ONLINE APPENDIX: BIAS-AWARE INFERENCE IN FUZZY REGRESSION DISCONTINUITY DESIGNS

Claudia Noack Christoph Rothe

A. PROOF OF THEOREM 4

We begin by using arguments similar to that of Abadie and Imbens (2006, Theorem 6) to show that

$$s_M(h_M) = \widehat{s}_M(h_M)(1 + o_{P,\mathcal{F}}(1)). \tag{A.1}$$

To simplify the presentation, we suppress various quantities' dependence on c in this proof. For example, we write $\widehat{s}^2_M(h_M)$ instead of $\widehat{s}^2_M(h_M(c), c)$, etc. We also define

$$q_i(h_M) = \frac{w_i(h_M)^2}{\sum_{i=1}^n w_i(h_M)^2 \sigma^2_{M,i}},$$

so that $\sum_{i=1}^n q_i(h_M)\widehat{\sigma}^2_{M,i} = \widehat{s}^2_M(h_M)/s^2_M(h_M)$. We note that $\max_{i=1,\ldots,n} q_i(h_M) = o_{P,\mathcal{F}}(1)$ and $\sum_{i=1}^n q_i(h_M) = O_{P,\mathcal{F}}(1)$ by the same arguments as in the proof of Theorem 2, and the fact that the variance terms $\sigma^2_{M,i}$ are uniformly bounded and bounded away from zero, respectively.

The proof for the case that Assumption LL1 holds is rather straightforward. As the kernel has compact support by Assumption 1, and h_M is bounded as a function of n, the number of support points at which $q_i(h_M) > 0$ is finite. It follows that $\sum_{i=1}^n \mathbf{1}\{X_i = x\}$ tends to infinity for all support points x with $q_i(h_M) > 0$ if $X_i = x$. Moreover, it holds that

$$\max_{i:q_i(h_M)>0} |\widehat{\sigma}^2_{M,i} - \sigma^2_{M,i}| = o_{P,\mathcal{F}}(1).$$

Since $\sum_{i=1}^n q_i(h_M) = O_{P,\mathcal{F}}(1)$ and $q_i(h_M)$ is positive, the statement of the theorem then

First Version: June 11, 2019. This Version: August 9, 2022. Claudia Noack, Nuffield College and Department of Economics, University of Oxford, email: claudia.noack@economics.ox.ac.uk, website: http://claudianoack.github.io. Christoph Rothe, Department of Economics, University of Mannheim, 68131 Mannheim, Germany, email: rothe@vwl.uni-mannheim.de, website: http://www.christophrothe.net.

follows because

$$\left|\frac{\widehat{s}^2_M(h_M)}{s^2_M(h_M)} - 1\right| = \left|\sum_{i=1}^n q_i(h_M)(\widehat{\sigma}^2_{M,i} - \sigma^2_{M,i})\right| \leq \max_{i:q_i(h_M)>0} |\widehat{\sigma}^2_{M,i} - \sigma^2_{M,i}| \cdot \sum_{i=1}^n q_i(h_M) = o_{P,\mathcal{F}}(1).$$

Now suppose that Assumption LL2 holds. In this case there are no ties in the data, and each unit has exactly $R_i = R$ nearest neighbors, with probability 1. We thus define the $R \times 2$ matrix $\widetilde{X}_{-i} = (\widetilde{X}'_{r_1}, \ldots, \widetilde{X}'_{r_R})'$, where $r_1, \ldots, r_R$ are the indices of the R nearest neighbors of unit i, and $\widetilde{X}_i = (1, X_i)$, let $H_i = \widetilde{X}_i(\widetilde{X}'_{-i}\widetilde{X}_{-i})^{-1}\widetilde{X}'_i$, and write $v_j(X_i) = \widetilde{X}_i(\widetilde{X}'_{-i}\widetilde{X}_{-i})^{-1}\widetilde{X}'_{-i}e_j$ with e_j the jth R-dimensional unit-vector. With W_i a generic random variable, we also write $\breve{W}_i = W_i - \sum_{j\in\mathcal{R}_i} v_j(X_i)W_j$. In the following, we use repeatedly that

$$\sum_{j\in\mathcal{R}_i} v_j(X_i) = 1, \quad \sum_{j\in\mathcal{R}_i} v_j(X_i)(X_j - X_i) = 0, \text{ and } \sum_{j\in\mathcal{R}_i} v_j(X_i)^2 = H_i,$$

which follows from basic algebra. Next, note that the variance estimators $\widehat{\sigma}^2_{M,i}$, $i = 1, \ldots, n$, are all well-defined with probability one, as the running variable is continuously distributed with a bounded density function. Also, recall that $M_i = Y_i - cT_i$ and $\mathbb{E}(M_i|X_i) = \mu_M(X_i) = \mu_Y(X_i) - c\mu_T(X_i)$, put $\varepsilon_i = M_i - \mu_M(X_i)$, and note that $\varepsilon_i = \varepsilon_{Y,i} - c\varepsilon_{T,i} = (Y_i - \mu_Y(X_i)) - c(T_i - \mu_T(X_i))$. The variance estimators can then be written as

$$\widehat{\sigma}^2_{M,i} = \frac{\breve{M}^2_i}{1 + H_i} = \frac{1}{1 + H_i}\left(\breve{\mu}_M(X_i) + \varepsilon_i - \sum_{j\in\mathcal{R}_i} v_j(X_i)\varepsilon_i\right)^2.$$

It then suffices to show the following:

$$\left|\sum_{i=1}^n q_i(h_M)(\sigma^2_{M,i} - \mathbb{E}[\widehat{\sigma}^2_{M,i}|\mathcal{X}_n])\right| = o_{P,\mathcal{F}}(1) \text{ and} \tag{A.2}$$

$$\left|\sum_{i=1}^n q_i(h_M)(\widehat{\sigma}^2_{M,i} - \mathbb{E}[\widehat{\sigma}^2_{M,i}|\mathcal{X}_n])\right| = o_{P,\mathcal{F}}(1), \tag{A.3}$$

We begin by noting that (A.2) follows from the triangle inequality and the fact that $\sum_{i=1}^n q_i(h_M) = O_{P,\mathcal{F}}(1)$ if

$$\max_{i=1,\ldots,n} |\sigma^2_{M,i} - \mathbb{E}[\widehat{\sigma}^2_{M,i}|\mathcal{X}_n]| = o_{P,\mathcal{F}}(1). \tag{A.4}$$

