Supplemental Appendix for
SIMPLE AND HONEST CONFIDENCE INTERVALS IN NONPARAMETRIC REGRESSION

By

Timothy B. Armstrong and Michal Kolesár

June 2016
Revised October 2016

COWLES FOUNDATION DISCUSSION PAPER NO. 2044S

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

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

Timothy B. Armstrong* Michal Kolesár†

Yale University Princeton University

October 5, 2016

Appendix B Verification of regularity conditions in examples

We verify the conditions (1), (8) and (10) in some applications.

B.1 Gaussian white noise model

The approximation (8) holds as an exact equality (i.e. with the $o(1)$ term equal to zero) 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

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

*email: timothy.armstrong@yale.edu

†email: mkolesar@princeton.edu

We are interested in estimating the linear functional $T(f)$ where f is known to be in the class $\mathcal{F} = \{f\colon J_2(f) \le C\}$ 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

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

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

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

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

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

Thus, Equation (1) holds with $\gamma_b = s_2 - s_0$, $\gamma_s = s_1 - s_0$,

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

and $S(k) = \sigma\|Kk\|$ and with both $o(1)$ terms equal to zero. This implies that (8) holds with the $o(1)$ term equal to zero, since the estimator is normally distributed.

B.2 Local polynomial estimators in fixed design regression

This section proves Theorem 3.1 and Equation (13) in Section 3.

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 $\mathcal{F}_{\mathrm{T},p}(M)$ and $\mathcal{F}_{\mathrm{Höl},p}(M)$ 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 $j = 1, \ldots, p-1$, so let us assume that $w(\cdot)$ satisfies these conditions. For $f \in \mathcal{F}_{\mathrm{T},p}(M)$, we can write $f(x) = \sum_{j=0}^{p-1} x^j f^{(j)}(0)/j! + r(x)$ with $|r(x)| \le M|x|^p/p!$. 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 $\overline{\operatorname{bias}}_{\mathcal{F}_{\mathrm{T},p}}(\hat{T}) = \sum_{i=1}^n M|w(x_i)x|^p/p!$.

For $f \in \mathcal{F}_{\mathrm{Höl},p}(M)$, 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 $f(0) = f'(0) = \cdots = f^{(p-1)}(0)$. This allows us to apply the following lemma.

Lemma B.1. *Let ν be a finite measure on $\mathbb{R}$ (with the Lebesgue σ-algebra) with finite support and 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 $f(0) = f'(0) = f''(0) = \cdots = f^{(p-1)}(0) = 0$. 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

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

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

$$f(x) = \int_{t_1=0}^x \int_{t_2=0}^{t_1} \cdots \int_{t_p=0}^{t_{p-1}} f^{(p)}(t_p)\,dt_p \cdots dt_2 dt_1 = \int_{s=0}^x \frac{f^{(p)}(s)(x-s)^{p-1}}{(p-1)!}\,ds.$$

Thus, by Fubini's Theorem,

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

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.1 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 $f^{(p)}(s) = M\,\text{sign}(\bar{w}_{p,\nu}(s))$, which gives $\overline{\text{bias}}_{\mathcal{F}_{\text{Höl},p}(M)}(\hat{T}) = \int |\bar{w}_{p,\nu}(s)|\,ds$.

We collect these results in the following theorem.

Theorem B.1. *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 $j = 1, \ldots, p-1$,*

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

where $\bar{w}_{p,\nu}(s)$ is as defined in Lemma B.1 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,

$$\begin{aligned}\int_{s=t}^\infty \overline{w}_{q,\nu}(s)\,ds &= \int_{s=t}^\infty \int_{x=s}^\infty \frac{w(x)(x-s)^{q-1}}{(q-1)!}\,d\nu(x)ds = \int_{x=t}^\infty \int_{s=t}^x \frac{w(x)(x-s)^{q-1}}{(q-1)!}\,dsd\nu(x) \\ &= \int_{x=t}^\infty w(x)\left[\frac{-(x-s)^q}{q!}\right]_{s=t}^x d\nu(x) = \int_{x=t}^\infty \frac{w(x)(x-t)^q}{q!}\,d\nu(x) = \bar{w}_{q+1,\nu}(t).\end{aligned} \tag{S1}$$

