_i;\hat{h},k)]f(x_i)$$
$$+ \sqrt{nh_n}\sum_{i=1}^n [w_q^n(x_i;h_n,k) - w_q^n(x_i;\hat{h},k)]u_i. \quad \text{(S7)}$$

Using a Taylor approximation to $f(x_i)$ around $x = 0$ and the fact that $\sum_{i=1}^n w_q^n(x_i;h,k)x_i^j = 0$ for $j < p$, it follows that the first term is bounded by

$$\sqrt{nh_n}M_n\sum_{i=1}^n |w_q^n(x_i;h_n,k) - w_q^n(x_i;\hat{h},k)|\frac{|x_i|^p}{p!} = \frac{tS(k)}{B(k)p!}\sum_{i=1}^n |w_q^n(x_i;h_n,k) - w_q^n(x_i;\hat{h},k)|\left|\frac{x_i}{h_n}\right|^p,$$

where we substitute $M_n = tn^{-1/2}S(k)/(B(k)h_n^{p+1/2})$. Letting C be a bound on the support of the kernel k, we have $|x_i| \leq C\max\{\hat{h}, h_n\}$ for any x_i such that the summand is nonzero. Thus, on the event $\hat{h} \leq 2h_n$, the above display is bounded by $\frac{(2C)^p tS(k)}{B(k)p!}$ times

$$\sum_{i=1}^n |w_q^n(x_i;h_n,k) - w_q^n(x_i;\hat{h},k)|.$$

Using the fact that $w_q^n(x_i;h_n,k) - w_q^n(x_i;\hat{h},k) = a_n(h_n)'[b_n(x_i;h_n) - b_n(x_i;\hat{h})] + [a_n(h) - a_n(\hat{h})]'b_n(x_i;\hat{h})$, it follows that the above display is bounded by

$$\|a_n(h_n)\|\sum_{i=1}^n \|b_n(x_i;h_n) - b_n(x_i;\hat{h})\| + \|a_n(h_n) - a_n(\hat{h})\|\sum_{i=1}^n \|b_n(x_i;\hat{h})\|.$$

Similarly, the last term in (S7) is bounded by

$$\|a_n(h_n)\|\left\|\sqrt{nh_n}\sum_{i=1}^n [b_n(x_i;h_n) - b_n(x_i;\hat{h})]u_i\right\| + \|a_n(h_n) - a_n(\hat{h})\|\left\|\sqrt{nh_n}\sum_{i=1}^n b_n(x_i;\hat{h})u_i\right\|.$$

Both of these quantities converge in probability to zero by the following lemma.

Lemma E.1. *Suppose that Assumption 3.1 and Assumption E.1 hold. Let $\tilde{g}(x) = k(x)x^j$ or $\tilde{g}(x) = |k(x)x^j|$ for some $j \geq 0$. Then*

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\sup_{s\in[1-\delta,1+\delta]}\frac{1}{nh_n}\sum_{i=1}^n |\tilde{g}(x_i/(sh_n)) - \tilde{g}(x_i/h_n)| = 0.$$

and

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\sup_{s\in[1-\delta,1+\delta]}\left|\frac{1}{nsh_n}\sum_{i=1}^{n}\tilde{g}(x_i/(sh_n)) - d\int_{\mathcal{X}}\tilde{g}(u)\,du\right| = 0.$$

If, in addition, Assumption 3.2 holds, then, for all $\varepsilon > 0$*,*

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\sup_{s\in[1-\delta,1+\delta]} P\left(\sup_{s\in[1-\delta,1+\delta]}\left|\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}[\tilde{g}(x_i/(sh_n)) - \tilde{g}(x_i/h_n)]u_i\right| > \varepsilon\right) = 0.$$

Proof. By Assumption 3.1, the second display in the lemma follows from the first. By Assumption E.1, for large enough C, $|\tilde{g}(u) - \tilde{g}(u')| \leq C|u-u'|\,\mathrm{I}\{\max\{|u|,|u'|\} \leq C\}$. Thus, the first display in the lemma is bounded by

$$\begin{aligned}
&\lim_{\delta\to 0}\limsup_{n\to\infty}\sup_{s\in[1-\delta,1+\delta]}\frac{1}{nh_n}\sum_{i=1}^{n} C\cdot|s^{-1}-1|\,\mathrm{I}\{|x_i/h_n| \leq 2C\} \\
&\quad= \lim_{\delta\to 0}\left[\sup_{s\in[1-\delta,1+\delta]}|s^{-1}-1|\right]\limsup_{n\to\infty}\frac{1}{nh_n}\sum_{i=1}^{n} C\cdot\mathrm{I}\{|x_i/h_n| \leq 2C\} \\
&\quad= \lim_{\delta\to 0}\left[\sup_{s\in[1-\delta,1+\delta]}|s^{-1}-1|\right]\int_{\mathcal{X}}\mathrm{I}\{u \leq 2C\}\,du\cdot C = 0.
\end{aligned}$$

For the second part of the lemma, we have, for $s, \tilde{s}$ in a small enough neighborhood of 1, letting $\overline{\sigma}^2$ denote a bound on $\sigma^2(x)$ in a neighborhood of zero,

$$\begin{aligned}
E\left(\sum_{i=1}^{n}\left[\frac{1}{\sqrt{nh_n}}\tilde{g}(x_i/(sh_n)) - \tilde{g}(x_i/(\tilde{s}h_n))\right]u_i\right)^2 &\leq \overline{\sigma}^2\frac{1}{nh_n}\sum_{i=1}^{n}\left[\tilde{g}(x_i/(sh_n)) - \tilde{g}(x_i/(\tilde{s}h_n))\right]^2 \\
&\leq \overline{\sigma}^2\frac{1}{nh_n}\sum_{i=1}^{n} C^2|x_i/h_n|^2|s^{-1}-\tilde{s}^{-1}|^2\,\mathrm{I}\{|x_i/h_n| \leq 2C\}.
\end{aligned}$$

For large enough n, this is bounded by $|s^{-1}-\tilde{s}^{-1}|^2$ times a constant that does not depend on n. The result now follows from Example 2.2.12 in van der Vaart and Wellner (1996). □

Finally, for the last statement of the theorem, note that the length of the CI is given by $2\widehat{\mathrm{se}}(\hat{h};k)\,\mathrm{cv}_{1-\alpha}(t)$ which, under the sequence f_n, is equal to a $1+o_P(1)$ term times

$$2n^{-1/2}h_n^{-1/2}S(k)\,\mathrm{cv}_{1-\alpha}(t) = 2n^{-r/2}M_n^{1-r}S(k)^rB(k)^{1-r}t^{r-1}\,\mathrm{cv}_{1-\alpha}(t).$$

E.2 Bounds based on global polynomial approximations

We now verify the conditions of Theorem E.1 in a particular setting. In particular, we consider classes $\mathcal{G}$ that relate M to a global polynomial approximation to the regression function, along with a plug-in bandwidth $\hat{h}$ based on this assumption.