To show (A.4), note that

$$\begin{aligned}\mathbb{E}\left[\widehat{\sigma}^2_{M,i}|\mathcal{X}_n\right] &= \frac{1}{1+H_i}\mathbb{E}\left[\left(\breve{\mu}_M(X_i)+\varepsilon_i-\sum_{j\in\mathcal{R}_i} v_j(X_i)\varepsilon_j\right)^2|\mathcal{X}_n\right] \\ &= \frac{1}{1+H_i}\left(\breve{\mu}_M(X_i)^2+\sigma^2_{M,i}+\sum_{j\in\mathcal{R}_i} v_j(X_i)^2\sigma^2_{M,j}\right) \\ &= \sigma^2_{M,i}+\frac{1}{1+H_i}\left(\breve{\mu}_M(X_i)^2+\sum_{j\in\mathcal{R}_i} v_j(X_i)^2(\sigma^2_{M,j}-\sigma^2_{M,i})\right).\end{aligned}$$

Here the second equality holds because ε_i and ε_j are independent if $i \neq j$, and are zero in expectation; and the third equality holds because $\sum_{j\in\mathcal{R}_i} v_j(X_i)^2 = H_i$. As the running variable density is uniformly bounded away from zero, it follows from the proof of Theorem 6 in Abadie and Imbens (2006) that

$$x_{\max} \equiv \max_{i=1,\ldots,n}\max_{r\in\mathcal{R}_i}|X_i - X_r| = o_{P,\mathcal{F}}(1). \tag{A.5}$$

Since $\sigma^2_{M,i}$ is uniformly Lipschitz continuous with some constant L_σ by Assumption 1, we then have that

$$\begin{aligned}\max_i \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_j(X_i)^2(\sigma^2_{M,j}-\sigma^2_{M,i})\right) &\leq L_\sigma x_{\max}\max_i \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_j(X_i)^2\right) \\ &\leq L_\sigma x_{\max}\max_i \frac{H_i}{1+H_i} = o_{P,\mathcal{F}}(1).\end{aligned}$$

To show (A.4), it thus only remains to show that

$$\max_i \frac{1}{1+H_i}\breve{\mu}_M(X_i)^2 = o_{P,\mathcal{F}}(1). \tag{A.6}$$

To do so, note that

$$\begin{aligned}&\max_{i\in\{1,\ldots,n\}}\left(\mu_M(X_i)-\sum_{j\in\mathcal{R}_i} v_j(X_i)\mu_M(X_j)\right) \\ &= \max_{i\in\{1,\ldots,n\}}\left(\mu_M(X_i)-\sum_{j\in\mathcal{R}_i} v_j(X_i)\left(\mu_M(X_i)+\mu'_M(X_i)(X_j-X_i)+\frac{1}{2}\mu''_M(\mathring{X}_{i,j})(X_j-X_i)^2\right)\right) \\ &= \frac{1}{2}\max_{i\in\{1,\ldots,n\}}\sum_{j\in\mathcal{R}_i} v_j(X_i)\mu''_M(\mathring{X}_{i,j})(X_j-X_i)^2.\end{aligned}$$

Here the first equality follows from a second order expansion, with $\mathring{X}_{i,j}$ some value between X_i and X_j, where $j \in \mathcal{R}_i$; and the second equality follows as $\sum_{j\in\mathcal{R}_i} v_j(X_i) = 1$ and $\sum_{j\in\mathcal{R}_i} v_j(X_i)(X_j - X_i) = 0$. We then find that

$$\begin{aligned}
\max_{i\in\{1,\ldots,n\}} \frac{1}{1+H_i}\breve{\mu}_M(X_i)^2 &= \frac{1}{4}\max_{i\in\{1,\ldots,n\}} \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_j(X_i)\mu''_M(\mathring{X}_{i,j})(X_j - X_i)^2\right)^2 \\
&\leq \frac{R}{4}\max_{i\in\{1,\ldots,n\}} \frac{1}{1+H_i}\sum_{j\in\mathcal{R}_i} v_j(X_i)^2\mu''_M(\mathring{X}_{i,j})^2(X_j - X_i)^4 \\
&\leq \frac{RB_M^2 x_{\max}^4}{4}\max_{i\in\{1,\ldots,n\}} \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_j(X_i)^2\right) = o_{P,\mathcal{F}}(1).
\end{aligned}$$

Here first inequality follows from Cauchy-Schwarz as the cardinality of $\mathcal{R}_i$ is R; and the second inequality follows as all the terms of the sum are positive, $\mu''(\mathring{X}_{i,j})^2$ is bounded by B_M^2, and $(X_j - X_i)^4 \leq x_{\max}^4$ for all i and $j \in \mathcal{R}_i$. The final equality follows because $\sum_{j\in\mathcal{R}_i} v_j(X_i)^2 = H_i$, and $H_i/(1+H_i) \leq 1$ for all $i \in \{1,\ldots,n\}$, and $x_{\max} = o_{P,\mathcal{F}}(1)$. This completes the proof of the statement (A.2).

To show that (A.3) holds, write $\widetilde{q}_i(h_M) = q_i(h_M)(1+H_i)^{-1}$. Note that since $|\widetilde{q}_i(h_M)| \leq |q_i(h_M)|$, it follows from Theorem A.1 that $\max_{i=1,\ldots,n}\widetilde{q}_i(h_M) = o_{P,\mathcal{F}}(1)$ and $\sum_{i=1}^n \widetilde{q}_i(h_M) = O_{P,\mathcal{F}}(1)$. We write this quantity the sum of five terms:

$$\begin{aligned}
&\sum_{i=1}^n q_i(h_M)(\widehat{\sigma}^2_{M,i} - \mathbb{E}[\widehat{\sigma}^2_{M,i}|\mathcal{X}_n]) \\
&= \sum_{i=1}^n \widetilde{q}_i(h_M)(\varepsilon_i^2 - \sigma^2_{M,i}) + \sum_{i=1}^n \widetilde{q}_i(h_M)\sum_{j\in\mathcal{R}_i} v_j^2(X_i)(\varepsilon_j^2 - \sigma^2_{M,j}) \\
&\quad + 2\sum_{i=1}^n \widetilde{q}_i(h_M)\varepsilon_i\sum_{j\in\mathcal{R}_i} v_j(X_i)\varepsilon_j + 2\sum_{i=1}^n \widetilde{q}_i(h_M)\breve{\mu}_M(X_i)\varepsilon_i - 2\sum_{i=1}^n \widetilde{q}_i(h_M)\breve{\mu}_M(X_i)\sum_{j\in\mathcal{R}_i} v_j(X_i)\varepsilon_j \\
&\equiv G_1 + G_2 + 2G_3 + 2G_4 + 2G_5.
\end{aligned}$$

It is easy to see that these five terms all have mean zero conditional on $\mathcal{X}_n$. It thus suffices to show that their second moments converge uniformly over the function class $\mathcal{F}$ to zero. In the following derivations, we write C for a generic positive constant whose value might differ between equations.