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 (S1), 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 (S1) 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 $0 \leq \underline{s} < \overline{s} < \tilde{s}$. Since $w(x)$ is positive for small enough x and changes signs only once, this means that, for some $s^* \leq \tilde{s}$, we have $w(x) \geq 0$ for $0 \leq x \leq s^*$ 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 $j = 1, \ldots, p-1$. We further assume

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

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

and

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

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

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* (8) *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.1 and the second equality follows from the fact that for any function $g(u)$ that is bounded over u in compact sets,

$$\begin{aligned}
&\left| \frac{1}{nh} \sum_{i=1}^{n} \tilde{k}_n(x_i/h) g(x_i/h) - d \int_{\mathcal{X}} k(u) g(u)\, du \right| \\
&\leq \left| \frac{1}{nh} \sum_{i=1}^{n} \tilde{k}(x_i/h) g(x_i/h) - d \int_{\mathcal{X}} k(u) g(u)\, du \right| + \frac{1}{nh} \sum_{i=1}^{n} \left| \tilde{k}_n(x_i/h) g(x_i/h) - \tilde{k}(x_i/h) g(x_i/h) \right| \\
&\qquad \leq o(1) + \frac{1}{nh} \sum_{i=1}^{n} I(|x_i/h| \leq K_s) \sup_{u \in [-K_s, K_s]} |g(u)| \cdot \sup_{u \in [-K_s, K_s]} |\tilde{k}_n(u) - \tilde{k}(u)| = o(1), \quad \text{(S2)}
\end{aligned}$$

where the second line follows by triangle inequality, the third line by Assumption 3.1 applied to the first summand, and the last equality follows by Assumption 3.1 applied to the first term, and Assumption B.1 applied to the last term.

For Hölder classes,

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

by Theorem B.1 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 $s > 0$,

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

Thus, by Equation (S2), 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,

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

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

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

By the triangle inequality,

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

where the equality follows by Assumption 3.1 applied to the second summand and the second term of the first summand, and Assumptions 3.2 and B.1 applied to the first term of the first summand. This gives the second display in the theorem.

The last statement (verification of Equation (8)) follows immediately from continuity of $\tilde{R}$ for these performance criteria, since $\hat{T}$ is distributed normal with constant variance. □

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.1 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 Equation (12). Theorem 3.1 and Equation (13) then follow immediately by applying Theorem B.2 with this choice of $\tilde{k}_n$ and $\tilde{k}$.

Appendix C Regression discontinuity with different bandwidths on either side of the cutoff

This appendix calculates the efficiency gain from using different bandwidths on either side of the cutoff. We state a result in a more general setup than that considered in Section 4.

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

$$R(\hat{T}(h_+, h_-; k)) = \tilde{R}(B(k)(h_-^{\gamma_b} + h_+^{\gamma_b}), n^{-1/2}(S_+(k)^2 h_+^{2\gamma_s} + S_-(k)^2 h_-^{2\gamma_s})^{1/2})(1 + o(1)), \quad \text{(S3)}$$

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 $\tilde{R}$ scales linearly with its arguments.

In the RD application considered in Section C, if Assumptions 3.1 holds, u_i is normally distributed, and $\sigma_+^2(x)$ and $\sigma_-^2(0)$ are right- and left-continuous at 0, then Condition (S3) 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 $B(k) = -M \int_0^\infty u^2 k_1^*(u) \mathrm{d}u/2$.

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,

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

Substituting $h_+ = \rho h_-$ and $h_- = (tn^{-1/2}(S_+(k)^2 \rho^{2\gamma_s} + S_-(k)^2)^{1/2} B(k)^{-1}(1 + \rho^{\gamma_b})^{-1})^{1/(\gamma_b - \gamma_s)}$ into (S3) and using linearity of $\tilde{R}$ gives

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

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 $t_R^* = \text{argmin}_{t>0} \tilde{R}(t,1)$, and depends only on the performance criterion. The optimal bandwidth ratio ρ is given by

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

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

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

In the RD application in Section 4, $\varsigma(k) = \sigma_+(0)/\sigma_-(0)$, 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.

Appendix D Optimal kernels for inference at a point

Here we give details of optimal kernel calculations discussed in Section 3.1 in the main text.