Let $\mathcal{F}(M)$ be the Taylor or Hölder class of order p, and let $\mathcal{F}(M;\eta)$ denote the class that imposes this bound only over $x \in [-\eta, \eta]$. Let $\tilde{p} \geq p$ be given. Let $Q_{\tilde{p}}f$ denote the minimum mean squared error $\tilde{p}$th order polynomial predictor for the regression function f:

$$Q_{\tilde{p}}f = \arg\min_h \int (f(x) - h(x))^2 d(x)\sigma^2(x)\, dx$$

where the minimum is taken over polynomials of order $\tilde{p}$. Here, $d(x)$ is such that the x_i's behave as if drawn from a distribution with density x_i, as formalized in the Assumption E.2 below.

Let $x_{\min}, x_{\max}$ be given with $-\infty < x_{\min} < x_{\max} < \infty$. Let

$$J(f) = J(f; \tilde{p}, x_{\min}, x_{\max}) = \sup_{x \in [x_{\min}, x_{\max}]} |[Q_{\tilde{p}}f]^{(p)}(x)|$$

denote the maximum pth derivative of the minimum mean squared error $\tilde{p}$th order approximation of f.

Let $\varepsilon > 0$ be given. Let

$$\mathcal{Q}(M, \tilde{p}, x_{\min}, x_{\max}, \varepsilon) = \{f : J(f) = \varepsilon M\},$$
$$\mathcal{G}(M) = \mathcal{G}(M; \tilde{p}, \varepsilon, \eta, x_{\min}, x_{\max}) = \mathcal{F}(M, \eta) \cap \mathcal{Q}(M, \tilde{p}, x_{\min}, x_{\max}, \varepsilon) \cap \{f : \sup_x |f(x)| \leq K\},$$

where K is some large constant, and

$$\mathcal{H}(\underline{M}, \overline{M}) = \cup_{M \in [\underline{M}, \overline{M}]} \mathcal{G}(M; \tilde{p}, \varepsilon, \eta, x_{\min}, x_{\max}).$$

This class formalizes the notion that the pth derivative in a neighborhood of zero is bounded by ε^{-1} times the maximum pth derivative of a global $\tilde{p}$th order global polynomial approximation. Setting $\varepsilon = 1$ corresponds to the suggestion in the main text.

Let

$$\hat{Q}_{\tilde{p}} = \arg\min_h \sum_{i=1}^n (y_i - h(x_i))^2, \quad \hat{J} = \sup_{x \in [x_{\min}, x_{\max}]} |\hat{Q}_{\tilde{p}}^{(p)}(x)|$$

We make the following additional assumption on the x_i's.

Assumption E.2. *For some bounded function $d(x)$ and a sequence c_n with $c_n \to \infty$ and $c_n/\sqrt{n} \to 0$, we have, for each $j = 0, \ldots, \tilde{p}$,*

$$c_n \left| \frac{1}{n} \sum_{i=1}^n x_i^j f_n(x_i) - \int u^j f_n(u) d(u)\, du \right| \to 0$$

for any uniformly bounded sequence of functions f_n. Furthermore, the $\tilde{p}+1$ by $\tilde{p}+1$ matrix *with (j,ℓ)th element given by $\int u^{j+\ell-2}d(u)\,du$ is invertible.*

Given a sequence c_n satisfying the conditions of Assumption E.2, if the x_i's are drawn iid from a distribution with density $d(x)$ for which all moments are finite, then Assumption E.2 will hold with probability approaching one.

We note the following consistency result for $\hat{J}$.

Lemma E.2. *Suppose Assumption 3.2 holds with $\sigma^2(x)$ bounded and that Assumption E.2 holds. Then $c_n|\hat{J} - J(f)| \overset{p}{\to} 0$ uniformly over $\{f : \sup_x |f(x)| \le K\}$.*

Proof. Let A denote the $\tilde{p}+1$ by $\tilde{p}+1$ matrix with (j,ℓ)th element given by $\int u^{j+\ell-2}d(u)\,du$, and let $\hat{A}$ denote the sample analogue with (j,ℓ)th element given by $\frac{1}{n}\sum_{i=1}^n x_i^{j+\ell-2}$. Let b_f be the $(\tilde{p}+1)\times 1$ vector with jth element $\int u^j f(u)d(u)\,du$ and $\hat{b}$ be the sample analogue with jth element $\frac{1}{n}\sum_{i=1}^n x_i^{j-1}y_i$. Then $A^{-1}b_f$ gives the coefficients of the polynomial $Q_{\tilde{p}}f$, and $\hat{A}^{-1}\hat{b}$ gives the coefficients of the polynomial $\hat{Q}$. Let $s(A,b)$ denote the function that takes the maximum of the pth derivative of this polynomial over $[x_{\min}, x_{\max}]$, so that $J(f) = s(A, b_f)$ and $\hat{J} = s(\hat{A},\hat{b})$. Note that $|s(\hat{A},\hat{b}) - s(A,b_f)|$ is bounded by $\max\{\|\hat{A}-A\|, \|\hat{b}-b_f\|\}$ times a constant that does not depend on f, so it suffices to show that $c_n\max\{\|\hat{A}-A\|, \|\hat{b}-b_f\|\}$ converges in probability to zero uniformly over bounded f.

We have $c_n\|\hat{A}-A\| \to 0$ by Assumption E.2. The jth element of $c_n(\hat{b}-b_f)$ is given by

$$\frac{c_n}{n}\sum_{i=1}^n u_i x_i^{j-1} + c_n\left(\frac{1}{n}\sum_{i=1}^n f(x_i)x_i^{j-1} - \int f(u)u^{j-1}d(u)\,du\right).$$

The expectation of the square of the first term converges to zero, since it is bounded by c_n^2/n^2 times a sequence that converges to a constant by Assumption E.2. The last term converges to zero uniformly over bounded f by Assumption E.2. Thus, $c_n\|\hat{b}-b_f\| \overset{p}{\to} 0$ uniformly over bounded f. □

Let M_n and ε_n be given, and consider honesty over the sequence of classes $\mathcal{G}(M_n;\tilde{p},\varepsilon_n,\eta,x_{\min},x_{\max})$. Let t be given, and let $\hat{h} = (n^{-1/2}\hat{S}(k)t/(\hat{M}\hat{B}(k)))^{2/(2p+1)}$ where $\hat{S}(k)/S(k)$ and $\hat{B}(k)/B(k)$ converge in probability to one uniformly over $\mathcal{G}(M_n)$ (as discussed in Section 3.3, we can also directly minimize the sample analogue of the criterion such that t is the asymptotically optimal bias-sd ratio). Then $\hat{h}$ will satisfy the conditions of Theorem E.1 so long as $\hat{h}/h(M_n)$ converges in probability to one uniformly over $\mathcal{G}(M_n)$, where

$$h(M) = (n^{-1/2}S(k)t/(MB(k)))^{2/(2p+1)}.$$

For this, it suffices that $\hat{M}/M_n$ converges in probability to one uniformly over $\mathcal{G}(M_n)$.