For the first term, we have that

$$\mathbb{V}(G_1|\mathcal{X}_n) = \sum_{i=1}^{n} \widetilde{q}_i(h_M)^2 \mathbb{E}[(\varepsilon_i^2 - \sigma_{M,i}^2)^2|\mathcal{X}_n] \leq C \max_{i=1,\dots,n} \widetilde{q}_i(h_M) \cdot \sum_{i=1}^{n} \widetilde{q}_i(h_M) = o_{P,\mathcal{F}}(1),$$

where the inequality follows from the bound on the fourth moment of ε_i and $\widetilde{q}_i(h_M)$ being positive, and the last equality follows since $\max_{i=1,\dots,n} \widetilde{q}_i(h_M) \sum_{i=1}^{n} \widetilde{q}_i(h_M) = o_{P,\mathcal{F}}(1)$.

We now turn to the second term, and note that by independent sampling

$$\begin{aligned}
\mathbb{V}(G_2|\mathcal{X}_n) &= \sum_{i=1}^{n}\sum_{l=1}^{n} \widetilde{q}_l(h_M)\widetilde{q}_i(h_M) \sum_{j\in\mathcal{R}_i}\sum_{k\in\mathcal{R}_l} v_k^2(X_l)v_j^2(X_i)\mathbb{E}\left[(\varepsilon_j^2 - \sigma_{M,j}^2)(\varepsilon_k^2 - \sigma_{M,k}^2)|\mathcal{X}_n\right] \\
&= \sum_{i=1}^{n} \sum_{l:\mathcal{R}_i\cap R_l\neq\emptyset} \widetilde{q}_l(h_M)\widetilde{q}_i(h_M) \sum_{j\in\mathcal{R}_i}\sum_{k\in\mathcal{R}_l} v_k^2(X_l)v_j^2(X_i)\mathbb{E}\left[(\varepsilon_j^2 - \sigma_{M,j}^2)(\varepsilon_k^2 - \sigma_{M,k}^2)|\mathcal{X}_n\right] \\
&\leq \sum_{i=1}^{n} \sum_{l:\mathcal{R}_i\cap\mathcal{R}_l\neq\emptyset} \widetilde{q}_l(h_M)\widetilde{q}_i(h_M) \sum_{j\in\mathcal{R}_i}\sum_{k\in\mathcal{R}_l} v_k^2(X_l)v_j^2(X_i)\mathbb{E}\left[(\varepsilon_j^2 - \sigma_{M,j}^2)^2|\mathcal{X}_n\right].
\end{aligned}$$

Using that ε_i has bounded fourth moments, that $\sum_{k\in\mathcal{R}_l} v_k^2(X_i) = H_i$, and that $H_i/(1+H_i) \leq 1$ for all $i \in \{1,\dots,n\}$, we further deduce that

$$\mathbb{V}(G_2|\mathcal{X}_n) \leq C \sum_{i=1}^{n} q_i(h_M) \sum_{l:\mathcal{R}_i\cap\mathcal{R}_l\neq\emptyset} q_l(h_M).$$

Finally, note that the cardinality of the set $\{l : \mathcal{R}_i \cap \mathcal{R}_l \neq \emptyset\}$, which contains the indices of those units that share at least one common R-nearest neighbor with unit i, is bounded by $3R+1$ (this can be seen through a simple counting exercise). We thus have that

$$\mathbb{V}(G_2|\mathcal{X}_n) \leq C \sum_{i=1}^{n} q_i(h_M)(3R+1) \max_{j\in\{1,\dots,n\}} q_j(h_M) = o_{P,\mathcal{F}}(1).$$

We now consider the third term, which satisfies

$$\mathbb{V}(G_3|\mathcal{X}_n) = \sum_{i=1}^{n}\sum_{k=1}^{n} \widetilde{q}_i(h_M)\widetilde{q}_k(h_M) \sum_{j\in\mathcal{R}_i}\sum_{l\in\mathcal{R}_k} v_j(X_i)v_l(x_g)\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l|\mathcal{X}_n].$$

To proceed, note that $\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l|\mathcal{X}_n] = 0$ unless the four indices involved in this expression

can be grouped into two pairs that each have the same value. This means that

$$\begin{aligned}\mathbb{V}(G_3|\mathcal{X}_n) &\leq C\sum_{i=1}^{n}\left(\sum_{j\in\mathcal{R}_i}\widetilde{q}_i(h_M)^2 v_j(X_i)^2 + \sum_{j\in R_i: i\in R_j}\widetilde{q}_i(h_M)\widetilde{q}_j(h_M)v_i(X_j)v_j(X_i)\right)\\ &\leq C\max_{i\in\{1,\ldots,n\}}\widetilde{q}_i(h_M)\sum_{i=1}^{n}\widetilde{q}_i(h_M)\sum_{j\in\mathcal{R}_i}v_j(X_i)^2\\ &= C\max_{i\in\{1,\ldots,n\}}\widetilde{q}_i(h_M)\sum_{i=1}^{n}q_i(h_M)\frac{H_i}{1+H_i} = o_{P,\mathcal{F}}(1).\end{aligned}$$

For the fourth and fifth term, we can use arguments similar to those used for the three previous terms to show that that

$$\begin{aligned}\mathbb{V}(G_4|\mathcal{X}_n) &\leq CB_M^2 x_{\max}^4\sum_{i=1}^{n}\widetilde{q}_i(h_M)^2 = o_{P,\mathcal{F}}(1);\\ \mathbb{V}(G_5|\mathcal{X}_n) &\leq CB_M^2 x_{\max}^4\max_{i\in\{1,\ldots,n\}}\left(q_i(h_M)\frac{H_i}{1+H_i}\right)\sum_{i=1}^{n}\widetilde{q}_i(h_M) = o_{P,\mathcal{F}}(1).\end{aligned}$$

This completes the proof of the statement (A.3); and thus (A.1) holds, as claimed.

To complete our proof, we still need show that

$$\widehat{s}_M(\widehat{h}_M) = \widehat{s}_M(h_M)(1+o_{P,\mathcal{F}}(1)), \tag{A.7}$$

as this together with (A.1) implies the statement of Theorem 4. Under Assumption LL1, this follows from arguments similarly to those in the proof of Lemma A.1, and under Assumption LL2 follow from arguments analogous to those in the proof of Theorem E.1 in Armstrong and Kolesár (2020). We omit the details for brevity.