The optimal equivalent kernel under the Taylor class $\mathcal{F}_{\text{T},p}(M)$ solves Equation 15 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

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

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 $\mathcal{F}_{\mathrm{Höl},2}(M)$ 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 $f^{\mathrm{Int}}_{\mathrm{Höl},2}(u)/\int_{-\infty}^{\infty} f^{\mathrm{Int}}_{\mathrm{Höl},2}(u)\,\mathrm{d}u$, where

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

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 $f^{\mathrm{Bd}}_{\mathrm{Höl},2}(u)/\int_{-\infty}^{\infty} f^{\mathrm{Bd}}_{\mathrm{Höl},2}(u)\,\mathrm{d}u$, where

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

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.

References

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

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

			$\mathcal{B}^{\mathrm{T}}_{p,q}(k) = \int_0^1 \lvert u^p k^*_q(u)\rvert \,\mathrm{d}u$			$\mathcal{B}^{\mathrm{Höl}}_{p,q}(k)$		
Kernel ($k(u)$)	q	$\int_0^1 k^*_q(u)^2 \,\mathrm{d}u$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
Uniform	0	1	$\frac{1}{2}$			$\frac{1}{2}$		
$1(\lvert u\rvert \le 1)$	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 S1: Kernel constants for standard deviation and maximum bias of local polynomial regression estimators of order q for selected kernels. Functional of interest is value of f at a boundary point.

			$\mathcal{B}^{\mathrm{T}}_{p,q}(k) = \int_{-1}^1 \lvert u^p k^*_q(u)\rvert \,\mathrm{d}u$			$\mathcal{B}^{\mathrm{Höl}}_{p,q}(k)$		
Kernel	q	$\int_{-1}^1 k^*_q(u)^2 \,\mathrm{d}u$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
Uniform	0	$\frac{1}{2}$	$\frac{1}{2}$			$\frac{1}{2}$		
$1(\lvert u\rvert \le 1)$	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	$\frac{5}{128}$