According to Lemma E.2, we can use the estimate $\hat{M} = \varepsilon^{-1}\hat{J}$, which gives

$$\frac{\hat{M}}{M_n} - 1 = \frac{\varepsilon_n^{-1}[\hat{J} - J(f)]}{M_n} = o_P(1/(\varepsilon_n M_n c_n))$$

uniformly over $\mathcal{G}(M; \tilde{p}, \varepsilon_n, \eta, x_{\min}, x_{\max})$. If Assumption E.2 holds for any c_n with $c_n/\sqrt{n} \to 0$, then this can be made to go to zero so long as $\varepsilon_n M_n \sqrt{n} \to \infty$. Thus, the resulting CI is honest over the class $\mathcal{H}(\underline{M}_n, \overline{M}_n)$ so long as $\varepsilon_n \underline{M}_n \sqrt{n} \to \infty$, and such that Assumption 3.1 holds for the sequences $\underline{M}_n$ and $\overline{M}_n$. Note also that, if one uses $\hat{M} = \tilde{\varepsilon}^{-1}\hat{J}$ where $\tilde{\varepsilon} < \varepsilon$ (thereby choosing ε to be "too small"), then the resulting CI will be wider, but will still have correct coverage.

While Assumption 3.1 is stated as a high level condition, note that, in order for this condition to hold with probability approaching one when the x_i's are drawn iid from a distribution satisfying appropriate regularity conditions, we will need $nh_n \to \infty$ and $h_n \to 0$ for the given sequence h_n. This will be ensured for any sequence $M_n \in [\underline{M}_n, \overline{M}_n]$ iff. $\underline{M}_n$ satisfies $n\underline{M}_n^2 \to \infty$ and $\overline{M}_n$ satisfies $\overline{M}_n/n^p \to 0$ so that $n(n\overline{M}_n^2)^{-1/(2p+1)} = n^{2p/(2p+1)}\overline{M}_n^{-2/(2p+1)} \to \infty$. Also, note that we have assumed a uniform bound on the magnitude of the regression function, which means that $\varepsilon_n \overline{M}_n$ must be bounded uniformly over n (although this condition could likely be relaxed).

E.3 Lower bounds

The CI in Theorem E.1 has the property that the ratio of its length to the length of an "oracle" FLCI that uses the unknown true M converges to one. If the optimal kernel is used and the bias-sd ratio is chosen to be optimal for FLCI length, then this CI is efficient among FLCIs over the class $\mathcal{F}(M; \eta)$. Furthermore, it is highly efficient among all CIs that are honest over the class $\mathcal{F}(M; \eta)$, since one can apply bounds such as Corollary 3.3 in Armstrong and Kolesár (2018). However, these results do not apply to the class $\mathcal{G}(M)$ over which the feasible CI with estimated optimal bandwidth has coverage, since $\mathcal{G}(M) \subsetneq \mathcal{F}(M; \eta)$: they do not rule out the possibility that this restricted class might allow for a more informative CI. To address this, we now derive efficiency bounds for the class $\mathcal{G}(M) = \mathcal{G}(M; \tilde{p}, \varepsilon, \eta, x_{\min}, x_{\max})$ used in Appendix E.2.

Theorem E.2. *Let M, ε, η and $[x_{\min}, x_{\max}]$ be given. Suppose that Assumptions 3.1 and 3.2 hold with $\sigma(x)$ bounded from above and below away from zero and u_i following a normal distribution, and that Assumption E.2 holds with $d(x)$ strictly positive on some open set in $\mathbb{R}\backslash[-\eta, \eta]$. Then, if the constant K used to define $\mathcal{G}(M)$ is large enough, the following holds. For any sequence of CIs $\{\hat{T} \pm \hat{\chi}\}$ with asymptotic coverage at least $1-\alpha$ under $\mathcal{G}(M)$,*

$$\lim_{C\to\infty} \liminf_n \inf_{f\in\mathcal{G}(M)} E_{f_n} \min\{2n^{r/2}\hat{\chi}, C\} \geq \frac{2M^{1-r}S(k^*)^r B(k^*)^{1-r}}{r^r(1-r)^{r-1}} \int_{z=-\infty}^{z_{1-\alpha}} (z_{1-\alpha} - z)^r \, d\Phi(z)$$

where k^* *minimizes* $S(k^*)^r B(k^*)^{1-r}$.

If $\hat{h}$ and $\widehat{\text{se}}(h;k)$ satisfy the conditions of Theorem E.1, then, by Theorem E.2, the relative efficiency of any CI $\{\hat{T} \pm \hat{\chi}\}$ to $\{\hat{T}_q(\hat{h};k) \pm \widehat{\text{se}}(\hat{h};k)\,\text{cv}_{1-\alpha}(t)\}$ satisfies the lower bound

$$\lim_{C\to\infty} \liminf_{n} \sup_{f\in\mathcal{G}(M)} \frac{E_f \min\{2n^{r/2}\hat{\chi}, C\}}{E_f \min\{2n^{r/2}\widehat{\text{se}}(\hat{k};k)\,\text{cv}_{1-\alpha}(t), C\}} \geq \frac{\int_{z=-\infty}^{z_{1-\alpha}} (z_{1-\alpha}-z)^r\, d\Phi(z)}{r^r(1-r)^r \inf_{\tilde{t}} \tilde{t}^{r-1}\,\text{cv}_{1-\alpha}(\tilde{t})} \cdot \frac{S(k^*)^r B(k^*)^{1-r}}{S(k)^r B(k)^{1-r}} \cdot \frac{\inf_{\tilde{t}} \text{cv}_{1-\alpha}(\tilde{t})}{t^{r-1}\,\text{cv}_{1-\alpha}(t)}.$$

The first term is the lower bound in Theorem E.1 of Armstrong and Kolesár (2018), which corresponds to the lower bound in Corollary 3.3 of that paper applied to the case where the modulus $\omega(\delta)$ is proportional to δ^r (as is the case in the relevant limiting experiment in the present setting; see Appendix E.4). The second term is the relative efficiency of the kernel k, and the final term is the efficiency of the bias-sd ratio used in the bandwidth $\hat{h}$ relative to the optimal bias-sd ratio for FLCI construction.