B. MORE GENERAL BANDWIDTH CHOICES

In the main body of the paper, the local linear regression estimators $\widehat{\tau}_M(h,c) = \widehat{\tau}_Y(h) - c\widehat{\tau}_T(h)$ on which our bias-aware AR CSs are based use the same bandwidth on each side of the cutoff, and also the same bandwidth for estimating τ_Y and τ_T; and the second derivatives of μ_Y and μ_T are bounded in absolute value by the same respective constant on either side of the cutoff. These features can all easily be relaxed. In particular, we can define a more general Hölder-type class of functions as

$$\mathcal{F}_H(B_+,B_-) = \{f_1(x)\mathbf{1}\{x\geq 0\} - f_0(x)\mathbf{1}\{x<0\} : \|f_1''\|_\infty \leq B_+, \|f_0''\|_\infty \leq B_-\},$$

define the class $\mathcal{F}_H^{\delta}(B_+, B_-)$ similarly, and then seek to obtain bias-aware AR CSs that are honest uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_{Y+}, B_{Y-}) \times \mathcal{F}_H^0(B_{T+}, B_{T-})$, based on the local linear regression estimator

$$\widehat{\tau}_M(\mathbf{h}, c) = \sum_{i=1}^{n} \left(w_{i,+}(h_{Y+}) - w_{i,-}(h_{Y-})\right) Y_i - c \sum_{i=1}^{n} \left(w_{i,+}(h_{T+}) - w_{i,-}(h_{T-})\right) T_i,$$

where $\mathbf{h} = (h_{T+}, h_{T-}, h_{Y+}, h_{Y-})$ is a vector of side- and function-specific bandwidths, and the weights $w_{i,+}(h)$ and $w_{i,-}(h)$ are as defined in the beginning of Appendix A in the main body of the paper. With such a setup, the explicit expression for the bound on the absolute value of the conditional bias of $\widehat{\tau}_M(\mathbf{h}, c)$ is

$$\begin{aligned}\overline{b}_M(\mathbf{h}, c) = &-\frac{B_{Y+}}{2} \sum_{i=1}^{n} w_{i,+}(h_{Y+}) X_i^2 - \frac{|c| B_{T+}}{2} \sum_{i=1}^{n} w_{i,+}(h_{T+}) X_i^2 \\ &+ \frac{B_{Y-}}{2} \sum_{i=1}^{n} w_{i,-}(h_{Y-}) X_i^2 + \frac{|c| B_{T-}}{2} \sum_{i=1}^{n} w_{i,-}(h_{T-}) X_i^2,\end{aligned}$$

and the conditional standard deviation of $\widehat{\tau}_M(\mathbf{h}, c)$ is

$$\begin{aligned}s_M(\mathbf{h}, c) = \Bigg(&\sum_{i=1}^{n} \left(w_{i,+}(h_{Y,+}) - w_{i,-}(h_{Y-})\right)^2 \sigma_{Y,i}^2 + c^2 \sum_{i=1}^{n} \left(w_{i,+}(h_{T+}) - w_{i,-}(h_{T-})\right)^2 \sigma_{T,i}^2 \\ &-2c \sum_{i=1}^{n} \left(w_{i,+}(h_{Y+}) - w_{i,-}(h_{Y-})\right) \left(w_{i,+}(h_{T+}) - w_{i,-}(h_{T-})\right) \sigma_{YT,i} \Bigg)^{1/2},\end{aligned}$$

with $\sigma_{Y,i}^2 = \mathbb{V}(Y_i|X_i)$, $\sigma_{T,i}^2 = \mathbb{V}(T_i|X_i)$, and $\sigma_{YT,i} = \mathbb{C}(Y_i, T_i|X_i)$ being conditional variance and covariance terms. A feasible standard error $\widehat{s}_M(\mathbf{h}, c)$ can be obtained by substituting nearest-neighbor estimates of the latter terms into the above expression for $s_M(\mathbf{h}, c)$. Letting $\widehat{\mathbf{h}}_M(c)$ be a feasible estimate of $\mathbf{h}_M(c) = \operatorname{argmin}_{\mathbf{h}} \operatorname{cv}_{1-\alpha}(r_M(\mathbf{h}, c)) \cdot s_M(\mathbf{h}, c)$, with $r_M(\mathbf{h}, c) = \overline{b}_M(\mathbf{h}, c)/s_M(\mathbf{h}, c)$, a generalization of our proposed bias-aware AR CS for θ is then given by

$$\mathcal{C}_{\text{ar}}^{\alpha} = \left\{ c : |\widehat{\tau}_M(\widehat{\mathbf{h}}_M(c), c)| \leq \operatorname{cv}_{1-\alpha}(\widehat{r}_M(\widehat{\mathbf{h}}_M(c), c)) \widehat{s}_M(\widehat{\mathbf{h}}_M(c), c)) \right\}.$$

A theoretical analysis of this CS would follow arguments that are fully analogous to those in the analysis of the CS in the main body of this paper, which only uses a single bandwidth, and would yield fully analogous results.

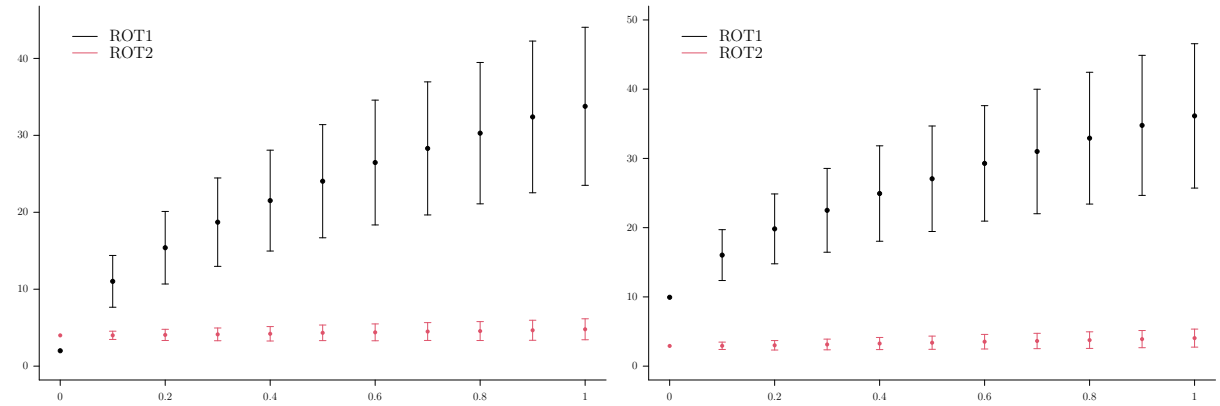

Figure 1: Mean (dots) and interquartile range (bars) of simulated ROT1 (black) and ROT2 (red) "rule-of-thumb" estimates of bound on absolute second derivative for $\sigma^2 \in \{0, .1, \ldots, 1\}$ and $\mu_Y(x) = x^2$ (left panel) and $\mu_Y(x) = x^2 - x^4$ (right panel)