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

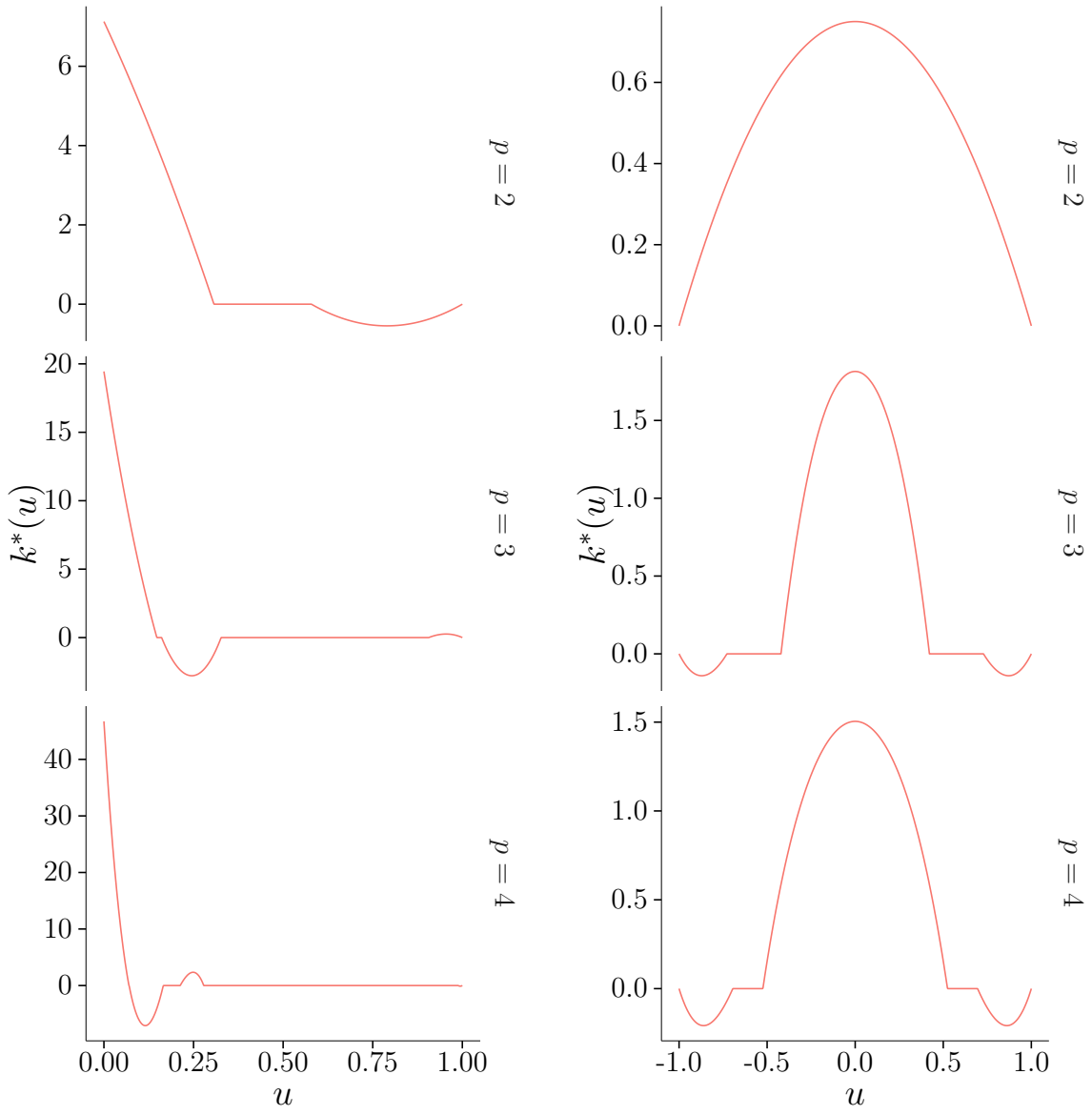

Figure S1: Optimal equivalent kernels for Taylor class $\mathcal{F}_{\mathrm{T},p}(M)$ on the boundary (left), and in the interior (right), rescaled to be supported on $[0, 1]$ on the boundary and $[-1, 1]$ in the interior.

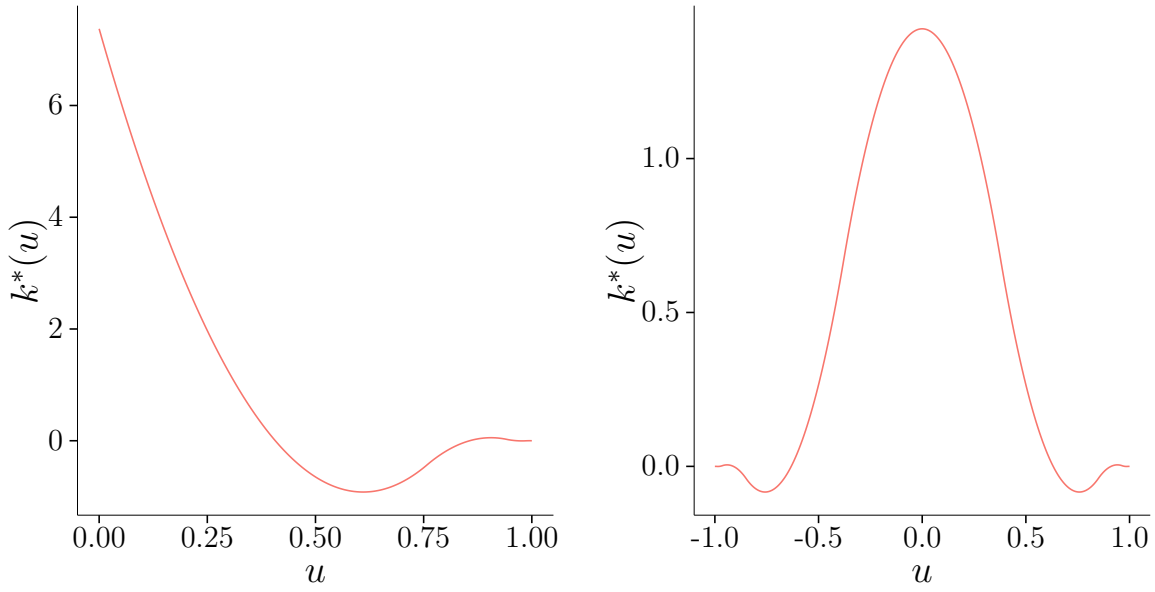

Figure S2: Optimal equivalent kernels for Hölder class $\mathcal{F}_{\text{Höl},2}(M)$ on the boundary (left), and in the interior (right), rescaled to be supported on $[0, 1]$ on the boundary and $[-1, 1]$ in the interior.