We now prove Theorem E.2. We begin by noting some properties of the optimal kernel k^*.

Lemma E.3. *Let* κ^* *solve*

$$\max_{\kappa} \kappa(0) \quad s.t. \quad \int_{\mathcal{X}} \kappa(u)^2\, du \leq 1,\; \kappa \in \mathcal{F}(1)$$

and let $k^*(x) = \kappa^*(x)/\int_{\mathcal{X}} \kappa(u)\,du$. *Then* k^* *has finite support, and it minimizes* $S(k)^r B(k)^{1-r}$ *over kernels* k. *Furthermore,* $S(k^*) = [\sigma^2(0)/d]^{1/2} r\kappa^*(0)$ *and* $B(k^*) = (1-r)\kappa^*(0)$, *so that* $S(k^*)^r B(k^*)^{1-r} = [\sigma^2(0)/d]^{r/2} r^r (1-r)^{1-r}\kappa^*(0)$.

Proof. The result follows from Low (1995) and Donoho and Low (1992). See Appendix E.4.3. □

The next lemma uses functions constructed from κ^* to derive testing bounds.

Lemma E.4. *Suppose that the conditions of Theorem E.2 hold. Given* $c \in \mathbb{R}$, *let* $\mathcal{K}_{c,n} = \{f\colon f(0) = cn^{-p/(2p+1)}\} \cap \mathcal{G}(M)$. *Then, if the constant* K *used to define* $\mathcal{G}(M)$ *is larger than a constant that depends only on* ε *and* M, *there exists a sequence of functions* $\tilde{\kappa}_{0,n} \in \mathcal{K}_{0,n}$ *such that the following holds. For any* $c \in \mathbb{R}$ *and any sequence of tests with asymptotic size* α *under* $\mathcal{K}_{c,n}$, *the asymptotic power under* $\tilde{\kappa}_{0,c}$ *is no greater than*

$$\Phi\left(|c/\kappa^*(0)|^{(2p+1)/(2p)} M^{-1/(2p)} [d/\sigma^2(0)]^{1/2} - z_{1-\alpha}\right).$$

Proof. It suffices to prove the result for $c > 0$. Let A and b_f be defined as in the proof of Lemma E.2, so that the coefficients of the minimum mean squared error $\tilde{p}$th order polynomial predictor are given by $A^{-1}b_f$. We first note that, under the conditions of the lemma, there exist bounded functions $f_1,\ldots,f_{\tilde{p}+1}$ supported on $\mathbb{R}\backslash[-\eta,\eta]$ such that the vectors $b_{f_1},\ldots,b_{f_{\tilde{p}+1}}$ are linearly independent. Thus, these vectors span $\mathbb{R}^{\tilde{p}+1}$, which means that there exist functions $g_1,\ldots,g_{f_{\tilde{p}+1}}$, which are linear combinations of the f_j's (and therefore also bounded and supported on $\mathbb{R}\backslash[-\eta,\eta]$) such that $b_{g_j}=e_j$ for each j, where e_j denotes the jth standard basis vector.

We construct functions in the sets $\mathcal{K}_{c,n}$ as follows. Let $\tilde{g}$ be a bounded function supported on $\mathbb{R}\backslash[-\eta,\eta]$ such that $J(\tilde{g})=\varepsilon M$. This function can be constructed by finding a polynomial such that the supremum of the pth derivative over $[x_{\min},x_{\max}]$ is equal to εM, and constructing a function with the given polynomial predictor coefficients as a linear combination of the g_js defined above. Given a function f supported on $[-\eta,\eta]$, the function $b_{f,1}g_1+b_{f,2}g_2+\cdots+b_{f,\tilde{p}+1}g_{\tilde{p}+1}$ is supported on $\mathbb{R}\backslash[-\eta,\eta]$ and has the same polynomial predictor coefficients as f. Thus, the function $f-(b_{f,1}g_1+b_{f,2}g_2+\cdots+b_{f,\tilde{p}+1}g_{\tilde{p}+1})+\tilde{g}$ has the same polynomial predictor coefficients as $\tilde{g}$. It therefore follows that, if $f\in\mathcal{F}(M;\eta)$ and K is larger than some constant that depends only on an upper bound for the elements of b_f and the functions $g_1,\ldots,g_{\tilde{p}+1}$ and $\tilde{g}$, this function will be in $\mathcal{G}(M)$.

Let $\tilde{\kappa}_{c,M,n}$ be defined in this way with the function $\kappa_{c,M,n}$ playing the role of f, where $\kappa_{c,M,n}(x)=Mh_{c,n}^p\kappa^*(x/h_{c,n})$ with $h_{c,n}=\tilde{c}n^{-1/(2p+1)}$ where $\tilde{c}=|c/[M\kappa^*(0)]|^{1/p}$. Note that $\kappa_{c,M,n}\in\mathcal{F}(M)$ by the renormalization property of Taylor and Hölder classes. Thus, once n is large enough that the support of $\kappa_{c,M,n}$ is contained in $[-\eta,\eta]$, we will have $\tilde{\kappa}_{c,M,n}\in\mathcal{K}_{c,n}$.

It follows that, for large enough n, the power under $\tilde{\kappa}_{0,M,n}$ of a level α_n test of $\mathcal{K}_{c,n}$ is bounded by the power under $\tilde{\kappa}_{0,M,n}$ of a test with rejection probability no greater than α_n under $\tilde{\kappa}_{c,M,n}$. By the Neyman-Pearson lemma and standard calculations, this is no greater than $\Phi(s_n-z_{1-\alpha_n})$ where

$$s_n^2=\sum_{i=1}^n\left[\tilde{\kappa}_{c,M,n}(x_i)-\tilde{\kappa}_{0,M,n}(x_i)\right]^2\sigma^{-2}(x_i)=M^2h_{c,n}^{2p}\sum_{i=1}^n\kappa^*(x_i/h_{c,n})^2\sigma^{-2}(x_i)$$
$$+\sum_{i=1}^n\left[\sum_{j=1}^{\tilde{p}+1}g_j(x_i)\sigma^{-2}(x_i)\int Mh_{c,n}^p\kappa^*(u/h_{c,n})u^{j-1}d(u)\,du\right]^2.$$