C. PROPERTIES OF RULE-OF-THUMB SMOOTHNESS BOUNDS

In this appendix, we study the properties of two data-driven rules-of-thumb (ROT) for selecting the smoothness constants B_Y and B_T, which are both based on fitting global polynomial specifications on either side of the cutoff. For simplicity, we focus on the case of B_Y, but the arguments apply analogously to the case of B_T. To describe the two methods, let $g_k(x) = (1, x, \ldots, x^k, \mathbf{1}\{x \geq 0\}, \mathbf{1}\{x \geq 0\}x, \ldots, \mathbf{1}\{x \geq 0\}x^k)^\top$ be a vector of polynomials, define the function

$$\widetilde{\mu}_{Y,k}(x) = g_k(x)^\top \widehat{\gamma}_k, \text{ with } \widehat{\gamma}_k = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - g_k(X_i)^\top \gamma)^2,$$

and write $\mathcal{X}$ for the range of the realizations of the running variable.Armstrong and Kolesár (2020) then consider fourth-order polynomials, and propose the ROT value

$$\widehat{B}_{Y,\text{ROT1}} = \sup_{x \in \mathcal{X}} |\widetilde{\mu}''_{Y,4}(x)|.$$

Imbens and Wager (2019) mention a ROT in which the maximal curvature implied by a quadratic fit is multiplied by some moderate factor, say 2, to guard against overly optimistic values, yielding the rule-of-thumb value

$$\widehat{B}_{Y,\text{ROT2}} = 2 \sup_{x \in \mathcal{X}} |\widetilde{\mu}''_{Y,2}(x)|.$$

We refer to these estimators ROT1 and ROT2 in the following. Both Armstrong and Kolesár (2020) and Imbens and Wager (2019) caution that the respective rules cannot be expected to

provide universally adequate smoothness bounds, and should rather serve as a first guidance that is complemented with other approaches in a sensitivity analysis.

To get a better understanding of the relative properties of these two rules, we conduct two small Monte Carlo experiments in which the conditional expectation function is either $\mu_Y(x) = x^2$ or $\mu_Y(x) = x^2 - x^4$. With each function and each $\sigma^2 \in \{0, .1, .2, \ldots, 1\}$, we conduct $10,000$ runs in which we simulate $n = 1,000$ realizations of (Y_i, X_i) according to

$$Y_i = \mu_Y(X_i) + \varepsilon_i, \quad X_i \sim U[-1, 1], \quad \varepsilon_i \sim N(0, \sigma^2), \quad X_i \perp \varepsilon_i,$$

and calculate both ROT values. If $\mu_Y(x) = x^2$, the true smallest upper bound on the absolute second derivative is $B_Y = 2$, whereas if $\mu_Y(x) = x^2 - x^4$, we have that $B_Y = 10$. In both cases, the corresponding values of "population R squared", defined as $R^2 = \mathbb{V}(\mu_Y(X_i))/\mathbb{V}(Y)$, are within the range typically encountered in empirical studies.

We start by considering the case $\mu_Y(x) = x^2$, for which both a second and a fourth order polynomial obviously constitute a correct specification, and thus $\widehat{B}_{Y,\text{ROT1}} \xrightarrow{p} B_Y = 2$ and $\widehat{B}_{Y,\text{ROT2}} \xrightarrow{p} 2B_Y = 4$ as $n \to \infty$. While one might therefore expect ROT1 to perform better than ROT2 rule in this setup, our results, summarized in the left panel of Figure 1, show that this is not the case. The distribution of ROT1's results depends strongly on the error variance, and tends to produce vast over-estimates of B_Y. For $\sigma^2 = 1$, for example, the average across simulation runs is 33.58, which exceeds the true bound by a factor of almost 17. ROT1's results are also quite volatile. ROT2, on the other hand, is much less affected by changes in the error variance: its mean across simulation runs increases from 4.01 for $\sigma^2 = 0.1$ to only 4.74 for $\sigma^2 = 1$, and its sampling variability is rather small.

Now consider the case $\mu_Y(x) = x^2 - x^4$. We have that $\widehat{B}_{Y,\text{ROT1}} \xrightarrow{p} 10 = B_Y$ and $\widehat{B}_{Y,\text{ROT2}} \xrightarrow{p} 2.753 \neq B_Y$ as $n \to \infty$, which means that ROT1 consistently estimates B_Y here, while the probability limit of ROT2 is about four times smaller than the true smoothness bound. Our simulation results for this setup are summarized in the right panel of Figure 1. Again, ROT1 estimates are highly variable, and tend to be much larger than the true smoothness bound. The discrepancy is not as pronounced as in the previous setup though: for $\sigma^2 = 1$, for example, the average across simulation runs is 36.86, which is only 3.6 times larger than B_Y. ROT2 is again much less affected by changes in the error variance: its mean across simulation runs increases from 2.78 for $\sigma^2 = 0.1$ to only 3.99 for $\sigma^2 = 1$, and its sampling variability is rather small. But due to the severe misspecification of a second-order polynomial, these values tend to severely under-estimate the true smoothness bounds.

These results first of all stress the theoretical point that no data-driven method for

choosing smoothness bounds can be expected to work well under all circumstances. Still, our exercise conveys some insight regarding under which condition one rule might be a better "first guess" than the other. Roughly speaking, the performance patterns of ROT1 can be explained by the fact that its underlying fourth order polynomial specification tends to produce erratic over-fits if the function $\mu_Y(x)$ is rather "simple", and there is a non-negligible level of noise in the data. This is much less of an issue with a quadratic model. In practice, we therefore recommend using ROT2 over ROT1 in settings where one believes that μ_Y is "close" to being a "moderately" convex or concave function. If this is not the shape one has in mind there is no obvious ordering of the ROTs, and both should be considered within a more extensive sensitivity analysis.

D. EXTENSION TO FUZZY REGRESSION KINK DESIGNS

D.1. **Description.** Our approach to FRD inference described in the main body of the paper can easily be extended to the cases in which the parameter of interest is the ratio of jumps in the derivatives (of some order $v \geq 0$) of two conditional expectation functions $\mu_Y(x) = \mathbb{E}(Y|X = x)$ and $\mu_T(x) = \mathbb{E}(T|X = x)$ at the threshold value zero.[1] The most prominent example of such a setup is the Fuzzy Regression Kink Designs (Card et al., 2015), where the goal is to estimate the ratio of jumps in the first derivatives of these functions. We now sketch our extension using notation analogous to that in Section 3.

For a generic random variable W_i, we write $\mu_W^{(v)}(x) = \partial^v \mathbb{E}(W_i|X_i = x)/(\partial x)^v$ for the vth derivative of its conditional expectation given X_i; $\mu_{W,+}^{(v)} = \lim_{x \downarrow 0} \mu_W^{(v)}(x)$ and $\mu_{W,-}^{(v)} = \lim_{x \uparrow 0} \mu_W^{(v)}(x)$ denote the left and right limits of the derivatives at the threshold; and $\tau_{W,v} = \mu_{W,+}^{(v)} - \mu_{W,-}^{(v)}$ denotes the corresponding jump in $\mu_W^{(v)}$. Our parameter of interest is $\theta_v = \tau_{Y,v}/\tau_{T,v}$, and the goal is again to construct CSs $\mathcal{C}^\alpha \subset \mathbb{R}$ with correct asymptotic coverage, uniformly in (μ_Y, μ_T) over some function class $\mathcal{F}$:

$$\liminf_{n \to \infty} \inf_{(\mu_Y, \mu_T) \in \mathcal{F}} \mathbb{P}(\theta_v \in \mathcal{C}^\alpha) \geq 1 - \alpha \tag{D.1}$$

for some $\alpha > 0$. We again define $\mathcal{F}$ as a smoothness class. Specifically, let

$$\mathcal{F}_{H,p}(B) = \{f_1(x)\mathbf{1}\{x \geq 0\} - f_0(x)\mathbf{1}\{x < 0\} : \|f_w^{(p+1)}\|_\infty \leq B, w = 0, 1\}$$

be the Hölder-type class of real functions that are potentially discontinuous at zero, $(p+1)$-times differentiable almost everywhere on either side of the threshold, and whose $(p+1)$th

[1] We could in principle allow the two derivatives to be of of different order, but as we are not aware of a setup that requires this we only consider identical orders here to keep the notation simple.

derivative is uniformly bounded by some constant $B > 0$. We also define the class

$$\mathcal{F}^{\delta}_{H,vp}(B) = \{f \in \mathcal{F}_{H,p}(B) : |f^{(v)}_{+} - f^{(v)}_{-}| > \delta\},$$

and assume that $(\mu_T, \mu_Y) \in \mathcal{F}^{0}_{H,vp}(B_T) \times \mathcal{F}_{H,p}(B_Y) \equiv \mathcal{F}$. Our CSs for the ratio of jumps in vth-order derivatives are based on pth order local polynomial regression, where $v \leq p$. Following standard results on the bias properties of local polynomial regression (Fan and Gijbels, 1996), it is generally recommended to use $p = v + 1$. For a generic dependent variable W_i, the local pth order polynomial estimator $\widehat{\tau}_{W,vp}(h)$ of $\tau_{W,v}$ is the $(p + v + 2)$th component of

$$\underset{\beta \in \mathbb{R}^{2p}}{\operatorname{argmin}} \sum_{i=1}^{n} K(X_i/h)(W_i - \beta^{\top}(1, X_i, X_i^2/2, \ldots, X_i^p/(p!), Z_i, Z_i X_i, \ldots, Z_i X_i^p/(p!))^2,$$

where $K(\cdot)$ is a kernel function with support $[-1, 1]$ and $h > 0$ is a bandwidth. It follows from standard least squares algebra that this estimator can be written as

$$\begin{aligned}
&\widehat{\tau}_{W,vp}(h) = \sum_{i=1}^{n} w_{vp,i}(h) W_i, \quad w_{vp,i}(h) = w_{vp,i,+}(h) - w_{vp,i,-}(h), \\
&w_{vp,i,+}(h) = e_{v+1}^{\top} Q_{p,+}^{-1} \widetilde{X}_{p,i} K(X_i/h) \mathbf{1}\{X_i \geq 0\}, \quad Q_{p,+} = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_{p,i} \widetilde{X}_{p,i}^{\top} \mathbf{1}\{X_i \geq 0\}, \\
&w_{vp,i,-}(h) = e_{v+1}^{\top} Q_{p,-}^{-1} \widetilde{X}_{p,i} K(X_i/h) \mathbf{1}\{X_i < 0\}, \quad Q_{p,-} = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_{p,i} \widetilde{X}_{p,i}^{\top} \mathbf{1}\{X_i < 0\},
\end{aligned}$$

with $\widetilde{X}_{p,i} = (1, X_i, X_i^2/2, \ldots, X_i^p/(p!))^{\top}$. We then obtain a bias-aware AR CS for θ_v by collecting those values of c for which an auxiliary bias-aware CI for $\tau_{M,v}(c) = \tau_{Y,v} - c\tau_{T,v}$ contains zero. To describe the construction, denote the conditional bias and standard deviation of $\widehat{\tau}_{M,vp}(h, c) = \sum_{i=1}^{n} w_{vp,i}(h) M_i(c)$ given $\mathcal{X}_n = (X_1, \ldots, X_n)'$ by $b_{M,vp}(h, c) = \mathbb{E}(\widehat{\tau}_{M,vp}(h, c)|\mathcal{X}_n) - \tau_{M,vp}(c)$ and $s_{M,vp}(h, c) = \mathbb{V}(\widehat{\tau}_{M,vp}(h, c)|\mathcal{X}_n)^{1/2}$, respectively. These quantities can be written more explicitly as

$$\begin{aligned}
b_{M,vp}(h, c) &= \sum_{i=1}^{n} w_{vp,i}(h) \mu_M(X_i, c) - (\mu^{(v)}_{M+}(c) - \mu^{(v)}_{M-}(c)), \\
s_{M,vp}(h, c) &= \left(\sum_{i=1}^{n} w_{vp,i}(h)^2 \sigma^2_{M,i}(c) \right)^{1/2},
\end{aligned}$$

with $\sigma^2_{M,i}(c) = \mathbb{V}(M_i(c)|X_i)$ the conditional variance of $M_i(c)$ given X_i. The bias depends on

(μ_Y, μ_T) through the transformation $\mu_M^{(v)} = \mu_Y^{(v)} - c \cdot \mu_T^{(v)}$ only, and $\mu_Y^{(v)} - c\mu_T^{(v)} \in \mathcal{F}_{H,vp}(B_Y + |c|B_T)$. Our main contribution is to show that one can bound $b_{M,vp}(h,c)$ in absolute value over the functions contained in $\mathcal{F}$ by

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |b_{M,vp}(h,c)| \leq \bar{b}_{M,vp}(h,c) \equiv (-1)^{p-v} \frac{B_Y + |c|B_T}{(p+1)!} \sum_{i=1}^{n} w_{vp,i}(h) X_i^{p+1} \text{sign}(X_i)^{v+1}, \tag{D.2}$$

assuming only that h is such that positive kernel weights are assigned to at least $(p+1)$ data points on either side of the threshold. An infeasible bias-aware AR CS for our parameter of interest θ_v is then given by

$$\mathcal{C}_{vp}^{\alpha} = \{c : |\widehat{\tau}_{M,vp}(h_{M,vp}(c),c)| \leq \text{cv}_{1-\alpha}(r_{M,vp}(h_{M,vp}(c),c)) s_{M,vp}(h_{M,vp}(c),c)\},$$

where $h_{M,vp}(c) = \text{argmin}_h \, \text{cv}_{1-\alpha}(r_{M,vp}(h,c)) s_{M,vp}(h,c)$ is again the efficiency-maximizing bandwidth and $r_{M,vp}(h,c) = \bar{b}_{M,vp}(h,c)/s_{M,vp}(h,c)$ the "worst case" bias to standard deviation ratio. We can then establish the following result.