Note that $h_{c,n}^{2p}=\tilde{c}^{2p}n^{-2p/(2p+1)}=n^{-1}\tilde{c}^{2p+1}n^{1/(2p+1)}\tilde{c}^{-1}=(n\tilde{c}n^{-1/(2p+1)})^{-1}\tilde{c}^{2p+1}$. Thus, the first term equals $\tilde{c}^{2p+1}M^2\frac{1}{nh_{c,n}}\sum_{i=1}^n\kappa^*(x_i/h_{c,n})^2\sigma^2(x_i)\to\sigma^{-2}(0)\tilde{c}^{2p+1}M^2d\int_{\mathcal{X}}\kappa^*(u)^2\,du$. The last term is bounded from above by a constant times

$$n\left[h_{c,n}^p\int\kappa^*(u/h_{c,n})\,du\right]^2=n\left[h_{c,n}^{p+1}\int\kappa^*(v)\,dv\right]=n^{1-(2p+2)/(2p+1)}\tilde{c}^{(2p+2)/p}\left[\int\kappa^*(u)\,du\right]^2\to 0.$$

The result then follows by plugging in $\tilde{c}$ and noting $\int\kappa^*(u)^2\,du=1$. □

To derive the lower bound on expected length, we argue as in the proof of Theorem C.2 in Armstrong and Kolesár (2019). Consider the set $\mathcal{I}(m) = \{\tilde{c}_n j/m : j \in \mathbb{Z}, |j| \leq m^2\}$ where $\tilde{c}_n = \kappa^*(0)M^{1/(2p+1)}[\sigma^2(0)/d]^{p/(2p+1)}n^{-p/(2p+1)}$. Let $\hat{T} \pm \hat{\chi}$ be a CI with asymptotic coverage at least $1-\alpha$ over $\mathcal{G}(M)$, and let $\mathcal{N}(n,m)$ denote the number of elements in $\mathcal{I}(m)$ that are in this confidence interval. Note that $\min\{2\hat{\chi}, 2\tilde{c}_n m\} \geq \tilde{c}_n[\mathcal{N}(m,n)-1]/m$. Let $\kappa_{0,n}$ and $\mathcal{K}_{c,n}$ be as defined in Lemma E.4. Let $\psi_{n,j}$ denote the test that rejects when the point $\tilde{c}_n j/m \in \mathcal{N}(n,m)$ is not in the CI $\hat{T} \pm \hat{\chi}$. Then $\psi_{n,j}$ is an asymptotically level α test of $\mathcal{K}_{c,n}$, so, by Lemma E.4,

$$E_{\kappa_{0,n}}\mathcal{N}(m,n) = \sum_{j=-m^2}^{m^2}(1 - E_{\kappa_{0,n}}\psi_{n,j}) \geq \sum_{j=-m^2}^{m^2}(1 - \Phi(|j/m|^{(2p+1)/2p} - z_{1-\alpha})) + o(1).$$

Thus, for all $m \in \mathbb{N}$, $\lim_{C\to\infty}\liminf_n E_{\kappa_{0,n}} \min\{2\tilde{c}_n^{-1}\hat{\chi}, C\}$ is bounded from below by

$$\begin{aligned}\frac{1}{m}\sum_{j=-m^2}^{m^2}\Phi(z_{1-\alpha} - |j/m|^{(2p+1)/(2p)}) &= \frac{1}{m}\sum_{j=-m^2}^{m^2}\int \mathrm{I}\{|j/m|^{(2p+1)/(2p)} \leq z_{1-\alpha} - z\}\,d\Phi(z)\\ &= \frac{1}{m}\sum_{j=-m^2}^{m^2}\int \mathrm{I}\{|j| \leq (z_{1-\alpha}-z)^{2p/(2p+1)}m\}\,d\Phi(z)\\ &\geq \int_{z=-\infty}^{z_{1-\alpha}}\frac{1}{m}\min\left\{2\left[(z_{1-\alpha}-z)^{2p/(2p+1)}m - 1\right], m\right\}\,d\Phi(z).\end{aligned}$$

This converges to $2\int_{z=-\infty}^{z_{1-\alpha}}(z_{1-\alpha}-z)^{2p/(2p+1)}\,d\Phi(z)$ by the Dominated Convergence Theorem. Thus,

$$\begin{aligned}&\lim_{C\to\infty}\liminf_n E_{\kappa_{0,n}}\min\{2n^{p/(2p+1)}\hat{\chi}, C\}\\ &\qquad\geq 2\kappa^*(0)M^{1/(2p+1)}[\sigma^2(0)/d]^{p/(2p+1)}\int_{z=-\infty}^{z_{1-\alpha}}(z_{1-\alpha}-z)^{2p/(2p+1)}\,d\Phi(z)\\ &\qquad= 2\kappa^*(0)M^{1-r}[\sigma^2(0)/d]^{r/2}\int_{z=-\infty}^{z_{1-\alpha}}(z_{1-\alpha}-z)^{r}\,d\Phi(z).\end{aligned}$$

Plugging in $S(k^*)^rB(k^*)^{1-r} = [\sigma^2(0)/d]^{r/2}r^r(1-r)^{1-r}\kappa^*(0)$ gives the result.

E.4 Limiting model and optimal kernel

In this appendix we derive the properties of the optimal kernel given in Lemma E.3. To do so, we apply results from Low (1995) and Donoho and Low (1992) to the limiting model

$$Y(dt) = f(t)\,dt + \lambda W(dt), \quad t \in \mathcal{X} \tag{S8}$$

where $\mathcal{X} = \mathbb{R}$ in the case where the point of interest is on the interior of the support of x_i and $\mathcal{X} = [0, \infty)$ when it is on the boundary. We also use this limiting model to give some intuitive motivation for the efficiency bound in Theorem E.2.

The white noise model (S8) is the same model as in Appendix B.2, with λ playing the role of $\sigma/\sqrt{n}$ in that appendix. Brown and Low (1996) establish a formal sense in which this white noise model, with λ replaced by the function $\lambda_n(t) = [\sigma^2(t)/(nd(t))]^{1/2}$, is asymptotically equivalent to the fixed design regression model. Since the asymptotic behavior of our estimators and bounds depends only on x_i in a shrinking neighborhood of zero, we then expect that $\lambda_n(t)$ can be replaced by the constant function $\lambda_n(0)$. For technical reasons, however, the proof of Theorem E.2 uses direct arguments, rather than appealing to the equivalence results of Brown and Low (1996) (in particular, these results do not apply immediately for Taylor classes, or when smoothness is only assumed in the neighborhood $[-\eta, \eta]$).