Theorem D.1. *Suppose that Assumptions 1 and either LL1 or LL2 hold. Then $\mathcal{C}_{vp}^{\alpha}$ is honest with respect to $\mathcal{F}$ in the sense of* (D.1).

It is also straightforward to obtain an analogous result for a feasible version of $\mathcal{C}_{vp}^{\alpha}$ that uses a valid standard error and an estimate of the optimal bandwidth, under appropriate regularity conditions.

D.2. **Proof of Theorem D.1.** The result follows from the same type of arguments as those used in the proof of Theorem 1 for the FRD case. The only step that requires particular attention is establishing the validity of the general bias bound in (D.2), as Armstrong and Kolesár (2020, Theorem B.3) give an explicit expression for the special case $p = 1$ and $v = 0$ only. We first establish a preliminary lemma. Let $\chi = \{x_0, x_1, \ldots, x_k\}$, with $0 \leq x_0 \leq x_1 \leq \ldots \leq x_k < h$ and $k \geq p$, be a generic set of at least $p+1$ constants from the interval $[0,h)$, write $\chi_{-i} = \chi \setminus \{x_i\}$ for the subset of χ that excludes its ith element, and define

$$\widehat{\beta}_{vp}(t,\chi) = \sum_{i=0}^{k} w_{vp,i,+}(h,\chi) \mathbf{1}\{x_i \geq t\}(x_i - t)^p,$$

where $w_{vp,i,+}(h,\chi)$ are local polynomial regression weights analogous to those defined above, but with χ taking the role of the data $\mathcal{X}_n$.[2] Put differently, the term $\widehat{\beta}_{vp}(t,\chi)$ is the $(v+1)$th

[2]A similar argument applies for the case that $-h < x_k \leq \cdots \leq x_1 \leq 0$.

coefficient in a weighted least squares regression of $\mathbf{1}\{x_i \geq t\}(x_i - t)^p$ on $(1, x_i, x_i^2, \ldots, x_i^p)^\top$. This term is well-defined as long as χ contains at least $p+1$ distinct elements.

Lemma D.1. *Suppose that either (i) χ has $(p+1)$ elements, all which are distinct; or (ii) χ has at least $(p+2)$ distinct elements, and $\widehat{\beta}_{vp}(t, \chi_{-i})$ satisfies* (D.3) *for all $i = 1, \ldots, |\chi|$. Then it holds for all $t \in \mathbb{R}$ that*

$$\widehat{\beta}_{vp}(t, \chi) \leq 0 \quad \textit{if } p - v \textit{ odd and } \quad \widehat{\beta}_{vp}(t, \chi) \geq 0 \quad \textit{if } p - v \textit{ even.} \tag{D.3}$$

To then establish the bias bound (D.2), note that the bias can be written as

$$\begin{aligned} b_{M,vp}(h, c) = &\left(\sum_{i: X_i \geq 0} w_{vp,i,+}(h) \mu_M(X_i, c) - \mu_{M+}^{(v)}(c) \right) \\ &- \left(\sum_{i: X_i < 0} w_{vp,i,-}(h) \mu_M(X_i, c) - \mu_{M-}^{(v)}(c) \right) \equiv T_1 + T_2. \end{aligned}$$

Since $\sum_{i: X_i \geq 0} w_{vp,i,+}(h) X_i^v = 1$ and $\sum_{i: X_i \geq 0} w_{vp,i,+}(h) X_i^j = 0$ for $j \neq v$ and $j \leq p$ by standard least squares algebra, it follows that

$$\begin{aligned} T_1 &= \sum_{i: X_i \geq 0} w_{vp,i,+}(h) \left(\sum_{j=0}^{p} \frac{1}{j!} X_i^j \mu_M^{(j)}(0, c) + \frac{1}{p!} \int_0^{X_i} \mu^{(p+1)}(X_i, c)(X_i - t)^j dt \right) - \mu_{M+}^{(v)}(c) \\ &= \frac{1}{p!} \int_0^\infty \mu_M^{(p+1)}(t, c) \widehat{\beta}_{vp}(t, \mathcal{X}_n^+), \end{aligned}$$

where $\mathcal{X}_n^+ = \{X_i \in \mathcal{X}_n : 0 \leq X_i \leq h\}$. This expression is clearly maximized in absolute value by any function $\mu_M(t, c)$ whose $(p+1)$th derivative is given by $\mu_M^{(p+1)}(t, c) = B_M \text{sign}(\widehat{\beta}_{vp}(t, \mathcal{X}_n^+))$ for $t \geq 0$.

We now construct a collection $\mathcal{X}_{n,k}^+$ of subsets of $\mathcal{X}_n^+$, with $k = p+1, \ldots, n$, as follows. Let $\mathcal{X}_{n,p+1}^+$ be an arbitrary subset of $p+1$ distinct elements of $\mathcal{X}_n^+$ (such a subset exists by assumption), and let $\mathcal{X}_{n,k}^+$, for $k > p+1$, be the union of $\mathcal{X}_{n,k-1}^+$ and an arbitrary element of $\mathcal{X}_n^+ \setminus \mathcal{X}_{n,k-1}^+$. Then Lemma D.1 implies that $\widehat{\beta}_{vp}(t, \mathcal{X}_{n,k}^+)$ satisfies (D.3) for any $k = p+1, \ldots, n$. Since $\mathcal{X}_{n,n}^+ = \mathcal{X}_n^+$, this means that $\text{sign}(\widehat{\beta}_{vp}(t, \mathcal{X}_n^+)) = (-1)^{p-v}$ for all t. The term T_1 is thus maximized in absolute value for any function μ_M such that $\mu_M(t, c) = (-1)^{p-v} B_M t^{p+1} \text{sign}(t)/((p+1)!)$ for $t \geq 0$. A similar reasoning implies that T_2 is maximized for any function μ_M such that $\mu_M(t, c) = (-1)^{p-v} B_M t^{p+1} \text{sign}(t)/((p+1)!)$ for $t < 0$. Together, these statements prove (D.2). □

D.3. **Proof of Lemma D.1.** To prove part (i), we denote the unique polynomial of order p that interpolates the points $\{(x, \mathbf{1}\{x \geq t\}(x-t)^p)\}_{x\in\chi}$ by $P(x, \chi_k)$. Our proof comes down to determining the sign of its coefficients. To do so, let $S(k) = k + |\{x \in \chi : x \leq t\}|$ be the sum of k and the number of elements of χ whose value does not exceed t, and consider subsets of χ of the form $\chi_k = \{x_i \in \chi : x_i \leq t\} \cup \{x_i \in \chi : S(1) \leq i \leq S(k)\}$ that contain those elements of χ whose value does not exceed t, and the k next largest ones. That is, $\chi_0 = \{x_i \in \chi : x_i \leq t\}$, and χ_1 is the union of χ_0 and the smallest element of χ that is larger than t, etc. We also note that if χ is such that $S(0) = 0$, then $\widehat{\beta}_{vp}(t, \chi) = (-1)^{p-v}\binom{p}{v}t^v$ clearly satisfies (D.3). It therefore suffices to restrict attention to sets χ such that $S(0) > 0$. It is also easy to see that $\widehat{\beta}_{vS(0)}(t, \chi_0) = 0$, and hence satisfies (D.3). It thus remains to show that if $\widehat{\beta}_{vS(k)}(t, \chi_k)$ satisfies (D.3), so does $\widehat{\beta}_{vS(k+1)}(t, \chi_{k+1})$. The statement of the lemma then follows by induction.

To show the last step, assume that $\widehat{\beta}_{vS(k)}(t, \chi_k)$ satisfies (D.3), and write the polynomial that interpolates the points $\{x, \mathbf{1}\{x \geq t\}(x-t)^{S(k+1)}\}_{x\in\chi_{k+1}}$ as

$$P(x, \chi_{k+1})x^v = (x-t)P(x, \chi_k) + \bar{\iota}_{p+1}\prod_{x_l\in\chi_k}(x - x_l), \text{ where} \tag{D.4}$$
$$\bar{\iota}_{k+1} = (x_{S(k+1)} - t)\left((x_{S(k+1)} - t)^{S(k)} - P(x_{S(k+1)}, \chi_k)\right)\prod_{x_l\in\chi_k}\frac{1}{x_{S(k+1)} - x_l}.$$

We can then express the $\widehat{\beta}_{vS(k+1)}(t, \chi_{k+1})$ in terms of the $\widehat{\beta}_{vS(k)}(t, \chi_k)$ by comparing the appropriate terms on both sides of equation (D.4). This yields that

$$\widehat{\beta}_{vS(k+1)}(t, \chi_{k+1}) = \begin{cases} \widehat{\beta}_{S(k)S(k)}(t, \chi_k) + \bar{\iota}_{k+1} & \text{if } v = S(k+1), \\ -t\widehat{\beta}_{0S(k)}(t, \chi_k) + (-1)^{S(k+1)}\bar{\iota}_{k+1}\prod_{0\leq j\leq S(k)} x_j & \text{if } v = 0, \\ \widehat{\beta}_{(v-1)S(k)}(t, \chi_k) - t\widehat{\beta}_{vS(k)}(t, \chi_k) + (-1)^{S(k+1)-v}\bar{\iota}_{k+1}\sum_{M\in\mathcal{M}_{S(k+1)-v}}\prod_{m_s\in M} x_{m_s} & \text{else.} \end{cases}$$

where $\mathcal{M}_v$ is the set of all subsets $M = \{m_1, \ldots, m_v\}$ of $\{1, \ldots, S(k+1)\}$ that contain exactly v elements. Careful inspection of the last display shows that $\widehat{\beta}_{vS(k+1)}(t, \chi_{k+1})$ satisfies (D.3) if $\bar{\iota}_{k+1} \geq 0$. We proof this claim by a simple argument about the number of zeros of polynomials. Let $\chi_{k\setminus 0} = \chi_k \backslash x_0$. We note that $\bar{\iota}_{k+1} \geq 0$ if

$$P(x, \chi_k) < P(x, \chi_{k\setminus 0} \cup x) \text{ for all } x > x_{S(k)}. \tag{D.5}$$

To show (D.5), we fix some arbitrary $x_l > x_{S(k)}$ and consider the two different polynomials $P(x, \chi_k)$ and $P(x, \chi_{k\setminus 0} \cup x_l)$. As these polynomials are of degree $S(k)$ and they intersect $S(k)-$times at all $x \in \chi_{k\setminus 0}$, they do not intersect for any $x \notin \chi_{k\setminus 0}$.

As the set χ_k was arbitrarily chosen, we note that by the induction argument the intercept of both polynomials has the same sign such that

$$\text{sign}(P(0, \chi_k)) = \text{sign}(P(0, \chi_{k\setminus 0} \cup x_l)). \tag{D.6}$$

Using (D.6) together with standard arguments of polynomials and their sign as $x \to \pm\infty$, (D.5) is satisfied if $|P(0, \chi_k)| \leq |P(0, \chi_{k\setminus 0} \cup x_l)|$. Polynomials of order $S(k)$, that are different from $(x-t)^{S(k)}$, can have at most $(S(k)+1)$ intersections with the function $g(x) = \mathbf{1}\{x \geq t\}(x-t)^{S(k)}$ for $t > 0$. This reasoning implies that the polynomial $P(x, \chi_{k\setminus 0} \cup x_l)$ does not have any intersections with the function $g(x)$ for $x \leq x_0$, and in particular it does not have any root for $x \leq x_0$, so that it has the same sign for all $0 \leq x \leq x_0$. As $P(x_0, \chi_k) = 0$, we can conclude that $|P(x, \chi_k)| \leq |P(x, \chi_{k\setminus 0} \cup x_l)|$ for any $x \leq x_0$. This completes our proof of part (i).

To prove part (ii) of the lemma, note that it follows from textbook arguments that

$$\widehat{\beta}_{vp}(t, \chi) = \widehat{\beta}_{vp}(t, \chi_{-i}) + (1 - l_i)^{-1} w_{vp,i,+}(h, \chi)\widehat{\epsilon}_i,$$

where $\widehat{\epsilon}_i = \mathbf{1}\{x_i \geq t\}(x_i - t)^p - \sum_{v=0}^{p} \widehat{\beta}_{vp}(t, \chi) x_i^v$ is the ith regression residual and $l_i = \sum_{j=0}^{p} w_{jp,i}(\chi) x_i^j$ is the leverage of the ith observation. We first consider the case that $\widehat{\beta}_{vp}(t, \chi_{-i}) \leq 0$ for all i, which implies that $\widehat{\beta}_{vp}(t, \chi) \leq (1 - l_i)^{-1} w_{vp,i,+}(h, \chi)\widehat{\epsilon}_i$. Since $\sum_{i=1}^{|\chi|} w_{vp,i,+}(h, \chi)\widehat{\epsilon}_i = 0$ and $0 \leq l_i < 1$ for all i by basic least squares algebra, we know that $(1 - l_i)^{-1} w_{vp,i,+}(h, \chi)\widehat{\epsilon}_i \leq 0$, for at least some i, which in turn means that $\widehat{\beta}_{vp}(t, \chi) \leq 0$. The same kind of argument applies to the case that $\widehat{\beta}_{vp}(t, \chi_{-i}) \geq 0$ for all i. □