E.4.1 Kernel estimators

Let k be a kernel with $\int_{\mathcal{X}} k(u)\, du = 1$ and $\int_{\mathcal{X}} k(u)u^j\, du = 0$ for $j = 1, \dots, p-1$. The kernel k will play the role of the equivalent kernel k_q^* in Section 3.2.1. A linear estimator in the white noise model takes the form

$$\hat{T}(h; k) = h^{-1} \int k(t)\, dY(t).$$

Since this falls into the Donoho and Low (1992) framework given in Appendix B.2, it follows that Eq. (2) holds with the $o(1)$ terms equal to zero. Indeed, under $f \in \mathcal{F}(M)$, $\hat{T}(h; k)$ follows a normal distribution with bias

$$h^{-1} \int_{\mathcal{X}} k(t/h)(f(t) - f(0))\, dt = \int_{\mathcal{X}} k(u)(f(hu) - f(0))\, du = Mh^p \int_{\mathcal{X}} k(u)(\tilde{f}(u) - \tilde{f}(0))\, du$$

where $\tilde{f}(u) = M^{-1}h^{-p}f(hu)$ is in $\mathcal{F}(1)$ iff. $f \in \mathcal{F}(M)$, by the renormalization property of the Hölder and Taylor class. The variance is given by

$$\lambda^2 h^{-2} \int_{\mathcal{X}} k(t/h)^2\, dt = \lambda^2 h^{-1} \int_{\mathcal{X}} k(u)^2\, du.$$

Thus, if we take $\lambda = [\sigma^2(0)/(nd)]^{1/2}$, Eq. (2) holds with $S(k) = \sigma(0)d^{-1/2}\sqrt{\int_{\mathcal{X}} k(u)\, du}$, $B(k) = \sup_{\tilde{f} \in \mathcal{F}(1)} \int_{\mathcal{X}} k(u)(\tilde{f}(u) - \tilde{f}(0))\, du$, $\gamma_b = p$ and $\gamma_s = -1/2$. Note that $S(k)$ matches Equation (5) with k playing the role of the equivalent kernel k_q^* in Equation (5). In addition, $B(k)$ matches the expression given in Theorem 3.1 (this can be shown by deriving $B(k)$ using the arguments in the proof of this theorem).

E.4.2 Modulus of continuity

The modulus of continuity for the limiting model, as defined in Donoho (1994), is given by

$$\omega(\delta) = 2\sup_f f(0) \quad \text{s.t.} \quad \int_{\mathcal{X}} f(x)^2\,dx \le \delta^2/4, \quad f \in \mathcal{F}(M).$$

Let $f^*_{\delta,M}$ denote the solution to this problem. Note that the function κ^* defined in Lemma E.3 is given by $f^*_{2,1}$. By Donoho and Low (1992), we have $f^*_{\delta,M}(x) = M\tilde{h}^p_{\delta,M}\kappa^*(x/\tilde{h}_{\delta,M})$ where $\tilde{h}_{\delta,M} = (\delta/(2M))^{2/(2p+1)}$, which gives

$$\omega(\delta) = 2M(\delta/(2M))^{2p/(2p+1)}\kappa^*(0) = (2M)^{1-r}\delta^r\kappa^*(0)$$

where $r = 2p/(2p+1)$ is the rate exponent. Note that

$$\omega'(\delta) = r(2M)^{1-r}\delta^{r-1}\kappa^*(0) = r\delta^{-1}\omega(\delta).$$

E.4.3 Optimal kernel

By Low (1995), the bias-sd optimizing kernel takes the form $t \mapsto f^*_{\delta,M}(t)/\int_{\mathcal{X}} f^*_{\delta,M}(u)\,du$ for some δ, so this implies that $k^*(t) = \kappa^*(t)/\int_{\mathcal{X}}\kappa^*(u)\,du$ is the optimal kernel. For Taylor classes, the support can be seen to be compact by examining the formula given in Section 3.2.1. For Hölder classes, this can be shown indirectly (see Lepski and Tsybakov, 2000). The worst-case bias of the estimate with bandwidth $h_{\delta,M}$ is given by

$$(1/2)(\omega(\delta) - \delta\omega'(\delta)) = (1/2)\omega(\delta)(1-r) = (1/2)(1-r)(2M)^{1-r}\delta^r\kappa^*(0) = M(1-r)\kappa^*(0)h^p_{\delta,M}$$

where we substitute $\delta = 2Mh^{(2p+1)/2}_{\delta,M}$ in the last step. This gives the formula $B(k^*) = (1-r)\kappa^*(0)$. The standard deviation is given by

$$\lambda\omega'(\delta) = \lambda r(2M)^{1-r}\delta^{r-1}\kappa^*(0) = \lambda r\kappa^*(0)h^{-1/2}_{\delta,M} = [\sigma^2(0)/d]^{1/2}r\kappa^*(0)n^{-1/2}h^{-1/2}_{\delta,M},$$

which gives $S(k^*) = [\sigma^2(0)/d]^{1/2}r\kappa^*(0)$. Thus, the leading term in the minimax performance is $S(k^*)^rB(k^*)^{1-r} = [\sigma^2(0)/d]^{r/2}r^r(1-r)^{1-r}\kappa^*(0)$.

E.4.4 Optimal FLCI and efficiency bound

We now show that the efficiency bound in Theorem E.2 corresponds to the bound given in Corollary 3.3 in Armstrong and Kolesár (2018), applied to the class $\mathcal{F}$ in the limiting model (S8). Thus, Theorem E.2 can be interpreted as showing that this efficiency bound holds in a formal asymptotic sense, with $\mathcal{F}(M;\eta)$ replaced by the smaller class $\mathcal{G}(M)$. We note that, for Taylor classes, such a bound is given for the class $\mathcal{F}(M)$ in Theorem E.1 in Armstrong and Kolesár (2018). Theorem E.2 shows that this efficiency bound holds for $\mathcal{G}(M)$.

First, we derive the length of the optimal FLCI, which is the denominator of the expression in Corollary 3.3 in Armstrong and Kolesár (2018). The bias-sd ratio is

$$t_\delta = \frac{(1/2)(1-r)(2M)^{1-r}\delta^r\kappa^*(0)}{\lambda r(2M)^{1-r}\delta^{r-1}\kappa^*(0)} = (1/2)(1/r-1)\delta/\lambda.$$

Since optimizing over the bandwidth is equivalent to optimizing over δ, it follows that the optimal FLCI has length

$$\begin{aligned}\inf_\delta 2\,\mathrm{cv}_{1-\alpha}(t_\delta)\cdot\lambda\omega'(\delta) &= \inf_\delta 2\,\mathrm{cv}_{1-\alpha}(t_\delta)\cdot\lambda r(2M)^{1-r}\delta^{r-1}\kappa^*(0)\\ &= \inf_\delta 2\,\mathrm{cv}_{1-\alpha}(t_\delta)\cdot\lambda r(2M)^{1-r}t_\delta^{r-1}\lambda^{r-1}(1/r-1)^{1-r}2^{r-1}\kappa^*(0)\\ &= \lambda^r M^{1-r}r(1/r-1)^{1-r}\kappa^*(0)\inf_\delta 2\,\mathrm{cv}_{1-\alpha}(t_\delta)\cdot t_\delta^{r-1}.\end{aligned}$$

Plugging in $\lambda = [\sigma^2(0)/(nd)]^{1/2}$ and $S(k^*)^n B(k^*)^{1-r} = [\sigma^2(0)/d]^{r/2}r^r(1-r)^{1-r}\kappa^*(0)$ gives $2n^{-r/2}M^{1-r}S(k^*)^r B(k^*)^{1-r}\inf_\delta \mathrm{cv}_{1-\alpha}(t_\delta)\cdot t_\delta^{r-1}$, which is the asymptotic length of the CI given in Theorem E.1 with k and h chosen optimally.

The lower bound given the numerator of the expression in Corollary 3.3 in Armstrong and Kolesár (2018) is

$$\int_{z=-\infty}^{z_{1-\alpha}}\omega(2\lambda(z_{1-\alpha}-z))\,dz = (2M)^{1-r}\kappa^*(0)2^r\lambda^r\int_{z=-\infty}^{z_{1-\alpha}}(z_{1-\alpha}-z)^r\,dz.$$

Plugging in $\lambda = [\sigma^2(0)/(nd)]^{1/2}$ and $S(k^*)^r B(k^*)^{1-r} = [\sigma^2(0)/d]^{r/2}r^r(1-r)^{1-r}\kappa^*(0)$ gives $2n^{-r/2}M^{1-r}\frac{S(k^*)^r B(k^*)^{1-r}}{r^r(1-r)^{1-r}}\int_{z=-\infty}^{z_{1-\alpha}}(z_{1-\alpha}-z)^r\,dz$, which is the asymptotic lower bound given in Theorem E.2.

References

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

Armstrong, T. B. and Kolesár, M. (2019). Sensitivity analysis using approximate moment condition models. ArXiv: 1808.07387.

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

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

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.

Guerre, E., Perrigne, I., and Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68(3):525–574.

Lepski, O. V. and Tsybakov, A. (2000). Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probability Theory and Related Fields*, 117(1):17–48.

Low, M. G. (1995). Bias-variance tradeoffs in functional estimation problems. *The Annals of Statistics*, 23(3):824–835.

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

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.

Sacks, J. and Ylvisaker, D. (1981). Asymptotically optimum kernels for density estimation at a point. *The Annals of Statistics*, 9(2):334–346.

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

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Table S1: Kernel constants for standard deviation and maximum bias of local polynomial regression estimators of order q for selected kernels. Inference at a boundary point

Kernel ($k(u)$)	q	$\int_0^1 k_q^*(u)^2\,\mathrm{d}u$	$\mathcal{B}^{\mathrm{T}}_{p,q}(k) = \int_0^1 \lvert u^p k_q^*(u)\rvert\,\mathrm{d}u$ $p=1$	$p=2$	$p=3$	$\mathcal{B}^{\mathrm{H\ddot{o}l}}_{p,q}(k)$ $p=1$	$p=2$	$p=3$
Uniform	0	1	$\frac{1}{2}$			$\frac{1}{2}$		
$\mathrm{I}\{\lvert u\rvert \le 1\}$	1	4	$\frac{16}{27}$	$\frac{59}{162}$		$\frac{8}{27}$	$\frac{1}{6}$	
	2	9	0.7055	0.4374	0.3294	0.2352	$\frac{216}{3125}$	$\frac{1}{20}$
Triangular	0	$\frac{4}{3}$	$\frac{1}{3}$			$\frac{1}{3}$		
$(1-\lvert u\rvert)_+$	1	$\frac{24}{5}$	$\frac{3}{8}$	$\frac{3}{16}$		$\frac{27}{128}$	$\frac{1}{10}$	
	2	$\frac{72}{7}$	0.4293	0.2147	0.1400	0.1699	$\frac{32}{729}$	$\frac{1}{35}$
Epanechnikov	0	$\frac{6}{5}$	$\frac{3}{8}$			$\frac{3}{8}$		
$\frac{3}{4}(1-u^2)_+$	1	4.498	0.4382	0.2290		0.2369	$\frac{11}{95}$	
	2	9.816	0.5079	0.2662	0.1777	0.1913	0.0508	$\frac{15}{448}$

Table S2: Kernel constants for standard deviation and maximum bias of local polynomial regression estimators of order q for selected kernels. Inference at an interior point.

Kernel	q	$\int_{-1}^1 k_q^*(u)^2\,\mathrm{d}u$	$\mathcal{B}^{\mathrm{T}}_{p,q}(k) = \int_{-1}^1 \lvert u^p k_q^*(u)\rvert\,\mathrm{d}u$ $p=1$	$p=2$	$p=3$	$\mathcal{B}^{\mathrm{H\ddot{o}l}}_{p,q}(k)$ $p=1$	$p=2$	$p=3$
Uniform	0	$\frac{1}{2}$	$\frac{1}{2}$			$\frac{1}{2}$		
$\mathrm{I}\{\lvert u\rvert \le 1\}$	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$		$\frac{1}{2}$	$\frac{1}{3}$	
	2	$\frac{9}{8}$	0.4875	0.2789	0.1975	0.2898	0.0859	$\frac{1}{16}$
Triangular	0	$\frac{2}{3}$	$\frac{1}{3}$			$\frac{1}{3}$		
$(1-\lvert u\rvert)_+$	1	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{6}$		$\frac{1}{3}$	$\frac{1}{6}$	
	2	$\frac{456}{343}$	0.3116	0.1399	0.0844	0.2103	0.0517	$\frac{8}{245}$
Epanechnikov	0	$\frac{3}{5}$	$\frac{3}{8}$			$\frac{3}{8}$		
$\frac{3}{4}(1-u^2)_+$	1	$\frac{3}{5}$	$\frac{3}{8}$	$\frac{1}{5}$		$\frac{3}{8}$	$\frac{1}{5}$	
	2	$\frac{5}{4}$	0.3603	0.1718	0.1067	0.2347	0.0604	$\frac{5}{128}$

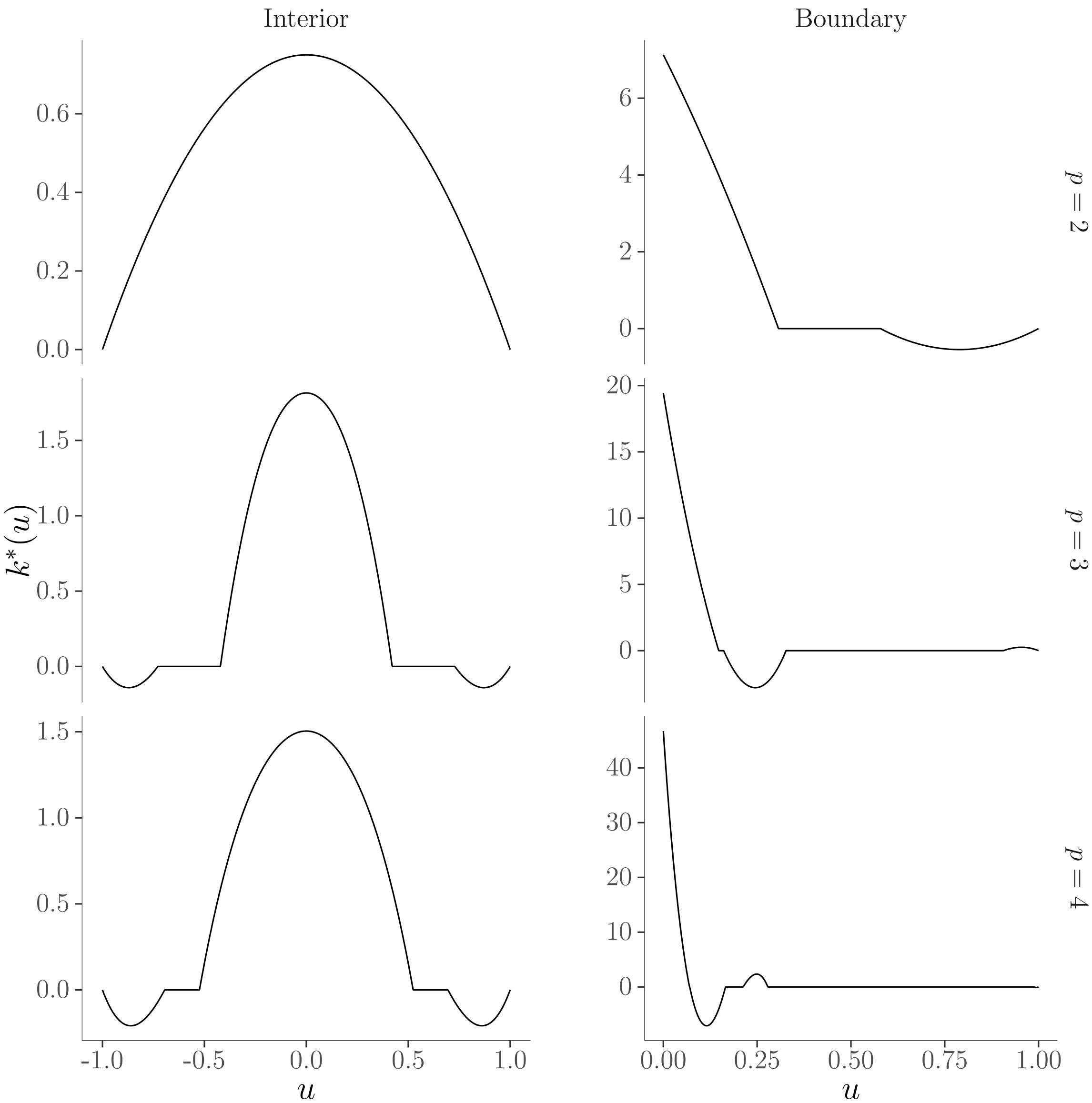

Figure S1: Optimal equivalent kernels for Taylor class $\mathcal{F}_{\mathrm{T},p}(M)$ on the interior, and in the boundary, rescaled to be supported on $[0, 1]$ on the boundary and $[-1, 1]$ in the interior.

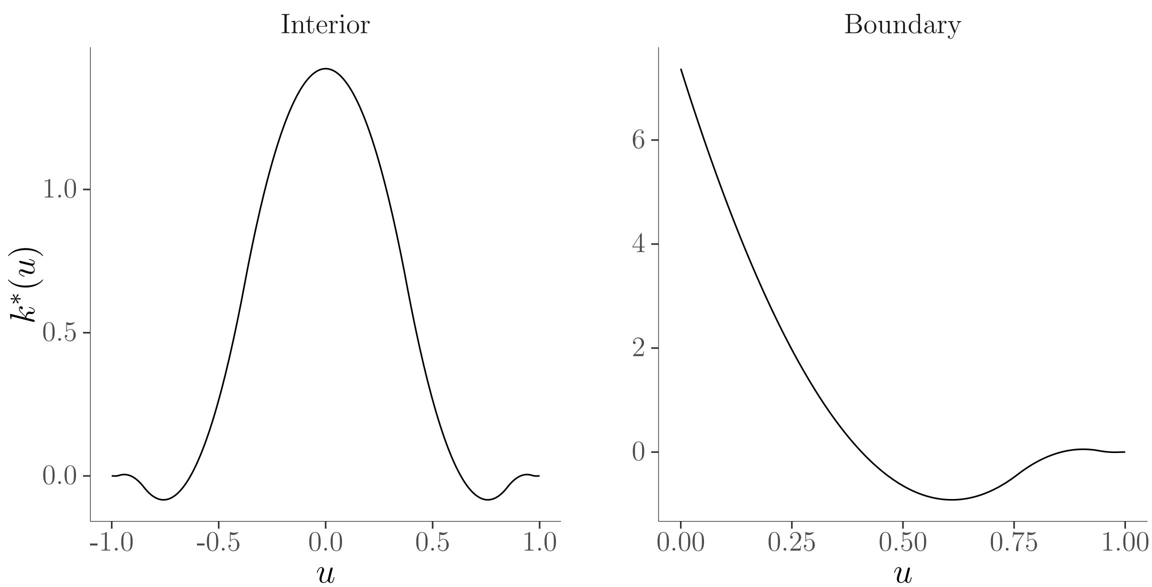

Figure S2: Optimal equivalent kernels for Hölder class $\mathcal{F}_{\text{Höl},2}(M)$ on the interior, and in the boundary, rescaled to be supported on $[0, 1]$ on the boundary and $[-1, 1]$ in the interior.